NOTES ON PROBABILITY AND STATISTICS (FOLLOWING WASSERMAN'S ALL OF STATISTICS)

ANTOINE REMOND-TIEDREZ

Disclaimer: These notes follow Larry Wasserman's *All of Statistics* [Was10]

Contents

Date: April 8, 2025.

1. Probability

1.1. Probability.

Definition 1.1 (Probability measure)**.** Let Ω be a set which we refer to as a *sample space* and whose elements $\omega \in \Omega$ we refer to as *outcomes*. Let $\mathcal{A}$ be a collection of subsets of Ω which satisfies the following properties.

1. $\emptyset \in \mathcal{A}$,
2. Closure under countable unions: if $A_1, A_2, \dots$ is a countable sequence of events in $\mathcal{A}$ then $\bigcup_{i=1}^{\infty} A_i$ also belongs to $\mathcal{A}$, and
3. Closure under complementation: if $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$.

We call $\mathcal{A}$ a *σ–algebra* and refer to the elements of $\mathcal{A}$ (which are subsets of the sample space) as *events*. A map $\mathbb{P} : \mathcal{A} \to [0, 1]$ is called a *probability measure* if it satisfies the following properties.

1. Non-negativity: $\mathbb{P}(A) \geqslant 0$ for any event A,
2. $\mathbb{P}(\Omega) = 1$, and
3. Countable additivity: if $A_1, A_2, \dots$ is a countable sequence of *pairwise disjoint* events in $\mathcal{A}$ then
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

1.2. Independent events.

Definition 1.2 (Independent events)**.** Two events A and B are *independent* if
$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B).$$
Note that here, and throughout in the sequel, we write $AB := A \cap B$.

Similarly, a set of events $(A_i)_{i \in I}$ is *independent* if
$$\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j)$$
for *any* finite subset J of the index set I.

1.3. Conditional probability.

Definition 1.3 (Conditional probability)**.** Let A and B be events. If $\mathbb{P}(B) > 0$ then we defined the *conditional probability of A given B* to be
$$\mathbb{P}(A|B) := \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}.$$

Remark 1.4 (Interpretation of conditional probability)**.** We can think of the conditional probability $\mathbb{P}(A|B)$ as the *fraction* of times A occurs among the times when B occurs.

Remark 1.5 (Conditional probability)**.** The following hold.

- For any event B with $\mathbb{P}(B) > 0$, $\mathbb{P}(\,\cdot\,|B)$ is itself a probability measure.
- Two events A and B with $\mathbb{P}(B) > 0$ are independent if and only if
$$\mathbb{P}(A|B) = \mathbb{P}(A),$$
i.e. knowing that B occurred does not give us any additional information about the likelihood of A occurring or not.

1.4. **Bayes' Theorem.**

Theorem 1.6 (Bayes)**.** *Let $A_1, \ldots, A_n$ be a partition of a sample space Ω such that $\mathbb{P}(A_i) > 0$ for every i. If B is an event such that $\mathbb{P}(B) > 0$ then, for each i,*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

Remark 1.7 (Bayes' Theorem and classification)**.** We call $\mathbb{P}(A_i)$ the *prior probability of A_i* and $\mathbb{P}(A_i|B)$ the *posterior probability of A_i*, meaning that the latter is an updated version of the former using the new information provided by the occurrence of B. This is why Bayes' Theorem forms the basis for classification algorithms.

2. Random variables

2.1. Introduction.

Definition 2.1 (Random variable)**.** Let Ω be a sample space. A *random variable is a map* $X : \Omega \to \mathbb{R}$ which assigns a real number $X(\omega)$ to every outcome ω.

2.2. Distribution functions and probability functions.

Definition 2.2 (Cumulative distribution function)**.** Let X be a random variable. The *cumulative distribution function*, or *CDF*, of X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = \mathbb{P}(X \leqslant x).$$

Definition 2.3 (Discrete random variable and probability mass function)**.** A random variable X is called *discrete* if it takes countably many values. The *probability mass function* of X is the function $f_X : \mathbb{R} \to [0, 1]$ defined by

$$f_X(x) = \mathbb{P}(X = x).$$

Remark 2.4 (Relation between the cumulative distribution function and the probability mass function)**.** Let X be a discrete random variable. Its CDF and its probability mass function are related via

$$F_X(x) = \sum_{x_i \leqslant x} f_X(x_i).$$

Definition 2.5 (Continuous random variable and probability density function)**.** A random variable X is called *continuous* if there exists a function $f_X : \mathbb{R} \to \mathbb{R}$, satisfying $f_X \geqslant 0$ and $\int_{-\infty}^{\infty} f_X = 1$, such that

$$\mathbb{P}(a < X < b) = \int_a^b f_X(y)dy$$

for every real numbers $a \leqslant b$. The function f_X is called the *probability density function*, or *PDF*, of X.

The PDF is related to the CDF via

$$F_X(x) = \int_{-\infty}^x f_X(y)dy$$

and $f_X = F'_X$ where F_X is differentiable.

Definition 2.6 (Quantile function)**.** Let X be a random variable with CDF F. The *inverse CDF*, or *quantile function*, of X is defined by

$$F^{-1}(q) = \inf\{x : F(x) > q\}$$

for all $q \in [0, 1]$, i.e. $F^{-1} : [0, 1] \to \mathbb{R}$. If F is strictly increasing and continuous then $F^{-1}(q)$ is the unique real number x such that $F(x) = q$ (i.e. in that case it is *actually* an inverse of F).

We call $F^{-1}(1/4)$ the *first quartile*, $F^{-1}(1/2)$ the *median*, or *second quartile* , and $F^{-1}(3/4)$ the *third quartile.*

Definition 2.7 (Equality in distribution)**.** Two random variables X and Y are said to be *equal in distribution* if their CDFs agree, i.e. if $F_X(x) = F_Y(x)$ for every $x \in \mathbb{R}$.o

Remark 2.8. Two random variables which are equal in distribution need not be equal. For example, consider a discrete random variable X whose probability mass function is symmetric about the origin and vanishes at zero. The random variable $Y := -X$ is then equal in distribution to X and yet $\mathbb{P}(X = Y) = 0$.

2.3. **Some important discrete random variables.**

Definition 2.9 (Point mass distribution)**.** A random variable X has a *point mass distribution at* a, written $X \sim \delta_a$, if $\mathbb{P}(X = 1) = 1$. In this case the CDF is

$$F(x) = \begin{cases} 0 & \text{if } x < a \text{ and} \\ 1 & \text{if } x \geqslant a \end{cases}$$

while the probability mass function is $f(x) = 1$ for $x = a$ and vanishes elsewhere.

Definition 2.10 (Discrete uniform distribution)**.** Let $k > 1$ be an integer. A random variable X has a *uniform distribution on* $\{1, \ldots, k\}$ if its probability mass function is given by

$$f(x) = \begin{cases} 1/k & \text{for } x = 1, \ldots, k \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Definition 2.11 (Bernoulli distribution)**.** Let $p \in [0, 1]$. We say that a random variable X has a *Bernoulli distribution*, written $X \sim \text{Bernoulli}(p)$, if

$$\mathbb{P}(X = 1) = p \text{ and } \mathbb{P}(X = 0) = 1 - p,$$

which corresponds to flipping a single (possibly biased) coin. The corresponding probability mass function is $f(x) = p^x(1-p)^{1-x}$ for $x = 0$ or 1.

Definition 2.12 (Binomial distribution)**.** Let $p \in [0, 1]$ and let n be a positive integer. We say that a random variable X has a *binomial distribution*, written $X \sim \text{Binomial}(n, p)$, if X counts the number of heads obtained by flipping n times a coin which falls heads up with probability p. In other words its probability mass function is given by

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, \ldots, n \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Definition 2.13 (Geometric distribution)**.** Let $p \in [0, 1]$. A random variable X has a *geometric distribution*, written $X \sim \text{Geom}(p)$, if

$$\mathbb{P}(X = k) = p(1-p)^{k-1}$$

for every integer $k \geqslant 1$. Such random variables can be thought of as the number of flips needed until a (possibly biased) coin shows heads for the first time.

Definition 2.14 (Poisson distribution)**.** Let $\lambda > 0$. A random variable X has a *Poisson distribution*, written $X \sim \text{Poisson}(\lambda)$, if its probability mass function is given by

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

for every non-negative integer x. This distribution describes the number of events occurring in a fixed time interval if these events occur with a known constant mean rate and independently of the time since the last event.

2.4. **Some important continuous random variables.**

Definition 2.15 (Uniform distribution)**.** Let $a < b$ be real numbers. A random variable X has a *uniform distribution*, written $X \sim \text{Unif}(a, b)$, if its PDF is given by

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a < x < b \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding CDF is

$$F(x) = \begin{cases} 0 & \text{if } x < a, \\ \dfrac{x-a}{b-a} & \text{if } a \leqslant x \leqslant b, \text{ and} \\ 1 & \text{if } x > b. \end{cases}$$

Definition 2.16 (Normal, or Gaussian, distribution)**.** Let $\mu \in \mathbb{R}$ and let $\sigma^2 > 0$. A random variable X has a *Normal*, or *Gaussian*, *distribution*, written $X \sim N(\mu, \sigma^2)$, if its PDF is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

for every $x \in \mathbb{R}$. The parameter μ is known as the *mean* and σ is known as the *standard deviation* (we will actually define both of these terms for general random variables later). We say that X has a *standard Normal distribution* if $\mu = 0$ and $\sigma = 1$.

Definition 2.17 (Exponential distribution)**.** Let $\beta > 0$. A random variable X has an *Exponential distribution*, written $X \sim \text{Exp}(\beta)$, if its PDF is given by

$$f(x) = \frac{1}{\beta} e^{-x/\beta}$$

for every $x > 0$. The exponential distribution is used to model waiting times between rare events (which are formalised using the theory of Poisson point processes).

Definition 2.18 (Gamma distribution)**.** Let $\alpha, \beta > 0$. A random variable X has a *Gamma distribution*, written $X \sim \text{Gamma}(\alpha, \beta)$, if its PDF is given by

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$$

for every $x > 0$, where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$ denotes the *Gamma function* (which is related to the factorial via $\Gamma(n) = (n-1)!$ for every positive integer n). Note that the $\text{Gamma}(1, \beta)$ distribution is exactly the same thing as the $\text{Exp}(\beta)$ distribution.

Definition 2.19 (Beta distribution)**.** Let $\alpha, \beta > 0$. A random variable X has a *Beta distribution*, written $X \sim \text{Beta}(\alpha, \beta)$, if its PDF is given by

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

for every $0 < x < 1$. Note that the PDF of the Beta distribution is similar to the PDFs of the Bernoulli, binomial, and geometric distributions. This algebraic similarity will be essential later when seeking closed-form expressions for certain inference methods.

Definition 2.20 (t and Cauchy distributions)**.** Let $\nu > 0$. A random variable X has a t *distribution with* ν *degrees of freedom*, written $X \sim t_\nu$, if its PDF is given by

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{(\nu+1)/2}}.$$

The t distribution is similar to the Normal distribution, but the t distribution has thicker tails (the Normal distribution actually corresponds to $\nu = \infty$).

A t distribution with $\nu = 1$ is called a *Cauchy distribution* and its corresponding PDF is

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Definition 2.21 (χ^2 distribution)**.** Let p be a positive integer. A random variable X has a χ^2 *distribution with* p *degrees of freedom*, written $X \sim \chi^2_p$, if its PDF is given by

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}$$

for every $x > 0$. Note that a χ^2 distribution is a special case of a gamma distribution.

2.5. **Bivariate distributions.**

Definition 2.22 (Joint mass distribution)**.** Let X and Y be discrete random variables. Their *joint mass distribution* is defined to be

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$$

where it is understood that both $X = x$ *and* $Y = y$ are asked to occur.

Definition 2.23 (Joint probability density function)**.** Let X and Y be continuous random variables. We say that f is a *joint probability density function*, or *joint PDF*, for the random variables (X, Y), if

(1) $f(x, y) \geqslant 0$ for all (x, y),
(2) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$, and
(3) for any (Borel) set $A \subseteq \mathbb{R} \times \mathbb{R}$, $\mathbb{P}((X, Y) \in A) = \int\int_A f(x, y) dx dy$.

Definition 2.24 (Joint cumulative distribution function)**.** Let X and Y be random variables which may be both discrete or both continuous. In both cases we define the *joint cumulative distribution function*, of *joint CDF*, as

$$F_{X,Y}(x, y) = \mathbb{P}(X \leqslant x, Y \leqslant y).$$

2.6. **Marginal distributions.**

Definition 2.25 (Marginal mass function)**.** Let X and Y be discrete random variables with joint mass distribution $f_{X,Y}$. The *marginal mass function* for X is defined by

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f_{X,Y}(x, y),$$

and similarly the marginal mass function for Y is defined by $f_Y(y) = \sum_x f_{X,Y}(x, y)$.

Definition 2.26 (Marginal probability density function)**.** Let X and Y be continuous random variables with joint probability density function $f_{X,Y}$. The *marginal probability density function*, or *marginal PDF*, for X and Y are defined respectively by

$$f_X(x) = \int f_{X,Y}(x, y)dy \text{ and } f_Y(y) = \int f_{X,Y}(x, y)dx.$$

Definition 2.27 (Marginal cumulative distribution function)**.** Let X and Y be random variables which may be both discrete or both continuous. Suppose they have marginal mass functions, or marginal PDFs, given by f_X and f_Y respectively. In both cases we define the *marginal cumulative distribution functions*, of *marginal CDFs*, F_X and F_Y to be the CDFs corresponding to the marginal PDFs f_X and f_Y, respectively.

2.7. **Independent random variables.**

Definition 2.28 (Independent random variables)**.** Two random variables X and Y are *independent* if, for every $A, B \subseteq \mathbb{R}$,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

Otherwise we say that they are *dependent*.

2.8. **Conditional distributions.**

Definition 2.29 (Conditional probability mass function)**.** Let X and Y de discrete random variables. For any y with $\mathbb{P}(Y = y) > 0$ we have that the *conditional probability mass function of X given Y* is

$$f_{X|Y}(x|y) = \mathbb{P}(X = x \,|\, Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Definition 2.30 (Conditional probability density function)**.** Let X and Y be continuous random variables. For any $y > 0$ we have that the *conditional probability density function*, or *conditional PDF*, *of X given Y* is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

such that, for any $A \subseteq \mathbb{R}$, $\mathbb{P}(X \in A \,|\, Y = y) = \int_A f_{X|Y}(x|y)dx$.

2.9. **Multivariate distributions and IID samples.**

Definition 2.31 (Random vector)**.** Let $X = (X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are random variables. We call X a *random vector*. We often refer to the joint probability mass function, or to the joint PDF , of X as its *probability mass function*, or *PDF* respectively, i.e. omitting the word "joint".

Definition 2.32 (IID samples and random samples)**.** Suppose that $X_1, \ldots, X_n$ are independent random variables such that each has the same marginal CDF F. We say that $X_1, \ldots, X_n$ are *IID* (Independent and Identically Distributed), we write $X_1, \ldots, X_n \sim F$, and we call $X_1, \ldots, X_n$ a *random sample of size n from F*. Moreover, if F has a probability mass function or PDF f we also write $X_1, \ldots, X_n \sim f$.

2.10. Two important multivariate distributions.

Definition 2.33 (Multinomial distribution)**.** Let $p = (p_1, \ldots, p_k)$ satisfy $p_j \geqslant 0$ for all j and $\sum_j p_j = 1$ and let n be a positive integer. We say that a random vector X has a *multinomial distribution*, written $X \sim \text{Multinomial}(n, p)$, if X_j counts the number of times color j appears when drawing n times with replacement from k balls with distinct colors, given that p_j is the probability of drawing a ball of color j. In other words its probability mass function is given by

$$f(x) = \binom{n}{x_1 \ldots x_k} p_1^{x_1} \ldots p_k^{x_k}$$

where the *multinomial coefficient* is defined as

$$\binom{n}{x_1 \ldots x_k} = \frac{n!}{x_1! \ldots x_k!}$$

for any collection of non-negative integers x_j satisfying $\sum_j x_j = n$.

Multinomial distributions are closely related to binomial distributions: the marginal distribution of X_j is a Binomial(n, p_j) distribution and, when $k = 2$, a multinomial distribution is precisely a binomial distribution.

Definition 2.34 (Multivariate Normal distribution)**.** Let $\mu \in \mathbb{R}^k$ and let Σ be a symmetric and positive-definite k-by-k matrix. A random vector X has a *multivariate Normal distribution*, written $X \sim N(\mu, \Sigma)$, if its PDF is given by

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\},$$

where $|A|$ denotes the determinant of a matrix A. The marginal distribution of X_j is $N(\mu_j, \Sigma_{jj})$.

2.11. Transformations of random variables.

Remark 2.35 (How to transform random variables)**.** Given a random variable X and its PDF f_X, if a random variable Y is defined via $Y = r(X)$ we may find its PDF as follows, in three steps.

(1) For each y we find the set $A_y = \{x : r(x) \leqslant y\}$.

(2) We find the CDF F_Y via

$$F_Y(y) = \mathbb{P}(Y \leqslant y) = \mathbb{P}(r(X) \leqslant y) = \mathbb{P}(X \in A_y) = \int_{A_y} f_X(x) dx.$$

(3) Provided that the CDF of Y is sufficiently regular we obtain the PDF by differentiating: $f_Y = F_Y'$.

2.12. Transformation of several random variables.

Remark 2.36 (How to transform several random variables)**.** Given two random variables X and Y with joint PDF $f_{X,Y}$, if a random variable Z is defined via $Z = r(X, Y)$ we may find its PDF as follows, in three steps mimicking those of Remark 2.35.

(1) For each z we find the set $A_z \{(x, y) : r(x, y) \leqslant z\}$.

(2) We find the CDF F_Z via

$$F_Z(z) = \mathbb{P}(Z \leqslant z) = \mathbb{P}(r(X, Y) \leqslant z) = \mathbb{P}((X, Y) \in A_z) = \int\int_{A_z} f_{X,Y}(x, y) dx dy.$$

(3) Provided that the CDF of Z is sufficiently regular we obtain the PDF by differentiating: $f_Z = F_Z'$.

2.13. **Bonus.**

Definition 2.37 (Categorical distribution)**.** Let $p = (p_1, \ldots, p_k)$ satisfy $p_j \geqslant 0$ for all j and $\sum_{j=1}^n p_j = 1$. We say that a random variable X has a *categorical distribution*, written $X \sim \text{Categorical}(p)$, if

$$\mathbb{P}(X = e_j) = p_j$$

where e_j is the j-th coordinate vector in $\mathbb{R}^k$. In other words X is assigned to the j-th of k–many categories with probability p_j.

Remark 2.38 (Categorical, Bernoulli, and Multinomial distributions)**.** The categorical distribution is related to several other important discrete distributions.

- If $X \sim \text{Categorical}(p)$ then $X_j = 1 \sim \text{Bernoulli}(p_j)$. In particular this means that when $k = 2$, a categorical or a Bernoulli distribution are the same thing (up to a bijective relabelling).
- If $X_1, \ldots, X_n \sim \text{Categorical}(p)$ then $\sum_{i=1}^n X_i \sim \text{Multinomial}(n, p)$.

Definition 2.39 (Mode)**.** Let X be a random variable. A *mode* of X is defined to be any

$$m(X) \in \arg\max_x f_X(x)$$

where f_X denotes the probability mass function or PDF of X (depending on whether X is discrete or continuous, respectively).

Definition 2.40 (Dirichlet distribution)**.** Let $k \geqslant 2$ be an integer and let $\alpha \in \mathbb{R}^k_{>0}$. A random vector X in Δ^{k-1} has a *Dirichlet distribution with parameter* α, written $X \sim \text{Dirichlet}(\alpha)$, if its PDF is given by

$$f(x; \alpha) = \frac{1}{B(\alpha)} x_1^{\alpha_1 - 1} \ldots x_k^{\alpha_k - 1}$$

where

$$B(\alpha) := \frac{\prod_{j=1}^k \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}$$

denotes the *Beta function*.

Remark 2.41 (Beta and Dirichlet distributions)**.** The Beta distribution is a special case of the Dirichlet distribution since

$$X \sim \text{Beta}(\alpha, \beta) \iff (X, 1 - X) \sim \text{Dirichlet}(\tilde{\alpha})$$

for $\tilde{\alpha} = (\alpha, \beta)$.

Theorem 2.42 (Conditional distribution of components of a multivariate Normal)**.** *Let $\mu \in \mathbb{R}^k$ and let Σ be a symmetric and positive-definite k-by-k matrix. Let $(Y, Z) \sim N(0, \Sigma)$ be a random vector for which the codomains of Y and Z are $\mathbb{R}^l$ and $\mathbb{R}^{k-l}$, respectively, for some $1 \leqslant l \leqslant k - 1$. Let us write*

$$\mu = (\mu_Y, \mu_Z) = (\mathbb{E}Y, \mathbb{E}Z) \;\; \textit{and}$$
$$\Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix} = \begin{pmatrix} \mathbb{V}Y & \text{Cov}(Y, Z) \\ \text{Cov}(Z, Y) & \mathbb{V}Z \end{pmatrix}.$$

Then

$$Z \,|\, Y = y \sim N\left(\bar{\mu}(y),\, \overline{\Sigma}\right)$$

for

$$\bar{\mu}(y) = \mu_Z + \Sigma_{ZY}\Sigma_{YY}^{-1}\,(y - \mu_Y)$$

and

$$\overline{\Sigma} = \Sigma_{ZZ} - \Sigma_{ZY}\Sigma_{YY}^{-1}\Sigma_{YZ}.$$

In particular if $l = k - l = 1$, i.e. if both Y and Z have codomain $\mathbb{R}$ then

$$\begin{aligned}\bar{\mu}(y) &= \mu_Z + \frac{\operatorname{Cov}(Y,\, Z)}{\sigma_Z^2}(y - \mu_Y) \text{ and}\\ \bar{\sigma}^2 &= \sigma_Z^2 - \frac{\operatorname{Cov}^2(Y,\, Z)}{\sigma_Y^2}.\end{aligned}$$

3. Expectation

3.1. Expectation of a random variable.

Definition 3.1 (Expected value)**.** Let X be a random variable. The *expected value*, or *mean*, or *first moment* of X is defined to be

$$\mathbb{E}(X) = \int x dF(x) = \begin{cases} \sum_x x f_X(x) & \text{if } X \text{ is discrete and} \\ \int x f_X(x) dx & \text{if } X \text{ is continuous,} \end{cases}$$

assuming that the integral (or sum) is well-defined. We write equivalently

$$\mathbb{E}(X) = \mathbb{E}X = \mu = \mu_X.$$

Theorem 3.2 (Rule of the Lazy Statistician)**.** *Let X be a random variable and let $Y = r(X)$. The expected value of Y is given by*

$$\mathbb{E}(Y) = \int r(x) dF(x).$$

Remark 3.3 (Probability is a special case of expectation)**.** For any $A \subseteq \mathbb{R}$ we have that

$$\mathbb{P}(X \in A) = \int_A dF(x) = \int I_A(x) dF(x) = \mathbb{E}\left(I_A(X)\right).$$

In other words: probability is a special case of expectation.

Definition 3.4 (Moments)**.** Let X be a random variable.

- The *k-th moment* of X is defined to be $\mathbb{E}\left(X^k\right)$, provided that $|X|^k$ is integrable.
- The *k-th central moment of X* is defined to be $\mathbb{E}\left((X-\mu)^k\right)$, where μ denotes the expected value of X, once again provided that $\mathbb{E}\left(|X-\mu|^k\right)$ is finite.

3.2. Properties of expectation.

Theorem 3.5 (Properties of expectation)**.** *Let $X_1, \ldots, X_n$ be random variables.*

(1) Expectation is linear, meaning that, for any real numbers $a_1, \ldots, a_n$,

$$\mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i \mathbb{E}(X_i).$$

(2) If moreover the X_i's are independent then

$$\mathbb{E}\left(\prod_i X_i\right) = \prod_i \mathbb{E}(X_i).$$

3.3. Variance and covariance.

Definition 3.6 (Variance and standard deviation)**.** Let X be a random variable with mean μ. The *variance* of X is defined to be

$$\mathbb{V}(X) = \mathbb{E}(X-\mu)^2 = \int (x-\mu)^2 dF(x),$$

assuming that this expectation is well-defined. We write equivalently

$$\mathbb{V}(X) = \mathbb{V}X = \sigma^2 = \sigma_X^2$$

and we call $\mathrm{sd}(X) = \sqrt{\mathbb{V}(X)}$ the *standard deviation* of X, such that $\mathrm{sd}(X) = \sigma = \sigma_X$.

Lemma 3.7 (Computing variances)**.** *Let X be a random variable and let μ be its mean. Then*

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2.$$

Definition 3.8 (Sample mean and sample variance)**.** Let $X_1, \ldots, X_n$ be random variables. We define the *sample mean* to be the random variable

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and we define the *sample variance* to be the random variable

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2.$$

Remark 3.9. It may at first glance seem odd that the pre-factor $\frac{1}{n-1}$ appears in the sample variance, instead of say $\frac{1}{n}$. As detailed in Exercise A.3.8 this pre-factor ensures that the sample variance approximates the variance of X_i in the case where the X_i's are IID; in that case $\mathbb{E}(S_n^2) = \mathbb{V}(X_i)$.

Definition 3.10 (Covariance and correlation)**.** Let X and Y be random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively. The *covariance between X and Y* is defined to be

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left((X - \mu_X)(Y - \mu_Y)\right)$$

and the *correlation between X and Y* is defined to be

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

We write equivalently

$$\rho(X, Y) = \rho_{X,Y} = \rho.$$

Lemma 3.11 (Computing covariances)**.** *Let X and Y be a random variables and let μ_X and μ_Y be their respective means. Then*

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mu_X \mu_Y.$$

3.4. **Expectation and Variance of Important Random Variables.** In this section we record the expectation and variance of some important random variables. Before we do so we must define these notions for *random vectors.*

Definition 3.12 (Expectation and variance for random vectors)**.** Let X be random vector. Its *mean (vector)* is defined to be

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_k) \end{pmatrix}$$

Distribution	Mean	Variance
Point mass at a	a	0
Bernoulli(p)	p	$p(1-p)$
Binomial$(n,\, p)$	np	$np(1-p)$
Geometric(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform$(a,\, b)$	$(a+b)/2$	$(b-a)^2/12$
Normal$(\mu,\, \sigma^2)$	μ	σ^2
Exponential(β)	β	β^2
Gamma$(\alpha,\, \beta)$	$\alpha\beta$	$\alpha\beta^2$
Beta(α, β)	$\alpha/(\alpha+\beta)$	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$
t_ν	0 (if $\nu > 1$)	$\nu/(\nu-2)$ if $\nu > 2$
χ^2_p	p	$2p$
Multinomial$(n,\, p)$	np	(see below)
Multivariate Normal$(\mu,\, \Sigma)$	μ	Σ

FIGURE 3.1. Expectation and variance of important random variables.

and its *variance–covariance matrix* is defined to be

$$\Sigma = \begin{pmatrix} \mathbb{V}(X_1) & \mathrm{Cov}(X_1,\, X_2) & \cdots & \mathrm{Cov}(X_1,\, X_k) \\ \mathrm{Cov}(X_2,\, X_1) & \mathbb{V}(X_2) & \cdots & \mathrm{Cov}(X_2,\, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_k,\, X_1) & \mathrm{Cov}(X_k,\, X_2) & \cdots & \mathbb{V}(X_k) \end{pmatrix}.$$

The variance–covariance matrix of a Multinomial$(n,\, p)$ random variable is

$$\begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_k \\ -np_2p_1 & np_2(1-p_2) & \cdots & -np_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -np_kp_1 & -np_kp_2 & \cdots & np_k(1-p_k) \end{pmatrix}.$$

3.5. **Conditional Expectation.**

Definition 3.13 (Conditional expectation)**.** Let X and Y be random variables. The *conditional expectation of X given $Y = y$* is

$$\mathbb{E}(X \,|\, Y = y) = \begin{cases} \displaystyle\sum_x x f_{X|Y}(x|y) & \text{if } X \text{ and } Y \text{ are discrete and} \\ \displaystyle\int x f_{X|Y}(x|y)dx & \text{if } X \text{ and } Y \text{ are continuous.} \end{cases}$$

Moreover, if $r(x,\, y)$ is a function of x and y then

$$\mathbb{E}(r(X,\, Y) \,|\, Y = y) = \begin{cases} \displaystyle\sum_x r(x,\, y) f_{X|Y}(x|y) & \text{if } X \text{ and } Y \text{ are discrete and} \\ \displaystyle\int r(x,\, y) f_{X|Y}(x|y)dx & \text{if } X \text{ and } Y \text{ are continuous.} \end{cases}$$

Remark 3.14 (Conditional expectations are random variables)**.** Since the conditional expectation $\mathbb{E}(X \,|\, Y = y)$ is a function of y, we often omit the mention of

y to describe the *random variable* $\mathbb{E}(X|Y)$. Indeed, for any given outcome ω, the value of $\mathbb{E}(X|Y)$ will be a function of the value of $Y(\omega)$, and so indeed $\mathbb{E}(X|Y)$ is a random variable.

Theorem 3.15 (Rule of Iterated Expectation)**.** *Let X and Y be random variables. Then, assuming all expectations below exist,*

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)).$$

Moreover, if $r(x, y)$ is a function of x and y then

$$\mathbb{E}(r(X, Y)) = \mathbb{E}(\mathbb{E}(r(X, Y) \mid Y)).$$

Definition 3.16 (Conditional variance)**.** Let X and Y be a random variables. The *conditional variance of X given $Y = y$* is defined to be

$$\mathbb{V}(X \mid Y = y) = \mathbb{E}\left[(X - \mu(y))^2 \middle| Y = y \right]$$

for $\mu(y) = \mathbb{E}(X \mid Y = y)$. In particular if X and Y are continuous random variables we have that

$$\mathbb{V}(X \mid Y = y) = \int (x - \mu(y))^2 f(x|y)dx$$

Theorem 3.17 (Law of Total Variance)**.** *Let X and Y be random variables. Provided that all of the expectation below exist,*

$$\mathbb{V}Y = \mathbb{E}\mathbb{V}(Y|X) + \mathbb{V}\mathbb{E}(Y|X).$$

Remark 3.18. The two terms in Theorem 3.17 above are sometimes known as the *unexplained* and *explained* components of the variance. To make sense of this we think of X as grouping outcomes into categories and think of Y as a random variable of interest. Knowing which category we are in (i.e. knowing the value of X) then gives us information about the value of Y. The variance in Y then has two components:

- within each category there is variance in Y due to inherent randomness in Y, this intra-category variance is the unexplained component of the variance, namely $\mathbb{E}\mathbb{V}(Y|X)$, and
- between each category there is variance in Y due to the simple fact that outcomes in different categories are meaningfully different from each other, this inter-category variance is the explained component of the variance, namely $\mathbb{V}\mathbb{E}(Y|X)$.

Note that, as shown in the proof of Exercise A.3.15, this decomposition of the variance holds because of *orthogonality*: in the notation of Theorem 3.17, $Y - \mu(X)$ and $\mu(X) - \mathbb{E}Y$ are orthogonal (where $\mu(X) = b(X)$ in the notation of Exercise A.3.15). In other words: the difference of Y with its conditional mean $\mu(X)$ is orthogonal to the difference of the conditional mean with the true mean $\mathbb{E}Y$; i.e. the intra-category variation is orthogonal to the inter-category variation.

3.6. **Moment generating functions.**

Definition 3.19 (Moment generating function)**.** Let X be a random variable. The *moment generating function*, or *MGF*, or *Laplace transform* of X is defined to be

$$\psi_X(t) = \mathbb{E}\left(e^{tX}\right).$$

Distribution	MGF
Bernoulli(p)	$pe^t + (1-p)$
Binomial$(n,\, p)$	$(pe^t + (1-p))^n$
Poisson(λ)	$e^{\lambda(e^t-1)}$
Normal$(\mu,\, \sigma^2)$	$\exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$
Gamma$(\alpha,\, \beta)$	$\left(\frac{1}{1-\beta t}\right)^\alpha$ for $t < 1/\beta$

FIGURE 3.2. Moment generating functions of important random variables.

Lemma 3.20 (Moment generating function and moments)**.** *Let X be a random variable whose MGF ψ_X is well-defined in an open neighbourhood of zero. The k-th moment of X is given by $\mathbb{E}(X^k) = \psi^{(k)}(0)$.*

Lemma 3.21 (Moment generating function of a sum)**.** *Let $X_1,\, \ldots,\, X_n$ be independent random variables with MGFs given by ψ_{X_i} for $i = 1,\, \ldots,\, n$. For $Y = \sum_{i=1}^n X_i$, the moment generating function of Y is given by $\psi_Y = \prod_{i=1}^n \psi_{X_i}$.*

3.7. Bonus.

Lemma 3.22 (Expectation and variance of a ratio of random variables)**.** *Let X and Z be random variables such that Z has codomain $[0,\, \infty)$ and*

$$\begin{aligned} \mu_X &:= \mathbb{E}X, & \sigma_X^2 &:= \mathbb{V}X, \\ \mu_Z &:= \mathbb{E}Z, \quad \textit{and} & \sigma_Z^2 &:= \mathbb{V}Z \end{aligned}$$

are well-defined. If the joint PDF of $(X,\, Z)$ is sufficiently regular then

$$\mathbb{E}\left(\frac{X}{Z}\right) \approx \frac{\mu_X}{\mu_Z} - \frac{\operatorname{Cov}(X,\, Z)}{\mu_Z^2} + \frac{\sigma_X^2 \mu_X}{\mu_Z^3}$$

and

$$\mathbb{V}\left(\frac{X}{Z}\right) \approx \frac{\mu_X^2}{\mu_Z^2}\left[\frac{\sigma_X^2}{\mu_X^2} + \frac{\sigma_Z^2}{\mu_Z^2} - 2\frac{\operatorname{Cov}(X,\, Z)}{\mu_X \mu_Z}\right].$$

Proof. This follows from a Taylor expansion of the joint PDF (see [Sel]). □

Definition 3.23 (Median absolute deviation)**.** Let X be a random variable and let m denote its median. The *median absolute deviation* of X is defined to be the median of $|X - m|$.

Lemma 3.24 (Identity for the median absolute deviation)**.** *Let X be a random variable whose PDF is symmetric about its median m and strictly positive everywhere. Then the median absolute deviation of X is*

$$F^{-1}\left(\frac{3}{4}\right) - m$$

where F denotes the CDF of X.

4. Inequalities

4.1. Probability inequalities.

Theorem 4.1 (Markov's inequality). *Let X be a non-negative random variable and suppose that $\mathbb{E}X$ exists. For any $t > 0$,*

$$\mathbb{P}(X > t) \leqslant \frac{\mathbb{E}(X)}{t}.$$

Theorem 4.2 (Chebyshev's inequality). *Let X be a random variable whose expectation μ and variance σ^2 exist. Then, for any $t,\, k > 0$,*

$$\mathbb{P}\left(|X-\mu| \geqslant t\right) \leqslant \frac{\sigma^2}{t^2} \text{ and } \mathbb{P}\left(|Z| \geqslant k\right) \leqslant \frac{1}{k^2}$$

for $Z := (X-\mu)/\sigma$.

Theorem 4.3 (Hoeffding's inequality, version 1). *Let $Y_1, \dots, Y_n$ be independent random variables which satisfy $\mathbb{E}(Y_i)$ and $a_i \leqslant Y \leqslant b_i$ for all i. Then, for any $\varepsilon,\, t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geqslant \varepsilon\right) \leqslant e^{-t\varepsilon} \prod_{i=1}^{n} e^{t^2(b_i-a_i)^2/8}.$$

Theorem 4.4 (Hoeffding's inequality, version 2). *Let $X_1, \dots, X_n \sim$ Bernoulli(p) be IID. For any $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| > \varepsilon\right) \leqslant 2e^{-2n\varepsilon^2}$$

for $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$.

Theorem 4.5 (Mill's inequality). *Let $Z \sim N(0,\, 1)$. For any $t > 0$,*

$$\mathbb{P}\left(|Z| > t\right) \leqslant \sqrt{\frac{2}{\pi}}\frac{e^{-t^2/2}}{t}.$$

4.2. Inequalities for Expectations.

Theorem 4.6 (Cauchy-Schwarz inequality). *Let X and Y be random variables with finite variance. Then*

$$\mathbb{E}|XY| \leqslant \sqrt{\mathbb{E}\left(X^2\right)\mathbb{E}\left(Y^2\right)}.$$

Theorem 4.7 (Jensen's inequality). *Let X be a random variable whose expectation exists and let $g : \mathbb{R} \to \mathbb{R}$ be convex, meaning that*

$$g((1-\theta)x + \theta y) \leqslant (1-\theta)g(x) + \theta g(y)$$

for any $\theta \in [0,\, 1]$ and any $x,\, y \in \mathbb{R}$. Then

$$g\left(\mathbb{E}X\right) \leqslant \mathbb{E}g(X).$$

5. Convergence of Random Variables

5.1. Types of Convergence.

Definition 5.1 (Convergence in probability)**.** Let $X_1, X_2, \dots$ be a sequence of random variables and let X be another random variable. We say that X_n *converges to X in probability*, written $X_n \xrightarrow{P} X$, if, for every $\varepsilon > 0$,

$$\mathbb{P}\left(|X_n - X| > \varepsilon\right) \to 0$$

as $n \to \infty$.

Definition 5.2 (Convergence in distribution)**.** Let $X_1, X_2, \dots$ be a sequence of random variables and let X be another random variable. Let F_n denote the CDF of X_n and F denote the CDF of X. We say that X_n *converges to X in distribution*, written $X_n \rightsquigarrow X$, if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

at every t where F is continuous.

Definition 5.3 (Convergence in quadratic mean)**.** Let $X_1, X_2, \dots$ be a sequence of random variables and let X be another random variable. We say that X_n *converges to X in quadratic mean*, written $X_n \xrightarrow{qm} X$, if

$$\mathbb{E}(X_n - X)^2 \to 0$$

as $n \to \infty$.

Theorem 5.4 (Relations between various modes of convergence)**.** *Let $X_1, X_2, \dots$ be a sequence of random variables and let X be another random variable. The following hold.*

1. *If X_n converges to X in quadratic mean then X_n converges to X in probability.*
2. *If X_n converges to X in probability then X_n converges to X in distribution.*
3. *If X_n converges to X in distribution and the distribution of X is a finite sum of point masses, then X_n also converges to X in probability.*

Theorem 5.5 (Closure of various modes of convergence)**.** *The following hold.*

1. *Convergence in quadratic mean is closed under linear combinations, i.e. if $X_n \xrightarrow{qm} X$ and $Y_n \xrightarrow{qm} Y$ then $aX_n + bY_n \xrightarrow{qm} aX + bY$ for any $a, b \in \mathbb{R}$.*
2. *Convergence in probability is closed under algebraic combinations (i.e. under sums and products) and under continuous functions i.e. if $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then $aX_n + bY_n \xrightarrow{P} aX + bY$, for any $a, b \in \mathbb{R}$, $X_nY_n \xrightarrow{P} XY$, and $g(X_n) \xrightarrow{P} g(X)$ for any continuous function $g : \mathbb{R} \to \mathbb{R}$.*
3. *Convergence in distribution is closed under continuous function, i.e. if $X_n \rightsquigarrow X$ then $g(X_n) \to g(X)$ for any continuous function $g : \mathbb{R} \to \mathbb{R}$. This includes $g(x) = cx$ and $g(x) = x + c$ for any constant c.*

5.2. The Law of Large Numbers.

Theorem 5.6 (Weak Law of Large Numbers)**.** *If $X_1, X_2, \dots$ are IID with mean μ then the sample mean $\overline{X}_n$ converges in probability to the mean μ.*

Remark 5.7 (Interpretation of the (weak) law of large numbers)**.** The weak law of large numbers tells us that the distribution of the sample mean becomes more concentrated around the true mean as the sample size increases. In other words the sample mean is eventually close to the true mean with high probability.

5.3. **The Central Limit Theorem.**

Theorem 5.8 (Central Limit Theorem)**.** *Let $X_1, \ldots, X_n$ be IID with mean μ and variance σ^2. Let $\overline{X}_n$ denote the sample mean. Then*

$$\frac{\overline{X}_n - \mu}{\sqrt{\mathbb{V}\left(\overline{X}_n\right)}} = \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \rightsquigarrow N(0,\, 1).$$

Remark 5.9 (Interpretation of the central limit theorem)**.** The central limit theorem tells us that probability statements about the sample mean can be approximate by probability statements about a Normal distribution. In particular it tells us that

$$\overline{X}_n \approx N\left(\mu,\, \frac{\sigma^2}{n}\right).$$

Remark 5.10 (Contrasting the weak law of large numbers and the central limit theorem)**.** The weak law of large numbers only tell us that the sample mean eventually approximates the true mean. The central limit theorem goes one step further and quantifies how far the sample mean will eventually be from the true mean. Of course this additional information comes at an additional cost: we need finite variance for the central limit theorem to hold, yet only finite expectation is required for the weak law of large numbers to hold.

Theorem 5.11 (Central Limit Theorem using the Sample Variance)**.** *Let $X_1, \ldots, X_n$ be IID with mean μ and finite variance. Let $\overline{X}_n$ denote the sample mean and let S_n^2 denote the sample variance. Then*

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{S_n} \rightsquigarrow N(0,\, 1).$$

Theorem 5.12 (Multivariate Central Limit Theorem)**.** *Let $X_1, \ldots, X_n$ be IID random vectors, where*

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{ki} \end{pmatrix},$$

with mean μ and variance–covariance matrix Σ. Let $\overline{X}$ denote the sample mean whose j-th coordinate is the sample mean of X_j, i.e.

$$\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ji}.$$

Then

$$\sqrt{n}\left(\overline{X} - \mu\right) \rightsquigarrow N(0,\, \Sigma).$$

5.4. **The Delta Method.**

Theorem 5.13 (Delta Method)**.** *Let $Y_1, Y_2, \ldots$ be a sequence of random variables and suppose that*

$$\frac{\sqrt{n}\,(Y_n - \mu)}{\sigma} \rightsquigarrow N(0, 1)$$

for some $\mu \in \mathbb{R}$ and $\sigma > 0$. For any differentiable function $g : \mathbb{R} \to \mathbb{R}$ satisfying $g'(\mu) \neq 0$ we have that

$$\frac{\sqrt{n}\,(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \rightsquigarrow N(0, 1).$$

In other words, if

$$Y_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

then

$$g(Y_n) \approx N\left(g(\mu), (g'(\mu))^2 \cdot \frac{\sigma^2}{n}\right).$$

Theorem 5.14 (Multivariate Delta Method)**.** *Let $Y_n = (Y_{1n}, \ldots, Y_{kn})$ be a sequence of random vectors such that*

$$\sqrt{n}\,(Y_n - \mu) \rightsquigarrow N(0, \Sigma)$$

for some $\mu \in \mathbb{R}^k$ and some positive-definite and symmetric $\Sigma \in \mathbb{R}^{k \times k}$. Suppose $g : \mathbb{R}^k \to \mathbb{R}$ is differentiable and that $\nabla g(\mu) \neq 0$. Then

$$\sqrt{n}\,(g(Y_n) - g(\mu)) \rightsquigarrow N\left(0, \nabla g(\mu)^T \Sigma \nabla g(\mu)\right).$$

6. Models, Statistical Inference, and Learning

Statistical inference attempts to answer the following question: given a sample $X_1, \ldots, X_n \sim F$, how do we infer F?

6.1. **Parametric and Nonparametric Models.**

Definition 6.1 (Statistical model)**.** A *statistical model* $\mathcal{F}$ is a set of distributions.

- It is called a *parametric model* if there exists a finite-dimensional set Θ, called the *parameter space*, such that $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$. The elements $\theta \in \Theta$ are called *parameters.*
- Otherwise it is called a *nonparametric model.*

If $\theta \in \Theta$ for some parameter space Θ of a parametric model $\mathcal{F}$ we then write

$$\mathbb{P}_\theta(X \in A) := \int_A f(x;\theta)dx$$

and similarly for $\mathbb{E}_\theta$ and $\mathbb{V}_\theta$.

Definition 6.2 (Statistical function)**.** Let $\mathcal{F}$ be a statistical model. A function $T : \mathcal{F} \to \mathbb{R}$ is called a *statistical functional.*

Definition 6.3 (Regression, regression models, and tasks)**.** Suppose X and Y are random variables. The function $r(x) := \mathbb{E}(Y|X = x)$ is called a *regression function*, in which case

- X is known as a *covariate*, *predictor*, *regressor*, *feature*, or *independent variable* and
- Y is known as a *response variable* or *dependent variable.*

Suppose moreover that the regression function r belongs to some set $\mathcal{F}$.

- If $\mathcal{F}$ is finite-dimensional we call it a *parametric regression model.*
- Otherwise we say that $\mathcal{F}$ is a *nonparametric regression model.*

We may then perform one of several *tasks.*

- If, given a new outcome ω, we seek to determine $Y(\omega)$ based on the observation of $X(\omega)$, then we say that we are performing a *prediction task.*
- If the response variable of a prediction task is discrete then we say that we are performing a *classification task.*
- If, given a sequence of samples (X_1, Y_1), (X_2, Y_2), $\ldots$ we seek to estimate the regression function $r \in \mathcal{F}$ then we say that we are performing a *regression task* or *curve estimation.*

Remark 6.4 (Characterisation of regression models)**.** We can always write a regression model in the form

$$Y = r(X) + \varepsilon$$

where ε is a random variable with mean zero.

Indeed: it suffices to define the $\varepsilon = Y - r(X)$ and then use the rule of iterated expectation to deduce that the expectation of ε vanishes.

6.2. **Point Estimation.**

Definition 6.5 (Point estimator)**.** Let $\mathcal{F}$ be a parametric model with parameter space Θ and let $\theta \in \Theta$ be a parameter. A *point estimator* $\hat{\theta}_n$ of the parameter θ is a random variable defined by

$$\hat{\theta}_n = g_n(X_1, \ldots, X_n)$$

where $X_1, \ldots, X_n$ are IID from $f(\,\cdot\,; \theta)$, for some function $g_n : \mathbb{R}^n \to \mathbb{R}$.

Definition 6.6 (Bias)**.** Let $\hat{\theta}_n$ be a point estimator of some parameter θ. The *bias* of $\hat{\theta}_n$ is defined as

$$\operatorname{bias}(\hat{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}_n) - \theta.$$

We say that a point estimator is *unbiased* if its bias is equal to zero for all n.

Definition 6.7 (Consistency)**.** Let $\hat{\theta}_n$ be a point estimator of some parameter θ. We say that $\hat{\theta}_n$ is *consistent* if $\hat{\theta}_n \xrightarrow{P} \theta$.

Definition 6.8 (Standard error)**.** Let $\hat{\theta}_n$ be a point estimator. The *standard error* of $\hat{\theta}_n$ is defined as

$$\operatorname{se}(\hat{\theta}_n) = \sqrt{\mathbb{V}_\theta(\hat{\theta}_n)}.$$

Definition 6.9 (Mean squared error)**.** Let $\hat{\theta}_n$ be a point estimator of some parameter θ. The *mean squared error* of $\hat{\theta}_n$ is defined as

$$\operatorname{MSE}(\hat{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}_n - \theta)^2.$$

Theorem 6.10 (Decomposition of the mean squared error)**.** *Let $\hat{\theta}_n$ be a point estimator of some parameter θ. The mean squared error of $\hat{\theta}_n$ may be written in terms of its bias and standard error as follows:*

$$\operatorname{MSE}(\hat{\theta}_n) = \operatorname{bias}^2(\hat{\theta}_n) + \operatorname{se}^2(\hat{\theta}_n).$$

Remark 6.11. The decomposition of the mean squared error in Theorem 6.10 above may be understood as follows. There are two sources of error in the point estimator: the first source of error is the bias, which is how far the point estimator is, on average, from the parameter it estimates, and the second source of error is the standard error, which is the inherent randomness of the point estimator itself.

Definition 6.12 (Asymptotic normality)**.** Let $\hat{\theta}_n$ be a point estimator of some parameter θ. We say that this estimator is *asymptotically Normal* if

$$\frac{\hat{\theta}_n - \theta}{\operatorname{se}(\hat{\theta}_n)} \rightsquigarrow N(0, 1).$$

6.3. **Confidence Sets.**

Definition 6.13 (Confidence interval and coverage)**.** Let $\mathcal{F}$ be a parametric model with parameter space Θ and let $\alpha \in [0, 1]$. A $1 - \alpha$ *confidence interval for a parameter* θ is an interval $C_n = (a_n, b_n)$ where a_n and b_n are random variables defined by

$$a_n = a(X_1, \ldots, X_n) \text{ and } b_n = b(X_1, \ldots, X_n),$$

where $X_1, \ldots, X_n$ are IID from $f(\,\cdot\,; \theta)$, such that

$$\mathbb{P}_\theta\,(\theta \in C_n) \geqslant 1 - \alpha \text{ for all } \theta \in \Theta.$$

We refer to $1-\alpha$ as the *coverage* of the confidence interval. Moreover if θ is a vector then we use the term *confidence set* instead of confidence interval.

Theorem 6.14 (Normal-based Confidence Interval)**.** *Let $\mathcal{F}$ be a parametric model with parameter space Θ, let $\theta \in \Theta$ be a parameter, and let $\alpha \in [0, 1]$. Suppose that $\hat{\theta}_n$ is a point estimator for θ which is asymptotically Normal, let $\widehat{\mathrm{se}}_n$ denote a consistent point estimator of the standard error of $\hat{\theta}_n$, and define $z := \Phi^{-1}(1-\alpha/2)$ for Φ denoting the CDF of the standard Normal distribution, which means that $\mathbb{P}(-z < Z < z) = 1-\alpha$ for $Z \sim N(0, 1)$. Define*

$$C_n = \left(\hat{\theta}_n - z\,\widehat{\mathrm{se}}_n,\, \hat{\theta}_n + z\,\widehat{\mathrm{se}}_n\right).$$

Then C_n is an asymptotic $1-\alpha$ confidence interval for θ, meaning that

$$\mathbb{P}_\theta(\theta \in C_n) \to 1-\alpha \textit{ as } n \to \infty \textit{ for every } \theta \in \Theta.$$

6.4. **Hypothesis Testing.** As will be discussed in more detail in Section 10 below, hypothesis proceeds as follows. We begin with a default theory, known as a *null hypothesis*, and then ask if the data gathered provides sufficient evidence to *reject* the theory. If not, we retain the null hypothesis.

7. Estimating the CDF and Statistical Functionals

7.1. The Empirical Distribution Function.

Definition 7.1 (Empirical distribution function)**.** Let $X_1, \ldots, X_n \sim F$ be IID with respect to a CDF F. The *empirical distribution function* is the CDF $\hat{F}_n$ defined by

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x)$$

for every $x \in \mathbb{R}$, where

$$I(X \leqslant x) = \begin{cases} 1 & \text{if } X \leqslant x \text{ and} \\ 0 & \text{if } X > x. \end{cases}$$

Note that, just like other point estimators, the empirical CDF is a *random variable*.

Theorem 7.2 (Dvoretzky-Kiefer-Wolfowitz inequality)**.** *Let $X_1, \ldots, X_n \sim F$ be IID with respect to some CDF F and let $\widehat{F}_n$ denote the empirical CDF of F. For any $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|F(x) - \widehat{F}_n(x)\right| > \varepsilon\right) \leqslant 2e^{-2n\varepsilon^2}.$$

Remark 7.3 (Nonparametric confidence bands for the CDF)**.** Let F be a CDF and let $\alpha \in (0, 1)$. Define, for any positive integer n,

$$\varepsilon_n := \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$$

and then define, for any $x \in \mathbb{R}$,

$$L_n(x) := \max\left\{\widehat{F}_n(x) - \varepsilon_n,\, 0\right\} \text{ and } U_n(x) := \min\left\{\widehat{F}_n(x) + \varepsilon_n,\, 1\right\}$$

where $\widehat{F}_n$ denotes the empirical CDF. It follows from the DKW inequality recorded in Theorem 7.2 that (L_n, U_n) forms a *uniform* $1 - \alpha$ confidence interval for F, meaning that, for every n,

$$\mathbb{P}\left(L_n(x) \leqslant F(x) \leqslant U_n(x) \text{ for every } x \in \mathbb{R}\right) \geqslant 1 - \alpha.$$

Definition 7.4 (Plug-in estimator)**.** Let T be a statistical functional and let $\widehat{F}_n$ be the empirical CDF of a CDF F. The point estimator $T(\widehat{F}_n)$ is known as the *plug-in estimator* for $T(F)$.

Definition 7.5 (Linear functional and Riesz representation theorem)**.** Let T be a linear statistical functional over a statistical model $\mathcal{F}$. T is called a *linear functional* and, by the Riesz representation theorem, there exists a function $r : \mathbb{R} \to \mathbb{R}$ such that, for any CDF $F \in \mathcal{F}$,

$$T(F) = \int r(x)dF(x).$$

We call r the *representation* of T.

Theorem 7.6 (Plug-in estimator for linear functionals)**.** *Let T be a linear functional over a statistical model $\mathcal{F}$ with representation r. For any CDF $F \in \mathcal{F}$ let $\widehat{F}_n$ denote its empirical CDF. The plug-in estimator for $T(F)$ may be written as*

$$T(\widehat{F}_n) = \int r(x) d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i).$$

8. The Bootstrap

Definition 8.1 (Statistical estimator)**.** A *statistical estimator* is a collection of maps $T_1, T_2, \ldots$ satisfying $T_n : \mathbb{R}^n \to \mathbb{R}$.

Remark 8.2 (Statistical estimator)**.** In practice the arguments of a statistical estimator are taken to be IID random variables $X_1, \ldots, X_n \sim F$ for some CDF F. This occurs when the statistical estimator T_n is chosen to approximate a given statistical functional T, in the sense that $T_n(X_1, \ldots, X_n)$ approaches $T(X)$ as $n \to \infty$, for X a random variable with CDF F.

For example: point estimators, and in particular plug-in estimators, are statistical estimators.

Remark 8.3 (The Bootstrap)**.** The bootstrap is used to estimate variances, and hence build confidence intervals, for statistical estimators. This is done in two steps.

(1) The true variance $\mathbb{V}_F(T_n)$ is approximated by $\mathbb{V}_{F_n}(T_n)$, where F denotes the true underlying CDF and F_n denotes the empirical CDF associated with some IID samples.
(2) The quantity $\mathbb{V}_{F_n}(T_n)$ is approximated by *simulation*.

8.1. Simulation.

Remark 8.4 (Simulation)**.** Let Y be a random variable with CDF G and let $r : \mathbb{R} \to \mathbb{R}$ be a function. Say we wish to approximate the expected value of $r(Y)$ (since it may, for example, be a statistical functional of interest) using only IID samples $Y_1, \ldots, Y_B$. By the Weak Law of Large Numbers and the Rule of the Lazy Statistician:

$$\frac{1}{B}\sum_{j=1}^{B} r(Y_j) \to \int r(y)dG(y) = \mathbb{E}(r(Y)).$$

In other words we can approximate the expected value of $r(Y)$ by computing the sample mean of the transformed samples, i.e. the sample mean of $r(Y_1), \ldots, r(Y_B)$. This is known as *simulation*.

8.2. Bootstrap Variance Estimation.

Remark 8.5 (Bootstrap variance estimation)**.** Let $T_n(X_1, \ldots, X_n)$ be a statistical estimator, for $X_1, \ldots, X_n$ IID random variables with CDF F. To estimate the variance of T_n using the bootstrap method we may proceed as follows.

- Repeat B times, for $j = 1, \ldots, B$, the following.
  (1) Draw $X^*_{1,j}, \ldots, X^*_{n,j} \sim \hat{F}_n$, where $\hat{F}_n$ denotes the empirical CDF associated with the original sample $X_1, \ldots, X_n$. This really means drawing n points at random, with replacement, from the original samples $X_1, \ldots, X_n$.
  (2) Compute $T^*_{n,j} := T_n(X^*_{1,j}, \ldots, X^*_{n,j})$.
- Finally compute

$$v_{boot} := \frac{1}{B}\sum_{j=1}^{B}\left(T^*_{n,j} - \frac{1}{B}\sum_{k=1}^{B} T^*_{n,k}\right)^2.$$

This quantity is the plug-in estimator for the variance and is known as the *population variance* of the samples $T^*_{n,1}, \ldots, T^*_{n,B}$.

Note that the population variance differs from the sample variance, which has a pre-factor of $\frac{1}{B-1}$ instead of $\frac{1}{B}$. As shown in Exercise A.3.8, this ensures that the sample variance is an unbiased estimator of the variance, whereas the population variance is therefore a biased estimator of the variance (although both are consistent).

8.3. **Bootstrap Confidence Intervals.**

Remark 8.6 (Bootstrap Normal interval)**.** Let $T_n(X_1, \ldots, X_n)$ be a statistical estimator with $X_1, \ldots, X_n$ IID random variables. Let v_{boot} denote the bootstrap variance estimate and $\widehat{se}_{boot} := \sqrt{v_{boot}}$ denote the bootstrap estimate of the standard error. Consider $0 < \alpha < 1$ and let $z_{\alpha/2} := \Phi^{-1}(1 - \alpha/2)$ where Φ denotes the CDF of the standard normal distribution, which means that $\mathbb{P}(-z < Z < z) = 1-\alpha$ for $Z \sim N(0, 1)$.

The $1 - \alpha$ *bootstrap Normal confidence interval* for T_n, which is only accurate if T_n has a distribution "close" to a Normal distribution, is given by

$$C_n = \left(T_n - z_{\alpha/2}\, \widehat{se}_{boot},\, T_n + z_{\alpha/2}\, \widehat{se}_{boot}\right).$$

Remark 8.7 (Bootstrap pivotal interval)**.** Let T be a statistical functional and let F be a CDF from the corresponding statistical model. Let $X_1, \ldots, X_n \sim F$ be IID random variables and let $\hat{F}_n$ be the corresponding empirical CDF.

We consider $\theta := T(F)$ and its plug-in estimator $\hat{\theta}_n := T(\hat{F}_n)$. Let $\hat{\theta}^*_{n,\,1}, \ldots, \hat{\theta}^*_{n,\,B}$ denote bootstrap replications of $\hat{\theta}_n$, meaning that, for $j = 1, \ldots, B$,

$$\hat{\theta}^*_{n,\,j} := T\left(X^*_{1,\,j}, \ldots, X^*_{n,\,j}\right) \text{ for } X^*_{1,\,j}, \ldots, X^*_{n,\,j} \sim \hat{F}_n \text{ IID}.$$

For $0 < \alpha,\, \beta < 1$ let θ^*_β denote the β–quantile of $\hat{\theta}^*_{n,\,1}, \ldots, \hat{\theta}^*_{n,\,B}$ and define

$$\begin{cases} \hat{\alpha}_n := \hat{\theta}_n - \left(\theta^*_{1-\alpha/2} - \hat{\theta}_n\right) & = 2\hat{\theta}_n - \theta^*_{1-\alpha/2} \text{ and} \\ \hat{\beta}_n := \hat{\theta}_n - \left(\theta^*_{\alpha/2} - \hat{\theta}_n\right) & = 2\hat{\theta}_n - \theta^*_{\alpha/2}. \end{cases}$$

The *bootstrap pivotal confidence interval* for θ is given by

$$C_n = (\hat{a}_n,\, \hat{b}_n).$$

Where does this come from? The lower bound $\hat{a}_n$ is an approximation of the *exact* value (but not computable using the samples)

$$a := \hat{\theta}_n - H^{-1}(1 - \alpha/2)$$

where H is the CDF of the *pivot* random variable $R_n := \hat{\theta}_n - \theta$. [1] Indeed, since bootstrap replications of R_n, namely $\hat{\theta}^*_{n,\,j} - \hat{\theta}_n$, have the *same* quantiles as the bootstrap replications of $\hat{\theta}_n$, up to a shift by a constant value (constant provided the samples are fixed, which is precisely the underlying assumption of the bootstrap method), we therefore have that, by a bootstrap approximation for R_n and hence H,

$$H^{-1}(\beta) \approx \theta^*_\beta - \hat{\theta}_n,$$

[1] Random variables are known as *pivots* when they are function of IID samples $X_1, \ldots, X_n$ and yet their distribution is *fixed* and does not depend on the values of $X_1, \ldots, X_n$. For example the random variable $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ appearing in the Central Limit Theorem is an asymptotic pivot. Here R_n is not quite a pivot since it would require an appropriate scaling. However only the *quantiles* associated with R_n matter and so the scaling is not required.

and so

$$a \approx \hat{\theta}_n - (\theta^*_\beta - \hat{\theta}_n) =: \hat{a}_n,$$

as desired, and similarly for $\hat{b}_n$. A simple direct computation verifies that $(a,\, b)$ is an exact $1-\alpha$ confidence interval for θ, and so the bootstrap pivotal confidence interval given above is indeed an *approximate* $1-\alpha$ confidence interval.

Remark 8.8 (Bootstrap percentile interval)**.** Let T be a statistical functional and let F be a CDF from the corresponding statistical model. Let $X_1,\, \dots,\, X_n \sim F$ be IID random variables and let $\hat{F}_n$ be the corresponding empirical CDF.

We consider $\theta := T(F)$ and its plug-in estimator $\hat{\theta}_n := T(\hat{F}_n)$. Let $\hat{\theta}^*_{n,\,1},\, \dots,\, \hat{\theta}^*_{n,\,B}$ denote bootstrap replications of $\hat{\theta}_n$, meaning that, for $j = 1,\, \dots,\, B$,

$$\hat{\theta}^*_{n,\,j} := T\left(X^*_{1,\,j},\, \dots,\, X^*_{n,\,j}\right) \text{ for } X^*_{1,\,j},\, \dots,\, X^*_{n,\,j} \sim \hat{F}_n \text{ IID }.$$

For $0 < \beta < 1$ let θ^*_β denote the β–quantile of $\hat{\theta}^*_{n,\,1},\, \dots,\, \hat{\theta}^*_{n,\,B}$. Then, for any $0 < \alpha < 1$, the $1-\alpha$ *bootstrap percentile confidence interval* for θ is given by

$$C_n = (\theta^*_{\alpha/2},\, \theta^*_{1-\alpha/2}).$$

This is justified as long as there exists a monotone transformation taking T_n to a known, and fixed, distribution (such as for example a normal distribution).

9. Parametric Inference

9.1. The Methods of Moments.

Recall that the moment of a random variable is introduced in Definition 3.4.

Definition 9.1 (Sample moments)**.** Let $X_1, \ldots, X_n$ be random variables. For any $k \geqslant 1$ the *k–th sample moment* is defined to be

$$\hat{m}_k := \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Definition 9.2 (Method of moments estimator)**.** Consider a parametric model with parameter space $\Theta \subseteq \mathbb{R}^k$. The *method of moments estimator* $\hat{\theta}_n \in \Theta$ is defined to be the solution of

$$\begin{aligned} m_1(\theta) &= \hat{m}_1, \\ m_2(\theta) &= \hat{m}_2, \\ &\vdots \\ m_k(\theta) &= \hat{m}_k \end{aligned}$$

where $m_k(\theta)$ denotes the k-th moment of X when the true parameter is θ.

Theorem 9.3 (Method of moments estimator)**.** *Let $\hat{\theta}_n$ denote the method of moments estimator for a sufficiently regular parametric model with parameter space Θ. The following hold.*

(1) The estimate $\hat{\theta}_n$ exists with probability tending to 1.
(2) The estimate is consistent.
(3) The estimate is asymptotically Normal with

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \rightsquigarrow N(0, \Sigma)$$

where

$$\Sigma = \left[D\left(m^{-1}\right)\right]^T \mathbb{E}\left(Y \otimes Y\right) D\left(m^{-1}\right)$$

for $Y_i := X^i$ and $m : \Theta \to \mathbb{R}^k$ the map obtained by viewing each moment m_i as a function $m_i : \Theta \to \mathbb{R}$.

9.2. Maximum Likelihood.

Definition 9.4 (Likelihood functions)**.** Let $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a parametric model and let $X_1, \ldots, X_n$ be IID random variables with distribution $f(\,\cdot\,;\theta)$. The *likelihood function* is defined to be

$$\mathcal{L}_n(\theta) := \prod_{i=1}^{n} f\left(X_i;\theta\right).$$

Note that the likelihood function is a random variable. The *log-likelihood function* is defined to be $l_n(\theta) := \log \mathcal{L}_n(\theta)$.

Definition 9.5 (Maximum likelihood estimator)**.** Let $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a parametric model, let $X_1, \ldots, X_n$ be IID from $f(\,\cdot\,;\theta)$, and let $\mathcal{L}_n$ denote the corresponding likelihood function. The *maximum likelihood (point) estimator* (MLE) of the parameter θ, denoted by $\hat{\theta}_n$, is the value of θ which maximizes $\mathcal{L}_n(\theta)$.

9.3. Consistency of Maximum Likelihood Estimators.

Definition 9.6 (Kullback-Leibler divergence)**.** Let f and g be PDFs. The *Kullback-Leibler divergence* is defined to be

$$D(f,\, g) := \int f log\left(\frac{f}{g}\right).$$

If f and g belong to some parametric model $\mathcal{F}$ with parameter space Θ then, if $f = f(\,\cdot\,; \theta)$ and $g = f(\,\cdot\,; \psi)$ for some $\theta,\, \psi \in \Theta$, we write

$$D(\theta,\, \psi) := D(f,\, g).$$

Definition 9.7 (Identifiability)**.** Let $\mathcal{F}$ be a parametric model with parameter space Θ. We say that $\mathcal{F}$ is *identifiable* if, for any two parameters $\theta,\, \psi \in \Theta$, if θ and ψ are distinct then $D(\theta,\, \psi) > 0$.

From now on we will assume that all parametric models are identifiable.

Theorem 9.8 (Consistency of the MLE)**.** *Let $\mathcal{F}$ be a sufficiently regular parametric model. The MLE is consistent.*

9.4. Equivariance of the MLE.

Theorem 9.9 (Equivariance of the MLE)**.** *Let $\mathcal{F}$ be a parametric model with parameter space Θ and parameter θ. Let $g : \Theta \to \mathbb{R}$ be a sufficiently regular function. If $\hat{\theta}_n$ denotes the MLE of θ then the MLE of $g(\theta)$ is $g(\hat{\theta}_n)$.*

9.5. Asymptotic Normality.

Definition 9.10 (Score function and Fisher information)**.** Let $\mathcal{F} := \{f(\,\cdot\,; \theta) : \theta \in \Theta\}$ be a parametric model. The *score function* is defined to be

$$s := \partial_\theta \log f(\,\cdot\,; \theta).$$

Now let $X_1, \ldots\ X_n$ be IID random variables with distribution $f(\,\cdot\,; \theta)$. The *Fisher information* is defined to be

$$I_n(\theta) := \mathbb{V}\left(\sum_{i=1}^n s\left(X_i; \theta\right)\right) = \sum_{i=1}^n \mathbb{V}\left(s\left(X_i; \theta\right)\right).$$

When $n = 1$ we write $I(\theta) := I_n(\theta)$. Note that the score function is a random variable but that the Fisher information is *not*, it is a map $I_n : \Theta \to \mathbb{R}$.

Theorem 9.11 (Alternate expression for the Fisher information)**.** *Consider a parametric model $\mathcal{F} := \{f(\,\cdot\,; \theta) : \theta \in \Theta\}$. The Fisher information satisfies*

$$I_n(\theta) = nI(\theta)$$

and moreover, if the score function has mean zero, then we have that

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log f(X; \theta)}{\partial \theta^2}\right]$$

for $X \sim f(\,\cdot\,; \theta)$.

Theorem 9.12 (Asymptotic Normality of the MLE)**.** *Consider a sufficiently regular parametric model $\mathcal{F} := \{f(\,\cdot\,; \theta) : \theta \in \Theta\}$ with MLE $\hat{\theta}_n$ and let $se := \sqrt{\mathbb{V}\hat{\theta}_n}$. The following hold.*

(1) $se \approx \sqrt{1/I_n(\theta)}$ *and*

$$\frac{\hat{\theta}_n - \theta}{se} \rightsquigarrow N(0,\, 1).$$

(2) *For* $\widehat{se} := \sqrt{1/I_n(\hat{\theta}_n)}$,

$$\frac{\hat{\theta}_n - \theta}{\widehat{se}} \rightsquigarrow N(0,\, 1).$$

Remark 9.13 (Intuition behind the asymptotic normality of the MLE)**.** Recall that l_n denotes the log-likelihood. Then, since the MLE maximizes the log-likelihood,

$$0 = l_n'(\hat{\theta}_n) \approx l_n'(\theta) + (\hat{\theta}_n - \theta) l_n''(\theta)$$

and so

$$\sqrt{n}(\hat{\theta}_n - \theta) \approx \frac{\frac{1}{\sqrt{n}} l_n'(\theta)}{-\frac{1}{n} l_n''(\theta)}.$$

Now observe the following.

- The numerator is precisely $\sqrt{n}$ times the sample mean of the (mean-free) score function, whose variance is, by definition, the Fisher information $I(\theta)$. So the numerator converges in distribution to $N(0,\, I(\theta))$.
- The denominator is the sample mean of $\frac{\partial^2 \log f}{\partial \theta^2}$, which therefore converges in probability to its expectation. By virtue of Theorem 9.11 this expectation is precisely the Fisher information $I(\theta)$.

So finally:

$$\sqrt{n}(\hat{\theta}_n - \theta) \approx \frac{\frac{1}{\sqrt{n}} l_n'(\theta)}{-\frac{1}{n} l_n''(\theta)} \rightsquigarrow \frac{N(0,\, I(\theta))}{I(\theta)} = N(0,\, 1/I(\theta)),$$

as desired.

Theorem 9.14 (Confidence intervals for the MLE)**.** *Let* $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ *be a sufficiently regular parametric model with MLE* $\hat{\theta}_n$*, let* $\widehat{se} := \sqrt{1/I_n(\hat{\theta}_n)}$*, and let, for any* $0 < \alpha < 1$*,* $z_{\alpha/2} := \Phi^{-1}(1-\alpha/2)$ *where* Φ *denotes the CDF of the standard Normal distribution, which means that* $\mathbb{P}(-z < Z < z) = 1 - \alpha$ *for* $Z \sim N(0,\, 1)$*. Let*

$$C_n := \left(\hat{\theta}_n - z_{\alpha/2}\widehat{se},\, \hat{\theta}_n + z_{\alpha/2}\widehat{se}\right).$$

C_n *is an asymptotic* $1 - \alpha$ *confidence interval for* θ*, meaning that*

$$\mathbb{P}(\theta \in C_n) \to 1 - \alpha \text{ as } n \to \infty.$$

9.6. **Optimality.**

Definition 9.15 (Asymptotic relative efficiency)**.** Let T_n and U_n, $n \geqslant 1$, be two sequences of random variables such that, for the same $\theta \in \mathbb{R}$, there exist $t,\, u > 0$ for which

$$\sqrt{n}(T_n - \theta) \rightsquigarrow N(0,\, t^2) \text{ and } \sqrt{n}(U_n - \theta) \rightsquigarrow N(0,\, u^2).$$

The *asymptotic relative efficiency* of U to T is defined to be

$$ARE(U,\, T) := \frac{t^2}{u^2}.$$

Remark 9.16 (Interpretation of the asymptotic relative efficiency)**.** Suppose that $ARE(U, T) \leqslant 1$. This is interpret as "the estimator U_n is using a smaller fraction of the data, quantified by the asymptotic relative efficiency, compared to the estimator T_n" (and that is why the estimator U_n has asymptotically more variance than the estimator T_n).

Theorem 9.17 (Asymptotic optimality of the MLE)**.** *Let $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a sufficiently regular parametric model with MLE $\hat{\theta}_n$. The MLE is* asymptotically optimal, *or* efficient, *meaning that*

$$ARE\left(\tilde{\theta}_n, \hat{\theta}_n\right) \leqslant 1$$

for any point estimator $\tilde{\theta}_n$ of the parameter θ.

9.7. The Delta Method.

Theorem 9.18 (The Delta Method)**.** *Let $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a sufficiently regular parametric model with MLE $\hat{\theta}_n$ and let $\widehat{se}(\hat{\theta}_n) := \sqrt{1/I_n(\hat{\theta}_n)}$. For any sufficiently regular $g : \Theta \to \mathbb{R}$, if we define $\tau := g(\theta)$ as well as*

$$\hat{\tau}_n := g(\hat{\theta}_n) \text{ and } \widehat{se}(\hat{\tau}_n) := |g'(\hat{\theta}_n)|\widehat{se}(\hat{\theta}_n),$$

then

$$\frac{\hat{\tau}_n - \tau}{\widehat{se}(\hat{\tau}_n)} \rightsquigarrow N(0, 1).$$

Moreover, for any $0 < \alpha < 1$, let $z_{\alpha/2} := \Phi^{-1}(1 - \alpha/2)$ where Φ denotes the CDF of the standard Normal distribution, which means that $\mathbb{P}(-z < Z < z) = 1 - \alpha$ for $Z \sim N(0, 1)$. Then

$$C_n := \left(\hat{\tau}_n - z_{\alpha/2}\widehat{se}(\hat{\tau}_n), \hat{\tau}_n + z_{\alpha/2}\widehat{se}(\hat{\tau}_n)\right)$$

is an asymptotic $1 - \alpha$ confidence interval for τ, meaning that

$$\mathbb{P}(\tau \in C_n) \to 1 - \alpha \text{ as } n \to \infty.$$

9.8. Multiparameter Models.

Definition 9.19 (Fisher information matrix)**.** Let $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a parametric model with parameter space $\Theta \subseteq \mathbb{R}^k$ and let $X_1, \ldots X_n$ be IID random variables with distribution $f(\,\cdot\,;\theta)$. The *Fisher information matrix* is the k-by-k matrix defined by

$$[I_n(\theta)]_{ij} := -\mathbb{E}\left[\frac{\partial^2 l_n}{\partial\theta_i\partial\theta_j}\right],$$

where l_n denotes the log-likelihood function associated with the distribution f and the random variables $X_1, \ldots X_n$.

Theorem 9.20 (Asymptotic normality of the MLE and delta method for multiparameter models)**.** *Let $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a sufficiently regular statistical model with parameter space $\Theta \subseteq \mathbb{R}^k$ and with MLE $\hat{\theta}_n$, and let J_n denote the inverse of the fisher information matrix. The following hold.*

(1) *Asymptotically, $\hat{\theta}_n - \theta \approx N(0, J_n)$.*

(2) Let $\hat{\theta}_{n,\,j}$ denote the j-th component of $\hat{\theta}_n$ and let $\widehat{se}_j^2 := (J_n)_{jj}$ denote the j-th diagonal element of J_n. Then

$$\frac{\hat{\theta}_{n,\,j} - \theta_j}{\widehat{se}_j} \rightsquigarrow N(0,\,1)$$

and, asymptotically,

$$\operatorname{Cov}(\hat{\theta}_{n,\,j},\,\hat{\theta}_{n,\,k}) \approx (J_n)_{jk}.$$

(3) Let $g : \Theta \to \mathbb{R}$ be a sufficiently regular function with $\nabla g|_{\hat{\theta}_n} \neq 0$ almost surely, let

$$\tau := g(\theta),\ \hat{\tau}_n := g(\hat{\theta}_n),\ \text{ and } \widehat{se}(\hat{\tau}_n) := \left.\sqrt{(\nabla g)^T J_n \nabla g}\right|_{\theta=\hat{\theta}_n}.$$

Then

$$\frac{\hat{\tau}_n - \tau}{\widehat{se}(\hat{\tau}_n)} \rightsquigarrow N(0,\,1).$$

9.9. The Parametric Bootstrap.

Remark 9.21 (Parametric bootstrap variance estimation)**.** Let $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a parametric model and let $\hat{\theta}_n := \hat{\theta}_n(X_1,\,\dots\,X_n)$ be a point estimator of θ. To estimate the variance of $\hat{\theta}_n$ using the *parametric* bootstrap method we may proceed as follows.

- Repeat B times, for $j = 1,\,\dots,\,B$, the following.
  (1) Draw $X^*_{1,\,j},\,\dots,\,X^*_{n,\,j} \sim f(\,\cdot\,;\hat{\theta}_n)$. In other words: resample assuming that the parameter estimate is the true parameter.
  (2) Compute $\hat{\theta}^*_{n,\,j} := \hat{\theta}_n(X^*_{1,\,j},\,\dots,\,X^*_{n,\,j})$.
- Finally compute

$$v_{boot} := \frac{1}{B}\sum_{j=1}^{B}\left(\hat{\theta}^*_{n,\,j} - \frac{1}{B}\sum_{k=1}^{B}\hat{\theta}^*_{n,\,k}\right)^2,$$

which is the population variance of the resample estimates $\hat{\theta}^*_{n,\,1},\,\dots,\,\hat{\theta}^*_{n,\,B}$.

Note that once we have an estimate for the variance we may then produce confidence intervals as described in Section 8.3.

10. Hypothesis Testing and p-values

Definition 10.1 (Hypothesis testing)**.** Let $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a parametric model and let Θ_0 and Θ_1 be disjoint non-empty subsets of the parameter space Θ. We call the statements

$$\theta \in \Theta_0 \text{ and } \theta \in \Theta_1,$$

denoted by H_0 and H_1 respectively, the *null hypothesis* and the *alternative hypothesis*, respectively. Let X denote a random variable (whose distribution depends on the parameter θ) and let R be a subset of the codomain of X. We call X the *test statistic* and R the *rejection region*.

Hypothesis testing proceeds as follows:

- if the test statistic X belongs to the rejection region R then we *reject* the null hypothesis H_0 and
- if the test statistic X does not belong to the rejection region R then we do not reject the null hypothesis H_0.

Keep in mind that " hypothesis testing is like a legal trial [...] we retain [the null hypothesis] H_0 unless there is strong evidence to reject [it] ".

We refer to the tuple $(\mathcal{F},\, \Theta_0,\, \Theta_1,\, H_0,\, H_1,\, X,\, R)$ as a *hypothesis test*.

Remark 10.2 (Interpretation of the null hypothesis)**.** In practice it is important to choose the null hypothesis to correspond to the *status quo*. This is because a hypothesis test can never prove that the null hypothesis holds true, it can only establish with sufficient evidence that the null hypothesis ought to be rejected. If the test fails to reject the null hypothesis, it does *not* mean that the null hypothesis is true!

Definition 10.3 (Error types)**.** Let $(\mathcal{F},\, \Theta_0,\, \Theta_1,\, H_0,\, H_1,\, X,\, R)$ be a hypothesis test. There are two types of errors.

(1) If $X \in R$ but H_0 is true, i.e. if we reject the null hypothesis H_0 even though H_0 is true, then we have a *type I error*.
(2) If $X \notin R$ but H_0 is false, i.e. if we do not reject the null hypothesis H_0 even though H_0 is false, then we have a *type II error*.

Remark 10.4 (Interpreting both error types)**.** If we follow the analogy of "hypothesis testing as a legal trial" again, then we can think of each error type as follows.

(1) A type I error corresponds to finding an innocent person guilty.
(2) A type II error corresponds to acquitting a guilty person.

Definition 10.5 (Power, size, and level)**.** Let $(\mathcal{F},\, \Theta_0,\, \Theta_1,\, H_0,\, H_1,\, X,\, R)$ be a hypothesis test where the parametric model $\mathcal{F}$ has parameter space Θ.

- The *power function* of this test is the function $\beta : \Theta \to [0,\,1]$ defined by
$$\beta(\theta) := \mathbb{P}_\theta(X \in R).$$
In other words $\beta(\theta)$ is the probability that the null hypothesis is rejected if θ is the true parameter.
- The *size* of this test is defined to be
$$\alpha := \sup_{\theta \in \Theta_0} \beta(\theta).$$
In other words the size is a sharp upper bound on the probability of a type I error.

- This test is said to have *level* α if its size is less than or equal to α.

Remark 10.6 (Interpretation of the size of a test)**.** Consider a hypothesis test $\mathcal{T} := (\mathcal{F}, \Theta_0, \Theta_1, H_0, H_1, X, R)$ with size α. We want to justify the comment above that " the size is a sharp upper bound on the probability of a type I error. "

So let β denote the power function of the test $\mathcal{T}$ and suppose that θ is the true parameter of this test. Note that if $\theta \notin \Theta_0$ then it is not possible to have a type I error (since the null hypothesis is false in such a case). Therefore

$$\begin{aligned}\mathbb{P}_\theta(\text{type I error}) &= \mathbb{P}_\theta(H_0 \text{ and } X \in R)\\ &= \begin{cases}\mathbb{P}_\theta(X \in R) & \text{if } \theta \in \Theta_0 \text{ and}\\ 0 & \text{if } \theta \notin \Theta_0\end{cases}\\ &\leqslant \sup_{\Theta_0} \beta =: \alpha.\end{aligned}$$

Definition 10.7 (Types of hypotheses and test)**.** Consider the hypothesis test $\mathcal{T} := (\mathcal{F}, \Theta_0, \Theta_1, H_0, H_1, X, R)$ whose parametric model $\mathcal{F}$ has parameter space denoted by Θ.

- If there exists a parameter $\theta_0 \in \Theta$ such that $\Theta_0 = \{\theta_0\}$ then we call H_0 a *simple* hypothesis. In this case H_0 is the statement $\theta = \theta_0$.
- If a hypothesis (null or alternative) is not simple then we call it a *composite* hypothesis.
- If the null hypothesis is simple and the alternative hypothesis is its negation, meaning that Θ_1 is the complement of Θ_0, then we call $\mathcal{T}$ a *two-sided* test. In this case H_0 and H_1 are the statements
$$\theta = \theta_0 \text{ and } \theta \neq \theta_0,$$
respectively, for some $\theta_0 \in \Theta$.
- If $\Theta_0 = (-\infty, \theta_0]$ or $[\theta_0, +\infty)$ for some $\theta_0 \in \Theta$ and if Θ_1 is the complement of Θ_0 then we call $\mathcal{T}$ a *one-sided* test. For example, if $\Theta_0 = (-\infty, \theta_0]$ then H_0 and H_1 are the statements
$$\theta \leqslant \theta_0 \text{ and } \theta > \theta_0,$$
respectively.

Remark 10.8 (Most powerful test)**.** The ideal scenario would be to fix a level α and to then seek, among all tests of level α, that which maximizes the power under the alternative hypothesis.

In other words we seek to solve the *maximin* problem

$$\sup_{\mathcal{T} \,:\, \mathsf{size}(\mathcal{T}) \leqslant \alpha} \inf_{\Theta_1} \beta.$$

This way we guarantee that the probability of type I errors is bounded above (see Remark 10.6) while also ensuring that the the worst-case probability of rejection under the alternative hypothesis (quantified by $\inf_{\Theta_1} \beta$) is as large as possible, thus minimizing the probability of type II errors. Remember: under the null hypothesis we want the power to be uniformly small, but under the alternative hypothesis we want the power to be uniformly large.

Such tests are known as *most powerful* tests, but are often difficult to identify in practice (when they even exist at all!)

10.1. The Wald Test.

Definition 10.9 (The Wald test)**.** Let $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a parametric model, let $\hat{\theta}_n$ be a point estimator of θ, let $\theta_0 \in \Theta$ be the true parameter, and let $\widehat{se}_n$ be an estimate of the standard error of $\hat{\theta}_n$ such that

$$\frac{\hat{\theta}_n - \theta_0}{\widehat{se}_n} \rightsquigarrow N(0,\, 1).$$

This asymptotic normality is not guaranteed: that it occurs is the major assumption underpinning the Wald test.

Let $\alpha \in (0,\, 1)$. The size α Wald test is constructed as follows.

- $\Theta_0 := \{\theta_0\}$ and $\Theta_1 := \Theta \setminus \Theta_0$ such that the null and alternative hypotheses are, respectively,
$$H_0 : \theta = \theta_0 \text{ and } H_1 : \theta \neq \theta_0.$$
- The test statistic is the random variable
$$W_n := \frac{\hat{\theta}_n - \theta_0}{\widehat{se}_n}.$$
- The rejection region is
$$R := \left\{ W \in \mathbb{R} : |W| > z_{\alpha/2} \right\}$$
where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ for Φ denoting the CDF of the standard normal distribution.

The tuple $(\mathcal{F},\, \Theta_0,\, \Theta_1,\, H_0,\, H_1,\, W_n,\, R)$ is called a *size α Wald test.*

Theorem 10.10 (Properties of Wald tests)**.** *Let $\alpha \in (0,\, 1)$ and consider the size α Wald test $(\mathcal{F},\, \Theta_0,\, \Theta_1,\, H_0,\, H_1,\, W_n,\, R)$.*

1. *Asymptotically, the Wald test has size α, i.e.*
$$\mathbb{P}_{\theta_0}\left(|W_n| > z_{\alpha/2}\right) \to \alpha \text{ as } n \to \infty,$$
where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ for Φ denoting the CDF of the standard normal distribution.
2. *When the null hypothesis is false the power (i.e. the probability of correctly rejecting the null hypothesis) is, approximately,*
$$\beta(\theta) = 1 - \Phi\left(\frac{\theta_0 - \theta}{\widehat{se}_n} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta}{\widehat{se}_n} - z_{\alpha/2}\right)$$
for all $\theta \in \Theta_1$, where θ_0 denotes the true parameter for which $\Theta_0 = \{\theta_0\}$, and $\widehat{se}_n$ is the estimate of the standard error appearing in the Wald test statistic W_n.

Remark 10.11 (How to approximate the standard error)**.** We note that Theorem 10.10 above does not tell us *how* to approximate the standard error. So how would we go about doing that? We have discussed four approaches so far.

1. Using the plug-in estimator, if the standard error can be written analytically as a function of the CDF.
2. Using the (nonparametric) bootstrap.
3. Using the Fisher information and the asymptotic normality of the MLE, if the estimator is an MLE (or a function thereof, in which case the delta method will help).

(4) Using the parametric bootstrap.

Theorem 10.12 (Wald test and confidence intervals)**.** *Let $\alpha \in (0,\, 1)$ and consider the size α Wald test $(\mathcal{F},\, \Theta_0,\, \Theta_1,\, H_0,\, H_1,\, W_n,\, R)$ where*

$$W_n = \frac{\hat{\theta}_n - \theta_0}{\widehat{se}_n} \text{ and } z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$$

for $\hat{\theta}_n$, θ_0, $\widehat{se}_n$, $z_{\alpha/2}$, and Φ as in Definition 10.9. The null hypothesis H_0 is rejected if and only if θ_0 is outside the confidence interval

$$C_n := \left(\hat{\theta}_n - \widehat{se}_n z_{\alpha/2},\, \hat{\theta}_n + \widehat{se}_n z_{\alpha/2}\right).$$

Remark 10.13 (Wald test and confidence intervals)**.** Theorem 10.12 above tells us that performing a Wald test is equivalent to checking whether or not the *null value* (i.e. the parameter characterizing a simple null hypothesis) belongs to a confidence interval.

Nonetheless, for practical purposes it is often more useful to construct a confidence interval than to perform a Wald test. This is because the confidence interval yields more information: it tells us *how far* the null value is from the confidence interval (relative to the length of the interval), which is more fine-grained information than the mere binary truth value used to reject, or not, the null hypothesis.

10.2. p–value.

Definition 10.14 (p–value)**.** Suppose that, for $\alpha \in (0,\, 1)$, we have a family of hypothesis tests $\mathcal{T}_\alpha$ with size α, rejection region R_α, and test statistic $T_{n,\alpha}$. The *p–value* is defined to be

$$p - \text{value} := \inf\left\{\alpha \in (0,\, 1) : T_{n,\alpha} \in R_\alpha\right\}.$$

In other words the p–value is the smallest level at which we can reject the null hypothesis. In practice the test statistics are often independent of α, i.e. $T_{n,\alpha} = T_n$. Finally, keep in mind that p–values are *random variables*.

Remark 10.15 (Evidence scale)**.** As a general rule, here is how various p–values are interpreted.

p–value	Evidence against the null hypothesis
< 0.01	Very strong evidence
$0.01 - 0.05$	Strong evidence
$0.05 - 0.1$	Weak evidence
> 0.1	Little or no evidence

Theorem 10.16 (Alternate expression for p–values)**.** *Suppose that for $\alpha \in (0,\, 1)$ we have a family of hypothesis tests $\mathcal{T}_\alpha$ with size α, test statistic T_n* independent of *α, and rejection region R_α given by*

$$R_\alpha = \{t \in \mathbb{R} : t \geqslant c(\alpha)\}$$

for some strictly decreasing *function $c : (0,\, 1) \to \mathbb{R}$. The p–value may be written as follows: for any outcome ω,*

$$p - \textit{value}(\omega) = \sup_{\theta_* \in \Theta_0} \mathbb{P}_{\theta_*}\left(T_n^* \geqslant T_n(\omega)\right),$$

where Θ_0 is the subset of the parameter space corresponding to the null hypotheses of $\mathcal{T}_\alpha$ (which are assumed to be independent of *α). Here the notation θ_* and T_n^* is used to indicate that T_n^* is another independent instance of the test statistic with true parameter θ_*.*

In particular if $\Theta_0 = \{\theta_0\}$ for some $\theta_0 \in \mathbb{R}$ then, for any outcome ω,

$$p - value(\omega) = \mathbb{P}_{\theta_0}\left(T_n^{(0)} \geqslant T_n(\omega)\right),$$

where now $T_n^{(0)}$ is another independent *instance of the test statistic with true parameter θ_0. In other words: the p–value is the probability, under the null hypothesis, to obtain an equally or more extreme value of the test statistic.*

Theorem 10.17 (p–value of the Wald test)**.** *Let W_n denote the test statistic used in the Wald test (see Definition 10.9). The p–value of the Wald test is*

$$p - value = \mathbb{P}(|Z| > |W_n|)$$

or

$$p - value = 2\Phi(-|W_n|)$$

where $Z \sim N(0,\,1)$, where the probability is computed under the assumption that the null hypothesis is correct, and where Φ denotes the CDF of Z.

Theorem 10.18 (Distribution of the p–value under the null hypothesis)**.** *Suppose that for $\alpha \in (0,\,1)$ we have a family of hypothesis tests $\mathcal{T}_\alpha$ as in Theorem 10.16. Suppose moreover that the null hypothesis (which is* independent *of α) is simple and that the test statistics have nowhere vanishing continuous distributions. Under the null hypothesis, i.e. provided that the null hypothesis is true, the p–value has a Uniform$(0,1)$ distribution.*

Therefore, if we reject the null hypothesis when the p–value is less than $\alpha \in (0,\,1)$, the probability of a type I error is α.

10.3. Pearson's χ^2 Test for Multinomial Data.

Definition 10.19 (χ^2 distribution quantile)**.** Recall that the χ^2 distribution is introduced in Definition 2.21. This distribution arises naturally as follows: if $Z_1, \ldots\, Z_k \sim N(0,\,1)$ then $V := \sum_{i=1}^k Z_i^2 \sim \chi_k^2$. For $\alpha \in (0,\,1)$ we define the upper α quantile by $\chi^2_{k,\,\alpha} := F^{-1}(1-\alpha)$ where F is the CDF of a χ^2_k distribution, i.e. $\mathbb{P}(\chi_k^2 > \chi^2_{k,\,\alpha}) = \alpha$.

Definition 10.20 (Simplex)**.** For any $k \geqslant 1$ let Δ^{k-1} denote the $(k-1)$–dimensional simplex, i.e.

$$\Delta^{k-1} := \left\{ p \in [0,\,1]^k : \sum_{i=1}^k p_i = 1 \text{ and } p_i \geqslant 0 \text{ for all } i \right\} \subseteq [0,\,1]^k.$$

This is the set where the parameter p of a Multinomial$(n,\,p)$ distribution lives.

Definition 10.21 (Pearson's χ^2 statistic)**.** Let

$$X = (X_1,\,\ldots\,,\,X_k) \sim \text{Multinomial}(n,\,p),$$

where $p \in \Delta^{k-1}$ and note that X is implicitly dependent on n. Consider $p_0 \in \Delta^{k-1}$ and let

$$E_j := np_{0,\,j}$$

such that, under the assumption that p_0 is the true parameter, $E_j = \mathbb{E}X_j$. *Pearson's* χ^2 *statistic with true parameter* p_0 is defined to be

$$T := T_{n,\,k} := \sum_{i=1}^{k} \frac{(X_i - E_j)^2}{E_j}.$$

Remark 10.22 (Multinomial distributions and simplices)**.** Since the parameter p of a multinomial distribution belongs to the $(k-1)$–simplex $\Delta^{k-1} \subseteq [0,\,1]^k$, keeping track of the dimension k is important since it informs the number of degrees of freedom of the limiting distribution of Pearson's χ^2 statistic. Unsurprisingly, since the $(k-1)$–simplex is $(k-1)$–dimensional, the limiting distribution will have $k-1$ degrees of freedom (see Theorem 10.24 below).

Definition 10.23 (Pearson's χ^2 test)**.** Let $\mathcal{F}$ be the parametric model consisting of the Multinomial$(n,\,p)$ distributions where $p \in \Delta^{k-1}$, let $p_0 \in \Delta^{k-1}$ be fixed, and let $\alpha \in (0,\,1)$. The *size* α *Pearson's* χ^2 *test* is a hypothesis test constructed as follows.

- Let
$$H_0 : p = p_0 \text{ and } H_1 : p \neq p_0$$
be the null and alternative hypotheses, respectively, and let $\Theta := \{p_0\}$ and $\Theta_1 := \Delta^{k-1} \setminus \{p_0\}$ be the corresponding subsets of parameter space.
- Let Pearson's χ^2 statistic with true parameter p_0, denoted by T, be the test statistic.
- Let the rejection region be
$$R_\alpha := \left\{t \in \mathbb{R} : t > \chi^2_{k-1,\,\alpha}\right\}.$$

The hypothesis test $(\mathcal{F},\,\Theta_0,\,\Theta_1,\,H_0,\,H_1,\,T,\,R_\alpha)$ is called a *size* α *Pearson's* χ^2 *test.*

Theorem 10.24 (Limiting behaviour of Pearson's χ^2 statistic)**.** *Consider Pearson's* χ^2 *test with test statistic* T *and true parameter* $p_0 \in \Delta^{k-1}$*. Under its null hypothesis,* $T \rightsquigarrow \chi^2_{k-1}$ *as* $n \to \infty$*.*

Corollary 10.25 (Asymptotic size of Pearson's χ^2 test)**.** *For any* $\alpha \in (0,\,1)$ *the size* α *Pearson's* χ^2 *test has asymptotic size* α *as* $n \to \infty$*. Moreover the p–value satisfies*

$$p - value(\omega) = \mathbb{P}\left(\chi^2_{k-1} > T(\omega)\right)$$

for any outcome ω*, where* T *denotes the test statistic and where the true parameter lies in* Δ^{k-1}*.*

10.4. **Permutation Test.**

Definition 10.26 (Permutation)**.** An invertible map σ from $\{1,\,\ldots,\,N\}$ to itself is called a *permutation.* The set of all such permutations is denoted by S_N.

Definition 10.27 (Permutation distribution)**.** Let $T(Z_1,\,\ldots,\,Z_N)$ be a statistic. The *permutation distribution* of T, denoted $\mathcal{P}(T)$, is the discrete uniform distribution on

$$\left\{T\left(Z_{\sigma(1)},\,\ldots,\,Z_{\sigma(N)}\right) : \sigma \in S_N\right\}.$$

Remark 10.28 (Sources of randomness in the permutation distribution)**.** There are two sources of randomness in the permutation distribution.

(1) The first source of randomness is the randomness inherent to the Z_i's.

(2) For each outcome ω, once $Z_1(\omega), \ldots, Z_N(\omega)$ are *fixed*, there is another layer of randomness. Indeed: there are now $N!$ permutations of $Z_1(\omega), \ldots, Z_N(\omega)$ which are equally likely according to the permutation distribution. This is the second source of randomness.

In other words we are defining $\mathcal{T} \sim \mathcal{P}(T)$ *conditionally* via

$$\mathbb{E}\left(\mathcal{T} \,|\, Z_1 = z_1, \ldots, Z_N = z_N\right) \sim \text{Uniform}\left(\left\{T\left(z_{\sigma(1)}, \ldots, z_{\sigma(N)}\right) : \sigma \in S_N\right\}\right).$$

Definition 10.29 (Permutation test)**.** Let F_X and F_Y be two CDFs. Given $X_1, \ldots, X_m \sim F_X$ and $Y_1, \ldots, Y_n \sim F_Y$ consider the statistic

$$T\left(X_1, \ldots, X_m, Y_1, \ldots, Y_n\right) := \left|\overline{X}_m - \overline{Y}_n\right| = \left|\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{n}\sum_{j=1}^{n} Y_j\right|.$$

We construct the *permutation test* as follows.

- Let the statements
$$H_0 : F_X = F_Y \text{ and } H_1 : F_X \neq F_Y$$
be the null and alternative hypothesis, respectively.
- Let $\mathcal{T} \sim \mathcal{P}(T)$, whose distribution is the permutation distribution of T, and let
$$t := \frac{1}{N!}\sum_{\sigma \in S_N} I\left(\mathcal{T}_\sigma > T\right)$$
be the test statistic.
- For any $\alpha \in (0, 1)$ let
$$R_\alpha := \{t \in \mathbb{R} : t < \alpha\}$$
be the rejection region.

The tuple (H_0, H_1, t, R_α) is called the *permutation test.*

Theorem 10.30 (p–value of the permutation test)**.** *The p–value of the permutation test* (H_0, H_1, t, R_α) *is precisely its test statistic* t.

Remark 10.31 (Permutation tests)**.** (1) By contrast with Definition 10.1 which introduces hyphothesis tests, the definition of a permutation test does not involve a parametric model or any subsets of the parameter space. This is because permutation tests are *nonparametric.*

(2) By contrast with other tests described in this chapter, such as the Wald test or Pearson's χ^2 test, the permutation test does not rely on the number of samples being sufficiently large: it is an *exact* test (and not an *asymptotic* one).

This means that permutation tests are particularly well-suited to settings with small sample sizes. Actually, in practice, pertmutation tests behave similarly to asymptotic tests when the sample size becomes large.

(3) Even when n and m are small, the *factorial* of $N := n + m$ may be very large. In practice it may thus not be feasible to compute the test statistic t exactly. Instead we may rely on *simulation* and use

$$t^* := \frac{1}{|S|}\sum_{\sigma \in S} I(\mathcal{T}_\sigma > T).$$

for some subset $S \subsetneq S_N$.

(4) We can build permutation tests on other statistics besides the sample mean. We can for example use the *empirical median* (i.e. the plug-in estimator for the median), the sample variance, etc. For example using the empirical median would mean using

$$T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) := \left| F_{X,m}^{-1}(1/2) - F_{Y,n}^{-1}(1/2) \right| .$$

10.5. The Likelihood Ratio Test.

Definition 10.32 (Likelihood ratio test)**.** Let $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a parametric model and let Θ_0 be a subset of the parameter space $\Theta \subseteq \mathbb{R}^r$ such that, for some $1 \leqslant q < r$ and some $\theta_0 \in \mathbb{R}^{r-q}$,

$$\Theta_0 := \{\theta \in \Theta : (\theta_{q+1}, \ldots, \theta_r) = (\theta_{0,q+1}, \ldots, \theta_{0,r})\} .$$

Note that $r-q = \dim\Theta - \dim\Theta_0$. The *likelihood ratio test* is constructed as follows.

- Let

$$H_0 : \theta \in \Theta_0 \text{ and } H_1 : \theta \notin \Theta_0$$

  be the null and alternative hypotheses, respectively.
- Let $\mathcal{L}_n$ denote the likelihood function associated with $\mathcal{F}$ and let

$$\lambda_n := 2\log\left(\frac{\sup_\Theta \mathcal{L}_n}{\sup_{\Theta_0}\mathcal{L}_n}\right) = 2\log\left(\frac{\mathcal{L}_n(\hat\theta_n)}{\mathcal{L}_n(\hat\theta_{0,n})}\right),$$

  where $\hat\theta_n$ and $\hat\theta_{0,n}$ denote the MLE over Θ and Θ_0 respectively, be the test statistic, which is called the *likelihood ratio statistic*.
- For $\alpha \in (0,\,1)$ let

$$R_\alpha := \left\{\lambda \in \mathbb{R} : \lambda > \chi^2_{r-q,\,\alpha}\right\}$$

  be the rejection region.

The hypothesis test $(\mathcal{F}, \Theta_0, \Theta_0^c, H_0, H_1, \lambda_n, R_\alpha)$ is called a *likelihood ratio test.*

Theorem 10.33 (Limiting behaviour of the likelihood ratio test)**.** *Using the notation of Definition 10.32, under the null hypothesis,*

$$\lambda_n \rightsquigarrow \chi^2_{r-q} \text{ as } n \to \infty$$

and so the p–value satisfies, asymptotically,

$$p - value \approx \mathbb{P}\left(\chi^2_{r-q} > \lambda_n\right) \text{ as } n \to \infty.$$

Remark 10.34 (Likelihood ratio test)**.** (1) By contrast with the Wald test, which is useful when testing *scalar* parameters, the likelihood ratio test is useful when testing *vector* parameters.

(2) As Wasserman puts it: "You might have expected to see the maximum of the likelihood over Θ_0^c instead of Θ in the numerator. In practice, replacing Θ_0^c with Θ has little effect on the test statistic. Moreover, the theoretical properties of λ are much simpler if the test statistic is defined this way."

10.6. **Multiple Testing.**

Remark 10.35 (Multiple testing problem)**.** Suppose we perform m independent hypothesis tests of size α. The chance that there is at least one *false* rejection, i.e. at least one type I error, satisfies

$$\begin{aligned}\mathbb{P}(\text{at least one false rejection}) &= 1 - \mathbb{P}(\text{no false rejections}) \\ &= 1 - \mathbb{P}(\text{no false rejection in one test})^m \\ &= 1 - [1 - \mathbb{P}(\text{type I error in one test})]^m \\ &= 1 - (1-\alpha)^m \to 1 \text{ as } m \to \infty.\end{aligned}$$

This means that as m becomes large we become guaranteed to make erroneous rejections. We therefore need to make it more difficult to reject each individual *null hypothesis*.

Definition 10.36 (Bonferroni method)**.** Consider m hypothesis tests with null hypotheses $H_{0,1}, \ldots, H_{0,m}$ and p–values $P_1, \ldots, P_m$ and fix $\alpha \in (0, 1)$. The *Bonferroni method* rejects the null hypothesis $H_{0,i}$ if

$$P_i < \frac{\alpha}{m}.$$

Theorem 10.37 (Bonferroni method)**.** *Using the Bonferroni method as in Definition 10.36, the probability of falsely rejecting any null hypothesis is less than or equal to α.*

Remark 10.38 (Conservatism of the Bonferroni method)**.** The Bonferroni method is *very* conservative because it attempts to avoid even *one* false rejection of a null hypothesis.

Definition 10.39 (Benjamini-Hochberg method)**.** Consider m hypothesis tests with null hypotheses $H_{0,1}, \ldots, H_{0,m}$ and p–values $P_1, \ldots, P_m$ and fix $\alpha \in (0, 1)$.

(1) Let $P_{(1)} \leqslant \cdots \leqslant P_{(m)}$ denote the *ordered* p–values.

(2) Define, for $i = 1, \ldots, m$,

$$l_i := \frac{i\alpha}{C_m m} \text{ and } R := \max\left\{i : P_{(i)} < l_i\right\},$$

where

$$C_m := \begin{cases} 1 & \text{if the } p\text{–values are independent and} \\ \displaystyle\sum_{i=1}^{n} \frac{1}{i} & \text{otherwise.}\end{cases}$$

(3) Define $T := P_{(R)}$ to be the *BH rejection treshold.*

The *Benjamini-Hochberg method*, or *BH method*, rejects the null hypothesis $H_{0,i}$ if

$$P_{(i)} \leqslant T.$$

Definition 10.40 (False discovery rate and proportion)**.** Consider m hypothesis tests with null hypotheses $H_{0,1}, \ldots, H_{0,m}$. For a fixed outcome ω define the *false discovery proportion*, denoted FDP, to be

$$FDP := \begin{cases} \dfrac{\#\text{ falsely rejected null hyp.}}{\#\text{ rejected null hyp.}} & \text{if at least one null hyp. is true and} \\ 0 & \text{otherwise.}\end{cases}$$

The *false discovery rate*, denoted FDR, is defined as

$$FDR := \mathbb{E}\left[FDP\right].$$

Theorem 10.41 (BH method)**.** *Using the BH method as in Definition 10.39, regardless of how many null hypotheses are true, and regardless of the distribution of the p–values when the null hypotheses are false,*

$$FDR \leqslant \frac{\#\ \textit{true null hypotheses}}{\#\ \textit{null hypotheses}} \cdot \alpha \leqslant \alpha.$$

10.7. **Goodness-of-fit Tests.** The theory behind the goodness-of-fit test presented here is discussed in Section 30.3 of [Cra99], pp. 424-434. It is then discussed in Example 2, page 437, how to use this method for a goodness-of-fit test for a Normal model.

Definition 10.42 (Neyman-Pearson χ^2 test)**.** Let $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a statistical model, let $X_1, \ldots, X_n \sim f(\,\cdot\,;\theta)$ for some parameter θ, and let $I_1, \ldots, I_k$ be a partition, i.e. a pairwise disjoint cover, of the real line. For $j = 1, \ldots, k$, let

$$p_j(\theta) := \int_{I_j} f(x;\theta)dx,$$

let

$$N_j := \#\text{ observations in } I_j = |\{1 \leqslant i \leqslant k : X_i \in I_j\}| = \int_{I_j} dF_n(x),$$

and let $\tilde{\theta}_n$ denote the maximizer of the multinomial likelihood

$$Q(\theta) := \prod_{j=1}^{k} p_j(\theta)^{N_j}.$$

The *Neyman-Pearson* χ^2 *test* is constructed as follows.

- Let
$$H_0 : X_1, \ldots, X_n \sim f(x;\theta) \text{ for some } \theta \in \Theta$$
be the null hypothesis and let its negation be the alternative hypothesis.
- Let
$$Q_n := \sum_{j=1}^{k} \frac{\left[N_j - E_j(\tilde{\theta}_n)\right]^2}{E_j(\tilde{\theta}_n)}$$
be the test statistic, where
$$E_j(\tilde{\theta}) := np_j(\tilde{\theta}_n).$$
- Let s be the dimension of the parameter space Θ. For $\alpha \in (0,\,1)$ let
$$R_\alpha := \left\{q \in \mathbb{R} : q > \chi^2_{k-1-s,\,\alpha}\right\}$$
be the *rejection region.*

The tuple $(\mathcal{F}, H_0, H_1, Q_n, R_\alpha)$ is called the *Neyman-Pearson* χ^2 *test*, or *NP* χ^2 *test.*

Theorem 10.43 (Limiting behaviour of the NP χ^2 test)**.** *Using the notation of Definition 10.42, under the null hypothesis,*

$$Q_n \rightsquigarrow \chi^2_{k-1-s} \text{ as } n \to \infty$$

and so the p–value satisfies, asymptotically,

$$p - value \approx \mathbb{P}(\chi^2_{k-1-s} > Q_n) \text{ as } n \to \infty.$$

Remark 10.44 (Limitations of goodness-of-fit tests)**.** Goodness-of-fit tests such as the NP χ^2 test can never guarantee that an IID sample comes from a given parametric model. Indeed: if the NP χ^2 test does *not* reject the null hypothesis then it does *not* mean that the parametric model is correct.

11. Bayesian Inference

11.1. The Bayesian Method.

Definition 11.1 (Bayesian method)**.** Let $\mathcal{F}$ be a statistical model with parameter θ.

- We choose a probability density $f(\theta)$, called the *prior distribution*, over the parameter space of $\mathcal{F}$.
- We observe an IID sample $X_1, \ldots, X_n$ from $\mathcal{F}$.
- We compute the *posterior distribution* $f(\theta \mid X_1, \ldots, X_n)$ to be proportional to
$$\mathcal{L}_n(\theta) f(\theta)$$
where $\mathcal{L}_n$ denotes the likelihood function (which depends on $X_1, \ldots, X_n$).
  Note that the posterior distribution must be a probability distribution and so proportionality fully characterizes it since it must integrate to one.

Remark 11.2 (Bayesian philosophy)**.** In the Bayesian method the prior distribution is only a *subjective degree of belief* (and *not* a limiting frequency or objective property of some underlying process).

Definition 11.3 (Posterior confidence interval)**.** Let $f(\theta \mid X_1, \ldots, X_n)$ be a posterior distribution. Let $\alpha \in (0, 1)$ and let $a, b \in R$ such that

$$\int_{-\infty}^{a} f(\theta \mid X_1, \ldots, X_n)\, d\theta = \int_{b}^{+\infty} f(\theta \mid X_1, \ldots, X_n)\, d\theta = \frac{\alpha}{2}.$$

Then $C := (a, b)$ is called a $1 - \alpha$ *posterior confidence interval.*

Definition 11.4 (Conjugate prior)**.** Let $\mathcal{F}$ be a statistical model with parameter θ and let $f(\theta)$ be a prior distribution. If, for every IID sample $X_1, \ldots, X_n$ from $\mathcal{F}$, there exists a parameter $\tilde{\theta}$ such that the posterior distribution is equal to the prior evaluated at $\tilde{\theta}$, i.e.

$$f(\theta \mid X_1, \ldots, X_n) = f(\tilde{\theta})$$

then we say that the prior is *conjugate* with respect to the model.

11.2. Functions of parameters.

Remark 11.5 (Posterior distribution for functions of parameters)**.** Given a parameter θ with posterior distribution $f(\theta \mid X_1, \ldots, X_n)$ and given $\tau = g(\theta)$ for some function g we may follow Remark 2.35 (which tells us how to compute the PDF of a transformed random variable) to compute the posterior distribution of τ.

(1) For each τ we find the set $A_\tau = \{\theta : g(\theta) \leqslant \tau\}$.
(2) We find the posterior CDF via
$$\begin{aligned} H(\tau \mid X_1, \ldots, X_n) &= \mathbb{P}(g(\theta) \leqslant \tau \mid X_1, \ldots, X_n) \\ &= \mathbb{P}(\theta \in A_\tau \mid X_1, \ldots, X_n) \\ &= \int_{A_\tau} f(\theta \mid X_1, \ldots, X_n)\, d\theta. \end{aligned}$$

(3) Provided that the posterior CDF of τ is sufficiently regular we obtain the posterior PDF by differentiating:
$$h(\tau \mid X_1, \ldots, X_n) = H'(\tau \mid X_1, \ldots, X_n).$$

11.3. Large Sample Properties of Bayes' Procedure.

Theorem 11.6 (Asymptotic behaviour of posterior distributions)**.** *Let $\mathcal{F}$ be a parametric model with parameter θ and let $f(\theta)$ be a prior distribution such that both $\mathcal{F}$ and f are sufficiently regular. Let $\hat{\theta}_n$ denote the MLE and let $\widehat{se}_n := \sqrt{1/nI(\hat{\theta}_n)}$ where I denotes the Fisher information. The following hold.*

(1) The posterior distribution is asympotically Normal, as $n \to \infty$, with mean $\hat{\theta}_n$ and standard deviation $\widehat{se}_n$.
(2) The posterior mean*, i.e. the mean of the posterior distribution, is asymptotically equal to the MLE.*
(3) The asymptotic frequentist $1-\alpha$ confidence interval

$$C_n := \left(\hat{\theta}_n - z_{\alpha/2}\widehat{se}_n,\, \hat{\theta}_n + z_{\alpha/2}\widehat{se}_n\right),$$

for $z_{\alpha/2} := \Phi^{-1}(1-\alpha/2)$ where Φ denotes the CDF of a standard normal distribution, is also, asymptotically as $n \to \infty$, an approximate $1-\alpha$ Bayesian posterior interval.

11.4. Flat Priors and Improper Priors.

Definition 11.7 (Flat and improper priors)**.** Let $\mathcal{F}$ be a parametric model with parameter θ.

- A *flat prior* is any prior distribution equal to a constant.
- An *improper prior* is any prior distribution which is not integrable over the parameter space.

Remark 11.8 (Using improper priors)**.** Generally speaking, using improper priors (such as flat priors on parameter spaces with infinite measure) is not an issue *as long as* the resulting posterior distribution is an honest–to–goodness distribution.

Definition 11.9 (Jeffrey's prior)**.** Let $\mathcal{F}$ be a parametric model with Fisher information I. *Jeffreys' prior* is the prior distribution proportional to $\sqrt{I}$, or to $\sqrt{\det(I)}$ for multiparameter models (where now I denotes the Fisher information matrix).

Remark 11.10 (Appeal of the Jeffreys' prior)**.** The parameter space Θ of a parametric model $\mathcal{F}$ may be *reparametrized* via any bijection $\Gamma : \Theta \to \Theta'$. We may then view Θ' as a *new* parameter space for $\mathcal{F}$. The beauty of the Jeffreys' prior is that it provides a recipe for constructing a prior distribution which is, in some appropriate sense, *equivariant* under reparametrization.

11.5. Multiparameter Problems.

Remark 11.11 (Computing marginal posterior distributions)**.** Computing posterior distributions in the multiparameter setting where parameters take the form $\theta = (\theta_1, \dots, \theta_k)$ is done as in the single-parameter case. Nonetheless a wrinkle may arise when we seek to compute marginal posterior distributions: given the joint posterior distribution

$$f(\theta_1, \dots, \theta_k \,|\, X_1, \dots, X_n)$$

the i–th *marginal* posterior distribution is

$$f(\theta_i \,|\, X_1, \dots, X_n) = \int f(\theta_1, \dots, \theta_k \,|\, X_1, \dots, X_n)\, d\theta_1 \dots d\theta_{i-1} d\theta_{i+1} \dots d\theta_k.$$

Even if the joint posterior distribution is known analytically, evaluating the $(k-1)$–fold integral above may prove very challenging. Once again, simulation comes to the rescue: we may simulate $f\,(\theta_i \,|\, X_1,\,\ldots,\,X_n)$ by drawing an IID sample $\theta^{(1)},\,\ldots,\,\theta^{(B)}$ from the *joint* posterior distribution and then use the empirical distribution of

$$\theta_i^{(1)},\,\ldots,\,\theta_i^{(B)}$$

as an approximation of $f\,(\theta_i \,|\, X_1,\,\ldots,\,X_n)$.

11.6. **Bayesian Testing.**

Remark 11.12 (Bayesian hypothesis testing)**.** In *Bayesian hypothesis testing* we must specify *two* prior distributions: one for the parameter θ and one for the discrete set $\{H_0,\,H_1\}$ where H_0 denotes the null hypothesis and H_1 denotes the alternative hypothesis. If the null hypothesis is simple and the alternative hypothesis is its negation, i.e. we are testing

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0$$

for some parameter θ_0, and if we choose the (agnostic) prior $\mathbb{P}(H_0) = \mathbb{P}(H_1) = \frac{1}{2}$ over the hypotheses, then Bayes' Theorem tells us that

$$\mathbb{P}\,(H_0 \,|\, X_1,\,\ldots,\,X_n) = \frac{\mathcal{L}_n(\theta_0)}{\mathcal{L}_n(\theta_0) + \int \mathcal{L}_n(\theta) f(\theta) d\theta}$$

where $f(\theta)$ denotes the prior distribution for θ. Note that here improper priors *cannot* be used otherwise the denominator above will faill to be finite and the resulting Bayesian hypothesis test will be ill-defined.

12. Statistical Decision Theory

12.1. Preliminaries.

Remark 12.1 (What is statistical decision theory?)**.** *Statistical decision theory* is a formal theory for comparing statistical procedures. For example statistical decision theory may be used to choose rationally among several point estimators of the same parameter.

Definition 12.2 (Loss function)**.** Let $\mathcal{F}$ be a parametric model with parameter space Θ. A positive-definite function $L : \Theta \times \Theta \to \mathbb{R}$, which means that

$$L(\theta_1, \, \theta_2) \geqslant 0$$

and

$$L(\theta_1, \, \theta_2) = 0 \iff \theta_1 = \theta_2$$

for all $\theta_1, \, \theta_2 \in \Theta$ is called a *loss function*.

Definition 12.3 (Important loss functions)**.** We define the following loss functions on $\mathbb{R}$.

(1) The *squared error loss* $L(\theta_1, \, \theta_2) := (\theta_1 - \theta_2)^2$.
(2) The *absolute error loss* $L(\theta_1, \, \theta_2) := |\theta_1 - \theta_2|$.
(3) The L_p *loss* $L(\theta_1, \, \theta_2) := |\theta_1 - \theta_2|^p$ for $p \geqslant 1$. (Note that the absolute error loss corresponds to $p = 1$ while the squared error loss corresponds to $p = 2$.)
(4) The *zero-one loss*

$$L(\theta_1, \, \theta_2) := \mathbb{1}(\theta_1 \neq \theta_2) = \begin{cases} 1 & \text{if } \theta_1 \neq \theta_2 \text{ and} \\ 0 & \text{if } \theta_1 = \theta_2. \end{cases}$$

(5) The *Kullback-Leibler loss*

$$L(\theta_1, \, \theta_2) := \int \log \left(\frac{f(x; \theta_1)}{f(x; \theta_2)} \right) f(x; \theta_1) dx = \mathbb{E}_{\theta_1} \log \frac{f(\,\cdot\,; \theta_1)}{f(\,\cdot\,; \theta_2)}$$

where $f(\,\cdot\,; \theta)$ is the PDF of the distribution of the underlying parametric model corresponding to the parameter θ.

Definition 12.4 (Risk)**.** Let $\mathcal{F}$ be a parametric model with parameter θ and parameter space Θ, let L be a loss function over $\mathcal{F}$, and let $\hat{\theta}$ be a point estimator of θ. The *risk of* $\hat{\theta}$ *with respect to* L is the function $R(\,\cdot\,, \hat{\theta}) : \Theta \to \mathbb{R}_{\geqslant 0}$ defined by

$$R(\theta, \, \hat{\theta}) = \int L\left(\theta, \, \hat{\theta}(x_1, \, \dots, \, x_n)\right) f(x_1; \theta) \dots f(x_n; \theta) dx_1 \dots dx_n,$$

or in other words

$$R(\theta, \, \hat{\theta}) = \mathbb{E}_\theta \, L(\theta, \, \hat{\theta}),$$

for any $\theta \in \Theta$.

Lemma 12.5 (The mean squared error is a risk)**.** *The risk with respect to the squared error loss is the mean squared error.*

12.2. **Comparing Risk Functions.**

Definition 12.6 (Maximum risk)**.** Let $\hat{\theta}$ be a point estimator with risk R associated to a loss function L. The *maximum risk* with respect to the loss L is defined to be the value

$$\overline{R}(\hat{\theta}) := \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$$

where Θ denotes the parameter space.

Definition 12.7 (Bayes risk)**.** Let $\hat{\theta}$ be a point estimator of a parameter θ with prior distribution f and risk R associated to a loss function L. The *Bayes risk with respect to the loss L and the prior f* is defined to be the value

$$r(f, \hat{\theta}) := \int R(\theta, \hat{\theta}) f(\theta) d\theta.$$

In other words the Bayes risk is the mean, under the prior distribution, of the risk.

Remark 12.8 (Decision rules)**.** In the context of statistical decision theory, point estimators are sometimes referred to as *decision rules*. This is done for example in Definitions 12.9 and 12.10 below.

Definition 12.9 (Bayes rule)**.** Let $\mathcal{F}$ be a parametric model with risk R associated to a loss function L and let f be a prior distribution over its parameter space. A point estimator $\hat{\theta}$ is called a *Bayes rule for the model $\mathcal{F}$ with respect to the loss L and the prior f* if it minimizes the Bayes risk with respect to the loss L and the prior f, i.e.

$$r(f, \hat{\theta}) = \inf_{\tilde{\theta}} r(f, \tilde{\theta})$$

where the infimum is taken over all point estimators $\tilde{\theta}$. The Bayes rule is also called the *Bayes estimator*.

Definition 12.10 (Minimax rule)**.** Let $\mathcal{F}$ be a parametric model with parameter θ, parameter space Θ, and risk R associated to a loss function L. Let $\hat{\theta}$ be a point estimator of θ. We say that $\hat{\theta}$ is a *minimax rule for the model $\mathcal{F}$ with respect to the loss L* if it minimizes the maximum risk with respect to the loss L, i.e.

$$\overline{R}(\hat{\theta}) = \inf_{\tilde{\theta}} \overline{R}(\tilde{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \tilde{\theta})$$

where the infimum is taken over all point estimators $\tilde{\theta}$. The minimax rule is also called the *minimax estimator*.

Remark 12.11 (Trivial point estimators are ill-defined)**.** A point estimator $\hat{\theta}$ of a parameter θ should really be thought of as a sequence of maps $\hat{\theta}_n : \mathbb{R}^n \to \Theta$, where Θ denotes the parameter space.

This is an important perspective to keep in mind when one is naively tempted to argue that the trivial estimator "$\hat{\theta} = \theta$" ought to be both a Bayes rule and a minimax rule. Indeed, for such an estimator, the risk satisfies

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta L(\theta, \hat{\theta}) = \mathbb{E}_\theta L(\theta, \theta) \equiv 0$$

by positive-definiteness of the loss function. So what is the problem? The problem is that this is *not* a well-defined point estimator! Indeed: θ is just the notation used for arbitrary elements of Θ, so we *cannot* define a map $\hat{\theta}_n : \mathbb{R}^n \to \Theta$ as "$\hat{\theta} = \theta$"; that "equality" is not a well-defined statement.

12.3. Bayes Estimators.

Definition 12.12 (Posterior risk)**.** Let $\hat{\theta}$ be a point estimator of a parameter θ with prior distribution f and let L be a loss function. The *posterior risk of θ with respect to the loss L and the prior f* is defined to be the function $r(\hat{\theta}| \cdot\,) : \mathbb{R}^n \to \mathbb{R}$ given by

$$r(\hat{\theta}\,|\,x^n) = \int L(\theta,\, \hat{\theta}(x^n)) f(\theta\,|\,x^n) d\theta$$

where $x^n = (x_1,\, \ldots,\, x_n) \in \mathbb{R}^n$ and where $f(\theta\,|\,x^n)$ denotes the posterior distribution.

Theorem 12.13 (Charaterization of Bayes estimators)**.** *Let $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a parametric model with risk R associated to a loss function L and let f be a prior distribution over its parameter space. The following hold.*

(1) For any point estimator $\hat{\theta}$ of θ, its Bayes risk may be written as

$$r(f,\, \hat{\theta}) = \int r(\hat{\theta}\,|\,x^n) m(x^n) dx^n$$

where $x^n = (x_1,\, \ldots,\, x_n)$, $r(\hat{\theta}\,|\,x^n)$ denotes the posterior risk, and m is the marginal distribution of $X^n = (X_1,\, \ldots,\, X_n)$ given by

$$m(x^n) = \int f(x^n,\, \theta) d\theta = \int f(x^n\,|\,\theta) f(\theta) d\theta.$$

(2) The Bayes estimator is the point estimator which minimizes the posterior risk pointwise. In other words the Bayes estimator $\hat{\theta}$ may be defined as

$$\hat{\theta}(x^n) = \underset{\theta\in\Theta}{\arg\min}\, r(\theta\,|\,x^n)$$

for all $x^n \in \mathbb{R}^n$.

Theorem 12.14 (Bayes estimators for some important loss functions)**.** *Consider a parametric model $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ with prior distribution f over its parameter space and let L be a loss function.*

(1) Suppose L is the squared error loss. The Bayes estimator $\hat{\theta}$ is the mean of the posterior distribution, i.e.

$$\hat{\theta}(x^n) = \int \theta f(\theta\,|\,x^n) d\theta$$

for every $x^n \in \mathbb{R}^n$, where $f(\theta\,|\,x^n)$ denotes the posterior distribution.

(2) Suppose L is the absolute error loss. The Bayes estimator is the median of the posterior distribution.

(3) Suppose L is the zero-one loss. Any mode of the posterior distribution is a Bayes estimator.

12.4. Minimax rules.

Definition 12.15 (Least favourable prior)**.** Let $\hat{\theta}^f$ be the Bayes rule for the parametric model $\mathcal{F} := \{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ with respect to the loss function L and the prior distribution f over Θ. If the risk of $\hat{\theta}^f$ is uniformly bounded above by its Bayes risk, i.e.

$$R(\theta,\, \hat{\theta}^f) \leqslant r(f,\, \hat{\theta}^f) \text{ for all } \theta \in \Theta,$$

then we call f a *least favourable prior.*

Theorem 12.16 (Necessary conditions for minimax rules)**.** *Any Bayes estimator associated to a least favourable prior is a minimax rule.*

Corollary 12.17 (Necessary condition for minimax rules)**.** *Any Bayes estimator with constant risk is a minimax rule.*

Theorem 12.18 (Minimax estimator for the Normal model with known variance)**.** *Consider the Normal parametric model with unit variance*

$$\mathcal{F} := \{\phi(\,\cdot\, - \theta) : \theta \in \mathbb{R}\}\,,$$

where ϕ is the PDF of a standard normal, which consists of all the distributions $N(\theta,\,1)$ for $\theta \in \mathbb{R}$. The sample mean $\hat{\theta}(x^n) := \frac{1}{n}\sum_{i=1}^{n} x_i$ is a minimax rule for any sufficiently regular loss function whose level sets are convex and symmetric about the origin. Moreover it is the only point estimator with this property.

12.5. Maximum Likelihood, Minimax, and Bayes.

Remark 12.19 (Maximum likelihood, minimax, and Bayes)**.** In most parametric models, with a sufficiently large sample size the MLE is approximately both a minimax estimator and a Bayes estimator. However this fails when the number of parameters is large.

12.6. Admissibility.

Definition 12.20 (Admissible and inadmissible)**.** Let $\mathcal{F}$ be a parametric model with parameter space Θ and risk R associated with a loss function L. A point estimator $\hat{\theta}$ of $\mathcal{F}$ is said to be *inadmissible with respect to the loss* L if there exists another point estimator $\hat{\theta}'$ such that

$$R(\,\cdot\,,\,\hat{\theta}') \leqslant R(\,\cdot\,,\,\hat{\theta}) \text{ everywhere on } \Theta \text{ and}$$
$$R(\theta,\,\hat{\theta}') < R(\theta,\,\hat{\theta}) \text{ for at least one } \theta \in \Theta.$$

If a point estimator is not inadmissible then it is called *admissible*.

Remark 12.21 (Good and bad estimators)**.** Bayes and minimax estimators are essentially "good" point estimators whereas inadmissible estimators are essentially "bad" estimators. Note, however, that admissible estimators may still be bad. For example constant estimators are often admissible with respect to the squared error loss, even though they are very poor estimators.

Theorem 12.22 (Admissibility of Bayes rule)**.** *Let $\mathcal{F}$ be a parametric model with parameter space $\Theta \subseteq \mathbb{R}$ which is open. Let L be a loss function such that for every sufficiently regular point estimator $\hat{\theta}$ the associated risk $R(\,\cdot\,,\,\hat{\theta})$ is continuous over Θ. Let f be a prior distribution over Θ with* full support*, meaning that*

$$\int_{\theta-\varepsilon}^{\theta+\varepsilon} f > 0$$

for every $\theta \in \Theta$ and every sufficiently small $\varepsilon > 0$. Let $\hat{\theta}^f$ denote the Bayes rule for the model $\mathcal{F}$ with respect to the loss L and the prior f. If the Bayes risk of $\hat{\theta}^f$ is finite then $\hat{\theta}^f$ is admissible.

Theorem 12.23 (Admissibility of the sample mean for the Normal model)**.** *Consider the Normal parametric model*

$$\mathcal{F} := \left\{\frac{1}{\sigma}\phi\left(\frac{\cdot - \mu}{\sigma}\right) : \mu \in \mathbb{R} \text{ and } \sigma > 0\right\},$$

where ϕ is the PDF of a standard normal, which consists of all Normal distributions. The sample mean $\hat{\theta}(x^n) := \frac{1}{n}\sum_{i=1}^{n} x_i$ is admissible with respect to the squared error loss.

Theorem 12.24 (Admissiblity and minimax rules)**.** *Let $\mathcal{F}$ be a parametric model with risk R associated to a loss function L and let $\hat{\theta}$ be a point estimator of $\mathcal{F}$. If $\hat{\theta}$ has constant risk and is admissible then it is a minimax estimator.*

Definition 12.25 (Strongly inadmissible)**.** Let $\mathcal{F}$ be a parametric model with parameter space Θ and risk R associated to a loss function L. A point estimator $\hat{\theta}$ of $\mathcal{F}$ is said to be *strongly inadmissible with respect to the loss L* if there exists $\varepsilon > 0$ and another point estimator $\hat{\theta}'$ such that

$$R(\,\cdot\,, \hat{\theta}') \leqslant R(\,\cdot\,, \hat{\theta}) - \varepsilon \text{ everywhere on } \Theta.$$

Theorem 12.26 (Strong inadmissibility and minimax rules)**.** *Minimax rules are not strongly inadmissible.*

13. Linear and Logistic Regression

Remark 13.1 (Estimating the regression function)**.** Recall that we introduced the notion of a *regression function* in Definition 6.3. In practice we seek to estimate the regression function between two random variables X and Y based on data of the form

$$(Y_1,\, X_1),\, \ldots,\, (Y_n,\, X_n) \sim F_{Y,\,X}.$$

13.1. Simple Linear Regression.

Definition 13.2 (Simple linear regression distribution)**.** Let $\beta_0,\, \beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$. We define the *simple linear regression distribution* $F_{Y,X}(\,\cdot\,; \beta_0, \beta_1, \sigma^2)$ as follows:

$$(Y,\, X) \sim F_{Y,\,X}\left(\,\cdot\,; \beta_0,\, \beta_1,\, \sigma^2\right)$$

if there exists a CDF F_X and a random variable ε with mean zero and variance σ^2 such that

$$X \sim F_X \text{ and } Y = \beta_0 + \beta_1 X + \varepsilon.$$

Lemma 13.3 (Alternate characterization of simple linear regression distributions)**.** *Let β_0, β_1 and $\sigma^2 \geqslant 0$. The random variables X and Y have a simple linear regression distribution $F_{Y,X}(\,\cdot\,; \beta_0, \beta_1, \sigma^2)$ if and only if there exists a CDF F_X such that*

$$X \sim F_X,\ \mathbb{E}(Y - \beta_0 - \beta_1 X) = 0,\ \ \textit{and } \mathbb{V}(Y - \beta_0 - \beta_1 X) = \sigma^2.$$

Proof. If $(Y,\, X)$ has a simple linear regression distribution $F_{Y,X}(\,\cdot\,; \beta_0, \beta_1, \sigma^2)$ then, by definition, $Y - \beta_0 - \beta_1 X = \varepsilon$ where ε has mean zero and variance σ^2. Conversely, if $Y - \beta_0 - \beta_1 X$ has mean zero and variance σ^2 then choosing $\varepsilon := Y - \beta_0 - \beta_1 X$ establishes that $(Y,\, X)$ has a simple linear regression distribution $F_{Y,X}(\,\cdot\,; \beta_0, \beta_1, \sigma^2)$. □

Theorem 13.4 (Any joint distribution is a simple linear regression distribution)**.** *Let X and Y be two* random variables. *There exist infinitely many $\beta_0,\, \beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$ such that $(Y,\, X)$ has a simple linear regression distribution* $F_{Y,X}(\,\cdot\,; \beta_0, \beta_1, \sigma^2)$.

Proof. Choose any $\beta_0,\, \beta_1 \in \mathbb{R}$ such that

$$\mathbb{E}Y - \beta_0 - \beta_1 \mathbb{E}X = 0,$$

namely

$$\beta_0 = \mathbb{E}Y - \alpha \mathbb{E}X \text{ and } \beta_1 = \alpha$$

for any $\alpha \in \mathbb{R}$, and choose

$$\sigma^2 := \mathbb{V}(Y - \beta_0 - \beta_1 X) \geqslant 0.$$

Then, trivially by the choice of β_0, β_1, and σ^2, $Y - \beta_0 - \beta_1 X$ has mean zero and variance σ^2. It thus follows from Lemma 13.3 that $(Y,\, X)$ has a simple linear regression distribution $F_{Y,X}(\,\cdot\,; \beta_0, \beta_1, \sigma^2)$. □

Remark 13.5 (The mean is a simple linear regression)**.** The trivial case $\beta_0 = \mathbb{E}Y$ and $\beta_1 = 0$ in Theorem 13.4 means that we may write Y as

$$Y = \mathbb{E}Y + \varepsilon$$

for $\varepsilon := Y - \mathbb{E}Y$, where ε is then treated as noise. This is a crude, but perfectly valid, way of writing Y.

Theorem 13.6 (Regression function of a simple linear regression distribution)**.** *Let β_0, $\beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$, let X and Y be two random variables drawn from the simple linear regression distribution $F_{Y,X}(\,\cdot\,; \beta_0, \beta_1, \sigma^2)$, and let $\varepsilon := Y - \beta_0 - \beta_1 X$. The regression function r between Y and X is*

$$r(x) = \beta_0 + \beta_1 x$$

if and only if

$$\mathbb{E}\left(\varepsilon \,|\, X = x\right) = 0$$

for every x. In particular this holds if ε and X are independent.

Proof. This follows from a direct computation since

$$r(x) = \mathbb{E}\left(Y \,|\, X = x\right) = \mathbb{E}\left(\beta_0 + \beta_1 X + \varepsilon \,|\, X = x\right) = \beta_0 + \beta_1 x + \mathbb{E}\left(\varepsilon \,|\, X = x\right).$$

In particular, if ε and X are independent then

$$\mathbb{E}\left(\varepsilon \,|\, X = x\right) = \int e f(e|x) de = \int e f(e) de = \mathbb{E}\varepsilon = 0,$$

as desired. □

Definition 13.7 (Simple linear regression model)**.** The parametric model

$$\mathcal{F} := \left\{ F_{Y,X}(\,\cdot\,; \beta_0, \beta_1, \sigma^2) : \beta_0,\ \beta_1 \in \mathbb{R} \text{ and } \sigma^2 \geqslant 0 \right\},$$

whose parameter space is $\mathbb{R}^2 \times [0, \infty)$, is called the *simple linear regression model.*

Remark 13.8 (Abusing the term "parametric model")**.** In Definition 13.7 above we are abusing the term parametric model. Indeed: the simple linear regression model is *not* actually a parametric model (in the precise sense of Definition 6.1). This is because, for any fixed set of parameters β_0, $\beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$, there are infinitely many simple linear regression distributions with those parameters (which are distinguished, in the notation of Definition 13.2, by having distinct distributions F_X).

Nonetheless, in practice when carrying out inference about a simple linear regression model we typically only seek to estimate the parameters β_0, β_1, and σ^2 and do *not* seek to estimate the distribution F_X. This is why we will, abusively, refer to this model as a parametric model.

Remark 13.9 (Interpretation of the simple linear regression model)**.** Since, as established in Theorem 13.4, *any* joint distribution may be viewed as a simple linear regression distribution, the simple linear regression model as a *set* is of little interest (it contains *all* joint distributions!) The interest of this model lies in how it *parametrizes* joint distributions.

Recall also that, while all joint distributions may be viewed as simple linear regression functions, they will typically *not* have a linear regression function. This is made precise in Theorem 13.6

Remark 13.10 (Why the model is called simple)**.** The term *simple* in the simple linear regression model refers to the fact that the feature X is *one-dimensional*, i.e. the codomain of X is $\mathbb{R}^k$ for $k = 1$.

Definition 13.11 (Standard noise assumptions for the simple linear regression model)**.** Let β_0, $\beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$, let X and Y be two random variables drawn from the simple linear regression distribution $F_{Y,X}(\,\cdot\,; \beta_0, \beta_1, \sigma^2)$, and let

$$\varepsilon := Y - \beta_0 - \beta_1 X.$$

We say that ε satisfies the *standard noise assumptions for the simple linear regression model* if

$$\mathbb{E}\left(\varepsilon \,|\, X = x\right) = 0 \text{ and } \mathbb{V}\left(\varepsilon \,|\, X = x\right) = \sigma^2$$

for all x.

Remark 13.12 (Standard noise assumptions for the simple linear regression model)**.** As proved in Theorem 13.6, and as already noted above in Remark 13.9, if ε satisfies the standard noise assumptions for the simple linear regression model then the regression function is indeed $r(x) = \beta_0 + \beta_1 x$.

Definition 13.13 (Terminology for the simple linear regression model)**.** Consider an IID sample $(Y_1, X_1), \dots, (Y_n, X_n)$ drawn from a distribution in the simple linear regression model with parameters $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$ and let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote point estimators of β_0 and β_1, respectively.

- The function $\hat{r}$ defined by
$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$
is called the *fitted line*.
- The values $\hat{Y}_i := \hat{r}(X_i)$ are called the *predicted values* or *fitted values*.
- The values
$$\hat{\varepsilon}_i := Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$
are called the *residuals*.
- The *residual sum of squares*, or *RSS*, is defined to be
$$RSS := \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

Definition 13.14 (Least squares estimates for the simple linear regression model)**.** Let $(Y_1, X_1), \dots, (Y_n, X_n)$ be an IID sample drawn from a distribution in the simple linear regression model with parameters $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$. The *least squares estimates* are the point estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ (of β_0 and β_1 respectively) which minimize the residual sum of squares. In other words

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Theorem 13.15 (Least squares estimates)**.** *Let $(Y_1, X_1), \dots, (Y_n, X_n)$ be an IID sample drawn from a distribution in the simple linear regression model with parameters $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$. The least squares estimates are*

$$\hat{\beta}_0 = \overline{Y}_n - \hat{\beta}_1 \overline{X}_n \text{ and } \hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)\left(Y_i - \overline{Y}_n\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2}$$

for $\overline{X}_n$ and $\overline{Y}_n$ denoting the sample means*. Moreover, if the standard noise assumptions for the simple linear regression model are satisfied then*

$$\hat{\sigma}^2 := \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{RSS}{n-2},$$

where $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$ denote the residuals, is an unbiased point estimator of σ^2.

Remark 13.16 (Form of the least squares estimates)**.** In the notation of Theorem 13.15 let $\hat{C}$ and S_P^2 denote the plug-in estimators of the covariance $\mathrm{Cov}(X,\, Y)$ and the variance $\mathbb{V}X$, respectively, such that S_P^2 is the *population variance*. Then we see that the least squares estimate for β_1 is

$$\hat{\beta}_1 = \frac{\hat{C}}{S_P^2}.$$

This should not be surprising: if $Y = \beta_0 + \beta_1 X + \varepsilon$ for ε independent of X (such that, as per Theorem 13.6, the regression function between X and Y is truly linear, i.e. $r(x) = \beta_0 + \beta_1 x$), then

$$\mathrm{Cov}(Y,\, X) = \mathrm{Cov}(\beta_0 + \beta_1 X + \varepsilon,\, X) = \beta_1 \mathbb{V}(X) + \underbrace{\mathrm{Cov}(\varepsilon,\, X)}_{=0}$$

and so

$$\beta_1 = \frac{\mathrm{Cov}(Y,\, X)}{\mathbb{V}(X)}.$$

Remark 13.17 (Bias correction for the noise variance σ^2)**.** In general, for k–dimensional linear regression where X has codomain $\mathbb{R}^k$, the unbiased point estimator of σ^2 takes the form

$$\frac{1}{n-k-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

This is consistent with the formula recorded in Theorem 13.15 above since in that case $k = 1$.

13.2. **Least Squares and Maximum Likelihood.**

Definition 13.18 (Normal noise assumption for the simple linear regression model)**.** Let $\beta_0,\, \beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$, let X and Y be two random variables drawn from the simple linear regression distribution $F_{Y,X}(\,\cdot\,; \beta_0, \beta_1, \sigma^2)$, and let

$$\varepsilon := Y - \beta_0 - \beta_1 X.$$

We say that ε satisfies the *Normal noise assumptions for the simple linear regression model* if

$$\mathbb{E}\left(\varepsilon \,|\, X = x\right) \sim N(0,\, \sigma^2)$$

for all x.

Remark 13.19 (Normal noise assumptions for the simple linear regression model)**.** While in practice we typically use the Normal noise assumptions for the simple linear regression model when establishing certain properties of least squares estimates, especially when a connection to the likelihood is desired or helpful, there is also a practical takeaway here: linear regression performs well when the residuals are normally distributed.

Remark 13.20 (Normal versus standard noise assumptions for the linear regression model)**.** We note immediately that, for the simple linear regression model, the Normal noise assumption implies the standard noise assumptions.

Theorem 13.21 (Least squares and maximum likelihood)**.** *Consider the simple linear regression model with parameters $\beta_0,\, \beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$. Under the Normal* *noise assumption the least squares estimates (for β_0 and β_1) are maximum likelihood estimates. Moreover, in that case, the MLE for σ^2 is*

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon_i}^2.$$

Note that this differs from the unbiased point estimator of σ^2 recorded in Theorem 13.15.

13.3. **Properties of Least Squares Estimators.**

Theorem 13.22 (Conditional means and variances of the least squares estimates)**.** *Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from a distribution in the simple linear regression model with parameters $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$ and let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the least squares estimates (for β_0 and β_1 respectively). Under the standard noise assumptions for the simple linear regression model the conditional means and variances of the least squares estimates given the feature data $X_1, \ldots, X_n$ are given by, for $\hat{\beta} := (\hat{\beta}_0, \hat{\beta}_1)$,*

$$\mathbb{E}(\hat{\beta} \,|\, X^n) = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \text{ and } \mathbb{V}(\hat{\beta} \,|\, X^n) = \frac{\sigma^2}{n s_X^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix}$$

where $X^n = (X_1, \ldots, X_n)$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean, and we denote by $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$ the population variance.

Remark 13.23 (Standard error estimates for the least squares estimates)**.** Under the assumptions and notation of Theorem 13.22, the standard error estimates for the least squares estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$\widehat{se}(\hat{\beta}_0) = \frac{\hat{\sigma}}{s_X\sqrt{n}}\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}} \text{ and } \widehat{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{s_X\sqrt{n}},$$

where we abused notation and wrote $\widehat{se}(\hat{\beta}_i)$ to denote the estimate of the *conditional* standard error $\widehat{se}(\hat{\beta}_i \,|\, X^n)$ for $i = 0, 1$, for $\hat{\sigma}$ as in Theorem 13.15.

Theorem 13.24 (Properties of the least squares estimates)**.** *Consider an IID sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ drawn from a distribution in the simple linear regression model $\mathcal{L}$ with parameters $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$, let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the least squares estimates (for β_0 and β_1 respectively), and let $\widehat{se}(\hat{\beta}_0)$ and $\widehat{se}(\beta_1)$ denote the corresponding conditional standard error estimates as introduced in Remark 13.23. The following hold.*

1. *The least squares estimates are consistent, i.e. $\hat{\beta}_0 \xrightarrow{P} \beta_0$ and $\hat{\beta}_1 \xrightarrow{P} \beta_1$ as $n \to \infty$.*
2. *The least squares estimates are asymptotically normal, i.e.*

$$\frac{\hat{\beta}_0 - \beta_0}{\widehat{se}(\hat{\beta}_0)} \rightsquigarrow N(0,\, 1) \text{ and } \frac{\hat{\beta}_1 - \beta_1}{\widehat{se}(\hat{\beta}_1)} \rightsquigarrow N(0,\, 1)$$

   as $n \to \infty$.

3. *For any $\alpha \in (0,\, 1)$, as $n \to \infty$ the intervals*

$$\hat{\beta}_0 \pm z_{\alpha/2}\widehat{se}(\hat{\beta}_0) \text{ and } \hat{\beta}_1 \pm z_{\alpha/2}\widehat{se}(\hat{\beta}_1)$$

are asymptotic $1-\alpha$ confidence intervals for β_0 and β_1, respectively, where $z_{\alpha/2} := \Phi^{-1}(1-\alpha/2)$ for Φ the CDF of a standard normal distribution.

(4) *For any $\alpha \in (0, 1)$, if we define the statements H_0 and H_1 as*

$$H_0 : \beta_1 = 0 \text{ and } H_1 : \beta_1 \neq 0$$

and let $W := \hat{\beta}_1 / \widehat{se}(\hat{\beta}_1)$ and $R_\alpha := \left\{ w \in \mathbb{R} : |w| > z_{\alpha/2} \right\}$, then the tuple $(\mathcal{L}, \{\beta_0\}, \mathbb{R} \setminus \{\beta_0\}, H_0, H_1, W, R_\alpha)$ is a size α Wald test.

13.4. **Prediction.**

Theorem 13.25 (Prediction interval for simple linear regression)**.** *Consider an IID sample $(Y_1, X_1), \dots, (Y_n, X_n)$ drawn from a distribution in the simple linear regression model with parameters $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma^2 \geqslant 0$, let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the least squares estimates (for β_0 and β_1 respectively), and let $\hat{\sigma}^2$ denote the consistent point estimator of σ^2 recorded in Theorem 13.15.*

Let (Y, X) be drawn from the same distribution and fix x_* in the codomain of X. An estimate of $Y_* := \mathbb{E}(Y \mid X = x_*)$ is

$$\hat{Y}_* := \hat{\beta}_0 + \hat{\beta}_1 x_*.$$

Moreover, under the Normal noise assumption for the simple linear regression model, if we define (for $X_* = x_*$)

$$\hat{\xi}_n^2 := \hat{\sigma}^2 \left(\frac{\sum_{i=1}^n (X_i - X_*)^2}{n \sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2} + 1 \right)$$

where $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ denotes the sample mean, then for any $\alpha \in (0, 1)$ the interval

$$\hat{Y}_* \pm z_{\alpha/2} \hat{\xi}_n,$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ for Φ the CDF of a standard normal, is asymptotically a $1-\alpha$ confidence interval for Y_* as $n \to \infty$.

13.5. **Multiple Regression.**

Definition 13.26 (Linear regression distribution)**.** Let $\beta \in \mathbb{R}^{k+1}$ and $\sigma^2 \geqslant 0$. We define the *linear regression distribution* $F_{Y,X}(\,\cdot\,; \beta, \sigma^2)$ as follows:

$$(Y, X) \sim F_{Y,X}(\,\cdot\,; \beta, \sigma^2),$$

where the codomains of Y and X are $\mathbb{R}$ and $\mathbb{R}^{k+1}$, respectively, if there exists a multivariate distribution F_X and a random variable ε with mean zero and variance σ^2 such that

$$X = (X_0, X_1, \dots, X_k) \sim F_X,\ X_0 = 1 \text{ always, and } Y = \beta \cdot X + \varepsilon.$$

Remark 13.27 (Formulation of linear regression)**.** Imposing $X_0 = 1$ in Definition 13.26 may appear strange. This is done for convenience. Instead of writing

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon \text{ for } \beta_0 \in \mathbb{R} \text{ and } \beta_1, X \in \mathbb{R}^k$$

we write, more concisely,

$$Y = \beta \cdot X + \varepsilon \text{ for } \beta \text{ and } X = (1, X_1, \dots, X_k) \in \mathbb{R}^{k+1}.$$

This trick essentially allows us to *pretend*, notationally speaking, that the *affine* map $X \mapsto \beta \cdot X$ is *linear*.

Remark 13.28. A linear regression distribution with $k = 1$ is a *simple* linear regression distribution.

Lemma 13.29 (Alternate characterization of linear regression distributions)**.** *Let $\beta \in \mathbb{R}^{k+1}$ and $\sigma^2 \geqslant 0$. The random variable Y and the random vector X in $\mathbb{R}^{k+1}$ have a linear regression distribution $F_{Y,X}(\,\cdot\,; \beta, \sigma^2)$ if and only if there exists a multivariate distribution F_X such that*

$$X = (X_0, X_1, \dots, X_k) \sim F_X,\, X_0 = 1 \text{ always,}$$
$$\mathbb{E}(Y - \beta \cdot X) = 0, \text{ and } \mathbb{V}(Y - \beta \cdot X) = \sigma^2.$$

Proof. If (Y, X) has a linear regression distribution $F_{Y,X}(\,\cdot\,; \beta, \sigma^2)$ then, by definition, $X_0 = 1$ always and $Y - \beta \cdot X = \varepsilon$ where ε has mean zero and variance σ^2. Conversely, if $X_0 = 1$ always and $Y - \beta \cdot X$ has mean zero and variance σ^2 then choosing $\varepsilon := Y - \beta \cdot X$ establishes that (Y, X) has a linear regression distribution $F_{Y,X}(\,\cdot\,; \beta, \sigma^2)$. □

Theorem 13.30 (Any joint distribution is a linear regression distribution)**.** *Let Y be a random variable, let X be a random vector in $\mathbb{R}^k$, and define the random vector $\widetilde{X} =: (1, X)$ in $\mathbb{R}^{k+1}$. There are infinitely many $\beta \in \mathbb{R}^{k+1}$ and $\sigma^2 \geqslant 0$ such that $(Y, \widetilde{X})$ has a linear regression distribution $F_{Y,X}(\,\cdot\,; \beta, \sigma^2)$.*

Proof. Note that, since $\widetilde{X}_0 \equiv 1$ and hence $\mathbb{E}\widetilde{X}_0 = 1$, $\mathbb{E}\widetilde{X}$ is a non-zero vector. We may therefore choose

$$\beta := \mathbb{E}Y \frac{\mathbb{E}\widetilde{X}}{|\mathbb{E}\widetilde{X}|^2} + \alpha \text{ for any } \alpha \perp \mathbb{E}\widetilde{X},\, \alpha \in \mathbb{R}^{k+1},$$

and

$$\sigma^2 := \mathbb{V}(Y - \beta \cdot \widetilde{X}).$$

Note that there are infinitely many possible choices for α since $\mathbb{E}\widetilde{X}$ is a non-zero vector in $\mathbb{R}^{k+1}$ and so $(\mathbb{E}\widetilde{X})^\perp$ is a k–dimensional subspace. Then

$$\mathbb{E}(Y - \beta \cdot \widetilde{X}) = \mathbb{E}Y - \underbrace{\beta \cdot \mathbb{E}\widetilde{X}}_{\mathbb{E}Y} - \underbrace{\alpha \cdot \mathbb{E}\widetilde{X}}_{0} = 0$$

while, trivially, $\mathbb{V}(Y - \beta \cdot \widetilde{X}) = \sigma^2$. By Lemma 13.29 this tells us that $(Y, \widetilde{X})$ has a linear regression distribution $F_{Y,X}(\,\cdot\,; \beta, \sigma^2)$. □

Theorem 13.31 (Regression function of a linear regression distribution)**.** *Let $\beta \in \mathbb{R}^{k+1}$ and $\sigma^2 \geqslant 0$, let (Y, X) be a random variable-random vector pair drawn from the linear regression distribution $F_{Y,X}(\,\cdot\,; \beta, \sigma^2)$, and let $\varepsilon := Y - \beta \cdot X$. The regression function r between Y and X is*

$$r(x) = \beta \cdot x \text{ when } x_0 = 1$$

if and only if

$$\mathbb{E}(\varepsilon \,|\, X = x) = 0 \text{ for every } x \in \mathbb{R}^{k+1} \text{ with } x_0 = 1.$$

Proof. This follows from a direct computation since, when $x_0 = 1$,

$$r(x) = \mathbb{E}\,(Y \,|\, X = x) = \mathbb{E}\,(\beta \cdot X + \varepsilon \,|\, X = x) = \beta \cdot x + \mathbb{E}\,(\varepsilon \,|\, X = x)\,. \qquad \square$$

Definition 13.32 (Linear regression model)**.** The parametric model

$$\mathcal{F} := \left\{ F_{Y,X}(\,\cdot\,;\beta,\sigma^2) : \beta \in \mathbb{R}^{k+1} \text{ and } \sigma^2 \geqslant 0 \right\},$$

whose parameter space is $\mathbb{R}^{k+1} \times [0,\infty)$, is called the *linear regression model.*

Remark 13.33 (Abusing the term "parametric model", again)**.** Once again, the term parametric model is abused in Definition 13.32 above. This is exactly the same situation as that discussed in Remark 13.8 so we refer to that previous remark and do not expand further here.

Remark 13.34. When $k = 1$ the linear regression model is the *simple* linear regression model.

Definition 13.35 (Standard noise assumptions for the linear regression model)**.** Let $\beta \in \mathbb{R}^{k+1}$ and $\sigma^2 \geqslant 0$, let (Y, X) be a random variable-random vector pair drawn from the linear regression distribution $F_{Y,X}(\,\cdot\,;\beta,\sigma^2)$, and let

$$\varepsilon := Y - \beta \cdot X.$$

We say that ε satisfies the *standard noise assumptions for the linear regression model* if

$$\mathbb{E}\,(\varepsilon \,|\, X = x) = 0 \text{ and } \mathbb{V}\,(\varepsilon \,|\, X = x) = \sigma^2$$

for all $x \in \mathbb{R}^{k+1}$ with $x_0 = 1$.

Remark 13.36 (On the noise term in the linear regression model)**.** In the linear regression model the response variable Y is a scalar. Therefore, even though the covariate X is a random *vector*, the noise term ε (in the notation of Definition 13.35) is also a scalar-valued random variable. In particular its variance is a scalar (and not a matrix).

When we later see expressions of the form $\mathbb{V}\,(\varepsilon^n \,|\, X^n) = \sigma^2 I$ the matrix comes from the fact that the ε_i's and X_i's are IID, and so independent from one another, where here $\varepsilon^n = (\varepsilon_1, \dots, \varepsilon_n)$ and $X^n = (X_1, \dots, X_n)$.

Definition 13.37 (Terminology for the linear regression model)**.** Consider an IID sample $(Y_1, X_1), \dots, (Y_n, X_n)$ drawn from a distribution in the linear regression model with parameters $\beta \in \mathbb{R}^{k+1}$ and $\sigma^2 \geqslant 0$ and let $\hat{\beta}$ denote a point estimator of β.

- Since $X_i \in \mathbb{R}^k$ for $1 \leqslant i \leqslant n$ we define the *design matrix* $\mathbb{X}$ via

  $$\mathbb{X}_{ij} := (X_i)_j$$

  such that $\mathbb{X}$ is an n-by-$(k+1)$ matrix.
- We call $\mathbb{Y} := (Y_1, \dots, Y_n) \in \mathbb{R}^n$ the *response vector.*
- The function $\hat{r}$ defined by

  $$\hat{r}(x) = \hat{\beta} \cdot x$$

  for every $x \in \mathbb{R}^{k+1}$ with $x_0 = 1$ is called the *fitted hyperplane.*
- The values $\hat{Y}_i := \hat{r}(X_i)$ are called the *predicted values* or *fitted values.*
- The values

  $$\hat{\varepsilon}_i := Y_i - \hat{Y}_i = Y_i - \hat{\beta} \cdot X_i$$

  are called the *residuals* and $\hat{\varepsilon} := (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)$ is called the *vector of residuals.*

- The *residual sum of squares*, or *RSS*, is defined to be

$$RSS := \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = |\hat{\varepsilon}|^2.$$

Lemma 13.38 (Linear regression model in terms of the design matrix)**.** *Consider an IID sample* $(Y_1, X_1), \ldots, (Y_n, X_n)$ *drawn from a distribution in the linear regression model with parameters* $\beta \in \mathbb{R}^{k+1}$ *and* $\sigma^2 \geqslant 0$*. Let* $\mathbb{X}$ *denote the design matrix, let* $\mathbb{Y}$ *denote the response vector, and let* $\hat{\varepsilon}$ *denote the vector of residuals We have that*

$$\mathbb{Y} = \mathbb{X}\beta + \hat{\varepsilon}.$$

Proof. This follows immediately since

$$Y_i = \beta \cdot X_i + \hat{\varepsilon}_i = \sum_{j=1}^{k} \beta_j (X_i)_j + \hat{\varepsilon}_i = (\mathbb{X}\beta)_i + \hat{\varepsilon}_i. \qquad \square$$

Definition 13.39 (Least squares estimate for the linear regression model)**.** Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from a distribution in the linear regression model with parameters $\beta \in \mathbb{R}^{k+1}$ and $\sigma^2 \geqslant 0$. The *least squares estimate* is the point estimator $\hat{\beta}$ of β which minimizes the residual sum of squares. In other words, for $\mathbb{X}$ denoting the design matrix and $\mathbb{Y}$ denoting the response vector,

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^{k+1}} |\mathbb{Y} - \mathbb{X}\beta|^2.$$

Remark 13.40 (Exact least squares estimate)**.** Note that if the linear regression model is *exact*, meaning in the notation of Definition 13.35 that the noise term has variance $\mathbb{V}(\varepsilon \,|\, X) \equiv 0$ (i.e. $\sigma^2 = 0$ under the standard noise assumptions for the linear regression model), then the least squares estimate $\hat{\beta}$ is simply the solution to the linear system

$$\mathbb{Y} = \mathbb{X}\beta$$

where $\mathbb{X}$ denotes the design matrix and $\mathbb{Y}$ denotes the response vector.

Typically, however, σ^2 is *strictly positive* and in that case the least squares estimate is *not* an exact solution of that linear system. (Typically the response vector $\mathbb{Y}$ will not not belong to the image of the design matrix $\mathbb{X}$, and so exact solutions of that linear system do not exist.)

Theorem 13.41 (Least squares estimate)**.** *Let* $(Y_1, X_1), \ldots, (Y_n, X_n)$ *be an IID sample drawn from a distribution in the linear regression model with parameters* $\beta \in \mathbb{R}^{k+1}$ *and* $\sigma^2 \geqslant 0$*, let* $\mathbb{X}$ *denote the associated design matrix, and let* $\mathbb{Y}$ *denote the associated response vector. If* $\mathbb{X}^T\mathbb{X}$ *is invertible then the least squares estimate is*

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}.$$

Moreover, if the standard noise assumptions for the linear regression model are satisfied then the following hold.

(1) The point estimator

$$\hat{\sigma}^2 := \frac{1}{n-k-1} \sum_{i=1}^{n} |\hat{\varepsilon}_i|^2 = \frac{RSS}{n-(k+1)},$$

where $\hat{\varepsilon}_i$ *denote the residuals, is an unbiased estimator of* σ^2*.*

(2) The conditional mean and variance of the least squares estimate given the feature data $X_1, \dots, X_n$ *(summarized in the design matrix $\mathbb{X}$) are*

$$\mathbb{E}(\hat{\beta} \,|\, \mathbb{X}) = \beta \text{ and } \mathbb{V}(\hat{\beta} \,|\, \mathbb{X}) = \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1}.$$

(3) The interval

$$\hat{\beta}_j \pm z_{\alpha/2} \widehat{se}(\hat{\beta}_j),$$

where $z_{\alpha/2} = \Phi(1 - \alpha/2)$ for Φ the CDF of a standard normal and where $\widehat{se}(\hat{\beta}_j)$ is the j–th diagonal element of $\hat{\sigma}^2 (\mathbb{X}^T \mathbb{X})^{-1}$, is asymptotically a $1-\alpha$ confidence interval for β_j as $n \to \infty$.

Remark 13.42 (Gram matrix and collinearity)**.** In the notation of Theorem 13.41 above the Gram matrix $\mathbb{X}^T \mathbb{X}$ plays an essential role.

- If $\mathbb{X}^T \mathbb{X}$ is invertible, then we may compute the least squares estimate in terms of its inverse.
- If $\mathbb{X}^T \mathbb{X}$ fails to be invertible then Lemma C.8 tells us that the design matrix $\mathbb{X}$ fails to have full rank, i.e. $\operatorname{rank} \mathbb{X} < k+1$. This means that the $(k+1)$–many features of X are, on that particular sample, *not* linearly independent. They actually live in some lower-dimensional subspace! We would therefore need to to first identify that subspace and then carry out a lower-dimensional linear regression restricted to that subspace.

13.6. Model Selection.

Remark 13.43 (Underfitting and overfitting)**.** Typically, as more covariates are used in a linear regression model, the bias of the predictions decreases while their variance increases.

- Too few covariates yield high bias; this is called *underfitting*.
- Too many covariates yield high variance; this is called *overfitting*.

To produce good predictions we therefore typically seek a balance between bias and variance (which is reminiscent of the bias-variance decomposition of the mean squared error recorded in Theorem 6.10).

Definition 13.44 (Affine reduction)**.** Let X be a random vector with codomain $\mathbb{R}^{k+1}$ written $X = (X_0, X_1, \dots, X_k)$ and let $S \subseteq \{1, \dots, k\}$. We define the *affine S–reduction* of X, denoted X^S, to be the random vector in $R^{|S|+1}$

$$X^S := (X_j : j = 0 \text{ or } j \in S)\,.$$

Remark 13.45. We use the notation $(X_j : j \in \mathcal{J})$ in Definition 13.44 instead of $\{X_j : j \in \mathcal{J}\}$ to indicate that X^S is an *ordered* tuple.

For example if $k = 4$ and $S = \{2, 4\}$, such that $|S| + 1 = 3$, then the *affine S–reduction* is

$$X^S = (X_0, X_1, X_2, X_3, X_4)^S = (X_0, X_2, X_4),$$

which is indeed a random vector in $\mathbb{R}^3$. Note that X^S differs from (X_0, X_4, X_2), i.e. indeed the order matters.

Remark 13.46. We single out the index $j = 0$ in the definition of an affine S–reduction because, *specifically* in the context of the linear regression model, the random variable X_0 must satisfy $X_0 \equiv 1$ in order to account for the affine term β_0 (see also Remark 13.27). Therefore we *cannot* drop X_0 when constructing the S–reduction of X.

Definition 13.47 (Dimension of the linear regression model)**.** The linear regression model with parameter space $\mathbb{R}^{k+1} \times [0, \infty)$ is said to have *dimension* k.

Definition 13.48 (Terminology for reduced linear regresson models)**.** Consider an IID sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ drawn from the linear regression model of dimension k, let $S \subseteq \{1, \ldots, k\}$, and write $l := |S|$.

- We define the *S–reduced design matrix* $\mathbb{X}_S$ to be the design matrix corresponding to the S–affine reductions $X_1^S, \ldots, X_n^S$, such that $\mathbb{X}_S$ is a n-by-$(l+1)$ matrix.
- We define the *S–reduced least squares estimate* $\hat{\beta}_S$ to be the least squares estimate of $(Y_1, X_1^S), \ldots, (Y_n, X_n^S)$ over the linear regression model of dimension l.
- The function $\hat{r}_S$ defined by
$$\hat{r}_S(x) := \hat{\beta}_S \cdot x$$
for every $x \in \mathbb{R}^{l+1}$ with $x_0 = 1$ is called the *S–reduced fitted hyperplane.*
- The values $\hat{Y}_i(S) := \hat{r}_S(X_i^S)$ are called the *S–reduced predicted values* or *S–reduced fitted values.*

Definition 13.49 (Prediction risk)**.** Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from a distribution in the linear regression model of dimension k, consider $S \subseteq \{1, \ldots, k\}$, and let $\hat{Y}_i(S)$ denote the S–reduced predicted values for $1 \leqslant i \leqslant n$. The *prediction risk* $R(S)$ is defined to be

$$R(S) := \sum_{i=1}^{n} \mathbb{E}\left([\hat{Y}_i(S) - Y_i^*]^2\right)$$

where $Y_i^* \sim Y \,|\, X = X_i$ for $1 \leqslant i \leqslant n$ where (Y, X) denotes a random variable–random vector pair drawn from the same distribution as above.

Remark 13.50 (Risk and prediction risk)**.** If we consider Y_i^* to be the baseline truth (akin to a true parameter, which Y_i^* is not since it is a random variable) and consider $\hat{Y}_i(S)$ to be a "point estimator" of Y_i^* then we can view the prediction risk as the sum of the risks with respect to the squared error loss given by

$$R(Y_i^*, \hat{Y}_i(S)) = \mathbb{E}\left([Y_i^* - \hat{Y}_i(S)]^2\right),$$

such that indeed

$$R(S) = \sum_{i=1}^{n} R(Y_i^*, \hat{Y}_i(S)).$$

Remark 13.51 (Prediction risk)**.** The prediction risk is a measure of how good our model is. In practice we seek to find $S \subseteq \{1, \ldots, k\}$ which minimizes the associated prediction risk $R(S)$. However, since the prediction risk cannot be computed from an IID sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ we must instead *approximate it.*

Definition 13.52 (Training error)**.** Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from the linear regression model of dimension k, let $S \subseteq \{1, \ldots, k\}$, and let $\hat{Y}_i(S)$ denote the S–reduced predicted values for $1 \leqslant i \leqslant n$. The *training error* $\hat{R}_{tr}(S)$ is defined to be

$$\hat{R}_{tr}(S) := \sum_{i=1}^{n} [\hat{Y}_i(S) - Y_i]^2.$$

Lemma 13.53 (Training error in terms of the reduced design matrix)**.** *Consider an IID sample* $(Y_1, X_1), \ldots, (Y_n, X_n)$ *drawn from the linear regression model of dimension* k *and let* $S \subseteq \{1, \ldots, k\}$*. The training error may be written as*

$$\hat{R}_{tr}(S) = \left|\mathbb{Y} - \mathbb{X}_S \hat{\beta}_S\right|^2$$

where $\mathbb{Y}$ *denotes the response vector,* $\mathbb{X}_S$ *denotes the S–reduced design matrix, and* $\hat{\beta}_S$ *denotes the S–reduced least squares estimate. In particular the training error is the residual sum of squares corresponding to the sample*

$$(Y_1, X_1^S), \ldots, (Y_n, X_n^S).$$

Proof. The first identity is immediate since

$$\begin{aligned}\hat{R}_{tr}(S) = \sum_{i=1}^{n} \left(\hat{Y}_i(S) - Y_i\right)^2 &= \sum_{i=1}^{n} \left(\hat{\beta}_S \cdot X_i^S - Y_i\right)^2 \\ &= \sum_{i=1}^{n} \left(\sum_{j=1}^{|S|} \hat{\beta}_{S,j} \left(X_i^S\right)_j - Y_i\right)^2 \\ &= \sum_{i=1}^{n} \left[\left(\mathbb{X}_S \hat{\beta}_S\right)_i - Y_i\right]^2 \\ &= \left|\mathbb{X}_S \hat{\beta}_S - \mathbb{Y}\right|^2.\end{aligned}$$

In light of Lemma 13.38 it then follows that the training error is indeed a residual sum of squares as claimed above. □

Theorem 13.54 (Bias of the training error)**.** *Let* $(Y_1, X_1), \ldots, (Y_n, X_n)$ *be an IID sample drawn from the linear regression model of dimension* k *and let us consider* $S \subseteq \{1, \ldots, k\}$*. As a point estimator of the prediction risk, the training error satisfies, under the standard noise assumptions for the simple linear regression model,*

$$\text{bias}[\hat{R}_{tr}(S)] = -2 \sum_{i=1}^{n} \text{Cov}(\hat{Y}_i(S), Y_i) < 0,$$

where $\hat{Y}_i(S)$ *denoted the S–reduced predicted values for* $1 \leqslant i \leqslant n$*.*

Remark 13.55 (Bias of the training error)**.** Theorem 13.54 tells us that the training error is a *downward-biased* estimate of the prediction risk, i.e. $\hat{R}_{tr}(S) < R(S)$. Moreover as the number of covariates grow, the predicted values $\hat{Y}_i$ will correlate strongly with the sample response Y_i and so $\text{Cov}(\hat{Y}_i, Y_i)$ will grow. (This is a manifestation of *overfitting* – see Remark 13.43.)

Definition 13.56 (Mallows' C_p statistic)**.** Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from a distribution in the linear regression model with parameters $\beta \in \mathbb{R}^{k+1}$ and $\sigma^2 \geqslant 0$ and let $S \subseteq \{1, \ldots, k\}$. *Mallows' C_p statistic* is defined to be

$$\hat{R}(S) := \hat{R}_{tr}(S) + 2|S|\hat{\sigma}^2$$

for $\hat{\sigma}^2$ as in Theorem 13.15 obtained from the full model (i.e. with $|S| = k$) and where $\hat{R}_{tr}(S)$ denotes the training error.

Remark 13.57 (Motivation of Mallows' C_p statistic)**.** Suppose that, for some $S \subseteq \{1, \ldots, k\}$ with $|S| = n$, i.e. the dimension of the model being equal to the number of data points, the S–reduced least squares estimate $\hat{\beta}_S$ is *exact* in the sense that it solves

$$\mathbb{Y} = \mathbb{X}_S \hat{\beta}_S$$

for $\mathbb{Y}$ the response vector and $\mathbb{X}_S$ the S–reduced design matrix. In some sense this corresponds to *maximal overfitting* since we have enough covariates to exactly match the data (see Remark 13.43). Then Theorem 13.54 tells us that, under the standard noise assumptions for the simple linear regression model, the training error and the prediction risk are related via

$$\mathbb{E}[\hat{R}_{tr}(S)] = R(S) - 2\sum_{i=1}^{n} \mathrm{Cov}(\hat{Y}_i,\, Y_i)$$

where, since $\hat{\beta}_S$ is exact,

$$\hat{Y}_i = Y_i \text{ for all } i.$$

Therefore the rule of iterated expectation tells us that

$$\begin{aligned}
\mathrm{Cov}(\hat{Y}_i,\, Y_i) &= \mathbb{E}\left[\mathrm{Cov}(Y_i,\, Y_i \,|\, \mathbb{X})\right] \\
&= \mathbb{E}\left[\mathbb{V}(Y_i \,|\, \mathbb{X})\right] \\
&= \mathbb{E}\left[\mathbb{V}(\beta \cdot X_i + \varepsilon_i \,|\, \mathbb{X})\right] \\
&= \mathbb{E}\left[\mathbb{V}(\varepsilon_i \,|\, \mathbb{X})\right] \\
&= \sigma^2.
\end{aligned}$$

So finally we conclude that, since $|S| = n$ by assumption,

$$R(S) = \mathbb{E}[\hat{R}_{tr}(S)] + 2\sum_{i=1}^{n} \mathrm{Cov}(\hat{Y}_i,\, Y_i) = \mathbb{E}[\hat{R}_{tr}(S)] + 2|S|\sigma^2.$$

In other words: in this instance the mean of Mallows' C_p statistic is *equal* to the prediction risk.

This is admittedly a contrived example since in practice $n \gg k > |S|$ (which is also why we do not expect *exact* least squares estimates in practice). Nonetheless, Mallows' C_p statistic remains a good estimate of the prediction risk in general.

Definition 13.58 (Akaike Information Criterion)**.** Let $(Y_1,\, X_1), \ldots, (Y_n,\, X_n)$ be an IID sample drawn from the linear regression model of dimension k and let $S \subseteq \{1, \ldots, k\}$. We define the *Akaike Information Criterion* to be

$$AIC(S) := l_S - |S|$$

where l_S denotes the log-likelihood of the model evaluated at the MLE. Note that this means that l_S is the log-likelihood function of the joint distribution of $(Y_i,\, X_i)_{i=1}^n$ (see Exercise A.13.6).

Remark 13.59 (Akaike Information Criterion)**.** In practice we use the AIC by selecting $S \subseteq \{1, \ldots, k\}$ such that the corresponding AIC is *maximal* (see Remark 13.60 and Exercise A.13.6).

Remark 13.60 (Good methods and trade-offs)**.** Recall (c.f. Remark 13.51) that we seek good models by seeking models with low prediction risk. In particular (c.f.

Remark 13.57) we know that Mallows' C_p statistic estimates the prediction risk. In light of Remark 13.55 we may interpret the two terms of Mallows' C_p statistic as

$$\hat{R}(S) = \text{ lack of fit } + \text{ complexity penalty.}$$

In other words: *good models trade-off fit and complexity.* Finally note that since Exercise A.13.6 tells us that, under the Normal noise assumptions for the simple linear regression model, maximizing the AIC is equivalent to minimizing Mallows' C_p statistic, we may interpret the AIC as

$$AIC = \text{ goodness of fit } - \text{ complexity.}$$

Definition 13.61 (Leave-one-out cross-validation)**.** Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from the linear regression model of dimension k, consider $S \subseteq \{1, \ldots, k\}$, and let $l := |S|$. We define the *leave-one-out cross-validation risk estimator* to be

$$\hat{R}_{CV}(S) := \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{(i)}(S)\right)^2$$

where $Y_{(i)}(S) := \hat{\beta}_{S,(i)} \cdot X_i$ for $\hat{\beta}_{S,(i)}$ the least squares estimate of

$$(Y_1, X_1^S), \ldots, (Y_{i-1}, X_{i-1}), (Y_{i+1}, X_{i+1}), \ldots, (Y_n, X_n^S),$$

i.e. the S–affine reduction of the original sample with (Y_i, X_i^S) removed, over the linear regression model of dimension l.

In some sense we compute, for each i, how to best predict Y_i from all the other data points except (Y_i, X_i).

Lemma 13.62 (Alternate expression for leave-one-out cross-validation)**.** *Consider an IID sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ drawn from the linear regression model of dimension k and let $S \subseteq \{1, \ldots, k\}$. The leave-one-out cross-validation risk estimator may be written as*

$$\hat{R}_{CV}(S) := \sum_{i=1}^{n} \left[\frac{Y_i - \hat{Y}_i(S)}{1 - \mathcal{U}_{ii}(S)}\right]^2$$

where $\mathcal{U}_{ii}(S)$ is the i-th diagonal element of the matrix $\mathbb{X}_S(\mathbb{X}_S^T\mathbb{X}_S)^{-1}\mathbb{X}_S^T$ where $\mathbb{X}_S$ denotes the S–reduced design matrix. (The interpretation of this matrix is recorded in Lemma C.8.)

In particular one need not drop each observation one at a time and perform n linear regressions: a single linear regression without omission suffices to compute $\hat{R}_{CV}(S)$.

Definition 13.63 (p–fold cross-validation)**.** Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from the linear regression model of dimension k, let $S \subseteq \{1, \ldots, k\}$, let $l := |S|$, and let $\mathcal{G}$ be a partition of $\{1, \ldots, n\}$ of size p.

For each group G in $\mathcal{G}$ let $\hat{\beta}_{S,G}$ denote the least squares estimate of

$$\left\{(Y_i, X_i^S) : i \in G\right\},$$

i.e. the S–affine reduction of the samples corresponding to group G, over the linear regression model of dimension l, and let

$$\hat{R}_{CV}(S, G) := \sum_{i \notin G} \left[Y_i - \hat{Y}_i(S, G)\right]^2$$

where $\hat{Y}_i(S, G) := \hat{\beta}_{S,G} \cdot X_i$, i.e. summing over data points *not* in group G. The *p–fold cross-validation risk estimator with respect to the partition $\mathcal{G}$* is

$$\hat{R}_{CV}(S; \mathcal{G}) := \sum_{G \in \mathcal{G}} \frac{|G|}{n} \hat{R}_{CV}(S, G).$$

In other words we compute, for each group, how well the linear model determined from that group predicts the *remaining* data points. We then obtain the estimator by averaging these quantities over all groups (weighted by the sizes of the group).

Remark 13.64 (Leave-one-out cross-validation is essentially a form of *improper* n–fold cross-validation)**.** Suppose that in Definition 13.63 we relax the requirement that $\mathcal{G}$ be a partition by no longer requiring that the groups $G \in \mathcal{G}$ be pairwise disjoint. An instance of such an *improper* "partition" is

$$\mathcal{G} = \{G_i : 1 \leqslant i \leqslant n\} \text{ where } G_i = \{1, \ldots, i-1, i+1, \ldots, n\}.$$

Then the corresponding *improper* n–fold cross-validation risk estimator would be

$$\begin{aligned} \hat{R}_{CV}(S; \mathcal{G}) &= \sum_{i=1}^{n} \frac{n-1}{n} \hat{R}_{CV}(S, G_i) \\ &= \frac{n-1}{n} \sum_{i=1}^{n} \left[Y_i - \hat{Y}_i(S, G_i)\right]^2 \\ &= \frac{n-1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_{S,G_i} \cdot X_i\right)^2. \end{aligned}$$

Crucially: the improper n–fold cross-validation least squares estimate $\hat{\beta}_{S,G_i}$ is exactly the same as the leave-one-out cross-validation least squares estimate $\hat{\beta}_{S,(i)}$! Therefore

$$\begin{aligned} \hat{R}_{CV}(S; \mathcal{G}) &= \frac{n-1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_{S,G_i} \cdot X_i\right)^2 \\ &= \frac{n-1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_{S,(i)} \cdot X_i\right)^2 = \frac{n-1}{n} \hat{R}_{CV}(S), \end{aligned}$$

i.e. the two risk estimators are the same up to a factor of $\frac{n-1}{n}$.

Remark 13.65 (Motivation for p–fold cross-validation)**.** p–fold cross-validation is another way to address the *downward-bias* of the training error which is due to the correlation between the sample values and the predicted values (see Remark 13.55). Mallows' C_p statistic (and AIC, since Exercise A.13.6 tells us that they may be comparable – see also Remark 13.60) addresses this issue by adding a *complexity penalty* to the training errors. Instead, p–fold cross-validation addresses this issue by evaluating the training error on subsets where the samples and the predictions are *decorrelated* since $\hat{Y}_i$ is predicted by a model that never saw (Y_i, X_i).

Remark 13.66 (Cross-validation and Mallows' C_p statistic)**.** In practice, when it comes to linear regression, Mallows' C_p statistic and cross-validation (either leave-one-out cross-validation or p–fold cross-validation) often yield similar results and so one might as well use the computationally simpler Mallows' C_p statistic.

When it comes to more complex models, however, cross-validation may be preferable.

Definition 13.67 (Bayesian Information Criterion)**.** Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from the linear regression model of dimension k and let $S \subseteq \{1, \ldots, k\}$. We define the *Bayesian Information Criterion* to be

$$BIC(S) := l_S - \frac{|S|}{2} \log n$$

where l_S denotes the log-likelihood of the model evaluated at the MLE.

Remark 13.68 (Bayesian information criterion)**.** In practice we use the BIC by selecting $S \subseteq \{1, \ldots, k\}$ such that the corresponding BIC is *maximal* (which is reminiscent of how the AIC is used – see Remark 13.59).

Remark 13.69 (Bayesian interpretation of the BIC)**.** Let $\mathcal{S} = \{S_1, \ldots, S_m\}$ denote a set of models, i.e. here in the context of linear regression $S_j \subseteq \{1, \ldots, k\}$ for $1 \leqslant j \leqslant m$, and suppose that we assign the uniform distribution $\mathbb{P}(S_j) = 1/m$ as a prior distribution on $\mathcal{S}$. Provided that the prior distributions for the parameters of each model are sufficiently regular, the posterior distribution is approximately

$$\mathbb{P}\left(S_j \,|\, \text{data}\right) \approx \frac{e^{BIC(S_j)}}{\sum_{r=1}^{m} e^{BIC(S_r)}}.$$

This means that choosing S to maximize the BIC is equivalent to choosing S to be the mode of the posterior distribution, i.e. the model for which the posterior probability $\mathbb{P}\left(S_j \,|\, \text{data}\right)$ is maximal.

Remark 13.70 (BIC, AIC, and Mallows' C_p statistic)**.** Recall that in Remark 13.60 we interpreted the AIC as

$$AIC = \text{ goodness of fit } - \text{ complexity.}$$

Since the AIC and BIC are defined in very similar ways, up to a factor of $\frac{1}{2}\log n$ multiplying the complexity penalty term $|S|$, we may therefore view the BIC as imposing a *stronger* complexity penalty and thus choosing *simpler* models.

Remark 13.71 (Model search)**.** If there are k covariates then there are 2^k possible reduced linear regression models (since there are 2^k subsets of $\{1, \ldots, k\}$).

If k is small then we may perform each of the 2^k possible linear regressions, assign a score to each model (such as Mallows' C_p statistic, the AIC, or the BIC), and select the model with the best score (smallest for Mallows' C_p statistic; largest for AIC and BIC).

When k is large, however, this is infeasible and we may only search over a *subset* of the set of possible models.

Definition 13.72 (Model score)**.** A *model score* for the linear regression model of dimension k is a map

$$\hat{R} : \mathcal{P}\left(\{1, \ldots, k\}\right) \to \mathbb{R}.$$

Remark 13.73 (Model scores)**.** As the notation $\hat{R}$ suggests (notation which is slightly overused since it describes both generic model scores as well as a specific example of one, namely Mallows' C_p statistic), model scores are typically viewed as point estimators of the prediction risk. The following are examples of model scores.

- Training error.
- Mallows' C_p statistic.
- AIC.

- Leave-one-out cross-validation risk estimator.
- p–fold cross-validation risk estimator.
- BIC.

Note that some of these scores are used to select models by *minimizing* the score (as is done with the training error, Mallows' C_p statistic, or cross-validation risk estimators) while other scores select models by *maximizing* the score (as is done with the AIC and BIC). This distinction is merely cosmetic: if a model score $\hat{R}$ selects models by minimization then we may simply use $-\hat{R}$ if we wish to instead select models by maximization. In the sequel (and notably in Definition 13.74 we will thus assume without loss of generality that all model scores select models by maximization.

Definition 13.74 (Forward and backward stepwise regression)**.** Consider an IID sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ drawn from the linear regression model of dimension k and let $\hat{R}$ be a model score.

Forward stepwise regression selects $S_* \subseteq \{1, \ldots, k\}$ as follows. We initialize with $S = \emptyset$ and repeat the following until a subset S_* is selected.

- Let $i_* \in \arg\max_{i \notin S} \hat{R}(S \cup \{i\})$.
- If $\hat{R}(S \cup \{i_*\}) \leqslant \hat{R}(S)$ then select $S_* = S$.
- Otherwise, if $\hat{R}(S \cup \{i_*\}) > \hat{R}(S)$ then update $S := S \cup \{i_*\}$. In particular, if $|S| = k$ then select $S_* = S$, otherwise continue.

Backward stepwise regression selects $S_* \subseteq \{1, \ldots, k\}$ as follows. We initialize with $S = \{1, \ldots, k\}$ and repeat the following until a subset S_* is selected.

- Let $i_* \in \arg\max_{i \in S} \hat{R}(S \setminus \{i\})$.
- If $\hat{R}(S \setminus \{i_*\}) \leqslant \hat{R}(S)$ then select $S_* = S$.
- Otherwise, if $\hat{R}(S \setminus \{i_*\}) > \hat{R}(S)$ then update $S := S \setminus \{i_*\}$. In particular, if $|S| = 0$ then select $S_* = S$, otherwise continue.

Remark 13.75 (Forward and backward stepwise regression)**.** Forward and backward stepwise regression are *greedy* procedures.

- Forward stepwise regression proceeds by *adding* covariates into the model, one at a time. At every step we choose the covariate whose addition will most increase the model score (and, if there are no covariates whose addition improves the score, we stop).
- Backward stepwise regression proceeds in reverse, *removing* covariates from the model one at a time. At every step we choose the covariate whose removal will most increase the model score (and, if there are no covariates whose removal improves the score, we stop).

Since these procedures are greedy there is *no guarantee* that they will select the model with the highest score. Indeed, a simple way to see this is to note that these stepwise regression procedures will only consider, at most,

$$\sum_{j=1}^{k} j = \frac{k(k-1)}{2} \approx k^2$$

of the models. But there are 2^k possible models and $k^2 \ll 2^k$, so most of the possible models will never be considered.

Definition 13.76 (Zheng-Loh model selection method)**.** Consider an IID sample $(Y_1, X_1), \dots, (Y_n, X_n)$ drawn from the linear regression model of dimension k. The *Zheng-Loh model selection method* selects $S_* \subseteq \{1, \dots, k\}$ as follows.

(1) Let $\hat{\beta}$ be the least squares estimate (for the full model, i.e. $S = \{1, \dots, k\}$) and let, for every $1 \leqslant j \leqslant k$,

$$W_j := \frac{\hat{\beta}_j}{\widehat{se}(\hat{\beta}_j)},$$

where $\widehat{se}(\hat{\beta}_j)$ is as in item 3 of Theorem 13.41, be the Wald test statistic testing

$$H_0 : \beta_j = 0 \text{ versus } H_1 : \beta_j \neq 0.$$

(2) Order the test statistics from largest to smallest in absolute value, i.e.

$$|W_{j_1}| \geqslant |W_{j_2}| \geqslant, \dots, \geqslant |W_{j_k}|.$$

(3) Choose

$$\hat{a} = \underset{1 \leqslant a \leqslant k}{\arg\min}\, \hat{R}_{tr}(S_a) + a\hat{\sigma}^2 \log n$$

where $S_a := \{j_1, \dots, j_a\}$, i.e. S_a retains only the first a covariates with the largest Wald test statistics (in absolute value), and where $\hat{\sigma}^2$ is as in Theorem 13.41. Recall in particular that the training error $\hat{R}_{tr}(S_a)$ is easily accessible since it corresponds to a residual sum of squares (see Lemma 13.53).

(4) Select $S_* = S_{\hat{a}}$.

Remark 13.77 (Zheng-Loh model selection method)**.** The Zheng-Log model selection method does not, by contrast with other model selection methods, seek to minimize the prediction risk (e.g. by way of minimizing a model score acting as a point estimator of the prediction risk).

Instead the Zheng-Log method *assumes* that the (true) parameter β satisfies $\beta_j = 0$ for some j's, and then seeks to find an estimate $\hat{\beta}$ also satisfying $\hat{\beta}_j = 0$ for all of those j's.

13.7. **Logistic Regression.**

Definition 13.78 (Logit and logistic functions)**.** The function logit : $(0, 1) \to \mathbb{R}$ is defined as

$$\operatorname{logit}(p) := \log\left(\frac{p}{1-p}\right).$$

Its inverse is the *logistic function* $\sigma : \mathbb{R} \to (0, 1)$ defined by

$$\sigma(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}.$$

Definition 13.79 (Logistic regression distribution)**.** Let $\beta \in \mathbb{R}^{k+1}$. We define the *logistic regression distribution* $F_{Y,X}(\,\cdot\,;\beta)$ as follows:

$$(Y, X) \sim F_{Y,X}(\,\cdot\,;\beta),$$

where the codomain of Y and X are $\{0, 1\}$ and $\mathbb{R}^{k+1}$, respectively, if there exists a multivariate distribution F_X such that

$$X = (X_0, X_1, \dots, X_k) \sim F_X,\ X_0 = 1 \text{ always, and } Y \,|\, X = x \sim \text{Bernoulli}(p)$$

for

$$p = p(x; \beta) = \sigma(\beta \cdot x) = \frac{e^{\beta \cdot x}}{1 + e^{\beta \cdot x}}$$

or, equivalently,

$$\operatorname{logit} p(x; \beta) = \beta \cdot x.$$

Remark 13.80. Once again, as for linear regression distributions, we impose the condition $X_0 = 1$ in order to account for the affine term β_0 in

$$\operatorname{logit} p(x; \beta) = \beta \cdot x = \beta_0 + \sum_{j=1}^{k} \beta_j x_j.$$

(See also Remark 13.27.)

Remark 13.81 (Where did the noise go?)**.** A noise term ε appears in the definition of a linear regression distribution but is seemingly absent in the definition of a logistic regression distribution. Yet some noise is present in both cases, as the following observation highlights.

Suppose the parameter β is known and suppose that, for a particular outcome ω, we know the value $X(\omega)$ of X. Then, in either case (linear or logistic) we *still* do not know the value of Y exactly.

- In the case of linear regression: $Y = \beta \cdot X + \varepsilon$ and so $Y(\omega) = \beta \cdot X(\omega) + \varepsilon(\omega)$. Since $\varepsilon(\omega)$ is not known, neither is $Y(\omega)$.
- In the case of logistic regression: $Y \mid X = X(\omega) \sim \text{Bernoulli}(p)$ where $p = p(X(\omega); \beta) = \sigma(\beta \cdot X(\omega))$ is known. Here too, now because of the randomness of the Bernoulli draw, $Y(\omega)$ is not known.

So in the case of logistic regression the noise term ε disappeared and yet the randomness/noise is still there!

To reconcile these ideas let us restrict our attention to the case of linear regression under the Normal noise assumption. Then

$$Y \mid X \sim \beta \cdot X + \varepsilon \sim N(\beta \cdot X,\, \sigma^2).$$

In other words, in the case of *linear* regression under the Normal noise assumption we have that

$$Y \mid X = x \sim N(\beta \cdot x,\, \sigma^2)$$

whereas in the case of *logistic* regression we have that

$$Y \mid X = x \sim \text{Bernoulli}\left(\sigma(\beta \cdot x)\right).$$

The punchline is this: the noise term ε did not disappear, instead it is implicit in the Bernoulli distribution!

Finally note that the Normal noise assumption is helpful to make this explanation cleaner, but is not required. Indeed, under the weaker standard noise assumptions we have that

$$Y \mid X = x \sim F_\varepsilon(\beta \cdot x,\, \sigma^2)$$

for some *fixed* distribution $F_\varepsilon(\beta \cdot x,\, \sigma^2)$ with mean $\beta \cdot x$ and variance σ^2 (which is not necessarily a normal distribution).

Remark 13.82 (Regression functions and the codomain of the response variable)**.** Let us discuss how much information is contained in the regression function depending on the codomain of the response variable. In both cases we consider a covariate $X \in \mathbb{R}^{k+1}$ with $X_0 \equiv 1$.

(1) Suppose first that the codomain of the response variable Y is $\mathbb{R}$. Then, given the regression function $r(x) := E(Y \mid X = x)$ we only have information about the mean of Y. Indeed, by the rule of iterated expectation,

$$\mathbb{E}Y = \mathbb{E}\left[E\left(Y \mid X\right)\right] = \mathbb{E}\left[r(X)\right].$$

However the *distribution* of Y remains unknown.

(2) Suppose now that the codomain of the response variable Y is $\{0, 1\}$. Then $Y \sim \mathrm{Bernoulli}(q)$ for

$$q := \mathbb{P}(Y = 1) = \mathbb{E}(Y).$$

Similarly: $Y \mid X = x \sim \mathrm{Bernoulli}(p(x))$ for

$$p(x) := \mathbb{P}\left(Y = 1 \mid X = x\right) = \mathbb{E}\left(Y \mid X = x\right) =: r(x).$$

In other words: since Y, and hence $Y \mid X = x$, has codomain $\{0, 1\}$, both Y and $Y \mid X = x$ *must* be Bernoulli distributions and so the *full distribution* of $Y \mid X = x$ is characterized by its regression function!

That is the motivation behind logistic regression. When the codomain of Y is $\{0, 1\}$, we may characterize the full distribution of Y from its regression function. To make it easier, since that regression function has codomain $(0, 1)$ we apply the logit function to it in order to work in the classical setting of (transformed) regression functions with codomain $\mathbb{R}$.

Theorem 13.83 (Any joint distribution is approximately a logistic regression distribution)**.** *Let Y be a random variable in $\{0, 1\}$ and let X be a random vector in $\mathbb{R}^{k+1}$ with $X_0 \equiv 1$. There are infinitely many $\beta \in \mathbb{R}^{k+1}$ such that*

$$Y \mid X \sim \mathit{Bernoulli}\left(\sigma\left(\beta \cdot X + \varepsilon\right)\right)$$

for some random variable ε with mean zero and finite variance, in which case we say that (Y, X) approximately *has a logistic regression distribution.*

Proof. Define $p(x) := \mathbb{E}\left(Y \mid X = x\right)$ (which is the regression function between Y and X) such that $p : \mathbb{R}^{k+1} \to (0, 1)$ is a function, and so $p(X)$ is a random variable in $(0, 1)$. Then $\mathrm{logit}(p(X))$ is a random variable in $\mathbb{R}$ and so, by Theorem 13.30, there are infinitely many $\beta \in \mathbb{R}^{k+1}$ such that

$$\mathrm{logit}\, p(X) = \beta \cdot X + \varepsilon$$

for ε a random variable with mean zero and finite variance. Since the inverse of the logit function is the logistic function σ we have that

$$\mathbb{P}\left(Y = 1 \mid X\right) = \mathbb{E}\left(Y \mid X\right) = p(X) = \sigma\left(\beta \cdot X + \varepsilon\right)$$

and so indeed

$$Y \mid X \sim \mathrm{Bernoulli}\left(\sigma(\beta \cdot X + \varepsilon\right). \qquad \square$$

Remark 13.84 (Generalized linear models)**.** Distributions of the form

$$Y \mid X = x \sim f\left(\,\cdot\,; g^{-1}(\beta \cdot x)\right)$$

for some parametric model $\{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ and some bijection $g : \Theta \to \mathbb{R}$ known as the *link function*, are known as *generalized linear regression distributions* (and so their collections are known as *generalized linear regression models*).

Their unifying idea is to perform (vanilla) linear regression on the transformed *parameter space* $g(\Theta)$ (and then use the inverse of the link function to recover the conditional distribution of Y given X).

- Logistic regression distributions correspond to using the Bernoulli model and the logit function as link function.
- Linear regression distributions under the Normal noise assumption with *known* variance correspond to using the Normal model with *fixed* variance and the identity as link function, i.e.
$$Y \mid X = x \sim N(\beta \cdot x,\, \sigma^2)$$
for $\sigma^2 \geqslant 0$ given.

Note that in some cases (see Remark 13.82) the codomain of the response variable Y *imposes* a choice of parametric model. For example (as discussed in Remark 13.82), if the codomain of Y is $\{0, 1\}$ then we may *without loss of generality* view Y, and $Y \mid X = x$, as a Bernoulli random variable.

Definition 13.85 (Logistic regression model)**.** The parametric model
$$\mathcal{F} := \left\{ F_{Y,X}(\,\cdot\,; \beta) : \beta \in \mathbb{R}^{k+1} \right\}$$
is called the *logistic regression model.*

Theorem 13.86 (MLE for the logistic regression model)**.** *Consider an IID sample* $(Y_1, X_1), \ldots, (Y_n, X_n)$ *drawn from a distribution in the logistic regression model with parameter* $\beta \in \mathbb{R}^{k+1}$*, let* $\mathbb{X}$ *denote the design matrix, let* $\mathbb{Y}$ *denote the response vector, and define* $\mathfrak{p} = \mathfrak{p}(\beta) \in \mathbb{R}^n$*,* $\mathbb{W} = \mathbb{W}(\beta) \in \mathbb{R}^{n \times n}$*, and* $\mathbb{Z} = \mathbb{Z}(\beta) \in \mathbb{R}^n$ *via*
$$\begin{aligned} &\operatorname{logit} p_i = \beta \cdot X_i \text{ such that } p_i = \sigma(\beta \cdot X_i) = \sigma(\mathbb{X}\beta)_i, \\ &\mathbb{W} = \operatorname{diag}\left[\mathfrak{p}(1-\mathfrak{p})\right], \text{ i.e. } \mathbb{W}_{ij} = p_i(1-p_i)\delta_{ij}, \text{ and} \\ &\mathbb{Z} = \mathbb{X}\beta + \mathbb{W}^{-1}\left(\mathbb{Y} - \mathfrak{p}\right). \end{aligned}$$
If $\mathbb{X}^T\mathbb{W}\mathbb{X}$ *is uniformly elliptic over* β*, i.e.*
$$\mathbb{X}^T\mathbb{W}\mathbb{X}\xi \cdot \xi \geqslant \lambda|\xi|^2 \text{ for all } \xi \in \mathbb{R}^{k+1}$$
for $\lambda > 0$ *independent of* β*, then the unique MLE* $\hat{\beta}$ *of* β *is the limit of the sequence* $(\hat{\beta}_k)_{k \geqslant 0}$ *defined by*
$$\hat{\beta}_0 := 0 \text{ and } \hat{\beta}_{k+1} := \left(\mathbb{X}^T\mathbb{W}\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{W}\mathbb{Z}\big|_{\beta = \hat{\beta}_k} \text{ for } k \geqslant 0.$$
Moreover
$$\hat{I} := \mathbb{X}^T\mathbb{W}\mathbb{X}\big|_{\beta=\hat{\beta}}$$
is an estimate of the Fisher information matrix such that, for
$$\widehat{se}(\hat{\beta}_j) := \sqrt{\hat{I}^{-1}_{jj}},$$
we have that
$$\frac{\hat{\beta}_j - \beta_j}{\widehat{se}(\hat{\beta}_j)} \rightsquigarrow N(0,\,1) \text{ as } n \to \infty.$$

Proof. See Exercise A.23.9. □

Remark 13.87 (Practical computation of the MLE for the logistic regression model)**.** Given $\beta \in \mathbb{R}^{k+1}$ the next iterate in the sequence described in Theorem 13.86 is constructed as follows. (See Exercise A.23.9 for details.)

(1) Let $p_i := \sigma(\beta \cdot X_i) = \sigma(\mathbb{X}\beta)_i$ for σ denoting the logistic function.
(2) Let $Z_i := (\mathbb{X}\beta)_i + \frac{Y_i - p_i}{p_i(1-p_i)}$.

(3) Let $\mathbb{W} := \operatorname{diag}[\mathfrak{p}(1-\mathfrak{p})]$, i.e. $\mathbb{W}_{ij} = p_i(1-p_i)\delta_{ij}$.
(4) Let $\beta_{new} := (\mathbb{X}^T\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\mathbb{W}\mathbb{Z}$, i.e. β_{new} is the $\mathbb{W}$–weighted least squares estimate

$$\beta_{new} = \underset{\beta\in\mathbb{R}^{k+1}}{\arg\min}\,||\mathbb{Z}-\mathbb{X}\beta||^2_{\mathbb{W}}$$

where $||\beta||^2_{\mathbb{W}} := \mathbb{W}\beta\cdot\beta$.

Remark 13.88 (Model selection for logistic regression)**.** Model selection for the logistic regression model is typically done using the AIC (since we have access to the conditional likelihood of Y given X when performing logistic regression – see Exercise A.23.9 where the log-likelihood is recorded).

13.8. **Bonus.**

Definition 13.89 (Normal noise assumption for the linear regression model)**.** Let $\beta_0 \in \mathbb{R}^{k+1}$ and $\sigma^2 \geqslant 0$, let $(Y,\, X)$ be a random variable-random vector pair drawwn from the linear regression distribution $F_{Y,X}(\,\cdot\,;\beta_0,\sigma^2)$, and let

$$\varepsilon := Y - \beta\cdot X.$$

We say that ε satisfies the *Normal noise assumptions for the linear regression model* if

$$\mathbb{E}\left(\varepsilon \,|\, X = x\right) \sim N(0,\,\sigma^2)$$

for all $x \in \mathbb{R}^{k+1}$ with $x_0 = 1$.

Theorem 13.90 (Prediction interval for multiple linear regression)**.** *Consider an IID sample* $(Y_1,\, X_1), \ldots, (Y_n,\, X_n)$ *drawn from a distribution in the linear regression model with parameters* $\beta \in \mathbb{R}^{k+1}$ *and* $\sigma^2 \geqslant 0$*, let* $\hat\beta$ *denote the least squares estimate (provided it exists – see Theorem 13.41), and let* $\hat\sigma^2$ *denote the consistent point estimator of* σ^2 *recorded in Theorem 13.41.*

Let $(Y,\, X)$ *be drawn from the same distribution and fix* x_* *in the codomain of* X*. An estimate of* $Y_* := \mathbb{E}(Y\,|\,X = x_*)$ *is*

$$\hat Y_* := \hat\beta\cdot x_*.$$

Moreover, under the Normal noise assumption for the linear regression model, if we define (for $X_* = x_*$*)*

$$\hat\xi_n^2 := \hat\sigma^2\left(X_*^T(\mathbb{X}^T\mathbb{X})^{-1}X_* + 1\right)$$

where $\mathbb{X}$ *denotes the design matrix, then for any* $\alpha \in (0,\,1)$ *the interval*

$$\hat Y_* \pm z_{\alpha/2}\hat\xi_n,$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ *for* Φ *the CDF of a standard normal, is asymptotically a* $1-\alpha$ *confidence interval for* Y_* *as* $n\to\infty$*.*

Remark 13.91. Theorem 13.90 is not found in [Was10]. It comes from section 3.5 of [Wei14].

Remark 13.92. The form of the prediction interval recorded in Theorem 13.90 for *multiple* linear regression reduces to the prediction interval recorded in Theorem 13.25 for *simple* linear regression when $k = 1$. This can be seen from a straightforward calculation which is recorded in Exercise A.23.10.

Theorem 13.93 (Parameter prediction interval for logistic regression)**.** *Consider an IID sample* $(Y_1, X_1), \ldots, (Y_n, X_n)$ *drawn from a distribution in the logistic regression model with parameter* $\beta \in \mathbb{R}^{k+1}$ *and let* $\hat{\beta}$ *denote the MLE (provided it exists – see Theorem 13.86).*

Let (Y, X) *be drawn from the same distribution and fix* x_* *in the codomain of* X*. An estimate of* $p_* := \mathbb{P}\,(Y = 1 \,|\, X = x_*)$ *is*

$$\hat{p}_* := \sigma(\hat{\beta} \cdot x_*),$$

where σ *is the logistic function. Moreover, if we define (for* $X_* = x_*$*)*

$$\widehat{se}(\hat{p}_*) := \sigma(\hat{p}_*)[1 - \sigma(\hat{p}_*)]\sqrt{\hat{J} X_* \cdot X_*}$$

where $\hat{J} := \hat{I}^{-1}$ *for* $\hat{I}$ *the estimate of the Fisher information matrix recorded in Theorem 13.86, then for any* $\alpha \in (0, 1)$ *the interval*

$$\hat{p}_* \pm z_{\alpha/2}\widehat{se}(\hat{p}_*),$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ *for* Φ *the CDF of a standard normal, is asymptotically a* $1 - \alpha$ *confidence interval for* p_* *as* $n \to \infty$*.*

Remark 13.94 (Parameter prediction instead of prediction)**.** Note that in Theorem 13.93 above we are *not* providing a prediction interval for $Y_* := \mathbb{E}\,(Y \,|\, X = X_*)$, as is done for example in Theorems 13.25 and 13.90. This is because the codomain of Y_* is $\{0, 1\}$, so providing a prediction interval for Y_* would not be particularly meaningful (nor would it be particularly feasible)! Instead we provide a prediction interval for its *Bernoulli parameter* $p_* := \mathbb{P}\,(Y = 1 \,|\, X = x_*)$.

14. Multivariate Models

14.1. Random Vectors.

Theorem 14.1 (Linear algebra of expectation and variance)**.** *Let X be a random vector in $\mathbb{R}^k$ with mean μ and variance-covariance matrix Σ, let $a \in \mathbb{R}^k$, and let A be a n-by-k matrix. The following hold.*

1. $\mathbb{E}(a \cdot X) = a \cdot X$.
2. $\mathbb{V}(a \cdot X) = \Sigma a \cdot a$.
3. $\mathbb{E}(AX) = A\mu$.
4. $\mathbb{V}(AX) = A\Sigma A^T$.

14.2. Estimating the Correlation.

Theorem 14.2 (Asymptotic confidence interval for the correlation via the Fisher transformation)**.** *Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be an IID sample from a multivariate distribution on $\mathbb{R}^2$. Consider the point estimator of the correlation between X and Y, $\rho := \rho(X, Y)$, given by*

$$\hat{\rho} = \frac{\frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{s_X s_Y}$$

(which is the plug-in estimator of the correlation) where $\overline{X}$ and $\overline{Y}$ denote the respective sample means and s_X^2 and s_Y^2 denote the respective sample variances. Define

$$\hat{\theta} := \operatorname{arctanh} \hat{\rho} \text{ and } \widehat{se}(\hat{\theta}) := \frac{1}{\sqrt{n-3}}$$

as well as, for any $\alpha \in (0, 1)$,

$$a := \hat{\theta} - z_{\alpha/2}\widehat{se}(\hat{\theta}) \text{ and } b := \hat{\theta} + z_{\alpha/2}\widehat{se}(\hat{\theta})$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ for Φ the CDF of a standard normal. Then (a, b) is asymptotically a $1-\alpha$ confidence interval for $\theta := \operatorname{arctanh}\rho$ as $n \to \infty$ and

$$(\tanh a, \tanh b)$$

is asymptotically a $1-\alpha$ confidence interval for ρ as $n \to \infty$.

Remark 14.3 (Alternate expressions for arctanh and tanh)**.** For any $\rho \in (-1, 1)$ we have that

$$\operatorname{arctanh}\rho = \frac{1}{2}\left[\log(1+\rho) - \log(1-\rho)\right]$$

while for any $\theta \in \mathbb{R}$ we have that

$$\tanh\theta = \frac{e^{2\theta}-1}{e^{2\theta}+1}.$$

Remark 14.4 (Alternate confidence interval for the correlation)**.** We can also obtain a confidence interval for the plug-in estimator of the correlation using the *bootstrap* (see Remark 8.5 and Section 8.3).

14.3. **Multivariate Normal.**

Theorem 14.5 (Linear algebra of multivariate normal distributions)**.** *Consider μ, $a \in \mathbb{R}^k$, Σ a symmetric positive-definite k-by-k matrix, and A an n-by-k matrix. The following hold.*

(1) If $Z \sim N(0,\, I_k)$ and $X := \mu + \Sigma^{1/2} Z$ then $X \sim N(\mu,\, \Sigma)$.
(2) If $X \sim N(\mu,\, \Sigma)$ and $Z := \Sigma^{1/2}(X - \mu)$ then $Z \sim N(0,\, I_k)$.
(3) If $X \sim N(\mu,\, \Sigma)$ then $a \cdot X \sim N(a \cdot \mu,\, \Sigma a \cdot a)$.
(4) If $X \sim N(\mu,\, \Sigma)$ then $AX \sim N(A\mu,\, A\Sigma A^T)$.
(5) If $X \sim N(\mu,\, \Sigma)$ and $V := \Sigma^{-1}(X - \mu) \cdot (X - \mu)$ then $V \sim \chi^2_k$.

Remark 14.6. Note that item 4 in Theorem 14.5 above is *not* found in [Was10]. It is recorded in Theorem 2.4.1 of [And03].

Definition 14.7 (Multivariate sample variance)**.** Let $X_1, \dots, X_n$ be random vectors in $\mathbb{R}^k$. The *multivariate sample variance* is the random k-by-k matrix defined by

$$S_{lm} := \frac{1}{n-1} \sum_{i=1}^{n} \left(X_{i,l} - \overline{X}_l\right) \left(X_{i,m} - \overline{X}_m\right)$$

where $\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$, such that $\overline{X}_l = \frac{1}{n} \sum_{i=1}^{n} X_{i,l}$, denotes the sample mean.

Remark 14.8 (Multivariate sample variance and empirical distribution)**.** The multivariate sample variance is nothing more than the variance–covariance matrix of the empirical distribution.

Theorem 14.9 (MLE for the multivariate Normal model)**.** *Consider the k–*dimensional multivariate Normal model

$$\mathcal{F} := \left\{ f(\,\cdot\,; \mu,\, \Sigma) : \mu \in \mathbb{R}^k,\, \Sigma \in \mathbb{R}^{k \times k} \text{ with } \Sigma = \Sigma^T \text{ and } \Sigma > 0 \right\}.$$

The MLE is

$$(\hat{\mu},\, \hat{\Sigma}) = \left(\overline{X},\, \frac{n-1}{n} S \right)$$

where $\overline{X}$ denote the sample mean and S denotes the multivariate sample variance.

15. Inference About Independence

Definition 15.1 (Notation for independence and dependence)**.** Let X and Y be two random variables.

- If X and Y are independent then we write $X \amalg Y$.
- If X and Y are dependent then we write $X \leftrightsquigarrow Y$.

15.1. Two Binary Variables.

Definition 15.2 (Binary random variable)**.** A random variable is called *binary* if its codomain is $\{0, 1\}$.

For example the response variable in logistic regression (see Section 13.7) is a binary random variable. Note also that random variable being binary is equivalent to a random variable being a Bernoulli random variable.

Definition 15.3 (Two-by-two tables)**.** Let $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ be an IID sample drawn from two binary random variables Y and Z. This sample may be represented as a two-by-two table

	$Y = 0$	$Y = 1$	
$Z = 0$	X_{00}	X_{01}	$X_{0\cdot}$
$Z = 1$	X_{10}	X_{11}	$X_{1\cdot}$
	$X_{\cdot 0}$	$X_{\cdot 1}$	$X_{\cdot\cdot} = n$

where

$$X_{ij} = \text{ number of observations where } Z = i \text{ and } Y = j$$

and where dotted subscripts denote sums such that

$$\begin{aligned} X_{i\cdot} &= \sum_{j=0}^{1} X_{ij} = \text{ number of observations where } Z = i, \\ X_{\cdot j} &= \sum_{i=0}^{1} X_{ij} = \text{ number of observations where } Y = j, \text{ and} \\ X_{\cdot\cdot} &= \sum_{i,j=0}^{1} X_{ij} = \text{ number of observations } = n. \end{aligned}$$

The underlying data-generating process may also be represented as a two-by-two table, namely

	$Y = 0$	$Y = 1$	
$Z = 0$	p_{00}	p_{01}	$p_{0\cdot}$
$Z = 1$	p_{10}	p_{11}	$p_{1\cdot}$
	$p_{\cdot 0}$	$p_{\cdot 1}$	$p_{\cdot\cdot} = 1$

where

$$\begin{aligned} p_{ij} &= \mathbb{P}\left(Z = i \text{ and } Y = j\right), \\ p_{i\cdot} &= \mathbb{P}\left(Z = i\right), \\ p_{\cdot j} &= \mathbb{P}\left(Y = j\right), \text{ and} \\ p_{\cdot\cdot} &= \sum_{i,j=0}^{1} p_{ij} = 1. \end{aligned}$$

In particular these two-by-two tables are related as follows: if we define

$$X := (X_{00},\, X_{01},\, X_{10},\, X_{11}) \text{ and } p := (p_{00},\, p_{01},\, p_{10},\, p_{11})$$

then we have that

$$X \sim \text{Multinomial}(n,\, p).$$

We call X the *two-by-two random variable associated with Y and Z* and call p the *two-by-two parameter associated with Y and Z.*

Definition 15.4 (Odds ratio)**.** Let Y and Z be binary random variables and let p be the corresponding two-by-two parameter.

- We define the *odds ratio* to be
$$\psi := \frac{p_{00}p_{11}}{p_{01}p_{10}}.$$
- We define the *log odds ratio* to be
$$\gamma := \log \psi.$$

Remark 15.5 (Odds and odds ratio)**.** For any binary random variable Y, its associated *odds* are defined to be

$$\mathcal{O}(Y) := \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}.$$

In particular if $q := \mathbb{P}(Y=1)$ is the Bernoulli parameter of Y then

$$\mathcal{O}(Y) = \frac{q}{1-q},$$

and so $\log \mathcal{O}(Y) = \text{logit}(q)$ is precisely the (transformed) parameter around which logistic regression is built. But where do these odds come from? Well: they are nothing more than the usual odds in everyday English. For example, if $q = 2/3$ then $Y = 1$ is twice as likely as $Y = 0$ and so

$$\mathcal{O}(Y) = \frac{q}{1-q} = \frac{2/3}{1/3} = 2.$$

Conversely, if $q = 1/3$ then $Y = 1$ is half as likely as $Y = 0$ and so

$$\mathcal{O}(Y) = \frac{1/3}{2/3} = \frac{1}{2}.$$

Generally speaking the odds actually *characterize* a binary random variable since

$$\mathcal{O}(Y) = \frac{q}{1-q} \iff q = \frac{\mathcal{O}(Y)}{1+\mathcal{O}(Y)}.$$

Now suppose that Z is another binary random variable which may or may not influence Y. Then we define the *conditional odds*

$$\mathcal{O}\left(Y \,|\, Z=1\right) := \frac{\mathbb{P}\left(Y=1 \,|\, Z=1\right)}{\mathbb{P}\left(Y=0 \,|\, Z=1\right)}$$

and

$$\mathcal{O}\left(Y \,|\, Z=0\right) := \frac{\mathbb{P}\left(Y=1 \,|\, Z=0\right)}{\mathbb{P}\left(Y=0 \,|\, Z=0\right)}.$$

The ratio of odds

$$\mathit{ratio}\left(Y \,|\, Z\right) := \frac{\mathcal{O}\left(Y \,|\, Z=1\right)}{\mathcal{O}\left(Y \,|\, Z=0\right)}$$

thus encapsulates the *impact of Z on Y*. (For example: Z does not have an impact on Y if and only if $\textit{ratio}\,(Y\,|\,Z) = 1$, and otherwise the sign of $\textit{ratio}\,(Y\,|\,Z) - 1$ tells us in which direction will observing $Z = 1$ impact the odds of Y.) In particular, in the notation of Definition 15.3 we may write

$$\mathcal{O}\,(Y\,|\,Z = 1) = \frac{\mathbb{P}\,(Y = 1\,|\,Z = 1)}{\mathbb{P}\,(Y = 0\,|\,Z = 1)} = \frac{\mathbb{P}\,(Y = 1 \text{ and } Z = 1)\,/\mathbb{P}\,(Z = 1)}{\mathbb{P}\,(Y = 0 \text{ and } Z = 1)\,/\mathbb{P}\,(Z = 1)} = \frac{p_{11}}{p_{10}}$$

and

$$\mathcal{O}\,(Y\,|\,Z = 0) = \frac{\mathbb{P}\,(Y = 1\,|\,Z = 0)}{\mathbb{P}\,(Y = 0\,|\,Z = 0)} = \frac{\mathbb{P}\,(Y = 1 \text{ and } Z = 0)\,/\mathbb{P}\,(Z = 0)}{\mathbb{P}\,(Y = 0 \text{ and } Z = 0)\,/\mathbb{P}\,(Z = 0)} = \frac{p_{01}}{p_{00}},$$

such that

$$\textit{ratio}\,(Y\,|\,Z) = \frac{\mathcal{O}\,(Y\,|\,Z = 1)}{\mathcal{O}\,(Y\,|\,Z = 0)} = \frac{p_{11}/p_{10}}{p_{01}/p_{00}} = \frac{p_{00}p_{11}}{p_{01}p_{10}} = \psi.$$

In other words: the odds ratio ψ is indeed a ratio of odds! Morever we can interpret the odds ratio as telling us about the impact that Z has on Y.

Theorem 15.6 (Characterization of the independence of two binary variables)**.** *Let Y and Z be two binary random variables. The following are equivalent.*

(1) $Y \perp\!\!\!\perp Z$, i.e. Y and Z are independent.
(2) The odds ratio ψ satisfies $\psi = 1$.
(3) The log odds ratio γ satisfies $\gamma = 0$.
(4) For $i, j \in \{0, 1\}$, $p_{ij} = p_{i\cdot}p_{\cdot j}$ where p denotes the two-by-two parameter associated with Y and Z.

Theorem 15.7 (Likelihood ratio test for the independence of two binary random variables)**.** *Let $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ be an IID sample drawn from two binary random variables Y and Z and let X be the corresponding two-by-two random variable. We consider the hypotheses*

$$H_0 : Y \perp\!\!\!\perp Z \textit{ versus } H_1 : Y \not\!\perp\!\!\!\perp Z.$$

Define the parametric model

$$\mathcal{F} := \left\{ \textit{Multinomial}(n,\, p) : p \in \Delta^3 \subseteq \mathbb{R}^4 \right\},$$

the sets

$$\Theta_0 := \left\{ p = (p_{00},\, p_{01},\, p_{10},\, p_{11}) \in \Delta^3 : \gamma(p) = 0 \right\},$$

where γ denotes the log odds ratio, and $\Theta_1 = \Delta^3$, the test statistic

$$\begin{aligned} T &:= 2 \sum_{i,j=0}^{1} X_{ij} \log \frac{X_{ij} X_{\cdot\cdot}}{X_{i\cdot} X_{\cdot j}} \\ &= 2 \left(X_{00} \log \frac{nX_{00}}{X_{0\cdot} X_{\cdot 0}} + X_{01} \log \frac{nX_{01}}{X_{0\cdot} X_{\cdot 1}} + X_{10} \log \frac{nX_{10}}{X_{1\cdot} X_{\cdot 0}} + X_{11} \log \frac{nX_{11}}{X_{1\cdot} X_{\cdot 1}} \right), \end{aligned}$$

and, for $\alpha \in (0, 1)$, the rejection region

$$R_\alpha := \left\{ t \in \mathbb{R} : t > \chi^2_{1,\,\alpha} \right\}.$$

The hypothesis test $(\mathcal{F},\, \Theta_0,\, \Theta_1,\, H_0,\, H_1,\, T,\, R_\alpha)$ is a likelihood ratio test. In particular its p–value satisfies

$$p - \textit{value}(\omega) \approx \mathbb{P}\left(\chi^2_1 > T(\omega) \right) \ \textit{ as } n \to \infty$$

for any outcome ω. (In the expression above the probability is taken over the distribution of χ^2_1, treating $T(\omega)$ as a fixed value.)

Theorem 15.8 (Pearson's χ^2 test for the independence of two binary random variables)**.** *Let $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ be an IID sample drawn from two binary random variables Y and Z and let X be the corresponding two-by-two random variable. We consider the hypotheses*

$$H_0 : Y \perp\!\!\!\perp Z \text{ versus } H_1 : Y \not\perp\!\!\!\perp Z.$$

Define the test statistic

$$\mathcal{U} := \sum_{i,j=0}^{1} \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$$

where

$$E_{ij} = \frac{X_{i\cdot} X_{\cdot j}}{n}$$

and, for $\alpha \in (0, 1)$, define the rejection region

$$R_\alpha := \left\{ u \in \mathbb{R} : u > \chi^2_{1,\,\alpha} \right\}.$$

The hypothesis test $(H_0, H_1, \mathcal{U}, R_\alpha)$, called Pearson's χ^2 test for the independence of two binary random variables, *has asymptotic size α as $n \to \infty$ and its p–value satisfies*

$$p - value(\omega) \approx \mathbb{P}\left(\chi^2_1 > \mathcal{U}(\omega)\right)$$

for any outcome ω.

Remark 15.9 (Pearson's χ^2 test for the independence of two binary random variables)**.** By contrast with a typical hypothesis test as introduced in Definition 10.1, Pearson's χ^2 test for the independence of two binary random variables introduced in Theorem 15.8 above does *not* involve a parametric model $\mathcal{F}$ or sets of parameters Θ_0 and Θ_1. In some sense we can thus view Pearson's χ^2 test for the independence of two binary random variables as a *non-parametric* test. This is to be nuanced by the fact that binary random variables can always by parametrized as Bernoulli random variables and by the parametric ideas used to motivate this test – see Remark 15.10.

Remark 15.10 (Motivating Pearson's χ^2 test for the independence of two binary random variables)**.** Under the null hypothesis that two binary random variables Y and Z are independent, Theorem 15.6 tells us that $p_{ij} = p_{i\cdot} p_{\cdot j}$. Under this constraint the MLE $\hat{p}$ for p is (see Exercise A.15.2)

$$\hat{p}_{ij} = \frac{X_{i\cdot} X_{\cdot j}}{n^2}.$$

We may thus construct a hypothesis test in the spirit of Pearson's χ^2 test for multinomial data, where we treat the MLE $\hat{p}$ as a true parameter. The corresponding test statistic would then indeed be

$$\sum_{i,j=0}^{1} \frac{(X_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} = \mathcal{U}$$

as desired since $n\hat{p}_{ij} = E_{ij}$.

Remark 15.11 (Limiting degrees of freedom of Pearson's χ^2 test for the independence of two binary random variables)**.** That the limiting distribution of the test statistic in Pearson's χ^2 test for the independence of two binary random variables be a χ^2 distribution is not particularly surprising since we might expect that

$$\frac{X_{ij} - E_{ij}}{\sqrt{E_{ij}}} \rightsquigarrow N(0, \sigma_{ij}^2) \text{ as } n \to \infty$$

for some $\sigma_{ij}^2 > 0$. $\mathcal{U}$ would then indeed be asymptotically a sum of squares of Normal distributions, i.e. a χ^2 distribution.

But why does this limiting χ^2 distribution only have *one* degree of freedom? To answer this we observe first that, asymptotically, tests built like Pearson's χ^2 test for multinomial data are often equivalent to likelihood ratio tests (see for example [Ken61] which discusses this in the chapters titled "Tests of Fit" and "Categorized Data").

The likelihood ratio test corresponding to Pearson's χ^2 test for the independence of two binary random variables considers, for $p := (p_{00}, p_{01}, p_{10}, p_{11})$ the two-by-two parameter associated with Y and Z, the full parameter space

$$\Theta_1 = \Delta^3 \subseteq \mathbb{R}^4$$

versus the restricted parameter space (see Theorem 15.7 and its proof in Exercise A.15.2 for details)

$$\Theta_0 = \left\{ p \in \Delta^3 : \gamma(p) = 0 \right\}$$

since indeed Theorem 15.6 tells us that Y and Z are independent if and only if the log odds ratio γ satisfies $\gamma = 0$. In particular, since $\gamma = 0$ is a single scalar constraint, we see that

$$\dim \Theta_0 = \dim \Delta^3 - 1 = 2$$

and hence the number of degrees of freedom of the limiting χ^2 distribution of the likelihood ratio statistic are

$$\dim \Theta_1 - \dim \Theta_0 = 3 - 2 = 1.$$

Since Pearson's χ^2 test for the independence of two binary random variables is asymptotically equivalent to this likelihood ratio test we conclude that indeed the limiting distribution of its test statistc will be a χ^2 distribution with *one* degree of freedom (as recorded in Theorem 15.8).

Crucially, note that this number of degrees of freedom corresponds here to the *number of constraints* imposed on p by the independence of Y and Z, which in the two-by-two case (i.e. for two binary random variables) comes down to a *single* scalar constraint (namely $\psi = 1$, where ψ is the odds ratio, or equivalently $\gamma = 0$, as established in Theorem 15.6).

Theorem 15.12 (MLE for the odds ratio)**.** *Let $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ be an IID sample drawn from two binary random variables Y and Z, let X be the corresponding two-by-two random variable, and let ψ and γ denote the corresponding odds ratio and log odds ratio, respectively. The MLE for ψ and γ are, respectively,*

$$\hat{\psi} = \frac{X_{00}X_{11}}{X_{01}X_{10}} \text{ and } \hat{\gamma} = \log \hat{\psi}.$$

Moreover

$$\widehat{se}(\hat{\psi}) = \hat{\psi}\widehat{se}(\hat{\gamma}) \text{ and } \widehat{se}(\hat{\gamma}) = \sqrt{\frac{1}{X_{00}} + \frac{1}{X_{01}} + \frac{1}{X_{10}} + \frac{1}{X_{11}}}$$

are asymptotic estimates of the standard errors of $\hat{\psi}$ and $\hat{\gamma}$, respectively, such that

$$\frac{\hat{\psi}-\psi}{\widehat{se}(\hat{\psi})} \rightsquigarrow N(0,\,1) \ \textit{and}\ \frac{\hat{\gamma}-\gamma}{\widehat{se}(\hat{\gamma})} \rightsquigarrow N(0,\,1) \ \textit{as}\ n \to \infty.$$

Remark 15.13 (MLE for the odds ratio)**.** The MLE for the odds ratio and log odds ratio recorded in Theorem 15.12 above may be used to determine not just whether or not two binary random variables are independent, but also *how strong the dependence* is.

Remark 15.14 (Modified estimator for the odds ratio for small sample size)**.** In the notation of Theorem 15.12, when the sample size n is small the MLE for ψ and γ can have very large variance. In that case we may then use the modified point estimators

$$\tilde{\psi} := \frac{\left(X_{00}+\frac{1}{2}\right)\left(X_{11}+\frac{1}{2}\right)}{\left(X_{01}+\frac{1}{2}\right)\left(X_{10}+\frac{1}{2}\right)}$$

for ψ and $\tilde{\gamma} := \log\tilde{\psi}$ for γ.

More precisely, these estimators may be obtained from a posterior mean $\tilde{p}$. Indeed, under the prior distribution

$$p \sim \mathrm{Dirichlet}(\alpha)$$

for $\alpha \in \mathbb{R}^k_{>0}$ the posterior distribution with respect to the categorical model is

$$p \,|\, X \sim \mathrm{Dirichlet}(\alpha + X).$$

In other words the Dirichlet distribution is conjugate with respect to the categorical model. The modified estimators $\tilde{\psi}$ and $\tilde{\gamma}$ then correspond to

$$\tilde{\psi} = \psi(\tilde{p}) \text{ and } \tilde{\gamma} = \gamma(\tilde{p})$$

for $\tilde{p}$ the posterior mean with respect to the Dirichlet prior with parameter

$$\alpha = \frac{1}{2}\mathbb{1} = \left(\frac{1}{2},\,\ldots,\,\frac{1}{2}\right) \in \mathbb{R}^k,$$

namely

$$\tilde{p} = \mathbb{E}\left[\mathrm{Dirichlet}(\alpha+X)\right] = \frac{\alpha+X}{\sum\limits_{i,j=0}^{1}(\alpha_{ij}+X_{ij})} =: \frac{\alpha+X}{c}$$

such that indeed

$$\tilde{\psi} = \frac{\tilde{p}_{00}\tilde{p}_{11}}{\tilde{p}_{01}\tilde{p}_{10}} = \frac{\frac{\alpha_{00}+X_{00}}{c}\cdot\frac{\alpha_{11}+X_{11}}{c}}{\frac{\alpha_{01}+X_{01}}{c}\cdot\frac{\alpha_{10}+X_{10}}{c}} = \frac{\left(\frac{1}{2}+X_{00}\right)\left(\frac{1}{2}+X_{11}\right)}{\left(\frac{1}{2}+X_{01}\right)\left(\frac{1}{2}+X_{10}\right)}$$

and $\tilde{\gamma} = \log\tilde{\psi}$. In particular note that $\mathrm{Dirichlet}(\mathbb{1})$ is the uniform distribution on the simplex Δ^{k-1}, so here the prior weighs more heavily the pure states e_j than the mixed states such as $\frac{1}{k}\mathbb{1}$.

Theorem 15.15 (Wald test for the independence of two binary random variables via the log odds ratio)**.** *Let $(Y_1,\,Z_1),\,\ldots,\,(Y_n,\,Z_n)$ be an IID sample drawn from two binary random variables Y and Z and let $\hat{\gamma}$ and $\widehat{se}(\hat{\gamma})$ denote the MLE for the log odds ratio γ and the asymptotic estimate of its standard error, respectively, as recorded in Theorem 15.12. We consider the hypotheses*

$$H_0 : Y \perp\!\!\!\perp Z \ \textit{versus}\ H_1 : Y \leftrightsquigarrow Z$$

and define the test statistic

$$W := \frac{\hat{\gamma}}{\widehat{se}(\hat{\gamma})}$$

and, for $\alpha \in (0, 1)$, the rejection region

$$R_\alpha := \left\{ w \in \mathbb{R} : |w| > z_{\alpha/2} \right\}$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ for Φ denoting the CDF of the standard normal distribution. The hypothesis test (H_0, H_1, W, R_α) is a size α Wald test.

Proof. This follows immediately from the asymptotic normality of the MLE (see Theorem 9.12) and from Theorem 15.6 which shows that Y and Z are independent if and only if the log odds ratio vanishes. □

Theorem 15.16 (Asymptotic confidence interval for the odds ratio and log odds ratio)**.** *Let $(Y_1, Z_1), \dots, (Y_n, Z_n)$ be an IID sample drawn from two binary random variables Y and Z, let ψ denote the corresponding odds ratio, let γ denote the corresponding log odds ratio, and let $\hat{\psi}$, $\hat{\gamma}$, $\widehat{se}(\hat{\psi})$, and $\widehat{se}(\hat{\gamma})$ denote their respective MLE and asymptotic estimates of their standard errors as recorded in Theorem 15.12. For any $\alpha \in (0, 1)$,*

$$\hat{\gamma} \pm z_{\alpha/2}\widehat{se}(\hat{\gamma}),$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ for Φ denoting the CDF of the standard normal distribution, is an asymptotic $1 - \alpha$ confidence interval for γ as $n \to \infty$. Moreover

$$\exp\left[\hat{\gamma} \pm z_{\alpha/2}\widehat{se}(\hat{\gamma})\right] \text{ and } \hat{\psi} \pm z_{\alpha/2}\widehat{se}(\hat{\psi})$$

are asymptotic $1 - \alpha$ confidence intervals for ψ as $n \to \infty$. In practice we often use the former since it is usually more accurate.

Remark 15.17 (Interpretation of the odds ratio confidence interval)**.** Recall from Remark 15.5 that we can write the odds ratio ψ as

$$\psi = \frac{\mathcal{O}\left(Y \mid Z = 1\right)}{\mathcal{O}\left(Y \mid Z = 0\right)} \text{ where } \mathcal{O}\left(Y \mid Z = i\right) = \frac{\mathbb{P}\left(Y = 1 \mid Z = i\right)}{\mathbb{P}\left(Y = 0 \mid Z = i\right)}.$$

So, for example, if a confidence interval for ψ is $(2, 4)$ then we can say (with some quantified confidence) that "given $Z = 1$, $Y = 1$ is two to four times more likely".

Remark 15.18 (Transformed confidence intervals)**.** In Theorem 14.2 we used

$$\tanh\left[\hat{\theta} \pm z_{\alpha/2}\widehat{se}(\hat{\theta})\right]$$

as a confidence interval for the correlation ρ and similarly in Theorem 15.16 we used

$$\exp\left[\hat{\gamma} \pm z_{\alpha/2}\widehat{se}(\hat{\gamma})\right]$$

as a confidence interval for the odds ratio ψ. In both cases a transformed interval of the form

$$f\left[\hat{\lambda} \pm z_{\alpha/2}\widehat{se}(\hat{\lambda})\right]$$

is used instead of an interval of the form

$$\hat{\mu} \pm z_{\alpha/2}\widehat{se}(\hat{\mu}).$$

The reasoning is similar in both cases: the original parameter ρ, ψ, or μ does *not* live in $\mathbb{R}$, but in a constrained subset, and yet its *transformation*

$$\theta = \operatorname{arctanh}\rho,\ \gamma = \exp\psi,\ \text{ or } \lambda = f(\mu)$$

does. Indeed: ρ lives in $[-1, 1]$ and ψ lives in $[0, \infty]$ but θ and γ live in $\mathbb{R}$. As long as the transformation is chosen appropriately, we may then often reasonably expect that the transformed parameter θ, γ, or λ asymptotically has a Normal distribution. We can then build asymptotic Normal confidence intervals

$$\hat{\theta} \pm z_{\alpha/2}\widehat{se}(\hat{\theta}),\ \hat{\gamma} \pm z_{\alpha/2}\widehat{se}(\hat{\gamma}),\ \text{and } \hat{\lambda} \pm z_{\alpha/2}\widehat{se}(\hat{\lambda})$$

and *then* transform these intervals to

$$\tanh\left[\hat{\theta} \pm z_{\alpha/2}\widehat{se}(\hat{\theta})\right],\ \exp\left[\hat{\gamma} \pm z_{\alpha/2}\widehat{se}(\hat{\gamma})\right],\ \text{and } f\left[\hat{\lambda} \pm z_{\alpha/2}\widehat{se}(\hat{\lambda})\right].$$

15.2. **Two Discrete Variables.**

Definition 15.19 (Contingency table)**.** Let $(Y_1, Z_1), \dots, (Y_n, Z_n)$ be an IID sample drawn from two discrete random variables Y and Z with finite codomains $\{0, \dots, J-1\}$ and $\{0, \dots, I-1\}$, respectively, where $I, J \geqslant 2$. This sample may be represented as an I-by-J table

	$Y=0$	$\dots$	$Y=j$	$\dots$	$Y=J$	
$Z=0$	X_{00}	$\dots$	X_{0j}	$\dots$	X_{0J}	$X_{0\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$Z=i$	X_{i0}	$\dots$	X_{ij}	$\dots$	X_{iJ}	$X_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$Z=I$	X_{I0}	$\dots$	X_{Ij}	$\dots$	X_{IJ}	$X_{I\cdot}$
	$X_{\cdot 0}$	$\dots$	$X_{\cdot j}$	$\dots$	$X_{\cdot J}$	$X_{\cdot\cdot} = n$

known as a *contingency table*, or as a *cross-tabulation*, where

$$X_{ij} = \text{ number of observations where } Z = i \text{ and } Y = j$$

and where dotted subscripts denote sums such that

$$X_{i\cdot} = \sum_{j=0}^{J-1} X_{ij},\ X_{\cdot j} = \sum_{i=0}^{I-1} X_{ij},\ \text{and } X_{\cdot\cdot} = \sum_{i,j=0}^{I-1,J-1} X_{ij} = n$$

(as in Definition 15.3). The underlying data-generating process may also be represented as an I-by-J table, namely

	$Y=0$	$\dots$	$Y=j$	$\dots$	$Y=J$	
$Z=0$	p_{00}	$\dots$	p_{0j}	$\dots$	p_{0J}	$p_{0\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$Z=i$	p_{i0}	$\dots$	p_{ij}	$\dots$	p_{iJ}	$p_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$Z=I$	p_{I0}	$\dots$	p_{Ij}	$\dots$	p_{IJ}	$p_{I\cdot}$
	$p_{\cdot 0}$	$\dots$	$p_{\cdot j}$	$\dots$	$p_{\cdot J}$	$p_{\cdot\cdot} = 1$

where

$$p_{ij} = \mathbb{P}(Z = i \text{ and } Y = j),\ p_{i\cdot} = \mathbb{P}(Z = i),\ p_{\cdot j} = \mathbb{P}(Y = j),\ \text{and } p_{\cdot\cdot} = 1.$$

In particular these two tables are related as follows: if we define

$$X := (X_{00}, \dots, X_{0J}, \dots, X_{I0}, \dots, X_{IJ}) \text{ and}$$
$$p := (p_{00}, \dots, p_{0J}, \dots, p_{I0}, \dots, p_{IJ}) \in \Delta^{IJ-1}$$

then we have that

$$X \sim \text{Multinomial}(n,\, p).$$

We call X the *I-by-J random variable associated with Y and Z* and call p the *I-by-J parameter associated with Y and Z.*

Remark 15.20 (Two-by-two tables are contingency tables)**.** When $I = J = 2$, an I-by-J contingency table is a two-by-two table (i.e. the terminology and notation are consistent).

Theorem 15.21 (Tests for the independence of two discrete random variables with finite codomain)**.** *Let $(Y_1,\, Z_1),\, \dots,\, (Y_n,\, Z_n)$ be an IID sample drawn from two discrete random variables Y and Z with finite codomains $\{0,\, \dots,\, J-1\}$ and $\{0,\, \dots,\, I-1\}$, respectively, where $I, J \geqslant 2$ and let X be the corresponding I-by-J random variable. We consider the hypotheses*

$$H_0 : Y \amalg Z \text{ versus } H_1 : Y \leftrightsquigarrow Z.$$

Define the parametric model

$$\mathcal{F} := \left\{ \textit{Multinomial}(n,\, p) : p \in \Delta^{IJ-1} \subseteq \mathbb{R}^{IJ} \right\},$$

the sets

$$\Theta_0 := \left\{ p = (p_{00},\, \dots,\, p_{0J},\, \dots,\, p_{I0},\, \dots,\, p_{IJ}) \in \Delta^{IJ-1} : p_{ij} = p_{i\cdot} p_{\cdot j} \text{ for all } i,\, j \right\},$$

and $\Theta_1 = \Delta^{IJ-1}$, the test statistics

$$T := 2 \sum_{i=0}^{I-1} \sum_{j=0}^{J-1} X_{ij} \log \frac{X_{ij} X_{\cdot\cdot}}{X_{i\cdot} X_{\cdot j}}$$

and

$$\mathcal{U} := \sum_{i=0}^{I-1} \sum_{j=0}^{J-1} \frac{\left(X_{ij} - E_{ij}\right)^2}{E_{ij}} \text{ where } E_{ij} = \frac{X_{i\cdot} X_{\cdot j}}{n},$$

and, for $\alpha \in (0,\, 1)$, the rejection region

$$R_\alpha := \left\{ s \in \mathbb{R} : s > \chi^2_{\nu,\, \alpha} \right\}$$

for $\nu := (I-1)(J-1)$.

- *The hypothesis test $(\mathcal{F},\, \Theta_0,\, \Theta_1,\, H_0,\, H_1,\, T,\, R_\alpha)$ is a likelihood ratio test. In particular its p–value satisfies*

  $$p - \textit{value}(\omega) \approx \mathbb{P}\left(\chi^2_\nu > T(\omega)\right) \text{ as } n \to \infty$$

  for any outcome ω.
- *The hypothesis test $(H_0,\, H_1,\, \mathcal{U},\, R_\alpha)$, called* Pearson's χ^2 test for the independence of two discrete random variables with finite codomains*, has asymptotic size α as $n \to \infty$ and its p–value satisfies*

  $$p - \textit{value}(\omega) \approx \mathbb{P}\left(\chi^2_\nu > \mathcal{U}(\omega)\right) \text{ as } n \to \infty$$

  for any outcome ω.

Remark 15.22 (Number of parameters characterizing the independence of discrete random variables)**.** Suppose that Y and Z are independent discrete random variables with finite codomains $\{0,\, \dots,\, J-1\}$ and $\{0,\, \dots,\, I-1\}$, respectively, where $I, J \geqslant 2$. Suppose moreover that they are *non-degenerate* in the sense that

$\mathbb{P}(Z = i)$, $\mathbb{P}(Y = j) > 0$ for all i and j. Then the joint distribution of (Y, Z) is characterized by

$$I + J - 2 \text{ parameters}$$

in [0, 1].

Indeed, suppose that p_{i0} and p_{0j} are given for all i, j. From these $I + J - 1$ parameters we will be able to determine all other probabilities p_{ij}. Actually, we will see that we can do so with *one redundancy*, which will thus lower the number of truly characteristic parameters to

$$(I + J - 1) - 1 = I + J - 2.$$

Well, note that for all $i, j > 0$, by independence we have that

$$p_{ij} = p_{i\cdot} p_{\cdot j} = \frac{p_{i0} p_{0j}}{p_{\cdot 0} p_{0\cdot}}$$

where *all* of the quantities on the right-hand side may be computed in terms of p_{i0} and p_{0j}. So indeed we can determine all of the probabilities p_{ij}. However there is a redundancy. Indeed, we must have that these probabilities sum to one and so

$$1 = \sum_{i,j} p_{ij} = p_{00} + p_{0\cdot} + p_{\cdot 0} + \sum_{i,j>0} p_{ij}$$

where

$$\begin{aligned} \sum_{i,j>0} p_{ij} &= \frac{\sum_{i,j=0} p_{i0} p_{0j}}{p_{\cdot 0} p_{0\cdot}} \\ &= \frac{\left(\sum_{i>0} p_{i0}\right) \left(\sum_{j>0} p_{0j}\right)}{p_{\cdot 0} p_{0\cdot}} \\ &= \frac{(p_{\cdot 0} - p_{00})(p_{0\cdot} - p_{00})}{p_{\cdot 0} p_{0\cdot}} \end{aligned}$$

such that

$$p_{00} + p_{0\cdot} + p_{\cdot 0} = 1 - \frac{(p_{\cdot 0} - p_{00})(p_{0\cdot} - p_{00})}{p_{\cdot 0} p_{0\cdot}}.$$

Crucially: the right-hand side is invariant under the rescaling

$$(p_{i0},\, p_{0j}) \mapsto (s p_{i0},\, s p_{0j}),$$

for $s > 0$, of the initial $I + J - 1$ parameters while the left-hand side scales proportionally to s. We may thus rescale these parameters to enforce the condition

$$\sum_{i,j} p_{ij} = 1,$$

showing that there was *one scalar redundancy* in our $I + J - 1$ parameters, which means that there are actually only

$$I + J - 2 \text{ parameters}$$

which may be freely chosen.

Pictorially we used the black dots below to determine the white dots: (here $I = 5$ and $J = 3$)

	$Y=0$	$Y=1$	$Y=2$
$Z=0$	●	●	●
$Z=1$	●	○	○
$Z=2$	●	○	○
$Z=3$	●	○	○
$Z=4$	●	○	○

but in doing so we *overdetermined* the problem and had to then rescale our initial choices (rescaling all values by the same *single* scalar multiple) in order to satisfy

$$\sum_{i,j} p_{ij} = 1.$$

Remark 15.23 (Asymptotic degrees of freedom in Pearson's test for the independence of two discrete random variables with finite codomains)**.** Why does the test statistic for Pearson's test for the independence of two discrete random variables with finite codomains, recorded in Theorem 15.21 above, asymptotically have under the null hypothesis a χ^2 distribution with $\nu = (I-1)(J-1)$ degrees of freedom?

Well, as discussed in Remark 15.11, tests built like Pearson's χ^2 test for multinomial data are often equivalent to likelihood ratio tests. Since the corresponding likelihood ratio test is also recorded in Theorem 15.21 it thus suffices to explain why the limiting distribution of the likelihood ratio statistic is a χ^2 distribution with ν degrees of freedom.

This is fairly straighforward in light of the counting carried out in Remark 15.22 above. Indeed, in the notation of Theorem 15.21 we have that

$$\Theta_1 = \Delta^{IJ-1} \subseteq \mathbb{R}^{IJ}$$

while, by Remark 15.22,

$$\dim \Theta_0 = I + J - 2.$$

Therefore

$$\dim \Theta_1 - \dim \Theta_0 = (IJ-1) - (I+J-2) = IJ - I - J + 1 = (I-1)(J-1).$$

As before (see Remark 15.11), $(I-1)(J-1)$ represents a *number of constraints* (see also Remark 15.24 below).

Remark 15.24 (Number of constraints characterizing the independence of discrete random variables)**.** In the two-by-two case, Theorem 15.6 told us that Y and Z were independent if p satisfied *one* scalar condition (e.g. $\psi = 1$, or equivalently $\gamma = 0$). Since otherwise $p \in \Delta^3$ this means that the space of independent two-by-two distributions is 2–dimensional since

$$2 = \dim \Delta^3 - \text{number of constraints} = 3 - 1.$$

We can use Remark 15.22 to reverse-engineer this process in the case of I-by-J distributions. In that case $p \in \Delta^{IJ-1}$ and the space of these distributions has dimension $I + J - 2$, so the number of constraints must be

$$\begin{aligned} &\text{number of constraints} \\ &= \dim \Delta^{IJ-1} - \text{ dimension of space of independent } I\text{-by-}J \text{ distributions} \\ &= (IJ-1) - (I+J-2) = IJ - I - J + 1 = (I-1)(J-1). \end{aligned}$$

In other words, if we ever found an analog of Theorem 15.6 for the I-by-J case then we know that independence would be characterized by $(I-1)(J-1)$–many scalar conditions (even though we do not know what these conditions would be, precisely).

15.3. **Two Continuous Variables.**

Remark 15.25 (Testing the independence of two continuous random variables via their correlation)**.** Given an IID sample $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ from two continuous random variables Y and Z then we may measure the dependence between Y and Z via their correlation, which may be estimated as in Theorem 14.2.

- If Y and Z are *correlated*, i.e. $\rho(Y, Z) \neq 0$, then Y and Z are *dependent*.
- If Y and Z are *uncorrelated*, i.e. $\rho(Y, Z) = 0$, then we *cannot* deduce that they are independent – see Example 15.26 below.

Example 15.26 (Uncorrelated but dependent random variables)**.** Consider the continuous random variable $Y \sim$ Uniform$(-1, 1)$ and define $Z := Y^2$. Then Y and Z are uncorrelated since $\mathbb{E}Y = 0$ and so

$$\mathrm{Cov}(Y, Z) = \mathbb{E}(YZ) - \mathbb{E}(Y)\mathbb{E}(Z) = \mathbb{E}\left(Y^3\right) = \int y^3 f(y)dy = \frac{1}{2}\int_{-1}^{1} y^3 dy = 0,$$

since $g(y) = y^3$ is an odd function. However Y and Z are *not* independent since, for example,

$$Z \leqslant \frac{1}{4} \iff |Y| \leqslant \frac{1}{2}$$

and so

$$\mathbb{P}\left(|Y| > \frac{1}{2} \,\middle|\, Z \leqslant \frac{1}{4}\right) = 0 \neq \frac{1}{2} = \mathbb{P}\left(|Y| > \frac{1}{2}\right).$$

15.4. **One Continuous Variable and One Discrete.**

Definition 15.27 (Two-sample Kolmogorov-Smirnov test)**.** Let Y and Z be continuous random variables with CDF F_Y and F_Z, respectively. Let $Y_1, \ldots, Y_n$ and $Z_1, \ldots, Z_m$ be IID samples from Y and Z, respectively, and let $\hat{F}_{Y,n}$ and $\hat{F}_{Z,n}$ denote the corresponding empirical distribution functions. Consider the hypotheses

$$H_0 : F_Y = F_Z \text{ versus } H_1 : F_Y \neq F_Z$$

and define the test statistic

$$D := \sup_{x \in \mathbb{R}} \left|\hat{F}_{Y,\,n}(x) - \hat{F}_{Z,\,m}(x)\right|$$

and, for $\alpha \in (0, 1)$, the rejection region

$$R_\alpha := \left\{ d \in \mathbb{R} : \sqrt{\frac{nm}{n+m}}\, d > H^{-1}(1-\alpha) \right\}$$

where

$$H(t) := 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2t^2}$$

denotes the CDF of the *Kolmogorov distribution*

$$K := \sup_{t \in [0,\,1]} |B_t|$$

where B_t denotes the *unit Brownian bridge*, for which $B_0 = B_1 = 0$. The tuple (H_0, H_1, D, R_α) is called the *two-sample Kolmogorov-Smirnov test.*

Theorem 15.28 (Asymptotic distribution of the two-sample Kolmogorov-Smirnov test statistic)**.** *In the notation of Definition 15.27 the two-sample Kolmogorov-Smirnov test statistic* D *satisfies*

$$\sqrt{\frac{nm}{n+m}}D \rightsquigarrow K \text{ as } n,\, m \to \infty \text{ under the null } H_0.$$

Corollary 15.29 (Properties of the two-sample Kolmogorov-Smirnov test)**.** *The two-sample Kolmogorov-Smirnov test has size* α *asymptotically as* $n,\, m \to \infty$ *and its* p*–value is, in the notation of Definition 15.27,*

$$p - value = 1 - H\left(\sqrt{\frac{nm}{n+m}}D\right),$$

i.e.

$$p - value(\omega) = \mathbb{P}\left(K > \sqrt{\frac{nm}{n+m}}D(\omega)\right)$$

for any outcome ω*.*

Theorem 15.30 (Kolmogorov-Smirnov test for the independence of a continuous and a binary random variable)**.** *Let* $(Y_1,\, Z_1),\, \dots,\, (Y_n,\, Z_n)$ *be an IID sample drawn from a binary random variable* Y *and a continuous random variable* Z*. For* $j = 0, 1$ *let* $\hat{F}_j$ *denote the empirical distribution function of* $Z \,|\, Y = j$*, i.e.*

$$\hat{F}_j(z) = \frac{1}{n_j}\sum_{i=1}^{n} \mathbb{1}(Z_i \leqslant z)\mathbb{1}(Y_i = j)$$

where

$$n_j = \sum_{i=1}^{n} \mathbb{1}(Y_i = j) = (\text{number of sample points with } Y = j).$$

We consider the hypotheses

$$H_0 : Y \perp\!\!\!\perp Z \text{ versus } H_1 : Y \not\perp\!\!\!\perp Z$$

and define the test statistic

$$D := \sup_{x \in \mathbb{R}} \left|\hat{F}_0(x) - \hat{F}_1(x)\right|$$

and, for $\alpha \in (0,\, 1)$*, the rejection region*

$$R_\alpha := \left\{d \in \mathbb{R} : \sqrt{\frac{n_0 n_1}{n_0 + n_1}}D > H^{-1}(1-\alpha)\right\}$$

for H *the CDF of the Kolmogorov distribution as in Definition 15.27. The tuple* $(H_0,\, H_1,\, D,\, R_\alpha)$ *is a two-sample Kolmogorov-Smirnov test for the null* $F_0 = F_1$*, which is equivalent to* H_0*. In particular its* p*–value satisfies*

$$p - value = 1 - H\left(\sqrt{\frac{n_0 n_1}{n_0 + n_1}}D\right).$$

Remark 15.31 (What about non-binary discrete random variables?)**.** Theorem 15.30 above tells us how to test whether or not a continuous random variable and a binary random variable are independent. But what if the discrete random variable in question is not merely binary, but instead has a codomain with more than two elements (while still finite)? In that case there are two approaches we can take, in either case denoting by Z the continuous variable and by Y the discrete variable with finite codomain $\{0,\, \dots,\, J-1\}$.

- We can define the conditional CDF
$$F_j := \mathbb{P}\left(Z \leqslant \cdot \,|\, Y = j\right)$$
for $0 \leqslant j \leqslant J-1$, then perform the $\binom{J}{2}$–many tests for $F_{j_1} = F_{j_2}$ by using the Kolmogorov-Smirnov test of Theorem 15.30 every time, with a Benjamini-Hochberg correction for multiple testing with *non-independent* hypotheses as in Theorem 10.41.
- We can use results from the modern litterature to extend the Kolmogorov-Smirnov test from two to J samples, as is done for example in [BH10].

16. Causal Inference

16.1. The Counterfactual Model.

Definition 16.1 (Consistency relationship and potential outcomes)**.** Let X, Y, C_0, and C_1 be binary random variables. We say that the tuple (X, Y, C_0, C_1) is *consistent* if the *consistency relationship*

$$Y = C_X$$

is satisfied. In that case we will call C_0 and C_1 *potential outcomes.*

Definition 16.2 (Counterfactual)**.** Let (X, Y, C_0, C_1) be a consistent tuple of binary random variables.

- If $X = 0$ then we say that C_1 is *counterfactual.*
- If $X = 1$ then we say that C_0 is *counterfactual.*

We refer to the random vector $C = (C_0, C_1)$ as the *counterfactual vector.*

Remark 16.3 (Interpretation of the potential outcomes)**.** Let (X, Y, C_0, C_1) be a consistent tuple of binary random variables and let us fix an outcome ω.

- If $X(\omega) = 0$ then $Y(\omega) = C_{X(\omega)}(\omega) = C_0(\omega)$. This means that $C_0(\omega)$ is the *actual* outcome while $C_1(\omega)$ is the counterfactual which is *not* observed (since in practice we only observe X and Y).
- If $X(\omega) = 1$ then $Y(\omega) = C_1(\omega)$. Now $C_1(\omega)$ is the actual outcome and $C_0(\omega)$ is the counterfactual which fails to be observed.

In general, since the value of X is not known, we do not know which of C_0 or C_1 will be observed (and which will be counterfactual). We thus refer to both C_0 and C_1 as *potential* outcomes.

In particular this terminology implicitly assumes that X is a treatment or exposure variable, which in practice can be interved upon (by treating/exposing subjects, or not), while Y is an outcome variable, where in practice one of the outcomes (say, $Y = 1$) is preferable to the other (say, $Y = 0$).

Remark 16.4 (Types of subjects)**.** Let (X, Y, C_0, C_1) be a consistent tuple of binary random variables which we interpret as follows:

$$\begin{aligned} X = 1 &\iff \text{subject is treated and} \\ X = 0 &\iff \text{subject is \textit{not} treated} \end{aligned}$$

while

$$\begin{aligned} Y = 1 &\iff \text{subject lives and} \\ Y = 0 &\iff \text{subject dies.} \end{aligned}$$

Now let us fix an outcome ω. Even though one of $C_0(\omega)$ or $C_1(\omega)$ is counterfactual (which we interpret as "not being observed"), both $C_0(\omega)$ and $C_1(\omega)$ remain well-defined values. We can then classify the outcomes (or *subjects*) ω according to the values of C_0 and C_1 (even though the true *type* of a subject can never actually be observed since the counterfactual can never be observed).

	C_0	C_1
survivor	1	1
responder	0	1
anti-responder	1	0
doomed	0	0

We now expound on each type of subject.

- Survivors will always live, whether they are treated or not.
- Responders will live if and only if they are treated.
- Anti-responders will (counter-intuitively) live if and only if they are *not* treated.
- Doomed subjects will always die, whether they are treated or not.

Remark 16.5 (Causal effect)**.** The interest in consistent tuples (X, Y, C_0, C_1) of binary random variables, interpreted in terms of treatment and survival as in Remark 16.4, lies in answering the following question.

Does treatment increase the probability of survival?

This question boils down to the following inequality:

$$\mathbb{P}(C_1 = 1) > \mathbb{P}(C_0 = 1)$$

(noting that $\mathbb{P}(C_j = 1)$ fully characterizes the distribution of C_j since C_j is binary.)

We therefore say that X *has a causal effect on* Y precisely when $\mathbb{P}(C_1 = 1)$ and $\mathbb{P}(C_0 = 1)$ are *distinct*. (Typically we ask more specifically that $\mathbb{P}(C_1 = 1) > \mathbb{P}(C_0 = 1)$, not merely that $\mathbb{P}(C_1 = 1) = \mathbb{P}(C_0 = 1)$.) To quantify this causal effect we thus seek a statistical functional $T = T(C_0, C_1)$ (which depends on the *distributions* of C_0 and C_1) such that, for some fixed critical value T_*,

$$T(C_0, C_1) > T_* \iff \mathbb{P}(C_1 = 1) > \mathbb{P}(C_0 = 1).$$

Definition 16.6 (Measures of causal effect)**.** Let (X, Y, C_0, C_1) be a consistent tuple of binary random variables.

- $\theta := \mathbb{E}(C_1) - \mathbb{E}(C_0)$ is called the *average causal effect*.
- $\theta_o := \frac{\mathbb{P}(C_1=1)}{\mathbb{P}(C_1=0)} \div \frac{\mathbb{P}(C_0=1)}{\mathbb{P}(C_0=0)}$ is called the *causal odds ratio*.
- $\theta_r := \frac{\mathbb{P}(C_1=1)}{\mathbb{P}(C_0=1)}$ is called the *causal relative risk*.

Remark 16.7 (Measures of causal effect quantify causal effect)**.** The average causal effect θ satisfies, since C_0 and C_1 are binary,

$$\theta = \mathbb{E}(C_1) - \mathbb{E}(C_0) = \mathbb{P}(C_1 = 1) - \mathbb{P}(C_0 = 1).$$

Therefore

$$\theta > 0 \iff \mathbb{P}(C_1 = 1) > \mathbb{P}(C_0 = 1)$$

and so the average causal effect θ does indeed quantify the causal effect of X on Y in the sense of Remark 16.5.

Similarly, since $l \mapsto \frac{l}{1+l}$ is the strictly increasing inverse of $p \mapsto \frac{p}{1-p}$ (as long as $0 < p < 1$) we deduce that the causal odds ratio θ_o satisfies

$$\begin{aligned}\theta_o > 1 &\iff \frac{\mathbb{P}(C_1 = 1)}{\mathbb{P}(C_1 = 0)} > \frac{\mathbb{P}(C_0 = 1)}{\mathbb{P}(C_0 = 0)} \\ &\iff \frac{\mathbb{P}(C_1 = 1)}{1 - \mathbb{P}(C_1 = 1)} > \frac{\mathbb{P}(C_0 = 1)}{1 - \mathbb{P}(C_0 = 1)} \\ &\iff \mathbb{P}(C_1 = 1) > \mathbb{P}(C_0 = 1),\end{aligned}$$

which verifies that the causal odds ratio θ_o also quantifies the causal effect of X on Y.

Finally we may observe immediately that the causal relative risk θ_r satisfies

$$\theta_r > 1 \iff \mathbb{P}(C_1 = 1) > \mathbb{P}(C_0 = 1),$$

which verifies that θ_r also quantifies the causal effect of X on Y.

Definition 16.8 (Association)**.** Let (X, Y, C_0, C_1) be a consistent tuple of binary random variables. We call

$$\alpha := \mathbb{E}\,(Y \,|\, X = 1) - \mathbb{E}\,(Y \,|\, X = 0)$$

the (average) *association*.

Remark 16.9 (Other measures of association)**.** Let (X, Y, C_0, C_1) be a consistent tuple of binary random variables and let α be the corresponding (average) association. Since Y is binary we may rewrite

$$\begin{aligned}\alpha &= \mathbb{E}\,(Y \,|\, X = 1) - \mathbb{E}\,(Y \,|\, X = 0)\\ &= \mathbb{P}\,(Y = 1 \,|\, X = 1) - \mathbb{P}\,(Y = 1 \,|\, X = 0)\end{aligned}$$

such that

$$\alpha > 0 \iff \mathbb{P}\,(Y = 1 \,|\, X = 1) > \mathbb{P}\,(Y = 1 \,|\, X = 0)\,.$$

In the spirit of Remarks 16.5 and 16.7 we may then proceed as in Definition 16.6 and find other *measures of association* which characterize

$$\mathbb{P}\,(Y = 1 \,|\, X = 1) > \mathbb{P}\,(Y = 1 \,|\, X = 0)\,.$$

For example we can mimick the causal odds ratio θ_o and define

$$\alpha_o := \frac{\mathbb{P}\,(Y = 1 \,|\, X = 1)}{\mathbb{P}\,(Y = 0 \,|\, X = 1)} \div \frac{\mathbb{P}\,(Y = 1 \,|\, X = 0)}{\mathbb{P}\,(Y = 0 \,|\, X = 0)}.$$

Proceeding as in Remark 15.5 we then see that α_r is precisely the odds ratio! (Where here (Y, X) is used instead of (Y, Z).)

Similarly we could mimick the causal relative risk θ_r and define the *relative risk*

$$\alpha_r := \frac{\mathbb{P}\,(Y = 1 \,|\, X = 1)}{\mathbb{P}\,(Y = 0 \,|\, X = 1)} = \frac{\mathbb{P}\,(Y = 1,\, X = 1)}{\mathbb{P}\,(Y = 0,\, X = 1)}.$$

Remark 16.10 (Causation and association are not generically related)**.** We provide here several examples of consistent tuples (X, Y, C_0, C_1) of binary random variables which show that the average causal effect θ is not related to the association α (barring additional assumptions). More precisely:

- if $\theta = 0$ then α need not vanish and may have either sign
- if $\alpha = 0$ then θ need not vanish and may have either sign
- if $\theta,\, \alpha \neq 0$ then their sign need not agree.

In other words:

- without causation, association may be present in either direction,
- with causation in either direction, association may not be present, and
- with causation in either direction, association may be present in the *other* direction.

The key observation used in constructing the examples below is that we may write the association α without any reference to Y since, in light of the consistency relationship $Y = C_X$,

$$\begin{aligned}\alpha &= \mathbb{E}\,(Y \,|\, X = 1) - \mathbb{E}\,(Y \,|\, X = 0)\\ &= \mathbb{E}\,(C_1 \,|\, X = 1) - \mathbb{E}\,(C_0 \,|\, X = 0)\\ &= \mathbb{P}\,(C_1 = 1 \,|\, X = 1) - \mathbb{P}\,(C_0 = 1 \,|\, X = 0)\,.\end{aligned}$$

This means that

$$\operatorname{sign}\alpha = \operatorname{sign}\Big[\mathbb{P}\left(C_1 = 1 \,|\, X = 1\right) - \mathbb{P}\left(C_0 = 1 \,|\, X = 0\right)\Big]$$

while

$$\operatorname{sign}\theta = \operatorname{sign}\Big[\mathbb{P}\left(C_1 = 1\right) - \mathbb{P}\left(C_0 = 1\right)\Big].$$

In particular the random variable Y is absent from these identities, which makes sense since $Y = C_X$ and so Y may be reconstructed from X and the potential outcomes C_0 and C_1. We will therefore omit Y in the examples below.

This is note a mere convenience: the omission of Y highlights that the relation between the causal effect (measured by θ) and the association (measured by α) comes down to how the distributions of X and (C_0, C_1) relate to each other.

- Example 1: $\theta = 0$ but $\alpha > 0$.
  Consider the following population of *two* subjects.

X	C_0	C_1
0	0	0
1	1	1

  Then

$$\theta = \mathbb{P}\left(C_1 = 1\right) - \mathbb{P}\left(C_0 = 1\right) = \frac{1}{2} - \frac{1}{2} = 0$$

  while

$$\alpha = \mathbb{P}\left(C_1 = 1 \,|\, X = 1\right) - \mathbb{P}\left(C_0 = 1 \,|\, X = 0\right) = 1 - 0 > 0.$$

- Example 2: $\theta = 0$ but $\alpha < 0$.
  Consider the following population of two subjects.

X	C_0	C_1
1	0	0
0	1	1

  Then $\theta = 0$ as in example 1 while now

$$\alpha = \mathbb{P}\left(C_1 = 1 \,|\, X = 1\right) - \mathbb{P}\left(C_0 = 1 \,|\, X = 0\right) = 0 - 1 < 0.$$

- Example 3: $\theta > 0$ but $\alpha = 0$.
  Consider the following population of two subjects.

X	C_0	C_1
0	0	1
1	0	0

  Then

$$\theta = \mathbb{P}\left(C_1 = 1\right) - \mathbb{P}\left(C_0 = 1\right) = \frac{1}{2} - 0 > 0$$

  while

$$\alpha = \mathbb{P}\left(C_1 = 1 \,|\, X = 1\right) - \mathbb{P}\left(C_0 = 1 \,|\, X = 0\right) = 0 - 0 = 0.$$

- Example 4: $\theta < 0$ but $\alpha = 0$.
  Consider the following population of two subjects.

X	C_0	C_1
0	0	0
1	1	0

Then

$$\theta = \mathbb{P}\left(C_1 = 1\right) - \mathbb{P}\left(C_0 = 1\right) = 0 - \frac{1}{2} < 0$$

while $\alpha = 0$ as in example 3.

- Example 5: $\theta < 0$ but $\alpha > 0$.
  Consider the following population of *three* subjects.

X	C_0	C_1
1	1	1
0	1	0
0	0	0

Then

$$\theta = \mathbb{P}\left(C_1 = 1\right) - \mathbb{P}\left(C_0 = 1\right) = \frac{1}{3} - \frac{2}{3} < 0$$

while

$$\alpha = \mathbb{P}\left(C_1 = 1 \,|\, X = 1\right) - \mathbb{P}\left(C_0 = 1 \,|\, X = 0\right) = 1 - \frac{1}{2} > 0.$$

- Example 6: $\theta > 0$ but $\alpha < 0$.
  Consider the following population of three subjects.

X	C_0	C_1
1	0	0
1	0	1
0	1	1

Then

$$\theta = \mathbb{P}\left(C_1 = 1\right) - \mathbb{P}\left(C_0 = 1\right) = \frac{2}{3} - \frac{1}{3} > 0$$

while

$$\alpha = \mathbb{P}\left(C_1 = 1 \,|\, X = 1\right) - \mathbb{P}\left(C_0 = 1 \,|\, X = 0\right) = \frac{1}{2} - 1 < 0.$$

Theorem 16.11 (Estimation of the causal effect)**.** *Let $(X,\, Y,\, C_0,\, C_1)$ be a consistent tuple of binary random variables, let θ be the corresponding average causal effect, and let α be the corresponding association. If X is independent of $(C_0,\, C_1)$ then $\alpha = \theta$.*

In that case, if $\left(X_1,\, Y_1\right),\, \dots,\, \left(X_n,\, Y_n\right)$ is an IID sample from $(X,\, Y)$ then

$$\hat{\theta} := \hat{\mathbb{E}}\left(Y \,|\, X = 1\right) - \hat{\mathbb{E}}\left(Y \,|\, X = 0\right),$$

where, for $j = 0, 1$,

$$\hat{\mathbb{E}}\left(Y \,|\, X = j\right) = \frac{1}{n_j} \sum_{i=1}^{n} Y_i \mathbb{1}(X_i = j) \text{ for } n_j := \sum_{i=1}^{n} \mathbb{1}(X_i = j),$$

is a consistent point estimator of θ. In particular note that

$$n_0 = \sum_{i=1}^{n} \left(1 - X_i\right),\, \hat{\mathbb{E}}\left(Y \,|\, X = 0\right) = \frac{1}{n_0} \sum_{i=1}^{n} Y_i \left(1 - X_i\right),$$

$$n_1 = \sum_{i=1}^{n} X_i, \text{ and } \hat{\mathbb{E}}\left(Y \,|\, X = 1\right) = \frac{1}{n_1} \sum_{i=1}^{n} Y_i X_i.$$

Remark 16.12 (Random assignment)**.** In practice the independence of X and (C_0, C_1) required by Theorem 16.11 is achieved assigning the subjects (who have an inherent and unknowable distribution of (C_0, C_1)) uniformly at random to the treatment $(X = 1)$ and non-treatment $(X = 0)$ groups. This is known as *random assignment.*

Definition 16.13 (Conditional causal effect)**.** Let (X, Y, C_0, C_1) be a consistent tuple of binary random variables and let Z be a random variable. For every $z \in \mathbb{R}$ we call

$$\theta_z := \mathbb{E}\left(C_1 \mid Z = z\right) - \mathbb{E}\left(C_0 \mid Z = z\right)$$

the *conditional causal effect.*

Remark 16.14 (Conditional causal effect and random assignment)**.** We may adapt Theorem 16.11 to show that if X and (C_0, C_1) are independent then the conditional causal effect satisfies

$$\theta_z = \mathbb{E}\left(Y \mid X = 1,\, Z = z\right) - \mathbb{E}\left(Y \mid X = 0,\, Z = z\right),$$

where the right-hand side can be estimated in the spirit of Theorem 16.11 with appropriate sample means. In particular, as discussed in Remark 16.12, the independence the covariate X and the counterfactual vector (C_0, C_1) can be enforced by *random assignment.*

16.2. **Beyond Binary Treatments.**

Definition 16.15 (Random function)**.** Let $\mathcal{X}$ and $\mathcal{Y}$ be sets and let $\mathcal{M}(\mathcal{Y})$ denote the set of random variables with codomain $\mathcal{Y}$. A map $f : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ is called a *random function from $\mathcal{X}$ to $\mathcal{Y}$*, or a *stochastic process.*

This means that, for every $x \in \mathcal{X}$, $f(x)$ is a random variable with codomain $\mathcal{Y}$. In particular $f(x)(\omega) \in \mathcal{Y}$ for every outcome ω and so $f(\,\cdot\,)(\omega)$ is a function from $\mathcal{X}$ to $\mathcal{Y}$; this is where the name *random function* comes from.

Definition 16.16 (Consistency relationship for non-binary random variables)**.** Let X and Y be random variables with codomains $\mathcal{X}$ and $\mathcal{Y}$, respectively, and let C be a random function from $\mathcal{X}$ to $\mathcal{Y}$. We say that the tuple (X, Y, C) is *consistent* if the *consistency relationship*

$$Y = C(X)$$

is satisfied, where $C(X)|_\omega = C\left(X(\omega)\right)(\omega)$ for every outcome ω. In that case we call C a *counterfactual function.*

Definition 16.17 (Causal regression function)**.** Let (X, Y, C) be a consistent tuple where X and Y have codomains $\mathcal{X}$ and $\mathcal{Y}$, respectively. The function $\theta : \mathcal{X} \to \mathcal{Y}$ defined by

$$\theta(x) := \mathbb{E}\left[C(x)\right]$$

is called the *causal regression function.*

Remark 16.18 (Causal regression function and average causal effect)**.** Let (X, Y, C) be a consistent tuple where X and Y are binary random variables. In order to compare the causal regression function with the average causal effect we will write interchangeably

$$C(0) =: C_0 \text{ and } C(1) =: C_1.$$

The key point is this: the causal regression function $\theta(x)$ is *not* a generalization of the average causal effect θ. Instead they are related, here, via

$$\theta = \mathbb{E}\left[C_1\right] - \mathbb{E}\left[C_0\right] = \mathbb{E}\left[C(1)\right] - \mathbb{E}\left[C(0)\right] = \theta(1) - \theta(0).$$

In particular the average causal effect may be recovered from the causal regression function but the converse fails.

Remark 16.19 (Association for non-binary random variables)**.** When dealing with binary random variables, the quantity of interest is the average causal effect. However, since that quantity typically cannot be estimated directly, a proxy is used instead: the association. (Careful: as Remark 16.10 shows, this proxy may in general have *no relation* with the true causal effect.)

Now, when dealing with *non-binary* random variables, the function of interest is the causal regression function

$$\theta(x) = \mathbb{E}\left[C(x)\right].$$

As is the case with the average causal effect, the causal regression function cannot be estimated directly and so we seek a proxy for it so that it *can* be estimated. We have already encountered such a proxy before, namely the regression function

$$r(x) = \mathbb{E}\left(Y \mid X = x\right).$$

In particular note that in light of the consistency relationship $Y = C(X)$ we can write the regression function in terms of the counterfactual function as

$$r(x) = \mathbb{E}\left[C(x) \mid X = x\right].$$

Theorem 16.20 (Estimation of the causal regression function)**.** *Let (X, Y, C) be a consistent tuple, let θ be the corresponding causal regression function, and let r be the regression function between Y and X. If X is independent of C then $r = \theta$.*

Remark 16.21 (Random "intervention")**.** As in the binary case (see Remark 16.12), the independence of X and C required by Theorem 16.20 may be achieved in practice by assigning the "intervention" values $X = x$ *uniformly at random.*

This is more complex that the random assignment discussed in Remark 16.12 since now X may take more than two values. In particular it is not clear how to do this when the codomain of X is infinite.

16.3. **Observational Studies and Confounding.**

Definition 16.22 (Observational study)**.** Let (X, Y, C) be a consistent tuple. If X is *not* independent of C then we say that this tuple is an *observational study.*

Remark 16.23 (Observational study)**.** Theorem 16.20 tells us that the regression function agrees with the causal regression function if the covariate X and the counterfactual function C are independent. In observational studies this fails, and so the causal regression function cannot, in general, be estimated.

Nonetheless, observational studies are common in practice, for example if random assignment (see Remark 16.12) is impossible or impractical. They rely on *confounding variables*, defined below, to be able to estimate the causal regression function.

Definition 16.24 (Confounding variable)**.** Let (X, Y, C) be a consistent tuple and let Z be a random variable. If

$$(C \amalg X) \mid Z$$

then we call Z a *confounding variable* and we say that *there is no unmeasured confounding.*

Theorem 16.25 (Estimation of the causal regression function when there is no unmeasured confounding)**.** *Let (X, Y, C) be a consistent tuple, let θ be the corresponding causal regression function, and let Z be a random variable under which there is no unmeasured confounding. Then*

$$\theta(x) = \int \mathbb{E}\left(Y \mid X = x,\, Z = z\right) dF_Z(z)$$

and we call this identity the adjusted treatment effect*.*

Now let $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$ be an IID sample drawn from (X, Y, Z). If $\hat{r}(x, z)$ is a consistent point estimator of the regression function

$$r(x, z) = \mathbb{E}\left(Y \mid X = x,\, Z = z\right)$$

then a consistent estimate of $\theta(x)$ is

$$\hat{\theta}(x) := \frac{1}{n} \sum_{i=1}^{n} \hat{r}\left(x,\, Z_i\right).$$

Proof. This follows from the consistency relationship $Y = C(X)$, the no measured confounding assumption $(C \amalg X) \mid Z$, and the rule of iterated expectation:

$$\begin{aligned}\int \mathbb{E}\left(Y \mid X = x,\, Z = z\right) dF_Z(z) &= \int \mathbb{E}\left[C(x) \mid X = x,\, Z = z\right] dF_Z(z) \\ &= \int \mathbb{E}\left[C(x) \mid Z = z\right] dF_Z(z) \\ &= \mathbb{E}\left[C(x)\right] = \theta(x),\end{aligned}$$

as desired. □

16.4. Simpson's Paradox.

Remark 16.26 (Simpson's Paradox)**.** Let (X, Y, C_0, C_1) be a consistent tuple of binary random variables. Simpson's paradox is what occurs when the inequality

$$\mathbb{P}\left(Y = 1 \mid X = 1\right) < \mathbb{P}\left(Y = 1 \mid X = 0\right)$$

is mistakenly believed to be equivalent to

$$\mathbb{P}\left(C_1 = 1\right) < \mathbb{P}\left(C_0 = 1\right).$$

The former inequality comes down precisely to the association α since, as shown in Remark 16.9,

$$\alpha = \mathbb{P}\left(Y = 1 \mid X = 1\right) - \mathbb{P}\left(Y = 1 \mid X = 0\right)$$

while the latter inequality comes down precisely to the average causal effect θ since, as shown in Remark 16.7,

$$\theta = \mathbb{P}\left(C_1 = 1\right) - \mathbb{P}\left(C_0 = 1\right).$$

Simpson's paradox is thus the name we give to the misconception that association ($\alpha < 0$) is equivalent to causation ($\theta < 0$). As detailed in Remark 16.10, this is emphatically *not* the case. (As noted in Remark 16.10, the relationship between association and causation comes down to the relationship between the distribution of the covariate X and the distribution of the counterfactual vector (C_0, C_1).)

So why is Simpson's paradox so prevalent, if it can so "easily" be debunked? Because the simple debunking above relies on the simple, but technical, notion of a counterfactual (see also Definition 16.1, which introduces required language); because it is very tempting to use the association α as a proxy for the average causal effect θ since the former can be estimated from observed quantities while the latter *cannot*.

17. Directed Graphs and Conditional Independence

17.1. Conditional Independence.

Definition 17.1 (Conditional independence)**.** Let X, Y, and Z be random variables. X and Y are *conditionally independent given* Z, written $X \amalg Y \mid Z$, if

$$f_{X,Y|Z}(x, y|z) = f_{X|Z}(x|z)f_{Y|Z}(y|z)$$

for all x, y, and z.

Theorem 17.2 (Alternate characterization of conditional independence)**.** *Let X, Y, and Z be random variables. X and Y are conditionally independent given Z if and only if*

$$f_{X|Y,Z}(x|y, z) = f_{X|Z}(x|z)$$

for all x, y, and z.

Theorem 17.3 (Characterization of conditional independence in terms of events)**.** *Let X, Y, and Z be random variables. X and Y are conditionally independent given Z if and only if*

$$\mathbb{P}(X \in A, Y \in B \mid Z = z) = \mathbb{P}(X \in A \mid Z = z)\, \mathbb{P}(Y \in B \mid Z = z)$$

for all events A and B.

Theorem 17.4 (Properties of conditional independence)**.** *Let W, X, Y, and Z be random variables and let $h : \mathbb{R} \to \mathbb{R}$ be a function. The following hold.*

1. *$X \amalg Y \mid Z$ implies $Y \amalg X \mid Z$.*
2. *$X \amalg Y \mid Z$ and $\mathcal{U} := h(X)$ implies $\mathcal{U} \amalg Y \mid Z$.*
3. *$X \amalg Y \mid Z$ and $\mathcal{U} := h(X)$ implies $X \amalg Y \mid (Z, \mathcal{U})$.*
4. *$X \amalg Y \mid Z$ and $X \amalg W \mid (Y, Z)$ implies $X \amalg (W, Y) \mid Z$.*
5. *$X \amalg Y \mid Z$ and $X \amalg Z \mid Y$ implies $X \amalg (Y, Z)$.*

17.2. DAGs.

Definition 17.5 (Directed graph)**.** A *directed graph* is a tuple (V, E) where V is a set whose elements are called *vertices* and E is a set of ordered pairs of vertices whose elements are called *edges*.

Definition 17.6 (Terminology for directed graphs)**.** Let $\mathcal{G}$ be a directed graph and let X and Y be vertices of $\mathcal{G}$.

- If $E = (X, Y)$ is an edge of $\mathcal{G}$ then we say that
  - E *starts* at X and *ends* at Y,
  - X is a *parent* of Y,
  - Y is a *child* of X, and
  - X and Y are *adjacent*.
- The set of all parents of X is denoted π_X.
- A *directed path from X to Y* is a sequence of edges $E_1, \ldots, E_k$ such that
  - E_1 starts at X,
  - E_k ends at Y, and
  - the end of E_j is the start of E_{j+1} for $1 \leqslant j \leqslant k - 1$.
- A directed path from X to X is called a *cycle*.
- If $X = Y$ or if there exists a directed path from X to Y then we say that
  - X is an *ancestor* of Y and
  - Y is a *descendant* of X.

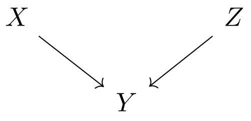


Figure 17.1. An unshielded collider.

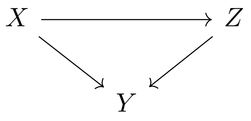


Figure 17.2. A shielded collider.

Definition 17.7 (Undirected path)**.** Let $\mathcal{G}$ be a directed graph and let X and Y be vertices of $\mathcal{G}$. An *undirected path from X to Y* is a sequence of edges $E_1, \ldots, E_k$ such that

- E_1 starts or ends at X,
- E_k starts or ends at Y, and
- E_j and E_{j-1} have a common vertex for every $1 \leqslant j \leqslant k-1$.

Definition 17.8 (Collider)**.** Let $\mathcal{G}$ be a directed graph and let X, Y, and Z be vertices of $\mathcal{G}$. If (X, Y) and (Z, Y) are edges of $\mathcal{G}$ then we say that *Y is a collider between X and Z.*

If moreover X and Z are adjacent then we say that the collider is *shielded*. Otherwise we say that the collider is *unshielded*.

Definition 17.9 (DAG)**.** A *directed acyclic graph*, or *DAG*, is a directed graph which does not have any cycles.

17.3. Probability and DAGs.

Definition 17.10 (DAG representation of a distribution)**.** Let $X = (X_1, \ldots, X_k)$ be a random vector in $\mathbb{R}^k$ with distribution $\mathbb{P}$ with PDF f and let $\mathcal{G} = (V, E)$ be a DAG where $V = \{X_1, \ldots, X_k\}$. We say that *$\mathbb{P}$ is Markov to $\mathcal{G}$*, or that *$\mathcal{G}$ represents $\mathbb{P}$*, if

$$f(x) = \prod_{j=1}^{k} f_{X_j \mid \pi_j}(x_j \mid \pi_j)$$

where π_j denotes the set of parents of X_j in $\mathcal{G}$. The set of distributions represented by $\mathcal{G}$ is denoted by $M(\mathcal{G})$.

Theorem 17.11 (Markov condition)**.** *Let $\mathcal{V}$ be a random vector in $\mathbb{R}^k$ with distribution $\mathbb{P}$ with PDF f and let $\mathcal{G} = (V, E)$ be a DAG where*

$$V = \{W : W = \mathcal{V}_i \text{ for some } i\}.$$

$\mathcal{G}$ represents $\mathbb{P}$ if and only if, for any random variable $W \in V$, if

$$\widetilde{W} := \{X \in V : \; X \text{ is neither a parent nor a descendent of } W\}$$

then W and $\widetilde{W}$ are conditionally independent given π_W. In other words

$$W \amalg \widetilde{W} \mid \pi_W \text{ for every } W \in V,$$

which is known as the Markov condition*.*

Remark 17.12 (Markov condition)**.** In the notation of Theorem 17.11 the Markov condition

$$W \amalg \widetilde{W} \mid \pi_W$$

may be thought of as saying that W is independent of the "past" given its parents.

17.4. More Independence Relations.

Definition 17.13 (d–separation)**.** Let $\mathcal{G}$ be a directed graph.

- Let X and Y be distinct vertices and let W be a set of vertices which does not contain X or Y. We say that X *and* Y *are d–connected given* W if there is an undirected path $\mathcal{U}$ from X to Y such that
  (1) every collider on $\mathcal{U}$ has a descendant in W and
  (2) the only vertices of $\mathcal{U}$ in W are X and Y.
  If X and Y are not d–connected given W then we say that they are *d–separated given* W*.*
- Let A, B, and W be sets of vertices where A and B are non-empty. We say that A *and* B *are d–connected given* W if there exist $X \in A$ and $Y \in B$ such that X and Y are d–connected given W.

  If A and B are not d–connected given W, i.e. if, for every $X \in A$ and $Y \in B$, X and Y are d–separated given W, then we say that A *and* B *are d–separated given* W*.*

Remark 17.14 (d–separation)**.** The "d" in "d–separation" stands for "directed" or "directional".

Definition 17.15 (Faithful representation)**.** Let $\mathcal{G}$ be a DAG which represents a distribution $\mathbb{P}$. We say that $\mathcal{G}$ is a *faithful* representation of $\mathbb{P}$ if the only (conditional) independence relations of $\mathbb{P}$ are those implied by the Markov conditions.

Example 17.16 (Faithful and unfaithful representations)**.** Let $\mathbb{P}$ be the distribution of a random vector (X, Y) with PDF

$$f(x, y) = f(x)f(y),$$

such that X and Y are independent. The DAGs

$$X \longrightarrow Y \qquad\qquad \text{and} \qquad\qquad Y \longleftarrow X$$

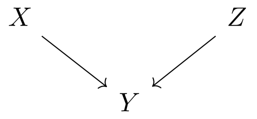
are unfaithful representations of $\mathbb{P}$ whereas

$$X \qquad\qquad Y,$$

the *empty graph* on $\{X, Y\}$, is a faithful representation of $\mathbb{P}$.

Theorem 17.17 (d–separation characterizes conditional independence)**.** *Let* $\mathcal{G}$ *be a DAG which represents faithfully a distribution* $\mathbb{P}$ *and let* A*,* B*, and* C *be disjoint sets of vertices where* A *and* B *are non-empty.* A *and* B *are d–separated* given C *if and only if* A *and* B *are conditionally independent* given C*.*

Remark 17.18 (Interpretation of DAG representations of distributions)**.** Theorem 17.17 helps us interpret the DAG representation of a distribution. Suppose for example we have three random variables X, Y, and Z whose joint distribution is represented by the DAG



This means that X and Z are independent (see Exercise A.17.5) and then have a *causal effect* on Y since

$$f(x,\,y,\,z) = f(y\,|\,x,\,z)f(x)f(z)$$

and so

$$f(y) = \int\int f(y\,|\,x,\,z)f(x)f(z)dxdz.$$

In particular, Theorem 17.17 tells us that, although X and Z are independent, they are *conditionally* independent given Y.

Definition 17.19 (Markov equivalence)**.** Two DAGs are said to be *Markov equivalent* if their Markov conditions imply the same set of (conditional) independence relations.

Definition 17.20 (Undirected graph)**.** An *undirected graph* is a tuple $(V,\,E)$ where V is a set whose elements are called *vertices* and E is a set of unordered pairs of vertices called *edges*.

Definition 17.21 (Skeleton)**.** Let $\mathcal{G} = (V,\,E)$ be a directed graph. The *skeleton* of $\mathcal{G}$ is the undirected graph $(V,\,\widetilde{E})$ where

$$\widetilde{E} := \{\{X,\,Y\} : (X,\,Y) \in E\}\,,$$

i.e. all "arrows" in $\mathcal{G}$ are replaced with undirected edges.

Theorem 17.22 (Characterization of Markov equivalence)**.** *Two DAGs are Markov equivalent if and only if they have the same skeletons and the same unshielded colliders.*

Remark 17.23 (Shielding colliders and conditional independence)**.** Consider the simplest DAGs which contain unshielded or shielded colliders, namely $\mathcal{G}_u$ as in Figure 17.1 and $\mathcal{G}_s$ as in Figure 17.2, and suppose that they are faithful representations of distributions $\mathbb{P}_u$ and $\mathbb{P}_s$, respectively, for the random vector $(X,\,Y,\,Z)$.

As shown in Exercise A.17.5, *unshielded* colliders lead to subtle conditionally independence relations since, according to $\mathcal{G}_u$ (or equivalently $\mathbb{P}_u$),

$$X \perp\!\!\!\perp Z \text{ but } X \not\!\perp\!\!\!\perp Z\,|\,Y.$$

Shielded colliders are not burdened with such subtlety. For example: $\mathcal{G}_s$ is a complete DAG and so the PDF of $\mathbb{P}_s$ is written

$$f_s(x,\,y,\,z) = f(y\,|\,x,\,z)f(z\,|\,x)f(x).$$

Since $\mathcal{G}_s$ is a *faithful* representation of $\mathbb{P}_s$, we then know that $\mathbb{P}_s$ does not have *any* (conditional) independence relations. Indeed (see Exercise A.23.21), we can always write f_s as above by applying the definition of a conditional distribution twice, namely

$$f(x,\,y,\,z) = f(y\,|\,x,\,z)f(x,\,z) = f(y\,|\,x,\,z)f(z\,|\,x)f(x).$$

17.5. **Estimation for DAGs.**

Remark 17.24 (Estimating the distribution given a faithful DAG representation)**.** Given a DAG which is a faithful representation of a distribution, how do we estimate that distribution? In such circumstances we typically use a parametric model for each conditional density, i.e. we assume that the PDF of the distribution takes the form

$$f(x) = f(x;\theta) = \prod_{j=1}^{k} f\left(x_j \,|\, \pi_j; \theta_j\right)$$

where, as in Definition 17.10, the vertex set of random variables is $V = \{X_1, \ldots, X_k\}$ and $\pi_j = \pi_{X_j}$, such that the parameter is $\theta = (\theta_j)_{j=1}^k$.

Note that here θ may belong to a space of (much) higher dimension than $\mathbb{R}^k$. For example: suppose X has codomain $\{0, \ldots, l-1\}$ and that the parents of X are $\pi_X = \{Y, Z\}$ where both Y and Z are binary random variables. Then, for each of the four possible values of (Y, Z), the distribution of X is specified by a parameter in Δ^{l-1}. This means that parametrizing *only* $X \,|\, \pi_X$ requires $4(l-1)$ parameters! The silver lining of this discussion is that if all the random variables in the vertex set are discrete with *finite* codomain then using a parametric model may be done without loss of generality (since a discrete random variable with finite codomain of size l may always be parametrized by an element of Δ^{l-1}), albeit possibly at the cost of a (very) high-dimensional parameter (see the example discussed above).

We can then (typically) estimate the parameter θ using maximum likelihood estimators. Indeed, given an IID sample $X^{(1)}, \ldots, X^{(n)} \sim f$, the likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f\left(X^{(i)}; \theta\right) = \prod_{i=1}^{n}\prod_{j=1}^{k} f\left(X_{ij} \,|\, \pi_j; \theta_j\right)$$

where $X_{ij} := X_j^{(i)}$ (reminiscent of the design matrix used in linear regression). What is particularly convenient is that this likelihood (or equivalently the log-likelihood) *splits* as a product (or sum) of terms with each depend on only *one* of the components of θ:

$$l(\theta) = \log \mathcal{L}(\theta) = \sum_{j=1}^{k} \underbrace{\sum_{i=1}^{n} \log f\left(X_{ij} \,|\, \pi_j; \theta\right)}_{\text{only depends on } \theta_j.}$$

This means that we can find the MLE for each component θ_j of θ *separately*.

Remark 17.25 (Estimating the structure of a DAG)**.** Can we estimate the *structure* of a DAG given data? Given a fixed set of vertices we could do this by first fitting *every possible* DAG over this set of vertices, proceeding as discussed in Remark 17.24 by using maximum likelihood estimators. We would then use a *model score* (see Remark 13.71 and Definition 13.72), such as the AIC, to choose the "best" DAG among those. There are two caveats:

(1) searching through all possible DAGs on a given vertex set is computationally challenging, and prohibitively expensive for large vertex sets, and
(2) we would need a lot of data to reliably fit several DAGs.

These caveats may be less severe if prior information is known about the DAG structure.

17.6. **Causation Revisited.**

Definition 17.26 (Intervention)**.** Let $(\mathcal{G}, f)$ be a pair where $\mathcal{G} = (V, E)$ is a DAG, $\mathbb{P}$ is a distribution such that $\mathbb{P}$ is Markov to $\mathcal{G}$, and f denotes the PDF of $\mathbb{P}$, i.e.

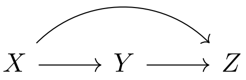
$$f(w) = \prod_{W \in V} f\left(w \,|\, \pi_w\right)$$

(as in Definition 17.10). Let $X \in V$ and let x^* be an element of the codomain of X. We define

- X^* to be the random variable defined by $X^* = x^*$,
- $\mathcal{G}^* := (V^*, E^*)$ where $V^* := \{X^*\} \cup V \setminus \{X\}$ and
  $$E^* := \left\{\left(\sigma\left(W_{out}\right), W_{in}\right) : \left(W_{out}, W_{in}\right) \in E \text{ and } W_{in} \neq X\right\}$$
  for $\sigma : V \to V^*$ the identity except for $\sigma(X) = X^*$, i.e. $\mathcal{G}^*$ is the DAG obtained by removing all edges *into* X and replacing X with X^*, and
- f^* to be the PDF defined by, writing $w = (\tilde{w}, x)$,
  $$f^*(\tilde{w}, x^*) := \frac{f(\tilde{w}, x^*)}{f(x^* \,|\, \pi_x)},$$
  i.e. f^* is the PDF obtained by removing the conditional PDF of $X \,|\, \pi_X$ and replacing x with x^*. In particular, since X^* is *deterministic*, meaning that $\mathbb{P}\left(X^* = x^*\right) = 1$ for x^* a constant, we often abusively write
  $$f^*(\tilde{w}) := f^*(\tilde{w}, x^*).$$

The pair $(\mathcal{G}^*, f^*)$ is called the *intervention "set $X = x^*$" on* $(\mathcal{G}, f)$.

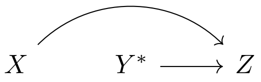
Example 17.27 (Intervention)**.** Consider the DAG $\mathcal{G}$ given by



which represents the PDF f given by

$$f(x, y, z) = f(z \,|\, x, y) f(y \,|\, x) f(x).$$

Let $y^* \in \mathbb{R}$ be *fixed* and let $(\mathcal{G}^*, f^*)$ denote the intervention "set $Y = y^*$" on $(\mathcal{G}, f)$. Then $\mathcal{G}^*$ is given by



where $Y^* \equiv y^*$ and, writing abusively $f^*(x, z)$ for $f^*(x, y^*, z)$,

$$f^*(x, z) = \frac{f(z \,|\, x, y^*) f(y^* \,|\, x) f(x)}{f(y^* \,|\, x)} = f(z \,|\, x, y^*) f(x).$$

Remark 17.28 (Active and passive conditioning)**.** Under the same setup as Example 17.27 let us assume furthermore that Z is discrete. We call

$$\mathbb{P}\left(Z = z \,|\, Y = y^*\right)$$

passive conditioning, or *conditioning by observation*, and this quantity is computed using the *original* PDF f as

$$\mathbb{P}\left(Z = z \,|\, Y = y^*\right) = f\left(z \,|\, y^*\right) = \frac{f\left(y^*, z\right)}{f\left(y^*\right)} = \frac{\int f\left(x, y^*, z\right) dx}{\int\int f\left(x, y^*, z\right) dx dz}.$$

We call the marginal distribution of Z under f^*, namely

$$\mathbb{P}(Z = z \,|\, Y := y) := f^*(z),$$

active conditioning, or *conditioning by intervention*, and this quantity is computed using the *new* PDF f^* as

$$\mathbb{P}(Z = z \,|\, Y := y) = f^*(z) = \int f^*(x, z) dx.$$

What are these different types of conditioning used for?

- Passive conditioning is used to answer *predictive* questions: given $Y = y^*$, what is the probability of $Z = z$?
- Active conditioning is used to answer *causal* questions: if we intervene and set $Y = y^*$, irrespectively of the random variables that Y depends on (i.e. irrespectively of its parents), then what is the probability of $Z = z$?

18. Undirected Graphs

18.1. Undirected Graphs.

Definition 18.1 (Terminology for undirected graphs)**.** Let $\mathcal{G} = (V, E)$ be an undirected graph.

- Two vertices X and Y are said to be *adjacent*, which we denote by $X \sim Y$, if $\{X, Y\} \in E$, i.e. if there is an edge between them.
- A *path* from a vertex X_1 to a vertex X_k is a sequence of vertices $X_1, \ldots, X_k$ such that $X_i \sim X_{i+1}$ for all $1 \leqslant i \leqslant k-1$.
- Let $A, B, C \subseteq V$ be distinct. We say that C *separates* A *and* B if every path from a vertex in A to a vertex in B contains a vertex in C.
- A *subgraph* of $\mathcal{G}$ is an undirected graph $(\mathcal{U}, F)$ where $\mathcal{U} \subseteq V$ and $F \subseteq E$.
- We say that $\mathcal{G}$ is *complete* if any two vertices of $\mathcal{G}$ are adjacent.

18.2. Probability and Graphs.

Definition 18.2 (Pairwise Markov graph)**.** Let $\mathcal{V}$ be a random vector in $\mathbb{R}^k$ with distribution $\mathbb{P}$ and let $\mathcal{G} = (V, E)$ be an undirected graph where

$$\mathcal{V} = \{W : W = \mathcal{V}_i \text{ for some } i\}.$$

We say that $\mathcal{G}$ is a *pairwise Markov graph* for $\mathbb{P}$ if

$$\{X, Y\} \notin E \iff X \amalg Y \,|\, V \setminus \{X, Y\} \text{ for every } X, Y \in V,$$

i.e. there is no edge between X and Y if and only if X and Y are conditionally independent given the remaining variables, and we call this equivalence the *pairwise Markov property.*

Theorem 18.3 (Global Markov property)**.** *Let $\mathcal{G} = (V, E)$ be a pairwise Markov graph for a distribution $\mathbb{P}$ and let A, B, and C be distinct subsets of V. If C separates A and B then $A \amalg B \,|\, C$ and we call this implication the* global Markov property*.*

Remark 18.4 (Separation by the empty set and independence)**.** Theorem 18.3 tells us that if two sets of vertices A and B are not *connected*, meaning that there is no path from a vertex in A to a vertex in B, or in other words A and B are separated by the empty set, then A and B are independent.

Corollary 18.5 (Equivalence of the Markov properties)**.** *Let $\mathcal{G}$ be an undirected graph whose vertices are random variables. A distribution $\mathbb{P}$ satisfies the pairwise Markov property if and only if it satisfies the global Markov property.*

In either (and hence both) case we therefore say that $\mathbb{P}$ is Markov *to $\mathcal{G}$.*

18.3. Cliques and Potentials.

Definition 18.6 (Clique)**.** Let $\mathcal{G}$ be an undirected graph. A *clique* is a set of pairwise adjacent vertices. It is said to be *maximal* if it is not possible to include another vertex and still be a clique.

Theorem 18.7 (Cliques and potentials)**.** *Let $\mathbb{P}$ be a distribution Markov to an undirected graph $\mathcal{G}$. Under suitable regularity conditions the PDF f of $\mathbb{P}$ may be written*

$$f(x) = \prod_C \psi_C(x_C)$$

where the product is over the maximal cliques of $\mathcal{G}$, each ψ_C is a positive function known as a potential*, and x_C denotes all the variables in the clique C. (Note that the maximal cliques need not form a partition of the vertex set, so some variables may appear more than once in the expansion above.)*

18.4. **Fitting graphs to data.**

Remark 18.8 (Estimating the structure of an undirected graph)**.** As with DAGs (see Remark 17.25), estimating the structure of an undirected graph is a serious challenge. However, when the vertices are discrete random variables, this can be done using *log-linear models*, as discussed in the next chapter.

19. Log-Linear Models

19.1. The Log-Linear Model.

Theorem 19.1 (Log-linear expansion)**.** *Let $X = (X_1, \ldots, X_m)$ be a discrete random vector with finite codomain such that, without loss of generality,*

$$X_j \in \{0, \ldots, r_j - 1\} \text{ for } 1 \leqslant j \leqslant m$$

for some integers $r_j \geqslant 2$. The PDF f of X may be written as

$$\log f(x) = \sum_{A \subseteq S} \psi_A(x)$$

where $S = \{1, \ldots, m\}$, and the ψ_A's satisfy

1. *$\psi_\emptyset$ is constant,*
2. *for every $A \subseteq S$, $\psi_A(x)$ is only a function of $x_A := (x_j : j \in A)$, and*
3. *if $j \in A$ and $x_j = 0$ then $\psi_A(x) = 0$.*

We call this expansion the log-linear expansion *of f.*

Remark 19.2 (Special element of the codomain)**.** The third condition in the definition of a log-linear expansion in Theorem 19.1 treats one of the elements of the codomains of X_j differently, namely *zero*. Making a *different* choice of which element of the codomains to treat in a special way would lead to *quantitative* differences in the ensuing log-linear expansion. Nonetheless, as shown in Exercise A.23.23, some *qualitative* properties of that log-linear expansion would remain unchanged.

Remark 19.3 (Multinomial versus log-linear parametrization)**.** Consider a discrete random vector $X = (X_1, \ldots, X_m)$ with finite codomain such that, without loss of generality,

$$X_j \in \{0, \ldots, r_j - 1\} \text{ for } 1 \leqslant j \leqslant m$$

for some integers $r_j \geqslant 2$. We may then view X as a categorical random variable with parameter

$$p_{i_1 \ldots i_m} = \mathbb{P}\left(X_1 = i_1, \ldots, X_m = i_m\right), \; 0 \leqslant i_j \leqslant r_j - 1 \text{ for } 1 \leqslant j \leqslant m$$

such that

$$p \in \Delta^{N-1} \subseteq \mathbb{R}^N \text{ for } N := \prod_{j=1}^m r_j.$$

The case $m = r_1 = r_2 = 2$ is precisely the case of *two-by-two tables* treated in Definition 15.3. Then the PDF f of X may be written

$$f(x) = \mathbb{P}\left(X_1 = x_1, \ldots, X_m = x_m\right) = p_{x_1 \ldots x_m}.$$

This is known as the *multinomial parametrization* of f. Using the log-linear expansion of f, another parametrization is possible.

Indeed, for $S = \{1, \ldots, m\}$ we may write

$$f(x) = \sum_{A \subseteq S} \psi_A(x)$$

for ψ_A as in Theorem 19.1. Moreover, for any $A \subseteq S$ we may write

$$A = \{j_k\}_{k=1}^{m_A} \text{ for } m_A = |A|$$

such that

$$\psi_A(x) = \sum_{v_{j_1}=1}^{r_{j_1}-1} \cdots \sum_{v_{j_k}=1}^{r_{j_k}-1} \beta_{A;v_{j_1},\ldots,v_{j_k}} \mathbb{1}\left(x_{j_1} = v_{j_1},\, \ldots,\, x_{j_k} = v_{j_k}\right)$$

for some parameters $\beta \in \mathbb{R}$. Note that the terms in the summation above do *not* contain any terms when the variables *vanish*, so as to comply with the third defining property of a log-linear expansion as in Theorem 19.1. For example if

$$S = \{1,\, 2,\, 3,\, 4\}\,,\; A = \{2,\, 4\}\,,\; r_2 = 2, \text{ and } r_4 = 3$$

then

$$\psi_A(x) = \beta_{A;1,1}\mathbb{1}\left(x_2 = 1,\, x_4 = 1\right) + \beta_{A;1,2}\mathbb{1}\left(x_2 = 1,\, x_4 = 2\right).$$

One may then verify that there are N such parameters β, so we may write

$$f = f(x;\beta) \text{ for } \beta \in \mathbb{R}^N.$$

This is known as the *log-linear parametrization* of f. Note that since $p = p(\beta)$ lives in the *constrained* set Δ^{N-1}, the parameters β are also constrained to live on some $(N-1)$–dimensional hypersurface in $\mathbb{R}^N$.

Definition 19.4 (Log-linear model)**.** Let $m \geqslant 1$ and $r_j \geqslant 2$ for $1 \leqslant j \leqslant m$ be integers, let $S \subseteq \{1,\, \ldots,\, m\}$, and let $N := \prod_{i=j}^{m} r_j$. The parametric model

$$\mathcal{F} := \left\{ f(\,\cdot\,;p) : f = \sum_{A \subseteq S} \psi_A \right\},$$

where we view $p \in \Delta^{N-1}$ as a parameter as in Remark 19.3, is called a *log-linear model.*

We often abuse terminology and refer to subsets of the log-linear model whose ditributions have log-linear expansions of a prescribed form as *models.* For example we may refer to the model

$$\psi_\emptyset + \psi_1 + \psi_{12}$$

to mean the subset of a log-linear model characterized by $\psi_A \neq 0$ if and only if A is $\emptyset$, $\{1\}$, or $\{1,\, 2\}$.

Remark 19.5 (Log-linear models are generic)**.** Theorem 19.1 tells us that *any* discrete random vector $X = (X_1,\, \ldots,\, X_m)$ whose components have finite codomains may be viewed, up to a bijection if necessary to ensure that the codomain of each X_j is precisely $\{0,\, \ldots,\, r_j - 1\}$, as drawn from a distribution in a log-linear model. This means that, similarly to the simple linear regression model (see Remark 13.9), the value of the log-linear model resides in how it *parametrizes* such random vectors – see for example Remark 19.3 which discusses an alternate parametrization of the log-linear model.

Lemma 19.6. *A partition $(X_A,\, X_B,\, X_C)$ of a random vector X satisfies the conditional independency $X_B \amalg X_C \,|\, X_A$ if and only if the PDF of X may be written as*

$$f\left(x_A,\, x_B,\, x_C\right) = g\left(x_A,\, x_B\right) h\left(x_A,\, x_C\right)$$

for some functions g and h.

Theorem 19.7 (Conditional independence and log-linear expansions)**.** *Consider a partition (X_1, X_B, X_C) of a random vector $X = (X_1, \ldots, X_m)$ and let ψ_T, $T \subseteq \{1, \ldots, m\}$, denote the log-linear expansion of the PDF f of X.*

$$X_B \perp\!\!\!\perp X_C \mid X_A \iff \psi_T = 0 \text{ whenever } T \text{ intersects both } B \text{ and } C.$$

In other words: all the ψ–terms with at least one coordinate in B and one coordinate in C must vanish.

19.2. **Graphical Log-Linear Models.**

Definition 19.8 (Graphical log-linear model)**.** Let f be a PDF from a log-linear model with log-linear expansion $f = \sum_{A \subseteq S} \psi_A$ for $S = \{1, \ldots, m\}$ for some $m \geqslant 1$. We say that f is *graphical* if there exists an undirected graph $\mathcal{G} = (S, E)$ such that $\psi_A \neq 0$ when $|A| = 0$ or 1 and, for $|A| \geqslant 2$,

$$\psi_A = 0 \iff \{i, j\} \notin E \text{ for some } i, j \in A$$

or equivalently

$$\psi_A \neq 0 \iff \{i, j\} \in E \text{ for every } i, j \in A.$$

The set of all graphical distributions from a log-linear model is called a *graphical log-linear model.* We often abuse terminology as is done with log-linear models – see Definition 19.4.

Example 19.9 (Graphical log-linear models on two variables)**.** There are two undirected graphs on two vertices (up to isomorphism) and so there are two graphical log-linear models on two random variables, namely

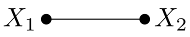
$$\psi_\emptyset + \psi_1 + \psi_2$$

for the empty graph

and

$$\psi_\emptyset + \psi_1 + \psi_2 + \psi_{12}$$

for the complete graph



These two models simply correspond to whether the two random variables are independent or dependent, respectively.

19.3. **Hierarchical Log-Linear Models.**

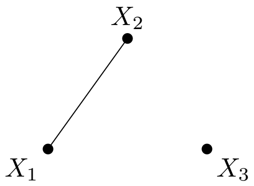
Definition 19.10 (Hierarchical log-linear model)**.** Let f be a PDF from a log-linear model with log-linear expansion $f = \sum_A \psi_A$. We say that f is *hierarchical* if

$$\psi_A = 0 \text{ and } A \subseteq B \text{ implies } \psi_B = 0.$$

The set of all hierarchical distributions from a log-linear model is called a *hierarchical log-linear model.* We often abuse terminology as is done with log-linear models – see Definition 19.4.

Lemma 19.11. *A graphical log-linear model is hierarchical but the converse does not hold in general.*

Example 19.12 (Types of log-linear models)**.** Here we consider three log-linear models which are Markov to the same undirected graph, namely



As per Theorem 18.3 this means that in all three cases $(X_1,\, X_2) \amalg X_3$. The graphical log-linear model

$$\psi_\emptyset + \psi_1 + \psi_2 + \psi_3 + \psi_{12}$$

imposes this independent relation and *nothing else*. This highlights that, given an undirected graph, there is only one corresponding graphical log-linear model but there may be (and typically are) more than one corresponding hierarchical, and hence log-linear, model.

Here for example we consider the hierarchical, but not graphical, model

$$\psi_\emptyset + \psi_1 + \psi_2 + \psi_{12},$$

which corresponds to removing the term ψ_3 from the graphical model above. This augments the independence relation $(X_1,\, X_2) \amalg X_3$ by also imposing

$$X_3 \,|\, (X_1,\, X_2) \sim \text{Uniform},$$

while the distribution of $(X_1,\, X_2)$ is still a generic log-linear distribution without any constraints.

Finally we consider the log-linear model

$$\psi_\emptyset + \psi_{12}$$

which further removes ψ_1 and ψ_2 from the hierarchical model above. This is not hierarchical since $\psi_1 = 0$ but $\psi_{12} \neq 0$. On top of the independence $(X_1,\, X_2) \amalg X_3$ and the condition $X_3 \,|\, (X_1,\, X_2) \sim$ Uniform this also imposes that

$$X_1 \,|\, X_2 = 0,\; X_2 \,|\, X_1 = 0 \sim \text{Uniform}.$$

19.4. **Model Generators.**

Definition 19.13 (Model generator)**.** Let $m,\, k \geqslant 1$, $1 \leqslant n_j$, $i_{j,l_j} \leqslant m$ for $1 \leqslant j \leqslant k$ and $1 \leqslant l_j \leqslant n_j$ be integers. The *formal* expression

$$\sum_{j=1}^{k} i_{j,1}.i_{j,2}.\cdots.i_{j,n_j}$$

is called the *model generator* corresponding to the *smallest* hierarchical log-linear model for which

$$\psi_{i_{j,1}i_{j,2}\ldots i_{j,n_j}} \neq 0 \text{ for every } 1 \leqslant j \leqslant k.$$

Example 19.14 (Model generator)**.** The model generator

$$1.2 + 1.3$$

corresponds to the smallest hierarchical log-linear model for which

$$\psi_{12} \text{ and } \psi_{13} \text{ are nonzero.}$$

This means that, necessarily,

$$\psi_\emptyset,\; \psi_1,\; \psi_2,\; \text{and } \psi_{13} \text{ are nonzero.}$$

Since it is the *smallest* such model we deduce that $\psi_A = 0$ for all other subsets A. In other words the model is precisely

$$\psi_\emptyset + \psi_1 + \psi_2 + \psi_3 + \psi_{12} + \psi_{13}.$$

19.5. **Fitting Log-Linear Models to Data.**

Remark 19.15 (Estimating the distribution given a log-linear model)**.** Given a log-linear model

$$\mathcal{F} := \left\{ f(\,\cdot\,;\beta) : f = \sum_{A \subseteq S} \psi_A \right\}$$

where we use the log-linear parameter $\beta \in \mathbb{R}^N$ as in Remark 19.3 and given an IID sample $X^{(1)}, \ldots, X^{(n)}$ drawn from a distribution in that model we may estimate the parameter β using maximum likelihood estimators. Indeed, the log-likelihood function is

$$l(\beta) = \sum_{i=1}^{n} f\left(X^{(i)}; \beta\right),$$

which in practice we can often maximize numerically in order to find the MLE $\hat{\beta}$. Typically the Fisher information matrix may then also be obtained numerically, from the inverse of which we can then deduce estimates for the standard error as well as confidence intervals.

Definition 19.16 (Saturated model)**.** The log-linear model on m random variables with model generator

$$1.2.\cdots.m$$

is called the *saturated model*. It corresponds to the graphical log-linear model whose corresponding undirected graph is the complete graph on m vertices, or in other words it is precisely the honest-to-goodness log-linear model with m variables of Definition 19.4, such that the saturated model corresponds to

$$\log f = \sum_{A} \psi_A$$

where $\psi_A \neq 0$ for *all* $A \subseteq \{1, \ldots, m\}$.

Remark 19.17 (Estimating the structure of a log-linear model)**.** In order to estimate the *structure* of a log-linear model, i.e. estimate for which A should ψ_A be included in the model, we must solve a *model selection* problem – see Section 13.6 where this is discussed for linear regression models. For log-linear models, model selection typically follows one of the following approaches.

- The first approach is to use a *model score* (see Definition 13.72 and Remark 13.73 in the context of linear regression), such as the Akaike Information Criterion defined below (and which is defined in exactly the same way as for linear regression – see Definition 13.58).

  When following this approach we often restrict the search to hierarchical log-linear models. This is done simply to reduce the search space but has the additional benefit that hierarchical models are often viewed as more *interpretable* than non-hierarchical log-linear models.

- The second approach is to rely on hypothesis testing, where we test, for each model M, the null hypothesis that the true model is M versus the alternative hypothesis that the true model is the saturated model.

  This approach will typically *reduce* the number of models under consideration but may not identify a *single* "best" model.

Definition 19.18 (Akaike Information Criterion)**.** Let $X^{(1)}, \ldots, X^{(n)}$ be an IID sample drawn from a distribution in a log-linear model M. We define the *Akaike Information Criterion* to be

$$AIC(M) := l_M - |M|$$

where l_M is the log-likelihood of the model M evaluated at the MLE and $|M|$ is the number of parameters of the model M.

Remark 19.19. The AIC of Definition 19.18 is defined in *exactly* the same way as the AIC of Definition 13.58. The only difference are the underlying sets of models.

Definition 19.20 (Deviance)**.** Let $m \geqslant 1$ be an integer, let M be a log-linear model on m random variables, and let $X^{(1)}, \ldots, X^{(n)}$ be an IID sample from a distribution in M. The *deviance* of M is defined to be

$$\operatorname{dev}(M) := 2\left(l_{sat} - l_M\right)$$

where l_{sat} is the log-likelihood of the saturated model on m variables evaluated at its MLE and l_M is the log-likelihood of the moel M evaluated at its MLE.

Note that the deviance is a random variable.

Theorem 19.21 (Deviance and log-linear model selection by hypothesis testing)**.** *Let $m \geqslant 1$ be an integer, let M be a log-linear model on m random variables, and let M_{sat} be the saturated model on m variables. The deviance of M is the likelihood ratio statistic for*

$$H_0 : \text{“the model is } M\text{” versus } H_1 : \text{“the model is } M_{sat}\text{”}.$$

In particular, under H_0, $dev\,(M) \rightsquigarrow \chi^2_\nu$ where the number of degrees of freedom is $\nu = |M_{sat}| - |M|$. Therefore the p–value of the corresponding likelihood ratio test satisfies, asymptotically as $n \to \infty$,

$$p - value(\omega) = \mathbb{P}\left(\chi^2_\nu > dev\,(M)|_\omega\right)$$

for any outcome ω.

20. Nonparametric Curve Estimation

20.1. The Bias-Variance Tradeoff.

Definition 20.1 (Mean integrated squared error)**.** Let $g : \mathbb{R} \to \mathbb{R}$ be a function and let $\hat{g}$ be a random function from $\mathbb{R}$ to $\mathbb{R}$. The *integrated squared error*, or *ISE*, is the random variable defined by

$$L(g, \hat{g}) := \int [g(x) - \hat{g}(x)]^2 dx$$

and the *mean integrated squared error*, or *MISE*, is the value

$$R(g, \hat{g}) := \mathbb{E}[L(g, \hat{g})].$$

Remark 20.2 (Integrated squared error, loss functions, and risk)**.** The integrated squared error $L(g, \hat{g})$ is precisely the integral over $\mathbb{R}$ of the squared error loss $L(g(x), \hat{g}(x))$. Variants of the integrated squared error which use other loss functions are thus often used.

Similarly, Fubini's Theorem (really, here, Tonelli's Theorem) tells us that

$$\begin{aligned} R(g, \hat{g}) = \mathbb{E}[L(g, \hat{g})] = \mathbb{E}\int L(g(x), \hat{g}(x))dx &= \int \mathbb{E}[L(g(x), \hat{g}(x))]\, dx \\ &= \int R[g(x), \hat{g}(x)]\, dx, \end{aligned}$$

i.e. the *mean* integrated squared error is precisely the integral of the risk $R[g(x), \hat{g}(x)]$. (also known, in this case, as the mean squared error – see Lemma 12.5). This is why the MISE itself is sometimes called the *risk*.

Lemma 20.3 (Bias-variance tradeoff)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a function and let $\hat{g}$ be a random function from $\mathbb{R}$ to $\mathbb{R}$ (which means that, for every $x \in \mathbb{R}$, $\hat{g}(x)$ is a random variable). The MISE may be written as*

$$R(g, \hat{g}) = \int b^2(x)dx + \int v(x)dx$$

where, for every x,

$$b(x) = \mathbb{E}[\hat{g}(x) - g(x)]$$

is the bias of $\hat{g}(x)$ as a point estimator of $g(x)$ and

$$v(x) = \mathbb{V}[\hat{g}(x)] = \mathbb{E}\left([\hat{g}(x) - \mathbb{E}\hat{g}(x)]^2\right)$$

is the variance of $\hat{g}(x)$.

Remark 20.4 (Bias-variance tradeoff and decomposition of the mean squared error)**.** Lemma 20.3 is nothing more than the integrated–in–x version of the decomposition of the mean squared error recorded in Theorem 6.10. This is not surprising in light of Remark 20.4, and actually this is precisely how Lemma 20.3 is proved in Exercise A.20.2.

Remark 20.5 (Undersmoothing and oversmoothing)**.** Lemma 20.3 quantifies a phenomenon similar to the underfitting and overfitting issues that can occur when selecting models (see for example Remark 13.71 in the context of linear regression). Here we assume that the random function $\hat{g}$ is an estimator of the function $g : \mathbb{R} \to \mathbb{R}$ and that this estimation process depends on a *smoothing parameter*. This smoothing parameter may be thought of as an *inversely proportional analog* of the number of parameters or covariates in model selection (once again, see Remark 13.71).

- When the smoothing parameter is small the bias term in Lemma 20.3 is small but the variance term is large; this is called *undersmoothing*.
- When the smoothing parameter is large the variance term is smal but the variance term is small; this is called *oversmoothing*.

Lemma 20.3 tells us that minimizing the MISE risk comes down to optimally trading-off bias and variance.

20.2. **Histograms.**

Definition 20.6 (Histogram estimator)**.** Let $X_1, \ldots, X_n$ be an IID sample satisfying $0 \leqslant X_i \leqslant 1$ for all i and let $m \geqslant 1$ be an integer.

- The intervals
$$B_1 = \left[0,\, \frac{1}{m}\right),\, \ldots,\, B_j = \left[\frac{j-1}{m},\, \frac{j}{m}\right),\, \ldots,\, B_m = \left[\frac{m-1}{m},\, 1\right],$$
for $1 \leqslant j \leqslant m$, are called *bins*.
- The value $h := \frac{1}{m}$ is called the *binwidth*.
- For $1 \leqslant j \leqslant m$, ν_j is the number of observations in the bin B_j, i.e.
$$\nu_j = \sum_{i=1}^{n} \mathbb{1}\left(X_i \in B_j\right),$$
and we define $\hat{p}_j := \frac{\nu_j}{n}$.

The *histogram estimator* is the random function $\hat{f}_n$ from $[0, 1]$ to $\mathbb{R}$ defined by
$$\hat{f}_n(x) := \sum_{j=1}^{m} \frac{\hat{p}_j}{h} \mathbb{1}\left(x \in B_j\right) = \begin{cases} \hat{p}_1/h & \text{if } x \in B_1, \\ \cdots & \\ \hat{p}_m/h & \text{if } x \in B_m. \end{cases}$$

Remark 20.7 (Histogram estimators and rescaling)**.** The assumption $0 \leqslant X_i \leqslant 1$ in Definition 20.6 is only there for convenience: if it fails then we simply shift and rescale our sample $X_i \mapsto \frac{X_i - \tau}{\lambda}$ for some appropriate $\tau \in \mathbb{R}$ and $\lambda > 0$. This does mean that histogram estimators apply to random variables with *bounded* codomains.

Remark 20.8 (Randomness of the histogram estimator)**.** The histogram estimator $\hat{f}_n$ is indeed a random function since it depends on the sample $X_1, \ldots, X_n$. Said more explicitly: for each outcome ω we obtain a different sample $X_1(\omega), \ldots, X_n(\omega)$, and hence a different function $\hat{f}_n(\omega) : [0, 1] \to \mathbb{R}$.

Theorem 20.9 (Properties of histogram estimators)**.** *Let $X_1, \ldots, X_n$ be an IID sample from a PDF f satisfying $0 \leqslant X_i \leqslant 1$ for all i, let $m \geqslant 1$ be an integer, and let $\hat{f}_n$ be the corresponding histogram estimator with bins $(B_j)_{j=1}^m$ and binwidth h. We define*
$$p_j := \int_{B_j} f \text{ for } 1 \leqslant j \leqslant m.$$
For every $x \in \mathbb{R}$, if $x \in B_j$ then
$$\mathbb{E}\left[\hat{f}_n(x)\right] = \frac{p_j}{h} \text{ and } \mathbb{V}\left[\hat{f}_n(x)\right] = \frac{p_j(1-p_j)}{nh^2}.$$

Theorem 20.10 (MISE of the histogram estimator)**.** *Let $X_1, \ldots, X_n$ be an IID sample from a PDF f satisfying $0 \leqslant X_i \leqslant 1$ for all i and*

$$\int (f')^2 < \infty$$

and let $\hat{f}_n$ be the histogram estimator with binwidth h. Provided f is sufficiently regular and h is sufficiently small, the MISE of the histogram estimator is given by

$$R\left(f, \hat{f}_n\right) \approx \frac{h^2}{12} \int (f')^2 + \frac{1}{nh},$$

which is minimized with respect to h at

$$h^* = \frac{1}{n^{1/3}} \left(\frac{6}{\int (f')^2} \right)^{1/3}.$$

If we denote by $\hat{f}_n^*$ the histogram estimator with binwidth h^* then its MISE is

$$R\left(f, \hat{f}_n^*\right) \approx \frac{C}{n^{2/3}} \text{ for } C := \left(\frac{3}{4}\right)^{2/3} \left(\int (f')^2\right)^{1/3}.$$

Remark 20.11 (Optimal MISE of the histogram estimator)**.** Theorem 20.10 tells us that with an optimal binwidth the MISE of the histogram estimator decays like $n^{-2/3}$. This is slower than the typical (and optimal) decay of n^{-1} for *parametric* density estimators. This is the price of being nonparametric.

Remark 20.12 (Optimal binwidth for the histogram estimator)**.** Theorem 20.10 tells us how to choose optimally the binwidth of a histogram estimator. This is of little practical use since that optimal binwidth depends on the *unknown* density f.

Definition 20.13 (Admissible binwidth)**.** We denote by $\mathfrak{h}$ the set

$$\mathfrak{h} := \left\{ h \in (0, 1] : \frac{1}{h} \text{ is an integer} \right\}$$

and we call any $h \in \mathfrak{h}$ an *admissible binwidth.*

Definition 20.14 (Histogram risk)**.** Let $X_1, \ldots, X_n$ be an IID sample from a PDF f satisfying $0 \leqslant X_i \leqslant 1$ for all i. For every admissible binwidth h let $\hat{f}_n$ denote the histogram estimator with binwidth h. The *histogram risk* is the random function from $\mathfrak{h}$ to $\mathbb{R}$ defined by

$$J(h) := \int \hat{f}_n^2 - 2 \int \hat{f}_n f$$

for every $h \in \mathfrak{h}$.

Remark 20.15 (Histogram risk)**.** For a fixed binwidth h the histogram risk is equal to the ISE of the histogram estimator up to an additive constant since

$$L\left(f, \hat{f}_n\right) = \int \left(f - \hat{f}_n\right)^2 = \int f^2 + J(h).$$

Since the expectation of the ISE, the MISE, is sometimes called the risk (see Remark 20.2) and since, as observed immediately above, J acts as a proxy for the ISE, and hence the MISE, we refer to J itself as a *risk*, namely the *histogram risk.*

We view the histogram risk as a function of the binwidth since, in practice, the histogram risk is used to find a binwidth with minimal MISE *without* appealing to Theorem 20.10 (see Remark 20.12).

Definition 20.16 (Leave-one-out cross-validation estimator of the histogram risk)**.** Let $X_1, \ldots, X_n$ be an IID sample satisfying $0 \leqslant X_i \leqslant 1$ for all i. We define the *leave-one-out cross-validation estimator of the histogram risk* to be the random function from $\mathfrak{h}$ to $\mathbb{R}$ defined by, for every admissible binwidth h,

$$\hat{J}(h) := \int \hat{f}_n^2 - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{(-i)}(X_i)$$

where $\hat{f}_n$ is the histogram estimator with binwidth h and $\hat{f}_{(-i)}$ is the histogram esimator with the same binwidth corresponding to the sample

$$X_1, \ldots X_{i-1}, X_{i+1}, \ldots, , X_n,$$

i.e. the original sample but with X_i removed.

Remark 20.17 (Leave-one-out cross-validation estimator of the histogram risk)**.** Why do we even have *any* hope that the leave-one-out cross-validation estimator of the histogram risk be a good estimator of the histogram risk? The key lies in

$$\frac{1}{n} \sum_{i=1}^{n} \hat{f}_{(-i)}(X_i)$$

estimating

$$\int \hat{f}_n f.$$

From the definition of $\hat{f}_n$ and $\hat{p}_j$ (see Definition 20.6) and using p_j as in Theorem 20.9 we may write the latter as

$$\int \hat{f}_n f = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} \underbrace{\int_{B_j} f}_{=p_j} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{p_j}{h} \mathbb{1}\left(X_i \in B_j\right)$$

while the former may be written

$$\frac{1}{n} \sum_{i=1}^{n} \hat{f}_{(-i)}(X_i) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\hat{p}_{(-i),j}}{h} \mathbb{1}\left(X_i \in B_j\right)$$

for

$$\hat{p}_{(-i),j} := \frac{1}{n-1} \sum_{\substack{l=1 \\ l \neq i}}^{n} \mathbb{1}\left(X_i \in B_j\right).$$

Crucially, since $\hat{p}_{(-i),j}$ estimates p_j, it follows that $\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i)$ estimates $\int \hat{f}_n f$, and so indeed $\hat{H}$ estimates J.

Theorem 20.18 (Asymptotic consistency of the leave-one-out cross-validation estimator of the histogram risk)**.** *Let $X_1, \ldots, X_n$ be an IID sample satisfying $0 \leqslant X_i \leqslant 1$ for all i, let J denote the histogram risk, and let $\hat{J}$ denote the leave-one-out cross-validation estimator of the histogram risk. For every admissible binwidth h, $\hat{J}(h)$ is an asymptotically consistent point estimator of $J(h)$ as $n \to \infty$, meaning that*

$$\mathbb{E}\hat{J}(h) \to \mathbb{E}J(h) \text{ as } n \to \infty.$$

Remark 20.19 (Asymptotic consistency of the leave-one-out cross-validation estimator of the histogram risk)**.** We can actually show more precisely that the bias is

$$\mathbb{E}\left[\hat{J}(h) - J(h)\right] = \frac{2}{n-1}\left(\frac{1}{h} - \mathbb{E}\int \hat{f}_n^2\right).$$

This is done in Exercise A.23.25. A practical takeaway from this identity is that we need $\frac{1}{nh} \ll 1$, or equivalently

$$n \gg \frac{1}{h},$$

in order for the leave-one-out cross-validation estimator of the histogram risk $\hat{J}(h)$ to be an approximately consistent point estimator of the histogram risk $J(h)$.

Theorem 20.20 (Identity for the leave-one-out cross-validation estimator of the histogram risk)**.** *Let $X_1, \ldots, X_n$ be an IID sample satisfying $0 \leqslant X_i \leqslant 1$ for all i and let $\hat{J}$ denote the leave-one-out cross-validation estimator of the histogram risk. For every admissible binwidth h let $\hat{p}_j$ be defined as in Definition 20.6 for $1 \leqslant j \leqslant m := 1/h$. For every $h \in \mathfrak{h}$,*

$$\hat{J}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{j=1}^{m}\hat{p}_j^2.$$

Definition 20.21 (Histogram projection)**.** Let f be an integrable function on $[0, 1]$, let $m \geqslant 1$ be an integer, let $h := 1/m$, and

$$B_j := \left[\frac{j-1}{m}, \frac{j}{m}\right) \text{ for } 1 \leqslant j \leqslant m-1 \text{ and } B_M := \left[\frac{m-1}{m}, 1\right]$$

as in Definition 20.6, and let $p_j := \int_{B_j} f$ a as in Theorem 20.9. We define the function $\bar{f} : [0, 1] \to \mathbb{R}$ by

$$\bar{f}(x) := \sum_{j=1}^{m}\frac{p_j}{h}\mathbb{1}\left(x \in B_j\right)$$

and call it a *histogram projection* of f.

Remark 20.22 (The histogram projection is a projection)**.** The histogram projection introduced in Definition 20.21 is actually an honest-to-goodness projection. Indeed, if we define the set of step functions

$$S_m := \left\{s : [0, 1] \to \mathbb{R} \text{ such that } s|_{B_j} \text{ is constant}\right\}$$

and assume furthermore that f is *square*–integrable then $\bar{f}$ is the L^2–projection of f onto S_m. This is proved in Exercise A.23.26.

Definition 20.23 (Confidence band)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be a function, let l and u be random functions from $\mathbb{R}$ to $\mathbb{R}$, and let $\alpha \in [0, 1]$. We say that (l, u) is a $1-\alpha$ *confidence band*, or $1-\alpha$ *confidence enevelope*, for f if

$$\mathbb{P}\left(l(x) \leqslant f(x) \leqslant u(x) \text{ for all } x\right) \geqslant 1-\alpha.$$

Theorem 20.24 (Confidence band using the histogram estimator)**.** *Let $X_1, \ldots, X_n$ be an IID sample from a PDF f supported on $[0, 1]$, let $m = m(n) \geqslant 1$ be an integer, and let $\hat{f}_n$ be the corresponding histogram estimator. For any $\alpha \in (0, 1)$ let*

$$c := \frac{z_{\alpha/(2m)}}{2}\sqrt{\frac{m}{n}}$$

where $z_\beta := \Phi^{-1}(1-\beta)$ *for any* $\beta \in (0,\, 1)$ *for* Φ *denoting the CDF of the standard Normal distribution and let, for* $x \in [0,\, 1]$,

$$l_n(x) := \max\left(\sqrt{\hat{f}_n(x)} - c,\, 0\right)^2 \textit{ and } u_n(x) := \left(\sqrt{\hat{f}_n(x)} + c\right)^2.$$

If $m(n) \to \infty$ *and* $\frac{m(n)}{n}\log n \to 0$ *as* $n \to \infty$ *then* $(l_n,\, u_n)$ *is an asymptotic* $1-\alpha$ *confidence band for the histogram projection* $\bar{f}_n$ *of* f *as* $n \to \infty$*, meaning that*

$$\liminf_{n\to\infty} \mathbb{P}\Big(l_n(x) \leqslant \bar{f}_n(x) \leqslant u_n(x) \textit{ for all } x \in [0,\, 1]\Big) \geqslant 1-\alpha.$$

Remark 20.25 (Confidence band using the histogram estimator)**.** A few remarks are in order concerning Theorem 20.24.

- The confidence band constructed in Theorem 20.24 is *not* a confidence band for the true density f. It is *only* a confidence band for its histogram projection $\bar{f}_n$.
- The upper and lower bounds in Theorem 20.24 are defined in terms of
$$\left(\sqrt{\hat{f}_n} \pm c\right)^2$$
instead of simply
$$\hat{f}_n \pm c$$
(with a correction to the lower bound to ensure that it remains non-negative). This is because, as noted in [Was10],
$$\hat{p}_j \sim N\left(p_j,\, \frac{p_j(1-p_j)}{n}\right)$$
but
$$\sqrt{\hat{p}_j} \sim N\left(\sqrt{p_j},\, \frac{1}{4n}\right),$$
both asymptotically as $n \to \infty$, for p_j as in Theorem 20.9. In other words: taking the square root of $\hat{f}_n$ ensures that the variance of $\sqrt{\hat{f}_n}$ is *uniform* across the whole domain $[0,\, 1]$. (This is also discussed in Chapter 3 of [Sco15].)
- The condition $\frac{m(n)}{n}\log n \to 0$ as $n \to \infty$ in Theorem 20.24 is reminiscent of Remark 20.19 where it was noted that issues arise when $m = \frac{1}{h}$ grows too fast as a function of n. Specifically in Remark 20.19 we noted that $n \gg m$ was needed to ensure that the bias of the leave-one-out cross-validation estimator of the histogram risk be small.

  In Theorem 20.24 we *quantify* how much smaller than n shoud m be. Indeed, we ask that $n \gg m\log n$ since
$$\frac{n}{m\log n} = \left(\frac{m}{n}\log n\right)^{-1} \to \infty \text{ as } n \to \infty.$$
- The constant c in Theorem 20.24 is far from optimal since, as discussed in [Was10], in its current form it relies on the crude fact that, if $Z_1,\, \ldots,\, Z_n$ are IID standard Normal random variables then
$$\mathbb{P}\left(\max_i |Z_i| > z\right) \leqslant \sum_i \mathbb{P}\left(|Z_i| > z\right).$$

A more careful characterization of the maximum absolute value of IID standard Normal random variables would therefore allow us to make c smaller, and thus make the confidence band tighter.

20.3. **Kernel Density Estimation.**

Definition 20.26 (Kernel)**.** A *kernel* is a function $K : \mathbb{R} \to \mathbb{R}$ which satisfies the following properties.

1. Non-negativity: $K \geqslant 0$.
2. Integrability: K is integrable.
3. Normalization: $\int K = 1$.
4. Mean zero: $\int xK(x)dx = 0$.
5. Finite variance: $\int x^2K(x)dx =: \sigma_K^2 < \infty$.

Example 20.27 (Kernels)**.** Here are example of kernels.

1. The *Epanechnikov kernel* is defined by
$$K(x) = \begin{cases} \frac{3}{4}\left(1 - \frac{x^2}{5}\right)/\sqrt{5} & \text{if } |x| < \sqrt{5} \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$
Note that $\frac{3}{4}\left[\left(1 - \frac{x^2}{c^2}\right)/\sqrt{c}\right] \cdot \mathbb{1}\left(|x| < \sqrt{c}\right)$ is also a valid kernel for any $c > 0$. The choice $c = 5$ ensures that the resulting kernel has *unit* variance, i.e.
$$\sigma_K^2 = \int x^2K(x)dx = 1.$$
2. The *Gaussian kernel*, or *Normal kernal*, is
$$K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$
3. The *boxcar kernel* is
$$K(x) = \mathbb{1}\left(|x| < 1/2\right) = \begin{cases} 1 & \text{if } -\frac{1}{2} < x < \frac{1}{2} \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Remark 20.28 (Kernels and PDFs)**.** The PDF of any continuous random variable with mean zero and finite variance is a kernel. For example:

1. the Epanechnikov kernel is the PDF of the *Epanechnikov distribution*, whose PDF may obtained by translating and scaling the domain of a Beta(2, 2) distribution,
2. the Gaussian kernel is the PDF of a standard Normal random variable, and
3. the boxcar kernel is the PDF of a Uniform $\left(-\frac{1}{2}, \frac{1}{2}\right)$ random variable.

The converse also holds: any kernel gives rise to a PDF with mean zero and variance σ_K^2. This latter point is precisely why we use the probabilistic vocabulary of means and variances when introducing kernels in Definition 20.26.

Definition 20.29 (Kernel density estimator)**.** Let $X_1, \ldots, X_n$ be an IID sample, let $h > 0$, and let K be a kernel. The *kernel density estimator* is the random function $\hat{f}_n$ from $\mathbb{R}$ to $\mathbb{R}$ defined by
$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

and we call h the *bandwidth*.

Remark 20.30 (Kernel, rescaling, and kernel density estimator)**.** For any kernel K and any $h > 0$ the function K^h defined by

$$K^h := \frac{1}{h} K\left(\frac{\cdot}{h}\right)$$

is *also* a kernel. It is a *rescaling* of the original kernel K. More precisely, its variance σ_h^2 is related to the variance σ_K^2 of the original kernel via

$$\sigma_h^2 = \int y^2 \cdot \frac{1}{h} K\left(\frac{y}{h}\right) dy = \int (hx)^2 \cdot \frac{1}{h} K(x)(hdx) = h^2 \sigma_K^2.$$

The bandwidth of a kernel density estimator is therefore a *smoothing parameter* (like the binwidth of a histogram estimator): when h is small then $\hat{f}_n$ is more jagged and when h is large then $\hat{f}_n$ is more flat.

Pushing this idea further: the kernel density estimator tends to the probability mass function associated with the empirical CDF when $h \to 0$ (the *least* smooth) and it tends to a uniform distribution as $h \to \infty$ (the *most* smooth).

Remark 20.31 (Choice of kernel and choice of bandwidth)**.** When working with histogram estimators, particular attention has to be paid to the choice of binwidth in order to minimize the MISE. Now, when working with kernel density estimators, we have two choices to make: the kernel and the bandwidth. It can be shown that the Epanechnikov kernel is optimal with respect to asymptotic MISE, but both in theory and in *practice* the choice of kernel is not crucial. The choice of bandwidth is very important, however.

Theorem 20.32 (MISE of kernel density estimators)**.** *Let $X_1, \ldots, X_n$ be an IID sample from a PDF f and let $\hat{f}_n$ be the kernel density estimator with kernel K and bandwidth h. Under suitable (weak) assumptions on f and K the MISE of the kernel density estimator is given by*

$$R\left(f, \hat{f}_n\right) \approx \frac{1}{4} \sigma_K^4 h^4 \int (f'')^2 + \frac{\int K^2}{nh},$$

which is minimized with respect to h at

$$h^* = \frac{1}{n^{4/5} \sigma_K^{4/5}} \left(\int K^2 \Big/ \int (f'')^2 \right)^{1/5}.$$

If we denote by $\hat{f}_n^*$ the kernel density estimator with the same kernel and bandwidth h^* then its MISE is

$$R\left(f, \hat{f}_n\right) \approx \frac{C}{n^{4/5}} \text{ for } C := \frac{5\sigma_K^{4/5}}{4n^{4/5}} \left(\int K^2 \right)^{4/5} \left(\int (f'')^2 \right)^{1/5}.$$

Remark 20.33 (Optimal MISE of kernel density estimators)**.** Theorem 20.32 tells us that with an optimal bandwidth the MISE of kernel density estimators decays like $n^{-4/5}$. This is faster than histogram estimators (see Theorem 20.10 and Remark 20.11), but still slower than parameter density estimators (see Remark 20.11 again). Nonetheless it can be shown under weak assumptions that $n^{-4/5}$ is the fastest decay rate that can be guaranteed for *nonparametric* density estimators.

Remark 20.34 (Optimal bandwidth of kernel density estimators)**.** Just like Theorem 20.10 told us how to choose optimally the binwidth of a histogram estimator, Theorem 20.32 tells us how to choose optimally the bandwidth of a kernel density estimator. As before, however (see Remark 20.12), this is of little practical use since the optimal bandwidth depends on the *unknown* density f.

Definition 20.35 (Kernel density estimator risk)**.** Let $X_1, \ldots, X_n$ be an IID sample from a PDF f and let K be a kernel. For every $h > 0$ let $\hat{f}_n$ denote the kernel density estimator with kernel k and bandwidth h. The *kernel density estimator risk associated with* K is the random function from $(0, \infty)$ to $\mathbb{R}$ defined by

$$J(h) := \int \hat{f}_n^2 - 2 \int \hat{f}_n f$$

for every $h > 0$.

Remark 20.36 (Kernel density estimator risk and histogram risk)**.** The kernel density estimator risk associated with a kernel K is defined just like the histogram risk, except that now $h > 0$ is *not* restricted to the set $\mathfrak{h}$ of *admissible* bindwidths.

Definition 20.37 (Leave-one-out cross-validation estimator of the kernel density estimator risk)**.** Let $X_1, \ldots, X_n$ be an IID sample and let K be a kernel. We define the *leave-one-out cross-validation estimator of the kernel density estimator risk associated with* K to be the random function from $(0, \infty)$ to $\mathbb{R}$ defined by, for every $h > 0$,

$$\hat{J}(h) := \int \hat{f}_n^2 - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{(-i)}(X_i)$$

where $\hat{f}_n$ is the kernel density estimator with kernel K and bandwidth h and $\hat{f}_{(-i)}$ is the kernel density estimator with the same kernel and the same bandwidth corresponding to the sample

$$X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n,$$

i.e. the original sample but with X_i removed.

Theorem 20.38 (Properties of the leave-one-out cross-validation estimator of the kernel density estimator risk)**.** *Let $X_1, \ldots, X_n$ be an IID sample, let K be a kernel, let J denote the associated kernel density estimator risk, and let $\hat{J}$ denote the associated leave-one-out cross-validation estimator of the kernel density estimator risk. For every bandwidth h, the following hold.*

(1) $\hat{J}(h)$ is a consistent estimator of $J(h)$, meaning that

$$\mathbb{E}\hat{J}(h) = \mathbb{E}J(h).$$

(2) If the kernel K is symmetric, *meaning that $K(-x) = K(x)$ for all $x \in \mathbb{R}$, then $\hat{J}(h)$ satisfies the approximate identity*

$$\hat{J}(h) = \frac{1}{n^2 h} \sum_{i,j=1}^{n} K^* \left(\frac{X_i - X_j}{h} \right) + \frac{2}{nh} K(0) + \mathcal{O}\left(\frac{1}{n^2 h} \right)$$

where $K^* := K^{(2)} - 2K$ for $K^{(2)} := K * K$ the iterated kernel *(see Exercises A.23.27 and A.23.28).*

Theorem 20.39 (Stone's Theorem)**.** *Let $X_1, \dots, X_n$ be an IID sample from a PDF f which is bounded, let K be a kernel, and let $\hat{J}$ denote the associated leave-one-out cross-validation estimator of the kernel density estimator risk. Let $\hat{f}^*$ denote the kernel density estimator with kernel K and bandwidth h^* selected by $\hat{J}$, i.e.*

$$h^* = \arg\min_{h>0} \hat{J}(h),$$

noting that h^* is a random variable, and, for any $h > 0$, let $\hat{f}_h$ denote the kernel density estimator with the same kernel K and with bandwidth h. Then, for R denoting the MISE,

$$\frac{R\left(f, \hat{f}^*\right)}{\inf_{h>0} R\left(f, \hat{f}_h\right)} \xrightarrow{P} 1 \text{ as } n \to \infty.$$

Remark 20.40 (Stone's Theorem and bandwidth selection for kernel density estimators)**.** Stone's Theorem provides rigorous grouding for, in practice, selecting the bandwidth of a kernel density estimator by minimizing the leave-one-out cross-validation estimator of the kernel density estimator risk.

Remark 20.41 (Discrete Fourier Transform)**.** We may actually use the Discrete Fourier Transform to compute kernel density estimators in practice. In particular, using the Discrete Fourier Transform means being able to use its fast implementation, the *Fast Fourier Transform*. A primer on Fourier analysis and the Discrete Fourier Transform is provided in Appendix D.

Let us write f_n for the kernel density estimator $\hat{f}_n$ in order to reserve the latter notation for the *Fourier transform* of f_n. Then

$$f_n = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h} K\left(\frac{\cdot - X_j}{h}\right) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h} K\left(\frac{\cdot}{h}\right) * \delta_{X_j} = K_h * \underbrace{\left(\frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}\right)}_{=:u_n}.$$

Therefore

$$\hat{f}_n = \hat{K}_h \hat{u}_n$$

where, for the Gaussian kernel,

$$\hat{K}_h = \sqrt{2\pi} K\left(2\pi h \,\cdot\,\right)$$

(see TODO:REF) and

$$\hat{u}_n(\xi) = \frac{1}{n}\sum_{j=1}^{n} \hat{\delta}_{X_j}(\xi)$$

where, by TODO:REF

$$\hat{\delta}_{X_j}(\xi) = \left[\delta_0\left(\cdot - X_j\right)\right]^{\wedge}(\xi) = e^{2\pi i X_j \xi} \underbrace{\hat{\delta}_0}_{=1}$$

such that

$$\hat{u}_n = \frac{1}{n}\sum_{j=1}^{n} e^{2\pi i X_j \xi}.$$

So finally

$$\hat{f}_n(\xi) = \sqrt{2\pi} K\left(2\pi h\xi\right) \cdot \frac{1}{n} \sum_{j=1}^{n} e^{2\pi i X_j \xi}.$$

Now choose

$$\begin{cases} T := X_{(n)} - X_{(1)} \text{ and} \\ B := \dfrac{2^J}{T} \text{ for some integer } J \geqslant 1\text{, such that } BT = 2^J, \end{cases}$$

and define

$$\begin{aligned} \hat{f}_n[k] &:= \frac{1}{T} \hat{f}\left(\frac{k}{T}\right) \\ &= \frac{\sqrt{2\pi}}{nT} K\left(\frac{2\pi hk}{T}\right) \sum_{j=1}^{n} e^{2\pi i \frac{X_j k}{T}} \text{ for } 0 \leqslant k \leqslant 2^J - 1 = BT - 1. \end{aligned}$$

We may then recover f_n via the inverse DFT (TODO:ADDREF), namely

$$f_n\left(\frac{m}{B}\right) = Bf_n[m] \iff f_n[m] = \frac{1}{B} f_n\left(\frac{m}{B}\right) \text{ for } 0 \leqslant m \leqslant 2^J - 1$$

where

$$f_n[m] = \frac{1}{B} \sum_{k=0}^{2^J-1} \hat{f}_n[k] e^{2\pi i \frac{km}{BT}}.$$

Lemma 20.42 (Expectation of kernel density estimator)**.** *Let $X_1, \ldots, X_n$ be an IID sample from a PDF f and let $\hat{f}_n$ be the kernel density estimator with kernel K and bandwidth h. Viewing $\mathbb{E}\hat{f}_n$ as a function from $\mathbb{R}$ to $\mathbb{R}$ we have that*

$$\mathbb{E}\hat{f}_n = \frac{1}{h} K\left(\frac{\cdot}{h}\right) * f.$$

Theorem 20.43 (Confidence band using kernel density estimators)**.** *Consider an IID sample $X_1, \ldots, X_n$ from a PDF f supported on a bounded interval of length l, let K be a kernel supported on a bounded interval of length ω, let $h = h(n) > 0$, and let $\hat{f}_n$ be the kernel density estimator with kernel K and bandwidth h. For any $\alpha \in (0, 1)$ let*

$$m := \frac{l}{\omega} \text{ and } q := \Phi^{-1}\left(\frac{1 + (1-\alpha)^{1/m}}{2}\right)$$

for Φ denoting the CDF of the standard Normal distribution. For $x \in \mathbb{R}$ let

$$Y_i(x) := \frac{1}{h} K\left(\frac{x - X_i}{h}\right),$$

noting that $\overline{Y}_n(x) = \hat{f}_n(x)$, let

$$s^2(x) := \frac{1}{n-1} \sum_{i=1}^{n} \left[Y_i(x) - \overline{Y}_n(x)\right]^2$$

be the sample variance of $Y_1, \ldots, Y_n$, let

$$se(x) := \frac{s(x)}{\sqrt{n}},$$

and let

$$l_n(x) := \max\left(\hat{f}_n(x) - q\, se(s),\, 0\right) \text{ and } u_n(x) := \hat{f}_n(x) + q\, se(s).$$

If $h(n) \to 0$ as $n \to \infty$ and $nh(n) \to \infty$ sufficiently fast as $n \to \infty$ then (l_n, u_n) is an asymptotic $1-\alpha$ confidence band for the smoothed *density*

$$\bar{f}_n := \frac{1}{h} K\left(\frac{\cdot}{h}\right) * f,$$

noting that $\bar{f}_n$ depends on n only through $h = h(n)$, meaning that

$$\liminf_{n\to\infty} \mathbb{P}\Big(l_n(x) \leqslant \bar{f}_n(x) \leqslant u_n(x) \text{ for all } x \in \mathbb{R}\Big) \geqslant 1-\alpha.$$

Remark 20.44 (Confidence bands for kernels with unbounded support)**.** Theorem 20.43 requires that the kernel used have bounded support. This fails for the Gaussian kernel, for example. In that specific case we would take

$$\omega = 3h,$$

i.e. three standard deviations, to ensure that the interval $\left(-\frac{\omega}{2}, \frac{\omega}{2}\right)$ captures all the regions where the kernel is non-negligible.

Remark 20.45 (Confidence bands using kernel density estimation and smoothed out density)**.** When using histogram estimators to construct a confidence band for the density in Theorem 20.24 we are only able to guarantee that the confidence band covers the *histogram projection* of the density, not the density itself. This is motivated by Remark 20.22 which tells us that this histogram projection is a genuine projection (as proved in Exercise A.23.26).

Here no such projection perspective characterizes the smoothed density

$$\bar{f} := K * f.$$

Indeed the operator $f \mapsto K * f$ is *not* idempotent since, as seen in Exercise A.23.27, applying it twice is the same as convolving with a kernel whose variance is *twice* that of K. In other words:

$$K * (K * f) = K^{(2)} * f \neq K * f \text{ since } \sigma^2_{K^{(2)}} = \sigma^2_K$$

for $K^{(2)}$ the *iterated kernel* defined in Exercise A.23.27 (and used in Theorem 20.38) and for σ^2_K denoting the variance of the kernel K (as in Definition 20.26). Since projections are necessarily idempotent this means that $f \mapsto K * f$ is *not* a projection.

So where does $\bar{f} = K * f$ come from? An answer is suggested by Lemma 20.42: the kernel density estimator *only sees* the smoothed out version of the density. More precisely, inspired by Lemma 20.42 we may write

$$\begin{aligned}\hat{f}_n &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{\cdot}{h}\right) * \delta_{X_i} \\ &= \frac{1}{h} K\left(\frac{\cdot}{h}\right) * \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}\end{aligned}$$

where δ_{X_i} is a point mass at X_i. Even with an "infinite amount" of data, such that, in the sense of distributions,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} = f,$$

the kernel density estimator would only recover

$$\begin{aligned}\hat{f}_\infty &= \frac{1}{h}K\left(\frac{\cdot}{h}\right) * \left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right)\\ &= \frac{1}{h}K\left(\frac{\cdot}{h}\right) * f\\ &= \bar{f}.\end{aligned}$$

Nonetheless, here is the saving grace: we ask that $h \to 0$ as $n \to \infty$. Therefore, as $n \to \infty$,

$$\hat{f} = \frac{1}{h}K\left(\frac{\cdot}{h}\right) * f \to f,$$

and so as $n \to \infty$ the confidence band of Theorem 20.43 covers an ever-closer approximation of the true density.

20.4. **Nonparametric Regression.**

Definition 20.46 (Nadaraya-Watson kernel estimator)**.** Consider an IID sample $(Y_1, X_1), \ldots, (Y_n, X_n)$, let $h > 0$, and let K be a kernel. The *Nadaraya-Watson kernel estimator* is the random function from $\mathbb{R}$ to $\mathbb{R}$ defined by, for every $x \in \mathbb{R}$,

$$\hat{r}(x) := \sum_{i=1}^{n} w_i(x)Y_i \text{ for } w_i(x) := \frac{K\left(\frac{x-X_i}{h}\right)}{\sum\limits_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)},$$

and we call h the *bandwidth*.

Remark 20.47 (Motivation for the Nadaraya-Watson kernel estimator)**.** Estimators of the regression function

$$r(x) = \mathbb{E}\left(Y \mid X = x\right)$$

often take the form $\tilde{r}(x) = \sum_{i=1}^{n} \tilde{w}_i(x)Y_i$ where $\tilde{r}$ is a weighted average of the Y_i's. Typically the weight $\tilde{w}_i$ give higher weights to values near X_i. The Nadaraya-Watson kernel estimator does so by first performing a kernel density estimation of the joint density of (X, Y) and then using the estimated density $\hat{f}$ to evaluate

$$\hat{r}(x) := \hat{\mathbb{E}}\left(Y \mid X = x\right) := \int y\hat{f}(y|x)dy = \frac{\int y\hat{f}(x,y)dy}{\int \hat{f}(x,y)dy}.$$

Indeed, if we use the kernel

$$K_h(x,y) := \frac{1}{h^2}K\left(\frac{x}{h}\right)K\left(\frac{y}{h}\right)$$

then we may compute that

$$\begin{aligned}\frac{\int y\hat{f}(x,y)dy}{\int \hat{f}(x,y)dy} &= \frac{\int \frac{y}{n}\sum_{i=1}^{n}\frac{1}{h^2}K\left(\frac{x-X_i}{h}\right)K\left(\frac{y-Y_i}{h}\right)dy}{\int \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h^2}K\left(\frac{x-X_j}{h}\right)K\left(\frac{y-Y_j}{h}\right)dy} \\ &= \frac{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\int yK\left(\frac{y-Y_i}{h}\right)dy}{\sum_{j=1}^{n}K\left(\frac{x-X_j}{h}\right)\underbrace{\int K\left(\frac{y-Y_j}{h}\right)dy}_{=1}} \\ &= \sum_{i=1}^{n}\left(\underbrace{\frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^{n}K\left(\frac{x-X_i}{h}\right)}}_{=w_i(x)}Y_i\right)\end{aligned}$$

as desired since K has mean zero and hence $K\left(\frac{\cdot - Y_i}{h}\right)$ has mean Y_i.

Theorem 20.48 (MISE of the Nadaraya-Watson kernel estimator)**.** *Consider an IID sample* $(Y_1, X_1), \ldots, (Y_n, X_n)$ *from a random vector* (Y, X) *with PDF* f*, let* $r(x) := \mathbb{E}(Y \mid X = x)$ *be the regression function between* Y *and* X*, consider* $\sigma^2(x) := \mathbb{V}(Y \mid X = x)$*, and let* $\hat{r}_n$ *be the Nadaraya-Watson kernel estimator with kernel* K *and bandwidth* h*. As an estimator of the regression function* r*, the MISE of* $\hat{r}_n$ *satisfies*

$$R(r, \hat{r}_n) \approx \frac{h^4\sigma_K^2}{4}\int\left(r'' + 2r'\frac{f'}{f}\right)^2 + \frac{\int \sigma^2/f}{nh}\int K^2.$$

In particular this is minimized with respect to h *at* h^* *proportional to* $n^{-1/5}$ *and if we denote by* $\hat{r}_n^*$ *the Nadaraya-Watson kernel estimator with the same kernel and bandwidth* h^* *then its MISE is proportional to* $n^{-4/5}$*.*

Definition 20.49 (Leave-one-out cross-validation estimator of the Nadaraya-Watson kernel estimator risk)**.** Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample and let K be a kernel. We define the *leave-out-out cross-validation estimator of the Nadaraya-Watson kernel estimator risk associated with* K to be the random function from $(0, \infty)$ to $(0, \infty)$ defined by, for every $h > 0$,

$$\hat{J}(h) := \sum_{i=1}^{n}\left[Y_i - \hat{r}_{(-i)}(X_i)\right]^2$$

where $\hat{r}_{(-i)}$ is the Nadaraya-Watson kernel estimator with kernel K and bandwidth h corresponding to the sample

$$(Y_1, X_1), \ldots, (Y_{i-1}, X_{i-1}), (Y_{i+1}, X_{i+1}), \ldots, (Y_n, X_n),$$

i.e. the original sample but with (Y_i, X_i) removed.

Theorem 20.50 (Identity for the leave-one-out cross-validation estimator of the Nadaraya-Watson kernel estimator risk)**.** *Let* $(Y_1, X_1), \ldots, (Y_n, X_n)$ *be an IID sample, let* K *be a kernel, and let* $\hat{r}$ *be the Nadaraya-Watson kernel estimator with* *kernel K and bandwidth h. The leave-one-out cross-validation estimator of the Nadaraya-Watson kernel estimator risk associated with K may be written as*

$$\hat{J}(h) = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{r}(X_i)}{1 - \frac{K(0)}{\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)}} \right)^2 .$$

Theorem 20.51 (Confidence band using the Nadayara-Watson kernel estimator)**.** *Consider an IID sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ from a PDF f whose marginal PDF for X is supported on a bounded interval of length l, let K be a kernel supported on a bounded interval of length ω, let $h = h(n) > 0$, and let $\hat{r}_n$ be the Nadaraya-Watson kernel estimator with kernel K and bandwidth h. Let*

$$\hat{\sigma}^2 := \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left(\widetilde{Y}_{i+1} - \widetilde{Y}_i \right)^2$$

where

$$\widetilde{Y}_j = Y_i \iff X_j = X_{(i)},$$

i.e. we order *$X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ and label the $\widetilde{Y}_j$'s accordingly. For any $\alpha \in (0, 1)$ let*

$$m := \frac{l}{\omega} \text{ and } q := \Phi^{-1}\left(\frac{1 + (1-\alpha)^{1/m}}{2} \right)$$

for Φ denoting the CDF of the standard Normal distribution. For $x \in \mathbb{R}$ let

$$se(x) := \hat{\sigma} \sqrt{\sum_{i=1}^{n} w_i^2(x)}$$

for w_i as in Definition 20.46, and let

$$l_n(x) := \hat{r}_n(x) - q\, se(s) \text{ and } u_n(x) := \hat{r}_n(x) + q\, se(s).$$

If $h(n) \to 0$ as $n \to \infty$ and $nh(n) \to \infty$ sufficiently fast as $n \to \infty$ then (l_n, u_n) is an asymptotic $1 - \alpha$ confidence band for the smoothed *version*

$$\bar{r}_n(x) := \mathbb{E}\left[\hat{r}_n(x)\right] \text{ for all } x,$$

of the true regression function $r(x) := \mathbb{E}\left(Y \mid X = x\right)$ between Y and X, meaning that

$$\liminf_{n \to \infty} \mathbb{P}\Big(l_n(x) \leqslant \bar{r}_n(x) \leqslant u_n(x) \text{ for all } x \in \mathbb{R} \Big) \geqslant 1 - \alpha.$$

As in Remark 20.44 we note that if the kernel does not *have bounded support then we take $\omega = 3h$.*

21. Smoothing Using Orthogonal Functions

21.1. Orthogonal Functions and L_2 Spaces.

Definition 21.1 (Inner product space)**.** An *inner product space* is a pair $(V, \langle\,\cdot\,,\,\cdot\,\rangle)$ where V is a vector space and $\langle\,\cdot\,,\,\cdot\,\rangle$ is a symmetric, bilinear, and positive-definite map $\langle\,\cdot\,,\,\cdot\,\rangle : V \times V \to \mathbb{R}$ which is called an *inner product*.

Example 21.2 (Inner product spaces)**.** Here are some examples of inner product spaces.

(1) $\mathbb{R}^d$ equipped with the *Euclidean inner product*

$$\langle x,\, y\rangle := \sum_{j=1}^{d} x_j y_j \text{ for all } x,\, y \in \mathbb{R}^d.$$

(2) The space $L_2[a,\, b]$ of square-integrable functions on $[a,\, b]$ may be equipped with the L_2 *inner product*

$$\langle f,\, g\rangle := \int_a^b f(x)g(x) \text{ for all } f,\, g \in L_2[a,\, b].$$

Definition 21.3 (Terminology for inner product spaces)**.** Let $(V, \langle\,\cdot\,,\,\cdot\,\rangle)$ be an inner product space.

- The map $||\cdot|| : V \to \mathbb{R}$ defined by

$$||v|| := \sqrt{\langle v,\, v\rangle}$$

  for any $v \in V$ is called the *norm*.
- Two vectors $v,\, w \in V$ are said to be *orthogonal* if $\langle v,\, w\rangle = 0$.
- A vector $v \in V$ is said to be *normal* if $||v|| = 1$.
- A (possibly infinite) set of mutually orthogonal normal vectors is called *orthonormal*.

Lemma 21.4 (Decomposition along an orthonormal basis)**.** *Let $\{\phi_j\}_{j=1}^n$ be an orthonormal basis of an inner product space* $(V, \langle\,\cdot\,,\,\cdot\,\rangle)$*. For any $v \in V$,*

$$v := \sum_{j=1}^{n} \langle v,\, \phi_j\rangle \phi_j.$$

Definition 21.5 (Hilbert space)**.** A *Hilbert space* is an inner product space which is complete as a metric space when using the norm as metric.

Definition 21.6 (Completeness)**.** A set $\{\phi_j\}_{j=0}^{\infty}$ in a Hilbert space is called *complete* if the only element of H orthogonal to every ϕ_j is the zero vector.

Theorem 21.7 (Characterization of orthonormal bases in Hilbert spaces)**.** *Let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal sequence in a Hilbert space H. The following are equivalent.*

(1) $(\phi_j)_{j=0}^{\infty}$ is an orthonormal basis, *i.e.*

$$f = \sum_{j=0}^{\infty} \langle f,\, \phi_j\rangle \phi_j \textit{ for all } f \in H,$$

meaning that

$$\left|\left|f - \sum_{j=0}^{n} \langle f,\, \phi_j\rangle \phi_j\right|\right|_H \to 0 \text{ as } n \to \infty.$$

(2) *For every $f \in H$, if we define $\beta_j := \langle f,\, \phi_j\rangle$ then*

$$||f||_H^2 = \sum_{j=0}^{\infty} \beta_j^2 = ||\beta||_{l^2}^2.$$

This is known as Parseval's identity.

(3) *$(\phi_j)_{j=0}^{\infty}$ is complete.*

Example 21.8 (Orthonormal bases of L_2)**.** Here are some examples of orthonormal bases of L_2.

(1) The *cosine basis* $(\phi_j)_{j=0}^{\infty}$ of $L_2[0,\, 1]$ is defined by, for $x \in [0,\, 1]$,

$$\phi_0(x) = 1 \text{ and } \phi_j(x) := \sqrt{2}\cos(j\pi x) \text{ for } j \geqslant 1.$$

It is an orthonormal basis of $L_2[0,\, 1]$.

(2) The *Legendre polynomials* on $[-1,\, 1]$ defined by

$$P_j(x) := \frac{1}{2^j j!} \frac{d^j}{dx^j} \left(x^2 - 1\right)^j \text{ for } x \in [-1,\, 1] \text{ and } j \geqslant 0$$

are complete and mutually orthogonal in $L_2[-1,\, 1]$. Moreover the sequence $(\phi_j)_{j=0}^{\infty}$ defined by

$$\phi_j := \sqrt{\frac{2j+1}{2}} P_j$$

is an orthonormal basis. The Legendre polynomials satisfy the recurrence relation

$$P_{j+1}(x) = \frac{(2j+1)xP_j(x) - jP_{j-1}(x)}{j+1} \text{ for } j \geqslant 1,$$

with

$$P_0(x) = 1 \text{ and } P_1(x) = x.$$

The next couple of Legendre polynomials are

$$P_2(x) = \frac{1}{2}\left(3x^2 - 1\right) \text{ and } P_3(x) = \frac{1}{2}\left(5x^3 - 3x\right).$$

Remark 21.9 (Cosine basis and Fourier basis)**.** We may relate the cosine basis to Fourier series as follows. Given a function $f \in L_2[0,\, 1]$ we extend it to an even function $\tilde{f}$ on $[-1,\, 1]$, namely

$$\tilde{f}(x) := \begin{cases} f(-x) & \text{if } x \in [-1,\, 0) \text{ and} \\ f(x) & \text{if } x \in [0,\, 1]. \end{cases}$$

The Fourier series of the extension $\tilde{f}$ of f is then

$$\tilde{f}(x) = \sum_{k\in\mathbb{Z}} c_k e^{\pi i k x} = a_0 + \sum_{k=1}^{\infty} a_k \cos(\pi k x) + \sum_{k=1}^{\infty} b_k \sin(\pi k x)$$

for

$$c_k = \frac{1}{2}\int_{-1}^{1} \tilde{f}(x)e^{-\pi ikx}dx$$

and

$$\begin{aligned} a_0 &= c_0 = \frac{1}{2}\int_{-1}^{1} \tilde{f}(x)dx, \\ a_k &= c_k + c_{-k} = \int_{-1}^{1} \tilde{f}(x)\cos(\pi kx)dx, \text{ and} \\ b_k &= i(c_k - c_{-k}) = \int_{-1}^{1} \tilde{f}(x)\sin(\pi kx)dx. \end{aligned}$$

In particular: $\tilde{f}$ is even but $\sin(k\pi\,\cdot\,)$ is odd,, and so $b_k \equiv 0$ for all k. This means that

$$\tilde{f}(x) = a_0 + \sum_{k=1}^{\infty} a_k \cos(\pi kx).$$

Restricting $\tilde{f}$ to $[0,\, 1]$ then yields the cosine basis for f:

$$f(x) = \tilde{f}|_{[0,\,1]}(x) = a_0 + \sum_{k=1}^{\infty} a_k\cos(\pi kx) = a_0 + \sum_{k=1}^{\infty}\frac{a_k}{\sqrt{2}}\cdot\sqrt{2}\cos(\pi kx) = \sum_{j=0}^{\infty}\beta_j\phi_j(x)$$

for

$$\beta_j = \begin{cases} a_0 & \text{if } j = 0 \text{ and} \\ \dfrac{a_j}{\sqrt{2}} & \text{if } j \geqslant 1. \end{cases}$$

The normalization $\alpha \mapsto \frac{\alpha_j}{\sqrt{2}} = \beta_j$ comes from the fact that the cosine functions used in the Fourier series are normal on the *extended* domain $[-1,\, 1]$, not on $[0,\, 1]$. We can also verify this directly since

$$\begin{aligned} a_k &= \int_{-1}^{1} \tilde{f}(x)\cos(\pi kx)dx \\ &= \int_{-1}^{0} f(-x)\cos(\pi kx)dx + \int_{0}^{1} f(x)\cos(\pi kx)dx \\ &= 2\int_{0}^{1} f(x)\cos(\pi kx)dx \\ &= \sqrt{2}\langle f,\, \phi_k\rangle \\ &= \sqrt{2}\beta_k. \end{aligned}$$

Remark 21.10 (Smoothness and decay of the coefficients along the cosine basis)**.** Since Remark 21.9 tells us that the cosine basis is related to Fourier series, it is not surprising that the decay of the coefficients β_j of a function along the cosine basis tell us something about the smoothness of f, and vice-versa. We can indeed verify this directly. Integration by parts tells us that

$$\langle f'',\, \phi_j\rangle = \langle f,\, \phi_j''\rangle = -(j\pi)^2\langle f,\, \phi_j\rangle.$$

Therefore

$$f'' \in L^2 \iff (j\pi)^2\beta_j \in l^2,$$

where as usual $\beta_j = \langle f,\, \phi_j\rangle$, and similarly for higher-order derivatives.

Careful: The smoothness in question is *not* the smoothness of f! Instead, the discussion above characterizes the smoothness of the *periodic extension* of f, namely

$$P_1\left(f|_{[0,\,1]}\right)$$

where P_1 denotes the periodic extension operator. If f is already 1–periodic, then there is no distinction to be made. Otherwise, beware. (See item 2 of Exercise A.21.5 for an example).

21.2. Density Estimation.

Definition 21.11 (Empirical estimates of the basis coefficients)**.** Let f be a PDF in $L_2[0,\,1]$, let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal basis of $L_2[0,\,1]$, and let $X_1,\,\dots\,,\,X_n\sim f$ be IID. For every $j\geqslant 0$

$$\hat{\beta}_j := \frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i),$$

which is the sample mean of $\phi_j(X_1),\,\dots,\,\phi_j(X_n)$, is called the *empirical estimate of the basis coefficient* β_j where $(\beta_j)_{j=0}^{\infty}$ is defined implicitly by

$$f=\sum_{j=0}^{\infty}\beta_j\phi_j.$$

Remark 21.12 (Empirical estimates of the basis coefficients)**.** Why are the estimates of Definition 21.11 called “empirical”? This is because if we formally apply the decomposition

$$f=\sum_{j=0}^{\infty}\langle f,\,\phi_j\rangle\phi_j$$

(see Theorem 21.7) to the *empirical* probability mass function (derived from the empirical CDF)

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$$

we then obtain

$$\left\langle\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i},\,\phi_j\right\rangle=\frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i)=\hat{\beta}_j$$

and so the empirical PMF may be written, formally,

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}=\sum_{j=0}^{\infty}\hat{\beta}_j\phi_j.$$

In other words the empirical estimates $\hat{\beta}_j$ are the *exact* (formal) coefficients of the *empirical PMF.*

Theorem 21.13 (Properties of the empirical estimates of the basis coefficients)**.** *Let f be a PDF in $L_2[0,\,1]$, let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal basis of $L_2[0,\,1]$, let $X_1,\,\dots\,,\,X_n\sim f$ be IID, and let $(\hat{\beta}_j)_{j=0}^{\infty}$ be the corresponding empirical estimates of the basis coefficients. Then*

$$\mathbb{E}\hat{\beta}_j=\beta_j:=\langle f,\,\phi_j\rangle \textit{ and}$$

$$\mathbb{V}\hat{\beta}_j=\frac{\sigma_j^2}{n}\textit{ for }\sigma_j^2:=\mathbb{V}\left[\phi_j(X)\right]=\int_0^1\left[\phi_j(x)-\beta_j\right]^2 f(x)dx.$$

Remark 21.14 (Basis coefficients of PDFs)**.** A key observation underpinning Theorem 21.13 is the following. For a generic $f \in L_2[0, 1]$ its basis coefficients are given by

$$\beta_j = \langle f, \phi_j \rangle.$$

However, if f is also a PDF, then its coefficients also satisfy

$$\beta_j = \mathbb{E}\left[\phi_j(X)\right]$$

for $X \sim f$. Indeed:

$$\mathbb{E}\left[\phi_j(X)\right] = \int_0^1 \phi_j(x) f(x) dx = \langle f, \phi_j \rangle = \beta_j,$$

as desired.

Definition 21.15 (Orthogonal function density estimator)**.** Let f be a PDF in $L_2[0, 1]$, let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal basis of $L_2[0, 1]$, let $X_1, \ldots, X_n \sim f$ be IID, let $(\hat{\beta}_j)_{j=0}^{\infty}$ be the corresponding empirical estimates of the basis coefficients, and let $1 \leqslant J \leqslant \sqrt{n}$ be an integer. The random function from $[0, 1]$ to $\mathbb{R}$ defined by

$$\hat{f}(x) := \sum_{j=0}^{J} \hat{\beta}_j \phi_j(x) \text{ for every } x \in [0, 1]$$

is called the *orthogonal function density estimator* with *smoothing parameter* J.

Remark 21.16 (Orthogonal function density estimator and the empirical PMF)**.** Remark 21.12 told us that the empirical PMF

$$f_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

could be written, formally,

$$\frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} = \sum_{j=0}^{\infty} \hat{\beta}_j \phi_j.$$

Therefore the orthogonal function density estimator $\hat{f}$ is precisely a smoothed approximation of the empirical PMF f_n where the smoothing comes from truncation the infinite series.

Theorem 21.17 (MISE of the orthogonal function density estimator)**.** *Let f be a PDF in $L_2[0, 1]$, let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal basis of $L_2[0, 1]$, consider IID $X_1, \ldots, X_n \sim f$, and let $\hat{f}$ be the orthogonal function density estimator with smoothing parameter J. The MISE of $\hat{f}$ is*

$$R\left(f, \hat{f}\right) = \sum_{j=0}^{J} \frac{\sigma_j^2}{n} + \sum_{j=J+1}^{\infty} \beta_j^2$$

for β_j and σ_j^2 as in Theorem 21.13.

Definition 21.18 (Estimate of the orthogonal function density estimator risk)**.** Let f be a PDF in $L_2[0, 1]$, let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal basis of $L_2[0, 1]$, let

$X_1, \ldots, X_n \sim f$ be IID, and let $(\hat{\beta}_j)_{j=0}^{\infty}$ be the corresponding empirical estimates of the basis coefficients. For every $j \geqslant 0$ we define

$$\hat{\sigma}_j^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left[\phi_j(X_i) - \hat{\beta}_j \right]^2,$$

which is the sample variance of $\phi_j(X_1), \ldots, \phi_j(X_n)$, and for integers $1 \leqslant J \leqslant \sqrt{n}$ we define

$$\hat{R}(J) := \sum_{j=0}^{J} \frac{\hat{\sigma}_j^2}{n} + \sum_{j=J+1}^{\sqrt{n}} \left(\hat{\beta}_j^2 - \frac{\hat{\sigma}_j^2}{n} \right)_+,$$

which we call the *estimate of the orthogonal function density estimator risk*, where $s_+ := \max(s, 0)$ is called the *positive part* of s.

Remark 21.19 (Estimate of the orthogonal function density estimator risk)**.** Why is $\hat{R}(J)$ in Definition 21.18 a good estimate of the MISE of the orthogonal function density estimator $\hat{f}$ given by

$$R\left(f, \hat{f}\right) = \sum_{j=0}^{J} \frac{\sigma_j^2}{n} + \sum_{j=J+1}^{\infty} \beta_j^2$$

(see Theorem 21.17)?

The variance term $\sum_{j=0}^{J} \frac{\sigma_j^2}{n}$ is easy to estimate: since σ_j^2 is the variance of $\phi_j(X)$ (see Theorem 21.13) we estimate it by its *sample* variance $\hat{\sigma}_j^2$.

The bias term $\sum_{j=J+1}^{\infty} \beta_j^2$, besides the sensible truncation (here at $j \leqslant \sqrt{n}$) and taking the positive part, requires a little bit more thought. Since β_j and σ_j^2 are the mean and variance of $\phi_j(X)$, respectively (see Theorem 21.13), and since $\hat{\beta}_j$ and $\hat{\sigma}_j^2$ are their sample counterparts, we deduce that

$$\mathbb{E}\left[\hat{\beta}_j^2\right] = \left(\mathbb{E}\hat{\beta}_j\right)^2 + \mathbb{V}\hat{\beta}_j = \beta_j^2 + \frac{\sigma_j^2}{n},$$

and so

$$\mathbb{E}\left[\hat{\beta}_j^2 - \frac{\hat{\sigma}_j^2}{n}\right] = \beta_j^2,$$

which warrants using $\hat{\beta}_j^2 - \frac{\hat{\sigma}_j^2}{n}$ as an estimate for β_j^2.

Remark 21.20 (Density estimation via orthogonal functions in practice)**.** In practice we use the risk estimate of Definition 21.18 to select the smoothing parameter J which minimizes $\hat{R}(J)$ over $1 \leqslant J \leqslant \sqrt{n}$.

Theorem 21.21 (Confidence bands for orthogonal function density estimation)**.** *Let f be a PDF in $L_2[0, 1]$, let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal basis of $L_2[0, 1]$, let $X_1, \ldots, X_n \sim f$ be IID, and let $\hat{f}$ be the orthogonal function density estimator with smoothing parameter J. For any $\alpha \in (0, 1)$ let*

$$c := K^2 \sqrt{\frac{J \chi^2_{J,\alpha}}{n}} \text{ for } K := \max_{1 \leqslant j \leqslant J} \max_{0 \leqslant x \leqslant 1} |\phi_j(x)|$$

(such that $K = \sqrt{2}$ for the cosine basis whenever $J \geqslant 2$) and let

$$l(x) := \hat{f}(x) - c \text{ and } u(x) := \hat{f}(x) + c$$

for every $x \in [0,\, 1]$*. Then* $(l,\, u)$ *is an asymptotic* $1-\alpha$ *confidence band for the* smoothed *version*

$$f_J := \sum_{j=0}^{J} \beta_j \phi_j \text{ for } \beta_j := \langle f,\, \phi_j \rangle$$

of f*, meaning that*

$$\liminf_{n\to\infty} \mathbb{P}\left(l(x) \leqslant f_J(x) \leqslant u(x) \text{ for all } x \in [0,\, 1]\right) \geqslant 1-\alpha.$$

21.3. Regression.

Remark 21.22 (Assumption underlying nonparametric regression)**.** In this section we seek to estimate the regression function $r(x) := \mathbb{E}\left(Y \mid X = x\right)$ given samples from a random vector $(X,\, Y)$. Throughout this section we will *assume* that

$$X \sim \text{Uniform}(0,\, 1).$$

In some sense, this assumption is *generic*: if we know the CDF F_Z of a random variable Z then Exercise A.2.14 tells us that

$$X := F_Z(Z) \sim \text{Uniform}(0,\, 1).$$

In other words *any* random variable can be turned into a Uniform(0, 1) random variable for free using its CDF.

Of course, in *practice* we are only given samples of $(Y,\, Z)$ and do not know the CDF F_Z of Z. This is where Section 21.2 comes in: we can *estimate* the PDF of Z, deduce an estimate for its CDF, and then transform Z using that CDF estimate to obtain a Uniform(0, 1) random variable.

Definition 21.23 (Empirical estimates of the regression function basis coefficients)**.** Let f be the PDF of a random vector $(Y,\, X)$ whose marginal for X is a Uniform(0, 1) distribution, let the regression function $r(x) := \mathbb{E}\left(Y \mid X = x\right)$ between Y and X be in $L_2[0,\, 1]$, let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal basis of $L_2[0,\, 1]$, and let $(Y_1,\, X_1)\,, \ldots,\, (Y_n,\, X_n) \sim f$ be IID. For every $j \geqslant 0$

$$\hat{\beta}_j := \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(X_i),$$

which is the sample mean of $Y_1\phi_j(X_1),\, \ldots,\, Y_n\phi_j(X_n)$, is called the *empirical estimate of the regression function basis coefficients* β_j where $(\beta_j)_{j=0}^{\infty}$ is defined implicitly by

$$r = \sum_{j=0}^{\infty} \beta_j \phi_j.$$

Remark 21.24 (Empirical estimates of the regression function basis coefficients and uniformity assumption)**.** In *general* if we write the regression function as

$$r = \sum_{j=0}^{\infty} \beta_j \phi_j$$

then we can compute the regression function basis coefficients β_j as follows:

$$\begin{aligned}\beta_j &= \int_0^1 r(x)\phi_j(x)dx \\ &= \int_0^1 \mathbb{E}\left(Y \mid X = x\right)\phi_j(x)dx \\ &= \int_0^1 \left[\int_{\mathbb{R}} yf(y|x)dy\right]\phi_j(x)dx \\ &= \int_0^1 \int_{\mathbb{R}} y\frac{f(x,y)}{f_X(x)}\phi_j(x)dydx \\ &= \int_0^1 \int_{\mathbb{R}} \frac{y\phi_j(x)}{f_X(x)}f(x,y)dydx \\ &= \mathbb{E}\left[\frac{Y\phi_j(X)}{f_X(X)}\right].\end{aligned}$$

We may then approximate β_j empirically via its sample mean

$$\tilde{\beta}_j := \frac{1}{n}\sum_{i=1}^n \frac{Y_i\phi_j(X_i)}{f_X(X_i)}.$$

We encounter the same issue discussed in Remark 21.22: we now need the marginal f_X. This is where the uniformity assumption

$$X \sim \text{Uniform}(0,\, 1)$$

comes in. Under that assumption $f_X \equiv 1$ and so $\tilde{\beta}_j$ simplifies to

$$\tilde{\beta}_j = \frac{1}{n}\sum_{i=1}^n Y_i\phi_j(X_i) = \hat{\beta}_j$$

for $\hat{\beta}_j$ as in Definition 21.23. In other words: the empirical estimates $\hat{\beta}_j$ do indeed estimate the coefficients of the regression function r, under the assumption that $X \sim \text{Uniform}(0,\, 1)$.

Theorem 21.25 (Limiting behaviour of the empirical estimates of the regression function basis coefficients)**.** *Let f be the PDF of a random vector $(Y,\, X)$ whose marginal for X is a Uniform(0, 1) distribution, let the regression function $r(x) := \mathbb{E}\left(Y \mid X = x\right)$ between Y and X be in $L_2[0,\, 1]$, let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal basis of $L_2[0,\, 1]$, let $(Y_1,\, X_1),\, \dots,\, (Y_n,\, X_n) \sim f$ be IID, and let $(\hat{\beta}_j)_{j=0}^{\infty}$ be the corresponding empirical estimates of the regresion function basis coefficients. Assume that*

$$\mathbb{V}\left(Y \mid X = x\right) = \sigma^2 \textit{ for every } x \in [0,\, 1].$$

This assumption is known as homoscedasticity*. Then, for every $j \geqslant 0$,*

$$\mathbb{E}\left[Y\phi_j(X)\right] = \beta_j \textit{ and } \mathbb{V}\left[Y\phi_j(X)\right] = \sigma^2.$$

such that

$$\mathbb{E}\left(\hat{\beta}_j\right) = \beta_j \textit{ and } \mathbb{V}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{n},$$

and so

$$\hat{\beta}_j \sim N\left(\beta_j,\, \frac{\sigma^2}{n}\right)$$

in distribution as $n \to \infty$ for $\beta_j := \langle r, \phi_j \rangle$.

Remark 21.26 (Homoscedasticity)**.** The homoscedasticity assumption

$$\mathbb{V}(Y \mid X = x) = \sigma^2 \text{ for all } x \in [0, 1]$$

is an assumption about the *homogeneity in x* of the conditional variance $\mathbb{V}(Y \mid X = x)$.

Definition 21.27 (Orthogonal function regression estimator)**.** Let f be the PDF of a random vector (Y, X) whose marginal for X is a Uniform$(0, 1)$ distribution, let the regression function $r(x) := \mathbb{E}(Y \mid X = x)$ between Y and X be in $L_2[0, 1]$, let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal basis of $L_2[0, 1]$, let $(Y_1, X_1), \dots, (Y_n, X_n) \sim f$ be IID, let $(\hat{\beta}_j)_{j=0}^{\infty}$ be the corresponding empirical estimates of the regression function basis coefficients, and let $1 \leqslant J \leqslant n$ be an integer. The random function from $[0, 1]$ to $\mathbb{R}$ defined by

$$\hat{r}(x) := \sum_{j=0}^{J} \hat{\beta}_j \phi_j(x)$$

is called the *orthogonal function regression estimator* with *smoothing parameter J*.

Theorem 21.28 (MISE of the orthogonal function regression estimator)**.** *Let f be the PDF of a random vector (Y, X) whose marginal for X is a Uniform$(0, 1)$ distribution, let the regression function $r(x) := \mathbb{E}(Y \mid X = x)$ between Y and X be in $L_2[0, 1]$, let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal basis of $L_2[0, 1]$, consider an IID sample $(Y_1, X_1), \dots, (Y_n, X_n) \sim f$, and let $\hat{r}$ be the orthogonal function regression estimator with smoothing parameter J. Assume homoscedasticity, i.e.*

$$\mathbb{V}(Y \mid X = x) = \sigma^2 \textit{ for all } x \in [0, 1].$$

Then the MISE of $\hat{r}$ is

$$R(r, \hat{r}) = \frac{J\sigma^2}{n} + \sum_{j=J+1}^{\infty} \beta_j^2$$

for β_j as in Theorem 21.25.

Definition 21.29 (Estimate of the orthogonal function regression estimator risk)**.** Let f be the PDF of a random vector (Y, X) whose marginal for X is a Uniform$(0, 1)$ distribution, let the regression function $r(x) := \mathbb{E}(Y \mid X = x)$ between Y and X be in $L_2[0, 1]$, let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal basis of $L_2[0, 1]$, consider an IID sample $(Y_1, X_1), \dots, (Y_n, X_n) \sim f$, and let $(\hat{\beta}_j)_{j=0}^{\infty}$ be the corresponding empirical estimates of the regression function basis coefficients. Let $k = \lfloor n/4 \rfloor$ and define

$$\hat{\sigma}^2 := \frac{n}{k} \sum_{j=n-k+1}^{n} \hat{\beta}_j^2,$$

i.e. summing over the last k terms in $1 \leqslant j \leqslant n$, and for any integer $1 \leqslant J \leqslant n$, define

$$\hat{R}(J) := \frac{J\hat{\sigma}^2}{n} + \sum_{j=J+1}^{n} \left(\hat{\beta}_j^2 - \frac{\hat{\sigma}^2}{n} \right)_+,$$

where $s_+ := \max(s, 0)$ is the *positive part* of s. We call $\hat{\sigma}$ the *orthogonal function regression variance estimate* and call $\hat{R}$ the *estimate of the orthogonal function regression estimator risk.*

Remark 21.30 (Estimate of the orthogonal function regression estimator risk)**.** Why is $\hat{R}(J)$ in Definition 21.29 a good estimate of the MISE of the orthogonal function regression estimator $\hat{r}$ given by

$$R(r,\,\hat{r}) := \frac{J\sigma^2}{n} + \sum_{j=J+1}^{\infty} \beta_j^2$$

(see Theorem 21.28)?

To estimate the contribution from variance, namely $\sum_{j=J+1}^{\infty} \beta_j^2$, we proceed as in Remark 21.19, now using Theorem 21.25 for the mean and variance of $\hat{\beta}_j$ such that the homoscedasticity (see Remark 21.26) tells us that

$$\mathbb{E}\hat{\beta}_j^2 = \left(\mathbb{E}\hat{\beta}_j\right)^2 + \mathbb{V}\hat{\beta}_j = \beta_j^2 + \frac{\sigma^2}{n}$$

and so indeed

$$\mathbb{E}\left(\hat{\beta}_j^2 - \frac{\hat{\sigma}^2}{n}\right) \approx \beta_j^2$$

provided that $\hat{\sigma}^2$ is a consistent estimator of the variance σ^2.

So now we turn our attention to the contribution fron the variance, namely $\frac{J\sigma^2}{n}$, which also comes down to estimating the variance σ^2. To estimate the variance σ^2 we make one key observation: for a basis like the cosine basis we expect the coefficients $(\beta_j)_{j=0}^{\infty}$ of the regression function to decay rapidly as $j \to \infty$, provided r is sufficiently regular – this is discussed in Remark 21.10. For $j \geqslant n-k+1$, i.e. among the *last k modes* among $1 \leqslant j \leqslant n$, we may thus expect $\beta_j \approx 0$, from which Theorem 21.25 tells us that

$$\hat{\beta}_j \sim N\left(0,\,\frac{\sigma^2}{n}\right) \text{ for } j \geqslant n-k+1.$$

Therefore $\hat{\beta}_j \approx \frac{\sigma}{\sqrt{n}} Z_j$ for IID $N(0,1)$ random variables $Z_{n-k+1},\,\dots,\,Z_n$ and so

$$\hat{\sigma}^2 = \frac{n}{k} \sum_{j=n-k+1}^{n} \hat{\beta}_j^2 \approx \frac{\sigma^2}{k} \sum_{j=n-k+1}^{n} Z_j^2 \sim \frac{\sigma^2}{k} \chi_k^2.$$

Since $\mathbb{E}\chi_k^2 = k$ we conclude that $\mathbb{E}\hat{\sigma}^2 \approx \sigma^2$, as desired.

Note also that $\mathbb{V}\chi_k^2 = 2k$ and so $\mathbb{V}\hat{\sigma}^2 \approx \frac{\sigma^4}{k^2}(2k) = \frac{2\sigma^4}{k} \to 0$ as $n \to \infty$ since k grows like $n/4$.

Remark 21.31 (Regression via orthogonal functions in practice)**.** In practice we use the risk estimate of Definition 21.29 to select the smoothing parameter J which minimizes $\hat{R}(J)$ over $1 \leqslant J \leqslant n$.

Theorem 21.32 (Confidence bands for orthogonal function regression)**.** *Let f be the PDF of a random vector $(Y,\,X)$ whose marginal for X is a Uniform$(0,\,1)$ distribution, let the regression function $r(x) := \mathbb{E}\,(Y \mid X = x)$ between Y and X be in $L_2[0,\,1]$, let $(\phi_j)_{j=0}^{\infty}$ be an orthonormal basis of $L_2[0,\,1]$, consider an IID sample $(Y_1,\,X_1)\,,\,\dots,\,(Y_n,\,X_n) \sim f$, let $\hat{r}$ be the orthogonal function regression estimator with smoothing parameter J, and let $\hat{\sigma}^2$ be the orthogonal function regression variance estimate with $k = \lfloor n/4 \rfloor$. Assume homoscedasticity, i.e.*

$$\mathbb{V}\,(Y \mid X = x) = \sigma^2 \textit{ for all } x \in [0,\,1]$$

and assume that

$$J < n - k + 1.$$

We define

$$\tilde{a}(x) := \sum_{j=0}^{J} \phi_j^2(x) \text{ for } x \in [0,\, 1]$$

and for any $\alpha \in (0,\, 1)$ *we define*

$$c(x) := \sqrt{\frac{\tilde{a}(x)\hat{\sigma}^2\chi^2_{J,\,\alpha}}{n}}$$

and let

$$l(x) := \hat{r}(x) - c(x) \text{ and } u(x) := \hat{r}(x) + c(x).$$

Then $(l,\, u)$ *is an asymptotic* $1-\alpha$ *confidence band for the* smoothed *version*

$$r_J := \sum_{j=0}^{J} \beta_j \phi_j \text{ for } \beta_j := \langle r,\, \phi_j \rangle$$

of r*, meaning that*

$$\liminf_{n\to\infty} \mathbb{P}\left(l(x) \leqslant r_J(x) \leqslant u(x) \text{ for all } x \in [0,\, 1]\right) \geqslant 1 - \alpha.$$

Remark 21.33 (Identity for confidence bands with respect to the cosine basis)**.** When using the cosine basis the function $\tilde{a}$ used in Theorem 21.32 to construct confidence bands has a simple closed form, namely

$$\tilde{a}(x) = 1 + \frac{1}{2}\left(2J - 1 + \frac{\sin\left((2J+1)\pi x\right)}{\sin\left(\pi x\right)}\right)$$

21.4. **Wavelets.**

Definition 21.34 (Haar wavelets)**.** The function $\phi : \mathbb{R} \to \mathbb{R}$ defined by

$$\phi(x) = \begin{cases} 1 & \text{if } 0 \leqslant x < 1 \text{ and} \\ 0 & \text{otherwise} \end{cases}$$

is called the *Haar father wavelet*, or *Haar scaling function*. The function $\psi : \mathbb{R} \to \mathbb{R}$ defined by

$$\psi(x) := \begin{cases} -1 & \text{if } 0 \leqslant x \leqslant \dfrac{1}{2}, \\ 1 & \text{if } \dfrac{1}{2} < x \leqslant 1, \text{ and} \\ 0 & \text{otherwise} \end{cases}$$

is called the *Haar mother wavelet*. For $j,\, k \in \mathbb{Z}$ the functions $\psi_{j,k} : \mathbb{R} \to \mathbb{R}$ defined by

$$\psi_{j,k}(x) := 2^{j/2}\psi\left(2^j x - k\right)$$

are called *Haar children wavelets*.

Remark 21.35 (Location, scale, and normalization of the children wavelets)**.** The Haar mother wavelet is centered at $x = \frac{1}{2}$, has support $[0,\, 1]$, so we say it has *unit scale*, and its L_2 norm is

$$\int_0^1 \psi^2 = 1,$$

i.e. ψ is normal. The Haar children wavelet $\psi_{j,k}$ is centered at $x = \frac{k+1/2}{2^j}$ since

$$2^j x - k = \frac{1}{2} \iff x = \frac{k+1/2}{2^j}$$

has support $\left[\frac{k}{2^j}, \frac{k+1}{2^j}\right]$ since

$$0 \leqslant 2^j x - k \leqslant 1 \iff \frac{k}{2^j} \leqslant x \leqslant \frac{k+1}{2^j},$$

so we say that it has *scale* 2^{-j}, and the change of variables $x = 2^j - k$ tells us that its L_2 norm is

$$\int_{k/2^j}^{(k+1)/2^j} \psi_{j,k}^2(y)dy = \int_0^1 \left[2^{j/2}\psi(x)\right]^2 \frac{dx}{2^j} = \int_0^1 \psi^2 = 1,$$

i.e. $\psi_{j,k}$ is also normal. These properties are summarized in the table below.

Name	Symbol	Center	Scale	Norm
Mother	ψ	$1/2$	1	1
Child	$\psi_{j,k}$	$(k+1/2)/2^j$	2^{-j}	1

Remark 21.36 (Location of the Haar wavelets)**.** Here we follow the convention of [Was10] which centers the Haar mother wavelet at $\frac{1}{2}$. This is because we will chiefly use the Haar wavelets on the interval $[0, 1]$ and this convention makes it easier to work with that particular interval.

Theorem 21.37 (The Haar system is an orthonormal basis)**.** *Let ϕ denote the Haar father wavelet and let $(\psi_{j,k})_{j=0,\,k=0}^{\infty,\,2^j-1}$ denote Haar children wavelets. The set*

$$\{\phi\} \cup \left\{\psi_{j,k} : j \geqslant 1 \text{ and } 0 \leqslant k \leqslant 2^j - 1\right\}$$

is an orthonormal basis of $L_2[0, 1]$ called the Haar system *or* Haar basis*.*

Remark 21.38 (Children wavelets omitted from the Haar basis)**.** When constructing the Haar basis in Theorem 21.37 we omit children wavelets when $j < 0$ or when $j \geqslant 0$ but $k < 0$ or $k \geqslant 2^j$. This is because we construct the Haar basis to be a basis of $L_2[0, 1]$, not of $L_2(\mathbb{R})$. Children wavelets whose support is *not* contained in $[0, 1]$ are therefore discarded.

Corollary 21.39 (Expansion along the Haar basis)**.** *Let ϕ denote the Haar father wavelet and let $(\psi_{j,k})_{j=0,\,k=0}^{\infty,\,2^j-1}$ denote Haar children wavelets. For every $f \in L_2[0, 1]$ there exist uniquely determined $\alpha \in \mathbb{R}$ and $(\beta_{j,k})_{j=0,\,k=0}^{\infty,\,2^j-1}$ in $\mathbb{R}$ such that*

$$f = \alpha\phi + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k},$$

meaning that

$$\alpha\phi + \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k} \to f \text{ in } L_2 \text{ as } J \to \infty.$$

Moreover

$$\alpha = \int_0^1 f\phi \text{ and } \beta_{j,k} = \int_0^1 f\psi_{j,k}$$

for every $j \geqslant 1$ and $0 \leqslant k \leqslant 2^j - 1$.

Definition 21.40 (Terminology for Haar basis expansions)**.** Let ϕ denote the Haar father wavelet, let $(\psi_{j,k})_{j=0,\,k=0}^{\infty,\,2^j-1}$ denote Haar children wavelets, let $f \in L_2[0,\,1]$, and let $\alpha \in \mathbb{R}$ and $(\beta_{j,k})_{j=0,\,k=0}^{\infty,\,2^j-1}$ in $\mathbb{R}$ be as in Corollary 21.39 such that

$$f = \alpha\phi + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k}.$$

We call α the *scaling coefficient* and we call $(\beta_{j,k})_{j=0,\,k=0}^{\infty,\,2^j-1}$ the *detail coefficients*. For any integer $J \geqslant 0$ the finite sum

$$\alpha\phi + \sum_{j=0}^{J}\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k}$$

is called the *resolution* J approximation of f.

Note that the resolution J approximation is a sum of 2^J terms since $1+\sum_{j=0}^{J} 2^j = 2^J$.

Definition 21.41 (Haar wavelet regression estimator)**.** Let f be the PDF of a random vector $(Y,\,X)$ whose marginal for X is a Uniform(0, 1) distribution, let the regression function $r(x) := \mathbb{E}\,(Y \,|\, X = x)$ between Y and X be in $L_2[0,\,1]$, let ϕ denote the Haar father wavelet, let $(\psi_{j,k})_{j=0,\,k=0}^{\infty,\,2^j-1}$ denote Haar children wavelets, and let $(Y_1,\,X_1)\,,\,\dots,\,(Y_n,\,X_n) \sim f$ be IID. Let $J := \lfloor \log_2 n \rfloor$. Define

$$\hat{\alpha} := \frac{1}{n}\sum_{i=1}^{n} Y_i\phi(X_i) \text{ and } D_{j,k} := \frac{1}{n}\sum_{i=1}^{n} Y_i\psi_{j,k}(X_i)$$

for $0 \leqslant j \leqslant J-1$ and $0 \leqslant k \leqslant 2^{j-1}$,

$$\hat{\sigma} := \frac{\text{median }\left(|D_{J-1,k}| : 0 \leqslant k \leqslant 2^{J-1}-1\right)}{\Phi^{-1}\left(\frac{3}{4}\right)}$$

where Φ denotes the CDF of a standard normal distribution, and

$$\hat{\beta}_{j,k} := D_{j,k}\mathbb{1}\left(|D_{j,k}| > \hat{\sigma}\sqrt{2\log n}\right)$$

for $0 \leqslant j \leqslant J-1$ and $0 \leqslant k \leqslant 2^{j-1}$. The random function $\hat{r}$ from $[0,\,1]$ to $\mathbb{R}$ defined by

$$\hat{r} = \hat{\alpha}\phi + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\hat{\beta}_{j,k}\psi_{j,k}$$

is called the *Haar wavelet regression estimator with universal tresholding*.

Remark 21.42 (Truncation versus tresholding)**.** When we use orthogonal functions (such as the cosine basis) in Definition 21.27 to construct a regression estimator we drop *all* coefficient estimates $\hat{\beta}_j$ past a certain *truncation* treshold J known as the smoothing parameter. This parameter J is selected by first estimating the homoscedastic variance σ^2 (see Definition 21.29) and then minimizing the risk estimator $\hat{R}(J)$ with respect to J.

By contrast, when we use the Haar wavelets in Definition 21.41 to construct a regression estimator we drop coefficients $\hat{\beta}_{j,k}$ by *tresholding* instead of truncation. This treshold is determined by first estimating the homoscedastic variance σ^2 (see

Definition 21.41) and then dropping all coefficients below $\sqrt{2\hat{\sigma}^2 \log n}$ (in absolute value).

Note that when using wavelets the series is always truncated at the same J, namely $J = \lfloor \log_2 n \rfloor$ to ensure that $n \approx 2^J$, where 2^J is the number of parameters α and $\beta_{j,k}$ to fit *before* tresholding.

Remark 21.43 (Robustness of the median absolute deviation)**.** Why do we use the *sample median*

$$\operatorname{median}\left(|X_1|, \ldots, |X_n|\right)$$

to estimate the variance in Definition 21.41 instead of, say the usual sample variance. This is because the former is more *robust*, meaning that it is less sensitive to outliers, as demonstrated numerically in Exercise A.21.8.

More generally, note that the median of $|X|$ is also knows as the median absolute deviation (since here X has median zero). So Exercise A.21.8 tells us this: the median absolute deviation is a more robust measure of spread than the standard deviation (or, equivalently, the variance). This is why Definition 21.41 uses the *sample median* of $|X|$, which essentially coincides with the *sample median absolute deviation* of X, since X has median zero, instead of the sample variance.

Remark 21.44 (Empirical choice of the resolution)**.** In practice choosing $J \approx \log_2 n$ as suggested in Definition 21.41 can lead to estimators with very high variance. Choosing a coarser resolution, i.e. a smaller value of J, can help in such cases.

22. Classification

22.1. Introduction.

Remark 22.1 (The many names of classification)**.** The problem of predicting a discrete random variable Y with finite codomain from another random variable X is called

- *classification*,
- *supervised learning*,
- *discrimination*, or
- *pattern recognition*.

Definition 22.2 (Classification rule)**.** Let Y be a random variable with finite codomain $\mathcal{Y}$ and let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$. A map from $\mathcal{X}$ to $\mathcal{Y}$ is called a *classification rule for* Y *given* X or a *classifier.*

Definition 22.3 (Classification estimator)**.** Let Y be a random variable with finite codomain $\mathcal{Y}$ and let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$. A random function from $\mathcal{X}$ to $\mathcal{Y}$ is called a *classification estimator*

Remark 22.4 (Classification estimator)**.** In practice the randomness in a classification estimator comes from the fact that it is constructed from an IID sample $(Y_1, X_1), \ldots, (Y_n, X_n)$.

22.2. Error Rates and the Bayes Classifier.

Definition 22.5 (Error rates)**.** Let Y be a random variable with finite codomain $\mathcal{Y}$ and let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$. For any classification rule h for Y given X we define the following.

(1) The *true error rate* of h is defined to be

$$L(h) := \mathbb{P}\left(h(X) \neq Y\right).$$

(2) The *empirical error rate*, or *training error rate*, is defined to be

$$\hat{L}_n(h) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(h(X_i) \neq Y_i\right).$$

Remark 22.6 (Error rates with respect to other loss functions)**.** The true error rate L implicitly relies on the zero-one loss L_0 since

$$L(h) = \mathbb{P}\left(h(X) \neq Y\right) = \mathbb{E}\left[\mathbb{1}\left(h(X) \neq Y\right)\right] = \mathbb{E}\left[L_0\left(h(X), Y\right)\right].$$

We may therefore define an error rate $\widetilde{L}$ with respect to *any* loss function L via

$$\widetilde{L}(h) := \mathbb{E}\left[L\left(h(X), Y\right)\right].$$

Definition 22.7 (Decision boundary)**.** Let Y be a random variable with finite codomain $\mathcal{Y} = \{0, 1\}$, i.e. a binary random variable, let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$, and let $h : \mathcal{X} \to \{0, 1\} = \mathcal{Y}$ be a classification rule. If the (topological) boundaries of the sets $\{h = 0\}$ and $\{h = 1\}$ exist and agree with each other then we call

$$\mathcal{D}(h) := \partial\left\{h = 0\right\} = \partial\left\{h = 1\right\}$$

the *decision boundary* of h.

Definition 22.8 (Bayes classification rule)**.** Let Y be a binary random variable, let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$, and let $r(x) := \mathbb{E}\,(Y \,|\, X = x)$ be the regression function between Y and X. The classification rule h^* defined by

$$h^* = \mathbb{1}\left(r > \frac{1}{2}\right)$$

is called the *Bayes classification rule* or the *Bayes classifier.*

Remark 22.9 (Decision boundary of the Bayes classification rule)**.** Let Y be a binary random variable, let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$, and let $r(x) := \mathbb{E}\,(Y \,|\, X = x)$ be the regression function between Y and X. If the regression function is sufficiently regular then the decision boundary of the Bayes classification rule h^* is

$$\mathcal{D}(h^*) = \partial\,\{h^* = 1\} = \partial\left\{r > \frac{1}{2}\right\} = \left\{r = \frac{1}{2}\right\}.$$

In particular, since

$$r(x) = \mathbb{E}\,(Y \,|\, X = x) = \mathbb{P}\,(Y = 1 \,|\, X = x)$$

for $\mathcal{Y} = \{0,\, 1\}$ we observe that

$$\begin{aligned} r(x) = \frac{1}{2} &\iff \mathbb{P}\,(Y = 1 \,|\, X = x) = \frac{1}{2} \\ &\iff \mathbb{P}\,(Y = 1 \,|\, X = x) = \mathbb{P}\,(Y = 0 \,|\, X = x) = \frac{1}{2}. \end{aligned}$$

In other words the decision boundary of the Bayes classification rule is

$$\mathcal{D}(h^*) = \{x \in \mathcal{X} : \mathbb{P}\,(Y = 1 \,|\, X = x) = \mathbb{P}\,(Y = 0 \,|\, X = x)\}\,.$$

This is the set of points where, even with perfect information (i.e. even knowing the joint distribution of (Y, X)), with the value of X in hand we cannot determine which of $Y = 0$ and $Y = 1$ is more likely to occur.

Remark 22.10 (Alternate formulation of the Bayes classification rule)**.** Proceeding as in Remark 22.9 we observe that

$$r(x) > \frac{1}{2} \iff \mathbb{P}\,(Y = 1 \,|\, X = x) > \mathbb{P}\,(Y = 0 \,|\, X = x)$$

and so we may write the Bayes classification rule as

$$h^*(x) = \begin{cases} 1 & \text{if } \mathbb{P}\,(Y = 1 \,|\, X = x) > \mathbb{P}\,(Y = 0 \,|\, X = x) \text{ and} \\ 0 & \text{if } \mathbb{P}\,(Y = 0 \,|\, X = x) \geqslant \mathbb{P}\,(Y = 1 \,|\, X = x)\,. \end{cases}$$

Theorem 22.11 (The Bayes classification rule is optimal)**.** *Let Y be a binary random variable and let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$. The Bayes classification rule h^* is optimal with respect to the true error rate in the sense that, for every classification rule h for Y given X,*

$$L(h^*) \leqslant L(h).$$

Remark 22.12 (The regression function of a binary random variable)**.** If Y is a binary random variable then its regression function with respect to any random variable X may be written as follows using Bayes' Theorem:

$$\begin{aligned} r(x) &= \mathbb{E}\left(Y \,|\, X = x\right) \\ &= \mathbb{P}\left(Y = 1 \,|\, X = x\right) \\ &= \frac{f\left(x \,|\, Y = 1\right)\mathbb{P}\left(Y = 1\right)}{f\left(x \,|\, Y = 1\right)\mathbb{P}\left(Y = 1\right) + f\left(x \,|\, Y = 0\right)\mathbb{P}\left(Y = 0\right)} \\ &= \frac{\pi f_1(x)}{\pi f_1(x) + (1 - \pi) f_0(x)} \end{aligned}$$

for

$$\pi := \mathbb{P}\left(Y = 1\right)$$

characterizing the distribution of the binary response Y,

$$f_1(x) := f\left(x \,|\, Y = 1\right)$$

the conditional PDF of the distribution of the covariate X given $Y = 1$, which is sometimes viewed as the distribution of X among the *class* $Y = 1$, and

$$f_0(x) := f\left(x \,|\, Y = 0\right)$$

the conditional PDF of the distribution of the covariate X given $Y = 0$, which is sometimes viewed as the distribution of X among the *class* $Y = 0$. (Note that this latter perspective as *classes* gets muddier if the conditional distribution of Y given X is *not* deterministic.)

Remark 22.13 (Approaches to classification)**.** For a binary response random variable Y and any covariate random vector X the Bayes classification rule is optimal – see Theorem 22.11. However the Bayes classification rule *cannot* be determined solely from IID samples

$$\left(Y_1, X_1\right), \ldots, \left(Y_n, X_n\right),$$

so it must be *approximated*. Broadly speaking there are three approaches to approximating the Bayes classification rule.

- **Empirical risk minimization**. We choose a set of classifiers $\mathcal{H}$ and find $\hat{h} \in \mathcal{H}$ which minimizes some estimate of the true error rate over $\mathcal{H}$.
- **Regression**. We find an estimate $\hat{r}$ of the regression function and define the corresponding classification estimator

  $$\hat{h} := \mathbb{1}\left(\hat{r} > \frac{1}{2}\right),$$

  i.e. proceeding as in Definition 22.8 where we introduced the Bayes classifier.
- **Density estimation**. For π, f_0, and f_1 as in Remark 22.12 we estimate f_0 from the samples X_i for which $Y_i = 0$, estimate f_1 from the samples X_i for which $Y_i = 1$, and estimate π via $\hat{\pi} := \frac{1}{n}\sum_{i=1}^{n} Y_i$. We may then construct a regression estimator using the functional form of Remark 22.12 as

  $$\hat{r} := \frac{\hat{\pi}\hat{f}_1}{\hat{\pi}\hat{f}_1 + (1 - \hat{\pi})\hat{f}_0}$$

  and define the corresponding classification estimator as in the *regression* approach, namely

  $$\hat{h} := \mathbb{1}\left(\hat{r} > \frac{1}{2}\right).$$

Theorem 22.14 (Optimal classifier for non-binary responses)**.** *Consider a random variable Y with finite codomain $\mathcal{Y}$ and let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$. Define, for $x \in \mathcal{X}$,*

$$h^*(x) := \underset{y \in \mathcal{Y}}{\arg\max}\, \mathbb{P}\left(Y = y \,|\, X = x\right) = \underset{y \in \mathcal{Y}}{\arg\max}\, \pi_y f_y(x)$$

where

$$\pi_y := \mathbb{P}\left(Y = y\right) \text{ and } f_y(x) := f\left(x \,|\, Y = y\right).$$

The classification rule h^* is optimal with respect to the true error rate in the sense that

$$L(h^*) \leqslant L(h)$$

for any classification rule h for Y given X.

Remark 22.15 (Optimal classifier for non-binary responses)**.** The two formulations of the optimal classifier for non-binary random variables recorded in Theorem 22.14 are indeed equivalent since Bayes' Theorem tells us that

$$\mathbb{P}\left(Y = y \,|\, X = x\right) = \frac{f\left(x \,|\, Y = y\right) \mathbb{P}\left(Y = y\right)}{\sum_{\tilde{y} \in \mathcal{Y}} f\left(x \,|\, Y = \tilde{y}\right) \mathbb{P}\left(Y = \tilde{y}\right)} = \frac{\pi_y f_y(x)}{\sum_{\tilde{y} \in \mathcal{Y}} \pi_{\tilde{y}} f_{\tilde{y}}(x)}.$$

The denominator of the right-hand side does not depend on y and so indeed

$$\underset{y \in \mathcal{Y}}{\arg\max}\, \mathbb{P}\left(Y = y \,|\, X = x\right) = \underset{y \in \mathcal{Y}}{\arg\max}\, \pi_y f_y(x).$$

22.3. **Gaussian and Linear Classifiers.**

As discussed in Remark 22.13, there are several ways to construct classification estimators. In this section we discuss some estimators constructed by following the *density estimation* approach.

Theorem 22.16 (Gaussian Bayes classifier)**.** *Let Y be a binary random variable and let X be a random vector in $\mathbb{R}^d$. Suppose that*

$$X \,|\, Y = 0 \sim N\left(\mu_0, \Sigma_0\right) \text{ and } X \,|\, Y = 1 \sim N\left(\mu_1, \Sigma_1\right).$$

Then the Bayes classifier for Y given X is

$$h^*(x) = \mathbb{1}\left(r_1^2(x) < r_0^2(x) + 2\log\frac{\pi_1}{\pi_0} + \log\frac{|\Sigma_0|}{|\Sigma_1|}\right)$$

for $x \in \mathbb{R}$, where $|A| := \det A$ and, for $i = 0, 1$,

$$\pi_i := \mathbb{P}\left(Y = i\right) \text{ and } r_i^2(x) := \Sigma_i^{-1}\left(x - \mu_i\right) \cdot \left(x - \mu_i\right).$$

r_i is called the Mahalanobis distance*. Equivalently*

$$h^*(x) = \underset{i=0,1}{\arg\max}\, \delta_i(x) = \mathbb{1}\left(\delta_1(x) > \delta_0(x)\right)$$

where, for $i = 0, 1$,

$$\begin{aligned} \delta_i(x) &:= -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}\Sigma_i^{-1}\left(x - \mu_i\right) \cdot \left(x - \mu_i\right) + \log\pi_i \\ &= -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}r_i^2(x) + \log\pi_i \end{aligned}$$

are called (quadratic) discriminant function*.*

Remark 22.17 (Interpretation of the Mahalanobis distance)**.** In one dimension the Mahalanobis distance is

$$r(x) = \frac{|x-\mu|}{\sigma}.$$

It is therefore a *dimensionless* distance which measures how many standard deviations a point x is from the mean.

Definition 22.18 (Quadratic discriminant analysis)**.** Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample from (Y, X) where Y is a binary random variable and X is a random vector in $\mathbb{R}^d$. We define, for $x \in \mathbb{R}^d$,

$$\begin{aligned}
&\hat{\pi}_1 := \frac{1}{n}\sum_{i=1}^{n} Y_i, && \hat{\pi}_0 := \frac{1}{n}\sum_{i=1}^{n}(1-Y_i) = 1-\hat{\pi}_1,\\
&n_1 := \sum_{i=1}^{n} Y_i, && n_0 := \sum_{i=1}^{n}(1-Y_i) = n - n_0,\\
&\hat{\mu}_1 := \frac{1}{n_1}\sum_{\substack{i=1\\ Y_i=1}}^{n} X_i, && \hat{\mu}_0 := \frac{1}{n_0}\sum_{\substack{i=1\\ Y_i=0}}^{n} X_i,\\
&\hat{S}_1 := \frac{1}{n_1}\sum_{\substack{i=1\\ Y_i=1}}^{n} (X_i-\hat{\mu}_1)\otimes(X_i-\hat{\mu}_1), && \hat{S}_0 := \frac{1}{n_0}\sum_{\substack{i=1\\ Y_i=0}}^{n} (X_i-\hat{\mu}_0)\otimes(X_i-\hat{\mu}_0),\\
&\hat{r}_1(x) := \hat{S}_1^{-1}(x-\hat{\mu}_1)\cdot(x-\hat{\mu}_1), && \hat{r}_0(x) := \hat{S}_0^{-1}(x-\hat{\mu}_0)\cdot(x-\hat{\mu}_0),\\
&\hat{\delta}_1(x) := \log\hat{\pi}_1 - \frac{1}{2}\log|\hat{S}_1| - \frac{1}{2}\hat{r}_1(x), \text{ and} && \hat{\delta}_0(x) := \log\hat{\pi}_0 - \frac{1}{2}\log|\hat{S}_0| - \frac{1}{2}\hat{r}_0(x),
\end{aligned}$$

where $|A| := \det A$. The classification estimator

$$\hat{h}(x) := \mathbb{1}\left(\hat{\delta}_1(x) > \hat{\delta}_0(x)\right)$$

is called the *quadratic discriminant analysis*, or *QDA*, *classification estimator.*

Remark 22.19 (Motivation for quadratic discriminant analysis)**.** Let Y be a binary random variable and let X be a random vector in $\mathbb{R}^d$. If $X \mid Y$ is Normal, i.e.

$$X \mid Y = 0 \sim N(\mu_0, \Sigma_0) \text{ and } X \mid Y = 1 \sim N(\mu_1, \Sigma_1)$$

then the QDA classification estimator may be derived by the *density estimation* approach of Remark 22.13.

Indeed, under the Gaussian assumption Theorem 22.16 tells us that the (optimal) Bayes classifier is

$$h^*(x) = \mathbb{1}(\delta_1(x) > \delta_0(x))$$

where δ_0 and δ_1 denote the quadratic discriminant functions. The QDA classification estimator is thus obtained by estimating δ_0 and δ_1, which in turn is done by estimating the parameters that characterize them, namely

$$\pi_0,\ \pi_1,\ \mu_0,\ \mu_1,\ \Sigma_0, \text{ and } \Sigma_1.$$

As seen in Exercise A.22.2,

$$\delta_i = \log(\pi_i f_i) + C,$$

where $f_i := f(\,\cdot \mid Y = i)$, and so estimating δ_0 and δ_1 does indeed come down to a *density estimation* problem. Here this problem is resolved *parametrically* by estimating the parameters μ and Σ of $f(\,\cdot \mid Y = i) \sim N(\mu, \Sigma)$.

Theorem 22.20 (Linear Bayes classifier). *Let Y be a binary random variable and let X be a random vector in $\mathbb{R}^d$. Suppose that*

$$X \,|\, Y = 0 \sim N\left(\mu_0,\, \Sigma\right) \text{ and } X \,|\, Y = 1 \sim N\left(\mu_1,\, \Sigma\right),$$

i.e. $\mathbb{V}\left(X \,|\, Y = 0\right) = \mathbb{V}\left(X \,|\, Y = 1\right)$. *Then the Bayes classifier for Y given X is*

$$h^*(x) := \mathbb{1}\left(\delta_1(x) > \delta_0(x)\right) = \operatorname*{arg\,max}_{i=0,1} \delta_i(x)$$

for $x \in \mathbb{R}$ where, for $i = 0, 1$,

$$\delta_i(x) := \Sigma^{-1}\mu_i \cdot x - \frac{1}{2}\Sigma^{-1}\mu_i \cdot \mu_i + \log \pi_i$$

for $\pi := \mathbb{P}\left(Y = i\right)$. The functions δ_0 and δ_1 are called (linear) discriminant functions.

Definition 22.21 (Linear discriminant analysis). Let $\left(Y_1,\, X_1\right), \dots, \left(Y_n,\, X_n\right)$ be an IID sample from $(Y,\, X)$ where Y is a binary random variable and X is a random vector in $\mathbb{R}^d$. We define

$$\hat{\pi}_0,\, \hat{\pi}_1,\, n_0,\, n_1,\, \hat{\mu}_0,\, \hat{\mu}_1,\, \hat{S}_0,\, \hat{S}_1,$$

as in Definition 22.18 and we define

$$\hat{S} := \frac{n_0 \hat{S}_0 + n_1 \hat{S}_1}{n_0 + n_1}$$

as well as, for $x \in \mathbb{R}^d$,

$$\begin{aligned}\hat{\delta}_0(x) &:= \hat{S}^{-1}\hat{\mu}_0 \cdot x - \frac{1}{2}\hat{S}^{-1}\hat{\mu}_0 \cdot \hat{\mu}_0 + \log \hat{\pi}_0 \text{ and} \\ \hat{\delta}_1(x) &:= \hat{S}^{-1}\hat{\mu}_1 \cdot x - \frac{1}{2}\hat{S}^{-1}\hat{\mu}_1 \cdot \hat{\mu}_1 + \log \hat{\pi}_1.\end{aligned}$$

The classification estimator

$$\hat{h}(x) := \mathbb{1}\left(\hat{\delta}_1(x) > \hat{\delta}_0(x)\right)$$

is called the *linear discriminant analysis*, or *LDA*, *classification estimator*.

Remark 22.22 (Motivation for linear discriminant analysis). As discussed in Remark 22.19, the QDA classification estimator is derived from the *Gaussian* Bayes classifier (recorded in Theorem 22.16 when $X \,|\, Y = y$ is Normal for $y = 0, 1$) by *density estimation* (see also Remark 22.13).

In exactly the same way the LDA classification estimator is derived from the *linear* Bayes classifier (recorded in Theorem 22.20 when $X \,|\, Y = y$ is Normal for $y = 0, 1$) *and* $\mathbb{V}\left(X \,|\, Y = 0\right) = \mathbb{V}\left(X \,|\, Y = 1\right)$) by density estimation.

In particular, as shown in Exercise A.23.35 the assumption that

$$\mathbb{V}\left(X \,|\, Y = 0\right) = \mathbb{V}\left(X \,|\, Y = 1\right)$$

means that the discriminant functions (see Theorem 22.16 and Theorem 22.20) are now *linear*.

Theorem 22.23 (Gaussian and linear Bayes classifiers for non-binary responses). *Let Y be a random variable with finite codomain $\{0, \dots, K-1\}$ and let X be a random vector in $\mathbb{R}^d$. Suppose that*

$$X \,|\, Y = k \sim N\left(\mu_k,\, \Sigma_k\right) \text{ for every } 0 \leqslant k \leqslant K-1.$$

Then the Bayes classifier for Y given X is

$$h^*(x) = \operatorname*{arg\,max}_{0 \leqslant k \leqslant K-1} \delta_k(x)$$

for the quadratic discriminant functions

$$\begin{aligned}\delta_k(x) &:= -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}\Sigma_k^{-1}(x-\mu_k)\cdot(x-\mu_k) + \log\pi_k \\ &= -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}r_k^2(x) + \log\pi_k\end{aligned}$$

for $x \in \mathbb{R}$ where r_k is the Mahalanobis distance and $\pi_k := \mathbb{P}(Y = k)$. Suppose moreover that

$$\Sigma_0 = \Sigma_1 = \cdots = \Sigma_{K-1} =: \Sigma.$$

Then the Bayes classifier for Y given X may also be written as

$$h^*(x) = \operatorname*{arg\,max}_{0 \leqslant k \leqslant K-1} \bar{\delta}_k(x)$$

for the linear discriminant functions

$$\bar{\delta}_k(x) := \Sigma^{-1}\mu_k \cdot x - \frac{1}{2}\Sigma^{-1}\mu_k\cdot\mu_k + \log\pi_k$$

for $x \in \mathbb{R}$.

Remark 22.24 (Quadratic and linear discriminant analysis for non-binary responses)**.** Proceeding as in Definitions 22.18 and 22.21, which leverage Theorems 22.16 and 22.20, respectively, we may leverage Theorem 22.23 to carry out either quadratic or linear discriminant analysis on a sample

$$(Y_1, X_1), \ldots, (Y_n, X_n) \sim (Y, X)$$

where now Y is *not* binary and instead has codomain $\{0, \ldots, K-1\}$.

This is done by estimating π_k, μ_k, and Σ_k for $0 \leqslant k \leqslant K-1$ and then plugging these estimates into the functional forms of the quadratic and linear discriminant functions of Theorem 22.23 to construct a classification estimator (as is done in Definitions 22.18 a 22.21).

Definition 22.25 (Between-class and within-class variance)**.** Let Y be random variable with finite codomain $\{0, \ldots, K-1\}$ and let X be a random vector in $\mathbb{R}^d$. We call

- $\mathbb{V}[\mathbb{E}(X \mid Y)]$ the *between-class variance* of X and
- $\mathbb{E}[\mathbb{V}(X \mid Y)]$ the *within-class variance* of X.

Remark 22.26 (Between-class and within-class variance)**.** By the Law of Total Variance the between-class variance and the within-class variance add up to the variance. The usefulness of this decomposition is discussed in Remark 3.18.

Definition 22.27 (Fisher discriminant ratio)**.** Let Y be a random variable with finite codomain $\{0, \ldots, K-1\}$ and let X be random vector in $\mathbb{R}^d$. The ratio

$$\frac{\mathbb{V}[\mathbb{E}(X \mid Y)]}{\mathbb{E}[\mathbb{V}(X \mid Y)]}$$

of the between-class variance over the within-class variance is called the *Fisher discriminant ratio* of X given Y.

Definition 22.28 (Fisher's linear discriminant function)**.** Let Y be a random variable with finite codomain $\{0, \ldots, K-1\}$ and let X be a random vector in $\mathbb{R}^d$. We define

$$w := \underset{w \in \mathbb{R}^d}{\arg\max} \frac{\mathbb{V}\left[\mathbb{E}\left(X \,|\, Y\right)\right] v \cdot v}{\mathbb{E}\left[\mathbb{V}\left(X \,|\, Y\right)\right] v \cdot v}$$

and call the function $u : \mathbb{R}^d \to \mathbb{R}$ defined by

$$u(x) := w \cdot x$$

Fisher's linear discriminant function.

Remark 22.29 (Fisher's linear discriminant function and generalized Rayleigh quotients)**.** The ratio characterizing Fisher's linear discriminant function is known as a *generalized Rayleigh quotient* since both the between-class variance $\mathbb{V}\left[\mathbb{E}\left(X \,|\, Y\right)\right]$ and the within-class variance $\mathbb{E}\left[\mathbb{V}\left(X \,|\, Y\right)\right]$ are symmetric positive-definite matrices.

The maximization of this ratio is then equivalent to a generalized eigenvalue problem, which in turn is equivalent to both the maximization of a standard Rayleigh quotient and to a standard eigenvalue problem. This is proved in Theorem C.10.

Remark 22.30 (Motivating Fisher's linear discriminant function)**.** Fisher asks the following question: which *linear* functional allows us to best discriminate between classes? To answer this we must first *quantify* discrimination.

This quantification is done using the Fisher discriminant ratio. Since linear functionals are necessarily of the form $x \mapsto v \cdot x$ for some v we are thus looking for $v \in \mathbb{R}^d$ which maximizes the Fisher discriminant ratio of $v \cdot X$, namely

$$\frac{\mathbb{V}\left[\mathbb{E}\left(v \cdot X \,|\, Y\right)\right]}{\mathbb{E}\left[\mathbb{V}\left(v \cdot X \,|\, Y\right)\right]}.$$

In light of Theorem 14.1 we may write this ratio as

$$\frac{\mathbb{V}\left[\mathbb{E}\left(v \cdot X \,|\, Y\right)\right]}{\mathbb{E}\left[\mathbb{V}\left(v \cdot X \,|\, Y\right)\right]} = \frac{\mathbb{V}\left[v \cdot \mathbb{E}\left(X \,|\, Y\right)\right]}{\mathbb{E}\left[\mathbb{V}\left(X \,|\, Y\right) v \cdot v\right]} = \frac{\mathbb{V}\left[\mathbb{E}\left(X \,|\, Y\right)\right] v \cdot v}{\mathbb{E}\left[\mathbb{V}\left(X \,|\, Y\right)\right] v \cdot v},$$

which is precisely the ratio maximized to define Fisher's linear discriminant function.

Definition 22.31 (Fisher's linear classifier)**.** Let Y be a binary random variable, let X be a random vector, and let us denote

$$\mu_0 := \mathbb{E}\left(X \,|\, Y = 0\right) \text{ and } \mu_1 := \mathbb{E}\left(X \,|\, Y = 1\right).$$

Suppose that

$$\mathbb{P}\left(Y = 0\right) = \mathbb{P}\left(Y = 1\right) = \frac{1}{2} \text{ and } \mathbb{V}\left(X \,|\, Y = 0\right) = \mathbb{V}\left(X \,|\, Y = 1\right) =: \Sigma.$$

Let $u(x) := w \cdot x$ denote Fisher's linear discriminant function where we specifically choose

$$w := \Sigma^{-1}(\mu_1 - \mu_0)$$

and let

$$m := \frac{\mu_0 + \mu_1}{2}.$$

The classifier

$$h := \mathbb{1}\left(u > w \cdot m\right)$$

is called *Fisher's linear classifier.*

Remark 22.32 (Fisher's linear classifier)**.** Exercise A.23.38 shows that the choice of w in Definition 22.31 is indeed a minimizer of the Fisher discriminant ratio, such that $x \mapsto w \cdot x$ is indeed Fisher's linear discriminant function.

Remark 22.33 (Another path to Fisher's linear discriminant function)**.** Here we discuss another way to motivate the specific discriminant function used in Fisher's linear classifier, namely

$$u(x) = \Sigma^{-1}(\mu_1 - \mu_0) \cdot x.$$

We may start by asking the following question: which linear functional best discriminates along the $\mu_1 - \mu_0$ direction in X–space with respect to the intrinsic geometry of X–space? For now "X–space" just means the codomain of X. Even if X is not Normal, Mahalanobis suggests that we use the variance of X to describe X–space by using the *Mahalanobis inner product*

$$\langle x_1,\, x_2\rangle := \Sigma^{-1}x_1 \cdot x_2 \text{ for } x_1,\, x_2 \in \mathbb{R}^d,$$

whose corresponding norm is precisely the squred Mahalanobis distance. "X–space" now more precisely refers to the inner product space $\big(\mathbb{R}^d,\, \langle\,\cdot\,,\,\cdot\,\rangle\big)$.

To discriminate along $\mu_1 - \mu_0$ in a linear fashion while using the Mahalanobis inner product we therefore naturally use

$$u(x) := \langle \mu_1 - \mu_0,\, x\rangle = \Sigma^{-1}(\mu_1 - \mu_0) \cdot x,$$

which is precisely the discriminant function used in Fisher's linear classifier. The cutoff $w \cdot m$ is then simply chosen to be the midpoint between $u(\mu_0)$ and $u(\mu_1)$ since

$$\frac{u(\mu_0) + u(\mu_1)}{2} = w \cdot \frac{\mu_0 + \mu_1}{2} = w \cdot m.$$

Finally note that, since Σ^{-1} is positive-definite,

$$u(\mu_1) - u(\mu_0) = w \cdot (\mu_1 - \mu_0) = \Sigma^{-1}(\mu_1 - \mu_0) \cdot (\mu_1 - \mu_0) > 0,$$

meaning that $u(\mu_1) > u(\mu_0)$, and so we choose the class $Y = 1$ when u is *above* the cutoff $w \cdot m$, i.e.

$$h = \mathbb{1}(u > w \cdot m).$$

Definition 22.34 (Fisher's linear classification estimator)**.** Let $(Y_1,\, X_1)\,,\, \ldots,\, (Y_n,\, X_n)$ be an IID sample from $(Y,\, X)$ where Y is a binary random variable and X is a random vector in $\mathbb{R}^d$. We define

$$\hat{\mu}_0,\; \hat{\mu}_1,\; \text{and } \hat{S}$$

as in Definition 22.21,

$$\hat{w} := \hat{S}^{-1}(\hat{\mu}_0 - \hat{\mu}_1)\,,\; \hat{m} := \frac{\hat{\mu}_0 + \hat{\mu}_1}{2},$$

and, for $x \in \mathbb{R}^d$,

$$\hat{u}(x) := \hat{S}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) \cdot x.$$

The classification estimator

$$\hat{h}(x) := \mathbb{1}\left(\hat{u}(x) > \hat{w} \cdot \hat{m}\right)$$

is called *Fisher's linear classification estimator*.

Remark 22.35 (Fisher's linear discriminant classification estimator and LDA)**.** As shown in Exercise A.23.39, Fisher's linear classification estimator agrees with the LDA classificatin estimator when $\hat{\pi}_0 = \hat{\pi}_1 = \frac{1}{2}$.

Remark 22.36 (Fisher's linear classification estimator for more than two classes)**.** To construct classification estimators similar to Fisher's linear classification estimator when there are more than two classes we go back to the Fisher discriminant ratio of $v \cdot X$, namely

$$\frac{\mathbb{VE}\left(X \,|\, Y\right) v \cdot v}{\mathbb{EV}\left(X \,|\, Y\right) v \cdot v} =: R(v)$$

(as used in Definition 22.28). Since maximizing this generalized Rayleigh quotient is equivalent to an eigenvalue problem for a symmetric matrix (see Remark 22.29 and Theorem C.10) we can extract an orthonormal basis $v_1, \ldots, v_d$ of $\mathbb{R}^d$ from this ratio via

$$\begin{aligned} v_1 &:= \arg\max_{\mathbb{R}^d} R, \\ v_2 &:= \arg\max_{\mathbb{R}^d \setminus \mathrm{span}(v_1)} R, \\ &\vdots \\ v_j &:= \arg\max_{\mathbb{R}^d \setminus \mathrm{span}(v_1,\, \ldots,\, v_{j-1})} R, \\ &\vdots \\ v_d &:= \arg\max_{\mathbb{R}^d \setminus \mathrm{span}(v_1,\, \ldots,\, v_{d-1})} R. \end{aligned}$$

This basis is known as a basis of *discriminant coordinates* (see [HTF09]).

We may then use up to d–many linear discriminant functions

$$u_j(x) := v_j \cdot x$$

to distinguish the K classes. In particular, estimating the ratio R from data comes down to estimating the between-class variance $\mathbb{VE}\left(X \,|\, Y\right)$ and the within-class variance $\mathbb{EV}\left(X \,|\, Y\right)$, which may be done for example by first estimating the parameters

$$\pi_k := \mathbb{P}\left(Y = k\right), \; \mu_k := \mathbb{E}\left(X \,|\, Y = k\right), \text{ and } \Sigma_k := \mathbb{V}\left(X \,|\, Y = k\right)$$

and then using the identities of Exercise A.23.36.

22.4. **Linear and Logistic Regression.** In the terminology of Remark 22.13 we spent Section 22.3 discussing classification estimators constructed via the *density estimation* approach. In this section we discuss some classification estimators constructed via the *regression* approach.

Definition 22.37 (Linear regression classification estimator)**.** Consider an IID sample $\left(Y_1,\, X_1\right), \ldots, \left(Y_n,\, X_n\right)$ from $\left(Y,\, X\right)$ where Y is a binary random variable and X is a random vector in $\mathbb{R}^d$, let $\widetilde{X}_i := \left(1,\, X_i\right) \in \mathbb{R}^{d+1}$ for $1 \leqslant i \leqslant n$, and let $\hat{\beta} \in \mathbb{R}^{d+1}$ be the least squares estimate for the sample

$$\left(Y_1,\, \widetilde{X}_1\right), \ldots, \left(Y_n,\, \widetilde{X}_n\right).$$

We define

$$\hat{r}(x) := \hat{\beta}_0 + \sum_{j=1}^{d} \hat{\beta}_j x_j \text{ for } x \in \mathbb{R}^d.$$

The classification estimator

$$\hat{h}(x) := \mathbb{1}\left(\hat{r}(x) > \frac{1}{2}\right)$$

is called the *linear regression classification estimator.*

Remark 22.38 (Implementing the linear regression classification estimator)**.** Recall from Definition 13.39 and Theorem 13.41 that the least squares estimate $\hat{\beta}$ satisfies

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{d+1}} |\mathbb{Y} - \mathbb{X}\beta|^2$$

or equivalently

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y},$$

where $\mathbb{X}$ is the design matrix

$$\mathbb{X}_{ij} = \left(\widetilde{X}_i\right)_j = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \in \mathbb{R}^{n\times(d+1)}$$

and $\mathbb{Y} = (Y_1, \dots, Y_n)$ is the response vector.

Definition 22.39 (Logistic regression classification estimator)**.** Consider an IID sample $(Y_1, X_1), \dots, (Y_n, X_n)$ from (Y, X) where Y is a binary random variable and X is a random vector in $\mathbb{R}^d$ such that, for $\widetilde{X}_i := (1, X_i) \in \mathbb{R}^{d+1}$ for $1 \leqslant i \leqslant n$,

$$\left(Y_1, \widetilde{X}_1\right), \dots, \left(Y_n, \widetilde{X}_n\right).$$

is drawn from a distribution in the logistic regression model whose parameter is $\beta \in \mathbb{R}^{d+1}$, and let $\hat{\beta}$ denote the MLE of β constructed as in Theorem 13.86 (see also Remark 13.87). We define

$$\hat{r}(x) := \sigma\left(\hat{\beta}_0 + \sum_{j=1}^{d} \hat{\beta}_j x_j\right) \quad \text{for } x \in \mathbb{R}^d$$

where σ is the logistic function. The classification estimator

$$\hat{h}(x) := \mathbb{1}\left(\hat{r}(x) > \frac{1}{2}\right)$$

is called the *logistic regression classification estimator.*

22.5. Relationship Between Logistic Regression and LDA.

Remark 22.40 (Relationship between logistic regression and LDA)**.** Let Y be a binary random variable and let X be a random vector in $\mathbb{R}^d$. On the one hand, if $\left(Y, \widetilde{X}\right)$ has a logistic regression distribution for some $(\beta_0, \beta) \in \mathbb{R}^{1+d}$, where $\widetilde{X} = (1, X)$, then *by definition*

$$\operatorname{logit} \mathbb{P}(Y = 1 \mid X = x) = \beta_0 + \beta \cdot x.$$

On the other hand, if

$$X \mid Y = 0 \sim N(\mu_0, \Sigma) \text{ and } X \mid Y = 1 \sim N(\mu_1, \Sigma),$$

i.e. $\mathbb{V}(X \mid Y=0) = \mathbb{V}(X \mid Y=1)$ as in Theorem 22.20, then Exercises A.22.2 and A.23.35 tell us that

$$\begin{aligned}
\operatorname{logit} \mathbb{P}(Y=1 \mid X=x) &= \log \frac{\mathbb{P}(Y=1 \mid X=x)}{1-\mathbb{P}(Y=1 \mid X=x)} \\
&= \log \frac{\mathbb{P}(Y=1 \mid X=x)}{\mathbb{P}(Y=0 \mid X=x)} \\
&= \log \frac{\pi_1 f(x \mid Y=1)}{\pi_0 f(x \mid Y=0)} \\
&= \delta_1(x) - \delta_0(x) \\
&= \Sigma^{-1}(\mu_1 - \mu_0) \cdot x + \frac{1}{2}\Sigma^{-1}\mu_0 \cdot \mu_0 - \frac{1}{2}\Sigma^{-1}\mu_1 \cdot \mu_1 + \log \frac{\pi_1}{\pi_0} \\
&=: \alpha \cdot x + \alpha_0 .
\end{aligned}$$

In other words the logistic regression classification estimator and the LDA classification estimator both lead to *linear* classification estimators of the form

$$\hat{h}(x) = \mathbb{1}(\hat{\gamma}_0 + \hat{\gamma} \cdot x > 0).$$

The only, but essential, difference lies in *how* the parameters are estimated.

To contrast these two approaches we note that they factor the joint PDF differently.

- When constructing the LDA classification estimator we wrote
$$f(x,y) = f(x|y)f(y)$$
and made the assumption that
$$X \mid Y \text{ is Normal.}$$
We then estimated both the Normal parameters of the conditional PDF $f(x|y)$ and the Bernoulli parameter of the marginal $f(y)$.
- When constructing the logistic regression classification estimator we wrote the joint PDF the *other way around*, i.e.
$$f(x,y) = f(y|x)f(x),$$
and made the assumption that
$$Y \mid X \text{ is logistic,}$$
completely disregarding the distribution of the marginal PDF of X. This is because only the conditional PDF $f(y|x)$ matters for classification. We then estimated the logistic parameters, i.e. the linear parameters of the logit of the Bernoulli parameter of $Y \mid X$ (see also Remark 13.82).

In particular this makes it clear that the logistic regression classification estimator is *more nonparametric* than the LDA classification estimator.

22.6. **Density Estimation and Naive Bayes.** Ater constructing classification estimators following the *regression* approach in Section 22.4 (in the terminology of Remark 22.13) here we go back, as in Section 22.3, to following the *density estimation* approach.

Theorem 22.41 (Naive Bayes classifier)**.** *Let Y be a random variable with finite codomain $\{0, \ldots, K-1\}$ and let X be a random vector in $\mathbb{R}^d$. Suppose that the components $X_1, \ldots, X_d$ of X are independent. Then the Bayes classifier for Y given X is, for $x \in \mathbb{R}^d$,*

$$h^*(x) := \underset{0 \leqslant k \leqslant K-1}{\arg\max} \left(\pi_k \prod_{j=1}^{d} f_{kj}(x_j) \right)$$

and it is called the naive Bayes classifier*, where*

$$\pi_k := \mathbb{P}(Y = k) \text{ and } f_{kj} := f_{X_j}(x \mid Y = k)$$

such that f_{kj} is the conditional marginal PDF of X_j given $Y = k$.

Proof. By independence of the components of X this follows immediately from Theorem 22.14. □

Definition 22.42 (Naive Bayes classification estimator)**.** Consider an IID sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ from (Y, X) where Y is a random variable with finite codomain $\{0, \ldots, K-1\}$ and $X = \left(X^{(1)}, \ldots, X^{(d)}\right)$ is a random vector in $\mathbb{R}^d$. Let

$$\hat{\pi}_k := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Y_i = k)$$

and let $\hat{f}_{jk}$ be estimators of the conditional marginal PDF $f_{X^{(j)}}(x \mid Y = k)$ of the j-th component of X given $Y = k$ for $0 \leqslant k \leqslant K-1$ and $1 \leqslant j \leqslant d$. The classification estimator

$$\hat{h}(x) := \underset{0 \leqslant k \leqslant K-1}{\arg\max} \left(\pi_k \prod_{j=1}^{d} \hat{f}_{jk}(x_j) \right) \text{ for } x \in \mathbb{R}^d$$

is called a *naive Bayes classification estimator*.

Remark 22.43 (Non-uniqueness of naive Bayes classification estimators)**.** There are many ways to perform density estimation and so there are many ways to estimate the conditional marginal PDF

$$f_{jk}(x) := f_{X^{(j)}}(x \mid Y = k).$$

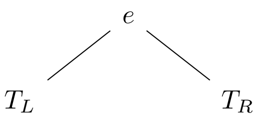
There are therefore many different ways to construct naive Bayes classification estimators.

Remark 22.44 (Practical use of naive Bayes classification estimators)**.** In practice naive Bayes classification estimators are often used when the covariate X is *discrete* and *high-dimensional*, for example when X has a finite codomain subset of $\mathbb{R}^d$ for d large.

22.7. **Trees.**

Definition 22.45 (Binary rooted tree)**.** Let S be a set. The set of *binary rooted trees* over S is defined recursively as follows.

- **Basis step.** There is a binary rooted tree called the *empty tree*. We say that this tree does not have a *root*.
- **Recursive step.** For any two binary rooted trees T_L and T_R and any $e \in S$ we define the binary rooted tree



as follows.

1. e is called the *root* of the tree.
2. If T_L is the empty tree then we say that e does not have a *left child*. Otherwise we call the root of T_L the left child of e.
3. If T_R is the empty tree then we say that e does not have a *right child*. Otherwise we call the root of T_R the right child of e.
4. We call e and the roots of T_L and T_R, if they exist, *vertices* of the tree.

Definition 22.46 (Leaf)**.** A vertex of a binary rooted tree which does not have any children is called a *leaf*. We denote the set of leaves of a binary rooted tree $\mathcal{T}$ by $\mathcal{L}(\mathcal{T})$.

Definition 22.47 (Power set)**.** Let S be a set. The *power set* of S, denoted $\mathcal{P}(S)$, is the set of all subsets of S.

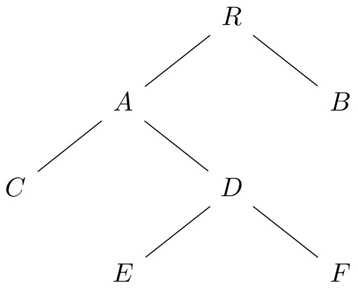
Definition 22.48 (Partition)**.** Let S be a set. A collection of pairwise disjoint subsets of S whose union covers S is called a *partition*.

Definition 22.49 (Tree partition)**.** Let $R \subseteq \mathbb{R}^d$. A *tree partition* $\mathcal{T}$ of R is a binary rooted tree over the power set of R satisfying the following conditions.

- The root of $\mathcal{T}$ is R.
- For every vertex V of $\mathcal{T}$, if V has children then it has exactly two children and they form a partition of V. Moreover the two children are separated by a *coordinate hyperplane*
$$\{x \in \mathbb{R}^d : x_j = c\}$$
for some $1 \leqslant j \leqslant d$ and some $c \in \mathbb{R}$.

Lemma 22.50. *The set of leaves of a tree partition of $R \subseteq \mathbb{R}^d$ is a partition of R.*

Example 22.51 (Tree partition)**.** The binary rooted tree



where $R = [0, 1]^2$ and where A, B, C, D, E, and F are as in Figure 22.1, is a tree partition of R. The leaves $\{B, C, E, F\}$ form a partition of R, as per Lemma 22.50.

Definition 22.52 (Tree classification map)**.** Let $K \geqslant 2$ be an integer, let $R \subseteq \mathbb{R}^d$, let $\mathcal{T}$ be a tree partition of R, and let $\mathcal{H} : \mathcal{L}(\mathcal{T}) \to \{0, \ldots, K-1\}$ be a map. The map
$$H : R \to \{0, \ldots, K-1\}$$

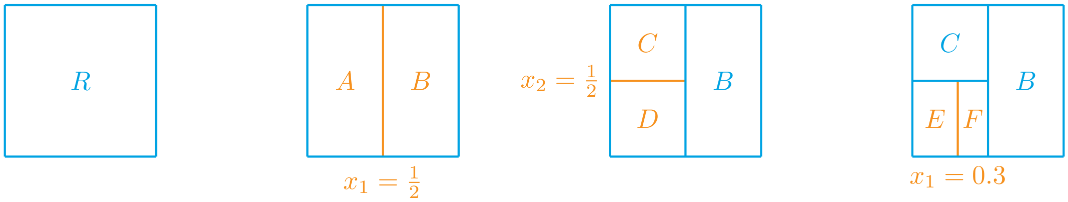


FIGURE 22.1. A tree partition of $R = [0, 1]^2$.

characterized by

$$H(x) = k \iff x \in L \text{ and } \mathcal{H}(L) = k \text{ for } L \in \mathcal{L}(\mathcal{T})$$

is called a *tree classification map.*

Remark 22.53 (Tree classification maps are well-defined)**.** Lemma 22.50 tells us that the leaves $\mathcal{L}(\mathcal{T})$ of $\mathcal{T}$ form a partition of R. Therefore every $x \in R$ belongs to one, and only one, leaf L. Defining H via

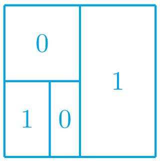
$$H(x) = k \iff x \in L \text{ and } \mathcal{H}(L) = k$$

is therefore perfectly unambiguous (i.e. well-defined).

Example 22.54 (Tree classification map)**.** Using the tree partition of Example 22.51 we may define the tree classification map $H : [0, 1]^2 \to \{0, 1\}$ (i.e. $K = 2$) by

$$H(x) = \begin{cases} 0 & \text{if } x \in C \text{ or } x \in F \text{ and} \\ 1 & \text{if } x \in B \text{ or } x \in D. \end{cases}$$

This is represented diagramatically below.



Definition 22.55 (Tree classification estimator)**.** Let Y be a random variable with finite codomain $\{0, \ldots, K-1\}$ and let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$. A random function h from $\mathcal{X}$ to $\{0, \ldots, K-1\}$ (i.e. a classification estimator) such that, for every outcome ω, $h(\cdot)(\omega)$ is a tree classification map is called a *tree classification estimator.*

Remark 22.56 (Constructing a tree classification estimator)**.** Let Y be a random variable with finite codomain $\{0, \ldots, K-1\}$ and let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$. Given IID samples

$$(Y_1, X_1), \ldots, (Y_n, X_n)$$

how do we construct a tree classification estimator? There are two stages to the construction of a tree classification estimator.

(1) First we *grow* the tree. Given a tree classification estimator h whose corresponding tree partition for this particular outcome is $\mathcal{T}$ we consider, for every leaf L of $\mathcal{T}$, if there is a hyperplane of the form

$$\{x \in \mathbb{R}^d : x_j \leqslant c\}$$

for $1 \leqslant j \leqslant d$ and $c \in \mathbb{R}$ with which it is worth *splitting* L into two, thus creating two children, C_L and C_R, of L.

We would then update the classification estimator by setting $h|_{C_L}$ and $h|_{C_R}$ to the sample value of Y most present in C_L and C_R, respectively, i.e.

$$h|_{C_L} := \underset{0 \leqslant k \leqslant K-1}{\arg\max} \sum_{i=1}^{n} \mathbb{1}\left(Y_i = k \text{ and } X_i \in C_L\right) \text{ and}$$

$$h|_{C_R} := \underset{0 \leqslant k \leqslant K-1}{\arg\max} \sum_{i=1}^{n} \mathbb{1}\left(Y_i = k \text{ and } X_i \in C_R\right).$$

How do we assess if it is worth splitting a leaf L into two? The *misclassification rate* for a region R, where here $R = C_L$, C_R, defined by

$$\hat{L}(h) := \frac{1}{n_R} \sum_{\substack{i=1 \\ X_i \in R}}^{n} \mathbb{1}\left(h\left(X_i\right) \neq Y_i\right) \text{ for } n_R := \sum_{i=1}^{n} \mathbb{1}\left(X_i \in R\right),$$

which estimates the true error rate $\mathbb{P}\left(h(X) \neq Y \mid X \in R\right)$ for the region R, is an appealing option to determine if L is worth splitting into two or not. However the misclassification rate is not sensitive to how "pure" a region is: splitting a region with (400, 400) samples, meaning 400 samples with $Y_i = 0$ and 400 samples with $Y_i = 1$, into

- (300, 100) labelled as 0 and (100, 300) labelled as 1, so making $100 + 100 = 200$ errors out of 800,
- (400, 200) labelled as 0 and (0, 200) labelled as 1, so making $200 + 0 = 200$ errors out of 800,

will produce the *same* misclassification rate, namely $\frac{1}{4}$, even though the latter subdivision is typically preferable because it produces a "pure" child where only one class is present.

That is why other measures than the misclassification rate are often used, such as the *Gini index* discussed below.

See also [HTF09] for additional details on tree-based methods.

Definition 22.57 (Gini index)**.** Let Y be a random variable with finite codomain $\{0, \ldots, K-1\}$, let X be a random vector in $\mathbb{R}^d$, and let R be a measurable region of $\mathbb{R}^d$. The *Gini index* of Y given X over R is defined to be

$$\mathcal{G}(R) := \sum_{k=0}^{K-1} p_k\left(1 - p_k\right) = 1 - \sum_{k=0}^{K-1} p_k^2 = 1 - p \cdot p$$

for

$$p_k := \mathbb{P}\left(Y = k \mid X \in R\right), 0 \leqslant k \leqslant K-1,$$

such that $p \in \Delta^{K-1} \subseteq \mathbb{R}^K$.

Lemma 22.58 (Range of the Gini index)**.** *Let Y be a random variable with finite codomain $\{0, \ldots, K-1\}$, let X be a random vector in $\mathbb{R}^d$, and let R be a measurable region of $\mathbb{R}^d$. The Gini index satisfies*

$$0 \leqslant \mathcal{G}(R) \leqslant 1 - \frac{1}{K}.$$

Moreover, for $p \in \Delta^{K-1} \subseteq \mathbb{R}^K$ as in Definition 22.57,

- *the minimum $\mathcal{G} = 0$ is achieved when $p = e_k$ for some $0 \leqslant k \leqslant K-1$, i.e. when p is* pure*, and*
- *the maximum $\mathcal{G} = 1 - \frac{1}{K}$ is achieved when $p = \frac{1}{K}\mathbb{1}$, i.e. when p is* perfectly mixed*, for $\mathbb{1} = (1, \ldots, 1) \in \mathbb{R}^K$.*

Definition 22.59 (Categorical version of a discrete random variable with finite codomain)**.** Let Y be a random variable with finite codomain $\{0, \ldots, K-1\}$. The random variable Z with codomain $\mathbb{R}^K$ defined by

$$Z_k := \mathbb{1}(Y = k)$$

is called the *categorical version*, or *one-hot encoding*, of Y.

Remark 22.60 (Categorical version of a discrete random variable with finite codomain)**.** In Definition 22.59 above Z is called the *categorical* version of Y because

$$Z \sim \mathrm{Categorical}(p)$$

for $\pi_k := \mathbb{P}(Y = k)$, $0 \leqslant k \leqslant K-1$. Moreover we can recover Y from Z via

$$Y = k \iff e_k \cdot Z = 1$$

since $Z \in \{e_0, e_1, \ldots, e_{k-1}\}$, where $e_0, \ldots, e_{k-1}$ denotes the canonical basis of $\mathbb{R}^K$.

Lemma 22.61 (Gini index and one-hot encoding)**.** *Let Y be a random variable with finite codomain $\{0, \ldots, K-1\}$, let Z be its categorical version, let X be a random vector in $\mathbb{R}^d$, and let R be a measurable region of $\mathbb{R}^d$. The Gini index of Y satisfies*

$$\mathcal{G}(R) = \operatorname{tr} \mathbb{V}(Z \mid X \in R).$$

Definition 22.62 (Gini dimension)**.** Let $p \in \Delta^{K-1}$. The *Gini dimension* of p is defined to be

$$\dim_{\mathcal{G}} p := \frac{1}{1-g} - 1 \text{ for } g := 1 - p \cdot p.$$

Remark 22.63 (Gini dimension)**.** The Gini dimension is defined to satisfy

$$\dim_{\mathcal{G}} p = \frac{1}{1-g} - 1 \iff g = 1 - \frac{1}{1 + \dim_{\mathcal{G}} p}.$$

The motivation for this is discussed in Remark 22.64 below.

Remark 22.64 (Interpretation of the Gini index and Gini dimension)**.** The Gini index measures the extent to which

$$Y \mid X \in R$$

is *pure* (i.e. Y only takes *one* value when X is in R) or *mixed* (we say that Y is *perfectly mixed* if it takes all values $0, \ldots, K-1$ with equal probability). We can make this idea precise using the categorical version, or one-hot encoding, Z of Y since Lemma 22.61 tells us that the Gini index satisfies

$$\mathcal{G}(R) = \operatorname{tr} \mathbb{V}(Z \mid X \in R).$$

Crucially: Theorem C.12 tells us that $\operatorname{tr} \mathbb{V}$ is a *"linear count"* of the dimension of $\operatorname{im} \mathbb{V}$. Since

$$\mathbb{V}(Z \mid X \in R) = \operatorname{diag} p - p \otimes p$$

(as per Exercise A.23.12) this comes down to a "linear count" of the "dimension" of the parameter p. This is made precise by the result of Lemma 22.58 and the terminology of Definition 22.62:

- if $p = e_k$ for some k, i.e. $p \in \{e_k\}$ where $\dim \{e_k\} = 0$, then the Gini index is $\mathcal{G} = 0$ and the Gini dimension is $\dim_{\mathcal{G}} p = 0$,
- if $p = \frac{1}{K}\mathbb{1} = \left(\frac{1}{K}, \ldots, \frac{1}{K}\right) \in \mathbb{R}^K$, i.e. $p \in \Delta^{K-1}$ where $\dim \Delta^{K-1} = K-1$ then the Gini index is $\mathcal{G} = 1 - \frac{1}{K}$ and the Gini dimension is $\dim_{\mathcal{G}} p = K-1$,
- if
$$p = \left(\underbrace{\frac{1}{k}, \ldots, \frac{1}{k}}_{k \text{ times}} 0, \ldots, 0\right) \in \mathbb{R}^K$$
for some $0 \leqslant k \leqslant K-1$, or any permutation thereof, i.e. "$p \in \Delta^{k-1}$" where $\dim \Delta^{k-1} = k-1$, then the Gini index is
$$\mathcal{G} = 1 - p \cdot p = 1 - \sum_{j=0}^{k-1} \left(\frac{1}{k}\right)^2 = 1 - \frac{k}{k^2} = 1 - \frac{1}{k}$$
and the Gini dimension is
$$\dim_{\mathcal{G}} p = \frac{1}{1-\mathcal{G}} - 1 = k - 1.$$

Note that the Gini dimension is, like the trace (see Theorem C.12), a "linear count" of the dimension. For example if
$$p = \left(1 - \frac{\varepsilon}{K-1}, \varepsilon, \ldots, \varepsilon\right) \in \mathbb{R}^K$$
for $0 < \varepsilon << 1$, such that $p \approx e_1$, then
$$g := 1 - p \cdot p = 1 - \left(1 - \frac{\varepsilon}{K-1}\right)^2 - (K-1)\,\varepsilon^2 = \frac{2}{K-1}\varepsilon + O(\varepsilon^2)$$
such that
$$\dim_{\mathcal{G}} p = \frac{1}{1-\mathcal{G}} - 1 \approx \frac{1}{1 - \frac{2}{K-1}\varepsilon} = \frac{1 - 1 + \frac{2}{K-1}\varepsilon}{1 - \frac{2}{K-1}\varepsilon} \approx \frac{2}{K-1}\varepsilon.$$
This makes sense: $p \approx e_1$, so $\dim_{\mathcal{G}} p \approx \dim_{\mathcal{G}} e_1 = 0$ *even though* p itself belongs to $(K-1)$–dimensional open subset of Δ^{K-1} and all K of the components of p are nonzero.

In summary: the Gini index measures the extent to which
$$Y \mid X \in R$$
is pure, i.e. $\dim_{\mathcal{G}} = 0$, versus perfectly mixed, i.e. $\dim_{\mathcal{G}} p = K-1$, allowing a *meaningful interpretation* of values in between by way of the *Gini dimension.* For example if $K = 10$ and
$$\mathcal{G} \approx \frac{3}{4} \iff \dim_{\mathcal{G}} p \approx 3$$
then we know that there are only about *three classes* of Y meaningfully left in R.

22.8. Assessing Error Rates and Choosing a Good Classifier.

Remark 22.65 (Choosing a good classifier)**.** Recall that classification estimators are *random* in the sense that, if Y is a random variable with finite codomain $\mathcal{Y}$ and X is a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$ then a classification estimator $\hat{h}$ for Y given X is a random function from $\mathcal{X}$ to $\mathcal{Y}$. Therefore the true error rate of a classification estimator $\hat{h}$ is itself a *random variable*: for every outcome ω,

$$L\left(\hat{h}(\,\cdot\,)(\omega)\right) = \mathbb{P}\left(\hat{h}(X)(\omega) \neq Y\right)$$

where the probability on the right-hand side is taken over the joint distribution of $(Y,\, X)$ while keeping the outcome ω *fixed*.

In practice we are given an IID sample $(Y_1,\, X_1)\,,\, \ldots,\, (Y_n,\, X_n)$ from $(Y,\, X)$. Note that fixing the sample is tantamount to fixing the outcome ω, such that $\hat{h}(\,\cdot\,)(\omega)$ is an honest-to-goodness classification rule, for any classification estimator $\hat{h}$. We then typically seek to construct a classification estimator $\hat{h}$ which minimizes the *mean* true error rate

$$\mathbb{E}\left[L\left(\hat{h}\right)\right].$$

A natural starting point for the estimation of this mean true error rate is to use the training error rate $\hat{L}_n\left(\hat{h}\right)$. However, as in the case of *regression* (see Theorem 13.54 and Remark 13.55), the training error rate is a *downward-biased* estimate of the mean true error rate.

Broadly speaking there are then two approaches to estimating the mean true error rate.

- **Cross-validation**. We estimate the mean true error rate $\mathbb{E}L\left(\hat{h}\right)$ by
  (1) replacing the expectation $\mathbb{E}$ with an *average over batches* (see Definition 22.70 below) and,
  (2) for each batch, replacing the true error rate L with the *training* error rate.
- **Probability inequalities.** This is typically used in the context of *empirical risk minimization* (see Remark 22.13) where, for every outcome ω, the classification rule $\hat{h}(\,\cdot\,)(\omega)$ belongs to a set $\mathcal{H}$ specified ahead of time.

  In that context we may construct $1-\alpha$ confidence intervals for the true error rate $L\left(\hat{h}(\,\cdot\,)(\omega)\right)$, and hence for the *mean* true error rate $\mathbb{E}L(\hat{h})$, of the form

  $$\hat{L}_n\left(\hat{h}(\,\cdot\,)(\omega)\right) \pm \varepsilon$$

  where $\varepsilon = \varepsilon\,(n,\, \alpha,\, \mathcal{H})$.

22.8.1. *Cross-validation.*

Definition 22.66 (Classification method)**.** Let Y be a random variable with finite codomain $\mathcal{Y}$ and let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$. A *classification method* for Y given X is a collection of maps

$$\mathfrak{h}_n : (\mathcal{Y} \times \mathcal{X})^n \to \mathcal{Y}^{\mathcal{X}}$$

for $n \geqslant 1$. In other words: for any $n \geqslant 1$, $Y_1,\, \ldots,\, Y_n \in \mathcal{Y}$, and $X_1,\, \ldots,\, X_n \in \mathcal{X}$, $\mathfrak{h}_n\left((Y_1,\, X_1)\,,\, \ldots,\, (Y_n,\, X_n)\right)$ is a classification rule for Y given X.

Remark 22.67 (Classification estimators from classification methods)**.** Let $\mathfrak{h}_n$ be a classification method. We may construct a classification estimator as follows: for every outcome ω let $\hat{h}$ be the classification estimator defined by

$$\hat{h}(\,\cdot\,)(\omega) := \mathfrak{h}_n\left((Y_1(\omega),\, X_1(\omega))\,,\, \ldots,\, (Y_n(\omega),\, X_n(\omega))\right)$$

for $(Y_1,\, X_1)\,,\, \ldots,\, (Y_n,\, X_n)$ IID with the distribution of $(Y,\, X)$.

In other words: in practice classification methods are nothing more than the *recipe* we use to systematically turn a sample into a classification rule.

For instance Definitions 22.18, 22.21, 22.34, and 22.39 are implicitly defining classification *methods* corresponding to quadratic discriminant analysis, linear discriminant analysis, Fisher's linear classification estimator, and logistic regression classification, respectively.

Definition 22.68 (Validation set cross-validation)**.** Let Y be a random variable with finite codomain $\mathcal{Y}$, let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$, let $(Y_1,\, X_1)\,,\, \ldots,\, (Y_n,\, X_n)$ be an IID sample from $(Y,\, X)$, and let $\mathfrak{h}$ be a classification method for Y given X. Let

$$\{\mathcal{T},\, \mathcal{V}\}$$

be a partition of $\{1,\, \ldots,\, n\}$ where we call $\mathcal{T}$ the *training set* and $\mathcal{V}$ the *validation set.* Let

$$\hat{h} := \mathfrak{h}\left((Y_i,\, X_i) : i \in \mathcal{T}\right)$$

be the classification rule produced by the classification method $\mathfrak{h}$ when considering only the samples in the *training* set. The *observed* error rate of that classifier on the *validation* set $\mathcal{V}$, namely

$$\hat{L}\left(\mathfrak{h}\right) := \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \mathbb{1}\left(\hat{h}(X_i) \neq Y_i\right)$$

is called the *validation set cross-validation* estimate of the mean true error rate of the classification method $\mathfrak{h}$ associated with the training set $\mathcal{T}$ and the validation set $\mathcal{V}$.

Remark 22.69 (Sizes of the training and validation sets)**.** In practice in Definition 22.68 above we typically choose

$$|\mathcal{T}| \approx \frac{9n}{10} and |\mathcal{V}| \approx \frac{n}{10},$$

i.e. taking the validation set to contain about one tenth of the total samples.

Definition 22.70 (K-fold cross-validation)**.** Let Y be a random variable with finite codomain $\mathcal{Y}$, let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$, consider an IID sample $(Y_1,\, X_1)\,,\, \ldots,\, (Y_n,\, X_n)$ drawn from $(Y,\, X)$, let $\mathfrak{h}$ be a classification method for Y given X, and let $K \geqslant 2$ be an integer. Let

$$\mathcal{S} = \{S_1,\, \ldots,\, S_K\}$$

be a partition of $[n] := \{1,\, \ldots,\, n\}$ into K subsets, which we call *batches*, satisfying $\max|S_i| \leqslant 1 + \min|S_i|$ (i.e. all the batches have essentially the same size). For $1 \leqslant k \leqslant K$ let

$$\hat{h}_{(k)} := \mathfrak{h}\left((Y_i,\, X_i) : i \in [n] \setminus S_k\right)$$

be the classification rule produced by the classification method $\mathfrak{h}$ when considering all batches *except* the k–th one and let

$$\hat{L}_{(k)} := \frac{1}{|S_k|} \sum_{i \in S_k} \mathbb{1}\left(\hat{h}_{(k)}(X_i) \neq Y_i\right)$$

be the *observed* error rate of that classification rule on the omitted batch S_k. We call

$$\hat{L}\left(\mathfrak{h}\right) := \frac{1}{K} \sum_{k=1}^{K} \hat{L}_{(k)}$$

the *K–fold cross-validation* estimate of the mean true error rate of the classification method $\mathfrak{h}$ associated with the partition $\mathcal{S}$.

Remark 22.71 (K-fold cross-validation in practice)**.** In practice we often choose $K \approx 10$ in Definition 22.70 above.

Remark 22.72 (K-fold and leave-one-out cross-validation)**.** K-fold cross-validation as introduced in Definition 22.70 is reminiscent of Definitions 13.61, 20.16, 20.37, and 20.49 where leave-one-out cross-validation is introduced in a variety of contexts. More precisely: leave-one-out cross-validation is *exactly* n-fold cross-validation (where n denotes the number of samples, as usual).

Nonetheless leave-one-out cross-validation does often have one benefit that generic K-fold cross-validation does not possess for $K < n$, namely identities that allow us to evaluate the leave-one-out cross-validation estimate *without* constructing n separate estimators.

Remark 22.73 (K-fold cross-validation and mean true error rate)**.** Recall from Remark 22.67 that, given a classification method $\mathfrak{h}$, we may construct the associated classification estimator $\hat{h}$ via

$$\hat{h}(\,\cdot\,)(\omega) = \mathfrak{h}\left((Y_1(\omega),\, X_1(\omega))\,,\,\dots\,,\,(Y_n(\omega),\, X_n(\omega))\right)$$

for any outcome ω. This means that, in Definitions 22.68 and 22.70, when we refer to the mean true error rate of the classification *method* it is implicitly understood that we are referring to the mean true error rate of the corresponding classification *estimator*, namely

$$\mathbb{E}\left[L\left(\hat{h}\right)\right]$$

where the expectation is taken over the sample space where the outcomes ω reside.

22.8.2. *Probability inequalities.*

Theorem 22.74 (Bound on the true error rate for finite sets of classifiers)**.** *Let Y be a discrete random variable with finite codomain, let X be a random vector, let $(Y_1,\, X_1)\,,\,\dots\,,\,(Y_n,\, X_n)$ be an IID sample drawn from $(Y,\, X)$, and let $\mathcal{H}$ be a finite set of classification rules for Y given X of size m. For every $\varepsilon > 0$,*

$$\mathbb{P}\left(\max_{h \in \mathcal{H}} \left|\hat{L}_n(h) - L(h)\right| > \varepsilon\right) \leqslant 2me^{-2n\varepsilon^2}.$$

Proof. The sub-additivity of probability measures tells us that

$$\mathbb{P}\left(\max_{h\in\mathcal{H}}\left|\hat{L}_n(h)-L(h)\right|>\varepsilon\right)\leqslant\mathbb{P}\left(\bigcup_{h\in\mathcal{H}}\left|\hat{L}_n(h)-L(h)\right|>\varepsilon\right)$$
$$\leqslant\sum_{h\in\mathcal{H}}\mathbb{P}\left(\left|\hat{L}_n(h)-L(h)\right|>\varepsilon\right).$$

We note that, for

$$p:=\mathbb{P}\left(h(X)\neq Y\right)=L(h),$$

$\hat{L}_n(h)$ is a sample average of IID Bernoulli(p) random variables since

$$\hat{L}_n(h)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(h(X_i)\neq Y_i\right)$$

and $\mathbb{P}\left(\mathbb{1}\left(h(X_i)\neq Y_i\right)=1\right)=\mathbb{P}\left(h(X_i)\neq Y_i\right)=p$. To conclude we then use Hoeffding's inequality (Theorem 4.4). □

Corollary 22.75 (Confidence interval for the true error rate for finite sets of classifiers)**.** *Let Y be a discrete random variable with finite codomain, let X be a random vector, let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from (Y, X), and let $\mathcal{H}$ be a finite set of classification rules for Y given X of size m. For any $\alpha\in(0, 1)$ we define*

$$\varepsilon:=\sqrt{\frac{2}{n}\log\frac{2m}{\alpha}}.$$

Then

$$\hat{L}_n(h)\pm\varepsilon$$

is a $1-\alpha$ confidence interval for the true error rate $L(h)$ for any $h\in\mathcal{H}$. In particular this holds if we choose

$$\hat{h}\in\arg\min_{\mathcal{H}}\hat{L}_n,$$

i.e. choose $\hat{h}$ by empirical risk minimization *(see Remark 22.13).*

Remark 22.76 (What about infinite sets of classifiers?)**.** Corollary 22.75 above provides confidence intervals for classifiers built by *empirical risk minimization* (see Remark 22.13) when the set of admissible classifiers is *finite*. In practice this is seldom the case and so we need to appeal instead to so-called *Vapnik–Chervonenkis theory*, or *VC theory*, to be able to handle the case where there are infinitely many admissible classifiers.

Definition 22.77 (Shattering)**.** Let $\mathcal{A}$ be a collection of subsets of $\mathbb{R}^d$ and let $F\subseteq\mathbb{R}^d$. We say that $\mathcal{A}$ *shatters* F if, for every subset $S\subseteq F$, there exists $A\in\mathcal{A}$ such that

$$A\cap F=S,$$

in which case we say that A *picks out* S.

Example 22.78 (Shattering)**.** For every $a\in\mathbb{R}$ we define $I_a:=(-\infty, a]$ and consider $\mathcal{A}:=\{I_a : a\in\mathbb{R}\}$ (i.e. the dimension is $d=1$). Any subset $F\subseteq\mathbb{R}$ containing *one* element x is shattered by $\mathcal{A}$ since

$$I_{x-1}\cap F=\emptyset\text{ and }I_x\cap F=\{x\}.$$

However, any subset F of $\mathbb{R}$ containing *two* elements $x < y$ *cannot* be shattered by $\mathcal{A}$ since no set in $\mathcal{A}$ picks out $\{y\}$. Indeed: if $y \in I_a$ for some a then I_a necessarily contains every real number smaller than y, including x. In other words: if $y \in I_a$, then $F \cap I_a = \{x, y\} \neq \{y\}$, i.e. indeed $\mathcal{A}$ cannot pick out $\{y\}$.

Definition 22.79 (Shatter coefficient)**.** Let $\mathcal{A}$ be a collection of subsets of $\mathbb{R}^d$ and let $F \subseteq \mathbb{R}^d$. We define $N_{\mathcal{A}}(F)$ to be the number of subsets of F shattered by $\mathcal{A}$ and we call

$$s(\mathcal{A}, n) := \max\left\{N_{\mathcal{A}}(F) : F \subseteq \mathbb{R}^d \text{ contains exactly } n \text{ elements}\right\}$$

the *n–th shatter coeffificent of $\mathcal{A}$*.

Example 22.80 (Shatter coefficient)**.** Note that, for any $n \geqslant 0$, the n–th shatter coefficient is bounded by

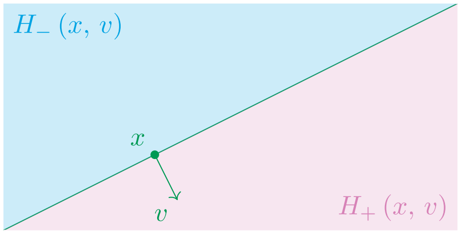
$$0 \leqslant s(\mathcal{A}, n) \leqslant 2^n.$$

How close the shatter coefficient is to 2^n measures how well sets in $\mathcal{A}$ are able to discriminate among n points.

(1) Consider $\mathcal{A} = \{I_a : a \in R\}$ for $I_a = (-\infty, a]$ as in Example 22.78. As observed in Example 22.78,
   - $N_{\mathcal{A}}(F) = 2 = 2^1$ if F has one element and
   - $N_{\mathcal{A}}(F) = 3 < 2^2$ if F has two elements. Indeed: if $F = \{x, y\}$ with $x < y$ then, as discussed in Example 22.78, $\mathcal{A}$ can only pick out $\emptyset$, $\{x\}$, and $\{x, y\}$, but *not* $\{y\}$.

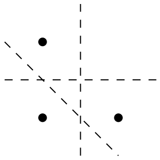
   Therefore
   $$s(\mathcal{A}, 1) = 2 \text{ and } s(\mathcal{A}, 2) = 3.$$

(2) Consider the collection of affine half-spaces of the plane $\mathbb{R}^2$, i.e.
   $$\mathcal{A} = \left\{H_\pm(x, v) : x, v \in \mathbb{R}^2\right\}$$
   where
   $$H_\pm(x, v) = \left\{v \in \mathbb{R}^2 : \pm v \cdot (y - x) \geqslant 0\right\}.$$
   Pictorially:



   Clearly $\mathcal{A}$ can shatter any one–element or two–element set, so we now ask ourselves: what are the 3–rd and 4–th shatter coefficvients of $\mathcal{A}$?
   - If $F \subseteq \mathbb{R}^2$ contains *three* elements, then $\mathcal{A}$ may or may not shatter F. If the three elements of F are *colinear*, or equivalently one element, say y, is a convex combination of the other two, say x and z, then $\mathcal{A}$ does *not* shatter F since it cannot pick out $\{x, z\}$. Indeed: the sets in $\mathcal{A}$ are convex and so any set in $\mathcal{A}$ containing both x and z also contain y.

     However, if the tree elements of F are *not* colinear, then without loss of generality (up to an affine transformation) they look like this



and so the lines depicted above may be used to shatter F with $\mathcal{A}$. Note that since $\mathcal{A}$ shatters *some* sets with three elements we deduce that

$$s\left(\mathcal{A},\,3\right) = 2^3 = 8.$$

- If $F \subseteq \mathbb{R}^2$ contains *four* elements then it can *never* be shattered by $\mathcal{A}$. This is proved in Exercise A.23.43. In particular this only shows that

  $$s\left(\mathcal{A},\,4\right) < 2^4 = 16,$$

  i.e. $s\left(\mathcal{A},\,4\right) \leqslant 15$. As we will see below in Definition 22.86 and Theorem 22.89, knowing that $s\left(\mathcal{A},\,n\right) < 2^n$ is crude but sufficient for our purposes here.

Definition 22.81 (Empirical probability measure)**.** Let $\mathbb{P}$ be a probability measure on $\mathbb{R}^d$ and let $X_1,\,\ldots\,,\,X_n \sim \mathbb{P}$ be IID. The probability measure $\mathbb{P}_n$ on $\mathbb{R}^d$ defined by

$$\mathbb{P}_n(A) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(X_i \in A\right) \text{ for every } A \subseteq \mathbb{R}^d$$

is called the *empirical probability measure*.

Theorem 22.82 (Vapnik and Chervonenkis)**.** *Let $\mathbb{P}$ be a probability measure on $\mathbb{R}^d$ and let $\mathcal{A}$ be a collection of subsets of $\mathbb{R}^d$. For any integer $n \geqslant 1$ and any $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} \left|\mathbb{P}_n(A) - \mathbb{P}(A)\right| > \varepsilon\right) \leqslant 8\, s\left(\mathcal{A},\,n\right) e^{-n\varepsilon^2/32}.$$

Definition 22.83 (Shatter coefficient for a collection of binary classifiers)**.** Let Y be a binary random variable, let X be a random vector, and let $\mathcal{H}$ be a set of classification rules for Y given X. Let

$$\mathcal{A} := \left\{h^{-1}(1) : h \in \mathcal{H}\right\}.$$

For any integer $n \geqslant 0$ we call

$$s\left(\mathcal{H},\,n\right) := s\left(\mathcal{A},\,n\right)$$

the *n–th shatter coefficient of $\mathcal{H}$*.

Corollary 22.84 (Bounds and confidence intervals for the true error rate for infinite sets of classifiers)**.** *Let Y be a binary random variable, let X be a random vector in $\mathbb{R}^d$, let $\left(Y_1,\,X_1\right),\,\ldots,\,\left(Y_n,\,X_n\right)$ be an IID sample drawn from $(Y,\,X)$, and let $\mathcal{H}$ be a set of classification rules. For any $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} \left|\hat{L}_n(h) - L(h)\right| > \varepsilon\right) \leqslant 8\, s\left(\mathcal{H},\,n\right) e^{-n\varepsilon^2/32}.$$

and, for any $\alpha \in (0,\,1)$, if we define

$$\varepsilon := \frac{32}{n}\log\left(\frac{s\left(\mathcal{H},\,n\right)}{\alpha}\right)$$

then

$$\hat{L}_n(h) \pm \varepsilon$$

is a $1-\alpha$ *confidence interval for the true error rate* $L(h)$ *for any* $h \in \mathcal{H}$*. In particular this holds if we choose*

$$\hat{h} \in \arg\sup_{h\in\mathcal{H}} \hat{L}_n,$$

i.e. choose $\hat{h}$ *by* empirical risk minimization *(see Remark 22.13).*

Remark 22.85 (Confidence intervals for the true error rate of finite and infinite sets of classifiers)**.** If we compare ε in Corollary 22.84 with ε in Corollary 22.75 we see that the shatter coefficient $s(\mathcal{H}, n)$ plays in Corollary 22.84 the role played by $m = |\mathcal{H}|$ in Corollary 22.75, thus answering the question posed in Remark 22.76.

Definition 22.86 (Vapnik-Chervnonenkis dimension)**.** Let $\mathcal{A}$ be a collection of subsets of $\mathbb{R}^d$. The *Vapnik–Chervonenkis dimension*, or *VC dimension*, of $\mathcal{A}$ is defined to be

$$VC(\mathcal{A}) := \log_2\left(\sup\left\{n \geqslant 0 : s(\mathcal{A}, n) = 2^n\right\}\right),$$

following the conventions that $\log_2(\sup\emptyset) = -\infty$ and $\log_2\infty = \infty$.

Remark 22.87 (Interpretation of the VC dimension)**.** If $VC(\mathcal{A}) = k$ then we are guaranteed that

- *some* sets with k elements can be shattered by $\mathcal{A}$ and
- *no* sets with $k+1$ elements can be shattered by $\mathcal{A}$.

It does *not* guaranteed that *all* sets with k elements may be shattered by $\mathcal{A}$ – see Example 22.88 below.

Example 22.88 (VC dimension)**.** Here are some examples of collections of subsets of $\mathbb{R}^d$ and their VC dimension.

(1) If $\mathcal{A} = \emptyset$ then $s(\mathcal{A}, n) = 0 < 2^n$ for all n and so $VC(\mathcal{A}) = -\infty$.
(2) If $\mathcal{A} = \{\emptyset\}$ then $s(\mathcal{A}, n) \equiv 1$ for all n and so $VC(\mathcal{A}) = 0$.
(3) Consider $\mathcal{A} = \{I_a : a \in \mathbb{R}\}$ for $I_a = (-\infty, a]$ as in Examples 22.78 and Example 22.80. We have shown in Example 22.80 that

$$s(\mathcal{A}, 1) = 2^1 \text{ but } s(\mathcal{A}, 2) < 2^2,$$

which means that $VC(\mathcal{A}) = 1$.
(4) Consider the collection $\mathcal{A}$ of affine half-spaces introduced in Example 22.80. Since

$$s(\mathcal{A}, 3) = 2^3 \text{ but } s(\mathcal{A}, 4) < 2^4$$

we deduce that $VC(\mathcal{A}) = 3$. Note that here, as discussed in Remark 22.87, $VC(\mathcal{A}) = 3$ even though some 3–element subsets of $\mathbb{R}^2$ *cannot* be shattered by $\mathcal{A}$ (see Example 22.80 for which ones).

Theorem 22.89 (VC dimension and bounds on the shatter coefficients)**.** *Let* $\mathcal{A}$ *be a collection of subsets of* $\mathbb{R}^d$ *with finite VC dimension* k*. Then the shatter coefficients of* $\mathcal{A}$ *satisfy*

$$s(\mathcal{A}, n) \leqslant n^k + 1$$

for all $n \geqslant 0$.

Lemma 22.90 (VC dimension of affine half-spaces)**.** *Let $\mathcal{A}$ be the collection of affine half-spaces of $\mathbb{R}^d$, i.e.*

$$\mathcal{A} = \left\{ H_\pm (x,\, v) : x,\, v \in \mathbb{R}^2 \right\}$$

where

$$H_\pm (x,\, v) = \left\{ v \in \mathbb{R}^2 : \pm v \cdot (y - x) \geqslant 0 \right\}.$$

Then $VC(\mathcal{A}) = d + 1$.

Remark 22.91 (VC dimension of affine half-spaces)**.** Lemma 22.90 is proved in the case $d = 2$ in Example 22.80, with the help of Exercise A.23.43. The general case $d > 2$ follows similarly, critically leveraging Radon's Theorem and the fact that both affine half-spaces and their complements are convex sets.

Corollary 22.92 (Confidence intervals for the true error rate of linear classifiers)**.** *Let Y be a binary random variable, let X be a random vector in $\mathbb{R}^d$, let $(Y_1,\, X_1), \ldots, (Y_n,\, X_n)$ be an IID sample drawn from $(Y,\, X)$, and let $\mathcal{H}$ be the set of* linear *classification rules, i.e.*

$$\mathcal{H} = \left\{ h : h = \mathbb{1}(r > c) \text{ for some } c \in \mathbb{R} \text{ and some linear functional } r : \mathbb{R}^d \to \mathbb{R} \right\}.$$

For any $\alpha \in (0,\, 1)$, if we define

$$\varepsilon := \frac{32}{n} \log \left(\frac{8\left(n^{d+1} + 1\right)}{\alpha} \right)$$

then, for any $h \in \mathcal{H}$,

$$\hat{L}_n(h) \pm \varepsilon$$

is a $1 - \alpha$ confidence interval for the true error rate $L(h)$.

Proof. This follows from combining Corollary 22.84, Theorem 22.89, and Lemma 22.90. □

22.9. Support Vector Machines.

Remark 22.93 (Terminology of Support Vector Machines)**.** The use of the term "Support Vector Machine" is a bit more loose in [Was10] that in the broader litterature. Following [HTF09] we make the following distinctions between several closely related *empirical risk minimization* classification estimators (see also Remark 22.13).

- **Optimal Separating Hyperplane.** When the data is *linearly separable* (see Definition 22.95) we seek the linear classification estimator which "best" separates the data. The resulting classifier is known as the Optimal Separating Hyperplane classifier, or OSH classifier.
- **Support Vector Classifier.** When the data is *not* linearly separable we seek the linear classifier which "best" separates the data *while* allowing some misclassification. The resulting classification estimator is known as the Support Vector Classifier, or SVC.

  To make the relation between OSH classifiers and SVC more explicit the terms *"hard-margin" SVC* (for OSH classisfiers) and *"soft-margin" SVC* (for SVC) are sometimes used.

- **Support Vector Machine.** Linear classifiers can be turned into nonlinear classifiers using the famous "kernel trick" (discussed in Remark 22.114). The name Support Vector Machine, or SVM, is then used to denote a Support Vector Classifier which is using the kernel trick.

  This trick can also be used for other linear classifiers. We discuss its application to Fisher's linear classification estimator and to the logistic regression classification estimator in Section 22.10 below.

In particular, some of the notes in this sectin are not found in [Was10] and are taken from [HTF09] instead.

Remark 22.94 (Empirical risk minimization over linear classifiers and rewriting binary responses)**.** So far we have considered binary random variables to have codomain $\{0, 1\}$. As seen in Definition 22.8, by way of Theorem 22.11, this means that we construct classification rules to be of the form

$$h = \mathbb{1}\left(r > \frac{1}{2}\right).$$

In particular when we restrict our attention to classifiers where r is *linear* (or, really, *affine*), as is done in empirical risk minimization (see Remark 22.13), the constant $\frac{1}{2}$ is a little clunky.

This is why, in this section, we consider instead binary random variables to have codomain $\{-1, +1\}$. Classifiers then take the form

$$h = \mathbb{1}\left(r > 0\right) \approx \operatorname{sign} r.$$

Definition 22.95 (Separating affine hyperplanes)**.** Let $A, B \subseteq \mathbb{R}^d$.

(1) Let $H : \mathbb{R}^d \to \mathbb{R}$ be an affine functional. We say that A and B are *separated by the affine hyperplane* $\{H = 0\}$ if

$$H|_A > 0 \text{ and } H|_B < 0,$$

meaning that $H(a) > 0 > H(b)$ for every $a \in A$ and $b \in B$.

(2) If there exists an affine functional H such that A and B are separated by the affine hyperplane $\{H = 0\}$ then we say that A and B can be *separated by an affine hyperplane.*

Definition 22.96 (Distance between sets)**.** Let $A \subseteq \mathbb{R}^d$.

(1) For any $B \subseteq \mathbb{R}^d$ the *distance between A and B* is defined to be

$$\begin{aligned} d\left(A, B\right) &:= \inf\left\{d\left(a, b\right) : a \in A \text{ and } b \in B\right\} \\ &= \inf\left\{|a - b| : a \in A \text{ and } b \in B\right\} \end{aligned}$$

(2) For any $x \in \mathbb{R}^d$ the *distance between x and A* is defined to be

$$d\left(x, A\right) := \inf_{x \in A} d\left(x, a\right).$$

Definition 22.97 (Separation margin and support vectors)**.** Let $A, B \subseteq \mathbb{R}^d$ be separated by the affine hyperplane $\{H = 0\} =: P$.

(1) The *separation margin* of H for A and B is

$$\operatorname{margin}\left(H; A, B\right) := d\left(A, P\right) + d\left(P, B\right).$$

(2) Any vector $x \in A$ satisfying
$$d\,(x,\,P) = d\,(A,\,P) = \inf_{\bar{x}\in A} d\,(\bar{x},\,P)$$
is called a *support vector* for A. Similarly any $x \in B$ satisfying
$$d\,(x,\,P) = d\,(B,\,P) = \inf_{\bar{x}\in B} d\,(\bar{x},\,P)$$
is called a *support vector* for B.

Definition 22.98 (Normalized separating affine hyperplanes)**.** Let $H : \mathbb{R}^d \to \mathbb{R}$ be an affine functional whose zero level set separates $A,\, B \subseteq \mathbb{R}^d$. We say that H is *normalized* if
$$\min_A H = -\max_B = 1.$$

Lemma 22.99 (Normalizing separating affine hyperplanes)**.** *Suppose that two compact sets $A,\, B \subseteq \mathbb{R}^d$ are separated by an affine hyperplane. Then there exists a normalized affine functional $H : \mathbb{R}^d \to \mathbb{R}$ whose zero level set separates A and B.*

Moreover if we write $H(x) = a_0 + a \cdot x$ for every $x \in \mathbb{R}^d$ where $a_0 \in \mathbb{R}$ and $a \in \mathbb{R}^d$ then
$$\textit{margin}\,(H;\,A,\,B) = \frac{2}{|a|}.$$

Proof. Since A and B are separated by an affine hyperplane there exists an affine functional $\widetilde{H} : \mathbb{R}^d \to \mathbb{R}$ for which
$$\widetilde{H}|_A > 0 \text{ and } \widetilde{H}|_B < 0.$$
In particular, since $\widetilde{H}$ is continuous and A is compact, $\widetilde{H}$ attains its minimum $c_+ > 0$ at some point $a_* \in A$. Similarly, since B is also compact, $\widetilde{H}$ attains its maximum $-c_- < 0$ at some point $b_* \in B$. Now let $\varphi : \mathbb{R} \to \mathbb{R}$ denote the affine function satisfying
$$\varphi(c_+) = 1 \text{ and } \varphi(-c_-) = -1,$$
namely
$$\varphi(s) = \frac{s - c_+}{c_+ + c_-} + \frac{s + c_-}{c_+ + c_-}.$$
Then
$$H := \varphi \circ \widetilde{H}$$
is an affine functional satisfying
$$\min_A H = \varphi\left(\min_A \widetilde{H}\right) = \varphi(c_+) = 1 \text{ and}$$
$$\max_B H = \varphi\left(\max_B \widetilde{H}\right) = \varphi(-c_-) = -1,$$
i.e. H is normalized as desired.

Morever Figure 22.2 tells us that
$$\text{margin}\,(H;\,A,\,B) = (a_* - b_*) \cdot \frac{a}{|a|}$$
where
$$a_* \cdot a = H(a_*) - a_0 \text{ and } b_* \cdot a = H(b_*) - a_0$$
such that, since $H(a_*) = -H(b_*) - 1$,
$$\text{margin}\,(H;\,A,\,B) = \frac{H(a_*) - H(b_*)}{|a|} = \frac{2}{|a|}$$

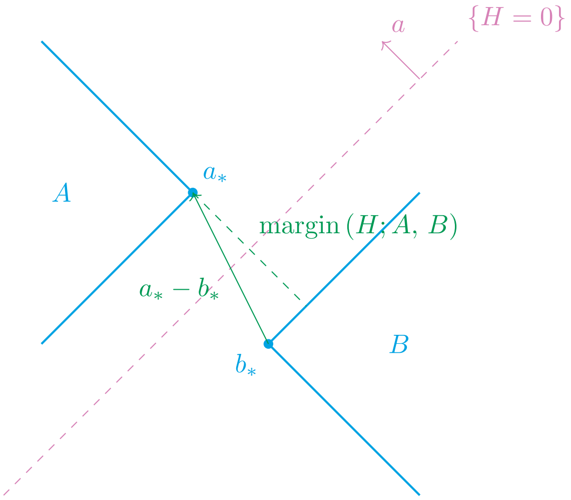


FIGURE 22.2. The diagram supporting the proof that margin $(H; A,\, B) = \frac{2}{|a|}$ when $H(x) = a_0 + a \cdot x$ for every $x \in \mathbb{R}^d$ when $a_0 \in \mathbb{R}$ and $a \in \mathbb{R}^d$.

as desired. □

Lemma 22.100 (Characterizing separating affine hyperplanes for compact sets)**.** *Let $D \subseteq \mathbb{R}^d$, let $Y : D \to \{-1,\, 1\}$ be a function (*not *a random variable), and suppose that the level sets*

$$\mathcal{C}_- := \{x \in D : Y(x) = -1\} \text{ and } \mathcal{C}_+ := \{x \in D : Y(x) = +1\}$$

are compact. Then $\mathcal{C}_-$ and $\mathcal{C}_+$ can be separated by an affine hyperplane if and only if there exists an affine functional $H : \mathbb{R}^d \to \mathbb{R}$ such that

$$YH \geqslant 1 \text{ on } D.$$

Proof. Suppose first that $\mathcal{C}_-$ and $\mathcal{C}_+$ can be separated by an affine hyperplane. Then Lemma 22.99 tells us that there exists an affine functional $H : \mathbb{R}^d \to \mathbb{R}$ such that

$$\min_{\mathcal{C}_+} H = 1 \text{ and } \min_{\mathcal{C}_-} H = -1.$$

In particular, since $Y = 1$ on $\mathcal{C}_+$ this means that

$$YH \geqslant 1 \text{ on } C_+$$

while, since $Y = -1$ on $\mathcal{C}_-$,

$$YH = -H \geqslant 1 \text{ on } C_-.$$

Since $\mathcal{C}_+ \cup \mathcal{C}_- = D$ this proves that $YH \geqslant 1$ on D, as desired.

Now suppose conversely that $YH \geqslant 1$. By definition of $\mathcal{C}_+$ this means that

$$H \geqslant 1 > 0 \text{ on } C_+$$

and by definition of $\mathcal{C}_-$ this means that

$$H \leqslant -1 < 0 \text{ on } C_-.$$

In other words: $\{H = 0\}$ separates $\mathcal{C}_-$ and $\mathcal{C}_+$, as desired. □

Corollary 22.101 (Characterizing separating affine hyperplanes for binary responses)**.** *Let Y be a random variable with codomain $\{-1, +1\}$, let X be a random vector in $\mathbb{R}^d$, and let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from (Y, X). The sets*

$$\mathcal{C}_- := \{X_i : Y_1 = -1\} \text{ and } \mathcal{C}_+ := \{X_i : Y_1 = +1\}$$

can be separated by an affine hyperplane if and only if there exists an affine function $H : \mathbb{R}^d \to \mathbb{R}$ such that

$$Y_i H(X_i) \geqslant 1 \text{ for all } i = 1, \ldots, n.$$

Proof. This follows from Lemma 22.100 since, for $D = \bigcup_{i=1}^n \{X_i\}$, the level sets of $\overline{Y} : D \to \{-1, +1\}$ defined by

$$\overline{Y}(X_i) = Y_i \text{ for all } i$$

are precisely $\mathcal{C}_-$ and $\mathcal{C}_+$. □

Corollary 22.102 (Characterizing separating affine hyperplanes for binary responses with maximal separation margin)**.** *Let Y be a random variable with codomain $\{-1, +1\}$, let X be a random vector in $\mathbb{R}^d$, and let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from (Y, X). Suppose that the sets*

$$\mathcal{C}_- := \{X_i : Y_1 = -1\} \text{ and } \mathcal{C}_+ := \{X_i : Y_1 = +1\}$$

can be separated by an affine hyperplane and let $\mathfrak{H}$ denote the set of affine functionals on $\mathbb{R}^d$ whose zero level sets separate $\mathcal{C}_-$ and $\mathcal{C}_+$. The maximizer $\widehat{H}$ of

$$\text{margin}\,(H; A, B) \text{ over } H \in \mathfrak{H}$$

is given by

$$\widehat{H}(x) = \hat{a}_0 + \hat{a} \cdot x \text{ for every } x \in \mathbb{R}^d$$

where $(\hat{a}_0, \hat{a}) \in \mathbb{R} \times \mathbb{R}^d$ is the minimizer of

$$|a| \text{ over } (a_0, a) \in \mathbb{R} \times \mathbb{R}^d \text{ subject to } Y_i\,(a_0 + a \cdot X_i) \geqslant 1 \text{ for all } i.$$

Proof. By the proof of Lemma 22.99 we know that every $\widetilde{H} \in \mathfrak{H}$ may be written as

$$\widetilde{H} = \varphi \circ H$$

where $\varphi : \mathbb{R} \to \mathbb{R}$ is affine and H is a normalized affine functional. In particular $\widetilde{H}$ and H have the same level sets, up to the relabelling carried out by φ, and so they have the same separation margin for $\mathcal{C}_+$ and $\mathcal{C}_-$. In order to maximize the separation margin over $\mathfrak{H}$ it this sufficies to maximize it over the smaller subset of *normalized* affine functionals separating $\mathcal{C}_+$ and $\mathcal{C}_-$.

For such an affine functional H, if we write $H(x) = a_0 + a \cdot x$ for $x \in \mathbb{R}^d$, then Lemma 22.99 tells us that

$$\text{margin}\,(H; \mathcal{C}_-, \mathcal{C}_+) = \frac{2}{|a|},$$

so indeed *maximizing* the margin is equivalent to *minimizing* $|a|$.

Finally we note from Lemma 22.100 that if H separates $\mathcal{C}_-$ and $\mathcal{C}_+$ then

$$Y_i\,(a_0 + a \cdot X_i) = Y_i H\,(X_i) \geqslant 1 \text{ for all } i.$$

To conclude: maximizing the margin over functionals separating $\mathcal{C}_-$ and $\mathcal{C}_+$ is equivalent to minimizing $|a|$ over the constraints $Y_i\,(a_0 + a \cdot X_i) \geqslant 1$ for all i. □

Lemma 22.103 (Wolfe dual of the maximal separation margin problem)**.** *Let $y_1, \ldots, y_n \in \{-1, +1\}$ and $x_1, \ldots, x_n \in \mathbb{R}^d$ be fixed. Consider the optimization problem*

$$\min_{(a_0,\, a) \in \mathbb{R} \times \mathbb{R}^d} \frac{1}{2}|a|^2 \text{ subject to } y_i\left(a_0 + a \cdot x_i\right) \geqslant 1 \textit{ for all } 1 \leqslant i \leqslant n.$$

Suppose that this problem is strictly feasible, meaning that

$$y_i\left(a_0 + a \cdot x_i\right) > 1 \textit{ for all } i$$

for some $(a_0,\, a) \in \mathbb{R} \times \mathbb{R}^d$. Then any minimizer $(a_0,\, a) \in \mathbb{R} \times \mathbb{R}^d$ satisfies

$$a = \sum_{i=1}^{n} y_i \lambda_i x_i$$

where $\lambda \in \mathbb{R}^n$ is a maximizer of the parametrized Wolfe dual *(see Remark B.23)*

$$\max_{\lambda \in \mathbb{R}^n} \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,k=1}^{n} y_i y_k \left(x_i \cdot x_k\right) \lambda_i \lambda_k \text{ subject to } \sum_{i=1}^{n} y_i \lambda_i = 0 \textit{ and } \lambda \succcurlyeq 0$$

and a_0 is recovered from the complementary slackness condition

$$y_i\left(a_0 + a \cdot x_i\right) = 1 \iff a_0 = y_i - a \cdot x_i$$

(since $y_i = \pm 1$) for any i for which $\lambda_i > 0$.

Proof. The original optimization problem has a quadratic objective function, affine inequality constraints, and is strictly feasible by assumption. Therefore it satisfies the assumptions of Corollary B.22 (and Remark B.23). To see what this result would tell us here we compute the stationarity condition of the original problem. Its Lagrangian is

$$\begin{aligned} L\left(a_0,\, a,\, \lambda\right) &= \frac{1}{2}|a|^2 + \sum_{i=1}^{n} \lambda_i \left[1 - y_i\left(a_0 + a \cdot x_i\right)\right] \\ &= \frac{1}{2}|a|^2 + \sum_{i=1}^{n} \lambda_i - \left(\sum_{i=1}^{n} y_i \lambda_i\right) a_0 - a \cdot \left(\sum_{i=1}^{n} y_i \lambda_i x_i\right) \end{aligned}$$

and so

$$\partial_{a_0} L = -\sum_{i=1}^{n} y_i \lambda_i = -y \cdot \lambda$$

while

$$\nabla_a L = a - \sum_{i=1}^{n} y_i \lambda_i x_i.$$

So the stationarity conditions of the original problem read

$$y \cdot \lambda = 0 \text{ and } a = \sum_{i=1}^{n} y_i \lambda_i x_i.$$

Therefore its Wolfe dual is

$$\max_{(a_0,\, a,\, \lambda) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^n} L\left(a_0,\, a,\, \lambda\right) \text{ subject to } \lambda \succcurlyeq 0 \text{ and } \nabla_{(a_0,\, a)} L = 0,$$

or equivalently

$$\max_{(a_0,\, a,\, \lambda)\in\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}^n} \frac{1}{2}|a|^2 + \sum_{i=1}^n \lambda_i - (y\cdot\lambda)\, a_0 - a\cdot\left(\sum_{i=1}^n y_i\lambda_i x_i\right)$$

$$\text{subject to } \lambda \succcurlyeq 0,\ y\cdot\lambda - 0, \text{ and } a = \sum_{i=1}^n y_i\lambda_i x_i.$$

We may now proceed as in Remark B.23 and use the stationarity conditions to *parametrize* the Wolfe dual's objective. We obtain

$$\begin{aligned}\frac{1}{2}|a|^2 + \sum_{i=1}^n \lambda_i - (y\cdot\lambda)\, a_0 + a\cdot\left(\sum_{i=1}^n y_i\lambda_i x_i\right) &= \frac{1}{2}|a|^2 + \sum_{i=1}^n \lambda_i - 0 - |a|^2\\ &= \sum_{i=1}^n \lambda_i - \frac{1}{2}|a|^2\\ &= \sum_{i=1}^n \lambda_i - \frac{1}{2}\sum_{i,k=1}^n y_i y_k\,(x_i\cdot x_k)\,\lambda_i\lambda_k.\end{aligned}$$

This means that any maximizer $(a_0,\, a,\, \lambda)$ of the Wolfe dual satisfies

$$a = \sum_{i=1}^n y_i\lambda_i x_i$$

for λ a maximizer of the *parametrized* Wolfe dual

$$\max_{\lambda\in\mathbb{R}^n} \sum_{i=1}^n \lambda_i - \frac{1}{2}\sum_{i,k=1}^n y_i y_k\,(x_i\cdot x_k)\,\lambda_i\lambda_k \text{ subject to } \lambda\succcurlyeq 0 \text{ and } y\cdot\lambda = 0.$$

To conclude we recall from Corollary B.22 that the maximizer λ obtained from the Wolfe dual agrees with the optimal Lagrange multiplier of the original problem. But then the complementary slackness condition of Lemma B.12 (which holds since the assumptions of Theorem B.11 hold, and so strong duality holds) tells us that, for this optimal Lagrange multiplier λ,

$$\lambda_i\,[1 - y_i\,(a_0 + a\cdot x_i)] = 0 \text{ for all } 1\leqslant i\leqslant n.$$

In particular we may recover a_0 from this equation for any i for which $\lambda_i > 0$, as then necessarily $y_i\,(a_0 + a\cdot x_i) = 1$, i.e.

$$a_0 = y_i - a\cdot x_i$$

(since $y_i = \pm 1$), as desired. □

Corollary 22.104 (Constructing maximally separating affine hyperplanes from data)**.** *Let Y be a random variable with codomain $\{-1,\, +1\}$, let X be a random vector in $\mathbb{R}^d$, and let $(Y_1,\, X_1)\,,\,\dots,\,(Y_n,\, X_n)$ be an IID sample drawn from $(Y,\, X)$. Suppose that the sets*

$$\mathcal{C}_- := \{X_i : Y_1 = -1\} \text{ and } \mathcal{C}_+ := \{X_i : Y_1 = +1\}$$

can be separated by an affine hyperplane and let $\widehat{H}$ be an affine functional with largest separation margin among those whose zero level sets separate $\mathcal{C}_-$ and $\mathcal{C}_+$. If we write

$$\widehat{H}(x) = \hat{a}_0 + \hat{a}\cdot x \text{ for every } x\in\mathbb{R}^d$$

then $\hat{a}_0 \in \mathbb{R}$ and $\hat{a} \in \mathbb{R}^d$ may be obtained as follows.

(1) *Let $\hat{\lambda} \in \mathbb{R}^n$ be a maximizer of*

$$\max_{\lambda \in \mathbb{R}^n} \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i,k=1}^n Y_i Y_k \left(X_i \cdot X_k\right) \lambda_i \lambda_k$$

$$\text{subject to } \sum_{i=1}^n Y_i \lambda_i = 0 \textit{ and } \lambda_i \geqslant 0 \textit{ for all } 1 \leqslant i \leqslant n.$$

(2) *Let*

$$\hat{a} := \sum_{i=1}^n Y_i \hat{\lambda}_i X_i.$$

(3) *For any i for which $\hat{\lambda}_i > 0$ let $\hat{a}_0$ satisfy*

$$Y_i \left(\hat{a}_0 + \hat{a} \cdot X_i\right) = 1 \iff \hat{a}_0 = Y_i - \hat{a} \cdot X_i$$

(since $Y_i = \pm 1$). In particular: for any i for which $\hat{\lambda}_i > 0$, X_i is a support vector of the separating affine hyperplane $\left\{\widehat{H} = 0\right\}$ for $\mathcal{C}_{\operatorname{sign}(Y_i)}$.

In particular this means that

$$\widehat{H}(x) = \hat{a}_0 + \sum_{i=1}^n Y_i \hat{\lambda}_i \left(X_i \cdot x\right)$$

for every $x \in \mathbb{R}^d$ where we need only sum over i's for which $\hat{\lambda}_i > 0$.

Remark 22.105 (Number of support vectors in practice)**.** Note that in practice in Corollary 22.104 we expect the number of support vectors to be *small*. More precisely: if the distribution of the covariate X is not too degenerate then there will typically be *two* support vectors, one for each class. Writing $i \in \{-, +\}$ for these two support vectors, meaning that $Y_\pm = \pm 1$, we then have that

$$\widehat{H}(x) = \hat{a}_0 + \hat{\lambda}_+ \left(X_+ \cdot x\right) - \hat{\lambda}_- \left(X_- \cdot x\right) = \hat{a}_0 + \left(\hat{\lambda}_+ X_+ - \hat{\lambda}_- X_-\right) \cdot x.$$

Proof of Corollary 22.104. As in the proof of Corollary 22.102 we may without loss of generality restrict our attention to normalized affine functionals separating $\mathcal{C}_-$ and $\mathcal{C}_+$. Writing such affine functionals as

$$H(x) = a_0 + a \cdot x$$

for $(a_0, a) \in \mathbb{R} \times \mathbb{R}^d$, Corollary 22.102 tells us that the maximal separation margin corresponds to the minimizer of

$$\min_{(a_0,\, a) \in \mathbb{R} \times \mathbb{R}^d} \frac{1}{2} |a|^2 \text{ subject to } Y_i \left(a_0 + a \cdot X_i\right) \geqslant 1 \text{ for all } i.$$

Since $\mathcal{C}_-$ and $\mathcal{C}_+$ can be separated by an affine hyperplane, Corollary 22.101 tells us that this problem is feasible, i.e. there exists an affine functional H and $(a_0, a) \in \mathbb{R} \times \mathbb{R}$ such that

$$Y_i H(X_i) = Y_i \left(a_0 + a \cdot X_i\right) \geqslant 1 \text{ for all } i.$$

In particular, since the constraints are *affine*, feasibility implies strict feasibility: if (a_0, a) is feasible then $(2a_0, 2a)$ is *strictly* feasible. Therefore the assumptions of Lemma 22.103 are satisfied, which establishes (1)–(3) *except* the part about support vectors.

That last past follows from complementary slackness: if $\hat{\lambda}_i > 0$ then the corresponding inequality constraint holds with *equality*, i.e.

$$Y_i \left(\hat{a}_0 + \hat{a} \cdot X_i \right) = 1.$$

Since $Y_i = \pm 1$ this means that

$$\widehat{H}(X_i) = \operatorname{sign}(Y_i) = \pm 1.$$

Moreover, since the zero level set $\{H = 0\}$ is affine, we can view H as a multiple of the signed distance function from that level set. But remember that, since $\widehat{H}$ is *normalized*,

$$\widehat{H}|_{\mathcal{C}_-} \leqslant -1 \text{ and } \widehat{H}|_{\mathcal{C}_+} \geqslant +1.$$

So the only way for $\widehat{H}(X_i)$ to *equal* ± 1 is for X_i to be one of the points in $\mathcal{C}_\pm$ *closest* to $\{H = 0\}$, i.e. indeed X_i is a support vector. □

Definition 22.106 (Optimal Separating Hyperplane classifier)**.** Under the assumptions of Corollary 22.104 the classification estimator

$$\hat{h} := \mathbb{1}\left(\widehat{H} > 0 \right),$$

for $\widehat{H}$ constructed as in Corollary 22.104, is called the *Optimal Separating Hyperplane classifier*, or *OSH classifier*.

Remark 22.107 (What if the data is not linearly separable?)**.** Corollary 22.104 tells us how to construct an Optimal Separating Hyperplane if the sample can be separated by an affine hyperplane. What if that is not possible?

A common approach in this case is to continue maximizing the separation margin while allowing some samples to be on the *wrong* side of the separating affine hyperplane. Lemma 22.108 below suggests how to do this while controlling the number of misclassifications.

Lemma 22.108 (Controlling the number of misclassifications)**.** *Consider some fixed points $y_1, \dots, y_n \in \{-1, +1\}$ and $x_1, \dots, x_n \in \mathbb{R}^d$ and let $H : \mathbb{R}^d \to \mathbb{R}$ be an affine functional separating*

$$\mathcal{C}_- := \{x_i : y_1 = -1\} \text{ and } \mathcal{C}_+ := \{x_i : y_1 = +1\}.$$

Then

$$\sum_{i=1}^{n} \mathbb{1}\left(\operatorname{sign} H(X_i) \neq Y_i\right) \leqslant \sum_{i=1}^{n} \left[1 - Y_i H(X_i)\right] \mathbb{1}\left(\operatorname{sign} H(X_i) \neq Y_i\right) =: (\star).$$

In other words the number of points misclassified *by H is bounded above by $(\star)$.*

Proof. It suffices to show that, if X_i is misclassified, meaning that $\operatorname{sign} H(X_i) \neq Y_i$, then

$$1 - Y_i H(X_i) \geqslant 1.$$

So let i correspond to such a point. Crucially, we observe that

$$\operatorname{sign}\left[Y_i H(X_i)\right] = \operatorname{sign}(Y_i) \operatorname{sign} H(X_i) = Y_i \operatorname{sign} H(X_i) \leqslant 0,$$

i.e. $Y_i H(X_i) \leqslant 0$. From this inequality it follows immediately that

$$1 - Y_i H(X_i) \geqslant 1$$

as desired. □

Definition 22.109 (Support Vector Classifier)**.** Let Y be a random variable with codomain $\{-1,\, +1\}$, let X be a random vector in $\mathbb{R}^d$, and let $(Y_1,\, X_1)\,,\, \dots,\, (Y_n,\, X_n)$ be an IID sample drawn from $(Y,\, X)$. Let $N \geqslant 0$ and consider the affine functional $\hat{H} : \mathbb{R}^d \to \mathbb{R}$ defined by, for every $x \in \mathbb{R}^d$,

$$\widehat{H}(x) = a_0 + a \cdot x$$

where $(a_0,\, a,\, \xi) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^n$ is a minimizer of

$$|a| \text{ subject to } Y_i\,(a_0 + a \cdot X_i) \geqslant 1 - \xi_i \text{ and } \xi_i \geqslant 0 \text{ for all } i \text{ and } \sum_{i=1}^{n} \xi_i \leqslant N.$$

The classification estimator

$$\hat{h} := \mathbb{1}\left(\widehat{H} > 0\right)$$

is called the *Support Vector Classifier*, or *SVC*, with parameter N.

Remark 22.110 (Motivation for the SVC)**.** As noted in Remark 22.107, when the sample is *not* separable by an affine hyperplane we typically continue maximizing the separation margin while allowing some samples to be on the wrong side of the separating affine hyperplane. This is precisely what the constraints

$$Y_i \underbrace{(a_0 + a \cdot X_i)}_{=H(X_i)} \geqslant 1 - \xi_i \text{ and } \xi_i \geqslant 0$$

allow, where ξ_i measures the failure of X_i to be outside the margin set

$$\left\{x \in \mathbb{R}^d : -1 \leqslant H(x) \leqslant 1\right\}.$$

Meanwhile the constraint

$$\sum_{i=1}^{n} \xi_i \leqslant N$$

ensures that the number of misclassified samples is under control (as motivated in Lemma 22.108 and proved in Corollary 22.111 below).

Corollary 22.111 (Number of misclassifications of the SVC)**.** *Under the assumptions of Definition 22.109,*

$$\sum_{i=1}^{n} \mathbb{1}\left(\hat{h}(X_i) \neq Y_i\right) \leqslant N.$$

In other words the SVC will misclassify no more than N samples.

Proof. This follows from Lemma 22.108 and the constraints on ξ:

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{1}\left(\hat{h}(X_i) \neq Y_i\right) &\leqslant \sum_{i=1}^{n} \mathbb{1}\,\mathrm{sign}\left(\widehat{H}(X_i) \neq Y_i\right)\\ &\leqslant \sum_{i=1}^{n} \left[1 - Y_i H(X_i)\right] \mathbb{1}\left(\mathrm{sign}\, H(X_i) \neq Y_i\right)\\ &\leqslant \sum_{i=1}^{n} \xi_i \mathbb{1}\left(\mathrm{sign}\, H(X_i) \neq Y_i\right)\\ &\leqslant \sum_{i=1}^{n} \xi\\ &\leqslant N.\end{aligned}$$

□

Lemma 22.112 (Equivalent formulation of the SVC optimization problem)**.** *Let $y_1, \ldots, y_n \in \{-1, +1\}$ and $x_1, \ldots, x_n \in \mathbb{R}^d$ be fixed. For any $C, N \geqslant 0$ we denote by $P(N)$ the optimization problem*

$$\min_{(a_0,\, a,\, \xi) \in \mathbb{R}\times\mathbb{R}^d\times\mathbb{R}^n} \frac{1}{2}|a|^2$$

$$\text{subject to } y_i\left(a_0 + a \cdot x_i\right) \geqslant 1 - \xi_i \textit{ and } \xi_i \geqslant 0 \textit{ for all } 1 \leqslant i \leqslant n \textit{ and } \sum_{i=1}^{n} \xi_i \leqslant N$$

and by $P(C)$ the optimization problem

$$\min_{(a_0,\, a,\, \xi) \in \mathbb{R}\times\mathbb{R}^d\times\mathbb{R}^n} \frac{1}{2}|a|^2 + C\sum_{i=1}^{n} \xi_i$$
$$\text{subject to } y_i\left(a_0 + a \cdot x_i\right) \geqslant 1 - \xi_i \textit{ and } \xi_i \geqslant 0 \textit{ for all } 1 \leqslant i \leqslant n.$$

These two problems are equivalent *in the following sense.*

1. *For all $C > 0$, if (a_0, a, ξ) is a minimizer of $P(C)$ then there exists $\overline{N} \geqslant 0$ such that (a_0, a, ξ) is also a minimizer of $P(N)$ for every $N \geqslant \overline{N}$.*
2. *There exists $N_* \geqslant 0$ such that, for every $N \geqslant N_*$, $P(N)$ is strictly feasible and, if (a_0, a, ξ) is a minimizer of $P(N)$ then there exists $\overline{C} \geqslant 0$ such that (a_0, a, ξ) is also a minimizer of $P(C)$ for every $C \geqslant \overline{C}$.*

Proof. We may prove this result by using Corollary B.19 and comparing the KKT conditions of both problems. In order to use Corollary B.19 we only need to verify that each problem is strictly feasible. Since

$$\xi_i := 1 + \left[1 - y_i\left(a_0 + a \cdot x_i\right)\right]_+$$

satisfies both $\xi \geqslant 0$ and

$$\xi > 1 - y_i\left(a_0 + a \cdot x_i\right) \iff y_i\left(a_0 + a \cdot x_i\right) > 1 - \xi_i,$$

we deduce that $P(C)$ is strictly feasible for all $C \geqslant 0$. The same ξ shows that $P(N)$ is strictly feasible *provided* N is sufficiently large. (This makes sense since, for small N, we do not expect $P(N)$ to be feasible. For example if $N = 0$ then the problem is feasible if and only if the samples x_i corresponding to $y_i = -1$ and $y_i = +1$ are *linearly separable*.) Comparing the KKT conditions of both problems is then straightforward but lengthy and so we omit it here. □

Remark 22.113 (Wolfe dual of the SVC optimization problem)**.** In practice the SVC is computed using the Wolfe dual of the problem $P(C)$ recorded in Lemma 22.112. This Wolfe dual is

$$\max_{\lambda\in\mathbb{R}^n}\sum_{i=1}^n \lambda_i - \frac{1}{2}\sum_{i,k=1}^n Y_iY_k\left(X_i\cdot X_k\right)\lambda_i\lambda_k \text{ subject to } 0\preccurlyeq\lambda\preccurlyeq C \text{ and } \mathbb{Y}\cdot\lambda=0,$$

meaning that

$$0\leqslant\lambda_i\leqslant C \text{ for all } i \text{ and } \sum_{i=1}^n Y_i\lambda_i=0.$$

The minimizer $(a_0,\, a)\in\mathbb{R}\times\mathbb{R}^d$ is the recovered as in Lemma 22.103 via

$$a=\sum_{i=1}^n Y_i\lambda_iX_i$$

while a_0 is recovered via complementary slackness.

22.10. **Kernelization.**

Remark 22.114 (The kernel trick)**.** We can obtain *nonlinear* decision boundaries with *linear* classification rules by nonlinearly embedding the original space $\mathbb{R}^d$, where the covariate X lives, into another (typially) higher-dimensional) space.

The prototypical example is to construct the unit circle as a decision boundary. To do so we define $\phi:\mathbb{R}^2\to\mathbb{R}$ by

$$\phi\left(x_1,\, x_2\right):=x_1^2+x_2^2$$

and consider the classification rule over its *codomain* defined by

$$h(s):=\mathbb{1}\left(s>1\right).$$

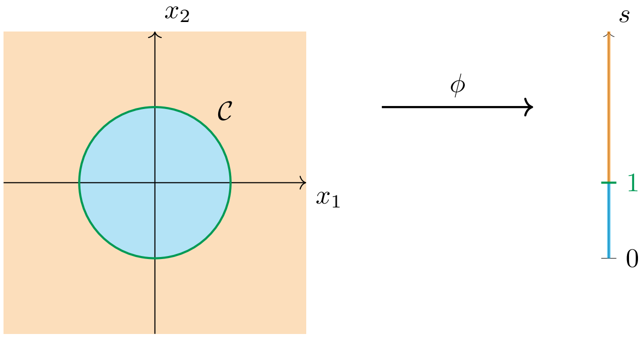
Then the *pullback* $h^*:=h\circ\phi$, such that

$$h^*\left(x_1,\, x_2\right)=h\left(\phi(x_1,\, x_2)\right)=\mathbb{1}\left(x_1^2+x_2^2>1\right),$$

has decision boundary the unit circle

$$\mathcal{C}=\left\{(x_1,\, x_2)\in\mathbb{R}^2 : x_1^2+x_2^2=1\right\}.$$

Pictorially:



In general this procedure comes down to finding an inner product space H (in the example above $H=\mathbb{R}$) and a map $\phi:\mathbb{R}^d\to H$, which is called a *feature map*. Given a random variable Y with finite codomain $\mathcal{Y}$, a random vector X with codomain $\mathbb{R}^d$, and an IID sample $\left(Y_1,\, X_1\right),\, \ldots,\, \left(Y_n,\, X_n\right)$ drawn from $(Y,\, X)$ we then seek to construt a classification estimator for Y given $\phi(X)$.

Where do kernels come in? For kernels to come in we must make the following observation: several classification methods do not require full knowledge of the covariate samples $X_1, \ldots, X_n$ and instead only require their *dot products*

$$X_i \cdot X_j \text{ for } 1 \leqslant i,\, j \leqslant n.$$

This is the case for the OSH classifier and the SVC, as made clear by Corollary 22.104 and Remark 22.113, respectively. We will see in Remarks 22.122 and 22.124 below that this is also the case for Fisher's linear classification estimator, and for the logistic regression classification estimator.

Combining these two observations, namely

- the use of a feature map $\phi : \mathbb{R}^d \to V$ and
- the focus on dot products $X_i \cdot X_j$,

we see that we need only consider the inner products

$$\left\langle \phi\left(X_i\right),\, \phi\left(X_j\right)\right\rangle_H.$$

This is why in practice we define the *kernel*

$$K\left(x,\, \tilde{x}\right) := \left\langle \phi(x),\, \phi(\tilde{x})\right\rangle_H$$

for every $x,\, \tilde{x} \in \mathbb{R}^d$. Crucially: once we have the kernel K in hand we do not need the feature map ϕ anymore. Theorem 22.118 below goes even further: if the kernel K is sufficiently nice then it implicitly determines the feature map ϕ.

The punchline is that we can obtain nonlinear decision boundaries with linear classifiers provided we specify a kernel. That is what we call the *"kernel trick"*.

Definition 22.115 (Kernel)**.** Let $D \subseteq \mathbb{R}^d$ be compact. A map $K : D \times D \to \mathbb{R}$ which is continuous, symmetric, and positive-definite, meaning that

$$\int_{D\times D} K\left(x,\, \tilde{x}\right) f\left(x\right) f\left(\tilde{x}\right) \geqslant 0$$

for every square-integrable function f over D, is called a *kernel.*

Remark 22.116 (Overloading the term "kernel")**.** We already defined kernels in Definition 20.26 when discussing nonparametric regression in Chapter 20. Should both a kernel in the sense of Definition 20.26 and a kernel in the sense of Definition 22.115 need to coexist, we may refer to the former as a *regression* kernel and to the latter as a *classification* kernel.

Example 22.117 (Kernels)**.** Here are some kernels commonly used in classification.

(1) The *polynomial* kernel of *degree* r (a positive integer) and with *coefficient* $a > 0$ is

$$K\left(x,\, \tilde{x}\right) := \left(a + x \cdot \tilde{x}\right)^r$$

for every $x,\, \tilde{x} \in \mathbb{R}^d$.

(2) The *Gaussian*, *radial basis function*, or *RBF* kernel with *covariance* $\sigma^2 > 0$, or equivalently *parameter* $\gamma := \frac{1}{2\sigma^2} > 0$, is

$$K\left(x,\, \tilde{x}\right) := \exp\left(-\frac{\left|x - \tilde{x}\right|^2}{2\sigma^2}\right) = \exp\left(-\gamma\left|x - \tilde{x}\right|^2\right)$$

for every $x,\, \tilde{x} \in \mathbb{R}^d$.

(3) The *sigmoid*, or *neural network* kernel with *parameters* a, $b > 0$ is

$$K(x,\,\tilde{x}) := \tanh(ax \cdot \tilde{x} + b)$$

for every x, $\tilde{x} \in \mathbb{R}^d$.

Theorem 22.118 (Mercer)**.** *Let $D \subseteq \mathbb{R}^n$ and let K be a kernel on D. Define the operator $T_K : L_2(D) \to L_2(D)$ via*

$$(T_K f)(x) := \int_D K(x,\,\tilde{x})\, f(\tilde{x})\, d\tilde{x}$$

for every $f \in L_2(D)$ and every $x \in D$. There exists an orthonormal basis $(\psi_j)_{j=1}^{\infty}$ of $L_2(D)$ of eigenfunctions of T_K with non-negative eigenvalues $(\lambda_j)_{j=1}^{\infty}$, i.e.

$$T_K \psi_j = \lambda_j \psi_j,$$

such that eigenfunctions with non-zero eigenvalues are continuous and

$$K(x,\,\tilde{x}) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x)\, \psi_j(\tilde{x})$$

for every x, $\tilde{x} \in D$, where the convergence of the series is absolute and uniform. In particular if we define $\phi : D \to l^2$, where l^2 denotes the space of square-summable sequences, via

$$\phi(x) := \left(\sqrt{\lambda_j}\psi_j(x)\right)_{j=1}^{\infty}$$

for every $x \in D$ then

$$K(x,\,\tilde{x}) = \langle \phi(x),\, \phi(\tilde{x}) \rangle_{l^2}$$

for every x, $\tilde{x} \in D$, i.e, the kernel K necessarily arises from a feature map ϕ.

Definition 22.119 (Multi-index notation)**.** Any element $p \in \mathbb{N}^d$, $d \geqslant 2$, is called a *multi-index*, where here we follow the convention that zero is a natural number.

- The *length* of p is defined to be $|p| := \sum_{i=1}^{n} p_i$.
- We define $p! := p_1! \dots p_d! = \prod_{i=1}^{d} p_i!$.
- For $k = |p|$ we write $\binom{k}{p}$ for the multinomial coefficient

$$\binom{k}{p} := \binom{k}{p_1,\,\dots,\,p_d} = \frac{k!}{p!} = \frac{(p_1 + \cdots + p_d)!}{p_1! \dots p_d!}$$

- For any $x \in \mathbb{R}^d$ we write x^p for the *monomial*

$$x^p := \prod_{i=1}^{d} x_i^{p_i}.$$

Definition 22.120 (Monomial basis)**.** Let $r \geqslant 0$ be an integer. The basis of monomials

$$\left\{x^p : p \in \mathbb{N}^d \text{ and } |p| \leqslant r\right\},$$

where p is a multi-index, of the space of polynomials over $x \in \mathbb{R}^d$ of degree at most r is called the *monomial basis* of that space.

Remark 22.121 (Applications of Mercer's Theorem)**.** For sufficiently simple kernels, such as the polynomial kernel and the Gaussian kernel, we can write down explicitly the induced feature map whose existence is guaranteed by Theorem 22.118.

(1) For the polynomial kernel

$$K\left(x,\,\tilde{x}\right)=\left(1+x\cdot\tilde{x}\right)^{r}$$

with $x,\,\tilde{x}\in\mathbb{R}^{d}$ the corresponding feature map is

$$\phi(x)=\left(\binom{r}{r-|p|,\,p}\,x^{p}:p\in\mathbb{N}^{d}\text{ and }|p|\leqslant r\right)$$

where p is a multi-index and, for $l=r-|p|$,

$$\binom{r}{r-|p|,\,p}=\binom{r}{l,\,p_1,\,\ldots,\,p_d}=\frac{r!}{l!p_1!\ldots p_d!}$$

is a multinomial coefficient.

For example if $d=r=2$ then

$$\phi(x)=\left(1,\,\sqrt{2}x_1,\,\sqrt{2}x_2,\,x_1^2,\,\sqrt{2}x_1x_2,\,x_2^2\right)$$

and if $d=2$ and $r=3$ then $\phi(x)$ is

$$\left(1,\,\sqrt{3}x_1,\,\sqrt{3}x_2,\,\sqrt{3}x_1^2,\,\sqrt{6}x_1x_2,\,\sqrt{3}x_2^2,\,x_1^3,\,\sqrt{3}x_1^2x_2,\,\sqrt{3}x_1x_2^2,\,x_2^3\right)$$

These results are proved in Exercise A.23.45–A.23.47.

In particular this motivates calling such a kernel *"polynomial"*: the corresponding feature map is (up to scaling) precisely the monomial basis of the space of polynomials of degree at most r.

Note that this means that the decision boundary of a linear classifier in the codomain of ϕ corresponds to a polynomial curve of degree at most r in the original domain $\mathbb{R}^d$.

(2) For the Gaussian kernel with unit variance

$$K\left(x,\,\tilde{x}\right)=\exp\left(-\frac{|x-\tilde{x}|^2}{2}\right)$$

with $x,\,\tilde{x}\in\mathbb{R}^{d}$ the corresponding feature map is $\phi:\mathbb{R}^d\to l^2\left(\mathbb{N}^d\right)$ given by

$$\phi(x)=\left(\frac{x^p}{p!}e^{-\frac{|x|^2}{2}}:p\in\mathbb{N}^d\right)$$

where p is a multi-index. This is proved in Exercise A.23.48 (see also Remark A.8).

Since *every* monomial x^p, $p\in\mathbb{N}^d$, appears in the feature map we sometimes consider intuitively the Gaussian kernel to be akin to a polynomial kernel with degree $r=\infty$.

Remark 22.122 (Kernelization of Fisher's linear classification estimator)**.** By Exercise A.23.36, if there are *two* classes then the Fisher discriminant ratio of $v\cdot X$ may be written, for any $v\in\mathbb{R}^d$, as

$$\begin{aligned}R(v)&:=\frac{\mathbb{V}\mathbb{E}\left(X\,|\,Y\right)v\cdot v}{\mathbb{E}\mathbb{V}\left(X\,|\,Y\right)v\cdot v}\\&=\frac{\left[\pi_0\mu_0\otimes\mu_0+\pi_1\mu_1\otimes\mu_1-\left(\pi_0\mu_0+\pi_1\mu_1\right)\otimes\left(\pi_0\mu_0+\pi_1\mu_1\right)\right]v\cdot v}{\left(\pi_0\Sigma_0+\pi_1\Sigma_1\right)v\cdot v}.\end{aligned}$$

In particular if $\pi_0 = \pi_1 = \frac{1}{2}$ then the solution of Exercise A.23.37 tells us that the ratio simplifies to

$$R(v) = \frac{1}{2} \cdot \frac{(\mu_0 - \mu_1) \otimes (\mu_0 - \mu_1)\, v \cdot v}{(\Sigma_0 + \Sigma_1)\, v \cdot v},$$

which may be estimated by (up to a factor of a half)

$$\frac{\left(\overline{X}_0 - \overline{X}_1\right) \otimes \left(\overline{X}_0 - \overline{X}_1\right) v \cdot v}{(S_0 + S_1)\, v \cdot v}$$

where $\overline{X}_j$ and S_j denote the sample mean and sample variance of the samples

$$\{X_i : Y_i = j\},$$

respectively.

In particular, since $\pi_0 = \pi_1$ we expect that

$$n_0 := \sum_{i=1}^{n} \mathbb{1}\,(Y_j = 0) \approx \sum_{i=1}^{n} \mathbb{1}\,(Y_j = 0) =: n_1$$

and hence

$$S_0 + S_1 \approx \frac{1}{n_0 - 1}\left[(n_0 - 1)\, S_0 + (n_1 - 1)\, S_1\right].$$

Since the prefactor of $\frac{1}{n_0-1}$ is irrelevant when maximizing the ratio we may drop it and hence use

$$\widetilde{S}_j := (n_j - 1)\, S_j = \sum_{i=1}^{n} \left(X_i - \overline{X}_j\right) \otimes \left(X_i - \overline{X}_j\right) \mathbb{1}\,(Y_i = j)$$

instead of S_j. In summary we seek to maximize

$$\hat{R}(v) := \frac{\left(\overline{X}_0 - \overline{X}_1\right) \otimes \left(\overline{X}_0 - \overline{X}_1\right) v \cdot v}{\left(\widetilde{S}_0 + \widetilde{S}_1\right) v \cdot v}$$

over $v \in \mathbb{R}^d$.

In particular: in order to use the kernel trick (see Remark 22.114) we must write the ratio $\hat{R}$ solely in terms of *dot products* between the covariate samples $X_1, \ldots, X_n$. A key step in that direction comes from Exercise A.23.49 which tells us that the minimizer w of $\hat{R}$ is a linear combination of the covariate samples $X_1, \ldots, X_n$, i.e.

$$w = \sum_{i=1}^{n} \alpha_i X_i$$

for some $\alpha \in \mathbb{R}^n$. We now wish to write $\hat{R}(w)$ in terms of α, which means writing

$$\overline{X}_j \cdot w \text{ and } \widetilde{S} w \cdot w$$

in terms of α.

We compute that

$$\overline{X}_j \cdot w = \sum_{i=1}^{n} \alpha_i X_i \cdot \overline{X}_j$$

while

$$\widetilde{S}_j w \cdot w = \sum_{i,k=1}^{n} \alpha_i \alpha_k \widetilde{S}_j X_i \cdot X_k$$

and so we define the vectors $M_j \in \mathbb{R}^n$ and the n-by-n matrices N_j by

$$\left(M_j\right)_i := X_i \cdot \overline{X}_j \text{ and } \left(N_j\right)_{ik} := \widetilde{S}_j X_i \cdot X_k$$

such that

$$\overline{X}_j \cdot w = M_j \cdot \alpha \text{ and } \widetilde{S}_j w \cdot w = N_j \alpha \cdot \alpha.$$

To write M_j and N_j explicitly in terms of dot products between the covariate samples we introduce the *non-symmetric* n-by-n matrices K_j defined by

$$\left(K_j\right)_{ik} := \left(X_i \cdot X_k\right) \mathbb{1}\left(Y_k = j\right).$$

We then compute that

$$\left(M_j\right)_i = X_i \cdot \overline{X}_j = \frac{1}{n_j} \sum_{k=1}^{n} \left(X_i \cdot X_k\right) \mathbb{1}\left(Y_k = j\right) = \frac{1}{n_j} \sum_{k=1}^{n} \left(K_j\right)_{ik} = \frac{1}{n_j} \left(K_j \mathbb{1}\right)_i$$

where $\mathbb{1} = (1, \ldots, 1) \in \mathbb{R}^n$, i.e.

$$M_j = \frac{1}{n_j} K_j \mathbb{1}.$$

We may also compute that

$$\begin{aligned}
\left(N_j\right)_{ik} = \widetilde{S}_j X_i \cdot X_k &= \sum_{l=1}^{n} \left(X_l - \overline{X}_j\right) \otimes \left(X_l - \overline{X}_j\right) X_i \cdot X_k \; \mathbb{1}\left(Y_l = j\right) \\
&= \sum_{l=1}^{n} \Big[\left(X_l \cdot X_i\right)\left(X_l \cdot X_k\right) - \left(X_l \cdot X_i\right)\left(\overline{X}_j \cdot X_k\right) \\
&\qquad - \left(\overline{X}_j \cdot X_i\right)\left(X_l \cdot X_k\right) + \left(\overline{X}_j \cdot X_i\right)\left(\overline{X}_j \cdot X_k\right) \Big] \mathbb{1}\left(Y_l = j\right) \\
&= \sum_{l=1}^{n} \left(X_l \cdot X_i\right)\left(X_l \cdot X_k\right) \mathbb{1}\left(Y_l = j\right) - n_j \left(\overline{X}_j \cdot X_i\right)\left(\overline{X}_j \cdot X_k\right) \\
&= \sum_{l=1}^{n} (K_j)_{il} (K_j)_{kl} - n_j (M_j)_i (M_j)_k,
\end{aligned}$$

i.e.

$$N_j = K_j K_j^T - n_j M_j \otimes M_j.$$

In particular since $M_j = \frac{1}{n_j} K_j \mathbb{1}$ we may deduce that

$$n_j M_j \otimes M_j = \frac{1}{n_j} \left(K_j \mathbb{1}\right) \otimes \left(K_j \otimes \mathbb{1}\right) = \frac{1}{n_j} K_j \left(\mathbb{1} \otimes \mathbb{1}\right) K_j^T$$

and so we may rewrite N_j as

$$N_j = K_j \left(I - \frac{1}{n_j} \mathbb{1} \otimes \mathbb{1} \right) K_j^T.$$

In conclusion if we define the n-by-n matrices

$$M := \left(M_0 - M_1\right) \otimes \left(M_0 - M_1\right) \text{ and } N := N_0 + N_1$$

then we may write $\hat{R}$ as follows:

$$\hat{R}(w) = \frac{\left(\overline{X}_0 - \overline{X}_1\right) \otimes \left(\overline{X}_0 - \overline{X}_1\right) w \cdot w}{\left(\widetilde{S}_0 + \widetilde{S}_1\right) w \cdot w} = \frac{\left(\overline{X}_0 \cdot w - \overline{X}_1 \cdot w\right)^2}{\widetilde{S}_0 w \cdot w + \widetilde{S}_1 w \cdot w}$$
$$= \frac{\left(M_0 \cdot \alpha - M_1 \cdot \alpha\right)^2}{N_0 \alpha \cdot \alpha + N_1 \alpha \cdot \alpha} = \frac{\left(M_0 - M_1\right) \otimes \left(M_0 - M_1\right) \alpha \cdot \alpha}{\left(N_0 + N_1\right) \alpha \cdot \alpha} = \frac{M\alpha \cdot \alpha}{N\alpha \cdot \alpha}$$

over $\alpha \in \mathbb{R}^n$. Crucially: M and N are fully determined by $\left(K_j\right)_{j=0,1}$ and hence fully determined by the dot products between the covariate samples.

Given a kernel K we would then define

$$\left(\mathbb{K}_j\right)_{ik} := K\left(X_i,\, X_k\right) \mathbb{1}\left(Y_k = j\right)$$

as well as

$$\mathbb{M} := \left(\mathbb{M}_0 - \mathbb{M}_1\right) \otimes \left(\mathbb{M}_0 - \mathbb{M}_1\right) \text{ for } \mathbb{M}_j := \frac{1}{n_j} \mathbb{K}_j \mathbb{1}$$

and

$$\mathbb{N} := \mathbb{N}_0 + \mathbb{N}_1 \text{ for } \mathbb{N}_j := \mathbb{K}_j \left(I - \frac{1}{n_j} \mathbb{1} \otimes \mathbb{1}\right) \mathbb{K}_j^T.$$

The *kernelized Fisher's linear classification estimator* would then be

$$\hat{h}(x) := \mathbb{1}\left(w \cdot x > s\right)$$

for

$$w = \sum_{i=1}^{n} \alpha_i^* X_i \text{ where } \alpha^* = \underset{\alpha \in \mathbb{R}^n}{\arg\max} \frac{\mathbb{M}\alpha \cdot \alpha}{\mathbb{N}\alpha \cdot \alpha}$$

and for an appropriately chosen scalar $s \in \mathbb{R}$ (say, chosen to minimize the *empirical error rate*).

Remark 22.123 (Kernelization of OSH classifiers and SVC: Support Vector Machines)**.** Applying the kernel trick (see Remark 22.114) to OSH classifiers and to SVC is straightforward since their constructions are already characterized by dot products between covariate samples.

So let K be a kernel. The *kernelized OSH classification estimator* is constructed by solving

$$\max_{\lambda \in \mathbb{R}^n} \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,k=1}^{n} Y_i Y_k K\left(X_i,\, X_k\right) \lambda_i \lambda_k$$
$$\text{subject to } \lambda_i \geqslant 0 \text{ for all } i \text{ and } \sum_{i=1}^{n} Y_i \lambda_i = 0$$

and recovering $a_0 \in \mathbb{R}$ from

$$Y_i \left(a_0 + \sum_{k=1}^{n} Y_k \lambda_k K\left(X_k,\, X_i\right)\right) = 1 \text{ for any } i \text{ where } \lambda_i > 0$$

such that the classification estimator is

$$\hat{h}(x) := \mathbb{1}\left(a_0 + \sum_{i=1}^{n} Y_i \lambda_i K\left(X_i,\, x\right) > 0\right).$$

Recall that we expect a small number of the λ_i's to be non-zero, and so this sum only involves a small number of terms. (See Definition 22.106 and Corollary 22.104 for the version of this involving only dot products.)

Similarly, using now Remark 22.113 and Definition 22.109 for inspiration, the *kernelized SVC*, also known as a *Support Vector Machine*, or *SVM*, is constructed by fixing a *parameter* $C > 0$, solving

$$\begin{aligned} &\max_{\lambda \in \mathbb{R}^n} \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i,k=1}^n Y_i Y_k K\left(X_i,\, X_k\right) \lambda_i \lambda_k \\ &\text{subject to } 0 \leqslant \lambda_i \leqslant C \text{ for all } i \text{ and } \sum_{i=1}^n Y_i \lambda_i = 0, \end{aligned}$$

and recovering $a_0 \in \mathbb{R}$ from

$$Y_i \left(a_0 + \sum_{k=1}^n Y_k \lambda_k K\left(X_k,\, X_i\right) \right) = 1 \text{ for any } i \text{ where } 0 < \lambda_i < C$$

(see Chapter 12 of [HTF09]) such that the classification estimator is

$$\hat{h}(x) := \mathbb{1}\left(a_0 + \sum_{i=1}^n Y_i \lambda_i K\left(X_i,\, x\right) > 0 \right).$$

Remark 22.124 (Kernelization of logistic regression classification)**.** We seek to apply the kernel trick (see Remark 22.114) to the logistic regression classification estimator. Recall that if $(Y,\, (1,\, X))$ has a logistic regression distribution with parameter $(\beta_0,\, \beta) \in \mathbb{R}^{1+d}$ and if $(Y_1,\, X_1), \ldots, (Y_n,\, X_n)$ is an IID sample drawn from $(Y,\, X)$ then the log-likelihood function is, up to terms independent of $(\beta_0,\, \beta)$ which we discard as they appear,

$$\begin{aligned} l\left(\beta_0,\, \beta\right) &= \sum_{i=1}^n \log f\left(X_i,\, Y_i; \beta_0,\, \beta\right) \\ &= \sum_{i=1}^n \log f\left(Y_i \,|\, X_i; \beta_0,\, \beta\right) + \log f\left(X_i\right) \\ &= \sum_{i=1}^n \log \left[p\left(X_i; \beta_0,\, \beta\right)^{Y_i} \left(1 - p\left(X_i; \beta_0,\, \beta\right)\right)^{1-Y_i} \right] \\ &= \sum_{i=1}^n Y_i \log p\left(X_i; \beta_0,\, \beta\right) + \left(1 - Y_i\right) \log\left(1 - p\left(X_i; \beta_0,\, \beta\right)\right) \end{aligned}$$

where

$$p\left(X_i; \beta_0,\, \beta\right) = \frac{e^{\gamma_i}}{1 + e^{\gamma_i}} \iff 1 - p\left(X_i; \beta_0,\, \beta\right) = \frac{1}{1 + e^{\gamma_i}}$$

for $\gamma_i := \beta_0 + \beta \cdot X_i$ and so

$$\begin{aligned} l(\beta_0, \beta) &= \sum_{i=1}^{n} Y_i \log \frac{e^{\gamma_i}}{1+e^{\gamma_i}} + (1 - Y_i) \log \frac{1}{1+e^{\gamma_i}} \\ &= \sum_{i=1}^{n} \gamma_i Y_i - Y_i \log(1+e^{\gamma_i}) - (1-Y_i)\log(1+e^{\gamma_i}) \\ &= \sum_{i=1}^{n} \gamma_i Y_i - \log(1+e^{\gamma_i}) \\ &= \sum_{i=1}^{n} (\beta_0 + \beta \cdot X_i) Y_i - \log[1 + \exp(\beta_0 + \beta \cdot X_i)]. \end{aligned}$$

Crucially: for

$$L(y, s) := -(\beta_0 + s) y + \log[1 + \exp(\beta_0 + s)]$$

and $r \equiv 0$ we see that the Representer Theorem of Exercise A.23.50 applies to the maximization of the log-likelihood function and so the MLE β is a linear combination of the covariate samples, i.e.

$$\beta = \sum_{i=1}^{n} \alpha_i X_i$$

for some $\alpha \in \mathbb{R}^n$. In particular if we write

$$\mathbb{K}_{ij} := X_i \cdot X_j$$

then

$$\beta \cdot X_i = \left(\sum_{j=1}^{n} \alpha_j X_j \right) \cdot X_i = (\mathbb{K}\alpha)_i$$

and so the log-likelihood may be written as

$$l(\beta_0, \beta) = \sum_{i=1}^{n} [\beta_0 + (\mathbb{K}\alpha)_i] - \log[1 + \exp(\beta_0 + (\mathbb{K}\alpha)_i)].$$

The optimal $\alpha \in \mathbb{R}^n$ may then be found iteratively by (see [ZH05])

$$\alpha_0 := 0 \text{ and } \alpha_{k+1} := \left(\mathbb{K}^T \mathbb{W} \mathbb{K}\right)^{-1} \mathbb{K}^T \mathbb{W} \mathbb{Z}\big|_{\alpha=\hat{\alpha}_k}$$

where

$$\mathbb{Z} := \mathbb{K}\alpha + \mathbb{W}^{-1}(\mathbb{Y} - \mathfrak{p})$$

for $\mathbb{W}$, $\mathbb{Y}$, and $\mathfrak{p}$ as in Theorem 13.86. In particular this is the *same* iteration as that of Theorem 13.86 (where we compute the MLE for the standard logistic regression model) except that now we use $\mathbb{K}$ instead of the design matrix $\mathbb{X}$.

This continues to hold if we replace $\mathbb{K}$ with

$$\mathbb{K}_{ij} := K(X_i, X_j)$$

for *any* kernel K. In other words: to kernelize the logistic regression classification estimator we simply apply the standard logistic regression classification estimator to $\mathbb{K}$ instead of $\mathbb{X}$. (See the jupyter notebook “Chapter 22 – Bonus 02” for a practical implementation of this idea.)

Remark 22.125 (On empirical risk minimization versus regression)**.** In the terminology of Remark 22.13, OSH classifiers, SVC, and SVM are all *empirical risk minimization* classifiers. Given a point $x \in \mathbb{R}^d$ they only label x as $y = 0$ or $y = 1$, but do not provide any information as to how *likely* either option is.

This is where *regression* classifiers come into play. Such classifiers, such as the logistic regression classification estimator or its kernelized cousin (see Remark 22.124), tell us not only which class the covariate should be labelled as, but also how likely it is to be in one or the other class. This does not come for free and there are typically trade-offs: for example SVM may be computed fast and require little memory to store since they are *sparse* and only require the *support vectors* to be stored.

22.11. **Other Classifiers.**

Definition 22.126 (K-nearest neighborhood)**.** Let $D \subseteq \mathbb{R}^d$ be a finite set, consider an integer $K \geqslant 1$, and let $x \in \mathbb{R}^d$. We say that $N \subseteq D$ is a *K-nearest neighborhood of x with respect to D* if

1. N has K elements and
2. $d(x, x_{in}) \leqslant d(x, x_{out})$ for every $x_{in} \in N$ and $x_{out} \in D \setminus N$.

We denote by $\mathcal{N}(x; D)$ the set of K-nearest neighborhoods of x with respect to D.

Remark 22.127 (K-nearest neighborhood)**.** Suppose that $x \in \mathbb{R}^d$ and $D \subseteq \mathbb{R}^d$ is a finite set such that

$$d(x, a) \neq d(x, b)$$

for every distinct $a, b \in D$. Then $N \subseteq D$ is a K-nearest neighborhood of x with respect to D if and only if N contains the K points in D *closest* to x.

Otherwise if the condition above is not satisfied then x may have several *distinct* K-nearest neighborhoods.

Definition 22.128 (K-nearest neighbors classification estimator)**.** Let Y be a random variable with finite codomain $\mathcal{Y}$, let X be a random vector in $\mathbb{R}^d$, let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from (Y, X), and let $K \geqslant 1$ be an integer. For any $x \in \mathbb{R}^d$ the *K-nearest neighbors classification estimator* $\hat{h}(x)$ is constructed as follows.

1. Choose a K-nearest neighborhood $N \in \mathcal{N}(x; X_1, \ldots, X_n)$ uniformly at random.
2. Majority vote: let $\mathcal{Y}_N \subseteq \mathcal{Y}$ be the subset of $\mathcal{Y}$ whose elements are the most frequent in $\{Y_i : X_i \in N\}$.
3. Break ties randomly: choose $y \in \mathcal{Y}_N$ uniformly at random.

Finally set $\hat{h}(x) := y$.

Remark 22.129 (K-nearest neighbors classification estimator)**.** The construction of $\mathcal{Y}_N$ in Definition 22.128 above will ideally produce a singleton. However, if for example

$$\{Y_i : X_i \in N\} = \{0, 0, 1, 1, 2, 3\},$$

this will not be the case since then $\mathcal{Y}_N = \{0, 1\}$. The K-nearest neighbors classification estimator would thus take the value 0 or 1, uniformly at random, at that point.

Definition 22.130 (Bagging)**.** Let Y be a binary random variable, let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$, let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from (Y, X), and let $\mathfrak{h}_n$ be a classification method for Y given X. We define the *bagged classification estimator* h^* based on $\mathfrak{h}_n$ with parameter $B \geqslant 2$, an integer, as follows:

$$h^* := \mathbb{1}\left(\frac{1}{B}\sum_{j=1}^{B}\mathfrak{h}_n\left(\left(Y^*_{1,j},\, X^*_{1,j}\right),\, \ldots,\, \left(Y^*_{n,j},\, X^*_{n,j}\right)\right) > \frac{1}{2}\right)$$

where, for $j = 1, \ldots, B$,

$$\left(Y^*_{1,j},\, X^*_{1,j}\right),\, \ldots,\, \left(Y^*_{n,j},\, X^*_{n,j}\right)$$

is an IID sample drawn from the empirical CDF of the original samples

$$\left(Y_1,\, X_1\right),\, \ldots,\, \left(Y_n,\, X_n\right),$$

i.e. they are *bootstrap samples* (see Chapter 8).

Definition 22.131 (Boosting)**.** Let Y be a random variable with codomain $\{-1, +1\}$, let X be a random vector with codomain $\mathcal{X} \subseteq \mathbb{R}^d$, let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an IID sample drawn from (Y, X), and let $\mathfrak{h}_n$ be a classification method for Y given X which *accepts samples weights.* We define the *boosted classification estimator* h^* based on $\mathfrak{h}_n$ with parameter $J \geqslant 2$, an integer, as follows.

1. Set the *sample weights* to be $w_i := \frac{1}{n}$ for $i = 1, \ldots, n$.
2. For $j = 1, \ldots, J$, do the following.
   1. (a) Let h_j be the classification estimator built from $\mathfrak{h}_n$ applied to the original sample with weights $w_1, \ldots, w_n$.
   2. (b) Define the *weighted empirical error rate*
   $$\widehat{L}_j := \frac{\sum_{i=1}^{n} w_i \mathbb{1}\left(h_j(X_i) \neq Y_i\right)}{\sum_{i=1}^{n} w_i}.$$
   3. (c) Define the *classifier weights*
   $$\alpha_j := \log\frac{1-\widehat{L}_j}{\widehat{L}_j},$$
   noting that $\alpha_j \in \mathbb{R}$ since $\widehat{L}_j \in (0, 1)$ with $\alpha_j \uparrow +\infty$ as $\widehat{L}_j \downarrow 0$ and $\alpha_j \downarrow -\infty$ as $\widehat{L}_j \uparrow 1$.
   4. (d) Update the sample weights via
   $$w_i := w_i \exp\left[\alpha_j \mathbb{1}\left(h_j(X_i) \neq Y_i\right)\right],$$
   such that w_j *only* changes if the i-th sample is misclassified by h_j (in which case it will typically increase since $\alpha_j > 0$ whenever $\widehat{L}_j < 1/2$).
3. Finally define
$$h^* := \mathbb{1}\left(\sum_{j=1}^{J}\alpha_j h_j > 0\right)$$

Remark 22.132 (AdaBoost)**.** The algorithm provided in Definition 22.131 above is one of many for boosting a classification method known as *AdaBoost* (short for *Adaptive Boosting*).

Remark 22.133 (Classification methods accepting weights)**.** In Definition 22.131 we ask for the classification method under consideration to *accept sample weights.* This means that the classification methods must take into account weights $w_i \geqslant 0$, one for each sample. The classification method will then be more likely to classify samples with higher weights, but the details of how that is implemented vary from one classification method to the next and we omit the details here.

Remark 22.134 (Bagging and boosting)**.** Heuristically we can view bagging as a *variance–reduction* procedure and boosting as a *bias–reduction* procedure.

Appendix A. Exercises

A.1. **Probability.**

Exercise A.1.1 (Continuity of probability measures)**.** Prove the following statements (known as the *continuity of probability measures*).

(1) Suppose that A_n is a monotone increasing sequence of sets and let us define $A_\infty = \bigcup_n A_n$. Prove that, for any probability measure $\mathbb{P}$, $\mathbb{P}(A_n) \to \mathbb{P}(A_\infty)$ as $n \to \infty$.

(2) Prove that the result above also holds for monotone *decreasing* sequences, provided the union is replaced with an intersection.

Solution. First we prove (1). We construct the sequence of sets B_i as follows:

$$B_1 = A_1,\, B_2 = A_2 \setminus A_1,\, \dots,\, B_n = A_n \setminus A_{n-1}.$$

This produces a sequence of pairwise disjoint sets B_i such that $A_n = \bigcup_{i=1}^n B_i$, and so $A_\infty = \bigcup_{i=1}^\infty B_i$. Therefore, by countable additivity of any probability measure $\mathbb{P}$ we deduce that

$$\mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{i=1}^n B_i\right) = \sum_{i=1}^n \mathbb{P}(B_i) \to \sum_{i=1}^\infty \mathbb{P}(B_i) = \mathbb{P}\left(\bigcup_{i=1}^\infty B_i\right),$$

thus proving the claim.

We then conclude that (2) follows from (1) and the fact that, if A_n is decreasing then A_n^c is increasing. Indeed, since $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ for any probability measure $\mathbb{P}$ and any event A, we conclude that if the sequence A_n is decreasing then

$$\mathbb{P}(A_n) = 1 - \mathbb{P}(A_n^c) \to 1 - \mathbb{P}\left(\bigcup_{n=1}^\infty A_n^c\right) = \mathbb{P}\left(\bigcap_{n=1}^\infty A_n\right)$$

as desired.

Exercise A.1.2 (Elementary properties of probability measures)**.** Let $\mathbb{P}$ be a probability measure over a sample space Ω and let A and B be events. Prove that

(1) $\mathbb{P}(\emptyset) = 0$,
(2) if $A \subseteq B$ then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$,
(3) $0 \leqslant \mathbb{P}(A) \leqslant 1$,
(4) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, and
(5) if $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Solution. We proceed in order.

(1) This follows from countable additivity: $\mathbb{P}(\Omega) = \mathbb{P}(\Omega) + \mathbb{P}(\emptyset)$ and hence $\mathbb{P}(\emptyset) = 0$.

(2) We may write B as a union of disjoint sets, namely $B = A \cup (B \setminus A)$, from which it follows from countable additivity and non-negativity that

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geqslant \mathbb{P}(A).$$

(3) This follows immediately from (1) and (2) since $\emptyset \subseteq A \subseteq \Omega$.

(4) This follows immediately from countable additivity since it suffices to rearrange the identity $1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c)$.

(5) This follows immediately from countable additivity.

Exercise A.1.3. Let Ω be a sample space and let $A_1, A_2, \dots$ be a countable sequence of events. Define $B_n = \bigcup_{i=n}^\infty A_i$ and $C_n = \bigcap_{i=n}^\infty A_i$.

(1) Show that B_n is a monotone decreasing sequence of sets and that C_n is a monotone increasing sequence of sets.
(2) Show that $\omega \in \bigcap_{n=1}^{\infty} B_n$ if and only if ω belongs to infinitely many of the events A_1, A_2, ...
(3) Show that $\omega \in \bigcup_{n=1}^{\infty} C_n$ if and only if ω belongs to all but finitely many of the events A_1, A_2, ...

Solution. We proceed in order.

(1) The sequence B_n is monotone decreasing since it satisfies

$$B_n = \bigcup_{i=n}^{\infty} A_i = A_n \cup \bigcup_{i=n+1}^{\infty} A_i = A_n \cup B_{n+1}$$

and hence $B_{n+1} \subseteq B_n$. Similarly the sequence C_n is monotone increasing since it satisfies

$$C_{n+1} = \bigcap_{i=n+1}^{\infty} A_i = A_{n+1} \cap \bigcap_{i=n}^{\infty} A_i = A_{n+1} \cap C_n$$

and hence $C_n \subseteq C_{n+1}$.

(2) This follows from the following chain of equivalences:

$$\begin{aligned}
&\omega \text{ belongs to } \textit{finitely} \text{ many } A_i\text{'s} \\
&\Leftrightarrow \exists I : \forall i \geqslant I,\, \omega \notin A_i \\
&\Leftrightarrow \exists I : \forall i \geqslant I,\, \omega \in A_i^c \\
&\Leftrightarrow \exists I : \omega \in \bigcap_{i \geqslant I} A_i^c \\
&\Leftrightarrow \omega \in \bigcup_{I=1}^{\infty} \bigcap_{i \geqslant I} A_i^c \\
&\Leftrightarrow \omega \notin \bigcap_{I=1}^{\infty} \bigcup_{i \geqslant I} A_i = \bigcap_{n=1}^{\infty} B_n.
\end{aligned}$$

(3) As in (2), this follows from a simple chain of equivalences:

$$\begin{aligned}
&\omega \text{ belongs to } \textit{all but finitely many } A_i\text{'s} \\
&\Leftrightarrow \exists I : \forall i \geqslant I,\, \omega \in A_i \\
&\Leftrightarrow \omega \in \bigcup_{I=1}^{\infty} \bigcap_{i \geqslant I} A_i = \bigcup_{n=1}^{\infty} C_n.
\end{aligned}$$

Exercise A.1.4. Let $(A_i)_{i \in I}$ be an arbitrary collection of sets (this collection may be uncountable, so these sets should not be thought of as events). Prove that $\left(\bigcup_{i \in I} A_i\right)^c = \bigcap_{i \in I} A_i^c$ and $\left(\bigcap_{i \in I} A_i\right)^c = \bigcup_{i \in I} A_i^c$.

Solution. This is precisely De Morgan's Laws.

Exercise A.1.5 (Coin tosses 1)**.** Suppose we toss a fair coin until we get exactly two heads. Describe the sample space S. What is the probability that exactly k tosses are required?

Solution. The sample space S is the set of finite strings on $\{H, T\}$ ending in H and containing exactly two H's. We can also view it as $S = S' \oplus \{H\}$ where S' is the set of finite strings containing exactly *one* H. In other words, a string in S can be characterised as $s = s'H$ for a string $s' \in S'$.

Then the probability that k tosses are needed may be computed as follows:

$$\begin{aligned}\mathbb{P}\,(k \text{ tosses}) &= \mathbb{P}\left([\text{a permutation of } H\underbrace{T\cdots T}_{k-2}]H\right)\\ &= (k-1)\mathbb{P}\left(H\underbrace{T\cdots T}_{k-2}H\right)\\ &= \frac{k-1}{2^k}.\end{aligned}$$

Exercise A.1.6 (Impossibility of a uniform distribution on the naturals)**.** Let $\Omega = \mathbb{N} = \{0, 1, 2, dots\}$, i.e. let us take the natural numbers to be our sample space. Prove that there does not exist a uniform distribution on Ω, i.e. that there does exist a probability measure $\mathbb{P}$ such that $\mathbb{P}(A) = \mathbb{P}(B)$ whenever $|A| = |B|$.

Solution. We suppose for the sake of contradiction that such a probability measure $\mathbb{P}$ exists. Let us denote $p := \mathbb{P}(0)$. Then, for any natural number n, $\mathbb{P}(n) = p$. Therefore, by countable additivity, we have that

$$1 = \mathbb{P}(\mathbb{N}) = \sum_{n=1}^{\infty} \mathbb{P}(n) = \sum_{n_1}^{\infty} p = \infty,$$

which is a contradiction.

Exercise A.1.7 (Subadditivity of probability measures)**.** Let $A_1, A_2, \ldots$ be a countable sequence of events. Show that $\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leqslant \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.

Solution. We construct the sets B_n via

$$B_1 = A_1,\, B_2 = A_2 \setminus A_2,\, B_3 = A_3 \setminus (A_1 \cup A_2),\, \ldots,\, B_n = A_n \setminus \left(\bigcup_{i=1}^{n-1} A_i\right),\, \ldots$$

such that the countable sequence of events B_n is pairwise disjoint and satisfies $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$. Therefore it follows from countably additivity that

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(B_n).$$

Since moreover $B_n \subseteq A_n$ for every n we conclude that indeed

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(B_n). \leqslant \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Exercise A.1.8. Let A_i be a countable sequence of events for which $\mathbb{P}(A_i) = 1$ for all i. Prove that $\mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = 1$.

Solution. This follows from item (4) of Exercise A.1.2 and from sub-additivity as recorded in Exercise A.1.7:

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \geqslant 1 - \sum_{i=1}^{\infty} \mathbb{P}(A_i) = 1 - 0 = 1.$$

Exercise A.1.9 (Conditional probability is a probability measure)**.** For a fixed event B with $\mathbb{P}(B) > 0$, show that $\mathbb{P}(\,\cdot\,|B)$ is a probability measure.

Solution. Non-negativity is immediate since the original probability measure $\mathbb{P}$ is non-negative. Moreover we have that

$$\mathbb{P}(\Omega|B) = \frac{\mathbb{P}(B\Omega)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$$

as desired. Finally, if A_1, A_2, $\ldots$ is a countable sequence of pairwise disjoint events then the events $B \cap A_i$ are also pairwise disjoint and so

$$\mathbb{P}\left(\bigcup_i A_i|B\right) = \frac{\mathbb{P}\left(B \cap \bigcup_i A_i\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(\bigcup_i B \cap A_i\right)}{\mathbb{P}(B)} = \sum_i \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} = \sum_i \mathbb{P}(A_i|B),$$

which verifies that countable additivity holds and concludes the proof.

Exercise A.1.10 (Monty Hall)**.** A prize is placed at random behind one of three doors. You pick a door. To be concrete, let's suppose you always pick door 1. Now Monty Hall chooses one of the other two doors, opens it, and shows you that it is empty. He then gives you the opportunity to keep your door or switch to the other unopened door. Should you stay or switch? Intuition suggests it does not matter. The correct answer is that you should switch. Prove it. It will help to specify the sample space

$$\Omega = \{(p,\, m) : p,\, m = 1,\, 2, \text{ or } 3\}$$

where p is the prize's door and m is the door opened by Monty.

Solution. We assume without loss of generality that

(a) we initially choose door 1 and that
(b) Month opens door 2.

We thus want to compute

$$\mathbb{P}(p = 1|m = 2) \text{ and } \mathbb{P}(p = 3|m = 2).$$

First we observe that

$$\mathbb{P}(p = 1|m = 2) = \frac{\mathbb{P}(p = 1,\, m = 2)}{\mathbb{P}(m = 2)} = \frac{\mathbb{P}(m = 2|p = 1)\mathbb{P}(p = 1)}{\mathbb{P}(m = 2)} = \frac{1/2 \cdot 1/3}{1/2} = 1/3$$

where we have used assumption (a), and thus the fact that Monty must open either door 2 or 3, to compute $\mathbb{P}(m = 2|p = 1)$ and $\mathbb{P}(m = 2)$.

Now we observe that

$$\mathbb{P}(p = 3|m = 2) = \frac{\mathbb{P}(m = 2|p = 3)\mathbb{P}(p = 3)}{\mathbb{P}(m = 2)} = \frac{1 \cdot 1/3}{1/2} = 2/3$$

where, as above, we have used assumption (a) to compute $\mathbb{P}(m = 2)$ and, here, also used it to compute $\mathbb{P}(m = 2|p = 3)$.

In conclusion we have that

$$\mathbb{P}(p = 3|m = 2) = \frac{2}{3} > \frac{1}{3} = \mathbb{P}(p = 1|m = 2)$$

and so we ought indeed to switch!

The key is that, for any two equally likely events A_1 and A_2, comparing

$$\mathbb{P}(A_1|B) \text{ and } \mathbb{P}(A_2|B)$$

boils down to comparing

$$\mathbb{P}(B|A_1) \text{ and } \mathbb{P}(B|A_2)$$

by virtue of Bayes' Theorem. Since here

$$\mathbb{P}(m=2|p=3) = 1 > \frac{1}{2} = \mathbb{P}(m=2|p=1)$$

we deduce that indeed we ought to switch.

Exercise A.1.11 (Independence and complements)**.** Suppose that A and B are independent events. Prove that A^c and B^c are independent events.

Solution. This follows from a direct computation:

$$\begin{aligned}\mathbb{P}(A^c \cap B^c) &= 1 - \mathbb{P}(A \cup B)\\ &= 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(A \cap B)\\ &= 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(A)\mathbb{P}(B)\\ &= (1 - \mathbb{P}(A))(1 - \mathbb{P}(B))\\ &= \mathbb{P}(A^c)\mathbb{P}(B^c).\end{aligned}$$

Exercise A.1.12 (Coloured cards)**.** There are three cards. The first is green on both sides, the second is red on both sides, and the third is green on one side and red on the other. We choose a card at random and we see one side (also chosen at random). If the side we see is green, what is the probability that the other side is also green? Many people intuitively answer 1/2. Show that the correct answer is 2/3.

Solution. This follows immediately from Bayes' Theorem, noting that the only card for which it is possible to initially see a green side and then see another green side upon returning it is precisely card 1. Then

$$\mathbb{P}(\text{ card 1 } | \text{ see green }) = \frac{\mathbb{P}(\text{ see green } | \text{ card 1 })\mathbb{P}(\text{ card 1 })}{\mathbb{P}(\text{ see green })}.$$

There are three cards, so

$$\mathbb{P}(\text{ card 1 }) = \frac{1}{3}$$

and there are six card faces, three of which are green, so

$$\mathbb{P}(\text{ see green }) = \frac{3}{6} = \frac{1}{2}.$$

Therefore

$$\mathbb{P}(\text{ card 1 } | \text{ see green }) = \frac{1 \cdot 1/3}{1/2} = 2/3$$

as desired.

Exercise A.1.13 (Coin tosses 2)**.** Suppose that a fair coin is tossed repeatedly until both a head and a tail have appeared at least once.

(1) Describe the sample space Ω.
(2) What is the probability that three tosses will be required?

Solution. We proceed as follows.

(1) The sample space contains two types of strings.
  - Strings whose initial segment contains one-to-finitely many H's, and then exactly one T.

- Strings whose initial segment contains one-to-finitely many T's, and then exactly one H.

In other words, for $\mathbb{N} = \mathbb{N} \setminus \{0\} = \{1, 2, \ldots\}$ and $\oplus$ denoting string concatenation, we have that

$$\Omega = (H\mathbb{N}_* \oplus \{T\}) \cup (T\mathbb{N}_* \oplus \{H\}) .$$

(2) We compute that

$$\begin{aligned}\mathbb{P}(\text{at least 3 tosses required}) &= 1 - \mathbb{P}(\text{2 tosses required}) \\ &= 1 - \mathbb{P}(HT) - \mathbb{P}(TH) \\ &= 1 - 2 \cdot \left(\frac{1}{2}\right)^2 = 1 - \frac{1}{2} = \frac{1}{2}.\end{aligned}$$

Exercise A.1.14 (Self-independence)**.** Show that if $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = q$ then A is independent of every other event. Show that if A is independent of itself then $\mathbb{P}(A)$ is either 0 or 1.

Solution. If $\mathbb{P}(A) = 0$ then, for any other event B,

$$\mathbb{P}(A \cap B) \leqslant \mathbb{P}(A) = 0 = \mathbb{P}(A)\mathbb{P}(B),$$

i.e. A and B are indeed independent. If $\mathbb{P}(A) = 1$ then, for any other event B, $\mathbb{P}(A \cup B) \geqslant \mathbb{P}(A) = 1$ and so

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = 1 + \mathbb{P}(B) - 1 = \mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B).$$

Finally, suppose that A is independent of itself. This means that

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2,$$

i.e.

$$0 = \mathbb{P}(A)^2 - \mathbb{P}(A) = \mathbb{P}(A)(\mathbb{P}(A) - 1),$$

which means that indeed $\mathbb{P}(A)$ must be equal to 0 or 1.

Exercise A.1.15 (Eye colour)**.** The probability that a child has blue eyes is 1/4. Assume independence between children. Consider a family with 3 children.

(1) If it is known that at least one child has blue eyes, what is the probability that at least two children have blue eyes?

(2) If it is known that the youngest child has blue eyes, what is the probability that at least two children have blue eyes?

Solution. We proceed as follows.

(1) First we compute that the probability that none of the three children has blue eyes is

$$\begin{aligned}\mathbb{P}(\text{none}) &= \mathbb{P}(\text{youngest does not})\mathbb{P}(\text{cadet does not})\mathbb{P}(\text{eldest does not}) \\ &= \left(\frac{3}{4}\right)^3 = \frac{27}{64}.\end{aligned}$$

Then we compute that the probability that exactly one of the three children has blue eyes is, by independence,

$$\begin{aligned}\mathbb{P}(\text{exactly 1}) &= 3 \cdot \mathbb{P}(\text{youngest has blue eyes})\mathbb{P}(\text{cadet does not})\mathbb{P}(\text{eldest does not}) \\ &= 3 \cdot \frac{1}{4} \cdot \left(\frac{3}{4}\right)^2 = \frac{27}{64}.\end{aligned}$$

So finally:

$$\begin{aligned}\mathbb{P}(\text{at least } 2|\text{at least } 1) &= \frac{\mathbb{P}(\text{at least } 2)}{\mathbb{P}(\text{ at least } 1)} = \frac{1-\mathbb{P}(\text{at most } 1)}{1-\mathbb{P}(\text{none})} \\ &= \frac{1-\mathbb{P}(\text{none})-\mathbb{P}(\text{exactly } 1)}{1-\mathbb{P}(\text{none})} \\ &= \frac{1-\frac{27}{64}-\frac{27}{64}}{1-\frac{27}{64}} = \frac{64-54}{64-27} = \frac{10}{37}.\end{aligned}$$

(2) First we compute that the probability that the youngest, and at least one other child, have blue eyes is

$$\begin{aligned}&\mathbb{P}(\text{youngest has blue eyes and at least one other does}) \\ &= \mathbb{P}(\text{youngest has blue eyes and exactly one another does}) \\ &\quad + \mathbb{P}(\text{youngest has blue eyes and both other do}) \\ &= 2\cdot\left(\frac{1}{4}\right)^2\cdot\frac{3}{4}+\left(\frac{1}{4}\right)^3 = \frac{6}{64}+\frac{1}{64} = \frac{7}{64}.\end{aligned}$$

Therefore

$$\begin{aligned}&\mathbb{P}(\text{at least } 2|\text{youngest has blue eyes}) \\ &= \frac{\mathbb{P}(\text{youngest has blue eyes and at least one other does})}{\mathbb{P}(\text{youngest has blue eyes})} \\ &= \frac{7/64}{1/4} = \frac{7}{16}.\end{aligned}$$

Note that the probability increases in the second case compared to the first: this makes sense intuitively since we are in a realm where more information is given about children have blue eyes (knowing exactly which one of the three has blue eyes is more information that knowing that one of them has blue eyes, but not knowing which one), and so the probability that at least two children have blue eyes increases.

Exercise A.1.16 (Independence and conditional probability)**.** If A and B are independent events then $\mathbb{P}(A|B) = \mathbb{P}(A)$. Also, for any pair of events A and B with positive probabilities,

$$\mathbb{P}(AB) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A).$$

Solution. The latter equation follows immediately from the definition of conditional probability. The former equation then follows immediately from the latter since, for independent events A and B we have that $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$ and so we may cancel out $\mathbb{P}(B)$ from each side of the latter equation.

Exercise A.1.17 (Chain Rule)**.** Show that $\mathbb{P}(ABC) = \mathbb{P}(A|BC)\mathbb{P}(BC|C)\mathbb{P}(C)$.

Solution. This follows from applying the definition of conditional expectation twice:

$$\mathbb{P}(ABC) = \frac{\mathbb{P}(ABC)}{\mathbb{P}(BC)}\cdot\frac{\mathbb{P}(BC)}{\mathbb{P}(C)}\cdot\mathbb{P}(C) = \mathbb{P}(A|BC)\mathbb{P}(B|C)\mathbb{P}(C).$$

Above we have assumed that $\mathbb{P}(BC)$ and $\mathbb{P}(C)$ are nonzero. If either of these probabilities vanishes then $\mathbb{P}(ABC) = 0$ and so the identity also (trivially) holds.

Exercise A.1.18 (Investigating Bayes' Theorem)**.** Suppose that $A_1, \dots, A_k$ form a partition of a sample space Ω and assume that B is an event with non-zero probability. Prove that if $\mathbb{P}(A_1|B) < \mathbb{P}(A_1)$ then $\mathbb{P}(A_i|B) > \mathbb{P}(A_i)$ for some $i \geqslant 2$.

Solution. Suppose that $\mathbb{P}(A_1|B) < \mathbb{P}(A_1)$ and suppose, for the sake of contradiction, that $\mathbb{P}(A_i|B) \leqslant \mathbb{P}(A_i)$ for all $i \geqslant 2$. Then

$$\sum_{i=1}^{k} \mathbb{P}(A_i|B) < \sum_{i=1}^{k} \mathbb{P}(A_i),$$

but, since $\mathbb{P}(\,\cdot\,|B)$ is a probability measure, both sides of that inequality sum up to one – this is a contradiction, proving that indeed $\mathbb{P}(A_i|B) > \mathbb{P}(A_i)$ for some $i \geqslant 2$.

Exercise A.1.19 (Application of Bayes' Theorem)**.** Suppose that 30 percent of computer owners use a Macintosh, 50 percent use Windows, and 20 percent use Linux. Suppose that 65 percent of the Mac users have succumbed to a computer virus, 82 percent of the Windows users get the virus, and 50 percent of the Linux users get the virus. We select a person at random and learn that her system was infected with the virus. What is the probability that she is a Windows user?

Solution. This follows from a direct application of Bayes' Theorem:

$$\begin{aligned}\mathbb{P}(W|V) &= \frac{\mathbb{P}(V|W)\mathbb{P}(W)}{\mathbb{P}(V|W)\mathbb{P}(W) + \mathbb{P}(V|M)\mathbb{P}(M) + \mathbb{P}(V|L)\mathbb{P}(L)} \\ &= \frac{82 \cdot 50}{82 \cdot 50 + 65 \cdot 30 + 50 \cdot 20} = \frac{4100}{7050} \approx 58\%.\end{aligned}$$

Exercise A.1.20 (Unfair coins)**.** A box contains 5 coins and each has a different probability of showing heads. Let $p_1, \dots, p_5$ denote the probability of heads on each coin. Suppose that

$$p_1 = 0,\ p_2 = 1/4,\ p_3 = 1/2,\ p_4 = 3/4, \text{ and } p_5 = 1.$$

Let H denote "heads is obtained" and let C_i denote the event that coin i is selected.

(1) Select a coin at random and toss it. Suppose a head is obtained. What is the posterior probability that coin i was selected ($i = 1, \dots, 5$)? In other words, find $\mathbb{P}(C_i|H)$ for $i = 1, \dots, 5$.
(2) Toss the coin again. What is the probability of another head? In other words, find $\mathbb{P}(H_2|H_1)$ where H_j = "heads on toss j".

Now suppose that the experiment was carried out as follows: we select a coin at random and toss it until a head is obtained.

(3) Find $\mathbb{P}(C_i|B_4)$ where B_4 = "first head is obtained on toss 4".

Solution. In each part the crux of the argument is Bayes' Theorem.

(1) We know that $\mathbb{P}(H|C_j) = p_j$ for $j = 1, \dots, 5$. We may then use Bayes' Theorem to compute that

$$\mathbb{P}(C_i|H) = \frac{\mathbb{P}(H|C_i)\mathbb{P}(C_i)}{\sum_j \mathbb{P}(H|C_j)\mathbb{P}(C_j)}{}_i = \frac{p_i}{\sum_j p_j}$$

since $\mathbb{P}(C_j) = 1/5$ for all j. Since

$$\sum_j p_j = \frac{1+2+3+4}{4} = \frac{10}{4} \left(= \frac{5}{2}\right)$$

we conclude that

$$\mathbb{P}(C_i|H) = \begin{cases} 0 & \text{if } i = 1, \\ \dfrac{1/4}{10/4} = \dfrac{1}{10} & \text{if } i = 2, \\ \dfrac{2/4}{10/4} = \dfrac{2}{10} & \text{if } i = 3, \\ \dfrac{3/4}{10/4} = \dfrac{3}{10} & \text{if } i = 4, \text{ and} \\ \dfrac{4/4}{10/4} = \dfrac{4}{10} & \text{if } i = 5. \end{cases}$$

(2) Once again we proceed (essentially) via Bayes' Theorem:

$$\begin{aligned}\mathbb{P}(H_2|H_1) = \frac{\mathbb{P}(H_2 \cap H_1)}{\mathbb{P}(H_1)} &= \frac{\sum_j \mathbb{P}(H_2 \cap H_1|C_j)\mathbb{P}(C_j)}{\sum_j \mathbb{P}(H_1|C_j)\mathbb{P}(C_j)} \\ &= \frac{\sum_j \mathbb{P}(H_2 \cap H_1|C_j)}{\sum_j \mathbb{P}(H_1|C_j)} = \frac{\sum_j p_j^2}{\sum_j p_j},\end{aligned}$$

once again using the fact that $\mathbb{P}(C_j)$ has the same value for all j (i.e. each coin is equally likely to be chosen). The denominator was already computed in part 1 above so we only need to compute the numerator:

$$\sum_j p_j^2 = \frac{1 + 4 + 9 + 16}{16} = \frac{15}{8}.$$

So finally

$$\mathbb{P}(H_2 \cap H_1) = \frac{15/8}{5/2} = \frac{15}{8} \cdot \frac{2}{5} = \frac{3}{4}.$$

(3) Once more, we use Bayes' Theorem and the fact that each coin is equally likely to be chosen. This yields

$$\mathbb{P}(C_i|B_4) = \frac{\mathbb{P}(B_4|C_i)\mathbb{P}(C_i)}{\sum_j \mathbb{P}(B_4|C_j)\mathbb{P}(C_j)} = \frac{\mathbb{P}(B_4|C_i)}{\sum_j \mathbb{P}(B_4|C_j)} = \frac{(1-p_i)^3 p_i}{\sum_j (1-p_j)^3 p_j}.$$

We thus compute that

$$(1-p_i)^3 p_i = \begin{cases} 0 & \text{if } i = 1, \\ \dfrac{3^3 \cdot 1}{2^8} = \dfrac{27}{256} & \text{if } i = 2, \\ \dfrac{2^3 \cdot 2}{2^8} = \dfrac{16}{256} & \text{if } i = 3, \\ \dfrac{1^3 \cdot 3}{2^8} = \dfrac{3}{256} & \text{if } i = 4, \\ 0 & \text{if } i = 5, \end{cases}$$

and so

$$\sum_j (1-p_j)^3 p_j = \frac{27 + 16 + 3}{256} = \frac{46}{256}.$$

We therefore conclude that

$$\mathbb{P}(C_i|B_4) = \begin{cases} 0 & \text{if } i = 1, \\ \dfrac{27/256}{46/256} = \dfrac{27}{46} & \text{if } i = 2, \\ \dfrac{16/256}{46/256} = \dfrac{16}{46} & \text{if } i = 3, \\ \dfrac{3/256}{46/256} = \dfrac{3}{46} & \text{if } i = 4, \\ 0 & \text{if } i = 5. \end{cases}$$

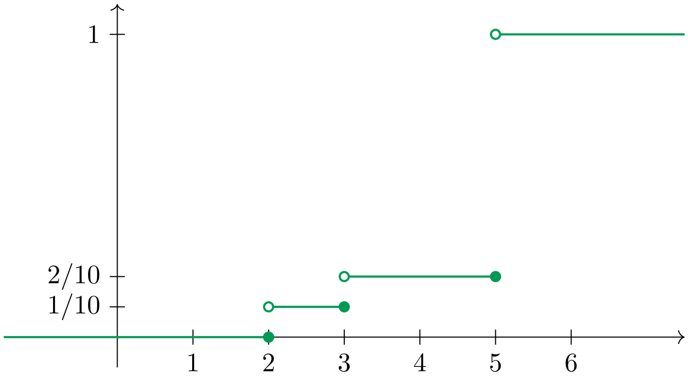


FIGURE A.1. CDF of Exercise A.2.2.

A.2. **Random variables.**

Exercise A.2.1 (Properties of CDFs, part 1)**.** Show that, for any random variable X, $\mathbb{P}(X = x) = F(x^+) - F(x^-)$.

Solution. First, recall that as per Theorem 2.8 of [Was10], CDFs are right-continuous and so $F(x^+) = F(x)$. Therefore it suffices to prove that $F(x^-) = \mathbb{P}(X < x)$ since then we may conclude that

$$F(x^+) - F(x^-) = F(x) - F(x^-) = \mathbb{P}(X \leqslant x) - \mathbb{P}(X < x) = \mathbb{P}(X = x).$$

So fix $x \in \mathbb{R}$ and let $y_n \uparrow x$ as $n \to \infty$. Since y_n is increasing, the sets $A_n = \{x < y_n\}$ are also increasing, meaning that $A_n \subseteq A_{n+1}$ for all n. Moreover we see that

$$\bigcup_{n=1}^{\infty} A_n = \{X < \sup y_n\} = \{X < x\}\,.$$

So finally, by the continuity of probability measures (Theorem 1.8 in [Was10]), we conclude that

$$F(x^-) = \lim_{n\to\infty} F(y_n) = \lim_{n\to\infty} \mathbb{P}(X < y_n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}(X < x)$$

as desired.

Exercise A.2.2 (Computations using CDFs)**.** Let X be a discrete random variable such that $\mathbb{P}(X = 2) = \mathbb{P}(X = 3) = 1/10$ and $\mathbb{P}(X = 5) = 8/10$. Plot the CDF F. Use F to find $\mathbb{P}(2 < X \leqslant 4.8)$ and $\mathbb{P}(2 \leqslant X \leqslant 4.8)$.

Solution. The CDF is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 2, \\ 1/10 & \text{if } 2 \leqslant x < 3, \\ 2/10 & \text{if } 3 \leqslant x < 5, \text{ and} \\ 1 & \text{if } 5 \leqslant x \end{cases}$$

and so its graph is as in Figure A.1. We now compute that

$$\mathbb{P}(2 < X \leqslant 4.8) = \mathbb{P}(X \leqslant 4.8) - \mathbb{P}(X \leqslant 2) = F(4.8) - F(2) = \frac{2}{10} - \frac{1}{10} = \frac{1}{10},$$

which makes sense since $\mathbb{P}(2 < X \leqslant 4.8) = \mathbb{P}(X = 3) = \frac{1}{10}$. Similarly, using Exercise A.2.1 immediately above we compute that

$$\begin{aligned}
\mathbb{P}(2 \leqslant X \leqslant 4.8) &= \mathbb{P}(X = 2) + \mathbb{P}(2 < X \leqslant 4.8) \\
&= \left[F(2) - F(2^-)\right] + [F(4.8) - F(2)] \\
&= F(4.8) - F(2^-) = \frac{2}{10} - 0 = \frac{2}{10}.
\end{aligned}$$

Once again this makes sense: $\mathbb{P}(2 \leqslant X \leqslant 4.8) = \mathbb{P}(X = 2) + \mathbb{P}(X = 3) = 2/10$.

Exercise A.2.3 (Properties of CDFs, part 2)**.** Let F be the CDF of a random variable X. The following hold.

(1) $\mathbb{P}(X = x) = F(x) - F(x^-)$.
(2) $\mathbb{P}(x < X \leqslant y) = F(y) - F(x)$.
(3) $\mathbb{P}(X > x) = 1 - F(x)$.
(4) If X is continuous then

$$\begin{aligned}
F(b) - F(a) &= \mathbb{P}(a < X < b) = \mathbb{P}(a \leqslant X < b) \\
&= \mathbb{P}(a < X \leqslant b) = \mathbb{P}(a \leqslant X \leqslant b).
\end{aligned}$$

Solution. We proceed as follows.

(1) We have shown in Exercise A.2.1 that $F(x^-) = \mathbb{P}(X < x)$ and so

$$\mathbb{P}(X = x) = \mathbb{P}(X \leqslant x) - \mathbb{P}(X < x) = F(x) - F(x^-).$$

(2) This follows from a direct computation:

$$\mathbb{P}(x < X \leqslant y) = \mathbb{P}(X \leqslant y) - \mathbb{P}(X \leqslant x) = F(y) - F(x).$$

(3) This follows from item 3 above and the fact that $F(\infty) = 1$ since we may write $\mathbb{P}(X > x) = \mathbb{P}(x < X \leqslant \infty)$.

(4) If X is continuous then there exists a function f, namely its PDF, such that $\mathbb{P}(a < X < b) = \int_a^b f$ for every $a < b$. In particular, for almost surely every real number c we have that, since the sets $\{c - 1/n < X < c + 1/n\}$ are decreasing,

$$\begin{aligned}
\mathbb{P}(X = c) &= \lim_{n\to\infty} \mathbb{P}\left(c - \frac{1}{n} < X < c + \frac{1}{n}\right) \\
&= \lim_{n\to\infty} \int_{c-1/n}^{c+1/n} f \\
&= \lim_{n\to\infty} \frac{2}{n} \fint_{c-1/n}^{c+1/n} f \\
&= 0
\end{aligned}$$

(where we have also used the Lebesgue Differentiation Theorem to deduce that, since f is integrable, its averages about a point converge to its value at that point almost surely). So finally, since

$$\begin{aligned}
&\mathbb{P}(a \leqslant X < b) = \mathbb{P}(X = a) + \mathbb{P}(a < X < b), \\
&\mathbb{P}(a < X \leqslant b) = \mathbb{P}(a < X < b) + \mathbb{P}(X = b), \text{ and} \\
&\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(X = a) + \mathbb{P}(a < X < b) + \mathbb{P}(X = b),
\end{aligned}$$

we conclude that indeed all of these three expressions are equal to the same thing, namely $\mathbb{P}(a < X < b)$.

Exercise A.2.4 (Computations using PDFs)**.** Let X have probability density function

$$f_X(x) = \begin{cases} 1/4 & \text{if } 0 < x < 1, \\ 3/8 & \text{if } 3 < x < 5, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

(1) Find the cumulative distribution function of X.
(2) Let $Y = 1/X$. Find the probability density function $f_Y(y)$ for Y. Hint: consider three cases, namely $\frac{1}{5} \leqslant y \leqslant \frac{1}{3}$, $\frac{1}{3} \leqslant y \leqslant 1$, and $y \geqslant 1$.

Solution. We proceed as follows.

(1) The CDF of X is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x \leqslant 0, \\ x/4 & \text{if } 0 < x \leqslant 1, \\ 1/4 & \text{if } 1 < x \leqslant 3, \\ 1/4 + 3(x-3)/8 & \text{if } 3 < x \leqslant 5, \text{ and} \\ 1 & \text{if } 5 < x. \end{cases}$$

(2) In order to compute the PDF of Y we first compute its CDF and then differentiate. We split into *four* cases.

- Case 1: $y \leqslant 1/5$. In this case $\mathbb{P}(Y \leqslant y) = \mathbb{P}(X \geqslant 5) = 0$.
- Case 2: $1/5 < y \leqslant 1/3$. In this case, since $3 \leqslant 1/y < 5$, we have that

$$\begin{aligned} \mathbb{P}(Y \leqslant y) &= \mathbb{P}(X \geqslant 1/y) \\ &= 1 - F_X(1/y) \\ &= 1 - \left[\frac{1}{4} + \frac{3\left(\frac{1}{y} - 3\right)}{8} \right] \\ &= \dots \\ &= \frac{15}{8} - \frac{3}{8y}. \end{aligned}$$

- Case 3: $1/3 < y \leqslant 1$. In this case

$$\mathbb{P}(Y \leqslant y) = \mathbb{P}(X \geqslant 1/y) = \mathbb{P}(X \geqslant 1)$$

  since X never take values between 1 and 3 and so

$$\mathbb{P}(Y \leqslant y) = 1 - F_X(1) = 3/4.$$

- Case 4: $1 < y$. Now, since $0 < \frac{1}{y} < 1$ we may proceed as in case 2 and compute that

$$\mathbb{P}(Y \leqslant y) = 1 - F_X(1/y) = 1 - \frac{1}{4y}.$$

In other words we have established that

$$F_Y(y) = \begin{cases} 0 & \text{if } y \leqslant \dfrac{1}{5}, \\ \dfrac{15}{8} - \dfrac{3}{8y} & \text{if } \dfrac{1}{5} < y \leqslant \dfrac{1}{3}, \\ \dfrac{3}{4} & \text{if } \dfrac{1}{3} < y \leqslant 1, \text{ and} \\ 1 - \dfrac{1}{4y} & \text{if } 1 < y. \end{cases}$$

Therefore since $f_y = F_y'$ we conclude that

$$f_Y(y) = \begin{cases} \dfrac{3}{8y^2} & \text{if } \dfrac{1}{5} < y < \dfrac{1}{3}, \\ \dfrac{1}{4y^2} & \text{if } 1 < y, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Exercise A.2.5 (Characterisation of independent for discrete random variables)**.** Let X and Y be discrete random variables. Show that X and Y are independent if and only if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$.

Solution. First, suppose that $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for every x and y. Then, for any A, $B \subseteq \mathbb{R}$ we have that

$$\begin{aligned} \mathbb{P}(X \in A, Y \in B) = \sum_{x\in A}\sum_{y\in B} f_{X,Y}(x, y) &= \sum_{x\in A}\sum_{y\in B} f_X(x)f_Y(y) \\ &= \sum_{x\in A} f_X(a) \underbrace{\left[\sum_{y\in B} f_Y(y)\right]}_{\mathbb{P}(Y\in B)} \\ &= \mathbb{P}(X \in A)\mathbb{P}(Y \in B), \end{aligned}$$

as desired. Conversely, suppose that X and Y are independent. Then, for any x and y, taking $A = \{x\}$ and $B = \{y\}$ we see that

$$\begin{aligned} f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) &= \mathbb{P}(X \in A, Y \in B) \\ &= \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \\ &= \mathbb{P}(X = x)\mathbb{P}(Y = y) \\ &= f_X(x)f_Y(y) \end{aligned}$$

as claimed.

Exercise A.2.6 (Indicator functions)**.** Let X have distribution F and density function f and let A be a subset of the real line. Let $I_A(x)$ be the indicator function for A:

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \text{ and} \\ 0 & \text{if } x \notin A. \end{cases}$$

Let $Y = I_A(X)$. Find an expression for the cumulative distribution function of Y. Hint: first find the probability mass function for Y.

Solution. This follows from a simple computation since Y only takes the values zero and one. In particular

$$\mathbb{P}(Y = 1) = \mathbb{P}(I_A(X) = 1) = \mathbb{P}(X \in A) = \int_A f_X$$

while

$$\mathbb{P}(Y = 0) = \mathbb{P}(X \notin A) = \int_{A^c} f_X = 1 - \int_A f_X.$$

Therefore

$$F_Y(y) = \begin{cases} 0 & \text{if } y \leqslant 0, \\ \displaystyle\int_{A^c} f_X = 1 - \int_A f_X & \text{if } 0 < y \leqslant 1, \text{ and} \\ 1 & \text{if } 1 < y \end{cases}$$

Exercise A.2.7 (Composition of independent random variables)**.** Let X and Y be independent and suppose that each has a Uniform(0, 1) distribution. We consider $Z = \min\{X, Y\}$. Find the density $f_Z(z)$ for Z. Hint: it might be easier to first find $\mathbb{P}(Z > z)$.

Solution. First, note that since $X, Y \sim$ Uniform(0, 1) we may compute that $F(\alpha) := F_X(\alpha) = F_Y(\alpha) = \alpha$ for $0 \leqslant \alpha \leqslant 1$. Therefore, using the independence of X and Y,

$$\begin{aligned} F_Z(z) = 1 - \mathbb{P}(Z > z) &= 1 - \mathbb{P}(\min\{X, Y\} > z) \\ &= 1 - \mathbb{P}(X > z, Y > z) \\ &= 1 - \mathbb{P}(X > z)\mathbb{P}(Y > z) \\ &= 1 - (1 - F(z))(1 - F(z)) \\ &= 1 - (1 - z)^2 \\ &= 2z - z^2 \end{aligned}$$

for $0 \leqslant z \leqslant 1$.

Exercise A.2.8 (CDF of the positive part)**.** Let X have CDF F. Find the CDF of $X^+ = \max\{0, X\}$.

Solution. This follows from a simple computation:

$$\begin{aligned} F_{X^+}(x) = \mathbb{P}(X^+ \leqslant x) &= \mathbb{P}(\max\{0, X\} \leqslant x) \\ &= \mathbb{P}(0 \leqslant x, X \leqslant x) \\ &= I_{\mathbb{R}_+}(x)\mathbb{P}(X \leqslant x) \\ &= I_{\mathbb{R}_+}(x)F_X(x), \end{aligned}$$

where $\mathbb{R}_+ = [0, \infty)$. In other words $F_{X^+} = I_{\mathbb{R}_+} F_x$.

Exercise A.2.9 (CDF and inverse CDF of the Exponential distribution)**.** Consider $X \sim \text{Exp}(\beta)$. Find $F(x)$ and $F^{-1}(q)$.

Solution. Since $X \sim \text{Exp}(\beta)$ we know that its PDF is given by $f(x) = \frac{1}{\beta}e^{-x/\beta}$. Therefore the CDF is given by, for any $x > 0$,

$$F(x) = \int_0^x \frac{1}{\beta} e^{-u/\beta} du = -e^{-u/\beta}|_{u=0}^{u=x} = 1 - e^{-x/\beta}.$$

Since the CDF is strictly increasing and continuous, the quantile function is precisely the inverse of the CDF. We compute that

$$q = 1 - e^{-x/\beta} \iff e^{-x/\beta} = 1 - q \iff -\frac{x}{\beta} = \ln(1-\beta) \iff x = -\beta\ln(1-q)$$

and so the quantile function is given by

$$F^{-1}(q) = -\beta\ln(1-q).$$

for $q \in (0,\, 1)$.

Exercise A.2.10 (Properties of independent random variables)**.** Let X and Y be independent random variables and let $g,\, h : \mathbb{R} \to \mathbb{R}$ be functions. Show that $g(X)$ and $h(Y)$ are independent.

Solution. For any two subsets $A,\, B \subseteq \mathbb{R}$ we have that

$$\begin{aligned}\mathbb{P}(g(X) \in A,\, h(Y) \in B) &= \mathbb{P}(X \in g^{-1}(A),\, Y \in h^{-1}(B))\\ &= \mathbb{P}(X \in g^{-1}(A))\mathbb{P}(Y \in h^{-1}(B))\\ &= \mathbb{P}(g(X) \in A)\mathbb{P}(h(Y) \in B),\end{aligned}$$

thus proving that $g(X)$ and $h(Y)$ are independent.

Exercise A.2.11 (Poisson-many coin tosses)**.** Suppose we toss a coin and let p be the probability of heads. Let X denote the number of heads and let Y denote the number of tails.

(1) Prove that X and Y are dependent.
(2) Let $N \sim \text{Poisson}(\lambda)$ and suppose that we toss a coin N times. Let X and Y be the number of heads and tails. Show that X and Y are independent.

Solution. We note that if p is either zero or one then, even in part 1, X and Y are *independent*. So we assume from now on that $0 < p < 1$.

(1) Say we tossed the coin n times. Then

$$P(X = n,\, Y = n) = 0$$

even though

$$P(X = n)P(Y = n) = p^n(1-p)^n \neq 0,$$

which verifies that X and Y are *dependent*.

(2) This follows from a simple, but careful, computation. On the one hand we have that

$$\begin{aligned}\mathbb{P}(X = k,\, Y = l) &= \mathbb{P}(X = k,\, Y = l,\, N = k + l)\\ &= \mathbb{P}(X = k,\, Y = l \,|\, N = k + l)\mathbb{P}(N = k + l)\\ &= \mathbb{P}(X = l \,|\, N = k + l)\mathbb{P}(N = k + l)\\ &= \binom{k+l}{k} p^k(1-p)^l \cdot e^{-\lambda}\frac{\lambda^{k+l}}{(k+l)!}\\ &= \frac{1}{k!l!}p^k(1-p)^l e^{-\lambda p}e^{-\lambda(1-p)}\lambda^k\lambda^l\\ &= e^{-\lambda p}\frac{(\lambda p)^k}{k!} \cdot e^{-\lambda(1-p)}\frac{[(1-p)]^l}{l!}.\end{aligned}$$

On the other hand we have that

$$\begin{aligned}\mathbb{P}(X = k) &= \sum_{n \geqslant k} \mathbb{P}(X = k \mid N = n)\mathbb{P}(N = n)\\ &= \sum_{n \geqslant k} \binom{n}{k} p^k (1-p)^{n-k} \cdot e^{-\lambda} \frac{\lambda^n}{n!}\\ &= \sum_{i \geqslant 0} \binom{k+i}{k} p^k (1-p)^i e^{-\lambda} \frac{\lambda^{k+i}}{(k+i)!}\\ &= \frac{1}{k!} (p\lambda)^k e^{-\lambda p} \underbrace{\sum_{i \geqslant 0} \frac{1}{i!} [(1-p)\lambda]^i e^{-(1-p)\lambda}}_{=1}\end{aligned}$$

where the summation adds up to one since we are summing up the probability mass function of a Poisson$((1-p)\lambda)$ distribution. In the same way we deduce that

$$\mathbb{P}(Y = l) = \frac{1}{l!} [(1-p)]\, \lambda^l e^{-\lambda(1-p)}$$

and so we may combine these three identities to conclude that indeed

$$\mathbb{P}(X = l,\, Y = l) = \mathbb{P}(X = k)\mathbb{P}(Y = l),$$

proving that X and Y are independent.

Exercise A.2.12 (Sufficient PDF condition for the independence of random variables)**.** Let X and Y be random variables such that the range of (X, Y) is a (possibly infinite) rectangle. Suppose that $f_{X,Y}(x, y) = g(x)h(y)$ for some functions g and h which are not necessarily PDFs. Prove that X and Y are independent.

Solution. This follows from direct computations. We have that, for any $A, B \subseteq \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(X \in A,\, Y \in B) &= \int_{A \times B} f_{X,Y} = \int_A g \int_B h,\\ \mathbb{P}(X \in A) &= \int_{A \times \mathbb{R}} f_{X,Y} = \int_A g \int_{\mathbb{R}} h, \text{ and}\\ \mathbb{P}(Y \in B) &= \int_{\mathbb{R} \times B} f_{X,Y} = \int_{\mathbb{R}} g \int_B h.\end{aligned}$$

In particular, since

$$\int_{\mathbb{R}} g \int_{\mathbb{R}} h = \int_{\mathbb{R} \times \mathbb{R}} f_{X,Y} = 1$$

we conclude that

$$\mathbb{P}(X \in A,\, Y \in B) = \int_A g \int_B h = \frac{\mathbb{P}(X \in A)}{\int_{\mathbb{R}} h} \cdot \frac{\mathbb{P}(Y \in B)}{\int_{\mathbb{R}} g} = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

as desired.

Exercise A.2.13 (Uniform distribution on the unit disk)**.** Let (X, Y) be uniformly distributed on the unit disk $D = \{(x, y) : x^2 + y^2 \leqslant 1\}$. Let $R = \sqrt{X^2 + Y^2}$. Find the CDF and PDF of R.

Solution. For $r \in [0,\, 1]$ we compute that

$$\begin{aligned} F_R(r) &= \mathbb{P}(\sqrt{X^2+Y^2} \leqslant r) \\ &= \text{Area}(rD)/\text{Area}(D) \\ &= \text{Area}\left(\left\{(x,\, y) : \sqrt{x^2+y^2} \leqslant r\right\}\right)/\text{Area}(D) \\ &= \pi r^2/\pi = r^2 \end{aligned}$$

and hence $f_R(r) = F_R'(r) = 2r$.

Exercise A.2.14 (A universal random number generator)**.** Let X have a continuous and strictly increasing CDF F. Let $Y = F(X)$. Find the density of Y. This is called the probability integral transform. Now let $U \sim \text{Uniform}(0,\, 1)$ and let $Z = F^{-1}(U)$. Show that $Z \sim F$.

Solution. First we note that since F is continuous and strictly increasing, it has a continuous and strictly increasing inverse. Moreover, since F maps to $[0,\, 1]$ we know that $F_Y = 0$ outside $[0,\, 1]$. Then, for $y \in [0,\, 1]$,

$$F_Y(y) = \mathbb{P}(F(X) \leqslant y) = \mathbb{P}(X \leqslant F^{-1}(y)) = F(F^{-1}(y)) = y.$$

In other words $Y \sim \text{Uniform}(0,\, 1)$.

We now compute that

$$F_Z(z) = \mathbb{P}(F^{-1}(U) \leqslant z) = \mathbb{P}(U \leqslant F(z)) = F(z)$$

since $\mathbb{P}(U \leqslant \alpha) = \alpha$ for any α in $[0,\, 1]$ and since $F(z)$ always belongs to $[0,\, 1]$.

Exercise A.2.15 (Sum of Poisson random variables)**.** Consider $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ for $\lambda,\, \mu > 0$ and assume that X and Y are independent. Show that the distribution of X given that $X + Y = n$ is $\text{Binomial}(n,\, \pi)$ where $\pi = \lambda/(\lambda+\mu)$.

Hint 1: You may use the following fact: If $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$, and X and Y are independent then $X + Y \sim \text{Poisson}(\mu + \lambda)$.

Hint 2: Note that $\{X = x,\, X + Y = n\} = \{X = x,\, Y = n - x\}$.

Solution. This follows from a direct computation since

$$\begin{aligned} \mathbb{P}(X = x \,|\, X + Y = n) &= \frac{\mathbb{P}(X = x,\, X + Y = n)}{\mathbb{P}(X + Y = n)} \\ &= \frac{\mathbb{P}(X = x,\, Y = n - x)}{\mathbb{P}(X + Y = n)} \\ &= \frac{\mathbb{P}(X = x)\mathbb{P}(Y = n - x)}{\mathbb{P}(X + Y = n)} \\ &= \frac{e^{-\lambda}\frac{\lambda^x}{x!} \cdot e^{-\mu}\frac{\mu^{n-x}}{(n-x)!}}{e^{-\lambda+\mu}\frac{(\lambda+\mu)^n}{n!}} \\ &= \binom{n}{x}\frac{\lambda^x\mu^{n-x}}{(\lambda+\mu)^n} \\ &= \binom{n}{x}\left(\frac{\lambda}{\lambda+\mu}\right)^x\left(\frac{\mu}{\lambda+\mu}\right)^{n-x} \\ &= \mathbb{P}(\text{Binomial}(n,\, \lambda/(\lambda+\mu)) = x) \end{aligned}$$

as desired since $1 - \frac{\lambda}{\lambda+\mu} = \frac{\mu}{\lambda+\mu}$.

Exercise A.2.16 (Working with PDFs)**.** Consider the joint PDF

$$f_{X,\,Y}(x,\,y) = \begin{cases} c(x+y^2) & \text{if } 0 \leqslant x \leqslant 1 \text{ and } 0 \leqslant y \leqslant 1 \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{P}(X < 1/2 \,|\, Y = 1/2)$.

Solution. We begin by computing the marginal PDF of Y:

$$f_Y(y) = \int_0^1 f_{X,\,Y}(x,\,y)dx = \int_0^1 c(x+y^2)dx = c\left(\frac{1}{2}+y^2\right).$$

We then compute the conditional PDF of X given Y to be

$$f_{X|Y}(x|y) = \frac{f_{X,\,Y}(x,\,y)}{f_Y(y)} = \frac{c\left(x+y^2\right)}{c\left(\frac{1}{2}+y^2\right)} = \frac{x+y^2}{\frac{1}{2}+y^2}.$$

In particular we may conclude that

$$\begin{aligned} \mathbb{P}\left(X < \frac{1}{2} \,\middle|\, Y = \frac{1}{2}\right) &= \int_0^{1/2} f_{X|Y}\left(x \,\middle|\, y = \frac{1}{2}\right) dx \\ &= \int_0^{1/2} \frac{x+\frac{1}{4}}{\frac{1}{2}+\frac{1}{4}} dx \\ &= \frac{4}{3}\left(\frac{x^2}{2}+\frac{x}{4}\right)\Big|_{x=0}^{x=1/2} \\ &= \frac{4}{3}\left(\frac{1}{8}+\frac{1}{8}\right) = \frac{1}{3}. \end{aligned}$$

Exercise A.2.17 (Formula for transforming random variables)**.** Let X be a random variable with continuous PDF f_X and let r be a diffeomorphism (r and its inverse are differentiable) which is either strictly monotone increasing and strictly monotone decreasing. Let Y be a random variable defined by $Y = r(X)$. Prove that the PDF f_Y of Y satisfies

$$f_Y = (f_X \circ r^{-1}) \left|(r^{-1})'\right|.$$

Solution. This follows from the Leibniz integral rule which states that, provided that a and b are differentiable and that f is continuous,

$$\frac{d}{dt}\int_{a(t)}^{b(t)} f = (f \circ b)b' - (f \circ a)a'.$$

So let us suppose first that r is strictly increasing. Then, for any y,

$$A_y = \{x : r(x) \leqslant y\} = \left\{x : x \leqslant r^{-1}(y)\right\}$$

and so

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\int_{A_y} f_X = \frac{d}{dy}\int_{-\infty}^{r^{-1}(y)} f_X = (f_X \circ r^{-1})(r^{-1})',$$

where $(r^{-1})' > 0$. Similarly, if r is strictly decreasing, $A_y = \left\{x \geqslant r^{-1}(y)\right\}$ and hence

$$f_Y(y) = \frac{d}{dy}\int_{r^{-1}(y)}^{\infty} f_X = -(f_X \circ r^{-1})(r^{-1})',$$

where now $(r^{-1})' < 0$. In general we therefore conclude that $f_Y = (f_X \circ r^{-1})|(r^{-1})'|$ as desired.

Exercise A.2.18 (Algebraic operations on uniform random variables)**.** Consider $X,\, Y \sim \text{Uniform}(0,\, 1)$ be independent. Find the PDF for $X - Y$ and X/Y.

Solution. We simply compute, following the recipe of Remark 2.36. First we note that $f_{X,\,Y} = \mathbb{1}_{[0,\,1]^2}$ and so, for any $A \subseteq [0,\,1]^2$,

$$\mathbb{P}((X,\,Y) \in A) = \int\int_A f_{X,\,Y} = |A|$$

where $|A|$ denotes the area of A.

Now we turn our attention to $Z := X - Y$. For any $-1 \leqslant z \leqslant 1$ we see that

$$A_z = \{(x,\,y) : x - y \leqslant z\} = \{(x,\,y) : y \geqslant x - z\}$$

such that

$$\mathbb{P}(Z \leqslant z) = \int\int_{A_z} f_{X,\,Y} = |A_z| = \begin{cases} \dfrac{(z+1)^2}{2} & \text{if } -1 \leqslant z \leqslant 0 \text{ and} \\ 1 - \dfrac{(1-z)^2}{2} & \text{if } 0 < z \leqslant 1. \end{cases}$$

This is easiest to see by observing that the regions A_z are triangles contained in $[0,\,1]^2$ when $-1 \leqslant z \leqslant 0$, or that their complement in $[0,\,1]^2$ is a triangle when $0 < z \leqslant 1$. So finally we deduce that

$$f_Z(z) = F_Z'(z) = \begin{cases} z+1 & \text{if } -1 \leqslant z \leqslant 0 \text{ and} \\ 1-z & \text{if } 0 < z \leqslant 1, \end{cases}$$

i.e. $f_Z(z) = 1 - |z|$.

Now we turn our attention to $W = x/y$. For any $w > 0$ we have that

$$A_w = \left\{(x,\,y) : \frac{x}{y} \leqslant w\right\} = \left\{(x,\,y) : y \geqslant \frac{x}{w}\right\}$$

and so

$$F_W(w) = \int\int_{A_w} f_{X,\,Y} = |A_w| = \begin{cases} \dfrac{w}{2} & \text{if } 0 < w \leqslant 1 \text{ and} \\ 1 - \dfrac{1}{2w} & \text{if } w > 1. \end{cases}$$

Once again, computing the areas of A_w above amounts to computing the area of triangles in $[0,\,1]^2$. So finally:

$$f_W(w) = F_W'(w) = \begin{cases} \dfrac{1}{2} & \text{if } 0 < w \leqslant 1 \text{ and} \\ \dfrac{1}{2w^2} & \text{if } w > 1. \end{cases}$$

Exercise A.2.19 (Maximum of IID exponential random variables)**.** Consider the IID random variables $X_1,\, \dots,\, X_n \sim \text{Exp}(\beta)$, for some $\beta > 0$. Define the random variable $Y = \max\{X_1,\, \dots,\, X_n\}$. Find the PDF of Y.
Hint: $Y \leqslant y$ if and only if $X_i \leqslant y$ for all $i = 1,\, \dots,\, n$.

Solution. This follows from a direct computation, using the CDF of the exponential distribution recorded in Exercise A.2.9. We have that

$$\begin{aligned} F_y(y) = \mathbb{P}(Y \leqslant y) &= \mathbb{P}(X_1 \leqslant y, \ldots, X_n \leqslant y) \\ &= \mathbb{P}(X_1 \leqslant y) \ldots \mathbb{P}(X_n \leqslant y) = \left(1 - e^{-y/\beta}\right)^n \end{aligned}$$

and hence

$$f_Y(y) = F_Y'(y) = \frac{n}{\beta} e^{-y/\beta} \left(1 - e^{-y/\beta}\right)^{n-1}.$$

A.3. **Expectation.**

Exercise A.3.1 (Gambling)**.** Suppose we play a game where we start with c dollars. On each play of the game you either double or halve your money, with equal probability. What is your expected fortune after n trials?

Solution. Let X denote your fortune after n trial. We can readily verify that the probability mass function of X is given by

$$f_X\left(2^{-2n+k}\right) = \binom{n}{k} 2^{-n}.$$

This can be done by computing by hand the probability mass function of X for small values of n (e.g. $n = 1, 2, 3, 4$) and could then be established for arbitrary n by induction. It then follows that the expected fortune after n trials is, by virtue of the binomial theorem,

$$\begin{aligned}\mathbb{E}(X) = \sum_{k=0}^{n} 2^{-2n+k} f_X\left(2^{-2n+k}\right) &= \sum_{k=0}^{n} 2^{-2n+k} \cdot \binom{n}{k} 2^{-n} \\ &= 4^{-n} \sum_{k=0}^{n} \binom{n}{k} 4^k \\ &= 4^{-n}(1+4)^n = \left(\frac{5}{4}\right)^n.\end{aligned}$$

Exercise A.3.2 (Characterisation of vanishing variance)**.** Show that $\mathbb{V}(X) = 0$ if and only if there is a constant c such that $\mathbb{P}(X = c) = 1$.

Solution. Suppose that $\mathbb{V}(X) = 0$, which means that $\mathbb{E}(X-\mu)^2 = 0$. Since $(X-\mu)^2 \geqslant 0$ this means that $X - \mu = 0$ $\mathbb{P}$–almost surely, i.e. $X = \mu$ $\mathbb{P}$–almost surely, and so we may take $c = \mu$.

Conversely, suppose that $\mathbb{P}(X = c) = 1$ for some $c \in \mathbb{R}$. Then

$$\mathbb{E}(X) = c \cdot \mathbb{P}(X = c) = 1$$

and

$$\mathbb{V}(X) = \mathbb{E}(X - c)^2 = 0$$

as desired.

Exercise A.3.3 (Expectation of the maximum of uniform random variables)**.** Let $X_1, \ldots, X_n \sim \text{Uniform}(0, 1)$ and let $Y_n = \max\{X_1, \ldots, X_n\}$. Find $\mathbb{E}(Y_n)$.

Solution. We begin by computing the CDF of Y_n:

$$\mathbb{P}(Y_n \leqslant y) = \mathbb{P}(X_1 \leqslant y, \ldots, X_n \leqslant y) = \mathbb{P}(X_1 \leqslant y) \ldots \mathbb{P}(X_n \leqslant y) = y^n$$

for y in $[0, 1]$. Therefore $f_Y(y) = ny^{n-1}$ for y in $[0, 1]$. So finally

$$\mathbb{E}(Y_n) = \int_0^1 y f_Y(y) dy = n \int_0^1 y^n dy = \frac{n}{n+1}.$$

Exercise A.3.4 (Random walk)**.** A particle starts at the origin of the real line and moves along the line in jumps of one unit. For each jump the probability is p that the particle will jump one unit to the left and the probability is $1-p$ that the particle will jump one unit to the right. Let X_n be the position of the particle at after n units. Find $\mathbb{E}(X_n)$ and $\mathbb{V}(X_n)$. (This is known as a *random walk.*)

Solution. We define, for $i = 1, \ldots, n$,

$$Y_i = \begin{cases} -1 & \text{with probability } p \text{ and} \\ +1 & \text{with probability } 1-p \end{cases}$$

to be IID random variables such that $X_n = \sum_{i=1}^n Y_i$. Then

$$\mathbb{E}(Y_i) = (-1) \cdot p + 1 \cdot (1 - p_{=}1 - 2p$$

and so

$$\mathbb{E}(X_n) = n(1 - 2p).$$

Similarly

$$\mathbb{V}(Y_i) = \mathbb{E}(Y_i^2) - \mu_{Y_i}^2 = 1 - (1-2p)^2 = 4p(1-p)$$

and so

$$\mathbb{V}(X_n) = \mathbb{V}\left(\sum_{i=1}^n Y_i\right) + \underbrace{\sum_{i,\,j} \mathrm{Cov}\,(Y_i,\, Y_j)}_{=0} = 4np(1-p).$$

Exercise A.3.5 (Coin toss 3)**.** A fair coin is tossed until a head is obtained. What is the expected number of tosses that will be required?

Solution. Let X be the random variable counting when the first head obtains, which means that

$$\mathbb{P}(X = 1) = 1/2,\, \mathbb{P}(X = 2) = 1/4,\, \mathbb{P}(X = 3) = 1/8,$$

i.e. $\mathbb{P}(X = k) = 2^{-k}$ for any integer $k \geqslant 1$. Therefore

$$\mathbb{E}(X) = \sum_{n \geqslant 1} n2^{-n} = 2.$$

Indeed, we know that for any $x \in (0,\, 1)$

$$\sum_{n \geqslant 0} x^n = \frac{1}{1-x}$$

and hence differentiating tells us that

$$\sum_{n \geqslant 1} nx^{n-1} = \frac{1}{(1-x)^2}.$$

Therefore

$$\sum_{n \geqslant 1} nx^n = \frac{x}{(1-x)^2}$$

such that in particular for $x = 1/2$ we have that

$$\sum_{n \geqslant 1} n2^{-n} = \frac{1/2}{1/4} = 2,$$

as claimed above.

Exercise A.3.6 (Rule of the Lazy Statistician for discrete random variables)**.** Let X be a discrete random variable and define $Y = r(X)$. Prove that

$$\mathbb{E}(Y) = \mathbb{E}(r(X)).$$

Solution. This follows from a direct computation, using the fact that all terms are positive and so the orderings of (possibly countable) sums can be interchanged. We compute that

$$\begin{aligned}\mathbb{E}(Y) = \sum_{y \in \operatorname{im} Y} y f_Y(y) &= \sum_{y \in \operatorname{im} Y} y \sum_{x\colon r(x)=y} f_X(x) \\ &= \sum_{y \in \operatorname{im} Y} \sum_{x\colon r(x)=y} r(x) f_X(x) \\ &= \sum_x r(x) f_X(x) = \mathbb{E}(r(X)).\end{aligned}$$

Exercise A.3.7 (Alternate integral formula for the expectation)**.** Let X be a continuous random variable with CDF F. Suppose that $\mathbb{P}(X > 0) = 1$ and that $\mathbb{E}(X)$ exists. Show that $\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > x)dx$. Hint: Consider integrating by parts. The following fact is helpful: if $\mathbb{E}(X)$ exists then $\lim_{n\to\infty} x\,[1 - F(x)] = 0$.

Solution. As per the hint, this follows from integrating by parts

$$\begin{aligned}\int_0^\infty \mathbb{P}(X > x)dx &= \int_0^\infty [1 - F(x)]\,dx \\ &= x\,[1 - F(x)]\Big|_{x=0}^{x=\infty} - \int_0^\infty x\,[-F'(x)]\,dx \\ &= 0 - 0 + \int_0^\infty x f_X(x)dx = \mathbb{E}(X).\end{aligned}$$

where we have used that fact that $x\,[1 - F(x)] \to 0$ as $x \to \infty$ and where, in the last line, we have used the fact that $\mathbb{P}(X > 0) = 1$ to write

$$\int_0^\infty x f_X(x)dx = \int_{\mathbb{R}} x f_X(x)dx = \mathbb{E}(X).$$

Exercise A.3.8 (Expectation and variance of sample mean and sample variance)**.** Let $X_1, \ldots, X_n$ be IID and let $\mu = \mathbb{E}(X_i)$ and $\sigma^2 = \mathbb{V}(X_i)$. Prove that

$$\mathbb{E}\left(\overline{X}_n\right) = \mu,\ \mathbb{V}\left(\overline{X}_n\right) = \frac{\sigma^2}{n},\ \text{and } \mathbb{E}(S_n^2) = \sigma^2.$$

Solution. We proceed in order.

(1) First we compute that, by the linearity of expectation,

$$\mathbb{E}\overline{X}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i = \frac{1}{n}\cdot n\mu = \mu.$$

(2) Second, using the fact that $\operatorname{Cov}(X, X) = \mathbb{V}(X)$ and the bilinearity of covariance we obtain that

$$\mathbb{V}\overline{X}_n = \frac{1}{n^2}\left[\sum_{i=1}^n \mathbb{V}X_i + \sum_{i\neq j}\operatorname{Cov}(X_i, X_j)\right] = \frac{1}{n^2}\cdot n\sigma^2 + 0 = \frac{\sigma^2}{n}.$$

(3) The last identity requires a bit more work. First, we note that for any random variable X, $\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$ and so, if $\mathbb{E}X = \mathbb{E}Y$ it follows that

$$\mathbb{E}(X^2) - \mathbb{E}(Y^2) = \mathbb{V}X - \mathbb{V}Y.$$

Therefore, using this identity, the fact that $\sum_i X_i = n\overline{X}_n$, and item 2 above we compute that

$$\begin{aligned}
\mathbb{E}\left[\sum_i \left(X_i - \overline{X}_n\right)^2\right] &= \mathbb{E}\left[\sum_i \left(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2\right)\right] \\
&= \mathbb{E}\left[\sum_i X_i^2 - 2n\overline{X}_n^2 + n\overline{X}_n^2\right] \\
&= \mathbb{E}\left[\sum_i X_i^2 - n\overline{X}_n^2\right] \\
&= \sum_i \left(\mathbb{E}X_i^2 - \mathbb{E}\overline{X}_n^2\right) \\
&= \sum_i \left(\mathbb{V}X_i - \mathbb{V}\overline{X}_n\right) \\
&= n\left(\sigma^2 - \frac{\sigma^2}{n}\right) = n \cdot \frac{(n-1)\sigma}{n} = (n-1)\sigma.
\end{aligned}$$

So finally we conclude that indeed

$$\mathbb{E}\left(S_n^2\right) = \frac{1}{n-1}\mathbb{E}\left[\sum_i \left(X_i - \overline{X}_n\right)^2\right] = \sigma.$$

Exercise A.3.9 (Mean and variance of a log-normal distribution)**.** Let $X \sim N(0, 1)$ and let $Y = e^X$ (which is called a *log–normal* distribution). Find $\mathbb{E}Y$ and $\mathbb{V}Y$.

Solution. First we compute that

$$\mathbb{E}Y = \mathbb{E}e^X = \int e^x f_X(x)dx = \int e^x \cdot \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx$$

where $x - x^2/2 = -(x-1)^2/2 + 1/2$ and so, since $\frac{1}{\sqrt{2\pi}}e^{-(x-1)^2/2}$ is the PDF of a $N(1, 1)$ random variable and hence integrates to one, we deduce that

$$\mathbb{E}Y = \int \frac{1}{\sqrt{2\pi}}e^{-(x-1)^2/2}e^{1/2}dx = e^{1/2}.$$

Similarly, since $\mathbb{E}Y^2 = \mathbb{E}e^{2X}$ where now $2x - x^2/2 = -(x-2)^2/2 + 2$, we have that

$$\mathbb{E}Y^2 = \int \frac{1}{\sqrt{2\pi}}e^{-(x-1)^2/2}e^2dx = e^2$$

and so

$$\mathbb{V}Y = \mathbb{E}Y^2 - \left(\mathbb{E}Y\right)^2 = e^2 - e = e(e-1).$$

Exercise A.3.10 (Mean and variance of important distributions)**.** Prove the formulas given in Figure 3.1 for the Bernoulli, Poisson, Uniform, Exponential, Gamma, and Beta distributions. Here are some hints. For the mean of the Poisson distribution, use the fact that $e^a = \sum_{x=0}^{\infty} a^x/x!$. To compute the variance, first compute $\mathbb{E}(X(X-1))$. For the mean of the Gamma distribution, it will help to multiply and divide by $\Gamma(\alpha+1)/\beta^{\alpha+1}$ and use the fact that a Gamma density integrates to one. For the Beta distribution, multiply by $\Gamma(\alpha+1)\Gamma(\beta)/\Gamma(\alpha+\beta+1)$.

Solution. We proceed in order.

- We begin with the Bernoulli distribution. Its expectation is given by
$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot (1-p) = p$$
and, since
$$\mathbb{E}(X^2) = 1 \cdot p + 0 \cdot (1-p) = p$$
we deduce that its variance is given by
$$\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = p - p^2 = p(1-p).$$
- We now turn our attention to the Poisson distribution. The expectation is given by
$$\begin{aligned}\mathbb{E}X &= \sum_{k \geqslant 0} k e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \sum_{k \geqslant 1} \dots \\ &= \sum_{k \geqslant 1} e^{-\lambda} \frac{\lambda^{k-1} \cdot \lambda}{(k-1)!} \\ &= \lambda \sum_{l \geqslant 0} e^{-\lambda} \frac{\lambda^l}{l!} = 1.\end{aligned}$$
To compute the variance we follow the hint and first compute
$$\begin{aligned}\mathbb{E}X(X-1) &= \sum_{k \geqslant 0} k(k-1) e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \sum_{k \geqslant 2} \dots \\ &= \sum_{k \geqslant 2} e^{-\lambda} \frac{\lambda^{k-2} \cdot \lambda^2}{(k-2)!} \\ &= \lambda^2 \sum_{l \geqslant 0} e^{-\lambda} \frac{\lambda^l}{l!} \\ &= \lambda^2.\end{aligned}$$
It then follows that
$$\begin{aligned}\mathbb{V}X &= \mathbb{E}X^2 - (\mathbb{E}X)^2 \\ &= \mathbb{E}X(X-1) + \mathbb{E}X - (\mathbb{E}X)^2 \\ &= \lambda^2 + \lambda - \lambda^2 \\ &= \lambda.\end{aligned}$$
- We now turn our attention to the Uniform distribution. First we compute that
$$\frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \cdot \frac{b^2 - a^2}{2} = \frac{b+a}{2}$$

and hence $\mathbb{E}(X) = \frac{b+a}{2}$. Second we compute that

$$\frac{1}{b-a}\int_a^b x^2 dx = \frac{1}{b-a} \cdot \frac{b^2 - a^2}{2} = \frac{b^2+ab+a^2}{3}$$

and hence

$$\mathbb{V}X = \frac{b^2+ab+a^2}{3} - \frac{b^2+2ab+a^2}{4} = \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}.$$

- We now turn our attention to the Exponential distribution. First we compute that

$$\begin{aligned}\mathbb{E}X &= \int_0^\infty \frac{x}{\beta} e^{-x/\beta} dx \\ &= x \cdot \left(-e^{-x/\beta}\right)\Big|_{x=0}^{x=\infty} - \int_0^\infty \left(-e^{-x/\beta}\right) dx \\ &= 0 - 0 + \beta = \beta.\end{aligned}$$

  Second we compute that

$$\begin{aligned}\mathbb{E}X^2 &= \int_0^\infty \frac{x^2}{\beta} e^{-x/\beta} dx \\ &= x^2 \cdot \left(-e^{-x/\beta}\right)\Big|_{x=0}^{x=\infty} - \int_0^\infty 2x \cdot \left(-e^{-x/\beta}\right) dx \\ &= 0 - 0 + 2\beta \cdot \int_0^\infty \frac{x}{\beta} e^{-x/\beta} dx = 2\beta^2\end{aligned}$$

  and hence

$$\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = 2\beta^2 - \beta^2 = \beta^2.$$

- We now turn our attention to the Gamma distribution. First we compute that

$$\begin{aligned}\mathbb{E}X &= \int_0^\infty \frac{x^\alpha e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} dx \\ &= \frac{\beta\,\Gamma(\alpha+1)}{\Gamma(\alpha)} \int_0^\infty \frac{x^\alpha e^{-x/\beta}}{\beta^{\alpha+1}\Gamma(\alpha+1)} dx \\ &= \alpha\beta\end{aligned}$$

  since $\Gamma(y+1)/\Gamma(y) = y$ for any $y > 0$. Second we compute that

$$\begin{aligned}\mathbb{E}X^2 &= \int_0^\infty \frac{x^{\alpha+1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} dx \\ &= \frac{\beta^2\,\Gamma(\alpha+2)}{\Gamma(\alpha)} \int_0^\infty \frac{x^{\alpha+1} e^{-x/\beta}}{\beta^{\alpha+2}\Gamma(\alpha+2)} dx \\ &= \frac{\beta^2\,\Gamma(\alpha+2)}{\Gamma(\alpha+1)} \cdot \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \alpha(\alpha+1)\beta^2\end{aligned}$$

  and hence

$$\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2.$$

- Finally we turn our attention to the Beta distribution. First we compute that
$$\begin{aligned}\mathbb{E}X &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha}(1-x)^{\beta-1}dx\\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\cdot\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+\beta+1)}\int_0^1 \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)} x^{\alpha}(1-x)^{\beta-1}dx\\ &= \frac{\alpha}{\alpha+\beta}\end{aligned}$$
since $\Gamma(y+1)/\Gamma(y) = y$ for any $y > 0$. Second we compute that
$$\begin{aligned}\mathbb{E}X^2 &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha+1}(1-x)^{\beta-1}dx\\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\cdot\frac{\Gamma(\alpha+2)}{\Gamma(\alpha+\beta+2)}\int_0^1 \frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+2)\Gamma(\beta)} x^{\alpha+1}(1-x)^{\beta-1}dx\\ &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\end{aligned}$$
such that
$$\begin{aligned}\mathbb{V}X &= \mathbb{E}X^2 - (\mathbb{E}X^2)\\ &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2}\\ &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.\end{aligned}$$

Exercise A.3.11 (A hierarchical model)**.** Suppose that we generate a random variable X in the following way. First we flip a fair coin. If the coin is heads, take X to have a Uniform(0, 1) distribution. If the coin is tails, take X to have a Uniform(3, 4) distribution.

(1) Find the mean of X.
(2) Find the standard deviation of X.

Solution. We proceed in order.

(1) We compute that the expectation is given by
$$\begin{aligned}\mathbb{E}X &= \mathbb{E}(X \mid \text{heads})\mathbb{P}(\text{heads}) + \mathbb{E}(X \mid \text{tails})\mathbb{P}(\text{tails})\\ &= \frac{1}{2}\int_0^1 xdx + \frac{1}{2}\int_3^4 xdx\\ &= \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{7}{2} = 2.\end{aligned}$$
(2) First we compute that
$$\begin{aligned}\mathbb{E}X^2 &= \mathbb{E}(X^2 \mid \text{heads})\mathbb{P}(\text{heads}) + \mathbb{E}(X^2 \mid \text{tails})\mathbb{P}(\text{tails})\\ &= \frac{1}{2}\int_0^1 x^2dx + \frac{1}{2}\int_3^4 x^2dx\\ &= \frac{1}{2}\cdot\frac{1}{3} + \frac{1}{2}\cdot\frac{64-27}{3} = \frac{38}{6} = \frac{19}{3}.\end{aligned}$$

Therefore the variance is given by

$$\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{19}{3} - 4 = \frac{7}{3}.$$

Exercise A.3.12 (Bilinearity of the covariance)**.** Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be random variables and let $a_1, \ldots, a_m$ and $b_1, \ldots, b_n$ be constants. Show that

$$\operatorname{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{i=j}^{n} a_i b_j \operatorname{Cov}(X_i, Y_j).$$

Solution. This is precisely the bilinearity of the covariance, which follows immediately from the linearity of expectation.

Exercise A.3.13 (Variance of a joint distribution)**.** Consider

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{3}(x+y) & \text{if } 0 \leqslant x \leqslant 1 \text{ and } 0 \leqslant y \leqslant 2 \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Compute $\mathbb{V}(2X - 3Y + 8)$.

Solution. We proceed in several steps. First we compute the marginal distributions:

$$f_X(x) = \int_0^2 f_{X,Y}(x, y)dy = \frac{2x}{3} + \int_0^2 \frac{y}{3}dy = \frac{2}{3}(x+1)$$

while

$$f_Y(y) = \int_0^1 f_{X,Y}(x, y)dx = \frac{y}{3} + \int_0^1 \frac{x}{3}dx = \frac{1}{6}(2y+1).$$

This allows us to compute the first two moments of X and Y. We have that

$$\begin{aligned}
\mathbb{E}X &= \int_0^1 x f_X(x)dx = \frac{2}{3}\int_0^1 (x^2 + x)dx = \frac{5}{9}, \\
\mathbb{E}X^2 &= \int_0^1 x^2 f_X(x)dx = \frac{2}{3}\int_0^1 (x^3 + x^2)dx = \frac{7}{18}, \\
\mathbb{E}Y &= \int_0^2 y f_Y(y)dy = \frac{1}{6}\int_0^2 (2y^2 + y)dy = \frac{11}{9}, \text{ and} \\
\mathbb{E}Y^2 &= \int_0^2 y^2 f_Y(y)dy = \frac{1}{6}\int_0^2 (2y^3 + y^2)dy = \frac{16}{9}.
\end{aligned}$$

Therefore the variances of X and Y are given by

$$\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{13}{162} \text{ and } \mathbb{V}Y = \mathbb{E}Y^2 - (\mathbb{E}Y)^2 = \frac{23}{81}.$$

The last piece we need is the covariance between X and Y. So first we compute that

$$\begin{aligned}
\mathbb{E}(XY) &= \int_0^1 \int_0^2 xy f_{X,Y}(x, y)dydx \\
&= \frac{1}{3}\int_0^1 \int_0^2 (x^2 y + xy^2)dydx \\
&= \frac{1}{3}\int_0^1 \left(2x^2 + \frac{8}{3}x\right)dx = \frac{2}{3},
\end{aligned}$$

which tells us that

$$\operatorname{Cov}(X,\, Y) = \mathbb{E}XY - (\mathbb{E}X)(\mathbb{E}Y) = \frac{-1}{81}.$$

Finally we put all of these pieces together, noting that

$$\mathbb{V}(1) = \operatorname{Cov}(X,\, 1) = \operatorname{Cov}(Y,\, 1) = 0$$

and hence

$$\mathbb{V}(2X - 3Y + 8) = 4\mathbb{V}(X) + 9\mathbb{V}(Y) - 2 \cdot 6 \cdot \operatorname{Cov}(X,\, Y) = \frac{245}{81} \approx 3.02.$$

Exercise A.3.14 (Pulling expressions out of conditional expectations)**.** Let $r(x)$ be a function of x and let $s(y)$ be a function of y. Show that

$$\mathbb{E}(r(X)s(Y) \,|\, X) = r(X)\mathbb{E}(s(Y) \,|\, X).$$

Also show that $\mathbb{E}(r(X)|X) = r(X)$.

Solution. This follows immediately from the definition of conditional expectation. We only carry out the computation for continuous random variables but a similar computation produces the claim for discrete random variables. For any x we have that

$$\begin{aligned} \mathbb{E}(r(X)s(Y) \,|\, X = x) &= \int r(x)s(y)f_{Y|X}(y|x)dy \\ &= r(x) \int s(y)f_{Y|X}(y|x)dy \\ &= r(x)\,\mathbb{E}(s(Y) \,|\, X = x), \end{aligned}$$

and so indeed $\mathbb{E}(r(X)s(Y) \,|\, X) = r(X)\mathbb{E}(s(Y) \,|\, X)$ as desired. To deduce the second identity we simply take $s \equiv 1$.

Exercise A.3.15 (Proving the law of total variance)**.** Prove that

$$\mathbb{V}(Y) = \mathbb{E}\mathbb{V}(Y|X) + \mathbb{V}\mathbb{E}(Y|X).$$

Hint: Let $m = \mathbb{E}(Y)$ and let $b(x) = \mathbb{E}(Y \,|\, X = x)$. Note that

$$\mathbb{E}(b(X)) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y) = m.$$

Bear in mind that b is a function of x. Now write

$$\mathbb{V}(Y) = \mathbb{E}(Y - m)^2 = \mathbb{E}((Y - b(X)) + (b(X) - m))^2.$$

Expand the square and take expectations. You then have to take the expectation of three terms. In each case, use the rule of iterated expectation: namely that $\mathbb{E}(\text{stuff}) = \mathbb{E}(\mathbb{E}(\text{stuff}|X))$.

Solution. As per the hint we split

$$\begin{aligned} \mathbb{V}Y &= \mathbb{E}(Y - m)^2 \\ &= \mathbb{E}([Y - b(X)] + [b(X) - m])^2 \\ &= \mathbb{E}[Y - b(X)]^2 + 2\mathbb{E}\left[(Y - b(X))\,(b(X) - m)\right] + \mathbb{E}[b(X) - m]^2 \\ &=: A + 2C + B. \end{aligned}$$

First we observe that, by the rule of iterated expectation,

$$A = \mathbb{E}[Y - b(X)]^2 = \mathbb{E}\left[\mathbb{E}\left([Y - b(X)]^2 \middle| X\right)\right]$$

where, by definition of b and of conditional expectation,

$$\mathbb{E}\left([Y-b(X)]^2\middle|X\right)=\mathbb{E}\left([Y-\mathbb{E}(Y|X)]^2\middle|X\right)=\mathbb{V}(Y|X)$$

such that

$$A=\mathbb{E}\mathbb{V}(Y|X).$$

Second we note that, since $\mathbb{E}b(X)=m$, it follows immediately that

$$B=\mathbb{E}[b(X)-m]^2=\mathbb{E}[b(X)-\mathbb{E}b(X)]^2=\mathbb{V}b(X)$$

and so, by definition of b,

$$B=\mathbb{V}\mathbb{E}(Y|X).$$

Third, and last, it suffices to show that $Y-b(X)$ and $b(X)-m$ are orthogonal, i.e. that $C=0$. We compute that

$$C=\mathbb{E}\left[(Y-b(X))\,(b(X)-m)\right]=\mathbb{E}\left[Yb(X)\right]-m\mathbb{E}(Y)-\mathbb{E}\left[b(X)^2\right]+m\mathbb{E}\left[b(X)\right].$$

Since $m=\mathbb{E}Y=\mathbb{E}b(X)$ this means that

$$C=\mathbb{E}\left[Yb(X)\right]-\mathbb{E}\left[b(X)^2\right].$$

Crucially, in light of the rule of iterated expectation, Exercise A.3.14, and the definition of b, we deduce that

$$\begin{aligned}\mathbb{E}\left[Yb(X)\right]&=\mathbb{E}\left(\mathbb{E}\left[Yb(X)|X\right]\right)\\&=\mathbb{E}\left(b(X)\mathbb{E}(Y|X)\right)\\&=\mathbb{E}\left[b(X)^2\right]\end{aligned}$$

and so indeed $C=0$. Combining the expressions obtained for A, B, and C produces the claim.

Exercise A.3.16 (Criteria for uncorrelation)**.** Show that if $\mathbb{E}(X\,|\,Y=y)=c$ for some constant c then X and Y are uncorrelated, meaning that $\mathbb{E}(XY)=(\mathbb{E}X)(\mathbb{E}Y)$.

Solution. First we compute using the rule of iterated expectation and Exercise A.3.14 that

$$\mathbb{E}X=\mathbb{E}\mathbb{E}(X|Y)=\mathbb{E}(c)=c.$$

Using the same two properties of conditional expectation once more we conclude that indeed

$$\mathbb{E}(XY)=\mathbb{E}\mathbb{E}(XY|Y)=\mathbb{E}\left(Y\mathbb{E}(X|Y)\right)=\mathbb{E}(cY)=c\mathbb{E}Y=(\mathbb{E}X)(\mathbb{E}Y)$$

as desired

Exercise A.3.17 (Linearity and bilinearity of expectation and variance for random vectors)**.** Let a be a fixed vector and let X be a random vector with mean μ and variance-covariance matrix Σ. Prove that

$$\mathbb{E}(a^TX)=a^T\mu \text{ and } \mathbb{V}(a^tX)=a^T\Sigma a.$$

Moreover, if A is a fixed matrix, prove that

$$\mathbb{E}(AX)=A\mu \text{ and } \mathbb{V}(AX)=A\Sigma A^T.$$

Solution. First we compute that

$$\mathbb{E}(a^T X) = \mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i \mathbb{E}(X_i) = \sum_i a_i \mu_i = a^T \mu.$$

Similarly we have that

$$\mathbb{V}(a^T \mu) = \mathbb{V}\left(\sum_i a_i X_i\right) = \sum_{i,\, j} a_i a_j \operatorname{Cov}(X_i,\, X_j) = \sum_{i,\, j} a_i a_j \Sigma_{ij} = a^T \Sigma a.$$

We now turn our attention to the vector-valued expectations and variance-covariance matrices. First we see that

$$\mathbb{E}(AX)_i = \mathbb{E}\left(\sum_j A_{ij} X_j\right) = \sum_j A_{ij} \mathbb{E}(X_j) = \sum_j A_{ij} \mu_j = (A\mu)_i,$$

i.e. $\mathbb{E}(AX) = A\mu$ as desired. Finally we see that

$$\begin{aligned}\mathbb{V}(AX)_{ij} &= \operatorname{Cov}\Big((AX)_i,\, (AX)_j\Big)\\ &= \operatorname{Cov}\left(\sum_k A_{ik} X_k,\, \sum_l A_{jl} X_l\right)\\ &= \sum_{k,\, l} A_{ik} A_{jl} \operatorname{Cov}(X_k,\, X_l)\\ &= \sum_{k,\, l} A_{ik} \Sigma_{kl} A^T_{lj}\end{aligned}$$

such that indeed $\mathbb{V}(AX) = A\Sigma A^T$.

Exercise A.3.18 (Conditional expectation and covariance)**.** Let X and Y be random variables. Suppose that $\mathbb{E}(Y|X) = X$. Show that $\operatorname{Cov}(X,\, Y) = \mathbb{V}(X)$.

Solution. First we compute, using the rule of iterated expectation and Exercise A.3.14, that

$$\mathbb{E}(XY) = \mathbb{E}\left[\mathbb{E}(XY|X)\right] = \mathbb{E}\left[X\mathbb{E}(Y|X)\right] = \mathbb{E}(X^2)$$

while

$$\mathbb{E}(Y) = \mathbb{E}\left[E(Y|X)\right] = \mathbb{E}X.$$

So finally

$$\operatorname{Cov}(X,\, Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{V}(X)$$

as desired.

Exercise A.3.19 (Computing a conditional expectation)**.** Let $X \sim \text{Uniform}(0,\, 1)$. Let $0 < a < b < 1$. Let

$$Y = \begin{cases} 1 & \text{if } 0 < x < b \text{ and} \\ 0 & \text{if } b \leqslant x < 1 \end{cases}$$

and

$$Z = \begin{cases} 1 & \text{if } a < x < 1 \text{ and} \\ 0 & \text{if } 0 < x \leqslant a. \end{cases}$$

(1) Are Y and Z independent? Why/Why not?
(2) Find $\mathbb{E}(Y|Z)$. Hint: what values of z can Z take? Now find $\mathbb{E}(Y \mid Z = z)$.

Solution. We proceed in order.

(1) The random variables Y and Z are *not* independent. To see that, we first observe that $Y = \mathbb{1}_{(0,\,b)}(X)$ and $Z = \mathbb{1}_{(a,\,1)}(X)$ and so

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}\left[\mathbb{1}_{(0,\,b)}(X)\right] = \int_0^b dx = b,\\ \mathbb{E}(Z) &= \mathbb{E}\left[\mathbb{1}_{(a,\,1)}(X)\right] = \int_a^1 dx = 1-a, \text{ and}\\ \mathbb{E}(YZ) &= \mathbb{E}\left[\mathbb{1}_{(0,\,b)}(X)\mathbb{1}_{(a,\,1)}(X)\right] = \int_a^b dx = b-a.\end{aligned}$$

Therefore

$$\operatorname{Cov}(Y,\,Z) = \mathbb{E}(YZ) - (\mathbb{E}Y)(\mathbb{E}Z) = -a(1-b) \neq 0,$$

which verifies that Y and Z are not independent (since independent random variables are necessarily uncorrelated).

(2) We note that Z only takes the values $z = 0$ or $z = 1$ and so there are two case to consider.

Case 1: $Z = 0$. This means that $X < a$, and so necessarily $Y = 1$. Therefore $\mathbb{E}(Y|Z = 0) = 1$.

Case 2: $Z = 1$. We need to perform some simple computations in this case. On the one hand

$$\mathbb{P}(Y = 0\,|\,Z = 1) = \frac{\mathbb{P}(Y = 0,\,Z = 1)}{\mathbb{P}(Z = 1)} = \frac{\mathbb{P}(X > b)}{\mathbb{P}(X > a)} = \frac{1-b}{1-a}$$

while on the other hand

$$\mathbb{P}(Y = 1\,|\,Z = 1) = \frac{\mathbb{P}(Y = 1,\,Z = 1)}{\mathbb{P}(Z = 1)} = \frac{\mathbb{P}(a < X < b)}{\mathbb{P}(X > a)} = \frac{b-a}{1-a}.$$

Therefore

$$\mathbb{E}(Y\,|\,Z = 1) = 0 \cdot \frac{1-b}{1-a} + 1 \cdot \frac{b-a}{1-a} = \frac{b-a}{1-a}.$$

So finally we are able to conclude that

$$\mathbb{E}(Y|Z) = \begin{cases} 1 & \text{if } Z = 0 \text{ and}\\ \dfrac{b-a}{1-a} & \text{if } Z = 1.\end{cases}$$

Exercise A.3.20 (Moment generating functions of important distributions)**.** Find the moment generating function for the Poisson, Normal and Gamma distributions.

Solution. First we consider the Poisson distribution. We compute that

$$\mathbb{E}\left(e^{tX}\right) = \sum_{k\geqslant 0} e^{tk} e^{-\lambda}\frac{\lambda^k}{k!} = e^{\lambda(e^t-1)}\sum_{k\geqslant 0} e^{-\lambda e^t}\frac{\left(\lambda e^t\right)^k}{k!} = e^{\lambda(e^t-1)}.$$

We now turn our attention to the Normal distribution and compute that

$$\mathbb{E}\left(e^{tX}\right) = \int_{\mathbb{R}} \frac{1}{\sigma\sqrt{2\pi}} e^{tx} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$$

where

$$-\frac{(x-\mu)^2}{\sigma^2}+tx=-\frac{1}{2\sigma^2}\left(x^2-2\mu x+\mu^2-2\sigma^2 tx\right)$$
$$=-\frac{1}{2\sigma^2}\left[\left(x-(\mu+\sigma^2 t)\right)^2-2\mu\sigma^2 t-\sigma^4 t^2\right]$$
$$=-\frac{1}{2\sigma^2}\left(x-(\mu+\sigma^2 t)\right)^2+\mu t+\frac{\sigma^2 t^2}{2}$$

and hence

$$\mathbb{E}\left(e^{tX}\right)=e^{\mu t+\frac{\sigma^2 t^2}{2}}\int_{\mathbb{R}}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(x-\left(\mu+\sigma^2 t\right)\right)^2}{2\sigma^2}\right)dx=e^{\mu t+\frac{\sigma^2 t^2}{2}}$$

Finally we turn our attention to the Gamma distribution. We compute that

$$\mathbb{E}\left(e^{tX}\right)=\int_0^\infty\frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{tx-\frac{x}{\beta}}dx=\left(\frac{\beta_t}{\beta}\right)^\alpha\int_0^\infty\frac{1}{\beta_t^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta_t}dx$$

for

$$\beta_t=\frac{-1}{t-\frac{1}{\beta}}=\frac{\beta}{1-\beta t}.$$

In particular:

$$\beta_t>0\iff t<\frac{1}{\beta}$$

in which case

$$\mathbb{E}\left(e^{tX}\right)=\left(\frac{\beta_t}{\beta}\right)^\alpha=\frac{1}{(1-\beta t)^\alpha}.$$

Exercise A.3.21 (Moment generating function of a sum of IID exponentials)**.** Let $X_1,\,\dots,\,X_n\sim\text{Exp}(\beta)$. Find the moment generating function of X_i. Prove that $\sum_{i=1}^n X_i\sim\text{Gamma}(n,\,\beta)$.

Solution. We first compute the moment generating function of a single Exp(β) random variable:

$$\mathbb{E}\left(e^{tX_i}\right)=\int_0^\infty\frac{1}{\beta}e^{tx-\frac{x}{\beta}}dx=\frac{\beta_t}{\beta}\int_0^\infty\frac{1}{\beta_t}e^{-x/\beta_t}dx$$

for

$$\beta_t:=\frac{-1}{t-\frac{1}{\beta}}=\frac{\beta}{1-\beta t}.$$

Since $\beta_t>0$ if and only if $t<1/\beta$ we have that, for such t sufficiently near zero,

$$\mathbb{E}\left(e^{tX}\right)=\frac{\beta_t}{\beta}=\frac{1}{1-\beta}.$$

We may now use Lemma 3.21 to compute the moment generating function of the sum $Y=\sum_{i=1}^n X_i$ which is given by

$$\psi_Y(t)=\psi_X^n(t)=\frac{1}{(1-\beta t)^n}$$

and which we recognize from Exercise A.3.20 to indeed be the moment generating function of a Gamma$(n,\,\beta)$ distribution.

A.4. **Inequalities.**

Exercise A.4.1 (Chebyshev and exponential distributions)**.** Let $X \sim \text{Exponential}(\beta)$. Find $\mathbb{P}(|X-\mu| \geqslant k\sigma)$ for $k > 1$. Compare this to the bound you get from Chebyshev's inequality.

Solution. Recall that, for exponential random variables, $\mu = \sigma = \beta$. Therefore, since $\mu - k\sigma = (1-k)\beta < 0$ and since X is non-negative, we have that

$$\begin{aligned}\mathbb{P}(|X-\mu| \geqslant k\sigma) &= \mathbb{P}(X \leqslant \mu - k\sigma) + \mathbb{P}(X \geqslant \mu + k\sigma)\\ &= \mathbb{P}(X \geqslant (1+k)\beta)\\ &= \int_{(1+k)\beta}^{\infty} \frac{1}{\beta} e^{-x/\beta} dx\\ &= -e^{-x/\beta}\Big|_{x=(1+k)\beta}^{\infty}\\ &= e^{-1-k}.\end{aligned}$$

By contrast, the Chebyshev inequality would only tells us that

$$\mathbb{P}(|X-\mu| \geqslant k\sigma) \leqslant \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2},$$

which only ensures *polynomial* decay in k, whereas the inequality above guaranteed *exponential* decay in k.

Exercise A.4.2 (Chebyshev and Poisson distributions)**.** Let $X \sim \text{Poisson}(\lambda)$. Use Chebyshev's inequality to show that $\mathbb{P}(X \geqslant 2\lambda) \leqslant 1/\lambda$.

Solution. Since X is non-negative and since $\mu = \sigma^2 = \lambda$, Chebyshev's inequality tells us that indeed

$$\mathbb{P}(X \geqslant 2\lambda) = \mathbb{P}(|X-\lambda| \geqslant \lambda) = \mathbb{P}(|X-\mu| \geqslant \lambda) \leqslant \frac{\sigma^2}{\lambda^2} = \frac{1}{\lambda}.$$

Exercise A.4.3 (Comparing Chebyshev and Hoeffding)**.** Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Bound $\mathbb{P}\left(|\overline{X}_n - p| > \varepsilon\right)$ using Chebyshev's inequality and using Hoeffding's inequality. Show that, when n is large, the bound from Hoeffding's inequality is smaller than the bound from Chebyshev's inequality.

Proof. We recall that $\mathbb{E}(X_i) = p$ and $\mathbb{V}(X_i) = p(1-p)$ and hence, by Exercise A.3.8, $\mathbb{E}(\overline{X}_n) = p$ and $\mathbb{V}(\overline{X}_n) = p(1-p)/n^2$. Therefore Chebyshev's inequality tells us that

$$\mathbb{P}\left(|\overline{X}_n - p| > \varepsilon\right) \leqslant \frac{p(1-p)/n^2}{\varepsilon^2}$$

while Hoeffding's inequality tells us that

$$\mathbb{P}\left(|\overline{X}_n - p| > \varepsilon\right) \leqslant 2e^{-2n\varepsilon^2}.$$

Since the Hoeffding bound is exponential whereas the Chebyshev bound is polynomial, it follows that the Hoeffding bound decays faster as n tends to infinity (to prove this we could use the Taylor expansion $x = \sum_{k\geqslant 0} x^k/k!$ of the exponential). □

Exercise A.4.4 (Confidence intervals and Hoeffding's inequality)**.** Let $X_1, \ldots, X_n \sim$ Bernoulli(p).

(1) Let $\alpha > 0$ and define

$$\varepsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}.$$

Let $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Define $C_n = (\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n)$. Use Hoeffding's inequality to show that

$$\mathbb{P}(C_n \text{ contains } p) \geqslant 1 - \alpha.$$

(2) Suppose we want the length of the interval to be no more than l_*. How large should n be?

Solution. First we note that ε_n is chosen precisely such that

$$2e^{-2n\varepsilon_n^2} \leqslant \alpha.$$

Therefore Hoeffding's inequality tells us that

$$\begin{aligned}\mathbb{P}\left(C_n \text{ contains } \hat{p}_n\right) = \mathbb{P}\left(|\hat{p}_n - p| \leqslant \varepsilon_n\right) &= 1 - \mathbb{P}\left(|\hat{p}_n - p| \geqslant \varepsilon_n\right)\\ &\geqslant 1 - 2e^{-2n\varepsilon_n^2} = 1 - \alpha,\end{aligned}$$

as desired.

Now we turn our attention to the length of the interval. Since the length of the interval is $2\varepsilon_n$ we have that

$$\begin{aligned}2\varepsilon_n \leqslant l_* &\iff \sqrt{\frac{2}{n}\log\left(\frac{2}{\alpha}\right)} \leqslant l_*\\ &\iff \frac{2}{n}\log\left(\frac{2}{\alpha}\right) \leqslant l_*^2\\ &\iff n \geqslant \frac{2}{l_*^2}\log\left(\frac{2}{\alpha}\right),\end{aligned}$$

i.e. we must have n larger than $\frac{2}{l_*^2}\log\left(\frac{2}{\alpha}\right)$.

Exercise A.4.5 (Mill's inequality)**.** Prove that if $Z \sim N(0, 1)$ and $t > 0$ then

$$\mathbb{P}\left(|Z| > t\right) \leqslant \sqrt{\frac{2}{\pi}}\frac{e^{-t^2/2}}{t}$$

Hint: Note that $\mathbb{P}\left(|Z| > t\right) = 2\mathbb{P}\left(Z > t\right)$. Now write out what $\mathbb{P}(Z > t)$ means and note that $x/t > 1$ whenever $x > t$.

Solution. We compute that

$$\begin{aligned}\mathbb{P}\left(|Z| > t\right) &= 2\mathbb{P}(Z > t)\\ &= 2\int_t^\infty \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx\\ &= \sqrt{\frac{2}{\pi}}\int_t^\infty \frac{x}{x}e^{-x^2/2}dx\\ &\leqslant \sqrt{\frac{2}{\pi}} \cdot \frac{1}{t} \cdot \int_t^\infty xe^{-x^2/2}dx\end{aligned}$$

where substituting $u = x^2/2$ tells us that

$$\int_t^\infty xe^{-x^2/2}dx = \int_{t^2/2}^\infty e^{-u}du = \left(-e^{-u}\right)\Big|_{u=t^2/2}^{u=\infty} = e^{-t^2/2}$$

and so indeed

$$\mathbb{P}\left(|Z| > t\right) \leqslant \sqrt{\frac{2}{\pi}} \cdot \frac{e^{-t^2/2}}{t}.$$

Exercise A.4.6 (Comparing Chebyshev and Mill)**.** Let $X_1, \ldots, X_n \sim N(0, 1)$. Bound $\mathbb{P}\left(|\overline{X}_n| > t\right)$ using Mill's inequality, where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n$. Compare to the Chebyshev bound.

Solution. Since the X_i's are independent we know that $\sum_{i=1}^n X_i \sim N(0, n)$, and hence $\overline{X}_n \sim N(0, 1/n)$. In particular, to make Mill's inequality easiest to use we note that $\sqrt{n}\overline{X}_n \sim N(0, 1)$. Mill's inequality then tells us that

$$\mathbb{P}\left(|\overline{X}_n| > t\right) = \mathbb{P}\left(\sqrt{n}|\overline{X}_n| > \sqrt{n}t\right) \leqslant \sqrt{\frac{2}{\pi}}\frac{e^{-nt^2}}{\sqrt{n}t}.$$

By contrast, Chebyshev's inequality tells us that

$$\mathbb{P}\left(|\overline{X}_n| > t\right) \leqslant \frac{\mathbb{V}\left(\overline{X}_n\right)}{t^2} = \frac{1}{nt^2}.$$

We thus see that the decay provided by the Mill bound is much faster than the Chebyshev bound since decays exponentially instead of decaying polynomially.

A.5. **Convergence of Random Variables.**

Exercise A.5.1 (Convergence in probability of the sample variance)**.** Let $X_1, \ldots, X_n$ be IID with finite mean $\mu = \mathbb{E}(X_1)$ and finite variance $\sigma^2 = \mathbb{V}(X_1)$. Let $\overline{X}_n$ be the sample mean and let S_n^2 be the sample variance.

(1) Show that $\mathbb{E}(S_n^2) = \sigma^2$.

(2) Show that $S_n^2 \xrightarrow{P} \sigma^2$. Hint: show that $S_n^2 = \frac{c_n}{n}\sum_{i=1}^n X_i^2 - d_n\overline{X}_n$ where $c_n \to 1$ and $d_n \to 1$. Apply the law of large numbers to $\frac{1}{n}\sum_{i=1}^n X_i^2$ and to $\overline{X}_n$. Then use Slutzky's Theorem, which states that if $Y_n \rightsquigarrow Y$ and $Z_n \rightsquigarrow c$ for some constant c then $Y_n + Z_n \rightsquigarrow Y + c$.

Solution. We note that part 1 was taken care of in Exercise A.3.8 and so we turn our attention to part 2. First we observe that

$$\begin{aligned}
S_n^2 &= \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^n \left(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2\right) \\
&= \frac{1}{n-1}\left[\left(\sum_{i=1}^n X_i^2\right) - 2n\overline{X}_n^2 + n\overline{X}_n^2\right] \\
&= \frac{n}{n-1}\cdot\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) - \frac{n}{n-1}\overline{X}_n^2.
\end{aligned}$$

We thus take $c_n = d_n = \frac{n}{n-1}$ and have that

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} \mathbb{E}X^2$$

by the weak law of large numbers, where

$$\mathbb{E}X^2 = \mathbb{V}X + (\mathbb{E}X)^2 = \sigma^2 + \mu^2,$$

while the weak law of large numbers also tells us that

$$\overline{X}_n^2 \xrightarrow{P} \mu^2$$

since convergence in probability is closed under multiplication. So finally:

$$S_n^2 \xrightarrow{P} (\sigma^2 + \mu^2) - \mu^2 = \sigma^2$$

as desired.

Exercise A.5.2 (Convergence in quadratic mean to a constant)**.** Let $X_1, X_2, \ldots$ be a sequence of random variables and let b be a constant. Show that $X_n \xrightarrow{qm} b$ if and only if

$$\mathbb{E}(X_n) \to b \text{ and } \mathbb{V}(X_n) \to 0.$$

Solution. We observe that

$$\begin{aligned}
\mathbb{E}(X_n - b)^2 &= \mathbb{E}(X_n^2) - 2b\mathbb{E}(X_n) + b^2 \\
&= \mathbb{V}(X_n) + \mathbb{E}(X_n)^2 - 2b\mathbb{E}(X_n) + b^2 \\
&= \mathbb{V}(X_n) + [\mathbb{E}(X_n) - b]^2.
\end{aligned}$$

Since both $\mathbb{V}(X_n)$ and $[\mathbb{E}(X_n)-b]^2$ are non-negative we deduce that indeed

$$\begin{aligned} X_n \overset{qm}{\to} b &\iff \mathbb{E}(X_n-b)^2 \to 0 \\ &\iff \mathbb{V}(X_n) \to 0 \text{ and } [\mathbb{E}(X_n)-b]^2 \to 0 \\ &\iff \mathbb{V}(X_n) \to 0 \text{ and } \mathbb{E}(X_n) \to b. \end{aligned}$$

Exercise A.5.3 (Finite variance and convergence in quadratic mean)**.** Let $X_1, X_2, \ldots$ be IID with mean $\mu = \mathbb{E}(X_1)$. Suppose that the variance of X_1 is finite. Show that $\overline{X}_n \overset{qm}{\to} \mu$.

Solution. This follows immediately from the mean and variance of the sample mean (see Exercise A.3.8):

$$\begin{aligned} \mathbb{E}(\overline{X}_n-\mu)^2 &= \mathbb{E}\left(\overline{X}_n^2\right) - 2\mu\mathbb{E}\left(\overline{X}_n\right) + \mu^2 \\ &= \mathbb{E}\left(\overline{X}_n^2\right) - \mu^2 \\ &= \mathbb{V}\left(\overline{X}_n\right) + \mathbb{E}\left(\overline{X}_n\right)^2 - \mu^2 \\ &= \frac{\sigma^2}{n} \to 0 \text{ as } n \to \infty. \end{aligned}$$

Exercise A.5.4 (Convergence in various modes, example 1)**.** Let $X_1, \ldots, X_2$ be a sequence of random variables such that

$$\mathbb{P}\left(X_n = \frac{1}{n}\right) = 1 - \frac{1}{n^2} \text{ and } \mathbb{P}\left(X_n = n\right) = \frac{1}{n^2}.$$

Does X_n converge in probability? Does X_n converge in quadratic mean?

Solution. For any $0 < \varepsilon < 1$ we have that, for any $n \geqslant \frac{1}{\varepsilon}$,

$$\mathbb{P}\left(|X_n| > \varepsilon\right) = \mathbb{P}\left(X_n = n\right) = \frac{1}{n^2} \to 0 \text{ as } n \to \infty$$

and so $X_n \overset{P}{\to} 0$. However we have that

$$\mathbb{E}\left(X_n^2\right) = \frac{1}{n^2}\mathbb{P}\left(X_n = \frac{1}{n}\right) + n^2\mathbb{P}\left(X_n = n\right) = \frac{1}{n^2}\left(1 - \frac{1}{n^2}\right) + \frac{n^2}{n^2} \geqslant 1$$

which shows that X_n does *not* converge to zero in quadratic mean.

Exercise A.5.5 (Convergence in quadratic mean of sum of squares of Bernoulli)**.** Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Prove that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \overset{P}{\to} p \text{ and } \frac{1}{n}\sum_{i=1}^{n} X_i^2 \overset{qm}{\to} p.$$

Solution. Note that since convergence in quadratic mean implies convergence in probability, it suffices to establish the former. Now observe that if $X \sim \text{Bernoulli}(p)$ then X^2 has the same distribution (since X only takes values 0 and 1). The random variable of interest is therefore the sample mean of the X_i^2 random variables, which are Bernoulli distributed. Since $\text{Bernoulli}(p)$ random variables have finite variance and mean p we deduce from Exercise A.5.3 that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \overset{qm}{\to} p$$

as desired.

Exercise A.5.6 (Using the central limit theorem, 1)**.** Suppose that the height of men has mean 68 inches and standard deviation 2.6 inches. We draw 100 men at random. Find (approximately) the probability that the average height of men in our sample will be at least 68 inches.

Solution. We denote by $\mu = 68$ the mean and $\sigma = 2.6$ the standard deviation. The central limit theorem then tells us that, since the average height among our sample is precisely the sample mean,

$$\mathbb{P}\left(\overline{X}_n \geqslant \mu\right) = \mathbb{P}\left(\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \geqslant 0\right) \approx \mathbb{P}(Z \geqslant 0)$$

for $Z \sim N(0,\, 1)$. Therefore $\mathbb{P}(\overline{X}_n \geqslant \mu) \approx 1/2$.

Exercise A.5.7 (Convergence in various modes, example 2)**.** Let $\lambda = 1/n$ for integers $n = 1,\, 2,\, \ldots$ and let $X_n \sim \text{Poisson}(\lambda_n)$.

(1) Show that $X_n \overset{P}{\to} 0$.
(2) Let $Y_n = nX_n$. Show that $Y_n \overset{P}{\to} 0$.

Solution. We proceed in order.

(1) This follows from Markov's inequality: for any $\varepsilon > 0$,

$$\mathbb{P}\left(|X_n| > \varepsilon\right) = \mathbb{P}\left(X_n > \varepsilon\right) \leqslant \frac{\mathbb{E}X_n}{\varepsilon} = \frac{1/n}{\varepsilon} \to 0.$$

(2) This follows from the following observation: for any $0 < \varepsilon < 1$,

$$\begin{aligned}\mathbb{P}(Y_n > \varepsilon) &= \mathbb{P}(Y_n \neq 0)\\ &= 1 - \mathbb{P}(Y_n = 0)\\ &= 1 - \mathbb{P}(X_n = 0)\\ &= 1 - \left(e^{-\lambda_n}\frac{\lambda_n^x}{x!}\right)\Bigg|_{x=0}\\ &= 1 - e^{-1/n} \to 0.\end{aligned}$$

Note that Y_n converges to zero in probability even though Y_n does *not* converge to zero in quadratic mean since

$$\begin{aligned}\mathbb{E}Y_n^2 &= \mathbb{V}(Y_n) + \left(\mathbb{E}Y_n\right)^2\\ &= n^2\left(\mathbb{V}(X_n) + \mathbb{E}(X_n)^2\right)\\ &= n^2\left(\lambda_n + \lambda_n^2\right)\\ &= n + 1\end{aligned}$$

which most definitely does not go to zero as n goes to infinity.

Exercise A.5.8 (Using the central limit theorem, 2)**.** Suppose we have a computer program consisting of $n = 100$ pages of code. Let X_i be the number of errors on the i–th page of code. Suppose that the X_i's are Poisson with mean 1 and that they are independent. Let $Y = \sum_{i=1}^{n} X_i$ be the total number of errors. Use the central limit theorem to approximate $\mathbb{P}(Y < 90)$.

Solution. Recall that a Poisson(λ) random variable has mean λ and variance λ. Since $\mathbb{E}(X_i) = 1$ we thus deduce that $\mathbb{V}(X_i) = 1$. The central limit theorem therefore tell us that

$$\frac{\sqrt{n}\left(\overline{X}_n - \mathbb{E}(X_i)\right)}{\sqrt{\mathbb{V}(X_i)}} = \sqrt{n}\left(\overline{X}_n - 1\right)$$

converges in distribution to a standard normal. Therefore, for $n = 100$ we have that $10\left(\overline{X}_n - 1\right)$ is approximately a standard normal and hence

$$\begin{aligned}\mathbb{P}(Y < 90) &= \mathbb{P}\left(n\overline{X}_n < 90\right)\\ &= \mathbb{P}\left(100\overline{X}_n < 90\right)\\ &= \mathbb{P}\left(10\left(\overline{X}_n - 1\right) < -1\right)\\ &\approx \mathbb{P}(Z < -1) = \Phi(-1)\end{aligned}$$

for $Z \sim N(0,\, 1)$, where Φ denotes the CDF of the standard normal distribution. So finally

$$\mathbb{P}(Y < 90) \approx \Phi(-1) \approx 0.159.$$

Exercise A.5.9 (Convergence in various modes, example 3)**.** Consider a random variable X for which $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$. Define

$$X_n = \begin{cases} X & \text{with probability } 1 - 1/n \text{ and} \\ e^n & \text{with probability } 1/n. \end{cases}$$

Does X_n converge to X in probability? Does X_n converge to X in distribution? Does X_n converge to X in quadratic mean?

Solution. First we show that X_n *does* converge to X in probability. Observe that only $X_n = X$ ensures that $|X_n - X| \leqslant 1$ and so, for any $0 < \varepsilon < 1$,

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(X_n \neq X) = \frac{1}{n} \to 0.$$

In other words: X_n converges in probability to X, and hence it must also converge to X in distribution. Now we turn our attention to the matter of convergence in quadratic mean. To make the computation easier to carry out, we define Y to be the random variable which determines whether $X_n = X$ or $X_n = e^n$, i.e.

$$X_n = \begin{cases} X & \text{if } Y = 1 \text{ and} \\ e^n & \text{if } Y = 0 \end{cases}$$

where $\mathbb{P}(Y = 1) = 1 - \mathbb{P}(Y = 0) = 1 - 1/n$. Therefore, by the rule of iterated expectation,

$$\begin{aligned}\mathbb{E}(X_n - X)^2 &= \mathbb{E}\left(\mathbb{E}\left[(X_n - X)^2 \,\middle|\, Y\right]\right)\\ &= \mathbb{E}\left[(X_n - X)^2 \,\middle|\, Y = 1\right]\mathbb{P}(Y = 1) + \mathbb{E}\left[(X_n - X)^2 \,\middle|\, Y = 0\right]\mathbb{P}(Y = 0)\\ &= 0 \cdot \left(1 - \frac{1}{n}\right) + \left[(e^n - 1)^2 \cdot \frac{1}{2} + (e^n + 1)^2 \cdot \frac{1}{2}\right] \cdot \frac{1}{n}\\ &= \frac{e^{2n} + 1}{n},\end{aligned}$$

where we have used the algebraic identity $(a+b)^2 + (a-b)^2 = 2(a^2+b^2)$. Since $(e^{2n}+1)/n$ does not approach zero as $n \to \infty$ we conclude that X_n does not converge to X in quadratic mean.

Exercise A.5.10 (Bound for the standard normal)**.** Let $Z \sim N(0,\,1)$. Let $t > 0$. Show that, for any $k > 0$,

$$\mathbb{P}\left(|Z| > t\right) \leqslant \frac{\mathbb{E}|Z|^k}{t^k}.$$

Compare this to Mill's inequality.

Solution. This follows directly from Markov's inequality since

$$\mathbb{P}\left(|Z| > t\right) = \mathbb{P}\left(|Z|^k > t^k\right) \leqslant \frac{|Z|^k}{t^k}.$$

By contrast, Mill's inequality tells us that

$$\mathbb{P}\left(|z| > t\right) \leqslant \sqrt{\frac{2}{\pi}}\frac{e^{-t^2/2}}{t},$$

which decays much faster (exponentially instead of polynomially).

Exercise A.5.11 (Convergence in various modes, example 4)**.** Suppose that we have $X_n \sim N(0,\,1/n)$ and let X be a random variable with distribution $F(x) = 0$ if $x \leqslant 0$ and $F(x) = 1$ if $x \geqslant 0$. Does X_n converge to X in probability? Does X_n converge to X in distribution?

Solution. We note that X is a point mass at zero, i.e. $\mathbb{P}(X = 0) = 1$. We will therefore show that X_n converges to X in quadratic mean, and that hence it converges to X both in probability and in distribution. We compute that

$$\mathbb{E}\left(X_n^2\right) = \mathbb{V}\left(X_n\right) + \left(\mathbb{E}(X_n)\right)^2 = \frac{1}{n} + 0 \to 0,$$

as desired.

Exercise A.5.12 (Convergence in distribution of discrete random variables)**.** Let $X,\, X_1,\, X_2,\, \ldots$ be random variables that are positive and integer-valued. Show that $X_n \rightsquigarrow X$ if and only if

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$$

for every positive integer k.

Solution. This follows from simple computations. If $X_n \rightsquigarrow X$ then we have that, for every positive integer k,

$$\begin{aligned}\mathbb{P}(X_n = k) &= \mathbb{P}(X_n \leqslant k + 1/2) - \mathbb{P}(X_n \leqslant k - 1/2)\\ &\to \mathbb{P}(X \leqslant k + 1/2) - \mathbb{P}(X \leqslant k - 1/2) = \mathbb{P}(X = k),\end{aligned}$$

as desired. Conversely, suppose that $\mathbb{P}(X_n = k) \to \mathbb{P}(X = k)$ for every positive integer k. For any $x \in \mathbb{R}$, if $\lfloor x \rfloor$ denotes the integer part of x then

$$\mathbb{P}(X_n \leqslant x) = \sum_{k=0}^{\lfloor x \rfloor} \mathbb{P}(X_n = k) \to \sum_{k=0}^{\lfloor x \rfloor} \mathbb{P}(X = k) = \mathbb{P}(X \leqslant k),$$

proving that $X_n \rightsquigarrow X$ as desired.

Exercise A.5.13 (An explicit computation of convergence)**.** Let $Z_1, Z_2, \ldots$ be IID random variables with density f. Suppose that $\mathbb{P}(Z_i > 0) = 1$ and suppose that $\lim_{x \downarrow 0} f(x) =: \lambda > 0$. Let

$$X_n := n \min\{Z_1, \ldots, Z_n\}.$$

Show that $X_n \rightsquigarrow Z$ where Z has an exponential distribution with mean $1/\lambda$.

Solution. For any $x > 0$ we may compute that

$$\begin{aligned}
\mathbb{P}(X_n > x) &= \mathbb{P}\left(Z_i > \frac{x}{n} \text{ for all } i\right) \\
&= \mathbb{P}\left(Z_i > \frac{x}{n}\right)^n \\
&= \left[1 - \mathbb{P}\left(Z_i \leqslant \frac{x}{n}\right)\right]^n \\
&= \left(1 - \int_0^{x/n} f\right)^n \\
&= \left(1 - \frac{x}{n} \fint_0^{x/n} f\right)^n.
\end{aligned}$$

Defining

$$\lambda_n := \fint_0^{x/n} f$$

we see that $\lambda_n \to \lambda$. Since moreover $(1 + y/n)^n \to e^y$ we deduce that

$$\mathbb{P}(X_n > x) = \left(1 - \frac{x\lambda_n}{n}\right)^n \to e^{-x\lambda}.$$

In particular, the CDF of an Exponential$(1/\lambda)$ random variable is recorded in Exercise A.2.9 to be $F(x) = 1 - e^{-\lambda x}$, which proves that indeed

$$X_n \rightsquigarrow \text{Exponential}(1/\lambda),$$

an exponential distribution with mean $1/\lambda$ as desired.

Note that we have used the following result: if $a_n \to a$ then $\left(1 + \frac{a_n}{a}\right)^n \to e^a$, or equivalently $n \log\left(1 + \frac{a_n}{n}\right) \to a$. To verify this result we observe that $\log 1 = 0$ and $\log' 1 = 1$, hence $\log x = x + O(x^2)$ as $x \to 0$. Therefore, since $a_n/n \to 0$ and since a_n is bounded we have that

$$n \log\left(1 + \frac{a_n}{n}\right) = n\left[\frac{a_n}{n} + O\left(\frac{a_n^2}{n^2}\right)\right] = a_n + n \cdot O\left(\frac{1}{n^2}\right) = a_n + O\left(\frac{1}{n}\right) \to a,$$

as desired.

Exercise A.5.14 (Delta method)**.** Let $X_1, \ldots, X_n \sim \text{Uniform}(0, 1)$. Let $Y_n := \overline{X}_n^2$ for $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ denoting the sample mean. Find the limiting distribution of Y_n.

Solution. Recall that if $X \sim \text{Uniform}(0, 1)$ then X has mean $\mu = \frac{1}{2}$ and variance $\sigma^2 = \frac{1}{12}$. The central limit theorem therefore tells us that

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \rightsquigarrow N(0, 1)$$

and hence, for $g(x) := x^2$ where $g'(\mu) = 1 \neq 0$, the delta method allows us to conclude that

$$\frac{\sqrt{n}\,(Y_n - 1/4)}{1/12} = \frac{\sqrt{n}\left(\overline{X}_n^2 - \mu^2\right)}{|g'(\mu)|\sigma} \rightsquigarrow N(0,\, 1),$$

i.e.

$$Y_n \approx N\left(\frac{1}{4},\, \frac{1}{12n}\right),$$

meaning that Y_n has an asymptotic mean of $1/4$ and an asymptotic variance of $1/(12n)$.

Exercise A.5.15 (Multivariate delta method)**.** Let

$$\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix},\, \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix},\, \ldots,\, \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix}$$

be IID random vectors with mean $\mu = (\mu_1,\, \mu_2)$ and variance Σ. Let

$$\overline{X}_1^{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_{1i},\, \overline{X}_2^{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_{2i},$$

and define $Y_n = \overline{X}_1^{(n)} \big/ \overline{X}_2^{(n)}$. Find the limiting distribution of Y_n.

Solution. By the multivariate central limit theorem we know that

$$\sqrt{n}\left(\overline{X}^{(n)} - \mu\right) \rightsquigarrow N(0,\, \Sigma).$$

Using $g(x_1,\, x_2) := x_1/x_2$ for which

$$\nabla g(\mu) = \begin{pmatrix} 1/\mu_2 \\ -\mu_1/\mu_2^2 \end{pmatrix} = \frac{1}{\mu_2^2}\begin{pmatrix} \mu_2 \\ -\mu_1 \end{pmatrix}$$

we deduce from the multivariate delta method that, since $Y_n = g\left(\overline{X}^{(n)}\right)$,

$$\sqrt{n}\left(Y_n - \frac{\mu_1}{\mu_2}\right) \rightsquigarrow N\left(0,\, \nabla g(\mu)^T \Sigma \nabla g(\mu)\right).$$

In particular we may compute explicitly that

$$\begin{aligned} \nabla g(\mu)^T \Sigma \nabla g(\mu) &= \frac{1}{\mu_2^4}\begin{pmatrix} \mu_2 & -\mu_1 \end{pmatrix}\begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_{22}^2 \end{pmatrix}\begin{pmatrix} \mu_2 \\ -\mu_1 \end{pmatrix} \\ &= \frac{1}{\mu_2^4}\left(\sigma_{11}^2\mu_2^2 + \sigma_{22}^2\mu_1^2 - 2\sigma_{12}^2\mu_1\mu_2\right). \end{aligned}$$

Exercise A.5.16 (Failure of additivity of convergence in distribution)**.** Construct an example where $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ but $X_n + Y_n$ does not converge in distribution to $X + Y$.

Solution. Consider $X_1,\, X_2,\, \ldots$ to be IID $N(0,\, 1)$ random variables and define $Y_n = -X_n$ for all n. Then $X_n \rightsquigarrow N(0,\, 1)$ and $Y_n \rightsquigarrow N(0,\, 1)$ but $X_n + Y_n \equiv 0$ and so clearly it does not converge in distribution to $N(0,\, 2)$, which is the sum of two independent $N(0,\, 1)$ random variables.

A.6. Models, Statistical Inference, and Learning.

Exercise A.6.1 (Sample mean as an estimator for Poisson distributions)**.** Consider $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ and let $\hat{\lambda}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$. Find the bias, standard error, and mean squared error of this estimator.

Solution. We compute that

$$\mathbb{E}_\lambda(\hat{\lambda}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\lambda(X_i) = \lambda$$

and so the bias$(\hat{\lambda}_n) = \mathbb{E}_\lambda(\hat{\lambda}_n) - \lambda = 0$. Similarly we compute that, for $\overline{X}_n$ denoting the sample mean,

$$\mathbb{V}_\lambda(\hat{\lambda}_n) = \mathbb{V}_\lambda(\overline{X}_n) = \frac{\mathbb{V}_\lambda(X_1)}{n} = \frac{\lambda}{n},$$

i.e. $\text{se}(\hat{\lambda}_n) = \sqrt{\lambda/n}$. So finally

$$\text{MSE}(\hat{\lambda}_n) = \text{bias}^2(\hat{\lambda}_n) + \text{se}^2(\hat{\lambda}_n) = \frac{\lambda}{n}.$$

Exercise A.6.2 (An biased estimator for Uniform distributions)**.** Consider samples $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$ and let $\hat{\theta}_n = \max\{X_1, \ldots, X_n\}$. Find the bias, standard error, and mean squared error of this estimator.

Solution. We begin by mimicking Exercise A.3.3 to compute the CDF of $\hat{\theta}_n$:

$$\mathbb{P}(\hat{\theta}_n \leqslant x) = \mathbb{P}(X_1 \leqslant x)^n = \left(\frac{x}{\theta}\right)^n.$$

Therefore the PDF of $\hat{\theta}_n$ is given by nx^{n-1}/θ^n and so its expectation is

$$\mathbb{E}_\theta(\hat{\theta}_n) = \int_0^\theta x \cdot \frac{nx^{n-1}}{\theta^n}dx = \frac{n}{n+1}\cdot\frac{x^{n+1}}{\theta^n}\Big|_{x=0}^{x=\theta} = \frac{n\theta}{n+1}.$$

Therefore the bias of $\hat{\theta}_n$ is

$$\text{bias}(\hat{\theta}_n) = \frac{n\theta}{n+1} - \theta = -\frac{\theta}{n+1}.$$

We then compute the variance of $\hat{\theta}_n$ to be

$$\mathbb{V}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}_n^2) - \left(\frac{n\theta}{n+1}\right)^2$$

where

$$\mathbb{E}_\theta(\hat{\theta}_n^2) = \int_0^\theta x^2 \cdot \frac{nx^{n-1}}{\theta^n}dx = \frac{n}{n+2}\cdot\frac{x^{n+2}}{\theta^n}\Big|_{x=0}^{x=\theta} = \frac{n\theta^2}{n+2}$$

and hence

$$V_\theta(\hat{\theta}_n) = \left[\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right]\theta^2 = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

This means that the standard error is given by

$$\text{se}(\hat{\theta}_n) = \frac{\theta}{n+1}\sqrt{\frac{n}{n+2}}.$$

So finally the mean squared error is

$$\text{MSE}(\hat{\theta}_n) = \text{bias}^2(\hat{\theta}_n) + \text{se}^2(\hat{\theta}_n) = \frac{\theta^2}{(n+1)^2} + \frac{n\theta^2}{(n+1)^2(n+2)} = \frac{2\theta}{(n+1)(n+2)}.$$

Exercise A.6.3 (An unbiased estimator for Uniform distributions)**.** Consider samples $X_1, \ldots, X_n \sim \text{Uniform}(\theta)$ and let $\hat{\theta}_n := 2\overline{X}_n$. Find the bias, standard error, and mean squared error of this estimator.

Solution. First we compute that

$$\mathbb{E}_\theta(\hat{\theta}_n) = 2\mathbb{E}_\theta(\hat{\theta}_n) = 2\mathbb{E}_\theta(X_1) = 2 \cdot \frac{\theta}{2} = \theta$$

which tells us that this point estimator is unbiased. Second we compute that

$$\mathbb{V}_\theta(\hat{\theta}_n) = 4\mathbb{V}_\theta(\overline{X}_n) = \frac{4\mathbb{V}_\theta(\overline{X}_n)}{n} = \frac{\theta^2}{3n},$$

and hence $\text{se}(\hat{\theta}_n) = \theta/\sqrt{3n}$. So finally we conclude that

$$\text{MSE}(\hat{\theta}_n) = \text{bias}^2(\hat{\theta}_n) + \text{se}^2(\hat{\theta}_n) = \frac{\theta^2}{3n}.$$

Remark A.1 (Biased and unbiased operators)**.** In both Exercises A.6.2 and A.6.3 above we consider point estimators which are consistent, since their mean squared error converges to zero. However, the *biased* estimator of Exercise A.6.2 actually converges faster than the *unbiased* estimator of Exercise A.6.3, which serves as evidence that seeking unbiased estimators at all cost may be misguided.

A.7. **Estimating the CDF and Statistical Functionals.**

Exercise A.7.1 (Properties of the empirical distribution function)**.** Consider samples $X_1, \ldots, X_n \sim F$ be IID and let

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x)$$

for every $x \in \mathbb{R}$. Show that

(1) $\mathbb{E}(\widehat{F}_n(x)) = F(x)$,
(2) $\mathbb{V}(\widehat{F}_n(x)) = \frac{1}{n}F(x)(1 - F(x))$,
(3) as a point estimator of $F(x)$, $\mathrm{MSE}(\widehat{F}_n(x)) = \frac{1}{n}F(x)(1 - F(x))$, and
(4) $\widehat{F}_n(x) \xrightarrow{P} F(x)$.

Solution. We proceed in order and throughout we write $X = X_1$ for simplicity.

(1) This follows immediately from the following computation:

$$\begin{aligned}\mathbb{E}(\widehat{F}_n(x)) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(I(X_i \leqslant x)) \\ &= \mathbb{E}(I(X \leqslant x)) \\ &= 1 \cdot \mathbb{P}(X \leqslant x) + 0 \cdot \mathbb{P}(X > x) \\ &= \mathbb{P}(X \leqslant x) = F(x).\end{aligned}$$

(2) By independence we compute that

$$\mathbb{V}(\widehat{F}_n(x)) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(I(X_i \leqslant x)) = \frac{1}{n}\mathbb{V}(I(X \leqslant x))$$

where, using as derived in part 1 above that $\mathbb{E}(I(X \leqslant x)) = F(x)$,

$$\mathbb{E}\left(I(X \leqslant x)^2\right) = \mathbb{E}(I(X \leqslant x)) = F(x)$$

and hence

$$\mathbb{V}(I(X \leqslant x)) = F(x) - F(x)^2 = F(x)(1 - F(x))$$

such that finally

$$\mathbb{V}(\widehat{F}_n(x)) = \frac{1}{n}F(x)(1 - F(x)).$$

(3) Part 1 tells us that $\widehat{F}_n(x)$ is an unbiased point estimator of $F(x)$ and so its mean squared error is given by, using part 2,

$$\mathrm{MSE}(\widehat{F}_n(x)) = \mathrm{se}^2(\widehat{F}_n(x)) = \mathbb{V}(\widehat{F}_n(x)) = \frac{F(x)(1 - F(x))}{n}.$$

(4) The mean squared error converges to zero, which is equivalent to the convergence in quadratic mean of $\widehat{F}_n(x)$ to $F(x)$. This in turn implies that $\widehat{F}_n(x)$ converges in probability to $F(x)$.

Exercise A.7.2 (Plug-in estimators for Bernoulli random variables)**.** Consider samples $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(p)$ and $Y_1, \ldots, Y_n \sim \mathrm{Bernoulli}(q)$. Find the plug-in estimator and estimated standard error for p. Find an approximate 90 percent confidence interval for p. Find the plug-in estimator and estimate standard error for $p - q$. Find an approximate 90 percent confidence interval for $p - q$.

Solution. First we find the plug-in estimator for p. For $X \sim \text{Bernoulli}(p)$ we have that

$$p = \mathbb{P}(X = 1) = \mathbb{P}(X \leqslant 1) - \mathbb{P}(X \leqslant 0) = F(1) - F(0)$$

and so the plug-in estimator for p is

$$\hat{p}_n = \widehat{F}_n(1) - \widehat{F}_n(0).$$

In particular, for $X_i \sim \text{Bernoulli}(p)$ we see as above that

$$I(\mathbb{P}(X_i \leqslant 1)) - I(\mathbb{P}(X_i \leqslant 0)) = I(\mathbb{P}(X_i = 1)) = X_i$$

and so

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} \left[I(\mathbb{P}(X_i \leqslant 1)) - I(\mathbb{P}(X_i \leqslant 0)) \right] = \overline{X}_n,$$

i.e. the plug-in estimator for p is simply the sample mean. Moreover, since we are considering $X_i \sim \text{Bernoulli}(p)$, the sample mean can be interpreted as the *sample success frequency*.

We now estimate the standard error for p. Since $\hat{p}_n = \overline{X}_n$ we compute that

$$\text{se}^2(\hat{p}_n) = \mathbb{V}(\hat{p}_n) = \frac{\mathbb{V}(X)}{n}.$$

We can then estimate the variance $\sigma^2 = \mathbb{V}(X)$ in a few different ways:

- using the plug-in estimator $\hat{\sigma}_n^2 = \int x^2 d\widehat{F}_n(x) - \left(x d\widehat{F}_n(x) \right)^2$,
- using the sample variance $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n \right)^2$, or
- first computing that, for $X \sim \text{Bernoulli}(p)$, $\mathbb{V}(X) = p(1-p)$, and then using the corresponding plug-in estimator $\hat{p}_n(1 - \hat{p}_n)$.

We will use the third method. The estimated standard error is then given by

$$\widehat{\text{se}}_n = \sqrt{\frac{\hat{p}_n(1 - \hat{p}_n)}{n}}.$$

Finally we find an approximate 90 percent confidence intervfal for p. We assume that

$$\frac{\sqrt{n}(\hat{p}_n - p)}{\widehat{\text{se}}_n} \rightsquigarrow N(0,\, 1),$$

noting that if we had used the sample variance to define the estimated standard error for p we would then be guaranteed that this convergence holds. For $\alpha = 0.1$ and Φ denoting the CDF of the standard normal distribution we then define

$$z := \Phi^{-1}(1 - \alpha/2)$$

and consider

$$C_n := (\hat{p}_n - z\,\widehat{\text{se}}_n,\, \hat{p}_n + z\,\widehat{\text{se}}_n) \cap [0,\, 1].$$

Then

$$\begin{aligned} \mathbb{P}(p \in C_n) = \mathbb{P}\left(\left| \frac{\hat{p}_n - p}{\widehat{\text{se}}_n} \right| < z \right) &= 1 - 2\mathbb{P}\left(\frac{\hat{p}_n - p}{\widehat{\text{se}}_n} > z \right) \\ &\to 1 - 2(1 - \Phi(z)) = 1 - \alpha \end{aligned}$$

as desired.

We now turn our attention to estimating $p - q$. First we find the plug-in estimator for $p - q$. We proceed as above, writing F for the CDF of the Bernoulli(p) distribution and G for the CDF of the Bernoulli(q) distribution such that

$$p - q = F(1) - F(0) - G(1) + G(0)$$

and so the plug-in estimator for $p - q$ is

$$\hat{p}_n - \hat{q}_n = \widehat{F}_n(1) - \widehat{F}_n(0) + \widehat{G}_n(0) - \widehat{G}_n(1) = \overline{X}_n - \overline{Y}_n.$$

We can then determine its estimated standard as follows. First we compute that, by independence,

$$\mathrm{se}^2(\hat{p}_n - \hat{q}_n) = \mathbb{V}(\overline{X}_n - \overline{Y}_n) = \mathbb{V}(\overline{X}_n) + \mathbb{V}(\overline{Y}_n) = \frac{\mathbb{V}(X) + \mathbb{V}(Y)}{n}$$

and so we may once again use the plug-in estimator and consider

$$\widehat{\mathrm{se}}_n = \sqrt{\frac{1}{n}\left[\hat{p}_n(1 - \hat{p}_n) + \hat{q}_n(1 - \hat{q}_n)\right]}.$$

Finally we proceed as above to find an approximate 90 percent confidence interval for $p - q$ and define

$$D_n := (\hat{p}_n - \hat{q}_n - z\widehat{\mathrm{se}}_n,\, \hat{p}_n - \hat{q}_n + z\widehat{\mathrm{se}}_n) \cap [0,\, 1]$$

such that

$$\mathbb{P}(p - q \in D_n) \to 1 - \alpha.$$

Exercise A.7.3 (Central limit theorem for the empirical CDF)**.** Let $X_1, \ldots, X_n \sim F$ and let $\widehat{F}_n(x)$ be the empirical distribution function. For a fixed x, use the central limit theorem to find the limiting distribution of $\widehat{F}_n(x)$.

Solution. Fix $x \in \mathbb{R}$. We note that if we define

$$Y_i := I(X_i \leqslant x)$$

then $Y_1, \ldots, Y_n$ are IID Bernoulli(p) random variables for $p = \mathbb{P}(X_i \leqslant x) = F(x)$. Therefore

$$\widehat{F}_n(x) = \overline{Y}_n,$$

i.e. the empirical distribution function is precisely the sample mean of the Y_i's. Since

$$\mathbb{E}(Y_i) = p = F(x) \text{ and } \mathbb{V}(Y_i) = p(1 - p) = F(x)(1 - F(x))$$

we thus deduce from the central limit theorem that

$$\widehat{F}_n(x) = \overline{Y}_n \approx N\left(F(x),\, \frac{F(x)(1 - F(x))}{n}\right).$$

More precisely:

$$\frac{\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right)}{F(x)(1 - F(x))} \rightsquigarrow N(0,\, 1).$$

Exercise A.7.4 (Covariance of the empirical CDF)**.** Let x and y be two distinct points. Find $\mathrm{Cov}(\widehat{F}_n(x),\, \widehat{F}_n(y))$.

Solution. First we compute that

$$\operatorname{Cov}\left(\widehat{F}_n(x),\, \widehat{F}_n(y)\right) = \frac{1}{n^2}\sum_{i,j=1}^{n} \operatorname{Cov}\left(\mathbb{P}(X_i \leqslant x),\, \mathbb{P}(X_j \leqslant x)\right).$$

Proceeding as in Exercise A.7.1 we see that

$$\mathbb{E}(\mathbb{P}(X_i \leqslant x)) = \mathbb{P}(X_i \leqslant x) = F(x)$$

while, if $i \neq j$, independence tells us that

$$\begin{aligned}\mathbb{E}\left(\mathbb{P}(X_i \leqslant x)\mathbb{P}(X_j \leqslant y)\right) &= 1 \cdot \mathbb{P}(X_i \leqslant x,\, X_j \leqslant y) + 0 \cdot \mathbb{P}(X_i > x \text{ or } X_j > y)\\ &= \mathbb{P}(X_i \leqslant x)\mathbb{P}(X_j \leqslant y)\\ &= F(x)F(y),\end{aligned}$$

and, for $i = j$, we obtain similarly that

$$\begin{aligned}\mathbb{E}\left(\mathbb{P}(X_i \leqslant x)\mathbb{P}(X_i \leqslant y)\right) &= \mathbb{P}(X_i \leqslant x,\, X_i \leqslant y)\\ &= \mathbb{P}(X_i \leqslant \min\{x,\, y\})\\ &= F(\min\{x,\, y\})\end{aligned}$$

Therefore, if $i \neq j$ then

$$\operatorname{Cov}\left(\mathbb{P}(X_i \leqslant x),\, \mathbb{P}(X_j \leqslant y)\right) = F(x)F(y) - F(x)F(y) = 0$$

while if $i = j$ then

$$\begin{aligned}\operatorname{Cov}\left(\mathbb{P}(X_i \leqslant x),\, \mathbb{P}(X_i \leqslant y)\right) &= F(\min\{x,\, y\}) - F(x)F(y)\\ &= F(\min\{x,\, y\})(1 - F(\max\{x,\, y\})).\end{aligned}$$

In particular, by symmetry of the covariance we may assume without loss of generality that $x \leqslant y$, in which case

$$\operatorname{Cov}\left(\mathbb{P}(X_i \leqslant x),\, \mathbb{P}(X_j \leqslant y)\right) = F(x)(1 - F(y))\delta_{ij}.$$

So finally, for $x \leqslant y$, we conclude that

$$\operatorname{Cov}\left(\widehat{F}_n(x),\, \widehat{F}_n(y)\right) = \frac{1}{n^2}\sum_{i,j=1}^{n} F(x)(1 - F(y))\delta_{ij} = \frac{F(x)(1 - F(y))}{n}.$$

Exercise A.7.5 (Clinical trials, 1)**.** 100 people are given a standard antibiotic to treat an infection and another 100 are given a new antibiotic. In the first group, 90 people recover; in the second group, 85 people recover. Let p_1 be the probability of recovery under the standard treatment and let p_2 be the probability of recovery under the new treatment. We are interested in estimating $\theta = p_1 - p_2$. Provide an estimate, standard error, 80 percent confidence interval, and 95 percent confidence interval for θ.

Solution. The key is that this boils down to Exercise A.7.2 since we can view recovery under each of the treatments as independent Bernoulli random variables. Then

$$\hat{\theta} = \hat{p}_1 - \hat{p}_2 = \frac{90}{100} - \frac{85}{100} = 0.05$$

and, since $n = 100$,

$$\widehat{\text{se}} = \sqrt{\frac{1}{100}\left[\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)\right]} = \frac{\sqrt{87}}{200} \approx 0.047.$$

So finally the $1-\alpha$ confidence interval will be given by

$$C=\left(\hat{\theta}-z\,\widehat{\mathrm{se}},\,\hat{\theta}+z\,\widehat{\mathrm{se}}\right)$$

for $z=1-\Phi^{-1}(\alpha/2)$, where as usual Φ denotes the CDF of the standard Normal distribution. More precisely:

- for $\alpha=0.2$ this yields $\theta\in(5\pm 5.98)\%$ with 80 percent confidence and
- for $\alpha=0.05$ this yields $\theta\in(5\pm 9.14)\%$ with 95 percent confidence.

A.8. **The Bootstrap.**

Exercise A.8.1 (All three types of confidence intervals for empirical data)**.** Consider the data in Example 8.6. Find the plug-in estimate of the correlation coefficient. Estimate the standard error using the bootstrap. Find a 95 percent confidence interval using the Normal, pivotal, and percentile methods.

Solution. We are given data of the form $X_i = (Y_i,\, Z_i)$ for $i = 1,\, \dots,\, 15$. The correlation coefficient is

$$\rho(Y,\, Z) = \frac{\operatorname{Cov}(Y,\, Z)}{\sqrt{\mathbb{V}(Y)\mathbb{V}(Z)}} = \frac{\mathbb{E}\left[(Y-\mu_Y)(Z-\mu_Z)\right]}{\sqrt{\mathbb{E}\left[(Y-\mu_Y)^2\right]\mathbb{E}\left[(Z-\mu_Z)^2\right]}}.$$

To obtain the plug-in estimator we write the expression for the correlation coefficient above in terms of the CDFs F_Y, F_Z, and in terms of the *joint* CDF $F_{Y,\,Z}$ and then substitute the respective *empirical* CDFs. This yields

$$\hat{\rho}_n = \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y}_n)(Z_i - \overline{Z}_n)}{\sqrt{\frac{1}{n}\sum_{i=1}^n \left(Y_i - \overline{Y}_n\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^n \left(Z_i - \overline{Z}_n\right)^2}}.$$

To estimate the standard error using the bootstrap we compute, following the usual recipe:

$$\widehat{se}_{boot} = \sqrt{v_{boot}} = \sqrt{\frac{1}{B}\sum_{j=1}^B \left(\hat{\rho}^*_{n,\,j} - \frac{1}{B}\sum_{j=1}^B \hat{\rho}^*_{n,\,j}\right)^2} = \sqrt{\frac{1}{B}\sum_{j=1}^B \left(\rho^*_j - \overline{\rho}^*\right)^2}$$

where

$$\rho^*_j := \hat{\rho}^*_{n,\,j} \text{ and } \overline{\rho}^* := \frac{1}{B}\sum_{j=1}^B \hat{\rho}^*_j.$$

Finally, in order to compute the various bootstrap confidence intervals we use the usual recipes once more.

- The $1-\alpha$ bootstrap Normal interval is given by
$$\hat{\rho}_n \pm z_{\alpha/2}\, \widehat{se}_{boot}$$
where $z_{\alpha/2} := \Phi^{-1}(1-\alpha/2) = S^{-1}(\alpha/2)$ for Φ denoting the CDF of a standard normal distribution and $S := 1 - \Phi$ denoting the *survival function* of that same distribution.
- The $1-\alpha$ bootstrap pivotal interval is given by
$$C_n = 2\hat{\rho}_n - \left(\hat{\rho}^*_{1-\alpha/2}, \hat{\rho}^*_{\alpha/2}\right)$$
where $\hat{\rho}^*_\beta$ denotes the β–quantile of the bootstrap replications $\hat{\rho}^*_1,\, \dots,\, \hat{\rho}^*_B$.
- The $1-\alpha$ bootstrap percentile interval is given by
$$C_n = \left(\hat{\rho}^*_{\alpha/2}, \hat{\rho}^*_{1-\alpha/2}\right)$$
for $\hat{\rho}^*_\beta$ as above.

Exercise A.8.2 (How many distinct bootstrap samples are possible?)**.** Let $X_1,\, \dots,\, X_n$ be distinct observations (no ties). Show that there are

$$\binom{2n-1}{n}$$

distinct bootstrap samples.

Solution. This amounts to putting n indistinguishable balls into n distinct buckets (the balls correspond to the resampled data and the bucket each ball lands in corresponds to which of the *original* data values it is equal to).

This is equivalent to a "stars and stripes" problem: there are n stars (corresponding to the balls) and $n-1$ stripes (corresponding to the bucket *dividers*) to arrange in a line, where the stars are (still) indistinguishable and now so are the stripes. There are then

$$\frac{\overbrace{[n+(n-1)]!}^{\text{arranging the stars } \textit{and} \text{ stripes}}}{\underbrace{n!}_{\text{indistinguishable stars}}\underbrace{(n-1)!}_{\text{indistinguishable stripes}}} = \frac{(2n-1)!}{n![(2n-1)-n]!} = \binom{2n-1}{n}$$

such arrangements, as desired.

Exercise A.8.3 (Expectation and variance of the bootstrap sample mean)**.** Let $X_1, \ldots, X_n$ be distinct observations (no ties). Let $X_1^*, \ldots, X_n^*$ denote a bootstrap sample and let $\overline{X}_n^* := \frac{1}{n}\sum_{i=1}^n X_i^*$. Find

(1) $\mathbb{E}\left(\overline{X}_n^* \,\middle|\, X_1, \ldots, X_n\right)$,
(2) $\mathbb{V}\left(\overline{X}_n^* \,\middle|\, X_1, \ldots, X_n\right)$,
(3) $\mathbb{E}\left(\overline{X}_n^*\right)$, and
(4) $\mathbb{V}\left(\overline{X}_n^*\right)$.

Solution. (1) By linearity of conditional expectation we compute that

$$\mathbb{E}\left(\overline{X}_n^* \,\middle|\, X_1, \ldots, X_n\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(X_i^* \,|\, X_1, \ldots, X_n\right)$$

where

$$\mathbb{E}\left(X_i^* \,|\, X_1, \ldots, X_n\right) = \sum_{j=1}^n X_j \underbrace{\mathbb{P}\left(X_i^* = X_j\right)}_{=1/n} = \overline{X}_n.$$

Therefore

$$\mathbb{E}\left(\overline{X}_n^* \,\middle|\, X_1, \ldots, X_n\right) = \overline{X}_n.$$

(2) The key observation is that, *conditioned* on $X_1, \ldots, X_n$, any two bootstrap resamples X_i^* and X_j^* are independent when $i \neq j$ (by construction of the

bootstrap). Therefore, using item 1, we deduce that

$$\begin{aligned}\mathbb{V}\left(\overline{X}_n^* \,\middle|\, X_1,\,\dots,\,X_n\right) &= \operatorname{Cov}\left(\frac{1}{n}\sum_{i=1}^n X_i^*,\,\frac{1}{n}\sum_{j=1}^n X_j^* \,\middle|\, X_1,\,\dots,\,X_n\right)\\ &= \frac{1}{n^2}\sum_{i,j=1}^n \underbrace{\operatorname{Cov}\left(X_i^*,\,X_j^* \,\middle|\, X_1,\,\dots,\,X_n\right)}_{=0\text{ when }i\neq j}\\ &= \frac{1}{n^2}\sum_{i=1}^n \operatorname{Cov}\left(X_i^*,\,X_i^* \,|\, X_1,\,\dots,\,X_n\right)\\ &= \frac{1}{n}\mathbb{V}\left(X_1^* \,|\, X_1,\,\dots,\,X_n\right)\\ &= \frac{1}{n}\mathbb{E}\left[\left(X_1^*-\overline{X}_n\right)^2 \,\middle|\, X_1,\,\dots,\,X_n\right]\\ &= \frac{1}{n}\sum_{i=1}^n\left(X_i-\overline{X}_n\right)^2\underbrace{\mathbb{P}\left(X_1^*=X_i\right)}_{=1/n}\\ &= \frac{1}{n^2}\sum_{i=1}^n\left(X_i-\overline{X}_n\right)^2.\end{aligned}$$

We note that for $S_P^2 := \frac{1}{n}\sum_{i=1}^n\left(X_i-\overline{X}_n\right)^2$ denoting the *population* variance, this means that

$$\mathbb{V}\left(\overline{X}_n^* \,\middle|\, X_1,\,\dots,\,X_n\right) = \frac{1}{n}S_P^2.$$

(3) By linearity of the expectation and the rule of iterated expectation we deduce immediately from item 1 that

$$\mathbb{E}\left(\overline{X}_n^*\right) = \mathbb{E}\left[\mathbb{E}\left(\overline{X}_n^* \,\middle|\, X_1,\,\dots,\,X_n\right)\right] = \mathbb{E}\left[\overline{X}_n\right] = \mathbb{E}(X) =: \mu.$$

(4) The law of total variance saves the day: using items 1 and 3 tells us that

$$\mathbb{V}\left(\overline{X}_n\right) = \mathbb{E}\mathbb{V}\left(\overline{X}_n^* \,\middle|\, X_1,\,\dots,\,X_n\right) + \mathbb{V}\mathbb{E}\left(\overline{X}_n^* \,\middle|\, X_1,\,\dots,\,X_n\right) = \frac{1}{n}\mathbb{E}S_P^2 + \mathbb{V}\overline{X}_n.$$

We recall that the population variance is a *biased* estimator of the variance. More precisely: for the *sample* variance

$$S^2 := \frac{1}{n-1}\sum_{i=1}^n\left(X_i-\overline{X}_n\right)^2 = \frac{n}{n-1}S_P^2,$$

we have that $\mathbb{E}S^2 = \mathbb{E}(X-\mu)^2 =: \sigma^2$ and so, since Exercise A.3.8 tells us that $\mathbb{V}\overline{X}_n = \frac{\sigma^2}{n}$,

$$\mathbb{V}\overline{X}_n^* = \frac{1}{n}\cdot\frac{n-1}{n}\cdot\mathbb{E}S^2 + \frac{\sigma^2}{n} = \frac{(n-1)+n}{n^2}\sigma^2 = \frac{2n-1}{n^2}\sigma^2.$$

In particular we note that $\mathbb{V}\overline{X}_n^* \approx 2\mathbb{V}\overline{X}_n$, i.e. the variance of the *bootstrap* sample mean is about twice as large as the variance of the *original* sample mean.

Exercise A.8.4 (Poor performance of the bootstrap)**.** Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$ and let $\hat{\theta} := X_{\max} := \max\{X_1, \ldots, X_n\}$. Show that the bootstrap performs quite poorly since $\mathbb{P}(\hat{\theta}^*) = \hat{\theta} \to 1 - e^{-1} \approx 0.632$ even though $\mathbb{P}(\hat{\theta} = \theta) = 0$. Here $\hat{\theta}^*$ denotes the maximum value obtained after *one* bootstrap resampling, i.e. $\hat{\theta}^* := \max\{X_1^*, \ldots, X_n^*\}$.

Solution. First we show that $\mathbb{P}(\hat{\theta} = \theta) = 0$. We compute that

$$\mathbb{P}(\hat{\theta} = \theta) = \mathbb{P}(\max\{X_1, \ldots, X_n\} = \theta) = \mathbb{P}(X_i = \theta \text{ for some } i) \leqslant \sum_i \underbrace{\mathbb{P}(X_i = \theta)}_{=0} = 0.$$

Now we show that $\mathbb{P}(\hat{\theta}^* = \hat{\theta}) = 1 - \left(1 - \frac{1}{n}\right)^n$. We compute that

$$\begin{aligned}
\mathbb{P}\left(\hat{\theta}^* = \hat{\theta}\right) &= \mathbb{E}\left[\mathbb{P}\left(\max_j X_j^* = \max X_i \,\middle|\, X_1, \ldots, X_n\right)\right] \\
&= \mathbb{E}\left[1 - \mathbb{P}\left(\max_j X_j^* < \max X_i \,\middle|\, X_1, \ldots, X_n\right)\right] \\
&= \mathbb{E}\left[1 - \mathbb{P}\left(X_j^* < \max X_i \text{ for all } j \,|\, X_1, \ldots, X_n\right)\right]
\end{aligned}$$

and so, by the conditional independence of $X_j^* \,|\, X_1, \ldots, X_n$,

$$\begin{aligned}
\mathbb{P}\left(\hat{\theta}^* = \hat{\theta}\right) &= \mathbb{E}\left[1 - \mathbb{P}\left(X_1^* < \max X_i \,|\, X_1, \ldots, X_n\right)^n\right] \\
&= \mathbb{E}\left[1 - \left(1 - \mathbb{P}\left(X_1^* = \max X_i \,|\, X_1, \ldots, X_n\right)\right)^n\right] \\
&= \mathbb{E}\left[1 - \left(1 - \frac{1}{n}\right)^n\right] \\
&= 1 - \left(1 + \frac{(-1)}{n}\right)^n.
\end{aligned}$$

So finally:

$$\mathbb{P}\left(\hat{\theta}^* = \hat{\theta}\right) = 1 - \left(1 + \frac{(-1)}{n}\right)^n \to 1 - e^{-1} \approx 0.632 \text{ as } n \to \infty,$$

as desired.

Exercise A.8.5 (Exact formula for a bootstrap variance)**.** Consider $T_n := \overline{X}_n^2$ and define $\hat{\alpha}_k := \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^k$ for $k \geqslant 1$. Verify that, for $v_{boot} := \mathbb{V}_n(T_n^*)$, where $\mathbb{V}_n$ denotes variance with respect to the empirical measure $\hat{F}_n$,

$$v_{boot} = \frac{4\overline{X}_n^2 \hat{\alpha}_2}{n} + \frac{4\overline{X}_n \hat{\alpha}_3}{n^2} + \frac{\hat{\alpha}_4 + (2n-3)\hat{\alpha}_2^2}{n^3}.$$

Remark A.2. Previously, for example in Exercise A.8.3, we used the notation

$$\mathbb{V}\left(T_n^* \,|\, X_1, \ldots, X_n\right)$$

to mean

$$\mathbb{V}_{\hat{F}_n}\left(T_n^*\right).$$

Here we will write, instead,

$$\mathbb{V}_n\left(T_n^*\right).$$

Solution. We begin by observing that

$$\mathbb{V}_n\left(T_n^*\right)=\mathbb{E}_n\left[\left(T_n^*\right)^2\right]-\left(\mathbb{E}_n T_n^*\right)^2.$$

We have actually already established, in Exercise A.8.3, identities that let us compute the latter of these two terms. Indeed:

$$\begin{aligned}\mathbb{E}_n T_n^* &= \mathbb{E}_n \overline{X}_n^2\\ &= \mathbb{V}_n \overline{X}_n + \left(\mathbb{E}_n \overline{X}_n\right)^2\\ &= \frac{1}{n^2}\sum_{i=1}^n \left(X_i-\overline{X}_n\right)^2+\overline{X}_n^2\\ &= \frac{\hat{\alpha}_2}{n}+\overline{X}_n^2\end{aligned}$$

and so

$$\left(\mathbb{E}_n T_n^*\right)^2=\overline{X}_n^4+\frac{2\overline{X}_n^2\hat{\alpha}_2}{n}+\frac{\hat{\alpha}_2^2}{n^2}.$$

We now turn our attention to the first term, noting that

$$\mathbb{E}_n\left[\left(T_n^*\right)^2\right]=\mathbb{E}_n\left[\left(\overline{X}_n^*\right)^4\right].$$

In order to reduce this term to expressions like $\hat{\alpha}_k$ we will write

$$\overline{X}_n^*=\underbrace{\frac{1}{n}\sum_{i=1}^n\left(X_i^*-\overline{X}_n\right)}_{=:S_n}+\overline{X}_n=S_n+\overline{X}_n.$$

Crucially: $\overline{X}_n$ can be pulled out of expectations with respect to $\hat{F}_n$ (since, in that context, $\overline{X}_n$ is *constant*) while S_n will give rise to terms like $\hat{\alpha}_k$. (More generally, note that writing Y^k as $[(Y-\mathbb{E}Y)+\mathbb{E}Y]^k$, as is done here, is a common trick.) In particular

$$\begin{aligned}\mathbb{E}_n\left[\left(T_n^*\right)^2\right] &= \mathbb{E}_n\left[\left(\overline{X}_n^*\right)^4\right]\\ &= \mathbb{E}_n\left[\left(S_n+\overline{X}_n\right)^4\right]\\ &= \mathbb{E}_n S_n^4+4\overline{X}_n\mathbb{E}_n S_n^2+6\overline{X}_n^2\mathbb{E}_n S_n^2+4\overline{X}_n^3\mathbb{E}_n S_n+\overline{X}_n^4.\end{aligned}$$

We note that, by construction of S_n and by Exercise A.8.3,

$$\mathbb{E}_n S_n=\frac{1}{n}\sum_{i=1}^n\left[\underbrace{\left(\mathbb{E}_n X_i^*\right)}_{=\overline{X}_n}-\overline{X}_n\right]=0.$$

We now turn our attention to the computation of $\mathbb{E}_n S_n^k$ for $k=2$, 3, and 4.

First we note that

$$\begin{aligned}\mathbb{E}_n S_n^2 &= \frac{1}{n^2}\mathbb{E}_n\left[\left(\sum_{i=1}^n X_i^* - \overline{X}_n\right)^2\right]\\ &= \frac{1}{n^2}\mathbb{E}_n \sum_{i,j=1}^n \left(X_i^* - \overline{X}_n\right)\left(X_j^* - \overline{X}_n\right)\\ &= \frac{1}{n^2}\sum_{i,j=1}^n \underbrace{\mathbb{E}_n\left[\left(X_i^* - \overline{X}_n\right)\left(X_j^* - \overline{X}_n\right)\right]}_{=:B_{ij}}.\end{aligned}$$

Crucially, when $i \neq j$, by conditional independence of

$$X_i^* \,|\, X_1, \ldots, X_n \text{ and } X_j^* \,|\, X_1, \ldots, X_n$$

and since their means are both equal to $\overline{X}_n$, we deduce that $B_{ij} = 0$. Therefore

$$\mathbb{E}_n S_n^2 = \frac{1}{n^2}\sum_{i=1}^n B_{ii}$$

where

$$\begin{aligned}B_{ii} &= \mathbb{E}_n\left[\left(X_i^* - \overline{X}_n\right)^2\right]\\ &= \sum_{a=1}^n \left(X_a - \overline{X}_n\right)^2 \mathbb{P}\left(X_i^* = X_a\right)\\ &= \frac{1}{n}\sum_{a=1}^n \left(X_a - \overline{X}_n\right)^2\\ &= \hat{\alpha}_2\end{aligned}$$

and so

$$\mathbb{E}_n S_n^2 = \frac{1}{n^2}\sum_{i=1}^n \hat{\alpha}_2 = \frac{\hat{\alpha}_2}{n}.$$

We now evaluate $\mathbb{E}_n S_n^3$. Proceeding as above we see that

$$\begin{aligned}\mathbb{E}_n S_n^3 &= \frac{1}{n^3}\mathbb{E}_n\left[\left(\sum_{i=1}^n X_i^* - \overline{X}_n\right)^3\right]\\ &= \frac{1}{n^3}\sum_{i,j,k=1}^n \underbrace{\mathbb{E}_n\left[\left(X_i^* - \overline{X}_n\right)\left(X_j^* - \overline{X}_n\right)\left(X_k^* - \overline{X}_n\right)\right]}_{=:C_{ijk}}.\end{aligned}$$

To evaluate C_{ijk} we split into three cases.

- Case 1: All three indices are the same. Then

$$\begin{aligned} C_{iii} &= \mathbb{E}\left[\left(X_i^* - \overline{X}_n\right)^3\right] \\ &= \sum_{a=1}^{n} \left(X_a - \overline{X}_n\right)^3 \mathbb{P}\left(X_i^* = X_a\right) \\ &= \frac{1}{n}\sum_{a=1}^{n} \left(X_a - \overline{X}_n\right)^3 \\ &= \hat{\alpha}_3. \end{aligned}$$

- Case 2: One index is repeated twice and is distinct from the third index. Then, by mutual conditional independence,

$$\begin{aligned} C_{iij} &= \mathbb{E}_n\left[\left(X_i^* - \overline{X}_n\right)^2\left(X_j^* - \overline{X}_n\right)\right] \\ &= \mathbb{E}_n\left[\left(X_i^* - \overline{X}_n\right)^2\right]\underbrace{\mathbb{E}_n\left(X_j^* - \overline{X}_n\right)}_{=0} \\ &= 0. \end{aligned}$$

- Case 3: All three indices are distinct. Then, once again using mutual conditional independence,

$$C_{ijk} = \mathbb{E}_n\left(X_i^* - \overline{X}_n\right)\mathbb{E}_n\left(X_j^* - \overline{X}_n\right)\mathbb{E}_n\left(X_k^* - \overline{X}_n\right) = 0 \cdot 0 \cdot 0 = 0.$$

So finally, putting these three cases together we deduce that

$$\mathbb{E}_n S_n^3 = \frac{1}{n^3}\sum_{i=1}^{n} C_{iii} = \frac{\hat{\alpha}_3}{n^2}.$$

Finally we evaluate $\mathbb{E}_n S_n^4$, noting that

$$\begin{aligned} \mathbb{E}_n S_n^4 &= \frac{1}{n^4}\mathbb{E}_n\left[\left(\sum_{i=1}^{n} X_i^* - \overline{X}_n\right)^4\right] \\ &= \frac{1}{n^4}\sum_{i,j,k,l=1}^{n} \underbrace{\mathbb{E}_n\left[\left(X_i^* - \overline{X}_n\right)\left(X_j^* - \overline{X}_n\right)\left(X_k^* - \overline{X}_n\right)\left(X_l^* - \overline{X}_n\right)\right]}_{=:D_{ijkl}}. \end{aligned}$$

To evaluate D_{ijkl} we split into five cases.

- Case 1: All four indices agree. Then

$$\begin{aligned} D_{iiii} &= \mathbb{E}_n\left[\left(X_i^* - \overline{X}_n\right)^4\right] \\ &= \sum_{a=1}^{n} \left(X_a - \overline{X}_n\right)^4 \mathbb{P}\left(X_i^* = X_a\right) \\ &= \frac{1}{n}\sum_{a=1}^{n} \left(X_a - \overline{X}_n\right)^4 \\ &= \hat{\alpha}_4. \end{aligned}$$

- Case 2: One index is repeated three times and is distinct from the fourth index. Then, by mutual conditional independence,

$$D_{iiij} = \mathbb{E}_n \left[\left(X_i^* - \overline{X}_n \right)^3 \right] \underbrace{\mathbb{E}_n \left(X_j^* - \overline{X}_n \right)}_{=0} = 0.$$

- Case 3: Two indices are repeated twice each. Then mutual conditional independence tells us that

$$\begin{aligned} D_{iijj} &= \mathbb{E}_n \left[\left(X_i^* - \overline{X}_n \right)^2 \right] \mathbb{E}_n \left[\left(X_j^* - \overline{X}_n \right)^2 \right] \\ &= B_{ii} B_{jj} \\ &= \hat{\alpha}_2^2. \end{aligned}$$

- Case 4: One index is repeated twice and is distinct from the pairwise distinct remaining indices. In this case mutual conditional independence tells us that

$$D_{iijk} = \mathbb{E}_n \left[\left(X_i^* - \overline{X}_n \right)^2 \right] \underbrace{\mathbb{E}_n \left(X_j^* - \overline{X}_n \right)}_{=0} \underbrace{\mathbb{E}_n \left(X_k^* - \overline{X}_n \right)}_{=0} = 0.$$

- Case 5: All four indices are distinct, in which case mutual conditional independence tells us that

$$\begin{aligned} D_{ijkl} &= \mathbb{E}_n \left(X_i^* - \overline{X}_n \right) \mathbb{E}_n \left(X_j^* - \overline{X}_n \right) \mathbb{E}_n \left(X_k^* - \overline{X}_n \right) \mathbb{E}_n \left(X_l^* - \overline{X}_n \right) \\ &= 0 \cdot 0 \cdot 0 \cdot 0 \\ &= 0. \end{aligned}$$

Putting all five cases together we deduce that, since there are $\frac{4!}{2!2!} = 6$ ways to rearrange the "word" $iijj$,

$$\mathbb{E}_n S_n^4 = \frac{1}{n^4} \left(\sum_{i=1}^n D_{iiii} + \sum_{1 \leqslant i < j \leqslant n} 6 D_{iijj} \right).$$

Since there are $\binom{n}{2} = \frac{n(n-1)}{2}$ ways to choose i and j such that $1 \leqslant i < j \leqslant n$ this simplifies to

$$\begin{aligned} \mathbb{E}_n S_n^4 &= \frac{1}{n^4} \sum_{i=1}^n \hat{\alpha}_4 + \frac{6}{n^4} \sum_{1 \leqslant i < j \leqslant n} \hat{\alpha}_2^2 \\ &= \frac{\hat{\alpha}_4}{n^3} + \frac{6}{n^4} \cdot \frac{n(n-1)}{2} \cdot \hat{\alpha}_2^2 \\ &= \frac{\hat{\alpha}_4}{n^3} + \frac{3(n-1)\hat{\alpha}_2^2}{n^3}. \end{aligned}$$

We are now ready to conclude. Putting together all of the intermediate computations above we obtain that

$$\begin{aligned}
\mathbb{V}_n\left(T_n^*\right) &= \mathbb{E}_n\left[\left(T_n^*\right)^2\right] - \left(\mathbb{E}_n T_n^*\right)^2 \\
&= \mathbb{E}_n S_n^4 + 4\overline{X}_n \mathbb{E}_n S_n^3 + 6\overline{X}_n^2 \mathbb{E}_n S_n^2 + 4\overline{X}_n^3 \mathbb{E}_n S_n + \overline{X}_n^4 - \overline{X}_n^4 - \frac{2\overline{X}_n^2 \hat{\alpha}_2}{n} - \frac{\hat{\alpha}_2^2}{n^2} \\
&= \frac{\hat{\alpha}_4}{n^3} + \frac{3(n-1)\hat{\alpha}_2^2}{n^3} + \frac{4\overline{X}_n \hat{\alpha}_3}{n^2} + \frac{6\overline{X}_n^2 \hat{\alpha}_2}{n} + 0 - \frac{2\overline{X}_n^2 \hat{\alpha}_2}{n} - \frac{\hat{\alpha}_2^2}{n^2} \\
&= \frac{\hat{\alpha}_4}{n^3} + \frac{(2n-3)\hat{\alpha}_2^2}{n^3} + \frac{4\overline{X}_n \hat{\alpha}_3}{n^2} + \frac{4\overline{X}_n^2 \hat{\alpha}_2}{n},
\end{aligned}$$

as desired.

A.9. **Parametric Inference.**

Exercise A.9.1 (Method of moments for the Gamma distribution)**.** Consider random variables $X_1, \ldots, X_n \sim \text{Gamma}(\alpha, \beta)$. Find the method of moments estimator for α and β.

Solution. Recall that if $X \sim \text{Gamma}(\alpha, \beta)$ then

$$\mathbb{E}X = \alpha\beta \text{ and } \mathbb{V}X = \alpha\beta^2$$

and so

$$\mathbb{E}(X^2) = \mathbb{V}X + (\mathbb{E}X)^2 = \alpha\beta^2 + \alpha^2\beta^2 = \alpha\beta(1+\alpha).$$

We are thus looking for $\hat{\alpha}_n$ and $\hat{\beta}_n$ to be the solutions of

$$\alpha\beta = \overline{X}_n =: \hat{m}_1 \text{ and } \alpha\beta^2(1+\alpha) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 =: \hat{m}_2.$$

The first equation tells us that

$$\beta = \frac{\hat{m}_1}{\alpha}.$$

Plugging this into the second equation tells us that

$$\alpha\frac{\hat{m}_1^2}{\alpha^2}(1+\alpha) = \hat{m}_2 \Leftrightarrow \frac{1+\alpha}{\alpha} = \frac{\hat{m}_2}{\hat{m}_1^2} \Leftrightarrow \alpha = \frac{1}{\frac{\hat{m}_2}{\hat{m}_1^2} - 1} = \frac{\hat{m}_1^2}{\hat{m}_2 - \hat{m}_1^2}$$

and so

$$\beta = \frac{\hat{m}_2 - \hat{m}_1^2}{\hat{m}_1}.$$

In other words:

$$\hat{\alpha}_n = \frac{\hat{m}_1^2}{\hat{m}_2 - \hat{m}_1^2} \text{ and } \hat{\beta}_n = \frac{\hat{m}_2 - \hat{m}_1^2}{\hat{m}_1}.$$

Note that, if we write $\hat{\mu} := \overline{X}_n$ and notice that, for $\hat{\sigma}_P^2$ denoting the *population* variance, the identity $\hat{\sigma}_P^2 = \hat{m}_2 - \hat{m}_1^2$ holds, then we can rewrite

$$\hat{\alpha}_n = \frac{\hat{\mu}^2}{\hat{\sigma}_P^2} \text{ and } \hat{\beta}_n = \frac{\hat{\sigma}_P^2}{\hat{\mu}}.$$

This is not surprising since

$$\alpha = \frac{(\mathbb{E}X)^2}{\mathbb{V}X} \text{ and } \beta = \frac{\mathbb{V}X}{\mathbb{E}X}.$$

Exercise A.9.2 (Estimators for uniform distributions)**.** Let $X_1, \ldots, X_n \sim \text{Uniform}(a, b)$.

(1) Find the method of moments estimator for a and b.
(2) Find the MLE $\hat{a}$ and $\hat{b}$.
(3) Let $\tau := \int x dF(x)$. Find the MLE of τ.
(4) Let $\hat{\tau}$ be the MLE of τ. Let $\tilde{\tau}$ be the nonparametric plug-in estimator of $\tau = \int xF(x)$. Suppose that $a = 1$, $b = 3$, and $n = 10$. Find the MSE of $\hat{\tau}$ by simulation. Find the MSE of $\tilde{\tau}$ analytically. Compare.

Solution. (1) For $X \sim \text{Uniform}(a, b)$,

$$\mathbb{E}X = \frac{a+b}{2} \text{ and } \mathbb{V}X = \frac{(b-a)^2}{12}$$

and hence

$$\mathbb{E}(X^2) = \mathbb{V}X + (\mathbb{E}X)^2 = \frac{(b-a)^2}{12} + \frac{(a+b)^2}{4}.$$

Introducing

$$s := \frac{a+b}{2} \text{ and } d := \frac{b-a}{2}$$

such that

$$a = s + d \text{ and } b = s - d$$

we are looking to solve

$$s = \overline{X}_n =: m_1 \text{ and } s^2 + \frac{d^2}{3} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 =: m_2.$$

Therefore, since d must be positive,

$$s = m_1 \text{ and } d = \sqrt{3(m_2 - s^2)}.$$

So, finally,

$$\begin{aligned} \hat{a}_n &= s + d = m_1 + \sqrt{3(m_2 - m_1^2)} \text{ and} \\ \hat{b}_n &= s - d = m_1 - \sqrt{3(m_2 - m_1^2)}. \end{aligned}$$

In particular, for $\hat{\mu} := \overline{X}_n$ and $\hat{\sigma}_P^2$ denoting the population variance, such that $\hat{\sigma}_P^2 = m_2 - m_1^2$, we have that

$$\hat{a}_n = \hat{\mu} + \sqrt{3\hat{\sigma}_P^2} \text{ and } \hat{b}_n = \hat{\mu} - \sqrt{3\hat{\sigma}_P^2},$$

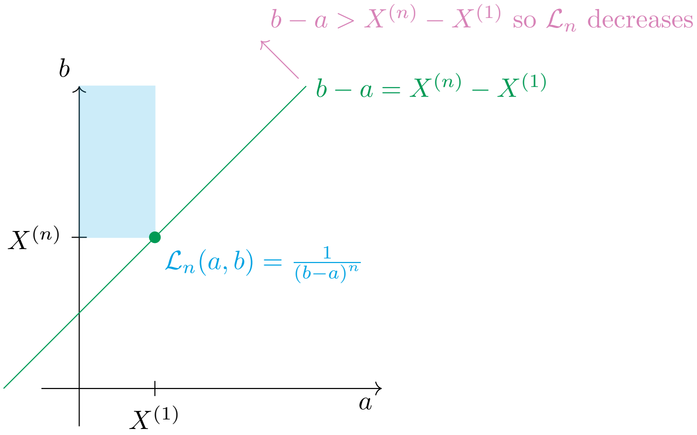
which makes sense since

$$a = \mu + \sqrt{3\sigma^2} \text{ and } b = \mu - \sqrt{3\sigma^2}.$$

(2) The likelihood function is given by

$$\begin{aligned} \mathcal{L}_n = \prod_{i=1}^{n} f\left(X_i; a, b\right) &= \prod_{i=1}^{n} \left[\frac{1}{b-a}\mathbb{1}(a \leqslant X_i \leqslant b)\right] \\ &= \frac{1}{(b-a)^n}\mathbb{1}(a \leqslant X_i \leqslant b \text{ for all } i) \\ &= \frac{1}{(b-a)^n}\mathbb{1}(a \leqslant X^{(1)} \text{ and } b \geqslant X^{(n)}) \end{aligned}$$

where $X^{(1)}$, ..., $X^{(n)}$ denote the *increasing reordering* of X_1, ..., X_n.


$b - a > X^{(n)} - X^{(1)}$ so $\mathcal{L}_n$ decreases
$b - a = X^{(n)} - X^{(1)}$
b
$X^{(n)}$
$\mathcal{L}_n(a, b) = \frac{1}{(b-a)^n}$
$X^{(1)}$
a

Therefore, as shown in the picture above, $\mathcal{L}_n$ attains its maximum at

$$\hat{a} = X^{(1)} \text{ and } \hat{b} = X^{(n)},$$

which are thus the MLE.

(3) By equivariance of the MLE, and since we have derived above the MLE for the parameters a and b, it suffices to write τ in terms of these parameters. Since

$$\tau = \int x dF(x) = \mathbb{E}X = \frac{a+b}{2}$$

it follows from equivariance that

$$\hat{\tau} = \frac{\hat{a} + \hat{b}}{2} = \frac{X^{(1)} + X^{(n)}}{2}.$$

(4) Since $\tau = \int x dF(x) = \mathbb{E}X$ it follows that its plug-in estimator is

$$\tilde{\tau} = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}_n.$$

Therefore

$$\begin{aligned} \text{MSE}(\tilde{\tau}) = \mathbb{E}(\tilde{\tau} - \tau)^2 &= \mathbb{V}\overline{X}_n \\ &= \frac{\mathbb{V}X}{n} \text{ by Exercise A.3.8} \\ &= \frac{(b-a)^2}{12n}. \end{aligned}$$

In particular, when $a = 1$, $b = 3$, and $n = 10$,

$$MSE(\tilde{\tau}) = \frac{4}{120} = \frac{1}{30} \approx 0.033.$$

By contrast, simulations show that

$$MSE(\hat{\tau}) \approx 0.015$$

We note that the MLE has lower mean-squared error than the plug-in estimator.

Exercise A.9.3 (Consistency of the MLE for uniform distributions)**.** Let $X_1, \ldots\ X_n \sim$ Uniform$(0, \theta)$. Show that the MLE is consistent.

Solution. Recall from Exercise A.9.2 that the MLE is $\hat{\theta}_n = \max\{X_1, \ldots\ X_n\}$. So we wish to show that, for every $\varepsilon > 0$,

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) \to 0 \text{ as } n \to \infty.$$

Note that the bound $\hat{\theta}_n \leqslant \theta$ always holds since θ is an upper bound for all of the X_i's. Note also that we may without loss of generality work with $0 < \varepsilon < \theta$ since the matter is trivial otherwise. So, for $0 < \varepsilon < \theta$, we may compute that

$$\begin{aligned}\mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = \mathbb{P}\left(\hat{\theta}_n < \theta - \varepsilon\right) &= \mathbb{P}\left(X_i < \theta - \varepsilon \text{ for all } i\right) \\ &= \mathbb{P}\left(X_1 < \theta - \varepsilon\right)^n = \left(\frac{\theta - \varepsilon}{\theta}\right)^n.\end{aligned}$$

Since $0 < \frac{\theta - \varepsilon}{\theta} < 1$, the claim follows.

Exercise A.9.4 (Estimators for Poisson distributions)**.** Let $X_1, \ldots\ X_n \sim \text{Poisson}(\lambda)$. Find the method of moments estimator, the maximum likelihood estimator, and the Fisher information $I(\lambda)$.

Solution. First we find the method of moments estimator. Since the mean of a Poisson(λ) random variable is precisely λ, the method of moments estimator $\tilde{\lambda}$ is simply $\tilde{\lambda} = \overline{X}_n$.

Now we turn our attention to the maximum likelihood estimator. We compute that the likelihood function is

$$\mathcal{L}_n = \prod_{i=1}^n f(X_i; \theta) = \prod_{i=1}^n \left(e^{-\lambda} \frac{\lambda^{X_i}}{X_i!}\right) = e^{-n\lambda} \lambda^{n\overline{X}_n} \prod_{i=1}^n \frac{1}{X_i!}$$

and so, up to an additive constant, the log-likelihood function is given by

$$l_n = -n\lambda + n\overline{X}_n \log \lambda + C.$$

Therefore

$$l_n' = -n + \frac{n\overline{X}_n}{\lambda}$$

such that

$$l_n'(\lambda) = 0 \iff \lambda = \overline{X}_n.$$

In other words: the MLE $\hat{\lambda}$ is *also* the sample mean, i.e. $\hat{\lambda} = \overline{X}_n$.

Finally we turn our attention to the Fisher information. First, we recall that $I(\lambda) = -\mathbb{E}\left[\frac{\partial^2 \log f}{\partial \lambda^2}\right]$. In particular we note that $\log f = l_1$ (the log-likelihood with $n = 1$) and so we have already computed above that

$$\partial_\lambda(\log f) = l_1' = -1 + \frac{x}{\lambda}.$$

Therefore

$$\partial_\lambda^2(\log f) = \partial_\lambda\left(-1 + \frac{x}{\lambda}\right) = \frac{-x}{\lambda^2}$$

and so the Fisher information is given by

$$I(\lambda) = -\mathbb{E}\left[\frac{-X}{\lambda^2}\right] = \frac{1}{\lambda^2}\mathbb{E}X = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}.$$

Exercise A.9.5 (Two estimators for normal distributions)**.** Let $X_1, \dots X_n \sim N(\theta, 1)$. Define

$$Y_i = \begin{cases} 1 & \text{if } X_i > 0 \text{ and} \\ 0 & \text{if } X_i \leqslant 0. \end{cases}$$

Let $\psi := \mathbb{P}(Y_1 = 1)$.

(1) Find the maximum likelihood estimator $\hat{\psi}_n$ of ψ.
(2) Find an approximate 95 percent confidence interval for ψ.
(3) Define $\tilde{\psi}_n := \frac{1}{n}\sum_{i=1}^n Y_i$. Show that $\tilde{\psi}_n$ is a consistent estimator of ψ.
(4) Compute the asymptotic relative efficiency of $\tilde{\psi}_n$ to $\hat{\psi}_n$. Hint: Use the delta method to get the standard error of the MLE. Then compute the standard error (i.e. the standard deviation) of $\tilde{\psi}_n$.
(5) Suppose that the data are not really normal. Show that $\hat{\psi}_n$ is not consistent. What, if anything, does $\hat{\psi}_n$ converge to?

Solution. (1) We wish to use the equivariance of the MLE and so we seek to write ψ as a function of θ. We compute that

$$\psi = \mathbb{P}(Y_1 = 1) = \mathbb{P}(X_1 > 0) = \mathbb{P}(\underbrace{X_1 - \theta}_{=:Z} > -\theta)$$

where Z is a standard normal. Therefore, for Ψ denoting the CDF of a standard normal distribution and for Z' denoting *another* standard normal,

$$\psi = \mathbb{P}(Z > -\theta) = \mathbb{P}(Z' < \theta) = \Phi(\theta).$$

Now we find the MLE for θ. The log-likelihood function is given by

$$\begin{aligned} l_n(\theta) &= \sum_{i=1}^n \log f(X_i; \theta) \\ &= \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(X_i - \theta)^2\right]\right) \\ &= -\sum_{i=1}^n \frac{1}{2}(X_i - \theta)^2 + C. \end{aligned}$$

Therefore

$$l_n'(\theta) = -\sum_{i=1}^n \frac{1}{2} \cdot (-2) \cdot (X_i - \theta) = \sum_{i=1}^n (X_i - \theta) = n(\overline{X}_n - \theta),$$

and so the MLE is simply the sample mean, i.e

$$\hat{\theta}_n = \overline{X}_n.$$

So finally, by equivariance, the MLE for ψ is

$$\hat{\psi}_n = \Phi(\hat{\theta}_n) = \Phi(\overline{X}_n).$$

(2) We will use the delta method, which means that we must first compute the Fisher information. Recall that

$$I(\theta) = -\mathbb{E}[\partial_\theta^2 \log f] = -\mathbb{E}[\partial_\theta l_1']$$

where l_1 is the log-likelihood for $n = 1$. We compute that

$$\partial_\theta l_1' = l_1'' = (X - \theta)' = -1,$$

which means that $I(\theta) = 1$, and that $I_n(\theta) = n$. To use the delta method we now define

$$\widehat{se}(\hat{\psi}_n) := |\Phi'(\overline{X}_n)|\widehat{se}(\overline{X}_n) = |\phi(\overline{X}_n)|\sqrt{1/I_n(\overline{X}_n)} = \frac{1}{\sqrt{n}}|\phi(\overline{X}_n)|,$$

where $\phi = \Phi'$ denotes the PDF of a standard normal distribution.

So finally Theorem 9.18 tells us that, for $\alpha = 0.05$, such that we have that $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2) \approx 1.96$,

$$C_n := \hat{\psi}_n + z_{\alpha/2}\widehat{se}\hat{\psi}_n(-1,\, 1) = \Phi(\overline{X}_n) + \frac{1.96}{\sqrt{n}}|\phi(\overline{X}_n)|(-1,\, 1)$$

is an asymptotic 95 percent confidence interval for ψ.

(3) The random variables $Y_1, \ldots\ Y_n$ are IID with finite mean, namely equal to ψ, since they are Bernoulli random variables. The estimator $\tilde{\psi}_n$, which is the sample mean of the Y_i's, thus converges to their mean ψ in probability by the Weak Law of Large Numbers (c.f. Theorem 5.6). In other words: the estimator $\tilde{\psi}_n$ is consistent.

(4) Using the delta method we computed in item 2 above that the asymptotic standard error of the MLE is

$$se(\hat{\psi}_n) = \frac{1}{\sqrt{n}}|\phi(\theta)|.$$

We now turn our attention to the standard error of $\tilde{\psi}_n$. As noted in item 3 above, the random variables $Y_1, \ldots\ Y_n$ are Bernoulli with

$$\mathbb{E}\tilde{\psi}_n = \psi \text{ and } \mathbb{V}\tilde{\psi}_n = \mathbb{V}Y_n = \frac{\mathbb{V}Y}{n}$$

where, using item 1, we see that

$$p := \mathbb{P}(Y = 1) = \mathbb{P}(X > 0) = \Phi(\theta)$$

and so

$$\mathbb{V}Y = p(1 - p) = \Phi(\theta)(1 - \Phi(\theta)).$$

So finally

$$\mathbb{V}\tilde{\psi}_n = \frac{\Phi(\theta)[1 - \Phi(\theta)]}{n}$$

and hence

$$se(\tilde{\psi}_n) = \sqrt{\mathbb{V}\tilde{\psi}_n} = \frac{1}{\sqrt{n}}\sqrt{\Phi(\theta)[1 - \Phi(\theta)]}.$$

We thus conclude that

$$ARE(\tilde{\psi}_n,\, \hat{\psi}_n) = \frac{\phi(\theta)^2}{\Phi(\theta)[1 - \Phi(\theta)]}.$$

(5) Note that, as long as the X_i's are IID with finite mean, then the Weak Law of Large Numbers tells us that

$$\hat{\theta}_n = \overline{X}_n \xrightarrow{P} \mathbb{E}X.$$

By continuity of the CDF of a standard normal distribution it follows that

$$\hat{\psi}_n = \Phi(\hat{\theta}_n) \xrightarrow{P} \Phi(\mathbb{E}X).$$

For $Z \sim N(0,\, 1)$ we compute that

$$\Phi(\mathbb{E}X) = \mathbb{P}(Z \leqslant \mathbb{E}X)$$

and so it follows that $\hat{\psi}_n$ is consistent if and only if

$$\mathbb{P}(Z \leqslant \mathbb{E}X) = \mathbb{P}(X > 0).$$

This is guaranteed to occur if $X \sim N(\theta,\, 1)$, as shown in item 1 above, but would be a rather miraculous coincidence otherwise!

Exercise A.9.6 (MLE confidence interval for the percentile of a Normal distribution)**.** Let $X_1, \ldots\, X_n \sim N(\mu,\, \sigma^2)$. Let τ be the 95th percentile, i.e. $\tau = F_X^{-1}(0.95)$ for F_X denoting the CDF of the $N(\mu,\, \sigma^2)$ distribution.

(1) Find the MLE of τ.
(2) Find an expression for an approximate $1 - \alpha$ confidence interval for τ.
(3) Suppose the data are:

3.23	−2.50	1.88	−0.68	4.43	0.17
1.03	−0.07	−0.01	0.76	1.76	3.18
0.33	−0.31	0.30	−0.61	1.52	5.43
1.54	2.28	0.42	2.33	−1.03	4.00
0.39					

Find the MLE $\hat{\tau}$. Find the standard error using the delta method. Find the standard error using the parametric bootstrap.

Solution. (1) First we find the MLE for μ and σ (assuming without loss of generality that $\sigma > 0$). The log-likelihood function is given by

$$\begin{aligned} l_n &= \sum_{i=1}^{n} \log f(X_i; \mu,\, \sigma^2) \\ &= \sum_{i=1}^{n} \log \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} (X_i - \mu)^2 \right) \right] \\ &= \sum_{i=1}^{n} \left(-\log\sigma - \frac{1}{2\sigma^2} (X_i - \mu)^2 \right) + C. \end{aligned}$$

Therefore

$$\partial_\mu l_n = \sum_{i=1}^{n} -\frac{(-2)}{2\sigma^2} (X_i - \mu) = \frac{n}{\sigma^2} (\overline{X}_n - \mu)$$

and

$$\partial_\sigma l_n = \sum_{i=1}^{n} \left[\frac{-1}{\sigma} + \frac{2}{2\sigma^3} (X_i - \mu)^2 \right] = \frac{-n}{\sigma^2} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (X_i - \mu)^2.$$

Since

$$\begin{aligned}
&\sum_{i=1}^{n} (X_i - \mu)^2 \\
&= \sum_{i=1}^{n} \left[(X_i - \overline{X}_n) + (\overline{X}_n - \mu)\right]^2 \\
&= \sum_{i=1}^{n} \left[(X_i - \overline{X}_n)^2 + 2(X_i - \overline{X}_n)(\overline{X}_n - \mu) + (\overline{X}_n - \mu)^2\right] \\
&= \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 + 2(\overline{X}_n - \mu)\underbrace{\sum_{i=1}^{n}(X_i - \overline{X}_n)}_{=0} + n(\overline{X}_n - \mu)^2 \\
&=: nS_P^2 + n(\overline{X}_n - \mu)^2
\end{aligned}$$

it follows that

$$\nabla l_n = n \begin{pmatrix} \frac{1}{\sigma^2}(\overline{X}_n - \mu) \\ -\frac{1}{\sigma} + \frac{1}{\sigma^3}\left[S_P^2 + (\overline{X}_n - \mu)^2\right] \end{pmatrix}$$

and so

$$\begin{aligned}
\nabla l_n = 0 &\Leftrightarrow \mu = \overline{X}_n \text{ and } -\frac{1}{\sigma} + \frac{S_P^2}{\sigma^3} = 0 \\
&\Leftrightarrow \mu = \overline{X}_n \text{ and } \sigma^2 = S_P^2.
\end{aligned}$$

In other words the MLE for μ is the sample mean and the MLE for σ^2 is the *population* variance.

We now seek to write τ as a function of μ and σ. We observe that, for $Z := \frac{X-\mu}{\sigma} \sim N(0,1)$ and for Φ denoting the its CDF, τ is characterized by

$$\begin{aligned}
\mathbb{P}(X < \tau) = 0.95 &\Leftrightarrow \mathbb{P}\left(\frac{X-\mu}{\sigma} < \frac{\tau-\mu}{\sigma}\right) = 0.95 \\
&\Leftrightarrow \Phi\left(\frac{X-\mu}{\sigma}\right) = 0.95 \\
&\Leftrightarrow \tau = \mu + \sigma\Phi^{-1}(0.95).
\end{aligned}$$

Therefore, by equivariance, the MLE for τ is

$$\hat{\tau} = \hat{\mu} + \hat{\sigma}\Phi^{-1}(0.95) = \overline{X}_n + \sqrt{S_P^2}\,\Phi^{-1}(0.95).$$

(2) To find an *asymptotic* $1-\alpha$ confidence interval for τ we use the multiparameter delta method. First we compute the Fisher information matrix. Since

$$\frac{1}{n}\nabla l_n = \begin{pmatrix} \frac{1}{\sigma^2}(\overline{X}_n - \mu) \\ -\frac{1}{\sigma} + \frac{1}{\sigma^3}\left[S_P^2 + (\overline{X}_n - \mu)^2\right] \end{pmatrix}$$

it follows that

$$\begin{aligned}
\frac{1}{n}\partial_\mu^2 l_n &= -\frac{1}{\sigma^2}, \\
\frac{1}{n}\partial_\mu\partial_\sigma l_n &= \frac{-2}{\sigma^3}(\overline{X}_n - \mu), \text{ and} \\
\frac{1}{n}\partial_\sigma^2 l_n &= \frac{1}{\sigma^2} - \frac{3}{\sigma^4}\left[S_P^2 + (\overline{X}_n - \mu)^2\right],
\end{aligned}$$

i.e.

$$\frac{1}{n}\nabla^2 l_n = \begin{pmatrix} -\frac{1}{\sigma^2} & -\frac{2}{\sigma^3}(\overline{X}_n - \mu) \\ -\frac{2}{\sigma^3}(\overline{X}_n - \mu) & \frac{1}{\sigma^2} - \frac{3}{\sigma^4}\left[S_P^2 + (\overline{X}_n - \mu)^2\right] \end{pmatrix}$$

Now recall that

$$\begin{aligned}
\mathbb{E}(\overline{X}_n - \mu) = 0,\, \mathbb{E}(S_P^2) &= \frac{n-1}{n}\sigma^2, \text{ and} \\
\mathbb{E}(\overline{X}_n - \mu)^2 = \mathbb{V}\overline{X}_n &= \frac{\mathbb{V}X}{n} = \frac{\sigma^2}{n}.
\end{aligned}$$

Therefore the Fisher information matrix is

$$\begin{aligned}
I_n(\mu,\, \sigma) &= -\mathbb{E}[\nabla^2 l_n] \\
&= \frac{n}{\sigma^2}\begin{pmatrix} 1 & 0 \\ 0 & -1 + \frac{3}{\sigma^2}\underbrace{\left(\frac{n-1}{n}\sigma^2 + \frac{\sigma^2}{n}\right)}_{=\sigma^2} \end{pmatrix} = \frac{n}{\sigma^2}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}
\end{aligned}$$

and so its inverse is given by

$$J_n = I_n(\mu,\, \sigma)^{-1} = \frac{\sigma^2}{n}\begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

Since $\tau = \mu + \sigma\Phi^{-1}(0.95) =: g(\mu, \sigma)$ where

$$\nabla g = \begin{pmatrix} 1 \\ \Phi^{-1}(0.95) \end{pmatrix}$$

we define

$$\begin{aligned}
\widehat{se}(\hat{\tau}) &:= \left.\sqrt{(\nabla g)^T J_n \nabla g}\,\right|_{(\mu,\,\sigma)=(\hat{\mu},\,\hat{\sigma})} \\
&= \left.\sqrt{\frac{\sigma^2}{n}\left(1 + \frac{1}{2}[\Phi^{-1}(0.95)]^2\right)}\,\right|_{(\mu,\,\sigma)=(\hat{\mu},\,\hat{\sigma})} \\
&= \sqrt{\frac{S_P^2}{n}\left(1 + \frac{1}{2}[\Phi^{-1}(0.95)]^2\right)}.
\end{aligned}$$

So finally the multiparameter delta method tells us that

$$C_n := \hat{\tau} \pm z_{\alpha/2}\widehat{se}(\hat{\tau}),$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$, is an asymptotic $1-\alpha$ confidence interval for τ.

(3) With the given data we may compute that

$$\overline{X}_n \approx 1.19 \text{ and } S_P^2 \approx 3.30.$$

Therefore

$$\hat{\tau} \approx 4.18$$

and the estimate of the standard error given by the delta method is

$$\widehat{se}(\hat{\tau}) \approx 0.558.$$

We may also use the parametric bootstrap to estimate this standard error. Performing 10,000 resamplings yields

$$se_{boot} \approx 0.554.$$

Exercise A.9.7 (Comparing two treatments)**.** n_1 people are given treatment 1 and n_2 people are given treatment 2. Let X_1 be the number of people on treatment 1 who respond favorably to the treatment and let X_2 be the number of people on treatment 2 who respond favorably. Assume that $X_1 \sim \text{Binomial}(n_1,\, p_1)$ and $X_2 \sim \text{Binomial}(n_2,\, p_2)$. Let $\psi := p_1 - p_2$.

(1) Find the MLE $\hat{\psi}$ for ψ.
(2) Find the Fisher information matrix $I(p_1,\, p_2)$.
(3) Use the multiparameter delta method to find the asymptotic standard error of $\hat{\psi}$.
(4) Suppose that $n_1 = n_2 = 200$, $X_1 = 160$, and $X_2 = 148$. Find $\hat{\psi}$. Find an approximate 90 percent confidence interval for ψ using
 (a) the delta method and
 (b) the parametric bootstrap.

Solution. (1) Consider $f(x; n, p) := \binom{n}{x} p^x (1-p)^{n-x}$ to be the CDF of a Binomial(n, p) distribution where n is known. The corresponding log-likelihood function is

$$l(p) := \log f = C + x \log p + (n-x)\log(1-p)$$

and so

$$l'(p) = \frac{x}{p} + \frac{x-n}{1-p}.$$

In particular

$$\begin{aligned} l'(p) = 0 &\Leftrightarrow \frac{x}{p} = \frac{n-x}{1-p} \\ &\Leftrightarrow x - xp = np - xp \\ &\Leftrightarrow p = \frac{x}{n}, \end{aligned}$$

i.e. the MLE for p is $\hat{p} = \frac{X}{n}$.

Since X_1 and X_2 are assumed to be independent, the log-likelihood is

$$\begin{aligned} l_2(p_1,\, p_2) &:= \log f_{X_1,\,X_2}(X_1, X_2; n_1, p_1, n_2, p_2) \\ &= \log f(X_1; n_1, p_1) + \log f(X_2; n_2, p_2), \end{aligned}$$

which is maximized when $p_1 = \frac{X_1}{n_1}$ and $p_2 = \frac{X_2}{n_2}$. In other words the MLE for $(p_1,\, p_2)$ is $(X_1/n_1,\, X_2/n_2)$. By equivariance of the MLE we deduce that

$$\hat{\psi} = \hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}.$$

(2) Recall that

$$I(p_1,\, p_2) = -\mathbb{E}\left[\nabla^2 l_2\right].$$

Since $l_2(p_1,\, p_2) = l(p_1)+l(p_2)$ we deduce from the expression for l' recorded above that

$$\nabla l_2 = \begin{pmatrix} l'(p_1) \\ l'(p_2) \end{pmatrix} = \begin{pmatrix} \frac{X_1}{p_1} + \frac{X_1-n_1}{1-p_1} \\ \frac{X_2}{p_2} + \frac{X_2-n_2}{1-p_2} \end{pmatrix}$$

and hence

$$\nabla^2 l_2(p_1,\, p_2) = \begin{pmatrix} -\frac{X_1}{p_1^2} + \frac{X_1-n_1}{(1-p_1)^2} & 0 \\ 0 & -\frac{X_2}{p_2^2} + \frac{X_2-n_2}{(1-p_2)^2} \end{pmatrix}.$$

Since $\mathbb{E}X_i = n_i p_i$ we then compute that

$$\begin{aligned} \mathbb{E}\left[-\frac{X_i}{p_i^2} + \frac{X_i - n_i}{(1-p_i)^2}\right] = \frac{-n_i}{p_i} + \frac{n_i(p_i-1)}{(1-p_i)^2} &= \frac{-n_i}{p_i} - \frac{n_i}{1-p_i} \\ &= \frac{-n_i}{p_i(1-p_i)}. \end{aligned}$$

So, finally, we conclude that the Fisher infromation matrix is given by

$$I(p_1,\, p_2) = \begin{pmatrix} \frac{n_1}{p_1(1-p_1)} & 0 \\ 0 & \frac{n_2}{p_2(1-p_2)} \end{pmatrix}.$$

(3) Since $\psi = p_1 - p_2 =: g(p_1,\, p_2)$ such that

$$\nabla g = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

and since the inverse of the Fisher information matrix is

$$J := I(p_1,\, p_2)^{-1} = \begin{pmatrix} \frac{p_1(1-p_1)}{n_1} & 0 \\ 0 & \frac{p_2(1-p_2)}{n_2} \end{pmatrix}$$

we define

$$\widehat{se}\hat{\psi} := \sqrt{(\nabla g)^T J \nabla g}\,\Big|_{(p_1,\, p_2)=(\hat{p}_1,\, \hat{p}_2)} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

which the delta method tells us is precisely the asymptotic standard error of $\hat{\psi}$.

(4) Using either method the 90 percent confidence interval takes the form

$$\hat{\psi} \pm z\widehat{se}$$

where $z := \Phi^{-1}(1-0.1/2)$ for Φ denoting the CDF of the standard normal distribution and where the estimator $\widehat{se}$ differs based on the method. Note that in either case

$$\hat{p}_1 = \frac{X_1}{n_1} = 0.8 \text{ and } \hat{p}_2 = \frac{X_2}{n_2} = 0.74$$

and hence

$$\hat{\psi} = \hat{p}_1 - \hat{p}_2 = 0.06,$$

while $z = \Phi^{-1}(1-0.05) \approx 1.645$.

(a) Using the delta method, as shown in item 3 above,

$$\widehat{se}(\hat{\psi}) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

and so, in this case,

$$\widehat{se}(\hat{\psi}) = \frac{\sqrt{0.8 \cdot 0.2 + 0.74 \cdot 0.36}}{\sqrt{200}} \approx 0.0420.$$

(b) Using the parametric bootstrap and performing 500,00 resamplings we obtain that

$$se_{boot} \approx 0.0420.$$

Exercise A.9.8 (Fisher information matrix of the Normal model)**.** Let $X_1, \ldots\ X_n \sim N(\mu,\, \sigma^2)$. Find the Fisher information matrix.

Solution. This is done in Exercise A.9.6 above where we obtain

$$I_n(\mu,\, \sigma^2) = \frac{n}{\sigma^2}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Exercise A.9.9 (Various confidence intervals for a log-normal model)**.** Consider random variables $X_1, \ldots\ X_n \sim \mathrm{Normal}(\mu,\, 1)$. Let $\theta := e^{\mu}$ and let $\hat{\theta} := e^{\overline{X}}$ be the MLE. Create a data set (using $\mu = 5$) consisting of $n = 100$ observations.

Use the delta method to get $\widehat{se}$ and a 95 percent confidence interval for θ. Use the parametric bootstrap to get $\widehat{se}$ and a 95 percent confidence interval for θ. Use the nonparametric bootstrap to get $\widehat{se}$ and a 95 percent confidence interval for θ. Compare your answers.

Solution. In all three cases the confidence interval takes the form

$$\hat{\theta} \pm z\widehat{se},$$

where $z := \Phi^{-1}(1 - 0.05/2)$ for Φ denoting the CDF of the standard normal distribution, the only difference being how $\widehat{se}$ is computed, either analytically or using the parametric/nonparametric bootstrap. We discuss here how $\widehat{se}$ is computed analytically using the delta method. The bootstrap methods are taken care of in Jupyter notebooks.

Since $\theta = e^{\mu} =: g(\mu)$ such that $g' = g$ and since $X_i \sim \mathrm{Normal}(\mu,\, 1)$ such that

$$I(\mu) = -\mathbb{E}\left[(\log f)''\right] = -\mathbb{E}\left[\left(-\frac{1}{2}(X-\mu)^2\right)''\right] = \frac{1}{2}\mathbb{E}\left[2(\mu - X)'\right] = \mathbb{E}1 = 1,$$

and so $I_n(\mu) = n$, we define

$$\widehat{se}(\hat{\theta}) := \left|e^{\overline{X}}\right| \cdot \sqrt{\frac{1}{I_n(\mu)}} = \frac{1}{\sqrt{n}}e^{\overline{X}}.$$

A.10. **Hypothesis Testing and p-values.**

Exercise A.10.1 (Power of the Wald test)**.** Prove that for the Wald test, when $\theta \neq \theta_0$, then

$$\beta(\theta) = 1 - \Phi\left(\frac{\theta_0 - \theta}{\widehat{se}_n} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta}{\widehat{se}_n} - z_{\alpha/2}\right).$$

Solution. By definition, the power function of the Wald test is

$$\beta(\theta) = \mathbb{P}_\theta\left(\left|\frac{\hat{\theta}_n - \theta_0}{\widehat{se}_n}\right| > z_{\alpha/2}\right)$$

where, in order for the Wald test to be valid,

$$z_n := \frac{\hat{\theta}_n - \theta}{\widehat{se}_n} \rightsquigarrow N(0,\,1) \text{ if } \theta \text{ is the true parameter.}$$

We may therefore write, for $Y_n := \frac{\theta_0 - \theta}{\widehat{se}_n}$,

$$\beta(\theta) = \mathbb{P}_\theta\left(\left|\frac{\hat{\theta}_n - \theta}{\widehat{se}_n} + \frac{\theta - \theta_0}{\widehat{se}_n}\right| > z_{\alpha/2}\right) = \mathbb{P}_\theta\left(|Z_n - Y_n| > z_{\alpha/2}\right).$$

Writing Φ_n for the CDF of Z_n we then compute that

$$\begin{aligned}\beta(\theta) &= \mathbb{P}\left(|Z_n - Y_n| > z_{\alpha/2}\right)\\ &= \mathbb{P}\left(Z_n - Y_n > z_{\alpha/2}\right) + \mathbb{P}\left(Z_n - Y_n < -z_{\alpha/2}\right)\\ &= \mathbb{P}\left(Z_n > Y_n + z_{\alpha/2}\right) + \mathbb{P}\left(Z_n < Y_n - z_{\alpha/2}\right)\\ &= 1 - \Phi_n\left(Y_n + z_{\alpha/2}\right) + \Phi_n\left(Y_n - z_{\alpha/2}\right).\end{aligned}$$

In particular, when n is sufficiently large to ensure that Z_n has distribution approximately $N(0,\,1)$, which means that $\Phi_n \approx \Phi$, we conclude that

$$\beta(\theta) \approx 1 - \Phi\left(Y_n + z_{\alpha/2}\right) + \Phi\left(Y_n - z_{\alpha/2}\right),$$

as desired.

Exercise A.10.2 (Distribution of the p–value under the null hypothesis)**.** Suppose that for $\alpha \in (0,\,1)$ we have a family of hypothesis tests $\mathcal{T}_\alpha$ as in Theorem 10.16. Suppose moreover that the null hypothesis (which is *independent* of α) is simple and that the test statistics have nowhere vanishing continuous distributions. Prove that, under the null hypothesis, i.e. provided that the null hypothesis is true, the p–value has a Uniform$(0, 1)$ distribution.

Solution. Since the null hypothesis is simple, Theorem 10.16 tells us that, for any outcome ω,

$$p - \text{value}(\omega) = \mathbb{P}_{\theta_0}\left(T_n^{(0)} \geqslant T_n(\omega)\right) = 1 - F_n(T_n(\omega))$$

where F_n denotes the CDF of T_n under the null hypothesis. As shown in Exercise A.2.14, since T_n has a nowhere vanishing PDF,

$$F_n(T_n) \sim \text{Uniform}(0,\,1).$$

In particular, for any $\mathcal{U} \sim \text{Uniform}(0,\,1)$ and any $\alpha \in (0,\,1)$,

$$\mathbb{P}(1 - \mathcal{U} < \alpha) = \mathbb{P}(\mathcal{U} > 1 - \alpha) = 1 - \mathbb{P}(\mathcal{U} \leqslant 1 - \alpha) = 1 - (1 - \alpha) = \alpha,$$

i.e. $1-\mathcal{U}$ is also a Uniform(0, 1) random variable. So finally:

$$p-\text{value} = 1 - F_n(T_n) = 1 - \mathcal{U} \sim \text{Uniform}(0,\, 1).$$

Exercise A.10.3 (Wald test and confidence intervals)**.** Prove Theorem 10.12.

Solution. The null hypothesis is *not* rejected by a Wald test if and only if

$$\begin{aligned}|W_n| \leqslant z_{\alpha/2} &\Leftrightarrow -z_{\alpha/2} \leqslant \frac{\hat{\theta}_n - \theta_0}{\widehat{se}_n} \leqslant z_{\alpha/2}\\ &\Leftrightarrow \hat{\theta}_n - \widehat{se}_n z_{\alpha/2} \leqslant \theta_0 \leqslant \hat{\theta}_n + \widehat{se}_n z_{\alpha/2}\\ &\Leftrightarrow \theta_0 \in \left[\hat{\theta}_n - \widehat{se}_n z_{\alpha/2},\, \hat{\theta}_n + \widehat{se}_n z_{\alpha/2}\right].\end{aligned}$$

Since

$$\theta_0 = \hat{\theta}_n \pm \widehat{se}_n z_{\alpha/2} \iff \hat{\theta}_n = \theta_0 \mp \widehat{se}_n z_{\alpha/2}$$

are both events with asymptotically vanishing probability (since W_n is approximately normal then) it follows that, for n sufficiently large, and with very high probability, the null hypothesis is *not* rejected if and only if $\theta_0 \in C_n$, as desired. (Note that the qualifier "asymptotically" and "with very high probability" may be removed if $\hat{\theta}_n$ and $\widehat{se}_n$ have a continuous joint PDF, which is often the case in practice).

Exercise A.10.4 (Alternate expression for p–values)**.** Prove Theorem 10.16.

Solution. First we note that since c is strictly decreasing it is also invertible. We may therefore rewrite the p–value as follows:

$$\begin{aligned}p-\text{value} &= \inf\left\{\alpha : T_n \in R_\alpha\right\}\\ &= \inf\left\{\alpha : T_n \geqslant c(\alpha)\right\}\\ &= \inf\left\{\alpha : c^{-1}(T_n) \leqslant \alpha\right\}\\ &= c^{-1}(T_n).\end{aligned}$$

In particular, for any fixed outcome ω,

$$p-\text{value}(\omega) = c^{-1}(T_n(\omega)).$$

So it remains to relate the inverse c^{-1} to $\sup \mathbb{P}\left(T_n^* \geqslant T_n(\omega)\right)$. To obtain such an expression we use the fact the each test $\mathcal{T}_\alpha$ has size α. By definition this means that

$$\alpha = \sup_{\theta_* \in \Theta_0} \mathbb{P}_{\theta_*}\left(T_n^* \in R_\alpha\right) = \sup_{\theta_* \in \Theta_0} \mathbb{P}_{\theta_*}\left(T_n^* \geqslant c(\alpha)\right).$$

In particular for any fixed $t \in \mathbb{R}$ we may choose $\alpha := c^{-1}(t)$ and deduce from these equations that

$$c^{-1}(t) = \alpha = \sup_{\theta_* \in \Theta_0} \mathbb{P}_{\theta_*}\left(T_n^* \geqslant c\left(c^{-1}(t)\right)\right) = \sup_{\theta_* \in \Theta_0} \mathbb{P}_{\theta_*}\left(T_n^* \geqslant t\right).$$

Choosing $t := T_n(\omega)$ for some fixed outcome ω allows us to conclude:

$$p-\text{value}(\omega) = c^{-1}(T_n(\omega)) = \sup_{\theta_* \in \Theta_0} \mathbb{P}_{\theta_*}\left(T_n^* \geqslant T_n(\omega)\right)$$

as desired.

Exercise A.10.5 (Testing for a uniform model)**.** Let $X_1, \dots, X_n \sim \text{Uniform}(0, \theta)$ and let $Y := \max\{X_1, \dots, X_n\}$. We want to test

$$H_0 : \theta = \frac{1}{2} \text{ versus } H_1 : \theta > \frac{1}{2}.$$

The Wald test is not appropriate since Y does not converge to a Normal. Suppose we decide to test this hypothesis by rejecting H_0 when $Y > c$.

(1) Find the power function.
(2) What choice of c will make the size of the test 0.05?
(3) In a sample of size $n = 20$ with $Y = 0.48$ what is the p–value? What conclusion about H_0 would you make?
(4) In a sample of size $n = 20$ with $Y = 0.52$ what is the p–value? What conclusion about H_0 would you make?

Solution. (1) We begin by computing the distribution of Y. Its CDF is

$$\mathbb{P}_\theta(Y \leqslant y) = \mathbb{P}_\theta(X_i \leqslant y \text{ for all } i) = \mathbb{P}_\theta(X_i \leqslant y)^n = \left(\frac{y}{\theta}\right)^n.$$

Therefore the power function is

$$\beta(\theta) := \mathbb{P}_\theta(Y > c) = 1 - \left(\frac{y}{\theta}\right)^n.$$

(2) The size of the test is

$$\alpha := \sup_{\theta = 1/2} \left[1 - \left(\frac{c}{\theta}\right)^n\right] = 1 - (2c)^n.$$

Inverting this relation lets us write c as a function of α:

$$\begin{aligned} \alpha = 1 - (2c)^n &\Leftrightarrow (2c)^n = 1 - \alpha \\ &\Leftrightarrow c = \frac{1}{2}(1 - \alpha)^{1/n}. \end{aligned}$$

In particular when $\alpha = 0.05$ we deduce that

$$c = \frac{1}{2}(0.95)^{1/n}.$$

When $n = 20$ this comes out to $c \approx 0.4987$.

(3) As we have seen in item 2 above the function $c_n(\alpha) := \frac{1}{2}(1 - \alpha)^{1/n}$ has inverse $c_n^{-1}(y) = 1 - (2c)^n$. Since c_n is strictly decreasing it follows that

$$\begin{aligned} p\text{–value} &= \inf\{\alpha : Y > c_n(\alpha)\} \\ &= \inf\left\{\alpha : c_n^{-1}(Y) < \alpha\right\} \\ &= \max(0, c_n^{-1}(Y)) \\ &= \max(0, 1 - (2Y)^n). \end{aligned}$$

Note that we had to invoke the function $\max\{0, \cdot\}$ since $c_n^{-1}(Y)$ may be negative, when $Y > 1/2$, even though α itself cannot be negative.

In particular when $n = 20$ and $Y = 0.48$ we have that

$$p\text{–value} = 1 - (2 \cdot 0.48)^{20} \approx 0.558$$

which means that we do not have sufficient evidence to reject the null hypothesis.

(4) Proceeding as in item 3 above:

$$\begin{aligned} p\text{–value} &= \max\{0,\, 1-(2Y)^n\} \\ &= \max\left\{0,\, 1-(2\cdot 0.52)^{20}\right\} \\ &\approx \max\{0,\, -1\} = 0 \end{aligned}$$

and so we may categorically, without a doubt, reject the null hypothesis. This is not surprising: if $Y > \frac{1}{2}$ then one of the X_i's is larger than $\frac{1}{2}$ and so it cannot have been generated by a Uniform$(0, 1/2)$ distribution.

Exercise A.10.6 (Postponing death)**.** There is a theory that people can postpone their deaths until after an important event. To test the theory, Phillips and King (1998) collected data on deaths around the Jewish holiday Passover. Of 1919 deaths, 922 died the week before the holiday and 997 died the week after. Think of this as a Binomial and test the null hypothesis that $\theta = 1/2$. Report and interpret the p–value. Also construct a confidence interval for θ.

Solution. We view $X = 922$ as a random variable with Binomial$(n,\, \theta)$ distribution where $n = 1919$. This means that we interpret θ as follows:

$$\theta = \mathbb{P}(\text{dying before Passover}) \text{ and } 1-\theta = \mathbb{P}(\text{dying after Passover}).$$

We wish to test the null hypothesis

$$H_0 : \theta = \frac{1}{2}.$$

Since binomial distributions approach Normal distributions as n increases (by virtue of the Central Limit Theorem, since binomial distributions are sums of IID Bernoulli random variables) it is sensible to seek to use a Wald test.

We will use $\hat{\theta}_n := \frac{X}{n}$ as test statistic since, as shown in Exercise A.9.7, this is the MLE for θ. That same exercise showed that the Fisher information is then

$$I(\theta) = \frac{1}{\theta(1-\theta)}$$

and so we may approximate the standard error of $\hat{\theta}_n$ with

$$\widehat{se}_n := \sqrt{\frac{1}{nI(\hat{\theta}_n)}} = \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}.$$

The asymptotic normality of the MLE then guarantees that employing the Wald test is valid here. We thus reject the null hypothesis when

$$\left|\frac{\hat{\theta}_n - 1/2}{\widehat{se}_n}\right| > z_{\alpha/2}$$

where $z_{\alpha/2} := \Phi^{-1}(1-\alpha/2)$ for Φ the CDF of the standard normal distribution. As computed in Exercise A.23.4 the p–value of the Wald test is then

$$p - \text{value} = 2\left[1 - \Phi\left(\left|\frac{\hat{\theta}_n - 1/2}{\widehat{se}_n}\right|\right)\right].$$

Here we have $n = 1919$ and $X = 922$ such that

$$\left|\frac{\hat{\theta}_n - 1/2}{\widehat{se}_n}\right| \approx 1.7$$

and the p-value is approximately 0.087, which is weak evidence against the null hypothesis (and is arguably not enough evidence to reject the null hypothesis).

Finally, as per Theorem 10.12, the Wald test is equivalent to checking whether or not $\theta_0 = \frac{1}{2}$ belongs to the confidence interval

$$C_{n,\,\alpha} = \hat{\theta}_n + \widehat{se}_n z_{\alpha/2}.$$

When $\alpha = 0.05$ the confidence interval in our case becomes

$$C_n \approx (0.458,\, 0.503),$$

which includes θ_0 (but only barely, which is another way of saying that the p–value is close to, but not above, 0.05).

Exercise A.10.7 (Testing a sum of Normals)**.** Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Consider testing

$$H_0 : \theta_0 = 0 \text{ versus } H_1 : \theta = 1.$$

Let the rejection region be $R = \{x \in \mathbb{R}^n : T(x) > c\}$ where $T(x) := \frac{1}{n}\sum_{i=1}^n x_i$.

(1) Find c such that the test has size α.

(2) Find the power under H_1, that is find $\beta(1)$. Show that $\beta(1) \to 1$ as $n \to \infty$.

Solution. (1) Recall that if $X, Y \sim N(\mu_X, \sigma_X^2), N(\mu_Y, \sigma_Y^2)$ are independent then

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Therefore

$$T_n := T(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i \sim \frac{1}{n} N(n\theta, n) = N(\theta, 1/n).$$

We may thus compute the size of the test to be

$$\begin{aligned}\mathbb{P}_{\theta=0}(T_n > c) &= \mathbb{P}\left(N(0, 1/n) > c\right)\\ &= \mathbb{P}\left(\sqrt{n}N(0, 1/n) > \sqrt{n}c\right)\\ &= \mathbb{P}\left(N(0, 1) > \sqrt{n}c\right)\\ &= 1 - \Phi(\sqrt{n}c).\end{aligned}$$

Inverting this relation yields

$$\alpha = 1 - \Phi(\sqrt{n}c) \Leftrightarrow c = \frac{1}{\sqrt{n}}\Phi^{-1}(1-\alpha) =: \frac{1}{\sqrt{n}} z_\alpha.$$

In other words this test has size α when $c = \frac{z_\alpha}{\sqrt{n}}$.

(2) We compute that

$$\begin{aligned}\beta(\theta) &= \mathbb{P}_\theta(T_n > c)\\ &= \mathbb{P}(N(\theta, 1/n) > c)\\ &= \mathbb{P}(N(0, 1) > \sqrt{n}(c - \theta))\\ &= 1 - \Phi\left[\sqrt{n}(c-\theta)\right].\end{aligned}$$

Since $c = \frac{z_\alpha}{\sqrt{n}}$ this simplifies to

$$\begin{aligned}\beta(\theta) = 1 - \Phi\left[\sqrt{n}\left(\frac{z_\alpha}{\sqrt{n}} - \theta\right)\right] &= 1 - \Phi\left(z_\alpha - \theta\sqrt{n}\right)\\ &= \mathbb{P}\left(N(0, 1) > z_\alpha - \theta\sqrt{n}\right).\end{aligned}$$

In particular when $\theta = 1$ we see that

$$\beta(1) = 1 - \Phi(z_\alpha - \sqrt{n}) \to 1 - \Phi(-\infty) = 1$$

as $n \to \infty$. Note that n need not be particularly large: for $\alpha = 0.05$, $\beta(1) > 0.9$ as soon as $n = 11$.

Exercise A.10.8 (Power under the alternative of a Wald test)**.** Let $\hat{\theta}_n$ be the MLE of a parameter θ and let $\widehat{se} = [nI(\hat{\theta}_n)]^{-1/2}$ where $I(\theta)$ is the Fisher information. Consider testing

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0.$$

Consider the Wald test with rejection region $\left\{|Z| > z_{\alpha/2}\right\}$ where $Z = (\hat{\theta}_n - \theta_0)/\widehat{se}$. Let $\theta_1 > \theta_0$ be some alternative. Show that $\beta(\theta_1) \to 1$.

Solution. Theorem 10.10 provides an asymptotic expression for the power of the Wald test when $\theta \neq \theta_0$ (see Exercise A.10.1 above for details) and so, when n is sufficiently large,

$$\beta(\theta_1) \approx 1 - \Phi\left(\frac{\theta_0 - \theta_1}{\widehat{se}} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_1}{\widehat{se}} - z_{\alpha/2}\right).$$

By consistency and asymptotic normality of the MLE we know that $\widehat{se} \to 0$ almost surely as $n \to \infty$. Therefore, since $\theta_1 > \theta_0$,

$$\frac{\theta_0 - \theta_1}{\widehat{se}} \pm z_{\alpha/2} \to -\infty \text{ almost surely as } n \to \infty$$

and so

$$\beta(\theta_1) \to 1 - \Phi(-\infty) + \Phi(-\infty) = 1 \text{ almost surely as } n \to \infty.$$

Exercise A.10.9 (Mark Twain and Quintus Curtius Snodgrass)**.** In 1861, 10 essays appeaed in the New Orleans Daily Crescent. They were signed "Quintus Curtius Snodgrass" and some people suspected they were actually written by Mark Twain. To investigate this, we will consider the proportion of three letter words found in an author's work. From eight Twain essays we have

$$0.225, 0.262, 0.217, 0.240, 0.230, 0.229, 0.235, \text{ and } 0.217.$$

From 10 Snodgrass essays we have

$$0.209, 0.205, 0.196, 0.210, 0.202, 0.207, 0.224, 0.223, 0.220, \text{ and } 0.201.$$

- Perform a Wald test for equality of the means. Use the nonparametric plug-in estimator. Report the p–value and a 95 percent confidence interval for the difference of means. What do you conclude?
- Now use a permutation test to avoid the use of large sample methods. What is your conclusion? (Brinegar (1963))

Solution.

- We perform a Wald test with null hypothesis

$$p_X - p_Y = 0$$

using the test statistic

$$W := \frac{(\overline{X}_m - \overline{Y}_n) - 0}{\widehat{se}}.$$

Here we choose

$$\widehat{se} := \sqrt{\widehat{se}_m^2(\overline{X}_m) + \widehat{se}_n^2(\overline{Y}_n)} =: \sqrt{\frac{S_{X,P}^2}{m} + \frac{S_{Y,P}^2}{n}}$$

for

$$S_{X,P}^2 := \frac{1}{m}\sum_{i=1}^{m}\left(X_i - \overline{X}_m\right)^2 \text{ and } S_{Y,P}^2 := \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - \overline{Y}_n\right)^2$$

denoting the *population* variances. This is indeed the plug-in estimator of the variance since

$$\mathbb{V}(\overline{X}_m) = \frac{\mathbb{V}X}{m} \overset{\text{plug-in}}{\mapsto} \frac{S_{X,P}^2}{m} =: \widehat{se}_m^2(\overline{X}_m).$$

In other words the Wald statistic is

$$W = \frac{\overline{X}_m - \overline{Y}_n}{\sqrt{\frac{S_{X,P}^2}{m} + \frac{S_{Y,P}^2}{n}}}.$$

With the given data, where the X_i's correspond to the Mark Twain essays while the Y_j's correspond to the Snodgrass essays, we obtain

$$W \approx 3.9$$

and so the p–value is

$$p - \text{value} = 2\Phi(-|W|) \approx 0.00008.$$

We may also compute a 95 percent confidence interval to be, for $\alpha = 0.05$, $C_n : W \pm \widehat{se} z_{\alpha/2}$ such that here

$$C_n \approx (0.01, 0.03).$$

While the p–value is very low, the confidence interval is very close to zero and so it is difficult to argue that there is enough evidence to conclude that the two series of essays were written by different authors (or at least, in different styles).

- We randomly choose a subset $S \subsetneq S_N$, where $N = 8 + 10 = 18$ (such that $18! > 10^{15}$). We then compute

$$t^* := \frac{1}{|S|}\sum_{\sigma \in S} I(\mathcal{T}_\sigma > T)$$

where

$$T\left(X_1, \ldots, X_m, Y_1, \ldots, Y_n\right) := \left|\overline{X}_m - \overline{Y}_n\right|$$

and, for $Z_1, \ldots, Z_N$ denoting $X_1, \ldots, X_m, Y_1, \ldots, Y_n$,

$$\mathcal{T}_\sigma = T(Z_1, \ldots, Z_N).$$

With one such subset $S \subsetneq S_N$ of size $|S| = 30,000$ we obtain that

$$t^* = 0.0009.$$

Since, for a permutation test, t^* is the p–value, this is strong evidence that the X_i's and Y_j's come from different distributions.

Exercise A.10.10 (Mortality patterns)**.** Here are the number of elderly Jewish and Chinese women who died just before and after the Chinese Harvest Moon Festival.

Week	Chinese	Jewish
-2	55	141
-1	33	145
1	70	139
2	49	161

Compare the two mortality patterns. (Phillips and Smith (1990)).

Solution. For each ethnicity $e = C, J$ we can view the number of deaths over four weeks as a single sample from a Multinomial(n_e, p_e) distribution where p_e belongs to the 3–simplex $\Delta^3 \subseteq \mathbb{R}^4$. We will then take two different approaches.

(1) We will use a likelihood ratio test to consider the null hypothesis $p_C = p_J$.
(2) We will use *two* Pearson's χ^2 tests, one for each ethnicity, to test the null hypothesis $p_0 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ in each case. We will use the Bonferroni method to account for the multiple testing that occurs.

We denote by X_1, X_2, X_3, X_4 the number of deaths in weeks -2, -1, 1, and 2 respectively of elderly *Chinese* women. We denote by Y_i the corresponding values for elderly *Jewish* women.

(1) First we use a likelihood ratio test. The full parameter space is

$$\Theta = \Delta^3 \times \Delta^3 \subseteq \mathbb{R}^4 \times \mathbb{R}^4$$

while

$$\Theta_0 = \left\{(p_C, p_J) \in \Delta^3 \times \Delta^3 : p_C = p_J\right\}.$$

The likelihood function is

$$\mathcal{L}_n(p_C, p_J) = \binom{n_C}{X_1, \dots, X_4} p_{C,1}^{X_1} \dots p_{C,4}^{X_4} \binom{n_J}{Y_1, \dots, Y_4} p_{J,1}^{Y_1} \dots p_{J,4}^{Y_4}.$$

As noted in Remark 2.38, a Multinomial(n, p) distribution is the sum of n IID Categorical(p) random variables. Since we have computed in Exercise A.23.13 (see also Exercise A.23.14 and Remark A.5) that the MLE of a Categorical(p) distribution to be the (vector) sample mean, we deduce that the MLE here takes the same form, i.e.

$$(\hat{p}_C, \hat{p}_J) = \left(\frac{X}{n_C}, \frac{Y}{n_J}\right).$$

Therefore

$$\mathcal{L}_n\left(\hat{p}_C, \hat{p}_J\right)$$
$$= \binom{n_C}{X_1, \dots, X_4} \left(\frac{X_1}{n_C}\right)^{X_1} \dots \left(\frac{X_4}{n_C}\right)^{X_4} \binom{n_J}{Y_1, \dots, Y_4} \left(\frac{Y_1}{n_J}\right)^{Y_1} \dots \left(\frac{Y_4}{n_J}\right)^{Y_4}.$$

We now compute the MLE over Θ_0. The key difference is that now the likelihood function takes the form

$$\mathcal{L}_n(p, p) = \binom{n_C}{X_1, \dots, X_4} \binom{n_J}{Y_1, \dots, Y_4} p_1^{X_1+Y_1} \dots p_4^{X_4+Y_4},$$

which, up to a prefactor independent of p, is the likelihood of the random variable $Z := X + Y \sim$ Multinomial$(n_C + n_J, p)$ (note that this is only the distribution *under the null hypothesis*, as is under consideration here). This means that the MLE over Θ_0 is

$$\hat{p}_0 = \frac{Z}{n_Z} = \frac{X+Y}{n_C + n_J}.$$

Therefore

$$\mathcal{L}_n\left(\hat{p}_0,\, \hat{p}_0\right) = \binom{n_C}{X_1,\, \ldots,\, X_4}\binom{n_J}{Y_1,\, \ldots,\, Y_4}\left(\frac{X_1+Y_1}{n_C+n_J}\right)^{X_1+Y_1}\cdots\left(\frac{X_4+Y_4}{n_C+n_J}\right)^{X_4+Y_4}.$$

So finally the likelihood ratio statistic is

$$\begin{aligned}\lambda &:= 2\log\frac{\left(\frac{X_1}{n_C}\right)^{X_1}\cdots\left(\frac{X_4}{n_C}\right)^{X_4}\left(\frac{Y_1}{n_J}\right)^{Y_1}\cdots\left(\frac{Y_4}{n_J}\right)^{Y_4}}{\left(\frac{X_1+Y_1}{n_C+n_J}\right)^{X_1+Y_1}\cdots\left(\frac{X_4+Y_4}{n_C+n_J}\right)^{X_4+Y_4}}\\ &= 2\sum_{i=1}^{4}\log\frac{\left(\frac{X_i}{n_C}\right)^{X_i}\left(\frac{Y_i}{n_J}\right)^{Y_i}}{\left(\frac{X_i+Y_i}{n_C+n_J}\right)^{X_i+Y_i}},\end{aligned}$$

which has rejection region

$$\left\{\lambda \in \mathbb{R} : \lambda > \chi^2_{1,\,\alpha}\right\}$$

since $\dim\Theta - \dim\Theta_0 = 6 - 5 = 1$. Recall also that the p–value is

$$p - \text{value} = \mathbb{P}(\chi^2_1 > \lambda) = 1 - F_1(\lambda)$$

for F_1 denoting the CDF of χ^2_1.

With the data we are given we compute that the test statistic is $\lambda \approx 13$ while the cutoff for rejection is at $\chi^2_{1,\,\alpha} \approx 4$ when $\alpha = 0.05$. We therefore reject the null hypothesis: there is statistically significant evidence that the distribution of deaths over those four weeks differ by ethnicity. This is further corroborated by a p–value of approximately 0.0004.

Note that, however, the multinomial assumption could nonetheless be challenged: for example, the numbers of deaths in consecutive weeks are *not* independent from each other, but the multinomial assumption treats them as such.

(2) We now use Pearson's χ^2 test separately for each ethnicity, with a Bonferroni correction for multiple testing. So we choose two of Pearson's χ^2 statistics, namely

$$\begin{aligned}T_C &:= \frac{(X_1 - \frac{n_C}{4})^2}{n_C/4} + \cdots + \frac{(X_4 - \frac{n_C}{4})^2}{n_C/4}\\ &= \frac{4}{n_C}\left[\left(X_1 - \frac{n_C}{4}\right)^2 + \cdots + \left(X_4 - \frac{n_C}{4}\right)^2\right]\end{aligned}$$

and

$$T_J := \frac{4}{n_J}\left[\left(Y_1 - \frac{n_J}{4}\right)^2 + \cdots + \left(Y_4 - \frac{n_J}{4}\right)^2\right]$$

to be the test statistics in each case. For each ethnicity the rejection region is

$$\left\{t \in \mathbb{R} : t > \chi^2_{3,\,\alpha}\right\}$$

since the parameter space is $\Theta = \Delta^3 \subseteq \mathbb{R}^4$. Recall that the p–value is given by

$$p - \text{value} = \mathbb{P}(\chi^2_3 > T_e) = 1 - F_3(T_e)$$

for $e = C, J$ and where F_3 denotes the CDF of a χ^2_3 distribution.

With the data we are given we compute that, for a cutoff of $\chi^2_{3,\,\alpha} \approx 8$ when $\alpha = 0.05$,

$$T_C \approx 14 \text{ and } T_J \approx 2.$$

This corresponds to p–values of

$$p - \text{value}_C \approx 0.0035 \text{ and } p - \text{value}_J \approx 0.57.$$

Since we are using the Bonferroni method with two hypothesis tests we note that

$$p - \text{value}_C < \frac{\alpha}{2} = 0.025 < p - \text{value}_J$$

and so we reject the null hypothesis for Chinese elderly women but *not* for Jewish elderly women. In other words: there is evidence that the probability that an elderly Chinese woman dies is affecte by the proximity of the Chinese Harvest Moon Festival. There is no such evidence for Jewish elderly women.

Exercise A.10.11 (Drug testing)**.** A randomized, double-blind experiment was conducted to assess the effectivenes of several drugs for reducing postoperative nausea. The data are as follows.

	Number of Patients	Incidence of Nausea
Placebo	80	45
Chlorpromazine	75	26
Dimenhydrinate	85	52
Pentobarbital (100 mg)	67	35
Pentobarbital (150 mg)	85	37

(1) Test each drug verus the placebo at the 5 per cent level. Summarize your findings.

(2) Use the Bonferroni and the FDR method to adjust for multiple testing. (Beecher (1959)).

Solution. (1) For each drug we can compare it to the placebo using a Wald test, i.e. testing

$$H_0 : p_{drug} = p_{placebo} \text{ versus } H_1 : p_{drug} \neq p_{placebo}.$$

Writing $p_d := p_{drug}$ and $p_p = p_{placebo}$ we know that the MLE for a Bernoulli(p) model is

$$\hat{p} = \overline{X}_n$$

with Fisher information given by

$$I(p) = \frac{1}{p(1-p)}$$

such that, for

$$\widehat{se} := \sqrt{I(\hat{p})/n} = \sqrt{\hat{p}(1-\hat{p})/n},$$

the asymptotic normality of the MLE guarantees that

$$\frac{\hat{p} - p}{\widehat{se}} = \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{\hat{p}(1-\hat{p})}} \rightsquigarrow N(0,\, 1).$$

We may therefore choose

$$W := \frac{\hat{p}_d - \hat{p}_p}{\sqrt{\widehat{se}_d^2 + \widehat{se}_p^2}} = \frac{\sqrt{n}(\hat{p}_d - \hat{p}_p)}{\sqrt{\hat{p}_e(1-\hat{p}_e) + \hat{p}_p(1-\hat{p}_p)}}$$

as the statistic to use in the Wald test. The rejection region is then

$$R_\alpha := \left\{ w \in \mathbb{R} : |w| > z_{\alpha/2} \right\}$$

with

$$p - \text{value} = 2\Phi(-|w|).$$

We compute, recording approximate values below, that

	$\hat{p}$	W	p–value
Placebo	0.56	–	–
Chlorpromazine	0.35	-2.72	0.07
Dimenhydrinate	0.61	0.65	0.5
Pentobarbital (100 mg)	0.52	-0.47	0.6
Pentobarbital (150 mg)	0.44	-1.67	0.1

At level $\alpha = 0.05$ only the first drug, Chlorpromazine, shows a statistically significant effect.

(2) Using the Bonferroni method, at level $\alpha = 0.05$ we reject all null hypotheses with p–values below $\frac{\alpha}{4} = 0.0125$. As in item 1 above this means that only Chlorpromazine shows a statistically significant effect.

Using the BH method we note that all four tests are independent and so

$$l_i = i\frac{\alpha}{4}.$$

Comparing the ordered p–values to l_i, namely

Ordered p–values	l_i
0.007	0.0125
0.1	0.025
0.5	0.0375
0.6	0.05

we see that the BH rejection treshold is $T = P_{(1)} = 0.007$, i.e. it is still only Chlorpromazine which shows a statistcally significant effect.

Exercise A.10.12 (Wald test for Poisson distributions)**.** Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. Let $\lambda > 0$. Find the size α Wald test for

$$H_0 : \lambda = \lambda_0 \text{ versus } H_1 : \lambda \neq \lambda_0.$$

Solution. As shown in Exercise A.9.4 the MLE for λ is $\hat{\lambda} = \overline{X}_n$ and the Fisher information is $I(\lambda) = 1/\lambda$. Therefore, by asymptotic normality of the MLE, if we define $\widehat{se} := [nI(\hat{\lambda})]^{-1/2} = \sqrt{\hat{\lambda}/n}$, then the Wald test statistic is

$$W := \frac{\hat{\lambda} - \lambda_0}{\widehat{se}} = \frac{\sqrt{n}\left(\hat{\lambda} - \lambda_0\right)}{\hat{\lambda}} = \sqrt{n}\left(1 - \frac{\lambda_0}{\hat{\lambda}}\right)$$

and the rejection region is

$$R_\alpha = \left\{ w \in \mathbb{R} : w > z_{\alpha/2} \right\}.$$

Exercise A.10.13 (Likelihood ratio and Wald test for the mean of a Normal model)**.** Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Construct the likelihood ratio test for

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0.$$

Compare to the Wald test.

Solution. First we construct the likelihood ratio test. The likelihood function is

$$\mathcal{L}_n = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n \left(X_i - \mu\right)^2\right)$$

while

$$\Theta = \left\{(\mu, \sigma^2) : \mu \in \mathbb{R} \text{ and } \sigma^2 > 0\right\} \text{ and } \Theta_0 = \left\{(\mu_0, \sigma^2) : \sigma^2 > 0\right\}.$$

We have computed in item 1 of Exercise A.9.6 that the MLE over Θ is

$$\hat{\mu} = \overline{X}_n \text{ and } \hat{\sigma}^2 = S_P^2 := \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$$

and so

$$\mathcal{L}_n(\hat{\mu}, \hat{\sigma}^2) = \left(2\pi S_P^2\right)^{-n/2} \exp\left(-\frac{1}{2S_P^2}\underbrace{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}_{=nS_P^2}\right) = \left(2\pi S_P^2\right)^{-n/2} e^{-n/2}.$$

We now compute the MLE for σ^2 over Θ_0. Since

$$l_n(\mu_0, \sigma^2) = \log \mathcal{L}_n(\mu_0, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(X_i - \mu_0\right)^2 + C$$

it follows that

$$\partial_{\sigma^2} l_n = -\frac{n}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n \left(X_i - \mu_0\right)^2 = \frac{1}{2\sigma^2}\left(-n + \frac{1}{\sigma^2}\sum_{i=1}^n \left(X_i - \mu_0\right)^2\right).$$

The solution $\hat{\sigma}_0^2$ of $\partial_{\sigma^2} l_n(\mu_0, \sigma^2) = 0$ is therefore

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \mu_0\right)^2 =: S_{0,P}^2.$$

Therefore

$$\mathcal{L}_n(\mu_0, \hat{\sigma}_0^2) = \left(2\pi S_{0,P}^2\right)^{-n/2} \exp\left(-\frac{1}{2S_{0,P}^2}\underbrace{\sum_{i=1}^n \left(X_i - \mu_0\right)^2}_{=nS_{0,P}^2}\right) = \left(2\pi S_{0,P}^2\right)^{-n/2} e^{-n/2}$$

and so the likelihood ratio is

$$\frac{\mathcal{L}_n(\hat{\mu}, \hat{\sigma}^2)}{\mathcal{L}_n(\mu_0, \hat{\sigma}_0^2)} = \frac{\left(S_P^2\right)^{-n/2}}{\left(S_{0,P}^2\right)^{-n/2}} = \left(\frac{S_{0,P}^2}{S_P^2}\right)^{n/2}.$$

So finally the likelihood ratio statistic is

$$\lambda = n \log \frac{S_P^2}{S_{0,P}^2},$$

and recall that the rejection region is given by, since $\dim\Theta - \dim\Theta_0 = 1$,

$$\left\{\lambda \in \mathbb{R} : \lambda > \chi^2_{1,\,\alpha}\right\}.$$

We now turn our attention to the Wald test. We have computed in item 3 of Exercise A.9.6 that the inverse of the Fisher information matrix is

$$J_n(\mu,\,\sigma^2) = \frac{\sigma^2}{n}\begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

Therefore, for

$$\widehat{se} := \left[J_n(\hat\mu,\,\hat\sigma^2)\right]_{11}^{1/2} = \sqrt{\frac{\hat\sigma^2}{n}} = \sqrt{\frac{S_P^2}{n}},$$

the asymptotic normality of the MLE for multiparameter models tells us that

$$W := \frac{\hat\mu - \mu_0}{\widehat{se}} = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sqrt{S_P^2}} \rightsquigarrow N(0,\,1)$$

is a valid Wald test statistic. (Appealing to the asymptotic normality of the MLE is actually overkill: since the MLE are the sample mean and population variance, the convergence of W in distribution to a standard normal follows directly from the Central Limit Theorem.) Recall that the rejection region is

$$\left\{w \in \mathbb{R} : |w| > z_{\alpha/2}\right\}.$$

Exercise A.10.14 (Likelihood ratio and Wald test for the variance of a Normal model)**.** Let $X_1,\,\ldots\,,\,X_n \sim N(\mu,\,\sigma^2)$. Construct the likelihood ratio test for

$$H_0 : \sigma = \sigma_0 \text{ versus } H_1 : \sigma \neq \sigma_0.$$

Compare to the Wald test.

Solution. The likelihood function $\mathcal{L}_n$ and the parameter space Θ are as in Exercise A.10.13 above, but now

$$\Theta_0 := \left\{(\mu,\,\sigma_0^2) : \mu \in \mathbb{R}\right\}.$$

Therefore the MLE over Θ is as above, namely $\hat\mu = \overline{X}_n$ and $\hat\sigma^2 = S_P^2$, such that, as above,

$$\mathcal{L}_n(\hat\mu,\,\hat\sigma^2) = \left(2\pi S_P^2\right)^{-n/2} e^{-n/2}.$$

We now compute the MLE over Θ_0. Since

$$l_n(\mu,\,\sigma_0^2) = \log\mathcal{L}_n(\mu,\,\sigma_0^2) = -\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i - \mu)^2 + C$$

it follows that

$$\partial_\mu l_n = -\frac{1}{\sigma_0^2}\sum_{i=1}^n (X_i - \mu) = -\frac{n}{\sigma_0^2}\left(\overline{X}_n - \mu\right)$$

and so

$$\hat\mu_0 = \overline{X}_n.$$

In particular note that the MLE $\hat\mu_0$ over Θ_0 is equal to the MLE $\hat\mu$ over the full parameter space Θ. Evaluating the likelihood function at the MLE over Θ_0 now yields

$$\mathcal{L}_n(\hat{\mu}_0,\, \sigma_0^2) = \left(2\pi\sigma_0^2\right)^{-n/2} \exp\left(- \frac{1}{2\sigma_0^2} \overbrace{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}^{=nS_P^2} \right)$$

$$= \left(2\pi\sigma_0^2\right)^{-n/2} \exp\left(- \frac{nS_P^2}{2\sigma_0^2} \right).$$

Therefore the likelihood ratio is

$$\frac{\mathcal{L}_n(\hat{\mu},\, \hat{\sigma}^2)}{\mathcal{L}_n(\hat{\mu}_0,\, \sigma_0^2)} = \frac{\left(S_P^2\right)^{-n/2} \exp\left(-\frac{n}{2}\right)}{\left(\sigma_0^2\right)^{-n/2} \exp\left(-\frac{nS_P^2}{2\sigma_0^2}\right)} = \left(\frac{S_P^2}{\sigma_0^2}\right)^{-n/2} \exp\left[\frac{n}{2}\left(\frac{S_P^2}{\sigma_0^2} - 1\right)\right]$$

and so the likelihood ratio statistic is

$$\lambda = 2 \log \frac{\mathcal{L}_n(\hat{\mu},\, \hat{\sigma}^2)}{\mathcal{L}_n(\hat{\mu}_0,\, \sigma_0^2)} = -n \left(\log \frac{S_P^2}{\sigma_0^2} - \frac{S_P^2}{\sigma_0^2} + 1 \right).$$

In particular note that $\log x - (x - 1)$ is precisely subtracting the first order Taylor polynomial of $\log x$ about $x = 1$. Finally recall that the rejection region is

$$\left\{ \lambda \in \mathbb{R} : \lambda > \chi^2_{1,\, \alpha} \right\}.$$

We now turn our attention to the Wald test. Recall from item 3 of Exercise A.9.6 that the inverse of the Fisher information matrix is

$$J_n(\mu,\, \sigma^2) = \frac{\sigma^2}{n} \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

Therefore, for

$$\widehat{se} := \left[J_n(\hat{\mu},\, \hat{\sigma}^2)\right]_{22}^{1/2} = \sqrt{\frac{\hat{\sigma}^2}{2n}} = \sqrt{\frac{S_P^2}{2n}},$$

the asymptotic normality of the MLE for multiparameter models tells us that

$$W := \frac{\hat{\sigma} - \sigma_0}{\widehat{se}} = \frac{\sqrt{2n}\left(\sqrt{S_P^2} - \sigma_0\right)}{\sqrt{S_P^2}} = \sqrt{2n}\left(1 - \sqrt{\frac{\sigma_0^2}{S_P^2}}\right) \rightsquigarrow N(0,\, 1)$$

and thus that it is a valid Wald test statistic, where recall that the rejection region is

$$\left\{ w \in \mathbb{R} : |w| > z_{\alpha/2} \right\}.$$

Exercise A.10.15 (Likelihood ratio and Wald test for a Binomial model)**.** Consider $X_1,\, \dots,\, X_n \sim \text{Binomial}(n,\, p)$. Construct the likelihood ratio test for

$$H_0 : p = p_0 \text{ versus } H_1 : p \neq p_0.$$

Compare to the Wald test.

Remark A.3 (Likelihood ratio and Wald test for a Binomial model)**.** Exercise A.10.15 immediately above should really be thought of as studying a *Bernoulli*(p) model where we are sampling n IID Bernoulli random variables and recording their sum as a Binomial$(n,\, p)$.

This means that the full parameter space is $\Theta = [0,\, 1] \ni p$, i.e. n is *not* a parameter, instead it is thought of as given and approaching infinity. In particular we are only taking *one* sample from the Binomial$(n,\, p)$ distribution.

Solution. First we construct the likelihood ratio test. The likelihood function is

$$\mathcal{L} = \mathcal{L}_1 = f(X; n, p) = \binom{n}{X} p^X (1-p)^{n-X}$$

and the parameter space is

$$\Theta = [0,\, 1] \ni p$$

while $\Theta_0 = \{p_0\}$. To compute the MLE over Θ we note that the log-likelihood function is

$$l(p) = \log \mathcal{L} = C + X \log p + (n - X) \log(1-p)$$

and so

$$l' = \frac{X}{p} - \frac{n-X}{1-p} = \frac{X - np}{p(1-p)}.$$

Therefore the MLE is

$$\hat{p} = \frac{X}{n}.$$

Note that, trivially, the MLE over Θ_0 is $\hat{p}_0$. Therefore the likelihood ratio is

$$\frac{\mathcal{L}(\hat{p})}{\mathcal{L}(\hat{p}_0)} = \frac{\hat{p}^X (1-\hat{p})^{n-X}}{p_0^X (1-p_0)^{n-X}} = \left(\frac{\hat{p}}{p_0}\right)^X \left(\frac{1-\hat{p}}{1-p_0}\right)^{n-X}$$

and so the likelihood ratio statistic is

$$\lambda := 2 \log \frac{\mathcal{L}(\hat{p})}{\mathcal{L}(\hat{p}_0)} = 2 \left[X \log \frac{\hat{p}}{p_0} + (n-X) \log \frac{1-\hat{p}}{1-p_0} \right],$$

with rejection region given by, since $\dim \Theta - \dim \Theta_0 = 1 - 0 = 1$,

$$\left\{ \lambda \in \mathbb{R} : \lambda > \chi^2_{1,\,\alpha} \right\}.$$

We now turn our attention to the Wald test. Since the Fisher information is

$$I_n(p) = -\mathbb{E}\left[l''(p)\right] = -\mathbb{E}\left[\frac{X(2p-1) - np^2}{p^2(1-p)^2}\right] = -\frac{np(1-p)}{p^2(1-p)^2} = \frac{n}{p(1-p)}$$

we define

$$\widehat{se} := \sqrt{1/I_n(\hat{p})} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

such that, by asymptotic normality of the MLE,

$$W := \frac{\hat{p} - p_0}{\widehat{se}} = \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} \rightsquigarrow N(0,\, 1)$$

and is thus a valid Wald test statistic, with rejection region

$$\left\{ w \in \mathbb{R} : |w| > z_{\alpha/2} \right\}.$$

Exercise A.10.16 (Comparing the Wald and likelihood ratio statistics)**.** Let θ be a scalar parameter and suppose we test

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0.$$

Let W be the Wald test statistic based on the MLE and let λ be the likelihood ratio test statistic. Show that these tests are equivalent in the sense that

$$\frac{W^2}{\lambda} \xrightarrow{P} 1$$

as $n \to \infty$. Hint: Use a Taylor expansion of the log-likelihood $l(\theta)$ to show that

$$\lambda \approx \left[\sqrt{n}(\hat{\theta} - \theta_0)\right]^2 \left(-\frac{1}{n} l''(\theta)\right)$$

where $\hat{\theta}$ denotes the MLE.

Solution. We note that, in this case, the likelihood ratio statistic takes the form

$$\lambda = 2 \log \frac{\mathcal{L}_n(\hat{\theta})}{\mathcal{L}_n(\theta_0)} = 2 l_n(\hat{\theta}) - 2 l_n(\theta_0).$$

In particular a Taylor expansion of l_n about $\hat{\theta}$ tells us that

$$l_n(\theta_0) \approx l_n(\hat{\theta}) + \underbrace{l_n'(\hat{\theta})}_{=0}(\theta_0 - \hat{\theta}) + \frac{1}{2} l_n''(\hat{\theta})(\theta_0 - \hat{\theta})^2$$

and so

$$\lambda = 2\left[l_n(\hat{\theta}) - l_n(\theta_0)\right] \approx -l_n''(\hat{\theta})(\theta_0 - \hat{\theta})^2 = \left[\sqrt{n}(\hat{\theta} - \theta_0)\right]^2 \left(-\frac{1}{n} l_n''(\hat{\theta})\right).$$

In particular since

$$l_n = nl \text{ and } I(\theta) = -\mathbb{E}[l''(\theta)]$$

it follows that

$$\frac{-\frac{1}{n} l_n''(\hat{\theta})}{I(\hat{\theta})} = \frac{-l''(\hat{\theta})}{I(\hat{\theta})} \xrightarrow{P} 1.$$

Since the MLE–based Wald statistic is

$$W = \frac{\hat{\theta} - \theta_0}{\sqrt{1/I_n(\hat{\theta})}} = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{1/I(\hat{\theta})}},$$

we may therefore conclude that

$$\frac{W^2}{\lambda} \approx \frac{\frac{\left[\sqrt{n}(\hat{\theta} - \theta_0)\right]^2}{1/I(\hat{\theta})}}{\left[\sqrt{n}(\hat{\theta} - \theta_0)\right]^2 \left(-\frac{1}{n} l_n''(\hat{\theta})\right)} = \frac{I(\hat{\theta})}{-\frac{1}{n} l_n''(\hat{\theta})} \xrightarrow{P} 1$$

as $n \to \infty$ as desired.

A.11. **Bayesian Inference.**

Exercise A.11.1 (Normals are conjugate priors)**.** Let $X_1, \dots, X_n \sim N(\theta, \sigma^2)$. For simplicity let us assume that σ is known. Suppose that we take as prior $\theta \sim N(a, b^2)$. Show that the posterior for θ is $N(\bar{\theta}, \tau^2)$ where

$$\bar{\theta} = w\overline{X} + (1-w)a \text{ and } \frac{1}{\tau^2} = \frac{1}{se^2} + \frac{1}{b^2},$$

where $\overline{X}$ denotes the sample mean, for

$$se = \frac{\sigma}{\sqrt{n}} \text{ and } w = \frac{\tau^2}{se^2}.$$

Solution. The likelihood is

$$\mathcal{L}_n \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \theta)^2\right]$$

and so the posterior distribution is proportional to

$$\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \theta)^2\right]\exp\left[-\frac{1}{2b^2}(\theta - a)^2\right].$$

First we compute that

$$\sum_{i=1}^n (X_i - \theta)^2 = \sum_{i=1}^n \left(X_i^2 - 2X_i\theta + \theta^2\right) = \sum_{i=1}^n X_i^2 - 2n\overline{X}\theta + n\theta^2$$

and so, since factors independent of θ may be dismissed, the posterior distribution is proportional to

$$\exp\left[-\frac{1}{2\sigma^2}(n\theta^2 - 2n\overline{X}\theta)^2 - \frac{1}{2b^2}(\theta^2 - 2a\theta)^2\right].$$

We now rearrange

$$\begin{aligned}
-\frac{1}{2\sigma^2}(n\theta^2 - 2n\overline{X}\theta)^2 - \frac{1}{2b^2}(\theta^2 - 2a\theta)^2 &= -\frac{1}{2}\theta^2\left(\frac{n}{\sigma^2} + \frac{1}{b^2}\right) + \frac{n\overline{X}\theta}{\sigma^2} + \frac{a\theta}{b^2} \\
&= -\frac{1}{2\tau^2}\theta^2 + \theta\left(\frac{n\overline{X}}{\sigma^2} + \frac{a}{b^2}\right) \\
&= -\frac{1}{2\tau^2}\theta^2 + \frac{1}{2\tau^2}\cdot 2\cdot\theta\cdot\tau^2\left(\frac{n\overline{X}}{\sigma^2} + \frac{a}{b^2}\right)
\end{aligned}$$

where

$$\tau^2\left(\frac{n\overline{X}}{\sigma^2} + \frac{a}{b^2}\right) = \frac{\tau^2}{se^2}\overline{X} + \frac{\tau^2}{b^2}a = w\overline{X} + (1-w)a = \bar{\theta}$$

since, by definition of τ,

$$\frac{\tau^2}{b^2} = \tau^2\left(\frac{1}{\tau^2} - \frac{1}{se^2}\right) = 1 - w.$$

So finally the posterior distribution is proportional to

$$\exp\left(-\frac{1}{2\tau^2}\theta^2 + \frac{1}{2\tau^2}\cdot 2\cdot\theta\cdot\bar{\theta}\right) \propto \exp\left[-\frac{1}{2\tau^2}(\theta - \bar{\theta})^2\right],$$

which verifies that indeed the posterior distribution is $N(\bar{\theta}, \tau^2)$.

Exercise A.11.2 (Bayesian inference and simulation for normal and log-normal models)**.** Let $X_1, \ldots, X_n \sim N(\mu, 1)$.

(1) Simulate a data set (using $\mu = 5$) consisting of $n = 100$ observations.
(2) Take $f(\mu) = 1$ and find the posterior density. Plot the density.
(3) Simulate 1,000 draws from the posterior. Plot a histogram of the simulated values and compare the histogram to the answer in item 2.
(4) Let $\theta = e^\mu$. Find the posterior density for θ analytically and by simulation.
(5) Find a 95 percent posterior interval for μ.
(6) Find a 95 percent posterior interval for θ.

Solution. (1) See Jupyter notebook.

(2) Since the prior is flat, the posterior is proportional to the likelihood, namely

$$(2\pi)^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2\right]$$

where

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n} X_i^2 - 2n\overline{X}_n\mu + n\mu^2$$

and so the posterior is proportional to

$$\begin{aligned}\exp\left[-\frac{1}{2}\left(n\mu^2 - 2n\overline{X}_n\mu\right)\right] &= \exp\left[-\frac{n}{2}\left(\mu^2 - 2\overline{X}_n\mu\right)\right]\\ &= \exp\left[-\frac{n}{2}\left(\mu - \overline{X}_n\right)^2 + C\right]\\ &\propto \exp\left[-\frac{n}{2}\left(\mu - \overline{X}_n\right)^2\right],\end{aligned}$$

which means that the posterior is a $N(\overline{X}_n, 1/n)$ distribution.

(3) See Jupyter notebook.

(4) We note that

$$e^\mu \leqslant \theta \iff \mu \leqslant \log\theta$$

and so the posterior CDF of θ is

$$\begin{aligned}H(\theta \mid X_1, \ldots, X_n) &= \int_{\{e^\mu \leqslant \theta\}} f(\mu \mid X_1, \ldots, X_n)\, d\mu\\ &= \int_{-\infty}^{\log\theta} f(\mu \mid X_1, \ldots, X_n)\, d\mu.\end{aligned}$$

Therefore the posterior CDF of θ is

$$\begin{aligned}h(\theta \mid X_1, \ldots, X_n) &= \frac{d}{d\theta}\int_{-\infty}^{\log\theta} f(\mu \mid X_1, \ldots, X_n)\, d\mu\\ &= \frac{1}{\theta} f(\theta \mid X_1, \ldots, X_n)\\ &= \frac{1}{\theta}\sqrt{\frac{n}{2\pi}} \exp\left[-\frac{n}{2}\left(\log\theta - \overline{X}_n\right)^2\right],\end{aligned}$$

which is indeed the PDF of a log-normal distribution characterised by $\log\theta \sim N(\overline{X}_n, 1/n)$.

(5) If we denote by F the posterior CDF of μ then we are looking for a and b such that

$$\begin{aligned} F(a) &= \int_{-\infty}^{a} f\left(\mu \,|\, X_1,\, \ldots,\, X_n\right) d\mu \\ &= 1 - F(b) = \int_{b}^{\infty} f\left(\mu \,|\, X_1,\, \ldots,\, X_n\right) d\mu = \frac{\alpha}{2}. \end{aligned}$$

In other words

$$a = F^{-1}\left(\frac{\alpha}{2}\right) \text{ and } b = F^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Since the posterior of μ is a $N(\overline{X}_n,\, \frac{1}{n})$ distribution its CDF is

$$F(\mu) = \Phi\left(\sqrt{n}(\mu - \overline{X}_n)\right)$$

for Φ denoting the CDF of a standard normal. Therefore

$$F^{-1}(q) = \overline{X}_n + \frac{1}{\sqrt{n}}\Phi^{-1}(q)$$

and so

$$a = \overline{X}_n + \frac{1}{\sqrt{n}}\Phi^{-1}\left(\frac{\alpha}{2}\right) \text{ and } b = \overline{X}_n + \frac{1}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

where $-\Phi^{-1}(\alpha/2) = \Phi^{-1}(1 - \alpha/2) = z_{\alpha/2}$. So the 95 percent posterior interval for μ is

$$\overline{X}_n \pm \frac{z_{\alpha/2}}{\sqrt{n}}.$$

(6) We denote by H the posterior CDF of θ. Then, proceeding as in item 5 above, we choose

$$a = H^{-1}\left(\frac{\alpha}{2}\right) \text{ and } b = H^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Crucially we have (essentially) obtained in item 4 that

$$H(\theta) = F(\log\theta)$$

and so, by item 5 where F was obtained,

$$H(\theta) = \Phi\left(\sqrt{n}(\log\theta - \overline{X}_n)\right).$$

Therefore

$$H^{-1}(q) = \exp\left(\overline{X}_n + \frac{1}{\sqrt{n}}\Phi^{-1}(q)\right)$$

and so

$$\begin{aligned} a &= H^{-1}\left(\frac{\alpha}{2}\right) = \exp\left(\overline{X}_n - \frac{z_{\alpha/2}}{\sqrt{n}}\right) \text{ and} \\ b &= H^{-1}\left(1 - \frac{\alpha}{2}\right) = \exp\left(\overline{X}_n + \frac{z_{\alpha/2}}{\sqrt{n}}\right). \end{aligned}$$

In other words: if we denote by C and D the posterior intervals for μ and θ, respectively, then since $\theta = e^{\mu}$ it follows that $D = e^{C}$.

Exercise A.11.3 (Improper prior for a uniform model)**.** Consider an IID sample $X_1,\, \ldots,\, X_n \sim \text{Uniform}(0,\, \theta)$. Let $f(\theta) \propto \frac{1}{\theta}$. Find the posterior density.

Solution. As computed in Exercise A.9.2 the likelihood function is

$$\mathcal{L}_n(\theta) = \frac{1}{\theta^n} \mathbb{1}(X_{(n)} \leqslant \theta)$$

where $X_{(n)} := \max\{X_1, \ldots, X_n\}$. Therefore the posterior distribution is proportional to

$$\mathcal{L}_n(\theta) f(\theta) = \frac{1}{\theta^{n+1}} \mathbb{1}(\theta \geqslant X_{(n)}).$$

Since, for any $x > 0$,

$$\int_x^\infty \frac{1}{\theta^{n+1}} d\theta = \frac{-1}{n\theta^n} |_{\theta = x}^{\theta = \infty} = \frac{1}{nx^n}$$

it follows that

$$f(\theta \,|\, X_1, \ldots, X_n) = \frac{nX_{(n)}}{\theta^{n+1}} \mathbb{1}(\theta \geqslant X_{(n)}).$$

Exercise A.11.4 (Comparing two binomials)**.** Suppose that 50 people are given a placebo and 50 are given a new treatment. 30 placebo patients show improvement while 40 treated patients show improvement. Let $\tau = p_2 - p_1$, where p_2 is the probability of improving under treatment and p_1 is the probability of improving under placebo.

(1) Find the MLE of τ. Find the standard error and 90 percent confidence interval using the delta method.
(2) Find the standard error and 90 percent confidence interval using the parametric bootstrap.
(3) Use the prior $f(p_1, p_2) = 1$. Use simulation to find the posterior mean and posterior 90 percent interval for τ.
(4) Let

$$\psi = \log\left(\frac{p_1}{1 - p_1} \div \frac{p_2}{1 - p_2}\right)$$

be the log-odds ratio. Note that $\psi = 0$ if $p_1 = p_2$. Find the MLE of ψ. Use the delta method to find a 90 percent confidence interval for ψ.
(5) Use simulation to find the posterior mean and posterior 90 percent interval for ψ.

Solution. Let us write $X_1, \ldots, X_{50} \sim \text{Bernoulli}(p_1)$ and $Y_1, \ldots, Y_{50} \sim \text{Bernoulli}(p_2)$ such that

$$\sum_{i=1}^{50} X_i = 30 \text{ and } \sum_{i=1}^{50} Y_i = 40.$$

(1) For Bernoulli models we know that the MLE is $\hat{p} = \overline{X}_n$, i.e.

$$\hat{p}_1 = \overline{X}_n \text{ and } \hat{p}_2 = \overline{Y}_n$$

for $n = 50$. Therefore equivariance of the MLE tells us that the MLE for τ is

$$\hat{\tau} = \hat{p}_2 - \hat{p}_1 = \overline{Y}_n - \overline{X}_n.$$

As computed in Exercise A.9.7 the Fisher information matrix is

$$I(p_1, p_2) = \begin{pmatrix} \frac{1}{p_1(1-p_1)} & 0 \\ 0 & \frac{1}{p_2(1-p_2)} \end{pmatrix}$$

and so, since $\tau = g(p_1,\, p_2)$ for $g(p_1,\, p_2) := p_2 - p_1$ with

$$\nabla g = \begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

the delta method tells us that if we define

$$\begin{aligned} \widehat{se}(\hat{\tau}) &:= \sqrt{(\nabla g)^T [nI(p_1,\, p_2)]^{-1} \nabla g}\,\Big|_{(p_1,\, p_2)=(\hat{p}_1,\, \hat{p}_2)} \\ &= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)}{n}} \end{aligned}$$

then

$$\frac{\hat{\tau} - \tau}{\widehat{se}} \rightsquigarrow N(0,\, 1)$$

as $n \to \infty$. The corresponding 90 percent confidence interval is thus, for $\alpha = 0.1$,

$$\hat{\tau} \pm z_{\alpha/2}\widehat{se}$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ for Φ the CDF of a standard Normal.

(2) To estimate the standard error of $\hat{\tau}$ via the parametric bootstrap we fix $B \geqslant 1$ and sample, for $1 \leqslant j \leqslant B$,

$$\begin{aligned} &X^*_{1,j},\, \ldots,\, X^*_{n,j} \sim \text{Bernoulli}(\overline{X}_n) \text{ and} \\ &Y^*_{1,j},\, \ldots,\, Y^*_{n,j} \sim \text{Bernoulli}(\overline{Y}_n). \end{aligned}$$

We then compute

$$\hat{\tau}^*_j := \overline{Y}^*_j - \overline{X}^*_j$$

and

$$\widehat{se}^2_{boot}(\hat{\tau}) := \frac{1}{B}\sum_{j=1}^{B}\left(\hat{\tau}^*_j - \frac{1}{B}\sum_{k=1}^{B}\hat{\tau}^*_k\right)^2,$$

which is really the population variance of $\hat{\tau}^*_1,\, \ldots,\, \hat{\tau}^*_B$. The corresponding Normal 90 percent confidence interval is then

$$\hat{\tau} \pm z_{\alpha/2}\widehat{se}_{boot}(\hat{\tau})$$

for $\alpha = 0.1$ and $z_{\alpha/2}$ as above.

Alternatively we may obtain a bootstrap percentile interval directly as

$$\left(\tau^*_{(\alpha/2)},\, \tau^*_{(1-\alpha/2)}\right)$$

where, for any $\beta \in (0,\, 1)$, τ^*_β denotes the β–quantile of $\hat{\tau}^*_1,\, \ldots,\, \hat{\tau}^*_B$.

(3) Since the prior is constant, the posterior is directly proportional to the likelihood. Here the likelihood is

$$\begin{aligned} \mathcal{L}_n(p_1,\, p_2) &= \left[\prod_{i=1}^{n} p_1^{X_i}(1-p_1)^{1-X_i}\right]\left[\prod_{i=1}^{n} p_2^{Y_i}(1-p_2)^{1-Y_i}\right] \\ &= p_1^{n\overline{X}_n}(1-p_1)^{n(1-\overline{X}_n)} p_2^{n\overline{Y}_n}(1-p_2)^{n(1-\overline{Y}_n)}, \end{aligned}$$

which is proportional to the product of two Beta distributions. In other words, if we denote by $f(\,\cdot\,;\alpha,\, \beta)$ the PDF of a Beta$(\alpha,\, \beta)$ distribution, then the joint posterior is

$$f_{post}(p_1,\, p_2) = f\left(p_1; n\overline{X}_n, n(1-\overline{X}_n)\right) f\left(p_2; n\overline{Y}_n, n(1-\overline{Y}_n)\right).$$

To estimate the posterior mean and a posterior 90 percent confidence interval by simulation we draw

$$\left(p_1^{(1)},\, p_n^{(1)}\right),\, \ldots,\, \left(p_1^{(B)},\, p_n^{(B)}\right) \sim f_{post}(p_1,\, p_2),$$

or equivalently draw

$$\begin{aligned} p_1^{(1)},\, \ldots,\, p_1^{(B)} &\sim \text{Beta}\left(n\overline{X}_n,\, n(1-\overline{X}_n)\right) \text{ and} \\ p_2^{(1)},\, \ldots,\, p_2^{(B)} &\sim \text{Beta}\left(n\overline{Y}_n,\, n(1-\overline{Y}_n)\right) \end{aligned}$$

and define

$$\tau^{(j)} := p_2^{(j)} - p_1^{(j)}$$

for $1 \leqslant j \leqslant B$. The estimate of the posterior mean is then

$$\hat{\tau}_{sim} := \frac{1}{B}\sum_{j=1}^{B} \tau^{(j)}$$

and the estimate of the standard error is

$$\widehat{se}^2_{sim}(\hat{\tau}) := \frac{1}{B}\sum_{j=1}^{B}\left(\tau^{(j)} - \frac{1}{B}\sum_{k=1}^{B}\tau^{(k)}\right)^2,$$

which, as in the case of the bootstrap, is really just the population variance of $\tau^{(1)},\, \ldots,\, \tau^{(B)}$. The corresponding Normal 90 percent confidence interval is then

$$\hat{\tau}_{sim} \pm z_{\alpha/2}\widehat{se}_{sim}(\hat{\tau})$$

for $\alpha = 0.1$ and $z_{\alpha/2}$ as above while the percentile confidence interval is

$$\left(\tau_{(\alpha/2)},\, \tau_{(1-\alpha/2)}\right)$$

where $\tau_{(\beta)}$ denotes the β–quantile of $\tau^{(1)},\, \ldots,\, \tau^{(B)}$ for any $\beta \in (0,\, 1)$.

(4) By equivariance of the MLE, the MLE for ψ is

$$\hat{\psi} = \log\left(\frac{\hat{p}_1}{1-\hat{p}_1} \div \frac{\hat{p}_2}{1-\hat{p}_2}\right) = \log\left(\frac{\overline{X}_n}{1-\overline{X}_n} \div \frac{\overline{Y}_n}{1-\overline{Y}_n}\right).$$

Since $\psi = h(p_1,\, p_2)$ for

$$\begin{aligned} h(p_1,\, p_2) &= \log\left(\frac{p_1}{1-p_1} \div \frac{p_2}{1-p_2}\right) \\ &= \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right), \end{aligned}$$

such that

$$\nabla h = \begin{pmatrix} \frac{1}{p_1(1-p_1)} \\ \frac{1}{p_2(1-p_2)} \end{pmatrix},$$

we define

$$\begin{aligned} \widehat{se}(\hat{\psi}) &:= \sqrt{(\nabla h)^T (nI)^{-1} \nabla h}\,\Big|_{(p_1,\, p_2) = (\hat{p}_1,\, \hat{p}_2)} \\ &= \sqrt{\frac{1}{n}\left(\frac{1}{\hat{p}_1(1-\hat{p}_1)} + \frac{1}{\hat{p}_2(1-\hat{p}_2)}\right)}. \end{aligned}$$

By the delta method a 90 percent confidence interval for ψ is thus

$$\hat{\psi} \pm z_{\alpha/2}\widehat{se}(\psi).$$

(5) We proceed as in item 3 and define

$$\psi^{(j)} := h\left(p_1^{(j)},\, p_2^{(j)}\right) = \log\left(\frac{p_1^{(j)}}{1-p_1^{(j)}} \div \frac{p_2^{(j)}}{1-p_2^{(j)}}\right)$$

for $1 \leqslant j \leqslant B$. The posterior mean is then

$$\hat{\psi}_{sim} := \frac{1}{B}\sum_{j=1}^{B} \psi^{(j)}$$

and the estimate of the standard error is

$$\widehat{se}^2_{sim}(\hat{\psi}) := \frac{1}{B}\sum_{j=1}^{B}\left(\psi^{(j)} - \frac{1}{B}\sum_{j=k}^{B}\psi^{(k)}\right)^2,$$

i.e. the population variance of $\psi^{(1)}, \dots, \psi^{(B)}$. The corresponding Normal 90 percent confidence interval is then

$$\hat{\psi}_{sim} \pm z_{\alpha/2}\widehat{se}_{sim}(\hat{\psi})$$

for $\alpha = 0.1$ and $z_{\alpha/2}$ as above while the percentile confidence interval is

$$\left(\psi_{(\alpha/2)},\, \psi_{(1-\alpha/2)}\right)$$

where $\psi_{(\beta)}$ denotes the β–quantile of $\psi^{(1)}, \dots, \psi^{(B)}$ for any $\beta \in (0,\, 1)$.

Exercise A.11.5 (Beta prior for a Bernoulli model)**.** Consider the Bernoulli(p) observations

$$0 \qquad 1 \qquad 0 \qquad 1 \qquad 0 \qquad 0 \qquad 0 \qquad 0 \qquad 0 \qquad 0.$$

Plot the posterior for p using these priors: Beta(0.5, 0.5), Beta(1, 1), Beta(10, 10), and Beta(100, 100).

Solution. Recall that the PDF of a Beta(α, β) distribution is

$$f(p; \alpha, \beta) = C(\alpha, \beta)p^{\alpha-1}(1-p)^{\beta-1}$$

for $x \in (0,\, 1)$. The likelihood function for Bernoulli random variables is

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} p^{X_i}(1-p_i)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p_i)^{n-\sum_{i=1}^{n} X_i} = p^s(1-p)^{n-s}$$

for $s := \sum_{i=1}^{n} X_i$. Therefore, if the prior distribution is Beta(α, β), then the posterior distribution is proportional to

$$\begin{aligned}\mathcal{L}_n(p)f(p; \alpha, \beta) &\propto p^s(1-p)^{n-s}p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto f(p; \alpha+s-1, \beta+n-s-1).\end{aligned}$$

In other words the Bernoulli posterior of a Beta(α, β) prior is

$$\text{Beta}(\alpha + s,\, \beta + n - s).$$

In particular, here

$$s = 2 \text{ and } n = 10, \text{ such that } n - s = 8,$$

and so the priors and posteriors are as follow:

Prior	Posterior
Beta(0.5, 0.5)	Beta(2.5, 8.5)
Beta(1, 1)	Beta(3, 9)
Beta(10, 10)	Beta(12, 18)
Beta(100, 100)	Beta(102, 108)

Exercise A.11.6 (Gamma and Jeffreys' prior for a Poisson model)**.** Consider a sample $X_1, \dots, X_n \sim \text{Poisson}(\lambda)$.

(1) Let $\lambda \sim \text{Gamma}(\alpha, \beta)$ be the prior. Show that the posterior is also a Gamma. Find the posterior mean.
(2) Find the Jeffrey's prior. Find the posterior.

Solution. (1) As computed in Exercise A.9.4 the likelihood is

$$\mathcal{L}_n = e^{-n\lambda}\lambda^{n\overline{X}}\prod_{i=1}^{n}\frac{1}{X_i!},$$

where $\overline{X}$ denotes the sample mean, and recall that the PDF of a Gamma(α, β) distribution takes the form

$$f(x;\alpha,\beta) = C(\alpha,\beta)\lambda^{\alpha-1}e^{-\lambda/\beta}.$$

Therefore the posterior distribution is proportional to, for $s := \sum_{i=1}^{n} X_i$,

$$e^{-n\lambda}\lambda^{s}\lambda^{\alpha-1}e^{-\lambda/\beta} = \lambda^{\alpha+s-1}e^{-\lambda\left(\frac{1}{\beta}+n\right)}$$

and so the posterior distribution is

$$\text{Gamma}\left(\alpha+s,\ \frac{1}{\frac{1}{\beta}+n}\right)$$

distribution. Since the mean of a Gamma(α, β) distribution is $\alpha\beta$ it follows that the posterior mean is

$$\frac{\alpha+s}{\frac{1}{\beta}+n} = \frac{1}{\beta n+1}\alpha\beta + \left(1-\frac{1}{\beta n+1}\right)\frac{s}{n}$$

i.e., since $\frac{s}{n} = \overline{X} =: \hat{\mu}$, which is the MLE,

$$\mu_{post} = \frac{1}{\beta n+1}\mu_{prior} + \left(1-\frac{1}{\beta n+1}\right)\hat{\mu}.$$

(2) As computed in Exercise A.9.4 the Fisher information is

$$I(\lambda) = \frac{1}{\lambda}.$$

The Jeffreys' (improper) prior is therefore

$$f(\lambda) \propto \sqrt{I(\lambda)} = \lambda^{-1/2}.$$

The posterior is then proportional to

$$e^{-n\lambda}\lambda^{s}\cdot\lambda^{-1/2} = e^{-n\lambda}\lambda^{s-1/2}$$

which means that the posterior is a

$$\text{Gamma}\left(s+\frac{1}{2},\ \frac{1}{n}\right)$$

distribution. Note that we could also have viewed the prior formally as a degenerate Gamma $\left(\frac{1}{2}, \infty\right)$ distribution, such that the computations of item 1 above would immediately tell us that the posterior is a

$$\text{Gamma}\left(\frac{1}{2}+s,\ \lim_{\beta\to\infty}\frac{1}{\frac{1}{\beta}+n}\right) = \text{Gamma}\left(\frac{1}{2}+s,\ \frac{1}{n}\right)$$

distribution.

Exercise A.11.7 (Horwitz-Thompson estimator)**.** Let

$$\theta = (\theta_1, \ldots, \theta_B)$$

be a vector of unkown parameters such that $0 \leqslant \theta_j \leqslant 1$ for $1 \leqslant j \leqslant B$. Let

$$\xi = (\xi_1, \ldots, \xi_B)$$

be a vector of *known* numbers such that $0 < \delta \leqslant \xi \leqslant 1-\delta < 1$ for $1 \leqslant j \leqslant B$ where δ is some, small, positive number. Each data point is drawn in the following way (for $1 \leqslant i \leqslant n$).

(1) Draw X_i uniformly from $\{1, \ldots, B\}$.
(2) Draw $R_i \sim \text{Bernoulli}(\xi_{X_i})$.
(3) If $R_i = 1$, then draw $Y_i \sim \text{Bernoulli}(\theta_{X_i})$. If $R_i = 0$, do not draw Y_i.

The model may seem a little artificial but, in fact, it is a caricature of some real *missing data* problems in which some data points are not observed. In this example, $R_i = 0$ can be thought of as meaning "missing". Our goal is to estimate

$$\psi = \mathbb{P}(Y=1).$$

Note that

$$\psi = \mathbb{P}(Y=1) = \frac{1}{B}\sum_{j=1}^{B}\mathbb{P}(Y=1|X=j)\mathbb{P}(X=j) = \frac{1}{B}\sum_{j=1}^{B}\theta_j =: g(\theta)$$

so $\psi = g(\theta)$ is a function of θ. Define the *Horwitz-Thompson estimator*

$$\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_iY_i}{\xi_{X_i}}.$$

Show that

$$\mathbb{E}\hat{\psi} = \psi \text{ and } \mathbb{V}\hat{\psi} \leqslant \frac{1}{n\delta^2}.$$

Solution. First we note that

$$\mathbb{E}\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\frac{R_iY_i}{\xi_{X_i}}\right)$$

and so it suffices to compute $\mathbb{E}\left(\frac{RY}{\xi_X}\right)$. The rule of iterated expectation tells us that

$$\mathbb{E}\left(\frac{RY}{\xi_X}\right) = \mathbb{E}\left[\mathbb{E}\left(\frac{RY}{\xi_X}\,\middle|\,X\right)\right] = \frac{1}{B}\sum_{j=1}^{B}\mathbb{E}\left(\frac{RY}{\xi_j}\,\middle|\,X=j\right)$$

where, using the rule of iterated expectation again,

$$\begin{aligned}
\mathbb{E}\left(\frac{RY}{\xi_j}\,\middle|\, X=j\right) &= \frac{1}{\xi_j}\mathbb{E}\left[\mathbb{E}\left(RY \mid R,\, X=j\right)\right] \\
&= \frac{1}{\xi_j}\left[\mathbb{E}\left(Y \mid R=1,\, X=j\right)\mathbb{P}\left(R=1 \mid X=j\right) + 0\cdot\mathbb{P}\left(R=0 \mid X=j\right)\right] \\
&= \frac{1}{\xi_j}\mathbb{E}\left(Y \mid R=1,\, X=j\right)\cdot\xi_j \\
&= \mathbb{E}\left(Y \mid R=1,\, X=j\right) \\
&= 1\cdot\theta_j + 0\cdot(1-\theta_j) \\
&= \theta_j.
\end{aligned}$$

So finally we conclude that

$$\mathbb{E}\hat{\psi} = \mathbb{E}\left(\frac{RY}{\xi_X}\right) = \frac{1}{B}\sum_{j=1}^{B}\theta_j = \psi,$$

as desired. This establishes that $\hat{\psi}$ is a consistent estimator of ψ.

Now we turn our attention to the variance of this estimator. First note that, for $Z_i := \frac{R_i Y_i}{\xi_{X_i}}$, we may write $\hat{\psi} = \overline{Z}_n$. Therefore

$$\mathbb{V}\hat{\psi} = \frac{\mathbb{V}Z}{n}.$$

Since, as compute above, $\mathbb{E}Z = \psi$, we deduce from the rule of iterated expectation that

$$\begin{aligned}
\mathbb{V}Z &= \mathbb{E}(Z-\psi)^2 \\
&= \mathbb{E}\left[\left(\frac{RY}{\xi_X} - \psi\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{RY}{\xi_X} - \psi\right)^2 \,\middle|\, X\right]\right] \\
&= \frac{1}{B}\sum_{j=1}^{B}\mathbb{E}\left[\left(\frac{RY}{\xi_X} - \psi\right)^2 \,\middle|\, X=j\right]
\end{aligned}$$

where, since $\theta_j \geqslant 0$ and $0 < \delta < \xi_j$,

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{RY}{\xi_j} - \psi\right)^2 \middle| X = j\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{RY}{\xi_j} - \psi\right)^2 \middle| R,\, X = j\right]\right] \\
&= \mathbb{E}\left[\left(\frac{Y}{\xi_j} - \psi\right)^2 \middle| R = 1,\, X = j\right] \mathbb{P}\left(R = 1 \,|\, X = j\right) \\
&\quad + \mathbb{E}(\psi^2)\mathbb{P}\left(R = 0 \,|\, X = j\right) \\
&= \mathbb{E}\left[\left(\frac{Y}{\xi_j} - \psi\right)^2 \middle| R = 1,\, X = j\right] \cdot \xi_j + \psi^2(1 - \xi_j) \\
&= \left[\left(\frac{1}{\xi_j} - \psi\right)^2 \theta_j + \psi^2(1 - \theta_j)\right] \xi_j + \psi^2(1 - \xi_j) \\
&= \left(\frac{1}{\xi_j^2} - 2\frac{\psi}{\xi_j} + \psi^2\right) \theta_j\xi_j + \psi^2\xi_j - \psi^2\theta_j\xi_j + \psi^2 - \psi^2\xi_j \\
&= \frac{\theta_j}{\xi_j} - 2\psi\theta_j + \psi^2\theta_j\xi_j - \psi^2\theta_j\xi_h + \psi^2 \\
&= \frac{\theta_j}{\xi_j} - 2\psi\theta_j + \psi^2 \\
&\leqslant \frac{\theta_j}{\delta} - 2\psi\theta_j + \psi^2.
\end{aligned}$$

Therefore, since $\theta_j \leqslant 1$, which means that $\psi \leqslant 1$, we obtain that

$$\begin{aligned}
\mathbb{V}Z &\leqslant \frac{1}{B}\sum_{j=1}^{B}\left(\frac{\theta_j}{\delta} - 2\psi\theta_j + \psi^2\right) \\
&= \frac{\psi}{\delta} - 2\psi^2 + \psi^2 \\
&= \frac{\psi}{\delta} - \psi^2 \\
&\leqslant \frac{1}{\delta}.
\end{aligned}$$

So finally we conclude that

$$\mathbb{V}\hat{\psi} = \frac{1}{n}\mathbb{V}Z \leqslant \frac{1}{n\delta}.$$

Note that we actually obtain a *sharper* bound than asked for since $0 < \delta < 1$ and so $\frac{1}{\delta} < \frac{1}{\delta^2}$.

Exercise A.11.8 (Bayesian hypothesis testing)**.** Let $X \sim N(\mu,\, 1)$. Consider testing

$$H_0 : \mu = 0 \text{ versus } H_1 : \mu \neq 0.$$

Take $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$. Let the prior for μ under H_1 be $\mu \sim N(0,\, b^2)$. Find an expression for $\mathbb{P}(H_0|X = x)$. Compare $\mathbb{P}(H_0|X = x)$ to the p–value of the Wald test. Do the comparison numerically for a variety of values of x and b. Now repeat the problem using a sample of size n. You will see that the posterior probability of H_0 can be large even when the p–value is small, especially when n

is large. This disagreement between Bayesian and frequentist testing is called the Jeffreys–Lindley paradox.

Solution. Let ϕ denote the PDF of a standard Normal. By Bayes' Theorem and the fact that $\mathbb{P}(H_0) = \mathbb{P}(H_1)$ under our prior we compute that, since $X \sim N(0,\,1)$ under the null hypothesis,

$$\begin{aligned}\mathbb{P}\left(H_0 \,|\, X = x\right) &= \frac{P\left(H_0 \text{ and } X = x\right)}{\mathbb{P}\left(X = x\right)} \\ &= \frac{\mathbb{P}\left(X = x \,|\, H_0\right)\mathbb{P}(H_0)}{\mathbb{P}\left(X = x \,|\, H_0\right)\mathbb{P}(H_0) + \mathbb{P}\left(X = x \,|\, H_1\right)\mathbb{P}(H_1)} \\ &= \frac{\phi(x)}{\phi(x) + \mathbb{P}\left(X = x \,|\, H_1\right)}\end{aligned}$$

where, by the rule of iterated expectation (where the expectation is taken over the prior distribution of μ),

$$\begin{aligned}\mathbb{P}\left(X = x \,|\, H_1\right) &= \mathbb{E}\left[\mathbb{P}\left(X = x \,|\, H_1,\, \mu\right)\right] \\ &= \mathbb{E}\left[\mathbb{P}\left(X - \mu = x - \mu \,|\, H_1,\, \mu\right)\right] \\ &= \mathbb{E}\left[\phi(x - \mu)\right] \\ &= \int_{-\infty}^{+\infty} \phi(x-\mu) f(\mu) d\mu\end{aligned}$$

for f denoting the PDF of the prior distribution of μ. We compute that

$$\phi(x-\mu)f(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\mu)^2\right] \frac{1}{\sqrt{2\pi b^2}} \exp\left[-\frac{\mu^2}{2b^2}\right]$$

where, inspired by Exercise A.11.1, we define $\bar{\mu}$ and τ^2 via

$$\bar{\mu} = \tau^2 X \text{ and } \frac{1}{\tau^2} = 1 + \frac{1}{b^2}$$

such that

$$\begin{aligned}-\frac{1}{2}(x-\mu)^2 - \frac{\mu^2}{2b^2} &= -\frac{x^2}{2} - \frac{1}{2}\left(1 + \frac{1}{b^2}\right)\mu^2 + \frac{1}{2}\cdot 2x\mu \\ &= -\frac{x^2}{2} - \frac{1}{2\tau^2}(\mu - \tau^2 x)^2 + \frac{1}{2\tau^2}(\tau^2 x)^2 \\ &= (\tau^2 - 1)\frac{x^2}{2} - \frac{1}{2\tau^2}(\mu - \bar{\mu})^2.\end{aligned}$$

Therefore, since as shown in Exercise A.11.1, the posterior distribution of μ is $N(\bar{\mu},\, \tau^2)$,

$$\begin{aligned}\phi(x-\mu)f(\mu) &= \sqrt{\frac{2\pi\tau^2}{4\pi^2 b^2}} \exp\left[(\tau^2 - 1)\frac{x^2}{2}\right] \cdot \underbrace{\frac{1}{\sqrt{2\pi\tau^2}} \exp\left[-\frac{(\bar{\mu} - \mu)^2}{2\tau^2}\right]}_{f_{post}(\mu)} \\ &= \frac{\tau}{b} \exp\left(\frac{\tau^2 x^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \cdot f_{post}(\mu) \\ &= \frac{\tau}{b} \exp\left(\frac{\tau^2 x^2}{2}\right) \cdot \phi(x) \cdot f_{post}(\mu).\end{aligned}$$

So finally:

$$\mathbb{P}\left(H_0 \,|\, X = x\right) = \frac{\phi(x)}{\phi(x) + \frac{\tau}{b}\exp\left(\frac{\tau^2 x^2}{2}\right)\phi(x)} = \frac{1}{1 + \frac{\tau}{b}\exp\left(\frac{\tau^2 x^2}{2}\right)}$$
$$= \frac{1}{1 + \frac{\tau}{b}\exp\left(\frac{\bar{\mu}^2}{2\tau^2}\right)}.$$

By contrast, since $n = 1$ and since $\mathbb{V}_{H_0} X = 1$, the Wald test statistic is simply

$$W := X.$$

The p–value of the Wald test is thus

$$p - \text{value} = 2\Phi(-|X|)$$

(which is of course independent of b).

We now consider the case where $n > 1$. As before, since $\mathbb{P}(H_0) = \mathbb{P}(H_1)$, Bayes' Theorem tells us that

$$\begin{aligned}&\mathbb{P}\left(H_0 \,|\, X_1 = x_1, \ldots, X_n = x_n\right)\\ &= \frac{\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n \,|\, H_0\right)}{\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n \,|\, H_0\right) + \mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n \,|\, H_1\right)}\\ &= \frac{\prod_{i=1}^n \phi(x_i)}{\prod_{i=1}^n \phi(x_i) + \mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n \,|\, H_1\right)}\end{aligned}$$

where we may use the rule of iterated expectation once again to see that, taking expectation over the prior distribution of μ,

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n \,|\, H_1\right) = \mathbb{E}\left[\prod_{i=1}^n \phi(x_i - \mu)\right]$$
$$= \int_{-\infty}^{+\infty} \left[\prod_{i=1}^n \phi(x_i - \mu)\right] f(\mu) d\mu.$$

Similarly to how we proceeded above, where now $\bar{\mu}_n$ and τ_n^2 are defined by

$$\bar{\mu}_n = n\tau_n^2 \overline{X}_n \text{ and } \frac{1}{\tau_n^2} = n + \frac{1}{b^2},$$

we compute that

$$\left[\prod_{i=1}^n \phi(x_i - \mu)\right] f(\mu) = (2\pi)^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2\right] \left(2\pi b^2\right)^{-1/2} \exp\left(-\frac{\mu^2}{2b^2}\right)$$

where

$$\begin{aligned}
-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2-\frac{\mu^2}{2b^2} &= -\frac{1}{2}\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}x_i\mu-\frac{n\mu^2}{2}-\frac{\mu^2}{2b^2}\\
&= -\frac{1}{2}\sum_{i=1}^{n}x_i^2+n\bar{x}_n\mu-\frac{1}{2\tau_n^2}\mu^2\\
&= -\frac{1}{2}\sum_{i=1}^{n}x_i^2-\frac{1}{2\tau_n^2}\left(\mu-n\tau_n^2\bar{x}_n\right)^2+\frac{1}{2\tau_n^2}\left(n\tau_n^2\bar{x}_n\right)^2\\
&= -\frac{1}{2}\sum_{i=1}^{n}x_i^2-\frac{1}{2\tau_n^2}\left(\mu-\bar{\mu}_n\right)^2+\frac{n^2\tau_n^2\bar{x}_n^2}{2}
\end{aligned}$$

and so, since Exercise A.11.1 tells us that the posterior distribution of μ is $N(\bar{\mu}_n,\, \tau_n^2)$,

$$\begin{aligned}
&\left[\prod_{i=1}^{n}\phi(x_i-\mu)\right]f(\mu)\\
&= \sqrt{\frac{2\pi\tau_n^2}{(2\pi)^{2n}2\pi b^2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}x_i^2+\frac{n^2\tau_n^2\bar{x}_n^2}{2}\right)\cdot\frac{1}{\sqrt{2\pi\tau_n^2}}\exp\left(-\frac{\left(\mu-\bar{\mu}_n\right)^2}{2\tau_n^2}\right)\\
&= \frac{\tau_n}{b}\exp\left(\frac{n^2\tau_n^2\bar{x}_n^2}{2}\right)\cdot(2\pi)^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}x_i^2\right)\cdot f_{post}(\mu)\\
&= \frac{\tau_n}{b}\exp\left(\frac{n^2\tau_n^2\bar{x}_n^2}{2}\right)\cdot\left[\prod_{i=1}^{n}\phi(x_i)\right]\cdot f_{post}(\mu).
\end{aligned}$$

So finally

$$\begin{aligned}
\mathbb{P}\left(X_1=x_1,\,\dots,\,X_n=x_n\,|\,H_1\right) &= \frac{1}{1+\frac{\tau_n}{b}\exp\left(\frac{n^2\tau_n^2\bar{x}_n^2}{2}\right)}\\
&= \frac{1}{1+\frac{\tau_n}{b}\exp\left(\frac{\bar{\mu}_n^2}{2\tau_n^2}\right)}.
\end{aligned}$$

When $n>1$ the Wald statistic is now

$$W_n := \sqrt{n}\overline{X}_n$$

and the p–value the form

$$p-\text{value}=2\Phi(-|W_n|)=2\Phi(-\sqrt{n}|\overline{X}_n|).$$

A.12. **Statistical Decision Theory.**

Exercise A.12.1 (Bayes estimators for some conjugate priors)**.** In each of the following models, find the Bayes risk and the Bayes estimator using squared error loss.

(1) $X \sim$ Binomial$(n,\, p)$, $p \sim$ Beta$(\alpha,\, \beta)$.
(2) $X \sim$ Poisson(λ), $p \sim$ Gamma$(\alpha,\, \beta)$.
(3) $X \sim N(\theta,\, \sigma^2)$ where σ^2 is known and $\theta \sim N(a,\, b^2)$.

Solution. (1) We begin by observing that the posterior distribution is proportional to

$$\begin{aligned}\mathcal{L}(p)f(p) &\propto p^X(1-p)^{1-X} \cdot p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{\alpha+X-1}(1-p)^{\beta+n-X-1}\end{aligned}$$

and so the posterior is a Beta$(\alpha+X,\, \beta+n-X)$ distribution. With respect to the squared error loss the Bayes estimator is then the posterior mean. Since the mean of a Beta$(\alpha,\, \beta)$ distribution is $\alpha/(\alpha/\beta)$, the Bayes estimator is

$$\hat{p} = \frac{\alpha+X}{(\alpha+X)+(\beta+n-X)} = \frac{\alpha+X}{\alpha+\beta+n}$$

which we may write

$$\hat{p} = \underbrace{\frac{n}{\alpha+\beta+n}}_{=:\theta} \cdot \underbrace{\frac{X}{n}}_{\hat{p}_{MLE}} + \underbrace{\frac{\alpha+\beta}{\alpha+\beta+n}}_{1-\theta} \cdot \underbrace{\frac{\alpha}{\alpha+\beta}}_{p_{prior}}.$$

In other words the posterior mean is a convex combination of the MLE and the prior mean.

We now turn our attention to the Bayes risk. Since the MLE is an unbiased estimator and since $\mathbb{V}\hat{p}_{MLE} = \frac{\mathbb{V}X}{n^2} = \frac{p(1-p)}{n}$ we compute that the risk is

$$\begin{aligned}R(p,\, \hat{p}) &= \mathbb{E}_p\left[(p-\hat{p})^2\right] \\ &= \mathbb{E}_p\left(\left[\theta(p-\hat{p}_{MLE}) + (1-\theta)(p-p_{prior})\right]^2\right) \\ &= \theta^2\mathbb{E}_p\left[(p-\hat{p}_{MLE})^2\right] + 2\theta(1-\theta)(p-p_{prior})\underbrace{\mathbb{E}_p\left(p-\hat{p}_{MLE}\right)}_{=0} \\ &\quad + (1-\theta)^2(p-p_{prior})^2 \\ &= \theta^2\mathbb{V}\hat{p}_{MLE} + (1-\theta)^2(p-p_{prior})^2 \\ &= \frac{1}{n}\theta^2 p(1-p) + (1-\theta)^2(p-p_{prior})^2.\end{aligned}$$

Now recall that the variance of a Beta$(\alpha,\, \beta)$ distribution is

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

and so the Bayes risk is

$$\begin{aligned} r(f,\,\hat{p}) &= \int R(p,\,\hat{p}) f(p) dp \\ &= \frac{\theta^2}{n} \int p(1-p) f(p) dp + (1-\theta)^2 \int (p - p_{prior})^2 f(p) dp \\ &= \frac{\theta^2}{n} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int p^{\alpha} (1-p)^{\beta} dp + (1-\theta)^2 \cdot \frac{\alpha\beta}{(\alpha+\beta)^2 (\alpha+\beta+1)} \\ &= \frac{\theta^2}{n} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)} + \frac{(\alpha+\beta)^2}{(\alpha+\beta+n)^2} \cdot \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ & \frac{n}{(\alpha+\beta+n)^2} \cdot \frac{(\alpha+1)(\beta+1)}{(\alpha+\beta+2)(\alpha+\beta+1)} + \frac{\alpha\beta}{(\alpha+\beta+n)^2(\alpha+\beta+1)} \\ &= \frac{1}{(\alpha+\beta+n)^2(\alpha+\beta+1)} \left(\frac{n(\alpha+1)(\beta+1)}{\alpha+\beta+2} + \alpha\beta \right), \end{aligned}$$

noting that

$$r(f,\,\hat{p}) \sim \frac{1}{n} \text{ as } n \to \infty.$$

(2) As seen in Exercise A.11.6 we know that a Gamma(α, β) prior is conjugate with respect to a Poisson model. In particular, here $n = 1$ and so the posterior is a Gamma($\alpha + X$, $\frac{\beta}{\beta+1}$) distribution with posterior mean

$$\hat{\lambda} = \underbrace{\frac{\beta}{\beta+1}}_{=:\theta} \cdot \underbrace{X}_{\hat{\lambda}_{MLE}} + \underbrace{\frac{1}{\beta+1}}_{1-\theta} \cdot \underbrace{\alpha\beta}_{\lambda_{prior}}.$$

So once again the posterior mean is a convex combination of the MLE and the prior mean. We then compute as in item 1 that, since the MLE is an unbiased estimator, the risk is given by

$$R(\lambda,\,\hat{\lambda}) = \mathbb{E}_{\lambda} \left[(\lambda - \hat{\lambda})^2 \right] = \theta^2 \mathbb{V} \hat{\lambda}_{MLE} + (1-\theta)^2 (\lambda - \lambda_{prior})^2$$

where

$$\mathbb{V}\hat{\lambda}_{MLE} = \mathbb{V}X = \lambda.$$

Since a Gamma(α, β) distribution has mean $\alpha\beta$ and variance $\alpha\beta^2$ we deduce that the Bayes risk is

$$\begin{aligned} r(f,\,\hat{\lambda}) &= \int R(\lambda,\,\hat{\lambda}) f(\lambda) d\lambda \\ &= \theta^2 \int \lambda f(\lambda) d\lambda + (1-\theta)^2 \int (\lambda - \lambda_{prior})^2 f(\lambda) d\lambda \\ &= \frac{\beta^2}{(\beta+1)^2} \cdot \alpha\beta + \frac{1}{(\beta+1)^2} \cdot \alpha\beta^2 \\ &= \frac{\alpha\beta^2}{(\beta+1)^2} (\beta+1) \\ &= \frac{\alpha\beta^2}{\beta+1}. \end{aligned}$$

(3) As shown in Exercise A.11.1 the posterior distribution is $N(\bar{\theta},\, \tau^2)$ where

$$\bar{\theta} = \frac{\tau^2}{\sigma^2} X + \left(1 - \frac{\tau^2}{\sigma^2}\right) a \text{ and } \frac{1}{\tau^2} = \frac{1}{\sigma^2} + \frac{1}{b^2}.$$

In particular, for $\varphi := \frac{\tau^2}{\sigma^2}$, the posterior mean is precisely

$$\bar{\theta} = \varphi X + (1 - \varphi)a$$

such that, once again, the posterior mean is a convex combination of the MLE $\hat{\theta}_{MLE} = X$ and the prior mean $\theta_{prior} = a$. Since the MLE is an unbiased estimator we proceed as in item 1 and write the risk as

$$R(\theta,\, \bar{\theta}) = \mathbb{E}_\theta \left[(\theta - \bar{\theta})^2 \right] = \varphi^2 \mathbb{V} \hat{\theta}_{MLE} + (1 - \varphi)^2 (\theta - \theta_{prior})^2$$

where

$$\mathbb{V} \hat{\theta}_{MLE} = \mathbb{V} X = \sigma^2.$$

Therefore the Bayes risk is

$$\begin{aligned} r(f,\, \bar{\theta}) &= \varphi^2 \int \sigma^2 f(\theta) d\theta + (1 - \varphi)^2 \int (\theta - \theta_{prior})^2 f(\theta) d\theta \\ &= \varphi^2 \sigma^2 + (1 - \varphi)^2 b^2. \end{aligned}$$

To simplify this expression we note that

$$\tau^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{b^2}} = \frac{\sigma^2 b^2}{\sigma^2 + b^2}$$

and so

$$\varphi = \frac{\tau^2}{\sigma^2} = \frac{b^2}{\sigma^2 + b^2} \text{ and } 1 - \varphi = \frac{\sigma^2}{\sigma^2 + b^2}.$$

Therefore the Bayes risk is

$$\begin{aligned} r(f,\, \bar{\theta}) &= \left(\frac{b^2}{\sigma^2 + b^2} \right)^2 \sigma^2 + \left(\frac{\sigma^2}{\sigma^2 + b^2} \right)^2 b^2 \\ &= \frac{b^2 \sigma^2}{\sigma^2 + b^2} \\ &= \tau^2. \end{aligned}$$

Exercise A.12.2 (Admissibility and minimax estimator for a Normal model with squared error loss)**.** Let $X_1,\, \ldots\,,\, X_n \sim N(\theta,\, \sigma^2)$ where σ^2 is known and suppose we estimate θ with loss function $L(\theta,\, \hat{\theta}) = (\theta - \hat{\theta})^2 / \sigma^2$. Show that $\overline{X}$ is admissible and minimax.

Solution. Theorem 12.23 tells us that the sample mean is admissible since the loss function here is merely a fixed multiple of the squared error loss (and so any point estimator witnessing the inadmissibility of the sample mean here would also witness the inadmissibility of the sample mean for the squared error loss). Since the sample mean is an unbiased estimator of θ with variance $\frac{\sigma^2}{n}$ we compute that the risk is given by, for any $\theta \in \mathbb{R}$,

$$R(\theta,\, \overline{X}) = \mathbb{E}_\theta \left[\frac{(\theta - \overline{X})^2}{\sigma^2} \right] = \frac{1}{\sigma^2} \mathbb{V}_\theta \overline{X} = \frac{1}{n}.$$

The risk of this admissible estimator is constant and so we deduce from Theorem 12.24 that the sample mean is a minimax rule.

Exercise A.12.3 (Bayes estimator under the zero-one loss is the mode)**.** Consider a finite parameter space $\Theta = \{\theta_1, \ldots, \theta_k\}$. Prove that the posterior mode is the Bayes estimator under zero-one loss.

Solution. Note that the posterior distribution $f(\cdot \,|\, x^n)$ is a distribution over the finite set $\Theta = \{\theta_1, \ldots, \theta_k\}$, so we can really view $f(\cdot \,|\, x^n)$ as a map from Θ to $[0, 1]$ such that

$$f(\theta_i \,|\, x^n) \geqslant 0 \text{ for all } i \text{ and } \sum_{i=1}^{k} f(\theta_i \,|\, x^n) = 1.$$

Now we turn our attention to the Bayes estimator. Recall that the Bayes estimator minimizes the posterior risk pointwise. Here with the zero-one loss, which for convenience we may write as

$$L(\theta_1, \theta_2) = \mathbb{1}(\theta_1 \neq \theta_2) = 1 - \mathbb{1}(\theta_1 = \theta_2),$$

the posterior risk is

$$\begin{aligned} r(\theta_i \,|\, x^n) &= \int L(\theta, \theta_i) f(\theta \,|\, x^n) d\theta \\ &= \sum_{j=1}^{k} L(\theta_j, \theta_i) f(\theta_j \,|\, x^n) \\ &= 1 - \sum_{j=1}^{k} \mathbb{1}(\theta_j = \theta_i) f(\theta_j \,|\, x^n) \\ &= 1 - f(\theta_i \,|\, x^n). \end{aligned}$$

Minimizing the posterior risk is thus the same as maximizing the posterior distribution, which is another way of saying that the Bayes estimator is precisely the posterior mode. Indeed, the Bayes estimator $\hat{\theta}$ satisfies

$$\hat{\theta}(x^n) = \arg\min_{\theta_i} r(\theta_i \,|\, x^n) = \arg\min_{\theta_i} \left[1 - f(\theta_i \,|\, x^n)\right] = \arg\max_{\theta_i} f(\theta_i \,|\, x^n),$$

which is precisely the mode of the posterior distribution.

Exercise A.12.4 (Multiple of the sample variance minimizing risk)**.** Let $X_1, \ldots, X_n$ be a sample from a distribution with variance σ^2. Consider estimators of the form bS^2 where S^2 is the sample variance. Let the loss function for estimating σ^2 be

$$L(\sigma^2, \hat{\sigma}^2) = \frac{\hat{\sigma}^2}{\sigma^2} - 1 - \log\left(\frac{\hat{\sigma}^2}{\sigma^2}\right).$$

Find the optimal value of b that minimizes the risk for all σ^2.

Solution. Since the sample variance is an unbiased estimator of the variance we compute that the risk is

$$\begin{aligned}R(\sigma^2,\, bS^2) &= \mathbb{E}_{\sigma^2}\left[L(\sigma^2,\, bS^2)\right]\\ &= \mathbb{E}_{\sigma^2}\left[\frac{bS^2}{\sigma^2} - 1 - \log\left(\frac{bS^2}{\sigma^2}\right)\right]\\ &= \frac{b}{\sigma^2}\mathbb{E}_{\sigma^2}(S^2) - 1 - \mathbb{E}_{\sigma^2}\left(\log b + \log S^2 - \log\sigma^2\right)\\ &= b - \log b + C\end{aligned}$$

for some constant C which depends on σ^2 and S^2, i.e. which depends on the underlying distribution, but which does *not* depend on b. Since

$$(b - \log b)' = 1 - \frac{1}{b}$$

which vanishes when $b = 1$ and since

$$(b - \log b)'' = \frac{1}{b^2} > 0$$

such that $b - \log b$ is strictly convex, we deduce that $b = 1$ is the global minimizer.

Exercise A.12.5 (A trivial unique minimax rule)**.** Let $X \sim \text{Binomial}(n,\, p)$ and suppose that the loss function is

$$L(p,\, \hat{p}) = \left(1 - \frac{\hat{p}}{p}\right)^2$$

where $0 < p < 1$. Consider the estimator $\hat{p}(x) = 0$. This estimator falls outside the parameter space $(0,\, 1)$ but we will allow this. Show that $\hat{p}$ is the unique minimax rule.

Solution. First we compute the maximum risk of $\hat{p} = 0$. This is immediate since, for any $p \in (0,\, 1)$,

$$R(p,\, \hat{p}) = \mathbb{E}_p\left[\left(1 - \frac{\hat{p}}{p}\right)^2\right] = 1$$

and so

$$\overline{R}(\hat{p}) = \sup_{p\in(0,\,1)} R(p,\, \hat{p}) = 1.$$

We now show that for any point estimator $\tilde{p} : \{0,\, \ldots,\, n\} \to [0,\, 1]$, if $\tilde{p} \neq 0$, meaning that $\tilde{p}(j) > 0$ for some $1 \leqslant j \leqslant n$, then it has larger maximum risk, i.e.

$$\overline{R}(\tilde{p}) > 1.$$

To do this it suffices to show that, for such estimators $\tilde{p}$, there is some $p \in (0,\, 1)$ for which

$$R(p,\, \tilde{p}) > 1.$$

So let $\tilde{p} : \{0, \ldots, n\} \to [0, 1]$ be a non-zero point estimator. Then, for all $p \in (0, 1)$ the risk satisfies

$$\begin{aligned} R(p, \tilde{p}) &= \mathbb{E}_p \left[\left(1 - \frac{\tilde{p}}{p} \right)^2 \right] \\ &= \mathbb{E}_p \left[\left(1 - \frac{\tilde{p}}{p} \right)^2 \mathbb{1}(\tilde{p} = 0) + \left(1 - \frac{\tilde{p}}{p} \right)^2 \mathbb{1}(\tilde{p} > 0) \right] \\ &= \mathbb{E}_p \left[\mathbb{1}(\tilde{p} = 0) + \left(1 - \frac{\tilde{p}}{p} \right)^2 \mathbb{1}(\tilde{p} > 0) \right] \end{aligned}$$

where

$$\mathbb{1}(\tilde{p} = 0) = 1 - \mathbb{1}(\tilde{p} > 0)$$

and so

$$\begin{aligned} R(p, \tilde{p}) &= \mathbb{E}_p \left[1 - \mathbb{1}(\tilde{p} > 0) + \left(1 - \frac{\tilde{p}}{p} \right)^2 \mathbb{1}(\tilde{p} > 0) \right] \\ &= 1 + \mathbb{E}_p \left(\left[\left(1 - \frac{\tilde{p}}{p} \right)^2 - 1 \right] \mathbb{1}(\tilde{p} > 0) \right). \end{aligned}$$

Crucially, since $\tilde{p} \neq 0$ we know that

$$p_* := \min_{i : \tilde{p}(i) > 0} \tilde{p}(i) > 0.$$

Therefore, for all $p \in (0, p_*)$,

$$\begin{aligned} \mathbb{E}_p \left(\left[\left(1 - \frac{\tilde{p}}{p} \right)^2 - 1 \right] \mathbb{1}(\tilde{p} > 0) \right) &\geqslant \mathbb{E}_p \left(\left[\left(1 - \frac{p_*}{p} \right)^2 - 1 \right] \mathbb{1}(\tilde{p} > 0) \right) \\ &= \left[\left(1 - \frac{p_*}{p} \right)^2 - 1 \right] \mathbb{P}(\tilde{p} > 0) =: c(p). \end{aligned}$$

In other words, for all $p \in (0, p_*)$, we know that

$$R(p, \hat{p}) = 1 + c(p).$$

Crucially:

$$\begin{aligned} c(p) > 0 &\Leftrightarrow \left(1 - \frac{p_*}{p} \right)^2 - 1 > 0 \\ &\Leftrightarrow \left(\frac{p_*}{p} \right)^2 - 2 \cdot \frac{p_*}{p} > 0 \\ &\Leftrightarrow \frac{p_*}{p} - 2 > 0 \\ &\Leftrightarrow p < \frac{p_*}{2}, \end{aligned}$$

which means that, for any $p \in \left(0, \frac{p_*}{2}\right)$, $c(p) > 0$ and thus the risk satisfies

$$R(p, \tilde{p}) = 1 + c(p) > 1,$$

and so the maximum risk satisfies

$$\overline{R}(\tilde{p}) > 1$$

as desired.

A.13. **Linear and Logistic Regression.**

Exercise A.13.1 (Least squares estimates)**.** Prove Theorem 13.15 where we record the formulae for the least squares estimates and an unbiased estimator of σ^2 (under the standard noise assumptions) for the simple linear regression model.

Solution. We begin by verifying the formulae for the least squares estimates. First we compute that

$$\partial_{\beta_0} RSS = \sum_{i=1}^{n} (-2) \cdot (Y_i - \beta_0 - \beta_1 X_i) = -2n(\overline{Y} - \beta_0 - \beta_1 \overline{X}),$$

where $\overline{X}$ and $\overline{Y}$ denote the sample means, while

$$\partial_{\beta_1} RSS = \sum_{i=1}^{n} (-2X_i) \cdot (Y_i - \beta_0 - \beta_1 X_i) = -2X \cdot (Y - \beta_0 \mathbb{1} - \beta_1 X)$$

where $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$, and $\mathbb{1} = (1, \ldots, 1)$ live in $\mathbb{R}^n$. Since the residual sum of squares is a strictly convex function of (β_0, β_1), we deduce that the unique global minimum occurs when the gradient of the RSS vanishes, i.e.

$$\begin{cases} \overline{Y} - \beta_0 - \beta_1 \overline{X} = 0 \text{ and} \\ X \cdot (Y - \beta_0 \mathbb{1} - \beta_1 X) = 0. \end{cases}$$

To make this easier to work with we define

$$\widetilde{X} := X - \overline{X}\mathbb{1} \text{ and } \widetilde{Y} := Y - \overline{Y}\mathbb{1},$$

meaning that

$$\widetilde{X}_i = X_i - \overline{X} \text{ and } \widetilde{Y}_i = Y_i - \overline{Y}.$$

Crucially: $\widetilde{X}$ and $\widetilde{Y}$ have average zero and so they are orthogonal to the constant vector $\mathbb{1}$ (and its multiples). Since $\overline{Y} - \beta_0 - \beta_1 \overline{X} = 0$ we may thus use this orthogonality to rewrite the second equation as

$$\begin{aligned} 0 &= X \cdot (Y - \beta_0 - \beta_1 X) \\ &= X \cdot (\widetilde{Y} - \beta_1 \widetilde{X}) + X \cdot [\underbrace{(\overline{Y} - \beta_0 - \beta_1 \overline{X})}_{=0} \mathbb{1}] \\ &= \widetilde{X} \cdot (\widetilde{Y} - \beta_1 \widetilde{X}) \\ &= \widetilde{X} \cdot \widetilde{Y} - \beta_1 |\widetilde{X}|^2. \end{aligned}$$

In other words

$$\hat{\beta}_1 = \frac{\widetilde{X} \cdot \widetilde{Y}}{|\widetilde{X}|^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

while

$$\overline{Y} - \hat{\beta}_0 - \hat{\beta}_1 \overline{X} = 0 \iff \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

We now turn our attention to the unbiased estimator for σ^2. We begin by computing that

$$\mathbb{E}\left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2\right) = \mathbb{E}\left(|\hat{\varepsilon}|^2\right) = \mathbb{E}\left(|Y - \hat{\beta}_0 \mathbb{1} - \hat{\beta}_1 X|^2\right).$$

By orthogonality in $\mathbb{R}^n$ and the definitions of $\hat{\beta}_0$ and $\hat{\beta}_1$ we have that

$$|Y - \hat{\beta}_0 \mathbb{1} - \hat{\beta}_1 X|^2 = |\widetilde{Y} - \beta_1 \widetilde{X}|^2 + |\underbrace{(\overline{Y} - \hat{\beta}_0 - \beta_1 \overline{X})}_{=0} \mathbb{1}|^2 = \left| \widetilde{Y} - \frac{\widetilde{Y} \cdot \widetilde{X}}{|\widetilde{X}|^2} \widetilde{X} \right|^2 .$$

Crucially we may write the term $\widetilde{Y} - \frac{\widetilde{Y} \cdot \widetilde{X}}{|\widetilde{X}|^2} \widetilde{X}$ as the result of two consecutive orthogonal projections, namely

$$\begin{aligned}\widetilde{Y} - \frac{\widetilde{Y} \cdot \widetilde{X}}{|\widetilde{X}|^2} \widetilde{X} &= \left(I - \frac{\widetilde{X} \otimes \widetilde{X}}{|\widetilde{X}|^2} \right) \widetilde{Y} \\ &= \left(I - \frac{\widetilde{X} \otimes \widetilde{X}}{|\widetilde{X}|^2} \right) (Y - \overline{Y} \mathbb{1}) \\ &= \left(I - \frac{\widetilde{X} \otimes \widetilde{X}}{|\widetilde{X}|^2} \right) \left(I - \frac{\mathbb{1} \otimes \mathbb{1}}{|\mathbb{1}|^2} \right) Y\end{aligned}$$

since $|\mathbb{1}|^2 = \sqrt{\sum_{i=1}^n 1} = n$. Another crucial observation is that $\widetilde{X}$ and $\mathbb{1}$ are, by definition of $\widetilde{X}$, orthogonal. We may thus lump these two projections together and write

$$\left(I - \frac{\widetilde{X} \otimes \widetilde{X}}{|\widetilde{X}|^2} \right) \left(I - \frac{\mathbb{1} \otimes \mathbb{1}}{|\mathbb{1}|^2} \right) = I - \frac{\widetilde{X} \otimes \widetilde{X}}{|\widetilde{X}|^2} - \frac{\mathbb{1} \otimes \mathbb{1}}{|\mathbb{1}|^2} =: A(\widetilde{X}).$$

Note that writing it this way really only serves to make it clear that $A(\widetilde{X})$ is an orthogonal projection with a two-dimensional kernel. Using the rule of iterated expectation to condition on X we deduce that

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^n \hat{\varepsilon}_i^2 \right) &= \mathbb{E}\left(\left| \widetilde{Y} - \frac{\widetilde{Y} \cdot \widetilde{X}}{|\widetilde{X}|^2} \widetilde{X} \right|^2 \right) \\ &= \mathbb{E}\left(A(\widetilde{X}) Y \cdot Y \right) \\ &= \mathbb{E}\left[\mathbb{E}\left(A(\widetilde{X}) Y \cdot Y \,\middle|\, X \right) \right]\end{aligned}$$

where, using Exercise A.23.6,

$$\mathbb{E}\left(A(\widetilde{X}) Y \cdot Y \,\middle|\, X \right) = A(\widetilde{X}) : \mathrm{Cov}(Y, Y \,|\, X) + A(\widetilde{X}) \mathbb{E}(Y \,|\, X) \cdot \mathbb{E}(Y \,|\, X).$$

In light of the standard noise assumptions for the simple linear regression model we note that

$$\mathbb{E}\,(Y \,|\, X) = \mathbb{E}\,(\beta_0 \mathbb{1} + \beta_1 X + \varepsilon \,|\, X) = \beta_0 \mathbb{1} + \beta_1 X + 0$$

while

$$\begin{aligned}\mathrm{Cov}\,(Y_i, Y_j \,|\, X) &= \mathrm{Cov}\,(\beta_0 + \beta_1 X_i + \varepsilon_i, \beta_0 + \beta_1 X_j + \varepsilon_j \,|\, X) \\ &= \mathrm{Cov}\,(\varepsilon_i, \varepsilon_j \,|\, X) \\ &= \delta_{ij} \mathbb{V}\,(\varepsilon \,|\, X) = \sigma^2 \delta_{ij}.\end{aligned}$$

We may decompose

$$\mathbb{E}\,(Y \,|\, X) = \left(\beta_0 + \beta_1 \overline{X} \right) \mathbb{1} + \beta_1 \widetilde{X}$$

and so, since $A(\widetilde{X})$ annihilates both $\mathbb{1}$ and $\widetilde{X}$, it must also anihilate $\mathbb{E}\,(Y \,|\, X)$. Therefore

$$\mathbb{E}\left(A(\widetilde{X})Y \cdot Y \,\middle|\, X\right) = A(\widetilde{X}) : \sigma^2 I + 0 = \sigma^2 \operatorname{tr} A(\widetilde{X}).$$

Since $A(\widetilde{X})$ is an n-by-n orthogonal projection matrix with a two-dimensional kernel we know that

$$\operatorname{tr} A(\widetilde{X}) = \operatorname{rank} A(\widetilde{X}) = n - 2.$$

So finally we conclude that

$$\mathbb{E}\left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2\right) = \mathbb{E}\left[\mathbb{E}\left(A(\widetilde{X})Y \cdot Y \,\middle|\, X\right)\right] = \sigma^2 \mathbb{E}\left[\operatorname{tr} A(\widetilde{X})\right] = \sigma^2(n-2),$$

as desired.

Exercise A.13.2 (Conditional means and variances of the least squares estimates)**.** Prove Theorem 13.22 where we record the formulae for the conditional means and variances of the least squares estimates given the feature data.

Solution. Recall from Theorem 13.15 that

$$\hat{\beta}_0 = \overline{Y}_n - \hat{\beta}_1 \overline{X}_n \text{ and } \hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}.$$

First we compute the conditional means. The key observation comes from Theorem 13.6: under the standard noise assumptions we know that

$$\mathbb{E}(Y \,|\, X) = \beta_0 + \beta_1 X.$$

In particular, here, this means that

$$\mathbb{E}(Y_i \,|\, X^n) = \beta_0 + \beta_1 X_i$$

for all $1 \leqslant i \leqslant n$. Therefore

$$\mathbb{E}(\overline{Y}_n \,|\, X^n) = \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 X_i) = \beta_0 + \beta_1 \overline{X}_n$$

and so

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1 \,|\, X^n) &= \frac{1}{\sum_{i=1}^n (X_i - \overline{X}_n)^2} \sum_{i=1}^n (X_i - \overline{X}_n)\mathbb{E}(Y_i - \overline{Y}_n \,|\, X^n) \\
&= \frac{1}{n s_X^2} \sum_{i=1}^n (X_i - \overline{X}_n) \cdot \beta_1 (X_i - \overline{X}_n) \\
&= \frac{\beta_1 n s_X^2}{n s_X^2} \\
&= \beta_1.
\end{aligned}$$

So finally

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_0 \,|\, X^n) &= \mathbb{E}(\overline{Y}_n \,|\, X^n) - \overline{X}_n \cdot \mathbb{E}(\hat{\beta}_1 \,|\, X^n) \\
&= (\beta_0 + \beta_1 \overline{X}_n) - \beta_1 \overline{X}_n \\
&= \beta_0.
\end{aligned}$$

We now turn our attention to the conditional variance-covariance matrix. First we compute the conditional variance of $\hat{\beta}_1$, namely

$$\mathbb{V}(\hat{\beta}_1 \,|\, X^n) = \mathbb{E}\left[(\hat{\beta}_1 - \beta_1)^2 \,\middle|\, X^n\right].$$

We may expand

$$\begin{aligned}
\hat{\beta}_1 - \hat{\beta} &= \frac{1}{ns_X^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n) - \beta_1 \\
&= \frac{1}{ns_X^2}\left[\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n) - \beta_1 \sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right] \\
&= \frac{1}{ns_X^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n - \beta_1 X_i + \beta_1 \overline{X}_n) \\
&= \frac{1}{ns_X^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \beta_0 - \beta_1 X_i - \overline{Y}_n + \beta_0 + \beta_1 \overline{X}_n) \\
&= \frac{1}{ns_X^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)(\varepsilon_i - \overline{\varepsilon}_n)
\end{aligned}$$

where $\overline{\varepsilon}_n := \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$. In particular, since generally speaking

$$\mathbb{V}\left(\sum_i \alpha_i Z_i\right) = \sum_{i,j}\alpha_i\alpha_j \operatorname{Cov}(Z_i,\, Z_j)$$

we deduce that, here,

$$\mathbb{V}(\hat{\beta}_1 \,|\, X^n) = \frac{1}{(ns_X^2)^2}\sum_{i,j=1}^{n}(X_i - \overline{X}_n)(X_j - \overline{X}_n)\operatorname{Cov}(\varepsilon_i - \overline{\varepsilon}_n,\, \varepsilon_j - \overline{\varepsilon}_n \,|\, X^n).$$

Since ε satisfies the standard noise assumptions and since $\varepsilon_1, \ldots, \varepsilon_n$ are IID we compute that

$$\operatorname{Cov}(\varepsilon_i,\, \varepsilon_j \,|\, X^n) = \sigma^2\delta_{ij}$$

while

$$\operatorname{Cov}(\varepsilon_i,\, \overline{\varepsilon}_n \,|\, X^n) = \frac{1}{n}\sum_{j=1}^{n}\operatorname{Cov}(\varepsilon_i,\, \varepsilon_j \,|\, X^n) = \frac{1}{n}\sum_{j=1}^{n}\sigma^2\delta_{ij} = \frac{\sigma^2}{n}$$

and

$$\operatorname{Cov}(\overline{\varepsilon}_n,\, \overline{\varepsilon}_n \,|\, X^n) = \mathbb{V}(\overline{\varepsilon}_n \,|\, X^n) = \frac{\mathbb{V}(\varepsilon \,|\, X)}{n} = \frac{\sigma^2}{n}.$$

Therefore

$$\begin{aligned}
&\operatorname{Cov}(\varepsilon_i - \overline{\varepsilon}_n,\, \varepsilon_j - \overline{\varepsilon}_n \,|\, X^n) \\
&= \operatorname{Cov}(\varepsilon_i,\, \varepsilon_j) - \operatorname{Cov}(\varepsilon_i,\, \overline{\varepsilon}_n) - \operatorname{Cov}(\varepsilon_j,\, \overline{\varepsilon}_n) + \operatorname{Cov}(\overline{\varepsilon}_n,\, \overline{\varepsilon}_n) \\
&= \sigma^2\delta_{ij} - \frac{\sigma^2}{n}
\end{aligned}$$

and so

$$\begin{aligned}\mathbb{V}(\hat{\beta}_1 \,|\, X^n) &= \frac{1}{\left(ns_X^2\right)^2} \sum_{i,j=1}^{n} (X_i - \overline{X}_n)(X_j - \overline{X}_n) \cdot \sigma^2 \left(\delta_{ij} - \frac{1}{n}\right) \\ &= \frac{\sigma^2}{\left(ns_X^2\right)^2} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2 - \frac{\sigma^2}{n\left(ns_X^2\right)^2} \sum_{i=1}^{n} (X_i - \overline{X}_n) \underbrace{\left[\sum_{j=1}^{n} (X_j - \overline{X}_n)\right]}_{=0} \\ &= \frac{\sigma^2}{ns_X^2},\end{aligned}$$

as desired.

We now turn our attention to the conditional covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$. Since

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}_0,\, \hat{\beta}_1 \,|\, X^n) &= \mathrm{Cov}(\overline{Y}_n - \hat{\beta}_1 \overline{X}_n,\, \hat{\beta}_1 \,|\, X^n) \\ &= \mathrm{Cov}(\overline{Y}_n,\, \hat{\beta}_1 \,|\, X^n) - \overline{X}_n \mathbb{V}(\hat{\beta}_1,\, X^n)\end{aligned}$$

where, as computed above, $\mathbb{V}(\hat{\beta}_1 \,|\, X^n) = \frac{\sigma^2}{ns_X^2}$, it suffices for us to show that

$$\mathrm{Cov}(\overline{Y}_n,\, \hat{\beta}_1 \,|\, X^n) = 0.$$

We compute that

$$\begin{aligned}\mathrm{Cov}(\overline{Y}_n,\, \hat{\beta}_1 \,|\, X^n) &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{Cov}(Y_i,\, \hat{\beta}_1 \,|\, X^n) \\ &= \frac{1}{n^2 s_X^2} \sum_{i,j=1}^{n} (X_j - \overline{X}_n) \,\mathrm{Cov}(Y_i,\, Y_j - \overline{Y}_n \,|\, X^n)\end{aligned}$$

where

$$\begin{aligned}\mathrm{Cov}(Y_i,\, Y_j \,|\, X^n) &= \mathrm{Cov}(\beta_0 + \beta_1 X_i + \varepsilon_i,\, \beta_0 + \beta_1 X_j + \varepsilon_j \,|\, X^n) \\ &= \mathrm{Cov}(\varepsilon_i,\, \varepsilon_j \,|\, X^n) \\ &= \sigma^2 \delta_{ij}\end{aligned}$$

and so

$$\mathrm{Cov}(Y_i,\, \overline{Y}_n \,|\, X^n) = \frac{1}{n} \sum_{j=1}^{n} \mathrm{Cov}(Y_i,\, Y_j \,|\, X^n) = \frac{1}{n} \sum_{j=1}^{n} \sigma^2 \delta_{ij} = \frac{\sigma^2}{n}.$$

Therefore

$$\begin{aligned}\mathrm{Cov}(\overline{Y}_n,\, \hat{\beta}_1 \,|\, X^n) &= \frac{1}{n^2 s_X^2} \sum_{i,j=1}^{n} (X_j - \overline{X}_n) \cdot \sigma^2 \left(\delta_{ij} - \frac{1}{n}\right) \\ &= \frac{\sigma^2}{n^2 s_X^2} \sum_{j=1}^{n} (X_j - \overline{X}_n) \underbrace{\left[\sum_{i=1}^{n} \left(\delta_{ij} - \frac{1}{n}\right)\right]}_{=1-\frac{n}{n}=0},\end{aligned}$$

as desired.

Finally we compute the conditional variance of $\hat{\beta}_0$. Using the computations carried out above we see that

$$\begin{aligned}\mathbb{V}(\hat{\beta}_0 \,|\, X^n) &= \mathbb{V}(\overline{Y}_n - \hat{\beta}_1 \overline{X}_n \,|\, X^n) \\ &= \mathbb{V}(\overline{Y}_n \,|\, X^n) - 2\overline{X}_n \underbrace{\mathrm{Cov}(\overline{Y}_n,\, \hat{\beta}_1 \,|\, X^n}_{=0} + \overline{X}_n^2 \mathbb{V}(\hat{\beta}_1 \,|\, X^n) \\ &= \mathbb{V}(\overline{Y}_n \,|\, X^n) + \frac{\sigma^2 \overline{X}_n^2}{n s_X^2}\end{aligned}$$

where

$$\mathbb{V}(\overline{Y}_n \,|\, X^n) = \mathbb{V}(\beta_0 + \beta_1 \overline{X}_n + \overline{\varepsilon}_n \,|\, X^n) = \mathbb{V}(\overline{\varepsilon}_n \,|\, X^n) = \frac{\sigma^2}{n}.$$

Therefore

$$\mathbb{V}(\hat{\beta}_0 \,|\, X^n) = \frac{\sigma^2}{n} \left(1 + \frac{\overline{X}_n^2}{s_X^2} \right).$$

We note that

$$\begin{aligned}s_X^2 &= \frac{1}{n} \sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i^2 - 2X_i \overline{X}_n + \overline{X}_n^2) \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - 2\overline{X}_n^2 + \overline{X}_n^2 \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \overline{X}_n^2\end{aligned}$$

and so we conclude that

$$\mathbb{V}(\hat{\beta}_0 \,|\, X^n) = \frac{\sigma^2}{n s_X^2} (s_X^2 + \overline{X}_n^2) = \frac{\sigma^2}{n s_X^2} \cdot \frac{1}{n} \sum_{i=1}^n X_i^2$$

as desired.

Exercise A.13.3 (Regression through the origin)**.** Consider the *regression through the origin* model

$$Y_i = \beta X_i + \varepsilon_i.$$

Find the least squares estimates for β. Find the standard error of the estimate. Find conditions that guarantee that the estimate is consistent.

Solution. Given an IID sample $(Y_1,\, X_1),\, \ldots,\, (Y_n,\, X_n)$ drawn from a distrbituion in the regression through the origin model and given a point estimator $\hat{\beta}$ the residuals are now

$$\hat{\varepsilon}_i := Y_i - \hat{\beta} X_i$$

and so the residual sum of squares is

$$RSS := \sum_{i=1}^n \left(Y_i - \hat{\beta} X_i\right)^2.$$

We compute that

$$\partial_{\hat{\beta}} RSS = \sum_{i=1}^{n} -2X_i \cdot (Y_i - \hat{\beta} X_i) = -2X \cdot (Y - \hat{\beta} X)$$

and so the least squares estimate $\hat{\beta}$ must satisfy

$$X^n \cdot (Y^n - \hat{\beta} X^n) = 0 \iff \hat{\beta} = \frac{X^n \cdot Y^n}{|X^n|^2} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$

We now compute the conditional mean and variance of the least squares estimate given the feature data $X^n = (X_1, \ldots, X_n)$. Mimicking the standard noise assumption for the simple linear regression model we *assume* that, for β the true parameter, $\varepsilon := Y - \beta X$ satisfies

$$\mathbb{E}(\varepsilon \,|\, X) = 0 \text{ and } \mathbb{V}(\varepsilon \,|\, X) = \sigma^2.$$

Then we may compute that

$$\mathbb{E}(Y_i \,|\, X^n) = \mathbb{E}(\beta X_i - \varepsilon_i \,|\, X^n) = \beta X_i + \underbrace{\mathbb{E}(\varepsilon_i \,|\, X^n)}_{=0}$$

and so the conditional mean is

$$\mathbb{E}(\hat{\beta} \,|\, X^n) = \frac{\sum_{i=1}^n X_i \mathbb{E}(Y_i \,|\, X^n)}{\sum_{i=1}^n X_i^2} = \frac{\sum_{i=1}^n \beta X_i^2}{\sum_{i=1}^n X_i^2} = \beta.$$

Therefore the conditional variance is

$$\mathbb{V}(\hat{\beta} \,|\, X^n) = \mathbb{E}\left[(\hat{\beta} - \beta)^2 \,\middle|\, X^n \right]$$

where, proceeding as in Exercise A.13.2,

$$\hat{\beta} - \beta = \frac{1}{\sum_{i=1}^n X_i^2} \sum_{i=1}^n X_i \underbrace{(Y_i - \beta X_i)}_{=\varepsilon_i}$$

and so

$$\mathbb{V}(\hat{\beta} \,|\, X^n) = \frac{1}{\left(\sum_{i=1}^n X_i^2\right)^2} \sum_{i,j=1}^n X_i X_j \underbrace{\operatorname{Cov}(\varepsilon_i,\, \varepsilon_j \,|\, X^n)}_{=\sigma^2 \delta_{ij}} = \frac{\sum_{i=1}^n \sigma^2 X_i^2}{\left(\sum_{i=1}^n X_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n X_i^2}.$$

Consequently the conditional standard error of $\hat{\beta}$ is

$$\widehat{se}(\hat{\beta} \,|\, X^n) = \sqrt{\mathbb{V}(\hat{\beta} \,|\, X^n)} = \frac{\sigma}{\sqrt{\sum_{i=1}^n X_i^2}}.$$

Finally we wish to show that, under the assumptions above, namely

$$\mathbb{E}(\varepsilon \,|\, X) = 0 \text{ and } \mathbb{V}(\varepsilon \,|\, X) = \sigma^2,$$

$\hat{\beta}$ is consistent. Well we compute that

$$\hat{\beta} = \frac{X^n \cdot Y^n}{|X^n|^2} = \frac{X^n \cdot (\beta X^n + \varepsilon^n)}{|X^n|^2} = \beta + \frac{X^n \cdot \varepsilon^n}{|X^n|^2}$$

where the Weak Law of Large Numbers tells us that

$$\frac{X^n \cdot \varepsilon^n}{|X^n|^2} = \frac{\frac{1}{n}\sum_{i=1}^n X_i \varepsilon_i}{\frac{1}{n}\sum_{i=1}^n X_i^2} \xrightarrow{P} \frac{\mathbb{E}(X\varepsilon)}{\mathbb{E}(X^2)} \text{ as } n \to \infty.$$

Crucially we deduce from our noise assumptions and the rule of iterated expectation that

$$\mathbb{E}(X\varepsilon) = \mathbb{E}[\mathbb{E}(X\varepsilon \,|\, X)] = \mathbb{E}[X \underbrace{\mathbb{E}(\varepsilon \,|\, X)}_{=0}] = 0.$$

Therefore $\frac{X^n \cdot \varepsilon^n}{|X^n|^2} \xrightarrow{P} 0$ as $n \to \infty$ and so

$$\hat{\beta} \xrightarrow{P} \beta \text{ as } n \to \infty$$

as desired.

Note that the same reasoning tells us that, for a sample $(Y_1, X_1), \dots, (Y_n, X_n)$ drawn from *any* distribution, not necessarily one in the regression through the origin model,

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i}{\frac{1}{n}\sum_{i=1}^n X_i^2} \xrightarrow{P} \frac{\mathbb{E}(XY)}{\mathbb{E}(X^2)}$$

by the Weak Law of Large Numbers. However, in order to establish *consistency* there must be a *true* parameter to converge to. In that case, meaning if $Y = \beta X + \varepsilon$ for some ε satisfying our assumptions, we compute that

$$\frac{\mathbb{E}(XY)}{\mathbb{E}(X^2)} = \frac{\mathbb{E}(X(\beta X + \varepsilon))}{\mathbb{E}(X^2)} = \frac{\beta\mathbb{E}(X^2) + \mathbb{E}(X\varepsilon)}{\mathbb{E}(X^2)} = \beta$$

as desired since $\mathbb{E}(X\varepsilon) = 0$. Put simply: $\hat{\beta}$ *always* converges (as long as the relevant expectations are finite) but its limit may not have anything to do with the regression through the origin model.

Exercise A.13.4 (Bias of the training error)**.** Prove the identity recorded in Theorem 13.54, i.e. prove that

$$\text{bias}[\hat{R}_{tr}(S)] = -2\sum_{i=1}^n \text{Cov}(\hat{Y}_i, Y_i).$$

Solution. We compute that, as an estimator of the prediction risk $R(S)$,

$$\text{bias}[\hat{R}_{tr}(S)] = \mathbb{E}[\hat{R}_{tr}(S) - R(S)] = \sum_{i=1}^n \mathbb{E}[(\hat{Y}_i - Y_i)^2] - \sum_{i=1}^n \mathbb{E}[(\hat{Y}_i - Y_i^*)^2].$$

In particular, for any i, some simple algrebra reveals that

$$\begin{aligned}(\hat{Y}_i - Y_i)^2 - (\hat{Y}_i - Y_i^*)^2 &= (\hat{Y}_i^2 - 2\hat{Y}_i Y_i + Y_i^2) - (\hat{Y}_i^2 - 2\hat{Y}_i Y_i^* + (Y_i^*)^2) \\ &= -2\hat{Y}_i Y_i + Y_i^2 + 2\hat{Y}_i Y_i^* - (Y_i^*)^2 \\ &= 2\hat{Y}_i(Y_i^* - Y_i) + Y_i^2 - (Y_i^*)^2 \\ &= 2\hat{Y}_i(Y_i^* - Y_i) + (Y_i - Y_i^*)(Y_i + Y_i^*) \\ &= (Y_i^* - Y_i)(2\hat{Y}_i - Y_i - Y_i^*)\end{aligned}$$

and so

$$\text{bias}[\hat{R}_{tr}(S)] = \sum_{i=1}^n \mathbb{E}[(Y_i^* - Y_i)(2\hat{Y}_i - Y_i - Y_i^*)].$$

We may therefore write, by the rule of iterated expectation,

$$\text{bias}[\hat{R}_{tr}(S)] = \sum_{i=1}^n \mathbb{E}\left(\mathbb{E}[(Y_i^* - Y_i)(2\hat{Y}_i - Y_i - Y_i^*) \,|\, \mathbb{X}]\right).$$

Crucially, we may verify that, conditioned on the feature data summarized in $\mathbb{X}$, Y_i^* and Y_i have the same mean. Indeed we always have that, since $Y_i^* \sim Y \mid X = X_i$,

$$\mathbb{E}Y_i^* = \mathbb{E}(Y \mid X = X_i) = \mathbb{E}(\beta \cdot X + \varepsilon \mid X = X_i) = \beta \cdot X_i + 0$$

while, conditioning on $\mathbb{X}$,

$$\mathbb{E}(Y_i \mid \mathbb{X}) = \mathbb{E}(\beta \cdot X_i + \varepsilon_i \mid \mathbb{X}) = \beta \cdot X_i + 0.$$

So let us write $Z_i := 2\hat{Y}_i - Y_i - Y_i^*$ and $\mu_i := \mathbb{E}(Z_i \mid \mathbb{X})$. Since $\mathbb{E}(Y_i^* - Y_i \mid \mathbb{X}) = 0$ the orthogonality of constants and random variables with mean zero tells us that

$$\mathbb{E}[(Y_i^* - Y_i)Z_i \mid \mathbb{X}] = \mathbb{E}[(Y_i^* - Y_i)(Z_i - \mu_i) \mid \mathbb{X}] = \mathrm{Cov}(Y_i^* - Y_i,\, Z_i \mid \mathbb{X}).$$

Another crucial observation at this stage is that Y_i^* is independent from both Y_i and $\hat{Y}_i$. Therefore

$$\begin{aligned}\mathrm{Cov}(Y_i^* - Y_i,\, Z_i \mid \mathbb{X}) &= \mathrm{Cov}(Y_i^* - Y_i,\, 2\hat{Y}_i - Y_i - Y_i^* \mid \mathbb{X}) \\ &= -\mathbb{V}(Y_i^* \mid \mathbb{X}) + \mathbb{V}(Y_i \mid \mathbb{X}) - 2\,\mathrm{Cov}(Y_i,\, \hat{Y}_i \mid \mathbb{X}).\end{aligned}$$

The first two conditional variances are equal since, on the one hand, we know that $Y_i^* \sim Y \mid X = X_i$ and so

$$\begin{aligned}\mathbb{V}(Y_i^* \mid \mathbb{X}) = \mathbb{V}(Y_i^*) &= \mathbb{V}(Y \mid X = X_i) \\ &= \mathbb{V}(\beta \cdot X + \varepsilon \mid X = X_i) = \mathbb{V}(\varepsilon \mid X = X_i) = \sigma^2\end{aligned}$$

while, on the other hand,

$$\mathbb{V}(Y_i \mid \mathbb{X}) = \mathbb{V}(\beta \cdot X_i + \varepsilon_i \mid \mathbb{X}) = \mathbb{V}(\varepsilon_i \mid \mathbb{X}) = \sigma^2.$$

Therefore

$$\mathrm{Cov}(Y_i^* - Y_i,\, Z_i \mid \mathbb{X}) = -2\,\mathrm{Cov}(Y_i,\, \hat{Y}_i \mid \mathbb{X}).$$

So finally we may put it all together and use the rule of iterated expectation one last time to conclude that

$$\begin{aligned}\mathrm{bias}[\hat{R}_{tr}(S)] &= \sum_{i=1}^n \mathbb{E}\left(\mathbb{E}[(Y_i^* - Y_i)Z_i \mid \mathbb{X}]\right) \\ &= \sum_{i=1}^n \mathbb{E}[\mathrm{Cov}(Y_i^* - Y_i,\, Z_i \mid \mathbb{X})] \\ &= -2\sum_{i=1}^n \mathbb{E}[\mathrm{Cov}(Y_i,\, \hat{Y}_i \mid \mathbb{X})] \\ &= -2\sum_{i=1}^n \mathrm{Cov}(Y_i,\, \hat{Y}_i).\end{aligned}$$

By symmetry of the covariance this is precisely the desired expression.

Exercise A.13.5 (Wald test for a linear combination of the simple linear regression parameters)**.** In the simple linear regression model, construct a Wald test for

$$H_0 : \beta_1 = 17\beta_0 \text{ versus } H_1 : \beta_1 \neq 17\beta_0.$$

Solution. Here we assume that the standard noise assumptions for the simple linear regression model hold in order for Theorem 13.22 to be applicable. We need to find the asymptotic variance ν^2 for which

$$\sqrt{n}\left[(\hat{\beta}_1 - 17\hat{\beta}_0) - (\beta_1 - 17\beta_0)\right] \rightsquigarrow N(0,\, \nu^2).$$

We may do this using the multivariate delta method for the Central Limit Theorem since Theorem 13.22 tells us that, for $\beta = (\beta_0,\, \beta_1)$,

$$\sqrt{n}(\hat{\beta}_n - \beta) \rightsquigarrow N(0,\, \Sigma)$$

for

$$\Sigma = \frac{\sigma^2}{\mathbb{V}(X)} \begin{pmatrix} \mathbb{E}(X^2) & -\mathbb{E}(X) \\ -\mathbb{E}(X) & 1 \end{pmatrix}.$$

Therefore, writing $\hat{\beta}_1 - 17\hat{\beta}_0 = g(\hat{\beta})$ for $g(\beta_0,\, \beta_1) := -17\beta_0 + \beta_1$ where

$$\nabla g = \begin{pmatrix} -17 \\ 1 \end{pmatrix}$$

we compute that

$$\nu^2 = (\nabla g)^T \Sigma\, \nabla g = \frac{\sigma^2}{\mathbb{V}(X)} \left[17^2 \mathbb{E}(X^2) + 2 \cdot (-17) \cdot (-\mathbb{E}X) + 1\right].$$

So for the estimate

$$\hat{\nu}_n^2 = \frac{\sigma^2}{s_X} \left(17^2 \cdot \frac{1}{n} \sum_{i=1}^n X_i^2 + 2 \cdot 17 \cdot \overline{X}_n + 1 \right)$$

we deduce that

$$W_n := \frac{\sqrt{n}\left[g(\hat{\beta}_n) - g(\beta)\right]}{\hat{\nu}_n} \rightsquigarrow N(0,\, 1)$$

and so W_n is a valid Wald test statistic. We then rejet the null $\beta_1 = 17\beta_0$ at size $\alpha \in (0,\, 1)$ when

$$|W_n| > \Phi(1 - \alpha/2)$$

for Φ denoting the CDF of a standard Normal.

Exercise A.13.6 (Comparing the AIC and Mallow's C_p statistic)**.** Assume a linear regression model with Normal errors. Take σ known. Show that the model with highest AIC is the model with the lowest Mallows' C_p statistic.

Solution. We will show more precisely that, under the Normal noise assumptions, the AIC value and Mallow's C_p statistic are related via

$$AIC(S) = -\frac{1}{2\sigma^2} \hat{R}(S) + C$$

for some constant C independent of S. On the one hand the training error is

$$\hat{R}_{tr}(S) = \sum_{i=1}^n \left(\hat{Y}_i - Y_i\right)^2 = \sum_{i=1}^n \left(\hat{\beta}_S \cdot X_i - Y_i\right)^2$$

and so Mallows' C_p statistic is

$$\hat{R}(S) = \hat{R}_{tr}(S) + 2|S|\sigma^2 = \sum_{i=1}^n \left(\hat{\beta} \cdot X_i - Y_i\right)^2 + 2|S|\sigma^2.$$

On the other hand, up to additive terms independent of β (and hence independent of S) which are irrelevant when maximizing over $\hat{\beta}_S$ (or equivalently over S) and which we denote as constants, the log-likelihood is

$$l = \sum_{i=1}^n \log f(Y_i,\, X_i; \beta,\, \sigma^2) = \sum_{i=1}^n \log f(Y_i \,|\, X_i; \beta,\, \sigma^2) + C$$

where, by the Normal noise assumption,

$$Y_i \mid X_i \sim N(\beta \cdot X_i,\, \sigma^2)$$

and so, for ϕ denoting the PDF of a standard normal,

$$l = \sum_{i=1}^{n} \log \phi \left(\frac{Y_i - \beta \cdot X_i}{\sigma} \right) + C = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta \cdot X_i)^2 + C.$$

In particular, under the Normal noise assumptions, Theorem 13.21 tells us that the least squares estimate $\hat{\beta}_S$ is the MLE and so the AIC is

$$AIC(S) = l(\hat{\beta}_S) - |S| = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_S \cdot X_i\right)^2 - |S| + C = -\frac{1}{2\sigma^2} \hat{R}(S) + C.$$

This verifies that maximizing the AIC is thus equivalent to minimizing Mallows' C_p statistic $\hat{R}$.

Exercise A.13.7 (AIC and model selection)**.** In this question we will take a closer look at the AIC method. Let $X_1, \dots, X_n$ be IID observations. Consider two models $\mathcal{M}_0$ and $\mathcal{M}_1$. Under $\mathcal{M}_0$ the data are assumed to be $N(0,\, 1)$ while under $\mathcal{M}_1$ the data are assumed to be $N(\theta,\, 1)$ for some unknown $\theta \in \mathbb{R}$:

$$\begin{aligned} \mathcal{M}_0 &: X_1, \dots, X_n \sim N(0,\, 1) \\ \mathcal{M}_1 &: X_1, \dots, X_n \sim N(\theta,\, 1),\, \theta \in \mathbb{R}. \end{aligned}$$

This is just another way to view the hypothesis testing problem

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta \neq 0.$$

Let $l_n(\theta)$ be the log-likelihood function. The AIC score for a model is the log-likelihood at the MLE minus the number of parameters. (Some people multiply this score by 2 but that is irrelevant.) Thus, the AIC score for $\mathcal{M}_0$ is $AIC_0 = l_n(0)$ and the AIC score for $\mathcal{M}_1$ is $AIC_1 = l_n(\hat{\theta}) - 1$. Suppose we choose the model with the highest AIC score. Let J_n denote the selected model:

$$J_n = \begin{cases} 0 & \text{if } AIC_0 > AIC_1 \\ 1 & \text{if } AIC_1 \geqslant AIC_0. \end{cases}$$

(1) Suppose that $\mathcal{M}_0$ is the true model, i.e. $\theta = 0$. Find

$$\lim_{n \to \infty} \mathbb{P}(J_n = 0).$$

Now compute $\lim_{n\to\infty} \mathbb{P}(J_n = 0)$ when $\theta \neq 0$.

(2) The fact that $\lim_{n\to\infty} \mathbb{P}(J_n = 0) \neq 1$ when $\theta = 0$ is why some people say that AIC "overfits". But this is not quite true as we shall now see. Let $\phi_\theta(x)$ denote a Normal density function with mean θ and variance 1. Define

$$\hat{f}_n(x) = \begin{cases} \phi_0(x) & \text{if } J_n = 0 \\ \phi_{\hat{\theta}}(x) & \text{if } J_n = 1. \end{cases}$$

If $\theta = 0$, show that $D(\phi_0,\, \hat{f}_n) \xrightarrow{P} 0$ as $n \to \infty$ where

$$D(f,\, g) = \int f(x) \log \left(\frac{f(x)}{g(x)} \right) dx$$

is the Kullback-Leibler distance. Show also that $D(\phi_\theta, \hat{f}_n) \overset{P}{\to} 0$ if $\theta \neq 0$. Hence, AIC consistently esimates the true density even if it "overshoots" the correct model.

(3) Repeat this analysis for BIC which is the log-likelihood minus $(p/2)\log n$ where p is the number of parameters and n is the sample size.

Solution. (1) By definition

$$\mathbb{P}(J_n = 0) = \mathbb{P}(AIC_0 > AIC_1) = \mathbb{P}(l_n(0) > l_n(\hat{\theta}) - 1).$$

Up to constants independent of θ the log-likelihood is given by

$$l_n(\theta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(X_i - \theta)^2\right]\right) = C - \frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2.$$

For X^* denoting a bootstrap sample and $\mathbb{E}_n$ denoting expectation with respect to the empirical distribution induced by $X_1, \dots, X_n$ we compute that, since random variables with mean zero are orthogonal to constants,

$$\sum_{i=1}^{n}(X_i - \theta)^2 = n\mathbb{E}_n(X^* - \theta)^2 = n\mathbb{E}_n(X^* - \overline{X})^2 + n(\overline{X} - \theta)^2$$

for $\overline{X}$ denoting the sample mean, and so

$$l_n(\theta) = C - \frac{n}{2}(\overline{X} - \theta)^2,$$

which means that $\hat{\theta} = \overline{X}$.

First we consider the case $\theta = 0$, i.e. work under $\mathcal{M}_0$. In that case we have that

$$\overline{X} \sim N\left(0, \frac{1}{n}\right)$$

and so, for Z denoting a standard Normal and Φ denoting its CDF,

$$\begin{aligned}
\mathbb{P}(J_n = 0) &= \mathbb{P}\left(-\frac{n}{2}\overline{X} > -1\right) \\
&= \mathbb{P}\left(|\sqrt{n}\,\overline{X}| < \sqrt{2}\right) \\
&= \mathbb{P}\left(|Z| < \sqrt{2}\right) \\
&= \Phi(\sqrt{2}) - \Phi(-\sqrt{2}) \\
&= 2\Phi(\sqrt{2}) - 1 \\
&\approx 0.84.
\end{aligned}$$

So in particular

$$\lim_{n\to\infty} \mathbb{P}(J_n = 0) = 2\Phi(\sqrt{2}) - 1 < 1 \text{ under } \mathcal{M}_0.$$

Now we turn our attention to the case $\theta \neq 0$, i.e. work under $\mathcal{M}_1$. Then

$$\begin{aligned}\mathbb{P}\left(J_n = 0\right) &= \mathbb{P}\left(n\overline{X}^2 < 2\right)\\ &= \mathbb{P}\left(n\big(\overline{X} - \theta\big)^2 < 2 - 2n\theta\overline{X} + n\theta^2\right)\\ &= \mathbb{P}\left(n\big(\overline{X} - \theta\big)^2 < 2 + n\theta(\theta - 2\overline{X})\right)\\ &\leqslant \mathbb{P}\left(2 + n\theta(\theta - 2\overline{X}) > 0\right)\\ &= \mathbb{P}\left(\overline{X} - \theta < \frac{1}{n\theta} - \frac{\theta}{2}\right).\end{aligned}$$

We can now either use the fact that, under $\mathcal{M}_1$, $\overline{X} \sim N(\theta,\, 1/n)$, such that, for Φ denoting the CDF of a standard normal,

$$\begin{aligned}\mathbb{P}\left(\overline{X} - \theta < \frac{1}{n\theta} - \frac{\theta}{2}\right) &= \mathbb{P}\left(\sqrt{n}(\overline{X} - \theta) < \frac{1}{\sqrt{n}\theta} - \frac{\sqrt{n}\theta}{2}\right)\\ &= \Phi\left(\frac{1}{\sqrt{n}\theta} - \frac{\sqrt{n}\theta}{2}\right) \to \Phi(-\infty) = 0 \text{ as } n \to \infty\end{aligned}$$

or use Chebyshev's inequality such that

$$\begin{aligned}\mathbb{P}\left(\overline{X} - \theta < \frac{1}{n\theta} - \frac{\theta}{2}\right) &\leqslant \mathbb{P}\left(|\overline{X} - \theta| > \frac{\theta}{2} - \frac{1}{n\theta}\right)\\ &\leqslant \frac{\mathbb{V}\overline{X}}{\left(\frac{\theta}{2} - \frac{1}{n\theta}\right)^2} = \frac{1}{n\left(\frac{\theta}{2} - \frac{1}{n\theta}\right)^2} \to 0 \text{ as } n \to \infty.\end{aligned}$$

Either way we conclude that

$$\lim_{n\to\infty} \mathbb{P}\left(J_n = 0\right) = 0 \text{ under } \mathcal{M}_1.$$

(2) First we compute that, for any $\theta_1,\, \theta_2 \in \mathbb{R}$,

$$D\left(\phi_{\theta_1},\, \phi_{\theta_2}\right) = \int \phi_{\theta_1} \log \frac{\phi_{\theta_1}(x)}{\phi_{\theta_2}(x)} dx = \mathbb{E}_{\theta_1}\left[\log \frac{\phi_{\theta_1}(X)}{\phi_{\theta_2}(X)}\right]$$

where $\mathbb{E}_\theta[f(X)]$ means that we evaluate the expectation under $X \sim N(\theta,\, 1)$. Since

$$\begin{aligned}\log \frac{\phi_{\theta_1}(X)}{\phi_{\theta_2}(X)} &= \log \frac{\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(X - \theta_1)^2\right]}{\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(X - \theta_2)^2\right]}\\ &= -\frac{1}{2}\left[(X - \theta_1)^2 - (X - \theta_2)^2\right]\\ &= -\frac{1}{2}\left[-2X(\theta_1 - \theta_2) + \theta_1^2 - \theta_2^2\right]\\ &= -\frac{1}{2}(\theta_1 + \theta_2 - 2X)(\theta_1 - \theta_2)\end{aligned}$$

we obtain that

$$\begin{aligned} D\left(\phi_{\theta_1},\, \phi_{\theta_2}\right) &= -\frac{1}{2}(\theta_1-\theta_2)\mathbb{E}_{\theta_1}(\theta_1+\theta_2-2X) \\ &= -\frac{1}{2}(\theta_1-\theta_2)(\theta_2-\theta_1) \\ &= \frac{1}{2}(\theta_1-\theta_2)^2. \end{aligned}$$

We now focus on the case $\theta = 0$. Since the Kullback-Leibler divergence is positive-definite we see that

$$\begin{aligned} |D(\phi_0,\, \hat{f}_n)| &= D(\phi_0,\, \hat{f}_n) \\ &= \mathbb{1}(J_n=0)\underbrace{D(\phi_0,\, \phi_0)}_{=0} + \mathbb{1}(J_n=1)D(\phi_0,\, \phi_{\hat{\theta}}) \\ &= \mathbb{1}(J_n=1)\cdot\frac{\theta^2}{2}. \end{aligned}$$

Since $\hat{\theta} = \overline{X}$ with $\sqrt{n}\,\overline{X} \sim N(0,\,1)$ we deduce that, under $\mathcal{M}_0$ (i.e. when $\theta = 0$), for any $\varepsilon > 0$ and for $Z \sim N(0,\,1)$,

$$\begin{aligned} \mathbb{P}\left(|D(\phi_0,\, \hat{f}_n)| > \varepsilon\right) &= \mathbb{P}\left(\frac{\overline{X}^2}{2} > \varepsilon \text{ and } J_n = 1\right) \\ &\leqslant \mathbb{P}\left(|\sqrt{n}\,\overline{X}| > \sqrt{2n\varepsilon}\right) \\ &= \mathbb{P}\left(|Z| > \sqrt{2n\varepsilon}\right) \to 0 \text{ as } n \to \infty, \end{aligned}$$

verifying that indeed

$$|D(\phi_0,\, \hat{f}_n)| \xrightarrow{P} 0 \text{ as } n \to \infty \text{ under } \mathcal{M}_0.$$

We now focus on the case $\theta \neq 0$. Proceeding as above we see that

$$\begin{aligned} |D(\phi_\theta,\, \hat{f}_n)| &= D(\phi_\theta,\, \hat{f}_n) \\ &= \mathbb{1}(J_n=0)D(\phi_\theta,\, \phi_0) + \mathbb{1}(J_n=1)D(\phi_\theta,\, \phi_{\hat{\theta}}) \\ &= \mathbb{1}(J_n=0)\cdot\frac{\theta^2}{2} + \mathbb{1}(J_n=1)\cdot\frac{(\theta-\hat{\theta})^2}{2}. \end{aligned}$$

Since the pair of sets $\{J_n = 0\}$ and $\{J_n = 1\}$ forms a partition we may deduce that, for any $\varepsilon > 0$,

$$\begin{aligned} &\mathbb{P}\left(|D(\phi_\theta,\, \hat{f}_n)| > \varepsilon\right) \\ &= \mathbb{P}\left(|D(\phi_\theta,\, \hat{f}_n)| > \varepsilon \text{ and } J_n = 0\right) \\ &\quad + \mathbb{P}\left(|D(\phi_\theta,\, \hat{f}_n)| > \varepsilon \text{ and } J_n = 1\right) \\ &\leqslant \mathbb{P}(J_n = 0) + \underbrace{\mathbb{P}\left(\frac{(\theta-\hat{\theta})^2}{2} > \varepsilon \text{ and } J_n = 1\right)}_{(\star)} \end{aligned}$$

where Chebyshev's inequality tells us that

$$(\star) \leqslant \mathbb{P}\left((\overline{X}-\theta)^2 > 2\varepsilon\right) \leqslant \frac{\mathbb{V}\overline{X}}{2\varepsilon} = \frac{1}{2n\varepsilon}.$$

Since we have showed in item 1 that $\mathbb{P}(J_n = 0) \to 0$ as $n \to \infty$ under $\mathcal{M}_1$ we conclude that

$$\mathbb{P}\left(|D(\phi_\theta,\, \hat{f}_n)| > \varepsilon\right) \to 0 \text{ as } n \to \infty \text{ for any } \varepsilon > 0,$$

i.e.

$$|D(\phi_\theta,\, \hat{f}_n)| \xrightarrow{P} 0 \text{ as } n \to \infty \text{ under } \mathcal{M}_1.$$

(3) We now use the BIC such that

$$BIC_0 > BIC_1 \iff l_n(0) > l_n(\hat{\theta}) - \frac{1}{2}\log n.$$

So let us define

$$\widetilde{J}_n := \begin{cases} 0 & \text{if } BIC_0 > BIC_1 \\ 1 & \text{if } BIC_1 \geqslant BIC_0 \end{cases}$$

Then, under $\mathcal{M}_0$, $\overline{X} \sim N(0,\, 1/n)$ and so

$$\begin{aligned} \mathbb{P}\left(\widetilde{J}_n = 0\right) &= \mathbb{P}\left(-\frac{n\overline{X}^2}{2} > -\frac{1}{2}\log n\right) \\ &= \mathbb{P}\left(|\sqrt{n}\,\overline{X}| < \sqrt{\log n}\right) \\ &= \mathbb{P}\left(|Z| < \sqrt{\log n}\right) \to 1 \text{ as } n \to \infty \end{aligned}$$

for $Z \sim N(0,\, 1)$. Meanwhile, under $\mathcal{M}_1$, proceeding as in item 1 tells us that, using the approach leveraging Chebyshev's inequality,

$$\begin{aligned} \mathbb{P}\left(\widetilde{J}_n = 0\right) &= \mathbb{P}\left(n\overline{X}^2 < \log n\right) \\ &= \mathbb{P}\left(n(\overline{X}-\theta)^2 < \log n + n\theta(\theta - 2\overline{X})\right) \\ &\leqslant \mathbb{P}\left(\log n + n\theta(\theta - 2\overline{X}) > 0\right) \\ &= \mathbb{P}\left(\overline{X} - \theta < \frac{\log n}{n\theta} - \frac{\theta}{2}\right) \\ &\leqslant \mathbb{P}\left(|\overline{X} - \theta| > \frac{\theta}{2} - \frac{\log n}{n\theta}\right) \\ &\leqslant \frac{\mathbb{V}\overline{X}}{\left(\frac{\theta}{2} - \frac{\log n}{n\theta}\right)^2} \\ &= \frac{1}{n\left(\frac{\theta}{2} - \frac{\log n}{n\theta}\right)^2} \to 0 \text{ as } n \to \infty. \end{aligned}$$

In summary: by contrast with using AIC, with BIC we now have that

$$\mathbb{P}\left(\widetilde{J}_n = 0\right) \to 1 \text{ as } n \to \infty \text{ under } \mathcal{M}_0,$$

while the behaviour under $\mathcal{M}_1$ is comparable since

$$\mathbb{P}\left(\widetilde{J}_n = 0\right) \to 0 \text{ as } n \to \infty \text{ under } \mathcal{M}_1,$$

as before.

Now we turn our attention to the Kullback-Leibler divergence between the true density and the *new* estimated density

$$\tilde{f}_n(x) = \begin{cases} \phi_0(x) & \text{if } \widetilde{J}_n = 0 \\ \phi_{\hat{\theta}}(x) & \text{if } \widetilde{J}_n = 1 \end{cases}$$

when using the BIC. Crucially, we may reason as in item 2 above and derive that

$$D(\phi_\theta, \tilde{f}_n) = \mathbb{1}(\widetilde{J}_n = 0) D(\phi_\theta, \phi_0) + \mathbb{1}(\widetilde{J}_n = 1) D(\phi_\theta, \phi_{\hat{\theta}})$$

for all $\theta \in \mathbb{R}$ (including $\theta = 0$). Since $\mathbb{P}(\widetilde{J}_n = 0) \to 0$ as $n \to \infty$ under $\mathcal{M}_1$ we may then conclude as in item 2 that

$$D(\phi_0, \tilde{f}_n) \xrightarrow{P} 0 \text{ as } n \to \infty \text{ under } \mathcal{M}_0 \text{ and}$$
$$D(\phi_\theta, \tilde{f}_n) \xrightarrow{P} 0 \text{ as } n \to \infty \text{ under } \mathcal{M}_1.$$

Exercise A.13.8 (Prediction intervals for simple linear regression)**.** In this question we take a closer look at prediction intervals. Let $\theta = \beta_0 + \beta_1 X_*$ and let $\hat{\theta} = \hat{\beta}_0 + \hat{\beta}_1 X_*$. Thus, $\hat{Y}_* = \hat{\theta}$ while $Y_* = \theta + \varepsilon$. Now, $\hat{\theta} \approx N(\theta, se^2)$ where

$$se^2 = \mathbb{V}(\hat{\theta}) = \mathbb{V}(\hat{\beta}_0 + \hat{\beta}_1 x_*).$$

Note that $\mathbb{V}(\hat{\theta})$ is the same as $\mathbb{V}(\hat{Y}_*)$. Now, $\hat{\theta} \pm 2\sqrt{\mathbb{V}(\hat{\theta})}$ is an approximate 95 percent confidence interval for $\theta = \beta_0 + \beta_1 x_*$ using the usual argument for a confidence interval. But, as you shall now show, it is not a valid confidence interval for Y_*. Note that here we require the Normal noise assumption for the simple linear regression model to hold.

(1) Let $s = \sqrt{\mathbb{V}(\hat{Y}_*)}$. Show that

$$\mathbb{P}(\hat{Y}_* - 2s < Y_* < \hat{Y}_* + 2s) \approx \mathbb{P}(-2 < N(0,\, 1 + \sigma^2/s^2) < 2) \neq 0.95$$

(2) The problem is that the quantity of interest Y_* is equal to a parameter θ plus a random variable. We can fix this by defining

$$\xi_n^2 = \mathbb{V}(\hat{Y}_*) + \sigma^2 = \left[\frac{\sum_{i=1}^n (x_i - x_*)^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} + 1 \right] \sigma^2.$$

In practice, we substitute $\hat{\sigma}$ for σ and we denote the resulting quantity by $\hat{\xi}_n$. Now consider the interval $\hat{Y}_* \pm 2\hat{\xi}_n$. Show that

$$\mathbb{P}(\hat{Y}_* - 2\hat{\xi}_n < Y_* < \hat{Y}_* + 2\hat{\xi}_n) \approx \mathbb{P}(-2 < N(0,\, 1) < 2) \approx 0.95.$$

Solution. (1) We may rearrange

$$\mathbb{P}(\hat{Y}_* - 2s < Y_* < \hat{Y}_* + 2s) = \mathbb{P}\left(\left| \frac{\hat{Y}_* - Y_*}{s} \right| < 2 \right)$$

where

$$\hat{Y}_* - Y_* = (\theta + \varepsilon) - \hat{\theta}$$

and so

$$\mathbb{P}(\hat{Y}_* - 2s < Y_* < \hat{Y}_* + 2s) = \mathbb{P}\left(\left| \frac{(\hat{\theta} - \theta) + \varepsilon}{s} \right| < 2 \right).$$

Crucially: $\hat{\theta}$ and ε are conditionally *independent* given the feature data X^n with, by the Normal noise assumption,

$$(\hat{\theta}-\theta)\,|\,X^n \approx N(0,\, s^2) \text{ and } \varepsilon\,|\,X^n \sim N(0,\, \sigma^2)$$

such that

$$\left.\frac{(\hat{\theta}-\theta)+\varepsilon}{s}\,\right|\,X^n \approx N\left(0,\, \frac{s^2+\sigma^2}{s^2}\right) = N\left(0,\, 1+\frac{\sigma^2}{s^2}\right).$$

Therefore, by the rule of iterated expectation,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{(\hat{\theta}-\theta)+\varepsilon}{s}\right| < 2\right). &= \mathbb{E}\left[\mathbb{P}\left(\left|\frac{(\hat{\theta}-\theta)+\varepsilon}{s}\right| < 2 \,\middle|\, X^n\right)\right] \\ &\approx \mathbb{E}\left[\mathbb{P}\left(\left|N\left(0,\, 1+\frac{\sigma^2}{s^2}\right)\right| < 2 \,\middle|\, X^n\right)\right].\end{aligned}$$

In particular, since $2 = z_{\alpha/2}$ for $\alpha = 2(1-\Phi(2)) \approx 0.46$, we may compute that, for any $\nu > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|N(0,\, \nu^2)\right| < z_{\alpha/2}\right) &= \mathbb{P}\left(|N(0,\, 1)| < z_{\alpha/2}/\nu\right) \\ &= \mathbb{P}\left(-z_{\alpha/2}/\nu < N(0,\, 1) < z_{\alpha/2}/\nu\right) \\ &= \Phi(z_{\alpha/2}/\nu) - \Phi(-z_{\alpha/2}/\nu) \\ &= 2\Phi(z_{\alpha/2}/\nu) - 1 \\ &= 2\Phi\left[\Phi^{-1}(1-\alpha/2)/\nu\right] - 1.\end{aligned}$$

So finally, since $\nu^2 := 1 + \sigma^2/s^2 > 1$ always (remember: s, and hence ν, are random variables) and since Φ is strictly increasing we conclude that

$$\begin{aligned}\mathbb{P}(\hat{Y}_* - 2s < Y_* < \hat{Y}_* + 2s) &\approx \mathbb{E}\left[\mathbb{P}\left(\left|N\left(0,\, 1+\frac{\sigma^2}{s^2}\right)\right| < 2 \,\middle|\, X^n\right)\right] \\ &= \mathbb{E}\left(2\Phi\left[\Phi^{-1}(1-\alpha/2)/\nu\right] - 1 \,\middle|\, X^n\right) \\ &< 2\Phi\left[\Phi^{-1}(1-\alpha/2)\right] - 1 \\ &= 1 - \alpha = 2\Phi(2) - 1 \approx 0.954.\end{aligned}$$

In other words there is some value $p_* < 0.954$ such that

$$\mathbb{P}(\hat{Y}_* - 2s < Y_* < \hat{Y}_* + 2s) \to p_* \text{ as } n \to \infty.$$

(2) This follows from exactly the same reasoning as that which was used in item 1:

$$\mathbb{P}(\hat{Y}_* - 2\hat{\xi}_n < Y_* < \hat{Y}_* + 2\hat{\xi}_n) = \mathbb{P}\left(\left|\frac{\hat{Y}_* - Y_*}{\hat{\xi}_n}\right| < 2\right)$$

where

$$\frac{\hat{Y}_* - Y_*}{\hat{\xi}_n} = \frac{(\hat{\theta}-\theta)+\varepsilon}{\sqrt{s^2+\hat{\sigma}^2}} \approx \frac{(\hat{\theta}-\theta)+\varepsilon}{\sqrt{s^2+\sigma^2}} \approx N\left(0,\, \frac{s^2+\sigma^2}{s^2+\sigma^2}\right) = N(0,\, 1)$$

and so indeed

$$\mathbb{P}(\hat{Y}_* - 2\hat{\xi}_n < Y_* < \hat{Y}_* + 2\hat{\xi}_n) = \mathbb{P}\left(|N(0,\, 1)| < 2\right) = 2\Phi(2) - 1 \approx 0.954.$$

Finally we derive the desired formula for $\mathbb{V}(\hat{Y}_* \,|\, X^n)$. We compute that

$$\begin{aligned}\mathbb{V}(\hat{Y}_* \,|\, X^n) &= \mathbb{V}(\hat{\theta} \,|\, X^n)\\ &= \mathbb{V}(\hat{\beta}_0 + x_* \hat{\beta}_1 \,|\, X^n)\\ &= \mathbb{V}(\hat{\beta}_0 \,|\, X^n) + 2x_* \operatorname{Cov}(\hat{\beta}_0,\, \hat{\beta}_1 \,|\, X^n) + x_*^2 \mathbb{V}(\hat{\beta}_1 \,|\, X^n)\\ &= \mathbb{V}(\hat{\beta} \,|\, X^n) \begin{pmatrix} 1 \\ x_* \end{pmatrix} \cdot \begin{pmatrix} 1 \\ x_* \end{pmatrix}.\end{aligned}$$

So Theorem 13.22 tells us that

$$\begin{aligned}\mathbb{V}(\hat{Y}_* \,|\, X^n) &= \frac{\sigma^2}{n s_X^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix} \begin{pmatrix} 1 \\ x_* \end{pmatrix} \cdot \begin{pmatrix} 1 \\ x_* \end{pmatrix}\\ &= \frac{\sigma^2}{n s_X^2} \left(\frac{1}{n} \sum_{i=1}^n X_i^2 - 2x_* \overline{X}_n + x_*^2 \right)\\ &= \frac{\sigma^2}{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2} \cdot \frac{1}{n} \sum_{i=1}^n (X_i^2 - 2x_* X_i + x_*^2)\\ &= \frac{\sigma^2 \sum_{i=1}^n \left(X_i - x_*\right)^2}{n \sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2},\end{aligned}$$

as desired.

A.14. **Multivariate Models.**

Exercise A.14.1 (Linear algebra of expectation and variance)**.** Prove Theorem 14.1 where we record how expectations and variances behave under vector dot products and matrix–vector multiplication.

Solution. The first item follows from the linearity of expectation.

$$\mathbb{E}(a \cdot X) = \mathbb{E}\left(\sum_{i=1}^{k} a_i X_i\right) = \sum_{i=1}^{k} a_i \mathbb{E}\left(X_i\right) = a \cdot \mu.$$

The second item follows from the bilinearity of the covariance.

$$\begin{aligned}\mathbb{V}(a \cdot X) = \mathbb{V}\left(\sum_{i=1}^{k} a_i X_i\right) &= \operatorname{Cov}\left(\sum_{i=1}^{k} a_i X_i,\, \sum_{j=1}^{k} a_j X_j\right)\\ &= \sum_{i,j=1}^{k} a_i a_j \operatorname{Cov}(X_i,\, X_j) = \sum_{i,j=1}^{k} a_i a_j \Sigma_{ij} = \Sigma a \cdot a.\end{aligned}$$

Now let A_i denote the i-th row of A. The third item then follows from the first.

$$\mathbb{E}(AX)_i = \mathbb{E}(A_i \cdot X) = A_i \cdot \mu = (A\mu)_i.$$

Finally the last item follows from using the bilinearity of the covariance once more.

$$\begin{aligned}\mathbb{V}(AX)_{ij} = \operatorname{Cov}\left((AX)_i,\, (AX)_j\right) &= \operatorname{Cov}\left(\sum_{l=1}^{k} A_{il} X_l,\, \sum_{m=1}^{k} A_{jm} X_m\right)\\ &= \sum_{l,m=1}^{k} A_{il} A_{jm} \operatorname{Cov}\left(X_l,\, X_m\right)\\ &= \sum_{l,m=1}^{k} A_{il} A_{jm} \Sigma_{lm} = \left(A \Sigma A^T\right)_{ij}.\end{aligned}$$

Exercise A.14.2 (Fisher information for Multinomial)**.** Find the Fisher information matrix for the MLE of a Multinomial.

Solution. The key observation is that $X \sim \text{Multinomial}(n,\, p)$ if and only if

$$X = \sum_{i=1}^{n} Y_i$$

for $Y^{(1)},\, \dots,\, Y^{(n)} \sim \text{Categorical}(p)$ IID. As shown in Exercise A.23.14 (see also Remark A.5 and Exercise A.23.13), in order to compute the Fisher information matrix of the Categorical model we must *parametrize* that model by using the constraint $\sum_{j=1}^{k} p_j = 1$ to write one of the components of p in terms of the remaining $k-1$ components.

So let us do so by writing $p_k = \sum_{j=1}^{k-1} p_j$ and defining $q := (p_1,\, \dots,\, p_{k-1})$. As seen in Exercise A.23.14 the MLE is then

$$\hat{q} = \left(\overline{Y}_1,\, \dots,\, \overline{Y}_{k-1}\right)$$

where

$$\overline{Y}_j := \frac{1}{n} \sum_{i=1}^{n} Y_j^{(i)} = \frac{1}{n} X_j$$

such that

$$\hat{q} = \frac{1}{n}\left(X_1, \ldots, X_{k-1}\right).$$

Exercise A.23.14 also tells us that the Fisher information is

$$\begin{aligned} I(q) &= \operatorname{diag}\left(\frac{1}{q}\right) + \frac{1}{1 - q_1 - \cdots - q_{k-1}}\mathbb{1} \\ &= \begin{pmatrix} \frac{1}{q_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{q_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{q_{k-1}} \end{pmatrix} + \frac{1}{1 - q_1 - \cdots - q_{k-1}} \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{pmatrix}. \end{aligned}$$

In particular when we evaluate the Fisher information at the MLE this becomes, for $\widetilde{X} := (X_1, \ldots, X_{k-1})$,

$$\begin{aligned} I(\hat{q}) &= \operatorname{diag}\left(\frac{n}{\widetilde{X}}\right) + \left(1 - \frac{1}{n}\sum_{j=1}^{k-1} X_j\right)^{-1}\mathbb{1} \\ &= \operatorname{diag}\left(\frac{n}{\widetilde{X}}\right) + \frac{1}{1 - X_k/n}\mathbb{1}. \end{aligned}$$

A.15. **Inference About Independence.**

Exercise A.15.1 (Characterization of the independence of two binary variables)**.** Prove Theorem 15.6 where we establish various characterizations of the independence between two binary random variables Y and Z, namely

(1) $Y \amalg Z$
(2) The odds ratio $\psi = 1$.
(3) The log odds ratio $\gamma = 0$.
(4) $p_{ij} = p_{i\cdot}p_{\cdot j}$ for all i, j, where p denotes the two-by-two parameter associated with Y and Z.

Solution. It follows immediately from the definition of the independence of random variables that (1) and (4) are equivalent since

$$\begin{aligned} &\iff p_{ij} = p_{i\cdot}p_{\cdot j} \text{ for all } i,\, j \\ &\iff \mathbb{P}\left(Y = j,\, Z = i\right) = \mathbb{P}\left(Y = j\right)\mathbb{P}\left(Z = i\right) \text{ for all } i,\, j \\ &\iff Y \amalg Z. \end{aligned}$$

The equivalence of (3) and (2) is also immediate since $\gamma = \log \psi$ and the since logarithm is a bijection which vanishes at one.

So finally we show that (2) and (4) are equivalent. To do so we introduce

$$\delta := p_{00}p_{11} - p_{01}p_{10}$$

such that

$$\psi = 1 \iff \frac{p_{00}p_{11}}{p_{01}p_{10}} = 1 \iff p_{00}p_{11} = p_{01}p_{10} \iff \delta = 0.$$

It thus suffices to show that

$$(4) \iff \delta = 0.$$

This equivalence comes down to elementary algebra. Since

$$p_{00} + p_{01} + p_{10} + p_{11} = 1$$

we may compute that

$$\begin{aligned} p_{0\cdot}p_{\cdot 0} &= \left(p_{00} + p_{01}\right)\left(p_{00} + p_{10}\right) = p_{00}\left(1 - p_{11}\right) + p_{01}p_{10} = p_{00} - \delta, \\ p_{0\cdot}p_{\cdot 1} &= \left(p_{00} + p_{01}\right)\left(p_{01} + p_{11}\right) = p_{01}\left(1 - p_{10}\right) + p_{00}p_{11} = p_{01} + \delta, \\ p_{1\cdot}p_{\cdot 0} &= \left(p_{10} + p_{11}\right)\left(p_{00} + p_{10}\right) = p_{10}\left(1 - p_{01}\right) + p_{11}p_{00} = p_{10} + \delta, \text{ and} \\ p_{1\cdot}p_{\cdot 1} &= \left(p_{10} + p_{11}\right)\left(p_{01} + p_{11}\right) = p_{11}\left(1 - p_{00}\right) + p_{10}p_{01} = p_{11} - \delta. \end{aligned}$$

In other words:

$$\begin{aligned} \delta &= p_{00} - p_{0\cdot}p_{\cdot 0} \\ &= p_{0\cdot}p_{\cdot 1} - p_{01} \\ &= p_{1\cdot}p_{\cdot 0} - p_{10} \\ &= p_{11} - p_{1\cdot}p_{\cdot 1} \end{aligned}$$

and so

$$\delta = 0 \iff (4)$$

as desired, which concludes the proof.

Exercise A.15.2 (Likelihood ratio test for the independence of two binary random variables)**.** Prove Theorem 15.7 where we derive a likelihood ratio test for the independence of two binary random variables.

Solution. As noted in Definition 15.3, under either the null or the alternative hypothesis the two-by-two random variable X satisfies

$$X \sim \text{Multinomial}(n,\, p)$$

for $p \in \Delta^3 \subseteq \mathbb{R}^4$ denoting the two-by-two parameter. Moreover, under the null hypothesis Theorem 15.6 tells us that p must satisfy

$$\gamma(p) = 0.$$

This verifies that, for a likelihood ratio test of

$$H_0 : Y \amalg Z \text{ versus } H_1 : Y \leftrightsquigarrow Z,$$

we may indeed use the model

$$\mathcal{F} := \left\{\text{Multinomial}(n,\, p) : p \in \Delta^3\right\}$$

and the sets

$$\Theta_0 := \left\{p = (p_{00},\, p_{01},\, p_{10},\, p_{11}) \in \Delta^3 : \gamma(p) = 0\right\}$$

and $\Theta_1 := \Delta^3$. Moreover, since Θ_0 differs from Θ_1 by the imposition of one scalar constraint, we know that

$$\dim \Theta_1 - \dim \Theta_0 = 1$$

and so indeed the rejection region is, for any $\alpha \in (0,\, 1)$,

$$R_\alpha := \left\{t \in \mathbb{R} : t > \chi^2_{1,\,\alpha}\right\}.$$

So all that remains to do is to verify that the likelihood ratio statistic takes the desired form. From Exercise A.23.13 we know that, since

$$\begin{aligned} & X \sim \text{Multinomial}(n,\, p) \\ \iff & X = \sum_{i=1}^{n} X^{(i)} \text{ where } X^{(1)},\, \ldots,\, X^{(n)} \sim \text{Categorical}(p) \text{ IID}, \end{aligned}$$

the likelihood function is

$$\mathcal{L}(p) = p_{00}^{X_{00}} p_{01}^{X_{01}} p_{10}^{X_{10}} p_{11}^{X_{11}}$$

and the MLE $\hat{p}$ over the *full* parameter space Θ_1 is

$$\hat{p} = \frac{X}{n}.$$

Now we compute the MLE $\hat{q}$ over the *restricted* parameter space Θ_0. We proceed as in Exercise A.23.13 and use Lagrange multipliers, noting that there are now *two* constraints to satisfy:

$$q_{00} + q_{01} + q_{10} + q_{11} = 1 \text{ and } \gamma(q) = 0.$$

So we define, for the Lagrange multiplier $\lambda \in \mathbb{R}^2$,

$$f(q,\, \lambda) := \log \mathcal{L}(q) - \lambda \cdot \left(\sum_{i,j=0}^{1} q_{ij} - 1,\, \gamma(q)\right)$$

and seek to maximize f over $\mathbb{R}^4 \times \mathbb{R}^2$. Since

$$\log \mathcal{L}(q) = X_{00} \log q_{00} + X_{01} \log q_{01} + X_{10} \log q_{10} + X_{11} \log q_{11}$$

and since

$$\gamma(q) = \log \frac{q_{00} q_{11}}{q_{01} q_{10}} = \log q_{00} - \log q_{01} - \log q_{10} + \log q_{11}$$

we compute that

$$\nabla f = \begin{pmatrix} X_{00}/q_{00} - \lambda_1 - \lambda_2/q_{00} \\ X_{01}/q_{01} - \lambda_1 + \lambda_2/q_{01} \\ X_{10}/q_{10} - \lambda_1 + \lambda_2/q_{10} \\ X_{11}/q_{11} - \lambda_1 - \lambda_2/q_{11} \\ 1 - q_{00} - q_{01} - q_{10} - q_{11} \\ -\log q_{00} + \log q_{01} + \log q_{10} - \log q_{11} \end{pmatrix}.$$

In particular

$$\begin{aligned} 0 &= q \cdot \nabla_q f \\ &= (X_{00} - \lambda_1 q_{00} - \lambda_2) + (X_{01} - \lambda_1 q_{01} + \lambda_2) \\ &\quad + (X_{10} - \lambda_1 q_{10} + \lambda_2) + (X_{11} - \lambda_1 q_{11} - \lambda_2) \\ &= (X_{00} + X_{01} + X_{10} + X_{11}) - (q_{00} + q_{01} + q_{10} + q_{11})\,\lambda_1 \\ &= n - \lambda_1, \end{aligned}$$

from which we deduce that

$$\lambda_1 = n.$$

To compute λ_2 we note that $q_{ij}\partial_{q_{ij}} f = 0$ for all i, j, i.e.

$$\begin{aligned} 0 &= X_{00} - nq_{00} - \lambda_2 \\ &= X_{01} - nq_{01} + \lambda_2 \\ &= X_{10} - nq_{10} + \lambda_2 \\ &= X_{11} - nq_{11} - \lambda_2 \\ \iff q_{00} &= (X_{00} - \lambda_2)/n, \\ q_{01} &= (X_{01} + \lambda_2)/n, \\ q_{10} &= (X_{10} + \lambda_2)/n, \text{ and} \\ q_{11} &= (X_{11} - \lambda_2)/n. \end{aligned}$$

Since $\gamma = 0$, or equivalently $q_{00}q_{11} = q_{01}q_{10}$, we deduce that

$$\begin{aligned} &(X_{00} - \lambda_2)(X_{11} - \lambda_2) = (X_{01} + \lambda_2)(X_{10} + \lambda_2) \\ &\iff -(X_{00} + X_{11})\,\lambda_2 + X_{00}X_{11} = (X_{01} + X_{10})\,\lambda_2 + X_{01}X_{10} \\ &\iff \lambda_2 = \frac{X_{00}X_{11} - X_{01}X_{10}}{X_{00} + X_{01} + X_{10} + X_{11}} = \frac{X_{00}X_{11} - X_{01}X_{10}}{n}. \end{aligned}$$

Inspired by Exercise A.15.1 and the four identities obtained therein equating the quantity $\delta := p_{00}p_{11} - p_{01}p_{10}$ to $(-1)^{i-j}(p_{ij} - p_{i\cdot}p_{\cdot j})$ we observe that, in exactly the same way, we may compute that

$$\begin{aligned} X_{00}X_{11} - X_{01}X_{10} &= nX_{00} - X_{0\cdot}X_{\cdot 0} \\ &= X_{0\cdot}X_{\cdot 1} - nX_{01} \\ &= X_{1\cdot}X_{\cdot 0} - nX_{10} \\ &= nX_{11} - X_{1\cdot}X_{\cdot 1}. \end{aligned}$$

Using these identities when plugging the expression for λ_2 into the expressions for q_{ij} in terms of X_{ij} and λ_2 recorded above we deduce that

$$\begin{aligned}
q_{00} &= \frac{X_{00}}{n} - \frac{X_{00}X_{11} - X_{01}X_{10}}{n^2} = \frac{X_{00}}{n} - \frac{X_{00}}{n} + \frac{X_{0\cdot}X_{\cdot 0}}{n^2} = \frac{X_{0\cdot}X_{\cdot 0}}{n^2}, \\
q_{01} &= \frac{X_{01}}{n} + \frac{X_{00}X_{11} - X_{01}X_{10}}{n^2} = \frac{X_{01}}{n} + \frac{X_{0\cdot}X_{\cdot 1}}{n^2} - \frac{X_{01}}{n} = \frac{X_{0\cdot}X_{\cdot 1}}{n^2}, \\
q_{10} &= \frac{X_{10}}{n} + \frac{X_{00}X_{11} - X_{01}X_{10}}{n^2} = \frac{X_{10}}{n} + \frac{X_{1\cdot}X_{\cdot 0}}{n^2} - \frac{X_{10}}{n} = \frac{X_{1\cdot}X_{\cdot 0}}{n^2}, \\
q_{11} &= \frac{X_{11}}{n} - \frac{X_{00}X_{11} - X_{01}X_{10}}{n^2} = \frac{X_{11}}{n} - \frac{X_{11}}{n} + \frac{X_{1\cdot}X_{\cdot 1}}{n^2} = \frac{X_{1\cdot}X_{\cdot 1}}{n^2}.
\end{aligned}$$

In other words

$$\hat{q}_{ij} = \frac{X_{i\cdot}X_{\cdot j}}{n^2}.$$

So finally we conclude that the likelihood ratio statistic is

$$\begin{aligned}
2\log\frac{\mathcal{L}(\hat{p})}{\mathcal{L}(\hat{q})} &= 2\log\frac{\hat{p}_{00}^{X_{00}}\hat{p}_{01}^{X_{01}}\hat{p}_{10}^{X_{10}}\hat{p}_{11}^{X_{11}}}{\hat{q}_{00}^{X_{00}}\hat{q}_{01}^{X_{01}}\hat{q}_{10}^{X_{10}}\hat{q}_{11}^{X_{11}}} \\
&= 2X_{00}\log\frac{\hat{p}_{00}}{\hat{q}_{00}} + 2X_{01}\log\frac{\hat{p}_{01}}{\hat{q}_{01}} + 2X_{10}\log\frac{\hat{p}_{10}}{\hat{q}_{10}} + 2X_{11}\log\frac{\hat{p}_{11}}{\hat{q}_{11}}
\end{aligned}$$

where, since $X_{\cdot\cdot} = n$,

$$\frac{\hat{p}_{ij}}{\hat{q}_{ij}} = \frac{X_{ij}/n}{X_{i\cdot}X_{\cdot j}/n^2} = \frac{X_{ij}X_{\cdot\cdot}}{X_{i\cdot}X_{\cdot j}}$$

such that indeed

$$2\log\frac{\mathcal{L}(\hat{p})}{\mathcal{L}(\hat{q})} = \sum_{i,j=0}^{1} 2X_{ij}\log\frac{\hat{p}_{ij}}{\hat{q}_{ij}} = \sum_{i,j=0}^{1} 2X_{ij}\log\frac{X_{ij}X_{\cdot\cdot}}{X_{i\cdot}X_{\cdot j}} = T,$$

as desired.

Exercise A.15.3 (MLE for the odds ratio)**.** Prove Theorem 15.12 where we record the MLE and estimates of the standard error for the odds ratio and log odds ratio.

Solution. We have estaslished in Exercise A.15.2 that the MLE for the two-by-two parameter p is

$$\hat{p} = \frac{X}{n}.$$

By equivariance of the MLE it then follows that the MLE for ψ and γ are given by

$$\hat{\psi} = \psi(\hat{p}) = \frac{\hat{p}_{00}\hat{p}_{11}}{\hat{p}_{01}\hat{p}_{10}} = \frac{\frac{X_{00}}{n}\cdot\frac{X_{11}}{n}}{\frac{X_{01}}{n}\cdot\frac{X_{10}}{n}} = \frac{X_{00}X_{11}}{X_{01}X_{10}}$$

and

$$\hat{\gamma} = \log\hat{\psi},$$

as desired.

Now we derive estimators of the standard errors. Note that

$$X = \sum_{i=1}^{n} X^{(i)} \text{ for } X^{(1)}, \ldots, X^{(n)} \sim \text{Categorical}(p)$$

where Exercise A.23.12 tells us that

$$\mathbb{E}X^{(i)} = p \text{ and } \mathbb{V}X^{(i)} = \operatorname{diag} p - p \otimes p.$$

Since the MLE

$$\hat{p} = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}$$

is a sample mean of this categorical sample, the Central Limit Theorem tells us that

$$\sqrt{n}(\hat{p} - p) \rightsquigarrow N\left(0,\, \operatorname{diag} p - p \otimes p\right).$$

In particular the MLE for the log odds ratio is

$$\hat{\gamma} = \log \hat{\psi} = \log \hat{p}_{00} - \log \hat{p}_{01} - \log \hat{p}_{10} + \log \hat{p}_{11} = \gamma(\hat{p})$$

with

$$\nabla\gamma(p) = \begin{pmatrix} 1/p_{00} \\ -1/p_{01} \\ -1/p_{10} \\ 1/p_{11} \end{pmatrix}$$

and so, since $v^T(w \otimes w)v = (v \cdot w)^2$, we may compute that

$$\begin{aligned}
\sigma_\gamma^2(p) &:= (\nabla\gamma)^T \left(\operatorname{diag} p - p \otimes p\right) \nabla\gamma \\
&= \left[\sum_{i,j=0}^{1} \frac{(-1)^{i-j}}{p_{ij}} \cdot p_{ij} \cdot \frac{(-1)^{i-j}}{p_{ij}}\right] - (p \cdot \nabla\gamma)^2 \\
&= \left(\sum_{i,j=0}^{1} \frac{1}{p_{ij}}\right) - \underbrace{\left(\frac{p_{00}}{p_{00}} - \frac{p_{01}}{p_{01}} - \frac{p_{10}}{p_{10}} + \frac{p_{11}}{p_{11}}\right)}_{=0} \\
&= \frac{1}{p_{00}} + \frac{1}{p_{01}} + \frac{1}{p_{10}} + \frac{1}{p_{11}}.
\end{aligned}$$

Therefore the delta method tells us that

$$\sqrt{n}(\hat{\gamma} - \gamma) \rightsquigarrow N\left(0,\, \sigma_\gamma^2(p)\right),$$

or equivalently

$$\frac{\hat{\gamma} - \gamma}{\sqrt{\sigma_\gamma^2(p)/n}} \rightsquigarrow N(0,\, 1),$$

which verifies that a suitable estimate for the standard error of $\hat{\gamma}$ is

$$\frac{\sigma_\gamma^2(\hat{p})}{n} = \frac{1}{n}\sum_{i,j=0}^{1} \frac{1}{\hat{p}_{ij}} = \frac{1}{n}\sum_{i,j=0}^{1} \frac{n}{X_{ij}} = \sum_{i,j=0}^{1} \frac{1}{X_{ij}} =: \widehat{se}^2(\hat{\gamma}),$$

as desired. Finally, since $\hat{\psi} = e^{\hat{\gamma}}$ with

$$\frac{\sqrt{n}(\hat{\gamma} - \gamma)}{\sigma_\gamma(p)} \rightsquigarrow N(0,\, 1)$$

we may use the delta method once again to compute that

$$\frac{\sqrt{n}(\hat{\psi} - \psi)}{e^{\hat{\gamma}}\sigma_\gamma(\hat{p})} = \frac{\hat{\psi} - \psi}{\hat{\psi}\widehat{se}(\hat{\gamma})} \rightsquigarrow N(0,\, 1),$$

i.e. indeed $\widehat{se}(\hat{\psi}) := \hat{\psi}\widehat{se}(\hat{\gamma})$ is an appropriate estimate of the standard error for $\hat{\psi}$.

A.16. **Causal Inference.**

Exercise A.16.1 (Association and causation are not related)**.** Create an example in which the association satisfies $\alpha > 0$ and the average causal effect satisfies $\theta < 0$.

Solution. This is example 5 in Remark 16.10.

Exercise A.16.2 (Estimation of the causal regression function)**.** Prove Theorem 16.20 where we show that, if the covariate X is independent from the counterfactual function then the regression function is equal to the causal regression function.

Solution. By the consistency relationship $Y = C(X)$ and the independence between X and C we see immediately that

$$r(x) = \mathbb{E}\left(Y \,|\, X = x\right) = \mathbb{E}\left[C(X) \,|\, X = x\right] = \mathbb{E}\left[C(x) \,|\, X = x\right] = \mathbb{E}\left[C(x)\right] = \theta(x)$$

as desired.

Exercise A.16.3 (Estimated bounds on the causal effect)**.** Suppose you are given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ from an observational study, where $X_i \in \{0, 1\}$ and $Y_i \in \{0, 1\}$. Although it is not possible to estimate the causal effect θ, it is possible to put bounds on θ. Find upper and lower bounds on θ that can be consistently estimated from the data. Show that the bounds have width 1.

Hint: Note that

$$\mathbb{E}(C_1) = \mathbb{E}\left(C_1 \,|\, X = 1\right) \mathbb{P}\left(X = 1\right) = \mathbb{E}\left(C_1 \,|\, X = 0\right) \mathbb{P}\left(X = 0\right).$$

Solution. Using the hint and the consistency relationship $Y = C_X$ we may write the causal effect θ as

$$\begin{aligned}
\theta &= \mathbb{E}\left(C_1\right) - \mathbb{E}\left(C_0\right) \\
&= \mathbb{E}\left(C_1 \,|\, X = 1\right) \mathbb{P}\left(X = 1\right) + \mathbb{E}\left(C_1 \,|\, X = 0\right) \mathbb{P}\left(X = 0\right) \\
&- \mathbb{E}\left(C_0 \,|\, X = 1\right) \mathbb{P}\left(X = 1\right) - \mathbb{E}\left(C_0 \,|\, X = 0\right) \mathbb{P}\left(X = 0\right) \\
&= \underbrace{\mathbb{E}\left(Y \,|\, X = 1\right) \mathbb{P}\left(X = 1\right) - \mathbb{E}\left(Y \,|\, X = 0\right) \mathbb{P}\left(X = 0\right)}_{=:\beta} \\
&+ \underbrace{\mathbb{E}\left(C_1 \,|\, X = 0\right) \mathbb{P}\left(X = 0\right) - \mathbb{E}\left(C_0 \,|\, X = 1\right) \mathbb{P}\left(X = 1\right)}_{=:\gamma}.
\end{aligned}$$

In particular, since Y is binary we may write

$$\begin{aligned}
\beta &= \mathbb{P}\left(Y = 1 \,|\, X = 1\right) \mathbb{P}\left(X = 1\right) - \mathbb{P}\left(Y = 1 \,|\, X = 0\right) \mathbb{P}\left(X = 0\right) \\
&= \mathbb{P}\left(Y = 1,\, X = 1\right) - \mathbb{P}\left(Y = 1,\, X = 0\right),
\end{aligned}$$

and so β is the contribution to the causal effect θ which may be consistently estimated in terms of (X_i, Y_i). The term γ is the contribution to the causal effect θ which *cannot* be estimated since it involves

$$C_1 \,|\, X = 0 \text{ and } C_0 \,|\, X = 1,$$

which are never observed with the sample (X_i, Y_i). Nonetheless we can find *bounds* on γ which may be estimated since C_0 and C_1 are binary and so

$$\begin{aligned}
\gamma &= \mathbb{P}\left(C_1 = 1 \,|\, X = 0\right) \mathbb{P}\left(X = 0\right) - \mathbb{P}\left(C_0 = 1 \,|\, X = 1\right) \mathbb{P}\left(X = 1\right) \\
&= \mathbb{P}\left(C_1 = 1,\, X = 0\right) - \mathbb{P}\left(C_0 = 1,\, X = 1\right).
\end{aligned}$$

Therefore

$$\begin{aligned}-\mathbb{P}\left(X=1\right) &\leqslant \mathbb{P}\left(C_0=1,\, X=1\right)\\ &\leqslant \gamma\\ &\leqslant \mathbb{P}\left(C_1=1,\, X=0\right)\\ &\leqslant \mathbb{P}\left(X=0\right)\end{aligned}$$

and so, since $\theta=\beta+\gamma$, we deduce that

$$\underbrace{\beta-\mathbb{P}\left(X=1\right)}_{=:\theta_l}\leqslant\theta\leqslant\underbrace{\beta+\mathbb{P}\left(X=0\right)}_{=:\theta_u}.$$

In particular these bounds have unit width as expected since

$$\theta_u-\theta_l=\mathbb{P}\left(X=0\right)+\mathbb{P}\left(X=1\right)=1.$$

Finally we record estimators for θ_l and θ_u which are guaranteed to be consistent by the Weak Law of Large Numbers. We define

$$\begin{aligned}\hat{\theta}_l &:= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{1}\left(Y_i=1,\, X_i=1\right)-\mathbb{1}\left(Y_i=1,\, X_i=0\right)-\mathbb{1}\left(X_i=1\right)\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\left[Y_iX_i-Y_i(1-X_i)-X_i\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(2Y_iX_i-Y_i-X_i\right)\\ &= \frac{-1}{n}\sum_{i=1}^{n}\left(X_i-Y_i\right)^2\\ &= \frac{-1}{n}\sum_{i=1}^{n}\left|X_i-Y_i\right|\end{aligned}$$

and

$$\begin{aligned}\hat{\theta}_u &:= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{1}\left(Y_i=1,\, X_i=1\right)-\mathbb{1}\left(Y_i=1,\, X_i=0\right)+\mathbb{1}\left(X_i=0\right)\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\left[Y_iX_i-Y_i(1-X_i)+(1-X_i)\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(2Y_iX_i-Y_i-X_i+1\right)\\ &= \hat{\theta}_l+\frac{1}{n}\sum_{i=1}^{n}1\\ &= \hat{\theta}_l+1.\end{aligned}$$

Unfortunately these bounds are of limited practical value since they do not enable us to test for the sign of θ. Indeed, from the estimate $\hat{\theta}_l$ we will only be able to construct a confidence interval of the form

$$(\hat{\theta}_l-\delta_1,\, \hat{\theta}_l+\delta_2)$$

for $\delta_1,\,\delta_2 > 0$, which leads to a confidence interval for $\theta_u = 1 + \theta_l$ of the form

$$(\hat{\theta}_u - \delta_1,\, \hat{\theta}_u + \delta_2).$$

Here is the issue: $-1 \leqslant \hat{\theta}_l \leqslant 0 \leqslant \hat{\theta}_u \leqslant 1$ and so the confidence interval

$$(\hat{\theta}_l - \delta_1,\, \hat{\theta}_u + \delta_2)$$

for θ will always contain the open interval $(-\delta_1,\, \delta_2)$, which is an open interval about 0. In other words: we cannot construct a confidence interval for θ which excludes zero! This means that we are not able to use $\hat{\theta}_l$ and $\hat{\theta}_u$ to offer statistical evidence that θ is nonzero, i.e. we are not able to offer statistical evidence that there is a causal effect (in one direction or another).

Exercise A.16.4 (Association and causation are not related – General case)**.** Suppose that $X \in \mathbb{R}$ and that, for each subject i, $C_i(x) = \beta_{1,i}x$. Each subject has their own slope $\beta_{1,i}$. Construct a joint distribution on $(\beta_1,\, X)$ such that $\mathbb{P}\left(\beta_1 > 0\right) = 1$ but $\mathbb{E}\left(Y \,|\, X = x\right)$ is a decreasing function of x, where $Y = C(X)$. Interpret.

Solution. We may choose X to be *any* random variable with codomain $(0,\, \infty)$ and then define

$$\beta_1 := \frac{\gamma(X)}{X}$$

for *any* strictly decreasing function $\gamma : (0,\, \infty) \to (0,\, \infty)$. Then indeed, since $\gamma > 0$ everywhere,

$$\mathbb{P}\left(\beta_1 > 0\right) = \mathbb{P}\left(\frac{\gamma(X)}{X} > 0\right) = \mathbb{P}\left(X > 0\right) = 1$$

as desired while

$$\begin{aligned}\mathbb{E}\left(Y \,|\, X = x\right) = \mathbb{E}\left[C(X) \,|\, X = x\right] &= \mathbb{E}\left(\beta_1 X \,|\, X = x\right)\\ &= \mathbb{E}\left[\frac{\gamma(X)}{X} \cdot X \,\middle|\, X = x\right] = \gamma(x),\end{aligned}$$

which is indeed decreasing.

The key is that, altough the slope β_1 is positive for every subject, meaning that there is a *positive* causal effect, that slope is *not* independent from the covariate X, leading to a *negative* association. More specifically: the slope β_1 decreases sufficiently fast, as a function of the covariate X, to make it so that the association is negative, i.e. the regression function is strictly decreasing, even though there is a positive causal effect.

More specifically we may choose $X \sim$ Exponential(1) and $\beta_1 := 1/X^2$ (i.e. choosing $\gamma(x) = 1/x$). Then $\beta_1 > 0$ always and yet

$$\mathbb{E}\left(Y \,|\, X = x\right) = \mathbb{E}\left(\beta_1 X \,|\, X = x\right) = \mathbb{E}\left(\frac{1}{X} \,\middle|\, X = x\right) = \frac{1}{x},$$

which is indeed strictly decreasing.

Exercise A.16.5 (Estimation of the median causal effect)**.** Let $X \in \{0,\, 1\}$ be a binary treatment variable and let $(C_0,\, C_1)$ denote the corresponding potential outcomes. Let $Y = C_X$ denote the observed response. Let F_0 and F_1 be the cumulative distribution functions for C_0 and C_1. Assume that F_0 and F_1 are both continuous and strictly increasing. Let $\theta = m_1 - m_0$ where $m_0 := F_0^{-1}(1/2)$ is the median of C_0 and $m_1 := F_1^{-1}(1/2)$ is the median of C_1. Suppose that the treatment X is assigned randomly. Find an expression for θ involving only the joint distribution of X and Y.

Solution. The key observation is that, since the covariate X and the counterfactual vector (C_0, C_1) are independent (as is implied by the random assignment of X), the consistency relationship $Y = C_X$ tells us that

$$F_1(y) = \mathbb{P}\left(C_1 \leqslant y\right) = \mathbb{P}\left(C_1 \leqslant y \,|\, X = 1\right) = \mathbb{P}\left(Y \leqslant y \,|\, X = 1\right) = F_{Y|X=1}(y),$$

and similarly $F_0 = F_{Y|X=0}$. Therefore

$$\theta = F_1^{-1}\left(\frac{1}{2}\right) - F_0^{-1}\left(\frac{1}{2}\right) = F_{Y\,|\,X=1}^{-1}\left(\frac{1}{2}\right) - F_{Y\,|\,X=0}^{-1}\left(\frac{1}{2}\right),$$

i.e., under random assignment, the median causal effect θ is the difference between the *conditional* medians of Y given X.

A.17. Directed Graphs and Conditional Independence.

Exercise A.17.1 (Alternate characterization of conditional independence)**.** Prove Theorem 17.2 where we record an alternate characterization of conditional independence.

Solution. This ultimately boils down to Remark 1.5 which tells us that two events A and B are independent if and only if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. Here:

$$\begin{aligned} X \amalg Y \mid Z &\iff f_{X,Y|Z}(x,y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z) \\ &\iff \frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)} = \frac{f_{X,Z}(x,z)}{f_Z(z)} \cdot \frac{f_{Y,Z}(y,z)}{f_Z(z)} \\ &\iff \frac{f_{X,Y,Z}(x,y,z)}{f_{Y,Z}(y,z)} = \frac{f_{X,Z}(x,z)}{f_Z(z)} \\ &\iff f_{X|Y,Z}(x|y,z) = f_{X|Z}(x|z), \end{aligned}$$

as desired.

Exercise A.17.2 (Properties of conditional independence)**.** Prove Theorem 17.4 where we record various properties of conditional independence.

Solution.

(1) This is immediate since the defining identity of conditional independence is symmetric in (X,Y). Indeed:

$$\begin{aligned} X \amalg Y \mid Z &\iff f_{X,Y|Z}(x,y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z) \\ &\iff f_{Y,X|Z}(y,x|z) = f_{Y|Z}(y|z) f_{X|Z}(x|z) \\ &\iff Y \amalg X \mid Z. \end{aligned}$$

(2) This follows from Theorem 17.3 since, if $X \amalg Y \mid Z$ then

$$\begin{aligned} \mathbb{P}(\mathcal{U} \in A,\, Y \in B \mid Z = z) &= \mathbb{P}\left(X \in h^{-1}(A),\, Y \in B \,\middle|\, Z = z\right) \\ &= \mathbb{P}\left(X \in h^{-1}(A) \,\middle|\, Z = z\right) \mathbb{P}(Y \in B \mid Z = z) \\ &= \mathbb{P}(\mathcal{U} \in A \mid Z = z)\, \mathbb{P}(Y \in B \mid Z = z) \end{aligned}$$

for all events A and B and every $z \in \mathbb{R}$, which proves that $\mathcal{U} \amalg Y \mid Z$.

(3) Proceeding as in Exercise A.23.18 we see that

$$\mathbb{P}(X \in A,\, Y \in B \mid \mathcal{U} = u,\, Z = z) = \frac{\mathbb{P}(X \in A,\, Y \in B,\, \mathcal{U} = u \mid Z = z)}{\mathbb{P}(\mathcal{U} = u,\, Z = z)}$$

where

$$\mathcal{U} = u \iff h(X) = u \iff X \in h^{-1}(u)$$

and so, since $X \amalg Y \mid Z$, Theorem 17.3 tells us that

$$\begin{aligned} &\mathbb{P}(X \in A,\, Y \in B \mid \mathcal{U} = u,\, Z = z) \\ &= \frac{\mathbb{P}\left(X \in A \cap h^{-1}(u),\, Y \in B \,\middle|\, Z = z\right)}{\mathbb{P}(\mathcal{U} = u,\, Z = z)} \\ &= \frac{\mathbb{P}\left(X \in A \cap h^{-1}(u) \,\middle|\, Z = z\right) \mathbb{P}(Y \in B \mid Z = z)}{\mathbb{P}(\mathcal{U} = u,\, Z = z)} \\ &= \mathbb{P}(X \in A \mid \mathcal{U} = u,\, Z = z)\, \mathbb{P}(Y \in B \mid Z = z)\,. \end{aligned}$$

In particular, for $A' := h^{-1}(u)$ we may use Exercise A.23.18 and Theorem 17.3 once again to compute that

$$\begin{aligned}\mathbb{P}\left(Y \in B \,|\, X \in A',\, Z = z\right) &= \frac{\mathbb{P}\left(Y \in B,\, X \in A' \,|\, Z = z\right)}{\mathbb{P}\left(X \in A' \,|\, Z = z\right)} \\ &= \mathbb{P}\left(Y \in B \,|\, Z = z\right)\end{aligned}$$

which verifies that

$$\begin{aligned}&\mathbb{P}\left(X \in A,\, Y \in B \,|\, \mathcal{U} = u,\, Z = z\right) \\ &= \mathbb{P}\left(X \in A \,|\, \mathcal{U} = u,\, Z = z\right) \mathbb{P}\left(Y \in B \,|\, X \in A',\, Z = z\right) \\ &= \mathbb{P}\left(X \in A \,|\, \mathcal{U} = u,\, Z = z\right) \mathbb{P}\left(Y \in B \,|\, \mathcal{U} = u,\, Z = z\right),\end{aligned}$$

i.e. indeed $X \amalg Y \,|\, (\mathcal{U},\, Z)$.

(4) We proceed as in Exercise A.23.19, with additional conditioning on Z everywhere. In particular we use a PDF-version of Exercise A.23.18, namely the fact that

$$f_{E,F|G} = f_{E|F,G} f_{F|G}$$

for E, F, and G *random variables* (and not *events* as in Exercise A.23.18). As in Exercise A.23.19, for simplicity we omit the arguments of the PDFs below and compute:

$$\begin{aligned}f_{X,W,Y|Z} &= f_{X,W|Y,Z} f_{Y|Z} \\ &= f_{X|Y,Z} f_{W|Y,Z} f_{Y|Z} \\ &= f_{X|Z} f_{W|Y,Z} f_{Y|Z} \\ &= f_{X|Z} f_{W,Y|Z},\end{aligned}$$

which proves that $X \amalg (W,\, Y) \,|\, Z$.

(5) Since $X \amalg Z \,|\, Y$ we have that

$$f_{X,Y,Z}(x, y, z) = f_{X|Y,Z}(x|y, z) f_{Y,Z}(y, z) = f_{X|Y}(x|y) f_{Y,Z}(y, z)$$

where, since $X \amalg Y \,|\, Z$,

$$f_{X|Y}(x|y) = f_{X|Y,Z}(x|y, z) f_Z(z) = f_{X|Z}(x|z) f_Z(z) = f_{X,Z}(x, z)$$

and so, since this holds for *all* z,

$$f_{X|Y}(x|y) = \int f_{X,Z}(x, z) dz = f_X(x).$$

Putting it all together we conclude that

$$f_{X,Y,Z} = f_{X|Y} f_{Y,Z} = f_X f_{Y,Z},$$

which verifies that $X \amalg (Y,\, Z)$, as desired.

Exercise A.17.3 (Conditional independence without independence)**.** Let X, Y, and Z have the following joint distribution:

$Z = 0$	$Y = 0$	$Y = 1$
$X = 0$	0.405	0.045
$X = 1$	0.045	0.005

and

$Z = 1$	$Y = 0$	$Y = 1$
$X = 0$	0.125	0.125
$X = 1$	0.125	0.125

(1) Find the conditional distribution of X and Y given $Z = 0$ and the conditional distribution of X and Y given $Z = 1$.
(2) Show that $X \amalg Y \mid Z$.
(3) Find the marginal distribution of X and Y.
(4) Show that X and Y are not marginally independent.

Solution. (1) We compute that

$$f_{X,Y|Z}(x, y|0) = \frac{f_{X,Y,Z}(x, y, 0)}{f_Z(0)}$$

where

$$f_Z(0) = 0.405 + 0.045 + 0.045 + 0.005 = 0.5$$

and so the conditional distribution of (X, Y) given $Z = 0$ is

	$Y = 0$	$Y = 1$
$X = 0$	$\frac{0.405}{0.5} = \frac{81}{100}$	$\frac{0.045}{0.5} = \frac{9}{100}$
$X = 1$	$\frac{0.045}{0.5} = \frac{9}{100}$	$\frac{0.005}{0.5} = \frac{1}{100}$

Similarly, since

$$f_Z(1) = 0.125 + 0.125 + 0.125 + 0.125 = 0.5$$

(or simply $f_Z(1) = 1 - f_Z(0) = 0.5$, since Z is binary), the conditional distribution of (X, Y) given $Z = 1$ is

	$Y = 0$	$Y = 1$
$X = 0$	$\frac{0.125}{0.5} = \frac{1}{4}$	$\frac{0.125}{0.5} = \frac{1}{4}$
$X = 1$	$\frac{0.125}{0.5} = \frac{1}{4}$	$\frac{0.125}{0.5} = \frac{1}{4}$

(2) Since

$$f_{X|Z}(x|z) = \sum_{y=0,1} f_{X,Y|Z}(x, y|z) \text{ and } f_{Y|Z}(y|z) = \sum_{x=0,1} f_{X,Y|Z}(x, y|z),$$

inspecting the tables recording the conditional distribution of (X, Y) given Z tells us immediately that

$$f_{X,Y|Z} = f_{X|Z} f_{Y|Z}$$

since

$Z = 0$	$f_{Y|Z}(0) = \frac{81+9}{100} = \frac{9}{10}$	$f_{Y|Z}(1) = \frac{9+1}{100} = \frac{1}{10}$
$f_{X|Z}(0) = \frac{81+9}{100} = \frac{9}{10}$	$\frac{9}{10} \cdot \frac{9}{10} = \frac{81}{100}$	$\frac{9}{10} \cdot \frac{1}{10} = \frac{9}{100}$
$f_{X|Z}(1) = \frac{9+1}{100} = \frac{1}{10}$	$\frac{1}{10} \cdot \frac{9}{10} = \frac{9}{100}$	$\frac{1}{10} \cdot \frac{1}{10} = \frac{1}{100}$

while

$Z = 1$	$f_{Y\|Z}(0) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$	$f_{Y\|Z}(1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$
$f_{X\|Z}(0) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$
$f_{X\|Z}(1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

such that indeed, by comparing the two tables depicting $f_{X|Z} f_{Y|Z}$ immediately above with the two tables depicting $f_{X,Y|Z}$ in item 1, we see that

$$f_{X|Z} f_{Y|Z} = f_{X,Y|Z}$$

for both $Z = 0$ and $Z = 1$. This verifies that X and Y are conditionally independent given Z.

(3) Since

$$f_{X,Y}(x, y) = \sum_{z=0,1} f_{X,Y,Z}(x, y, z)$$

the marginal distribution of (X, Y) is

	$Y = 0$	$Y = 1$
$X = 0$	$0.405 + 0.125 = 0.53$	$0.045 + 0.125 = 0.17$
$X = 1$	$0.045 + 0.125 = 0.17$	$0.005 + 0.125 = 0.13$

(4) Since

$$f_X(x) = \sum_{y=0,1} f_{X,Y}(x, y) \text{ and } f_Y(y) = \sum_{x=0,1} f_{X,Y}(x, y)$$

we compute the marginal distributions of X and Y to be

	$f_Y(0) = 0.53 + 0.17 = 0.7$	$f_Y(1) = 0.17 + 0.13 = 0.3$
$f_X(0) = 0.53 + 0.17 = 0.7$		
$f_X(1) = 0.17 + 0.13 = 0.3$		

However

$$f_X f_Y \neq f_{X,Y}$$

since the table for $f_X(x) f_Y(y)$ is

	$y = 0$	$y = 1$
$x = 0$	$0.7 \cdot 0.7 = 0.49$	$0.7 \cdot 0.3 = 0.21$
$x = 1$	$0.2 \cdot 0.7 = 0.21$	$0.3 \cdot 0.3 = 0.09$

which differs *everywhere* from the table for $f_{X,Y}(x, y)$ recorded in item 3. This means that indeed X and Y are *not* independent.

Exercise A.17.4 (Applying the Markov condition to simple DAGs)**.** Consider the three DAGs

$$X \longrightarrow Y \longrightarrow Z \qquad X \longleftarrow Y \longleftarrow Z \qquad X \longleftarrow Y \longrightarrow Z$$

Prove that $X \perp\!\!\!\perp Z \,|\, Y$.

Solution. First we consider

$$X \longrightarrow Y \longrightarrow Z.$$

In this case

$$\pi_Z = \{Y\} \text{ and } \widetilde{Z} = \{X\}$$

so the Markov condition for Z tells us that

$$Z \amalg X \,|\, Y.$$

Now we consider

$$X \longleftarrow Y \longleftarrow Z.$$

We proceed as above, considering now the Markov condition for X which tells us that $X \amalg Z \,|\, Y$. Finally we consider

$$X \longleftarrow Y \longrightarrow Z.$$

Now we have that

$$\pi_X = \{Y\} \text{ and } \widetilde{X} = \{Z\}$$

and so the Markov condition for X tells us that

$$X \amalg Z \,|\, Y.$$

(Note that here we could have also used the Markov condition for Z to deduce the same conditional independence.)

Exercise A.17.5 (Applying the d–separation characterization to a simple DAG)**.** Consider the faithful DAG

$$X \longrightarrow Y \longleftarrow Z.$$

Prove that $X \amalg Z$ and $X \not\amalg Z \,|\, Y$.

Solution. By the Markov condition (Theorem 17.11), since

$$\pi_X = \emptyset \text{ while } \widetilde{X} = \{W\},$$

we deduce that indeed $X \amalg Z$. Now Theorem 17.17 tells us that

$$X \amalg Z \,|\, Y \iff X \text{ and } Z \text{ are d–separated given } Y.$$

Since $\mathcal{U} = (X, Y, Z)$ is an undirected path from X to Z which contains the collider Y, it witnesses the d–connectedness of X and Z given Y. Therefore X and Z are conditionally *dependent* given Y, as desired.

Exercise A.17.6 (DAG estimation)**.** Consider some random variables $X \in \{0, 1\}$, $Y \in \{0, 1\}$, and $Z \in \{0, 1, 2\}$, Suppose the distribution of (X, Y, Z) is Markov to:

$$X \to Y \to Z.$$

Create a joint distribution $f(x, y, z)$ that is Markov to this DAG. Generate 1000 random vectors from this distribution. Estimate the distribution from the data using maximum likelihood. Compare the estimated distribution to the true distribution. Let $\theta = (\theta_{000}, \theta_{001}, \ldots, \theta_{112})$ where $\theta_{rst} = \mathbb{P}(X = r, Y = s, Z = t)$. Use the bootstrap to get standard errors and 95 percent confidence intervals for these 12 parameters.

Solution. A distribution Markov to

$$X \to Y \to Z$$

must have the form

$$f(x,\, y,\, z) = f(z \,|\, y) f(y \,|\, x) f(x).$$

For the sake of cleaner notation let us write

$$g(x; p) = p\mathbb{1}(x = 1) + (1 - p)\mathbb{1}(x = 0) \text{ for } p \in [0,\, 1]$$

and

$$h(x; p) = p_0\mathbb{1}(x = 0) + p_1\mathbb{1}(x = 1) + p_2\mathbb{1}(x = 2) \text{ for } p \in \Delta^2 \subseteq \mathbb{R}^3,$$

where we use $(e_0,\, e_1,\, e_2)$ as the standard ordered basis of $\mathbb{R}^3$, such that

- $g(\,\cdot\,; p)$ is the PDF of a Bernoulli(p) random variable and
- $h(\,\cdot\,; p)$ is the PDF of a Categorical(p) random variable transformed by $e_i \mapsto i$ for $i = 0,\, 1,\, 2$.

Since $X,\, Y \in \{0,\, 1\}$ and $Z \in \{0,\, 1,\, 2\}$ we may *without any additional assumptions* write the joint PDF of $(X,\, Y,\, Z)$ parametrically in terms of g and h as

$$f(x,\, y,\, z) = f(x,\, y,\, z; q) = f_{Z|Y}(z \,|\, y; q_Z) f_{Y|X}(y \,|\, x; q_Y) f_X(x; q_X)$$

where $q = (q_X,\, q_Y,\, q_Z) \in [0,\, 1] \times [0,\, 1]^2 \times \left(\Delta^2\right)^2$ such that

$$\begin{aligned}
f_X(\,\cdot\,; q_X) &= g(\,\cdot\,; q_X), \\
f_{Y|X}(\,\cdot \,|\, x; q_Y) &= \begin{cases} g(\,\cdot\,; q_{Y,0}) & \text{if } x = 0 \text{ and} \\ g(\,\cdot\,; q_{Y,1}) & \text{if } x = 1, \text{ and} \end{cases} \\
f_{Z|Y}(\,\cdot \,|\, y; q_Z) &= \begin{cases} h(\,\cdot\,; q_{Z,0}) & \text{if } y = 0 \text{ and} \\ h(\,\cdot\,; q_{Z,1}) & \text{if } y = 1. \end{cases}
\end{aligned}$$

The likelihood function is then

$$\mathcal{L}(q) = \prod_{i=1}^{n} f_{Z|Y}\left(Z_i \,|\, Y_i; q_Z\right) f_{Y|X}\left(Y_i \,|\, X_i; q_Y\right) f_X\left(X_i; q_X\right)$$

and so the log-likelihood function takes the form

$$\begin{aligned}
l(q) &= \sum_{i=1}^{n} \log f_{Z|Y}\left(Z_i \,|\, Y_i; q_Z\right) + \sum_{i=1}^{n} \log f_{Y|X}\left(Y_i \,|\, X_i; q_Y\right) + \sum_{i=1}^{n} \log f_X\left(X_i; q_X\right) \\
&= \sum_{i=1}^{n} \left[\mathbb{1}(Y_i = 0) \log h\left(Z_i; q_{Z,0}\right) + \mathbb{1}(Y_i = 1) \log h\left(Z_i; q_{Z,1}\right)\right] \\
&\quad + \sum_{i=1}^{n} \left[\mathbb{1}(X_i = 0) \log g\left(Y_i; q_{Y,0}\right) + \mathbb{1}(X_i = 1) \log g\left(Y_i; q_{Y,1}\right)\right] \\
&\quad + \sum_{i=1}^{n} \log g\left(X_i; q_X\right) \\
&=: l_{h,0}(q_{Z,0}) + l_{h,1}(q_{Z,1}) + l_{g,0}(q_{Y,0}) + l_{g,1}(q_{Y,1}) + l_g(q_X).
\end{aligned}$$

Crucially: the log-likelihood splits as a sum of five terms each depending on a *different* component of q. This means that we can maximize the log-likelihood by maximizing each term separately.

Since

$$l_g(q_X) = \sum_{i=1}^{n} \log g\left(X_i; q_X\right),$$

which is the log-likelihood corresponding to an IID sample $X_1, \ldots, X_n$ drawn from a Bernoulli(q_X) distribution, we know that the MLE for q_X is the MLE of the Bernoulli model, i.e. the sample mean

$$\hat{q}_X := \overline{X}.$$

Similarly, since

$$l_{g,0}(q_{Y,0}) = \sum_{i=1}^{n} \mathbb{1}(X_i = 0) \log g\left(Y_i; q_{Y,0}\right) = \sum_{j=1}^{n_{X,0}} \log g\left(Y_{i_j}; q_{Y,0}\right),$$

where

$$n_{X,0} := \#\left\{i : X_i = 0\right\} = \sum_{i=1}^{n} \mathbb{1}(X_i = 0)$$

and where Y_{i_j} is the subsequence of Y_i characterized by $X_{i_j} = 0$, we deduce in the same way that the MLE for $q_{Y,0}$ is the *conditional* sample mean

$$\hat{q}_{Y,0} := \frac{1}{n_{X,0}} \sum_{j=1}^{n_{X,0}} Y_{i_j} := \frac{1}{n_{X,0}} \sum_{i=1}^{n} \mathbb{1}(X_i = 0) Y_i.$$

In exactly the same way we deduce that the MLE for $q_{Y,1}$ is

$$\hat{q}_{Y,1} := \frac{1}{n_{X,q}} \sum_{i=1}^{n} \mathbb{1}(X_1 = 1) Y_i \text{ where } n_{X,1} = \sum_{i=1}^{n} \mathbb{1}(X_i = 1).$$

Finally, since

$$l_{h,0} = \sum_{i=1}^{n} \mathbb{1}(Y_i = 0) \log h\left(Z_i; q_{Z,0}\right) = \sum_{j=1}^{n_{Y,0}} \log h\left(Z_{i_j}; q_{Z,0}\right),$$

where

$$n_{Y,0} := \#\left\{i : Y_i = 0\right\} = \sum_{i=1}^{n} \mathbb{1}(Y_i = 0)$$

and where Z_{i_j} is the subsequence of Z_i characterized by $Y_{i_j} = 0$, we observe that $l_{h,0}$ is the log-likelihood corresponding to the IID sample

$$Z_{i_1}, \ldots, Z_{i_{n_{Y,0}}}$$

drawn from $h(\,\cdot\,; q_{Z,0})$, a transformed Categorical$(q_{Z,0})$ distribution. We computed in Exercises A.23.13 and A.23.14 the MLE of the Categorical model. Since h is a known transformation of a Categorical distribution we deduce that if

$$W_1, \ldots, W_n \sim h(\,\cdot\, p) \text{ IID}$$

then the MLE for $p \in \Delta^2$ is

$$\left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(W_i = 0),\ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(W_i = 1),\ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(W_i = 2)\right).$$

So here the MLE for $q_{Z,0}$ is

$$\hat{q}_{Z,0} := \begin{pmatrix} \frac{1}{n_{Y,0}} \sum_{j=1}^{n_{Y,0}} \mathbb{1}(Z_{i_j} = 0) \\ \\ \frac{1}{n_{Y,0}} \sum_{j=1}^{n_{Y,0}} \mathbb{1}(Z_{i_j} = 1) \\ \\ \frac{1}{n_{Y,0}} \sum_{j=1}^{n_{Y,0}} \mathbb{1}(Z_{i_j} = 2) \end{pmatrix} = \begin{pmatrix} \frac{1}{n_{Y,0}} \sum_{i=1}^{n} \mathbb{1}(Y_i = 0)\mathbb{1}(Z_{i_j} = 0) \\ \\ \frac{1}{n_{Y,0}} \sum_{i=1}^{n} \mathbb{1}(Y_i = 0)\mathbb{1}(Z_{i_j} = 1) \\ \\ \frac{1}{n_{Y,0}} \sum_{i=1}^{n} \mathbb{1}(Y_i = 0)\mathbb{1}(Z_{i_j} = 2) \end{pmatrix}.$$

In exactly the same way we deduce that the MLE for $q_{Z,1}$ is

$$\hat{q}_{Z,1} := \begin{pmatrix} \frac{1}{n_{Y,1}} \sum_{i=1}^{n} \mathbb{1}(Y_i = 1)\mathbb{1}(Z_{i_j} = 0) \\ \\ \frac{1}{n_{Y,1}} \sum_{i=1}^{n} \mathbb{1}(Y_i = 1)\mathbb{1}(Z_{i_j} = 1) \\ \\ \frac{1}{n_{Y,1}} \sum_{i=1}^{n} \mathbb{1}(Y_i = 1)\mathbb{1}(Z_{i_j} = 2) \end{pmatrix} \text{ where } n_{Y,1} := \sum_{i=1}^{n} \mathbb{1}(Y_i = 1).$$

We now turn our attention to θ. The plug-in estimator for θ is

$$\hat{\theta}_{rst} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i = r)\mathbb{1}(Y_i = s)\mathbb{1}(Z_i = z).$$

Since all the MLE for components of q recorded above are also plug-in estimators (this is not guaranteed in general, but happens to be true here), we can actually relate the MLE for q to the plug-in estimator for θ. Indeed, since

$$\mathbb{P}\left(Z = 0 \,|\, Y = 0\right) = \frac{\mathbb{P}\left(Z = 0,\, Y = 0\right)}{\mathbb{P}\left(Y = 0\right)} = \frac{\theta_{\cdot 00}}{\theta_{\cdot 0 \cdot}},$$

where dotted subscripts denote sums as in Definition 15.3, with similar identities when $Z = 1$ and $Z = 2$, or when $Y = 1$, we observe that

$$\begin{aligned} q_{Z,0} &= \left(\mathbb{P}\left(Z = 0 \,|\, Y = 0\right), \mathbb{P}\left(Z = 1 \,|\, Y = 0\right), \mathbb{P}\left(Z = 2 \,|\, Y = 0\right)\right) \\ &= \frac{1}{\theta_{\cdot 0 \cdot}} \left(\theta_{\cdot 00},\, \theta_{\cdot 01},\, \theta_{\cdot 02}\right) \end{aligned}$$

while

$$q_{Z,1} = \frac{1}{\theta_{\cdot 1 \cdot}} \left(\theta_{\cdot 10},\, \theta_{\cdot 11},\, \theta_{\cdot 12}\right).$$

Similarly:

$$q_{Y,0} = \frac{\theta_{01\cdot}}{\theta_{0\cdot\cdot}},\; q_{Y,1} = \frac{\theta_{11\cdot}}{\theta_{1\cdot\cdot}}, \text{ and } q_X = \frac{\theta_{1\cdot\cdot}}{\theta_{\cdot\cdot\cdot}},$$

where $\theta_{\cdot\cdot\cdot} = 1$. Therefore the MLE for q is related to the plug-in estimator for θ in the same way! To make the ensuing identities (even) easier to evaluate numerically we define

$$N_{rst} := \sum_{i=1}^{n} \mathbb{1}(X_i = r)\mathbb{1}(Y_i = s)\mathbb{1}(Z_i = z),$$

such that

$$\hat{\theta}_{rst} = \frac{N_{rst}}{n}.$$

Then the MLE for the components of q may be written as follows:

$$\begin{aligned} &q_X = \frac{N_{1\cdot\cdot}}{n},\ q_{Y,0} = \frac{N_{01\cdot}}{N_{0\cdot\cdot}},\ q_{Y,1} = \frac{N_{11\cdot}}{N_{1\cdot\cdot}}, \\ &q_{Z,0} = \frac{1}{N_{\cdot 0\cdot}}\left(N_{\cdot 00},\, N_{\cdot 01},\, N_{\cdot 02}\right),\ \text{and } q_{Z,1} = \frac{1}{N_{\cdot 1\cdot}}\left(N_{\cdot 10},\, N_{\cdot 11},\, N_{\cdot 12}\right). \end{aligned}$$

When implementing this numerically, as a sanity check, what do we expect? Recall that, since $X \to Y \to Z$ is a representation of the distribution,

$$\theta_{rst} = \mathbb{P}\left(X = r,\, Y = s,\, Z = t\right) = \mathbb{P}\left(Z = t \,|\, Y = s\right) \mathbb{P}\left(Y = s \,|\, X = r\right) \mathbb{P}\left(X = r\right).$$

To make this easier to relate to q let us introduce

$$\bar{q} = \left(\bar{q}_X,\, \bar{q}_Y,\, \bar{q}_Z\right) \in \Delta^1 \times \left(\Delta^1\right)^2 \times \left(\Delta^2\right)^2$$

via, for $j = 0, 1$,

$$\begin{cases} \bar{q}_X = \left(1 - q_X,\, q_X\right), \\ \bar{q}_{Y,j} = \left(1 - q_{Y,j},\, q_{Y,j}\right),\ \text{and} \\ \bar{q}_{Z,j} = q_{Z,j} \end{cases}$$

such that

$$\begin{cases} \mathbb{P}\left(X = r\right) = \bar{q}_{X,r}, \\ \mathbb{P}\left(Y = s \,|\, X = r\right) = \bar{q}_{Y,r,s},\ \text{and} \\ \mathbb{P}\left(Z = t \,|\, Y = s\right) = \bar{q}_{Z,s,t}. \end{cases}$$

Therefore

$$\theta_{rst} = \bar{q}_{X,r}\bar{q}_{Y,r,s}\bar{q}_{Z,s,t}.$$

Say we choose

$$\begin{aligned} &q_X = 0.3, \\ &q_{Y,0} = 0.6 \text{ and } q_{Y,1} = 0.8,\ \text{and} \\ &q_{Z,0} = (0.5, 0.2, 0.3) \text{ and } q_{Z,1} = (0.25, 0.65, 0.1), \end{aligned}$$

which is equivalent to

$$\begin{aligned} &\bar{q}_X = (0.7,\, 0.3), \\ &\bar{q}_Y = ((0.4, 0.6), (0.2, 0.8)),\ \text{and} \\ &\bar{q}_Z = ((0.5, 0.2, 0.3), (0.25, 0.65, 0.1)). \end{aligned}$$

Note that we have only *seven* free parameters to choose, not twelve, since being Markov to $X \to Y \to Z$ imposes some conditional independence relations on (X, Y, Z). Then the values of θ_{rst} are

$Z = t = 0$	$Y = s = 0$	$Y = s = 1$
$X = r = 0$	$\theta_{000} = 0.7 \times 0.4 \times 0.5 = 0.140$	$\theta_{010} = 0.7 \times 0.6 \times 0.25 = 0.105$
$X = r = 1$	$\theta_{100} = 0.3 \times 0.2 \times 0.5 = 0.030$	$\theta_{110} = 0.3 \times 0.8 \times 0.25 = 0.060$

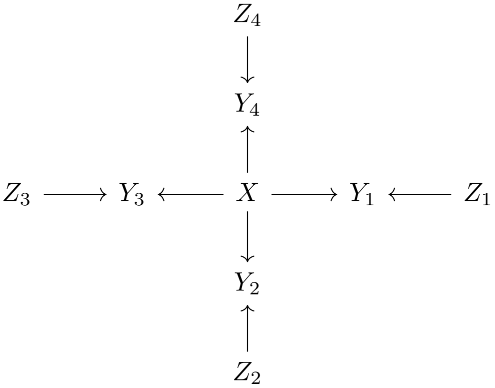
as well as

$Z = 1$	$Y = 0$	$Y = 1$
$X = 0$	$\theta_{001} = 0.7 \times 0.4 \times 0.2 = 0.056$	$\theta_{011} = 0.7 \times 0.6 \times 0.65 = 0.273$
$X = 1$	$\theta_{101} = 0.3 \times 0.2 \times 0.2 = 0.012$	$\theta_{111} = 0.3 \times 0.8 \times 0.65 = 0.156$

and

$Z = 2$	$Y = 0$	$Y = 1$
$X = 0$	$\theta_{002} = 0.7 \times 0.4 \times 0.3 = 0.084$	$\theta_{012} = 0.7 \times 0.6 \times 0.1 = 0.042$
$X = 1$	$\theta_{102} = 0.3 \times 0.2 \times 0.3 = 0.018$	$\theta_{112} = 0.3 \times 0.8 \times 0.1 = 0.024$

Exercise A.17.7 (d–separation for more complicated graphs)**.** Consider the DAG



(1) Write down the factorization of the joint density.
(2) Prove that $X \amalg Z_j$ for all j.

Solution. (1) Applying Definition 17.10 we see that the joint density is

$$f(x) \cdot \prod_{j=1}^{4} f(z_j) f\left(y_j \,|\, x,\, z_j\right).$$

(2) For each j we see immediately that $\pi_{Z_j} = \emptyset$ and $\widetilde{Z}_j = \{X_j\}$ and so the Markov condition for Z_j reads $Z_j \amalg X$, as desired. Alternatively we may compute directly, e.g. for $j = 1$, that

$$\begin{aligned} f(x,\, z_1) &= \int f(x) \prod_{j=1}^{4} f(z_j) f(y_j \,|\, x,\, z_j) dy_1 \dots dy_4 dz_2 \dots dz_4 \\ &= \int f(x) f(z_1) f(z_2) f(z_3) f(z_4) dz_2 dz_3 dz_4 \\ &= f(x) f(z_1). \end{aligned}$$

Exercise A.17.8 (Intervention – Binary case)**.** Let $V = (X,\, Y,\, Z)$ have the following joint distribution

$$\begin{aligned} &X \sim \text{Bernoulli}\left(\frac{1}{2}\right) \\ &Y \,|\, X = x \sim \text{Bernoulli}\left(\frac{e^{4x-2}}{1 + e^{4x-2}}\right) \\ &Z \,|\, X = x,\, Y = y \sim \text{Bernoulli}\left(\frac{e^{2(x+y)-2}}{1 + e^{2(x+y)-2}}\right). \end{aligned}$$

(1) Find an expression for

$$\mathbb{P}\left(Z = z \,|\, Y = y\right).$$

In particular, find $\mathbb{P}\left(Z=1 \,|\, Y=1\right)$.

(2) Write a program to simulate the model. Conduct a simulation and compute $\mathbb{P}\left(Z=1 \,|\, Y=1\right)$ empirically. Plot this as a function of the simulation size N. It should converge to the theoretical value you computed in (1).

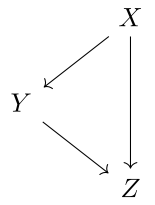
(3) Write down an expression for

$$\mathbb{P}\left(Z=z \,|\, Y:=1\right).$$

In particular, find $\mathbb{P}\left(Z=1 \,|\, Y:=1\right)$.

(4) Modify your program to simulate the intervention "set $Y=1$". Conduct a simulation and compute $\mathbb{P}\left(Z=1 \,|\, Y:=1\right)$ empirically. Plot this as a function of the simulation size N. It should converge to the theoretical value you computed in (3).

Solution. (1) The DAG is



and note that using the logistic function $\sigma(s)=\frac{e^s}{1+e^s}$ we may write

$$\begin{aligned}
&Y \,|\, X=x \sim \text{Bernoulli}\left(\sigma(4x-2)\right) \text{ and}\\
&Z \,|\, X=x,\, Y=y \sim \text{Bernoulli}\left(\sigma(2(x+y)-2)\right).
\end{aligned}$$

In particular, since $X\in\{0,\,1\}$ and since

$$\begin{aligned}
&\sigma(4\cdot 0-2)=\sigma(-2)\approx 0.12 \text{ and}\\
&\sigma(4\cdot 1-2)=\sigma(2)\approx 0.88,
\end{aligned}$$

we have that

$$Y \,|\, X=x \sim \begin{cases}\text{Bernoulli}\left(\sigma(-2)\right) & \text{if } x=0 \text{ and}\\ \text{Bernoulli}\left(\sigma(2)\right) & \text{if } x=1.\end{cases}$$

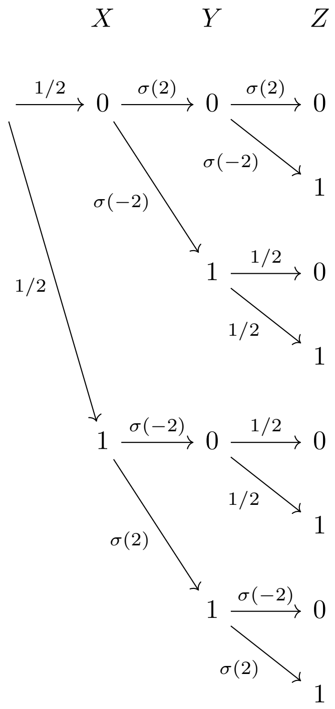
Similarly: $Y\in\{0,\,1\}$ with

$$\begin{aligned}
&\sigma(2\cdot 0-2)=\sigma(-2)\approx 0.12,\\
&\sigma(2\cdot 1-2)=\sigma(0)=\frac{1}{2} \text{ and}\\
&\sigma(2\cdot 2-2)=\sigma(2)\approx 0.88,
\end{aligned}$$

we have that

$$Z \,|\, X=x,\, Y=y \sim \begin{cases}\text{Bernoulli}\left(\sigma(-2)\right) & \text{if } x=y=0,\\ \text{Bernoulli}\left(1/2\right) & \text{if } (x,\,y)=(0,\,1) \text{ or } (1,\,0), \text{ and}\\ \text{Bernoulli}\left(\sigma(2)\right) & \text{if } x=y=1.\end{cases}$$

Since X, Y, and Z are all *binary* random variables we may summarize the situation with the following *binary* tree:



Therefore

$$
\begin{aligned}
&\mathbb{P}\left(Z = z \,|\, Y = y\right) \\
&\quad = \frac{\mathbb{P}\left(Z = z,\, Y = y\right)}{\mathbb{P}\left(Y = y\right)} \\
&\quad = \frac{\mathbb{P}\left(Z = z,\, Y = y,\, X = 0\right) + \mathbb{P}\left(Z = z,\, Y = y,\, X = 1\right)}{\mathbb{P}\left(Y = y,\, X = 0\right) + \mathbb{P}\left(Y = y,\, X = 1\right)} \\
&\quad = \Big[\mathbb{P}\left(Z = z \,|\, X = 0,\, Y = y\right) \mathbb{P}\left(Y = y \,|\, X = 0\right) \mathbb{P}\left(X = 0\right) \\
&\qquad + \mathbb{P}\left(Z = z \,|\, X = 1,\, Y = y\right) \mathbb{P}\left(Y = y \,|\, X = 1\right) \mathbb{P}\left(X = 1\right)\Big] \\
&\qquad \div \Big[\mathbb{P}\left(Y = y \,|\, X = 0\right) \mathbb{P}\left(X = 0\right) + \mathbb{P}\left(Y = y \,|\, X = 1\right) \mathbb{P}\left(X = 1\right)\Big] \\
&\quad = \Big[\mathbb{P}\left(Z = z \,|\, X = 0,\, Y = y\right) \mathbb{P}\left(Y = y \,|\, X = 0\right) \\
&\qquad + \mathbb{P}\left(Z = z \,|\, X = 1,\, Y = y\right) \mathbb{P}\left(Y = y \,|\, X = 1\right)\Big] \\
&\qquad \div \Big[\mathbb{P}\left(Y = y \,|\, X = 0\right) + \mathbb{P}\left(Y = y \,|\, X = 1\right)\Big]
\end{aligned}
$$

since $\mathbb{P}(X=0) = (X=1) = 1/2$, and so

$$\begin{aligned}
\mathbb{P}(Z=0\,|\,Y=0) &= \frac{\sigma(2)\cdot\sigma(2)+\frac{1}{2}\sigma(-2)}{\sigma(2)+\sigma(-2)} = \sigma(2)^2 + \frac{1}{2}\sigma(-2),\\
\mathbb{P}(Z=0\,|\,Y=1) &= \frac{\frac{1}{2}\sigma(-2)+\sigma(-2)\cdot\sigma(2)}{\sigma(-2)+\sigma(2)} = \sigma(-2)\left[\frac{1}{2}+\sigma(2)\right],\\
\mathbb{P}(Z=1\,|\,Y=0) &= \frac{\sigma(-2)\cdot\sigma(2)+\frac{1}{2}\sigma(-2)}{\sigma(2)+\sigma(-2)} = \sigma(-2)\left[\frac{1}{2}+\sigma(2)\right], \text{ and}\\
\mathbb{P}(Z=1\,|\,Y=1) &= \frac{\frac{1}{2}\sigma(-2)+\sigma(2)\cdot\sigma(2)}{\sigma(-2)+\sigma(2)} = \sigma(2)^2 + \frac{1}{2}\sigma(-2).
\end{aligned}$$

In particular we have, as expected, that

$$\mathbb{P}(Z=0\,|\,Y=j) + \mathbb{P}(Z=1\,|\,Y=j) = 1$$

for $j = 0, 1$. Surprisingly, however, we also have that

$$\mathbb{P}(Z=0\,|\,Y=0) = \mathbb{P}(Z=1\,|\,Y=1)$$

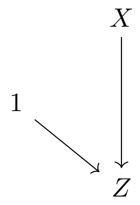
and

$$\mathbb{P}(Z=0\,|\,Y=1) = \mathbb{P}(Z=1\,|\,Y=0).$$

The quantity of interest is

$$\mathbb{P}(Z=1\,|\,Y=1) = \sigma(2)^2 + \frac{1}{2}\sigma(-2) \approx 0.84.$$

(3) The intervention "set $Y = 1$" creates the DAG



with PDF

$$f^*(x,\, z) = f(z\,|\,x,\, 1)f(x).$$

In other words, since $2(x+y)-2|_{y=1} = 2x$, the joint distribution f^* of $(X,\, Z)$ corresponds to

$$\begin{aligned}
&X \sim \text{Bernoulli}\left(\frac{1}{2}\right) \text{ and}\\
&Z \sim \text{Bernoulli}\left(\sigma(2x)\right).
\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{P}\left(Z = z \,|\, Y := 1\right) = f^*(z) &= \sum_{x=0}^{1} f^*(z \,|\, x) f^*(x) \\ &= \frac{1}{2} \sum_{x=0}^{1} f^*(z \,|\, x) \\ &= \frac{1}{2} \sum_{x=0}^{1} f(z \,|\, x,\, 1) \\ &= \frac{1}{2} \sum_{x=0}^{1} \mathbb{P}\left(Z = z \,|\, X = x,\, Y = 1\right),\end{aligned}$$

or alternatively we can derive the same identity via

$$\begin{aligned}&\mathbb{P}\left(Z = z \,|\, Y := 1\right) \\ &= \mathbb{P}\left(Z = z \,|\, X = 0,\, Y := 1\right) \mathbb{P}\left(X = 0\right) \\ &\quad + \mathbb{P}\left(Z = z \,|\, X = 1,\, Y := 1\right) \mathbb{P}\left(X = 1\right) \\ &= \frac{1}{2}\left[\mathbb{P}\left(Z = z \,|\, X = 0,\, Y = 1\right) + \mathbb{P}\left(Z = z \,|\, X = 1,\, Y = 1\right)\right].\end{aligned}$$

Either way we obtain that

$$\begin{aligned}\mathbb{P}\left(Z = 0 \,|\, Y := 1\right) &= \frac{1}{2}\left(\frac{1}{2} + \sigma(-2)\right) = \frac{1}{4} + \frac{1}{2}\sigma(-2) \text{ and} \\ \mathbb{P}\left(Z = 1 \,|\, Y := 1\right) &= \frac{1}{2}\left(\frac{1}{2} + \sigma(2)\right) = \frac{1}{4} + \frac{1}{2}\sigma(2)\end{aligned}$$

(which, as expected, add up to one). The quantity of interest is

$$\mathbb{P}\left(Z = 1 \,|\, Y := 1\right) = \frac{1}{4} + \frac{1}{2}\sigma(2) \approx 0.65.$$

How do we interpret this? $\mathbb{P}\left(Z = 1 \,|\, Y := 1\right)$ is a *causal* quantity: if we intervene and set $Y = 1$ then $Z = 1$ occurs with probability

$$\mathbb{P}\left(Z = 1 \,|\, Y := 1\right) \approx 0.65.$$

Why is $\mathbb{P}\left(Z = 1 \,|\, Y = 1\right)$ different (and, in this case, larger)? This comes down to the same issue as that which was illustrated using counterfactuals in Chapter 16 (see Remark 16.10 in particular): the distribution of the covariate Y is *not* independent from the distribution of how subjects *react* to the "treatment" Y.

More specifically, here, when we *observe* $Y = 1$ in a subject we are much more likely to have a subject with $X = 1$. Indeed

$$\mathbb{P}\left(X = 0,\, Y = 1\right) = \frac{\sigma(-2)}{2} \text{ while } \mathbb{P}\left(X = 1,\, Y = 1\right) = \frac{\sigma(2)}{2}$$

and so the odds of $X = 1$ over $X = 0$, given $Y = 1$, are

$$\frac{\mathbb{P}\left(X = 1 \,|\, Y = 1\right)}{\mathbb{P}\left(X = 0 \,|\, Y = 1\right)} = \frac{\mathbb{P}\left(X = 1,\, Y = 1\right)}{\mathbb{P}\left(X = 0,\, Y = 1\right)} = \frac{\sigma(2)}{\sigma(-2)} \approx 7.$$

This matters because subjects with $X = 1$ reach much better to treatment than subjects with $X = 0$ (when we restrict our attention to subjects with $Y = 1$, which we do when intervening and setting $Y = 1$). Indeed:

$$\mathbb{P}\left(Z = 1 \,|\, X = 1,\, Y = 1\right) = \sigma(2) > \sigma(-2) = \mathbb{P}\left(Z = 1 \,|\, X = 0,\, Y = 1\right).$$

In summary: when oberving $Y = 1$ we are likely to be observing a subject with $X = 1$, and such subjects react better to treatment. Therefore

$$\mathbb{P}\left(Z = 1 \,|\, Y = 1\right) \approx 0.84 > 0.69 \approx \mathbb{P}\left(Z = 1 \,|\, Y := 1\right),$$

i.e. conditioning by *observation* leads to a higher "survival" ($\iff Z = 1$) probability than conditioning by *intervention*.

Exercise A.17.9 (Intervention – Continuous case)**.** This is a continuous, Gaussian version of the last question. Let $V = (X,\, Y,\, Z)$ have the following joint distribution

$$\begin{aligned} &X \sim N(0,\, 1) \\ &Y \,|\, X = x \sim N(\alpha x,\, 1) \\ &Z \,|\, X = x,\, Y = y \sim N(\beta y + \gamma x,\, 1). \end{aligned}$$

Here α, β, and γ are fixed parameters. Economists refer to models like this as structural equation models.

(1) Find an explicit expression for $f(z \,|\, y)$ and

$$\mathbb{E}\left(Z \,|\, Y = y\right) = \int z f\left(z \,|\, y\right) dz.$$

(2) Find an explicit expression for $f\left(z \,|\, Y := y\right)$ and then find

$$\mathbb{E}\left(Z \,|\, Y := y\right) = \int z f\left(z \,|\, Y := y\right) dz.$$

Compare to item 1.

(3) Find the joint distribution of $(Y,\, Z)$. Find the correlation ρ between Y and Z.

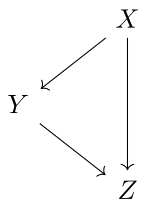
(4) Suppose that X is not observed and we try to make causal conclusions from the marginal distribution of $(Y,\, Z)$. (Think of X as unobserved confounding variables.) In particular, suppose we declare that Y causes Z if $\rho \neq 0$ and we declare that Y does not cause Z if $\rho = 0$. Show that this will lead to erroneous conclusions.

(5) Suppose we conduct a randomized experiement in which Y is randomly assigned. To be concrete, suppose that

$$\begin{aligned} &X \sim N(0,\, 1) \\ &Y \,|\, X = x \sim N(\alpha x,\, 1) \\ &Z \,|\, X = x,\, Y = y \sim N(\beta y + \gamma x,\, 1). \end{aligned}$$

Show that the method in item 4 now yields correct conclusions (i.e., $\rho = 0$ if and only if $f\left(z \,|\, Y := y\right)$ does not depend on y).

Solution. (1) The DAG is



and the PDF is

$$\begin{aligned} f(x,\, y,\, z) &= f(z \,|\, x,\, y) f(y \,|\, x) f(x) \\ &= \varphi(z - (\beta y + \gamma x))\varphi(y - \alpha x)\varphi(x) \end{aligned}$$

where φ denotes the PDF of a standard Normal distribution. In particular we may write

$$\begin{aligned} &\varphi(z - \beta y - \gamma x)\varphi(y - \alpha x)\varphi(x) \\ &= (2\pi)^{-3/2} \exp\left[-\frac{1}{2}(z - \beta y - \gamma x)^2 - \frac{1}{2}(y - \alpha x)^2 - \frac{x^2}{2} \right] \end{aligned}$$

where we may collect the various quadratic terms in the exponential and observe that

$$(z - \beta y - \gamma x)^2 + (y - \alpha x)^2 + x^2 = \Sigma^{-1}(x,\, y,\, z) \cdot (x,\, y,\, z)$$

for

$$\Sigma^{-1} = \begin{pmatrix} 1 + \alpha^2 + \beta^2 & \beta\gamma - \alpha & -\gamma \\ \beta\gamma - \alpha & 1 + \beta^2 & -\beta \\ -\gamma & -\beta & 1 \end{pmatrix}$$

such that

$$\Sigma = \begin{pmatrix} 1 & \alpha & \alpha\beta + \gamma \\ \alpha & 1 + \alpha^2 & \left(1 + \alpha^2\right)\beta + \alpha\gamma \\ \alpha\beta + \gamma & \left(1 + \alpha^2\right)\beta + \alpha\gamma & 1 + \left(1 + \alpha^2\right)\beta + 2\alpha\beta\gamma + \gamma^2 \end{pmatrix}.$$

In particular this means that $(X,\, Y,\, Z) \sim N(0,\, \Sigma)$. Therefore Theorem 2.42 tells us that

$$Z \,|\, Y = y \sim N\left(\mu(y),\, \sigma^2\right)$$

for

$$\mu(y) = \frac{\Sigma_{23}}{\Sigma_{22}} y = \left(\beta + \frac{\alpha\gamma}{1 + \alpha^2}\right) y$$

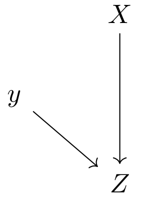
and

$$\sigma^2 = \Sigma_{33} - \frac{\Sigma_{23}^2}{\Sigma_{22}} = 1 + \frac{\gamma^2}{1 + \alpha^2}.$$

So finally

$$\mathbb{E}\left(Z \,|\, Y = y\right) = \mu(y) = \left(\beta + \frac{\alpha\gamma}{1 + \alpha^2}\right) y.$$

(2) After the intervention "set $Y = y$" the DAG becomes



and the PDF becomes

$$f^*(x,\, z) = f(z \,|\, x,\, y) f(x) = \varphi(z - \beta y - \gamma x)\varphi(x).$$

We proceed as above and note that now

$$\begin{aligned}
&(z-\beta y-\gamma x)^2+x^2\\
&=\begin{pmatrix}1+\gamma^2 & -\gamma\\ -\gamma & 1\end{pmatrix}\begin{pmatrix}x\\ z\end{pmatrix}\cdot\begin{pmatrix}x\\ z\end{pmatrix}+2\begin{pmatrix}\beta\gamma y\\ -\beta y\end{pmatrix}\cdot\begin{pmatrix}x\\ z\end{pmatrix}+\beta^2y^2\\
&=\begin{pmatrix}1+\gamma^2 & -\gamma\\ -\gamma & 1\end{pmatrix}\begin{pmatrix}x\\ z-\beta y\end{pmatrix}\cdot\begin{pmatrix}x\\ z-\beta y\end{pmatrix}.
\end{aligned}$$

Since

$$\begin{pmatrix}1+\gamma^2 & -\gamma\\ -\gamma & 1\end{pmatrix}^{-1}=\begin{pmatrix}1 & \gamma\\ \gamma & 1+\gamma^2\end{pmatrix}$$

this means that

$$(X,\,Z)\,|\,Y:=y\sim N\left(\begin{pmatrix}0\\ \beta y\end{pmatrix},\begin{pmatrix}1 & \gamma\\ \gamma & 1+\gamma^2\end{pmatrix}\right),$$

and so in particular

$$Z\,|\,Y:=y\sim N\left(\beta y,\,1+\gamma^2\right).$$

So finally

$$\mathbb{E}\left(Z\,|\,Y:=y\right)=\beta y.$$

This differs from the *passive* conditional expectation since we saw in item 1 that

$$\mathbb{E}\left(Z\,|\,Y=y\right)=\left(\beta+\frac{\alpha\gamma}{1+\alpha^2}\right)y.$$

In other words:

$$\mathbb{E}\left(Z\,|\,Y=y\right)-\mathbb{E}\left(Z\,|\,Y:=y\right)=\frac{\alpha\gamma}{1+\alpha^2}y.$$

(3) We discovered in item 1 that $(X,\,Y,\,Z)\sim N(0,\,\Sigma)$. Therefore the correlation ρ between Y and Z is

$$\begin{aligned}
\rho=\frac{\operatorname{Cov}(Y,\,Z)}{\sigma_Y\sigma_Z}&=\frac{\Sigma_{23}}{\sqrt{\Sigma_{22}\Sigma_{33}}}\\
&=\frac{\left(1+\alpha^2\right)\beta+\alpha\gamma}{\sqrt{(1+\alpha^2)\left[1+(1+\alpha^2)\,\beta+2\alpha\beta\gamma+\gamma^2\right]}}.
\end{aligned}$$

(4) If we *choose* $\alpha=1$ and $\gamma=-2\beta$ then, for *any* $\beta\in\mathbb{R}$ we will have that

$$\rho=0\text{ but }\mathbb{E}\left(Z\,|\,Y:=y\right)=\beta y.$$

In other words: the *causal* effect of Y on Z could be in either direction, and of any strength, and we would *always* erroneously declare that Y does not cause Z because the correlation vanishes.

Conversely, when $\beta=0$ we have that

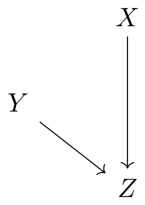
$$Z\,|\,Y=y\sim N(0,\,1+\gamma^2),$$

i.e. the distribution of $Z\,|\,Y=y$ is independent of y, which means that there is no causal effect, and so we may choose $\gamma=\pm\alpha$ such that the correlation is

$$\rho=\frac{\pm\alpha^2}{\left(1+\alpha^2\right)^2}.$$

We would then erroneously declare that Y causes Z for any $\alpha \neq 0$ (seeing association in either direction, depending on the sign of γ/α) even though there is no causal effect.

(5) The DAG is now



with PDF

$$g(x,\, y,\, z) = f(z \,|\, x,\, y) f(y) f(x) = \varphi(z - \beta y - \gamma x)\varphi(y - \alpha)\varphi(x).$$

To write this PDF as the PDF of a multivariate Normal we note that now $\mathbb{E}Y = \alpha$ and so, by the rule of iterated expectation,

$$\mathbb{E}Z = \mathbb{E}\left[E\left(Z \,|\, X,\, Y\right)\right] = \mathbb{E}\left(\beta Y + \gamma X\right) = \beta \mathbb{E}Y + 0 = \alpha\beta.$$

We are therefore looking for a symmetric and positive-definite 3-by-3 matrix Q such that

$$\begin{aligned} &(z - \beta y - \gamma x)^2 + (y - \alpha)^2 + x^2 \\ &= Q\left(x,\, y - \alpha,\, z - \alpha\beta\right) \cdot \left(x,\, y - \alpha,\, z - \alpha\beta\right). \end{aligned}$$

Careful inspection shows that this indeed holds for

$$Q = \begin{pmatrix} 1 + \gamma^2 & \beta\gamma & -\gamma \\ \beta\gamma & 1 + \beta^2 & -\beta \\ -\gamma & -\beta & 1 \end{pmatrix},$$

whose inverse is

$$V = \begin{pmatrix} 1 & 0 & \gamma \\ 0 & 1 & \beta \\ \gamma & \beta & 1 + \beta^2 + \gamma^2 \end{pmatrix}.$$

(Note that $V = \Sigma|_{\alpha=0}$.) Writing $m := (1,\, \alpha,\, \alpha\beta)$ we deduce that

$$(X,\, Y,\, Z) \sim N\left(m,\, V\right).$$

In order to use Theorem 2.42 once again we compute that

$$\frac{V_{23}}{V_{22}} y = \beta y \text{ and } V_{33} - \frac{V_{23}^2}{V_{22}} = 1 + \gamma^2,$$

which means that

$$Z \,|\, Y = y \sim N\left(\beta y,\, 1 + \gamma^2\right).$$

Proceeding as in item 3 we now compute the correlation between Y and Z:

$$\rho = \frac{\mathrm{Cov}(Y,\, Z)}{\sigma_Y \sigma_Z} = \frac{V_{23}}{\sqrt{V_{22} V_{33}}} = \frac{\beta}{\sqrt{1 + \beta^2 + \gamma^2}}.$$

Crucially:

$$Z \,|\, Y = y \text{ depends on } y \iff \beta \neq 0 \iff \rho \neq 0,$$

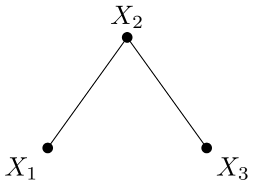
which shows that here the random assignment of Y guarantees that the correlation vanishes if and only if there is no causal effect of Y on Z.

A.18. **Undirected Graphs.**

Exercise A.18.1 (Independence relations and graphs)**.** Consider random variables (X_1, X_2, X_3). In each of the following cases, draw a graph that has the given independence relations.

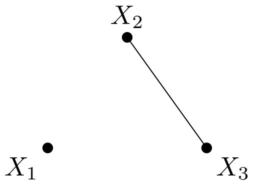
(1) $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$.
(2) $X_1 \perp\!\!\!\perp X_2 \,|\, X_3$ and $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$.
(3) $X_1 \perp\!\!\!\perp X_2 \,|\, X_3$ and $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$ and $X_2 \perp\!\!\!\perp X_3 \,|\, X_1$.

Solution. (1) By the pairwise Markov property, $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$ tells us that there is no edge between X_1 and X_3, i.e. $\{X_1, X_3\} \notin E$. One suitable graph is



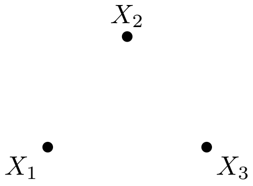
but any subgraph also works.

(2) Proceeding as in item 1 we deduce that $\{X_1, X_3\}$ and $\{X_1, X_2\}$ are not edges. A suitable graph is



and once again any subgraph works (here there is only one subgraph, namely the empty graph). In particular we deduce that, necessarily, the independence relation $X_1 \perp\!\!\!\perp (X_2, X_3)$ must hold.

(3) Proceeding as above we deduce that the graph cannot have *any* edges. The only suitable graph is thus the *empty* graph

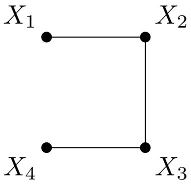


In particular we observe that X_1, X_2, and X_3 are necessarily pairwise independent.

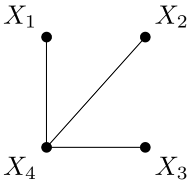
Exercise A.18.2 (More independence relations and graphs)**.** Consider random variables (X_1, X_2, X_3, X_4). In each of the following cases, draw a graph that has the given independence relations.

(1) $X_1 \perp\!\!\!\perp X_3 \,|\, X_2, X_4$ and $X_1 \perp\!\!\!\perp X_4 \,|\, X_1, X_3$ and $X_2 \perp\!\!\!\perp X_4 \,|\, X_1, X_3$.
(2) $X_1 \perp\!\!\!\perp X_2 \,|\, X_3, X_4$ and $X_1 \perp\!\!\!\perp X_3 \,|\, X_2, X_4$ and $X_2 \perp\!\!\!\perp X_3 \,|\, X_1, X_4$.
(3) $X_1 \perp\!\!\!\perp X_3 \,|\, X_2, X_4$ and $X_2 \perp\!\!\!\perp X_4 \,|\, X_1, X_3$.

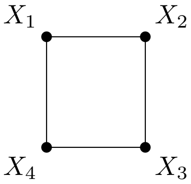
Solution. (1) By the pairwise Markov condition this means that $\{X_1, X_3\}$, $\{X_1, X_4\}$, and $\{X_2, X_4\}$ cannot be edges. Therefore the graph below, or any subgraph thereof, is suitable.



(2) Proceeding as in item 1 we see that $\{X_1, X_2\}$, $\{X_1, X_3\}$, and $\{X_2, X_3\}$ cannot be edges. The graph below, or any of its subgraphs, is suitable.

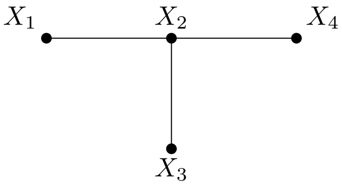


(3) Proceeding as above we see that $\{X_1, X_3\}$ and $\{X_2, X_4\}$ cannot be edges, so the graph below or any of its subgraphs does the trick.

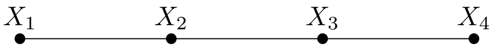


Exercise A.18.3 (Minimal conditional independence)**.** A conditional independece between a pair of variables is *minimal* if it is not possible to use the Separation Theorem (Theorem 18.3) to eliminate any variable from the conditioning set, i.e. from the right hand side of the bar. Write down the minimal conditional independencies from:

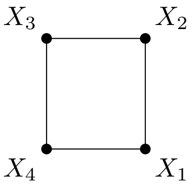
(1) the graph



(2) the graph

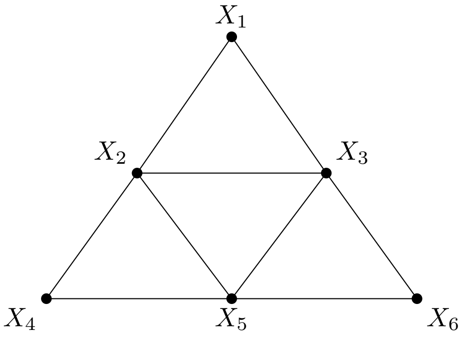


(3) the graph



and

(4) the graph



Solution. In order to determine the minimal conditional indepedencies, given a pair of vertices we find the smallest set of vertices separating them.

(1) Here this yields

$$\begin{aligned} &X_1 \perp\!\!\!\perp X_3 \,|\, X_2, \\ &X_3 \perp\!\!\!\perp X_4 \,|\, X_2, \text{ and} \\ &X_4 \perp\!\!\!\perp X_1 \,|\, X_2. \end{aligned}$$

(2) We obtain

$$\begin{aligned} &X_1 \perp\!\!\!\perp X_3 \,|\, X_2, \\ &X_2 \perp\!\!\!\perp X_4 \,|\, X_2, \text{ and} \\ &X_1 \perp\!\!\!\perp X_4 \,|\, (X_2, X_3)\,. \end{aligned}$$

(3) Here we have

$$\begin{aligned} &X_1 \perp\!\!\!\perp X_3 \,|\, (X_2, X_4) \text{ and} \\ &X_2 \perp\!\!\!\perp X_4 \,|\, (X_1, X_3)\,. \end{aligned}$$

(4) In this case we obtain

$$\begin{array}{ll} X_1 \perp\!\!\!\perp X_4 \,|\, (X_2, X_3)\,, & X_1 \perp\!\!\!\perp X_4 \,|\, (X_2, X_5)\,, \\ X_4 \perp\!\!\!\perp X_6 \,|\, (X_5, X_2)\,, & X_4 \perp\!\!\!\perp X_6 \,|\, (X_5, X_3)\,, \\ X_6 \perp\!\!\!\perp X_1 \,|\, (X_3, X_2)\,, & X_6 \perp\!\!\!\perp X_1 \,|\, (X_3, X_5)\,, \end{array}$$

as well as

$$\begin{aligned} &X_1 \perp\!\!\!\perp X_5 \,|\, (X_2, X_3)\,, \\ &X_2 \perp\!\!\!\perp X_6 \,|\, (X_3, X_5)\,, \text{ and} \\ &X_3 \perp\!\!\!\perp X_4 \,|\, (X_5, X_2)\,. \end{aligned}$$

Note that in this case, given a pair of vertices, several minimal conditional independencies may be found.

Exercise A.18.4 (Likelihood ratio test for the conditional independence of binary random variables)**.** Let X_1, X_2, and X_3 be binary random variables. Construct the likelihood ratio test for

$$H_0 : X_1 \perp\!\!\!\perp X_2 \,|\, X_3 \text{ versus } H_1 : X_1 \not\!\perp\!\!\!\perp X_2 \,|\, X_3.$$

Solution. We can view $X := (X_1,\, X_2,\, X_3)$ as a random vector with codomain $\{0,\, 1\}^3$ and parameter

$$p \in \Theta_1 := \Delta^7 \subseteq [0,\, 1]^8 \cong [0,\, 1]^{2\times 2\times 2}$$

such that

$$\mathbb{P}\left(X_1 = a,\, X_2 = b,\, X_3 = c\right) = p_{abc} \in [0,\, 1]$$

with

$$\sum_{a,b,c=0}^{1} p_{abc} = 1.$$

Inspired by Theorem 15.6 (and its proof in Exercise A.15.1) we will now show that

$$X_1 \perp\!\!\!\perp X_2 \,|\, X_3 \iff p_{00c}p_{11c} = p_{01c}p_{10c} \text{ for } c = 0 \text{ and } c = 1.$$

Here is the proof. First we define

$$\delta_c := p_{00c}p_{11c} - p_{01c}p_{10c}$$

and then compute as in Exercise A.15.1 that, where dotted subscripts denote summations as in Definition 15.3,

$$\begin{aligned} p_{0\cdot c}p_{\cdot 0c} &= \left(p_{00c} + p_{01c}\right)\left(p_{00c} + p_{10c}\right) \\ &= p_{00c}\left(p_{\cdot\cdot c} - p_{11c}\right) + p_{01c}p_{10c} \\ &= p_{00c}p_{\cdot\cdot c} - \delta_c, \end{aligned}$$

and so, similarly,

$$\begin{aligned} p_{0\cdot c}p_{\cdot 1c} &= p_{01c}\left(p_{\cdot\cdot c} - p_{10c}\right) + p_{00c}p_{11c} = p_{01c}p_{\cdot\cdot c} + \delta_c, \\ p_{1\cdot c}p_{\cdot 0c} &= p_{10c}\left(p_{\cdot\cdot c} - p_{01c}\right) + p_{11c}p_{00c} = p_{10c}p_{\cdot\cdot c} + \delta_c, \text{ and} \\ p_{1\cdot c}p_{\cdot 1c} &= p_{11c}\left(p_{\cdot\cdot c} - p_{00c}\right) + p_{10c}p_{01c} = p_{11c}p_{\cdot\cdot c} - \delta_c. \end{aligned}$$

We may therefore verify that

$$\begin{aligned} & X_1 \perp\!\!\!\perp X_2 \,|\, X_3 \\ \iff & \mathbb{P}\left(X_1 = a,\, X_2 = b \,|\, X_3 = c\right) = \mathbb{P}\left(X_1 = a \,|\, X_3 = c\right)\mathbb{P}\left(X_2 = b \,|\, X_3 = c\right) \\ \iff & \mathbb{P}\frac{\mathbb{P}\left(X_1 = a,\, X_2 = b,\, X_3 = c\right)}{\mathbb{P}\left(X_3 = c\right)} = \frac{\mathbb{P}\left(X_1 = a,\, X_3 = c\right)\mathbb{P}\left(X_2 = b,\, X_3 = c\right)}{\mathbb{P}\left(X_3 = c\right)^2} \\ \iff & p_{abc}p_{\cdot\cdot c} = p_{a\cdot c}p_{\cdot bc} \\ \iff & \delta_c = 0. \end{aligned}$$

This concludes the proof. In particular, with this result in hand we may then use the *conditional* log odds ratio

$$\gamma_c := \log \frac{p_{00c}p_{11c}}{p_{01c}p_{10c}}$$

and characterize

$$\begin{aligned} X_1 \amalg X_2 \,|\, X_3 &\iff \gamma_0 = \gamma_1 = 0 \\ &\iff p \in \Theta_0 := \{p \in \Theta_1 : \gamma_0(p) = \gamma_1(p) = 0\} \,. \end{aligned}$$

Since $\dim \Theta_1 - \dim \Theta_0 = 2$ (there are *two* scalar constraints imposed on p in Θ_0) the rejection region is

$$R_\alpha = \left\{ t \in \mathbb{R} : t > \chi^2_{2,\,\alpha} \right\} .$$

Finally we compute the likelihood ratio statistic. This is done by defining the multinomial random variable Y as in Definition 15.3 by

$$Y = (Y_{000},\, Y_{001},\, Y_{010},\, \dots) \in \mathbb{R}^8$$

where

$$Y_{abc} = \#\{\text{observations where } X_1 = a,\ X_2 = b, \text{ and } X_3 = c\}$$

such that $Y \sim \text{Multinomial}(n,\, p)$. Over Θ_1 the MLE may be found from Exercise A.23.13 to be

$$\hat{p} = \frac{Y}{n}.$$

To compute the MLE $\hat{q}$ over Θ_0 we proceed as in Exercise A.15.2 and seek to maximize, for λ now in $\mathbb{R}^3$,

$$f(q,\, \lambda) := \log \mathcal{L}(q) - \lambda \cdot \left(\gamma_0(q),\, \gamma_1(q),\, \sum_{a,b,c=0}^{1} q_{abc} - 1 \right)$$

where

$$\log \mathcal{L}(q) = \sum_{a,b,c=0}^{1} Y_{abc} \log q_{abc}$$

and, for $c = 0, 1$,

$$\gamma_c(q) = \log \frac{q_{00c} q_{11c}}{q_{01c} q_{10c}} = \log q_{00c} - \log q_{01c} - \log q_{10c} + \log q_{11c}.$$

Therefore

$$\nabla f = \begin{pmatrix} Y_{000}/q_{000} - \lambda_0/q_{000} - \lambda_2 \\ Y_{010}/q_{010} + \lambda_0/q_{010} - \lambda_2 \\ Y_{100}/q_{100} + \lambda_0/q_{100} - \lambda_2 \\ Y_{110}/q_{110} - \lambda_0/q_{110} - \lambda_2 \\ Y_{001}/q_{001} - \lambda_1/q_{001} - \lambda_2 \\ Y_{011}/q_{011} + \lambda_1/q_{011} - \lambda_2 \\ Y_{101}/q_{101} + \lambda_1/q_{101} - \lambda_2 \\ Y_{111}/q_{111} - \lambda_1/q_{111} - \lambda_2 \\ -\log q_{000} + \log q_{010} + \log q_{100} - \log q_{110} \\ -\log q_{001} + \log q_{011} + \log q_{101} - \log q_{111} \\ 1 - q_{000} - q_{001} - q_{010} - q_{011} - q_{100} - q_{101} - q_{110} - q_{111} \end{pmatrix} .$$

Proceeding as in Exercise A.15.2 we see that

$$0 = q \cdot \nabla_q f = n - \lambda_2$$

and so $\lambda_2 = n$. Similarly we deduce that

$$n q_{abc} = Y_{abc} + (-1)^{a+b+1} \lambda_c$$

and so

$$\lambda_c = \frac{\delta_c}{Y_{\cdot\cdot c}}.$$

Since (omitting the details of the computation)

$$\delta_c = (-1)^{a+b}\left(Y_{\cdot\cdot c}Y_{abc} - Y_{a\cdot c}Y_{\cdot bc}\right)$$

we conclude that

$$nY_{\cdot\cdot c}\hat{q}_{abc} = Y_{\cdot\cdot c}Y_{abc} + (-1)^{a+b+1}\delta_c = Y_{a\cdot c}Y_{\cdot bc},$$

i.e.

$$\hat{q}_{abc} = \frac{Y_{a\cdot c}Y_{\cdot bc}}{nY_{\cdot\cdot c}}.$$

So finally we conclude that the likelihood ratio statistic is

$$T := 2\log\frac{\mathcal{L}(\hat{p})}{\mathcal{L}(\hat{q})} = 2\log\prod_{a,b,c=0}^{1}\left(\frac{\hat{p}_{abc}}{\hat{q}_{abc}}\right)^{Y_{abc}} = 2\sum_{a,b,c=0}^{1} Y_{abc}\log\frac{Y_{abc}Y_{\cdot\cdot c}}{Y_{a\cdot c}Y_{\cdot bc}}.$$

Recall in particular from Theorem 10.33 that the asymptotic p–value of this likelihood ratio test is

$$p - \text{value}(\omega) = \mathbb{P}\left(\chi_2^2 > T(\omega)\right)$$

for any outcome ω.

Note that we may write the likelihood ratio statistic as

$$T = T_0 + T_1$$

where

$$T_c = 2\sum_{a,b,c=0}^{1} Y_{abc}\log\frac{Y_{abc}Y_{\cdot\cdot c}}{Y_{a\cdot c}Y_{\cdot bc}}$$

is the likelihood ratio statistic testing

$$H_0 : X_1 \perp\!\!\!\perp X_2 \,|\, X_3 = c \text{ versus } H_1 : X_1 \not\!\perp\!\!\!\perp X_2 \,|\, X_3 = c$$

as recorded in Theorem 15.7.

Exercise A.18.5 (Cancer data and undirected graphs)**.** Here are breast cancer data from Morrison et al. (1973) on diagnostic center (X_1), nuclear grade (X_2), and survival (X_3):

	X_2	maligant	malignant	benign	benign
	X_3	died	survived	died	survival
X_1	Boston	35	59	47	112
	Glamorgan	42	77	26	76

(1) Treat this as a multinomial and find the maximum likelihood ratio.
(2) If someone has a tumor classified as benign at the Glamorgan clinic, what is the estimated probability that they will die? Find the standard error for this estimate.
(3) Test the following hypotheses:

$$X_1 \perp\!\!\!\perp X_2 \,|\, X_3 \text{ versus } X_1 \leftrightsquigarrow X_2 \,|\, X_3,$$

$$X_1 \perp\!\!\!\perp X_3 \,|\, X_2 \text{ versus } X_1 \leftrightsquigarrow X_3 \,|\, X_2 \text{ and}$$

$$X_2 \perp\!\!\!\perp X_3 \,|\, X_1 \text{ versus } X_2 \leftrightsquigarrow X_3 \,|\, X_1.$$

Use the test from Exercise A.18.4. Based on the results of your tests, draw and interpret the resulting graph.

Solution. (1) We use the notation of Exercise A.18.4 such that

$$\hat{p} = \frac{Y_{abc}}{n},$$

where here $n = 474$. We use the convention

$$\begin{aligned} a = 0 &\iff \text{Boston}, & a = 1 &\iff \text{Glamorgan}, \\ b = 0 &\iff \text{benign}, & b = 1 &\iff \text{malignant, and} \\ c = 0 &\iff \text{died}, & c = 1 &\iff \text{survived}. \end{aligned}$$

The MLE is

$$\hat{p} = \frac{Y}{n}$$

such that

$$\begin{aligned} \hat{p} &= (\hat{p}_{000},\, \hat{p}_{001},\, \hat{p}_{010},\, \dots) \\ &\approx (0.1,\, 0.24,\, 0.07,\, 0.12,\, 0.05,\, 0.16,\, 0.09,\, 0.16)\,. \end{aligned}$$

(2) We note that

$$\begin{aligned} \theta &:= \mathbb{P}\left(\text{die} \,|\, \text{benign, Glamorgan}\right) \\ &= \frac{\mathbb{P}\left(\text{Glamorgan, benign, die}\right)}{\mathbb{P}\left(\text{Glamorgan, benign}\right)} \\ &= \frac{p_{100}}{p_{10\cdot}} = \frac{p_{100}}{p_{100} + p_{101}} \end{aligned}$$

and so by equivariance the MLE is

$$\hat{\theta} = \hat{\mathbb{P}}\left(\text{die} \,|\, \text{benign, Glamorgan}\right) = \frac{\hat{p}_{100}}{\hat{p}_{100} + \hat{p}_{101}} \approx 0.25.$$

Now to make the notation less burdensome we write

$$r_0 := p_{100},\, r_1 := p_{101}, \text{ and } r_\cdot := p_{10\cdot} = r_0 + r_1$$

such that $\theta = r_0/r_1$ and $\hat{\theta} = \hat{r}_0/\hat{r}_1$. We computed in Exercise A.23.14 the Fisher information matrix of the Categorical model and so we know by asymptotic normality of the MLE that

$$\sqrt{n}\left[(\hat{r}_0,\, \hat{r}_1) - (r_0,\, r_1)\right] \rightsquigarrow N\left(0,\, \overline{J}\right)$$

where

$$\overline{J} = \begin{pmatrix} r_0(1 - r_0) & -r_0 r_1 \\ -r_0 r_1 & r_1(1 - r_1) \end{pmatrix}.$$

Since $\theta = g(r_0,\, r_1)$ for $g(x,\, y) = \frac{x}{x+y}$ such that

$$\nabla g = \frac{1}{(x+y)^2} \begin{pmatrix} y \\ -x \end{pmatrix}$$

the delta method tells us that, for

$$\begin{aligned}\sigma &:= \overline{J}\nabla g \cdot \nabla g|_{(r_0,\, r_1)}\\ &= \frac{1}{r_\cdot^4}\left[r_0(1-r_0)r_1^2 + 2r_0^2 r_1^2 + r_1(1-r_1)r_0^2\right]\\ &= \frac{r_0 r_1^2 + r_0^2 r_1}{r_\cdot^4}\\ &= \frac{r_0 r_1}{r_\cdot^3},\end{aligned}$$

then

$$\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow N(0,\, \sigma).$$

We may thus use

$$\widehat{se} := \sqrt{\frac{\sigma}{n}}\Big|_{r=\hat{r}} = \sqrt{\frac{\hat{r}_0 \hat{r}_1}{n\hat{r}_\cdot^3}}$$

as an estimate for the standard error of $\hat{\theta}$. Here

$$\widehat{se} \approx 0.43.$$

In particular

$$(0.17,\, 0.34)$$

is an asymptotic 95% confidence interval for θ.

(3) As obtained in Exercise A.18.4 the likelihood ratio statistic for these three tests are, respectively,

$$\begin{aligned}T_3 &:= 2\sum_{a,b,c=0}^{1} Y_{abc} \log \frac{Y_{abc}Y_{\cdot\cdot c}}{Y_{a\cdot c}Y_{\cdot bc}},\\ T_2 &:= 2\sum_{a,b,c=0}^{1} Y_{abc} \log \frac{Y_{abc}Y_{\cdot b\cdot}}{Y_{ab\cdot}Y_{\cdot bc}}, \text{ and}\\ T_1 &:= 2\sum_{a,b,c=0}^{1} Y_{abc} \log \frac{Y_{abc}Y_{a\cdot\cdot}}{Y_{ab\cdot}Y_{a\cdot b}},\end{aligned}$$

with asymptotic p–values

$$p - \text{value}_j = \mathbb{P}\left(\chi_2^2 > T_j\right).$$

We obtain

$$\begin{cases} p - \text{value}_1 \approx 0.1,\\ p - \text{value}_2 \approx 0.7, \text{ and}\\ p - \text{value}_3 \approx 0.001,\end{cases}$$

which means that the *only* null hypothesis we may reject is

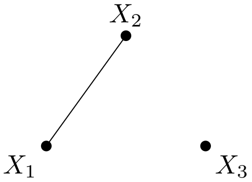
$$X_1 \amalg X_2 \,|\, X_3.$$

(Note that the conclusion is the same whether or not we account for multiple *independent* tests since $p - \text{value}_3 \approx 0.001 = \frac{0.05}{50}$, which is much smaller than the usual $\alpha = 0.05$ rejection treshold.) Note also that there is some, albeit somewhat weak, evidence that the null hypothesis $X_2 \amalg X_3 \,|\, X_1$ ought to be rejected as well.

The graph corresponding to

$$X_2 \perp\!\!\!\perp X_3 \mid X_1 \text{ and } X_1 \perp\!\!\!\perp X_3 \mid X_2 \text{ but } X_1 \not\!\perp\!\!\!\perp X_2 \mid X_3$$

is

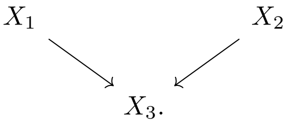


Based on this graph we see that, given this data, the only conditional *dependence* we can establish is

$$X_1 \not\!\perp\!\!\!\perp X_2 \mid X_3.$$

This is not necessarily useful since X_3 corresponds to "survival versus death", which is usually viewed as a *response* variable, not as a covariate. Conditioning on a response variable is very delicate business. Indeed, it can even be *misleading* if the response is a *collider*, as discussed below.

Indeed, suppose that the DAG underlying these variables is



Note that we are *not* inferring this DAG from the data presented here; we are only invoking this DAG for *illustrative* purposes. Then, as shown in Exercise A.17.5,

$$X_1 \perp\!\!\!\perp X_2 \text{ but } X_1 \not\!\perp\!\!\!\perp X_2 \mid X_3.$$

This highlights that *undirected* graphs are easier to interpret when all the variables are *covariates*, as opposed to some of them being covariates and other being responses (as is the case here).

As a side note: the likelihood ratio test of Theorem 15.7 for $H_0 : X_1 \perp\!\!\!\perp X_2$ has a p–value of approximately 0.0002. This is very strong evidence *rejecting* the (unconditional) independence of X_1 and X_2.

A.19. **Log-Linear Models.**

Exercise A.19.1 (Log-linear parametrization for a simple case)**.** Let $X = (X_1, X_2)$ where $X_1 \in \{0, 1\}$ and $X_2 \in \{0, 1, 2\}$. The log-linear expansion of the PDF of X is

$$\psi_\emptyset(x) + \psi_1(x) + \psi_2(x) + \psi_{12}(x)$$

for

$$\begin{aligned}
\psi_\emptyset(x) &= \log p_{00}, \\
\psi_1(x) &= x_1 \log \frac{p_{10}}{p_{00}}, \\
\psi_2(x) &= \mathbb{1}(x_2 = 1) \log \frac{p_{01}}{p_{00}} + \mathbb{1}(x_2 = 2) \log \frac{p_{02}}{p_{00}}, \text{ and} \\
\psi_{12}(x) &= \mathbb{1}(x_1 = 1,\, x_2 = 1) \log \frac{p_{00}p_{11}}{p_{01}p_{10}} + \mathbb{1}(x_1 = 1,\, x_2 = 2) \log \frac{p_{00}p_{12}}{p_{02}p_{10}}.
\end{aligned}$$

In particular we may write

$$\begin{aligned}
\psi_\emptyset(x) &= \beta_1, \\
\psi_1(x) &= \beta_2 x_1, \\
\psi_2(x) &= \beta_3 \mathbb{1}(x_2 = 1) + \beta_4 \mathbb{1}(x_2 = 2), \text{ and} \\
\psi_{12}(x) &= \beta_5 \mathbb{1}(x_1 = 1,\, x_2 = 1) + \beta_6 \mathbb{1}(x_1 = 1,\, x_2 = 2)
\end{aligned}$$

for

$$\begin{aligned}
\beta_1 &= \log p_{00}, & \beta_2 &= \log \frac{p_{10}}{p_{00}}, \\
\beta_3 &= \log \frac{p_{01}}{p_{00}}, & \beta_4 &= \log \frac{p_{02}}{p_{00}}, \\
\beta_5 &= \log \frac{p_{00}p_{11}}{p_{01}p_{10}}, \text{ and} & \beta_6 &= \log \frac{p_{00}p_{12}}{p_{02}p_{10}}.
\end{aligned}$$

Solve for the p_{ij}'s in terms of the β's.

Solution. Using the expression for β_1 we solve for p_{00} and obtain

$$p_{00} = e^{\beta_1}.$$

Plugging this into the expression for β_2, β_3, and β_4 tells us that

$$p_{10} = e^{\beta_1 + \beta_2},\ p_{01} = e^{\beta_1 + \beta_3}, \text{ and } p_{02} = e^{\beta_1 + \beta_4}.$$

Finally we plug all of this into the expressions for β_5 and β_6 to conclude that

$$\begin{aligned}
& p_{01}p_{10}e^{\beta_5} = p_{00}p_{11} \text{ and} & p_{02}p_{10}e^{\beta_6} &= p_{00}p_{12} \\
\iff\quad & e^{\beta_1+\beta_3}e^{\beta_1+\beta_2}e^{\beta_5} = e^{\beta_1}p_{11} \text{ and} & e^{\beta_1+\beta_4}e^{\beta_1+\beta_2}e^{\beta_6} &= e^{\beta_1}p_{12} \\
\iff\quad & p_{11} = e^{\beta_1+\beta_2+\beta_3+\beta_5} \text{ and} & p_{12} &= e^{\beta_1+\beta_2+\beta_4+\beta_6}.
\end{aligned}$$

In summary

$$\begin{aligned}
p_{00} &= e^{\beta_1}, & p_{10} &= e^{\beta_1+\beta_2}, \\
p_{01} &= e^{\beta_1+\beta_3}, & p_{11} &= e^{\beta_1+\beta_2+\beta_3+\beta_5}, \\
p_{02} &= e^{\beta_1+\beta_4}, \text{ and} & p_{12} &= e^{\beta_1+\beta_2+\beta_4+\beta_6}.
\end{aligned}$$

Note that a more suggestive indexing scheme for the β's, which departs from the notation used in the book, is

$$\begin{aligned}
\beta_{**} &= \log p_{00}, & \beta_{1*} &= \log \frac{p_{10}}{p_{00}}, \\
\beta_{*1} &= \log \frac{p_{01}}{p_{00}}, & \beta_{*2} &= \log \frac{p_{02}}{p_{00}}, \\
\beta_{11} &= \log \frac{p_{00}p_{11}}{p_{01}p_{10}}, \text{ and} & \beta_{12} &= \log \frac{p_{00}p_{12}}{p_{02}p_{10}}.
\end{aligned}$$

This leads to the much cleaner formulae

$$\begin{aligned}
&\psi_\emptyset(x) = \beta_{**}, \\
&\psi_1(x) = \beta_{1*}\mathbb{1}(x_1 = 1), \\
&\psi_2(x) = \beta_{*1}\mathbb{1}(x_2 = 1) + \beta_{*2}\mathbb{1}(x_2 = 2), \text{ and} \\
&\psi_{12}(x) = \beta_{11}\mathbb{1}(x_1 = 1,\, x_2 = 1) + \beta_{12}\mathbb{1}(x_1 = 1,\, x_2 = 2)
\end{aligned}$$

as well as

$$\begin{aligned}
p_{00} &= e^{\beta_{**}}, & p_{10} &= e^{\beta_{**}+\beta_{1*}}, & p_{11} &= e^{\beta_{**}+\beta_{1*}+\beta_{*1}+\beta_{11}}, \\
& & p_{01} &= e^{\beta_{**}+\beta_{*1}}, & p_{12} &= e^{\beta_{**}+\beta_{1*}+\beta_{*2}+\beta_{12}}, \text{ and} \\
& & p_{02} &= e^{\beta_{**}+\beta_{*2}}.
\end{aligned}$$

Exercise A.19.2 (A characterization of conditional independence via PDFs)**.** Prove Lemma 19.6.

Solution. If $X_B \amalg X_C \mid X_A$ then

$$f(x_A,\, x_B,\, x_C) = f(x_B,\, x_C \mid x_A)f(x_A) = \underbrace{f(x_B \mid x_A)}_{g(x_A,\, x_B)}\underbrace{f(x_C \mid x_A)f(x_A)}_{h(x_A,\, x_C)}$$

as desired. Conversely, if

$$f(x_A,\, x_B,\, x_C) = g(x_A,\, x_B)h(x_A,\, x_C)$$

then

$$\begin{aligned}
f(x_B \mid x_A,\, x_C) &= \frac{f(x_A,\, x_B,\, x_C)}{f(x_A,\, x_C)} \\
&= \frac{g(x_A,\, x_B)h(x_A,\, x_C)}{h(x_A,\, x_C)\int g(x_A,\, x_B)dx_B} \\
&= \frac{g(x_A,\, x_B)}{\int g(x_A,\, x_B)dx_B} \\
&= \frac{g(x_A,\, x_B)\int h(x_A,\, x_c)dx_C}{\int g(x_A,\, x_B)dx_B \int h(x_A,\, x_c)dx_C} \\
&= \frac{f(x_A,\, x_B)}{f(x_A)} \\
&= f(x_B \mid x_A).
\end{aligned}$$

By Theorem 17.2 this verifies that $X_B \amalg X_C \mid X_A$.

Exercise A.19.3 (Graphical implies hierarchical)**.** Prove Lemma 19.11 where we show that graphical models are hierarchical but the converse fails.

Solution. Suppose $f = \sum_A \psi_A$ is graphical with respect to some undirected graph $\mathcal{G}$ and let $A \subseteq B$ with $\psi_A = 0$. This means that $|A| \geqslant 2$ and that $\{i, j\} \subseteq A$ for some unordered pair $\{i, j\}$ which is *not* an edge in $\mathcal{G}$. Therefore $\{i, j\} \subseteq B$ as well, and, since f is graphical, this means that $\psi_B = 0$, proving that f is hierarchical.

The converse fails since, as seen in Example 19.12, the model

$$\psi_\emptyset + \psi_1 + \psi_2 + \psi_{12}$$

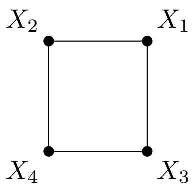
on *three* variables is hierarchical but not graphical (since $\psi_3 = 0$).

Exercise A.19.4 (A log-linear model)**.** Consider the random variables X_1, X_2, X_3, and X_4. Suppose the log-density is

$$\log f(x) = \psi_\emptyset(x) + \psi_{12}(x) + \psi_{13}(x) + \psi_{24}(x) + \psi_{34}(x).$$

(1) Draw the graph $\mathcal{G}$ for these variables.
(2) Write down all independence and conditional independence relations implied by the graph.
(3) Is this model graphical? Is it hierarchical?

Solution. (1) f is Markov to the undirected graph



(2) f is Markov to an undirected graph with only one (connected) component so it does not have any indepence relations implied by its graph. To find *conditional* independence relations we use the global Markov property, which tells us that

$$X_1 \amalg X_4 \mid (X_2, X_3) \text{ and } X_2 \amalg X_3 \mid (X_1, X_4).$$

This is, up to relabelling the random variables, the same set of conditional independence relations as in item 3 of Exercise A.18.2.

(3) This model is not hierarchical since $\psi_1 = 0$ and yet $\psi_{12} \neq 0$. Consequently Lemma 19.11 tells us that this model is not graphical either (which we can also verify directly since $\psi_1 = 0$).

Exercise A.19.5 (A log-linear expansion)**.** Suppose that parameters $p(x_1, x_2, x_3)$ are proportional to the following values:

	x_2	0	0	1	1
	x_3	0	1	0	1
x_1	0	2	8	4	16
	1	16	128	32	256

Find the ψ–terms for the log-linear expansion. Comment on the model.

Solution. We have computed in Exercise A.23.22 the log-linear expansion of the PDF of (X_1, X_2, X_3), whenever X_1, X_2, and X_3 are binary random variables, in terms of the parameters

$$p_{abc} = \mathbb{P}(X_1 = a, X_2 = b, X_3 = c).$$

Note that, apart from $\psi_\emptyset$, the ψ–terms of the log-linear expansion only depend on *ratios* of the parameters p_{abc}, so we only need to normalize these parameters in order to computer $\psi_\emptyset$. We denote

$$\begin{aligned} &N_{000} = 2, && N_{001} = 8, && N_{010} = 4, && N_{011} = 16, \\ &N_{100} = 16, && N_{101} = 128, && N_{110} = 32, \text{ and} && N_{011} = 256 \end{aligned}$$

and obtain

$$\psi_\emptyset = \log \frac{N_{000}}{\sum_{a,b,c=0}^{1} N_{abc}} = \log \frac{2}{462} = -\log 231.$$

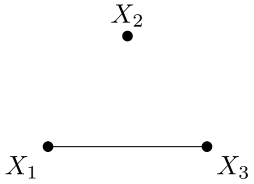
The other (non-constant) ψ–terms are given by

$$\begin{aligned} \psi_1 &= x_1 \log \frac{N_{100}}{N_{000}} = x_1 \log \frac{16}{2} = 3x_1 \log 2, \\ \psi_2 &= x_2 \log \frac{N_{010}}{N_{000}} = x_2 \log \frac{4}{2} = x_2 \log 2, \\ \psi_3 &= x_3 \log \frac{N_{001}}{N_{000}} = x_3 \log \frac{8}{2} = 2x_3 \log 2, \\ \psi_{12} &= x_1x_2 \log \frac{N_{000}N_{110}}{N_{010}N_{100}} = x_1x_2 \log \frac{2 \cdot 32}{4 \cdot 16} = 0, \\ \psi_{13} &= x_1x_3 \log \frac{N_{000}N_{101}}{N_{001}N_{100}} = x_1x_3 \log \frac{2 \cdot 128}{8 \cdot 16} = x_1x_3 \log 2, \\ \psi_{23} &= x_2x_3 \log \frac{N_{000}N_{011}}{N_{001}N_{010}} = x_2x_3 \log \frac{2 \cdot 16}{8 \cdot 4} = 0, \text{ and} \\ \psi_{123} &= x_1x_2x_3 \log \frac{N_{001}N_{010}N_{100}N_{111}}{N_{110}N_{101}N_{011}N_{000}} = x_1x_2x_3 \log \frac{8 \cdot 4 \cdot 16 \cdot 256}{32 \cdot 128 \cdot 16 \cdot 2} = 0. \end{aligned}$$

In particular we note that the only vanishing ψ–terms are

$$\psi_{12},\ \psi_{23}, \text{ and } \psi_{123}.$$

This means that this model is graphical and Markov to the undirected graph



This is actually, up to relabelling the random variables, the same model as the first example in Example 19.12. In particular, since this model is graphical it is also hierarchical. Finally, since this model is graphical, its only (conditional) independence relations are those we may read from its graph, namely

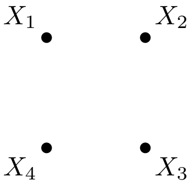
$$X_2 \amalg (X_1,\, X_3)\,.$$

Exercise A.19.6 (More log-linear expansions)**.** Let X_1, X_2, X_3, and X_4 be binary. Draw the independence graphs corresponding to the following log-linear models. Also, identify whether each is graphical and/or hierarchical (or neither).

(1) $\log f = 7 + 11x_1 + 2x_2 + 1.5x_3 + 17x_4$.
(2) $\log f = 7 + 11x_1 + 2x_2 + 1.5x_3 + 17x_4 + 12x_2x_3 + 78x_2x_4 + 3x_3x_4 + 32x_2x_3x_4$.
(3) $\log f = 7 + 11x_1 + 2x_2 + 1.5x_3 + 17x_4 + 12x_2x_3 + 3x_3x_4 + x_1x_4 + 2x_1x_2$.

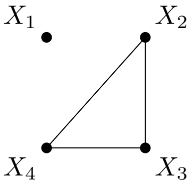
(4) $\log f = 7 + 5055 x_1 x_2 x_3 x_4$.

Solution. (1) This distribution is Markov to the *empty* undirected graph



It is graphical, and hence hierarchical. Note that all random variables are independent here.

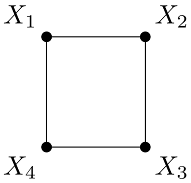
(2) This distribution is Markov to the undirected graph



It is graphical, and hence hierarchical. Here the only independence relation is

$$X_1 \amalg (X_2,\, X_3,\, X_4)\,.$$

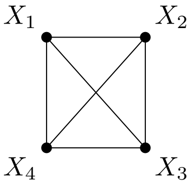
(3) This distribution is Markov to the undirected graph



It is graphical, and hence hierarchical. The conditional independence relations are

$$X_1 \amalg X_3 \mid (X_2,\, X_4) \text{ and } X_2 \amalg X_4 \mid (X_1,\, X_3)\,.$$

(4) This distribution is Markov to the *complete* undirected graph



It is *not* hierarchical since $\psi_{1234} \neq 0$ even though $\psi_1 = 0$. It is therefore not graphical either.

A.20. **Nonparametric Curve Estimation.**

Exercise A.20.1 (Properties of kernel density estimation using the boxcar kernel)**.** Let $X_1, \ldots, X_n \sim f$ and let $\hat{f}_n$ be the kernel density estimator using the boxcar kernel:

$$K(x) = \begin{cases} 1 & \text{if } -\dfrac{1}{2} < x < \dfrac{1}{2} \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

(1) Show that

$$\mathbb{E}\hat{f}_n(x) = \frac{1}{h}\int_{x-h/2}^{x+h/2} f$$

and

$$\mathbb{V}\hat{f}_n(x) = \frac{1}{nh^2}\left[\int_{x-h/2}^{x+h/2} f - \left(\int_{x-h/2}^{x+h/2} f\right)^2\right].$$

(2) Show that if $h \to 0$ and $nh \to \infty$ as $n \to \infty$ then, for every $x \in \mathbb{R}$, $\hat{f}_n(x) \xrightarrow{P} f(x)$.

Solution. (1) Recall that

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

where

$$\begin{aligned} K\left(\frac{x - X_i}{h}\right) &= \mathbb{1}\left(\left|\frac{x - X_i}{h}\right| < \frac{1}{2}\right) \\ &= \mathbb{1}\left(x - h/2 < X_i < x + h/2\right) =: Y_i \end{aligned}$$

such that $\hat{f}_n(x)$ is $\frac{1}{h}$ times the sample mean of IID Bernoulli random variables with Bernoulli parameter

$$\mathbb{P}\left(Y = 1\right) = \mathbb{P}\left(x - h/2 < X < x + h/2\right) = \int_{x-h/2}^{x+h/2} f =: p(x).$$

Therefore

$$\mathbb{E}\hat{f}_n(x) = \frac{1}{h}\mathbb{E}\overline{Y}_n = \frac{\mathbb{E}Y}{h} = \frac{p(x)}{h} = \frac{1}{h}\int_{x-h/2}^{x+h/2} f$$

and

$$\begin{aligned} \mathbb{V}\hat{f}_n(x) = \frac{1}{h^2}\mathbb{V}\overline{Y}_n = \frac{\mathbb{V}Y}{nh^2} &= \frac{p(x)(1 - p(x))}{nh^2} \\ &= \frac{1}{nh^2}\left[p(x) - p(x)^2\right] \\ &= \frac{1}{nh^2}\left[\int_{x-h/2}^{x+h/2} f - \left(\int_{x-h/2}^{x+h/2} f\right)^2\right]. \end{aligned}$$

(2) It follows immediately from the bias-variance decomposition of the mean squared error (Theorem 6.10) that

$$\mathbb{E}\left(\left[\hat{f}_n(x) - f(x)\right]^2\right) = \left(\mathbb{E}\left[\hat{f}_n(x) - f(x)\right]\right)^2 + \mathbb{V}\hat{f}_n(x).$$

Provided that f is continuous we have that

$$\frac{1}{h}\int_{x-h/2}^{x+h/2} f = \frac{p(x)}{h} \to f(x) \text{ as } h \to 0.$$

(otherwise this convergence only happens at Lesbesgue points of f, which have full measure in $\mathbb{R}$ since f is integrable). Therefore

$$\mathbb{E}\left[\hat{f}_n(x) - f(x)\right] = \frac{1}{h}\int_{x-h/2}^{x+h/2} f - f(x) \to 0 \text{ as } h \to 0$$

while

$$\begin{aligned} \mathbb{V}\hat{f}_n(x) &= \frac{1}{nh}\cdot\frac{p(x)\left[1-p(x)\right]}{h} \\ &= \frac{1}{nh}\cdot\frac{p(x)}{h}\cdot\left[1 - \frac{p(x)}{h}\cdot h\right] \to 0\cdot f(x)\cdot 1 \text{ as } n\to\infty \end{aligned}$$

since $nh \to \infty$ as $n \to \infty$. This shows that

$$\mathbb{E}\left(\left[\hat{f}_n(x) - f(x)\right]^2\right) \to 0 \text{ as } n \to \infty,$$

or in other words that $\hat{f}_n(x)$ converges to $f(x)$ in quadratic mean. It then follows immediately (see Theorem 5.4) that $\hat{f}_n(x)$ also converges in probability to $f(x)$, as desired.

Exercise A.20.2 (Bias-variance tradeoff)**.** Prove Lemma 20.3 where we record a precise form of the bias–variance tradeoff for the mean integrated squared error.

Solution. This follows from Remark 20.2 and Theorem 6.10. Indeed, inspired by Remark 20.2 we use Tonelli's Theorem to write

$$\begin{aligned} R(g,\,\hat{g}) = \mathbb{E}\left[L(g,\,\hat{g})\right] = \mathbb{E}\int L(g(x),\,\hat{g}(x))dx &= \int \mathbb{E}\left[L(g(x),\,\hat{g}(x))\right]dx \\ &= \int R\left[g(x),\,\hat{g}(x)\right]dx. \end{aligned}$$

In particular the decomposition of the mean squared error in Theorem 6.10 tells us that

$$R\left[g(x),\,\hat{g}(x)\right] = MSE\left[\hat{g}(x)\right] = \left(\mathbb{E}\left[\hat{g}(x)\right] - g(x)\right)^2 + \mathbb{V}\left[\hat{g}(x)\right] = b^2(x) + v(x).$$

Combining these two observations we deduce that, indeed,

$$R(g,\,\hat{g}) = \int b^2(x)dx + \int v(x)dx.$$

Exercise A.20.3 (Properties of histogram estimators)**.** Prove Theorem 20.9 where we record the expectation and variance of the histogram estimator.

Solution. Fix j and $x \in B_j$. The key observation is that $\hat{f}_n(x)$ is a multiple of the sample mean of IID Bernoulli random variables since

$$\hat{f}_n(x) = \frac{\hat{p}_j}{h} = \frac{1}{h}\cdot\frac{1}{n}\sum_{i=1}^{n}\underbrace{\mathbb{1}\left(X_i \in B_j\right)}_{=:Y_i}$$

where the Bernoulli parameter of Y is

$$\mathbb{P}(Y=1)=\mathbb{P}(X\in B_j)=\int_{B_j} f = p_j.$$

Therefore Theorem 3.5 and Exercise A.3.8 tell us that

$$\mathbb{E}\hat{f}_n(x)=\frac{1}{h}\mathbb{E}\overline{Y}_n=\frac{1}{h}\mathbb{E}Y=\frac{p_j}{h}$$

while

$$\mathbb{V}\hat{f}_n(x)=\frac{1}{h^2}\mathbb{V}\overline{Y}_n=\frac{1}{nh^2}\mathbb{V}Y=\frac{p_j(1-p_j)}{nh^2},$$

as desired.

Exercise A.20.4 (Identity for the leave-one-out cross-validation estimator of the histogram risk)**.** Prove Theorem 20.20 where we record an identity for the leave-one-out cross-validation estimator of the histogram risk.

Solution. Recall that the leave-one-out cross-validation estimator for the histogram risk is

$$\hat{J}=\int \hat{f}_n^2-\frac{2}{n}\sum_{i=1}^n \hat{f}_{(-i)}(X_i).$$

Since, by definition,

$$\hat{f}_n(x)=\sum_{j=1}^m \frac{\hat{p}_j}{h}\mathbb{1}(x\in B_j)$$

and since $(B_j)_{j=1}^m$ is a partition of $[0,1]$ with $|B_j|=h$ we may write the first term in $\hat{J}$ as

$$\int \hat{f}_n^2=\sum_{j=1}^m \frac{\hat{p}_j^2}{h^2}|B_j|=\frac{1}{h}\sum_{j=1}^m \hat{p}_j^2.$$

We now turn our attention to the second term in $\hat{J}$. We may write

$$\hat{f}_{(-i)}(x)=\sum_{j=1}^m \frac{\hat{p}_{(-i),j}}{h}\mathbb{1}(x\in B_j)$$

for

$$\begin{aligned}\hat{p}_{(-i),j} &:= \frac{1}{n-1}\sum_{\substack{l=1\\ l\neq i}}^n \mathbb{1}(X_l\in B_j)\\ &= \frac{1}{n-1}\left[\sum_{l=1}^n \mathbb{1}(X_l\in B_j)-\mathbb{1}(X_i\in B_j)\right]\\ &= \frac{n}{n-1}\hat{p}_j-\frac{1}{n-1}\mathbb{1}(X_i\in B_j).\end{aligned}$$

Therefore, since $\mathbb{1}(X_i\in B_j)^2=\mathbb{1}(X_i\in B_j)$, the second term in $\hat{J}$ is

$$-\frac{2}{n}\sum_{i=1}^n \hat{f}_{(-i)}(X_i)=-\frac{2}{nh}\sum_{i=1}^n\sum_{j=1}^m\left[\frac{n}{n-1}\hat{p}_j\mathbb{1}(X_i\in B_j)-\frac{1}{n-1}\mathbb{1}(X_i\in B_j)\right].$$

Now we observe that, since $X_i \in [0,\, 1]$ and since $(B_j)_{j=1}^m$ partitions $[0,\, 1]$,

$$\frac{1}{n}\sum_{i=1}^n \sum_{j=1}^m \mathbb{1}\left(X_i \in B_j\right) = \frac{1}{n}\sum_{i=1}^n 1 = 1.$$

By definition of $\hat{p}_j$ we also observe that

$$\frac{1}{n}\sum_{i=1}^n \hat{p}_j \mathbb{1}\left(X_i \in B_j\right) = \hat{p}_j^2.$$

Plugging these two observations into the expression above for the second term in $\hat{J}$ we deduce that

$$-\frac{2}{n}\sum_{i=1}^n \hat{f}_{(-i)}(X_i) = -\frac{2n}{(n-1)h}\sum_{j=1}^m \hat{p}_j + \frac{2}{(n-1)h}.$$

So finally we put the two terms of $\hat{J}$ back together and conclude that

$$\begin{aligned}
\hat{J} &= \frac{1}{h}\sum_{j=1}^m \hat{p}_j - \frac{2n}{(n-1)h}\sum_{j=1}^m \hat{p}_j + \frac{2}{(n-1)h}\\
&= \left(1 - \frac{2n}{n-1}\right)\frac{1}{h}\sum_{j=1}^m \hat{p}_j + \frac{2}{(n-1)h}\\
&= -\frac{n+1}{n-1}\cdot\frac{1}{h}\sum_{j=1}^m \hat{p}_j^2 + \frac{2}{(n-1)h},
\end{aligned}$$

as desired.

Exercise A.20.5 (Properties of the leave-one-out cross-validation estimator of the kernel density estimator risk)**.** Prove Theorem 20.38 where we record the consistency and an approximate identiy for the leave-one-out cross-validation estimator of the kernel density estimator risk.

Solution. (1) It suffices to show that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \hat{f}_{(-i)}(X_i)\right] = \mathbb{E}\left[\int \hat{f}_n f\right].$$

Moreover, since

$$\int \hat{f}_n f = \frac{1}{n}\sum_{i=1}^n \int \frac{1}{h} K\left(\frac{x - X_i}{h}\right) f(x)dx$$

and

$$\hat{f}_{(-i)}(X_i) = \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^n \frac{1}{h}K\left(\frac{X_i - X_j}{h}\right),$$

it suffices to show that

$$\mathbb{E}\left[\frac{1}{h}K\left(\frac{X_i - X_j}{h}\right)\right] = \mathbb{E}\left[\int \frac{1}{h}K\left(\frac{x - X_i}{h}\right) f(x)dx\right]$$

for all $i \neq j$. Since $X_i, X_j \sim f$ this follows immediately from the Rule of the Lazy Statistician:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{h}K\left(\frac{X_i - X_j}{h}\right)\right] &= \int\int \frac{1}{h}K\left(\frac{x-y}{h}\right) f(x)f(y)dxdy \\ &= \mathbb{E}\left[\int \frac{1}{h}K\left(\frac{x-X_i}{h}\right) f(x)dx\right].\end{aligned}$$

(2) Recall that

$$\hat{J}(h) = \int \hat{f}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^n \hat{f}_{(-i)}(X_i).$$

Since

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h}K\left(\frac{x-X_i}{h}\right)$$

we deduce that

$$\hat{f}_n^2(x) = \frac{1}{n^2h^2}\sum_{i,j=1}^n K\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_j}{h}\right)dx$$

where the change of variables $y = \frac{x-X_j}{h}$ and the symmetry of the kernel tell us that

$$\begin{aligned}\int K\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_j}{h}\right)dx &= \int K\left(\frac{X_j - X_i}{h} + y\right)K(y)hdy \\ &= h\int K\left(\frac{X_i - X_j}{h} - y\right)K(y)dy \\ &= hK^{(2)}\left(\frac{X_i - X_j}{h}\right)\end{aligned}$$

such that

$$\int \hat{f}_n^2(x)dx = \frac{1}{n^2h}\sum_{i,j=1}^n K^{(2)}\left(\frac{X_i - X_j}{h}\right).$$

Now, since

$$\begin{aligned}\hat{f}_{(-i)}(x) &= \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^n \frac{1}{h}K\left(\frac{x-X_j}{h}\right) \\ &= \frac{1}{n-1}\sum_{j=1}^n \frac{1}{h}K\left(\frac{x-X_j}{h}\right) - \frac{1}{(n-1)h}K\left(\frac{x-X_i}{h}\right)\end{aligned}$$

we deduce that

$$\begin{aligned}&-\frac{2}{n}\sum_{i=1}^n \hat{f}_{(-i)}(X_i) \\ &= \frac{-2}{(n-1)nh}\sum_{i,j=1}^n K\left(\frac{X_i - X_j}{h}\right) + \frac{2}{(n-1)h}\cdot \underbrace{\frac{1}{n}\sum_{i=1}^n K(0)}_{K(0)}\end{aligned}$$

So finally

$$\begin{aligned}\hat{J}(h) &= \int \hat{f}_n(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i) \\ &= \frac{1}{n^2h}\sum_{i,j=1}^{n} K^{(2)}\left(\frac{X_i - X_j}{h}\right) \\ &\quad - \frac{2}{(n-1)nh}\sum_{i,j=1}^{n} K\left(\frac{X_i - X_j}{h}\right) + \frac{2}{(n-1)h}K(0)\end{aligned}$$

where

$$\frac{1}{(n-1)n} = \frac{1}{n^2} + \frac{1}{n^2(n-1)} \text{ and } \frac{1}{n-1} = \frac{1}{n} + \frac{1}{n(n-1)}$$

such that

$$\begin{aligned}&\hat{J}(h) \\ &= \frac{1}{n^2h}\sum_{i,j=1}^{n}\left[K^{(2)}\left(\frac{X_i - X_j}{h}\right) - 2K\left(\frac{X_i - X_j}{h}\right)\right] + \frac{2}{nh}K(0) \\ &\quad - \frac{2}{n^2(n-1)h}\sum_{i,j=1}^{n} K\left(\frac{X_i - X_j}{h}\right) + \frac{2}{n(n-1)h}K(0) \\ &= \frac{1}{n^2h}\sum_{i,j=1}^{n} K^{*}\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0) + \mathcal{O}\left(\frac{1}{n^2h}\right).\end{aligned}$$

Exercise A.20.6 (Histogram regression estimator)**.** Consider regression data

$$(X_1, Y_1), \ldots, (X_n, Y_n).$$

Suppose that $0 \leqslant X_i \leqslant 1$. Define bins B_j as in Definition 20.6. For $x \in B_j$ define

$$\hat{r}_n(x) := \overline{Y}_j$$

where $\overline{Y}_j$ is the mean of all the Y_i's corresponding to those X_i's in B_j. Find the approximate risk of this estimator. From this expression for the risk, find the optimal bandwidth. At what rate does the risk go to zero?

This estimator is sometimes known as a regressogram. Note that the proof below requires that $f_X \geqslant c > 0$ on the interval $[0, 1]$ for some fixed value of c.

Solution. Recall that

$$B_j = \left[\frac{j-1}{m}, \frac{j}{m}\right) \text{ for } 1 \leqslant j \leqslant m \text{ and } B_m = \left[\frac{m-1}{m}, 1\right]$$

for some fixed integer $m \geqslant 1$. We seek to compute

$$R(r, \hat{r}_n) = \mathbb{E}\left(\int [r(x) - \hat{r}_n(x)]^2 dx\right).$$

So let us fix j and fix $x \in B_j$. The key observation is to write, for $h := 1/m$,

$$\begin{aligned}\hat{r}_n(x) &= \overline{Y}_j \\ &= \frac{\sum_{i=1}^n Y_i \mathbb{1}\left(X_i \in B_j\right)}{\sum_{i=1}^n \mathbb{1}\left(X_i \in B_j\right)} \\ &= \frac{\frac{1}{nh}\sum_{i=1}^n Y_i \mathbb{1}\left(X_i \in B_j\right)}{\frac{1}{nh}\sum_{i=1}^n \mathbb{1}\left(X_i \in B_j\right)} \\ &=: \frac{\hat{g}_n(x)}{\hat{f}_n(x)}\end{aligned}$$

where, for $x \in B_j$,

$$\hat{f}_n(x) = \frac{1}{h}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left(X_i \in B_j\right)\right)$$

is the histogram estimator of the marginal PDF f_X of X and

$$\hat{g}_n(x) = \frac{1}{h}\left(\frac{1}{n}\sum_{i=1}^n Y_i \mathbb{1}\left(X_i \in B_j\right)\right)$$

is an estimator of rf_X.

First we record some properties of $\hat{f}_n$ and $\hat{g}_n$. Recall that

$$\hat{f}_n(x) = \frac{\hat{p}_j}{h} \text{ where } \hat{p}_j := \frac{1}{n}\sum_{i=1}^n \underbrace{\mathbb{1}\left(X_i \in B_j\right)}_{=:C_i^j}.$$

Crucially: $C_1^j, \ldots, C_n^j$ are IID Bernoulli(p_j) random variables where

$$p_j = \mathbb{P}\left(C^j = 1\right) = \mathbb{P}\left(X \in B_j\right) = \int_{B_j} f.$$

Therefore

$$\mathbb{E}C^j = p_j \text{ and } \mathbb{V}C^j = p_j(1-p_j)$$

such that

$$\mathbb{E}\hat{f}_n(x) = \frac{\mathbb{E}C^j}{h} = \frac{p_j}{h} \text{ and } \mathbb{V}\hat{f}_n(x) = \frac{\mathbb{V}C^j}{nh^2} = \frac{p_j(1-p_j)}{nh^2}.$$

Similarly we may write

$$\hat{g}_n(x) = \frac{1}{h}\left(\frac{1}{n}\sum_{i=1}^n \underbrace{Y_i \mathbb{1}\left(X_i \in B_j\right)}_{Y_i C_i^j}\right)$$

where

$$\begin{aligned}\mathbb{E}\left[YC^j\right] &= \mathbb{E}\left[YC^j \,\middle|\, C^j = 1\right]\mathbb{P}\left(C^j = 1\right) + \mathbb{E}\left[YC^j \,\middle|\, C^j = 0\right]\mathbb{P}\left(C^j = 0\right) \\ &= \mathbb{E}\left(Y \,\middle|\, C^j = 1\right) \cdot p_j + 0 \cdot (1-p_j) \\ &= p_j \mathbb{E}\left(Y \mid X \in B_j\right) \\ &= p_j r_j\end{aligned}$$

for $r_j := \mathbb{E}\left(Y \mid X \in B_j\right)$ and, by the law of total variance,

$$\mathbb{V}\left[YC^j\right] = \mathbb{E}\left[\mathbb{V}\left(YC^j \,\middle|\, C^j\right)\right] + \mathbb{V}\left[\mathbb{E}\left(YC^j \,\middle|\, C^j\right)\right]$$

where

$$\mathbb{V}\left(YC^j \,\middle|\, C^j\right) = \begin{cases} \mathbb{V}\left(Y \,|\, X \in B_j\right) =: \sigma_j^2 \text{ with probability } p_j \text{ and} \\ \mathbb{V}\left(0 \,|\, X \notin B_j\right) = 0 \text{ with probability } 1 - p_j, \end{cases}$$

i.e. it is σ_j^2 times a Bernoulli(p_j) random variable, while

$$\mathbb{E}\left(YC^j \,\middle|\, C^j\right) = \begin{cases} r_j \text{ with probability } p_j \text{ and} \\ 0 \text{ with probability } 1 - p_j, \end{cases}$$

i.e. it is r_j times a Bernoulli(p_j) random variable, such that finally

$$\mathbb{V}\left[YC^j\right] = \sigma_j^2 p_j + r_j^2 p_j (1 - p_j).$$

Therefore

$$\mathbb{E}\hat{g}_n(x) = \frac{p_j r_j}{h} \text{ and } \mathbb{V}\hat{g}_n(x) = \frac{1}{nh^2}\left[\sigma_j^2 p_j + r_j^2 p_j (1 - p_j)\right].$$

Since we can write $\hat{r}_n$ in terms of $\hat{f}_n$ and $\hat{g}_n$ and since we have computed the means of $\hat{f}_n$ and $\hat{g}_n$ we may take *inspiration* from the bias–variance decomposition and write

$$\begin{aligned} \hat{r}_n - r &= \frac{\hat{g}_n}{\hat{f}_n} - \frac{r f_X}{f_X} \\ &= \frac{\hat{g}_n}{\hat{f}_n} - \frac{p_j r_j / h}{p_j / h} + \frac{p_j r_j / h}{p_j / h} - \frac{r f_X}{f_X} \\ &= \frac{\hat{g}_n}{\hat{f}_n} - r_j + r_j - r. \end{aligned}$$

The term $\frac{\hat{g}_n}{\hat{f}_n}$ is essentially the variance while the term $r_j - r$ is essentially the bias.

We turn our attention first to the bias term $r_j - r(x)$. We observe that

$$r(x) = \mathbb{E}\left(Y \,|\, X = x\right) = \int y f(y|x) dy = \int y \frac{f(x,y)}{f_X(x)} dy$$

while

$$\begin{aligned} r_j = \mathbb{E}\left(Y \,|\, X \in B_j\right) = \int y f\left(y \,|\, X \in B_j\right) dy &= \int y \frac{\int_{B_j} f(x,y) dx}{\int_{B_j} f_X} dy \\ &= \int y \frac{⨍_{B_j} f(x,y) dx}{⨍_{B_j} f_X} dy \end{aligned}$$

where $⨍_{B_j} = \frac{1}{|B_j|} \int_{B_j} = \frac{1}{h} \int_{B_j}$ is the *average* integral over B_j. Therefore

$$r_j - r(x) = \int y \left[\frac{⨍_{B_j} f(x,y) dx}{⨍_{B_j} f_X} - \frac{f(x,y)}{f_X(x)} \right] dy.$$

Since

$$\frac{a'}{b'} - \frac{a}{b} = \frac{a'b - ab'}{bb'} = \frac{a' - a}{b'} + \frac{a(b - b')}{bb'}$$

we may write

$$\begin{aligned}
&r_j - r(x)\\
&= \int y \left[\frac{\fint_{B_j} f(x,y)dx - f(x,y)}{\fint_{B_j} f_X} + \frac{f(x,y)\left(f_X(x) - \fint_{B_j} f_X\right)}{f_X(x)\fint_{B_j} f_X} \right] dy\\
&= \frac{1}{\fint_{B_j} f_X}\left[\int\fint_{B_j} yf(x,y)dxdy - \int yf(x,y)dy\right] + \frac{f_X(x) - \fint_{B_j} f_X}{f_X(x)\fint_{B_j} f_X}\int yf(x,y)dy
\end{aligned}$$

where

$$\int yf(x,y)dy = \int y\frac{f(x,y)}{f_X(x)}f_X(x)dy = f_X(x)\int yf(y|x)dy = f_X(x)r(x)$$

and so

$$\int\fint_{B_j} yf(x,y)dy = \fint_{B_j} f_X r$$

such that

$$r_j - r(x) = \frac{1}{\fint_{B_j} f_X}\left[\fint_{B_j} rf_X - r(x)f_X(x)\right] + \frac{r(x)\left(f_X(x) - \fint_{B_j} f_X\right)}{\fint_{B_j} f_X}.$$

Wasserman ([Was10], p. 308) tells us that, for any sufficiently regular function d and any $x \in B_j$,

$$d(x) - \fint_{B_j} d \approx d'(x)\left[h\left(j - \frac{1}{2}\right) - x\right].$$

Therefore

$$\begin{aligned}
r_j - r(x) &= \frac{1}{\fint_{B_j} f_X}\left[-\left(rf_X\right)'(x) + r(x)f_X'(x)\right]\left[h\left(j - \frac{1}{2}\right) - x\right]\\
&= \frac{-r'(x)f_X(x)}{\fint_{B_j} f_X}\left[h\left(j - \frac{1}{2}\right) - x\right].
\end{aligned}$$

Wasserman ([Was10], also p. 308) then tells us that

$$\int_{B_j} d(x)^2\left[h\left(j - \frac{1}{2}\right) - x\right]^2 dx \approx d(\tilde{x}_j)^2\frac{h^3}{12}$$

for $\tilde{x}_j = \frac{j-1/2}{m}$ the center of the bin B_j. Therefore, since

$$f_X(\tilde{x}_j) \approx \fint_{B_j} f \text{ as } h \to 0,$$

we deduce that

$$\int_{B_j}\left[r_j - r(x)\right]^2 dx \approx r'(\tilde{x}_j)^2\frac{h^3}{12}.$$

Finally we proceed as in [Was10] (you guessed it, p. 308) and conclude that, for $\bar{r} := \sum_{j=1}^m r_j \mathbb{1}_{B_j}$,

$$\int_0^1\left[\bar{r}(x) - r(x)\right]^2 dx = \sum_{j=1}^m\int_{B_j}\left[r_j - r(x)\right]^2 dx = \frac{h^2}{12}\sum_{j=1}^m hr'(\tilde{x}_j)^2 \approx \frac{h^2}{12}\int_0^1 (r')^2.$$

We now turn our attention to the variance. Recall that, for $x \in B_j$,

$$\mu_f := \mathbb{E}\hat{f}_n(x) = \frac{p_j}{h}, \qquad \sigma_f^2 := \mathbb{V}\hat{f}_n(x) = \frac{p_j(1-p_j)}{nh^2},$$
$$\mu_g := \mathbb{E}\hat{g}_n(x) = \frac{p_j r_j}{h}, \qquad \sigma_g^2 := \mathbb{V}\hat{g}_n(x) = \frac{\sigma_j^2 p_j + r_j^2 p_j (1-p_j)}{nh^2},$$

and, by independence,

$$\begin{aligned}\operatorname{Cov}\left(\hat{f}_n, \hat{g}_n\right) = \frac{1}{n^2h^2}\sum_{i,l=1}^{n} \operatorname{Cov}\left(C_i^j, Y_l C_l^j\right) &= \frac{1}{n^2h^2}\sum_{i,l=1}^{n} \mathbb{V}\left(Y_i C_i^j\right)\delta_{il} \\ &= \frac{1}{nh^2}\mathbb{V}\left(YC^j\right) = \sigma_g^2.\end{aligned}$$

Therefore Lemma 3.22 tells us that, using also the fact that $1 - p_j \to 1$ as $h \to 0$,

$$\begin{aligned}\mathbb{E}\left(\frac{\hat{g}_n}{\hat{f}_n}\right) &\approx \frac{\mu_g}{\mu_f} - \frac{\sigma_g^2}{\mu_f^2} + \frac{\sigma_g^2 \mu_g}{\mu_f^3} \\ &= \frac{1}{\mu_f^3}\left[\mu_g \mu_f^2 + \sigma_g^2\left(\mu_g - \mu_f\right)\right] \\ &\approx \frac{h^3}{p_j^3}\left[\frac{p_j r_j}{h} \cdot \frac{p_j^2}{h^2} + \frac{\sigma_j^2 p_j + r_j^2 p_j}{nh^2}\left(\frac{p_j r_j}{h} - \frac{p_j}{h}\right)\right] \\ &= r_j + \frac{\left(\sigma_j^2 + r_j^2\right)}{np_j}\left(r_j - 1\right) \\ &= r_j + \frac{1}{nh} \cdot \frac{\sigma_j^2 + r_j^2}{p_j/h}\left(r_j - 1\right)\end{aligned}$$

while

$$\begin{aligned}\mathbb{V}\left(\frac{\hat{g}_n}{\hat{f}_n}\right) &\approx \frac{\mu_g^2}{\mu_f^2}\left(\frac{\sigma_g^2}{\mu_g^2} + \frac{\sigma_f^2}{\mu_f^2} - \frac{2\sigma_g^2}{\mu_f \mu_g}\right) \\ &\approx \frac{p_j^2 r_j^2}{h^2} \cdot \frac{h^2}{p_j^2}\left(\sigma_g^2 \cdot \frac{h^2}{p_j^2 r_j^2} + \frac{p_j}{nh^2} \cdot \frac{h^2}{p_j^2} - 2\sigma_g^2 \cdot \frac{h}{p_j} \cdot \frac{h}{p_j r_j}\right) \\ &= r_j^2\left[\sigma_g^2 \cdot \frac{h^2}{p_j^2}\left(\frac{1}{r_j^2} - \frac{2}{r_j}\right) + \frac{1}{np_j}\right] \\ &= r_j^2\left[\frac{\sigma_j^2 p_j + r_j^2 p_j}{nh^2} \cdot \frac{h^2}{p_j^2}\left(\frac{1 - 2r_j}{r_j^2}\right) + \frac{1}{np_j}\right] \\ &= \frac{\sigma_j^2 + r_j^2}{np_j}\left(1 - 2r_j\right) + \frac{r_j^2}{np_j} \\ &= \frac{1}{np_j}\left[\sigma_j^2\left(1 - 2r_j\right) + 2r_j^2\left(1 - r_j\right)\right].\end{aligned}$$

Since, for $x \in B_j$,

$$p_j \approx hf_X(x),\ r_j \approx r(x),\ \text{and } \sigma_j^2 \approx \mathbb{V}\left(Y \mid X = x\right) =: \sigma^2(x)$$

we deduce that

$$\mathbb{V}\left(\frac{\hat{g}_n}{\hat{f}_n}\right) \approx \frac{1}{nhf_X(x)} \left(\sigma^2(x)\left[1-2r(x)\right] + 2r^2(x)\left[1-r(x)\right]\right)$$

and so

$$\int_0^1 \mathbb{V}\left(\frac{\hat{g}_n}{\hat{f}_n}\right) dx \approx \frac{1}{nh}\int_0^1 \frac{\sigma^2(1-2r)+2r^2(1-r)}{f_X} dx.$$

We are now ready to conclude. We note that the decomposition, for $x \in B_j$,

$$\hat{r}_n(x) - r(x) = \frac{\hat{g}_n(x)}{\hat{f}_n(x)} - r_j + r_j - r(x)$$

is *not* a true bias-variance decomposition since

$$\mu(x) := \mathbb{E}\left[\frac{\hat{g}_n(x)}{\hat{f}_n(x)}\right] \neq r_j.$$

Thankfully, however, it is *approximately* a bias-variance decomposition since the computation of the mean of $\hat{g}_n(x)/\hat{f}_n(x)$ above showed that the remainder

$$\mu(x) - r_j$$

is smaller than the leading-order terms. More precisely, we have that, if we define as we did above $\bar{r} := \sum_{j=1}^m r_j \mathbb{1}_{B_j}$, then

$$\begin{aligned} R\left(r, \hat{r}_n\right) &= \mathbb{E}\left(\int \left[\hat{r}_n(x) - r(x)\right]^2 dx\right) \\ &= \mathbb{E}\left(\int \left[\frac{\hat{g}_n(x)}{\hat{f}_n(x)} - \mu(x) + \mu(x) - \bar{r}(x) + \bar{r}(x) - r(x)\right]^2 dx\right). \end{aligned}$$

Since $\mu(x) = \mathbb{E}\left(\hat{g}_n(x)/\hat{f}_n(x)\right)$ and since mean-zero random variables are orthogonal to constants (which is the fact underpinning the bias–variance decomposition theorem) this simplifies to

$$\begin{aligned} R\left(r, \hat{r}_n\right) &= \int \left(\mathbb{E}\left[\frac{\hat{g}_n(x)}{\hat{f}_n(x)} - \mu(x)\right]^2\right) dx + \int \left[\underbrace{\mu(x) - \bar{r}(x)}_{=:\tilde{b}(x)} + \underbrace{\bar{r}(x) - r(x)}_{=:b(x)}\right]^2 dx \\ &= \int \underbrace{\mathbb{V}\left[\frac{\hat{g}_n(x)}{\hat{f}_n(x)}\right]}_{=:v(x)} dx + \int \tilde{b}^2 + 2\int \tilde{b}b + \int b^2. \end{aligned}$$

We handle each term one at a time. We have computed above that

$$\int_0^1 v \approx \frac{1}{nh}\int \frac{\sigma^2(1-2r)+2r^2(1-r)}{f_X},$$

that, for $x \in B_j$,

$$\tilde{b}(x) = \frac{1}{nh}\cdot\frac{\sigma_j^2 + r_j^2}{p_j/h}(r_j - 1) \approx \frac{1}{nh}\cdot\frac{\sigma^2(x)+r^2(x)}{f_X(x)}\left[r(x)-1\right],$$

such that crucially

$$\tilde{b}(x) \in O\left(\frac{1}{nh}\right),$$

and we have computed that

$$\int b^2 \approx \frac{h^2}{12}\int_0^1 (r')^2.$$

In particular we note that the terms

$$\int \tilde{b}^2 \in O\left(\frac{1}{n^2h^2}\right) \text{ and } \int \tilde{b}b \in O\left(\frac{h}{nh}\right) = O\left(\frac{1}{n}\right)$$

decay faster than the leading orders $\frac{1}{nh}$ and h, so we may discard them. We are left with

$$R\left(r,\, \hat{r}_n\right) \approx \frac{1}{nh}\int_0^1 \frac{\sigma^2(1-2r)+2r^2(1-r)}{f_X} + \frac{h^2}{12}\int_0^1 (r')^2.$$

Since the bias and variance scale at leading-order just like the bias and variance of the histogram estimator we deduce as in Theorem 20.10 that the MISE is minimized with respect to h when h is of order $n^{-1/3}$, in which case the MISE decays like $n^{-2/3}$. This is slower than the decay of the MISE for *kernel* regression estimators, namely slower than the Nadaraya-Watson kernel estimator (which decays like $n^{-4/5}$).

Exercise A.20.7 (Consistency of the estimate of the variance for Nadaraya-Watson kernel estimation)**.** Show that with suitable smoothness assumptions on $r(x)$, $\hat{\sigma}^2$ as defined in Theorem 20.51 is a consistent estimator of

$$\sigma^2 := \mathbb{V}\left(Y - r(X)\right).$$

It turns out that we need to assume that

(1) the marginal PDF of X satisfies

$$\frac{\mathbb{E}\left(\max X_i - \min X_i\right)}{n} \to 0 \text{ as } n \to \infty$$

for any IID sample $X_1, \ldots, X_n$ and that

(2) the regression function is differentiable with bounded derivative.

Solution. As in Theorem 20.51 we order $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ and label $\widetilde{Y}_1, \ldots, \widetilde{Y}_n$ accordingly, i.e.

$$\widetilde{Y}_j = Y_i \iff X_{(j)} = X_i.$$

Defining $\varepsilon_i := Y_i - r(X_i)$ such that

$$\mathbb{V}\varepsilon_i = \sigma^2$$

we obtain that

$$\mathbb{V}\left(\widetilde{Y}_{i+1} - \widetilde{Y}_i\right) = \mathbb{V}\left(\varepsilon_{i+1}\right) + \mathbb{V}\left(\varepsilon_i\right) + \mathbb{V}\left[r\left(X_{(i)}\right) - r\left(X_{(i+1)}\right)\right].$$

Therefore, since $\mathbb{E}\left(\widetilde{Y}_{i+1}-\widetilde{Y}_i\right)=0$,

$$\begin{aligned}\mathbb{E}\hat{\sigma}^2 &= \frac{1}{2(n-1)}\sum_{i=1}^{n-1}\mathbb{E}\left[\left(\widetilde{Y}_{i+1}-\widetilde{Y}_i\right)^2\right]\\ &= \frac{1}{2(n-1)}\sum_{i=1}^{n-1}\mathbb{V}\left(\widetilde{Y}_{i+1}-\widetilde{Y}_i\right)\\ &= \sigma^2+\underbrace{\frac{1}{2(n-1)}\sum_{i=1}^{n-1}\mathbb{V}\left[r\left(X_{(i+1)}\right)-r\left(X_{(i)}\right)\right]}_{=:b_n.}\end{aligned}$$

So now we show that the bias satisfies $b_n\to 0$ as $n\to\infty$. If we assume that r is differentiable with bounded derivative then then mean value theorem tells us that, since $\mathbb{E}r\left(X_{(i+1)}\right)=\mathbb{E}r\left(X_{(i)}\right)$,

$$\begin{aligned}\mathbb{V}\left[r\left(X_{(i+1)}\right)-r\left(X_{(i)}\right)\right] &= \mathbb{E}\left(\left[r\left(X_{(i+1)}\right)-r\left(X_{(i)}\right)\right]^2\right)\\ &= \mathbb{E}\left[r'\left(\xi_i\right)\left(X_{(i+1)}-X_{(i)}\right)^2\right]\end{aligned}$$

for some random variables ξ_i satisfying $X_{(i)}\leqslant\xi_i\leqslant X_{(i+1)}$. Therefore, since $\sum_i|\alpha|^2\leqslant\left(\sum_i|\alpha_i|\right)^2$ and since $\mathbb{E}$ is linear and monotone,

$$\begin{aligned}|b_n| &\leqslant \frac{1}{2(n-1)}||r'||_\infty^2\sum_{i=1}^{n-1}\mathbb{E}\left[|X_{(i+1)}-X_{(i)}|^2\right]\\ &\leqslant \frac{1}{2(n-1)}||r'||_\infty^2\mathbb{E}\left[\left(\sum_{i=1}^{n-1}|X_{(i+1)}-X_{(i)}|\right)^2\right]\\ &= \frac{||r'||_\infty^2}{2(n-1)}\mathbb{E}\left(|X_{(n)}-X_{(1)}|^2\right).\end{aligned}$$

As long as

$$\frac{\mathbb{E}\left(|X_{(n)}-X_{(1)}|^2\right)}{n}\to 0 \text{ as } n\to\infty,$$

which is a property of the marginal PDF of X (guaranteed if that marginal PDF is compactly supported, for example) and *not* of the regression function r, then the bias satisfies $b_n\to 0$ as $n\to\infty$ and so $\hat{\sigma}^2$ is a consistent estimator of σ^2. Note that this condition on the marginal PDF is required even if r is the identity, so this condition is *not* an artefact of our approach here.

Exercise A.20.8 (Identity for the leave-one-out cross-validation estimator of the Nadaraya-Watson kernel estimator risk)**.** Prove Theorem 20.50 where we record an identity for the leave-one-out cross-validation estimator of the Nadaraya-Watson kernel estimator risk.

Solution. This follows from a straightforward computation. First we may write the *full* estimator

$$\hat{r}(X_i)=\sum_{j=1}^n w_j(X_i)Y_j=\frac{\sum_{j=1}^n Y_jK\left(\frac{X_i-X_j}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_i-X_j}{h}\right)}$$

such that

$$Y_i - \hat{r}(X_i) = \frac{\sum_{j=1}^{n} (Y_i - Y_j) K\left(\frac{X_i - X_j}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)} =: \frac{A}{B}.$$

We may then write similarly the leave-one-out cross-validation estimator as

$$\hat{r}_{(-i)}(X_i) = \sum_{j \neq i} w_{(-i),j}(X_i) Y_j = \frac{\sum_{j \neq i} Y_j K\left(\frac{X_i - X_j}{h}\right)}{\sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right)}$$

such that, since $Y_i - Y_j = 0$ when $i = j$,

$$\begin{aligned} Y_i - \hat{r}_{(-i)}(X_i) &= \frac{\sum_{j \neq i} (Y_i - Y_j) K\left(\frac{X_i - X_j}{h}\right)}{\sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right)} \\ &= \frac{\sum_{j=1}^{n} (Y_i - Y_j) K\left(\frac{X_i - X_j}{h}\right)}{\left[\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)\right] - K(0)} = \frac{A}{B - K(0)}. \end{aligned}$$

It then follows that

$$Y_i - \hat{r}_{(-i)}(X_i) = \frac{A}{B - K(0)} = \frac{A/B}{1 - \frac{K(0)}{B}} = \frac{Y_i - \hat{r}(X_i)}{1 - \frac{K(0)}{\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)}},$$

as desired. Squaring and summing over $1 \leqslant i \leqslant n$ then yields the result.

A.21. **Smoothing Using Orthogonal Functions.**

Exercise A.21.1 (MISE of the orthogonal function density estimator)**.** Prove Theorem 21.17 where we record the MISE (also known as the *risk*) of the orthogonal function density estimator.

Solution. The basis expansion of $f - \hat{f}$ is

$$f - \hat{f} = \sum_{j=1}^{\infty} \beta_j \phi_j - \sum_{j=1}^{J} \hat{\beta}_j \phi_j = \sum_{j=1}^{J} \left(\beta_j - \hat{\beta}_j\right) \phi_j + \sum_{j=J+1}^{\infty} \beta_j \phi_j.$$

Since Theorem 21.13 tells us that

$$\mathbb{E}\hat{\beta}_j = \beta_j \text{ and } \mathbb{V}\hat{\beta}_j = \frac{\sigma_j^2}{n}$$

it then follows from Parseval's identity (see Theorem 21.7) that

$$\begin{aligned} R\left(f, \hat{f}\right) &= \mathbb{E}\left[\int \left(f - \hat{f}\right)^2\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{J} \left(\beta_j - \hat{\beta}_j\right)^2 + \sum_{j=J+1}^{\infty} \beta_j^2\right] \\ &= \sum_{j=1}^{J} \mathbb{V}\hat{\beta}_j + \sum_{j=J+1}^{\infty} \beta_j^2 \\ &= \sum_{j=1}^{J} \frac{\sigma_j^2}{n} + \sum_{j=J+1}^{\infty} \beta_j^2, \end{aligned}$$

as desired.

Exercise A.21.2 (MISE of the orthogonal function regression estimator)**.** Prove Theorem 21.28 where we record the MISE (also known as the *risk*) of the orthogonal function regression estimator.

Solution. Theorem 21.25 tells us that

$$\mathbb{E}\hat{\beta}_j = \beta_j \text{ and } \mathbb{V}\hat{\beta}_j = \frac{\sigma^2}{n}.$$

The result then follows from combining these identities with Parseval's identity (see Theorem 21.7):

$$\begin{aligned} R\left(r, \hat{r}\right) &= \mathbb{E}\left[\int \left(r - \hat{r}\right)^2\right] \\ &= \mathbb{E}\left[\sum_{j=0}^{J} \left(\beta_j - \hat{\beta}_j\right)^2 + \sum_{j=J+1}^{\infty} \beta_j^2\right] \\ &= \sum_{j=0}^{J} \mathbb{V}\hat{\beta}_j + \sum_{j=J+1}^{\infty} \beta_j^2 \\ &= \frac{J\sigma^2}{n} + \sum_{j=J+1}^{\infty} \beta_j^2, \end{aligned}$$

as desired.

Exercise A.21.3 (Orthonormal vectors in $\mathbb{R}^3$)**.** Let

$$\psi_1 = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right), \psi_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right), \text{ and } \psi_3 = \left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}\right).$$

Show that these vectors have norm 1 and are orthogonal.

Solution. First we compute the norms:

$$|\psi_1|^2 = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1, \ |\psi_2|^2 = \frac{1}{2} + \frac{1}{2} = 1, \text{ and } |\psi_3|^2 = \frac{1}{6} + \frac{1}{6} + \frac{4}{6} = 1.$$

Then we compute the inner products:

$$\begin{aligned} \psi_1 \cdot \psi_2 &= \frac{1}{\sqrt{6}}(1 - 1 + 0) = 0, \\ \psi_2 \cdot \psi_3 &= \frac{1}{\sqrt{12}}(1 - 1 + 0) = 0, \text{ and} \\ \psi_3 \cdot \psi_1 &= \frac{1}{\sqrt{18}}(1 + 1 - 2) = 0. \end{aligned}$$

Exercise A.21.4 (Parseval's identity)**.** Prove that Parseval's identity holds for any orthonormal basis of a Hilbert space (i.e. prove that item 1 implies item 2 in Theorem 21.7).

Solution. By item 1 we know that

$$||f||^2 = \lim_{n \to \inf} \left|\left|\sum_{j=1}^{n} \beta_j \phi_j\right|\right|^2$$

where, for every n, the orthonormality of the ϕ_j's tells us that

$$\left|\left|\sum_{j=1}^{n} \beta_j \phi_j\right|\right|^2 = \sum_{i,j=1}^{n} \beta_i \beta_j \underbrace{\langle \phi_i, \phi_j \rangle}_{\delta_{ij}} = \sum_{j=1}^{n} \beta_j^2$$

as desired.

Exercise A.21.5 (Cosine decompositions of functions)**.** Expand the following functions in the cosine basis on [0, 1].

(1) $f(x) = \sqrt{2}\cos(3\pi x)$.
(2) $f(x) = \sin(\pi x)$.

Solution. (1) f is one of the basis vectors, namely $f = \phi_3$. Therefore $\beta_j = \delta_{3j}$, i.e.

$$\beta_3 = 1 \text{ and } \beta_j = 0 \text{ for } j \neq 3.$$

(2) First we compute that

$$\beta_0 = \int_0^1 f(x)dx = \int_0^1 \sin(\pi x)dx = \frac{-\cos(\pi x)}{\pi}\bigg|_{x=0}^{x=1} = \frac{2}{\pi}.$$

Then, for $j \geqslant 1$,

$$\begin{aligned}
\int_0^1 f(x)\cos(j\pi x) &= \int_0^1 \sin(\pi x)\cos(j\pi x)dx \\
&= \frac{-1}{\pi}\cos(\pi x)\cos(j\pi x)\Big|_{x=0}^{x=1} \\
&\quad - \int_0^1 \left(\frac{-1}{\pi}\right)\cos(\pi x)(-j\pi)\sin(j\pi x)dx \\
&= \frac{1+\cos(j\pi)}{\pi} - j\int_0^1 \cos(\pi x)\sin(j\pi x)dx \\
&= \frac{1+\cos(j\pi)}{\pi} - \frac{j}{\pi}\sin(\pi x)\sin(j\pi x)\Big|_{x=0}^{x=1} \\
&\quad + j^2\int_0^1 \frac{1}{\pi}\sin(\pi x)(j\pi)\cos(j\pi x)dx \\
&= \frac{1+\cos(j\pi)}{\pi} - 0 + j^2\int_0^1 \sin(\pi x)\cos(j\pi x)dx.
\end{aligned}$$

Since

$$\begin{aligned}
\cos(j\pi) &= \begin{cases} -1 & \text{if } j \text{ is odd and} \\ 1 & \text{if } j \text{ is even} \end{cases} \\
&= (-1)^j
\end{aligned}$$

it follows that, for $j \geqslant 1$,

$$\begin{aligned}
&\frac{\beta_j}{\sqrt{2}} = \frac{1+(-1)^j}{\pi} + j^2\frac{\beta_j}{\sqrt{2}} \\
\iff &\beta_j = \frac{\sqrt{2}\left(1+(-1)^j\right)}{\pi(1-j^2)} = \begin{cases} 0 & \text{if } j \text{ is odd and} \\ \dfrac{2\sqrt{2}}{\pi(1-j^2)} & \text{if } j \text{ is even.} \end{cases}
\end{aligned}$$

Note that $f(x) = \sin(\pi x)$ is *smooth*, i.e. infinitely differentiable, and yet its coefficients along the cosine basis decay only algebraically. More precisely:

$$|\beta_j| \sim \frac{1}{j^2},\ j|\beta_j| \sim \frac{1}{j},\ \text{and } j^2|\beta_j| \sim 1,$$

which means that

$$\beta_j \in l^2,\ j\beta_j \in l^2,\ \text{ but } j^2\beta_j \notin l^2.$$

Translating this back into the spatial domain, this means that

$$f \in L^2,\ f' \in L^2,\ \text{ but } f'' \notin L^2!$$

To understand this we must go back to Remark 21.9: the cosine basis is derived from the Fourier basis. Therefore an implicit periodization occurs and the cosine basis does not "see" the smoothness of f, but rather it "sees" the smoothness of its *periodic extension*

$$P_1\left(f|_{[0,\,1]}\right),$$

where P_1 denotes the operator mapping a function on $[0,\, 1]$ to its 1–periodic extension. Since

$$f'(x) = \frac{d}{dx}\sin(\pi x) = \frac{1}{\pi}\cos(\pi x),$$

which has a jump when looping back around from $f'(1) = -\frac{1}{\pi}$ to $f'(0) = \frac{1}{\pi}$, it follows that

$$P_1(f)' \in L^2 \text{ but } P_1(f)'' \notin L^2$$

since $P_1(f)''$ has a Dirac mass (of strength $2/\pi$) at zero. This is what the fact that $j^2\beta_j \notin l^2$ was really telling us.

Exercise A.21.6 (Orthonormality of the Haar system)**.** Show that the Haar wavelets are orthonormal.

Solution. Let ϕ be the Haar father wavelet and let $\psi_{j,k}$ and $\psi_{i,l}$ be distinct Haar children wavelets where

$$i,\, j \geqslant 0,\; 0 \leqslant k \leqslant 2^j - 1, \text{ and } 0 \leqslant l \leqslant 2^i - 1$$

and where, without loss of generality,

$$j \leqslant i.$$

First we compute that, since $\phi|_{[0,\,1]} = 1$ and since

$$\psi_{j,k} = \begin{cases} -\,2^{j/2} & \text{if } \dfrac{k}{2^j} \leqslant x \leqslant \dfrac{k+1/2}{2^j}, \\ 2^{j/2} & \text{if } \dfrac{k+1/2}{2^j} < x \leqslant \dfrac{k+1}{2^j}, \text{ and} \\ 0 & \text{otherwise,} \end{cases}$$

it follows that

$$\int_0^1 \phi\psi_{j,k} = -2^{j/2}\cdot\frac{1/2}{2^j} + 2^{j/2}\cdot\frac{1/2}{2^j} = 0,$$

which verifies that the Haar father wavelet is orthogonal to all Haar children wavelets.

Now we compute $\langle\psi_{j,k},\, \psi_{i,l}\rangle$. If $i = j$ then necessarily $k \neq l$ since these two children wavelets are distinct, and so by Remark 21.35 they have disjoint support, which ensures that their inner product vanishes.

Now suppose that $j < i$, which means that $\psi_{i,l}$ is at a finer scale than $\psi_{j,k}$. If their supports are disjoint then we are done, so assume WLOG that their supports intersect (at more than a single point), such that necessarily

$$\operatorname{supp}\psi_{i,l} \subseteq \operatorname{supp}\psi_{j,k}$$

since the support of the fine-scale wavelet is necessarily a dyadic sub-interval of the support of the coarse-scale wavelet.

Crucially: the center of the support of the coarse-scale wavelet is

$$\frac{k+\frac{1}{2}}{2^j} = \frac{2^{i-j}(k+1/2)}{2^i} = \frac{\tilde{l}}{2^i}$$

where $\tilde{l} := 2^{i-j}\left(k+\frac{1}{2}\right)$ is an integer (see also Figure A.2). So the support of the fine-scale wavelet must be entirely contained in either

$$\left[\frac{k}{2^j},\, \frac{k+\frac{1}{2}}{2^j}\right] \text{ or } \left[\frac{k+1/2}{2^j},\, \frac{k+1}{2^j}\right].$$

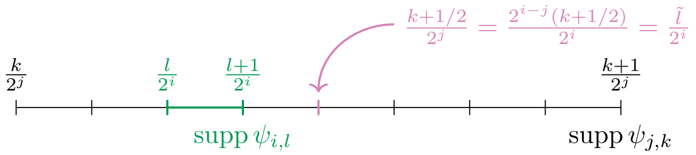


FIGURE A.2. A pictorial representation of the nested supports of a fine-scale wavelet $\psi_{i,l}$ and a coarse-scale wavelet $\psi_{j,k}$ where $j < i$.

But the coarse-scale wavelet $\psi_{j,k}$ is *constant* on these sub-intervals (since they lie to one side of the center of its support), while the fine-scale wavelet $\psi_{i,l}$ changes sign. Proceeding as in the computation of $\langle \phi,\, \psi_{j,k} \rangle$ we may thus deduce that, in either case,

$$\langle \psi_{j,k},\, \psi_{i,l} \rangle = 0.$$

This verifies that the Haar children wavelets are mutually orthogonal. To conclude the proof we simply note that the normality of the children wavelets was verified in Remark 21.35, while the normality of the father wavelet is trivial.

Exercise A.21.7 (Haar density estimation)**.** Let $X_1, \ldots, X_n \sim f$ for some density f on $[0,1]$. Let us consider constructing a wavelet histogram. Let ϕ and ψ be the Haar father and mother wavelet. Write

$$f = \phi + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \beta_{j,k} \psi_{j,k}$$

where the scaling coefficient satisfies $\alpha = 1$ since $\int_0^1 f = 1$. Let

$$\hat{\beta}_{j,k} := \frac{1}{n} \sum_{i=1}^{n} \psi_{j,k}(X_i)$$

be the sample mean of $\psi_{j,k}(X_1), \ldots, \psi_{j,k}(X_n)$.

(1) Show that $\hat{\beta}_j$ is an unbiased estimator of $\beta_{j,k}$.
(2) Define the Haar histogram

$$\hat{f} = \phi + \sum_{j=0}^{B} \sum_{k=0}^{2^j-1} \hat{\beta}_{j,k} \psi_{j,k}$$

for $0 \leqslant B \leqslant \log_2 n$. Find an approximate expression for the MISE, or risk, as a function of B.

Solution. (1) Theorem 21.37 tells us that the Haar system is an orthonormal basis. It then follows immediately from Theorem 21.13 that

$$\mathbb{E}\hat{\beta}_{j,k} = \mathbb{E}\beta_{j,k}.$$

(2) Since the Haar system is an orthonormal basis, Theorem 21.17 tells us that the risk of the Haar histogram satisfies

$$R\left(f,\, \hat{f}\right) = \sum_{j=0}^{B} \sum_{k=0}^{2^j-1} \frac{\sigma_{j,k}^2}{n} + \sum_{j=B+1}^{\infty} \sum_{k=0}^{2^j-1} \beta_{j,k}^2$$

for

$$\sigma_{j,k}^2 := \mathbb{V}\left[\psi_{j,k}(X)\right] \text{ where } X \sim f.$$

Proceeding as in Remark 21.19 we may then estimate the risk as in Definition 21.18 by

$$\hat{R}(B) := \sum_{j=0}^{B} \sum_{k=0}^{2^j-1} \frac{\sigma_{j,k}^2}{n} + \sum_{j=B+1}^{J_*} \sum_{k=0}^{2^j-1} \left(\hat{\beta}_{j,k}^2 - \frac{\hat{\sigma}_{j,k}^2}{n} \right)_+ ,$$

where $J_* = \lfloor \log_2 n \rfloor$, for

$$\hat{\sigma}_{j,k}^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left[\psi_{j,k}(X_i) - \hat{\beta}_{j,k} \right]^2$$

the sample variance of $\psi_{j,k}(X_1), \ldots, \psi_{j,k}(X_n)$.

We record a result found in [Mil] that will be used in Exercise A.21.8 below.

Theorem A.4 (Median Theorem)**.** *Let $Y_1, \ldots, Y_n$ where n is odd be IID from a PDF f with median μ such that $f(\mu) > 0$ and f is continuously differentiable in a neighbourhood of μ. Then*

$$\textit{median}\, (Y_1, \ldots, Y_n) \sim N\left(\mu, \sigma_n^2\right)$$

in distribution as $n \to \infty$ where

$$\sigma_n^2 = \frac{1}{8 f(\mu)^2 m} \textit{ for } n = 2m + 1.$$

In particular $\sigma_n \to 0$ as $n \to \infty$.

Exercise A.21.8 (Using the median of the absolute values to estimate the variance)**.** Let $X_1, \ldots, X_n \sim N\left(0, \sigma^2\right)$. Let

$$\hat{\sigma} := \frac{\text{median}\, (|X_1|, \ldots, |X_n|)}{0.6745}.$$

Show that $\mathbb{E}\hat{\sigma} = \sigma$.

NB: The denominator should more precisely be $\Phi^{-1}\left(\frac{3}{4}\right) \approx 0.6745$, where Φ denotes the CDF of a standard Normal distribution.

Solution. We will actually show that

$$\mathbb{E}\left[\frac{\text{median}\, (|X_1|, \ldots, |X_n|)}{\Phi^{-1}\left(\frac{3}{4}\right)}\right] \to \sigma \text{ as } n \to \infty.$$

Since

$$X \sim N(0, \sigma^2) \iff Z := \frac{X}{\sigma} \sim N(0, 1)$$

we may write $Z_i := \frac{X_i}{\sigma}$ and observe that

$$\mathbb{E}\, \text{median}\, (|X_1|, \ldots, |X_n|) = \sigma \mathbb{E}\left(|Z_1|, \ldots, |Z_n|\right).$$

It therefore suffices to show that

$$\mathbb{E}\, \text{median}\, (|Z_1|, \ldots, |Z_n|) \to \Phi^{-1}\left(\frac{3}{4}\right) \text{ as } n \to \infty.$$

Since the Median Theorem tells us that

$$\mathbb{E}\, \text{median}\, (|Z_1|, \ldots, |Z_n|) \to \text{median}\, (|Z|) \text{ as } n \to \infty$$

for $Z \sim N(0, 1)$, it is enough to show that

$$\text{median}\,(|Z|) = \Phi^{-1}\left(\frac{3}{4}\right).$$

This follows from a direct computation: writing $\nu :=$ median $(|Z|)$ we have that

$$\begin{aligned}\frac{1}{2} = \mathbb{P}\left(|Z| \leqslant \nu\right) = \mathbb{P}\left(-\nu \leqslant Z \leqslant \nu\right) &= \mathbb{P}\left(Z \leqslant \nu\right) - \mathbb{P}\left(Z \leqslant -\nu\right) \\ &= \Phi(\nu) - \Phi(-\nu) = 2\Phi(\nu) - 1\end{aligned}$$

and so

$$\Phi(\nu) = \frac{3}{4} \iff \nu = \Phi^{-1}\left(\frac{3}{4}\right),$$

as desired.

Exercise A.21.9 (Haar decompositions of functions)**.** Repeat Exercise A.21.5 using the Haar basis.

Solution. We begin with $f(x) = \sin(\pi x)$, using the expressions for α and $\beta_{j,k}$ recorded in Corollary 21.39. Proceeding exactly as in Exercise A.21.5 we obtain that

$$\alpha = \int_0^1 f\phi = \int_0^1 f = \frac{2}{\pi}.$$

To compute $\beta_{j,k}$ we note that

$$\psi_{j,k} = \begin{cases} -\,2^{j/2} & \text{if } \dfrac{k}{2^j} \leqslant x \leqslant \dfrac{k+1/2}{2^j}, \\ 2^{j/2} & \text{if } \dfrac{k+1/2}{2^j} < x \leqslant \dfrac{k+1}{2^j}, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, since $\int \sin \pi x = -\frac{\cos \pi x}{\pi}$,

$$\begin{aligned}\beta_{j,k} &= \int_0^1 f\psi_{j,k} \\ &= -2^{j/2}\int_{k/2^j}^{(k+1/2)/2^j} f + 2^{j/2}\int_{(k+1/2)/2^j}^{(k+1)/2^j} f \\ &= 2^{j/2}\frac{\cos \pi x}{\pi}\bigg|_{k/2^j}^{(k+1/2)/2^j} - 2^{j/2}\frac{\cos \pi x}{\pi}\bigg|_{(k+1/2)/2^j}^{(k+1)/2^j} \\ &= \frac{2^{j/2}}{\pi}\left[-\cos\frac{\pi(k+1)}{2^j} + 2\cos\frac{\pi(k+1/2)}{2^j} - \cos\frac{\pi k}{2^j}\right].\end{aligned}$$

We now consider $f(x) = \sqrt{2}\cos(3\pi x)$, which is no longer a basis function (by contrast with Exercise A.21.5), and so we must suffer through the computation. First we compute that

$$\alpha = \int_0^1 \sqrt{2}\cos(3\pi x)dx = \frac{\sqrt{2}}{2\pi}\sin(3\pi x)\Big|_{x=0}^{x=1} = 0.$$

Then

$$\begin{aligned}\beta_{j,k} &= -2^{j/2}\int_{k/2^j}^{(k+1/2)/2^j} f + 2^{j/2}\int_{(k+1/2)/2^j}^{(k+1)/2^j} f\\ &= -2^{j/2}\cdot\frac{\sqrt{2}}{3\pi}\sin(3\pi x)\bigg|_{k/2^j}^{(k+1/2)/2^j} + 2^{j/2}\cdot\frac{\sqrt{2}}{3\pi}\sin(3\pi x)\bigg|_{(k+1/2)/2^j}^{(k+1)/2^j}\\ &= \frac{2^{j/2}\sqrt{2}}{3\pi}\left[\sin\frac{3\pi(k+1)}{2^j} - 2\sin\frac{3\pi(k+1/2)}{2^j} + \sin\frac{3\pi k}{2^j}\right].\end{aligned}$$

A.22. **Classification.**

Exercise A.22.1 (The Bayes classification rule is optimal)**.** Prove Theorem 22.11 where we establish that the Bayes classification rule is optimal for binary classification.

Solution. For any classification rule h we may use the rule of iterated expectation to write its true error rate as

$$L(h) = \mathbb{P}\left(h(X) \neq Y\right) = \mathbb{E}\left[\mathbb{P}\left(h(X) \neq Y \,|\, X\right)\right] = \mathbb{E}\left[l(X; h)\right]$$

where, for any $x \in \mathcal{X}$,

$$l(x; h) := \mathbb{P}\left(h(x) \neq Y \,|\, X = x\right)$$

denotes the *pointwise true error rate at* x. In particular we may write this pointwise error rate as

$$l(x; h) = \begin{cases} \mathbb{P}\left(Y = 1 \,|\, X = x\right) =: \pi_x & \text{if } h(x) = 0 \text{ and} \\ \mathbb{P}\left(Y = 0 \,|\, X = x\right) = 1 - \pi_x & \text{if } h(x) = 1. \end{cases}$$

For any $x \in \mathcal{X}$ the Bayes classification rule then minimizes the pointwise true error rate at x by selecting precisely π_x or $1 - \pi_x$ as the pointwise true error rate depending on which of these two values is smaller!

In other words, reasoning as in Remark 22.10,

$$\begin{aligned} l(x; h^*) &= \begin{cases} \pi_x & \text{if } h^*(x) = 0 \text{ and} \\ 1 - \pi_x & \text{if } h^*(x) = 1 \end{cases} \\ &= \begin{cases} \pi_x & \text{if } \pi_x \leqslant 1 - \pi_x \text{ and} \\ 1 - \pi_x & \text{if } 1 - \pi_x < \pi_x \end{cases} \\ &\leqslant l(x; h) \end{aligned}$$

for *any* classification rule h and any $x \in \mathcal{X}$ since the pointwise error rate $l(x; h)$ must always be either π_x or $1 - \pi_x$. The monotonicity of expectation then concludes the proof:

$$L(h^*) = \mathbb{E}\left[l(x; h^*)\right] \leqslant \mathbb{E}\left[l(x; h)\right] = L(h).$$

Exercise A.22.2 (Gaussian Bayes classifier)**.** Prove Theorem 22.16 where we record the form of the Bayes classifier when $X \,|\, Y = y$ is assumed to be Gaussian for $y = 0$, 1.

Solution. Recall that the Bayes classifier is

$$h^*(x) = \begin{cases} 1 & \text{if } \mathbb{P}\left(Y = 1 \,|\, X = x\right) > \mathbb{P}\left(Y = 0 \,|\, X = x\right) \text{ and} \\ 0 & \text{if } \mathbb{P}\left(Y = 0 \,|\, X = x\right) \geqslant \mathbb{P}\left(Y = 1 \,|\, X = x\right). \end{cases}$$

So now we compute that, by Bayes' Theorem,

$$\mathbb{P}\left(Y = 1 \,|\, X = x\right) = \frac{\pi_1 f\left(x \,|\, Y = 1\right)}{\pi_1 f\left(x \,|\, Y = 1\right) + \pi_0 f\left(x \,|\, Y = 0\right)}$$

while

$$\mathbb{P}\left(Y = 0 \,|\, X = x\right) = \frac{\pi_0 f\left(x \,|\, Y = 0\right)}{\pi_1 f\left(x \,|\, Y = 1\right) + \pi_0 f\left(x \,|\, Y = 0\right)}$$

and so

$$\begin{aligned} & \mathbb{P}\left(Y=1 \mid X=x\right)>\mathbb{P}\left(Y=0 \mid X=x\right) \\ \Longleftrightarrow\; & \pi_1 f\left(x \mid Y=1\right)>\pi_0 f\left(x \mid Y=0\right) \\ \Longleftrightarrow\; & \log\left[\pi_1 f\left(x \mid Y=1\right)\right]>\log\left[\pi_0 f\left(x \mid Y=0\right)\right]. \end{aligned}$$

In particular the Gaussian assumption tells us that

$$\begin{aligned} f\left(x \mid Y=1\right) &= \frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}} \exp\left[-\frac{1}{2}\Sigma_1^{-1}\left(x-\mu_1\right)\cdot\left(x-\mu_1\right)\right] \\ &= \frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}} \exp\left[-\frac{1}{2}r_1^2(x)\right] \end{aligned}$$

and so

$$\begin{aligned} \log\left[\pi_1 f\left(x \mid Y=1\right)\right] &= \log \pi_1 - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}r_1^2(x) \\ &= \delta_1(x) - \frac{d}{2}\log 2\pi, \end{aligned}$$

and similarly for $\log\left[\pi_0 f\left(x \mid Y=0\right)\right]$. In summary:

$$\begin{aligned} & h^*(x)=1 \\ \Longleftrightarrow\; & \mathbb{P}\left(Y=1 \mid X=x\right)>\mathbb{P}\left(Y=0 \mid X=x\right) \\ \Longleftrightarrow\; & \log\left[\pi_1 f\left(x \mid Y=1\right)\right]>\log\left[\pi_0 f\left(x \mid Y=0\right)\right] \\ \Longleftrightarrow\; & \delta_1(x)>\delta_0(x). \end{aligned}$$

To conclude we simply verify that

$$2\left[\delta_1(x)-\delta_0(x)\right]=r_0^2(x)-r_1^2(x)+\log\frac{|\Sigma_0|}{|\Sigma_1|}+2\log\frac{\pi_1}{\pi_0}$$

and so indeed

$$\begin{aligned} h^*(x) &= \mathbb{1}\left(\delta_1(x)>\delta_0(x)\right) \\ &= \mathbb{1}\left(r_0^2(x)+\log\frac{|\Sigma_0|}{|\Sigma_1|}+2\log\frac{\pi_1}{\pi_0}>r_1^2(x)\right). \end{aligned}$$

Exercise A.22.3 (VC dimension of two-dimensional spheres)**.** Let $\mathcal{A}$ be the set of two-dimensional spheres. That is, $A\in\mathcal{A}$ if

$$A=\left\{(x,\,y):(x-a)^2+(y-b)^2\leqslant c^2\right\}$$

for some a, b, and c. Find the VC dimension of $\mathcal{A}$.

Solution. We will show that $VC\left(\mathcal{A}\right)=3$. First we show that there is a 3–element set in the plan which may be shattered by $\mathcal{A}$, namely the set

$$F=\{0,\,e_1,\,e_2\}$$

where e_1 and e_2 denote the canonical basis vectors in the plane. To make this easier to follow let us write

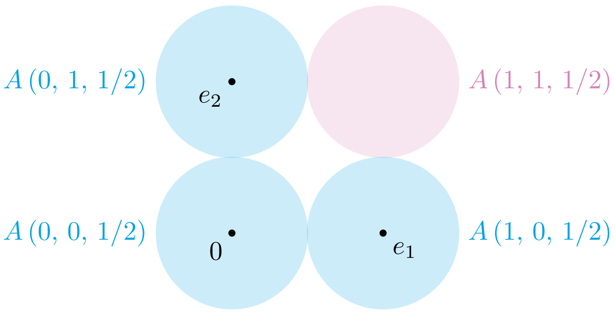
$$A\left(a,\,b,\,c\right):=\left\{(x,\,y)\in\mathbb{R}^2:(x-a)^2+(y-b)^2\leqslant c^2\right\}=\overline{B\left((a,\,b),\,c\right)}$$

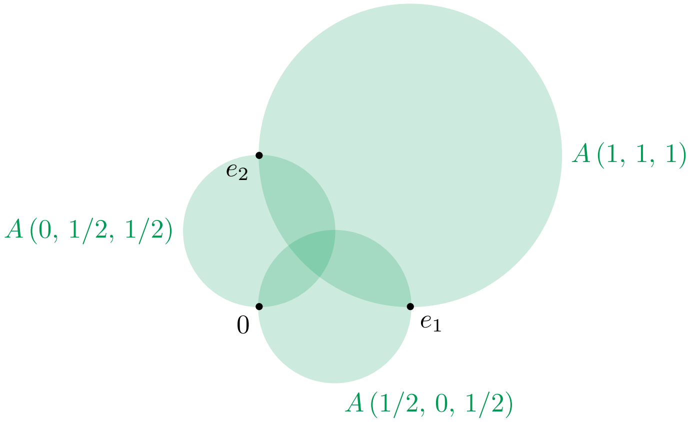
for the closed ball of radius $|c|$ centered at $(a,\,b)$. We then verify directly that all eight subsets of F may be picked out by $\mathcal{A}$.

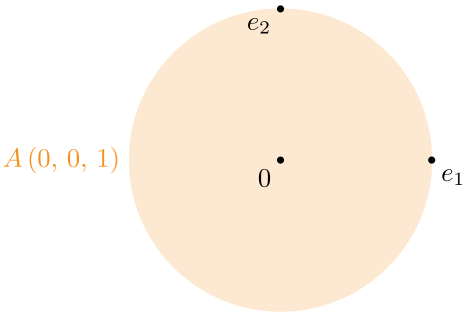
(1) $\emptyset$ is picked out by $A\left(1,\,1,\,1/2\right)$.

(2) $\{0\}$ is picked out by $A\,(0,\,0,\,1/2)$.
(3) $\{e_1\}$ is picked out by $A\,(1,\,0,\,1/2)$.
(4) $\{e_2\}$ is picked out by $A\,(0,\,1,\,1/2)$.
(5) $\{0,\,e_1\}$ is picked out by $A\,(1/2,\,0,\,1/2)$.
(6) $\{0,\,e_2\}$ is picked out by $A\,(0,\,1/2,\,1/2)$.
(7) $\{e_1,\,e_2\}$ is picked out by $A\,(1,\,1,\,1)$.
(8) $\{0,\,e_1,\,e_2\}$ is picked out by $A\,(0,\,0,\,1)$.

Pictorially:







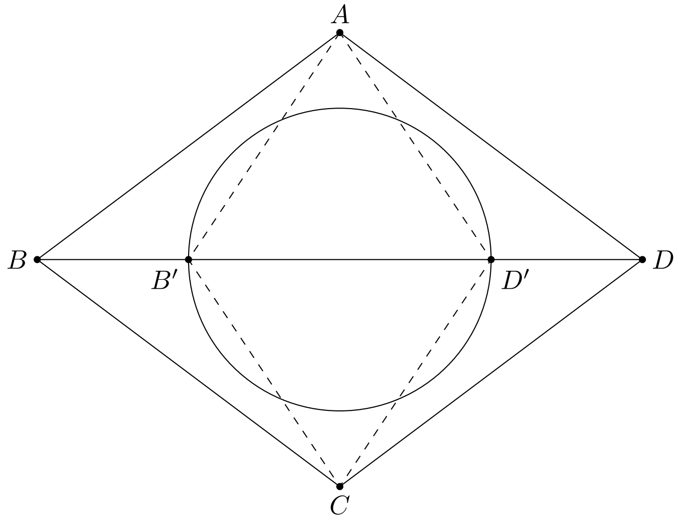
We now show that *no* 4–element set in the plane may be shattered by $\mathcal{A}$. So let A, B, C, and D be four distinct points in the plane. Radon's Theorem tells us that we may partition these two sets into two parts whose convex hulls intersect. In particular these two parts must have sizes either 1 and 3 or 2 and 2.

- **Case 1.** The two parts have sizes 1 and 3, which means that one of the four points, say A, is a convex combination of the other three, say B, C, and D.

Since all sets in $\mathcal{A}$ are convex, this means that any set in $\mathcal{A}$ containing B, C, and D *must* also contain A. In other words $\mathcal{A}$ cannot pick out $\{B, C, D\}$.

- **Case 2.** The two parts have sizes 2 and 2. Then, up to relabelling the points, $ABCD$ forms a quadrilateral whose interior angles add up to 360°, i.e.
$$\hat{A} + \hat{B} + \hat{C} + \hat{D} = 360^\circ.$$
In particular, up to relabelling $(A, C) \leftrightarrow (B, D)$, we have that $\hat{A} + \hat{C} \leqslant 180^\circ$. We will now show that $\mathcal{A}$ cannot pick out $\{A, C\}$.

  Suppose for the sake of contradiction that $\mathcal{A}$ *did* pick out $\{A, C\}$, which means that some closed ball contains A and C but not B and D. Then there exists a possibly smaller such ball for which A and C lie on its boundary. Let B' and D' denote the points obtained by retracting B and D, respectively, along the line segment BD until they enter the closed ball (which must eventually happen, since by virtue of Radon's Theorem the line segments AC and BD must intersect, and in particular since closed balls are convex and since both A and C belong to it, that intersection must occur inside the ball). Pictorially:



  Then $AB'CD'$ is inscribed in a circle and so
$$\widehat{AB'C} + \widehat{CD'A} = 180^\circ.$$
  But then, since $\hat{B} < \widehat{AB'C}$ and $\hat{D} < \widehat{CD'A}$, we have that
$$\hat{B} + \hat{D} < 180^\circ$$
  such that
$$\hat{A} + \hat{B} + \hat{C} + \hat{D} < 360^\circ,$$
  a contradiction. Therefore $\mathcal{A}$ cannot pick out $\{A, C\}$.

In either case there is a subset of $\{A, B, C, D\}$ not picked out by $\mathcal{A}$, so $\mathcal{A}$ does *not* shatter $\{A, B, C, D\}$. This means that $s(\mathcal{A}, 4) < 2^4$.

In conclusion: since $s(\mathcal{A}, 2) = 2^3$ but $s(\mathcal{A}, 4) < 2^4$ we deduce that $VC(\mathcal{A}) = 3$.

Exercise A.22.4 (Kernelization)**.** Suppose that $X_i \in \mathbb{R}$ and that $Y_i = 1$ whenever $|X_i| \leqslant 1$ and $Y_i = 0$ whenever $|X_i| > 0$. Show that no linear classifier can perfectly classify these data. Show that the kernelized data $Z_i = \left(X_i,\, X_i^2\right)$ can be linearly separated.

We will really show the following. Consider the following subsets of $\mathbb{R}$:

$$\mathcal{C}_0 := (-\infty,\, -1) \cup (1,\, \infty) = \{x \in \mathbb{R} : |x| > 1\} \text{ and}$$
$$\mathcal{C}_1 := (-1,\, 1) = \{x \in \mathbb{R} : |x| < 1\}\,.$$

Then we will show that the following hold.

(1) There is no point in $\mathbb{R}$ (affine hyperplanes in $\mathbb{R}$ are points) which separates $\mathcal{C}_0$ and $\mathcal{C}_1$.

(2) There is a vector space V and a map $\phi : \mathbb{R} \to V$, called a *feature map*, such that $\phi\left(\mathcal{C}_0\right)$ and $\phi\left(\mathcal{C}_1\right)$ may be separated by an affine hyperplane in V.

Solution. (1) For any point $x \in \mathbb{R}$ there exist points in $\mathcal{C}_0$ to the left *and* to the right of x, i.e. $\mathcal{C}_0$ intersects both $(-\infty,\, x)$ and $(x,\, \infty)$. Therefore $\{x\}$ does *not* separate $\mathcal{C}_0$ and $\mathcal{C}_1$. Since $x \in \mathbb{R}$ is arbitrary this means that $\mathcal{C}_0$ and $\mathcal{C}_1$ are not linearly separable.

(2) We may choose $V := \mathbb{R}$ and $\phi(x) := x^2$. Then the affine hyperplane $\{1\}$ separates $\phi\left(\mathcal{C}_0\right)$ and $\phi\left(\mathcal{C}_1\right)$ since, for every $y \in V$,

- if $y \in \phi\left(\mathcal{C}_0\right)$ then $y = \phi(x)$ for some $x \in \mathbb{R}$ for which $|x| > 1$ and so $y = x^2 > 1$, while
- if $y \in \phi\left(\mathcal{C}_1\right)$ then $y = x^2$ for some $x \in \mathbb{R}$, and so $y < 1$.

In summary:

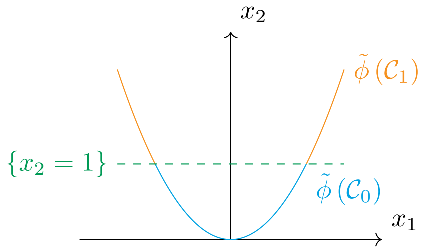
$$y|_{\phi(\mathcal{C}_0)} > 1 \text{ and } y|_{\phi(\mathcal{C}_1)} < 1,$$

which verifies that the affine hyperplane $\{1\}$ separates $\phi\left(\mathcal{C}_0\right)$ and $\phi\left(\mathcal{C}_1\right)$.

In particular this shows that if we define $\tilde{\phi} : \mathbb{R} \to \mathbb{R}^2$, as is suggested in the exercise prompt, then $\tilde{\phi}\left(\mathcal{C}_0\right)$ and $\tilde{\phi}\left(\mathcal{C}_1\right)$ are linearly separable, namely by the hyperplane

$$\left\{(x_1,\, x_2) \in \mathbb{R}^2 : x_2 = 1\right\}.$$

Pictorially:



A.23. **Bonus.** In this section we collect exercises not found in Wasserman's book but whose solutions either shed light on the material therein or help solve problems in that book in a clean way.

For example the problem below comes in handy when computing confidence intervals. Often these require an expression of the form $\Phi^{-1}(1-\alpha)$ to be evaluated, where Φ denotes the CDF of the standard normal distribution. The problem below provides an alternate way to evaluate this expression using the *survival function* instead of the CDF.

Exercise A.23.1 (Inverse of the survival function)**.** Let Φ be an invertible function and define $S(x) := 1 - \Phi(x)$. Prove that S is invertible with inverse given by $S^{-1}(\alpha) = \Phi^{-1}(1-\alpha)$.

Note that when Φ is a CDF, S is known as the corresponding *survival function.*

Solution. This follows from the observation that $S = \tau \circ \Phi$ for τ defined by

$$\tau(\alpha) = 1 - \alpha.$$

Since τ is invertible, it follows that S is invertible as well. Moreover, since the inverse of τ is τ itself, we can readily compute the inverse of S:

$$S^{-1}(\alpha) = \Phi^{-1}(\tau^{-1}(\alpha)) = \Phi^{-1}(1-\alpha),$$

as desired.

Exercise A.23.2 (Expectation of the score function)**.** Let $\{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ be a parametric model. The score function has vanishing mean, i.e

$$\mathbb{E}(\partial_\theta \log f) = 0.$$

Solution. This follows from a direct computation:

$$\mathbb{E}(\partial_\theta \log f) = \int (\partial_\theta \log f) f = \int \frac{\partial_\theta f}{f} f = \int \partial_\theta f = \partial_\theta \int f = \partial_\theta 1 = 0.$$

Exercise A.23.3 (Alternate expression for the Fisher information)**.** Prove that

$$\mathbb{E}\left[\partial_\theta^2 (\log f)\right] = -\mathbb{V}\left[\partial_\theta (\log f)\right].$$

Solution. First we compute that

$$\begin{aligned}\partial_\theta^2(\log f) = \partial_\theta\left(\frac{\partial_\theta f}{f}\right) &= \frac{(\partial_\theta^2 f)f - (\partial_\theta f)^2}{f^2} \\ &= (\partial_\theta^2 f)\frac{1}{f} - \left(\frac{\partial_\theta f}{f}\right)^2 \\ &= (\partial_t^2 f)\frac{1}{f} - \left[\partial_\theta(\log f)\right]^2.\end{aligned}$$

The first term may be integrated as follows:

$$\mathbb{E}\left[(\partial_\theta^2 f)\frac{1}{f}\right] = \int (\partial_\theta^2 f)\frac{1}{f} \cdot f = \int \partial_\theta^2 f = \partial_\theta^2 \int f = \partial_\theta^2 1 = 0.$$

Since the score function $\partial_\theta \log f$ has mean zero (see Exercise A.23.2), the second term may be integrated as follows:

$$\mathbb{E}\left[(\partial_\theta \log f)^2\right] = \mathbb{V}\left[(\partial_\theta \log f)^2\right].$$

Putting it all together we obtain the claim.

Exercise A.23.4 (p-value for the Wald test)**.** Let W_n be the test statistic used in the Wald test (see Definition 10.9). The p–value of the Wald test is

$$p - \text{value} = \mathbb{P}\left(|Z| > |W_n|\right)$$

or

$$\begin{aligned} p - \text{value} &= 2\left[1 - \Phi\left(|W_n|\right)\right] \\ &= 2\Phi\left(-|W_n|\right) \end{aligned}$$

where $Z \sim N(0,\,1)$ and Φ denotes its CDF.

Solution. This follows from a direct computation using the fact that the CDF Φ is strictly increasing:

$$p - \text{value} = \inf\left\{\alpha : |W_n| > z_{\alpha/2}\right\}$$

where $z_{\alpha/2} := \Phi^{-1}(1-\alpha/2)$ and so

$$\begin{aligned} |W_n| > z_{\alpha/2} &\Leftrightarrow |W_n| > \Phi^{-1}(1-\alpha/2) \\ &\Leftrightarrow \Phi(|W_n|) > 1-\alpha/2 \\ &\Leftrightarrow \alpha > 2\left[1-\Phi(|W_n|)\right] \end{aligned}$$

from which it follows that

$$p - \text{value} = 2\left[1-\Phi(|W_n|)\right].$$

To obtain the other expression for the p–value we note that, since

$$\mathbb{P}(Z \geqslant c) = \mathbb{P}(-Z \leqslant -c),$$

it follows that, for any $c \geqslant 0$,

$$2\left[1-\Phi(c)\right] = 2\mathbb{P}(Z \geqslant c) = \mathbb{P}(Z \geqslant c) + \mathbb{P}(-Z \leqslant -c) = \mathbb{P}(|Z| > c)$$

as desired. The last equality follows from the fact that, for any $x \in \mathbb{R}$,

$$1-\Phi(x) = \mathbb{P}(Z > x) = \mathbb{P}(Z < -x) = \Phi(-x).$$

Exercise A.23.5 (Properties of the permutation test)**.** Let $(H_0,\,H_1,\,t,\,R_\alpha)$ be a permutation test. Using the notation of Definition 10.29 and writing $Z_1,\,\ldots\,,\,Z_N$ for $X_1,\,\ldots\,,\,X_m,\,Y_1,\,\ldots\,,\,Y_n$, where $N = n+m$, prove the following.

(1) Under the null hypothesis the conditional distribution of the test statistic t given $Z_1,\,\ldots\,,\,Z_N$ is Uniform$(0,1)$.
(2) This test has size α.
(3) The p–value is equal to the test statistic t.

Solution. (1) Since $\mathcal{T} \sim \mathcal{P}(T)$ such that

$$\mathbb{E}\left(\mathcal{T} \,|\, Z_1 = z_1,\,\ldots,\,Z_N = z_N\right) \sim \text{Uniform}\left(\left\{T\left(z_{\sigma(1)},\,\ldots,\,z_{\sigma(N)}\right) : \sigma \in S_N\right\}\right)$$

it follows that

$$\begin{aligned} &\mathbb{P}_{H_0}\left(\mathcal{T} > T \,|\, Z_1 = z_1,\,\ldots,\,Z_N = z_N\right) \\ &= \frac{1}{N!}\sum_{\sigma \in S_N} I(\mathcal{T}_\sigma > T) \text{ where } Z_1 = z_1,\,\ldots,\,Z_N = z_N. \end{aligned}$$

In particular, using Exercise A.2.14 and proceeding as in Exercise A.10.2 we deduce that, as desired

$$\begin{aligned}\mathbb{E}\left(t \,|\, Z_1 = z_1, \dots, Z_N = z_N\right) &= \frac{1}{N!}\sum_{\sigma \in S_N} I(\mathcal{T}_\sigma > T) \text{ where } Z_1 = z_1, \dots, Z_N = z_N \\ &= \mathbb{P}_{H_0}\left(\mathcal{T} > T \,|\, Z_1 = z_1, \dots, Z_N = z_N\right) \\ &= 1 - \mathcal{F}(\mathcal{T}) \sim \text{Uniform}(0, 1),\end{aligned}$$

where $\mathcal{F}$ denotes the *conditional* CDF of $\mathcal{T}$ given $Z_1, \dots, Z_N$.

(2) Using the rule of iterated expectation as well as the usual trick, rewriting $\mathbb{P}(A|X) = \mathbb{E}(\mathbb{1}_A|X)$ to allow the rule of iterated expectation to also apply to conditional probabilities, we see that

$$\begin{aligned}\beta(H_0) &:= \mathbb{P}_{H_0}(t \in R_\alpha) \\ &= \mathbb{P}_{H_0}(t < \alpha) \\ &= \mathbb{E}_{H_0}\left[\mathbb{P}_{H_0}\left(t < \alpha \,|\, Z_1 = z_1, \dots, Z_N = z_N\right)\right]\end{aligned}$$

where item 1 above tells us that

$$\mathbb{E}\left(t \,|\, Z_1 = z_1, \dots, Z_N = z_N\right) \sim \text{Uniform}(0, 1)$$

and so

$$\mathbb{P}_{H_0}\left(t < \alpha \,|\, Z_1 = z_1, \dots, Z_N = z_N\right) = \alpha.$$

So finally

$$\beta(H_0) = \mathbb{E}\alpha = \alpha,$$

as desired.

(3) This is immediate since

$$p - \text{value} = \inf\{\alpha : t \in R_\alpha\} = \inf\{\alpha : t < \alpha\} = t.$$

Exercise A.23.6 (Expectations of simple tensors)**.** Let Y be a random vector of dimension n.

(1) Prove that $\text{Cov}(Y, Y) = \mathbb{E}(Y \otimes Y) - \mathbb{E}Y \otimes \mathbb{E}Y$.
(2) Deduce that, for any n-by-n matrix A, $\mathbb{E}(AY \cdot Y) = A : \text{Cov}(Y, Y) + A\mathbb{E}(Y) \cdot \mathbb{E}(Y)$.

Solution. (1) This is nothing more than a tensor-form of Lemma 3.11 since it tells us that

$$\text{Cov}(Y_i, Y_j) = \mathbb{E}(Y_i Y_j) - \mathbb{E}(Y_i)\mathbb{E}(Y_j).$$

(2) This follows immediately from item 1:

$$\begin{aligned}\mathbb{E}(AY \cdot Y) &= A : \mathbb{E}(Y \otimes Y) \\ &= A : \text{Cov}(Y, Y) + A : (\mathbb{E}Y \otimes \mathbb{E}Y) \\ &= A : \text{Cov}(Y, Y) + A\mathbb{E}(Y) \cdot \mathbb{E}(Y).\end{aligned}$$

Exercise A.23.7 (Characterizing summary statistics via loss functions)**.** Let X be a random variable, let L be a loss function, and let

$$m_L(X) := \operatorname*{arg\,min}_{\alpha \in \mathbb{R}} \mathbb{E}\left[L(X, \alpha)\right]$$

when this minimum is attained.

(1) If L is the squared error loss and X has finite variance prove that $m_L(X)$ is the mean of X.

(2) If L is the zero-one loss function and X is a discrete random variable prove that $m_L(X)$ is the mode of X.

Solution. (1) This is immediate since $\alpha \mapsto \mathbb{E}\left[(X-\alpha)^2\right]$ is well-defined (since X has finite variance) and strictly convex with

$$\frac{d}{d\alpha}\mathbb{E}\left[(X-\alpha)^2\right] = -2\mathbb{E}(X-\alpha),$$

which vanishes when $\alpha = \mathbb{E}X$.

(2) Let $(x_i)_{i=1}^{\infty}$ denote the points in the support of X and let f denote its PDF. Then, since $\mathbb{1}(x \neq \alpha) = 1 - \mathbb{1}(x = \alpha)$,

$$\begin{aligned}\mathbb{E}\left[L(X,\,\alpha)\right] &= \sum_{i\geqslant 1}\mathbb{1}(x_i \neq \alpha) f(x_i)\\ &= 1 - \sum_{i\geqslant 1}\mathbb{1}(x_i = \alpha) f(x_i)\\ &= \begin{cases} 1 - f(x_i) & \text{if } \alpha = x_i \text{ for some } i \text{ and}\\ 1 & \text{otherwise.}\end{cases}\end{aligned}$$

Therefore

$$\begin{aligned}m_L(X) &= \arg\min_{\alpha\in\mathbb{R}} \mathbb{E}\left[L(X,\,\alpha)\right]\\ &= \arg\min_{i\geqslant 1}\left[1 - f(x_i)\right]\\ &= \arg\max_{i\geqslant 1} f(x_i),\end{aligned}$$

which is indeed the mode of X.

Exercise A.23.8 (Least squares estimate for multiple linear regression)**.** Prove Theorem 13.41 where we record the formulae for the least squares estimate $\hat{\beta}$, an unbiased estimator of σ^2, and the conditional mean and variance of $\hat{\beta}$ given the feature data for the linear regression model. (You do not need to prove item 3 of Theorem 13.41 where asymptotic confidence intervals are recorded.)

Solution. First we compute the least squares estimate. Since the residual sum of squares is

$$RSS = \left|\mathbb{Y} - \mathbb{X}\beta\right|^2 = \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{k}\mathbb{X}_{ij}\beta_j\right)^2$$

we compute that

$$\partial_{\beta_k} RSS = \sum_{i=1}^{n} 2\left(Y_i - \sum_{j=1}^{k}\mathbb{X}_{ij}\beta_j\right)(-\mathbb{X}_{ik}) = -2(\mathbb{X}^T\mathbb{Y})_k + 2(\mathbb{X}^T\mathbb{X}\beta)_j$$

and so

$$\nabla^2 RSS = 2\mathbb{X}^T\mathbb{X}.$$

Since $\mathbb{X}^T\mathbb{X}$ is a positive-definite matrix (it is positive by definition and definite since it is assumed to be invertible) we deduce that RSS is a stricly convex function of β. Its global minimizer, the least squares estimate, thus occurs when

$$\nabla_\beta RSS = 0 \iff \mathbb{X}^T\mathbb{X}\hat{\beta} - \mathbb{X}^T\mathbb{Y} = 0 \iff \hat{\beta} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y},$$

as desired.

We now turn our attention to the unbiased estimator of σ^2. By the rule of iterated expectation

$$\mathbb{E}(RSS) = \mathbb{E}\left[|\mathbb{Y} - \mathbb{X}\beta|^2\right] = \mathbb{E}\left[\mathbb{E}\left(|\mathbb{Y} - \mathbb{X}\beta|^2 \,\middle|\, \mathbb{X}\right)\right]$$

where

$$|\mathbb{Y} - \mathbb{X}\beta|^2 = \left|\mathbb{Y} - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}\right|^2 = \left|\underbrace{(I - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T)}_{=:A(\mathbb{X})}\mathbb{Y}\right|^2.$$

Now Lemma C.8 tells us that $\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ is an orthogonal projection matrix onto the image of $\mathbb{X}$, which means that $A(\mathbb{X})$ is the orthgonal projection matrix onto the *kernel* of $\mathbb{X}$. In particular $A(\mathbb{X})$ satisfies $A^TA = A^2 = A$ and so

$$|\mathbb{Y} - \mathbb{X}\beta|^2 = |A(\mathbb{X})\mathbb{Y}|^2 = A(\mathbb{X})\mathbb{Y} \cdot \mathbb{Y}.$$

By Exercise A.23.6 we may then compute that

$$\mathbb{E}\left[A(\mathbb{X})\mathbb{Y} \cdot \mathbb{Y} \,|\, \mathbb{X}\right] = A(\mathbb{X}) : \mathrm{Cov}\left(\mathbb{Y},\, \mathbb{Y} \,|\, \mathbb{X}\right) + A(\mathbb{X})\mathbb{E}\left(\mathbb{Y} \,|\, \mathbb{X}\right) \cdot \mathbb{E}\left(\mathbb{Y} \,|\, \mathbb{X}\right).$$

We note that, by the IID and standard noise assumptions,

$$\mathrm{Cov}\left(\mathbb{Y},\, \mathbb{Y} \,|\, \mathbb{X}\right) = \mathrm{Cov}\left(\mathbb{X}\beta + \hat{\varepsilon},\, \mathbb{X}\beta + \hat{\varepsilon} \,|\, \mathbb{X}\right) = \mathrm{Cov}\left(\hat{\varepsilon},\, \hat{\varepsilon} \,|\, \mathbb{X}\right) = \sigma^2 I$$

while

$$\mathbb{E}\left(\mathbb{Y} \,|\, \mathbb{X}\right) = \mathbb{E}\left(\mathbb{X}\beta + \hat{\varepsilon} \,|\, \mathbb{X}\right) = \mathbb{X}\beta.$$

In particular, since Lemma C.8 tells us that $A(\mathbb{X})$ is an orthogonal projection matrix onto $\ker \mathbb{X}$, it must annihilate its orthogonal complement $\mathrm{im}\, \mathbb{X}$. Since $\mathbb{X}\beta$ belongs to that image we deduce that $A(\mathbb{X})$ annihilates $\mathbb{E}\left(\mathbb{Y} \,|\, \mathbb{X}\right) = \mathbb{X}\beta$. Meanwhile, Lemma C.8 also tells us that, by linearity of the trace

$$\mathrm{tr}\, A(\mathbb{X}) = \mathrm{tr}\, I_n - (k+1) = n - k - 1$$

since $\mathbb{X}$ is a n-by-$(k+1)$ matrix. So finally:

$$\mathbb{E}\left[A(\mathbb{X})\mathbb{Y} \cdot \mathbb{Y} \,|\, \mathbb{X}\right] = A(\mathbb{X}) : \sigma^2 I + \underbrace{A(\mathbb{X})\mathbb{X}\beta}_{=0} \cdot \mathbb{X}\beta = \sigma^2\, \mathrm{tr}\, A(\mathbb{X}) = \sigma^2(n - k - 1).$$

Therefore

$$\mathbb{E}(RSS) = \mathbb{E}\left(\mathbb{E}\left[A(\mathbb{X})\mathbb{Y} \cdot \mathbb{Y} \,|\, \mathbb{X}\right]\right) = \sigma^2(n - k - 1),$$

which proves that

$$\hat{\sigma}^2 := \frac{RSS}{n - k - 1}$$

in an unbiased estimator of σ^2, as desired.

Finally we turn our attention to the conditional mean and variance of $\hat{\beta}$ given the feature data $\mathbb{X}$. The conditional mean is easy to compute: by the standard noise assumptions we have that

$$\begin{aligned}
\mathbb{E}(\hat{\beta} \,|\, \mathbb{X}) &= \mathbb{E}\left[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y} \,\middle|\, \mathbb{X}\right] \\
&= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{E}\left(\mathbb{X}\beta + \hat{\varepsilon} \,|\, \mathbb{X}\right) \\
&= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\beta + \mathbb{E}\left(\hat{\varepsilon} \,\middle|\, \overline{X}_n\right) \\
&= \beta.
\end{aligned}$$

To compute the conditional variance we note that, for any random vector Z and any constant matrix A,

$$\mathbb{V}(AZ)_{ij} = \operatorname{Cov}\Big((AZ)_i,\, (AZ)_j\Big) = \sum_{l,m} A_{il}A_{jm} \operatorname{Cov}(Z_l,\, Z_m) = A\mathbb{V}(Z)A^T.$$

So here

$$\mathbb{V}(\hat{\beta} \,|\, \mathbb{X}) = \mathbb{V}\left[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y} \,\Big|\, \mathbb{X}\right] = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{V}\left(\mathbb{Y} \,|\, \mathbb{X}\right)\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-T}$$

where, by the IID and standard noise assumptions,

$$\mathbb{V}\left(\mathbb{Y} \,|\, \mathbb{X}\right) = \mathbb{V}\left(\mathbb{X}\beta + \hat{\varepsilon} \,|\, \mathbb{X}\right) = \mathbb{V}\left(\varepsilon \,|\, \mathbb{X}\right) = \sigma^2 I$$

and so, since $\mathbb{X}^T\mathbb{X}$ is symmetric,

$$\mathbb{V}(\hat{\beta} \,|\, \mathbb{X}) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T I\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$$

as desired.

Exercise A.23.9 (MLE for logistic regression)**.** Let $(Y_1,\, X_1),\, \ldots,\, (Y_n,\, X_n)$ be an IID sample drawn from a distribution in the logistic regression model with parameter $\beta \in \mathbb{R}^{k+1}$, let $\mathbb{X}$ denote the design matrix, and let $\mathbb{Y}$ denote the response vector.

(1) Prove that, up to additive constants independent of β, the log-likelihood is

$$l = \mathbb{Y}^T\mathbb{X}\beta - \operatorname{tr}\phi(\mathbb{X}\beta),$$

for $\phi(s) := \log(1 + e^s)$ such that ϕ is applied component-wise to the vector $\mathbb{X}\beta$ and where $\operatorname{tr} v := \sum_i v_i$.

(2) Deduce that the gradient of the log-likelihood is

$$\nabla_\beta l = \left[\mathbb{Y} - \sigma(\mathbb{X}\beta)\right]^T\mathbb{X} = (\mathbb{Y} - \mathfrak{p})^T\mathbb{X}$$

for σ denoting the logistic function and for $\mathfrak{p} = \mathfrak{p}(\beta) = \sigma(\mathbb{X}\beta)$ where σ is applied component-wise to the vector $\mathbb{X}\beta$, or equivalently defined implicitly via $\operatorname{logit}\mathfrak{p} = \mathbb{X}\beta$.

(3) Deduce that the Hessian of the log-likelihood is

$$\nabla_\beta^2 l = -\mathbb{X}^T\mathbb{W}\mathbb{X}$$

for $\mathbb{W} = \mathbb{W}(\beta) = \operatorname{diag}\left[\mathfrak{p}(1 - \mathfrak{p})\right]$, i.e. $\mathbb{W}_{ij} = p_i(1 - p_i)\delta_{ij}$.

(4) Conclude that the Newton ascent step for the maximization of the log-likelihood is

$$\beta_{k+1} = \beta_k + (\mathbb{X}^T\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T(\mathbb{Y} - \mathfrak{p}) = (\mathbb{X}^T\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\mathbb{W}\mathbb{Z}$$

for $\mathbb{Z} := \mathbb{X}\beta_k + \mathbb{W}^{-1}(\mathbb{Y} - \mathfrak{p})$.

(5) Verify that the Newton ascent iterate β_{k+1} is the $\mathbb{W}$–weighted least squares estimate

$$\beta_{k+1} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{k+1}} ||\mathbb{Z} - \mathbb{X}\beta||_{\mathbb{W}}^2 = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{k+1}} ||\mathbb{Y} - \mathfrak{p}||_{\mathbb{W}^{-1}}^2$$

where $||\beta||_{\mathbb{W}}^2 := \mathbb{W}\beta \cdot \beta$ (recall that $\mathbb{W}$, $\mathbb{Z}$, and $\mathfrak{p}$ are functions of β which are evaluated at $\beta = \beta_k$ in the minimization problems above).

Solution. (1) Recall that $Y \mid X = x \sim \text{Bernoulli}\,(\sigma(\beta \cdot X_i))$. Therefore, for

$$p_i = \sigma(\beta \cdot X_i) = \sigma(\mathbb{X}\beta)_i$$

the likelihood is, up to factors independent of β,

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^n f\left(Y_i \mid X_i; \beta\right) f(X_i) \\ &\propto \prod_{i=1}^n p_i^{Y_i} (1-p_i)^{1-Y_i} \\ &= \prod_{i=1}^n (1-p_i) \left(\frac{p_i}{1-p_i} \right)^{Y_i}.\end{aligned}$$

Therefore the log-likelihood is, up to additive terms independent of β,

$$l = \sum_{i=1}^n \log(1-p_i) + Y_i \log\left(\frac{p_i}{1-p_i}\right)$$

where

$$1 - p_i = 1 - \sigma(\beta \cdot X_i) = \sigma(-\beta \cdot X_i) = \frac{1}{1+e^{\beta \cdot X_i}},$$

since

$$1 - \sigma(s) = 1 - \frac{1}{1+e^s} = \frac{e^{-s}}{1+e^{-s}} = \sigma(-s),$$

such that

$$\log(1-p_i) = -\log\left(\frac{1}{1-p_i}\right) = -\log\left(1 + e^{\beta \cdot X_i}\right) = -\phi\left(\beta \cdot X_i\right)$$

while

$$\log\left(\frac{p_i}{1-p_i}\right) = \operatorname{logit} p_i = \beta \cdot X_i$$

and so

$$\begin{aligned} l &= \sum_{i=1}^n Y_i(\beta \cdot X_i) - \phi(\beta \cdot X_i) \\ &= \sum_{i=1}^n \left(\sum_{j=0}^k Y_i \beta_j \mathbb{X}_{ij} - \phi(\mathbb{X}\beta)_i \right) \\ &= \mathbb{Y}^T \mathbb{X}\beta - \operatorname{tr} \phi(\mathbb{X}\beta) \end{aligned}$$

as desired.

(2) Since

$$\phi'(s) = \frac{e^s}{1+e^s} = \sigma(s)$$

we may compute the gradient of the log-likelihood to be

$$\begin{aligned}\nabla_\beta l &= \nabla_\beta(\mathbb{Y}^T\mathbb{X}\beta) - \phi'(\mathbb{X}\beta)\nabla_\beta(\mathbb{X}\beta)\\ &= \mathbb{Y}^T\mathbb{X} - \sigma(\mathbb{X}\beta)^T\mathbb{X}\\ &= (\mathbb{Y} - \mathfrak{p})^T\mathbb{X}\end{aligned}$$

as desired.

(3) Since

$$1 - \sigma(s) = \sigma(-s) = \frac{1}{1+e^s}$$

it follows that

$$\sigma'(s) = \left(\frac{e^s}{1+e^s}\right)' = \frac{e^s}{(1+e^s)^2} = \frac{e^s}{1+e^s}\cdot\frac{1}{1+e^s} = \sigma(s)\left[1-\sigma(s)\right].$$

Therefore the Hessian of the log-likelihood is

$$\begin{aligned}(\nabla^2_\beta l)_{ij} &= \nabla_{\beta_j}(\nabla_\beta l)_i\\ &= -\nabla_{\beta_j}\left[\sigma(\mathbb{X}\beta)^T\mathbb{X}\right]_i\\ &= -\nabla_{\beta_j}\left[\sum_{m=1}^n \sigma\left(\sum_{p=0}^k \mathbb{X}_{mp}\beta_p\right)\mathbb{X}_{mi}\right]\\ &= -\sum_{m=1}^n \sigma'(\mathbb{X}\beta)_m\mathbb{X}_{mj}\mathbb{X}_{mi}\\ &= -\sum_{m=1}^n \sigma(\mathbb{X}\beta)_m\left[1-\sigma(\mathbb{X}\beta)_m\right]\mathbb{X}_{mj}\mathbb{X}_{mi}\\ &= -\sum_{m=1}^n p_m(1-p_m)\mathbb{X}_{mj}\mathbb{X}_{mi}\\ &= -(\mathbb{X}^T\mathbb{W}\mathbb{X})_{ij}\end{aligned}$$

as desired.

(4) The Newton ascent step for the maximization of l is, by definition,

$$\begin{aligned}\beta_{k+1} &= \beta_k - (\nabla^2_\beta l)^{-1}\nabla_\beta l\\ &= \beta_k + (\mathbb{X}^T\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T(\mathbb{Y}-\mathfrak{p}).\end{aligned}$$

We may then check that, by definition of $\mathbb{Z}$,

$$\begin{aligned}&(\mathbb{X}^T\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\mathbb{W}\mathbb{Z}\\ &= (\mathbb{X}^T\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\mathbb{W}\mathbb{X}\beta_k + (\mathbb{X}^T\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\mathbb{W}\mathbb{W}^{-1}(\mathbb{Y}-\mathfrak{p})\\ &= \beta_k + (\mathbb{X}^T\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T(\mathbb{Y}-\mathfrak{p})\\ &= \beta_{k+1}.\end{aligned}$$

(Note that there is no issue with signs here: for both minimization *and* maximization the Newton direction is $-(\nabla^2 f)^{-1}\nabla f$. Indeed: if f is stricly convex and to be minimized then $\nabla^2 f$ is stricly elliptic and the descent direction is $(\nabla^2 f)^{-1}(-\nabla f) = -(\nabla^2 f)^{-1}\nabla f$; while if f is strictly *concave* and to be *maximized* then the *negative* Hessian $-\nabla^2 f$ is strictly elliptic and so the ascent direction is $(-\nabla^2 f)^{-1}(\nabla f) = -(\nabla^2 f)^{-1}\nabla f$.)

(5) We may compute that

$$\nabla_\beta \left(||\mathbb{Z} - \mathbb{X}\beta||_{\mathbb{W}}^2 \right) = 2\mathbb{W}(\mathbb{Z} - \mathbb{X}\beta) \cdot (-\mathbb{X}) = -2(\mathbb{X}^T\mathbb{W}\mathbb{Z} - \mathbb{X}^T\mathbb{W}\mathbb{X}\beta)$$

and so indeed the $\mathbb{W}$–weighted least squares estimate $\hat{\beta}_{k+1}$ is

$$\hat{\beta}_{k+1} = (\mathbb{X}^T\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\mathbb{W}\mathbb{Z}.$$

Moreover it follows from the definition of $\mathbb{Z}$ that, for any β,

$$\begin{aligned} ||\mathbb{Z} - \mathbb{X}\beta||_{\mathbb{W}}^2 &= ||\mathbb{W}^{-1}(\mathbb{Y} - \mathfrak{p})||_{\mathbb{W}}^2 \\ &= \mathbb{W}\mathbb{W}^{-1}(\mathbb{Y} - \mathfrak{p}) \cdot \mathbb{W}^{-1}(\mathbb{Y} - \mathfrak{p}) \\ &= \mathbb{W}^{-1}(\mathbb{Y} - \mathfrak{p}) \cdot (\mathbb{Y} - \mathfrak{p}) \\ &= ||\mathbb{Y} - \mathfrak{p}||_{\mathbb{W}^{-1}}^2 \end{aligned}$$

as desired.

Exercise A.23.10 (Prediction intervals for multiple linear regressoin)**.** Verify that when $k = 1$, if $\mathbb{X}$ denotes the design matrix for linear regression and X_* denotes a new observation then

$$\widetilde{X}_*^T(\mathbb{X}^T\mathbb{X})^{-1}\widetilde{X}_* = \frac{\sum_{i=1}^n (X_i - X_*)^2}{n\sum_{i=1}^n (X_i - \overline{X}_n)^2}$$

where $\widetilde{X}_* := (1, X_*)$.

This shows that Theorem 13.90 for *multiple* linear regression reduces to reduces to Theorem 13.25 for *simple* linear regression when $k = 1$.

Solution. Since $k = 1$ the design matrix takes the form

$$\mathbb{X} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$$

and so

$$\mathbb{X}^T\mathbb{X} = \begin{pmatrix} 1 & \dots & 1 \\ X_1 & \dots & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} n & n\overline{X}_n \\ n\overline{X}_n & \sum_{i=1}^n X_i^2 \end{pmatrix}.$$

In particular

$$\det \mathbb{X}^T\mathbb{X} = n\sum_{i=1}^n X_i^2 - n^2\overline{X}_n^2.$$

If we denote by $\mathbb{E}_n$ expectation with respect to the empirical measure induced by $X_1, \dots, X_n$ and by X^* a corresponding bootstrap sample (which is *very* different from X_*, despite the similar notation), such that for example

$$\mathbb{E}_n[f(X^*)] = \frac{1}{n}\sum_{i=1}^n f(X_i),$$

then we observe that, since $\mathbb{E}_n X^* = \overline{X}_n$,

$$\begin{aligned}\det \mathbb{X}^T\mathbb{X} &= n^2\mathbb{E}_n[(X^*)^2] - n^2[\mathbb{E}_n(X^*)]^2\\ &= n^2\mathbb{V}_n(X^*)\\ &= n^2\mathbb{E}_n[(X^* - \overline{X}_n)^2]\\ &= n\sum_{i=1}^n (X_i - \overline{X}_n)^2\\ &= n^2 s_X^2\end{aligned}$$

for

$$s_X^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$$

which denotes the *population variance* in notation borrowed from Theorem 13.22. Therefore, since

$$\begin{pmatrix} a & b\\ c & d\end{pmatrix}^{-1} = \frac{1}{\det}\begin{pmatrix} d & -b\\ -c & a\end{pmatrix}$$

we deduce that

$$\begin{aligned}(\mathbb{X}^T\mathbb{X})^{-1} &= \frac{1}{n^2 s_X^2}\begin{pmatrix}\sum_{i=1}^n X_i^2 & -n\overline{X}_n\\ -n\overline{X}_n & n\end{pmatrix}\\ &= \frac{1}{ns_X^2}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n\\ -\overline{X}_n & 1\end{pmatrix}\end{aligned}$$

In particular we have shown in item 2 of Exercise A.13.8 that

$$\frac{1}{ns_X^2}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n\\ -\overline{X}_n & 1\end{pmatrix}\begin{pmatrix}1\\ X_*\end{pmatrix}\cdot\begin{pmatrix}1\\ X_*\end{pmatrix} = \frac{\sum_{i=1}^n (X_i - X_*)^2}{n\sum_{i=1}^n (X_i - \overline{X}_n)^2}$$

and so indeed

$$\begin{aligned}\widetilde{X}_*^T(\mathbb{X}^T\mathbb{X})^{-1}\widetilde{X}_* &= (\mathbb{X}^T\mathbb{X})^{-1}\begin{pmatrix}1\\ X_*\end{pmatrix}\cdot\begin{pmatrix}1\\ X_*\end{pmatrix}\\ &= \frac{1}{ns_X^2}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n\\ -\overline{X}_n & 1\end{pmatrix}\begin{pmatrix}1\\ X_*\end{pmatrix}\cdot\begin{pmatrix}1\\ X_*\end{pmatrix}\\ &= \frac{\sum_{i=1}^n (X_i - X_*)^2}{n\sum_{i=1}^n (X_i - \overline{X}_n)^2}\end{aligned}$$

as desired.

Exercise A.23.11 (Parameter prediction intervals for logistic regression)**.** Prove Theorem 13.93 where we record a prediction interval for the Bernoulli parameter $p := \mathbb{P}(Y = 1 \,|\, X)$ in logistic regression.

Solution. The key observation is that, by contrast with prediction intervals for *linear* regression (see Theorem 13.25 and Exercise A.13.8), the random variable $\operatorname{logit} p_*$ we seek to estimate satisfies

$$\operatorname{logit} p_* = \beta \cdot X_*,$$

i.e. there is no noise term ε as in linear regression where $Y_* = \beta \cdot X_* + \varepsilon$. By asymptotic normality of the MLE and the delta method the result then follows.

Indeed we may compute that

$$\mathbb{V}(\operatorname{logit} p_*) = \mathbb{V}(\hat{\beta} \cdot X_*) = \mathbb{V}(\hat{\beta}) X_* \cdot X_*$$

and so, since $\hat{p}_* = \sigma(\operatorname{logit} \hat{p}_*)$ with $\sigma' = \sigma(1-\sigma) \geqslant 0$, the delta method allows us to compute that, asymptotically,

$$\mathbb{V}(p_*) \approx |\sigma'(\hat{p}_*)|^2 \mathbb{V}(\operatorname{logit} p_*) = \sigma(p_*)^2 [1 - \sigma(\hat{p})]^2 \, \mathbb{V}(\hat{\beta}) X_* \cdot X_*.$$

Since $\mathbb{V}(\hat{\beta}) \approx \hat{J}$ asymptotically (the variance of the MLE is the inverse of its Fisher information matrix) we conclude that

$$\frac{\hat{p}_* - p_*}{\sqrt{\sigma(p_*)^2 [1 - \sigma(\hat{p})]^2 \hat{J} X_* \cdot X_*}} = \frac{\hat{p}_* - p_*}{\widehat{se}(\hat{p}_*)} \rightsquigarrow N(0,\, 1),$$

from which the form of the confidence interval follows.

Exercise A.23.12 (Expectation and variance of the categorical distribution)**.** Let $X \sim \text{Categorical}(p)$ for $p \in \Delta^{k-1} \subseteq \mathbb{R}^k$. Prove that

$$\mathbb{E}X = p \text{ and } \mathbb{V}X = \operatorname{diag}(p) - p \otimes p.$$

Solution. Recall that a categorical random variable is defined by

$$\mathbb{P}(X = e_j) = p_j.$$

Therefore

$$\mathbb{E}X = \sum_{j=1}^{k} e_j \mathbb{P}(X = e_j) = \sum_{j=1}^{k} e_j p_j = p.$$

In particular, each component of X has a Bernoulli distribution since $X_j \in \{0,\, 1\}$ with

$$\mathbb{P}(X_j = 1) = \mathbb{P}(X = e_j) = p_j$$

and so $X_j \sim \text{Bernoulli}(p)$. Thus

$$\mathbb{V}X_j = p_j(1 - p_j) = p_j - p_j^2.$$

Now for $j \neq l$ we have that

$$\begin{aligned}
\operatorname{Cov}(X_j,\, X_l) &= \mathbb{E}\left[(X_j - p_j)(X_l - p_l)\right] \\
&= (1 - p_j)(0 - p_l)\mathbb{P}(X = e_j) \\
&\quad + (0 - p_j)(1 - p_l)\mathbb{P}(X = e_l) \\
&\quad + (0 - p_j)(0 - p_l)\mathbb{P}(X \neq e_j,\, e_l) \\
&= -p_l \mathbb{P}(X = e_j) - p_j \mathbb{P}(X = e_l) \\
&\quad + p_j p_l \underbrace{\left[\mathbb{P}(X = e_j) + \mathbb{P}(X = e_l) + \mathbb{P}(X \neq e_j,\, e_l)\right]}_{=1} \\
&= -p_j p_l - p_j p_l + p_j p_l \\
&= -p_j p_l.
\end{aligned}$$

So indeed

$$(\mathbb{V}X)_{ij} = p_i \delta_{ij} - p_i p_j = (\operatorname{diag} p - p \otimes p)_{ij}.$$

Remark A.5 (MLE for the categorical distribution)**.** There are two approaches we can take to compute the MLE of a categorical distribution and its standard error.

(1) We can use a naive *improper* parametrization with k parameters, namely $p_1, \dots, p_k$ for $p \in \Delta^{k-1}$.

In this case we are dealing with a *constrained* problem when maximizing the likelihood and so we need to appeal to Lagrange multipliers. The key issue arises next, when we attempt to compute the Fisher information matrix. Put it simply: there are redundancies among the k parameters due to the constraint $\sum_{j=1}^{k} p_j = 1$ and so we may *not* compute a valid Fisher information matrix using this *improper* parametrization (see Exercise A.23.13 for details).

To compute the asymptotic variance of the MLE, and hence an estimate of its standard error, we must therefore proceed *directly*, without appealing to any properties of maximum likelihood estimators. To do so we leverage the fact that, in this case, the MLE is the sample mean.

(2) We can use a *proper* parametrization with $k-1$ parameters, such as for example $p_1, \dots, p_{k-1}$ for $p \in \Delta^{k-1}$. (Alhough note that which of the k parameters we omit is actually irrelevant, as shown in Exercise A.23.14).

This leads to a formula for the (log-)likelihood which is messier than above (when working with an improper parametrization), but which now leads to a valid computation of the Fisher information.

A crucial observation is the following: whichever approach we take, the MLE obtained for p is the same either way, and so is its asymptotic variance!

Exercise A.23.13 (MLE for the categorical distribution with an improper parametrization)**.** Let $X^{(1)}, \dots, X^{(n)} \sim \text{Categorical}(p)$ for $p \in \Delta^{k-1} \subseteq \mathbb{R}^k$.

(1) Verify that the MLE is $\hat{p} = \overline{X}$.

(2) Verify that

$$\frac{1}{n}\mathbb{V}\left[\nabla l_n(p)\right] = \operatorname{diag}\left(\frac{1}{p}\right) - \mathbb{1}.$$

Explain why this is *not* the Fisher information matrix and why neither is

$$-\frac{1}{n}\mathbb{E}\left[\nabla^2 l_n(p)\right] = \operatorname{diag}\left(\frac{1}{p}\right).$$

(3) Verify that if we define

$$J := \operatorname{diag} p - p \otimes p$$

then

$$\sqrt{n}(\hat{p} - p) \rightsquigarrow N(0,\, J).$$

(4) Deduce that if we define

$$\widehat{se}_j := \sqrt{\frac{\hat{p}_j(1-\hat{p}_j)}{n}}$$

then

$$\frac{\hat{p}_j - p_j}{\widehat{se}_j} \rightsquigarrow N(0,\, 1).$$

Solution. (1) The log-likelihood is

$$\begin{aligned} l &= \log \left(\prod_{i=1}^{n} p_1^{X_1^{(i)}} \dots p_k^{X_k^{(i)}} \right) \\ &= \left(\sum_{i=1}^{n} X_1^{(i)} \right) \log p_1 + \cdots + \left(\sum_{i=1}^{n} X_k^{(i)} \right) \log p_k \\ &= n\overline{X}_1 \log p_1 + \cdots + n\overline{X}_k \log p_k \end{aligned}$$

for

$$\overline{X}_j := \frac{1}{n} \sum_{i=1}^{n} X_j^{(i)}.$$

Since the parameter p lives in the constraint set Δ^{k-1}, (the simplex of dimension $k-1$) we use a Lagrange multiplier to find the MLE maximizing l over Δ^{k-1}. The simplex is the zero level is of

$$g(p) := \sum_{j=1}^{n} p_j - 1$$

and so we consider

$$l(p) - \lambda g(p).$$

We compute that

$$\nabla l - \lambda \nabla g = \begin{pmatrix} \frac{n\overline{X}_1}{p_1} \\ \vdots \\ \frac{n\overline{X}_k}{p_k} \end{pmatrix} - \lambda \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

and so the MLE $\hat{p}$ satisfies

$$\begin{aligned} & \nabla l - \lambda \nabla g \Big|_{\hat{p}} = 0 \\ \iff & \frac{\overline{X}_1}{p_1} = \cdots = \frac{\overline{X}_k}{p_k} = \frac{\lambda}{n} \\ \iff & \hat{p}_j = \frac{n\overline{X}_j}{\lambda}. \end{aligned}$$

Since $\hat{p} \in \Delta^{k-1}$ we observe that

$$1 = \sum_{j=1}^{k} \hat{p}_j = \frac{n}{\lambda} \sum_{j=1}^{k} \overline{X}_j = \frac{1}{\lambda} \sum_{j=1}^{k} \sum_{i=1}^{n} X_j^{(i)} = \frac{n}{\lambda},$$

i.e. $\lambda = n$. So finally

$$\hat{p} = \overline{X}.$$

(2) Exercise A.23.12 tells us that $\mathbb{V}X = \operatorname{diag} p - p \otimes p$ and so, since we know that, in general $\mathbb{V}(AY) = A\mathbb{V}(Y)A^T$ and $A(v \otimes w)B = (Av) \otimes (B^T w)$,

$$\begin{aligned}
\frac{1}{n}\mathbb{V}\left[\nabla l(p)\right] &= \frac{1}{n}\mathbb{V}\left[n \operatorname{diag}\left(\frac{1}{p}\right)\overline{X}\right] \\
&= n\mathbb{V}\left[\operatorname{diag}\left(\frac{1}{p}\right)\overline{X}\right] \\
&= n \operatorname{diag}\left(\frac{1}{p}\right)\mathbb{V}\left(\overline{X}\right)\operatorname{diag}\left(\frac{1}{p}\right) \\
&= n \operatorname{diag}\left(\frac{1}{p}\right)\frac{\mathbb{V}(X)}{n}\operatorname{diag}\left(\frac{1}{p}\right) \\
&= \operatorname{diag}\left(\frac{1}{p}\right)(\operatorname{diag} p - p \otimes p)\operatorname{diag}\left(\frac{1}{p}\right) \\
&= \operatorname{diag}\left(\frac{1}{p}\right) - \operatorname{diag}\left(\frac{1}{p}\right)p \otimes \operatorname{diag}\left(\frac{1}{p}\right)p \\
&= \operatorname{diag}\left(\frac{1}{p}\right) - \mathbb{1} \times \mathbb{1} \\
&= \operatorname{diag}\left(\frac{1}{p}\right) - \mathbb{1}
\end{aligned}$$

where $\mathbb{1}$ denotes both a constant vector or matrix of ones, as appropriate, such that $\mathbb{1} \otimes \mathbb{1} = \mathbb{1}$. In other words we have that

$$\frac{1}{n}\mathbb{V}\left[\nabla l(p)\right] = \operatorname{diag}\left(\frac{1}{p}\right) - \mathbb{1} = \begin{pmatrix} \frac{1}{p_1} - 1 & -1 & \dots & -1 \\ -1 & \frac{1}{p_2} - 1 & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & \frac{1}{p_k} - 1 \end{pmatrix}.$$

Nonetheless this is *not* the Fisher information matrix since it is *not* invertible. Indeed one may compute (e.g. assisted by Mathematica) that

$$\det\left(\operatorname{diag}\left(\frac{1}{p}\right) - \mathbb{1}\right) = \frac{1 - \sum_{j=1}^k p_j}{\prod_{j=1}^k p_j} = 0.$$

Another way to see that this is not the Fisher information matrix is to note that the score function, i.e. the gradient of the log-likelihood, does not have mean zero. Indeed:

$$\mathbb{E}\left[\nabla l(p)\right] = n \operatorname{diag}\left(\frac{1}{p}\right)\mathbb{E}\left(\overline{X}\right) = n \operatorname{diag}\left(\frac{1}{p}\right)p = n\mathbb{1} \neq 0.$$

In particular, since the score function does *not* have average zero, the usual identity

$$-\frac{1}{n}\mathbb{E}\left[\nabla^2 l\right] = \frac{1}{n}\mathbb{V}\left[\nabla l(p)\right]$$

fails in this case. One can indeed compute directly that

$$\nabla^2 l = -n \operatorname{diag}\left(\frac{\overline{X}}{p^2}\right)$$

and hence

$$-\frac{1}{n}\mathbb{E}\left[\nabla^2 l\right] = \operatorname{diag}\mathbb{E}\left(\frac{\overline{X}}{p^2}\right) = \operatorname{diag}\left(\frac{1}{p}\right) \neq \operatorname{diag}\left(\frac{1}{p}\right) - \mathbb{1} = \frac{1}{n}\mathbb{V}\left[\nabla l(p)\right].$$

(3) The key here is that the MLE is the sample mean $\overline{X}$ and so this follows from the multivariate Central Limit Theorem for $J := \mathbb{V}(X) = \operatorname{diag} p - p \otimes p$. Indeed, since $\mathbb{E}X = p$ (see Exercise A.23.12) we know that

$$\sqrt{n}(\hat{p} - p) = \sqrt{n}\left(\overline{X} - \mathbb{E}X\right) \rightsquigarrow N(0,\, \mathbb{V}(X)) = N(0,\, J),$$

as desired.

(4) We may write

$$\widehat{se}_j = \sqrt{\frac{\hat{p}_j(1-\hat{p}_j)}{n}} = \sqrt{\frac{J_{jj}}{n}}$$

while we deduce from item 3 and Theorem 14.5 that, if we use the matrix $P_j := e_j^T$,

$$\sqrt{n}(\hat{p}_j - p_j) = P_j\left[\sqrt{n}(\hat{p} - p)\right] \rightsquigarrow N\left(0,\, P_j J P_j^T\right)$$

where $P_j J P_j^T = e_j J e_j^T = J_{jj} = n\widehat{se}_j^2$. So indeed

$$\frac{\hat{p}_j - p_j}{\widehat{se}_j} = \frac{\sqrt{n}(\hat{p}_j - p_j)}{\sqrt{n\widehat{se}_j^2}} \rightsquigarrow N\left(0,\, \frac{P_j J P_j^T}{n\widehat{se}_j^2}\right) = N(0,\, 1),$$

as desired.

Exercise A.23.14 (MLE for the categorical distribution with a proper parametrization)**.** Let $X^{(1)}, \ldots, X^{(n)} \sim \text{Categorical}(p)$ for $p \in \Delta^{k-1} \subseteq \mathbb{R}^k$. We will now estimate

$$q := (p_1,\, \ldots,\, p_{k-1}) \in \mathbb{R}^{k-1}$$

instead of estimating p, using the constraint $p \in \Delta^{k-1}$ to write

$$p_k = 1 - \sum_{i=1}^{k-1} p_j.$$

(1) Verify that the MLE is

$$\hat{q} = \left(\overline{X}_1,\, \ldots,\, \overline{X}_{k-1}\right).$$

(2) Verify that the Fisher informatin matrix is given by

$$I(q) = \operatorname{diag}\left(\frac{1}{q}\right) + \frac{1}{1 - q_1 - \cdots - q_{k-1}}\mathbb{1}.$$

(3) Verify that if we define

$$J := \operatorname{diag} p - p \otimes p$$

then

$$\sqrt{n}(\hat{q} - q) \rightsquigarrow N(0,\, J).$$

(4) Deduce that if we define

$$\widehat{se}_j := \sqrt{\frac{\hat{q}_j(1-\hat{q}_j)}{n}}$$

then

$$\frac{\hat{q}_j - q_j}{\widehat{se}_j} \rightsquigarrow N(0,\, 1).$$

(5) Deduce that if we define

$$\hat{p}_k := \overline{X}_k \text{ and } \widehat{se}_k := \sqrt{\frac{\hat{p}_k(1-\hat{p}_k)}{n}}$$

then

$$\frac{\hat{p}_k - p_k}{\widehat{se}_k} \rightsquigarrow N(0,\,1).$$

Solution. (1) The log-likelihood is

$$\begin{aligned} l &= \log\left[\prod_{i=1}^n q_1^{X_1^{(i)}} \dots q_{k-1}^{X_{k-1}^{(i)}} (1-q_1-\dots-q_{k-1})^{X_k^{(i)}}\right] \\ &= \left(\sum_{i=1}^n X_1^{(i)}\right)\log q_1 + \dots + \left(\sum_{i=1}^n X_{k-1}^{(i)}\right)\log q_{k-1} \\ &\quad + \left(\sum_{i=1}^n X_k^{(i)}\right)\log(1-q_1-\dots-q_{k-1}) \\ &= n\overline{X}_1 \log q_1 + \dots + n\overline{X}_{k-1}\log q_{k-1} + \overline{X}_k \log(1-q_1-\dots-q_{k-1}) \end{aligned}$$

for

$$\overline{X}_j := \frac{1}{n}\sum_{i=1}^n X_j^{(i)}.$$

Applying derivatives yields

$$\partial_j l = \frac{n\overline{X}_j}{q_j} - \frac{n\overline{X}_k}{1-q_1-\dots-q_{k-1}},$$

i.e. the gradient is

$$\nabla l = n\frac{\overline{X}}{q} - n\frac{\overline{X}_k}{1-q_1-\dots-q_{k-1}}\mathbb{1}.$$

We may write $\overline{X}_k = 1-\overline{X}_1-\dots-\overline{X}_{k-1}$, which tells us that the MLE $\hat{q}$ must satisfy

$$n\frac{\overline{X}_j}{q_j} = n\frac{1-\overline{X}_1-\dots-\overline{X}_{k-1}}{1-q_1-\dots-q_{k-1}} \text{ for all } j.$$

We see immediately, by inspection, that

$$\hat{q} = \overline{X}$$

is *one* possible solution. To rule out any other solutions we compute the Hessian, verifying that $\nabla^2 < 0$.

First we compute that

$$\partial_j^2 l = -n\frac{\overline{X}_j}{q_j^2} - n\frac{\overline{X}_k}{(1-q_1-\dots-q_{k-1})^2}$$

while, if $j \neq m$,

$$\partial_j\partial_m l = -n\frac{\overline{X}_k}{(1-q_1-\dots-q_{k-1})^2}.$$

Therefore

$$\nabla^2 l = -n \operatorname{diag}\left(\frac{\overline{X}}{q^2}\right) - n \frac{\overline{X}_k}{\left(1-q_1-\cdots-q_{k-1}\right)^2}\mathbb{1}.$$

In particular, any matrix of the form

$$\operatorname{diag} r + \alpha \mathbb{1} \text{ with } r_j,\ \alpha > 0$$

is positive-definite since, for any $\xi \in \mathbb{R}^{k-1}$,

$$\begin{aligned}
\left[\operatorname{diag} r + \alpha\mathbb{1}\right] \xi \cdot \xi &= \sum_{i,j} \left(r_i \delta_{ij} \xi_i \xi_j + \alpha \xi_i \xi_j\right) \\
&= \sum_i r_i \xi_i^2 + \alpha \left(\sum_i \xi_i\right)^2 \\
&> \alpha |\xi|^2.
\end{aligned}$$

Therefore $\nabla^2 l < 0$ and so $\hat{q} = \overline{X}$ is indeed the *only* solution of $\nabla l = 0$.

(2) We compute that, since $\mathbb{E}\overline{X} = q$ (see Exercise A.23.12), the Fisher information matrix is given by

$$\begin{aligned}
I(q) &= -\frac{1}{n}\mathbb{E}\left[\nabla^2 l(q)\right] \\
&= \mathbb{E}\left[\operatorname{diag}\frac{\overline{X}}{q^2} + \frac{\overline{X}_k}{\left(1-q_1-\cdots-q_{k-1}\right)^2}\mathbb{1}\right] \\
&= \operatorname{diag}\left(\frac{1}{q}\right) + \mathbb{E}\left[\frac{1-\overline{X}_1-\cdots-\overline{X}_{k-1}}{\left(1-q_1-\cdots-q_{k-1}\right)^2]}\right]\mathbb{1} \\
&= \operatorname{diag}\left(\frac{1}{q}\right) + \frac{1}{1-q_1-\cdots-q_{k-1}}\mathbb{1}.
\end{aligned}$$

(3) We may compute explicitly that the inverse of the Fisher information matrix is

$$J := I^{-1} = \operatorname{diag} q - q \otimes q.$$

Indeed, since we may write $1 - q_1 - \cdots - q_{k-1} = 1 - \mathbb{1} \cdot q$ we have that (recall that $\mathbb{1}$ denotes both constant vectors and constant matrices full of ones, as appropriate, such that $\mathbb{1} \otimes \mathbb{1} = \mathbb{1}$)

$$\begin{aligned}
IJ &= \left(\operatorname{diag}\left(\frac{1}{q}\right) + \frac{1}{1-\mathbb{1}\cdot q}\mathbb{1}\right)\left(\operatorname{diag} q - q \otimes q\right) \\
&= \mathrm{id} - \underbrace{\left[\operatorname{diag}\left(\frac{1}{q}\right) q\right]}_{\mathbb{1}} \otimes q + \frac{1}{1-\mathbb{1}\cdot q} \underbrace{\mathbb{1}\operatorname{diag} q}_{\mathbb{1}\otimes q} - \frac{1}{1-\mathbb{1}\cdot q} \underbrace{\mathbb{1} q \otimes q}_{(\mathbb{1}\cdot q)\mathbb{1}\otimes q)} \\
&= \mathrm{id} + \underbrace{\left(-1 + \frac{1}{1-\mathbb{1}\cdot q} - \frac{\mathbb{1}\cdot q}{1-\mathbb{1}\cdot q}\right)}_{=0} \mathbb{1} \otimes q \\
&= \mathrm{id}
\end{aligned}$$

while similarly

$$\begin{aligned}
JI &= (\operatorname{diag} q - q \otimes q)\left(\operatorname{diag}\left(\frac{1}{q}\right) + \frac{1}{1 - \mathbb{1}\cdot q}\mathbb{1}\right)\\
&= \mathrm{id} + \frac{1}{1-\mathbb{1}\cdot q}\underbrace{(\operatorname{diag} q)\mathbb{1}}_{q\otimes\mathbb{1}} - q \otimes \underbrace{\left[\operatorname{diag}\left(\frac{1}{q}\right) q\right]}_{\mathbb{1}} - \frac{1}{1-\mathbb{1}\cdot q}\underbrace{q\otimes q\mathbb{1}}_{(q\cdot\mathbb{1})q\otimes\mathbb{1})}\\
&= \mathrm{id} + \underbrace{\left(\frac{1}{1-\mathbb{1}\cdot q} - 1 - \frac{q\cdot\mathbb{1}}{1-\mathbb{1}\cdot q}\right)}_{=0} q\otimes\mathbb{1}\\
&= \mathrm{id}\,.
\end{aligned}$$

The asymptotic normality of the MLE then tells us that

$$\sqrt{n}(\hat{p}-p) \rightsquigarrow N\left(0,\, I^{-1}\right) = N(0,\, J),$$

as desired.

(4) Since

$$\widehat{se}_j := \sqrt{\frac{\hat{q}_j(1-\hat{q}_j)}{n}} = \sqrt{\frac{J_{jj}}{n}}$$

we may then proceed exactly as in item 4 of Exercise A.23.13.

(5) Since

$$p_k = 1 - q_1 - \cdots - q_{k-1} = 1 - \mathbb{1}\cdot q$$

the equivariance of the MLE tells us that

$$\hat{p}_k = 1 - \mathbb{1}\cdot\hat{q} = 1 - \hat{q}_1 - \cdots - \hat{q}_{k-1} = 1 - \overline{X}_1 - \cdots - \overline{X}_{k-1} = \overline{X}_k.$$

Moreover, since asymptotically

$$\mathbb{V}(\hat{q}) \approx \frac{1}{n}J$$

we deduce that, asymptotically,

$$\mathbb{V}(\hat{p}_k) = \mathbb{V}(\mathbb{1}\cdot\hat{q}) = \mathbb{V}(\hat{q})\mathbb{1}\cdot\mathbb{1} \approx \frac{1}{n}J\mathbb{1}\cdot\mathbb{1}\Big|_{q=\hat{q}}$$

where

$$\begin{aligned}
J\mathbb{1}\cdot\mathbb{1}\Big|_{q=\hat{q}} &= \sum_{i,j=1}^{k-1} J_{ij}\\
&= \left(\sum_{i=1}^{k-1}\hat{q}_i\right) - \left(\sum_{i=1}^{k-1}\hat{q}_i\right)\left(\sum_{j=1}^{k-1}\hat{q}_j\right)\\
&= \left(\sum_{i=1}^{k-1}\hat{q}_i\right)\left(1 - \sum_{j=1}^{k-1}\hat{q}_j\right)\\
&= (1-\hat{p}_k)\hat{p}_k.
\end{aligned}$$

In other words, for

$$\widehat{se}_k := \sqrt{\frac{\hat{p}_k(1-\hat{p}_k)}{n}}$$

we have that

$$\frac{\hat{p}_k - p_k}{\widehat{se}_k} \rightsquigarrow N(0,\, 1)$$

as desired.

Exercise A.23.15 (Independence of binary random variables and correlation)**.** Let Y and Z be two *binary* random variables. They are independent if and only if they are uncorrelated, i.e.

$$Y \amalg Z \iff \operatorname{Cov}(Y,\, Z) = 0.$$

Solution. First we compute that

$$\mu_Z := \mathbb{E}Z = \mathbb{P}(Z=1) = p_{1\cdot} \text{ and } \mu_Y := \mathbb{E}Y = \mathbb{P}(Y=1) = p_{\cdot 1}$$

for p denoting the associated two-by-two parameter. Therefore

$$\begin{aligned}
\operatorname{Cov}(Z,\, Y) &= \mathbb{E}(Z-\mu_Z)(Y-\mu_Y) \\
&= (0-\mu_Z)(0-\mu_Y)p_{00} + (0-\mu_Z)(1-\mu_Y)p_{01} \\
&\quad + (1-\mu_Z)(0-\mu_Y)p_{10} + (1-\mu_Z)(1-\mu_Y)p_{11} \\
&= \mu_Z\mu_Y \underbrace{(p_{00}+p_{01}+p_{10}+p_{11})}_{1} - \mu_Z p_{01} - \mu_Y p_{10} - (\mu_Y+\mu_Z)p_{11} + p_{11} \\
&= \mu_Z\mu_Y - \mu_Z \underbrace{(p_{01}+p_{11})}_{\mu_Y} - \mu_Y \underbrace{(p_{10}+p_{11})}_{\mu_Z} + p_{11} \\
&= p_{11} - \mu_Z\mu_Y \\
&= p_{11} - (p_{10}+p_{11})(p_{01}+p_{11}) \\
&= p_{11} - [p_{11}(1-p_{00}) + p_{10}p_{01}] \\
&= p_{11}p_{00} - p_{10}p_{01}.
\end{aligned}$$

In other words, for ψ denoting the odds ratio, Theorem 15.6 tells us that

$$\begin{aligned}
\operatorname{Cov}(Z,\, Y) = 0 &\iff p_{11}p_{00} = p_{01}p_{10} \\
&\iff \psi = 1 \\
&\iff Y \amalg Z,
\end{aligned}$$

as desired.

Exercise A.23.16 (Limiting behaviour of Pearson's χ^2 statistic)**.** Prove Theorem 10.24 where we establish that the limiting behaviour of Pearson's χ^2 statistic is, under the null hypothesis of Pearson's χ^2 test for multinomial data in $\mathbb{R}^k$, a χ^2 distribution with $k-1$ degrees of freedom.

Solution. Recall that Pearson's χ^2 statistic is

$$T = \sum_{j=1}^{k} \frac{(X_k - E_j)^2}{E_j}$$

where $X \sim \text{Multinomial}(n,\, p_0)$ under the null, such that

$$\mathbb{E}X = np_0 \text{ and } \mathbb{V}X = n\operatorname{diag} p_0 - np_0 \otimes p_0,$$

and where $E_j = np_{0,j} = \mathbb{E}X_j$. In particular if we define

$$Z_j := \frac{X_j - E_j}{\sqrt{E_j}}$$

such that

$$T = \sum_{j=1}^{k} Z_j = |Z|^2$$

then we may compute that

$$\mathbb{E}Z_j = \frac{\mathbb{E}(X_j) - E_j}{\sqrt{E_j}} = 0$$

while

$$\begin{aligned}\operatorname{Cov}(Z_j,\, Z_l) &= \operatorname{Cov}\left(\frac{X_j - E_j}{\sqrt{E_j}},\, \frac{X_l - E_l}{\sqrt{E_l}}\right)\\ &= \frac{1}{\sqrt{E_j E_l}} \operatorname{Cov}(X_j,\, X_l)\\ &= \frac{1}{\sqrt{p_{0,j} p_{0,l}}} \left(p_{0,j}\delta_{jl} - p_{0,j} p_{0,l}\right)\\ &= \delta_{jl} - \sqrt{p_{0,j} p_{0,l}},\end{aligned}$$

i.e.

$$\mathbb{V}Z = I_k - \sqrt{p_0} \otimes \sqrt{p_0}.$$

Since $\det(I - v \otimes w) = 1 - v \cdot w$ we may compute that

$$\det\left(\alpha I_k - v \otimes w\right) = \alpha^k \det\left(I - \frac{v}{\alpha} \otimes w\right) = \alpha^k \left(1 - \frac{v \cdot w}{\alpha}\right)$$

and so the characteristic polynomial of $\mathbb{V}Z$ is

$$\det\left[\mathbb{V}(Z) - \lambda I_k\right] = \det\left[(1-\lambda) I_k - \sqrt{p_0} \otimes \sqrt{p_0}\right] = (1-\lambda)^k \left(1 - \frac{\sqrt{p_0} \cdot \sqrt{p_0}}{1-\lambda}\right)$$

where

$$\sqrt{p_0} \cdot \sqrt{p_0} = \sum_{j=1}^{k} p_{0,j} = 1$$

and so

$$\det\left[\mathbb{V}(Z) - \lambda I_k\right] = (1-\lambda)^k \left(1 - \frac{1}{1-\lambda}\right) = -\lambda (1-\lambda)^{k-1}.$$

This means that the eigenvalues of $\mathbb{V}Z$ are

$$\lambda_1 = \cdots = \lambda_{k-1} = 1 \text{ and } \lambda_k = 0.$$

Since $\mathbb{V}Z$ is symmetric it is diagonalizable by an orthogonal matrix Q for which

$$Q\mathbb{V}(Z)Q^T = I_{k-1} \oplus 0 = \operatorname{diag}(\underbrace{1,\, \ldots,\, 1}_{k\,-\,1 \text{ times}},\, 0).$$

So finally we can discuss the asymptotic distribution of the test statistic T. The Central Limit Theorem tells us that

$$Z \rightsquigarrow N(0,\, \mathbb{V}Z) \text{ as } n \to \infty.$$

We introduce

$$W := QZ.$$

From Theorem 14.1 we deduce that

$$W \rightsquigarrow N\left(0,\, Q\mathbb{V}(Z)Q^T\right) = N\left(0,\, I_{k-1} \oplus 0\right) \text{ as } n \to \infty,$$

which means that

$$|W|^2 \rightsquigarrow \chi^2_{k-1} \text{ as } n \to \infty.$$

So finally we conclude that, since Q is orthogonal and hence an isometry,

$$T = |Z|^2 = |QZ|^2 = |W|^2 \rightsquigarrow \chi^2_{k-1} \text{ as } n \to \infty.$$

Exercise A.23.17 (Properties of the two-sample Kolmogorov-Smirnov test)**.** Deduce Corollary 15.29, where we derive the size and p–value of the two-sample Kolmogorov-Smirnov test, from Theorem 15.28 where the asymptotic distribution of the corresponding test statistic is recorded to be the Kolmogorov distribution.

Solution. We may compute directly the asymptotic size to be

$$\begin{aligned}\mathbb{P}_{H_0}\left(D \in R_\alpha\right) &= \mathbb{P}_{H_0}\left(\sqrt{\frac{nm}{n+m}}D > H^{-1}(1-\alpha)\right) \\ &\to \mathbb{P}_{H_0}\left(K > H^{-1}(1-\alpha)\right) \\ &= \mathbb{P}_{H_0}\left(H(K) > 1-\alpha\right).\end{aligned}$$

Since H is the CDF of the random variable K, Exercise A.2.14 tells us that $H(K)$ is a Uniform(0, 1) random variable and so indeed the asymptotic size is

$$\mathbb{P}_{H_0}\left(D \in R_\alpha\right) \to \mathbb{P}_{H_0}\left(\text{Uniform}(0,\,1) > 1-\alpha\right) = \alpha.$$

We may then use the fact that H is strictly increasing to compute the p–value:

$$\begin{aligned}D \in R_\alpha &\iff \sqrt{\frac{nm}{n+m}}D > H^{-1}(1-\alpha) \\ &\iff H\left(\sqrt{\frac{nm}{n+m}}D\right) > 1-\alpha \\ &\iff \alpha > 1 - H\left(\sqrt{\frac{nm}{n+m}}D\right)\end{aligned}$$

and so

$$p-\text{value} = \inf\left\{\alpha \in (0,\,1) : D \in R_\alpha\right\} = 1 - H\left(\sqrt{\frac{nm}{n+m}}D\right).$$

In particular, for any outcome ω,

$$p-\text{value}(\omega) = 1 - H\left(\sqrt{\frac{nm}{n+m}}D(\omega)\right) = \mathbb{P}\left(K \geqslant \sqrt{\frac{nm}{n+m}}D(\omega)\right)$$

as desired – recall that K is a *continuous* random variable, and so, for every $x \in \mathbb{R}$, $\mathbb{P}(K \geqslant x) = \mathbb{P}(K > x)$.

Exercise A.23.18 (Chain rule under conditional independence)**.** Prove that

$$\mathbb{P}\left(E \,|\, FG\right) = \frac{\mathbb{P}\left(EF \,|\, G\right)}{\mathbb{P}\left(F \,|\, G\right)}$$

for any events E, F, and G.

Solution. If we define a *new* probability measure $\mathbb{P}_G$ via

$$\mathbb{P}_G(A) := \mathbb{P}(A \,|\, G)$$

for any event A then the identity we seek follows immediately from the definition of conditional independence with respect to the new measure $\mathbb{P}_G$ since

$$\mathbb{P}_G\left(E \,|\, F\right) = \frac{\mathbb{P}_G(EF)}{\mathbb{P}_G(F)} = \frac{\mathbb{P}(EFG)/\mathbb{P}(G)}{\mathbb{P}(FG)/\mathbb{P}(G)} = \mathbb{P}\left(E \,|\, FG\right)$$

and so

$$\mathbb{P}_G\left(E \,|\, F\right) = \frac{\mathbb{P}_G(EF)}{\mathbb{P}_G(F)} \iff \mathbb{P}\left(E \,|\, FG\right) = \frac{\mathbb{P}\left(EF \,|\, G\right)}{\mathbb{P}\left(F \,|\, G\right)}.$$

Alternatively we may proceed more prosaically and compute directly, without any reference to $\mathbb{P}_G$, that

$$\mathbb{P}\left(E \,|\, FG\right) = \frac{\mathbb{P}\left(EFG\right)}{\mathbb{P}\left(FG\right)} = \frac{\mathbb{P}\left(EFG\right)}{\mathbb{P}\left(G\right)} \cdot \frac{\mathbb{P}\left(G\right)}{\mathbb{P}\left(FG\right)} = \frac{\mathbb{P}\left(EF \,|\, G\right)}{\mathbb{P}\left(F \,|\, G\right)}.$$

Exercise A.23.19 (Chained conditional independence)**.** Let W, X, and Y be random variables such that $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp W \,|\, Y$. Prove that $X \perp\!\!\!\perp (W,\, Y)$.

Solution. For simplicity we omit the arguments of the PDFs below. We may then directly compute that

$$f_{X,W,Y} = f_{X,W|Y} f_Y = f_{X|Y} f_{W|Y} f_Y = f_X f_{W|Y} f_Y = f_X f_{W,Y},$$

verifying that indeed $X \perp\!\!\!\perp (W,\, Y)$.

Exercise A.23.20 (Direct proof of the Markov condition)**.** Consider the DAG

$$X \longleftarrow Y \longrightarrow Z$$

which represents a distribution with PDF

$$f(x,\, y,\, z) = f(y) f(x \,|\, y) f(z \,|\, y).$$

Without appealing to the Markov condition, prove that $X \perp\!\!\!\perp Z \,|\, Y$.

Solution. Since

$$f(y,\, z) = \int f(x,\, y,\, z) dx = f(y) f(z \,|\, y) \underbrace{\int f(x \,|\, y) dx}_{=1}$$

we deduce that

$$f(x \,|\, y,\, z) = \frac{f(x,\, y,\, z)}{f(y,\, z)} = f(x \,|\, y),$$

which by Theorem 17.2 proves that $X \perp\!\!\!\perp Z \,|\, Y$.

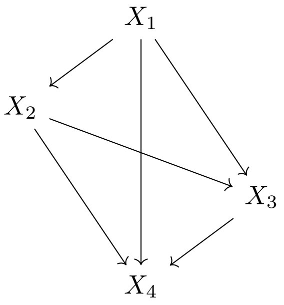
Exercise A.23.21 (Any distribution can be represented as a complete graph)**.** Let $X = (X_1,\, \ldots,\, X_k)$ be any random vector in $\mathbb{R}^k$. Prove that its distribution may be represented by the *complete DAG*, also known as a *tournament*, $(V,\, E)$ where $V = \{X_1,\, \ldots,\, X_k\}$ and

$$E = \{(X_i,\, X_j) : i < j\}.$$

Note that this graph is indeed complete since it contains

$$\sum_{i=1}^{k} (k - i) = \frac{k(k-1)}{2} = \binom{k}{2}$$

edges. For example if $k = 4$ the complete DAG is



Note also that, in general, this DAG representation may not be *faithful*.

Solution. This follows immediately from the chain rule (see Exercise A.1.17), which here is really just an iterated application of the definition of a conditional distribution. It tells us that

$$\begin{aligned} f(x_1, \dots, x_k) &= f(x_k \,|\, x_1, \dots, x_{k-1})\, f(x_1, \dots, x_{k-1}) \\ &= \dots \\ &= f(x_k \,|\, x_1, \dots, x_{k-1})\, f(x_{k-1} \,|\, x_1, \dots, x_{k-2}) \dots f(x_2 \,|\, x_2)\, f(x_1) \\ &= \prod_{j=1}^{k} f(x_j \,|\, x_1, \dots, x_{j-1}), \end{aligned}$$

which is indeed represented by the complete DAG (V, E) above since

$$\begin{aligned} \prod_{j=1}^{k} f(x_j \,|\, x_1, \dots, x_{j-1}) &= \prod_{j=1}^{k} f(x_j \,|\, x_i \text{ for } i < j) \\ &= \prod_{j=1}^{k} f(x_j \,|\, x_i \text{ for } (X_i, X_j) \in E) \\ &= \prod_{j=1}^{k} f(x_j \,|\, \pi_j). \end{aligned}$$

Exercise A.23.22 (Log-linear expansion for three binary random variables)**.** Let $X = (X_1, X_2, X_3)$ be a random vector where X_1, X_2, and X_3 are binary random variables. Verify that the log-linear expansion of the PDF f of X is given by

$$\begin{aligned}
&\psi_\emptyset = \log p_{000},\\
&\psi_1(x_1) = x_1 \log \frac{p_{100}}{p_{000}},\\
&\psi_2(x_2) = x_2 \log \frac{p_{010}}{p_{000}},\\
&\psi_3(x_3) = x_3 \log \frac{p_{001}}{p_{000}},\\
&\psi_{12}(x_1,\, x_2) = x_1 x_2 \log \frac{p_{000}p_{110}}{p_{010}p_{100}},\\
&\psi_{13}(x_1,\, x_3) = x_1 x_3 \log \frac{p_{000}p_{101}}{p_{001}p_{100}},\\
&\psi_{23}(x_2,\, x_3) = x_2 x_3 \log \frac{p_{000}p_{011}}{p_{001}p_{010}}, \text{ and}\\
&\psi_{123}(x_1,\, x_2,\, x_3) = x_1 x_2 x_3 \log \frac{p_{001}p_{010}p_{100}p_{111}}{p_{110}p_{101}p_{011}p_{000}}.
\end{aligned}$$

Solution. Writing

$$\mathbb{1}(a,\, b,\, c) := \mathbb{1}(x_1 = a,\, x_2 = b,\, x_3 = c)$$

we may observe that

$$\begin{aligned}
\log f = {}& \mathbb{1}(0,\,0,\,0)p_{000} + \mathbb{1}(0,\,0,\,1)p_{001}\\
&+ \mathbb{1}(0,\,1,\,0)p_{010} + \mathbb{1}(0,\,1,\,1)p_{011}\\
&+ \mathbb{1}(1,\,0,\,0)p_{100} + \mathbb{1}(1,\,0,\,1)p_{101}\\
&+ \mathbb{1}(1,\,1,\,0)p_{110} + \mathbb{1}(1,\,1,\,1)p_{111}.
\end{aligned}$$

We write

$$\begin{aligned}
\mathbb{1}(0,\,0,\,0) = {}& 1 - \mathbb{1}(0,\,0,\,1)\\
&-\mathbb{1}(0,\,1,\,0) - \mathbb{1}(0,\,1,\,1)\\
&-\mathbb{1}(1,\,0,\,0) - \mathbb{1}(1,\,0,\,1)\\
&-\mathbb{1}(1,\,1,\,0) - \mathbb{1}(1,\,1,\,1),
\end{aligned}$$

which yields

$$\begin{aligned}
\log f = {}& \log p_{000} && + \mathbb{1}(0,\,0,\,1)\log \frac{p_{001}}{p_{000}}\\
&+ \mathbb{1}(0,\,1,\,0)\log \frac{p_{010}}{p_{000}} && + \mathbb{1}(0,\,1,\,1)\log \frac{p_{011}}{p_{000}}\\
&+ \mathbb{1}(1,\,0,\,0)\log \frac{p_{100}}{p_{000}} && + \mathbb{1}(1,\,0,\,1)\log \frac{p_{101}}{p_{000}}\\
&+ \mathbb{1}(1,\,1,\,0)\log \frac{p_{110}}{p_{000}} && + \mathbb{1}(1,\,1,\,1)\log \frac{p_{111}}{p_{000}}.
\end{aligned}$$

We now use dashes to indicate summations such that, for example

$$\mathbb{1}(-,\, -,\, 1) = \mathbb{1}(x_3 = 1) \text{ and } \mathbb{1}(-,\, 1,\, 1) = \mathbb{1}(x_2 = x_3 = 1).$$

We now write

$$\begin{aligned}
\mathbb{1}(0,\,0,\,1) = \mathbb{1}(-,\,-,\,1) &- \mathbb{1}(0,\,1,\,1)\\
&- \mathbb{1}(1,\,0,\,1)\\
&- \mathbb{1}(1,\,1,\,1),\\
\mathbb{1}(0,\,1,\,0) = \mathbb{1}(-,\,1,\,-) &- \mathbb{1}(0,\,1,\,1)\\
&- \mathbb{1}(1,\,1,\,0)\\
&- \mathbb{1}(1,\,1,\,1), \text{ and}\\
\mathbb{1}(1,\,0,\,0) = \mathbb{1}(1,\,-,\,-) &- \mathbb{1}(1,\,0,\,1)\\
&- \mathbb{1}(1,\,1,\,0)\\
&- \mathbb{1}(1,\,1,\,1)
\end{aligned}$$

to obtain

$$\begin{aligned}
\log f = \log p_{000} & & &+ \mathbb{1}(x_3 = 1)\log\frac{p_{001}}{p_{000}}\\
&+ \mathbb{1}(x_2 = 1)\log\frac{p_{010}}{p_{000}} & &+ \mathbb{1}(0,\,1,\,1)\log\frac{p_{011}}{p_{000}}\cdot\frac{p_{000}}{p_{001}}\cdot\frac{p_{000}}{p_{010}}\\
&+ \mathbb{1}(x_1 = 1)\log\frac{p_{100}}{p_{000}} & &+ \mathbb{1}(1,\,0,\,1)\log\frac{p_{101}}{p_{000}}\cdot\frac{p_{000}}{p_{001}}\cdot\frac{p_{000}}{p_{100}}\\
&+ \mathbb{1}(1,\,1,\,0)\log\frac{p_{110}}{p_{000}}\cdot\frac{p_{000}}{p_{010}}\cdot\frac{p_{000}}{p_{100}} & &+ \mathbb{1}(1,\,1,\,1)\log\frac{p_{111}}{p_{000}}\cdot\frac{p_{000}}{p_{001}}\cdot\frac{p_{000}}{p_{010}}\cdot\frac{p_{000}}{p_{100}}
\end{aligned}$$

or equivalently

$$\begin{aligned}
\log f = \log p_{000} & & &+ \mathbb{1}(x_3 = 1)\log\frac{p_{001}}{p_{000}}\\
&+ \mathbb{1}(x_2 = 1)\log\frac{p_{010}}{p_{000}} & &+ \mathbb{1}(0,\,1,\,1)\log\frac{p_{000}p_{011}}{p_{001}p_{010}}\\
&+ \mathbb{1}(x_1 = 1)\log\frac{p_{100}}{p_{000}} & &+ \mathbb{1}(1,\,0,\,1)\log\frac{p_{000}p_{101}}{p_{001}p_{100}}\\
&+ \mathbb{1}(1,\,1,\,0)\log\frac{p_{000}p_{110}}{p_{010}p_{100}} & &+ \mathbb{1}(1,\,1,\,1)\log\frac{p_{000}^2p_{111}}{p_{001}p_{010}p_{100}}.
\end{aligned}$$

Finally we write

$$\begin{aligned}
\mathbb{1}(0,\,1,\,1) &= \mathbb{1}(-,\,1,\,1) - \mathbb{1}(1,\,1,\,1),\\
\mathbb{1}(1,\,0,\,1) &= \mathbb{1}(1,\,-,\,1) - \mathbb{1}(1,\,1,\,1), \text{ and}\\
\mathbb{1}(1,\,1,\,0) &= \mathbb{1}(1,\,1,\,-) - \mathbb{1}(1,\,1,\,1)
\end{aligned}$$

to obtain

$$\begin{aligned}
\log f = \log p_{000} & & &+ \mathbb{1}(x_3 = 1)\log\frac{p_{001}}{p_{000}}\\
&+ \mathbb{1}(x_2 = 1)\log\frac{p_{010}}{p_{000}} & &+ \mathbb{1}(x_2 = x_3 = 1)\log\frac{p_{000}p_{011}}{p_{001}p_{010}}\\
&+ \mathbb{1}(x_1 = 1)\log\frac{p_{100}}{p_{000}} & &+ \mathbb{1}(x_1 = x_3 = 1)\log\frac{p_{000}p_{101}}{p_{001}p_{100}}\\
&+ \mathbb{1}(x_1 = x_2 = 1)\log\frac{p_{000}p_{110}}{p_{010}p_{100}} & &+ \mathbb{1}(1,\,1,\,1)\log\frac{p_{000}p_{000}p_{111}}{p_{001}p_{010}p_{100}}\cdot\frac{p_{001}p_{010}}{p_{000}p_{011}}\cdot\frac{p_{010}p_{100}}{p_{000}p_{110}}\cdot\frac{p_{001}p_{100}}{p_{000}p_{101}},
\end{aligned}$$

which may be simplified to

$$\begin{aligned}\log f = \log p_{000} \qquad\qquad & + \mathbb{1}(x_3 = 1) \log \frac{p_{001}}{p_{000}} \\ + \mathbb{1}(x_2 = 1) \log \frac{p_{010}}{p_{000}} \qquad\qquad & + \mathbb{1}(x_2 = x_3 = 1) \log \frac{p_{000}p_{011}}{p_{001}p_{010}} \\ + \mathbb{1}(x_1 = 1) \log \frac{p_{100}}{p_{000}} \qquad\qquad & + \mathbb{1}(x_1 = x_3 = 1) \log \frac{p_{000}p_{101}}{p_{001}p_{100}} \\ + \mathbb{1}(x_1 = x_2 = 1) \log \frac{p_{000}p_{110}}{p_{010}p_{100}} \qquad\qquad & + \mathbb{1}(1,\, 1,\, 1) \log \frac{p_{001}p_{010}p_{100}p_{111}}{p_{110}p_{101}p_{011}p_{000}}\end{aligned}$$

In particular note that the terms

$$\log \frac{p_{000}p_{011}}{p_{001}p_{010}},\ \log \frac{p_{000}p_{101}}{p_{001}p_{100}},\ \text{ and } \log \frac{p_{000}p_{110}}{p_{010}p_{100}},$$

which appear in ψ_{23}, ψ_{13}, and ψ_{12} respectively, are precisely the log odds ratio characterizing the pairwise conditional independencies

$$X_2 \amalg X_3 \,|\, X_1 = 0,\ X_1 \amalg X_3 \,|\, X_2 = 0,\ \text{ and } X_1 \amalg X_2 \,|\, X_3 = 0,$$

respectively (as per Theorem 15.6) – see also Exercise A.18.4.

Exercise A.23.23 (Special element of the codomain in log-linear expansions)**.** Let $X = (X_1,\, X_2)$ be a random vector where X_1 and X_2 are binary random variables. Verify that the log-linear expansion of the PDF of X, when treating *one* as the special element of the codomain (and not *zero*), which means that instead of the third defining property of the log-linear expansion in Theorem 19.1 we impose

(3) if $j \in A$ and $x_j = 1$ then $\psi_A(x) = 0$,

is given by

$$f(x_1,\, x_2) = \log p_{11} + (1 - x_1) \log \frac{p_{01}}{p_{11}} + (1 - x_2) \log \frac{p_{10}}{p_{11}} + (1 - x_1)(1 - x_2) \log \frac{p_{00}p_{11}}{p_{01}p_{10}}.$$

Deduce that, as with the usual log-linear expansion of f,

$$\tilde{\beta}_4 = 0 \iff X_1 \amalg X_2$$

(where $\tilde{\beta}_4$ denotes the parameter corresponding to the usual log-linear parameter β_4, but now with respect to the expansion above).

Solution. We observe that

$$\begin{aligned}\log f = \mathbb{1}(x_1 = 0,\, x_2 = 0) \log p_{00} \qquad & + \mathbb{1}(x_1 = 0,\, x_2 = 1) \log p_{01} \\ + \mathbb{1}(x_1 = 1,\, x_2 = 0) \log p_{10} \qquad & + \mathbb{1}(x_1 = 1,\, x_2 = 1) \log p_{11}.\end{aligned}$$

We now write

$$\begin{aligned}\mathbb{1}(x_1 = 1,\, x_2 = 1) = 1 & - \mathbb{1}(x_1 = 0,\, x_2 = 0) \\ & - \mathbb{1}(x_1 = 0,\, x_2 = 1) \\ & - \mathbb{1}(x_1 = 1,\, x_2 = 0)\end{aligned}$$

such that

$$\begin{aligned}\log f = \mathbb{1}(x_1 = 0,\, x_2 = 0) \log \frac{p_{00}}{p_{11}} \qquad & + \mathbb{1}(x_1 = 0,\, x_2 = 1) \log \frac{p_{01}}{p_{11}} \\ + \mathbb{1}(x_1 = 1,\, x_2 = 0) \log \frac{p_{10}}{p_{11}} \qquad & + \log p_{11}.\end{aligned}$$

We then write

$$\begin{aligned}\mathbb{1}(x_1 = 0,\, x_2 = 1) &= \mathbb{1}(x_1 = 0) - \mathbb{1}(x_1 = 0,\, x_2 = 0) \text{ and}\\ \mathbb{1}(x_1 = 1,\, x_2 = 0) &= \mathbb{1}(x_2 = 0) - \mathbb{1}(x_1 = 0,\, x_2 = 0)\end{aligned}$$

such that

$$\begin{aligned}\log f = \mathbb{1}(x_1 = 0,\, x_2 = 0) \log \frac{p_{00}}{p_{11}} \cdot \frac{p_{11}}{p_{01}} \cdot \frac{p_{11}}{p_{10}} \qquad &+ \mathbb{1}(x_1 = 0) \log \frac{p_{01}}{p_{11}}\\ + \mathbb{1}(x_1 = 1) \log \frac{p_{10}}{p_{11}} \qquad &+ \log p_{11},\end{aligned}$$

or in other words

$$\log f = \log p_{11} + (1 - x_1) \log \frac{p_{01}}{p_{11}} + (1 - x_2) \log \frac{p_{10}}{p_{11}} + (1 - x_1)(1 - x_2) \log \frac{p_{00} p_{11}}{p_{01} p_{10}},$$

as desired. In particular we note that

$$\tilde{\psi}_{12} = (1 - x_1)(1 - x_2) \log \frac{p_{00} p_{11}}{p_{01} p_{10}},$$

which means that

$$\tilde{\beta}_4 = \frac{p_{00} p_{11}}{p_{01} p_{10}}.$$

In other words: $\tilde{\beta}_4$ is *precisely* the log odds ratio characterizing the independence of the two binary random variables X_1 and X_2 (see Theorem 15.6), just as happens in the case where *zero* is treated as the special element of the codomain. Indeed, in that latter case the (standard) log-linear expansion of f is

$$f(x_1,\, x_2) = \underbrace{\log p_{00}}_{\beta_1} + x_1 \underbrace{\log \frac{p_{01}}{p_{00}}}_{\beta_2} + x_2 \underbrace{\log \frac{p_{10}}{p_{00}}}_{\beta_3} + x_1 x_2 \underbrace{\log \frac{p_{00} p_{11}}{p_{01} p_{10}}}_{\beta_4}.$$

Exercise A.23.24 (Characterization of conditional independence via log-linear expansions)**.** Prove Theorem 19.7.

Solution. Since $\{A,\, B,\, C\}$ is a partition of $S = \{1,\, \ldots,\, m\}$ we see that, for any $T \subseteq S$, either

$$(T \subseteq A \cup B \text{ or } T \subseteq A \cup C) \text{ or } \underbrace{T \text{ intersects both } B \text{ and } C}_{I(T)},$$

and both propositions cannot occur simultaneously. Therefore

$$\log f = \sum_{\substack{T \subseteq A \cup B \\ \text{or } T \subseteq A \cup C}} \psi_T + \sum_{T : I(T)} \psi_T$$

where, by inclusion–exclusion and the fact that $\{A,\, B,\, C\}$ is a partition,

$$\sum_{\substack{T \subseteq A \cup B \\ \text{or } T \subseteq A \cup C}} \psi_T = \sum_{T \subseteq A \cup B} \psi_T + \sum_{T \subseteq A \cup C} \psi_T - \sum_{T \subseteq A} \psi_T.$$

Exponentiating then yields

$$f = \underbrace{\exp\left(\sum_{T \subseteq A \cup B} \psi_T \right)}_{\text{depends on } (x_A,\, x_B)} \underbrace{\exp\left(\sum_{T \subseteq A \cup C} \psi_T \right)}_{\text{depends on } (x_A,\, x_C)} \underbrace{\exp\left(- \sum_{T \subseteq A} \psi_T \right)}_{\text{depends on } x_A} \underbrace{\exp\left(\sum_{T : I(T)} \psi_T \right)}_{\substack{\text{depends on } (x_B,\, x_C) \\ \text{and possibly } x_A \text{ too.}}}$$

So finally we conclude that

$$\begin{aligned} X_B \amalg X_C \,|\, X_A &\iff f = g(x_A,\, x_B)h(x_A,\, x_C) \\ &\iff \exp\left(\sum_{T:I(T)} \psi_T\right) = 1 \\ &\iff \psi_T = 0 \text{ whenever } T \text{ intersects both } B \text{ and } C. \end{aligned}$$

Exercise A.23.25 (Bias of the leave-one-out cross-validation estimator of the histogram risk)**.** Prove that, for any admissible binwidth h, the bias of the leave-one-out cross-validation estimator of the histogram risk $J(h)$ is

$$\mathbb{E}\left[\hat{J}(h) - J(h)\right] = \frac{2}{n-1}\left(\frac{1}{h} - \mathbb{E}\int \hat{f}_n^2\right).$$

Solution. Introducing $\hat{p}_{(-i),j}$ as in Exercise A.20.4 we may proceed as in Remark 20.17 and compute that

$$\begin{aligned} \frac{1}{2}\left[\hat{J}(h) - J(h)\right] &= \int \hat{f}_n f - \frac{1}{n}\sum_{i=1}^n \hat{f}_{(-i)}(X_i) \\ &= \frac{1}{nh}\sum_{i=1}^n\sum_{j=1}^m \left[\hat{p}_j - \hat{p}_{(-i),j}\right] \mathbb{1}\left(X_i \in B_j\right) \end{aligned}$$

where Exercise A.20.4 tells us that

$$\begin{aligned} \hat{p}_j - \hat{p}_{(-i),j} &= \left(1 - \frac{n}{n-1}\right)\hat{p}_j + \frac{1}{n-1}\mathbb{1}\left(X_i \in B_j\right) \\ &= \frac{-1}{n-1}\hat{p}_j + \frac{1}{n-1}\mathbb{1}\left(X_i \in B_j\right). \end{aligned}$$

Therefore, since

$$\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^m \mathbb{1}\left(X_i \in B_j\right) = 1 \text{ and } \frac{1}{n}\sum_{i=1}^n \hat{p}_j \mathbb{1}\left(X_i \in B_j\right) = \hat{p}_j^2$$

as in Exercise A.20.4, we deduce that

$$\begin{aligned} \frac{1}{2}\left[\hat{J}(h) - J(h)\right] &= \frac{1}{nh}\sum_{i=1}^n\sum_{j=1}^m \left[\frac{1}{n-1}\mathbb{1}\left(X_i \in B_j\right) - \frac{\hat{p}_j}{n-1}\mathbb{1}\left(X_i \in B_j\right)\right] \\ &= \frac{1}{(n-1)h} - \frac{1}{(n-1)h}\sum_{j=1}^m \hat{p}_h^2. \end{aligned}$$

As obtained in Exercise A.20.4,

$$\frac{1}{h}\sum_{j=1}^m \hat{p}_j^2 = \int \hat{f}_n^2$$

and so we conclude that

$$\mathbb{E}\left[\hat{J}(h) - J(h)\right] = \frac{2}{n-1}\left(\frac{1}{h} - \mathbb{E}\int \hat{f}_n^2\right).$$

Exercise A.23.26 (The histogram projection is a projection)**.** Let $f : [0,\, 1] \to \mathbb{R}$ be square-integrable, let $m \geqslant 1$ be an integer, let

$$B_j := \left[\frac{j-1}{m},\, \frac{j}{m}\right) \text{ for } 1 \leqslant j \leqslant m-1 \text{ and } B_M := \left[\frac{m-1}{m},\, 1\right]$$

as in Definition 20.6, and let $\bar{f}$ be the histogram projection of f. If we define the set of step functions

$$S_m := \left\{ s : [0,\, 1] \to \mathbb{R} \text{ such that } s|_{B_j} \text{ is constant} \right\}$$

then $\bar{f}$ is the L^2–projection of f onto S_m in the sense that

$$\bar{f} = \operatorname*{arg\,min}_{g \in S_m} \int (f-g)^2.$$

Solution. For any $g \in S_m$ we have that

$$\int (f-g)^2 = \sum_{j=1}^{m} \int_{B_j} (f-g_j)^2$$

where g_j denotes the constant value of g on B_j. The minimization problem thus splits into m independent subproblems, one for each interval B_j. For each interval B_j we may then proceed as in item 1 of Exercise A.23.7 and observe that

$$\operatorname*{arg\,min}_{s \in \mathbb{R}} \int_{B_j} (f-s)^2 = \frac{1}{|B_j|} \int_{B_j} f.$$

Indeed $s \mapsto \int_{B_j} (f-s)^2$ is strictly convex with derivative

$$\frac{d}{ds} \int_{B_j} (f-s)^2 = -2 \int_{B_j} (f-s)\,,$$

which vanishes when

$$|B_j| s = \int_{B_j} f \iff s = \frac{1}{|B_j|} \int_{B_j} f.$$

In other words the minimizer g^* satisfies

$$g_j^* = \frac{1}{|B_j|} \int_{B_j} f = \frac{p_j}{1/m} = \frac{p_j}{h} = \bar{f}|_{B_j},$$

i.e. the minimizer is $g^* = \bar{f}$ as desired.

Exercise A.23.27 (Iterated kernel)**.** The *iterated kernel* is defined to be

$$K^{(2)} := K * K$$

for any kernel K, i.e.

$$K^{(2)}(z) = \int K(z-y)K(y)dy.$$

Prove that $K^{(2)}$ is a kernel with variance $2\sigma_K^2$. Moreover prove that if K is *symmetric*, meaning that $K(-x) = K(x)$ for all $x \in \mathbb{R}$, then so is $K^{(2)}$.

Solution. Since K is non-negative, so is $K^{(2)}$. Since K is integrable, non-negative, and normalized, Tonelli's Theorem and the change of variable $\bar{z} = z - y$ tell us that $K^{(2)}$ is also integrable and normalized since

$$\begin{aligned}\int K^{(2)}(z)dz &= \int\int K(z-y)K(y)dydz \\ &= \int K(y)\left(\int K(z-y)dz\right)dy \\ &= \int K(y)\left(\int K(\bar{z})d\bar{z}\right)dy \\ &= \int K(y)dy \\ &= 1.\end{aligned}$$

The same change of variable tells us that $K^{(2)}$ has mean zero:

$$\begin{aligned}\int zK^{(2)}(z)dz &= \int\int zK(z-y)K(y)dydz \\ &= \int\int(\bar{z}+y)K(\bar{z})K(y)dyd\bar{z} \\ &= \int(0+y)K(y)dy \\ &= 0.\end{aligned}$$

Similarly, this same change of variables tells us that the variance of $K^{(2)}$ is

$$\begin{aligned}&\int z^2K^{(2)}(z)dz \\ &= \int\int z^2K(z-y)K(y)dydz \\ &= \int\int(\bar{z}+y)^2K(\bar{z})K(y)dyd\bar{z} \\ &= \int \bar{z}^2K(\bar{z})d\bar{z} + 2\int\left(\int \bar{z}K(\bar{z})d\bar{z}\right)yK(y)dy + \int y^2K(y)dy \\ &= \sigma_K^2 + 0 + \sigma_K^2 \\ &= 2\sigma_K^2,\end{aligned}$$

as desired.

Finally, if K is symmetric, then so is $K^{(2)}$ since changing variables from y to $-y$ and using the symmetry of K yields

$$\begin{aligned}K^{(2)}(-z) &= \int K(-z-y)K(y)dy \\ &= \int K(-z+y)K(-y)dy \\ &= \int K(z-y)K(y)dy \\ &= K(z).\end{aligned}$$

Exercise A.23.28 (Examples of iterated kernels)**.** Recall from Exercise A.23.27 that, for any kernel K, the *iterated* kernel $K^{(2)}$ is defined to be $K^{(2)} := K * K$.

(1) Let $K = \varphi$ where φ is the PDF of a standard Normal random variable, i.e. K is the Gaussian kernel. Verify that $K^{(2)} = \frac{1}{\sqrt{2}}\varphi\left(\frac{\cdot}{\sqrt{2}}\right)$, i.e. the iterated kernel is a Gaussian kernel with variance $\sigma^2 = 2$.

(2) Let K be the boxcar kernel. Verify that

$$K^{(2)}(z) = (1 - |z|)\,\mathbb{1}\,(|z| < 1)\,,$$

which is known as the *triangular kernel*.

Solution. (1) We compute that

$$K^{(2)}(z) = \int \frac{1}{\sqrt{2\pi}}e^{-(z-y)^2/2} \cdot \frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy$$

where

$$y^2 + (z-y)^2 = 2y^2 - 2yz + z^2 = 2\Big(y - \frac{z}{2}\Big)^2 + \frac{z^2}{2}$$

such that

$$K^{(2)}(z) = \frac{1}{\sqrt{4\pi}}e^{-z^2/4}\underbrace{\int \frac{1}{\sqrt{\pi}}e^{-(y-z/2)^2}dy}_{=1} = \frac{1}{\sqrt{2}}\varphi\left(\frac{z}{\sqrt{2}}\right),$$

as desired.

(2) We compute that

$$\begin{aligned} K^{(2)}(z) &= \int \mathbb{1}\left(|z-y| < \frac{1}{2}\right)\mathbb{1}\left(|y| < \frac{1}{2}\right)dy \\ &= \int \mathbb{1}_{\left(z-\frac{1}{2},\, z+\frac{1}{2}\right)}(y)\mathbb{1}_{\left(-\frac{1}{2},\, \frac{1}{2}\right)}(y)dy. \end{aligned}$$

Since the intervals $\left(z - \frac{1}{2},\, z + \frac{1}{2}\right)$ and $\left(-\frac{1}{2},\, \frac{1}{2}\right)$ intersect if and only if $|z| < 1$ we deduce that $K^{(2)}(z) = 0$ if $|z| \geqslant 1$. For $|z| < 1$ we split the computation into two cases. If $z \in (-1,\, 0]$ then

$$K^{(2)}(z) = \int_{-\frac{1}{2}}^{z+\frac{1}{2}} dy = \left(z + \frac{1}{2}\right) - \left(-\frac{1}{2}\right) = z + 1$$

and if $z \in [0,\, 1)$ then

$$K^{(2)}(z) = \int_{z-\frac{1}{2}}^{\frac{1}{2}} dy = \frac{1}{2} - \left(z - \frac{1}{2}\right) = 1 - z.$$

In summary:

$$K^{(2)}(z) = (1 - |z|)\,\mathbb{1}\,(|z| < 1)\,,$$

as desired.

Exercise A.23.29 (Expectation of kernel density estimator)**.** Prove Lemma 20.42 where we record the expectation of kernel density estimators.

Solution. Fix $x \in \mathbb{R}$. Since

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

it follows from the Rule of the Lazy Statistician and the IID assumption that

$$\begin{aligned}\mathbb{E}\hat{f}_n(x) &= \mathbb{E}\left[\frac{1}{h}K\left(\frac{x - X_i}{h}\right)\right] \\ &= \int \frac{1}{h}K\left(\frac{x-y}{h}\right) f(y)dy \\ &= \left[\frac{1}{h}K\left(\frac{\cdot}{h}\right) * f\right](x),\end{aligned}$$

as desired.

Exercise A.23.30 (Identity for the variance of the regression error term)**.** Let $(Y,\, X)$ be a random vector, let $r(x) := \mathbb{E}\left(Y \,|\, X = x\right)$ be the regression function between Y and X, and let $\varepsilon := Y - r(X)$ be the "error" term. Then

$$\mathbb{V}\varepsilon = \mathbb{E}\mathbb{V}\left(Y \,|\, X\right).$$

Solution. The rule of iterated expectation tells us that Y and $r(X) = \mathbb{E}\left(Y \,|\, X\right)$ have the same expectation. Therefore

$$\begin{aligned}\mathbb{V}\varepsilon = \mathbb{V}\left[Y - r(X)\right] = \mathbb{E}\left([Y - r(X)]^2\right) &= \mathbb{E}\left([Y - \mathbb{E}\left(Y \,|\, X\right)]^2\right) \\ &= \mathbb{E}\left(\mathbb{E}\left([Y - \mathbb{E}\left(Y \,|\, X\right)]^2 \,\middle|\, X\right)\right)\end{aligned}$$

where the inner expectation is, by definition, the conditional variance:

$$\mathbb{E}\left([Y - \mathbb{E}\left(Y \,|\, X\right)]^2 \,\middle|\, X\right) = \mathbb{V}\left(Y \,|\, X\right).$$

Therefore

$$\mathbb{V}\varepsilon = \mathbb{E}\mathbb{V}\left(Y \,|\, X\right)$$

as desired.

Exercise A.23.31 (Limiting behaviour of the empirical estimates of the regressoin function basis coefficients)**.** Prove Theorem 21.25 where we establish the limiting behaviour of the empirical estimates of the regression function basis coefficients.

Solution. Since $\hat{\beta}_j$ is a sample mean, by the Central Limit Theorem it suffices to show that

$$\mathbb{E}\left[Y\phi_j(X)\right] = \beta_j \text{ and } \mathbb{V}\left[Y\phi_j(X)\right] = \sigma^2.$$

We first compute the expectation. Since $X \sim \text{Uniform}(0,\, 1)$ we observe that

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = f(x,y).$$

Therefore

$$\begin{aligned}\mathbb{E}\left[Y\phi_j(X)\right] &= \int_0^1 \int_{\mathbb{R}} y\phi_j(x)f(x,y)dydx\\ &= \int_0^1 \left[\int_{\mathbb{R}} yf(y|x)dy\right]\phi_j(x)dx\\ &= \int_0^1 \mathbb{E}\left(Y \mid X = x\right)\phi_j(x)dx\\ &= \int_0^1 r(x)\phi_j(x)dx\\ &= \beta_j,\end{aligned}$$

as desired. We now compute the variance. The rule of iterated expectation and homoscedasticity tells us that

$$\begin{aligned}\mathbb{V}\left[Y\phi_j(X)\right] &= \mathbb{E}\left(\mathbb{V}\left[Y\phi_j(X) \mid X\right]\right)\\ &= \mathbb{E}\left[\phi_j^2(X)\mathbb{V}\left(Y \mid X\right)\right]\\ &= \sigma^2\mathbb{E}\left[\phi_j^2(X)\right].\end{aligned}$$

Since ϕ_j is normal and since $X \sim \text{Uniform}(0,\, 1)$ we deduce that

$$\mathbb{E}\left[\phi_j^2(X)\right] = \int_0^1 \phi_j^2(x)f_X(x)dx = \int_0^1 \phi_j^2(x)dx = 1,$$

such that indeed $\mathbb{V}\left[Y\phi_j(X)\right] = \sigma^2$, as desired.

Exercise A.23.32 (The Haar system is an orthonormal basis)**.** Prove Theorem 21.37 where we show that the Haar system is an orthonormal basis.

Solution. A dyadic step function of resolution J is the name we give to a step function on $[0,\, 1]$ which is constant on the dyadic sub-intervals

$$\left[0,\, \frac{1}{2^J}\right),\, \ldots,\, \left[\frac{k}{2^J},\, \frac{k+1}{2^J}\right),\, \ldots,\, \left[\frac{2^J-1}{2^J},\, 1\right)$$

for $0 \leqslant k \leqslant 2^J - 1$. Since any $f \in L_2[0,\, 1]$ may be approximated by a sequence $(f_J)_{J=0}^{\infty}$ where each f_J is a dyadic step function of resolution J, it suffices to show that every dyadic step function may be written as a finite linear combination of the Haar father wavelet and Haar children wavelets.

We prove this by inducting on J, proving that for every J the dyadic step functions

$$s_{J,k} := \mathbb{1}_{[k/2^{-J},\, (k+1)/2^{-J}]},\, 0 \leqslant k \leqslant 2^J - 1,$$

may be written as a linear combination of the Haar wavelets. It then immediately follows that *every* dyadic step function of resolution J may be written as a linear combination of Haar wavelets.

The key idea is encapsulated in the step going from $J = 0$ to $J = 1$:

$$\frac{\phi+\psi}{2} = \mathbb{1}_{[1/2,\, 1)} = s_{1,1} \text{ and } \frac{\phi-\psi}{2} = \mathbb{1}_{[0,\, 1/2)} = s_{1,0}.$$

Then adding and subtracting $\psi_{j,k}$ for larger j produces finer and finer dyadic step functions. Let us now be more precise.

The base case $J = 0$ is trivial: $\phi = \mathbb{1}_{[0,\,1)} = s_{0,0}$. Now for the induction step suppose that, for some $J \geqslant 0$, the step functions

$$s_{J,k} \text{ for } 0 \leqslant k \leqslant 2^J - 1$$

may be written as linear combinations of father and children wavelets. then

$$\frac{1}{2}\left(s_{J,k} + 2{-}(J+1)/2\psi_{J+1,k}\right) = s_{J+1,k+1} \text{ and}$$
$$\frac{1}{2}\left(s_{J,k} - 2{-}(J+1)/2\psi_{J+1,k}\right) = s_{J+1,k}$$

and so the step functions

$$s_{J+1,k} \text{ for } 0 \leqslant k \leqslant 2^{J+1} - 1$$

may *also* all be written as linear combinations of father and children wavelets. This concludes the induction argument, which in turn concludes the proof.

Exercise A.23.33 (CDF of a symmetric PDF)**.** Let f be a PDF symmetric about some $x_0 \in \mathbb{R}$ and let F be the corresponding CDF. Then

$$F(x_0 - x) = 1 - F(x_0 + x)$$

for every $x \in \mathbb{R}$.

Proof. We compute that, by symmetry of f about x_0,

$$\begin{aligned}
F(x - x_0) + F(x + x_0) &= \int_{-\infty}^{x_0+x} f + \int_{-\infty}^{x_0+x} f \\
&= \int_{-\infty}^{x_0+x} f + \int_{-\infty}^{x_0} f + \int_{x_0}^{x_0+x} f \\
&= \int_{-\infty}^{x_0+x} f + \int_{x_0}^{+\infty} f + \int_{x_0-x}^{x_0} f \\
&= \int_{-\infty}^{+\infty} f = 1,
\end{aligned}$$

from which the claim follows. □

Exercise A.23.34 (Identity for the median absolute deviation)**.** Prove Lemma 3.24 where we establish an identity for the median absolute deviation in terms of the median and the quantile function.

Solution. The median absolute deviation of X is any number $\nu \in \mathbb{R}$ which satisfies

$$\mathbb{P}\left(|X - m| \leqslant \nu\right) = \frac{1}{2}.$$

Writing this equation in terms of the CDF F and using Exercise A.23.33 we deduce that

$$\frac{1}{2} = \mathbb{P}\left(m - \nu \leqslant X \leqslant m + \mu\right) = F(m+\nu) - F(m-\nu) = 2F(m+\nu) - 1.$$

In particular since f is strictly positive everywhere we know that F is strictly increasing, and hence invertible. Therefore ν satisfies

$$F(m+\nu) = \frac{3}{4} \iff \nu = F^{-1}\left(\frac{3}{4}\right) - m,$$

as claimed.

Exercise A.23.35 (Linear Bayes classifier)**.** Deduce Theorem 22.20 from Theorem 22.16, i.e. show that when $\mathbb{V}(X \mid Y=0) = \mathbb{V}(X \mid Y=1)$ the quadratic Bayes classifier becomes linear.

Solution. Theorem 22.16 tells us that

$$h^*(x) = \mathbb{1}\left(\tilde{\delta}_0(x) > \tilde{\delta}_1(x)\right)$$

where

$$\begin{aligned}
\tilde{\delta}_1(x) - \tilde{\delta}_0(x) &= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}r_1^2(x) + \log\pi_1 \\
&\quad + \frac{1}{2}\log|\Sigma| + \frac{1}{2}r_0^2(x) - \log\pi_0 \\
&= -\frac{1}{2}\Sigma^{-1}(x-\mu_1)\cdot(x-\mu_1) + \log\pi_1 \\
&\quad + \frac{1}{2}\Sigma^{-1}(x-\mu_0)\cdot(x-\mu_0) - \log\pi_0 \\
&= -\frac{1}{2}\Sigma^{-1}x\cdot x + \Sigma^{-1}x\cdot\mu_1 - \frac{1}{2}\Sigma^{-1}\mu_1\cdot\mu_1 + \log\pi_1 \\
&\quad + \frac{1}{2}\Sigma^{-1}x\cdot x - \Sigma^{-1}x\cdot\mu_0 + \frac{1}{2}\Sigma^{-1}\mu_0\cdot\mu_0 - \log\pi_0 \\
&= \delta_1(x) - \delta_0(x),
\end{aligned}$$

as desired. In other words: the difference between two quadratic forms with the same leading-order tems is a linear function.

Exercise A.23.36 (Between-class and within-class variances for classification)**.** Let Y be a random variable with finite codomain $\{0, \ldots, K-1\}$ and let X be a random vector in $\mathbb{R}^d$. We denote

$$\pi_k := \mathbb{P}(Y=k),\ \mu_k := \mathbb{E}(X \mid Y=k),\ \text{and } \Sigma_k := \mathbb{V}(X \mid Y=k)$$

for $k = 0, \ldots, K-1$. Prove that the between-class variance of X is

$$\mathbb{V}\mathbb{E}(X \mid Y) = \sum_{k=0}^{K-1} \pi_k\mu_k \otimes \mu_k - \left(\sum_{k=0}^{K-1}\pi_k\mu_k\right) \otimes \left(\sum_{l=0}^{K-1}\pi_l\mu_l\right)$$

and that the within-class variance of X is

$$\mathbb{E}\mathbb{V}(X \mid Y) = \sum_{k=0}^{K-1}\pi_k\Sigma_k.$$

Solution. We begin with the within-class variance. Since Y has codomain $\{0, \ldots, K-1\}$ we compute that

$$\mathbb{E}\mathbb{V}(X \mid Y) = \sum_{k=0}^{K-1}\mathbb{V}(X \mid Y=k)\,\mathbb{P}(Y=k) = \sum_{k=0}^{K-1}\Sigma_k\pi_k,$$

as desired. We now turn our attention to the between-class variance. We note that

$$\mathbb{E}(X \mid Y) \text{ takes value } \mu_k \text{ with probability } \pi_k$$

or, in other words,

$$\mathbb{E}(X \mid Y) = \mu Z \text{ for } Z \sim \text{Categorical}(\pi)$$

and for μ the d-by-K matrix whose k-th column is μ_k. We may then use Theorem 14.1 to compute that

$$\mathbb{V}\mathbb{E}\left(X \,|\, Y\right) = \mathbb{V}\left(\mu Z\right) = \mu \mathbb{V}\left(Z\right) \mu^T = \mu \left(\operatorname{diag} \pi - \pi \otimes \pi\right) \mu^T$$

where

$$\left[\mu \left(\operatorname{diag} \pi\right) \mu^T\right]_{ij} = \sum_{k,l} \mu_{ik} \pi_k \delta_{kl} \mu_{jl} = \sum_k \pi_k \mu_{ik} \mu_{jk} = \sum_k \pi_k (\mu_k \otimes \mu_k)_{ij}$$

and

$$\begin{aligned}\left[\mu \left(\pi \otimes \pi\right) \mu^T\right]_{ij} = \sum_{k,l} \mu_{ik} \pi_k \pi_l \mu_{jl} &= \left(\sum_k \pi_k \mu_{ik}\right) \left(\sum_l \pi_l \mu_{jl}\right) \\ &= \left(\sum_k \pi_k \mu_k\right)_i \left(\sum_l \pi_l \mu_l\right)_j .\end{aligned}$$

So indeed

$$\mathbb{V}\mathbb{E}\left(X \,|\, Y\right) = \mu \left(\operatorname{diag} \pi - \pi \otimes \pi\right) \mu^T = \sum_k \pi_k \mu_k \otimes \mu_k - \left(\sum_k \pi_k \mu_k\right) \otimes \left(\sum_l \pi_l \mu_l\right).$$

Exercise A.23.37 (Fisher's linear discriminant function and a generalized Rayleigh quotient)**.** Let Y be a binary random variable and let X be a random vector in $\mathbb{R}^d$. Suppose that

$$\mathbb{P}\left(Y = 0\right) = \mathbb{P}\left(Y = 1\right) = \frac{1}{2} \text{ and } \mathbb{V}\left(X \,|\, Y = 0\right) = \mathbb{V}\left(X \,|\, Y = 1\right) = \Sigma$$

and let

$$\mu_0 := \mathbb{E}\left(X \,|\, Y = 0\right) \text{ and } \mu_1 := \mathbb{E}\left(X \,|\, Y = 1\right).$$

Prove that Fisher's linear discriminant function $u(x) = w \cdot x$ is characterized by

$$w = \operatorname*{arg\,max}_{v \in \mathbb{R}^d} \frac{\left(\mu_0 - \mu_1\right) \otimes \left(\mu_0 - \mu_1\right) v \cdot v}{\Sigma v \cdot v}.$$

Solution. As observed in Remark 22.30 w is the maximizer of the Fisher discriminant ratio of $v \cdot X$ over $v \in \mathbb{R}^d$, i.e.

$$w = \operatorname*{arg\,max}_{v \in \mathbb{R}^d} \frac{\mathbb{V}\mathbb{E}\left(X \,|\, Y\right) v \cdot v}{\mathbb{E}\mathbb{V}\left(X \,|\, Y\right) v \cdot v}.$$

Exercise A.23.36 allows us to compute both the between-class variance and the within-class variance: writing $\pi_i := \mathbb{P}\left(Y = i\right) = \frac{1}{2}$ and $\Sigma_i := \mathbb{V}\left(X \,|\, Y = i\right) = \Sigma$, we obtain that

$$\begin{aligned}\mathbb{V}\mathbb{E}\left(X \,|\, Y\right) &= \pi_0 \mu_0 \otimes \mu_0 + \pi_1 \mu_1 \otimes \mu_1 - \left(\pi_0 \mu_0 + \pi_1 \mu_1\right) \otimes \left(\pi_0 \mu_0 + \pi_1 \mu_1\right) \\ &= \frac{1}{2} \mu_0 \otimes \mu_0 + \frac{1}{2} \mu_1 \otimes \mu_1 - \frac{1}{4} \left(\mu_0 + \mu_1\right) \otimes \left(\mu_0 + \mu_1\right) \\ &= \frac{1}{4} \left(\mu_0 \otimes \mu_0 + \mu_1 \otimes \mu_1 - \mu_0 \otimes \mu_1 - \mu_1 \otimes \mu_0\right) \\ &= \frac{1}{4} \left(\mu_0 - \mu_1\right) \otimes \left(\mu_0 - \mu_1\right)\end{aligned}$$

and

$$\mathbb{E}\mathbb{V}\left(X \,|\, Y\right) = \pi_0 \Sigma_0 + \pi_1 \Sigma_1 = \frac{1}{2} \left(\Sigma_0 + \Sigma_1\right) = \Sigma.$$

Therefore the Fisher discriminant ratio of $v \cdot X$ is

$$\frac{\mathbb{V}\mathbb{E}\left(X \,|\, Y\right) v \cdot v}{\mathbb{E}\mathbb{V}\left(X \,|\, Y\right) v \cdot v} = \frac{\frac{1}{4}\left(\mu_0 - \mu_1\right) \otimes \left(\mu_0 - \mu_1\right) v \cdot v}{\Sigma v \cdot v},$$

as desired since the factor of $1/4$ is irrelevant when it comes to maximizing the ratio.

Exercise A.23.38 (Validity of Fisher's linear classifier)**.** Prove that the choice of w in Definition 22.31 is valid, i.e. prove that

$$\Sigma^{-1}\left(\mu_1 - \mu_0\right) \in \operatorname*{arg\,max}_{v \in \mathbb{R}^d} \frac{\mathbb{V}\mathbb{E}\left(X \,|\, Y\right) v \cdot v}{\mathbb{E}\mathbb{V}\left(X \,|\, Y\right) v \cdot v}.$$

Solution. Exercise A.23.37 tells us that Fisher's linear discriminant function is characterized by

$$w = \operatorname*{arg\,max}_{v \in \mathbb{R}^d} \frac{\left(\mu_0 - \mu_1\right) \otimes \left(\mu_0 - \mu_1\right) v \cdot v}{\Sigma v \cdot v}$$

and Corollary C.11 tells us precisely that the maximizer of this ratio is any multiple of

$$\Sigma^{-1}\left(\mu_1 - \mu_0\right),$$

as desired.

Exercise A.23.39 (Fisher's linear classification estimator and LDA)**.** Show that Fisher's linear classification estimator agrees with the LDA classification estimator when $\hat{\pi}_0 = \hat{\pi}_1 = \frac{1}{2}$.

Proof. When $\hat{\pi}_0 = \hat{\pi}_1 = \frac{1}{2}$ the linear discriminant function estimates $\hat{\delta}_0$ and $\hat{\delta}_1$ in Definition 22.21 simplify to

$$\begin{aligned}\hat{\delta}_1(x) - \hat{\delta}_0(x) &= \hat{S}^{-1}\left(\hat{\mu}_1 - \hat{\mu}_0\right) \cdot x + \frac{1}{2}\hat{S}^{-1}\hat{\mu}_0 \cdot \hat{\mu}_0 - \frac{1}{2}\hat{S}^{-1}\hat{\mu}_1 \cdot \hat{\mu}_1 \\ &= \hat{S}^{-1}\left(\hat{\mu}_1 - \hat{\mu}_0\right) \cdot x + \frac{1}{2}\hat{S}^{-1}\left(\hat{\mu}_0 - \hat{\mu}_1\right) \cdot \left(\hat{\mu}_0 + \hat{\mu}_1\right).\end{aligned}$$

In the notation of Fisher's linear classification estimator from Definition 22.34 this reads

$$\hat{\delta}_1(x) - \hat{\delta}_0(x) = \hat{w} \cdot x - \hat{w} \cdot \hat{m} = \hat{u}(x) - \hat{w} \cdot \hat{m}$$

and so indeed

$$\hat{\delta}_1 > \hat{\delta}_0 \iff \hat{u} > \hat{w} \cdot \hat{m},$$

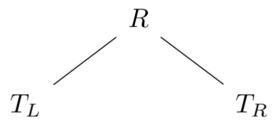
which proves that Fisher's linear classification estimator agrees with the LDA classification estimator. □

Exercise A.23.40 (Tree partition)**.** Prove Lemma 22.50 which states that the leaves of a tree partition of $R \subseteq \mathbb{R}^d$ form a partition of R.

Solution. This follows from the definition of a tree partition by structural induction.

The claim holds when the partition tree only has one vertex, namely R, so the base case holds.

Now suppose that the claim holds for two tree partitions T_L and T_R such that



is also a tree partition (see Definition 22.45 for the meaning of this diagram). Then the leaves of T_L partition the root of T_L and the leaves of T_R partition the root of T_R (by the recursive hypothesis). But, by definition of a partition tree,

$$\{\text{root of } T_L, \text{ root of } T_R\}$$

is itself a partition of R. Therefore

$$\{\text{leaves of } T_L\} \cup \{\text{leaves of } T_R\}$$

is a partition of R, which proves the recursive step and thus concludes the proof.

Exercise A.23.41 (Range of the Gini index)**.** Prove Lemma 22.58 where we record that the range of the Gini index is

$$0 \leqslant \mathcal{G} \leqslant 1 - \frac{1}{K}.$$

Solution. Since $\mathcal{G}(R) = 1 - p \cdot p$ with $p \in \Delta^{K-1} \subseteq \mathbb{R}^K$ we see immediately that

$$\mathcal{G}(R) \geqslant 1 - |p|^2 \geqslant 1 - 1 = 0$$

with equality when $p \cdot p = 1$, which occurs precisely when p is a pure state, i.e. when $p = e_k$ for some $0 \leqslant k \leqslant K-1$.

To find the maximum of $\mathcal{G}$ we use the method of Lagrange multipliers and consider

$$f\left(p,\, \lambda\right) := \frac{1}{2} p \cdot p - \lambda \left(p \cdot \mathbb{1} - 1\right) \text{ for } p \in \mathbb{R}^d_{>0} \text{ and } \lambda \in \mathbb{R},$$

since maximizing $1 - p \cdot p$ is equivalent to minimizing $\frac{1}{2} p \cdot p$ and since the constraint $p \in \Delta^{K-1}$ may be written as $p \cdot \mathbb{1} = 1$. Then

$$\nabla_p f = p - \lambda \mathbb{1} \text{ and } \partial_\lambda f = p \cdot \mathbb{1} - 1.$$

Equating $\nabla p f$ to zero tells us that the minimizer is $p = \lambda \mathbb{1}$ for some $\lambda \in \mathbb{R}$, and plugging that into $\partial_\lambda f = 0$ allows us to conclude that

$$\lambda \mathbb{1} \cdot \mathbb{1} = 1 \iff \lambda = \frac{1}{K}, \text{ since } \mathbb{1} \cdot \mathbb{1} = K,$$

and so $p_* = \frac{1}{K} \mathbb{1}$ as claimed. Finally: when $p_* = \frac{1}{K} \mathbb{1}$ we obtain that

$$\mathcal{G} = 1 - \frac{1}{K^2} \mathbb{1} \cdot \mathbb{1} = 1 - \frac{K}{K^2} = 1 - \frac{1}{K},$$

i.e. indeed the maximum value of $\mathcal{G}$ is $1 - \frac{1}{K}$.

Exercise A.23.42 (Gini index and one-hot encoding)**.** Prove Lemma 22.61 where we establish that

$$\mathcal{G}(R) = \operatorname{tr} \mathbb{V}\left(Z \,|\, X \in R\right)$$

where $\mathcal{G}$ is the Gini index of Y and Z is the categorical version of Y.

Solution. By definition of one-hot encoding,

$$Z \,|\, X \in R \sim \text{Categorical}(p)$$

for $p_k := \mathbb{P}\left(Y = k \,|\, X \in R\right)$ and so Exercise A.23.12 tells us that

$$\mathbb{V}\left(Z \,|\, X \in R\right) = \operatorname{diag} p - p \otimes p.$$

Therefore

$$\operatorname{tr} \mathbb{V}\left(Z \,|\, X \in R\right) = \sum_{k=0}^{K-1} \left(p_k - p_k^2\right) = \sum_{k=0}^{K-1} p_k \left(1 - p_k\right) = \mathcal{G}(R),$$

as desired.

Theorem A.6 (Radon)**.** *Any set of $d+2$ points in the $\mathbb{R}^d$ can be partitioned into two sets whose convex hulls intersect.*

Exercise A.23.43 (Shattering four points in the plan with affine half-spaces)**.** Prove that any set of four points in the plane cannot be shattered by the collection of affine half-spaces.

Solution. Let w, x, y, and z be distinct points in the plane $\mathbb{R}^2$. By hese four points may be partitioned into two sets whose convex hulls intersect.

- **Case 1:** One of the four points, say w, is in the convex hull of the other three, say x, y, and z. Affine half-spaces are convex and so any half-space containing $\{x,\, y,\, z\}$ must also contain $\{w\}$. Therefore the subset $\{x,\, y,\, z\}$ cannot be picked out by half-spaces.
- **Case 2:** None of the four points lie in the convex hull of the other four, so the partition produced by Radon's Theorem consists of two sets with two points, say without loss of generality $\{w,\, x\}$ and $\{y,\, z\}$. The convex hull of these sets are then nothing more than line segments S and T.

  We claim that affine half-spaces cannot pick out $\{w,\, x\}$. For the sake of contradiction, suppose they did. Then S, and hence $S \cap T$, would be contained in an affine half-space. But T lies entirely outside the half-space since both y and z lie there. This is a contradiction, which proves that $\{w,\, x\}$ cannot be picked out by half-spaces.

Either way we have found a subset of $\{w,\, x,\, y,\, z\}$ which cannot be picked out by affine half-spaces. This proves that these half-spaces do *not* shatter $\{w,\, x,\, y,\, z\}$.

Exercise A.23.44 (Distance between sets)**.** Prove that $d\,(A,\, B) = \inf\limits_{a\in A} d\,(a,\, B)$ for any $A, B \subseteq \mathbb{R}^d$.

Solution. Let us denote $D := \inf\limits_{a\in A} (A,\, B)$. First we show that $d\,(A,\, B) \geqslant D$. By definition of $d\,(A,\, B)$ we see that, for every $a \in A$ and $b \in B$,

$$d\,(a,\, b) \geqslant d\,(a,\, B) \geqslant D.$$

Taking the infimum over $A \times B$ on the left-hand side yields $d\,(A,\, B) \geqslant D$.

Now we show that $d\,(A,\, B) \leqslant D$. It is enough to show that $d\,(A,\, B) \leqslant D + \varepsilon$ for every $\varepsilon > 0$. So let us fix $\varepsilon > 0$. By definition of D there exists $a_\varepsilon \in A$ such that

$$d\,(a_\varepsilon,\, B) \leqslant D + \frac{\varepsilon}{2}.$$

By definition of $d\,(a_\varepsilon,\, B)$ there exists $b_\varepsilon \in B$ such that

$$d\,(a_\varepsilon,\, b_\varepsilon) \leqslant d\,(a_\varepsilon,\, B) + \frac{\varepsilon}{2}.$$

Therefore

$$d\,(A,\, B) \leqslant d\,(a_\varepsilon,\, b_\varepsilon) \leqslant D + \varepsilon,$$

as desired.

Exercise A.23.45 (Feature map of the polynomial kernel of degree two)**.** Prove that in dimension $d = 2$ the feature map corresponding to the polynomial kernel of degree 2,

$$K\,(x,\, \tilde{x}) := (1 + x \cdot \tilde{x})^2$$

where $x,\, \tilde{x} \in \mathbb{R}^d$, is

$$\phi\left(x\right) = \left(1,\, \sqrt{2}x_1,\, \sqrt{2}x_2,\, x_1^2,\, \sqrt{2}x_1x_2,\, x_2^2\right).$$

Solution. This follows from a direct computation

$$\begin{aligned} K\left(x,\, \tilde{x}\right) &= \left(1 + x_1\tilde{x}_1 + x_2\tilde{x}_2\right)^2 \\ &= 1 + x_1^2\tilde{x}_1^2 + x_2^2\tilde{x}_2^2 + 2\left(x_1\tilde{x}_1 + x_2\tilde{x}_2 + x_1\tilde{x}_1x_2\tilde{x}_2\right) \\ &= \left(1,\, \sqrt{2}x_1,\, \sqrt{2}x_2,\, x_1^2,\, \sqrt{2}x_1x_2,\, x_2^2\right) \cdot \left(1,\, \sqrt{2}\tilde{x}_1,\, \sqrt{2}\tilde{x}_2,\, \tilde{x}_1^2,\, \sqrt{2}\tilde{x}_1\tilde{x}_2,\, \tilde{x}_2^2\right) \\ &= \phi(x) \cdot \phi(\tilde{x}), \end{aligned}$$

as desired.

Theorem A.7 (Multinomial expansion)**.** *For any integer $r \geqslant 0$ and any $a \in \mathbb{R}^n$*

$$\left(\sum_{i=1}^{n} a_i\right)^r = \sum_{\substack{p \in \mathbb{N}^d \\ |p| = r}} \binom{r}{p} a^r,$$

where each p is a multi-index.

Exercise A.23.46 (Feature map of the polynomial kernel of degree three)**.** Prove that in dimension $d = 2$ the feature map corresponding to the polynomial kernel of degree 3,

$$K\left(x,\, \tilde{x}\right) := \left(1 + x \cdot \tilde{x}\right)^3$$

where $x,\, \tilde{x} \in \mathbb{R}^d$, is

$$\phi\left(x\right) = \left(1,\, \sqrt{3}x_1,\, \sqrt{3}x_2,\, \sqrt{3}x_1^2,\, \sqrt{6}x_1x_2,\, \sqrt{3}x_2^2,\, x_1^3,\, \sqrt{3}x_1^2x_2,\, \sqrt{3}x_1x_2^2,\, x_2^3\right).$$

Solution. By the multinomial expansion theorem

$$\begin{aligned} \left(1 + a + b\right)^4 &= \sum_{|p| = 3} \binom{3}{p} a^{p_2} b^{p_3} \\ &= \binom{3}{3,\, 0,\, 0} + \binom{3}{0,\, 3,\, 0} a^3 + \binom{3}{0,\, 0,\, 3} b^3 \\ &+ \binom{3}{2,\, 1,\, 0} a + \binom{3}{0,\, 2,\, 1} a^2 b + \binom{3}{1,\, 0,\, 2} b^2 \\ &+ \binom{3}{2,\, 0,\, 1} b + \binom{3}{1,\, 2,\, 0} a^2 + \binom{3}{0,\, 1,\, 2} ab^2 \\ &+ \binom{3}{1,\, 1,\, 1} ab \\ &= 1 + a^3 + b^3 + \binom{3}{1}\left(a + a^2 b + b^2 + b + a^2 + ab^2\right) + \frac{3!}{(1!)^3} ab \\ &= 1 + a^3 + b^3 + 3\left(a + a^2 b + b^2 + b + a^2 + ab^2\right) + 6ab. \end{aligned}$$

Therefore, for $a = x_1\tilde{x}_1$ and $b = x_2\tilde{x}_2$,

$$\begin{aligned}
K(x,\, \tilde{x}) &= (1 + x \cdot \tilde{x})^3 \\
&= (1 + a + b)^3 \\
&= 1 + x_1^3\tilde{x}_1^3 + x_2^3\tilde{x}_2^3 \\
&+ 3\left(x_1\tilde{x}_1 + x_1^2\tilde{x}_1^2x_2\tilde{x}_2 + x_2^2\tilde{x}_2^2 + x_2\tilde{x}_2 + x_1^2\tilde{x}_1^2 + x_1\tilde{x}_1x_2^2\tilde{x}_2^2\right) \\
&+ 6x_1\tilde{x}_1x_2\tilde{x}_2 \\
&= \left(1,\, \sqrt{3}x_1,\, \sqrt{3}x_2,\, \sqrt{3}x_1^2,\, \sqrt{6}x_1x_2,\, \sqrt{3}x_2^2,\, x_1^3,\, \sqrt{3}x_1^2x_2,\, \sqrt{3}x_1x_2^2,\, x_2^3\right) \\
&\cdot \left(1,\, \sqrt{3}\tilde{x}_1,\, \sqrt{3}\tilde{x}_2,\, \sqrt{3}\tilde{x}_1^2,\, \sqrt{6}\tilde{x}_1\tilde{x}_2,\, \sqrt{3}\tilde{x}_2^2,\, \tilde{x}_1^3,\, \sqrt{3}\tilde{x}_1^2\tilde{x}_2,\, \sqrt{3}\tilde{x}_1\tilde{x}_2^2,\, \tilde{x}_2^3\right) \\
&= \phi(x) \cdot \phi(\tilde{x}),
\end{aligned}$$

as desired.

Exercise A.23.47 (Feature map of the polynomial kernel)**.** Prove that in dimension d the feature map corresponding to the polynomial kernel of degree r,

$$K(x,\, \tilde{x}) = (1 + x \cdot \tilde{x})^r$$

where $x,\, \tilde{x} \in \mathbb{R}^d$, is

$$\phi(x) = \left(\sqrt{\binom{r}{r - |p|,\, p}}\, x^p : p \in \mathbb{N}^d \text{ and } |p| \leqslant r\right)$$

where p is a multi-index.

Solution. By the multinomial expansion theorem

$$\begin{aligned}
K(x,\, \tilde{x}) = \left(1 + \sum_{i=1}^{d} x_i\tilde{x}_i\right)^r &= \sum_{\substack{(l,p) \in \mathbb{N}^{1+d} \\ l + |p| = r}} \binom{r}{l,\, p} 1^l x^p \tilde{x}^p \\
&= \sum_{\substack{(l,p) \in \mathbb{N}^{1+d} \\ l + |p| = r}} \left(\sqrt{\binom{r}{l,\, p}}\, x^p\right)\left(\sqrt{\binom{r}{l,\, p}}\, \tilde{x}^p\right)
\end{aligned}$$

and so

$$\begin{aligned}
\phi(x) &= \left(\sqrt{\binom{r}{l,\, p}}\, x^p : (l,\, p) \in \mathbb{N}^{1+d} \text{ and } l + |p| = r\right) \\
&= \left(\sqrt{\binom{r}{r - |p|,\, p}}\, x^p : p \in \mathbb{N}^d \text{ and } |p| \leqslant r\right),
\end{aligned}$$

as desired.

Exercise A.23.48 (Feature map of the Gaussian kernel)**.** Prove that in dimension d the feature map corresponding to the Gaussian kernel with unit variance,

$$K(x,\, \tilde{x}) = \exp\left(-\frac{|x - \tilde{x}|}{2}\right)$$

where $x,\, \tilde{x} \in \mathbb{R}^d$, is $\phi : \mathbb{R}^d \to l^2\left(\mathbb{N}^d\right)$ defined by

$$\phi(x) = \left(\frac{x^p}{p!} e^{-\frac{|x|^2}{2}} : p \in \mathbb{N}^d \right)$$

where p is a multi-index.

Solution. Since

$$-\frac{1}{2}|x-\tilde{x}|^2 = -\frac{1}{2}|x|^2 + x\cdot\tilde{x} - \frac{1}{2}|\tilde{x}|^2$$

we have that

$$K\left(x,\, \tilde{x}\right) = e^{-\frac{|x|^2}{2}} e^{-\frac{|\tilde{x}|^2}{2}} e^{x\cdot\tilde{x}}.$$

In particular the series expansion of the exponential and the multinomial expansion theorem tell us that

$$\begin{aligned} e^{x\cdot\tilde{x}} = \sum_{j\geqslant 0} \frac{(x\cdot\tilde{x})^j}{j!} = \sum_{j\geqslant 0} \frac{1}{j!}\left(\sum_{i=1}^{d} x_i \tilde{x}_i\right)^j &= \sum_{j\geqslant 0} \frac{1}{j!} \sum_{|p|=j} \binom{j}{p} x^p \tilde{x}^p \\ &= \sum_{j\geqslant 0} \sum_{|p|=j} \frac{1}{p!} x^p \tilde{x}^p \\ &= \sum_{p\in\mathbb{N}^d} \left(\frac{1}{\sqrt{p!}} x^p\right)\left(\frac{1}{\sqrt{p!}} \tilde{x}^p\right). \end{aligned}$$

So finally

$$K\left(x,\, \tilde{x}\right) = \sum_{p\in\mathbb{N}^d} \left(\frac{1}{\sqrt{p!}} x^p e^{-|x|^2/2}\right)\left(\frac{1}{\sqrt{p!}} \tilde{x}^p e^{-|\tilde{x}|^2/2}\right) \qquad = \langle \phi(x),\, \phi(\tilde{x})\rangle_{l^2(\mathbb{N}^d)},$$

as desired.

Remark A.8 (Feature map of the Gaussian kernel)**.** In exactly the same way we may deduce that the non-unit variance Gaussian kernel may be written as

$$\begin{aligned} K_\gamma\left(x,\, \tilde{x}\right) &:= \exp\left(-\gamma|x-\tilde{x}|^2\right) \\ &= \sum_{p\in\mathbb{N}^d} \frac{(2\gamma)^{|p|}}{p!} \left(x^p e^{-\gamma|x|^2}\right)\left(\tilde{x}^p e^{-\gamma|\tilde{x}|^2}\right) \\ &= \langle \phi_\gamma(x),\, \phi_\gamma(\tilde{x})\rangle_{l^2} \end{aligned}$$

where the feature map is now

$$\phi_\gamma(x) = \sqrt{\frac{(2\gamma)^{|p|}}{p!}} x^p e^{-\gamma|x|^2}.$$

Exercise A.23.49 (A simple form of the Representer Theorem)**.** Let us consider $y_1,\, \dots,\, y_n \in \{0,\, 1\}$ and $x_1,\, \dots,\, x_n \in \mathbb{R}^d$ and define

$$n_j := \sum_{i=1}^{n} \mathbb{1}\left(y_i = j\right), \qquad \hat{\pi}_j := \frac{n_j}{n_0+n_1},$$

$$\bar{x}_j := \frac{1}{n_j}\sum_{i=1}^{n} x_i \mathbb{1}\left(y_i = j\right), \text{ and } \quad s_j := \frac{1}{n_j - 1}\sum_{i=1}^{n} \left(x_i - \bar{x}_j\right)\otimes\left(x_i - \bar{x}_j\right)\mathbb{1}\left(y_i = j\right)$$

for $j = 0, 1$. Suppose that $n_0, n_1 > 0$, and let us define, for every $v \in \mathbb{R}^d$,

$$R(v) := \frac{(\bar{x}_0 - \bar{x}_1) \otimes (\bar{x}_0 - \bar{x}_1)\, v \cdot v}{(\hat{\pi}_0 s_0 + \hat{\pi} s_1)\, v \cdot v}.$$

Prove that the minimum of R over $\mathbb{R}^d$ is a linear combination of $x_1, \ldots, x_n$.

Solution. Let $V_n := \operatorname{span}(x_1, \ldots, x_n)$. For any $v \in \mathbb{R}^d$ we may write

$$v = w + \xi$$

where $w \in V_n$ and $\xi \in V_n^\perp$. In particular this means that $x_i \cdot w = x_i \cdot w$ for every i. Moreover, since $\bar{x}_j \in V_n$,

$$\bar{x}_j \cdot v = \bar{x}_j \cdot w$$

while

$$\begin{aligned} s_j v \cdot v &= \frac{1}{n_j - 1} \sum_{i=1}^{n} \left[(x_i - \bar{x}_j) \cdot v\right]^2 \mathbb{1}\,(y_i = j) \\ &= \frac{1}{n_j - 1} \sum_{i=1}^{n} \left[(x_i - \bar{x}_j) \cdot w\right]^2 \mathbb{1}\,(y_i = j) = s_j w \cdot w. \end{aligned}$$

Therefore

$$R(v) = \frac{\left[(\bar{x}_0 - \bar{x}_1) \cdot v\right]^2}{\hat{\pi}_0\,(s_0 v \cdot v) + \hat{\pi}_1\,(s_1 v \cdot v)} = \frac{\left[(\bar{x}_0 - \bar{x}_1) \cdot w\right]^2}{\hat{\pi}_0\,(s_0 w \cdot w) + \hat{\pi}_1\,(s_1 w \cdot w)} = R(w),$$

which shows that

$$\min_{\mathbb{R}^d} R = \min_{V_n} R,$$

thus proving the claim.

Remark A.9 (Another form of the Represented Theorem)**.** The same reasoning, and hence the same result, as in Exercise A.23.49 holds if the numerator of R is replaced by

$$\hat{\pi}_0 \bar{x}_0 \otimes \bar{x}_0 + \hat{\pi}_1 \bar{x}_1 \otimes \bar{x}_1 - (\hat{\pi}_0 \bar{x}_0 + \hat{\pi}_1 \bar{x}_1) \otimes (\hat{\pi}_0 \bar{x}_0 + \hat{\pi}_1 \bar{x}_1)\,.$$

(Exercise A.23.36 tells us where this expression comes from: it is an estimator of the *between-class variance*.)

Exercise A.23.50 (Representer Theorem)**.** Let H be an inner product space, let $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a function, and let $r : [0, \infty) \to \mathbb{R}$ be non-decreasing. Consider $y_1, \ldots, y_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in H$ and define $f : H \to \mathbb{R}$ via, for every $\theta \in H$,

$$f(\theta) := \sum_{i=1}^{n} L\left(y_i, \langle \theta, x_i \rangle_H\right) + r\left(||\theta||_H\right).$$

Let $V_n := \operatorname{span}(x_1, \ldots, x_n)$. Then

$$\min_H f \geqslant \min_{V_n} f$$

and so if the minimizer of f over H exists then it is a linear combination of $x_1, \ldots, x_n$.

Remark A.10 (Representer Theorem)**.** There are no conditions imposed on L in Exercise A.23.50. Typically L is a loss function which is sufficiently nice to ensure that a minimizer of f *exist*, but a minimizer need not exist for the Representer Theorem to hold (possible vacuously).

Solution. Let us write, for any $\theta \in \mathbb{H}$,

$$\theta = \bar{\theta} + \xi \text{ where } \bar{\theta} \in V_n \text{ and } \xi \in V_n^{\perp}$$

where $V_n^{\perp}$ denotes the orthogonal complement of V_n in H. Then $||\theta||_H \geqslant ||\bar{\theta}||_H$ and so, since r is non-decreasing,

$$r\left(||\theta||_H\right) \geqslant r\left(||\bar{\theta}||_H\right).$$

Meanwhile, for every i, $x_i \in V$ and so $\langle \theta,\, x_i\rangle_H = \langle \bar{\theta},\, x_i\rangle_H$ such that

$$L\left(y_i,\, \langle \theta,\, x_i\rangle_H\right) = L\left(y_i,\, \langle \bar{\theta},\, x_i\rangle_H\right).$$

So finally

$$f(\theta) = \sum_{i=1}^{n} L\left(y_i,\, \langle \bar{\theta},\, x_i\rangle_H\right) + r\left(||\theta||_H\right) \geqslant \sum_{i=1}^{n} L\left(y_i,\, \langle \bar{\theta},\, x_i\rangle_H\right) + r\left(||\bar{\theta}||_H\right) = f(\bar{\theta}),$$

which proves the claim.

Appendix B. Facts from convex optimization

Much of this section is adapted from Chapter 5 of [BV09].

Definition B.1 (Lagrangian)**.** Let f, f_0, and $h : \mathbb{R}^d \to \mathbb{R}$ be functions and consider the optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } f(x) \leqslant 0 \text{ and } h(x) = 0.$$

The function $L : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined by

$$L\left(x,\, \lambda,\, \nu\right) := f_0(x) + \lambda f(x) + \nu h(x)$$

is called the *Lagrangian* associated with this optimization problem. The variables λ and ν are called *dual variables* or *Lagrange multipliers.*

Example B.2 (Lagrangian)**.** Consider the optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, \frac{1}{2}|x|^2 \text{ subject to } v \cdot x \leqslant c$$

for some $v \in \mathbb{R}^d$ and $c \in \mathbb{R}$. The associated Lagrangian is

$$L\left(x,\, \lambda\right) = \frac{1}{2}|x|^2 + \lambda\left(v \cdot x - c\right),$$

since $v \cdot x \leqslant c$ is equivalent to $v \cdot x - c \leqslant 0$.

Definition B.3 (Lagrange dual)**.** Let f_0, f and $h : \mathbb{R}^d \to \mathbb{R}$ be functions, consider the optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } f(x) \leqslant 0 \text{ and } h(x) = 0,$$

and let L be the associated Lagrangian. The function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \cup \{-\infty\}$ defined by

$$g\left(\lambda,\, \nu\right) := \inf_{x\in\mathbb{R}^d} L\left(x,\, \lambda,\, \nu\right)$$

is called the *Lagrange dual* associated with this optimization problem.

Example B.4 (Lagrange dual)**.** Consider the optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, \frac{1}{2}|x|^2 \text{ subject to } v \cdot x \leqslant c$$

for some $v \in \mathbb{R}^d$ and $c \in \mathbb{R}$. As observed in Example B.2 the associated Lagrangian is

$$L\left(x,\, \lambda\right) = \frac{1}{2}|x|^2 + \lambda\left(v \cdot x - c\right).$$

In order to derive the associated Lagrange dual we must minimize the Lagrangian with respect to x. Differentiating the Lagrangian with respect to x yields

$$\nabla_x L\left(x,\, \lambda\right) = x + \lambda v$$

and so the Lagrangian is minimized at

$$x_* = -\lambda v.$$

Inserting this expression into the Lagrangian tells us that the Lagrange dual is

$$g\left(\lambda\right) = L\left(x_*,\, \lambda\right) = \frac{\lambda^2}{2}|v|^2 - \lambda^2|v|^2 - \lambda c = -\frac{|v|^2}{2}\lambda^2 - c\lambda.$$

Definition B.5 (Primal and dual problem)**.** Let f_0, f and $h : \mathbb{R}^d \to \mathbb{R}$ be functions, consider the optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } f(x) \leqslant 0 \text{ and } h(x) = 0,$$

and let g be the associated Lagrange dual. The optimization problem

$$\underset{(\lambda,\,\nu)\in\mathbb{R}\times\mathbb{R}}{\text{maximize}}\, g\,(\lambda,\,\nu) \text{ subject to } \lambda \geqslant 0$$

is called the *dual problem* and the corresponding maximum value (possibly $\pm\infty$) is called the *dual optimal value.* The original optimization problem is called the *primal problem* and the corresponding minimum value (possible $\pm\infty$ as well) is called the *primal optimal value.*

Example B.6 (Primal and dual problem)**.** Consider the primal problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}|x|^2 \text{ subject to } v \cdot x \leqslant c$$

for some $v \in \mathbb{R}^d$ and $c \in \mathbb{R}$. Using the formula for its Lagrange dual recorded in Example B.4 we deduce that the dual problem is

$$\max_{\lambda\in\mathbb{R}} -\frac{|v|^2}{2}\lambda^2 - c\lambda \text{ subject to } \lambda \geqslant 0.$$

Lemma B.7 (Weak duality)**.** *Let f_0, f and $h : \mathbb{R}^d \to \mathbb{R}$ be functions, consider the optimization problem*

$$\underset{x\in\mathbb{R}^d}{\textit{minimize}}\, f_0(x) \textit{ subject to } f(x) \leqslant 0 \textit{ and } h(x) = 0,$$

and let p_* and d_* denote the primal and dual optimale values, respectively. Then

$$d_* \leqslant p_*.$$

This inequality is known as weak duality*.*

Proof. This follows from the choice that is made in the dual problem to restrict our attention to *non-negative* Lagrange multipliers λ corresponding to the *inequality* constraint. Indeed: for any *feasible* $x \in \mathbb{R}^d$, meaning that $f(x) \leqslant 0$ and $h(x) = 0$, the associated Lagrangian satisfies, when $\lambda \geqslant 0$,

$$L\,(x,\,\lambda,\,\nu) = f_0(x) + \underbrace{\lambda f(x)}_{\leqslant 0} + \underbrace{\nu h(x)}_{=0} \leqslant f_0(x)$$

and so the Lagrange dual g satisfies

$$g\,(\lambda,\,\nu) \leqslant f_0(x).$$

The claim then follows *if* such a feasible point exists. Otherwise, if there are no feasible points, then

$$\inf_{x\in\mathbb{R}^d} f_0(x) \text{ subject to } f(x) \leqslant 0 \text{ and } h(x) = 0 = \inf\emptyset = +\infty$$

and so the claim trivially holds. □

Definition B.8 (Strong duality)**.** Let f_0, f and $h : \mathbb{R}^d \to \mathbb{R}$ be functions, consider the optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } f(x) \leqslant 0 \text{ and } h(x) = 0,$$

and let p_* and d_* denote the primal and dual optimale values, respectively. If

$$d_* = p_*$$

then we say that *strong duality* holds for this optimization problem.

Definition B.9 (Strict feasibility)**.** Let f_0, f and $h : \mathbb{R}^d \to \mathbb{R}$ be functions and consider the optimization problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } f(x) \leqslant 0 \text{ and } h(x) = 0.$$

If there exists $x \in \mathbb{R}^d$ such that

$$f(x) < 0 \text{ and } h(x) = 0,$$

i.e. the inequality constraint holds in a *strict* sense, then we say that x is *strictly feasible* for this optimization problem.

Definition B.10 (Convex function)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex if, for every x, $y \in \mathbb{R}^d$ and every $\theta \in [0, 1]$,

$$f\left((1-\theta)\, x + \theta y\right) \leqslant (1-\theta)\, f(x) + \theta f(y).$$

Theorem B.11 (Slater's condition)**.** *Let f_0 and $f : \mathbb{R}^d \to \mathbb{R}$ be convex, let $v \in \mathbb{R}^d$, let $c \in \mathbb{R}$, and consider the optimization problem*

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } f(x) \leqslant 0 \textit{ and } v \cdot x = c$$

(i.e. we impose that the equality *constraint be affine). If this optimization problem admits a strictly feasible point $x \in \mathbb{R}^d$ then strong duality holds.*

Proof. See Section 5.3.2 of [BV09]. □

Lemma B.12 (Complementary slackness)**.** *Let f_0, f and $h : \mathbb{R}^d \to \mathbb{R}$ be functions and consider the optimization problem*

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } f(x) \leqslant 0 \textit{ and } h(x) = 0.$$

Suppose that this minimum is attained and that strong duality holds. Then, for any minimizer x_* of the primal problem and any maximizer (λ_*, ν_*) of the dual problem,

(1) x_* is also a minimizer of $L(\,\cdot\,, \lambda_*, \nu_*)$, where L denotes the Lagrangian, and
(2) $\lambda_* f(x_*) = 0$.

The latter identity is known as complementary slackness*.*

Remark B.13 (Complementary slackness)**.** Since $f(x) \leqslant 0$ and $\lambda \geqslant 0$ in general, complementary slackness

$$\lambda_* f(x_*) = 0$$

may be phrased equivalently as

$$\lambda_* > 0 \Rightarrow f(x_*) = 0$$

or

$$f(x_*) > 0 \Rightarrow \lambda_* = 0.$$

Intuitively this means that the Lagrange multiplier λ vanishes *unless* the inequality constraint is *active*, or *tight*, which means that $f = 0$.

Example B.14 (Complementary slackness)**.** Consider the optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\,\frac{1}{2}|x|^2 \text{ subject to } v\cdot x \leqslant c$$

for some $v\in\mathbb{R}^d$ and $c\in\mathbb{R}$. The complementary slackness condition reads

$$\lambda_*\left(v\cdot x_* - c\right) = 0$$

and so either

- $v\cdot x_* = 0$, which means that the minimizer x_* is on the boundary of the constraint set (i.e. the inequality constraint is *active*), or
- $\lambda_* = 0$, i.e. the optimal Lagrange multiplier vanishes, of
- *both* occur simultaneously.

Corollary B.15 (Stationarity)**.** *Let f_0, f and $h:\mathbb{R}^d\to\mathbb{R}$ be differentiable functions and consider the optimization problem*

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\,f_0(x) \text{ subject to } f(x)\leqslant 0 \textit{ and } h(x)=0.$$

Suppose that this minimum is attained and that strong duality holds. Then, for any minimizer x_* of the primal problem and any maximizer (λ_*, ν_*) of the dual problem,

$$\nabla f_0(x_*) + \lambda_*\nabla f(x_*) + \nu_*\nabla h(x_*) = 0.$$

This identity is known as stationarity*.*

Proof. Since f_0, f, and h are differentiable, so is the Lagrangian L. In particular Lemma B.12 tells us that x_* minimizes $L\left(\,\cdot\,, \lambda_*, \nu_*\right)$, and hence it must be a critical point such that

$$0 = \nabla_x L\left(x_*, \lambda_*, \nu_*\right) = \nabla f_0(x_*) + \lambda_*\nabla f(x_*) + \nu_*\nabla h(x_*),$$

as desired. □

Example B.16 (Stationarity)**.** Consider the optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\,\frac{1}{2}|x|^2 \text{ subject to } v\cdot x \leqslant c$$

for some $v\in\mathbb{R}^d$ and $c\in\mathbb{R}$. The stationarity condition reads

$$x_* + \lambda_* v = 0.$$

Proposition B.17 (Necessity of the KKT conditions)**.** *Let f_0, f and $h:\mathbb{R}^d\to\mathbb{R}$ be differentiable functions and consider the optimization problem*

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\,f_0(x) \text{ subject to } f(x)\leqslant 0 \textit{ and } h(x)=0.$$

Suppose that this minimum is attained and that strong duality holds. Then, for any minimizer x_* of the primal problem and any maximizer (λ_*, ν_*) of the dual problem, the following conditions hold.

(1) ***Primal feasibility.***

$$f(x_*)\leqslant 0 \textit{ and } h(x_*) = 0.$$

(2) ***Dual feasibility.***

$$\lambda_*\geqslant 0.$$

(3) ***Complementary slackness.***

$$\lambda_* f(x_*) = 0.$$

(4) ***Stationarity.***

$$\nabla f_0(x_*) + \lambda_* \nabla f(x_*) + \nu_* \nabla h(x_*) = 0.$$

The conditions (1)–(4) are known as the Karush-Kuhn-Tucker conditions, *or* KKT conditions.

Proof. Conditions (1) and (2) follow immediately from the definition of the primal and dual problems, respectively. Condition (3) follows from Lemma B.12. Condition (4) follows from Corollary B.15. □

Theorem B.18 (Sufficiency of the KKT conditions for convex problems)**.** *Let f_0 and $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable convex functions, let $h : \mathbb{R}^d \to \mathbb{R}$ be affine, and consider the optimization problem*

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } f(x) \leqslant 0 \textit{ and } h(x) = 0.$$

If any $(x_*, \lambda_*, \nu_*) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}$ satisfies the KKT conditions (1)–(4) of Proposition B.17 then x_* is a minimizer of the primal problem, (λ_*, ν_*) is a maximizer of the dual problem, and strong duality holds.

Proof. Suppose that $(x_*, \lambda_*, \nu_*) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}$ satisfies the KKT conditions. By condition (2), the dual feasibility, we know that $\lambda_* \geqslant 0$ and so $L(\,\cdot\,, \lambda_*, \nu_*)$ is convex (since f_0 and f are convex and h is affine), where L denotes the Lagrangian. But by assumption the Lagrangian is differentiable (since affine functions are differentiable) while condition (4), stationarity, tells us that

$$\nabla_x L(x_*, \lambda_*, \nu_*) = 0,$$

so we deduce that x_* is a minimizer of $L(\,\cdot\,, \lambda_*, \nu_*)$. Therefore conditions (1) and (3), primal feasibility and complementary slackness, tell us that, for the Lagrange dual g,

$$g(\lambda_*, \nu_*) = L(x_*, \lambda_*, \nu_*) = f_0(x_*) + \underbrace{\lambda_* f(x_*)}_{=0} + \nu_* \underbrace{h(x_*)}_{=0} = f_0(x_*).$$

This shows that x_* is a minimizer of the primal problem (using primal feasibility once more to deduce that x_* is feasible for the primal problem), that (λ_*, ν_*) is a maximizer of the dual problem, and that strong duality holds, as desired. □

Corollary B.19 (KKT)**.** *Let f_0 and $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable convex functions, let $h : \mathbb{R}^d \to \mathbb{R}$ be affine, consider the optimization problem*

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } f(x) \leqslant 0 \textit{ and } h(x) = 0,$$

and suppose that this problem admits a strictly feasible point. Then the KKT conditions (1)–(4) of Proposition B.17 hold at $(x_*, \lambda_*, \nu_*) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}$ if and only if x_* is a minimizer of the primal problem and (λ_*, ν_*) is a maximizer of the dual problem.

Proof. The "if" direction follows from combining Theorem B.11, which guarantees strong duality, with Proposition B.17. The "only if" direction is precisely Theorem B.18. □

Example B.20 (KKT)**.** Consider the primal problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}|x|^2 \text{ subject to } v\cdot x \leqslant c$$

for some $v \in \mathbb{R}^d$ and $c \in \mathbb{R}$ whose dual problem is observed in Example B.6 to be

$$\max_{\lambda\in\mathbb{R}} -\frac{|v|^2}{2}\lambda^2 - c\lambda \text{ subject to } \lambda \geqslant 0.$$

Suppose now that v is non-zero. Then the point $x = (c-1)\frac{v}{|v|^2}$ is strictly feasible for the primal problem since

$$v\cdot x = (c-1)\frac{|v|^2}{||^2} = c - 1 < c.$$

Therefore Corollary B.19 tells us that the KKT conditions

(1) $v \cdot x_* \leqslant c$ (primal feasibility),
(2) $\lambda_* \geqslant 0$ (dual feasibility),
(3) $\lambda_*\,(v\cdot x_* - c) = 0$ (complementary slackness – see also Example B.14) and
(4) $x_* + \lambda_* v = 0$ (stationarity – see also Example B.16)

characterize minimizers of x_* of the primal problem and maximizers λ_* of the dual problem.

Definition B.21 (Quadratic functions)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ given by

$$f(x) = Ax\cdot x + v\cdot x + c$$

for some d-by-d matrix A, $v \in \mathbb{R}^d$, and $c \in \mathbb{R}$ is called a *quadratic function.*

In particular note that A and v may vanish. This means that affine functions and constant functions are (degenerate) quadratic functions.

Corollary B.22 (Wolfe dual)**.** *Let f_0 and $f : \mathbb{R}^d \to \mathbb{R}$ be convex quadratic functions and consider that primal problem*

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } f(x) \leqslant 0.$$

Suppose that this problem admits a strictly feasible point $x \in \mathbb{R}^d$ where $f(x) < 0$. For any $x_* \in \mathbb{R}^d$, x_* is a minimizer of this problem if and only if it is a maximizer of the problem

$$\underset{(x,\,\lambda)\in\mathbb{R}^d\times\mathbb{R}}{\text{maximize}}\ f_0(x) + \lambda f(x) \text{ subject to } \lambda \geqslant 0 \textit{ and } \nabla f_0(x) + \lambda\nabla f(x) = 0.$$

This latter problem is called the Wolfe dual *of the original problem.*

Moreover the optimal λ_* from the Wolfe dual is exactly the same as the optimal Lagrange multiplier of the (standard) dual problem.

Proof. Since f_0 and f are quadratic they must be differentiable and their gradients are affine. Rewriting the Wolfe dual as

$$\underset{(x,\,\lambda)\in\mathbb{R}^d\times\mathbb{R}}{\text{minimize}}\ -f_0(x) - \lambda f(x) \text{ subject to } -\lambda \leqslant 0 \text{ and } \nabla f_0(x) + \lambda\nabla f(x) = 0,$$

and noting that, trivially, $\lambda = 1$ is strictly feasible, we observe that both the original optimization and its Wolfe dual satisfy the assumptions of Corollary B.19.

To show that these two problems are equivalent it then suffices to show that their KKT conditions are equivalent. Using the formulation of the Wolfe dual above, where it is viewed as a minimization problem, we see that the Lagrangian is

$$-f_0(x) - \lambda f(x) + \mu(-\lambda)$$

for a Lagrange multiplier $\mu \in \mathbb{R}$. The KKT conditions are therefore

$$\begin{cases} -\lambda \leqslant 0 \text{ and } \nabla f_0(x) + \lambda \nabla f(x) = 0, & \text{(primal feasibility)} \\ \mu \geqslant 0, & \text{(dual feasibility)} \\ \mu\lambda = 0, & \text{(complementary slackness), and} \\ -\nabla f_0(x) - \lambda \nabla f(x) = 0 \text{ and } -f(x) - \mu = 0 & \text{(stationarity).} \end{cases}$$

In particular the stationarity condition tells us that $\mu = -f(x)$ and so these conditions may equivalently be written as

$$\begin{cases} f(x) \leqslant 0, \\ \lambda \geqslant 0, \\ \lambda f(x) = 0, \text{ and} \\ \nabla f_0(x) + \lambda \nabla f(x) = 0. \end{cases}$$

These are precisely the KKT conditions for the original problem, thus verifying that the two problems are equivalent.

In particular this also verifies that the optimal λ_* from the Wolfe dual agrees with the optimal Lagrange multiplier of the (standard) dual problem. □

Remark B.23 (Wolfe dual)**.** In the Wolfe dual of Corollary B.22, the stationarity constraint

$$\nabla f_0(x) + \lambda \nabla f(x) = 0$$

is a *linear* system of equations. In practice, if f_0 and f are sufficiently nice then that system may be used to solve for x as a function of λ, i.e. writing

$$x = x(\lambda).$$

Inserting this explicit parametrization into the objective function of the Wolfe dual produces

$$\tilde{f}(\lambda) := f_0\left(x(\lambda)\right) + \lambda f\left(x(\lambda)\right)$$

and so the Wolfe dual problem becomes

$$\underset{\lambda \in \mathbb{R}}{\text{maximize}}\, \tilde{f}(\lambda) \text{ subject to } \lambda \geqslant 0,$$

which we refer to as the *parametrized Wolfe dual.* In other words we have traded the constraint $f(x) \leqslant 0$ for the *much* simpler constraint $\lambda \geqslant 0$, at the cost of a (possibly) messier objective function.

Example B.24 (Wolfe dual)**.** Consider the optimization problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{2}|x|^2 \text{ subject to } v \cdot x \leqslant c$$

for some *non-zero* $v \in \mathbb{R}^d$ and $c \in \mathbb{R}$. Since $x \mapsto \frac{1}{2}|x|^2$ is quadratic and convex while $x \mapsto v \cdot x - c$ is affine, Corollary B.22 tells us that the problem above is equivalent to its Wolfe dual, namely (using the stationarity condition recorded in Example B.16)

$$\max_{(x,\,\lambda)\in\mathbb{R}^d\times\mathbb{R}} \frac{1}{2}|x|^2 + \lambda\,(v\cdot x - c) \text{ subject to } \lambda \geqslant 0 \text{ and } x + \lambda v = 0.$$

Since the stationarity condition $x + \lambda v$ is particularly nice we may proceed as discussed in Remark B.23 and use it to solve for x as a function of λ, obtaining

$$x = -\lambda v.$$

Inserting this identity into the objective function of the Wolfe dual shows that the Wolfe dual may equivalently be written as

$$\begin{aligned}&\max_{\lambda\in\mathbb{R}} \frac{\lambda^2}{2}|v|^2 - \lambda^2|v|^2 - c\lambda \text{ subject to } \lambda \geqslant 0\\ &= \max_{\lambda\in\mathbb{R}} -\frac{|v|^2}{2}\lambda^2 - c\lambda\\ &= \max_{\lambda\geqslant 0} -\lambda\left(\frac{|v|^2}{2}\lambda + c\right).\end{aligned}$$

Since the quadratic $-\lambda\left(\frac{|v|^2}{2}\lambda + c\right)$ has roots at

$$\lambda = 0 \text{ and } \lambda = \frac{-2c}{|v|^2},$$

its maximum is attained at

$$\lambda = \frac{1}{2}\left(0 - \frac{2c}{|v|^2}\right) = \frac{-c}{|v|^2}.$$

In light of the dual constraint $\lambda \geqslant 0$ this means that the maximizer of the Wolfe dual is

$$\lambda_* = \max\left(0,\, \frac{-c}{|v|^2}\right) = -\frac{\min(0,\, c)}{|v|^2},$$

which we may also write as

$$\lambda_* = \frac{c_-}{|v|^2}$$

for $s_- := -\min(0,\, s) > 0$ denoting, for any $s \in \mathbb{R}$, the *negative part* of s.

So finally, using the stationarity condition one last time, we conclude that the minimizer x_* of the original problem is

$$x_* = -\lambda_* v = -c_-\frac{v}{|v|^2} = -\frac{c_-}{|v|}\hat{v}$$

for $\hat{v} := \frac{v}{|v|}$ denoting the unit vector in the direction of v.

In particular we can verify directly that strong duality holds since the primal optimal value is

$$p_* = \frac{1}{2}|x_*|^2 = \frac{c_-^2}{2|v|^2}$$

while the dual optimal value is, using the form of the Lagrange dual recorded in Example B.4,

$$d_* = -\frac{c_-^2}{2|v|^2} - \frac{cc_-}{|v|^2} = -\frac{c_-^2}{2|v|^2} + \frac{c_-^2}{|v|^2} = \frac{c_-^2}{2|v|^2} = p_*.$$

Remark B.25 (The economics interpretation of duality)**.** Suppose that

- $x \in \mathbb{R}^d$ denotes how a firm operates,
- $f_0(x)$ is the cost of operating according to x, and
- $f(x) \leqslant 0$ represents some limit on resources or regulatory limit.

The optimal operating condition x_* which maximizes profit while respecting these limits is the minimizer of the problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } f(x) \leqslant 0.$$

Now suppose that a regulatory body sets a price λ such that violations of the limit incur an additional cost for the firm proportional to λ and to the size of the violation. In other words

- if $f(x) > 0$, i.e. the firm violates the limit, then the firm *pays* $\lambda f(x)$ and
- if $f(x) \leqslant 0$, i.e. the firm respects the limit, then the firm *receives* $\lambda f(x)$ by selling its "unused" limit to another firm.

Then the total operating cost to the firm is

$$f_0(x) + \lambda f(x),$$

which is precisely the Lagrangian $L(x, \lambda)$. Because the firm minimizes cost, it operates at a cost of

$$g(\lambda) = \min_{x \in \mathbb{R}^d} L(x, \lambda)$$

where g is the Lagrange dual. The Lagrange dual therefore represents the optimal cost to the firm $g(\lambda)$ given the price λ set by the regulatory body. In particular the dual optimal value d_* is the optimal cost to the firm under the *least favourable* price λ_*.

The duality gap $p_* - d_*$, where p_* denotes the primal optimal value, is therefore the cost saving that the firm is *guaranteed* to make when paying for violations of the limit, as opposed to respecting the limit, *irrespectively* of the price λ.

In particular strong duality holds, i.e. there is no gap, precisely when some optimal price λ_* ensures that the firm cannot reduce its cost by being allowed to pay for limit violations.

Definition B.26 (Vector inequalities)**.** Let $v \in \mathbb{R}^d$.

- We write $v \preccurlyeq 0$ if and only if $v_i \leqslant 0$ for all $1 \leqslant i \leqslant d$.
- We write $v \prec 0$ if and only if $v_i < 0$ for all $1 \leqslant i \leqslant d$.

Remark B.27 (Multiple constraints)**.** Let f_0, f_i and $h_j : \mathbb{R}^d \to \mathbb{R}$ be functions for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant m$ and consider the optimization problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } f_i(x) \leqslant 0 \text{ and } h_j(x) = 0$$
$$\text{for every } 1 \leqslant i \leqslant m \text{ and } 1 \leqslant j \leqslant p.$$

This is a problem with *multiple constraints.* If we define the functions $F : \mathbb{R}^d \to \mathbb{R}^m$ and $H : \mathbb{R}^d \to \mathbb{R}^p$ via

$$F_i := f_i \text{ and } H_j := h_j$$

then we may write this problem equivalently as

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f_0(x) \text{ subject to } F(x) \preccurlyeq 0 \text{ and } H(x) = 0.$$

The *Lagrangian* is now a function $L : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ defined by

$$\begin{aligned} L\left(x,\, \lambda,\, \nu\right) &:= f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) = \sum_{j=1}^{p} \nu_j h_j(x) \\ &= f_0(x) + \lambda \cdot F(x) + \nu \cdot H(x) \end{aligned}$$

and the *Lagrangian dual* $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R} \cup \{-\infty\}$ is defined by

$$g\left(\lambda,\, \nu\right) := \inf_{x\in\mathbb{R}^d} L\left(x,\, \lambda,\, \nu\right).$$

Then, for p_* denoting the optimal value of the *primal problem*

$$\inf_{x\in\mathbb{R}^d} f_0(x) \text{ subject to } F(x) \preccurlyeq 0 \text{ and } H(x) = 0$$

and for d_* denoting the optimal value of the *dual problem*

$$\sup_{(\lambda,\, \nu)\in\mathbb{R}^m\times\mathbb{R}^p} g\left(\lambda,\, \nu\right) \text{ subject to } \lambda \succcurlyeq 0,$$

weak duality holds in the sense that

$$d_* \leqslant p_*$$

and we say that *strong duality* holds if $d_* = p_*$.

In particular suppose that f_0 and f_i are differentiable and convex for every $1 \leqslant i \leqslant m$, that H is affine, and that the primal problem admits a *strictly feasible* point $x \in \mathbb{R}^d$ where

$$F(x) \prec 0 \text{ and } H(x) = 0.$$

Then strong duality holds. Moreover x_* is a minimizer of the primal problem and $(\lambda_*,\, \nu_*)$ is a minimizer of the dual problem if and only if the following *KKT conditions* are satisfied.

(1) **Primal feasibility.**

$$F(x_*) \preccurlyeq 0 \text{ and } H(x_*) = 0.$$

(2) **Dual feasibility.**

$$\lambda_* \succcurlyeq 0.$$

(3) **Complementary slackness.**

$$\lambda_{*,i} f_i(x_*) = 0 \text{ for every } 1 \leqslant i \leqslant m.$$

(4) **Stationarity.**

$$\nabla f_0(x_*) + \lambda_* \cdot DF(x_*) + \nu_* \cdot DH(x_*) = 0.$$

If we assume furthermore that f_0 and every f_i are quadratic, that $H = 0$ (i.e. there is no equality constraint), and that the primal problem is strictly feasible, then the primal problem is equivalent to the *Wolfe dual*

$$\underset{(x,\, \lambda)\in\mathbb{R}^d\times\mathbb{R}^m}{\text{maximize}}\, f_0(x) + \lambda \cdot F(x) \text{ subject to } \lambda \succcurlyeq 0 \text{ and } \nabla f_0(x) + \lambda \cdot DF(x) = 0.$$

(Note that a Wolfe dual may still be written if equality constraints are present, they simply have to be kept as additional constraints for the Wolfe problem.) In particular the optimal λ_* from the Wolfe dual agrees with the optimal Lagrange multiplier of the standard dual problem.

Finally, if as discussed in Remark B.23 the stationarity condition

$$\nabla f_0(x) + \lambda \cdot DF(x) = 0$$

is nice enough to allow us to solve for $x = x(\lambda)$ then, for

$$\tilde{f}(\lambda) := f_0\left(x(\lambda)\right) + \lambda \cdot F\left(x(\lambda)\right)$$

the Wolfe dual becomes

$$\underset{\lambda \in \mathbb{R}^m}{\text{maximize}}\, \tilde{f}(\lambda) \text{ subject to } \lambda \succcurlyeq 0.$$

Remark B.28 (Maximization and concave constraints)**.** Since $\max f_0 = -\min\left(-f_0\right)$ and since $-f$ is convex whenever f is concave (by *definition* of concavity) it follows that everything discussed in this section also applies to *maximization* problems and to concave constraints $f \geqslant 0$ for f concave.

The only, but crucial modification, has to do with the *sign* of the Lagrange multipliers λ accounting for the inequality constraints.

Appendix C. Facts from linear algebra

Definition C.1 (Gram matrix)**.** Let $\mathbb{X}$ be an n-by-k matrix matrix. The matrix $\mathbb{X}^T\mathbb{X}$ is called the *Gram matrix* associated with $\mathbb{X}$.

Lemma C.2 (Kernel and Gram matrix)**.** *Let $\mathbb{X}$ be an n-by-k matrix matrix. The kernel of $\mathbb{X}$ is equal to the kernel of its Gram matrix, i.e.*

$$\ker \mathbb{X} = \ker \mathbb{X}^T\mathbb{X}.$$

Proof. Clearly $\ker \mathbb{X} \subseteq \ker \mathbb{X}^T\mathbb{X}$ so it suffices to show that the reverse inclusion holds. So let v be a non-zero vector in the kernel of the Gram matrix $\mathbb{X}^T\mathbb{X}$. Then, taking the dot product with v, we see that

$$0 = \mathbb{X}^T\mathbb{X}v \cdot v = |\mathbb{X}v|^2$$

and so it follows that $\mathbb{X}v = 0$. In other words v also belongs to the kernel of $\mathbb{X}$, which concludes the proof. □

Definition C.3 (Symmetric matrix)**.** A matrix A is called *symmetric* if $A^T = A$.

Lemma C.4 (Symmetric matrices and orthogonal decompositions)**.** *If a matrix A is symmetric then $\operatorname{im} A \perp \ker A$.*

Proof. Let A be a symmetric matrix. For any two vectors v and w where $w \in \ker A$ we have that

$$Av \cdot w = v \cdot Aw = 0,$$

which proves that $\operatorname{im} A \perp \ker A$. □

Remark C.5. The converse of Lemma C.4 is false since

$$A = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

which acts as a $\frac{\pi}{2}$–rotation in the (x, y)–plane while killing off the z–axis, is anti-symmetric (i.e. $A^T = -A$) and yet

$$\operatorname{im} A = \text{ } (x, y)\text{–plane } \perp \text{ } z\text{–axis } = \ker A.$$

Definition C.6 (Projection matrices)**.** Let P be a matrix.

- P is called a *projection matrix* if $P^2 = P$.
- P is called an *orthogonal projection matrix* if it is a projection matrix for which $\operatorname{im} P \perp \ker P$.

Lemma C.7 (Characterization of orthogonal projection matrices)**.** *Let P be a projection matrix. P is an orthogonal projection matrix if and only if P is symmetric.*

Proof. This "if" direction follows immediately from Lemma C.4 so we only need to prove the "only if" direction. To that effect, suppose that P is an orthogonal decomposition matrix. Since P must be square the rank-nullity theorem tells us that for every vector v there exist $v_0 \in \ker P$ and $\bar{v} \in \operatorname{im} P$, uniquely determined by v, such that $v = v_0 + \bar{v}$. Therefore, for any two vectors v and w, since $\operatorname{im} P \perp \ker P$ and since P is a projection matrix we deduce that $Pv = P\bar{v} = \bar{v}$ (and similarly for w) such that

$$Pv \cdot w = \bar{v} \cdot (w_0 + \bar{w}) = \bar{v} \cdot \bar{w} = (v_0 + \bar{v}) \cdot \bar{w} = v \cdot Pw,$$

which proves that P is symmetric. □

Lemma C.8 (Full rank, invertibility of the Gram matrix, and projection onto the image)**.** *Let $\mathbb{X}$ be an n-by-k matrix with $k < n$. The Gram matrix $\mathbb{X}^T\mathbb{X}$ is invertible if and only if $\mathbb{X}$ has full rank, i.e.* $\operatorname{rank}\mathbb{X} = k$*. Moreover, in either (and hence both) case the matrix $A := \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ is an orthogonal projection matrix onto the image of $\mathbb{X}$, with $\operatorname{tr} A = k$.*

Proof. Suppose that the Gram matrix $\mathbb{X}^T\mathbb{X}$ is *not* invertible. This means that some non-zero vector v belongs to the kernel of $\mathbb{X}^T\mathbb{X}$. By Lemma C.2 it follows that v also belongs to the kernel of $\mathbb{X}$. Since v is a non-zero vector this means that some of the columns of $\mathbb{X}$ are linearly dependent and so $\mathbb{X}$ does *not* have full rank, i.e. $\operatorname{rank}\mathbb{X} < k$.

Now suppose conversely that $\mathbb{X}^T\mathbb{X}$ is invertible. Note that $A^2 = A$ since

$$A^2 = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T = A$$

and so any eigenvalue λ of A must satisfy $\lambda^2 = \lambda$. This means that the only possible eigenvalues of A are 0 and 1. Note also that, since the trace is cyclic,

$$\operatorname{tr} A = \operatorname{tr}\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \operatorname{tr} I_k = k.$$

Since the trace is the sum of the eigenvalues weighted by their algebraic multiplicity we deduce that the eigenvalue 1 has algebraic multiplicity k while the eigenvalue 0 has multiplicity $n - k$. In particular, since $\mathbb{X}^T\mathbb{X}$ is symmetric it follows that

$$A^T = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-T}\mathbb{X}^T = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T = A,$$

i.e. A itself is symmetric. Therefore A is diagonalizable, which means that the *geometric* multiplicity of the eigenvalue 1 is k while the *geometric* multiplicity of the eigenvalue 0 is $n - k$. In particular: $\operatorname{rank} A = k$.

But $\operatorname{im} A \subseteq \operatorname{im}\mathbb{X}$ where $\mathbb{X}$ is a n-by-k matrix with $k < n$, which means that, in general, $\operatorname{rank}\mathbb{X} \leqslant k$. But here $k = \operatorname{rank} A \leqslant \operatorname{rank}\mathbb{X}$! This means that $\operatorname{im} A = \operatorname{im}\mathbb{X}$, i.e. indeed $\mathbb{X}$ has full rank and A is the projection matrix onto the image. Finally note that, as per Lemma C.7, since A is a symmetric projection matrix it must be an orthogonal projection matrix. □

Definition C.9 (Generalized eigenvalue problem)**.** Let A and B be d-by-d matrices, let $v \in \mathbb{R}^d$ and let $\lambda \in \mathbb{R}$. If

$$Av = \lambda Bv$$

then we say that v is a *B-eigenvector* of A with *B-eigenvalue* λ.

Theorem C.10 (Characterization of generalized Rayleigh quotients)**.** *Let A and Σ be symmetric and positive-definite d-by-d matrices and let $w_* \in \mathbb{R}$. The following are equivalent.*

(1) w_* maximizes the generalized Rayleigh quotient

$$\frac{Aw \cdot w}{\Sigma w \cdot w}$$

over $w \in \mathbb{R}^d$.

(2) w_* is a Σ-eigenvector of A corresponding to its largest Σ-eigenvalue.

(3) $w_* = \Sigma^{-1/2} z_*$ where z_* maximizes the Rayleigh quotient

$$\frac{\left(\Sigma^{-1/2} A \Sigma^{-1/2}\right) z \cdot z}{|z|^2}$$

over $z \in \mathbb{R}^d$.

(4) $w_* = \Sigma^{-1/2} z_*$ *where* z_* *is an eigenvector of* $\Sigma^{-1/2} A \Sigma^{-1/2}$ *corresponding to its largest eigenvalue.*

Proof. The equivalence of items 1 and 3 and of items 2 and 4 follows immediately from the change of variable $w = \Sigma^{-1/2} z$. The equivalence of items 3 and 4 is a classical result. It can for example be proved by reformulating the maximization of the Rayleigh quotient $\frac{Bz \cdot z}{|z|^2}$ as the *constrained* maximization of $Bz \cdot z$ subject to $|z| = 1$ and then using the method of Lagrange multipliers. □

Corollary C.11 (A generalized Rayleigh quotient)**.** *Let* Σ *be a symmetric and positive-definite d-by-d matrix. The maximizer of*

$$\frac{(v \otimes v) w \cdot w}{\Sigma w \cdot w} \text{ over } w \in \mathbb{R}^d$$

is

$$w_* = \Sigma^{-1} v$$

(or any scalar multiple thereof).

Proof. In order to use Theorem C.10 we first compute that

$$\Sigma^{-1/2} (v \otimes v) \Sigma^{-1/2} = \left(\Sigma^{-1/2} v \right) \otimes \left(\Sigma^{-1/2} v \right).$$

This is a rank-one symmetric matrix and so its only (non-zero) eigenvectors are proportional to

$$z_* = \Sigma^{-1/2} v.$$

Theorem C.10 then tells us that

$$w_* = \Sigma^{-1/2} z_* = \Sigma^{-1} v.$$

□

Theorem C.12 (The trace is a linear version of dimension-counting)**.** *Suppose that* $F : \mathbb{R}^{n \times n} \to \mathbb{R}$ *satisfies the following properties.*

(1) $F(P) = \dim \operatorname{im} P$ *for any projection matrix* P.

(2) F *is linear.*

Then F *is the trace.*

Remark C.13. Intuitively we may understand Theorem C.12 as telling us that the trace is a linear version of the dimension-counting operator.

Proof. Let E_{ij} denote the n-by-n matrix whose only nonzero entry is a one at the (i, j)–position. Since every matrix is a linear combination of E_{ij}'s it is enough to show that $F(E_{ij}) = \operatorname{tr} E_{ij}$ for every i and j.

If $i = j$ then E_{ii} is a projection matrix and so

$$F(E_{ii}) = \dim \operatorname{im} E_{ii} = 1 = \operatorname{tr} E_{ii}.$$

If $i \neq j$ we only consider the case $(i, j) = (1, 2)$ in detail. We may write

$$E_{12} = \left(\begin{array}{c|c} \begin{matrix} 0 & 1 \\ 0 & 0 \end{matrix} & 0 \\ \hline 0 & 0 \end{array}\right) = \left(\begin{array}{c|c} \begin{matrix} 0 & 1 \\ 0 & 1 \end{matrix} & 0 \\ \hline 0 & 0 \end{array}\right) - \left(\begin{array}{c|c} \begin{matrix} 0 & 0 \\ 0 & 1 \end{matrix} & 0 \\ \hline 0 & 0 \end{array}\right) =: A - B.$$

Note that A and B are both projection matrices of rank 1 since

$$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}^2 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

and this matrix projects onto

$$\left\{(x,\, y) \in \mathbb{R}^2 : x = y\right\}.$$

Therefore

$$F(E_{12}) = \operatorname{rank} A - \operatorname{rank} B = 0 = \operatorname{tr} E_{12}.$$

Proceeding similarly we can show that $F(E_{ij}) = \operatorname{tr} E_{ij} = 0$ for any $i \neq j$, which concludes the proof. □

Appendix D. Fourier analysis and the Discrete Fourier Transform

This chapter is admittedly a little out of place here. Nonetheless, my first encounter with the Discrete Fourier Transform and its fast implementation, the Fast Fourier Transform, occurred courtesy of kernel density estimator (see Section 20), which is why this chapter is here.

The key takeaways of this chapter are the following.

- If a function is *periodic* then its Fourier spectrum is *discrete* and inverse Fourier *series* recovers the function. We may view such a function as not having any *large-scale* structure.
- If a function is *band-limited*, i.e. its Fourier spectrum is *compact*, then it is characterized by spatial sampling at frequencies at or above its Nyquist frequency and the inverse *discrete-time* Fourier transform from these samples recovers the function. We may view such a function as not having any *small-scale* structure.
- The Poisson summation formula tells us that the Fourier transform turns *periodization* into evenly-spaced *sampling*, and vice-versa.
- If a function is *both* periodic and band-limited then its Fourier spectrum is *finite*. Such a function may be characterized in two ways:
  1. using finitely many coefficients of its Fourier series or
  2. using finitely many spatial samples.

  These characterizations turn out to be equivalent since (1) is the *Discrete* Fourier Transform of (2).
- In general the Discrete Fourier Transform (DFT) is used when we compute approximations of a function using finitely many values, disregarding both large-scales and small-scales, such that

  $$(P_T f)[n] \underset{DFT^{-1}}{\overset{DFT}{\rightleftarrows}} \left(P_B \hat{f}\right)[k]$$

  where $g[n] := \frac{1}{B} g\left(\frac{n}{B}\right)$ and $h[k] := \frac{1}{T} g\left(\frac{k}{T}\right)$ and where $P_T f$ denotes the T-periodic summation of f.

Appendix E. Reference notebooks

In this section we list Jupyter notebooks that provide (fairly) clean implementations of common inference techniques. All of these notebooks may be found at github.com/aremondtiedrez/all-of-statistics.

(1) Confidence bands for the empirical CDF: Chapter 07, Exercises 03 and 07.
(2) Bootstrap confidence intervals: Chapter 08, Exercise 03.
(3) Permutation test: Chapter 10, Bonus, Permutation Tests.
(4) Goodness-of-fit test: Chapter 10, Bonus, Goodness-of-fit Test (only for a Normal model).
(5) Confidence intervals (both normal and percentile) from both parametric bootstrap and posterior distributions: Chapter 11, Exercise 04, Version 2.
(6) James-Stein estimator (and comparison to the MLE) with a nice tabulated summary of the results: Chapter 12, Exercise 06.
(7) Simple linear regression, with plots: Chapter 13, Exercise 06.
(8) Multiple linear regression, with varying model selection procedures: Chapter 13, Exercise 07.
(9) Logistic regression, with varying model selection procedures: Chapter 13, Exercise 11, Version 3.
(10) Sampling uniformly at random from the $(k-1)$–dimensional simplex: Chapter 14, Exercise 03.
(11) Cheap way to generate symmetric positive-definite matrices (although the distribution is not guaranteed to be anything specific): Chapter 14, Exercise 04.
(12) Fisher confidence interval for the correlation: Chapter 14, Exercise 05.
(13) Testing the independence of two binary random variables: Chapter 15, Exercise 04.
(14) Testing the independence of two discrete random variables: Chapter 15, Exercise 05.
(15) Testing the independence of a continuous random variable and a binary random variable: Chapter 15, Exercise 07.
(16) Density estimation via histograms and kernel density estimators, with cross-validation selection of the amount of smoothing and confidence band: Chapter 20, Exercise 02, Versions 2, 3, and 4 (version 2 takes care of histogram estimation, version 3 takes care of kernel density estimation, and version 4 performs kernel density estimation, with given bandwidth, using the Fast Fourier Transform).
(17) Nonparametric regression via the Nadaraya-Watson kernel estimator: Chapter 20, Exercise 03.
(18) Cosine basis approximation: Chapter 21, Exercise 06.
(19) Haar wavelet basis approximation: Chapter 21, Exercise 12.
(20) Haar wavelet density estimation: Chapter 21, Exercise 10.
(21) Cosine basis regression: Chapter 21, Exercise 07, Version 2.
(22) Haar wavelet basis regression: Chapter 21, Exercise 07, Version 3.
(23) Classification via LDA, QDA, logistic regression, and trees, including K-fold cross-validation and very rudimentary feature selection (using a Wald test to test for significant intra-class mean differences). Chapter 22, Exercise 03, Version 3.

(24) Classification using a *linear* SVM, i.e. an Optimal Separating Hyperplane classifier (in the terminology of these notes): Chapter 22, Exercise 05.

(25) Classification using SVM with polynomial kernels of various degrees (including building an `sklearn` pipeline to normalize the data): Chapter 22, Exercise 08, Version 2.

(26) Classification using SVM with Gaussian kernel, including hyper-parameter search to minimize the cross-validation estimate of the true error rate: Chapter 22, Bonus 03.

(27) Classification using kernel logistic regression (with both polynomial kernels and the Gaussian kernel), including building a custom `KernelTransformer` to streamline that process: Chapter 22, Bonus 02.

(28) K-nearest neighbors classification, including cross-validation search for the best K: Chapter 22, Exercise 09.

(29) Bagging (based on trees): Chapter 22, Exercise 11.

(30) Boosting (using AdaBoost, and based on trees): Chapter 22, Exercise 12.

References

[And03] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 2003.

[BH10] W. Böhm and K. Hornik. A kolmogorov-smirnov test for r samples. *UW Wien Working Paper Series*, 2010.

[BV09] S. Boyd and L. Vandenberghe. Convex optimization, 2009.

[Cra99] H. Cramér. *Mathematical Methods of Statistics (PMS-9)*. Princeton University Press, 1999.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[Ken61] M. G. Kendall. *The Advanced Theory of Statistics, Volume 2*. Hafner Publishing Company, 1961.

[Mil] S. J. Miller. The probability lifesaver: Order statistics and the median theorem.

[Sco15] D. W. Scott. *Multivariate Density Estimation : Theory, Practice, and Visualization*. Wiley, 2015.

[Sel] H. Seltman. Approximations for mean and variance of ratio.

[Was10] L. Wasserman. *All of Statistics : a Concise Course in Statistical Inference*. Springer, 2010.

[Wei14] S. Weisberg. *Applied Linear Regression*. Wiley, 2014.

[ZH05] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 2005.