STAT 24410 Lecture Notes

Daniel Sanz-Alonso

11.14.22

Table of Contents

Foreword

These lecture notes are written with the purpose of helping you learn the material covered in the Autumn 2022 edition of STAT 24410/STAT 30030. The presentation is intended to be direct, informal, and fresh. The level of mathematical rigor is hopefully adequate to learn the material for the first time. I recommend you to use the textbook (Rice) as a complement, and on occasion I have written in parenthesis the sections of the textbook that correspond to a section in these notes. While I do encourage you to use the textbook and other resources, **these notes will be the basis for specifying the material that you should master.** Before each exam I will let you know which topics are covered by referring to these notes; however, you can choose to study those topics from these notes, from the textbook, or from any other resource that covers the same topics.

I want to thank Rina Foygel Barber for her generosity; many sections of these notes are based on personal notes that she developed and shared with me. I also want to thank Ran Dai for reading a previous version of these notes and providing extensive feedback that has greatly improved the presentation. I have made a significant effort to make the notes readable but they may still contain typos and inconsistencies. I appreciate it if you let me know of any by email.

Chapter 1

Introduction to Probability and Random Variables

1.1 Probability spaces, conditional probability, independence (1.2, 1.3, 1.4, 1.5)

A random experiment is a procedure with several possible outcomes.

- We call the set of all possible outcomes, Ω, a sample space.
- We call any possible outcome $\omega \in \Omega$ a sample point.
- We call any subset $E \subset \Omega$ an event.

A probability measure $\mathbb{P}$ on Ω is a rule that assigns to each event of interest a non-negative number. We ask that it satisfies three axioms:

1. $0 \leq \mathbb{P}(E) \leq 1$ for any event $E \subset \Omega$.
2. $\mathbb{P}(\Omega) = 1$.
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ for any disjoint events A and B.
   More generally, $\mathbb{P}(A \cup B \cup C \cup D \dots) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) + \mathbb{P}(D) + \cdots$ for any mutually disjoint events $A, B, C, D, \dots$ (holds for countable unions).

The following properties follow immediately from the axioms (check them!):

1. $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$ where $E^c = \Omega \setminus E$.
2. $\mathbb{P}(\varnothing) = 0$.
3. If $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
4. Addition law: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

A pair $(\Omega, \mathbb{P})$, where Ω is a sample space and $\mathbb{P}$ a probability on Ω, is called a probability space.
We define the conditional probability of A given B, provided that $\mathbb{P}(B) > 0$, by

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

We say that A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

We say that A, B, C are mutually independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C), \quad \mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C), \quad \mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C).$$

Two useful direct consequences (check them!) of the definition of conditional probability are:

- Product rule: $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) \quad [\mathbb{P}(B) > 0]$.
- Bayes' rule: $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \qquad [\mathbb{P}(A)\mathbb{P}(B) > 0]$.

1.2 Counting, permutations, & combinations (1.4)

1. **Ordered samples:** There are n^r ways to choose an ordered sample of size r from a set of size n, if sampling with replacement. There are $n(n-1)(n-2)\dots(n-r+1)$ if sampling without replacement.
2. **Unordered samples:** There are $\binom{n}{r} = \frac{n!}{r!(n-r)!}$ ways to choose an unordered sample of size r from a set of size n. Why? First, there are $n(n-1)\dots(n-r+1) = \frac{n!}{(n-r)!}$ many ways to choose an ordered list of r items. Then, each possible list has $r!$ many orderings—we have overcounted by a factor of $r!$
3. **Sorting into groups:** There are $\binom{n}{n_1,n_2,\dots,n_k} = \frac{n!}{n_1!\cdot n_2!\cdot\dots\cdot n_k!}$ ways to group n objects into k (unordered) groups of size $n_1 + n_2 + \dots + n_k = n$.

Sorting into groups can specialize to unordered samples—note that $\binom{n}{r} = \binom{n}{r,n-r}$, i.e. choosing a group of size r is equivalent to splitting into two groups of size r and $n-r$.

Count how many ways each is possible:

1. Split 10 students into Section 1 and Section 2, each with 5 students: $\binom{10}{5} = \frac{10!}{5!5!} = 252$.
2. Split 10 students into two groups, each with 5 students: $\binom{10}{5}/2 = 126$.
3. Split 10 students into two groups of any size (≥ 1 student in each group): $\frac{2^{10}-2}{2} = 511$
4. Scramble the word "students": 2 s's, 2 t's, 1 each u,d,e,n — $\binom{8}{2,2,1,1,1,1} = \frac{8!}{2!2!1!1!1!1!} = 10080$.

1.3 Discrete random variables (2.1–2.3)

1.3.1 Probability function and cumulative distribution function

After performing an experiment and observing the outcome, we might want to quantify some aspect of what we've observed. Formally, a random variable is a function from the sample space Ω to the real numbers. It assigns a numerical value to each possible outcome. Examples:

- Experiment: I roll one die, if it's a 6 I win \$10, otherwise I lose \$1. X is my net gain. If it's a fair die, then we know the probability measure:

Outcome	1	2	3	4	5	6
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
X	-1	-1	-1	-1	-1	10

Possible values of X are -1 and 10, with $\mathbb{P}(X = -1) = \frac{5}{6}$ and $\mathbb{P}(X = 10) = \frac{1}{6}$.

- Experiment: record sequence 3 of dice rolls. $\Omega = \{111; 112; 113; ....\}$. X = number of 1's. If the die is fair:

Outcome	111	112	113	...
Probability	$\frac{1}{6^3}$	$\frac{1}{6^3}$	$\frac{1}{6^3}$	...
X	3	2	2	...

For a random variable, we might be interested in probabilities that it takes a certain value or lies in a certain range. We should think of these as events. For example, if we ask "What is the probability that X is higher than 10?", written $\mathbb{P}(X > 10)$, we can equivalently consider the event

$$A = \{\text{all outcomes in } \Omega \text{ for which } X \text{ is higher than } 10\}$$

and can calculate $\mathbb{P}(A)$.

A discrete random variable is a random variable with finitely many, or countably-infinitely many, possible values.

- Probability function (a.k.a. probability mass function (PMF), frequency function): what is the probability that X will be equal to some particular value x?

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

  Sometimes we might write $p_X(\cdot)$ to emphasize that we are looking at the distribution of X.

- Cumulative distribution function (CDF): what is the probability that X will be less than or equal to some particular value x?

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

  Similarly we might write $F_X(\cdot)$.

1.3.2 Discrete distributions

For a discrete random variable X, what do we mean by the distribution of X? It's anything that specifies the exact probability function of X: the probability function itself, or the CDF, or some other description. If X and Y have the same probability function, i.e. $\mathbb{P}(X = x) = \mathbb{P}(Y = x)$ for any value x, then we say that X and Y have the same distribution. They do not have to be equal, e.g. X is the value on the red dice and Y is the value on the blue dice.

Any valid probability function specifies a distribution. But, some specific probability functions are common and come up in many scenarios, so they are named.

Bernoulli

X is a Bernoulli random variable if its only possible values are 0 and 1. It is parameterized by the probability p of "success", i.e. getting a 1. The distribution is written as Bernoulli(p) and we write $X \sim \text{Bernoulli}(p)$. The probability function is given by

$$\mathbb{P}(X = 1) = p, \qquad \mathbb{P}(X = 0) = 1 - p.$$

- Roll a dice once and record whether I get a 1 (success) or not (failure): Bernoulli$(\frac{1}{6})$.

Binomial

Binomial distribution: we count the number of successes from several trials. The Binomial(n, p) distribution, is the distribution of X which counts the number of successes from n trials, which are all independent and each have a probability p of success.

- Roll a dice 3 times and count the number of 1's: Binomial$(3, \frac{1}{6})$.
- Survey 100 people chosen at random in Chicago, and ask whether they support Rahm Emmanuel: Binomial$(100, p)$ where p is the true proportion of Chicago residents supporting R.E.

The probability function is

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

for each $k = 0, \ldots, n$. To understand this formula: There are $\binom{n}{k}$ sequences of 0's and 1's; each one has $p^k(1-p)^{n-k}$ probability. We can think of a Binomial as a sum of n Bernoulli's (which have to be independent!).

- Suppose I survey 100 Chicagoans, and in reality 80% of all Chicagoans support R.E. How likely is it that less than 70% of my sample will answer Yes, i.e. that I will underestimate the proportion as 0.7 or below? Take X = number of Yes's $\sim$ Binomial$(100, 0.8)$.

$$\mathbb{P}(X \leq 70) = \sum_{k=0}^{70} \mathbb{P}(X = x) = \sum_{k=0}^{70} \binom{100}{k} 0.8^k (1-0.8)^{100-k} = 0.011.$$

Poisson

The Poisson distribution is a natural distribution for data that is a count. For example, photon emission from an X-ray source roughly follows this distribution. The Poisson (λ) distribution has possible values $0, 1, 2, \ldots$, with probabilities

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} .$$

Larger λ means that the value is likely to be larger—we can think of λ as the intensity of the X-ray beam, and X is the (random) number of photons emitted by the beam during a fixed period of time.

- For a low intensity beam with $\lambda = 3$, what is the probability that at least one photon is emitted?

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - \frac{3^0 e^{-3}}{0!} = 0.95.$$

The Poisson distribution can be obtained as a limit of the Binomial distribution. Suppose that successes in an interval $[0, t)$ occur under the following premises:

1. The number of successes in disjoint subintervals are independent.
2. The rate of success λ is constant in time.
3. In an infinitesimal interval $[t, t + \Delta t)$ there is one success at most.

Let X be the number of successes in $[0, t)$. Then $X \sim$Poisson(λt).

Derivation of the Poisson from Binomial We break the interval $[0, t)$ into n subintervals of length Δt. By the model assumptions $1 - 2 - 3$ we can (approximately) compute the number of successes in $(0, t)$ by counting the number of successes in the subintervals:

- There are n subintervals.
- Probability of success in each of them is $\lambda \Delta t$.

Let X be the total number of successes. $X \sim$Binomial$(n, \lambda \Delta t)$.

$$\begin{aligned}
\mathbb{P}(X = k) &= \binom{n}{k} (\lambda \Delta t)^k (1 - \lambda \Delta t)^{n-k} \\
&= \frac{n!}{k!(n-k)!} \left(\lambda \frac{t}{n}\right)^k \left(1 - \lambda \frac{t}{n}\right)^{n-k} \\
&= \frac{(\lambda t)^k}{k!} \frac{n(n-1)\cdots(n-k+1)}{n^k} \left(1 - \frac{\lambda t}{n}\right)^{-k} \left(1 - \frac{\lambda t}{n}\right)^n \xrightarrow{n \to \infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!},
\end{aligned}$$

which is the probability function of a Poisson r.v. with parameter λt.

Geometric distribution

How many times do I have to flip a coin to get the first Heads? Suppose p = prob of Heads.

$$\mathbb{P}(k \text{ times}) = \mathbb{P}(\text{first } k-1 \text{ are T, } k\text{th one is H}) = (1-p)^{k-1} \cdot p\ .$$

So, the PMF is $p(k) = (1-p)^{k-1} \cdot p$ for the set of possible values $k = 1, 2, 3, \ldots$. This is the Geometric(p) distribution.

The geometric distribution is memoryless: given that I've flipped the coin 100 times with no successes so far, the distribution of the additional number of flips needed is unchanged. For example, if I've gotten 100 tails, what's the probability that the next one is heads?

$$\mathbb{P}(X = 101 \mid X > 100) = \frac{\mathbb{P}(X = 101 \& X > 100)}{\mathbb{P}(X > 100)} = \frac{\mathbb{P}(X = 101)}{\mathbb{P}(X > 100)} = \frac{(1-p)^{100} \cdot p}{(1-p)^{100}} = p.$$

1.3.3 Functions of a discrete r.v.

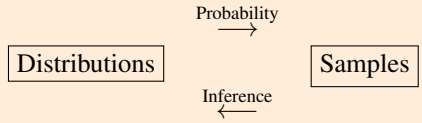
Suppose that X is a discrete r.v. with PMF $f_X(x)$, and $Y = g(X)$. What can we say about the distribution of Y?

- $X \sim$ Poisson(λ) is the X-ray beam passed into a tissue, Y = radiation dose received, then $Y = g(X)$ where $g(x) = c \cdot x$, c = radiation dose of one photon at that particular energy.

Y will certainly also be discrete, since it cannot have more possible values than X. PMF of Y:

$$p_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(X \text{ takes some value so that } g(X) = y) = \sum_{\substack{\text{possible values } x \text{ for } X \\ \text{such that } g(x) = y}} p_X(x).$$

Recess



- In most applications of statistics, our goal is to answer some questions about a population, and our available information is data sampled from the population
- Usually, if we know the distribution of the values in the population, then we'd have the answer to our question
- Modeling = deciding which types of distributions might be a good way to describe the underlying population
- Model fitting = estimating the parameters of the model, based on the data

1.4 Continuous random variables (2.2, 2.3)

Recall: a discrete random variable can take finitely many or countably infinitely many possible values. Other random variables might take values in a continuum. If a random variable X has no mass at any single value, i.e. $\mathbb{P}(X = x) = 0$ for any $x \in \mathbb{R}$, then it is a continuous random variable.

1.4.1 CDF & density

As with a discrete r.v., for a continuous r.v. we can describe its distribution via the CDF, $F(x) = \mathbb{P}(X \leq x)$. However, it would not make sense to use a PMF, since $\mathbb{P}(X = x) = 0$ for every value x. Instead we use a density function,

$f(x)$, which plays the same role:

$$\mathbb{P}(a < X < b) = \int_{x=a}^{b} f(x)\ dx.$$

Properties of a density function:

$$f(x) \geq 0 \text{ for all } x \in \mathbb{R}, \int_{x=-\infty}^{\infty} f(x)\ dx = 1,\ f \text{ is piecewise continuous.}$$

X is supported on the interval $I \Leftrightarrow f(x) = 0$ for any $x \notin I$. Sometimes it is convenient to think of f as a function with domain I rather than domain $\mathbb{R}$.

For any continuous distribution, closed vs open endpoints of intervals do not matter since $\mathbb{P}(X = x) = 0$ for any x, for example $\mathbb{P}(X \geq a) = \mathbb{P}(X > a)$, etc.

Now let's consider the CDF:

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(-\infty < X \leq x) = \int_{t=-\infty}^{x} f(t)\ dt.$$

Equivalently, $f(x) = F'(x)$: density is the derivative of the CDF.

1.4.2 Continuous distributions

As in the discrete case, any valid density function specifies a distribution. Some specific probability functions are common and come up in many scenarios, so they are named.

Uniform distribution

What do we mean when we "draw X at random from the interval $[a, b]$"? For a finite set (e.g. choose a random number between 1 and 10), the meaning is clear—every value should be equally likely. In the continuous case, we mean that the probability should be "evenly spread" across the interval, which corresponds to a flat density function:

$$f(x) = \frac{1}{b-a} \text{ for } x \in [a, b].$$

Exponential distribution

This continuous distribution supported on $[0, \infty)$, is memoryless (like the geometric distribution in the discrete case). Natural model for many physical phenomena that are memoryless: half-life decay, ion passing through a channel. Parameter: "rate" $\lambda > 0$. Larger λ means shorter time to decay etc.

$$f(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0.$$

Calculate the CDF:

$$F(x) = \mathbb{P}(X \leq x) = \int_{t=0}^{x} f(t)\ dt = \int_{t=0}^{x} \lambda e^{-\lambda t}\ dt = \left[-e^{-\lambda t}\right]_{t=0}^{x} = 1 - e^{-\lambda x}.$$

Gamma distribution

This is a generalization of the exponential distribution, also supported on $[0, \infty)$, which allows for a change in the rate of decay (e.g. a shell that is more likely to disintegrate if it's older, due to fragmentation; or, less likely, due to calcification that makes it stronger). Parameters: "shape" $\alpha > 0$, and "rate" $\lambda > 0$. Density:

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

Here the Γ function is a generalization of factorials: $\Gamma(k) = (k-1)!$, but it's also defined for non-integers.

Note that $\alpha = 1$ gives the exponential distribution. $\alpha < 1$ gives a shape for an event that becomes increasingly less likely, with the density declining sharply after zero. $\alpha > 1$ is for an event that becomes more likely as time goes on, and gives a shape that has a peak (mode) at some value above zero.

Note: the textbook calls λ the scale but this does not agree with standard terminology.

Normal distribution

Two parameters: "mean" μ, and "standard deviation" σ . (σ^2 is called the variance). Written as $\mathrm{N}(\mu, \sigma^2)$. Density:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

$\mathrm{N}(0,1)$ is called the standard normal distribution.

Symmetric "bell curve" shape, often used to approximate distribution of many measured quantities such as height. Most variables we can measure are not normally distributed, however, there is an important connection between the normal distribution and the process of sampling from a population: an average of a randomly chosen sample is approximately normally distributed, even if the individual values are not approximately normally distributed (Central Limit Theorem, later in the course).

1.4.3 Mixed random variables

Rainfall: we can think of the amount of rain that will tomorrow in Chicago is a random variable, which can take values in $[0, \infty)$. We know that $\mathbb{P}(X = 0)$ positive (there is a positive chance that there will be no rain) but $\mathbb{P}(X = x) = 0$ for any $x > 0$. We can express X as a mixture of a discrete r.v. and a continuous r.v., for example

$$X = 0 \text{ with probability } 0.6,\ X \sim \text{Exponential}(3) \text{ otherwise.}$$

We can think of this as a hierarchical model:

$$R \sim \text{Bernoulli}(0.4)$$
$$X \mid R \sim \begin{cases} 0, \text{ if } R = 0, \\ \text{Exponential}(3), \text{ if } R = 1. \end{cases}$$

Here R is a Bernoulli r.v., and can also be called an indicator variable which indicates whether a particular event occurs. If A = event that it does rain, then we can write $R = \mathbb{1}_A$, meaning that $R = 1$ for any outcome in the event A and $R = 0$ otherwise. Note that $\mathbb{1}_A$ is Bernoulli(p), with parameter $p = \mathbb{P}(A)$.

Any (univariate) r.v. can be decomposed as a mixture of a discrete r.v. and a continuous r.v. Note that for any type (discrete / continuous / mixed), we can always use the CDF $F_X(x)$. However, for mixed r.v.'s, there is no analogue of the PMF (as for discrete) or the PDF (as for continuous).

1.4.4 Functions of a continuous r.v. (2.3)

If X is a continuous random variable with density $f_X(x)$, and $Y = g(X)$, then what is the distribution of Y? Depending on the function g, Y might be discrete or continuous or mixed.

If g is a change of units, $g(x) = a \cdot x$ for some $a > 0$, then Y is a continuous r.v. and we can show its density is

$$f_Y(y) = f_X(y/a) \cdot 1/a.$$

Why? Dividing by a takes care of area integrating properly.

If g is a differentiable and strictly monotonic function, then Y will be a continuous r.v. with a density we can write in closed form:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

Distribution	How it arises?	Possible values
$X \sim$ Bernoulli(p)	One success/fail trial, p probability of success in it, $X =$ "number of successes".	0 or 1.
$X \sim$ Binomial(n,p)	n independent trials, p prob. of success in each $X =$ "number of successes".	$0, 1, \ldots, n$.
$X \sim$ Geometric(p)	Sequence of indep. trials, p prob. of success in each $X =$ "first success".	$0, 1, \ldots$
$X \sim$ Poisson(λ)	Poisson process with rate λ, $X =$ "# successes in $[0, t)$".	$0, 1, \ldots$
$X \sim$ Exponential(λ)	Poisson process with rate λ, $X =$ "# time of 1st success".	$(0, \infty)$.
$X \sim$ Gamma(λ, α)	Poisson process with rate λ, $X =$ "# time of α-th success".	$(0, \infty)$.
$X \sim$ N(μ, σ^2)	Ubiquitous. Central limit theorem.	$(-\infty, \infty)$.
$X \sim$ Uniform(a, b)	$X =$ "Point chosen at random in interval (a, b)".	(a, b).

Table 1.1: Brief, sketchy summary of how four important discrete distributions and four important continuous distributions may arise in practice.

We can check that the appropriate integrals are equal by taking the change of variables $x = g^{-1}(x)$.
Example:

- $X \sim$ Exponential(λ) and $Y = 1/X$. Then we have $g(x) = 1/x$ and $g^{-1}(y) = 1/y$, so $\left|\frac{d}{dy} g^{-1}(y)\right| = \frac{1}{y^2}$, and we have
$$f_Y(y) = f_X(1/y) \cdot \frac{1}{y^2} = \frac{\lambda}{y^2} e^{-\lambda/y}.$$

Important examples:

- If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$ then $Y \sim N(a\mu + b, a^2\sigma^2)$.
  Let $X \sim N(\mu, \sigma^2)$, and let $Y = a + bX$. We stated before that Y is also normally distributed but let's check:
$$g(x) = a + bx, g^{-1}(y) = \frac{y - a}{b}, \left|\frac{d}{dy} g^{-1}(y)\right| = \frac{1}{|b|}.$$
  And,
$$f_Y(y) = f_X(g^{-1}(y)) \cdot \frac{1}{|b|} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\frac{y-a}{b} - \mu)^2/2\sigma^2} \cdot \frac{1}{|b|} = \frac{1}{\sqrt{2\pi(b^2\sigma^2)}} e^{-(y-(a+b\mu))^2/2(b^2\sigma^2)}.$$
  So, $Y \sim N(a + b\mu, b^2\sigma^2)$.

- If X is continuous r.v. with CDF F, then $Y = F(X)$ has a Uniform$[0, 1]$ distribution.
  Sketch Proof: for any $u \in [0, 1]$, find $x = F^{-1}(u)$, i.e. $F(x) = u$
$$\mathbb{P}(Y \leq u) = \mathbb{P}(X \leq F^{-1}(u) = x) = F(x) = u.$$
  [Remark: the proof above works without further detail if F is strictly monotone in $(0, 1)$. Otherwise, care should be taken in defining F^{-1} through the "pseudoinverse" or quantile function.]

- If F is the CDF for some continuous distribution, and $U \sim$ Uniform$[0, 1]$, then $F^{-1}(U)$ is a continuous r.v. with CDF F. (A similar construction exists for a discrete or mixed distribution.)
$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$
  (Note that $F(x) \in [0, 1]$ always by definition of the CDF.)

Quiz: Show that if $X \sim$ Exponential(λ) and $Y = aX$ for $a > 0$, then $Y \sim$ Exponential(λ/a).

1.5 Expected value (4.1)

The expected value of a random variable X is the value we expect to get, on average, if we were able to repeat the random experiment or random process that generates X many times. It is written as $\mathbb{E}(X)$, μ_X, or just μ. For a discrete random variable it is calculated as

$$\mathbb{E}(X) = \sum_x x \cdot p(x),$$

which we can interpret as a "long-run average": $p(x)$ is the proportion of times that we'll get the value x. For a continuous random variable,

$$\mathbb{E}(X) = \int_{x=-\infty}^{\infty} x \cdot f(x)\, dx.$$

In both of these cases, we need to be careful to check that this value exists. For a finitely valued discrete random variable, this is fine. If X is discrete with infinitely many possible values, the summation exists as long as

$$\sum_x |x| \cdot p(x) < \infty$$

and for a continuous random variable,

$$\int_{x=-\infty}^{\infty} |x| \cdot f(x)\, dx < \infty.$$

If the appropriate condition is not satisfied, then we say that the expectation does not exist.

1.5.1 Examples

- Let X be Bernoulli(p). Then the possible values of X are 0 and 1, so $\mathbb{E}(X) = \sum_x x \cdot p(x) = 0 \cdot p(0) + 1 \cdot p(1) = p$. In particular, if $X = \mathbb{1}_A$ then $\mathbb{E}(X) = \mathbb{P}(A)$.
- Let X be Exponential(λ). Then

$$\mathbb{E}(X) = \int_{x=0}^{\infty} x \cdot \lambda e^{-\lambda x}\, dx = \left[-xe^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x}\right]_{x=0}^{\infty} = \frac{1}{\lambda}.$$

  For a time process, if the rate is $\lambda = 0.5(\text{seconds})^{-1}$, the expected value is 2 seconds (this is why it is common to refer to $1/\lambda$ as the scale).

1.5.2 Properties

Theorem 1 (Linearity)**.** *Let $X_1, \ldots, X_n$ be random variables. Let $Y = a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$. Then*

$$\mathbb{E}(Y) = a + b_1 \mathbb{E}(X_1) + b_2 \mathbb{E}(X_2) + \cdots + b_n \mathbb{E}(X_n),$$

as long as $\mathbb{E}(X_1), \ldots, \mathbb{E}(X_n)$ all exist.

Note that this doesn't assume anything about the relationship among the X_i's, e.g. they do not need to be independent.

Theorem 2 (Monotonicity)**.** *Let X and Y be two random variables such that $X \leq Y$ almost surely (meaning, $\mathbb{P}(X \leq Y) = 1$; implicitly, X and Y are defined on the same sample space). Then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.*

Theorem 3 (Transformations)**.** *Let X be a r.v. and let $Y = g(X)$. Then if X is discrete,*

$$\mathbb{E}(Y) = \sum_x g(x) \cdot p(x),$$

while if X is continuous,

$$\mathbb{E}(Y) = \int_{x=-\infty}^{\infty} g(x) \cdot f(x)\, dx.$$

1.5.3 More examples

- Expectation for $X \sim \text{Binomial}(n,p)$: $\mathbb{E}(X) = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k}$. Easier to do: $X = X_1 + \cdots + X_n$ where $X_i = \mathbb{1}\{i\text{th trial succeeds}\}$. Then $X_i \sim \text{Bernoulli}(p)$, so $\mathbb{E}(X_i) = p$, and therefore $\mathbb{E}(X) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = np$.
- Let X = number of red cards when I draw a hand of 20 cards. X is not binomial since the draws are not independent, however, we can still write $X = X_1 + \cdots + X_{20}$ where $X_i = \mathbb{1}\{i\text{th card is red}\}$. So, $\mathbb{E}(X) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_{20})$, and each card is equally likely to be red or black so $\mathbb{E}(X_i) = 0.5$ for each i. So, $\mathbb{E}(X) = 10$ as expected.
- Let (X,Y) be a point chosen uniformly at random in the unit square. What is the circumference of the rectangle defined by $(0,0)$ and (X,Y)? $C = 2X + 2Y$ so $\mathbb{E}(C) = 2\mathbb{E}(X) + 2\mathbb{E}(Y)$, since clearly X and Y are each Uniform$[0,1]$ we have $\mathbb{E}(X) = \mathbb{E}(Y) = 0.5$. To prove this last calculation, density of X is $f(x) = 1$ on $[0,1]$, so $\mathbb{E}(X) = \int_{x=0}^{1} x \cdot f(x)\, dx = \int_{x=0}^{1} x\, dx = \left[\frac{1}{2}x^2\right]_{x=0}^{1} = \frac{1}{2}$.
- In the same context, what is the expectation of the length from $(0,0)$ to (X,Y)? This is $L = \sqrt{X^2+Y^2}$ and we do not yet have a rule for calculating this—we'll need to use joint distributions. However, we can calculate things like $\mathbb{P}(L \le 1)$. Since the point is uniformly sampled, the probability of landing in the region is $\frac{\text{Area of the region}}{\text{Total area}} = \frac{\frac{1}{4}\cdot\pi\cdot 1^2}{1} = \pi/4$.

1.6 Variance (4.2)

The variance of a random variable X, measures how much it typically varies from its mean. Definition:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]\,,$$

as long as this expectation exists, and $\mathbb{E}(X)$ exists.
For a discrete random variable, we would calculate it as

$$\text{Var}(X) = \sum_x p(x) \cdot (x-\mu)^2$$

and for a continuous r.v.,

$$\text{Var}(X) = \int_{x=-\infty}^{\infty} f(x) \cdot (x-\mu)^2\, dx$$

where $\mu = \mathbb{E}(X)$. Variance is often denoted as σ^2 or σ_X^2. Also, σ is called the standard deviation.

1.6.1 Properties

1. Variance can be calculated with a simpler formula:

   $$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

   (Note: this also shows that $\mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$ in general.)

   Proof.

   $$\text{Var}(X) = \mathbb{E}((X-\mu)^2) = \mathbb{E}(X^2 - 2\mu\cdot X + \mu^2) = \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2 = \mathbb{E}(X^2) - \mu^2$$

   by linearity of the expected value. □

2. Degenerate case: $\text{Var}(X) = 0$ if and only if $\mathbb{P}(X = \mu) = 1$.

Proof. • If $\mathbb{P}(X = \mu) = 1$ then X is a discrete random variable, and so $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 1 \cdot \mu^2 - (1 \cdot \mu)^2 = 0$.

• If $\mathbb{P}(X = \mu) < 1$ then we must have $\mathbb{P}(|X - \mu| < \epsilon) < 1$ for some $\epsilon > 0$. Then let $Y = (X - \mu)^2$ and $Z = \epsilon^2 \cdot \mathbb{1}_{|X-\mu| \geq \epsilon}$. We have $Z \leq Y$ always, so

$$\mathrm{Var}(X) = \mathbb{E}((X - \mu)^2) = \mathbb{E}(Y) \geq \mathbb{E}(Z) = \epsilon^2 \cdot \mathbb{P}(|X - \mu| \geq \epsilon) > 0 \ .$$

□

1.6.2 Examples

1. X is Bernoulli(p). What is $\mathrm{Var}(X)$? We know that $\mathbb{E}(X) = p$ and
$$\mathbb{E}(X^2) = 1^2 \cdot p + 0^2 \cdot (1 - p) = p \ ,$$
and so
$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = p - p^2 = p(1 - p) \ .$$
2. X is Exponential(λ). What is $\mathrm{Var}(X)$? We know that $\mathbb{E}(X) = 1/\lambda$ and
$$\mathbb{E}(X^2) = \int_{x=0}^{\infty} x^2 \cdot \lambda e^{-\lambda x} \, dx = \left[-x^2 e^{-\lambda x} - \frac{2}{\lambda} x e^{-\lambda x} - \frac{2}{\lambda^2} e^{-\lambda x} \right]_{x=0}^{\infty} = \frac{2}{\lambda^2} \ ,$$
so
$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda} \right)^2 = \frac{1}{\lambda^2} \ .$$

1.7 Inequalities and limit theorems

1.7.1 Markov's and Chebyshev's inequality, Law of Large Numbers

Markov's inequality

Theorem 4 (Markov's inequality)**.** *Let X be any random variable supported on $[0, \infty)$. For any $t > 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mu}{t}.$$

Proof. Let $Y = t \cdot \mathbb{1}\{X \geq t\}$. Then $Y \leq X$ always, and so $\mathbb{E}(Y) \leq \mathbb{E}(X) = \mu$. But

$$\mathbb{E}(Y) = \sum_y y \cdot p_Y(y) = 0 \cdot \mathbb{P}(Y = 0) + t \cdot \mathbb{P}(Y = t) = t \cdot \mathbb{P}(X \geq t) \ .$$

□

Chebyshev's inequality

Theorem 5 (Chebyshev's inequality)**.**

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Proof. Let $Y = (X - \mu)^2$, then $\mathbb{E}(Y) = \sigma^2$ and apply Markov's inequality to get

$$\mathbb{P}(Y \geq t) \leq \frac{\sigma^2}{t}.$$

Replacing t with t^2, we prove the result. □

Law of Large Numbers

Theorem 6 (Law of large numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. with mean μ and finite variance σ^2. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\big(|\bar{X}_n - \mu| > \epsilon\big) \xrightarrow{n\to\infty} 0.$$

Proof. Let $\epsilon > 0$. Note that

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{X_1 + \cdots X_n}{n}\right] = \frac{1}{n}\mathbb{E}[X_1 + \cdots X_n] = \frac{1}{n}(\mu + \cdots + \mu) = \mu,$$

and similarly

$$\mathrm{Var}[\bar{X}_n] = \mathrm{Var}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{n^2}\mathrm{Var}[X_1 + \cdots X_n] = \frac{1}{n^2}(\sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}.$$

Therefore, applying Chebyshev's inequality gives

$$\mathbb{P}\big(|\bar{X}_n - \mu| > \epsilon\big) \le \frac{\sigma^2}{n\epsilon^2} \xrightarrow{n\to\infty} 0.$$

□

1.7.2 Central Limit Theorem

Standardization & calculating normal probabilities

We will denote by $\Phi(x)$ the CDF of a standard normal distribution, that is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right)\, dt.$$

How can we calculate CDF's and probabilities for another normal distribution?
Suppose that $X \sim N(\mu, \sigma^2)$. We have seen that any linear transformation of X is normal. In particular,

$$Z = \frac{X - \mu}{\sigma}$$

is standard normal since we can calculate $\mathbb{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$. This is called standardizing X.
Now let's look at the CDF of X.

$$\mathbb{P}(X \le x) = \mathbb{P}\left(\frac{X - \mu}{\sigma} \le \frac{x - \mu}{\sigma}\right) = \mathbb{P}\left(z \le \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

We can look up with software or in Table 2 in the back of the textbook.

The theorem

Let $X_1, X_2, \ldots$ be i.i.d. from some distribution with mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$ be the sum of the first n terms, and let $\bar{X}_n = S_n/n$ be the average of the first n terms.
Then the distribution of S_n is approximately $N(n\mu, n\sigma^2)$ and the distribution of $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$, as n grows large. More precisely, for any $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n}\cdot\sigma} \le x\right) = \Phi(x),$$

and equivalently

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \le x\right) = \Phi(x).$$

Normal approximation to the binomial

Let $X \sim \text{Binomial}(n, p)$. Then we know $\mathbb{E}(X) = np$, $\text{Var}(X) = np(1-p)$. In fact the distribution of X is well approximated by a normal distribution with these parameters; we can write $X = X_1 + \cdots + X_n$ where $X_i = \mathbb{1}_{\text{success on } i\text{th trial}}$ as usual. Then the X_i's are i.i.d. with mean p, variance $p(1-p)$, so by the CLT we see that the distribution of X is

$$X \approx N(np, np(1-p)).$$

A sample calculation: suppose we survey 1000 people with a binary question and the true proportion in the population is $p = 0.75$. What is the probability that less than 730 people in the sample say yes? Calculate: $np = 750$, $np(1-p) = 187.5$.

$$\mathbb{P}(X < 730) \approx \mathbb{P}(N(750, 187.5) < 730).$$

We calculate this with standardization. Let $Y \sim N(750, 187.5)$ and let $Z = \frac{Y-750}{\sqrt{187.5}}$. Then $Z \sim N(0,1)$. So,

$$\mathbb{P}(Y < 730) = \mathbb{P}\left(\frac{Y-750}{\sqrt{187.5}} < \frac{730-750}{\sqrt{187.5}}\right) = \mathbb{P}(Z < -1.46) = \Phi(-1.46) = 0.0721.$$

We will not worry about continuity corrections in this class.

More examples

1. NegativeBinomial(k, p): how many times do I need to flip a coin with chance p to get the kth Heads? We can write $X = X_1 + \cdots + X_k$ where X_i is how many flips after the $(i-1)$th Heads until the ith Heads, and $X_i \overset{\text{iid}}{\sim} \text{Geometric}(p)$. We know that $\mathbb{E}(X_i) = \frac{1}{p}$ and $\text{Var}(X_i) = \frac{1-p}{p^2}$. So,

$$X \approx N\left(\frac{k}{p}, \frac{k(1-p)}{p^2}\right).$$

2. A gambler plays a game where with probability 0.1 he wins \$8 while with probability 0.9 he loses \$1. What is the probability that he will be ahead after 20 rounds?
   Let X_i = winnings on game i.

$$\mathbb{E}(X_i) = 0.1 \cdot 8 + 0.9 \cdot (-1) = -1$$
$$\mathbb{E}(X_i^2) = 0.1 \cdot 8^2 + 0.9 \cdot (-1)^2 = 7.3$$
$$\text{Var}(X_i) = 7.3 - (-1)^2 = 6.3.$$

   So, for T = total winnings after 20 rounds = $X_1 + \cdots + X_{20}$,

$$T \approx N(20 \cdot (-1), 20 \cdot 6.3) = N(-20, 126).$$

$$\mathbb{P}(T > 0) = \mathbb{P}\left(N(-20, 126) > 0\right) = \mathbb{P}\left(N(0,1) > \frac{0-(-20)}{\sqrt{126}}\right) = \mathbb{P}(N(0,1) > 1.78) = 0.0375.$$

Exercises

Show all work and justify your answers!

Exercise 1 a) For two events A and B show using the axioms of probability that:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

b) For a bill to make it to the office of the president, it must be passed by both the House and the Senate. Suppose that of all bills, 60 percent pass the House, 80 percent pass the Senate, and 90 percent pass at least one of the two. What is the probability that the next bill will make it to the office of the president?

Exercise 2 Covid-19 is carried by 7% of the population. A diagnostic test for Covid-19 has the following accuracy. If an individual has Covid-19, the test correctly detects this 90% of the time. If an individual does not have Covid-19, the test incorrectly reads positive with probability 15%. Suppose person X is given the test and the outcome is positive. What is the probability that the person has Covid-19?

Exercise 3 An advertisement at a Casino claims that the expected value of times a person needs to play a game to win is 3 (the outcome of all games in this exercise are assumed to be independent). Suppose we have reason to believe it takes longer to win. We plan to keep playing the game until we win. If we still haven't won after four plays, we will reject the claim.

a) If the advertisement is true, what is the probability of not winning in the first four games? Namely, what is the chance we are wrong in rejecting the claim?

b) Suppose the true number of times needed to play on average in order to win was truly 5. What is the probability that you will incorrectly accept the claim of the advertisement?

Exercise 4 Let X be a random variable such that:

$$\mathbb{P}(X = -1) = \frac{1}{18}, \quad \mathbb{P}(X = 0) = \frac{16}{18}, \quad \mathbb{P}(X = 1) = \frac{1}{18}.$$

a) Compute $\mathbb{E}[X]$ and $\mathbb{V}[X]$.
b) Find a value $a > 0$ for which Chebyshev's bound is tight, i.e. equality is achieved:

$$\mathbb{P}\left(\left|X - \mathbb{E}[X]\right| \geq a\right) = \frac{\mathbb{V}[X]}{a^2}.$$

Exercise 5 Using Markov's inequality, bound the probability of:
a) Flipping ≤ 10 heads from 100 flips of a fair coin.
b) Flipping ≤ 10 heads from 100 flips of a biased coin with probability of heads equal to 0.2.

Using Chebyshev's inequality, bound the probability of:
c) Flipping between 40 and 60 heads from 100 flips of a fair coin.
d) Flipping between 10 and 30 heads from 100 flips of a biased coin with probability of heads equal to 0.2.

Exercise 6 Suppose that $X \sim N(0, 1)$. Calculate the following two density functions. For each one, be sure to specify the support i.e. the range of possible values of the random variable.

a) Calculate the density of $Y = X^3$.

b) Calculate the density of $Z = |X|$.

Exercise 7 Suppose a building has 10 floors. m people get into the elevator at level 0 and each one independently and uniformly at random chooses a floor between 1 and 10. What is the expected number of stops of the elevator? (Hint: it may be helpful to use linearity of expectation).

Exercise 8 Give an example of a probability space and events A, B, C such that it holds:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C), \quad \mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C)$$

but A, B, C, are not mutually independent.

Exercise 9 You roll a standard die repeatedly, and calculate the average value on the die after $n = 40$ rolls. Using the CLT, what is the probability (approximately) that this average is greater than 5?

Further Exercises

Exercise 1

i) Suppose that $A \cap B = \varnothing$ and $\mathbb{P}(A) > 0$. Can A and B be independent? State your reasons.
ii) Suppose that $A \subset B$ and that $\mathbb{P}(A) > 0$. Are A and B necessarily independent? If not, could they be independent? Give an example if so.

Exercise 2 Let X be a random variable, and define the random variable

$$Y = \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}}.$$

What is the expectation and variance of Y?

Exercise 3

a) A random variable X is always strictly larger than -120. You know that $\mathbb{E}[X] = -30$. Give the best upper bound you can on $\mathbb{P}(X \geq -10)$.

b) Let X be a Poisson random variable with parameter $\lambda = 3$. Let Y be a random variable only taking negative values with $\mathbb{E}[Y] = -2$. Show that $\mathbb{P}(X \geq 10 + Y) \leq \frac{1}{2}$.

Exercise 4 Consider the following generalization of a geometric random variable: suppose we have a positive integer r and a sequence of independent trials with probability of success p. Let X be the trial on which the r-th success occurs.

a) Find the probability distribution of X. (Hint: make sure that for the particular case $r = 1$ your answer agrees with the geometric distribution.)

b) Use linearity of expectation to compute the expected value of X.

Exercise 5 Let $X \sim \text{Exponential}(\lambda)$ and let t be a constant with $0 < t < \lambda$.

1. What is $\mathbb{E}\left[e^{tX}\right]$?
2. Use the Markov inequality to prove a bound on $\mathbb{P}(e^{tX} \geq a)$ (here $a > 0$ is any positive number, while we assume $0 < t < \lambda$ as before).
3. Now reformulate this into a bound on $\mathbb{P}(X \geq b)$ (here $b > 0$ is any positive number, and again $0 < t < \lambda$).

Chapter 2

Joint Distributions

Recall that a random variable is defined as a function of the outcome of some random experiment or random process. If we define multiple functions from the *same* random process, then this gives us two or more random variables whose distributions and probabilities are linked. The joint distribution of X and Y, or of a list of random variables $X_1, \ldots, X_n$, refers to the characterization of the joint probabilities.

- In general we're interested in calculating probabilities of the form $\mathbb{P}((X, Y) \in A)$ where A is any "reasonable" region in $\mathbb{R}^2$.
- The CDF is now defined in a joint way:

  $$F(x, y) = \mathbb{P}(X \leq x, Y \leq y) \ .$$

  (There is an implicit "and" in the probability.) We might write $F_{X,Y}$ instead of F.
- We can also ask questions about one variable on its own, for example: $\mathbb{P}(X \leq 3)$ implicitly means $\mathbb{P}(X \leq 3, \text{ and } Y \text{ takes any value})$. This is called the marginal distribution of X, and its CDF can be calculated as

  $$F_X(x) = \mathbb{P}(X \leq x) = \lim_{y \to +\infty} \mathbb{P}(X \leq x, Y \leq y) = \lim_{y \to +\infty} F(x, y).$$

- We can also ask questions about the conditional distribution, such as $\mathbb{P}(X \geq 3 \mid Y \leq 7)$.
- For joint distributions of more than two variables, everything is analogous, e.g. $F(x_1, x_2, x_3) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, X_3 \leq x_3)$.

2.1 Discrete

If the pair (X, Y) takes only finitely many or countably infinitely many values, then we can characterize its distribution by the probability mass function

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

for every possible value (x, y) for the pair. The marginal distribution for X is then calculated as

$$p_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y p(x, y) \ ,$$

and we could also do the same for Y.

2.1.1 Examples

1. Survey 20 students chosen at random from this class. X = number of second years, Y = number of third years, in my sample. Assuming that there are 100 students, among whom there are 10 second years and 30 third years, we have
$$\mathbb{P}(X = k, Y = \ell) = \frac{\binom{10}{k} \cdot \binom{30}{\ell} \cdot \binom{60}{20-k-\ell}}{\binom{100}{20}} .$$
We could also calculate, for example,
$$\mathbb{P}(X = k \mid Y = \ell) = \frac{\mathbb{P}(X = k, Y = \ell)}{\mathbb{P}(Y = \ell)} = \frac{\frac{\binom{10}{k} \cdot \binom{30}{\ell} \cdot \binom{60}{20-k-\ell}}{\binom{100}{20}}}{\frac{\binom{30}{\ell} \cdot \binom{70}{20-\ell}}{\binom{100}{20}}}$$
Note: in the denominator — this is called the Hypergeometric distribution, it's the distribution of how many successes you get when you sample *without* replacement.

2.2 Continuous

If the pair (X, Y) is continuously distributed, then there is a joint density $f(x, y)$ which is piecewise continuous, nonnegative, and integrates to 1, such that
$$\mathbb{P}((X, Y) \in A) = \int\int_A f(x, y)\, dy\, dx$$
for any "reasonable" region $A \subset \mathbb{R}^2$. Comparing to the CDF,
$$F(x, y) = \mathbb{P}(X \le x, Y \le y) = \mathbb{P}((X, Y) \in (-\infty, x] \times (-\infty, y]) = \int_{s=-\infty}^{x} \int_{t=-\infty}^{y} f(s, t)\, dt\, ds$$
and taking the derivative with respect to x and y, we see that
$$\frac{\partial^2}{\partial x \partial y} F(x, y) = f(x, y) .$$
Now let's calculate the marginal density for X:
$$F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}((X, Y) \in (-\infty, x] \times (-\infty, \infty)) \int_{s=-\infty}^{x} \int_{y=-\infty}^{\infty} f(s, y)\, dy\, ds$$
and since we know that $f_X(x) = \frac{d}{dx} F_X(x)$, we get
$$f_X(x) = \int_{y=-\infty}^{\infty} f(x, y)\, dy .$$

2.3 Discrete / continuous / mixed?

Unlike for a single random variable, for a joint distribution for (X, Y) (or for $(X_1, \ldots, X_n)$), it might be the case that (X, Y) has zero mass at any single point (i.e. $\mathbb{P}((X, Y) = (x, y)) = 0$ for every $(x, y) \in \mathbb{R}^2$), but the distribution is still not continuous. For example, suppose that we sample a person at random, and X = the height of the person while $Y = \mathbf{1}_{\{\text{male}\}}$, a binary variable indicating the sex of the person. Then $\mathbb{P}((X, Y) = (x, y)) = 0$ for any (X, Y) because X itself (i.e. height) is continuously distributed, but since Y is discrete, there cannot be a joint density—all the probability of the pair, is on the two lines $\{y = 0\} \cup \{y = 1\}$, a zero-area region of $\mathbb{R}^2$.

It may even be the case that X and Y are each continuous, but (X, Y) is not. For example, let $X \sim N(0, 1)$ and let $Y = X^2$. Then all the probability of the pair is on the curve $\{y = x^2\}$, a zero-area region of $\mathbb{R}^2$, so there cannot be a joint density.

We will consider a pair (or n-tuple) to be continuous if and only if there is a joint density.

2.3.1 Examples

1. Survey 20 students chosen at random from this class. X = number of second years, Y = number of third years, in my sample. Assuming that there are 100 students, among whom there are 10 second years and 30 third years, we have
$$\mathbb{P}(X = k, Y = \ell) = \frac{\binom{10}{k} \cdot \binom{30}{\ell} \cdot \binom{60}{20-k-\ell}}{\binom{100}{20}}$$
with possible values = pairs (k, ℓ) such that $k \in \{0, \dots, 10\}$, $\ell \in \{0, \dots, 30\}$, and $k + \ell \leq 20$.

2. Let (X, Y) be sampled uniformly at random from the unit square. Then the density of (X, Y) is
$$f(x, y) = \begin{cases} 1, & 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$
Therefore, for any region $A \subset [0, 1]^2$,
$$\mathbb{P}((X, Y) \in A) = \int \int_A f(x, y)\, dy\, dx = \text{Area of } A.$$

3. **Bivariate normal distribution**: take any means (μ_1, μ_2) and any variances (σ_1^2, σ_2^2) and any correlation $\rho \in [-1, 1]$. Density:
$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right)\right\}.$$
We could calculate that, marginally, $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$.

4. For bivariate or multivariate distributions, it's no longer true that any distribution is either discrete or continuous or a mixture of discrete & continuous. As an example, we might have X discrete and Y continuous. Suppose that windspeed on cloudy days is distributed as Exponential(5), and on sunny days is distributed as Exponential(2). 40% of days are cloudy and 60% are sunny. Let X = windspeed and $Y = \mathbb{1}_{\text{Sunny}}$. Then we can calculate e.g.
$$\mathbb{P}(Y = 0 \mid X \geq 4)$$
as we've done before. We can also calculate the marginal distribution of X:
$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \leq x, Y = 0) + \mathbb{P}(X \leq x, Y = 1) = (1 - e^{-5x}) \cdot 0.4 + (1 - e^{-2x}) \cdot 0.6$$
and so, taking the derivative,
$$f_X(x) = (-(-5) \cdot e^{-5x}) \cdot 0.4 = (-(-2) \cdot e^{-2x}) \cdot 0.6 = 2e^{-5x} + 1.2e^{-2x}\ .$$

2.4 Independent random variables

Random variables X and Y (implied: defined on the same sample space, i.e. (X, Y) has a joint distribution) are independent if
$$F_{X,Y}(x, y) = F_X(x) F_Y(y)$$
for all values (x, y). We write this as $X \perp\!\!\!\perp Y$.

2.4.1 Properties

1. If (X, Y) is discrete, then equivalently, the PMF can be written as a product of a function of x and a function of y, i.e.
$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$
2. If (X, Y) is continuous, then equivalently, the density can be written as a product of a function of x and a function of y,
$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$
3. For any rectangular region $A = B \times C \subset \mathbb{R}^2$, we have
$$\mathbb{P}((X, Y) \in A) = \mathbb{P}(X \in B)\mathbb{P}(Y \in C).$$

2.4.2 Examples

1. Bivariate normal distribution: if we set correlation $\rho = 0$, then X and Y are independent, because the density $f(x, y)$ factors over x and y.
2. Uniform distribution on the unit square: X and Y are independent, and in fact each one has a marginal distribution that is Uniform$[0, 1]$.
3. Earlier example: if (X, Y) is sampled uniformly at random from the unit square, what is the distribution of $L = \sqrt{X^2 + Y^2}$?
$$F_L(t) = \mathbb{P}(L \le t) = \mathbb{P}(\sqrt{X^2 + Y^2} \le t) = \int\int_{\sqrt{X^2+Y^2} \le t} f(x, y)\, dy\, dx\ ,$$
which is equal to the area of the unit square intersected with circle of radius t.
4. Draw a hand of 10 cards. Let X = number of red cards and Y = number of Kings. Possible values (k, ℓ) where $k = 0, \ldots, 10$, $\ell = 0, 1, 2, 3, 4$, but not all values are possible. For example if $k = 10$ then we can get at most 2 kings. Specifically, we must have $\ell \le \min\{2, k\} + \min\{2, 10 - k\}$. This already tells us we can't have independence, since the range of X values depends on Y, and vice versa. In other words, it's not possible to have
$$p(k, \ell) = p_X(k)p_Y(\ell)$$
because we know that $p_X(10) > 0$ and $p_Y(4) > 0$ but $p(10, 4) = 0$.
General principle: range of possible (X, Y) values must be (range of X values)$\times$(range of Y values), under independence.

2.5 Conditional distributions

As discussed before, we often want to ask about probabilities of X based on some (partial) knowledge of Y, e.g. $\mathbb{P}(X \ge 3 \mid Y \ge 7)$. A conditional distribution is slightly more specific—it asks about the distribution of X given knowledge of the *exact* value of Y. This is a bit counterintuitive for the continuous case, where $Y = y$ has probability 0 for any specific value y we might ask about, but we'll see that we can construct a sensible definition nonetheless.

2.5.1 Discrete case

If we know that $Y = y$, what is the distribution of X?

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)} = \frac{p_{X,Y}(x, y)}{\sum_{x'} p_{X,Y}(x', y)}\ .$$

We write this as $p_{X|Y}(x \mid y)$, read as "the conditional distribution of X given $Y = y$". This defines a valid PMF, for example,

$$\sum_x p_{X|Y}(x \mid y) = \sum_x \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{\mathbb{P}(Y = y)}{\mathbb{P}(Y = y)} = 1 \ .$$

One way to think of this concept is that there could be a new random variable, whose distribution is that of X when we condition on $Y = y$.

2.5.2 Continuous case

To think about the continuous case, let's first take a look at (univariate) density in a slightly different way. Since density is piecewise continuous, it's continuous near x for nearly any value x. Now take a value x where $f(x)$ is continuous and take some small $\epsilon > 0$. Then

$$\mathbb{P}(x < X < x + \epsilon) = \int_{t=x}^{x+\epsilon} f(t)\, dt \approx \epsilon \cdot f(x) \ ,$$

since $f(t) \approx f(x)$ for all $t \in [x, x + \epsilon]$ by continuity (since ϵ is small). So, we should think of density $f(x)$ roughly as the probability that $X \approx x$, but rescaling to compensate for the tiny width of the interval.

Similarly, we can't ask questions like $\mathbb{P}(X = x \mid Y = y)$ since $\mathbb{P}(Y = y)$ is just zero. But instead, let's try to take small intervals instead of points:

$$\begin{aligned}\mathbb{P}(x < X < x + \epsilon \mid y < Y < y + \epsilon) = \frac{\mathbb{P}(x < X < x + \epsilon, y < Y < y + \epsilon)}{\mathbb{P}(y < Y < y + \epsilon)} &= \frac{\int_{u=x}^{x+\epsilon} \int_{v=y}^{y+\epsilon} f_{X,Y}(u, v)\, dy\, dx}{\int_{v=y}^{y+\epsilon} f_Y(y)\, dy} \\ &\approx \frac{\epsilon^2 f_{X,Y}(x, y)}{\epsilon f_Y(y)} = \epsilon \cdot \frac{f_{X,Y}(x, y)}{f_Y(y)} \ .\end{aligned}$$

So by analogy, we should define the conditional density as

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

and we should think of this as representing the probabability that $X \approx x$ when $Y \approx y$ (but rescaling to compensate for the tiny width of the interval).

Law of total probability For probabilities, given a partition $B_1, B_2, \ldots$ we had

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i) \ .$$

Similarly, we can show that

$$f_Y(y) = \int_{x=-\infty}^{\infty} f_{Y|X}(y \mid x) f_X(x)\, dx \ .$$

To check why,

$$\int_{x=-\infty}^{\infty} f_{Y|X}(y \mid x) f_X(x)\, dx = \int_{x=-\infty}^{\infty} \frac{f_{X,Y}(x, y)}{f_X(x)} f_X(x)\, dx = f_Y(y) \ ,$$

as we calculated for marginal densities.

Independence If (X, Y) is continuous, $X \perp\!\!\!\perp Y$ is equivalent to the statement $f_{X|Y}(x \mid y) = f_X(x)$ for all (x, y) (and, same for reversing X and Y).

Examples:

1. (X, Y) is chosen uniformly at random from the unit disk, $\{x^2 + y^2 \leq 1\}$. Then

$$f(x,y) = \begin{cases} \frac{1}{\pi}, & x^2+y^2 \leq 1, \\ 0, & \text{otherwise} \end{cases}$$

Then,

$$f_Y(y) = \int_{x=-\infty}^{\infty} f(x,y)\, dx = \int_{x=-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \frac{1}{\pi}\, dx = \frac{2\sqrt{1-y^2}}{\pi}\ .$$

And so, for (x, y) in the support,

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\frac{1}{\pi}}{\frac{2\sqrt{1-y^2}}{\pi}} = \frac{1}{2\sqrt{1-y^2}}\ .$$

What is the support of $X \mid Y$? It's the interval $[-\sqrt{1-y^2}, \sqrt{1-y^2}]$. In other words,

$$X \mid Y \sim \text{Uniform}[-\sqrt{1-Y^2}, \sqrt{1-Y^2}]$$

or equally we could write

$$X \mid Y = y \sim \text{Uniform}[-\sqrt{1-y^2}, \sqrt{1-y^2}]$$

2.6 Functions of jointly distributed random variables

2.6.1 Expectations

If $Y = g(X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ have a joint PMF p, then

$$\mathbb{E}(Y) = \sum_{(x_1,\ldots,x_n)} g(x_1,\ldots,x_n) \cdot p(x_1,\ldots,x_n)\ .$$

If instead $X_1, \ldots, X_n$ have a joint density f, then

$$\mathbb{E}(Y) = \int_{x_1} \cdots \int_{x_n} g(x_1,\ldots,x_n) \cdot f(x_1,\ldots,x_n)\, dx_n\ \ldots\ dx_1\ .$$

Alternatively we could compute the density of Y from the densities of the X_i's, but we will not cover this in the course (except in the univariate case, $Y = g(X)$).

Independence

1. We saw that densities, PMFs, and CDFs factor across the two (or more) variables.

2. Expectation: if $X \perp\!\!\!\perp Y$,

$$\begin{aligned}\mathbb{E}(X \cdot Y) = \int_x \int_y x \cdot y \cdot f_{X,Y}(x,y)\, dy\, dx &= \int_x \int_y x \cdot y \cdot f_X(x) f_Y(y)\, dy\, dx \\ &= \left(\int_x x \cdot f_X(x)\, dx\right) \cdot \left(\int_y y \cdot f_Y(y)\, dy\right) = \mathbb{E}(X) \cdot \mathbb{E}(Y).\end{aligned}$$

More generally, if $X \perp\!\!\!\perp Y$, then for any functions g, h

$$\mathbb{E}(g(X) \cdot h(Y)) = \mathbb{E}(g(X)) \cdot \mathbb{E}(h(Y))\ .$$

2.6.2 Examples

1. (X, Y) uniform in unit square, $L = \sqrt{X^2 + Y^2}$. Then density is $f(x, y) = 1$ on the unit square $[0, 1]^2$, so

$$\mathbb{E}(L) = \int_{x=0}^{1} \int_{y=0}^{1} \sqrt{x^2 + y^2}\, dy\, dx\ .$$

   If $A = XY$ is the area, then $\mathbb{E}(A) = \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) = 0.25$, since $X \perp\!\!\!\perp Y$ and each is Uniform$[0, 1]$. We could alternatively do $\int_{x=0}^{1} \int_{y=0}^{1} xy \cdot 1\, dy\, dx$.

2.7 Covariance and correlation (4.3)

For random variables (X, Y), the covariance is

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

and the correlation is

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

2.7.1 Properties

1. Under linear transformations: First, let's recall mean and variance:

$$\mathbb{E}(a + bX) = a + b\mathbb{E}(X)$$

   and

$$\mathrm{Var}(a + bX) = \mathbb{E}((a + bX - \mu_{a+bX})^2) = \mathbb{E}((a + bX - (a + b\mu_X))^2) = b^2\mathbb{E}((X - \mu_X)^2) = b^2\,\mathrm{Var}(X)\ .$$

   Then for covariance:

$$\begin{aligned}&\mathrm{Cov}(a + bX, a' + b'Y) = \mathbb{E}((a + bX - \mu_{a+bX}) \cdot (a' + b'Y - \mu_{a'+b'Y}))\\ &= \mathbb{E}((a + bX - (a + b\mu_X)) \cdot (a' + b'Y - (a' + b'\mu_Y))) = bb'\mathbb{E}((X - \mu_X)(Y - \mu_Y)) = bb'\,\mathrm{Cov}(X, Y).\end{aligned}$$

   And for correlation:

$$\mathrm{Corr}(a + bX, a' + b'Y) = \frac{\mathrm{Cov}(a + bX, a' + b'Y)}{\sigma_{a+bX}\sigma_{a'+b'Y}} = \frac{bb'\,\mathrm{Cov}(X, Y)}{|b|\sigma_X |b'|\sigma_Y} = \mathrm{sign}(bb')\,\mathrm{Corr}(X, Y)\ .$$

2. Sums:

$$\begin{aligned}\mathrm{Var}(X + Y) &= \mathbb{E}((X + Y - \mu_{X+Y})^2) = \mathbb{E}((X + Y - \mu_X - \mu_Y)^2)\\ &= \mathbb{E}((X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)\ .\end{aligned}$$

3. Independence: if $X \perp\!\!\!\perp Y$, take functions $g(x) = x - \mu_X$ and $h(y) = y - \mu_Y$. Then,

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mu_X) \cdot (Y - \mu_Y)) = \mathbb{E}(X - \mu_X) \cdot \mathbb{E}(Y - \mu_Y) = 0 \cdot 0 = 0\ .$$

4. Independence & variance: if $X \perp\!\!\!\perp Y$,

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)\ .$$

   More generally, if $X_1, \ldots, X_n$ are pairwise independent,

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\ .$$

2.7.2 Examples

1. If (X, Y) has a bivariate normal distribution with parameters $\mu_1, \mu_2, \sigma_X^2, \sigma_Y^2, \rho$, then $\mathrm{Corr}(X, Y) = \rho$.
2. Let $X \sim \text{Binomial}(n, p)$. We can write $X = X_1 + \cdots + X_n$ where $X_i = \mathbb{1}_{\text{success on } i\text{th trial}}$. Then since $X_1, \ldots, X_n$ are mutually independent,
$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n \cdot p(1-p) \ .$$
3. It's possible to have $\mathrm{Corr}(X, Y) = 0$ but $X \not\perp\!\!\!\perp Y$. An example: $X \sim N(0, 1)$ and $Y = S|X|$ and $Z = T|X|$ where S, T are random signs (i.e. $\mathbb{P}(S = +1) = \mathbb{P}(S = -1) = 0.5$ and same for T), with S, T, X drawn independently. Then
$$\begin{aligned}\mathrm{Cov}(Y, Z) = \mathbb{E}(YZ) - \mathbb{E}(Y)\mathbb{E}(Z) &= \mathbb{E}(S|X| \cdot T|X|) - \mathbb{E}(S|X|)\mathbb{E}(T|X|) \\ &= \mathbb{E}(S)\mathbb{E}(T)\mathbb{E}(X^2) - \mathbb{E}(S)\mathbb{E}(T)\mathbb{E}(|X|)^2 = 0\end{aligned}$$
since $\mathbb{E}(S) = \mathbb{E}(T) = 0$. However, it's clear that $Y \not\perp\!\!\!\perp Z$. For example, $\mathbb{P}(Y > 3 \mid Z > 3) = 0.5$ while $\mathbb{P}(Y > 3)$ is very small.

   We can also check that $Y \sim N(0, 1)$ (and by symmetry, same for Z).
   Note that this is an example of a joint distribution on the pair of random variables (Y, Z) where Y is normally distributed and Z is normally distributed but (Y, Z) is not bivariate normal.

2.8 Conditional expectations

For a joint distribution on (X, Y), we can ask about the distribution of X conditional on observing $Y = y$. Like any distribution, we can calculate its expected value. Intuitively, we should think of $\mathbb{E}(X \mid Y = y)$ as a long-run average:

- Imagine drawing (X, Y) from its joint distribution many times
- Now throw out all trials except those for which we got $Y = y$
- Among those trials, what is the average X value?

Of course, this intuition relies somewhat on Y being discrete, but we can imagine taking limits for the continuous case.

Formulas Discrete case:
$$\mathbb{E}(X \mid Y = y) = \sum_x x \cdot p_{X|Y}(x \mid y),$$
and more generally
$$\mathbb{E}(g(X) \mid Y = y) = \sum_x g(x) \cdot p_{X|Y}(x \mid y).$$
Continuous case:
$$\mathbb{E}(X \mid Y = y) = \int_x x \cdot f_{X|Y}(x \mid y)\, dx,$$
and more generally
$$\mathbb{E}(g(X) \mid Y = y) = \int_x g(x) \cdot f_{X|Y}(x \mid y)\, dx.$$
We can also define conditional variance,
$$\mathrm{Var}(X \mid Y = y) = \mathbb{E}((X - \mathbb{E}(X \mid Y))^2 \mid Y),$$
or, equivalently,
$$\mathrm{Var}(X \mid Y = y) = \mathbb{E}(X^2 \mid Y = y) - \mathbb{E}(X \mid Y = y)^2.$$

- Among all trials with $Y = y$, what is the variability among the corresponding X values?

2.8.1 Tower law

The tower law, a.k.a. Law of Total Expectation:

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

What do we mean by these multiple expectations? We can write this as

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

That is, to get $\mathbb{E}(Y \mid X)$ we are taking the expectation over the conditional distribution of Y given X. The answer might be a function of X; then we take the expectation of that function.

2.8.2 Examples

1. Mossel's dice paradox. You roll a die and stop the first time you get a 6. Conditional on the event that all rolls were even, what's the expected value of # of rolls?
   We'll do a related problem: X = # of rolls, Y = # of odd numbers observed.
   Hierarchical model:
   $$\begin{cases} X \sim \text{Geometric}(1/6) \\ Y \mid X = k \sim \text{Binomial}(k-1, 3/5) \end{cases}$$
   Then we have
   $$\mathbb{E}(Y \mid X = k) = \mathbb{E}(\text{Binomial}(k-1, 0.6)) = 0.6(k-1).$$
   Another way to write this is
   $$\mathbb{E}(Y \mid X) = 0.6(X-1)\ .$$
   Then,
   $$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y \mid X)) = \mathbb{E}(0.6(X-1)) = 0.6\mathbb{E}(X) - 0.6 = 0.6 \cdot 6 - 0.6 = 3.$$
   (Next to last step: $\mathbb{E}(X) = 6$, by doing the calculation via the Geometric PMF.)

2. Let (X, Y) be supported on the unit square $[0,1]^2$ with density
   $$f(x, y) = (x + y)$$
   on this region (we have checked that this integrates to 1).

   - Marginal density:
     $$f_X(x) = \int_{y=0}^{1} (x+y)\, dy = x + 1/2\ ,$$
     and same for Y.
   - Conditional density: for $(x, y) \in [0,1]^2$,
     $$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{x+y}{y+1/2}\ .$$
   - Conditional expectation:
     $$\mathbb{E}(X \mid Y = y) = \int_{x=0}^{1} x \cdot f_{X|Y}(x \mid y)\, dx = \int_{x=0}^{1} x \cdot \frac{x+y}{y+1/2}\, dx = \left[\frac{x^2/2 \cdot y}{y+1/2} + \frac{x^3/3}{y+1/2}\right]_{x=0}^{1} = \frac{y/2 + 1/3}{y+1/2}\ .$$

2.8.3 Law of total variance

$$\mathrm{Var}(Y) = \mathbb{E}(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(\mathbb{E}(Y \mid X)).$$

Intuitively, if Y has high variance, it comes from one of two sources:

- Either Y is highly variable even when you already know the value of X (term 1).
- Or if not, then your expected value of Y must change a lot as you vary X (term 2).

In particular, this implies that

$$\mathbb{E}(\mathrm{Var}(Y \mid X)) \leq \mathrm{Var}(Y).$$

That is, *on average*, after conditioning on X we have less variability in Y (as compared to variability without conditioning). So conditioning on X cannot increase our uncertainty, on average. However, it is possible that for some values x we would have $\mathrm{Var}(Y \mid X = x) > \mathrm{Var}(Y)$.

2.8.4 Examples

1. Return to dice paradox. Recall that $Y \mid X \sim \text{Binomial}(X - 1, 0.6)$. Then conditional on X, we know that

   $$\mathbb{E}(Y \mid X) = 0.6(X - 1)$$

   and

   $$\mathrm{Var}(Y \mid X) = (X - 1) \cdot 0.6 \cdot (1 - 0.6) = 0.24(X - 1).$$

   It is also known that a Geometric(p) r.v. has mean $1/p$ and variance $\frac{1-p}{p^2}$.

   $$\begin{aligned}\mathrm{Var}(Y) = \mathbb{E}(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(\mathbb{E}(Y \mid X)) &= \mathbb{E}(0.24(X-1)) + \mathrm{Var}(0.6(X-1)) \\ &= 0.24(\mathbb{E}(X) - 1) + 0.36\,\mathrm{Var}(X) = 0.24(6-1) + 0.36\frac{5/6}{1/36} = 12\ .\end{aligned}$$

2. We could also use these laws to calculate $\mathrm{Cov}(X, Y)$ and $\mathrm{Corr}(X, Y)$. We have

   $$\mathbb{E}(XY) = \mathbb{E}(\mathbb{E}(XY \mid X))\ .$$

   Working from the inside out,

   $$\mathbb{E}(XY \mid X) = X \cdot \mathbb{E}(Y \mid X) = 0.6X(X - 1).$$

   Why does this work? If we condition on a value of X, then the value of X becomes a constant—this might be clearer using alternate notation,

   $$\mathbb{E}(XY \mid X = k) = \mathbb{E}(k \cdot Y \mid X = k) = k \cdot \mathbb{E}(Y \mid X = k)\ .$$

   So now,

   $$\mathbb{E}(XY) = \mathbb{E}(\mathbb{E}(XY \mid X)) = \mathbb{E}(0.6X^2 - 0.6X) = 0.6\mathbb{E}(X^2) - 0.6\mathbb{E}(X)\ .$$

   It's known that $\mathbb{E}(X) = 1/p = 6$ and $\mathrm{Var}(X) = \frac{1-p}{p^2} = 30$, so $\mathbb{E}(X^2) = \mathrm{Var}(X) + \mathbb{E}(X)^2 = 66$. So,

   $$\mathbb{E}(XY) = 0.6\mathbb{E}(X^2) - 0.6\mathbb{E}(X) = 0.6 \cdot 66 - 0.6 \cdot 6 = 36.$$

   And,

   $$\mathbb{E}(X)\mathbb{E}(Y) = 6 \cdot 3 = 18\ .$$

   So,

   $$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 18\ .$$

   And,

   $$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{18}{\sqrt{30} \cdot \sqrt{12}} = 0.95.$$

2.9 Bivariate normal

2.9.1 Linear transformations of a bivariate normal

If (X, Y) is a bivariate normal r.v. with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$, then any linear combination of (X, Y) is a (univariate) normal.

If $Z = aX + bY$, then

$$\mu_Z = a\mathbb{E}(X) + b\mathbb{E}(Y) = a\mu_1 + b\mu_2$$

and

$$\begin{aligned}\sigma_Z^2 &= \mathrm{Var}(aX + bY) = \mathrm{Var}(aX) + \mathrm{Var}(bY) + 2\,\mathrm{Cov}(aX, bY)\\ &= a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)\\ &= a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\sigma_1\sigma_2\rho\ .\end{aligned}$$

If (X, Y) is bivariate normal with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$, then any 2D linear transformation of (X, Y) is a bivariate normal.

Now suppose that U and V are standard normal. Define

$$X = U, Y = \rho U + \sqrt{1-\rho^2}V.$$

In other words,

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix}.$$

Then

$$\mathbb{E}(X) = \mathbb{E}(U) = 0, \mathrm{Var}(X) = \mathrm{Var}(U) = 1$$

and

$$\mathbb{E}(Y) = \rho\mathbb{E}(U) + \sqrt{1-\rho^2}\mathbb{E}(V) = 0, \mathrm{Var}(Y) = \rho^2\,\mathrm{Var}(U) + (1-\rho^2)\,\mathrm{Var}(V) = 1$$

where the last step holds because $U \perp\!\!\!\perp V$. Finally,

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(XY) = \mathbb{E}(\rho U^2 + \sqrt{1-\rho^2}UV) = \rho\mathbb{E}(U^2) + \sqrt{1-\rho^2}\mathbb{E}(UV) = \rho\ .$$

Since (X, Y) is a linear transformation of (U, V), its joint distribution is bivariate normal. So,

$$(X, Y) \sim N(0, 0, 1, 1, \rho).$$

Similarly if we defined

$$X = \mu_1 + \sigma_1 U, Y = \mu_2 + \sigma_2(\rho U + \sqrt{1-\rho^2}V)\ ,$$

we'd get a bivariate normal

$$(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho).$$

2.9.2 Conditionals

Take a bivariate normal distribution,

$$(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho).$$

What is the conditional distribution of Y given X? We saw above that we can represent (X, Y) as a linear transformation of (U, V). Conditioning on $X = x$ is the same as conditioning on $U = \frac{x-\mu_1}{\sigma_1}$. Hence, conditional on $X = x$,

$$Y = \mu_2 + \sigma_2(\rho U + \sqrt{1-\rho^2}V) = \mu_2 + \sigma_2\Big(\rho \cdot \frac{x-\mu_1}{\sigma_1} + \sqrt{1-\rho^2}V\Big).$$

This has mean

$$\mathbb{E}(Y \mid X = x) = \mu_2 + \sigma_2\rho\frac{x-\mu_1}{\sigma_1}.$$

As a sanity check, we can compute that

$$\mathbb{E}\big(\mathbb{E}(Y \mid X)\big) = \mathbb{E}\Big(\mu_2 + \sigma_2\rho\frac{X-\mu_1}{\sigma_1}\Big) = \mu_2,$$

so that indeed $\mathbb{E}\big(\mathbb{E}(Y \mid X)\big) = \mathbb{E}(Y)$. The conditional variance can be computed in the same way as the conditional expectation:

$$\mathrm{Var}(Y \mid X = x) = \sigma_2^2(1-\rho^2).$$

We will state as a fact without proof that we get a normal distribution. In other words,

$$Y \mid X \sim N\Big(\mu_2 + \sigma_2\rho\frac{X-\mu_1}{\sigma_1}, \sigma_2^2(1-\rho^2)\Big).$$

Note that the conditional variance, does not depend on the value of X. This is not true for other distributions.

2.10 Rejection sampling (example D in 3.5)

Suppose we want to generate data from a desired distribution. If we know the CDF of the distribution that we want, F, then we saw before that we can generate $U \sim \text{Uniform}[0,1]$ and set $X = F^{-1}(U)$, then X will have the distribution with the CDF F.

However, in some cases, this might not be practical. For example, we might have a scenario where we know the density $f(x)$ but cannot calculate $F(x)$. Or, we might be in a multivariate setting. Or, we might know the density $f(x)$ up to the normalizing constant, but it is computationally infeasible to calculate the normalizing constant; this is not a real issue in one dimension (we can just use numerical integration) but can become a massive problem in high dimensions.

2.10.1 The method

Suppose we have a desired density $f(x) = \frac{g(x)}{C}$ for a known function g and a (potentially unknown) normalizing constant C, and we also have a density $h(x)$, for which we have a sampler, such that

$$h(x) \geq c \cdot g(x)$$

for some constant $c > 0$. Then we:
(1) Sample Y from the density $h(x)$.
(2) Draw $U \sim \text{Uniform}[0,1]$.
(3) Accept, and set $X = Y$ t if $U \leq c \cdot g(Y)/h(Y)$ and reject and go back to (1) otherwise.

Now let's check that the distribution of our resulting sample X indeed has the density $f(x)$. We have

$$\begin{aligned}
\mathbb{P}(x < X < x+\epsilon) &= \mathbb{P}(x < Y < x+\epsilon \mid U \leq cg(Y)/h(Y)) \\
&= \frac{\mathbb{P}(x < Y < x+\epsilon, U \leq cg(Y)/h(Y))}{\mathbb{P}(U \leq cg(Y)/h(Y))} \\
&\approx \frac{\mathbb{P}(x < Y < x+\epsilon, U \leq cg(x)/h(x))}{\mathbb{P}(U \leq cg(Y)/h(Y))} \\
&= \frac{\mathbb{P}(x < Y < x+\epsilon) \cdot \mathbb{P}(U \leq cg(x)/h(x))}{\mathbb{P}(U \leq cg(Y)/h(Y))} \\
&= \frac{h(x)\epsilon \cdot cg(x)/h(x)}{\mathbb{P}(U \leq cg(Y)/h(Y))} \\
&= \frac{c\epsilon g(x)}{\mathbb{P}(U \leq cg(Y)/h(Y))}.
\end{aligned}$$

And,

$$\mathbb{P}(U \leq cg(Y)/h(Y)) = \mathbb{E}(\mathbb{1}\{U \leq cg(Y)/h(Y)\}) = \mathbb{E}(\mathbb{E}(\mathbb{1}\{U \leq cg(Y)/h(Y)\} \mid Y)) = \mathbb{E}(cg(Y)/h(Y))$$

$$= \int_y \frac{cg(y)}{h(y)} \cdot h(y)\, dy = \int_y cg(y)\, dy = \int_y cCf(y)\, dy = cC\,.$$

So,

$$\mathbb{P}(x < X < x + \epsilon) \approx c\epsilon g(x)/(cC) = \epsilon f(x),$$

and so we see that $f(x)$ is the density of X.

2.10.2 Examples

1. Suppose that you want to sample from some region A in $\mathbb{R}^2$. You could build a larger rectangle R around A, sample uniformly from R (which is easy since it's just independent uniform distributions for X and Y), then reject any samples that did not fall into A. Here $f(x) = 1/area(A)$ on $x \in A$, and so we can use $g(x) = 1$ on $x \in A$ and set $c = 1/area(R)$ (the normalizing constant $C = area(A)$ can be unknown).

2. In high dimensions, suppose that $X_1, \ldots, X_n$ are all Bernoulli (values 0 or 1) but not independent. Specifically suppose that there is some set $S \subset \{0,1\}^n$, where $(X_1, \ldots, X_n)$ is uniformly distributed across this set. If it's easy to check for any point $(x_1, \ldots, x_n)$ whether it lies in S, but it's hard to calculate $|S|$, then rejection sampling is our best option: we sample uniformly from $\{0,1\}^n$ and reject all samples which didn't fall into the set S. For instance, S might be the set of all points $x \in \{0,1\}^n$ satisfying a list of linear inequalities. Then $f(x) = 1/|S|$ on $x \in S$, and so we can set $g(x) = 1$ on $x \in S$, with $c = 1/2^n$. Note $C = |S|$ can be unknown.

2.10.3 Practical considerations

A critical consideration is how efficient the sampler is. Specifically, if $\mathbb{P}(\text{accept})$ is low (most samples get rejected) then in order to get n draws of X we'd need to have roughly $\frac{n}{\mathbb{P}(\text{accept})} \gg n$ draws of Y. For that reason, we want to have $h(x)$ lie as close as possible to $f(x)$, in order to raise the chance of acceptance.

Exercises

Show all work and justify your answers!

Exercise 1 Consider the following example: $X \sim \text{Poisson}(100)$ is the number of photons emitted by an X-ray beam, and then Y is the number of photons that successfully pass through an object (i.e. the person being imaged). Suppose that the distribution of Y, if we know how many photons X were sent into the object, is given by $Y \mid X \sim \text{Binomial}(X, 0.4)$, i.e. given that X many photons are sent into the object, each one has a 40% chance of making it through and passing out the other side.

Calculate the probability mass function for (X, Y), that is, $p(k, \ell) = \mathbb{P}(X = k, Y = \ell)$ (as a function of k and ℓ).

Exercise 2 Let

$$f(x) = \begin{cases} cx^3 & 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

a) Find the value of c that makes f a density function. For the remaining of the exercise use that value of c and assume f is the density of X.
b) Find $\mathbb{P}(0 \le X \le 3/4)$.
c) Find the expected value of X.
d) Find the variance of X.

Exercise 3 Suppose a random circle is formed by assuming the radius is uniformly distributed between 0 and 1.

a) What is the expected area of the circle?
b) What is the variance of the area of the circle?

Exercise 4 Either Bach or Lizhen will grade your homework, each with probability 0.5. If Bach grades it, he will give it a score $10Y$, where Y is drawn from a Binomial distribution with $n = 10, p = 0.7$. If Lizhen grades it, she will give it a score drawn from the continuous uniform distribution on $(60, 100)$.

a) What is the probability that you get a score ≤ 30?
b) What is the probability that you get a score ≥ 95?

Exercise 5 Let X and Y be random variables with variances σ_X^2 and σ_Y^2, and correlation ρ. Calculate $\text{Cov}(X+Y, X-Y)$.

Exercise 6 In this question we will use a basic example to learn about Bayesian statistics, which models parameters with prior distributions (sometimes to indicate uncertainty in our beliefs). Suppose that you have a coin which may not be fair. Its parameter P, which is the chance of landing Heads, could in theory lie anywhere in the range $[0, 1]$. You flip this coin one time and let $X = \mathbb{1}_{\text{Heads}}$, and now would like to draw some conclusions about P. Let's choose a prior distribution for this parameter, and assume that P is drawn from a Uniform$[0, 1]$ distribution.

1. Write down a hierarchical model for this scenario, which should be of the form

$$\begin{cases} \text{(some variable)} \sim \text{(some distribution)} \\ \text{(some other variable)} \mid \text{(the first variable)} \sim \text{(some distribution)} \end{cases}$$

2. Now calculate the following: for any $t \in [0, 1]$, find $\mathbb{P}(P \le t, X = 0)$ and $\mathbb{P}(P \le t, X = 1)$. To do these calculations, you are working with a joint distribution where one variable is discrete and one is continuous; intuitive rules will apply for combining integrals and sums, etc. For example it's fine to write $\mathbb{P}(X = 1 \mid P = p) = p$.

3. Finally calculate the conditional distribution of P, given that you observe $X = 1$. To do this, start with the conditional CDF, i.e. $\mathbb{P}(P \le t \mid X = k)$, then get the density.

   In Bayesian statistics, the conditional distribution of P given our observed value of X, is called the posterior distribution for P—meaning its distribution after observing the data (which in this case is X, the data from tossing the coin).

Exercise 7 Let X and Y be random variables supported on $[0, 1] \times [0, 1]$, with joint density

$$f(x, y) = C \cdot (x^2 + y^2)$$

on this region. Here C is a constant.

1. Calculate C.
2. Calculate the marginal density $f_X(x)$.
3. Calculate the conditional density $f_{Y|X}(y|x)$ for $(x, y) \in [0, 1]^2$.
4. Are the variables X and Y positively correlated, negatively correlated, or uncorrelated? For this portion of the problem, it is not necessary to calculate $\text{Corr}(X, Y)$, $\text{Cov}(X, Y)$ or to do any exact calculations (although this would also be fine)—it is sufficient to examine your calculations for $f_{Y|X}(y|x)$ and explain what you see. It may help to plot $f_{Y|X}(y|x)$ for various values of x.

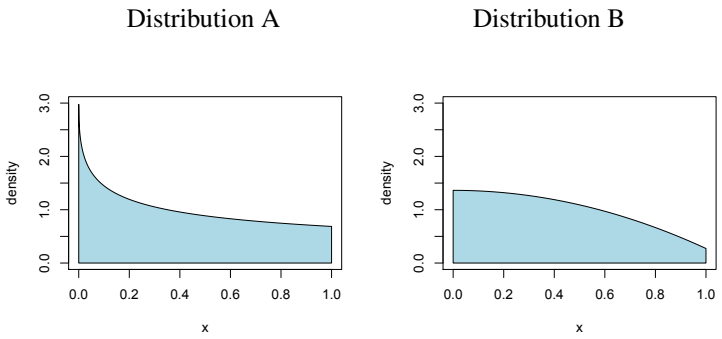
Exercise 8 Let U and V be independent Uniform$[0, 1]$ random variables.

1. Calculate $\mathbb{E}(U^k)$ where $k \geq 0$ is some fixed constant.
2. Calculate $\text{Cov}(UV, U + V)$.
3. Calculate $\mathbb{E}(V^U)$. (Hint: use the tower law and part 1.)

Exercise 9 Let X be uniform on $[0, 2]$. Conditional on $X = x$, let Y be uniform on $[0, x]$. Find the joint and marginal distributions of X and Y.

Exercise 10 Suppose a stick of length 1 is broken in two places, each chosen uniformly at random along the length of the stick. Find the probability that the three resulting pieces can be arranged to form a triangle (i.e. no piece is longer than the sum of the other two).

Exercise 11 Consider the following two density functions:



Consider two settings:

(1) You have access to samples from Distribution A, and you use rejection sampling to produce samples from Distribution B.

(2) You have access to samples from Distribution B, and you use rejection sampling to produce samples from Distribution A.

Which of these two implementations of rejection sampling will be more efficient, and which will be less efficient? Explain your answer thoroughly. You may use pictures to help explain your solution, but a picture alone without an explanation is not sufficient.

Exercise 12 Implement rejection sampling to obtain 1000 samples distributed with a Beta$(2, 2)$ distributions using samples from a Uniform$(0, 1)$ distribution. Plot a histogram with your 1000 samples and super-impose the density of a Beta$(2, 2)$ distribution. Include all the code you used.

Further Exercises

Exercise 1 Let the CDF of a random variable X be

$$F(x) = \begin{cases} 0 & x \leq 0, \\ \frac{x}{8} & 0 < x < 2, \\ \frac{x^2}{16} & 2 \leq x < 4, \\ 1 & x \geq 4. \end{cases}$$

a) Find the density function (PDF) of X.
b) Find $\mathbb{P}(1 \leq X \leq 3)$.
c) Find $\mathbb{P}(X \geq 1 | X \leq 3)$.

Exercise 2 Suppose that (X, Y) is a point chosen uniformly at random from the triangular region formed by connecting the points $(1, 0)$, $(0, 1)$ and $(0, -1)$.

1. Calculate $\mathbb{P}(X > 0.1 \mid Y > 0.1)$.
2. What is the CDF of the variable X?

Exercise 3

1. Let X and Y be random variables with $X \sim$ Exponential(λ_1) and $Y \sim$ Exponential(λ_2). Suppose that X and Y are independent. Let $Z = \max\{X, Y\}$. Calculate the CDF of Z.
2. Let $X \sim$ Exponential(1) and let $Y \sim$ Bernoulli(0.5). Again, X and Y are independent. Let $Z = X + Y$. Calculate the CDF of Z.

Exercise 4 Suppose that Y_1 and Y_2 are independent Poisson distributed random variables with means λ_1 and λ_2, respectively. Let $W = Y_1 + Y_2$. Use the fact that W has a Poisson distribution with mean $\lambda_1 + \lambda_2$ (you don't have to prove this fact) to show that the conditional distribution of Y_1, given that $W = w$, is a binomial distribution with $n = w$ and $p = \frac{\lambda_1}{\lambda_1 + \lambda_2}$.

Exercise 5 Let the **discrete** random variables Y_1 and Y_2 have joint probability distribution

$$p(y_1, y_2) = \frac{1}{3} \quad \text{for} \quad (y_1, y_2) = (-1, 1), (0, 1), (1, 0).$$

For example, there is a $\frac{1}{3}$ chance that $(Y_1, Y_2) = (-1, 1)$.
a) Find Cov(Y_1, Y_2).
b) Are Y_1 and Y_2 independent?

Chapter 3

Inference: Point Estimation (8.3, 8.5, etc.)

As noted in Chapter 2, probability theory provides the basis for statistical inference. Assuming a probability model for a population we can infer characteristics of that population from a random sample taken from it.

The type of problem we are considering here is the following: we have a family of distributions parametrized by θ—here θ might be one or more dimensional, e.g. θ could be (mean,variance) for the normal family. We observe data from the distribution and would like to estimate the parameter(s). Often we observe i.i.d. data points $X_1, \ldots, X_n$ drawn from the distribution. Some vocabulary & notation:

- The density or PMF for the distribution is often written as $f(x \mid \theta)$ or $p(x \mid \theta)$. This isn't really "conditional" in the true sense—we can think of θ as a fixed parameter value (in the frequentist framework), so we are not conditioning on θ strictly speaking. However if we think of θ as random (i.e. with a prior distribution / Bayesian), then this is a true conditional density/PMF.

- Any estimator of the parameters, $\hat{\theta}$, must be a function of $X_1, \ldots, X_n$ only—i.e. we must be able to calculate it using the data but not using the true value θ itself since that is unknown.

- The distribution of $\hat{\theta}$ (again treating θ as fixed and the X_i's as i.i.d. draws) is called the sampling distribution.

- The parameter θ may be multidimensional (or equivalently we can think of multiple parameters $\theta_1, \theta_2, \ldots$), but we won't cover theory for this case.

3.1 Sample mean & variance

Consider a sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where μ and σ^2 are unknown. How can we estimate the parameters μ and σ^2 from our sample?

Sample mean & sample variance:

$$\bar{X} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We can simplify a little bit:

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{2}{n-1}\sum_{i=1}^{n}X_i\bar{X} + \frac{1}{n-1}\sum_{i=1}^{n}\bar{X}^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{2n}{n-1}\bar{X}^2 + \frac{n}{n-1}\bar{X}^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{n}{n-1}\bar{X}^2. \end{aligned}$$

This is similar to $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$—we have $S^2 = \frac{n}{n-1} \cdot$ (sample mean of X^2 − (sample mean of $X)^2$).
Now let's check for "bias" in estimating μ and σ^2:

$$\mathbb{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mu$$

and

$$\mathbb{E}(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \frac{n}{n-1}\mathbb{E}(\bar{X}^2).$$

We have $\mathbb{E}(X_i^2) = \mathrm{Var}(X_i) + \mathbb{E}(X_i)^2 = \sigma^2 + \mu^2$. We also have $\bar{X} \sim N(\mu, \sigma^2/n)$ and so

$$\mathbb{E}(\bar{X}^2) = \mathrm{Var}(\bar{X}) + \mathbb{E}(\bar{X})^2 = \sigma^2/n + \mu^2 \ .$$

So,

$$\mathbb{E}(S^2) = \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left(\frac{1}{n}\sigma^2 + \mu^2\right)\right] = \sigma^2.$$

This explains dividing by $n-1$ in place of n.

Independence Remarkably, even though $\bar{X}$ and S^2 come from the same sample, they are independent when the data is normally distributed.

Let's check for the case $n = 2$. We have $\bar{X} = (X_1 + X_2)/2$ and

$$S^2 = \frac{1}{2-1}\sum_{i=1}^{2}(X_i - \bar{X})^2 = (X_1 - \frac{X_1+X_2}{2})^2 + (X_2 - \frac{X_1+X_2}{2})^2 = \frac{1}{2}(X_1 - X_2)^2$$

and we know that

$$\begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix}$$

is bivariate normal with covariance

$$\mathrm{Cov}(X_1+X_2, X_1-X_2) = \mathrm{Cov}(X_1,X_1) - \mathrm{Cov}(X_1,X_2) + \mathrm{Cov}(X_2,X_1) - \mathrm{Cov}(X_2,X_2) = \sigma^2 - 0 + 0 - \sigma^2 = 0.$$

Therefore, (see subsection 2.7.2) $X_1 + X_2 \perp\!\!\!\perp X_1 - X_2$ and so $\bar{X} \perp\!\!\!\perp S^2$.
Note also that $S^2 = \frac{1}{2}(X_1 - X_2)^2 \sim N(0, \sigma^2)^2$ since $X_1 - X_2 \sim N(0, 2\sigma^2)$.

3.2 Distributions derived from the normal distribution (6.2–6.3)

3.2.1 The χ^2 distribution

If $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} N(0,1)$ then the distribution of $V = Z_1^2 + \cdots + Z_n^2$ is called the χ^2_n distribution (χ^2 distribution with n degrees of freedom). It is actually a special case of the Gamma distribution: $\chi^2_n = \text{Gamma}(n/2, 1/2)$, so its density is

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

over $v \in (0, \infty)$. Calculations: for $V \sim \chi^2_n$,

$$\mathbb{E}(V) = \sum_{i=1}^{n} \mathbb{E}(Z_i^2) = n$$

and

$$\text{Var}(V) = \sum_{i=1}^{n} \text{Var}(Z_i^2) = 2n.$$

χ^2 distribution and sampling The centrality of the χ^2 distribution in statistics is largely due to the following result:

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Then

$$V := (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{3.1}$$

Important: the random variable V can be defined from the X_i without knowledge of μ and its distribution does not depend on μ.

3.2.2 The t distribution

If $Z \sim N(0,1)$ and $V \sim \chi^2_n$ and $Z \perp\!\!\!\perp V$, then the distribution of $T = Z/\sqrt{V/n}$ is called the t distribution with n degrees of freedom. Its density function can be calculated as

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}.$$

For a small n, the t_n distribution has heavy tails: $\mathbb{P}(T \geq x)$ is much larger than $1 - \Phi(x)$, as x grows large. For increasing n, the t_n distribution grows more similar to the normal distribution. We can see this in the density:

$$f(t) \propto \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} = \Bigg[\underbrace{\left(1 + \frac{t^2}{n}\right)^{n}}_{\to e^{t^2}}\Bigg]^{-\underbrace{(n+1)/2n}_{\to 1/2}} \to e^{-t^2/2}.$$

t distribution and sampling The centrality of the t distribution in statistics is largely due to the following result:

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Then

$$T := \sqrt{n}\frac{\bar{X} - \mu}{S} \sim t_{n-1}. \tag{3.2}$$

Important: the random variable T can be constructed from the X_i withouth knowledge of σ and its distribution does not depend on σ.

The result in equation (3.2) is a direct consequence of that in (3.1) and the independence of $\bar{X}$ and S^2. More precisely, rewrite

$$T = \sqrt{n}\frac{\bar{X}-\mu}{S} = \sqrt{n}\frac{\bar{X}-\mu}{\sigma\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\sqrt{n}\frac{\bar{X}-\mu}{\sigma}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{n-1}}},$$

and note that:

- $\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim N(0,1)$.
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.
- $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ and $\frac{(n-1)S^2}{\sigma^2}$ are independent because $\bar{X}$ is independent of S^2.

3.3 Method of moments

Let X be a r.v. and let $X_1, \ldots, X_n$ be i.i.d. with the same distribution as X. W need two definitions:

- $\mathbb{E}(X^k)$ is called the k-th moment of X.
- $\frac{\sum_{i=1}^n X_i^k}{n}$ is called the k-th sample moment of X.

Method of moments: to estimate k parameters equate the first k moments of X to the first k sample moments of X.

Uniform distribution in $(0,\theta)$. Let $X \sim U(0,\theta)$ with θ unknown.

Question: estimate θ using the method of moments from a sample $X_1, \ldots, X_n \sim U(0,\theta)$.

Answer: We have only one parameter (θ) to estimate, so $k = 1$. We thus equate the first moment of X to the first sample moment and solve for θ :

$$\mathbb{E}(X) = \sum_{i=1}^n X_i = \bar{X} \implies \frac{\theta}{2} = \bar{X} \implies \hat{\theta}_{\text{MoM}} = 2\bar{X}.$$

Gamma distribution with unknown rate λ and shape α parameters Now we need to estimate two parameters so $k = 2$. We equate the first moment to the first sample moment and the second moment to the second sample moment:

$$\mathbb{E}(X) = \bar{X}$$
$$\mathbb{E}(X^2) = \frac{\sum_{i=1}^n X_i^2}{n}.$$

Thus we get

$$\frac{\alpha}{\lambda} = \bar{X}$$
$$\mathbb{E}(X^2) = \text{Var}(X) + \mathbb{E}(X)^2 = \frac{\alpha}{\lambda^2} + \frac{\alpha^2}{\lambda^2}.$$

Solving for α and λ in the above system of two equations with two unknowns gives the method of moments estimator

$$\hat{\alpha}_{\text{MoM}} = \frac{\bar{X}^2}{\frac{\sum X_i^2}{n} - \bar{X}^2}, \qquad \hat{\lambda}_{\text{MoM}} = \frac{\bar{X}}{\frac{\sum X_i^2}{n} - \bar{X}^2}.$$

3.4 Maximum Likelihood Estimation

A common method is maximum likelihood estimation (MLE). From this point on let's use θ_0 for the true value of θ. The likelihood of θ given the data $X_1, \dots X_n$ is

$$\text{Likelihood}(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

i.e. the joint density (or PMF) of the data $X_1, \dots, X_n$ given the parameter θ.

Maximum likelihood estimation: choose as estimator the value of the parameters that maximizes the likelihood:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \prod_{i} f(X_i \mid \theta).$$

It is often more convenient to maximize the log-likelihood:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sum_{i} \log f(X_i \mid \theta).$$

MLE for mean of a normal distribution with known variance If $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then the density is $f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$. We have

$$\begin{aligned}
\hat{\mu} &= \arg\max_{\mu} \sum_{i} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i-\mu)^2/2\sigma^2} \right] \\
&= \arg\max_{\mu} \sum_{i} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - (X_i-\mu)^2/2\sigma^2 \right] \\
&= \arg\max_{\mu} \left[-\sum_{i} (X_i-\mu)^2 \right].
\end{aligned}$$

Taking the derivative of the term in brackets with respect to the parameter μ and equating to zero,

$$\sum_{i} 2(X_i - \mu) = 0$$

and so

$$\hat{\mu} = \bar{X},$$

the sample mean.

MLE for normal, unknown mean and variance If $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with μ, σ^2 unknown,

$$\begin{aligned}
(\hat{\mu}, \hat{\sigma}^2) &= \arg\max_{\mu,\sigma^2} \sum_{i} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i-\mu)^2/2\sigma^2} \right] \\
&= \arg\max_{\mu,\sigma^2} \sum_{i} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - (X_i-\mu)^2/2\sigma^2 \right] \\
&= \arg\max_{\mu,\sigma^2} \left[-\frac{n}{2}\log(\sigma^2) - \frac{\sum_i (X_i-\mu)^2}{2\sigma^2} \right].
\end{aligned}$$

Taking derivatives,

$$\frac{\partial}{\partial \mu} = -\frac{\sum_i 2(\mu - X_i)}{2\sigma^2}$$
$$\frac{\partial}{\partial \sigma^2} = -\frac{n}{2\sigma^2} - (-1) \cdot \frac{\sum_i (X_i - \mu)^2}{2(\sigma^2)^2}.$$

Setting $\frac{\partial}{\partial \mu}$ to zero, we get

$$\hat{\mu} = \bar{X}.$$

Plugging this in for μ when we set $\frac{\partial}{\partial \sigma^2}$ to zero, we get

$$0 = -\frac{n}{2\sigma^2} - (-1) \cdot \frac{\sum_i (X_i - \bar{X})^2}{2(\sigma^2)^2}$$

and so

$$\hat{\sigma}^2 = \frac{\sum_i (X_i - \bar{X})^2}{n}.$$

This is not exactly the same as S^2 from before — since S^2 divided by $n-1$ not by n. So we see that actually $\hat{\sigma}^2 = \frac{n-1}{n} S^2$ and so

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}(\bar{X}) = \mu, \qquad \mathbb{E}(\hat{\sigma}^2) = \frac{n-1}{n}\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2.$$

So this is nearly unbiased, but not quite.

3.5 Choice of estimators: bias, mean squared error (MSE), asymptotics

Throughout we let θ denote an unknown parameter and $\hat{\theta}$ be an estimator based on a sample $X_1, \ldots, X_n$.

3.5.1 Bias

The bias of $\hat{\theta}$ is defined by

$$\text{bias}(\hat{\theta}) := \mathbb{E}(\hat{\theta}) - \theta.$$

An estimator is called unbiased if $\text{bias}(\hat{\theta}) = 0$, that is, if $\mathbb{E}(\hat{\theta}) = \theta$.

3.5.2 Mean squared error

The mean squared error of $\hat{\theta}$ is defined by

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

If $\hat{\theta}$ is unbiased then this is equal to $\text{Var}(\hat{\theta})$. In general, we have the following simple but important result:

Theorem 7.

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2.$$

Proof.

$$\begin{aligned}\mathbb{E}((\hat{\theta}-\theta)^2) &= \mathbb{E}(([\hat{\theta}-\mathbb{E}(\hat{\theta})]+[\mathbb{E}(\hat{\theta})-\theta])^2) \\ &= \mathbb{E}([\hat{\theta}-\mathbb{E}(\hat{\theta})]^2)+\mathbb{E}([\mathbb{E}(\hat{\theta})-\theta]^2)+2\underbrace{\mathbb{E}([\hat{\theta}-\mathbb{E}(\hat{\theta})]\cdot[\mathbb{E}(\hat{\theta})-\theta])}_{=\mathbb{E}([\hat{\theta}-\mathbb{E}(\hat{\theta})])\cdot(\mathbb{E}(\hat{\theta})-\theta)=0} \\ &= \mathrm{Var}(\hat{\theta})+\mathrm{bias}(\hat{\theta})^2.\end{aligned}$$

□

This is the bias/variance tradeoff: sometimes we can reduce one at the cost of the other. For an extreme example, we could define $\hat{\theta} \equiv 0$, in which case $\mathrm{Var}(\hat{\theta}) = 0$ but bias is high.

Example Voltage reading is uniformly distributed over $(\theta, \theta+1)$, where θ is the true but uknown voltage. Let $X_1, \ldots, X_n$ be a random sample of independent readings, $X_i \sim U(\theta, \theta+1)$ and consider the estimator $\hat{\theta} = \bar{X}$ of θ. (This is a terrible choice of estimator, only chosen for the sake of illustrating the concepts.)

Q: What is the bias of $\hat{\theta}$?

A: We have that

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}(\bar{X}) = \theta + \frac{1}{2}$$

and so

$$\mathrm{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta = \frac{1}{2}.$$

Q: What is the mean squared error of $\hat{\theta}$?

A: Note that

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_i)}{n} = \frac{\theta+1-\theta}{12n} = \frac{1}{12n}.$$

Thus,

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{bias}(\hat{\theta})^2 = \frac{1}{12n} + \frac{1}{4}.$$

The more readings (that is, the larger n is) the smaller the MSE. However, the bias of $\hat{\theta}$ as an estimator of θ does not disappear by making more readings. In other words, in this example increasing the sample size reduces the variance of the estimator but not its bias.

Normal distribution $N(\mu, \sigma^2)$ with known σ. We have shown that the MoM and MLE for μ is $\hat{\mu} = \bar{X}$, the sample mean.This is unbiased, and

$$\mathrm{MSE} = \mathrm{Var}(\bar{X}) + 0 = \sigma^2/n.$$

Exponential distribution We saw that the MoM and MLE estimator of the rate λ of an exponential distribution is $\hat{\lambda} = \frac{1}{\bar{X}}$. We now investigate what is the bias and MSE of this estimator.

First note that $\mathbb{E}(X_i) = 1/\lambda$ for all i, so $\mathbb{E}(\bar{X}) = 1/\lambda$ meaning that in general we will *not* have $\mathbb{E}(1/\bar{X}) = \lambda$.

Some facts:

- If $X_i \sim \mathrm{Gamma}(\alpha_i, \lambda)$ independent for all i then $X_1 + \cdots + X_n \sim \mathrm{Gamma}(\alpha_1 + \cdots + \alpha_n, \lambda)$.
- For $Y \sim \mathrm{Gamma}(\alpha, \lambda)$, $\mathbb{E}\left(\frac{1}{Y}\right) = \frac{\lambda}{\alpha-1}$ as long as $\alpha > 1$ (otherwise not defined), and $\mathbb{E}\left(\frac{1}{Y^2}\right) = \frac{\lambda^2}{(\alpha-1)(\alpha-2)}$ as long as $\alpha > 2$.

So, $n \cdot \bar{X} \sim \mathrm{Gamma}(n, \lambda)$ and so

$$\mathbb{E}(\hat{\lambda}) = n \cdot \mathbb{E}\left(\frac{1}{n\bar{X}}\right) = \frac{n}{n-1}\lambda,$$

and

$$\mathbb{E}(\hat{\lambda}^2) = n^2 \cdot \mathbb{E}\left(\frac{1}{(n\bar{X})^2}\right) = \frac{n^2}{(n-1)(n-2)}\lambda^2.$$

This means that we have a slight bias. And the variance is

$$\mathrm{Var}(\hat{\lambda}) = \frac{n^2}{(n-1)(n-2)}\lambda^2 - \left(\frac{n}{n-1}\lambda\right)^2 = \lambda^2\frac{n^2}{(n-1)^2(n-2)}.$$

And so

$$\mathrm{MSE}(\hat{\lambda}) = \mathrm{Var}(\hat{\lambda}) + \mathrm{bias}(\hat{\lambda})^2 = \frac{\lambda^2 n^2}{(n-1)^2(n-2)} + \left(\frac{\lambda}{n-1}\right)^2.$$

We could try to reduce the bias. Let's define

$$\tilde{\lambda} = \hat{\theta} \cdot \frac{n-1}{n}$$

and then $\mathbb{E}(\tilde{\lambda}) = \frac{n-1}{n}\mathbb{E}(\hat{\lambda}) = \lambda$ so it's unbiased, and

$$\mathrm{Var}(\tilde{\lambda}) = \left(\frac{n-1}{n}\right)^2 \mathrm{Var}(\hat{\lambda}) = \frac{\lambda^2 n^2}{n^2(n-2)}.$$

So then we'd have a lower MSE,

$$\mathrm{MSE}(\tilde{\lambda}) = \mathrm{Var}(\tilde{\lambda}) + \mathrm{bias}(\tilde{\lambda})^2 = \frac{\lambda^2}{n-2}.$$

So in fact this reduces the bias and the variance! In other examples, there is a tradeoff in the sense that improving the bias may make the variance worse, and vice versa.

3.5.3 Efficiency

Compare $X \sim N(\theta, 1)$ with $\theta = -1$ or 0 or 1 vs $X \sim N(0.01\theta, 1)$ with same θ's. If the density $f(x \mid \theta)$ looks nearly identical, as a function of x, for the different values of θ, then we won't be able to distinguish whether our data was drawn from $f(x \mid \theta_1)$ vs $f(x \mid \theta_2)$. In other words, we need $f(x \mid \theta)$ to have a large nonzero derivative with respect to θ, across a range of typical x values. For mathematical reasons it's easier to work with log likelihood, i.e. we want $\frac{\partial}{\partial\theta}\log f(x \mid \theta)$ to be large (positive or negative), over the range of typical x values.

Fisher information To quantify the above, define Fisher information $I(\theta)$ as

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

(Expectation is taken with respect to $X \sim f(x \mid \theta)$; the equality holds under some regularity conditions.)

Theorem 8. *The Cramer-Rao Inequality and Lower Bound. Suppose $X_1, \ldots X_n$ form a random sample from the distribution with pdf $f(x;\theta)$. Subject to certain regularity conditions on $f(x;\theta)$, we have that for any unbiased estimator $\hat{\theta}$ for θ,*

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)},$$

where $I(\theta)$ is the Fisher Information about θ.

This is known as the Cramer-Rao lower bound and $\frac{1}{nI(\theta)}$ is the minimum variance achievable by any unbiased estimator of θ.

The larger $I(\theta)$ is, the more informative a typical observation is about the parameter θ, and the smaller the attainable variance of $\hat{\theta}$. Regularity conditions required to justify the exchange integration and differentiation in the proof include that the range of values of X must not depend on θ. An unbiased estimator $\hat{\theta}$ whose variance attains the Cramer-Rao lower bound is called **efficient**.

Example (Normal distribution) If $f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$ then

$$\frac{\partial^2}{\partial \mu^2} \log(f(x \mid \mu)) = \frac{\partial^2}{\partial \mu^2} \left[\text{constant} - \frac{(x-\mu)^2}{2\sigma^2} \right] = \frac{\partial}{\partial \mu} \left[\frac{2(x-\mu)}{2\sigma^2} \right] = -\frac{1}{\sigma^2}$$

so

$$I(\mu) = -\mathbb{E}\left(-\frac{1}{\sigma^2} \right) = \frac{1}{\sigma^2}.$$

Therefore, if $\hat{\mu}$ is any unbiased estimator for μ, the Cramer-Rao lower bound tells us that

$$\text{Var}(\hat{\theta}) \geq \frac{\sigma^2}{n}.$$

But, we know that $\bar{X}$ is unbiased for μ and its variance is σ^2/n. Hence, $\bar{X}$ is an efficient estimator for μ.

3.5.4 Consistency

We say that an estimator $\hat{\theta}_n$ of a parameter θ based on a sample of size n is consistent if, for any $\epsilon > 0$,

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) \xrightarrow{n \to \infty} 0.$$

If an estimator is consistent then the distribution of the estimator concentrates, as we obtain more and more samples, around the true unknown parameter θ. By the law of large numbers, the sample mean is a consistent estimator for the expected value. Under mild smoothness assumptions on $f(x|\theta)$ both the method of moments and MLE estimator for θ are consistent.

3.5.5 Asymptotic distribution of the MLE

For the normal sampling case, our MLE is $\hat{\mu} = \bar{X} \sim N(\mu, \sigma^2/n)$ — it is exactly normally distributed. One can rearrange this to show that $\frac{\bar{X}-\mu}{\sigma} \sim N(0,1)$. In other scenarios, it's very rare that we would be able to calculate an exact distribution result for $\hat{\theta} - \theta$ — but we may use approximations for the distribution of the MLE $\hat{\theta}$.

Fisher's theorem characterizes the asymptotic distribution of the MLE. Let $X_1, \ldots, X_n$ be i.i.d. draws from $f(x \mid \theta_0)$, with f satisfying some smoothness conditions. As $n \to \infty$, the distribution of $\hat{\theta}$ is approximately

$$\hat{\theta} \approx N\left(\theta_0, \frac{1}{nI(\theta_0)} \right). \tag{3.3}$$

More formally,

$$\sqrt{nI(\theta_0)} \cdot (\hat{\theta} - \theta_0) \to N(0,1)$$

(convergence in distribution, i.e. the CDF converges to Φ). We can also write

$$\sqrt{nI(\hat{\theta})} \cdot (\hat{\theta} - \theta_0) \to N(0,1)$$

which is useful since then the variance is no longer an unknown.

Note that equation (3.3) immediately implies that the MLE is (under smoothness conditions) *asymptotically* unbiased, and *asymptotically* efficient.

Remark: In the normal case, Fisher theorem gives $\hat{\mu} = \bar{X} \approx N(\mu, \sigma^2/n)$. In this case we know that this holds exactly, without approximation.

Exercises

Exercise 1
Note: This exercise concerns the multivariate Normal distribution that we have not covered in the lecture notes. As part of the exercise you may need to look up some definitions and figure out some details. This is deliberate on my side.
(i) Suppose X has distribution MVNn(μ,Σ) for positive definite Σ. Show that if A is an $n \times n$ matrix of full rank

$$Y = AX \sim \text{MVNn}(A\mu, A\Sigma A^T).$$

(ii) Now let A be orthogonal and suppose $Z_1, \ldots Z_n$ are i.i.d. $N(0,1)$ random variables. Show that $W = AZ$ consists of n i.i.d. standard normal random variables also. What is the distribution of $\sum_{i=2}^n W_i^2$?
(iii) Now consider A orthogonal as in (ii) with first row elements $A_{1i} = n^{-1/2},\ 1 \le i \le n$. Show that

$$\sum_{i=2}^n W_i^2 = \sum_{i=1}^n (Z_i - \bar{Z})^2.$$

(iv) Now suppose $X_1, \ldots, X_n$ are i.i.d. N(μ, σ^2) random variables. Calculate (using the above results) the distribution of $\sum_{i=1}^n (X_i - \bar{X})^2$ and show that it is independent of $\bar{X}$.

Exercise 2 Suppose X is a random variable with binomial distribution with parameters p and n. Find the bias and the MSE of the following estimators for p.

a) $\hat{p}_1 = X/n$.

b) $\hat{p}_2 = \frac{X+1}{n+2}$.

For which values of p is the MSE of $\hat{p}_1$ smaller (strictly) than the MSE of $\hat{p}_2$?

Exercise 3 Compute the asymptotic distribution of the MLE for λ, the parameter of an exponential distribution $f(x \mid \lambda) = \lambda e^{-\lambda x}$, $x > 0$. Is the MLE an unbiased estimator? Is it asymptotically unbiased?

Exercise 4 It is reasonable to assume the number of emails I receive in any given hour follows a Poisson distribution with parameter λ. In order to estimate the rate λ, I decide to check how many emails I got from 8am to 9am in the last 5 days. The results were $2, 7, 0, 5$, and 4. With these data compute the MLE for λ.

Exercise 5 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Uniform (a,b) with unknown a and b.

(i) Find the method of moment estimators of a and b.

(ii) Find the MLE for (a, b), show it is biased and find its bias.

Exercise 6. Let $Y_1, \ldots, Y_n$ be a random sample i.i.d. from the probability density function

$$f(y) = \begin{cases} \theta y^{\theta-1}, & 0 < y < 1, \\ 0, & \text{otherwise}, \end{cases}$$

where $\theta > 0$.

a) Find the bias for $\bar{Y}$ as an estimator for $\mathbb{E}[Y]$ (the expected value of the Y_i).

b) Find the MSE for $\bar{Y}$ as an estimator for $\mathbb{E}[Y]$ (the expected value of the Y_i).

c) Find the method of moments estimator for θ.

d) Find the MLE estimator for θ.

Exercise 7. Let $X_1, X_2, \ldots, X_n$ be i.i.d. with density function

$$f(x|\theta) = e^{-(x-\theta)}, \qquad x \ge \theta$$

and $f(x|\theta) = 0$ otherwise.

a) Find the method of moments estimator of θ.

b) Find the MLE of θ.

Exercise 8. Let $X_1, X_2, \ldots, X_n$ be i.i.d. Poisson(λ).

a) Does the MLE for λ achieve Cramer-Rao's lower bound?

b) Prove that the MLE is consistent for λ.

c) Find the asymptotic distribution of the MLE.

Exercise 9. In Exercise 3 you derived the asymptotic distribution of the MLE for the parameter λ of an exponential distribution. In this exercise you will validate your results empirically. Provide all code you use.

a) Choose $\lambda = 3$ and generate 10000 samples of 30 random variables $X_1, \ldots, X_{30}$ with distribution $f(x) = 3e^{-3x}$, $x > 0$.

b) For each of the 10000 samples, construct the MLE estimator for λ. This should give you 10000 MLE estimators for λ, each of them constructed using 30 independent draws with distribution $f(x) = 3e^{-3x}$, $x > 0$.

c) Plot a histogram with your 10000 estimators. Does the histogram agree with the asymptotic distribution of the MLE?

Exercise 10. a) Under the regularity conditions in the Cramer-Rao inequality, prove that there exists an unbiased estimator $\hat{\theta}(X)$ of θ whose variance attains the Cramer-Rao lower bound if and only if the score $\frac{\partial}{\partial\theta}\log f(X|\theta)$ can be expressed in the form

$$\frac{\partial}{\partial\theta}\log f(X|\theta) = I(\theta)\Big\{\hat{\theta}(X) - \theta\Big\},$$

or, equivalently, if and only if the function

$$\frac{\frac{\partial}{\partial\theta}\log f(X|\theta)}{I(\theta)} + \theta$$

does not depend on θ and is only dependent on X, in which case this statistic is the unbiased most efficient estimator of θ.

Hint: Cramer-Rao lower bound is attaind if and only if

$$\text{Corr}\left(\hat{\theta}(X), \frac{\partial}{\partial\theta}\log f(X|\theta)\right) = 1.$$

This can only happen if $\hat{\theta}(X)$ and $\frac{\partial}{\partial\theta}\log f(X|\theta)$ are linearly related random variables, i.e. if

$$\frac{\partial}{\partial\theta}\log f(X|\theta) = \alpha(\theta)\hat{\theta}(X) + \beta(\theta)$$

for some functions $\alpha(\theta)$ and $\beta(\theta)$ that do not depend on X.

b) Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with parameter λ. Find the Camer-Rao lower bound. Does there exist an unbiased estimator of λ whose variance is equal to the Cramer-Rao lower bound?

Further Exercises

Exercise 1 Suppose that X_1, X_2, X_3 are independent random variables from the same distribution with density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}.$$

a) We wish to construct an estimator for the parameter μ. Calculate the bias and mean squared error of the estimator $\hat{\mu} = \frac{3X_2+X_3}{4}$.

b) Find the likelihood function in terms of the three observations X_1, X_2, X_3 and use it to find the maximum likelihood estimator for μ.

Exercise 2 Suppose we have a circuit with voltage equal to μ. We have a voltage meter but its readings are uniformly distributed between μ and $\mu + 1$. Since we know the readings are greater than or equal to the true value, we decide to estimate μ by taking the minimum value of our readings.

a) Suppose we take two readings with our meter X_1 and X_2. What is the density function of our estimator $\hat{\theta} = \min(X_1, X_2)$?

b) What is $\mathbb{E}[\hat{\theta}]$?

Exercise 3 Compute the method of moments for λ, the parameter of an exponential distribution:

$$f(x|\lambda) = \lambda \exp(-\lambda x)$$

from a random sample of size n.

Exercise 4 Let $X_1, \ldots X_n$ be a sample from a normal distribution $f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$. Let $\hat{\mu}$ be the MLE estimator for μ. Show that $\hat{\mu}$ is an efficient estimator.

Chapter 4

Confidence Intervals

In this chapter we construct interval estimators (as opposed to point estimators) for an unknown parameter θ. These interval estimators are known as confidence intervals. The starting point, as always, is a sample $X_1, \ldots, X_n$ of i.i.d. random variables. We will focus on building confidence intervals for expected values and variances, that is, $\theta = \mathbb{E}(X_i)$ or $\theta = \mathrm{Var}(X_i)$. Confidence intervals are defined in terms of two point estimators (the end points of the interval): $(\hat{\theta}_L, \hat{\theta}_R)$. Thus, the end points of a confidence interval are random variables defined in terms of the sample. The intervals are constructed so that they have a certain (small) *level* α. The value of α represents how confident we are in our interval containing the unknown parameter θ. Precisely, the interval $(\hat{\theta}_L, \hat{\theta}_R)$ is said to have level α or to be a $1-\alpha$ confidence interval if

$$\mathbb{P}\Big(\theta \in (\hat{\theta}_L, \hat{\theta}_R)\Big) = 1 - \alpha. \tag{4.1}$$

Important: the parameter θ is a fixed unknown number. For any particular realization of the random experiment $(\hat{\theta}_L, \hat{\theta}_R)$ is a numeric interval (not random) that either contains or does not contain θ. However, if we repeat the random experiment zillions of times we obtain a corresponding zillion of numeric intervals $(\hat{\theta}_L, \hat{\theta}_R)$. The proportion of those intervals that will contain θ is (roughly) $1-\alpha$.

4.1 Confidence intervals for expected value of a normal sample

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ where μ and σ^2 are unknown (but the normal distribution is known/assumed to be true). We can estimate μ with the sample mean $\bar{X}$, but can we assess the accuracy of our estimate? We know that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

This means that we can calculate $\mathbb{P}\left(|\bar{X} - \mu| > \epsilon\right)$ for any ϵ, to see if it's likely that we are within ϵ of the true mean, i.e. if μ is in the interval $\bar{X} \pm \epsilon$. In particular suppose we want confidence level $1-\alpha$ (e.g. $\alpha = 0.05$). Let's take the tails of the t distribution and define the number $t_{\alpha/2}$ so that:

$$\mathbb{P}(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha, \qquad T \sim t_{n-1}.$$

To find this critical value $t_{\alpha/2}$ we use the CDF of the t_{n-1} distribution, given to us in table 4 of the book. Note that $t_{\alpha/2}$ depends on the desired confidence level α and on n. So,

$$\mathbb{P}\left(-t_{\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}\right) = 1 - \alpha$$

and

$$\mathbb{P}\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Note that $\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$ is a random interval; its center $\bar{X}$ and its width $t_{\alpha/2}\frac{S}{\sqrt{n}}$ both depend on the data. After generating the data and calculating $\bar{X}$ and S^2, this gives us a specific interval. It is no longer accurate to say that $\mathbb{P}(\mu \in \text{interval}) = 1 - \alpha$; the statement is either true or false (i.e. the event has already either occurred or not). However, we cannot see whether it's true or false; since before drawing the data we had a $1 - \alpha$ probability of constructing an interval for which the statement is true, we now say that we have $1 - \alpha$ confidence that the statement is true. This is a confidence interval for μ.

Remark More broadly, the interval

$$\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

is often used as an (approximate) $1 - \alpha$ confidence interval for the expected value of small populations (that is, when the sample size n is small).

4.2 Confidence intervals for variance of a normal sample

Let $X_1, \dots X_n \sim N(\mu, \sigma^2)$ with σ^2 unknown. We now that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Let $\chi^2_{L,n-1}$ and $\chi^2_{R,n-1}$ positive numbers such that

$$\mathbb{P}\left(\chi^2_{L,n-1} \le W \le \chi^2_{R,n-1}\right) = 1 - \alpha, \qquad W \sim \chi^2_{n-1}.$$

Then, replacing W by $\frac{(n-1)S^2}{\sigma^2}$ in the above, we obtain

$$\mathbb{P}\left(\chi^2_{L,n-1} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{R,n-1}\right) = 1 - \alpha.$$

Now, we massage until the parameter of interest σ^2 is on its own in the middle:

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{R,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{L,n-1}}\right) = 1 - \alpha.$$

Therefore

$$\left(\frac{(n-1)S^2}{\chi^2_{R,n-1}}, \frac{(n-1)S^2}{\chi^2_{L,n-1}}\right)$$

is a $1 - \alpha$ confidence interval for σ^2.

4.3 Confidence intervals for expected values, large sample size

Let now $X_1, \dots X_n$ be i.i.d. random variables with unknown expected value μ, and known variance σ^2. By the central limit theorem we know that, *approximately*,

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \sim N(0, 1).$$

Let $z_{\alpha/2}$ be such that

$$\mathbb{P}(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha \qquad Z \sim N(0, 1).$$

Then, by replacing Z by $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ we obtain that

$$\mathbb{P}\left(-z_{\alpha/2} \le \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \le z_{\alpha/2}\right) \approx 1 - \alpha.$$

The above is not an exact equality because of the approximation introduced through the use of the CLT. Now, rewriting the above so that μ (the parameter of interest) is in the middle, we obtain that

$$\mathbb{P}\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha.$$

Therefore,

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

is, approximately, a $1 - \alpha$ confidence interval for μ.

Remark: If σ is not known, then one may replace σ by S.

4.4 Confidence intervals using MLE

As before, we'll construct a confidence interval for θ_0 which is calculated using only observed values: picking z^* appropriately,

$$\mathbb{P}\left(\left|\sqrt{nI(\hat{\theta})} \cdot (\hat{\theta} - \theta_0)\right| \leq z^*\right) \approx 1 - \alpha$$

and in other words

$$\mathbb{P}\left(\hat{\theta} - z^*\frac{1}{\sqrt{nI(\hat{\theta})}} \leq \theta_0 \leq \hat{\theta} + z^*\frac{1}{\sqrt{nI(\hat{\theta})}}\right) \approx 1 - \alpha.$$

So, after observing the data, we have approximately $(1 - \alpha)$ confidence that θ_0 lies in this interval.

Examples For a 95% confidence interval ($\alpha = 0.05$) the critical z value is $z^* = 1.96$

1. For a normal distribution,
$$\hat{\mu} = \bar{X} \approx N(\mu, \sigma^2/n)$$
and so the 95% confidence interval for μ is
$$\bar{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}.$$

2. For exponential,
$$\hat{\lambda} = \frac{1}{\bar{X}} \approx N\Big(\lambda, \frac{\lambda^2}{n}\Big) \approx N\Big(\lambda, \frac{1}{\bar{X}^2 n}\Big)$$
and so the 95% confidence interval for λ is
$$\frac{1}{\bar{X}} \pm 1.96 \cdot \frac{1}{\bar{X} \cdot \sqrt{n}}.$$

4.5 Confidence intervals & multiple testing

Suppose that we are taking many different samples to answer a long list of questions: for each $j = 1, \ldots, p$, we measure a sample $X_{1,j}, \ldots, X_{n,j}$ which are i.i.d. from some distribution with mean μ_j and variance σ_j^2, and we build a confidence interval:

$$\text{95\% CI for } \mu_j : \bar{X}_j \pm t^* \cdot S_j/\sqrt{n}$$

where $\bar{X}_j$ and S_j are the sample mean & sample SD calculated for this question j. For example, we might be measuring gene expression level for p many genes, in a sample of patients with some particular disease. By definition

of confidence intervals, assuming that the data is not too far from normal and/or sample sizes n are large, and also the data sets for these p different questions is not overly dependent on each other, we should have

$$\mu_j \text{ lies in the confidence interval } \bar{X}_j \pm t^* \cdot S_j/\sqrt{n} \text{ for roughly 95\% of all the genes, i.e. } 0.95 \cdot p \text{ many genes}$$

Now, suppose we have a reference value for each μ_j, let's write μ_j^0, which is the average gene expression level for gene j among healthy patients. We're interested in finding which of the p genes (probably not too many of them) are associated with the disease, so we're interested in finding which genes j have means $\mu_j \neq \mu_j^0$. Suppose we find that m many genes appear to potentially have this difference, which we determine by checking whether the 95% CI for μ_j does *not* include the reference value μ_j^0. In general we would expect m to be much smaller than p — these m genes consist of (a) genes where truly $\mu_j \neq \mu_j^0$ and the sample revealed this difference; and (b) genes where actually $\mu_j = \mu_j^0$ but by random chance, $\bar{X}_j$ was unusually far from the population mean $\mu_j = \mu_j^0$ and so the CI does not contain $\mu_j = \mu_j^0$.

Traditionally, publications may sometimes report only these m results—the "significant" ones. Do we think that 95%, i.e. $0.95 \cdot m$ many, of these m genes have correct CIs, i.e. CIs containing the true population mean μ_j? No—in general, the percent will be much lower, because by subselecting the m "interesting" genes, we are biased towards picking exactly those genes where by random chance the sample mean landed far from the population mean.

This phenomenon is known as the multiple testing problem / problem of multiple comparisons. It arises whenever a statistical method designed for answering a single question is applied to a problem with multiple questions or with the potential for deciding between multiple options in your analysis.

4.6 Confidence intervals for difference of means

Suppose IQ among people who wear glasses is distributed $N(\mu_1, \sigma^2)$ and among people who don't wear glasses $N(\mu_2, \sigma^2)$, where μ_1, μ_2, σ^2 are unknown. We want to build a $1-\alpha$ confidence interval for the difference in average IQ between the two populations, based on two samples:

1. A sample $X_1^1, \ldots, X_{n_1}^1$ of IQs of people that wear glasses, so for $1 \leq i \leq n_1$ we have $X_i^1 \sim N(\mu_1, \sigma^2)$.
2. A sample $X_1^2, \ldots, X_{n_2}^2$ of IQs of people that don't wear glasses, so for $1 \leq i \leq n_2$ we have $X_i^2 \sim N(\mu_2, \sigma^2)$.

The way to construct a CI is similar as in the previous subsections, and we don't give the details. We define

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^1, \qquad \bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_i^2,$$

and note that $\bar{X}_1$ and $\bar{X}_2$ are estimators of μ_1 and μ_2, respectively. Similarly, we let S_1^2 and S_2^2 be the sample variances of $X_1^1, \ldots, X_{n_1}^1$ and $X_1^2, \ldots, X_{n_2}^2$, respectively, and define the *pooled variance*

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

Observe that S_1^2 and S_2^2 are estimators for σ^2, and the pooled variance S_p^2 combines these two estimators, weighing them accordingly to their sample size. It can then be shown that a $1-\alpha$ confidence interval for $\mu_1 - \mu_2$ is

$$\left(\bar{X}_1 - \bar{X}_2 - t^* S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \bar{X}_1 - \bar{X}_2 + t^* S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right),$$

where t^* is defined by

$$\mathbb{P}(-t^* \leq T \leq t^*) = 1-\alpha, \qquad T \sim t_{n_1+n_2-2}.$$

Exercises

Exercise 1 Suppose I want to estimate the average IQ in STAT 24410. I test 5 people from the class and the outcomes are $120, 115, 149, 112, 102$. (This is a small sample size.)

a) Give a 95% confidence interval (random interval) for the average IQ in STAT 24410.

b) Compute the sample variance S^2 with the five outcomes above, and give a numerical confidence interval for the average IQ.

c) Compute a 99% confidence interval for σ, the standard deviation of IQs in the class for the data above.

Exercise 2 We want to understand how much a typical UChicago student works. We poll 1000 students and find that the average number of hours spent studying per week is 40 hours with a standard deviation equal to 5. Give a 95% confidence interval for the average study time.

Exercise 3 A pollster would like to estimate the proportion of people who prefer cats over dogs. Suppose that the pollster wants an error of estimation less than 0.04 with probability 0.9. How many people do they need to poll? How is this influenced by the true value of p?

Exercise 4 In order to estimate the average number of hours slept by UChicago students the night before the final, I ask 1000 students how much they slept last night. On average they reported to have slept 7 hours, and the standard deviation was 2 hours. Find a confidence interval for the average number of hours of sleep with confidence coefficient 0.95. You can use that $z_{0.025} = 1.96$. Explain the choice of 0.025 briefly.

Chapter 5

Hypothesis Testing (9.1,9.2)

A hypothesis test is a procedure that allows us to “confidently" reject a hypothesis if it is clearly statistically inconsistent with data. We start with a non-mathy presentation to illustrate the ideas.

The typical setting for a hypothesis test is as follows:

1. Start with standing presupposition, steady state of affairs, etc. called the *null hypothesis* H_0.
2. Collect data or evidence, perform experiments.
3. Reject the null hypothesis if the data is *clearly* against H_0. Don’t reject H_0 otherwise.

Example 1 (A Trial). Let’s consider a trial.

1. The starting presupposition (null hypothesis) is that the defendant is not guilty.
2. Collect evidence, call witnesses, etc.
3. Reject the null hypothesis (that is, reject the hypothesis that the defendant is not guilty) if there is evidence against the defendant “beyond reasonable doubt". Otherwise they don’t reject the null hypothesis.

It is important to note that the jury can make two *different* types of mistake: decide that an innocent person is guilty (reject the null hypothesis when it is true), or decide that a guilty person is innocent (not reject the null hypothesis when it is false). These two different types of mistake are called Type I error and Type II error, respectively. □

Hypothesis tests are designed to avoid rejecting H_0 when it’s true. Therefore, when the test rejects H_0 one can be quite sure that H_0 is false. This motivates the following reasoning, widely used in science:

Example 2. Suppose a scientist wants to show that a drug is more effective than a placebo. They may proceed as follows:

1. Set H_0 to be the hypothesis that the drug is undistinguishable from the placebo.
2. Collect data: have people take the drug and the placebo and check the results.
3. Reject H_0 is there is strong evidence against it.

If the null hypothesis H_0 is rejected we can be sure that the drug is effective. □

5.1 Basic concepts

5.1.1 The four elements of a hypothesis test

A hypothesis test is made of four elements:

1. A null hypothesis, denoted H_0.
2. An alternative hypothesis, often denoted H_1 or H_A. The null and the alternative hypothesis play asymmetric roles.
3. A test statistic (think of it as an estimator, i.e. as a way to combine your random variables similarly as we did in previous chapters).
4. A rejection region.

The procedure is as follows: if the test statistic lies in the rejection region, then we reject H_0. In such a case we support the alternative hypothesis.

The distribution of the test statistic under the assumption that H_0 is true is called the null distribution.

5.1.2 The two type of errors

There are two types of error:

- Type I error: rejecting H_0 when it is true. Avoiding this type of error is the priority. The probability of Type I error is usually denoted by α, and referred to as the "significance level" of the test.
- Type II error: not rejecting H_0 when H_0 is false. The probability of this error is usually denoted by β.

It is in principle possible to avoid Type I error completely by never rejecting the null hypothesis, but that will lead to large Type II error. The way one typically operates is: i) fixing a tolerance to Type I error that one is willing to allow for (e.g. $\alpha = 0.05$); and ii) designing a test that has the smallest possible Type II error.

The **power** of a test is defined as $1 - \beta$. We want to have powerful tests with a given Type I error α.

Example 3. This example is meant to illustrate some of the above concepts, without suggesting any general theory yet.

We suspect a coin may be biased towards heads. Let p be its (unknown) probability of heads. We set

$$H_0 : p = 0.5, \qquad H_1 : p > 0.5.$$

In order to check if we can reject H_0 we will flip the coin 10 times and let the outcomes be expressed in terms of r.v.s. $X_1, \ldots, X_{10}$ where, for $i = 1, \ldots, 10$,

$$X_i = \begin{cases} 1 & \text{if the i-th toss is heads,} \\ 0 & \text{if the i-th toss is tails,} \end{cases}$$

and so $X_i \sim$Bernoulli(p). We may choose $\hat{p} = \bar{X}$ as our test statistic, and we may choose as our rejection region $RR = \{x : x \geq 0.7.\}$ Then we will reject H_0 if $\bar{X} \geq 0.7$. Let us compute the Type I error:

$$\begin{aligned} \alpha = \mathbb{P}(\text{Type I}) &= \mathbb{P}(\bar{X} \geq 0.7 \mid p = 0.5) \\ &= \mathbb{P}\Big(\sum_{i=1}^{10} X_i \geq 7 \mid p = 0.5\Big) \\ &= \sum_{k=7}^{10} \binom{10}{k} 0.5^k 0.5^{10-k}, \end{aligned}$$

where for the last equality we used that $\sum_{i=1}^{10} X_i \sim$Binomial$(10, p)$, and $p = 0.5$ under the null hypothesis. □

In order to provide more intuition let us check how α changes if we change the rejection region. Precisely, suppose that we have been given H_0, H_A and $\hat{\theta}$ and we consider two rejection regions RR_1 and RR_2 for a test with $RR_1 \subset RR_2$. If α_1 and α_2 are the associated Type I errors, what can we say about α_1 and α_2?

Note that since $RR_1 \subset RR_2$

$$\begin{aligned}\alpha_1 &= \mathbb{P}(\hat{\theta} \in RR_1 | H_0 \text{ true}) \\ &\leq \mathbb{P}(\hat{\theta} \in RR_2 | H_0 \text{ true}) = \alpha_2.\end{aligned}$$

This shows that we can always reduce the Type I error by making the rejection region smaller. This will be typically at the expense of larger Type II error.

5.2 Tests for expected value, large sample

Set-up $X_1, \ldots, X_n$ i.i.d. with expected value $\mathbb{E}(X_i) = \mu$ unknown, and $\mathrm{Var}(X_i) = \sigma^2$ known. Large sample size: $n \geq 30$.

Remark 1. If the variance σ^2 is unknown you can replace it in what follows by S^2. There is no much harm in doing so, since n is large and so S^2 approximates σ^2 well.

Aim: To test the hypothesis

$$H_0 : \mu = \mu_0$$

against one of these three alternative hypothesis:

1. $H_1 : \mu > \mu_0$; or
2. $H_1 : \mu < \mu_0$; or
3. $H_1 : \mu \neq \mu_0$.

We want to have (at least approximately) $\mathbb{P}(\text{Type I error}) = \alpha$, where α is a given tolerance to Type I error.

Weapons:

- By the Central Limit Theorem, since n is large we have that, approximately,

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \sim N(0, 1).$$

- The definition of z-score, i.e. if $Z \sim N(0, 1)$ then $z_{\alpha/2}$ is defined so that

$$\mathbb{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

Procedure: Choose $\bar{X}$ as your test statistic. Look for a rejection region of the form

1. $RR_1 = \{x > k_1\}$ for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.
2. $RR_2 = \{x < k_2\}$ for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$.
3. $RR_3 = \{|x - \mu_0| > k_3\}$ for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.

The numbers k_1, k_2 and k_3 will be chosen so that the test has $\alpha = \mathbb{P}(\text{Type I})$.

Implementation: We focus on testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$. The other two cases are analogous.

From the definition of the z-score we have that if $Z \sim N(0, 1)$ then

$$\alpha = \mathbb{P}(Z > z_\alpha). \tag{5.1}$$

But also, since we want α to be our probability of Type I error we may set

$$\alpha = \mathbb{P}(\text{Type I}) = \mathbb{P}(\bar{X} > k_1 | \mu = \mu_0) = \mathbb{P}\Big(\sqrt{n}\,\frac{\bar{X} - \mu_0}{\sigma} > \sqrt{n}\,\frac{k_1 - \mu_0}{\sigma}\Big|\mu = \mu_0\Big), \tag{5.2}$$

where by the CLT we have that under H_0, approximately, $\sqrt{n}\,\frac{\bar{X}-\mu_0}{\sigma} \sim N(0,1)$.
Therefore, combining (5.1) and (5.2) we obtain that

$$\sqrt{n}\,\frac{k_1-\mu_0}{\sigma} = z_\alpha,$$

that is,

$$k_1 = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

In this way, we derive the following hypothesis tests:

1. $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.
   Test statistic: $\bar{X}$.
   Rejection region $RR_1 = \{x : x > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\}$.
   Reject H_0 if $\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$.
2. $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$.
   Test statistic: $\bar{X}$.
   Rejection region $RR_2 = \{x : x < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}\}$.
   Reject H_0 if $\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$.
3. $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.
   Test statistic: $\bar{X}$.
   Rejection region $RR_3 = \{x : |x-\mu_0| > z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\}$.
   Reject H_0 if $|\bar{X}-\mu_0| > z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

Note that each of the three tests above can be defined equivalently by changing both the test statistic and the rejection region. Precisely, we can rewrite them as follows:

1. $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.
   Test statistic: $\sqrt{n}\frac{\bar{X}-\mu_0}{\sigma}$.
   Rejection region $RR_1 = \{x : x > z_\alpha\}$.
   Reject H_0 if $\sqrt{n}\frac{\bar{X}-\mu_0}{\sigma} > z_\alpha$.
2. $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$.
   Test statistic: $\sqrt{n}\frac{\bar{X}-\mu_0}{\sigma}$.
   Rejection region $RR_2 = \{x : x < -z_\alpha\}$.
   Reject H_0 if $\sqrt{n}\frac{\bar{X}-\mu_0}{\sigma} < z_\alpha$.
3. $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.
   Test statistic: $\sqrt{n}\frac{\bar{X}-\mu_0}{\sigma}$.
   Rejection region $RR_3 = \{x : |x| > z_{\alpha/2}\}$.
   Reject H_0 if $|\sqrt{n}\frac{\bar{X}-\mu_0}{\sigma}| > z_{\alpha/2}$.

Remark 2. The same idea applies in any scenario where there is guarantee that the test statistic is approximately Gaussian. The most notable example is the hypothesis test for the difference of expected values of two populations with large sample sizes.

5.3 Small sample test for expected values

In the exact same way as in the previous section we can construct the following tests for the small sample size case $n < 30$. We skip the details of the derivation.

Set-up $X_1, \ldots, X_n$ i.i.d. with expected value $\mathbb{E}(X_i) = \mu$ unknown, and unknown or known variance. Small sample size: $n < 30$.

1. $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.
   Test statistic: $\sqrt{n}\frac{\bar{X}-\mu_0}{S}$.
   Rejection region $RR_1 = \{x : x > t_{\alpha,n-1}\}$.
   Reject H_0 if $\sqrt{n}\frac{\bar{X}-\mu_0}{S} > t_{\alpha,n-1}$.
2. $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$.
   Test statistic: $\sqrt{n}\frac{\bar{X}-\mu_0}{S}$.
   Rejection region $RR_2 = \{x : x < t_{\alpha,n-1}\}$.
   Reject H_0 if $\sqrt{n}\frac{\bar{X}-\mu_0}{S} < t_{\alpha,n-1}$.
3. $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.
   Test statistic: $\sqrt{n}\frac{\bar{X}-\mu_0}{S}$.
   Rejection region $RR_3 = \{x : |x| > t_{\alpha/2,n-1}\}$.
   Reject H_0 if $|\sqrt{n}\frac{\bar{X}-\mu_0}{S}| > t_{\alpha/2,n-1}$.

In the above $t_{\alpha/2,n-1}$ is defined so that if $T \sim t_{n-1}$ then

$$\mathbb{P}(-t_{\alpha/2,n-1} \leq T \leq t_{\alpha/2,n-1}) = 1 - \alpha.$$

5.4 Tests for variance of a normal distribution

Set-up $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with σ^2 unknown.

1. $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma > \sigma_0$.
   Test statistic: $(n-1)S^2/\sigma_0^2$.
   Rejection region $RR_1 = \{x : x > \chi^2_{\alpha,R,n-1}\}$.
   Reject H_0 if $(n-1)S^2/\sigma_0^2 > \chi^2_{\alpha,R,n-1}$.
2. $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma < \sigma_0$.
   Test statistic: $(n-1)S^2/\sigma_0^2$.
   Rejection region $RR_2 = \{x : x < \chi^2_{\alpha,L,n-1}\}$.
   Reject H_0 if $(n-1)S^2/\sigma_0^2 < \chi^2_{\alpha,L,n-1}$.
3. $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma \neq \sigma_0$.
   Test statistic: $(n-1)S^2/\sigma_0^2$.
   Rejection region $RR_3 = \{x : x < \chi^2_{\alpha/2,L,n-1}\} \cup \{x : x > \chi^2_{\alpha/2,R,n-1}\}$.
   Reject H_0 if $(n-1)S^2/\sigma_0^2 < \chi^2_{\alpha/2,L,n-1}$ or $(n-1)S^2/\sigma_0^2 > \chi^2_{\alpha/2,R,n-1}$.

In the above $\chi^2_{\alpha,L,n-1}$ and $\chi^2_{\alpha,R,n-1}$ are numbers defined so that, if $W \sim \chi^2_{n-1}$, then

$$\mathbb{P}(W \leq \chi^2_{\alpha,L,n-1}) = \alpha,$$

and

$$\mathbb{P}(W \geq \chi^2_{\alpha,R,n-1}) = \alpha.$$

5.5 P-values & significance testing

P-values / significance testing is a method of testing H_0 against a (perhaps unspecified) alternative hypothesis, by fixing the Type I error rate. The P-value is the smallest α for which the given *observed* data (once you have done the random experiment) suggests rejection of H_0. Intuitively, a P-value tells you the probability, under the null, of observing an outcome that is more "extreme" than the data you observed.

To put this concept in a concrete setting, let's go back to Example 3:

Example 4. Consider once again testing whether a coin unbiased against the alternative hypothesis that it is biased towards heads. We had set:

$$H_0 : p = 0.5, \qquad H_1 : p > 0.5.$$

We flip the coin 10 times and let X be the number of heads. Suppose I flip the coin ten times and observe a total number of heads $X_1 + \cdots + X_{10} = 7$.
Question: Find the P-value for a test with RR of the form $\{x \geq k\}$ and test statistic $\bar{X}$.
Answer: Note that the observed value of the test statistic is $\bar{X} = 0.7$.

$$\begin{aligned}\text{P-value} &= \mathbb{P}(\bar{X} \geq 0.7 | p = 0.5) \\ &= \mathbb{P}(X_1 + \cdots + X_{10} \geq 7 | p = 0.5) = \sum_{k=7}^{10} 0.5^k 0.5^{10-k} = 0.178,\end{aligned}$$

since under the null hypothesis $X = X_1 + \cdots + X_{10} \sim$Binomial(10,0.5). □

To be a bit more general, suppose that we have a test statistic $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ and a rejection region of the form $\{x : x > k\}$. Suppose that we have observed $X_1 = x_1, \ldots, X_n = x_n$ where $x_1, \ldots, x_n$ are the numerical values that the random variables $X_1, \ldots, X_n$ took on a realization of our random experiment. Then we could define a p-value p

$$\begin{aligned}p &= \mathbb{P}(\hat{\theta}(X_1, \ldots, X_n) \geq \hat{\theta}(x_1, \ldots, x_n) | H_0) \\ &= 1 - F^-_{H_0}\big(\hat{\theta}(x_1, \ldots, x_n)\big),\end{aligned}$$

where F_{H_0} is the CDF of $\hat{\theta}$ according to H_0, the superscript $^-$ means to not include the endpoint in this CDF, and $\hat{\theta}(x_1, \ldots, x_n)$ is the actual value that the test statistic took on a realization of your random experiment.

5.6 Comparing simple hypotheses: likelihood ratio tests

A simple hypothesis is a hypothesis that fully specifies the distribution of the data. For example, suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, 1)$. Then we might have

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu = \mu_1.$$

In contrast, a composite hypothesis is anything that is less specific. For example,

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0.$$

Here H_0 is simple while H_1 is composite.

Suppose that our data is distributed as $X \sim f(x \mid \theta)$. We will often be thinking of i.i.d. sampling, i.e. $X = (X_1, \ldots, X_n)$ and $f(x \mid \theta) = f(x_1 \mid \theta) \times \cdots \times f(x_n \mid \theta)$, but we will use the more general notation now.

Suppose we believe that the true parameter θ has one of two values, θ_0 or θ_1, and we'd like to use the data to help decide which is more likely. We can write this as a test between two hypotheses:

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta = \theta_1.$$

How should we decide which is more likely? One way is to look at their likelihoods:

$$L(\theta_0) = \text{Likelihood}(\theta_0 \mid X) = f(X \mid \theta_0)$$

and

$$L(\theta_1) = \text{Likelihood}(\theta_1 \mid X) = f(X \mid \theta_1).$$

It makes sense that if $L(\theta_0)$ is very large and $L(\theta_1)$ is very small, then we would be quite certain that θ_0 is the right value of the parameter (and vice versa).

For comparing two simple hypotheses, we consider the *likelihood ratio* as a test statistic:

$$LR = \frac{\prod_i f(X_i \mid \theta_0)}{\prod_i f(X_i \mid \theta_1)},$$

and we reject H_0 if $LR \leq c$ for some predetermined threshold c.
In this case,

$$\mathbb{P}(\text{Type I error}) = \mathbb{P}\left(\frac{\prod_i f(X_i \mid \theta_0)}{\prod_i f(X_i \mid \theta_1)} \leq c \,\middle|\, \theta = \theta_0\right) = \int \ldots \int_{(x_1,\ldots,x_n) \text{ s.t. } LR \leq c} \prod_i f(x_i \mid \theta_0)\, dx_n \ldots dx_1.$$

For testing simple hypothesis, likelihood ratio tests are optimal in the sense that they are the most powerful tests. This is the content of the Neyman-Pearson lemma:

Neyman-Pearson lemma Suppose we are testing H_0 against H_1 and both are simple hypotheses. Let c be any threshold and let α = Type I error, β = Type II error, for the LR test with that threshold. Then for any other test with Type I error = α, its Type II error is $\geq \beta$. In other words, the LR test is the most powerful test, at any given Type I error level.

Example

1. Suppose we have one draw of a random variable $X \sim N(\mu, \sigma^2)$.

$$H_0 : X \sim N(1,1), H_1 : X \sim N(2,2).$$

- One choice of test might be to have a threshold $c \in [1,2]$. If $X \geq c$ we reject H_0. Then

$$\mathbb{P}(\text{Type I error}) = \mathbb{P}\big(X \geq c | X \sim N(1,1)\big) = 1 - \Phi\left(\frac{c-1}{1}\right),$$

$$\mathbb{P}(\text{Type II error}) = \mathbb{P}\big(X < c | X \sim N(2,2)\big) = \Phi\left(\frac{c-2}{\sqrt{2}}\right).$$

- LR test:

$$LR(X) = \frac{\frac{1}{\sqrt{2\pi}} e^{-(X-1)^2/2}}{\frac{1}{\sqrt{4\pi}} e^{-(X-2)^2/4}} = \sqrt{2} e^{(X-2)^2/4 - (X-1)^2/2} = \sqrt{2e} e^{-X^2/4}$$

$$LR(X) \leq c \Leftrightarrow |X| \geq 2\sqrt{\log\left(\frac{\sqrt{2e}}{c}\right)}.$$

Let's set threshold $c = 1.5$.

$$LR(X) \leq 1.5 \Leftrightarrow |X| \geq 1.33.$$

$$\mathbb{P}(\text{Type I error}) = \mathbb{P}(|X| \geq 1.33 | X \sim N(1,1)) = 0.38,$$

$$\mathbb{P}(\text{Type II error}) = \mathbb{P}(|X| < 1.33 | X \sim N(2,2)) = 0.309.$$

- Now let's go back to the first test. We need c such that $1 - \Phi\left(\frac{c-1}{1}\right) = 0.38$ i.e. $c = 1 + \Phi^{-1}(0.62) = 1.31$. And then

$$\mathbb{P}(\text{Type II error}) = \Phi\left(\frac{c-2}{\sqrt{2}}\right) = 0.313.$$

5.7 Generalized likelihood ratios

We are often interested in settings where H_0, H_1 or both H_0 and H_1 are composite hypotheses. Let us consider testing

$$H_0 : X \sim f(x \mid \theta) \text{ for some } \theta \in \Omega_0, \quad H_1 : X \sim f(x \mid \theta) \text{ for some } \theta \in \Omega_1$$

where Ω_0, Ω_1 are some sets of possible parameter values. For example,

$$H_0 : X \sim N(0,1), \quad H_1 : X \sim N(\mu, 1) \text{ for some } \mu > 1.$$

One approach in this setting is the generalized likelihood ratio test. We maximize the likelihood over each hypothesis.

$$LR = \frac{\frac{1}{\sqrt{2\pi}} e^{-(x-0)^2/2}}{\max_{\mu>1} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}}.$$

Due to issues of endpoints / closed sets, it's often easier to think of the denominator as a maximum over $H_0 \cup H_1$.

In general, we form two model spaces:

- Ω_0 is the model with all the possible parameter values from H_0.
- $\Omega = \Omega_0 \cup \Omega_1$ is the model with all possible parameter values from H_0 or H_1.

We define the generalized likelihood ratio

$$\Lambda(X) = \frac{\max_{\theta \in \Omega_0} f(X \mid \theta)}{\max_{\theta \in \Omega} f(X \mid \theta)}.$$

We will reject the null hypothesis if the generalized likelihood ratio is small. That is, we will consider rejection regions of the form $RR = \{x : x < k\}$ for some k.

Examples

1. Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, 1)$, and we are testing $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. Then

$$\prod_i f(X_i \mid \mu) = \prod_i \frac{1}{\sqrt{2\pi}} e^{-(X_i-\mu)^2/2} = \text{constant} \cdot e^{-n(\bar{X}-\mu)^2/2},$$

where the constant depends on X_i's but not on μ and so

$$\Lambda = \frac{\max_{\mu \in \{0\}} f(X \mid \theta)}{\max_{\mu \in \mathbb{R}} f(X \mid \theta)} = \frac{\max_{\mu \in \{0\}} e^{-n(\bar{X}-\mu)^2/2}}{\max_{\mu \in \mathbb{R}} e^{-n(\bar{X}-\mu)^2/2}} = \frac{e^{-n\bar{X}^2/2}}{1} = e^{-n\bar{X}^2/2}.$$

We will reject H_0 if Λ is small or, equivalently, if $|\bar{X}|$ is large. This is very intuitive, and agrees with the test derived in Section 5.2.

2. $H_0 : \mu = 0$ vs $H_1 : \mu > 0$. In this case, the denominator is maximized at $\mu = \bar{X}$ if $\bar{X} \geq 0$ or $\mu = 0$ otherwise, so

$$\Lambda = \frac{\max_{\mu \in \{0\}} f(X \mid \theta)}{\max_{\mu \in [0,\infty)} f(X \mid \theta)} = \frac{\max_{\mu \in \{0\}} e^{-n(\bar{X}-\mu)^2/2}}{\max_{\mu \in [0,\infty)} e^{-n(\bar{X}-\mu)^2/2}} = \begin{cases} \frac{e^{-n\bar{X}^2/2}}{1} = e^{-n\bar{X}^2/2}, & \bar{X} \geq 0, \\ \frac{e^{-n\bar{X}^2/2}}{e^{-n\bar{X}^2/2}} = 1, & \bar{X} \leq 0. \end{cases}$$

3. $H_0 : \mu = 0, \sigma^2 = 1$ vs H_1: anything else. Recall that the MLE is $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$, and so

$$\Lambda = \frac{\max_{\mu,\sigma^2 \in \{0,1\}} f(X \mid \theta)}{\max_{\mu,\sigma^2 \in \mathbb{R}\times\mathbb{R}_+} f(X \mid \theta)} = \frac{\max_{\mu,\sigma^2 \in \{0,1\}} \frac{1}{\sqrt{(2\pi\sigma^2)^n}} e^{-\sum_i (X_i-\mu)^2/2\sigma^2}}{\max_{\mu,\sigma^2 \in \mathbb{R}\times\mathbb{R}_+} \frac{1}{\sqrt{(2\pi\sigma^2)^n}} e^{-\sum_i (X_i-\mu)^2/2\sigma^2}} = \frac{e^{-\sum_i X_i^2/2}}{\frac{1}{\hat{\sigma}^n} e^{-n/2}}.$$

5.7.1 Asymptotic distribution

If Ω_0 and Ω can be represented as d_0 and d dimensional parameter spaces, then under some regularity conditions, if H_0 is true then

$$-2\log(\Lambda) \approx \chi^2_{d-d_0}.$$

Note that when using a generalized likelihood ratio test we want to reject when Λ is small, so when $-2\log(\Lambda)$ is large. Therefore, to turn a generalized likelihood ratio test into a significance test we can set a threshold

$$k^* = F^{-1}_{\chi^2_{d-d_0}}(1-\alpha)$$

and check whether $-2\log(\Lambda) > k^*$ (reject H_0) or $\leq k^*$ (do not reject H_0). To get a p-value we can do

$$1 - F_{\chi^2_{d-d_0}}(-2\log(\Lambda)).$$

Examples In our examples above,

1. In this case $dim(\Omega_0) = dim(\{0\}) = 0$, $dim(\Omega) = dim(\mathbb{R}) = 1$. So, $-2\log(\Lambda) \approx \chi^2_1$. In fact, $-2\log(\Lambda) = n\bar{X}^2$, since $\bar{X} \sim N(0, 1/n)$ then $\sqrt{n}\bar{X} \sim N(0,1)$ and $n\bar{X}^2 \sim \chi^2_1$—so this is exact.
2. This example doesn't satisfy regularity conditions since Ω has an "endpoint" right at $\mu = 0$ i.e. at Ω_0.
3. In this case $dim(\Omega_0) = 0$ and $dim(\Omega) = 2$. And

$$-2\log(\Lambda) = n\log(\hat{\sigma}^2) - n + \sum_i X_i^2 \approx \chi^2_2.$$

5.8 Multinomial data

Titanic \ Godfather	like	dislike	total
like	42	51	93
dislike	63	44	107
total	105	95	200

Consider testing one of these models:

- Model 1: A person's opinion about T & their opinion about G are independent, & both movies are equally popular.
- Model 2: A person's opinion about T & their opinion about G are independent.

5.8.1 Multinomial likelihoods & MLEs

Suppose that each individual's category is a multinomial draw with probability $p_1, \ldots, p_m$. Let $O_1, \ldots, O_m$ be the number of observed individuals in each category. The support is $\{O_1, \ldots, O_m \geq 0 : O_1 + \cdots + O_m = n\}$ with PMF

$$p_{O_1,\ldots,O_m}(n_1, \ldots, n_m) = \binom{n}{n_1, \ldots, n_m} p_1^{n_1} \cdot \cdots \cdot p_m^{n_m}.$$

Let Δ_m be the simplex, i.e. $\{p \in \mathbb{R}^m : p \geq 0, \sum_i p_i = 1\}$. We will learn to test

$$H_0 : p \in \Omega_0 \text{ vs } H_1 : p \in \Delta_m \backslash \Omega_0$$

where Ω_0 is some subset of the simplex.

What is the MLE over all of Δ_m: the observed proportions, $\hat{p}_i = \frac{O_i}{n}$.

Example for movies:

- Model 1: parametrized by one value q = proportion of people who like Titanic = proportion of people who like Godfather. Then the proportions for the four cells of the two-by-two table are $(q^2, q(1-q), q(1-q), (1-q)^2)$ for some $q \in [0,1]$. The likelihood is given by

$$\binom{n}{n_1, n_2, n_3, n_4} \cdot (q^2)^{n_{TG}} (q(1-q))^{n_T} (q(1-q))^{n_G} ((1-q)^2)^{n_0} = constant \cdot q^{2n_{TG}+n_T+n_G} (1-q)^{2n_0+n_T+n_G}.$$

  If we maximize this, we find that the MLE is

$$\hat{q} = \frac{2n_{TG} + n_T + n_G}{2n_{TG} + n_T + n_G + 2n_0 + n_T + n_G} = \frac{\text{\# who like Titanic + \# who like Godfather}}{2n} = \frac{105 + 93}{2 \cdot 200} = 0.495.$$

  So the MLE probability vector is

$$\hat{p} = (\hat{q}^2, \hat{q}(1-\hat{q}), \hat{q}(1-\hat{q}), (1-\hat{q})^2) = (.245, .250, .250, .255)$$

- Model 2: parametrized by q_T, q_G = proportion of people who like Titanic / Godfather. The proportions for the four cells are $(q_T q_G, q_T(1-q_G), (1-q_T)q_G, (1-q_T)(1-q_G))$. If we calculate the likelihood and maximize it, the MLE is

$$\hat{q}_T = \frac{\text{\# who like Titanic}}{n} = \frac{93}{200} = 0.465, \hat{q}_G = \frac{\text{\# who like Godfather}}{n} \frac{105}{200} = 0.525$$

  so

$$\hat{p} = (\hat{q}_T \hat{q}_G, \hat{q}_T(q - \hat{q}_G), (1-\hat{q}_T)\hat{q}_G, (1-\hat{q}_T)(1-\hat{q}_G)) = (.244, .221, .281, .254).$$

5.8.2 Pearson's χ^2 test

While we could apply a LR test, Pearson's χ^2 is a slightly different test, with a different test statistic to gain a bit more power:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

also compared against a χ^2 distribution with $m-1-$(# parameters fitted for H_0) d.f. Here O_i and E_i are the observed counts, and the expected counts according to a certain model.

For model 1: expected counts are

Titanic \ Godfather	like	dislike	total
like	49	50	
dislike	50	51	
total			200

(For instance, the 49 comes from multiplying 0.245×200, and so on.)

Then the χ^2 statistic is

$$X^2 = \frac{(42-49)^2}{49} + \frac{(51-50)^2}{50} + \frac{(63-50)^2}{50} + \frac{(44-51)^2}{51} = 5.36.$$

P-value $p = 1 - F_{\chi^2_2}(5.36) = 0.0686$.

For model 2: expected counts are

Titanic \ Godfather	like	dislike	total
like	48.8	44.2	
dislike	56.2	50.8	
total			200

(Don't round to integers!)
Then the χ^2 statistic is

$$X^2 = \frac{(42-48.8)^2}{48.8} + \frac{(51-44.2)^2}{44.2} + \frac{(63-56.2)^2}{56.2} + \frac{(44-50.8)^2}{50.8} = 3.73.$$

P-value $p = 1 - F_{\chi^2_1}(3.73) = 0.0534$.

5.8.3 χ^2 test of independence (13.4)

More generally, a χ^2 test of independence is checking whether the row category value is independent of the column category value, for a $r \times c$ contingency table. The space H_0 is therefore all $r \times c$ probability distributions p of the form:

	Col 1	Col 2	...	Col c
Row 1	$p_{11} = p_1^R p_1^C$	$p_{12} = p_1^R p_2^C$	...	$p_{1c} = p_1^R p_c^C$
Row 2	$p_{21} = p_2^R p_1^C$	$p_{22} = p_2^R p_2^C$	...	$p_{2c} = p_2^R p_c^C$
...				

The MLEs for H_0 are $\hat{p}_i^R = \frac{\text{\# samples in row } i}{n}$, $\hat{p}_j^C = \frac{\text{\# samples in col } j}{n}$.

To run a LRT or a Pearson's χ^2 test, the number of fitted parameters is $(r-1) + (c-1)$ (since p_r^R and p_c^C are determined by the rest of the vectors). So d.f. is

$$d.f. = rc - 1 - (r-1) - (c-1) = (r-1)(c-1).$$

Exercises

Exercise 1 Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$, with $H_0 : \lambda = 1$ vs $H_A : \lambda \neq 1$. Following the examples of Section 5.7, find the generalized likelihood ratio Λ and its asymptotic distribution.

Exercise 2 Suppose that the IQ of students in STAT 24410 is known to be normally distributed with a variance of 25. I sample 100 students and find an average IQ of 120. You can use that $z_{0.05} = 1.645$.

a) At the 0.05 level, is there evidence to suggest that the average IQ is < 130?

b) What is the p-value of this test?

Exercise 3 Suppose a researcher claims that the mean age of founders of start-up companies in Silicon Valley is less than 30 years. To test his claim, he randomly selected 8 start-up companies and obtained the following data on the age of their founders: $27, 29, 31, 33, 24, 28, 31, 33$ years. Is there significant evidence that the mean age of founders is less than 30 years? Test this at the 0.05 level of significance. You can use that $t_{0.05} = 1.895$ if the number of degrees of freedom is 7.

Exercise 4: Is this data Binomial? In this problem you will learn to test hypotheses about multinomial data—that is, data where each observation falls into one of several categories. In the book, this material is primarily in section 9.5 (but all the information you need is contained in the problem itself).

Suppose that 50 students each have 4 tries to throw a dart at a target. Here are the results:

# successes	0	1	2	3	4
# students	9	12	11	14	4

This data is multinomial—for each student, his result falls into one of these 5 categories.

Our goal is to answer the following question: is the distribution of this data characterized by a Binomial distribution with some common parameter p? That is, is it true that for each student, their 4 trials are independent and all have the same chance of success? One reason to believe that this hypothesis is not true, would be if we think that different students have substantially different skill levels for throwing darts. For example, if we believe that most students are all either experts at darts or very poor at throwing darts, then we'd expect to see lots of 4's (for the experts) and lots of 0's and 1's (for the students who are very bad at darts) and relatively few 2's and 3's, while for a Binomial distribution this could never be the case.

1. First, let X_i be the number of successes for the ith student, for $i = 1, \ldots, 50$. What is the likelihood for the observed data $X_1, \ldots, X_{50}$ using the parameter p?
2. Now calculate the MLE $\hat{p}$ as a function of $X_1, \ldots, X_{50}$.
3. Next, calculate the MLE for the actual observed data.
4. Next, given this MLE, how many students would you expect to fall into each of the five categories? That is, for each category, calculate $E(\#\text{ of students in this category})$ if in fact the true p were equal to your MLE $\hat{p}$:

# successes	0	1	2	3	4
# students	??	??	??	??	??

5. Now, we'll compute a statistic. If the numbers you observed (9, 12, etc) are very far from the expected values calculated in the previous part, then that suggests that the data might not be Binomial. However, the size of the discrepancy should be relative to the expected number (i.e. if you expect 10 and get 15, that's not as much of a discrepancy as if instead you expect 2 and get 8.) So you should calculate

$$\sum_{\text{every entry in the table}} \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}}.$$

6. The statistic you calculated is a χ^2 statistic, and should be compared against a χ^2 distribution where the number of degrees of freedom (d.f.) is (number of categories -1)$-$(number of parameters fitted in your model). Calculate this d.f. Then use the χ^2 table in the back of the textbook to get a p-value:

$$\text{p-value} = \mathbb{P}(\chi^2_{\text{d.f.}} \text{ is at least as large as the value that you obtained}).$$

Further Exercises

Exercise 1 (Likelihood ratio for variance). Suppose that we have data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(0, \nu)$ (here ν is the variance), and we are testing $H_0 : \nu = 1$ against $H_1 : \nu \neq 1$.

1. Compute the MLE $\hat{\nu}$.
2. Construct the generalized likelihood ratio statistic Λ for this test. Then calculate $-2\log(\Lambda)$, simplified into a clean form. You should be able to write it without any X_i's in the final answer, only $\hat{\nu}$ and constants such as e, n, etc.
3. Suppose you observe data $X_1, \ldots, X_n$ for sample size $n = 40$ with these summary statistics: sample mean $\bar{X} = 0.8$, sample variance $S^2 = 0.9$. What is the value of $-2\log(\Lambda)$, for this data?
4. Perform a χ^2 test (i.e. calculate a p-value—you can use the table in the back of the book).

Exercise 2 Suppose we flip a coin 10 times to determine whether or not it is biased. Let X denote the number of heads that appear. We wish to test $H_0 : p = 0.5$ versus $H_a : p > 0.5$.

a) If we use the rejection region $\{X > 7\}$, compute the probability α of a Type I error.

b) Suppose that there is reason to believe that $p = 0.7$. Compute the probability of Type II error for that specific value of the alternative hypothesis.

Exercise 3 Suppose we generated a sample X_1 from an exponential distribution with mean β, but we forgot if we used a mean $\beta_0 = 1$ or $\beta_a = 2$. Follow the steps below with the null hypothesis $\beta = \beta_0$ and alternative hypothesis $\beta = \beta_a$ to find a likelihood ratio test with a Type I error α.

a) Write the likelihood $L(\beta_0; X_1)$ which represents how likely β_0 (the null) is based on X_1. Write also the likelihood $L(\beta_a; X_1)$. Finally calculate the ratio $LR(X_1) = L(\beta_0; X_1)/L(\beta_a; X_1)$.

b) Given this form of the ratio (null over alternative), should we reject the null if $LR(X_1) \geq k$ (for some k) or if $LR(X_1) \leq k$. Why? Write the requirement for rejection in terms of X_1 e.g. reject if $X_1 \geq \cdots$ or reject if $X_1 \leq \cdots$.

c) Find the probability that X_1 lies in your rejection region assuming the null hypothesis is true, i.e. the exponential distribution had mean $\beta_0 = 1$. Your answer should depend on k. Set this probability equal to α and solve for k. Conclude by explicitly stating the test: "if ... reject the null. Otherwise, do not reject it".

Exercise 4 An advertisement at a Casino claims that a person only needs to play a slot machine 4 times on average to win. Suppose we have reason to believe it takes longer to win. We plan to play the slot machine. If we play at least 4 times before winning we will reject the advertisement.

a) What is the Type I error for this test?

b) Give a new rejection region that lowers Type I error.

c) Suppose we think it takes 6 tries on average to win. What is the Type II error under this assumption?

Chapter 6

Bayesian Statistics

6.1 Introduction

So far we have considered only the classical (a.k.a. frequentist) approach to statistical inference. In the classical paradigm, unknown parameters are treated as fixed but unknown constants which are to be estimated. Probabilistic statements are made only about "true" random variables and observed data is assumed to correspond to a sampled set of realisations of random variables. Bayesian inference provides an alternative approach.

In the Bayesian approach, parameters are treated as random variables and hence have a probability distribution. Prior information about θ is combined with information from sample data to compute a *posterior* distribution of θ. The posterior distribution contains all the available information about θ so should be used for making estimates or inferences. More formally, a Bayesian starts by using prior information about θ to define a *prior distribution*, $f_\Theta(\theta)$. Information from sample data is given by the likelihood $L(\theta; x) = f(x|\theta)$. By Bayes Theorem the conditional distribution of Θ given $X = x$ is

$$f(\theta|x) = \frac{f(x|\theta) f_\Theta(\theta)}{h(x)} \tag{6.1}$$

$$= \frac{L(\theta; x) f_\Theta(\theta)}{h(x)} \propto L(\theta; x)\, f_\Theta(\theta) \tag{6.2}$$

where $h(x)$ is the marginal distribution of x. The symbol $\propto$ indicates that the left-hand side and the right-hand side are equal up to a constant that does not depend on the variable θ. We call $f(\theta|x)$ the posterior distribution of θ. Actually, a Bayesian would most probably have written:

$$f(\theta|x) \propto f(x|\theta) f_\Theta(\theta).$$

There is no need to distinguish between parameters and random variables in notation and it's perfectly reasonable to condition upon the parameters within the Bayesian framework.

Note: A Bayesian statistician does not necessarily believe that all parameter values are classical random variables. The Bayesian interpretation of probability itself is different from the frequentist interpretation. Viewing probabilistic statements as quantifications of uncertainty without explicit reference to relative frequencies of occurrence allows their use much more widely. In the subjective Bayesian framework, probabilities are quantifications of personal belief in a statement.

Advantages and comments on the frequentist method:

- Confidence intervals have a nice definition.
- MLEs can "often" be done with pencil and paper.
- You don't have to specify a prior distribution which seems more sciency.
- Computing the MLE often requires to use some optimization algorithm.

Advantages and comments on the Bayesian method:

- You often have some prior beliefs about parameters.
- Using conjugate priors you can solve some models with pen and paper.
- Using computers you can deal with very complicated models involving missing data, lots of parameters and high-dimensional data.
- If you make a reasonable choice of prior distribution, and then collect enough data, your conclusions will be effectively independent of the choice of prior.
- Computing posterior probabilities and expectations often requires to use Monte Carlo sampling methods.

The important thing to remember about Bayesian statistics is:

$$\text{posterior} \propto \text{prior} \times \text{likelihood of the data.}$$

In words: the posterior distribution is proportional to the prior distribution times the likelihood of the observed data. We don't lose any information by putting in a proportional sign, as the missing constant is determined by the fact that the posterior is a probability distribution and so it must integrate to 1.

Another reason for losing the denominator in equation (**??**) is that in most cases it is either:

- Obvious, because the posterior is some recognizable distribution; or
- Too difficult to evaluate, as it may involve computing a complicated integral. In this case one needs to perform inference by sampling the posterior, which is done by using Monte Carlo methods that do not require to know the normalizing constant.

6.2 Set-up and examples

We consider the following setup:

$$\begin{cases} \Theta \sim f_\Theta(\theta) & \text{(Prior distribution)} \\ X_1, \dots, X_n \mid \Theta = \theta \overset{\text{iid}}{\sim} f(x \mid \theta) & \text{(Likelihood function)} \end{cases}$$

The posterior distribution of θ given $X_1 = x_1, \dots, X_n = x_n$ is

$$\begin{aligned} f_{\Theta \mid X_1,\dots,X_n}(\theta \mid x_1, \dots, x_n) &= \frac{f_{\Theta, X_1,\dots,X_n}(\theta, x_1, \dots, x_n)}{f_{X_1,\dots,X_n}(x_1, \dots, x_n)} \\ &= \underbrace{f_\Theta(\theta)}_{\text{Prior}} \underbrace{\prod_{i=1}^n f(x_i \mid \theta)}_{\text{Likelihood}} \cdot (\text{constant with respect to } \theta). \end{aligned}$$

In Bayesian statistics we think of Θ as random after observing the data, so it makes sense to talk about $\mathbb{P}(\Theta > 0)$ and other probability statements about Θ. Here is a very brief summary of point estimation, interval estimation, and hypothesis testing in Bayesian statistics:

- The posterior distribution of Θ treats it as a random variable, but what if we want a "point estimate", i.e. a single value which is a good estimate for the parameter? Two standard options:
  - Posterior mean: $\mathbb{E}(\Theta \mid X_1 = x_1, \dots, X_n = x_n)$ = expected value of the posterior distribution.

  - Posterior mode: $\arg\max_\theta f_{\Theta|X_1,\ldots,X_n}(\theta \mid x_1,\ldots,x_n)$ = mode of the posterior distribution.

- Credible interval: a $(1-\alpha)$ credible interval I is some random interval calculated as a function of the data $X_1,\ldots,X_n$ such that we have a $(1-\alpha)$ posterior probability that Θ lies in the interval, $\mathbb{P}(\Theta \in I \mid X_1,\ldots,X_n) = 1-\alpha$. There is no single way to construct a credible interval, but here are two common options:

  - Equal tailed interval: our interval is

$$F^{-1}_{\text{posterior distribution}}(\alpha/2) \le \theta \le F^{-1}_{\text{posterior distribution}}(1-\alpha/2).$$

  - High posterior density interval: our interval is given by

$$I = \{\theta : f_{\Theta|X_1,\ldots,X_n}(\theta \mid x_1,\ldots,x_n) \ge c\}$$

    where the density cutoff c is chosen so that the probability equals $1-\alpha$, i.e.

$$\int_{\theta\in I} f_{\Theta|X_1,\ldots,X_n}(\theta \mid x_1,\ldots,x_n)\, d\theta = 1-\alpha.$$

    Note that this region I might not be a single interval!

- Hypothesis testing under the Bayesian view of statistics is based on the posterior distribution. Consider for simplicity testing $H_0 : \theta > 0$ against $H_a : \theta \le 0$. A Bayesian would compute the posterior probability $\mathbb{P}(\Theta > 0 \mid X_1,\ldots,X_n)$ and reject the null hypotheses if said probability is below some threshold.

A "conjugate prior" distribution when combined with the likelihood function, produces a posterior distribution in the same family as the prior. If we find a conjugate prior distribution which adequately fits our prior beliefs regarding θ, we should use it because it will simplify computations considerably. However, one should not employ a conjugate prior distribution for computational convenience if it does not represent those prior beliefs reasonably closely.

In the following three subsections we show three examples of conjugate priors for the Bayesian estimation of the expected value of a normal distribution, the rate of an exponential distribution, and the probability of heads of a coin.

6.2.1 Bayesian inference for the mean of a normal distribution

$$\begin{cases} \mu \sim N(\mu_0, \nu^2) & \text{(Prior distribution)} \\ X_1,\ldots,X_n \mid \mu \overset{\text{iid}}{\sim} N(\mu,\sigma^2) & \text{(Likelihood function)} \end{cases}.$$

What is the posterior distribution for μ after observing samples $X_1,\ldots,X_n$?

$$\begin{aligned} f_{\mu|X_1,\ldots,X_n}(t|x_1,\ldots,x_n) &\propto f_\mu(t)\prod_{i=1}^n f_{X_i|\mu}(x_i \mid t) \\ &= \frac{1}{\sqrt{2\pi\nu^2}} e^{-(t-\mu_0)^2/2\nu^2} \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-t)^2/2\sigma^2} \\ &\propto \exp\left\{-\frac{1}{2}\left(t^2(\nu^{-2}+n\sigma^{-2}) - 2t(\mu_0\nu^{-2}+\sum_i x_i\sigma^{-2})\right)\right\} \\ &\propto \exp\left\{-\left(t - \frac{\mu_0\nu^{-2}+\sum_i x_i\sigma^{-2}}{(\nu^{-2}+n\sigma^{-2})}\right)^2 / 2(\nu^{-2}+n\sigma^{-2})^{-1}\right\}. \end{aligned}$$

Therefore, the posterior distribution of μ is a normal distribution

$$N\left(\frac{\mu_0\nu^{-2}+\sum_i x_i\sigma^{-2}}{(\nu^{-2}+n\sigma^{-2})}, (\nu^{-2}+n\sigma^{-2})^{-1}\right) = N\left(\bar{x}\cdot\frac{n/\sigma^2}{n/\sigma^2+1/\nu^2} + \mu_0\cdot\frac{1/\nu^2}{n/\sigma^2+1/\nu^2}, \frac{1}{n/\sigma^2+1/\nu^2}\right).$$

If n is very large, the mean is $\approx \bar{x}$ and the variance is ≈ 0. Note that as n grows the effect of the prior on the posterior is reduced.

Using the posterior distribution we can compute posterior probabilities for μ. For example, for $z^* = \Phi^{-1}(1-\alpha/2)$, $\mathbb{P}(N(0,1) \in \pm z^*) = 1-\alpha$ and so we have an interval for μ:

$$\mathbb{P}\left(\mu \in \left[\bar{x} \cdot \frac{n/\sigma^2}{n/\sigma^2+1/\nu^2} + \mu_0 \cdot \frac{1/\nu^2}{n/\sigma^2+1/\nu^2}\right] \pm z^* \cdot \sqrt{\frac{1}{n/\sigma^2+1/\nu^2}} \;\middle|\; X_1 = x_1, \ldots, X_n = x_n\right) = 1-\alpha.$$

This is called a $(1-\alpha)$ credible interval.

It is important to note that in Bayesian statistics we are giving a true probability — not a "confidence" statement. The reason is that we are now treating μ as random; even after observing the data μ is still random (with the distribution given by its posterior distribution) and so we can still talk about probabilities, in contrast to the frequentist interpretation.

Note also that for very large n, this credible interval is nearly equal to $\bar{x} \pm z^* \cdot \sigma/\sqrt{n}$, which is the frequentist confidence interval.

6.2.2 Bayesian inference for the rate of an exponential distribution

$$\begin{cases} \lambda \sim \text{Gamma}(k, r) & \text{(Prior distribution)} \\ X_1, \ldots, X_n \mid \lambda = t \overset{\text{iid}}{\sim} \text{Exponential}(t) & \text{(Likelihood function)} \end{cases}$$

where where k = shape, r = rate, both > 0. The posterior distribution is

$$\begin{aligned} f_{\lambda|X_1,\ldots,X_n}(t \mid x_1, \ldots, x_n) &\propto f_\lambda(t) \prod_i f(x_i \mid t) = \frac{r^k}{\Gamma(k)} t^{k-1} e^{-rt} \prod_i (te^{-tX_i}) \\ &\propto t^{k+n-1} e^{-t(r+\sum_i X_i)} \end{aligned}$$

and therefore, we see that

$$\lambda \mid X_1, \ldots, X_n \sim \text{Gamma}\left(k+n, r+\sum_i X_i\right).$$

Note: recalling that a Gamma distribution w/ a large shape parameter can be written as a sum of iid exponentials, the CLT tells us that for large n, this Gamma distribution is approximately normal. Recall also the following numerical characterizations of a Gamma(k, r) distribution:

- Mean: k/r; variance k/r^2.
- Mode: $\frac{k-1}{r}$ if $k \geq 1$ or 0 if $k < 1$.

Hence we obtain the following numerical characterizations of the posterior:

- Posterior mean: $\frac{k+n}{r+\sum_i X_i}$, which is $\approx \frac{1}{\bar{X}}$ when n large; variance $\frac{k+n}{(r+\sum_i X_i)^2}$.
- Posterior mode: $\frac{k+n-1}{r+\sum_i X_i}$, again $\approx \frac{1}{\bar{X}}$.
- Credible interval: for an equal-tailed interval, if we have access to the gamma CDF,

$$F^{-1}_{\text{Gamma}(k+n, r+\sum_i X_i)}(\alpha/2) \leq \lambda \leq F^{-1}_{\text{Gamma}(k+n, r+\sum_i X_i)}(1-\alpha/2).$$

- Since Gamma$(k+n, r+\sum_i X_i) \approx N(\frac{k+n}{r+\sum_i X_i}, \frac{k+n}{(r+\sum_i X_i)^2})$ according to CLT, this credible interval will be roughly

$$\approx \left[\frac{k+n}{r+\sum_i X_i} - z_* \sqrt{\frac{k+n}{(r+\sum_i X_i)^2}}, \frac{k+n}{r+\sum_i X_i} + z_* \sqrt{\frac{k+n}{(r+\sum_i X_i)^2}}\right].$$

For large n, this is roughly

$$\approx \left[\frac{1}{\bar{X}}\left(1 - z_*/\sqrt{n}\right), \frac{1}{\bar{X}}\left(1 + z_*/\sqrt{n}\right)\right]$$

which is equal to the $(1-\alpha)$ confidence interval.

6.2.3 Bayesian inference for the probability of heads

We have a coin with an unknown bias and we will flip the coin to try to determine its bias. We will use as prior distribution over the probability of heads the Beta distribution, that we now introduce:

The Beta distribution The Beta distribution: supported on $[0,1]$, depends on two parameters $\alpha, \beta > 0$.

$$f(t) = \frac{1}{B(\alpha,\beta)} t^{\alpha-1}(1-t)^{\beta-1}$$

where the normalizing constant is

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Beta$(1,1)$ is the uniform distribution. Beta(c,c) for $c<1$ is U-shaped. Beta(c,c) for $c>1$ is unimodal. Beta(α,β) is skewed for $\alpha \neq \beta$.

Expected value: $\frac{\alpha}{\alpha+\beta}$. [Check it using the fact from the Gamma function, $\Gamma(c+1) = c\Gamma(c)$].

Variance: $\frac{\alpha}{\alpha+\beta} \cdot \frac{\beta}{\alpha+\beta} \cdot \frac{1}{\alpha+\beta+1}$.

Hierarchical model and posterior

$$\begin{cases} \Theta \sim \mathsf{Beta}(\alpha,\beta) & \text{(Prior distribution)} \\ X \mid \Theta = \theta \sim \mathsf{Binomial}(n,\theta) & \text{(Likelihood function).} \end{cases}$$

We now find the posterior distribution:

$$\begin{aligned} f_{\Theta|X}(\theta \mid X=k) &\propto \underbrace{f_\Theta(\theta)}_{\text{prior}} \underbrace{\mathbb{P}(X=k|\Theta=\theta)}_{\text{likelihood}} \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^k(1-\theta)^{n-k} \\ &= \theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1} \end{aligned}$$

which is the density of the Beta$(\alpha+k, \beta+n-k)$ distribution. (If we want to be formal regarding PMFs & densities, we could work with probabilities only by looking at $\mathbb{P}(t \le \theta \le t+\epsilon, X=k)$ and then taking a limit as $\epsilon \to 0$.)

For example, we can calculate

$$\mathbb{E}(\theta \mid X=k) = \frac{\alpha+k}{\alpha+\beta+n},$$

derived from the expected value of the Beta distribution.

- The posterior distribution of Θ is Beta$(\alpha+X, \beta+n-X)$.
- The posterior mean of Θ is $\frac{\alpha+X}{\alpha+\beta+n}$.

In our example, if n is large, since k and $n-k$ are likely to be large, the new Beta distribution is likely to be (a) quite concentrated (low variance) with (b) mean $\approx k/n$. In other words our posterior belief is that θ is quite close to the observed fraction k/n. If instead n is small then our posterior may be quite similar to our prior; k and n are small relative to the original parameters α and β, so they don't change the distribution by much. A nice interpretation for α and β is that we previously observed α many Heads and β many Tails.

6.3 Uninformative priors

One possible criticism of Bayesian inference is that it's not clear what to do when we don't have any prior information. In fact, Bayesian inference is really just a rule for updating our beliefs in light of new data. A way to deal with this is to attempt to employ priors which encode our ignorance. An uninformative prior is one which attempts to encode as little information as possible about the value of a parameter. The simplest approach to the construction of uninformative priors is the flat prior which has $f_\Theta(\theta) = constant$ for all θ. Flat priors can sometimes be obtained as special or limiting cases of conjugate priors, e.g. – using a Beta$(1,1) \equiv U[0,1]$ prior for the Bernoulli parameter, or letting the variance of a normal prior tend to infinity. Other types of uninformative prior can also sometimes be obtained by procedures such as these. If the prior is approximately constant over the range of θ for which the likelihood is appreciable, then approximately $f(\theta|x) \propto L(\theta; x)$ and inference becomes similar to MLE estimation. Note that even here there is a difference as a Bayesian would make parametric inference on the basis of a loss function and it is not necessarily (or in fact often) the case that minimising expected loss occurs when one chooses the maximum of the posterior density as the estimator. In fact, the estimator which takes the maximum of the posterior density is the Bayesian estimator obtained as the limit of a sequence of loss functions and corresponds essentially to dealing with a loss which is zero if one estimates the parameter exactly correctly and one if there is any error at all. There are perhaps some situations in which this is justifiable, but one must think carefully about the choice of loss function when making decisions or inferences in a Bayesian framework.

Problems with Flat Priors If the prior range of θ is infinite, a flat prior cannot integrate to 1. Such an improper prior may lead to problems in finding a "proper" posterior (i.e. one which can be normalised to integrate to unity as any probability density must). We usually have some prior knowledge of θ and if we do we should use it, rather than claiming ignorance. In fact, one could argue that there are essentially no cases in which we really believe that any value in an unbounded set is not only possible but equally plausible.

Jeffreys Prior Another issue is whether an informative prior should be flat for θ or some function of θ, say θ^2 or $\log\theta$. We will draw different inferences for a prior which is flat over any one of these functions. In some sense, this demonstrates that flat priors do not really encode complete ignorance at all. One solution is to construct a prior which is flat for a function $\phi(\theta)$ whose Fisher information I_ϕ is constant. This leads to the Jeffreys prior which is proportional to $I_\theta^{1/2}$. Formally, we define the Jeffrey's prior for a parameter θ to be

$$f_\Theta^{\text{Jeff}}(\theta) \propto I_\theta^{1/2}.$$

We have the following result:

Proposition 1. *Suppose that $\Theta \sim f_\Theta^{Jeff}(\theta)$ is distributed according to the Jeffrey's prior of θ. Let $\Xi = \phi(\Theta)$. Then Ξ is distributed according to the Jeffrey's prior of $\xi = \phi(\theta)$.*

Proof. Using a change of variables,

$$\begin{aligned}
f_\Xi(\xi) &= f_\Theta(\theta)\left|\frac{d\theta}{d\xi}\right| \\
&= \sqrt{I_\theta\left(\frac{d\theta}{d\xi}\right)^2} \\
&= \sqrt{\mathbb{E}\Big[\Big(\frac{d\log f(X|\theta)}{d\theta}\Big)^2\Big]\left(\frac{d\theta}{d\xi}\right)^2} \\
&= \sqrt{\mathbb{E}\Big[\Big(\frac{d\log f(X|\theta)}{d\theta}\Big)^2\Big]\left(\frac{d\theta}{d\xi}\right)^2} \\
&= \sqrt{\mathbb{E}\Big[\Big(\frac{d\log f(X|\xi)}{d\xi}\Big)^2\Big]} = I_\xi^{1/2}.
\end{aligned}$$

□

We now give an example.

Example 5. Consider a Bernoulli likelihood $X|\Theta = \theta \sim \text{Bernoulli}(\theta)$, $0 < \theta < 1$. What is the Jeffrey's prior for θ?
Answer: We computed in class that $I_\theta = \frac{1}{\theta(1-\theta)}$. Therefore

$$f_\Theta^{\text{Jeff}}(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}} = \theta^{-1/2}(1-\theta)^{-1/2}, \quad 0 < \theta < 1.$$

Hence we see that the Jeffrey's prior is a Beta$(0.5, 0.5)$. Consider another parametrization, $X|\Theta = \theta \sim \text{Bernoulli}(\sin^2(\alpha))$, for $\alpha \in (0, \pi/2)$. Then $I_\alpha \equiv$ constant. Therefore, the Jeffrey's prior for α is the uniform prior in $(0, \pi/2)$.

Remark 3. Harold Jeffreys was a physicist who made substantial contributions to the theory and philosophy of Bayesian inference. His original motivation when developing this class of prior distributions was to develop a way of encoding a belief that the distribution should be invariant under a particular type of transformation: location parameters should have a prior invariant under shifts; scale parameters should have a prior invariant under scaling etc. Perhaps the biggest problem with uninformative priors is that there's really no way to represent total ignorance. Saying that a prior is flat over the real line is arguably a very strong statement. It says that a priori you believe there's a very significant possibility that the value is arbitrarily large, for example.

6.4 Decision theory and loss functions

6.4.1 Setting

Bayesian statistics can be used to help make choices in the face of uncertainty. Economists talk about making decisions to maximize your utility. Statisticians are more pessimistic, preferring to talk about minimizing losses. You are perceived to be losing money because if you knew the true parameters you could make better decisions. Here we present a classical setting for assessing the quality of estimators in frequentist and Bayesian settings, based on decision theory and loss functions.

1. Let Ω be the space of parameters θ.
2. Let $\mathcal{L} : \Omega \times \Omega \to \mathbb{R}$ denote the loss function.
3. Let $X_1, \dots X_n$ be a random sample and denote $X^n = (X_1, \dots, X_n) \in \mathbb{R}^n$.
4. Let $\hat{\theta}(X_1, \dots, X_n) \in \Omega$ denote, generically, an estimator of θ.

Example 6. Three important loss functions are:

1. $\mathcal{L}(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$. Known as modular loss or absolute error loss.
2. $\mathcal{L}(\theta, \hat{\theta}) = \big(\theta - \hat{\theta}\big)^2$. Squared error loss.
3. $\mathcal{L}(\theta, \hat{\theta}) = 1$ if $\theta \neq \hat{\theta}$ and $\mathcal{L}(\theta, \hat{\theta}) = 0$ if $\theta = \hat{\theta}$. Zero-one loss. □

We define the <u>risk</u> of an estimator $\hat{\theta} = \hat{\theta}(X^n)$ by

$$R(\theta, \hat{\theta}) := \mathbb{E}_{X^n|\theta}[\mathcal{L}(\theta, \hat{\theta})] = \int \mathcal{L}(\theta, \hat{\theta}(x^n)) \underbrace{f(x^n; \theta)}_{\text{likelihood} X^n|\theta} dx^n.$$

where the expectation $\mathbb{E}_{X^n|\theta}$ denotes expectation over X^n under the assumption that X^n is generated from $f(X^n; \theta)$. In particular, note that under square loss

$$R(\theta, \hat{\theta}) = \mathbb{E}_{X^n|\theta}\Big[\big(\theta - \hat{\theta}\big)^2\Big] = \text{MSE}(\hat{\theta}).$$

Example 7. Consider $X|\theta \sim N(\theta, 1)$ and squared error loss. Let's compare the risk of the estimators $\hat{\theta}_1(X) = X$ with $\hat{\theta}_2(X) = 5$.

$$R(\theta, \hat{\theta}_1) = \mathbb{E}_{X^n|\theta}[(X-\theta)^2] = 1,$$
$$R(\theta, \hat{\theta}_2) = \mathbb{E}_{X^n|\theta}[(5-\theta)^2] = (5-\theta)^2.$$

Neither estimator always wins.

There are two general principles for picking good estimators, leading to minimax estimators in frequentist settings and Bayes estimators in Bayesian settings. We formalize these concepts in the following two definitions:

Definition 1. The maximum risk is

$$\bar{R}(\hat{\theta}) = \max_{\theta} R(\theta, \hat{\theta}).$$

A minimax estimator minimizes the maximum risk: $\hat{\theta}$ is minimax if and only if

$$\bar{R}(\hat{\theta}) = \inf_{\widetilde{\theta}} \bar{R}(\widetilde{\theta}).$$

Definition 2. Given a prior $f_\Theta(\theta)$ on the parameter θ, the Bayes risk of an estimator $\hat{\theta}$ is defined by

$$r(\hat{\theta}) = \int R(\theta, \hat{\theta}) f_\Theta(\theta) d\theta.$$

A Bayes rule (or Bayes estimator) minimizes Bayes risk: $\hat{\theta}$ is called a Bayes rule if

$$r(\hat{\theta}) = \min_{\widehat{\theta}} r(\widehat{\theta}).$$

Example 8. Suppose $X_1, \ldots, X_n \sim$Bernoulli(θ) and consider squared error loss. Let $\hat{\theta}_1 := \bar{X}$, and let $\hat{\theta}_2^{\alpha,\beta} := \frac{\alpha + n\bar{X}}{\alpha+\beta+n}$ be the posterior mean when the prior is Beta(α, β). We then have (check it!) that

$$R(\theta, \hat{\theta}_1) = \mathrm{Var}(\hat{\theta}_1) = \theta(1-\theta)/n,$$
$$R(\theta, \hat{\theta}_2^{\alpha,\beta}) = \frac{n\theta(1-\theta) + (\alpha - (\alpha+\beta)\theta)^2}{(\alpha+\beta+n)^2}.$$

The estimator $\hat{\theta}_2$ with the choices $\alpha = \beta = \sqrt{n}/2$ is interesting as then

$$R(\theta, \hat{\theta}_2) = \frac{n}{4(n+\sqrt{n})^2},$$

which does not depend on θ. Taking maximums over θ we then get

$$\bar{R}(\hat{\theta}_1) = 1/(4n),$$
$$\bar{R}(\hat{\theta}_2^{\sqrt{n}/2, \sqrt{n}/2}) = \frac{n}{4(n+\sqrt{n})^2}.$$

In terms of maximum risk, $\hat{\theta}_2^{\sqrt{n}/2,\sqrt{n}/2}$ is slightly better. For large n the difference is very small. The danger of minimax is that you might choose a worse estimator due to a small part of the sample space. The advantage is you don't have to choose a prior.

Now consider Bayes risk with respect to a uniform prior, $f(\theta) = 1$, $0 < \theta < 1$. Then

$$r(\hat{\theta}_1) = \int R(\theta, \hat{\theta}_1)\, d\theta = \int \theta(1-\theta)/n \, d\theta = 1/(6n),$$
$$r(\hat{\theta}_2^{\sqrt{n}/2,\sqrt{n}/2}) = \int R(\theta, \hat{\theta}_2)\, d\theta = \frac{n}{4(n+\sqrt{n})^2}.$$

We shall see in the next subsection that the best Bayes risk with squared error loss and a given prior is achieved by the posterior mean computed with the same prior. In our case, we are considering the prior $f_\Theta(\theta) \equiv$Unif$(0,1) \equiv$Beta$(1,1)$. With this choice

$$R(\theta, \hat{\theta}^{1,1}) = \frac{(4-n)\theta^2 + (n-4)\theta + 1}{(2+n)^2},$$

and

$$r(\hat{\theta}_2^{1,1}) = \int R(\theta, \hat{\theta}^{1,1}) f_\Theta(\theta)\, d\theta = \frac{1}{6(n+2)}.$$

We can see that the Bayes risk with uniform prior of the estimator $\frac{1+n\bar{X}}{2+n}$ is slightly smaller than for the estimator $\bar{X}$.

6.4.2 Bayes estimators

The posterior risk of an estimator $\hat{\theta}$ is

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

Theorem 9. *Let $\hat{\theta} = \hat{\theta}(x^n)$ be the value of θ that minimizes the posterior risk $r(\hat{\theta}|x^n)$. Then $\hat{\theta}$ is the Bayes estimator.*

Proof.

$$\begin{aligned} r(\hat{\theta}) &= \int R(\theta, \hat{\theta}) f_\Theta(\theta)\, d\theta \\ &= \int\int \mathcal{L}(\theta, \hat{\theta}) f(x|\theta) f_\Theta(\theta)\, dx^n\, d\theta \\ &= \int\int \mathcal{L}(\theta, \hat{\theta}) f(\theta|x^n) f(x^n)\, d\theta\, dx^n \\ &= \int r(\hat{\theta}|x^n) f(x^n)\, dx^n. \end{aligned}$$

Choosing $\hat{\theta}$ to minimize the integrand minimizes the integral. □

Theorem 10. *Bayes estimators for some important loss functions:*

- *Under square error loss the Bayes estimator is the posterior mean.*
- *Under absolute loss the Bayes estimator is the median of the posterior distribution.*
- *Under zero-one loss the Bayes estimator is the mode of the posterior distribution*

Proof. We only prove the result for square error loss. The Bayes estimator $\hat{\theta}$ must minimize the posterior risk $r(\hat{\theta}|x^n)$. Therefore

$$\frac{\partial}{\partial\hat{\theta}} r(\hat{\theta}|x^n) = \frac{\partial}{\partial\hat{\theta}} \int (\theta - \hat{\theta})^2 f(\theta|x^n) d\theta = 0.$$

Thus,

$$2\int (\theta - \hat{\theta}) f(\theta|x^n) d\theta = 0,$$

and

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

□

Example 9. Let $X_1, \ldots X_n \sim$Bernoulli(θ) with Beta(α, β) prior belief. Under squared error loss the Bayes estimator is the posterior mean

$$\hat{\theta} = \frac{\alpha + n\bar{X}}{\alpha + \beta + n}.$$

Exercises

Exercise 1 Let $X_1, \ldots, X_n \sim N(\theta, \theta)$, $\theta > 0$, and consider squared error loss. Let

$$\hat{\theta}_1 = \bar{X} \qquad \hat{\theta}_2 = S^2$$

be the sample mean and the sample variance.

a) Calculate the risks of $\hat{\theta}_1$ and $\hat{\theta}_2$ as a function of θ.

b) Give an example of a prior distribution under which $\hat{\theta}_1$ has lower Bayes risk for all $n \geq 2$.

Exercise 2 Let $X_1, \ldots, X_n | \Theta = \theta$ be i.i.d. Geometric with parameter θ. Is Jeffrey's prior an improper prior?

Exercise 3 Recall that the $m-1$ simplex is the subset of $\mathbb{R}^m$ made of points $(\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$ that satisfy

$$\sum_{i=1}^{m} \theta_i = 1, \qquad \theta_i \geq 0.$$

Given an m-dimensional vector $\alpha = (\alpha_1, \ldots, \alpha_m)$ of parameters, we say that a random vector $(\Theta_1, \ldots, \Theta_n)$ taking values on the $m-1$ simplex has Dirichlet distribution with parameter α, written $\Theta \sim$Dir(α), if Θ has PDF in $\mathbb{R}^{m-1}$ given (up to normalization constant) by

$$f(\theta_1, \ldots, \theta_m | \alpha) \propto \theta_1^{\alpha_1 - 1} \times \cdots \times \theta_m^{\alpha_m - 1}.$$

Note that the Beta distribution is a particular case of the Dirichlet distribution, corresponding to $m = 2$.

Introduce the notation $\Theta := (\Theta_1, \ldots, \Theta_m)$ and $X := (X_1, \ldots, X_m)$ and consider the following Bayesian model:

$$\begin{aligned} \Theta &\sim \text{Dirichlet}(\alpha), \\ X | \Theta = (\theta_1, \ldots, \theta_m) &\sim \text{Multinomial}(n, \theta_1, \ldots, \theta_m). \end{aligned}$$

Find the posterior distribution on Θ.

Exercise 4 This exercise illustrates (once more) the concept of improper priors. Consider the prior beliefs given by $f_\Theta(\theta) = 1$ for all $\theta \in \mathbb{R}$. Clearly, $\int_{\mathbb{R}} f_\Theta(\theta) d\theta = \infty$. Consider the Bayesian model

$$\begin{aligned} &\text{Prior} : f_\Theta(\theta), \\ &\text{Likelihood} : X | \Theta = \theta \sim N(\theta, 1). \end{aligned}$$

Show that the posterior is a Normal distribution.

Exercise 5 Show that the median is the Bayes estimator under absolute error loss.

Exercise 6 Suppose that $X_1, \ldots, X_n \sim$Bernoulli(p) and your prior beliefs about p are that

$$p = \begin{cases} 0.25 & \text{with probability } 0.25, \\ 0.5 & \text{with probability } 0.5, \\ 0.75 & \text{with probability } 0.25. \end{cases}$$

Calculate the maximum a posteriori estimate, that is, the maximum value of the posterior PMF in terms of your sample.

Further Exercises

Exercise 1 Let $X_1, \ldots, X_n \sim$ Uniform(0,1/θ). Take the prior to be $f(\theta) = 1/\theta$ $(\theta > 0)$. Find the posterior distribution and the posterior mean.

Exercise 2 Your prior belief about θ is that $\theta \sim N(a, b^2)$. You will observe samples $X_1, \ldots X_n$ from the distribution $N(\theta, \sigma^2)$ where σ is known.

(i) Find the posterior distribution. Try to express your answer as simply as possible.

(ii) Calculate the limit of the posterior distribution as $n \to \infty$ with a, b and σ fixed.

(iii) Let $a = 0, b = 10$. Setting $\theta = 100, \sigma = 1$ and $n = 8$, generate a sample $X_1, \ldots, X_n \sim N(100, 1)$ using R. Calculate the posterior distribution for the observed sample. Find a 98% credible interval for θ by calculating the 1st and 99th quantiles of the posterior distribution.

(iv) Repeat part (iii) but $b = 1/10$. Comment on the difference.

Exercise 3 An experiment will produce a result X with distribution Binomial$(100, p)$ where p is unknown. A scientist tells you that their prior belief is $p \sim$ Beta(1000, 1000). You observe a sample $x = 2$.

(i) Calculate a posterior 95% credible interval for p.

(ii) Find numerically the lengths of the shortest and longest 95% credible intervals. [Hint: Use qbeta()]

Exercise 4 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Poisson(λ) where λ is unknown.

(a) Calculate the Jeffreys prior for the Poisson(λ) distribution. Calculate the posterior distribution with respect to the Jeffrey's prior.

(b) After taking expert advice, you decide instead that your prior beliefs about λ are given by a Gamma distribution with mean 100 and variance 200.

- Calculate the posterior distribution.
- Calculate the posterior mean.
- Calculate the Bayes estimator with respect to absolute error loss when $n = 4$, $X_1 = 57$, $X_2 = 104$, $X_3 = 97$ and $X_4 = 120$. (You may use R for this part.)

Exercise 5 Let $X_1, \ldots X_{100} \sim N(\theta, 1)$ and take your prior belief about θ to be that $\theta \sim N(0, 1)$. You observe a sample average of 0.2 . Test the hypothesis $H_0 : \theta < 0$ against $H_1 : \theta > 0$ with test sizes of $10\%, 5\%$ and 2%.