Imperial College London

M3/4S1 Statistical Theory

Kolyan Ray

Spring 2023

*Please send all comments and corrections to: kolyan.ray@ic.ac.uk

Contents

0 Review **1**
- 0.1 Probability and conditional distributions . . . 1
- 0.2 Order statistics . . . 2
- 0.3 Convergence of random variables . . . 2

1 Principles of point estimation **4**
- 1.1 Statistical models . . . 4
- 1.2 Estimators . . . 6
- 1.3 Method of Moments . . . 8
- 1.4 Sufficiency . . . 9
- 1.5 Rao-Blackwell theorem . . . 12

2 Likelihood-based estimation **15**
- 2.1 The likelihood Function . . . 15
- 2.2 Geometry of the likelihood: score and Fisher information . . . 18
- 2.3 The Cramer-Rao lower bound . . . 21
- 2.4 Numerical computation of MLEs (Non-examinable) . . . 23

3 Asymptotic theory for MLEs **25**
- 3.1 Consistency . . . 25
- 3.2 Asymptotic normality of the MLE . . . 27
- 3.3 Asymptotic efficiency and the delta method . . . 29

4 Bayesian inference **32**
- 4.1 Priors and posteriors . . . 32
- 4.2 Jeffreys priors . . . 35
- 4.3 Frequentist analysis of Bayesian methods . . . 36

5 Optimality in Estimation **38**
- 5.1 Decision Theory . . . 38
- 5.2 Bayes risk and minimax risk . . . 39
- 5.3 Minimum variance unbiased estimators . . . 42
- 5.4 Complete statistics . . . 44

6 Hypothesis testing and confidence intervals **50**
- 6.1 Hypothesis testing . . . 50
- 6.2 Uniformly most powerful tests . . . 51
- 6.3 Likelihood ratio tests . . . 56
- 6.4 Confidence Intervals . . . 59

Statistics is a set of principles and procedures for collecting and processing quantitative data in order to help make decisions. This course is concerned with presenting some of the mathematical principles behind formal statistical theory. We will assume we have some data generated from some unknown probability distribution, and we aim to use this data to learn certain properties of this distribution.

This course will concern *parametric inference*. We assume that we have a random variable X drawn from a member of a known family of probability distributions indexed by a finite-dimensional parameter (e.g. Poisson distributions). However, the parameter of the distribution is unknown and we aim to estimate it from the data. For example, we might know that $X \sim \text{Poisson}(\lambda)$ for some unknown $\lambda > 0$ and we wish to estimate λ.

Usually, we repeat the experiment or observations multiple times, and hence typically observe $X_1, \ldots, X_n$ independent and identically distributed (i.i.d) copies of X. Suppose we know the true distribution lies in a statistical model $\{P_\theta : \theta \in \Theta\}$. Some of the main goals of statistical inference about the parameter θ are:

- *Estimation*: construct an estimate $\hat{\theta}(X_1, \ldots, X_n)$ of the true value of θ.
- *Hypothesis testing*: construct a test to determine between two (or more) hypotheses concerning θ, e.g. whether $\theta = 0$ or $\theta \neq 0$.
- *Uncertainty quantification*: give an interval estimate or set of plausible values for θ, e.g.

$$\hat{C} = [\hat{\theta}_1(X_1, \ldots, X_n), \hat{\theta}_2(X_1, \ldots, X_n)].$$

0 Review

0.1 Probability and conditional distributions

For a real-valued random variable X on a probability space $(\Omega, \mathcal{F}, P)$, we define the cumulative distribution function $F : \mathbb{R} \to [0, 1]$ by

$$F(x) = P(X \leq x) = P(\omega \in \Omega : X(\omega) \leq x).$$

When X is discrete, we have

$$F(x) = \sum_{k \leq x} f(k),$$

where $f(k) = P(X = k)$ is the *probability mass function* (pmf) of X. When X is continuous and has a probability density function (pdf) $f : \mathbb{R} \to [0, \infty)$, then

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

The *conditional distribution* of random variables $X_1, \cdots, X_n$ given an event A satisfying $P(A) > 0$ is defined as

$$P(X_1 \leq x_1, \cdots, X_n \leq x_n | A) = \frac{P(X_1 \leq x_1, \cdots, X_n \leq x_n, A)}{P(A)}.$$

If A is the outcome of some random variable Y, then

$$P(X_1 \leq x_1, \cdots, X_n \leq x_n | Y = y) = \frac{P(X_1 \leq x_1, \cdots, X_n \leq x_n, Y = y)}{P(Y = y)},$$

whenever $P(Y = y) > 0$. The conditional expectation of X given $Y = y$ is defined as

$$E(X|Y = y) = \begin{cases} \sum_x xP(X = x|Y = y) & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} xf(x|y)\, \mathrm{d}x & \text{if } X \text{ is continuous,} \end{cases}$$

where $f(x|y) = \frac{f(x,y)}{f(y)}$ is the *conditional density* of X given $Y = y$, $f(x, y)$ is the *joint density* of X and Y, and $f(y)$ is the *marginal density* of Y.

We will use the *tower rule* for conditional expectations: if X, Y are random variables,

$$E[X] = E[E[X|Y]].$$

Moreover, if $Y = f(X)$ is a function of X, then for any random variable Z

$$E[Z|Y] = E[E[Z|X]|Y].$$

We will also use the *conditional variance formula*:

$$\text{var}(Y) = E[\text{var}(Y|X)] + \text{var}(E[Y|X]).$$

0.2 Order statistics

Suppose $X_1, \ldots, X_n$ are i.i.d random variables with common distribution function F. If we order the (X_i) so that

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)},$$

then these ordered variables are called the *order statistics* of $X_1, \ldots, X_n$. The distribution of $X_{(n)} = \max_i X_i$ is

$$F_{X_{(n)}}(t) = P(X_{(n)} \leq t) = P(X_1 \leq t, \ldots, X_n \leq t) = \prod_{i=1}^{n} P(X_i \leq t) = F(t)^n,$$

while the distribution of $X_{(1)} = \min_i X_i$ is

$$F_{X_{(1)}}(t) = P(X_{(1)} \leq t) = 1 - P(X_{(1)} > t) = 1 - (1 - F(t))^n.$$

For continuous variables, differentiating gives the density function of $X_{(n)}$,

$$f_{X_{(n)}}(t) = nF(t)^{n-1}f(t)$$

and of $X_{(1)}$,

$$f_{X_{(1)}}(t) = n(1 - F(t))^{n-1}f(t),$$

where f is the density of X_i.

0.3 Convergence of random variables

Let $X_1, X_2, \ldots$ and X be random variables having cumulative distribution functions $F_{X_1}, F_{X_2}, \ldots$ and F_X, respectively.

1. *Convergence in probability*: We say X_n converges to X in probability, written $X_n \to^P X$, if for every $\epsilon > 0$,
$$P(|X_n - X| \geq \epsilon) \to 0 \qquad \text{as } n \to \infty.$$
2. *Convergence in distribution*: We say X_n converges to X in distribution, written $X_n \to^d X$, if
$$F_{X_n}(x) \to F_X(x) \qquad \text{as } n \to \infty$$
for all $x \in \mathbb{R}$ at which F_X is *continuous*.

These notions of convergence are related via the following relationship:

$$X_n \to^P X \implies X_n \to^d X.$$

Remark. *These notions extend to vectors $(X_n), X$ in $\mathbb{R}^k$. For convergence in probability, replace $|X_n - X|$ by the Euclidean norm $\|X_n - X\|$. For convergence in distribution, writing $X = (X^1, \ldots, X^k)$, replace the cdf $P(X \leq x)$ with its vector analogue $P(X^1 \leq x_1, \ldots, X^k \leq x_k)$.*

We recall some useful results.

Lemma 0.1 (Markov's inequality). *If X is a real-valued random variable, then for any $t > 0$,*

$$P(|X| \geq t) \leq \frac{E|X|}{t}.$$

Theorem 0.1 (Slutsky's theorem). *Let $(X_n), (Y_n), X$ be random variables. If $X_n \to^d X$ and $Y_n \to^p c$ for a constant c, then*

$$X_n + Y_n \to^d X + c, \qquad X_n Y_n \to^d cX, \qquad X_n / Y_n \to^d X/c.$$

An important property of continuous functions is that they preserve limits. The same holds for probabilistic convergence.

Theorem 0.2 (Continuous mapping theorem). *Let $(X_n), X$ be random variables in $\mathbb{R}^k$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ be a continuous function such that $P(g \text{ is continuous at } X) = 1$. Then*

- *$X_n \to^d X$ implies $g(X_n) \to^d g(X)$,*
- *$X_n \to^P X$ implies $g(X_n) \to^P g(X)$.*

Many estimators in statistics are based on, or related to, the mean of i.i.d. random variables. A very important result regarding their convergence is the law of large numbers.

Proposition 0.1 (Weak law of large numbers). *Let $X_1, X_2, \ldots$ be i.i.d random variables with $E|X_i| < \infty$. Then as $n \to \infty$,*

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \to^P E[X_1].$$

One can also quantify the stochastic fluctuations of $\bar{X}_n$ around its expectation. These are of order $1/\sqrt{n}$ and look normally distributed whenever $\text{var}(X_1) = \sigma^2 < \infty$.

Theorem 0.3 (Central limit theorem). *Let $X_1, X_2, \ldots$ be i.i.d random variables with $\text{var}(X_i) = \sigma^2 < \infty$. Then as $n \to \infty$,*

$$\sqrt{n}(\bar{X}_n - E[X_1]) \to^d N(0, \sigma^2).$$

1 Principles of point estimation

1.1 Statistical models

Consider a random variable X taking values in some sample space $\mathcal{X}$ (e.g. $\mathcal{X} = \mathbb{R}$ or $\mathcal{X} = \mathbb{R}^k$) coming from some probability distribution P_θ, which is parametrized by an unknown parameter $\theta \in \mathbb{R}^p$.

Definition. *A statistical model for X is any family $\{P_\theta : \theta \in \Theta\}$ of probability distributions P_θ for the distribution of X. The set Θ is called the parameter space.*

If X is discrete/continuous with pmf/pdf f_θ, one can equivalently write the statistical model as

$$\{f_\theta : \theta \in \Theta\}.$$

In this course, we typically take $X = (X_1, \ldots, X_n)$ to be a *sample* of n independent and identically distributed (i.i.d) copies $X_1, \ldots, X_n$ of a real-valued random variable X_1. We then write

$$X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} P_\theta.$$

In the i.i.d. setting, it is equivalent to consider the pmf/pdf of X_1 instead of $X = (X_1, \ldots, X_n)$, since $f_{\theta,X}(x) = \prod_{i=1}^n f_{\theta,X_1}(x_i)$.

We will take $\Theta \subseteq \mathbb{R}^p$, so that θ can be a scalar or vector of parameters. When p is finite, this is known as a *parametric model*. There exist nonparametric models, where Θ is infinite-dimensional, but these are beyond the scope of this course.

Example. *If $X = (X_1, \cdots, X_n)$ with $X_i \overset{\text{i.i.d}}{\sim}$ Poisson(λ), then the joint pmf of X is*

$$f_\lambda(x) = \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}.$$

Here $\theta = \lambda$ with parameter space $\{\lambda : \lambda > 0\}$.

Example. *If $X = (X_1, \cdots, X_n)$ with $X_i \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$, then the joint pdf of X is*

$$f_{\mu,\sigma^2}(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}.$$

Here $\theta = (\mu, \sigma^2)$ is a two-dimensional vector of parameters with parameter space

$$\{(\mu, \sigma^2) : -\infty < \mu < \infty,\ \sigma^2 \geq 0\}.$$

A necessary condition to estimate θ based on data $X \sim P_\theta$ is that the model parameters θ can be identified from the probability distribution P_θ.

Definition. *A statistical model $\{P_\theta : \theta \in \Theta\}$ is identifiable if $P_{\theta_1} = P_{\theta_2}$ implies $\theta_1 = \theta_2$ for all $\theta_1, \theta_2 \in \Theta$.*

One should keep in mind we are really trying to estimate the distribution P_θ based on an observation $X \sim P_\theta$, and θ is just a convenient parametrization (labelling) of P_θ. Thus estimating θ (the label) only makes sense when the parametrization $\theta \mapsto P_\theta$ itself makes sense, i.e. is injective. Estimation in nonidentifiable models is not well-defined, since even if we exactly recover the true distribution P_θ, there are then multiple possible equivalent labels θ that can be used. All statistical models in this course are identifiable.

A natural and widely used class of parametric distributions are exponential families.

Definition. *A family of distributions $\{P_\theta : \theta \in \Theta\}$ is a k-parameter exponential family if its pmf/pdf takes the form:*

$$f_\theta(x) = \exp\left\{ \sum_{i=1}^{k} c_i(\theta) T_i(x) - d(\theta) + S(x) \right\},$$

where $c_i(\theta)$, $T_i(x)$, $d(\theta)$ and $S(x)$ are known functions, and the support of f_θ (i.e. the set of all x for which $f_\theta(x) > 0$) does not depend on θ.

When considering examples, it is important to specify which values of a distribution are known and which are considered parameters (i.e. allowed to vary).

Example. *If $X \sim Bin(n, \theta)$ with n known, then*

$$\begin{aligned} f_\theta(x) &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &= \binom{n}{x} \left(\frac{\theta}{1-\theta} \right)^x (1-\theta)^n \\ &= \binom{n}{x} \exp\left\{ x \log\left(\frac{\theta}{1-\theta} \right) \right\} \exp\left\{ n \log(1-\theta) \right\} \\ &= \exp\left\{ \underbrace{\log\left(\frac{\theta}{1-\theta} \right)}_{c(\theta)} \underbrace{x}_{T(x)} + \underbrace{n \log(1-\theta)}_{-d(\theta)} + \underbrace{\log \binom{n}{x}}_{S(x)} \right\}. \end{aligned}$$

Since $x \in \{0, 1, \cdots, n\}$ does not depend on θ, the Binomial distributions with n known form a (1-parameter) exponential family.

Example. *One can directly show that the normal distributions with parameters $\theta = (\mu, \sigma^2)$ form a (2-parameter) exponential family. More generally, let $X = (X_1, \cdots, X_n)$ with $X_i \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$. Then the joint pdf of X equals*

$$\begin{aligned} f_\theta(x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \\ &= \exp\left\{ \underbrace{-\frac{1}{2\sigma^2}}_{c_1(\theta)} \underbrace{\sum_{i=1}^{n} x_i^2}_{T_1(x)} + \underbrace{\frac{\mu}{\sigma^2}}_{c_2(\theta)} \underbrace{\sum_{i=1}^{n} x_i}_{T_2(x)} \underbrace{- \frac{n\mu^2}{2\sigma^2} - \frac{n}{2} \log(2\pi\sigma^2)}_{-d(\theta)} \right\}, \end{aligned}$$

which forms a 2-parameter exponential family.

Example. *If $X = (X_1, \cdots, X_n)$ with $X_i \overset{\text{i.i.d}}{\sim} U(0, \theta)$, then the joint density of X is*

$$f_\theta(x) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{(0,\theta)}(x_i),$$

where $\mathbb{1}_A(x_1)$ is the indicator function of a set A, i.e.

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Since

$$\mathbb{1}_{(0,\theta)}(x_1)\cdots\mathbb{1}_{(0,\theta)}(x_n) = \mathbb{1}_{(0,\theta)}(\min_i x_i)\mathbb{1}_{(0,\theta)}(\max_i x_i),$$

we can rewrite the joint density as

$$f_\theta(x) = \frac{1}{\theta^n}\mathbb{1}_{(0,\theta)}(\min_i x_i)\mathbb{1}_{(0,\theta)}(\max x_i).$$

We conclude that this is not an exponential family because the support of f_θ is $[0,\theta]^n$, which depends on θ.

Exponential families include many other common parametric families. We will return to them at various parts during the course.

1.2 Estimators

Given an i.i.d sample $X = (X_1, \ldots, X_n)$ and the knowledge their probability distribution takes the form P_θ for some unknown θ, the goal of estimation is to construct an estimator $\hat{\theta}(X) = \hat{\theta}(X_1, \ldots, X_n)$ for θ based on the data X. We often write $\hat{\theta}_n$ to make explicit the dependence on the sample size n. We now recall various definitions and facts from M2S2 Statistical Modelling I.

Definition. *A statistic is any function $T(X)$ of the observed data X. The distribution of $T(X)$ is called the sampling distribution of the statistic.*

Example. *$T(X) = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $T(X) = (X_{(1)}, \cdots, X_{(n)})$ are both statistics.*

Although a statistic is a function of the data, it may contain no information about the parameter θ.

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, 1)$. Writing $X_i = \mu + Z_i$ for $Z_i \overset{\text{i.i.d}}{\sim} N(0,1)$, we have*

$$T(X) = \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n (Z_i - \bar{Z}_n)^2,$$

which does not depend on μ.

Which statistic $\hat{\theta}(X)$ should one use to estimate θ? One way to decide if an estimator is reasonable is to look at its *bias*.

Definition. *Let $\hat{\theta} = T(X)$ be an estimator of θ. The bias of $\hat{\theta}$ is defined as*

$$b_\theta(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta.$$

If $b_\theta(\hat{\theta}) = 0$ for all $\theta \in \Theta$, then $\hat{\theta}$ is an unbiased estimator of θ.

The notation E_θ means taking the expectation under the probability distribution $X \sim P_\theta$ (recall $\hat{\theta} = \hat{\theta}(X)$ is a function of X). To evaluate $E_\theta[\hat{\theta}]$, we can either find the distribution of $\hat{\theta}$ and find its expected value, or evaluate T as a function of X directly, and find its expected value.

Not all estimators $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ are unbiased for a fixed sample size n. However, we would expect reasonable estimators to be asymptotically unbiased:

$$E_\theta[\hat{\theta}_n] \to \theta \qquad \text{as } n \to \infty.$$

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} U(0,\theta)$. The estimator $X_{(n)} = \max_{1\le i\le n} X_i$ is not unbiased for θ. Indeed,*

$$P(X_{(n)} \le t) = \prod_{i=1}^n P(X_i \le t) = (t/\theta)^n, \qquad 0 < t < \theta,$$

giving density function $f_{X_{(n)}}(t) = \frac{n}{\theta^n} t^{n-1} \mathbb{1}_{(0,\theta)}(t)$. *Therefore,*

$$E_\theta[X_{(n)}] = \int_0^\theta t \frac{n}{\theta^n} t^{n-1} \, \mathrm{d}t = \frac{n}{n+1}\theta \neq \theta,$$

so that $X_{(n)}$ *is biased for any* $n \geq 1$. *However, it is asymptotically unbiased since* $E_\theta[X_{(n)}] \to \theta$ *as* $n \to \infty$.

Being unbiased does not necessarily imply an estimator is good.

Definition. *The mean squared error (MSE) of an estimator* $\hat{\theta}$ *is*

$$\mathrm{MSE}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2].$$

We can rewrite the MSE using the *bias-variance decomposition* (see M2S2):

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{var}_\theta(\hat{\theta}) + b_\theta(\hat{\theta})^2.$$

A good unbiased estimator should also have small MSE. While unbiasedness is sometimes important, it is possible for a biased estimator to have smaller MSE than an unbiased estimator.

Example. *If* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, 1)$, *both the estimators* $T_1(X) = X_1$ *and* $T_2(X) = \frac{1}{n}\sum_{i=1}^n X_i$ *are unbiased. However,* $\mathrm{var}_\mu(T_1) = 1$, *while* $\mathrm{var}_\mu(T_2) = 1/n$, *so that* $\mathrm{MSE}_\mu(T_1) = 1 > 1/n = \mathrm{MSE}_\mu(T_2)$. *Clearly,* T_2 *makes better use of the data than* T_1, *even though they are both unbiased.*

Unbiasedness is not preserved by transformations, that is, if $\hat{\theta}$ is unbiased for θ, then $g(\hat{\theta})$ is not necessarily unbiased for $g(\theta)$.

Example. *If* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, 1)$, *we know* $\bar{X}_n$ *is unbiased for* θ. *However,* $\bar{X}_n^2$ *is biased for* θ^2 *since* $E[\bar{X}_n^2] = \mathrm{var}\,(\bar{X}) + (E(\bar{X}))^2 = \frac{1}{n} + \theta^2 \neq \theta^2$. *Similarly, if* $Y_1, \ldots, Y_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$, *one can show that* $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ *is unbiased for* σ^2, *but* $S = \sqrt{S^2}$ *is not unbiased for* σ.

Definition. *The standard error of an estimator* $\hat{\theta} = \hat{\theta}(X)$ *is the standard deviation of its sampling distribution:*

$$se(\hat{\theta}) = \sqrt{var(\hat{\theta})}.$$

The standard error often depends on the unknown parameter θ being estimated, but it can usually be estimated.

Example. *If* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$, *then the estimator* $\hat{\theta} = \bar{X}_n$ *has standard error* $se(\hat{\theta}) = se(\bar{X}_n) = \sigma/\sqrt{n}$. *If* σ^2 *is unknown, we can estimate the standard error by* $\hat{se} = S/\sqrt{n}$, *where* $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

We now run through these definitions with an example.

Example. *Suppose* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ *Binomial*$(1, \theta)$ *and let* $\hat{\theta}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

- *Bias:* $E_\theta\hat{\theta}_n = E_\theta[\bar{X}_n] = \theta$, *so* $\hat{\theta}_n$ *is unbiased.*
- *Standard error:* $\mathrm{var}_\theta(\hat{\theta}_n) = \theta(1-\theta)/n$, *so* $se(\hat{\theta}_n) = \sqrt{\theta(1-\theta)/n}$. *Since this depends on the unknown* θ, *we can estimate it by* $\hat{se}(\hat{\theta}_n) = \sqrt{\hat{\theta}_n(1-\hat{\theta}_n)/n} = \sqrt{\bar{X}_n(1-\bar{X}_n)/n}$.
- *MSE: using the bias-variance decomposition,* $MSE_\theta(\hat{\theta}_n) = \mathrm{var}_\theta(\hat{\theta}_n) = \theta(1-\theta)/n$.
- *Sampling distribution: in this example, we can exactly work out that the distribution of* $n\hat{\theta}$ *is Binomial*(n, θ). *We can also use the central limit theorem to work out the asymptotic sampling distribution: since*

  $$\sqrt{n}(\hat{\theta}_n - \theta) \to^d N(0, \theta(1-\theta)),$$

  we have $\hat{\theta}_n \approx N(\theta, \theta(1-\theta)/n)$ *for large* n.

1.3 Method of Moments

The method of moments (MM) is a well-known method for estimating the parameters in a parametric statistical model $\{P_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^p$. Define the k^{th} *population moment*

$$\mu_k = \mu_k(\theta) = E_\theta[X^k],$$

which is a function of θ computed based on the distribution P_θ. The corresponding k^{th} *sample moment* is

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

For example, μ_1 is the population mean, while $\hat{\mu}_1 = \bar{X}_n$ is the sample average. Note that $\hat{\mu}_k$ is a random variable, while μ_k is a constant depending on the population parameters. The method of moments (MM) estimator $\hat{\theta}_{MM}$ is obtained by equating the population and sample moments, i.e. solving the equations

$$\mu_k(\hat{\theta}_{MM}) = \hat{\mu}_k, \qquad k = 1, \ldots, p.$$

Note, there are p equations for p unknowns.

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$, we know that $E_\theta[X] = \mu$ and $E_\theta[X^2] = \mu^2 + \sigma^2$. Hence, the MM estimator of (μ, σ^2) satisfies the equations*

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

which are solved by

$$\hat{\mu}_{MM} = \bar{X}_n, \qquad \hat{\sigma}^2_{MM} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \Gamma(\alpha, \lambda)$, we know that $E_\theta[X] = \frac{\alpha}{\lambda}$ and $E_\theta[X^2] = \frac{\alpha^2}{\lambda^2} + \frac{\alpha}{\lambda^2}$. Hence, the MM estimator of (α, λ) satisfies the equations*

$$\frac{\alpha}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \frac{\alpha^2}{\lambda^2} + \frac{\alpha}{\lambda^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

These can be rewritten as

$$\alpha = \lambda\hat{\mu}_1, \qquad \alpha^2 + \alpha = \lambda^2\hat{\mu}_2,$$

which are solved by

$$\hat{\alpha}_{MM} = \frac{\bar{X}_n^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}, \qquad \hat{\lambda}_{MM} = \frac{\bar{X}_n}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}.$$

We will see later in the module that these estimators are not always optimal, but they are often easy to compute. They are also useful as initial values for other methods that require iterative numerical routines to compute. It can be proven that in general, method of moments estimators have a lot of interesting and desirable asymptotic properties, but this is beyond the scope of the module (see e.g. Chapter 4 of *Asymptotic Statistics* by A.W. van der Vaart).

1.4 Sufficiency

We often run experiments just to find out the value of θ. We might not be interested in the data points themselves and just want to understand the general population behaviour. This motivates the concept of a *sufficient statistic*, which contains all the information about θ present in our sample.

Definition. *A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of X given $T(X)$ does not depend on θ.*

This says that for any set A,

$$P(X \in A | T = t) = \text{free of } \theta.$$

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Poisson(θ), then $T(X) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ because*

$$\begin{aligned} P(X = x | T = t) &= \frac{P(X = x, T = t)}{P(T = t)} \\ &= \frac{P(X_1 = x_1, \cdots, X_n = x_n,\ X_1 + \cdots + X_n = t)}{P(T = t)} \\ &= \begin{cases} \dfrac{P(X_1 = x_1, \cdots, X_n = x_n)}{P(T = t)} & \text{if } t = x_1 + \cdots + x_n, \\ 0 & \text{otherwise}, \end{cases} \end{aligned}$$

and for $x_1 + \cdots + x_n = t$,

$$\frac{P(X_1 = x_1, \cdots, X_n = x_n)}{P(T = t)} = \frac{\prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!}}{\frac{e^{-n\theta}(n\theta)^t}{t!}} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i} \frac{1}{\prod_{i=1}^{n} x_i!}}{e^{-n\theta} n^t \theta^t \frac{1}{t!}} = \frac{t!}{n^t \prod_{i=1}^{n} x_i!},$$

which does not depend on θ. So if we know $T(X)$, additional knowledge of X does not give more information about θ.

A sufficient statistic allows us to keep all the information about θ while reducing the dimension of our data. In the previous example, the dimension of our full data $(X_1, \ldots, X_n)$ was n, while the dimension of the sufficient statistic $\sum_{i=1}^{n} X_i$ is 1. Using the full definition of sufficiency is not always easy, but there is a convenient theorem that allows us to find sufficient statistics.

Theorem 1.1 (Factorization criterion)**.** *Suppose that X has pmf/pdf $f_\theta(x)$. Then $T = T(X)$ is sufficient for θ if and only if*

$$f_\theta(x) = g(T(x), \theta) h(x)$$

for some functions g and h.

Proof. We only prove the discrete case. Suppose $f_\theta(x) = P_\theta(X = x) = g(T(x), \theta)h(x)$. If $T(x) = t$,

$$\begin{aligned} P_\theta(X = x | T = t) &= \frac{P_\theta(X = x, T(X) = t)}{P_\theta(T = t)} \\ &= \frac{P_\theta(X = x)}{\sum_{y:T(y)=t} P_\theta(X = y)} \\ &= \frac{g(T(x), \theta)h(x)}{\sum_{y:T(y)=t} g(T(y), \theta)h(y)} \\ &= \frac{g(t, \theta)h(x)}{g(t, \theta)\sum_{y:T(y)=t} h(y)} = \frac{h(x)}{\sum_{y:T(y)=t} h(y)}, \end{aligned}$$

which does not depend on θ. If $T(x) \neq t$, then $P_\theta(X = x|T = t) = 0$, which is again free of θ. Thus T is sufficient for θ.

Now suppose T is sufficient for θ, so that the conditional distribution of $X|T = t$ does not depend on θ. Then

$$P_\theta(X = x) = P_\theta(X = x, T = T(x)) = P_\theta(X = x|T = T(x))P_\theta(T = T(x)).$$

The first factor does not depend on θ by assumption; call it $h(x)$. Let the second factor be $g(T(x), \theta)$, and so we have the required factorisation.

[The continuous case is conceptually similar, but requires a bit of measure theory, hence we omit it]. □

Example. *Returning to the last example, if* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ *Poisson* (θ)*, then*

$$f_\theta(x) = \prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} = e^{-n\theta}\theta^{\sum_{i=1}^n x_i}\frac{1}{\prod_{i=1}^n x_i!} = g\left(\sum_{i=1}^n x_i, \theta\right)h(x),$$

for $g(t, \theta) = e^{-n\theta}\theta^t$ *and* $h(x) = 1/(\prod_{i=1}^n x_i!)$*. Thus by the factorization criterion,* $T(X) = \sum_{i=1}^n X_i$ *is sufficient for* θ*.*

Example. *If* $X_1, \cdots, X_n \overset{\text{i.i.d}}{\sim} U[0, \theta]$*, then*

$$f_\theta(x) = \prod_{i=1}^n \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n}\mathbb{1}_{\{\max_i x_i \leq \theta\}}\mathbb{1}_{\{\min_i x_i \geq 0\}}.$$

Taking $T(x) = \max_i x_i$*,* $g(t, \theta) = \frac{1}{\theta^n}\mathbb{1}_{\{t \leq \theta\}}$ *and* $h(x) = \mathbb{1}_{\{\min_i x_i \geq 0\}}$*, the factorization criterion gives that* $T = \max_i X_i$ *is sufficient for* θ*.*

Example. *If* $X_1, \cdots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$*, then*

$$\begin{aligned} f_\theta(x) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2 \right\} \\ &= \exp\left\{ -\frac{1}{2\sigma^2}\underbrace{\sum_{i=1}^n x_i^2}_{T_1} + \frac{\mu}{\sigma^2}\underbrace{\sum_{i=1}^n x_i}_{T_2} - \frac{n\mu^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2) \right\}. \end{aligned}$$

Therefore, by factorization criterion, $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ *is sufficient for* $\theta = (\mu, \sigma^2)$*.*

Example. *In general, if* X *has distribution belonging a* k*-parameter exponential family, that is,*

$$f_\theta(x) = \exp\left\{ \sum_{i=1}^k c_i(\theta)T_i(x) - d(\theta) + S(x) \right\},$$

then by the factorization criterion, $(T_1(X), \cdots, T_k(X))$ *is sufficient for* θ*.*

The last two examples show that sufficient statistics can be multidimensional. Note that sufficient statistics are <u>not</u> unique. Indeed, by the factorization criterion, any bijective function of a sufficient statistic is also sufficient. For instance, in the second last example, $(\bar{X}, S^2)$ is sufficient for (μ, σ^2). The full data X is also always sufficient for θ, though this is not much use. How can we decide if a sufficient statistic is "good"?

Definition. *A sufficient statistic* $T(X)$ *is <u>minimal</u> if it is a function of every other sufficient statistic, i.e. if* $T'(X)$ *is also sufficient, then* $T'(X) = T'(Y) \implies T(X) = T(Y)$*.*

Thus for any other sufficient statistic T', there exists a function h such that $T(x) = h(T'(x))$. A minimal sufficient statistic represents the maximal reduction of the data that contains as much information about the unknown parameter as the full data itself. This can be seen from the fact that T takes the minimal number of possible values as a function of x among all sufficient statistics, i.e. it keeps the least possible extra information.

Remark. *Minimal sufficient statistics are not unique. In fact, any bijective function of a minimal sufficient statistic is also minimal. However, this is coherent since these have the same dimension.*

Example. *Suppose* $X_1, \cdots, X_n \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ *and consider the statistics*

$$\begin{aligned} T_1(X) &= (X_1, \cdots, X_n), \\ T_2(X) &= (X_1^2, \cdots, X_n^2), \\ T_3(X) &= (X_1^2 + \cdots + X_m^2, X_{m+1}^2 + \cdots + X_n^2), \\ T_4(X) &= X_1^2 + \cdots + X_n^2. \end{aligned}$$

There are all sufficient statistics for σ^2 *(use the factorization criterion), while* T_i *provides greater reduction of the data as* i *increases. We will shortly see that* $T_4(X)$ *is a minimal sufficient statistic.*

Again, we have a useful theorem to find minimal sufficient statistics.

Theorem 1.2. *Suppose that* X *has pmf/pdf* $f_\theta(x)$ *and* $T = T(X)$ *is a statistic that satisfies*

$$\frac{f_\theta(x)}{f_\theta(x')} \text{ does not depend on } \theta \text{ if and only if } T(x) = T(x').$$

Then T *is minimal sufficient for* θ*.*

Proof. First we have to show that T is sufficient, for which we use the factorization criterion. For every possible value t of T, pick some x_t such that $T(x_t) = t$. Now for any x, let $t = T(x)$, so that $T(x) = T(x_t)$. By the theorem hypothesis, $\frac{f_\theta(x)}{f_\theta(x_t)}$ does not depend on θ, so call this ratio $h(x)$. Setting $g(t, \theta) = f_\theta(x_t)$ then gives

$$f_\theta(x) = f_\theta(x_t)\frac{f_\theta(x)}{f_\theta(x_t)} = g(t, \theta)h(x).$$

Thus T is sufficient for θ.

To show T is minimal, let S be any other sufficient statistic. By the factorization criterion, there exist functions g_S and h_S such that $f_\theta(x) = g_S(S(x), \theta)h_S(x)$. Let x and x' be any two sample points with $S(x) = S(x')$. Then

$$\frac{f_\theta(x)}{f_\theta(x')} = \frac{g_S(S(x), \theta)h_S(x)}{g_S(S(x'), \theta)h_S(x')} = \frac{h_S(x)}{h_S(x')}.$$

Since the ratio $\frac{h_S(x)}{h_S(x')}$ does not depend on θ, this implies $T(x) = T(x')$ by the assumption of the theorem. Thus $S(x) = S(x')$ implies $T(x) = T(x')$, i.e. T is a function of S. So T is minimal sufficient. □

Example. *Suppose* $X_1, \cdots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$*. For any* $x, x^* \in \mathbb{R}^n$*,*

$$\begin{aligned} \frac{f_\theta(x)}{f_\theta(x^*)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\}}{(2\pi\sigma^2)^{-n/2}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i^* - \mu)^2\}} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n (x_i^*)^2\right) + \frac{\mu}{\sigma^2}\left(\sum_{i=1}^n x_i - \sum_{i=1}^n x_i^*\right)\right\}. \end{aligned}$$

This ratio is a constant function of (μ, σ^2) *if and only if* $\sum_i x_i^2 = \sum_i (x_i^*)^2$ *and* $\sum_i x_i = \sum_i x_i^*$. *So* $T(X) = (\sum_i X_i^2, \sum_i X_i)$ *is a minimal sufficient statistic for* (μ, σ^2).

Writing $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^n X_i^2 - \frac{n}{n-1}\bar{X}^2$, *we see there is a bijection between* $T(X)$ *and* $T'(X) = (\bar{X}, S^2)$, *and hence* $T'(X)$ *is also minimal sufficient for* (μ, σ^2).

Example. *Suppose* $X_1, \cdots, X_n \overset{\text{i.i.d}}{\sim} U(\theta_1, \theta_2)$. *Arguing as above gives*

$$f_\theta(x) = \tfrac{1}{(\theta_2-\theta_1)^n}\mathbb{1}_{\{\max_i x_i < \theta_2\}}\mathbb{1}_{\{\min_i x_i > \theta_1\}}.$$

For any $x, x^* \in \mathbb{R}^n$,

$$\frac{f_\theta(x)}{f_\theta(x^*)} = \frac{(\theta_2-\theta_1)^{-n}\mathbb{1}_{\{\max_i x_i<\theta_2\}}\mathbb{1}_{\{\min_i x_i>\theta_1\}}}{(\theta_2-\theta_1)^{-n}\mathbb{1}_{\{\max_i x_i^*<\theta_2\}}\mathbb{1}_{\{\min_i x_i^*>\theta_1\}}} = \frac{\mathbb{1}_{\{\max_i x_i<\theta_2\}}\mathbb{1}_{\{\min_i x_i>\theta_1\}}}{\mathbb{1}_{\{\max_i x_i^*<\theta_2\}}\mathbb{1}_{\{\min_i x_i^*>\theta_1\}}}.$$

This ratio is constant as a function of (θ_1, θ_2) *if and only if* $\min_i x_i = \min_i x_i^*$ *and* $\max_i x_i = \max_i x_i^*$. *Therefore,* $(\min_i X_i, \max_i X_i)$ *is a minimal sufficient statistic for* $\theta = (\theta_1, \theta_2)$.

1.5 Rao-Blackwell theorem

It turns out we can use sufficient statistics to improve *any* estimator.

Theorem 1.3 (Rao-Blackwell theorem). *Let* $T = T(X)$ *be a sufficient statistic for* θ *and let* $\tilde{\theta}(X)$ *be an estimator for* θ *with* $E_\theta(\tilde{\theta}^2) < \infty$ *for all* $\theta \in \Theta$. *Let* $\hat{\theta}(X) = E[\tilde{\theta}(X)|T(X)]$. *Then for all* $\theta \in \Theta$,

$$b_\theta(\hat{\theta}) = b_\theta(\tilde{\theta}) \qquad \textit{and} \qquad \mathrm{var}_\theta(\hat{\theta}) \le \mathrm{var}_\theta(\tilde{\theta}).$$

In particular, $MSE(\hat{\theta}) \le MSE(\tilde{\theta})$. *The inequality is strict unless* $\tilde{\theta}$ *is a function of* T.

We must be careful here with the definition of $\hat{\theta}$. It is defined as the (conditional) expectation of $\tilde{\theta}(X)$, which could in principle depend on the unknown value of θ. However, since T is sufficient for θ, the conditional distribution of X given T does *not* depend on θ. Thus $\hat{\theta} = E[\tilde{\theta}(X)|T]$ does not depend on θ and is a genuine estimator.

The Rao-Blackwell theorem says that if an estimator is not a function of a sufficient statistic, then one can construct a new estimator with the same bias and smaller variance by taking the conditional expectation given a sufficient statistic. Note that if the original estimator $\tilde{\theta}$ is unbiased, so too is the new estimator $\hat{\theta}$. If $\tilde{\theta}$ is already a function of T, then $\hat{\theta} = \tilde{\theta}$.

Proof. For the bias, by the conditional expectation formula $E_\theta(\hat{\theta}) = E_\theta[E(\tilde{\theta}|T)] = E_\theta(\tilde{\theta})$, so that $b_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta = E_\theta(\tilde{\theta}) - \theta = b_\theta(\tilde{\theta})$.

By the conditional variance formula,

$$\mathrm{var}_\theta(\tilde{\theta}) = E_\theta[\mathrm{var}(\tilde{\theta}|T)] + \mathrm{var}_\theta(E[\tilde{\theta}|T]) = E_\theta[\underbrace{\mathrm{var}(\tilde{\theta}|T)}_{\ge 0}] + \mathrm{var}_\theta(\hat{\theta}),$$

and hence $\mathrm{var}_\theta(\tilde{\theta}) \ge \mathrm{var}_\theta(\hat{\theta})$. Using the bias-variance decomposition for the MSE, $MSE(\hat{\theta}) \le MSE(\tilde{\theta})$, with equality if and only if $\mathrm{var}(\tilde{\theta}|T) = 0$, i.e. $\tilde{\theta}$ is a function of T. □

Example. *Suppose* $X = (X_1, \ldots, X_n)$ *with* $X_i \overset{\text{i.i.d}}{\sim}$ *Poisson*(λ) *and let* $\theta = e^{-\lambda}$, *which is the probability that* $X_1 = 0$. *Using that* $\lambda = -\log\theta$,

$$f_\theta(x) = \prod_{i=1}^n \frac{\theta(-\log\theta)^{x_i}}{x_i!} = \frac{\theta^n(-\log\theta)^{\sum x_i}}{\prod x_i!}.$$

By the factorization criterion, $T = \sum_i X_i$ *is sufficient for* θ. *Note that* $\sum_i X_i \sim$ *Poisson*$(n\lambda)$.

We start with the rough estimator $\tilde{\theta} = \mathbb{1}_{\{X_1=0\}}$, which says if we observe nothing in the first observation period then we assume the event is impossible. One can check $\tilde{\theta}$ is unbiased. Using the independence of the (X_i),

$$\begin{aligned} E[\tilde{\theta}|T=t] &= P\left(X_1=0 \middle| \sum_{i=1}^n X_i = t\right) \\ &= \frac{P(X_1=0)P(\sum_{i=2}^n X_i = t)}{P(\sum_{i=1}^n X_i = t)} \\ &= \frac{e^{-\lambda} \times e^{-(n-1)\lambda}((n-1)\lambda)^t/t!}{e^{-n\lambda}(n\lambda)^t/t!} = \left(\frac{n-1}{n}\right)^t. \end{aligned}$$

The Rao-Blackwell estimator is thus $\hat{\theta} = (1-1/n)^{\sum X_i}$. Using the exponential approximation, $\hat{\theta} = (1-1/n)^{n\bar{X}} \approx e^{-\bar{X}} = e^{-\hat{\lambda}}$, which is more reasonable than $\tilde{\theta}$.

A natural question is which sufficient statistic to use in the conditional expectation.

Lemma 1.1. *Let T_1 and T_2 be two sufficient statistics for θ and let $\tilde{\theta}(X)$ be an estimator for θ with $E_\theta(\tilde{\theta}^2) < \infty$ for all $\theta \in \Theta$. Let $\hat{\theta}_1(X) = E[\tilde{\theta}(X)|T_1(X)]$ and $\hat{\theta}_2(X) = E[\tilde{\theta}(X)|T_2(X)]$. If $T_2 = h(T_1)$, then for all $\theta \in \Theta$,*

$$\mathrm{var}_\theta(\hat{\theta}_2) \leq \mathrm{var}_\theta(\hat{\theta}_1).$$

Proof. Since $T_2 = h(T_1)$, the tower rule for conditional expectations gives

$$E[\hat{\theta}_1|T_2] = E[E[\tilde{\theta}|T_1]|T_2] = E[\tilde{\theta}|T_2] = \hat{\theta}_2.$$

Applying the Rao-Blackwell theorem to $E[\hat{\theta}_1|T_2]$ then gives

$$\mathrm{var}_\theta(\hat{\theta}_2) \leq \mathrm{var}_\theta(\hat{\theta}_1).$$

□

This result says that the best variance reduction via the Rao-Blackwell theorem is achieved by conditioning on a sufficient statistic that is a function of any other sufficient statistic, i.e. a *minimal sufficient statistic*. Thus for a given a baseline estimator $\tilde{\theta}$ and minimal sufficient statistic T, the best (i.e. smallest variance) Rao-Blackwell estimator based on $\tilde{\theta}$ is $E[\tilde{\theta}|T]$.

However, that the variance of $E[\tilde{\theta}|T]$ does depend on the baseline estimator $\tilde{\theta}$, so that this estimator will not necessarily have smallest variance among all estimators with the same bias. Indeed, there may exist another baseline estimator θ^* for which $E[\theta^*|T]$ has smaller variance. In other words, the Rao-Blackwell Theorem does depend on the choice of baseline estimator.

Example. *Suppose X is a random variable with probability mass function*

$$f_\theta(x) = \begin{cases} \theta & \text{for } x = -1, \\ (1-\theta)^2\theta^x & \text{for } x = 0, 1, 2, \ldots, \end{cases}$$

where $0 < \theta < 1$. The full observation X is automatically a sufficient statistic, which can be shown to be minimal in this example. Two unbiased estimators of θ are

$$\begin{cases} \tilde{\theta}_1(X) = \mathbb{1}(X=-1), \\ \tilde{\theta}_2(X) = \mathbb{1}(X=-1) + X, \end{cases}$$

because $E_\theta \mathbb{1}(X=-1) = P_\theta(X=-1) = \theta$ and one can check that $E_\theta X = 0$ for all θ. Since $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are functions of the sufficient statistic X, the Rao-Blackwell estimators equal

$$\begin{cases} \hat{\theta}_1(X) = E[\tilde{\theta}_1|X] = \tilde{\theta}_1 \\ \hat{\theta}_2(X) = E[\tilde{\theta}_2|X] = \tilde{\theta}_2. \end{cases}$$

One has $\mathrm{var}_\theta(\tilde{\theta}_1) = \theta(1-\theta)$, *while*

$$\begin{aligned}\mathrm{var}_\theta(\tilde{\theta}_2) &= \mathrm{var}_\theta(\tilde{\theta}_1) + \mathrm{var}_\theta(X) + 2\mathrm{Cov}_\theta(\tilde{\theta}_1, X)\\ &= \theta(1-\theta) + \frac{2\theta}{1-\theta} - 2\theta\\ &= \frac{\theta(1+\theta^2)}{1-\theta} > \mathrm{var}_\theta(\tilde{\theta}_1)\end{aligned}$$

for all $0 < \theta < 1$.

This example shows that a minimal sufficient statistic is not enough to obtain the minimal variance unbiased estimator. In fact, the Rao-Blackwell theorem needs to be improved such that the conditional expectation $\hat{\theta} = E[\tilde{\theta}|T]$ is independent of $\tilde{\theta}$. For this, we will need the concept of a complete statistic, see Section 5.4 below.

2 Likelihood-based estimation

2.1 The likelihood Function

Definition. *Suppose $X = (X_1, \ldots, X_n)$ are random variables with joint pmf/pdf $f_\theta(x) = f_{n,\theta}(x)$, $\theta \in \Theta$. Consider observing a realization x of X. The likelihood function $L : \Theta \to \mathbb{R}$ is defined as*

$$L(\theta) = L_n(\theta) = f_{n,\theta}(x)$$

and the log-likelihood function $l : \Theta \to \mathbb{R}$ is defined as

$$l(\theta) = l_n(\theta) = \log L_n(\theta) = \log f_{n,\theta}(x).$$

One regards the likelihood as a function of the parameter θ for a *fixed* realization of the data $X = x$. The above definitions do not require an i.i.d. assumption and are well-defined for any joint pmf/pdf. However, in the i.i.d. setting considered in this course, the (log-)likelihood simplifies. If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} f_{1,\theta}$, then

$$L_n(\theta) = \prod_{i=1}^{n} f_{1,\theta}(x_i), \qquad l_n(\theta) = \sum_{i=1}^{n} \log f_{1,\theta}(x_i).$$

When studying the mathematical properties of likelihood-based estimators, it is often helpful to think of $L : \Theta \to \mathbb{R}$ as a random function, the randomness coming from X.

Definition. *A maximum likelihood estimator (MLE) is defined as any element $\hat{\theta} = \hat{\theta}_{ML} = \hat{\theta}_{ML}(X_1, \ldots, X_n) \in \Theta$ for which*

$$L_n(\hat{\theta}) = \max_{\theta \in \Theta} L_n(\theta).$$

Remark. *By definition of the MLE and the functions above:*

- *Since $t \mapsto \log t$ is a monotonically increasing function, it is equivalent to maximize L_n or l_n, so either function can be used in the definition of $\hat{\theta}_{ML}$.*
- *The estimator $\hat{\theta}_{ML}$ is a function of $X_1, \ldots, X_n$ only.*
- *The MLE need not be uniquely defined.*
- *If the log-likelihood is differentiable with respect to $\theta = (\theta_1, \ldots, \theta_p)$ over Θ, then the MLE $\hat{\theta}$ satisfies the equations*

  $$\frac{\partial}{\partial \theta_k} \log L_n(\hat{\theta}) = 0 \quad \text{for } k = 1, \ldots, p.$$

  These equations are often called the likelihood equations. Note that these equations can have multiple solutions (including multiple local maxima), so it is important to check a solution indeed maximizes the likelihood function.

Example. *If $X_1, \cdots, X_n \overset{\text{i.i.d}}{\sim} Exp(\lambda)$ with $\lambda > 0$, then the likelihood function is*

$$L_n(\lambda) = f_{n,\lambda}(x) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_i x_i}$$

and hence the log-likelihood function is

$$l_n(\lambda) = \log L_n(\lambda) = \log \left(\prod_{i=1}^{n} \lambda e^{-\lambda x_i} \right) = \sum_{i=1}^{n} \log(\lambda e^{-\lambda x_i}) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Differentiating this,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} l_n(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0,$$

which is solved by $\hat{\lambda} = 1/\bar{x}_n$*. To show this is indeed the maximum, note that*

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} l_n(\lambda) = -\frac{n}{\lambda^2} < 0,$$

so this is a global maximum. Therefore, $\hat{\lambda}_{ML} = \frac{1}{\bar{X}_n}$ *is the MLE of* λ*.*

Example. *If* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ *Poisson*(λ) *with* $\lambda > 0$*, then the likelihood function is*

$$L_n(\lambda) = f_{n,\lambda}(x) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

and hence the log-likelihood is

$$\begin{aligned} l_n(\lambda) = \log L_n(\lambda) = \log\left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) &= \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \\ &= -n\lambda + \log\lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!). \end{aligned}$$

Hence, if $\sum_{i=1}^{n} x_i > 0$*, then*

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} l_n(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0,$$

which has solution $\hat{\lambda} = \bar{x}_n$*. To show this is indeed the maximum, note that*

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} l_n(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i < 0$$

Therefore, $\hat{\lambda}_{ML} = \bar{X}_n$ *is the MLE of* λ*, provided that* $\sum_{i=1}^{n} X_i > 0$*.*

If $\sum_{i=1}^{n} X_i = 0$*, then* $l_n(\lambda) = -n\lambda$ *has no maximum on the parameter space* $(0, \infty)$ *and hence the MLE does not exist (*$l_n(\lambda)$ *is increasing as* $\lambda \downarrow 0$*, but* $0 \notin \Theta$ *is not a valid parameter choice).*

Example. *If* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$*, then the likelihood function is*

$$L_n(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},$$

and hence the log-likelihood is

$$\begin{aligned} l_n(\mu, \sigma^2) = \log L_n(\mu, \sigma^2) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \\ &= \sum_{i=1}^{n} \left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu)^2\right\} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \end{aligned}$$

Differentiating with respect to both μ and σ^2 (i.e. taking the gradient with respect to $\theta = (\mu, \sigma^2)$),

$$\frac{\partial}{\partial \mu} l_n(\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0, \qquad \frac{\partial}{\partial \sigma^2} l_n(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0,$$

which has solution

$$(\hat{\mu}, \hat{\sigma}^2) = \left(\bar{x}_n, \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \right).$$

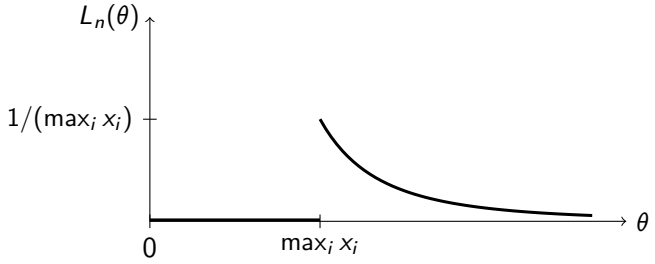
We need to prove this solution indeed maximizes the likelihood function. It is not difficult to show this: it suffices to show the matrix of second order partial derivatives, called the Hessian matrix, is negative definite. Therefore, $\hat{\mu}_{ML} = \bar{X}_n$ and $\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ are the MLEs of μ and σ^2, respectively. It can be checked that $E\hat{\sigma}^2_{ML} = \frac{n-1}{n} \sigma^2$, i.e. $\hat{\sigma}^2$ is biased.

It is not always possible to obtain MLEs by differentiating the likelihood function.

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} U[0, \theta]$, $\theta > 0$, then we saw earlier that*

$$L_n(\theta) = f_{n,\theta}(x) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n} \mathbb{1}_{\{\max_i x_i \leq \theta\}} \mathbb{1}_{\{\min_i x_i \geq 0\}}.$$

Since $\frac{1}{\theta^n}$ is a decreasing function of θ, the likelihood is maximized for the smallest possible value of θ such that the indicator functions are one, namely $\theta = \max_i x_i$. Therefore, $\hat{\theta}_{ML} = \max_i X_i$ is the MLE of θ. We can plot the likelihood function:



Remark. *For exponential families of distributions, one can often obtain the MLE directly. However, in more complicated settings, there is often no closed form for the MLE and we must find $\hat{\theta}_{ML}$ numerically. Some well-known numerical algorithms will be discussed in Section 2.4 below.*

Remark. *How does the MLE relate to sufficient statistics? If $T(X)$ is a sufficient statistic for θ, then by the factorization criterion, the likelihood equals*

$$L(\theta) = g(T(x), \theta) h(x).$$

To maximize this as a function of θ, we need only maximize g. Thus the MLE $\hat{\theta}_{ML}$ is a function of the sufficient statistic T [More precisely, if the MLE is not unique, then there exists an MLE $\hat{\theta}_{ML}$ which is a function of T].

2.1.1 Invariance of the MLE

An attractive property of MLEs is invariance: if $\hat{\theta}_{ML}$ is an MLE of θ and g is an *arbitrary* function of θ, then $g(\hat{\theta}_{ML})$ is an MLE of $g(\theta)$. The proof is straightforward if g is a bijection, but requires a little more care otherwise.

Write $\eta = g(\theta)$ and define the *induced likelihood function*

$$L^*(\eta) = \sup_{\theta : g(\theta) = \eta} L(\theta).$$

The value $\hat{\eta}$ that maximizes $L^*(\eta)$ is called the MLE of $\eta = g(\theta)$. From the last display, we see that the maxima of L and L^* coincide.

Theorem 2.1. *If $\hat{\theta}_{ML}$ is an MLE for θ and $g(\theta)$ is any function, then $g(\hat{\theta}_{ML})$ is an MLE for $g(\theta)$.*

Proof. We must show that $L^*(\hat{\eta}) = L^*(g(\hat{\theta}))$, where $\hat{\eta}$ is the maximizer of $L^*(\eta)$. Since the maxima of L and L^* coincide,

$$L^*(\hat{\eta}) = \sup_{\eta} \sup_{\theta: g(\theta)=\eta} L(\theta) = \sup_{\theta} L(\theta) = L(\hat{\theta}),$$

where the second equality is because the iterated maximization equals the unconditional maximization of θ. Moreover, since $\hat{\theta}$ is the MLE of θ,

$$L(\hat{\theta}) = \sup_{\theta: g(\theta)=g(\hat{\theta})} L(\theta) = L^*(g(\hat{\theta})).$$

We have thus shown that $L^*(\hat{\eta}) = L^*(g(\hat{\theta}))$, i.e. $g(\hat{\theta})$ is the maximizer of L^* and hence the MLE of $g(\theta)$. □

Example. *If $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} Exp(\lambda)$, we saw the MLE of λ is $\hat{\lambda}_{ML} = 1/\bar{X}_n$. Then the MLE of the variance $g(\lambda) = \text{var}_\lambda(X_1) = 1/\lambda^2$ is $g(\hat{\lambda}) = \bar{X}_n^2$.*

If general, if the MLE of the standard deviation σ is $\hat{\sigma}$, then the MLE of the variance σ^2 is $\hat{\sigma}^2$. This is useful in practice, since we can use this to simplify a lot of computations.

2.2 Geometry of the likelihood: score and Fisher information

It is important to remember that L_n is a function of the parameter θ, the randomness being in the values of $X_1, \dots, X_n$. Therefore, derivatives are taken with respect to θ not $x_1, \dots, x_n$. The idea behind maximum likelihood is to build a random function l_n based on the data and then maximize it. To understand the behaviour of the MLE, a natural first step is to study what happens if we were to maximize the *expectation* of this random function.

Lemma 2.1. *Consider a model $\{f_\theta : \theta \in \Theta\}$ such that $E_\theta|\log f_\theta(X)| < \infty$ for all $\theta \in \Theta$. If $X \sim f_{\theta_0}$ for some true $\theta_0 \in \Theta$, then for any $\theta \in \Theta$,*

$$E_{\theta_0}[l(\theta)] \leq E_{\theta_0}[l(\theta_0)],$$

i.e. $\theta \mapsto E_{\theta_0}[l(\theta)]$ is maximized at θ_0.

This lemma suggests that if we knew the function $\theta \mapsto E_{\theta_0}[l(\theta)]$, we could recover the true θ_0 exactly by maximization. Since we do not know this function, we instead use an approximation based on the i.i.d. data $X_1, \dots, X_n$:

$$\frac{1}{n} l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f_{1,\theta}(x_i),$$

the (normalized) log-likelihood. Since this is an average of i.i.d. random variables, the law of large numbers implies $\frac{1}{n} l_n(\theta) \to E_{\theta_0}[l(\theta)]$ for every $\theta \in \Theta$. This provides a heuristic motivation behind the idea of maximum likelihood estimation.

Proof. We consider the continuous case. For any $\theta \in \Theta$,

$$E_{\theta_0}[l(\theta) - l(\theta_0)] = E_{\theta_0}[\log f_\theta(X) - \log f_{\theta_0}(X)] = E_{\theta_0}\left[\log \frac{f_\theta(X)}{f_{\theta_0}(X)}\right].$$

Using the inequality $\log t \leq t - 1$ for all $t \geq 0$,

$$\begin{aligned} E_{\theta_0}[l(\theta) - l(\theta_0)] \leq E_{\theta_0}\left[\frac{f_\theta(X)}{f_{\theta_0}(X)} - 1\right] &= \int \left(\frac{f_\theta(x)}{f_{\theta_0}(x)} - 1\right) f_{\theta_0}(x)\, \mathrm{d}x \\ &= \int f_\theta(x)\, \mathrm{d}x - \int f_{\theta_0}(x)\, \mathrm{d}x = 1 - 1 = 0, \end{aligned}$$

since f_θ and f_{θ_0} are both pdfs. This proves the result in the continuous case. The discrete case follows identically upon replacing the integrals with sums. □

Remark. *The inequality in the last lemma is strict for all $\theta \neq \theta_0$ as long as the model is identifiable. This can be seen from the inequality $\log t \leq t-1$ used in the proof, which is strict except for equality at $t = 1$. Thus for identifiable models, θ_0 is the unique maximizer.*

We saw that in several examples (though not all), the MLE can be found as the unique zero of the gradient of the log-likelihood.

Definition. *For $\Theta \subseteq \mathbb{R}^p$ and $\theta \mapsto l_n(\theta)$ differentiable, the score function is defined as*

$$S_n(\theta) = \nabla_\theta l_n(\theta) = \left(\frac{\partial}{\partial \theta_1} l_n(\theta), \ldots, \frac{\partial}{\partial \theta_p} l_n(\theta) \right)^T.$$

One of the main uses of the score function is to find the MLE $\hat{\theta}$ by solving the likelihood equations, which can be rewritten as solving the (vector) equation $S_n(\hat{\theta}) = 0$. In the i.i.d. setting, this is equivalent to solving

$$\frac{1}{n} S_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \log f_\theta(x_i) = 0,$$

where we have exchanged the sum and derivatives. This method can be motivated in a similar way to the last lemma by replacing the empirical (data-dependent) sum with its expectation E_{θ_0}.

Lemma 2.2. *Consider a model $\{f_\theta : \theta \in \Theta\}$ that is regular enough that differentiation (in θ) and integration (in x) can be exchanged. Then for all $\theta \in int(\Theta)$,*

$$E_\theta[\nabla_\theta \log f_\theta(X)] = 0.$$

Proof. We consider only the continuous case. Computing the expectation directly,

$$\begin{aligned} E_\theta \left[\frac{\partial}{\partial \theta_i} \log f_\theta(X) \right] &= \int \left(\frac{\partial}{\partial \theta_i} \log f_\theta(x) \right) f_\theta(x) dx \\ &= \int \left(\frac{\partial}{\partial \theta_i} f_\theta(x) \frac{1}{f_\theta(x)} \right) f_\theta(x) dx \\ &= \int \frac{\partial}{\partial \theta_i} f_\theta(x) dx \\ &= \frac{\partial}{\partial \theta_i} \int f_\theta(x) dx = \frac{\partial}{\partial \theta_i} 1 = 0, \end{aligned}$$

where in the last line we have used the lemma hypothesis to exchange the order of integration and differentiation. □

Remark. *Conditions for interchanging the order of differentiation and integration can be found in a measure theory course. Note that when the support of f_θ depends on θ, this is generally not true. We saw this above in the example $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} U[0, \theta]$, where one cannot obtain the MLE for θ by differentiating the likelihood.*

In particular, this implies $E_{\theta_0}[\nabla_\theta \log f_{\theta_0}(X)] = 0$ at the true parameter θ_0.

Definition. *For a parameter space $\Theta \subseteq \mathbb{R}^p$, we define for all $\theta \in int(\Theta)$ the Fisher information matrix as*

$$I(\theta) = E_\theta[\nabla_\theta \log f_\theta(X) \nabla_\theta \log f_\theta(X)^T],$$

or written coordinate-wise:

$$I_{ij}(\theta) = E_\theta \left[\frac{\partial}{\partial \theta_i} \log f_\theta(X) \frac{\partial}{\partial \theta_j} \log f_\theta(X) \right], \qquad 1 \leq i, j \leq p.$$

Remark. *In dimension $p = 1$, we have*

$$I(\theta) = E_\theta\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right)^2\right] = \mathrm{var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right) = \mathrm{var}_\theta(l'(\theta)),$$

since the term in brackets is a centred random variable by the Lemma 2.2. Thus $I(\theta_0)$ measures the random variations of $S_n(\theta_0)$ about its mean, which equals zero. This helps to quantify the precision of $\hat{\theta}$ around θ_0 as a solution of $S_n(\hat{\theta}) = 0$.

Lemma 2.3. *Under the same regularity assumptions as Lemma 2.2, for all $\theta \in int(\Theta)$,*

$$I(\theta) = -E_\theta[\nabla_\theta^2 \log f_\theta(X)],$$

or written coordinate-wise:

$$I_{ij}(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_\theta(X)\right], \qquad 1 \le i, j \le p.$$

Proof. We again consider only the continuous case. The term in the expectation equals

$$\begin{aligned}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_\theta(x) &= \frac{\partial}{\partial\theta_i}\left(\frac{1}{f_\theta(x)}\frac{\partial}{\partial\theta_j}f_\theta(x)\right)\\ &= \frac{1}{f_\theta(x)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}f_\theta(x) - \frac{1}{f_\theta(x)^2}\frac{\partial}{\partial\theta_i}f_\theta(x)\frac{\partial}{\partial\theta_j}f_\theta(x).\end{aligned}$$

Taking expectations then gives

$$\begin{aligned}-E_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_\theta(X)\right] &= -\int\frac{1}{f_\theta(x)}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}f_\theta(x)\right)f_\theta(x)dx + E_\theta\left[\frac{1}{f_\theta(X)^2}\frac{\partial}{\partial\theta_i}f_\theta(X)\frac{\partial}{\partial\theta_j}f_\theta(X)\right]\\ &= -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\int f_\theta(x)dx + E_\theta\left[\left(\frac{1}{f_\theta(X)}\frac{\partial}{\partial\theta_i}f_\theta(X)\right)\left(\frac{1}{f_\theta(X)}\frac{\partial}{\partial\theta_j}f_\theta(X)\right)\right]\\ &= -\frac{\partial^2}{\partial\theta_i\partial\theta_j}1 + E_\theta\left[\left(\frac{\partial}{\partial\theta_i}\log f_\theta(X)\right)\left(\frac{\partial}{\partial\theta_j}\log f_\theta(X)\right)\right]\\ &= 0 + I_{ij}(\theta),\end{aligned}$$

where we have used the lemma hypothesis to exchange the order of integration and differentiation. □

Remark. *In dimension $p = 1$, the lemma becomes*

$$I(\theta) = E_\theta[(l'(\theta))^2] = -E_\theta[l''(\theta)].$$

Example. *If $X \sim Poisson(\theta)$, then the Fisher information is*

$$\begin{aligned}I(\theta) = E[(l'(\theta))^2] &= E\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right)^2\right]\\ &= E\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log\left(\frac{e^{-\theta}\theta^X}{X!}\right)\right)^2\right]\\ &= E\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}(-\theta + X\log\theta - \log(X!))\right)^2\right]\\ &= E\left[\left(-1 + \frac{X}{\theta}\right)^2\right] = E\left[1 - \frac{2X}{\theta} + \frac{X^2}{\theta^2}\right]\\ &= 1 - \frac{2}{\theta}E[X] + \frac{1}{\theta^2}E[X^2] = 1 - \frac{2}{\theta}\theta + \frac{(\theta^2+\theta)}{\theta^2} = \frac{1}{\theta}.\end{aligned}$$

Therefore, for a Poisson random variable, the Fisher information is the inverse of its variance.

Example. *Suppose $X \sim N(\mu, \sigma_0^2)$, where σ_0^2 is known (and hence not considered a model parameter). Then the Fisher information is*

$$\begin{aligned} I(\mu) &= E\left[\left(\frac{\mathrm{d}}{\mathrm{d}\mu}\log f_\mu(X)\right)^2\right] \\ &= E\left[\left(\frac{\mathrm{d}}{\mathrm{d}\mu}\log\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-(X-\mu)^2/(2\sigma_0^2)}\right)\right)^2\right] \\ &= E\left[\left(\frac{\mathrm{d}}{\mathrm{d}\mu}\left(-\frac{1}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}(X-\mu)^2\right)\right)^2\right] \\ &= E\left[\left(\frac{1}{\sigma_0^2}(X-\mu)\right)^2\right] = \frac{1}{\sigma_0^4}\mathrm{var}(X) = 1/\sigma_0^2. \end{aligned}$$

Note that if we instead considered $X \sim N(\mu, \sigma^2)$ with both (μ, σ^2) unknown, then the Fisher information would be a 2×2 matrix.

Proposition 2.1. *If $X = (X_1, \ldots, X_n)$ with X_i i.i.d. random variables, then the Fisher information $I_n(\theta)$ of X equals*

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

where $I(\theta)$ is the Fisher information of one random variable X_i.

Proof. See Problem Sheet 2. □

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma_0^2)$, where σ_0^2 is known. Using the previous example, the Fisher information of $(X_1, \ldots, X_n)$ is*

$$I_n(\mu) = nI(\mu) = \frac{n}{\sigma_0^2}.$$

The Fisher information increases with the sample size n and also when the variance σ_0^2 decreases. This makes sense since the sample $(X_1, \ldots, X_n)$ contains "more information" about μ when n is large and σ_0^2 is small.

2.3 The Cramer-Rao lower bound

The following result formalizes the link between Fisher information and precision of estimation. It provides a lower bound on the variance of any *unbiased* estimator of θ, stating that this variance is at least as high as the inverse of the Fisher information. Thus if the variance of an unbiased estimator achieves this lower bound, it has smallest possible variance among all *unbiased* estimators of θ.

Theorem 2.2 (Cramer-Rao lower bound)**.** *Consider a model $\{f_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ (i.e. $p = 1$) that is regular enough that differentiation (in θ) and integration (in x) can be exchanged. Let $\hat{\theta} = \hat{\theta}(X)$ be an unbiased estimator of θ based on an observation X from this model. Then for all $\theta \in int(\Theta)$,*

$$\mathrm{var}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2] \geq \frac{1}{I(\theta)},$$

where $I(\theta)$ is the Fisher information of X.

Proof. Recall that by the Cauchy Schwarz inequality, for any Y, Z,

$$\mathrm{Cov}_\theta(Y, Z)^2 \leq \mathrm{var}_\theta(Y)\mathrm{var}_\theta(Z).$$

Setting $Y = \hat{\theta}$ and $Z = \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X) = l'(\theta)$ gives

$$\mathrm{var}_\theta(\hat{\theta}) \geq \frac{\mathrm{Cov}_\theta(\hat{\theta}, Z)^2}{\mathrm{var}_\theta(Z)}. \tag{2.1}$$

We recall from the lemmas above that $E_\theta[Z] = 0$ and $\mathrm{var}_\theta(Z) = I(\theta)$. Therefore, in the continuous case,

$$\begin{aligned}
\mathrm{Cov}_\theta(\hat{\theta}, Z) = E_\theta[\hat{\theta}Z] - E_\theta[\hat{\theta}]E_\theta[Z] &= E_\theta\left[\hat{\theta}\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right] \\
&= \int \hat{\theta}(x)\left(\frac{1}{f_\theta(x)}\frac{\mathrm{d}}{\mathrm{d}\theta}f_\theta(x)\right) f_\theta(x)dx \\
&= \int \hat{\theta}(x)\frac{\mathrm{d}}{\mathrm{d}\theta}f_\theta(x)dx \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\int \hat{\theta}(x) f_\theta(x)dx \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}E_\theta[\hat{\theta}] = \frac{\mathrm{d}}{\mathrm{d}\theta}\theta = 1,
\end{aligned}$$

where we have used the regularity conditions to swap the derivative and integral, and the unbiasedness of $\hat{\theta}$ in the last line. Substituting in these formulas gives the result. The discrete case follows identically upon replacing the integrals with sums. □

Remark. *If $X = (X_1, \ldots, X_n)$ for X_i i.i.d., then for any unbiased estimator $\hat{\theta}$,*

$$\mathrm{var}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2] \geq \frac{1}{nI_{X_1}(\theta)},$$

where $I_{X_1}(\theta)$ is the Fisher information the random variable X_1.

Remark. *When $p = 1$, one can easily generalize the Cramer-Rao lower bound to the situation where one has a biased estimator $\hat{g}$ of $g(\theta)$ for some function $g : \Theta \to \mathbb{R}$. It then yields*

$$\mathrm{var}_\theta(\hat{g}) \geq \frac{(g'(\theta) + b'(\theta))^2}{I(\theta)}, \tag{2.2}$$

where $b(\theta) = E_\theta[\hat{g}] - g(\theta)$ is the bias of $\hat{g}$. This follows by substituting $E_\theta[\hat{\theta}] = \theta$ with $E_\theta[\hat{g}] = g(\theta) + b(\theta)$ in the last proof.

Remark. *The Cramer-Rao lower bound concerns the variance of an estimator, which is a univariate quantity. One can extend it to the multivariate setting $\theta \in \mathbb{R}^p$, $p \geq 1$ as follows. For any differentiable function $g : \Theta \to \mathbb{R}$, let $\hat{g}$ be an unbiased estimator of $g(\theta)$ based on an observation X from the model $\{f_\theta : \theta \in \Theta\}$. Then for all $\theta \in int(\Theta)$,*

$$\mathrm{var}_\theta(\hat{g}) \geq \nabla_\theta g(\theta)^T I^{-1}(\theta)\nabla_\theta g(\theta).$$

For example, if $g(\theta) = u^T\theta = \sum_{i=1}^p u_i\theta_i$, then $\nabla_\theta g(\theta) = u$, so the lower bound implies $\mathrm{var}_\theta(\hat{g}) \geq u^T I^{-1}(\theta)u$.

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma_0^2)$, where σ_0^2 is known. Then the Cramer-Rao lower bound for any unbiased estimator $\hat{\mu}$ of μ is*

$$\mathrm{var}_\mu(\hat{\mu}) \geq \frac{1}{I(\mu)} = \frac{1}{nI_{X_1}(\mu)} = \frac{1}{(n/\sigma_0^2)} = \frac{\sigma_0^2}{n}.$$

The unbiased estimator $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies $\mathrm{var}_\mu(\bar{X}_n) = \sigma_0^2/n$, which equals the lower bound. This estimator thus has minimal variance among all possible unbiased estimators of μ in this model.

Example. *Suppose* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \text{Poisson}(\lambda)$. *Then the Cramer-Rao lower bound for any unbiased estimator* $\hat{\lambda}$ *of* λ *is*

$$\text{var}_\lambda(\hat{\lambda}) \geq \frac{1}{I(\lambda)} = \frac{1}{nI_{X_1}(\lambda)} = \frac{1}{(n/\lambda)} = \frac{\lambda}{n}.$$

In this case, the unbiased estimator $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ *satisfies* $\text{var}_\lambda(\bar{X}) = \lambda/n$, *which again equals the lower bound. This estimator thus has minimal variance among all possible unbiased estimators of* λ *in this model.*

We next study when the Cramer-Rao lower bound can be attained.

Proposition 2.2. *Assume regularity conditions and* $p = 1$. *An unbiased statistic* $\hat{\theta}(X)$ *attains the Cramer-Rao lower bound if and only if* X *belongs to the exponential family*

$$f_\theta(x) = \exp\Big(A(\theta)\hat{\theta}(x) + B(\theta) + S(x)\Big)$$

for some functions A, B, S.

Proof. From the proof of the Cramer-Rao lower bound, we know that for $Z = \frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X) = l'(\theta)$ and $\hat{\theta}$ an unbiased estimator of θ, $\text{var}_\theta(Z) = I(\theta)$ and $\text{Cov}_\theta(\hat{\theta}, Z) = 1$. An estimator thus attains the Cramer-Rao lower bound if and only if

$$\text{var}_\theta(\hat{\theta}) = \frac{1}{I(\theta)} = \frac{\text{Cov}_\theta(\hat{\theta}, Z)^2}{\text{var}_\theta(Z)},$$

which is the equality case of the Cauchy-Schwarz inequality (see (2.1)). Now, equality holds for Cauchy-Schwarz if and only if Z is a linear function of $\hat{\theta}$, that is

$$Z = \frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(x) = A^*(\theta)\hat{\theta}(x) + B^*(\theta)$$

for all x. Integrating with respect to θ and exponentiating,

$$f_\theta(x) = \exp(A(\theta)\hat{\theta}(x) + B(\theta) + S(x)),$$

where the constant of integration $S(x)$ is a function of x. □

The Cramer-Rao lower bound cannot always be attained, but if an unbiased estimator attains the lower bound, then it has best possible variance among all unbiased estimators. We will see an example where it cannot be attained in Section 5.4.1.

2.4 Numerical computation of MLEs (Non-examinable)

Example. *Suppose* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \Gamma(\alpha, \lambda)$. *Then it is easy to show that the MLEs of* α *and* λ *satisfy the following likelihood equations:*

$$\frac{\partial}{\partial\alpha} l_n(\alpha, \lambda) = n\log\lambda - n\Gamma'(\alpha) + \sum_{i=1}^n \log x_i = 0$$

$$\frac{\partial}{\partial\lambda} l_n(\alpha, \lambda) = \frac{n\alpha}{\lambda} - \sum_{i=1}^n x_i = 0.$$

However, solving the above equations analytically is not generally possible due to the $\Gamma'(\alpha)$ *term.*

In general, there are many situations where it is not possible to obtain closed form expressions for MLEs. In these situations, it is necessary to calculate MLEs numerically. Three common numerical methods are the Newton-Raphson method, Fisher scoring method and EM algorithm. We shall summarize the first two.

2.4.1 The Newton-Raphson method

Suppose we want to solve $g(x_0) = 0$ for g a differentiable function (e.g. a likelihood equation). Given a value x close to x_0, a first-order Taylor expansion of $g(x_0)$ around x gives the approximation

$$0 = g(x_0) \approx g(x) + g'(x)(x_0 - x).$$

Rearranging,

$$x_0 \approx x - \frac{g(x)}{g'(x)}.$$

We iterate this procedure: given an estimate x_k, obtain the new estimate x_{k+1} by

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}$$

for $k = 1, 2, 3, \ldots$, terminating when the change in x value $|x_{k+1} - x_l| = \left|\frac{g(x_k)}{g'(x_k)}\right|$ is sufficiently small.

Geometrically, $(x_{k+1}, 0)$ is the intersection of the tangent of g at $(x_k, g(x_k))$ with the x-axis, so the improved update x_{k+1} is the unique root of the linear approximation of g at x_k. The idea is to perform a first order approximation to incorporate gradient information g' to improve on the previous guess for the root.

In maximum likelihood estimation, writing $l'(\theta) = S(\theta)$ (the score), we want to find the root $S(\theta) = 0$. Using the Newton-Raphson algorithm, we update

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + \frac{S(\hat{\theta}^{(k)})}{H(\hat{\theta}^{(k)})},$$

where

$$H(\theta) = -l''(\theta) = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f_\theta(x).$$

The procedure is iterated until convergence, i.e. when the absolute difference $|\hat{\theta}^{(k)} - \hat{\theta}^{(k+1)}|$ is sufficiently small. To use the Newton-Raphson method, an initial value $\hat{\theta}^{(0)}$ is required. Then, the computations can be done using standard software (like R).

2.4.2 Fisher scoring method

This is a simple modification of the Newton-Raphson method, where one replaces H by H^*, with

$$H^*(\theta) = -E_\theta \left[l''(\theta) \right].$$

This yields the Fisher scoring algorithm:

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + \frac{S(\hat{\theta}^{(k)})}{H^*(\hat{\theta}^{(k)})}.$$

Note that H^* is the Fisher information, where we take an expectation over the data x, whereas H is the *observed* Fisher information evaluated at the data points. An advantage of Fisher scoring is that by writing $-E_\theta[l''(\theta)] = E_\theta[(l'(\theta))^2]$, we need only evaluate first-derivatives, whereas H requires computing second derivatives. This can be computationally easier, especially in higher dimensions.

Both the Newton-Raphson and Fisher scoring algorithms are employed in standard statistical packages.

3 Asymptotic theory for MLEs

3.1 Consistency

Recall that a desirable property of an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ is that is be asymptotically unbiased:

$$E_\theta[\hat{\theta}_n] \to \theta \qquad \text{as } n \to \infty.$$

This says the estimator will be centred at the truth *on average*, though it may still fluctuate a lot. We now study a stronger concept that says an estimator itself should be increasingly close to the truth as the sample size n increases.

Definition. *Consider $(X_1, \ldots, X_n)$ with X_i i.i.d., arising from a statistical model $\{P_\theta : \theta \in \Theta\}$. A sequence of estimators $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ is said to be consistent if $\hat{\theta}_n \to \theta_0$ in probability as $n \to \infty$, whenever $(X_1, \ldots, X_n)$ are drawn from P_{θ_0}, i.e.*

$$P_{\theta_0}\big(|\hat{\theta}_n - \theta_0| > \epsilon\big) \to 0 \qquad \forall \epsilon > 0, \quad \text{as } n \to \infty.$$

We usually simply write $\hat{\theta}_n \to^P \theta_0$.

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$, then by the weak law of large numbers, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \to^P \mu$, i.e. $\bar{X}_n$ is consistent for μ.*

Remark. *By Markov's inequality,*

$$P_{\theta_0}\big(|\hat{\theta}_n - \theta_0| > \epsilon\big) = P_{\theta_0}\big((\hat{\theta}_n - \theta_0)^2 > \epsilon^2\big) \le \frac{E_{\theta_0}[(\hat{\theta}_n - \theta_0)^2]}{\epsilon^2} = \frac{\mathrm{MSE}_{\theta_0}(\hat{\theta}_n)}{\epsilon^2}.$$

Therefore, if $\mathrm{MSE}_{\theta_0}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is consistent for θ_0. In other words, $\hat{\theta}_n$ is consistent if both the bias $b_{\theta_0}(\hat{\theta}) \to 0$ and variance $\mathrm{var}_{\theta_0}(\hat{\theta}) \to 0$ asymptotically vanish as $n \to \infty$.

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$, then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is consistent for μ because it is unbiased and $\mathrm{var}(\bar{X}_n) = \sigma^2/n \to 0$. Similarly, $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is consistent for σ^2. (Hint: use the fact $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$).*

The maximum likelihood method often provides good estimators. To see what one can expect, consider the simple case where $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathrm{Poisson}(\theta)$. We saw above that the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is unbiased and attains the Cramer-Rao bound $1/I_n(\theta) = 1/(nI_{X_1}(\theta)) = \theta/n$ for this problem, and hence has best possible variance among unbiased estimators. For a general estimator $\hat{\theta}_n$, a reasonable optimality criterion might be *asymptotic efficiency*:

$$n\mathrm{var}_\theta(\hat{\theta}_n) \to I^{-1}_{X_1}(\theta),$$

that is, the variance *asymptotically* attains the Cramer-Rao bound. Moreover, by the central limit theorem,

$$\sqrt{n}(\bar{X}_n - \theta) \to^d N(0, \theta),$$

which says that $\bar{X}_n \approx N(\theta, \theta/n) = N(\theta, I^{-1}_{X_1}(\theta)/n)$. We will prove more generally that the MLE is asymptotically efficient and satisfies

$$\hat{\theta}_{MLE} \approx N(\theta, I^{-1}_{X_1}(\theta)/n)$$

under suitable regularity conditions.

We now show, under some regularity conditions on the model, that the maximum likelihood estimator $\hat{\theta}_{ML}$ is consistent. For the version of the theorem shown here, we use the following set of assumptions.

Assumption 3.1 (Model regularity). *Suppose $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} f_\theta$, where f_θ is a pmf/pdf on $\mathcal{X} \subseteq \mathbb{R}$ coming from a statistical model $\{f_\theta : \theta \in \Theta\}$ with log-likelihood $l_{X_1}(\theta) = \log f_\theta$, such that:*

1. *The parameter space Θ is an open subset of $\mathbb{R}$ (i.e. $p = 1$).*
2. *$\theta \mapsto l_{X_1}(\theta)$ is twice continuously differentiable in θ for all $x \in \mathcal{X}$.*
3. *$E_\theta[l''_{X_1}(\theta)] < \infty$ for all $\theta \in \Theta$.*
4. *We can exchange integration/summation in x with two-times differentiation in θ:*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int_{\mathcal{X}} f_\theta(x) dx = \int_{\mathcal{X}} \frac{\mathrm{d}}{\mathrm{d}\theta} f_\theta(x) dx, \qquad \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \int_{\mathcal{X}} f_\theta(x) dx = \int_{\mathcal{X}} \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} f_\theta(x) dx.$$

The following results hold under significantly weaker conditions than Assumption 3.1 at the expense of more complicated proofs. Since we are more interested in the results than the proofs, we do not provide more general conditions.

Theorem 3.1 (Consistency of the MLE). *Let $\{f_\theta : \theta \in \Theta\}$ be a statistical model satisfying Assumption 3.1 and suppose $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} f_{\theta_0}$ for some true $\theta_0 \in \Theta$. Then the MLE $\hat{\theta}$ satisfies, as $n \to \infty$,*

$$\hat{\theta}_{ML} \to^P \theta_0.$$

Sketch of proof. For simplicity, assume that Θ is compact and $\theta_0 \in \text{int}(\Theta)$ (so that θ_0 is not on the boundary of Θ). Let $\epsilon > 0$ be arbitrary and define $\Theta_\epsilon = \{\theta \in \Theta : |\theta - \theta_0| \geq \epsilon\}$ and $m(\theta) = E_{\theta_0}[l_1(\theta; X_1)]$. By identifiability of the model and Lemma 2.1, θ_0 is the unique maximizer of $\theta \mapsto m(\theta)$.

Since Θ_ϵ is the intersection of a closed set and Θ, it is compact. The function $m : \Theta_\epsilon \to \mathbb{R}$ is continuous by the regularity assumptions and thus attains its maximum on Θ_ϵ, say at θ_ϵ. Thus,

$$m(\theta_\epsilon) = \sup_{\theta \in \Theta_\epsilon} m(\theta) = c_\epsilon < m(\theta_0).$$

Therefore, there exists $\delta_\epsilon > 0$ such that $c_\epsilon + \delta_\epsilon < m(\theta_0) - \delta_\epsilon$. By the triangle inequality,

$$\sup_{\theta \in \Theta_\epsilon} \tfrac{1}{n} l_n(\theta) = \sup_{\theta \in \Theta_\epsilon} \left[\tfrac{1}{n} l_n(\theta) - m(\theta) + m(\theta)\right] \leq \sup_{\theta \in \Theta_\epsilon} m(\theta) + \sup_{\theta \in \Theta} \left|\tfrac{1}{n} l_n(\theta) - m(\theta)\right|.$$

Consider the events $A_n(\epsilon) = \{\sup_{\theta \in \Theta} |\frac{1}{n} l_n(\theta) - m(\theta)| < \delta_\epsilon\}$. On $A_n(\epsilon)$, it holds that

$$\sup_{\theta \in \Theta_\epsilon} \tfrac{1}{n} l_n(\theta) \leq c_\epsilon + \delta_\epsilon < m(\theta_0) - \delta_\epsilon.$$

But also on $A_n(\epsilon)$, $m(\theta_0) - \frac{1}{n} l_n(\theta_0) \leq \delta_\epsilon$, and so

$$\sup_{\theta \in \Theta_\epsilon} \tfrac{1}{n} l_n(\theta) < m(\theta_0) - \delta_\epsilon \leq \tfrac{1}{n} l_n(\theta_0).$$

Thus on the event $A_n(\epsilon)$, $\hat{\theta}_{ML}$ cannot lie in Θ_ϵ, otherwise we would have $\frac{1}{n} l_n(\hat{\theta}_{ML}) < \frac{1}{n} l_n(\theta_0)$ contradicting that $\hat{\theta}_{ML}$ is the MLE. Therefore, $A_n(\epsilon) \subseteq \{|\hat{\theta}_{ML} - \theta_0| < \epsilon\}$. To conclude, we need to show $P_{\theta_0}(A_n(\epsilon)) \to 1$ as $n \to \infty$. This follows from the *uniform law of large numbers*:

$$\sup_{\theta \in \Theta} \left|\tfrac{1}{n} l_n(\theta) - E_{\theta_0}[l_1(\theta; X_1)]\right| \to^P 0, \tag{3.1}$$

which holds for Θ compact (proving this is beyond the scope of this module). □

3.1.1 The uniform law of large numbers (Non-examinable)

We saw in the proof of Theorem 3.1 that it is not enough to have a pointwise law of large numbers such as the weak law of large numbers. We needed uniformity in (3.1), which is a significantly stronger notion. To highlight the difference, consider the following example.

Example. *Let $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. Rademacher variables, that is $P(\varepsilon_i = +1) = P(\varepsilon_i = -1) = 1/2$. For any $u_1, \dots, u_n \in \{-1, 1\}$, we again have*

$$P(u_i\varepsilon_i = +1) = P(u_i\varepsilon_i = -1) = 1/2,$$

so that by the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} u_i\varepsilon_i \to E[u_i\varepsilon_i] = 0, \qquad a.s.$$

But

$$\sup_{(u_1,\dots,u_n)\in\{-1,1\}^n} \left|\frac{1}{n}\sum_{i=1}^{n} u_i\varepsilon_i\right| = 1 \not\to 0$$

(set $u_i = \varepsilon_i$). The problem is the complexity of $\{-1, 1\}^n$ grows too fast with n.

To understand the uniform law of large numbers, consider first the case where Θ is finite. Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{X}$-valued random variables and let $h : \mathcal{X} \to \mathbb{R}$. Then $h(X_i)$ are also i.i.d., so if $E|h(X_i)| < \infty$, the strong law of large numbers implies

$$\frac{1}{n}\sum_{i=1}^{n}[h(X_i) - Eh(X_i)] \to 0 \qquad a.s.$$

If $h_1, \dots, h_M$ are functions for a *fixed* M, this also applies to each h_j. Thus on some events A_j with $P(A_j) = 1$,

$$\frac{1}{n}\sum_{i=1}^{n}[h_j(X_i) - Eh_j(X_i)] \to 0$$

[The above should be interpreted to mean that for any $\omega \in A_j$, the last convergence holds for the sequence $X_i(\omega)$ in the sense of convergence of real numbers]. On $A = \cap_{j=1}^{M} A_j$,

$$\max_{1\le j\le M}\left|\frac{1}{n}\sum_{i=1}^{n}[h_j(X_i) - Eh_j(X_i)]\right| \to 0.$$

Since $P(A^c) = P(\cup_{j=1}^{M} A_j^c) \le \sum_{j=1}^{M} P(A_j^c) = 0$, this convergence holds almost surely. Thus the uniform law of large numbers holds in the finite case.

For an infinite class of functions, such as $\{\log f_\theta(x) : \theta \in [0, 1]\}$, we can use compactness. The idea is that by compactness, Θ can be covered by a finite subset Θ' to arbitrary precision $\delta > 0$. We then use the uniform law of large numbers for the finite set Θ' and transfer this to Θ using the continuity of $\theta \mapsto \log f_\theta(x)$, i.e. any $\theta \in \Theta$ is δ-close to some $\theta' \in \Theta'$, so $\log f_\theta(X_i)$ is close to $\log f_{\theta'}(X_i)$.

3.2 Asymptotic normality of the MLE

The MLE $\hat\theta$ is therefore *consistent*, i.e. it converges in probability to the true value of the parameter θ_0. We can further quantify the size and shape of the stochastic fluctuations of $\hat\theta - \theta_0$ for large n.

Theorem 3.2 (Asymptotic normality of the MLE). *Let $\{f_\theta : \theta \in \Theta\}$ be a statistical model satisfying Assumption 3.1 and suppose $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} f_{\theta_0}$ for some true $\theta_0 \in \Theta$. Then the MLE $\hat{\theta}$ satisfies, as $n \to \infty$,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N\left(0, \frac{1}{I_{X_1}(\theta_0)}\right).$$

Proof. Since the MLE $\hat{\theta}$ is the maximizer of the log-likelihood $l_n(\theta) = \sum_{i=1}^n \log f_\theta(X_i)$, we have $l_n'(\hat{\theta}) = 0$. Applying the mean-value theorem to $l_n'(\theta)$, there exists ξ between θ_0 and $\hat{\theta}$ such that

$$l_n''(\xi) = \frac{l_n'(\hat{\theta}) - l_n'(\theta_0)}{\hat{\theta} - \theta_0},$$

so that rearranging,

$$0 = l_n'(\hat{\theta}) = l_n'(\theta_0) + l_n''(\xi)(\hat{\theta} - \theta_0),$$

i.e. a first order Taylor expansion. From this, we obtain

$$\hat{\theta} - \theta_0 = -\frac{l_n'(\theta_0)}{l_n''(\xi)}, \qquad \text{and so} \qquad \sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\frac{1}{\sqrt{n}} l_n'(\theta_0)}{\frac{1}{n} l_n''(\xi)}. \tag{3.2}$$

By Lemma 2.1, θ_0 is the maximizer of $\theta \mapsto E_{\theta_0}[l_1(\theta)]$ and so

$$\frac{\mathrm{d}}{\mathrm{d}\theta} E_{\theta_0}[l_1(\theta)]\Big|_{\theta=\theta_0} = E_{\theta_0}[l_1'(\theta_0)] = 0$$

using the model regularity conditions. Applying the central limit theorem to the numerator of (3.2),

$$\frac{1}{\sqrt{n}} l_n'(\theta_0) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X_i)\Big|_{\theta=\theta_0} - \underbrace{E_{\theta_0}[l_1'(\theta_0)]}_{=0}\right) \to^d N(0, \mathrm{var}_{\theta_0}(l_1'(\theta_0; X_1))),$$

where we recall $\mathrm{var}_{\theta_0}(l_1'(\theta_0)) = I_{X_1}(\theta_0)$ is the Fisher information for X_1.

For the denominator of (3.2), writing $l_1(\theta; X_i) = \log f_\theta(X_i)$ to make explicit the dependence on X_i,

$$\begin{aligned}\frac{1}{n} l_n''(\xi) = \frac{1}{n}\sum_{i=1}^n l_1''(\xi; X_i) = \frac{1}{n}\sum_{i=1}^n \left[l_1''(\xi; X_i) - E_{\theta_0}[l_1''(\xi; X_i)]\right] \\ + E_{\theta_0}\left[l_1''(\xi; X_1) - l_1''(\theta_0; X_1)\right] + E_{\theta_0}[l_1''(\theta_0, X_1)].\end{aligned}$$

Since $|\xi - \theta_0| \le |\hat{\theta} - \theta_0|$ and $\hat{\theta} \to^P \theta_0$ (consistency of the MLE, Theorem 3.1), we have $\xi \to^P \theta_0$. Since $\theta \mapsto E_{\theta_0}[l_1''(\theta; X_1)]$ is continuous under the model regularity conditions, the second term then satisfies $E_{\theta_0}[l_1''(\xi; X_1) - l_1''(\theta_0; X_1)] \to^P 0$ by the continuous mapping theorem. By Lemma 2.3, the third term equals $E_{\theta_0}[l_1''(\theta_0; X_1)] = -I_{X_1}(\theta_0)$. It remains to control the first term, whose proof we only sketch. Since $\xi \to^P \theta_0$, for any $\epsilon > 0$, we have $P_{\theta_0}(|\xi - \theta_0| \le \epsilon) \to 1$ as $n \to \infty$. On the last event, the first term is bounded by

$$\sup_{\theta : |\theta - \theta_0| \le \epsilon} \left|\frac{1}{n}\sum_{i=1}^n l_1''(\theta; X_i) - E_{\theta_0}[l_1''(\theta; X_1)]\right| \to^P 0,$$

where the last convergence follows from the uniform law of large numbers (see (3.1)), whose assumptions are satisfied in this case (but we do not prove this). In summary, the denominator of (3.2) satisfies $\frac{1}{n} l_n''(\xi) \to^P -I_{X_1}(\theta_0)$.

Combining the convergence of the numerator and denominator of (3.2) with Slutsky's theorem, we get

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\sqrt{n} l_n'(\theta_0)}{l_n''(\xi)} = -\frac{\frac{1}{\sqrt{n}} l_n'(\theta_0)}{\frac{1}{n} l_n''(\xi)} \to^d \frac{1}{I_{X_1}(\theta_0)} N(0, I_{X_1}(\theta_0)) = N(0, I_{X_1}^{-1}(\theta_0)).$$

□

Remark. *For $\theta \in \mathbb{R}^p$, $p \geq 1$, one has*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N_p(0, I_{X_1}^{-1}(\theta_0)),$$

where the inverse Fisher information matrix $I_{X_1}^{-1}(\theta_0)$ is now the limiting $p \times p$ covariance matrix.

This result has important implications constructing both confidence intervals and hypothesis testing based on the MLE, see Section 6.3.

3.3 Asymptotic efficiency and the delta method

It is often difficult to exactly compute the variance of an estimator for a finite sample size n. Instead, we can approximate this through its *asymptotic variance*.

Definition. *In a parametric model $\{f_\theta : \theta \in \Theta\}$, a consistent estimator $\hat{\theta}_n$ is asymptotically efficient if $n\text{var}_{\theta_0}(\hat{\theta}_n) \to I(\theta_0)^{-1}$ for all $\theta_0 \in int(\Theta)$ [or $n\text{Cov}_{\theta_0}(\hat{\theta}_n) \to I(\theta_0)^{-1}$ if $p > 1$].*

This says the variance asymptotically attains the Cramer-Rao bound, which is the best possible variance for an unbiased estimator.

Remark.

- *Under regularity conditions, Theorem 3.2 establishes that the MLE has asymptotic variance $n^{-1} I_{X_1}^{-1}(\theta_0)$ and is hence asymptotically efficient. In quite general situations, the MLE therefore performs very well for large sample sizes.*
- *However, this result does require certain regularity assumptions and otherwise the MLE can fail to be efficient, e.g. when $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} U[0, \theta]$ and the likelihood is discontinuous.*
- *For points θ_0 at the boundary of the parameter space Θ, the asymptotics of the MLE might not be normal. For example, in the model $\{N(\theta, 1) : \theta \in [0, \infty)\}$, the case $\theta_0 = 0$ is a counterexample (see Problem Sheet 3).*

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, 1)$, $\theta \in \mathbb{R}$, then the MLE of θ equals $\hat{\theta}_{ML} = \bar{X}_n$ and the Fisher information is $I_{X_1}(\theta) = 1$. If $\theta_0 \in \mathbb{R}$ is the true parameter, by Theorem 3.2, $\hat{\theta}_{ML}$ is asymptotically efficient for θ and*

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \to^d N(0, I_{X_1}^{-1}(\theta_0)) = N(0, 1).$$

Example. *In the previous example, if the parameter of interest is $g(\theta) = e^{t\theta}$, where t is known, then from the invariance of MLE (Theorem 2.1), the MLE of $g(\theta)$ is $g(\hat{\theta}_{ML}) = g(\bar{X}_n) = e^{t\bar{X}_n}$. As the MLE of $g(\theta)$, this is asymptotically efficient by Theorem 3.2.*

The invariance of MLE in the last example allows us to transfer statements about the asymptotic variance to the MLE of $g(\theta)$. In fact, we can transfer the entire limiting distribution.

Theorem 3.3 (Delta method). *Let $g : \Theta \to \mathbb{R}$ be continuously differentiable at θ_0 with gradient $\nabla_\theta g(\theta_0) \neq 0$. Let (Y_n) be random variables such that $\sqrt{n}(Y_n - \theta_0) \to^d Z$ for a random variable Z in $\mathbb{R}^p$. Then*

$$\sqrt{n}(g(Y_n) - g(\theta_0)) \to^d \nabla_\theta g(\theta_0)^T Z.$$

Proof. Set $h(t) = g(tY_n + (1-t)\theta_0)$ for $t \in [0,1]$. Applying the mean-value theorem to h, for some $\xi \in [0,1]$ and $\tilde{\theta}_n = \xi Y_n + (1-\xi)\theta_0$,

$$g(Y_n) - g(\theta_0) = h(1) - h(0) = h'(\xi)(1-0) = \nabla_\theta g(\tilde{\theta}_n)^T (Y_n - \theta_0).$$

Since $\sqrt{n}(Y_n - \theta_0) \to^d Z$, it follows that $Y_n \to^P \theta_0$. Using $\|\tilde{\theta}_n - \theta_0\| \leq \|Y_n - \theta_0\|$, we further have $\tilde{\theta}_n \to^P \theta_0$. Since $\nabla_\theta g$ is continuous at θ_0, the continuous mapping theorem implies $\nabla_\theta g(\tilde{\theta}_n) \to^P \nabla_\theta g(\theta_0)$. Finally, since $\sqrt{n}(Y_n - \theta_0) \to^d Z$, applying Slutsky's theorem gives

$$\sqrt{n}(g(Y_n) - g(\theta_0)) = \nabla_\theta g(\tilde{\theta}_n)^T \sqrt{n}(Y_n - \theta_0) \to^d \nabla_\theta g(\theta_0)^T Z.$$

□

The most important case is when the limiting random variable Z is normal.

Corollary 3.1. *Let $g : \Theta \to \mathbb{R}$ be continuously differentiable at θ_0 with gradient $\nabla_\theta g(\theta_0) \neq 0$. Let (Y_n) be random variables such that $\sqrt{n}(Y_n - \theta_0) \to^d N(0, \Sigma)$. Then*

$$\sqrt{n}(g(Y_n) - g(\theta_0)) \to^d N(0, \nabla_\theta g(\theta_0)^T \Sigma \nabla_\theta g(\theta_0)).$$

Proof. We need only simplify the limiting distribution. But if A is an $m \times p$ matrix and $Z \sim N_p(0, \Sigma)$ in $\mathbb{R}^p$, then $AZ \sim N_m(0, A\Sigma A^T)$ as required. □

Remark. *In dimension $p = 1$, the last corollary yields that if $\sqrt{n}(Y_n - \theta_0) \to^d N(0, \sigma^2)$, then*

$$\sqrt{n}(g(Y_n) - g(\theta_0)) \to^d N(0, g'(\theta_0)^2 \sigma^2).$$

Remark. *If the MLE $\hat{\theta}_{ML}$ is asymptotically normal as in Theorem 3.2, this implies*

$$\sqrt{n}(g(\hat{\theta}_{ML}) - g(\theta_0)) \to^d N(0, \nabla_\theta g(\theta_0)^T I^{-1}(\theta_0) \nabla_\theta g(\theta_0))$$

or when dimension $p = 1$,

$$\sqrt{n}(g(\hat{\theta}_{ML}) - g(\theta_0)) \to^d N(0, g'(\theta_0)^2 I^{-1}(\theta_0)).$$

Recall that $g(\hat{\theta}_{ML})$ is the MLE of $g(\theta)$ by the invariance of MLEs (Theorem 2.1). Since the limiting covariance above is the Cramer-Rao bound for unbiased estimation of $g(\theta)$ (see equation (2.2)), the plug-in MLE $g(\hat{\theta}_{ML})$ is asymptotically efficient.

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Binomial$(1, p)$, $p \in (0,1)$, then by the central limit theorem,*

$$\sqrt{n}(\bar{X}_n - p) \to^d N(0, p(1-p)).$$

Applying the Delta method with $g(\theta) = \log\theta$, so that $g'(\theta) = 1/\theta$, gives

$$\sqrt{n}\left(\log(\bar{X}_n) - \log p\right) \to^d N(0, p(1-p)(1/p)^2).$$

It is important that the centering for Y_n is θ_0, otherwise additional terms are required in the derivatives of g, as the next example shows.

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} Exp(\lambda)$, $\lambda > 0$, we know that the MLE $\hat{\lambda}_{ML} = \frac{1}{\bar{X}_n}$. We can treat this example directly without appealing to Theorem 3.2. By the central limit theorem,*

$$\sqrt{n}(\bar{X}_n - 1/\lambda) \to^d N(0, 1/\lambda^2).$$

Set $\theta = 1/\lambda$ and $g(\theta) = 1/\theta$, so that $g'(\theta) = -1/\theta^2$. Then the Delta method yields

$$\sqrt{n}\left(\frac{1}{\bar{X}_n} - \frac{1}{\theta}\right) \to^d N(0, 1/\theta^2),$$

or in λ parametrization

$$\sqrt{n}\left(\frac{1}{\bar{X}_n} - \lambda\right) \to^d N(0, \lambda^2).$$

3.3.1 Estimating the standard error

The Delta method is useful when one is interested in obtaining the asymptotic distribution of a function of the estimator, such as the standard error (i.e. the standard deviation of the estimation). In many situations, for example hypothesis testing or constructing confidence sets, one requires an estimate of the standard error.

Example. *Returning to the last example, the standard error of* $\hat{\lambda}_{ML} = 1/\bar{X}_n$ *is approximately* $se(\hat{\lambda}) = \lambda/\sqrt{n}$ *using the asymptotic distribution. However, this is still unknown since the true* λ *is unknown. A natural estimate is to replace* λ *by* $\hat{\lambda}_{ML}$ *to get*

$$\hat{se}(\hat{\lambda}_{ML}) = \frac{\hat{\lambda}_{ML}}{\sqrt{n}} = \frac{1}{\sqrt{n}\bar{X}_n},$$

which is a function of $\bar{X}_n$*. One could therefore apply the Delta method again to obtain the limiting distribution of* $\hat{se}(\hat{\lambda}_{ML})$*.*

However, in this case we can notice that $1/(\sqrt{n}\bar{X}_n)$ *is just a multiple of* $1/\bar{X}_n$*, which we have already studied. Therefore,*

$$n\left(\frac{1}{\sqrt{n}\bar{X}_n} - \frac{\lambda}{\sqrt{n}}\right) \to^d N(0, \lambda^2),$$

so for large n*,*

$$\hat{se}(\hat{\lambda}_{ML}) = \frac{1}{\sqrt{n}\bar{X}_n} \approx N\left(\frac{\lambda}{\sqrt{n}}, \frac{\lambda^2}{n^2}\right).$$

This tells us the estimated standard error $\hat{se}(\hat{\lambda}_{ML})$ *is (asymptotically) centered at the true standard error* $se(\lambda) = \lambda/\sqrt{n}$ *with standard deviation* λ/n*. This justifies using an estimated standard error instead of the true unknown standard error.*

Similarly, while the asymptotic normality result of Theorem 3.2 establishes that the MLE has best possible asymptotic variance, this variance depends on the Fisher information at the unknown true parameter θ_0 and is hence unknown. By Theorem 3.2, in dimension $p = 1$,

$$\mathrm{se}_{\theta_0}(\hat{\theta}_{ML}) = \mathrm{var}_{\theta_0}(\hat{\theta}_{ML})^{1/2} \approx \frac{1}{\sqrt{nI(\theta_0)}}.$$

To estimate the standard error of the MLE $\hat{\theta}_{ML}$, we need to estimate the Fisher information. Plugging in the estimate $\hat{\theta}_{ML}$,

$$\widehat{\mathrm{se}}(\hat{\theta}_{ML}) = \frac{1}{\sqrt{nI(\hat{\theta}_{ML})}},$$

where $nI(\hat{\theta}_{ML})$ is called the *expected Fisher information*. Since $I(\theta_0) = -E_{\theta_0}[l''(\theta_0)]$, we can alternatively estimate $I(\theta_0)$ by

$$\hat{I}(\hat{\theta}_{ML}) = -\frac{1}{n}\sum_{i=1}^{n} l''_{X_i}(\hat{\theta}_{ML}),$$

where $n\hat{I}(\theta_0)$ is called the *observed Fisher information*.

4 Bayesian inference

4.1 Priors and posteriors

In the frequentist approach to statistics, the parameter θ is treated as a fixed unknown. In contrast, in the the Bayesian approach, we consider θ as a random variable with its own distribution π on Θ. This can be motivated by randomness in the data-generating process, can represent subjective beliefs or outside information about the true value θ_0, or can simply be methodologically convenient for statistical decision making.

To explain the Bayesian approach, suppose $X \sim f_\theta$, where $\{f_\theta : \theta \in \Theta\}$ is a statistical model with $\Theta \subseteq \mathbb{R}^p$. The prior distribution $\pi(\theta)$ of θ is the probability distribution of θ *before* observing the data. It represents our beliefs or uncertainty about the parameter before collecting any data. After observing data $X = x$, we update the distribution of θ to obtain the posterior distribution $\pi(\theta|x)$ representing our updated beliefs in light of seeing x.

By Bayes' theorem, the posterior is obtained from the prior via

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f_X(x)} = \frac{f_\theta(x)\pi(\theta)}{\int_\Theta f_{\theta'}(x)\pi(\theta')\,\mathrm{d}\theta'}.$$

The integral (or sum) $f_X(x) = \int_\Theta f_{\theta'}(x)\pi(\theta')\,\mathrm{d}\theta'$ is the marginal probability of the observable X, often called the *evidence*. Since it does not depend on θ,

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta) = L(\theta)\pi(\theta).$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

The constant of proportionality is chosen to make the total mass of the posterior distribution equal to one. Usually, we use this form instead of attempting to calculate $f_X(x)$. The data enters through the likelihood $L(\theta)$, so by the factorization criterion, inference is automatically based on any sufficient statistic.

In Bayesian statistics, inference about θ is based on the posterior distribution. Commonly used point estimators are the posterior mean and mode, but one is often interested in more complex aspects of the posterior. To determine what is the "best" Bayesian estimator, one typically considers a *loss function*. This will be discussed in Chapter 5.2.

Example. *Suppose I have 3 coins in my pocket. One has probability of heads 0.25, one has 0.5 and one has 0.75. I randomly select one coin, flip it once and observe a head. What is the probability I chose the third coin?*

Let $X = 1$ denote the event I observe a head, $X = 0$ if a tail. Let $\theta \in \{0.25, 0.5, 0.75\}$ denote the probability of a head. Our prior for θ is $\pi(0.25) = \pi(0.5) = \pi(0.75) = 1/3$ since all coins are equally likely. The probability mass function is $P_\theta(X = x) = f_\theta(x) = \theta^x(1-\theta)^{1-x}$ for $x \in \{0, 1\}$. We can now compute the posterior based on observing $x = 1$:

Parameter θ	*Prior* $\pi(\theta)$	*Likelihood* $f_\theta(1)$	*Un-normalized posterior* $f_\theta(1)\pi(\theta)$	*Posterior* $\pi(\theta\|x=1)$
0.25	*0.33*	*0.25*	*0.0825*	*0.167*
0.50	*0.33*	*0.50*	*0.1650*	*0.333*
0.75	*0.33*	*0.75*	*0.2475*	*0.500*

So after observing a head, there is now a 50% chance we have picked the third coin.

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Bernoulli(θ) and assume a Beta(α, β) prior distribution for θ:*

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad 0 < \theta < 1,$$

where $\alpha, \beta > 0$ are known. Then the posterior distribution of θ given observations $X_1 = x_1, \ldots, X_n = x_n$ satisfies (keeping track of only the θ terms)

$$\pi(\theta|x) \propto f_\theta(x)\pi(\theta) \propto \left(\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}\right) \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n x_i+\beta-1}.$$

We recognize this as the density (as a function of θ) of a Beta($\sum x_i + \alpha, n - \sum x_i + \beta$) distribution. Thus from the formula for the Beta distribution, we can read off the normalizing constant:

$$\pi(\theta|x) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\sum_{i=1}^n x_i+\alpha)\Gamma(n-\sum_{i=1}^n x_i+\beta)} \theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n x_i+\beta-1}.$$

Therefore, the posterior distribution is also a Beta distribution, but with data dependent parameters.

To gain some understanding of the posterior in the last example, consider its mean and variance. Under the prior $\theta \sim$Beta(α, β), the mean and variance are $E[\theta] = \frac{\alpha}{\alpha+\beta}$ and $\text{var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Thus under the posterior, we have

$$E[\theta|x] = \frac{\sum x_i + \alpha}{n+\alpha+\beta} = \frac{n}{n+\alpha+\beta}\bar{X}_n + \frac{\alpha+\beta}{n+\alpha+\beta}E[\theta],$$

a weighted average of the prior mean and sample mean. So for small data sizes, the prior mean plays a significant role, but for large n, $E[\theta|x] \approx \bar{X}_n$. Similarly, $\text{var}(\theta|x) = \frac{(\sum x_i+\alpha)(n-\sum x_i+\beta)}{(n+\alpha+\beta)^2(n+\alpha+\beta+1)} \approx \frac{1}{n}\bar{X}_n(1-\bar{X}_n) \to 0$ as $n \to \infty$. This reflects that as we gain more data, our uncertainty about the parameter decreases.

We can plot the posterior. Suppose we draw $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Bernoulli(0.7) (so the true $\theta = 0.7$). We see in Figure 1 that as the sample size n increases, the posterior distribution becomes increasingly concentrated around the true parameter value 0.7. This is as we would expect, since for large n, the posterior should assign greater weight to the data.

To gain an idea of the effect of the prior choice, we plot 4 different Beta priors and their corresponding posteriors in Figure 2. For small n, the prior plays a significant role, but for large n, the posteriors are very similar. This is reassuring, since we expect the likelihood to dominate the prior when there is a lot of data.

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, 1)$. Assign to θ a normal prior distribution $\pi = N(0, \tau_0^2)$, for a known constant τ_0^2. Given observations $X_1 = x_1, \cdots, X_n = x_n$, the posterior distribution of θ satisfies (keeping track of only the θ terms)*

$$\begin{aligned}
\pi(\theta|x) \propto f_\theta(x)\pi(\theta) \propto e^{-\theta^2/(2\tau_0^2)} \prod_{i=1}^{n} e^{-(x_i-\theta)^2/2} &\propto \exp\left(-\frac{\theta^2}{2\tau_0^2} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \\
&\propto \exp\left(-\theta^2\left(\frac{1}{2\tau_0^2}+\frac{n}{2}\right) + \theta\sum_{i=1}^{n} x_i\right) \\
&\propto \exp\left(-\frac{1+n\tau_0^2}{2\tau_0^2}\left(\theta^2 - \frac{2\tau_0^2}{1+n\tau_0^2}n\bar{x}_n\theta\right)\right) \\
&\propto \exp\left(-\frac{1+n\tau_0^2}{2\tau_0^2}\left(\theta - \frac{\tau_0^2}{1+n\tau_0^2}n\bar{x}_n\right)^2\right),
\end{aligned}$$

where we have completed the square in the last line. We recognize this as the form of a Gaussian density and conclude

$$\pi(\theta|x) = N\left(\frac{\tau_0^2}{1/n+\tau_0^2}\bar{x}_n, \frac{\tau_0^2}{1+n\tau_0^2}\right).$$

We can immediately read off the posterior mean and variance from this formula.

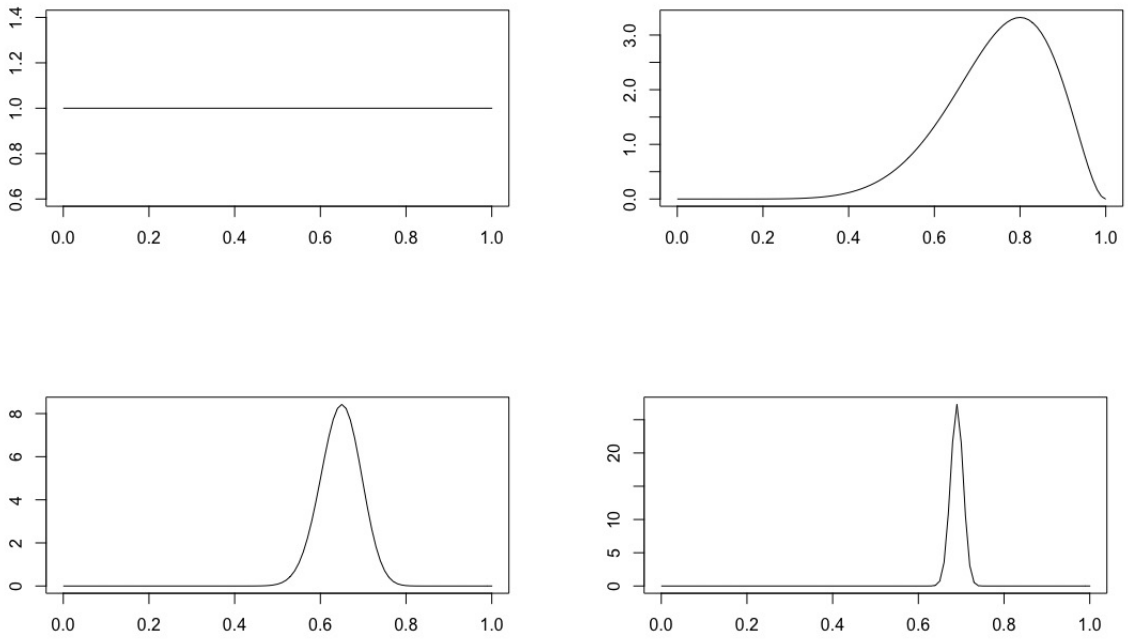


Figure 1: From left to right and top to bottom: posterior distributions based on a Beta(1,1) $= U[0,1]$ prior and observing $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Bernoulli(0.7) with (i) $n = 0$ (the prior); (ii) $n = 10$; (iii) $n = 100$; (iv) $n = 1000$.

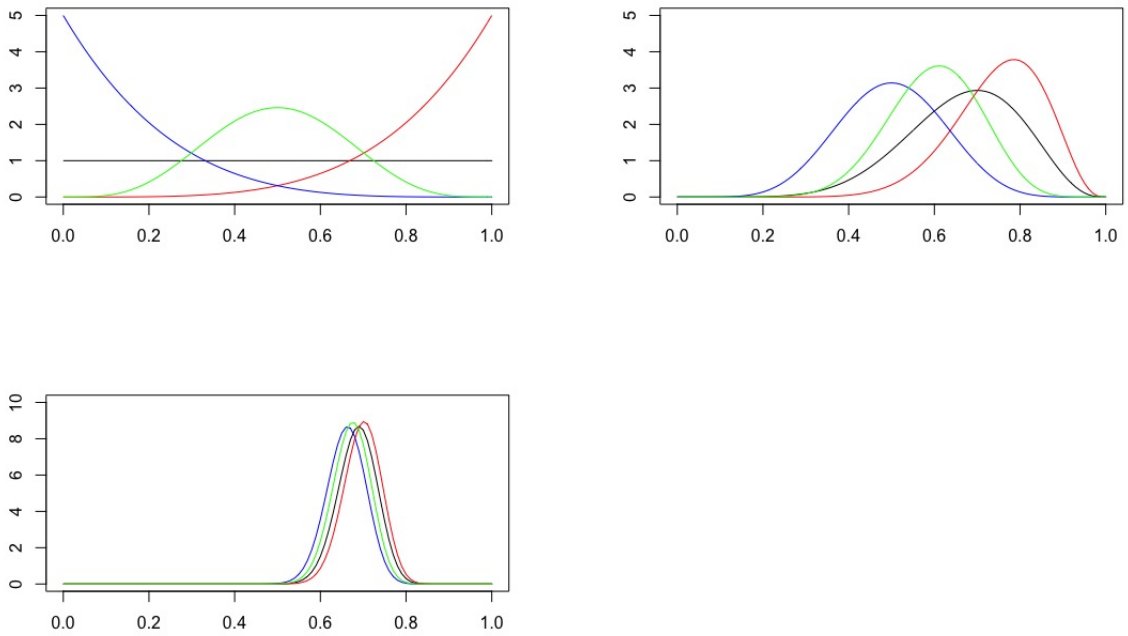


Figure 2: Effect of the choice of prior distribution: posterior distributions based Beta(1,1) (black), Beta(5,5) (green), Beta(1,5) (blue), Beta(5,1) (red) priors with (i) $n = 0$ (priors); (ii) $n = 10$; (iii) $n = 100$.

Remark. *In the last two examples, the posterior distribution is in the same family of distributions as the prior distribution. In both cases, the parameters have simply been updated based on the data. When this happens, the prior distribution is called a conjugate prior.*

Conjugate priors are typically used for computational convenience since the posterior can be computed analytically. In particular, they allow you to compute the integral $f_X(x) = \int_\Theta f_{\theta'}(x)\pi(\theta')\,\mathrm{d}\theta'$ in the denominator of Bayes' formula. For non-conjugate priors, this can be very difficult computationally.

The definition of the posterior can be extended to the case when π is not a probability distribution, i.e. does not integrate to 1. If $\int_\Theta \pi(\theta)d\theta < \infty$ is finite, then one can always renormalize π to get a proper prior distribution integrating to 1, without affecting the posterior. In general, the posterior is still well defined as long as $\int_\Theta f_\theta(x)\pi(\theta)d\theta < \infty$.

Definition. *A non-negative prior function π with $\int_\Theta \pi(\theta)d\theta = \infty$ is called an improper prior.*

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, 1)$. Assign to θ the prior $\pi(\theta) \propto 1$. This prior is improper since $\int_{-\infty}^{\infty} 1d\theta = \infty$ and hence there is no normalizing constant such that $\pi(\theta)$ integrates to one. Given observations $X_1 = x_1, \cdots, X_n = x_n$, the posterior distribution of θ satisfies*

$$\pi(\theta|x) \propto f_\theta(x)\pi(\theta) \propto \prod_{i=1}^{n} e^{-(x_i-\theta)^2/2} \propto e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2} \propto e^{-\frac{n}{2}(\theta-\bar{x}_n)^2},$$

where we have completed the square in the last step. Thus the posterior is the normal distribution $\pi(\theta|x) = N(\bar{x}_n, 1/n)$.

Note that this improper prior can be considered as the limit of the proper prior $N(0, \tau_0^2)$ we considered above as $\tau_0^2 \to \infty$. Indeed, taking $\tau_0^2 \to \infty$, the posterior distribution satisfies $N(\frac{\tau_0^2}{1/n+\tau_0^2}\bar{x}_n, \frac{\tau_0^2}{1+n\tau_0^2}) \to N(\bar{x}_n, 1/n)$.

4.2 Jeffreys priors

There may be compelling reasons to choose a particular prior, but often there are none. In this case, it is appealing to have a *default prior* or *objective prior* or *non-informative prior*. While there is not an accepted definition or choice for this, there are several possibilities.

One idea is to pick a "uniform prior" or "flat prior" that assigns equal weights to all possible values of θ and hence is as uninformative as possible. Indeed, this was a motivation for improper priors, for example taking $\pi(\theta) \propto 1$ when Θ is unbounded. However, this is not as straightforward as it seems.

Example. *Suppose $X \sim Binomial(n, p)$ and $p \sim U[0, 1]$ has a uniform prior. This seems reasonable if we are not certain about p. What if we used the reparametrization $q = \sqrt{p} \in [0, 1]$? The induced prior for q is then*

$$\Pi(q \le t) = \Pi(p < t^2) = t^2,$$

with prior density $\pi(q) = 2q$ on $[0, 1]$. This assigns a lot of mass near $q = 1$ and very little near $q = 0$.

The last example shows that a "uniform prior" in one parametrization is not necessarily so in another. For example, the odds ratio $p/(1-p)$ also gives a natural parametrization for the binomial model when considered as an exponential family, see the example in Section 1.1. One way around this is to construct a prior that is invariant under reparametrization.

Definition. *The prior $\pi(\theta) \propto \sqrt{det(I(\theta))}$ is called the Jeffreys prior. For $p = 1$, this is simply $\pi(\theta) \propto I(\theta)^{1/2}$.*

This prior might not be proper. The Cramer-Rao bound shows that the inverse of the Fisher information is a lower bound for the variance of an unbiased estimator of θ. Thus if $I(\theta)$ is large, the parameter is easy to estimate. The Jeffreys prior weighs the parameters according to the statistical difficulty with which they can be estimated, giving higher prior mass to easy to estimate parameters θ.

To give intuition for why the Jeffeys prior is "non-informative", note that large $I(\theta)$ makes the data informative about θ. If the data is informative, then it will have strong influence when forming the posterior distribution. The role of the prior is then diminished: the prior has less influence.

Since the Fisher information is additive over independent observations (Proposition 2.1), the Jeffreys prior for n i.i.d. observations is the same as for a single observation. Let us now see why the Jeffreys prior is invariant under reparametrization.

Lemma 4.1. *If θ has Jeffreys prior and $\varphi = h(\theta)$ is a smooth reparametrization, then φ also has Jeffreys prior.*

Proof. We prove this only in the case $p = 1$ and when h is a monotone increasing and differentiable function. Note that

$$\Pi(\varphi \leq t) = \Pi(h(\theta) \leq t) = \Pi(\theta \leq h^{-1}(t)).$$

Differentiating with respect to t gives prior density $\pi(\varphi) = \pi(h^{-1}(\varphi))\frac{\mathrm{d}}{\mathrm{d}\varphi}h^{-1}(\varphi) = \pi(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}\varphi}$, so

$$\begin{aligned}\pi(\varphi) = \pi(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}\varphi} \propto \left\{ I(\theta)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\varphi}\right)^2 \right\}^{1/2} &= \left\{ E_\theta\left[\left(\frac{\mathrm{d}\log f_\theta}{\mathrm{d}\theta}\right)^2\right]\left(\frac{\mathrm{d}\theta}{\mathrm{d}\varphi}\right)^2 \right\}^{1/2} \\ &= \left\{ E_\theta\left[\left(\frac{\mathrm{d}\log f_\theta}{\mathrm{d}\theta}\frac{\mathrm{d}\theta}{\mathrm{d}\varphi}\right)^2\right] \right\}^{1/2} \\ &= \left\{ E_\theta\left[\left(\frac{\mathrm{d}\log f_\theta}{\mathrm{d}\varphi}\right)^2\right] \right\}^{1/2} = I(\varphi)^{1/2}.\end{aligned}$$

The general case for p and h follows similarly using the change of variables formula for transforming probability densities: $\pi(\varphi) = \pi(\theta)|\det(\frac{\mathrm{d}\theta}{\mathrm{d}\varphi})|$, where $\frac{\mathrm{d}\theta}{\mathrm{d}\varphi}$ denotes the Jacobian matrix of the reparametrization. □

The Jeffreys prior is thus invariant under a change of coordinates of the parameter. This means the relative probability assigned to a volume of the parameter space is the same irrespective of the parametrization used to define the Jeffreys prior.

Example. *Suppose $X_1, \ldots, X_n \sim$ Poisson(λ), $\lambda > 0$. We computed above the Fisher information $I_{X_1}(\lambda) = 1/\lambda$ and so $I(\lambda) = n/\lambda$. The Jeffreys prior is then*

$$\pi(\lambda) \propto \sqrt{\frac{n}{\lambda}} \propto \frac{1}{\sqrt{\lambda}}, \qquad \lambda > 0.$$

Since $\int_0^\infty \pi(\lambda)d\lambda \propto \int_0^\infty \lambda^{-1/2}d\lambda = \infty$, the prior is improper. Note that for every $n \geq 1$, we obtain the same Jeffreys prior.

4.3 Frequentist analysis of Bayesian methods

We can analyze inference procedures based on Bayesian posterior distributions under the frequentist assumption that $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} f_{\theta_0}$ for some true θ_0. To be precise, we first compute the posterior as usual by conditioning on the data $X_1, \ldots, X_n$, and then treat (functions of) the posterior as a random quantity due to the randomness in $X_1, \ldots, X_n$.

Example. *Suppose* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, 1)$ *and we take prior* $\theta \sim N(0, 1)$*. From a previous example, we have posterior*

$$\pi(\theta|x_1, \ldots, x_n) = N\left(\frac{n}{n+1}\bar{x}_n, \frac{1}{n+1}\right).$$

The posterior mean $\bar{\theta}_n = \frac{n}{n+1}\bar{X}_n$ *is close, but not equal, to the MLE* $\hat{\theta}_{ML} = \bar{X}_n$*. Under the frequentist assumption* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta_0, 1)$*, the weak law of large numbers and Slutsky's theorem give*

$$\bar{\theta}_n = \frac{n}{n+1}\bar{X}_n \to^P \theta_0,$$

i.e. the posterior mean is consistent. We can also study the asymptotic normality of $\bar{\theta}_n$ *by writing*

$$\sqrt{n}(\bar{\theta}_n - \theta_0) = \sqrt{n}(\bar{\theta}_n - \hat{\theta}_{ML}) + \sqrt{n}(\hat{\theta}_{ML} - \theta_0).$$

By the central limit theorem (or Theorem 3.2), the second term satisfies $\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \to^d N(0, 1)$*. Expanding out the first term,*

$$\sqrt{n}(\bar{\theta}_n - \hat{\theta}_{ML}) = \sqrt{n}\left(\frac{n}{n+1} - 1\right)\bar{X}_n = -\frac{\sqrt{n}}{n+1}(\bar{X}_n - \theta_0 + \theta_0) \to^P 0,$$

using Slutsky's theorem. We can again apply Slutsky's theorem to show the sum of the two terms converges in distribution to $N(0, 1)$*. Thus the posterior mean is also asymptotically normal.*

Bayesian statistics is especially popular for *uncertainty quantification*. A $100(1-\alpha)\%$ *posterior credible set* is any set $C \subseteq \Theta$ such that

$$\Pi(C|X) = 1 - \alpha.$$

Such sets are often easier to compute than frequentist confidence sets (see Section 6.4) and are natural for *Bayesian uncertainty quantification*. In order to understand the performance of such credible sets, one needs to understand the behaviour of the posterior $\Pi(\cdot|X_1, \ldots, X_n)$ as a random probability distribution. This is beyond the scope of this module.

In fact, the whole posterior actually satisfies a much stronger form of asymptotic normality. The Bernstein-von Mises theorem says that for large n, the posterior behaves like a normal distribution centered at an efficient estimator (like the MLE) and variance equal to the Cramer-Rao bound. Let $\{f_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}$, be a parametric statistical model satisfying certain regularity conditions and suppose the prior has a pdf π that is continuous and positive at θ_0. Then if $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} f_{\theta_0}$,

$$\sup_A \left|\pi(A|X_1, \ldots, X_n) - P\left(N(\hat{\theta}_{ML}, I(\theta_0)^{-1}/n) \in A\right)\right| \to^P 0,$$

as $n \to \infty$, where the supremum is over all (Borel measurable) sets. Here, we assume $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} f_{\theta_0}$ and hence we treat the posterior $\Pi(\cdot|X_1, \ldots, X_n)$ as a random distribution. The distribution $N(\hat{\theta}_{ML}, I(\theta_0)^{-1}/n)$ is also random since it depends on the data through $\hat{\theta}_{ML}$, and this result says these two random distributions look increasingly alike as $n \to \infty$. For a precise statement and proof, see Chapter 10.2 of *Asymptotic Statistics* by A.W. van der Vaart).

As a consequence, it can be shown that any $100(1-\alpha)\%$ credible set C is also an asymptotic $100(1-\alpha)\%$ frequentist confidence set (see Section 6.4), which provides a frequentist justification for Bayesian uncertainty quantification.

5 Optimality in Estimation

5.1 Decision Theory

Thus far, we have discussed some properties that we expect good estimators to have, such as unbiasedness, small MSE, consistency and asymptotic normality. We are now going to study optimality in estimation in a general framework known as *decision theory*. Given a statistical model $\{f_\theta : \theta \in \Theta\}$ and an observation $X \in \mathcal{X}$, we can phrase many statistical problems as *decision problems* with an *action space* $\mathcal{A}$ and decision rules

$$\delta : \mathcal{X} \to \mathcal{A}.$$

Example. *Returning to some of the core problems of statistical inference:*

- *Estimation: $\mathcal{A} = \Theta$ and the decision $\delta(X) = \hat{\theta}(X)$ is an estimator.*
- *Hypothesis testing: $\mathcal{A} = \{0, 1\}$ and the decision $\delta(X)$ is a test.*
- *Uncertainty quantification: $\mathcal{A} =$ "subsets of Θ" and $\delta(X) = C(X)$ is a confidence set.*

We need a measure to assess the performance of a decision rule $\delta(X)$ and hence give a notion of optimality.

Definition. *A loss function $L : \mathcal{A} \times \Theta \to [0, \infty)$ is a non-negative function that determines the cost of a particular action for a given parameter θ.*

Example. *Some common loss functions:*

- *Estimation: we want the loss to measure the distance between our estimator and the true value. Two commonly used functions are the squared error loss and absolute error:*

$$L(a, \theta) = (a - \theta)^2 \quad \text{or} \quad L(a, \theta) = |a - \theta|.$$

- *Hypothesis testing: since we are either right or wrong, we incur cost 0 if we are right and cost 1 if we are wrong. Writing $a \in \{0, 1\}$ for the chosen hypothesis and $\theta \in \{0, 1\}$ for the true hypothesis,*

$$L(a, \theta) = \mathbb{1}_{\{a \neq \theta\}}.$$

Since a decision rule is typically based on an observation X, which incurs randomness, it is natural to consider the *expected* loss function under the distribution of X.

Definition. *For a loss function L, a decision rule δ and an observation $X \sim f_\theta$, the risk function is*

$$R(\delta, \theta) = E_\theta[L(\delta(X), \theta)] = \int_{\mathcal{X}} L(\delta(x), \theta) f_\theta(x) dx.$$

Minimizing this expected loss gives one notion of optimality.

Example. *Returning to our previous examples:*

- *Estimation: for squared error loss, the risk function $R(\delta, \theta) = E_\theta[(\delta(X) - \theta)^2] = MSE_\theta(\delta)$ is the mean-squared error. For absolute loss, $R(\delta, \theta) = E_\theta[|\delta(X) - \theta|]$ is the expected absolute error.*
- *Hypothesis testing: $R(\delta, \theta) = E_\theta[\mathbb{1}_{\{\delta(X) \neq \theta\}}] = P_\theta(\delta(X) \neq \theta)$ describes the probability of making a (type I/II) error.*

The risk function of a decision rule is a function of the parameter θ, and different decision rules can each perform better on different parts of the parameter space Θ.

Example. *If $X \sim Binomial(n, \theta)$ for $\theta \in [0,1]$, consider the estimators $\hat{\theta}_1(X) = X/n$ and $\hat{\theta}_2(X) = 1/2$ (i.e. we always guess $1/2$, irrespective of the observation). Then*

$$R(\hat{\theta}_1, \theta) = E_\theta[(\hat{\theta}_1(X) - \theta)^2] = \frac{1}{n^2}\text{var}_\theta(X) = \frac{\theta(1-\theta)}{n}$$

and

$$R(\hat{\theta}_2, \theta) = E_\theta[(\hat{\theta}_2(X) - \theta)^2] = (\theta - 1/2)^2.$$

Comparing the risk functions for $\theta \in [0,1]$, we see that $R(\hat{\theta}_2, \theta) \leq R(\hat{\theta}_1, \theta)$ if and only if $\theta \in [\frac{1}{2} - \frac{1}{\sqrt{n+1}}, \frac{1}{2} + \frac{1}{\sqrt{n+1}}]$, i.e. always guessing $1/2$ is indeed better if the true parameter is close to 1/2 (though it obviously does very badly otherwise!).

We see that in general, we cannot expect one estimator to always dominate, uniformly over the whole parameter space Θ. However, some estimators are genuinely worse than others.

Definition. *For a loss function L and parameter space Θ, a decision rule δ is inadmissible if there exists a decision rule $\delta^*(X)$ such that $R(\delta^*, \theta) \leq R(\delta, \theta)$ for all $\theta \in \Theta$, and the inequality is strict for some $\theta \in \Theta$. If no such δ^* exists, then δ is admissible.*

We rule out inadmissible decision rules, since one can then find a decision rule that performs at least as well for every parameter value, and strictly better for some parameter value.

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} Exp(\lambda)$, we know that the MLE of λ is $\hat{\lambda}_{ML} = 1/\bar{X}_n$, which is biased since $E_\lambda[\hat{\lambda}_{ML}] = \frac{n}{n-1}\lambda$ [this can be obtained using $\sum_{i=1}^n X_i \sim \Gamma(n, \lambda)$]. An unbiased estimator of λ is then*

$$\tilde{\lambda} = \frac{n-1}{n}\hat{\lambda}_{ML}.$$

Clearly,

$$\text{MSE}_\lambda(\tilde{\lambda}) < \text{MSE}_\lambda(\hat{\lambda}_{ML}) \qquad \forall \lambda > 0$$

(because $\tilde{\lambda}$ is unbiased and has a smaller variance). Therefore, the MLE $\hat{\lambda}_{ML} = \frac{1}{\bar{X}}$ is inadmissible under the squared error loss.

Remark. *In general, it is difficult to show that an estimator is admissible or not. However, some estimators are known to be admissible, as we will see below.*

5.2 Bayes risk and minimax risk

One way to compare risks over the whole parameter space is to weight the parameter space Θ using a Bayesian prior $\pi(\theta)$.

Definition. *Given a prior $\pi(\theta)$ on Θ and a loss function L, the π-Bayes risk for the decision rule δ is defined as*

$$R_\pi(\delta) = E_{\theta\sim\pi}[R(\delta, \theta)] = \int_\Theta R(\delta, \theta)\pi(\theta)d\theta = \int_\Theta \int_{\mathcal{X}} L(\delta(x), \theta) f_\theta(x)\pi(\theta)dx\, d\theta.$$

A π-Bayes decision rule δ_π is any decision rule that minimizes $R_\pi(\delta)$.

Note that the first expectation in the last definition is taken over θ, not the observation X. For estimation, where $\delta(X) = \hat{\theta}(X)$ is an estimator, the π-Bayes decision rule is often called the Bayes estimator of θ.

Example. *Consider $X \sim Binomial(n, \theta)$ with a $U[0,1]$ prior for θ. We saw above that for squared error loss, the risk function of the estimator $\hat{\theta}_1(X) = X/n$ is $R(X/n, \theta) = \theta(1-\theta)/n$. The Bayes risk is thus*

$$R_\pi(X/n) = E_\pi\left[\frac{\theta(1-\theta)}{n}\right] = \frac{1}{n}\int_0^1 \theta(1-\theta)d\theta = \frac{1}{6n}.$$

After observing $X \in \mathcal{X}$, a Bayesian updates their prior to the posterior using this data. It therefore makes sense to update the risk function based on the posterior.

Definition. *The posterior risk is defined as the average loss under the posterior distribution for an observation $X \in \mathcal{X}$:*

$$R_\pi(\delta(x)) = E_\pi[L(\delta(x), \theta)|x] = \int_\Theta L(\delta(x), \theta)\pi(\theta|x)d\theta.$$

Note the expectation in the last definition is taken over θ, and that $R_\pi(\delta(x))$ is a function of the observation $x \in \mathcal{X}$. One can sometimes minimize the posterior risk explicitly, as we now see.

Example. *Consider estimation with squared error loss $L(a, \theta) = (a-\theta)^2$. For an observation $x \in \mathcal{X}$,*

$$R_\pi(\delta(x)) = E_\pi[(\delta(x) - \theta)^2|x] = \int_\Theta (\delta(x)^2 - 2\delta(x)\theta + \theta^2)\pi(\theta|x)d\theta.$$

This is a quadratic in $\delta(x)$, which can be minimized by finding its stationary point:

$$\frac{\mathrm{d}}{\mathrm{d}\delta(x)} R_\pi(\delta(x)) = \frac{\mathrm{d}}{\mathrm{d}\delta(x)} \left[\delta(x)^2 - 2\delta(x) \int_\Theta \theta\pi(\theta|x)d\theta\right] = 2\delta(x) - 2\int_\Theta \theta\pi(\theta|x)d\theta = 0,$$

so that the minimizing decision rule is $\delta(x) = \int_\Theta \theta\pi(\theta|x)d\theta = E_\pi[\theta|x]$, the posterior mean. Note that we have done the above minimization for fixed $x \in \mathcal{X}$.

Proposition 5.1. *An estimator δ that minimizes the π-posterior risk also minimizes the π-Bayes risk.*

Proof. The π-Bayes risk equals

$$\begin{aligned}
R_\pi(\delta) &= \int_\Theta E_\theta[L(\delta(X), \theta)]\pi(\theta)d\theta \\
&= \int_\Theta \int_\mathcal{X} L(\delta(x), \theta) f_\theta(x)\pi(\theta)dx d\theta \\
&= \int_\mathcal{X} \int_\Theta L(\delta(x), \theta) \frac{f_\theta(x)\pi(\theta)}{\int_\Theta f_{\theta'}(x)\pi(\theta')d\theta'} \underbrace{\int_\Theta f_{\theta'}(x)\pi(\theta')d\theta'}_{\varphi(x)} d\theta dx \\
&= \int_\mathcal{X} E_\pi[L(\delta(x), \theta)|x]\varphi(x)dx.
\end{aligned}$$

Let δ_π be a decision rule that minimizes the *posterior risk*, i.e. for all $x \in \mathcal{X}$ and decision rules δ,

$$E_\pi[L(\delta_\pi(x), \theta)|x] \leq E_\pi[L(\delta(x), \theta)|x].$$

Multiplying both sides by the non-negative function $\varphi(x)$ and integrating over $\mathcal{X}$ gives the result. □

It is typically easier to minimize the *posterior risk* rather than the *Bayes risk*. This is convenient, since the last result shows that this immediately gives a minimizer to the Bayes risk too. The converse also holds under mild conditions: any minimizer of the Bayes risk also minimizes the posterior risk.

Proposition 5.2. *Suppose δ_π minimizes the Bayes risk $R_\pi(\delta)$ and $R_\pi(\delta_\pi) < \infty$. Then $\delta_\pi(x)$ minimizes the posterior risk $R_\pi(\delta(x))$ (with probability one under the prior predictive distribution $f_\pi(x) = \int f_\theta(x)\pi(\theta)d\theta$).*

The proof is not difficult but involves some measure theory, so we omit it. In the examples we consider here, one can ignore the reference to the prior predictive distribution in the last proposition.

Example. *Consider estimation with squared error loss $L(a,\theta) = (a-\theta)^2$. The π-Bayes estimator δ_π is the posterior mean $E_\pi[\theta|x]$ based on the prior π. Uniqueness follows from the uniqueness of the minimizer of the posterior risk.*

Example. *For absolute error $L(a,\theta) = |a-\theta|$, the π-Bayes estimator is any (not necessarily unique) posterior median (see Problem Sheet 4).*

Example. *Suppose $X_1,\dots,X_n \overset{\text{i.i.d}}{\sim}$ Bernoulli(θ) and assume a Beta(α,β) prior distribution for θ, i.e. $\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $0<\theta<1$. We saw in Section 4 that the posterior is $\theta|x \sim$ Beta$(\sum x_i+\alpha, n-\sum x_i+\beta)$. The Bayes estimator for θ under squared error loss is the posterior mean:*

$$E[\theta|x] = \frac{\sum x_i+\alpha}{n+\alpha+\beta} = \frac{n}{n+\alpha+\beta}\bar{X}_n + \frac{\alpha+\beta}{n+\alpha+\beta}\frac{\alpha}{\alpha+\beta}.$$

It is not exactly equal to the MLE $\hat{\theta}_{ML} = \bar{X}_n$, but it is close for large n.

Proposition 5.3. *If a Bayes estimator $\hat{\theta}_{Bayes}$ is unique, then it is admissible.*

Proof. See Problem Sheet 4. □

The Bayes risk allows us to consider the average loss of estimators over values of θ by considering a prior $\pi(\theta)$. Another approach is to consider the *worst case* risk over the entire parameter space Θ.

Definition. *The minimax risk is defined as the infimum ('min') over all decision rules δ of the maximal ('max') risk over the whole parameter space Θ:*

$$\inf_\delta \sup_{\theta\in\Theta} R(\delta,\theta).$$

A decision rule that attains the minimax risk is called minimax.

It is easy to show that a unique minimax estimator is admissible. Note that obtaining a minimax estimator is generally not easy. The Bayes risk, which represents an average, is never greater than the worst case risk.

Lemma 5.1. *For any decision rule δ and prior π for θ,*

$$R_\pi(\delta) \le \sup_{\theta\in\Theta} R(\delta,\theta).$$

Proof. Using the definitions,

$$R_\pi(\delta) = E_\pi[R(\delta,\theta)] \le \sup_\theta R(\delta,\theta).$$

□

Proposition 5.4. *Let π be a prior on Θ such that*

$$R_\pi(\delta_\pi) = \sup_{\theta\in\Theta} R(\delta_\pi,\theta),$$

where δ_π is a π-Bayes rule. Then δ_π is minimax. If in addition δ_π is the unique π-Bayes rule, then δ_π is unique minimax.

Proof. Let δ^* be any decision rule. Using Lemma 5.1 and that δ_π minimizes the π-Bayes risk,

$$\sup_{\theta\in\Theta} R(\delta^*,\theta) \ge R_\pi(\delta^*) \ge R_\pi(\delta_\pi) = \sup_{\theta\in\Theta} R(\delta_\pi,\theta).$$

Taking the infimum over δ^* gives

$$\inf_{\delta^*} \sup_{\theta\in\Theta} R(\delta^*,\theta) \ge \sup_{\theta\in\Theta} R(\delta_\pi,\theta),$$

i.e. δ_π is minimax. If δ_π is unique π-Bayes, then the second inequality in the first equation is strict for any $\delta^* \neq \delta_\pi$ and hence δ_π is unique minimax. □

Hence if the maximal risk of a Bayes rules equals the Bayes risk, the corresponding Bayes rule is minimax.

Corollary 5.1. *If a (unique) Bayes rule δ_π has constant risk in θ, then it is (unique) minimax.*

Proof. If a Bayes rule δ_π has constant risk, then

$$R_\pi(\delta_\pi) = E_\pi[\underbrace{R(\delta_\pi, \theta)}_{\text{constant in } \theta}] = \sup_{\theta \in \Theta} R(\delta_\pi, \theta).$$

It is thus minimax by Proposition 5.4. □

Example. *Returning to the last example, suppose we want to find a minimax estimator for θ. If we can find a prior π and corresponding Bayes rule δ_π such that the risk $R(\delta_\pi, \theta)$ is constant in θ, then δ_π is minimax.*

We saw the unique Bayes estimator under squared error loss with the prior $\theta \sim Beta(\alpha, \beta)$ is the posterior mean $\bar{\theta}_{\alpha,\beta} = E[\theta|x] = \frac{\sum x_i + \alpha}{n+\alpha+\beta}$. We can then try to solve

$$R(\bar{\theta}_{\alpha,\beta}, \theta) = constant \qquad \forall \theta \in [0, 1],$$

to find prior parameters α, β leading to constant risk, and hence a unique minimax estimator (see problem sheet). Note it is different from the MLE.

Lemma 5.2. *If δ is admissible and has constant risk, then it is minimax.*

Proof. See Problem Sheet. □

5.3 Minimum variance unbiased estimators

We have already seen some notions of optimal estimation via decision theory, where we find an optimal estimator by minimizing some concept of risk. For Bayes estimation, we choose the estimator that minimizes the average risk with respect to a prior distribution, while for minimax estimation, we pick an estimator that performs best in the worst case scenario.

We now restrict to the class of *unbiased* estimators of a parameter θ using squared error loss $L(a, \theta) = (a - \theta)^2$, so that the risk function $R(\hat{\theta}, \theta) = E_\theta[(\hat{\theta}(X) - \theta)^2] = \text{var}_\theta(\hat{\theta})$ reduces to the variance of the estimator. Minimizing the risk is thus equivalent to minimizing the estimator variance.

Definition. *Consider estimation of $g(\theta)$ based on data $X \sim P_\theta$ for a statistical model $\{P_\theta : \theta \in \Theta\}$. An unbiased estimator $\hat{g}(X)$ of $g(\theta)$ is a uniformly minimum variance unbiased estimator (UMVUE) if*

$$\text{var}_\theta(\hat{g}) \leq \text{var}_\theta(\tilde{g}) \qquad \forall \theta \in \Theta,$$

for any other unbiased estimator $\tilde{g}(X)$ of $g(\theta)$.

Since the true θ is unknown, it is crucial the above definition holds for all $\theta \in \Theta$.

Remark. *Recall that the Cramer-Rao bound provides a lower bound on the variance of any unbiased estimator. Therefore, if some estimator achieves this bound for every $\theta \in \Theta$, it is a UMVUE.*

5.3.1 Unbiased estimation

Although unbiased estimators are generally desirable, they do have potential drawbacks.

(1) Unbiased estimators may not exist.

Example. *For $X \sim Bin(n, \theta)$, $0 < \theta < 1$, no unbiased estimator exists for $g(\theta) = \frac{1}{\theta}$. Suppose such an estimator $T(X)$ existed, so that*

$$E_\theta[T(X)] = \sum_{x=0}^{n} T(x)\binom{n}{x}\theta^x(1-\theta)^{n-x} = \frac{1}{\theta}, \qquad \forall\, 0 < \theta < 1.$$

Letting $\theta \to 0$, the left-hand side tends to $T(0)$, while the right-hand side tends to infinity. But setting $T(0) = \infty$ means that $E_\theta[T(X)] = \infty \neq 1/\theta$ for all $0 < \theta < 1$, therefore no such unbiased $T(X)$ exists.

Note that the unbiasedness needs to hold over all the parameter space $\Theta = (0, 1)$.

(2) Unbiased estimators can be nonsense.

Example. *Suppose $X \sim Poisson(\lambda)$, $\lambda > 0$, and we seek an unbiased estimator for $g(\lambda) = e^{-2\lambda}$. Such an estimator $T(X)$ must satisfy*

$$E_\lambda[T(X)] = \sum_{x=0}^{\infty} T(x)\frac{e^{-\lambda}\lambda^x}{x!} = e^{-2\lambda}, \qquad \forall\, \lambda > 0.$$

Rearranging and using the exponential series for $e^{-\lambda}$,

$$\sum_{x=0}^{\infty} \frac{T(x)\lambda^x}{x!} = e^{-\lambda} = \sum_{x=0}^{\infty} \frac{(-1)^x\lambda^x}{x!}, \qquad \forall\, \lambda > 0.$$

Since this must hold for all $\lambda > 0$, this implies

$$T(x) = (-1)^x \quad x = 0, 1, 2, \ldots$$

i.e. $T(X)$ estimates $e^{-2\lambda}$ to be 1 if X is even and -1 if X is odd. Since $0 < e^{-2\lambda} < 1$, these estimates are not even in the range of the parameter $e^{-2\lambda}$ and this estimator is clearly nonsense.

In the previous example, if we instead have a sample $X_1, \ldots, X_n$, then using the same argument based on the sufficient statistic $n\bar{X}_n = \sum_{i=1}^n X_i$ yields the unbiased estimator $T(X) = (1 - \frac{2}{n})^{n\bar{X}_n}$ of $e^{-2\lambda}$. This is more sensible, since when the sample size n is large, $T(X) \approx e^{-2\bar{X}_n} \approx e^{-2\lambda}$ (one can show $T \xrightarrow{p} e^{-2\lambda}$ as $n \to \infty$).

(3) Unbiased estimators are not unique.

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} U(0, \theta)$, $\theta > 0$, then both $2\bar{X}_n$ and $\frac{n+1}{n}X_{(n)}$ are unbiased estimators of θ.*

In the last example, note that $2\bar{X}_n$ is the method of moments estimator of θ, while $\frac{n+1}{n}X_{(n)}$ is a function of a sufficient statistic. It is easy to show that $\frac{n+1}{n}X_{(n)}$ has smaller variance and hence is better than $2\bar{X}_n$ in terms of squared error loss. However, does $\frac{n+1}{n}X_{(n)}$ have minimum variance among *all* unbiased estimators of θ?

We know from the Rao-Blackwell theorem [which extends without modification from estimators of θ to estimators of $g(\theta)$] that if $T(X)$ is a sufficient statistic for θ and $\tilde{g}(X)$ is an unbiased estimator for $g(\theta)$, then $\hat{g}(X) = E[\tilde{g}(X)|T(X)]$ is unbiased for $g(\theta)$ and

$$\text{var}_\theta(\hat{g}) \leq \text{var}_\theta(\tilde{g}) \qquad \forall \theta \in \Theta,$$

with strict inequality for some $\theta \in \Theta$ unless $\tilde{g}$ is a function of T. Moreover, we know that the best possible Rao-Blackwell estimator is obtained by conditioning on a minimal sufficient statistic (Lemma 1.1). This implies that any candidate for UMVUE *must* be a function of a minimal sufficient statistic T, else we could further reduce the variance by conditioning on T.

5.4 Complete statistics

Definition. *Let X be a random variable with distribution P_θ from a parametric family $\{P_\theta : \theta \in \Theta\}$. A statistic $T = T(X)$ is said to be complete for θ if, for any (measurable) function g,*

$$\text{if } E_\theta[g(T)] = 0 \text{ for all } \theta \in \Theta, \text{ then } P_\theta(g(T) = 0) = 1 \text{ for all } \theta \in \Theta.$$

In other words, the only unbiased estimator of zero based on T is zero. To provide some intuition for this definition, recall that a set of vectors $v_1, \ldots, v_n$ is sometimes called *complete* (more commonly a basis) if they span $\mathbb{R}^n$, i.e. we can write any vector $v = \sum_{i=1}^n a_i v_i$. This is equivalent to the condition that if w is orthogonal to every $v_1, \ldots, v_n$, then $w = 0$. In the discrete case, T is complete means

$$\sum_{t \in \mathcal{T}} g(t) P_\theta(T = t) = 0 \quad \text{for all } \theta \in \Theta \implies g(t) = 0,$$

where the sum can be viewed as an inner product. Completeness thus means the family $\{P_\theta : \theta \in \Theta\}$ provides a sufficiently rich set of 'vectors' $(P_\theta(T = t))_{t \in \mathcal{T}}$ to give an orthogonality condition. This also shows that the definition of completeness depends on the *entire statistical model* $\{P_\theta : \theta \in \Theta\}$, not just T.

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Poisson(θ), $\theta > 0$, then the sufficient statistic $T = \sum_{i=1}^n X_i \sim$ Poisson$(n\theta)$ is also complete for θ. Indeed, suppose a function g satisfies*

$$E_\theta[g(T)] = \sum_{t=0}^{\infty} g(t) \frac{e^{-n\theta}(n\theta)^t}{t!} = e^{-n\theta} \sum_{t=0}^{\infty} \frac{g(t) n^t}{t!} \theta^t = 0 \qquad \forall\, \theta > 0.$$

The sum is a power series in θ, so if this is equal to zero identically, then all the coefficients must be zero, i.e. $g(t)n^t/t! = 0$ for $t = 0, 1, 2, \ldots$. This implies $g(t) = 0$ for all $t = 0, 1, 2, \ldots$, so that $P_\theta(g(T) = 0) = 1$ for all $\theta > 0$.

Example. *If $X_1, \cdots, X_n \overset{\text{i.i.d}}{\sim} U(0, \theta)$, $\theta > 0$, then the sufficient statistic $T = \max_i X_i$ is also complete for θ. Recall that T has probability density function $f_T(t) = \frac{n}{\theta^n} t^{n-1} \mathbb{1}_{(0,\theta)}(t)$. Suppose that a function g satisfies*

$$E_\theta[g(T)] = \int_0^\theta g(t) f_T(t) dt = \frac{n}{\theta^n} \int_0^\theta g(t) t^{n-1} dt = 0 \qquad \forall\, \theta > 0.$$

Differentiating the equation $\int_0^\theta g(t) t^n dt = 0$ with respect to θ yields $g(t)t^n = 0$ for all $t > 0$, so that $g(t) = 0$ for all $t > 0$. Thus $P_\theta(g(T) = 0) = 1$ for all $\theta > 0$.

Proposition 5.5. *Suppose $X = (X_1, \cdots, X_n)$ have joint distribution belonging to a k-parameter exponential family of distributions:*

$$f_\theta(x) = \exp\left\{ \sum_{i=1}^k c_i(\theta) T_i(x) - d(\theta) + S(x) \right\}.$$

If the exponential family has full rank (roughly speaking, if $c_1(\theta), \ldots, c_k(\theta)$ are linearly independent and $T_1(x), \ldots, T_k(x)$ are also linearly independent), then $T = (T_1(X), \cdots, T_k(X))$ is complete for θ.

Proof. See Lehmann, E.L. (1986), Testing Statistical Hypotheses, Theorem 4.3. □

Accordingly, many statistics and distributions we know are complete. However, the full rank condition should not be ignored. An example of a non full rank exponential family is $\{N(\theta, \theta^2) : \theta \neq 0\}$.

Example. *Suppose* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$. *Then* $T = (\sum_i X_i, \sum_i X_i^2)$ *is a complete statistic for* $\theta = (\mu, \sigma^2)$ *(because the dimensions of* T *and* θ *are equal). Similarly,* $(\bar{X}_n, S^2)$ *is also complete since it is a bijection of* T*.*

Theorem 5.1. *If a sufficient statistic* T *is complete, then it is minimal.*

Proof. Let S be an arbitrary sufficient statistic. We will show that T is a function of S. For simplicity, we assume that T is one-dimensional. If $T = (T_1, \ldots, T_d)$, the proof follows similarly by applying the following arguments coordinate-wise (i.e. replace T by T_i). Define

$$W(s) = E_\theta[T|S = s], \qquad Y(t) = E_\theta[W(S)|T = t].$$

By sufficiency of S, T, these conditional expectations do not depend on θ. We want to show that $P_\theta(T = W(S)) = 1$ for every θ, i.e. T is a function of S with probability 1. We do this in two steps.

(1) $P_\theta(T = Y(T)) = 1\ \forall\theta$: using the tower rule for conditional expectations,

$$E_\theta[Y(T)] = E_\theta[E_\theta[W(S)|T]] = E_\theta[W(S)] = E_\theta[E_\theta[T|S]] = E_\theta[T].$$

Hence for all θ, $E_\theta[Y(T) - T] = 0$, which implies $P_\theta(T = Y(T)) = 1$ for all θ by completeness of T.

(2) $P_\theta(W(S) = Y(T)) = 1\ \forall\theta$: By step (1), we have $P_\theta(E_\theta[Y(T)|S] = W(S)) = 1$. So if we can show that $P_\theta(\text{var}_\theta(Y(T)|S) = 0) = 1$, then with P_θ-probability one, the conditional distribution of $Y(T)|S$ has mean $W(S)$ and variance zero, and hence $P_\theta(Y(T) = W(S)) = 1$. By the conditional variance formula,

$$\begin{aligned}\text{var}_\theta(Y(T)) &= E_\theta[\text{var}_\theta(Y(T)|S)] + \text{var}_\theta(E_\theta[Y(T)|S]) \\ &\overset{(1)}{=} E_\theta[\text{var}_\theta(Y(T)|S)] + \text{var}_\theta(W(S)) \\ &= E_\theta[\text{var}_\theta(Y(T)|S)] + E_\theta[\text{var}_\theta(W(S)|T)] + \text{var}_\theta(Y(T)),\end{aligned}$$

which gives $E_\theta[\text{var}_\theta(Y(T)|S)] + E_\theta[\text{var}_\theta(W(S)|T)] = 0$. Because conditional variances are non-negative, we have $P_\theta(\text{var}_\theta(Y(T)|S) = 0) = 1$ as required. Combining (1) and (2) completes the proof. □

Remark. *The converse of the last theorem is <u>not</u> true, because one can have a minimal sufficient statistic that is not complete (see the example before Theorem 5.3).*

While sufficient statistics contain all the information about θ present in a sample, *ancillary* statistics contain no information about θ.

Definition. *A statistic is an <u>ancillary</u> statistic if its distribution does not depend on the parameter* θ*.*

Example. *If* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, 1)$*, then* $T(X) = \bar{X}_n$ *is <u>not</u> ancillary for* θ*, because the distribution* $\bar{X}_n \sim N\left(\theta, \frac{1}{n}\right)$ *depends on the parameter* θ*. However,* $T(X) = \max_i X_i - \min_i X_i$ *is ancillary for* θ*, since its distribution does not depend on* θ*.*

If T is complete, it turns out T contains no ancillary information for θ. In other words, a complete statistic T is independent of *any* ancillary statistic. This says T is a most efficient representation of the information about θ in the sample, since it is independent of any random variable whose distribution does not depend on θ.

Theorem 5.2 (Basu's Theorem)**.** *If* T *is a complete sufficient statistic for* θ*, then any ancillary statistic* V *is independent of* T*.*

Proof. It is enough to show that

$$P_\theta(V \in A|T) = P_\theta(V \in A)$$

for any (measurable) set A. Using the tower rule for conditional expectation, for all $\theta \in \Theta$ and A,

$$\begin{aligned} P_\theta(V \in A) &= E_\theta[\mathbb{1}(V \in A)] \\ &= E_\theta[E_\theta[\mathbb{1}(V \in A|T)]] \\ &= E_\theta[P_\theta(V \in A|T)]. \end{aligned}$$

Since V is ancillary, the probability $P_\theta(V \in A)$ is independent of θ for all sets A. Likewise, since T is sufficient, the conditional distribution of the statistic V given T, and hence $P_\theta(V \in A|T)$, does not depend on θ. Therefore,

$$E_\theta[\underbrace{P_\theta(V \in A|T) - P_\theta(V \in A)}_{g(T)}] = 0 \qquad \forall\, \theta \in \Theta.$$

By completeness, since g does not depend on θ, $P_\theta(g(T) = 0) = 1$ for all $\theta \in \Theta$. Thus with (P_θ-)probably one,

$$P_\theta(V \in A|T) - P_\theta(V \in A) = 0,$$

which completes the proof. □

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, \theta^2)$, $\theta > 0$. Then, $T = (\bar{X}_n, S^2)$ is sufficient for θ, but it is not complete. Indeed,*

$$E_\theta[\bar{X}_n] = \theta, \qquad E_\theta[cS] = \theta,$$

for some $c \neq 0$ (follows from $\frac{(n-1)S^2}{\theta^2} \sim \chi^2_{n-1}$). Hence, $E_\theta[\bar{X}_n - cS] = 0$, while $\bar{X}_n - cS \neq 0$. Therefore, $\bar{X}_n - cS$ is an unbiased estimator of zero with $P_\theta(\bar{X}_n - cS \neq 0) > 0$, which is a function of $T = (\bar{X}_n, S^2)$. In fact, $\bar{X}_n - cS$ is ancillary for θ. Note that $\{N(\theta, \theta^2) : \theta > 0\}$ is a curved exponential family and so does not have full rank.

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} U(\theta - 1, \theta + 1)$, $\theta \in \mathbb{R}$. Then, $T = (\min_i X_i, \max_i X_i)$ is (minimal) sufficient for θ, but it is not complete. One can check that $E_\theta[max_i X_i] = \theta + \frac{n-1}{n+1}$ and $E_\theta[\min_i X_i] = \theta - \frac{n-1}{n+1}$, so that*

$$E_\theta\left[\max_i X_i - \min_i X_i - \frac{2(n-1)}{n+1}\right] = 0,$$

while $P_\theta(\max_i X_i - \min_i X_i - \frac{2(n-1)}{n+1} \neq 0) = 1$. In fact, $\max_i X_i - \min_i X_i - \frac{2(n-1)}{n+1}$ is ancillary for θ and it is a function of T.

Using a complete and (minimal) sufficient statistic, we can find the best unbiased estimator (UMVUE).

Theorem 5.3 (Lehmann-Scheffe Theorem). *Let T be a sufficient and complete statistic for θ, and $\tilde{g}$ be an unbiased estimator of $g(\theta)$ with $\text{var}_\theta(\tilde{g}) < \infty$ for all $\theta \in \Theta$. If $\hat{g}(T(X)) = E[\tilde{g}(X)|T(X)]$, then $\hat{g}$ is the unique uniformly minimum variance unbiased estimator (UMVUE) of $g(\theta)$.*

Proof. Let V be another unbiased estimator of $g(\theta)$. By the Rao-Blackwell theorem, $V^*(T) = E[V|T]$ is unbiased and satisfies

$$\text{var}_\theta(V^*) \leq \text{var}_\theta(V) \qquad \forall\, \theta \in \Theta.$$

If we can show $P_\theta(\hat{g} = V^*) = 1$ for all $\theta \in \Theta$, this establishes that there is a unique $\hat{g}$ satisfying

$$\text{var}_\theta(\hat{g}) \leq \text{var}_\theta(V) \qquad \forall \theta \in \Theta$$

as required. Since both $\hat{g}$ and V^* are unbiased estimators of $g(\theta)$,

$$E_\theta[\hat{g} - V^*] = 0 \qquad \forall \theta \in \Theta.$$

Since they are both functions of the complete statistic T, this implies $P_\theta(\hat{g} - V^* = 0) = 1$ for all $\theta \in \Theta$, as desired. □

Remark.
- *The Lehmann-Scheffe Theorem states that if a complete and sufficient statistic T exists, then the UMVUE of $g(\theta)$ (if it exists) must be a function of T. In fact, any unbiased estimator based on T is the unique UMVUE.*
- *The UMVUE need not necessarily attain the Cramer-Rao bound, in which case no unbiased estimator achieves it uniformly over the parameter space. However, if an estimator attains the bound uniformly, it is the UMVUE.*

The Lehmann-Scheffe Theorem suggests two approaches to finding the UMVUE when a complete and sufficient statistic T exists:

(1) Given an unbiased estimator $\tilde{g}$ of $g(\theta)$, construct $\hat{g} = E[\tilde{g}|T]$, which is the unique UMVUE of $g(\theta)$.

(2) If we can find a function $h = h(T)$ such that $E_\theta[h(T)] = g(\theta)$, then $h(T)$ is the unique UMVUE of $g(\theta)$.

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Bernoulli(θ), $0 < \theta < 1$, we know $T = \sum_{i=1}^n X_i$ is complete and sufficient for θ. Since $E_\theta[T] = n\theta$, we have $E_\theta\left[\frac{T}{n}\right] = \theta$, and because $\frac{T}{n}$ is a function of the sufficient and complete statistic T, $\frac{T}{n} = \bar{X}_n$ is the UMVUE of θ.*

In this example, suppose we now want to find the UMVUE of $g(\theta) = \theta^2$. If $n \geq 2$, then $S = \mathbb{1}(X_1 + X_2 = 2)$ is an unbiased estimator of θ^2, because

$$E_\theta[\mathbb{1}(X_1 + X_2 = 2)] = P_\theta(X_1 + X_2 = 2) = \binom{2}{2}\theta^2(1-\theta)^{2-2} = \theta^2$$

$[X_1 + X_2 \sim \text{Binomial}\,(2, \theta)]$. By the Lehmann-Scheffe Theorem, $\hat{g} = E[S|T]$ is the UMVUE of θ^2, which we now compute:

$$\begin{aligned}
\hat{g}(t) &= E[\mathbb{1}(X_1 + X_2 = 2)|T = t] \\
&= P(X_1 + X_2 = 2|T = t) \\
&= \frac{P_\theta(X_1 + X_2 = 2, \sum_{i=1}^n X_i = t)}{P_\theta(\sum_{i=1}^n X_i = t)} \\
&= \frac{P_\theta(X_1 + X_2 = 2, \sum_{i=3}^n X_i = t-2)}{P_\theta(\sum_{i=1}^n X_i = t)} \\
&= \frac{P_\theta(X_1 + X_2 = 2)P_\theta(\sum_{i=3}^n X_i = t-2)}{P_\theta(\sum_{i=1}^n X_i = t)} \\
&= \begin{cases} \dfrac{\theta^2\binom{n-2}{t-2}\theta^{t-2}(1-\theta)^{(n-2)-(t-2)}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} & t \geq 2 \\ 0 & t < 2 \end{cases} \\
&= \begin{cases} \dfrac{\binom{n-2}{t-2}}{\binom{n}{t}} = \dfrac{t(t-1)}{n(n-1)} & t \geq 2 \\ 0 & t < 2 \end{cases} \\
&= \frac{t(t-1)}{n(n-1)}, \qquad t = 0, 1, 2, \ldots, n.
\end{aligned}$$

Therefore, $\hat{g}(T) = \frac{T(T-1)}{n(n-1)} = \frac{1}{n(n-1)}(\sum_{i=1}^n X_i)(\sum_{i=1}^n X_i - 1)$ is the UMVUE of θ^2.

Another way to obtain the UMVUE of θ^2 in this example is to find a function $h = h(T)$ such that

$$E_\theta[h(T)] = \sum_{t=0}^n h(t)\binom{n}{t}\theta^t(1-\theta)^{n-t} = \theta^2, \qquad \forall 0 < \theta < 1.$$

Setting $h(0) = h(1) = 0$, this is equivalent to

$$\begin{aligned} 1 &= \sum_{t=2}^{n} h(t)\binom{n}{t}\theta^{t-2}(1-\theta)^{n-t} \\ &= \sum_{k=0}^{n-2} h(k+2)\binom{n}{k+2}\theta^{k}(1-\theta)^{n-k+2} \\ &= \sum_{k=0}^{n-2} h(k+2)\frac{\binom{n}{k+2}}{\binom{n-2}{k}}\binom{n-2}{k}\theta^{k}(1-\theta)^{(n-2)-k}. \end{aligned}$$

Writing $1 = (1-\theta+\theta)^{n-2} = \sum_{k=0}^{n-2}\binom{n-2}{k}\theta^{k}(1-\theta)^{(n-2)-k}$, the last display is equivalent to

$$\sum_{k=0}^{n-2}\left(h(k+2)\frac{\binom{n}{k+2}}{\binom{n-2}{k}} - 1\right)\binom{n-2}{k}\theta^{k}(1-\theta)^{(n-2)-k} = 0. \tag{5.1}$$

Since this must hold for all $0 < \theta < 1$, this implies

$$h(k+2)\frac{\binom{n}{k+2}}{\binom{n-2}{k}} - 1 = 0, \qquad k = 0, 1, \ldots, n-2.$$

Rearranging,

$$h(t) = \frac{\binom{n-2}{t-2}}{\binom{n}{t}} = \frac{t(t-1)}{n(n-1)}, \qquad t = 2, 3, \ldots, n,$$

which matches the UMVUE of θ^2 derived above.

An alternative way to proceed from (5.1) is to note that the pmf in that equation is that of $K \sim \text{Binomial}(n-2, \theta)$, so (5.1) can be rewritten as $E_\theta[h(K+2)\frac{\binom{n}{K+2}}{\binom{n-2}{K}} - 1] = 0$. Since we know that K is complete for the statistical model $\{\text{Binomial}(n-2,\theta) : 0 < \theta < 1\}$ (the same model we are working in with $n-2$ instead of n), this implies $h(k+2)\frac{\binom{n}{k+2}}{\binom{n-2}{k}} - 1 = 0$ for $k = 0, 1, \ldots, n-2$.

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$, we know $T = (\bar{X}_n, S^2)$ is complete and sufficient for $\theta = (\mu, \sigma^2)$. Because $E_\theta[\bar{X}_n] = \mu$ and $\bar{X}_n$ is a function of T, $\bar{X}_n$ is the UMVUE of μ. Similarly, S^2 is the UMVUE of σ^2.*

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Poisson(θ), $\theta > 0$. It is easy to show that $\bar{X}_n$ is unbiased for θ and it is a function of the complete and sufficient statistic $T = \sum_{i=1}^{n} X_i$. Therefore, $\bar{X}_n$ is the UMVUE of θ.*

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} U(0, \theta)$, $\theta > 0$. We know $T = \max_i X_i$ is complete and sufficient for θ and that $\frac{n+1}{n}T = \frac{n+1}{n}\max_i X_i$ is unbiased for θ. Therefore, $\frac{n+1}{n}\max_i X_i$ is the UMVUE of θ. Similarly, the UMVUE of $E_\theta[X] = \frac{\theta}{2}$ is $\frac{n+1}{2n}\max_i X_i$.*

Suppose we now wish to find the UMVUE of θ^r, where $r \leq n$ is a constant. We will look for a function $h = h(T)$ such that $E_\theta[h(T)] = \theta^r$ for all $\theta > 0$. Recalling that T has density function $f_T(t) = \frac{n}{\theta^n}t^{n-1}\mathbb{1}_{(0,\theta)}(t)$, this is equivalent to

$$1 = \frac{1}{\theta^r}E_\theta[h(T)] = \int_0^\theta h(t)\frac{nt^{n-1}}{\theta^{n+r}}dt = \int_0^\theta \frac{h(t)n}{(n+r)t^r}\frac{(n+r)t^{n+r-1}}{\theta^{n+r}}dt.$$

The pdf in the right-hand integral is that of $\max_{1\leq i\leq n+r} X_i$ with $X_i \overset{\text{i.i.d}}{\sim} U(0,\theta)$, so that we can rearrange the last equation to give

$$\int_0^\theta \left(\frac{h(t)n}{(n+r)t^r} - 1\right)\frac{(n+r)t^{n+r-1}}{\theta^{n+r}}dt = 0.$$

Since $\max_{1\leq i\leq n+r} X_i$ *is complete in the model with sample size* $n+r$*, this implies* $\frac{h(t)n}{(n+r)t^r} - 1 = 0$ *and therefore*

$$h(t) = \frac{n+r}{n} t^r$$

Thus, $\frac{n+r}{n} \max_i X_i^r$ *is the UMVUE of* θ^r*. If* $r = 1$*, we already showed that* $\frac{n+1}{n} \max_i X_i$ *is the UMVUE of* θ*.*

Remark. *The UMVUE of a parameter does not necessarily exist. For example, if no unbiased estimator exists for a parameter, then the UMVUE cannot exist either. Also, if no complete statistic is available, then finding the UMVUE is not easy.*

Example. *If* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, \theta^2)$*, we know* $T = (\bar{X}_n, S^2)$ *is a minimal sufficient statistic for* θ*, but is not complete for* θ*. Therefore, we cannot use the Lehmann-Scheffe Theorem to find the UMVUE of* θ*. It has been shown that the UMVUE of* θ *does not exist actually exist for this model.*

5.4.1 Revisiting the Cramer-Rao lower bound

We now provide an example where the UMVUE does not attain the Cramer-Rao lower bound, hence no unbiased estimator attains it.

Example. *Suppose* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ *Bernoulli*(θ)*. Then*

$$f_\theta(x) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$
$$= \exp\left\{ \log\left(\frac{\theta}{1-\theta}\right) \sum_{i=1}^n x_i + n\log(1-\theta) \right\}.$$

This is an exponential family with natural statistic $T(X) = \sum_{i=1}^n X_i$*, so an estimator attains the Cramer-Rao lower bound if and only if it is of the form* $aT(X) + b$*, where* $a, b \in \mathbb{R}$*. In particular, the unbiased estimator* $\bar{X}_n$ *attains the lower bound and thus is the UMVUE (we saw this earlier using that* $T(X)$ *is sufficient and complete).*

Suppose we now wish to estimate $g(\theta) = \theta^2$*. The Cramer-Rao lower bound for any unbiased estimator* $\tilde{g}$ *of* θ^2 *(Remark after Theorem 2.2) is*

$$\text{var}_\theta(\tilde{g}) \geq \frac{g'(\theta)^2}{I(\theta)} = \frac{4\theta^2}{nI_{X_1}(\theta)} = \frac{4\theta^2}{\frac{n}{\theta(1-\theta)}}.$$

However, we saw above that the UMVUE is of the form

$$\hat{g}(T) = \frac{T(T-1)}{n(n-1)} = \frac{1}{n(n-1)} \left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n X_i - 1\right).$$

Since this is not of the form $a\sum_{i=1}^n X_i + b$*, it cannot attain the lower bound.*

6 Hypothesis testing and confidence intervals

6.1 Hypothesis testing

Often in statistics, we want to test between two hypotheses, a *null hypothesis* H_0 and *alternative hypothesis* H_1. Suppose we observe $X \sim P_\theta$, where $\{P_\theta : \theta \in \Theta\}$ is a statistical model, and we are interested in testing whether the true parameter lies in the set Θ_0 or $\Theta_1 = \Theta \backslash \Theta_0$. We formulate such hypotheses as

$$H_0 : \theta \in \Theta_0, \qquad H_1 : \theta \in \Theta_1.$$

Our goal is to test H_0 versus H_1 based on data $X \sim P_\theta$.

It is important to remember that the null hypothesis and alternative hypothesis are not considered equally. By default, we assume the null hypothesis is true. For us to reject the null hypothesis, we need a *lot* of evidence against it. This is because we consider incorrectly rejecting the null hypothesis to be a more serious error than accepting it when we should not.

Let $X \sim P_\theta$ take values in some sample space $\mathcal{X}$. We recall some definitions from M2S2 Statistical Modelling I.

Definition. *A test is a binary function $\phi : \mathcal{X} \to \{0, 1\}$ from the sample space. If $\phi(X) = \mathbb{1}_R(X)$ is an indicator function, then R is called the critical region or rejection region.*

For a test $\phi = \mathbb{1}_R$, whether the observation X falls in the rejection region R or its complement R^c determines which hypothesis we select:

$$\phi(X) = \mathbb{1}_R(X) = \begin{cases} 1 & \text{if } X \in R, \qquad (\text{reject } H_0 \implies H_1) \\ 0 & \text{if } X \notin R. \qquad (\text{do not reject } H_0 \implies H_0) \end{cases}$$

When performing a test, we may make two types of errors.

Type I error: reject H_0 when H_0 is true.

Type II error: reject H_1 when H_1 is true.

When both H_0 and H_1 are *simple* (i.e. $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$ contain a single point), we write

$$\alpha = P(\text{Type I error}) = P_{\theta_0}(X \in R),$$
$$\beta = P(\text{Type II error}) = P_{\theta_1}(X \notin R).$$

This can be summarized as follows:

	Accept H_0	Accept H_1
True H_0	correct decision	Type I error (α)
True H_1	Type II error (β)	correct decision

Type I (α) and Type II (β) errors are related: if one decreases the other typically increases. It is common to require α be smaller than some pre-specified level (e.g. $\alpha = 0.05$ or 0.01) and then try to construct a test minimizing β subject to the constraint on α.

When the alternative Θ_1 is *composite* (i.e. contains more than one point), the error probabilities do not have a single value.

Definition. *The power function $\pi_\phi : \Theta \to [0, 1]$ of a test $\phi = \mathbb{1}_R$ with rejection region R is*

$$\pi_\phi(\theta) = P_\theta(X \in R_\phi) = E_\theta[\phi(X)] = P_\theta(\textit{reject } H_0).$$

If $\Theta_1 = \{\theta_1\}$ is simple, then $\pi_\phi(\theta_1) = 1 - \beta$. A good test should have π_ϕ small for $\theta \in \Theta_0$ and large for $\theta \in \Theta_1$. However, these are not normally possible simultaneously.

Definition. *The size of a test ϕ is*

$$\alpha = \sup_{\theta \in \Theta_0} \pi_\phi(\theta).$$

This is the worst possible probability of making a Type I error over Θ_0.

Definition. *A test ϕ is a level α test if*

$$\sup_{\theta \in \Theta_0} \pi_\phi(\theta) \leq \alpha.$$

As mentioned above, one normally seeks a test ϕ maximizing $\pi_\phi(\theta)$ over the alternative hypothesis $\theta \in \Theta_1$, subject to ϕ being a level α test.

6.2 Uniformly most powerful tests

Definition. *A test ϕ is uniformly most powerful (UMP) at level α for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ if:*

(i) $\sup_{\theta \in \Theta_0} \pi_\phi(\theta) \leq \alpha$ *(level α test),*

(ii) for any other test level α test ϕ^*, we have $\pi_{\phi^*}(\theta) \leq \pi_\phi(\theta)$ for all $\theta \in \Theta_1$.

A UMP test has highest possible power, uniformly over the alternative Θ_1, subject to having Type I error bounded by α, uniformly over Θ_0. Note that UMP tests do not necessarily exist.

6.2.1 Simple hypotheses

We first consider *simple* hypotheses:

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta = \theta_1,$$

where θ_0 and θ_1 are known values. If f_{θ_i} is the pmf/pdf of X under H_i, the *likelihood ratio* of the two simple hypotheses H_0 and H_1 given data x is

$$\Lambda(x) = \Lambda(x; H_0, H_1) = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)}.$$

A *likelihood ratio test* (LRT) is one where the critical/rejection region takes the form

$$R = \{x : \Lambda(x; H_0, H_1) > k\}$$

for some k. These definitions will be presented more generally in Section 6.3. The following theorem states that the likelihood ratio test is UMP for simple hypotheses.

Lemma 6.1 (Neyman-Pearson lemma)**.** *Suppose $X \sim f_\theta(x)$ and consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Then among all tests of size α, the test with the largest power is the likelihood ratio test of size α:*

$$\phi(x) = \mathbb{1}\{x : \Lambda(x; H_0, H_1) > k\} = \begin{cases} 1 & \text{if } f_{\theta_1}(x) > k f_{\theta_0}(x), \\ 0 & \text{if } f_{\theta_1}(x) \leq k f_{\theta_0}(x), \end{cases}$$

where $k > 0$ is such that $E_{\theta_0}[\phi(X)] = P_{\theta_0}(f_{\theta_1}(X) > k f_{\theta_0}(X)) = \alpha$.

Remark. *We assume that there exists a k such that $E_{\theta_0}[\phi(X)] = \alpha$ exactly. Otherwise, we might have $E_{\theta_0}[\phi(X)] < \alpha$. This can be dealt with using the notion of a randomized test, which is discussed later.*

Proof. Let ϕ^* be any arbitrary level α test for which $E_{\theta_0}[\phi^*(X)] \leq \alpha$. It suffices to show the test $\phi(x)$ is more powerful than $\phi^*(x)$. For this, first note that

$$(\phi^*(x) - \phi(x))(\underbrace{f_{\theta_1}(x) - k f_{\theta_0}(x)}_{\begin{cases} > 0 & \text{if } \phi(x) = 1 \\ \leq 0 & \text{if } \phi(x) = 0 \end{cases}}) \leq 0 \qquad \forall\, x \in \mathcal{X}.$$

Integrating over $x \in \mathcal{X}$,

$$\int_{\mathcal{X}} (\phi^*(x) - \phi(x))(f_{\theta_1}(x) - k f_{\theta_0}(x))\, \mathrm{d}x \leq 0,$$

so that

$$\int_{\mathcal{X}} (\phi^*(x) - \phi(x)) f_{\theta_1}(x)\, \mathrm{d}x \leq k \int_{\mathcal{X}} (\phi^*(x) - \phi(x)) f_{\theta_0}(x)\, \mathrm{d}x.$$

By the assumption on the sizes of the tests, this yields

$$E_{\theta_1}[\phi^*(X)] - E_{\theta_1}[\phi(X)] \leq k(\underbrace{E_{\theta_0}[\phi^*(X)]}_{\leq \alpha} - \underbrace{E_{\theta_0}[\phi(X)]}_{\alpha}) \leq 0,$$

and thus

$$\pi_{\phi^*}(\theta_1) = E_{\theta_1}[\phi^*(X)] \leq E_{\theta_1}[\phi(X)] = \pi_{\phi}(\theta_1).$$

Therefore, ϕ is more powerful than ϕ^*. □

The Neyman-Pearson Lemma states that an optimal test statistic for these hypotheses is $\Lambda(X; H_0, H_1) = \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}$ and, for a given level α, we reject the null hypothesis H_0 if $\Lambda(X) \geq k$, where k is chosen so that the test has size α.

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, \sigma_0^2)$, where σ_0^2 is known. We want to find the best size α test of*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta = \theta_1,$$

where $\theta_1 > \theta_0$ are known values. The likelihood ratio equals

$$\begin{aligned} \Lambda(x) = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} &= \frac{\prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{1}{2\sigma_0^2}(x_i - \theta_1)^2 \right\} \right)}{\prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{1}{2\sigma_0^2}(x_i - \theta_0)^2 \right\} \right)} \\ &= \exp\left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \theta_1)^2 + \frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \theta_0)^2 \right\} \\ &= \exp\left\{ \frac{\theta_1 - \theta_0}{\sigma_0^2} \sum_{i=1}^n x_i + \frac{n(\theta_0^2 - \theta_1^2)}{2\sigma_0^2} \right\}. \end{aligned}$$

So by the Neyman-Pearson lemma, we reject H_0 if $\Lambda(x) \geq k$, where k is determined by the size α. But $\Lambda(x)$ above is an increasing function of $\bar{x}_n$, so $\Lambda(x) \geq k \iff \bar{x}_n \geq c$ for some c. Hence we reject H_0 if $\bar{x}_n \geq c$, where c is chosen such that $P(\bar{X}_n \geq c | H_0) = \alpha$.

Under H_0, $\bar{X}_n \sim N(\theta_0, \sigma_0^2/n)$, hence $Z = \sqrt{n}(\bar{X}_n - \theta_0)/\sigma_0 \sim N(0,1)$. Since $\bar{x}_n \geq c \iff z > c'$ for some c', the size α test rejects H_0 if

$$z = \frac{\sqrt{n}(\bar{x}_n - \theta_0)}{\sigma_0} \geq z_\alpha,$$

where $P(N(0,1) \geq z_\alpha) = \alpha$. Rearranging for $\bar{x}_n$, we get the test

$$\phi(x) = \begin{cases} 1 & \text{if } \bar{x}_n \geq \theta_0 + z_\alpha \frac{\sigma_0}{\sqrt{n}}, \\ 0 & \text{if } \bar{x}_n < \theta_0 + z_\alpha \frac{\sigma_0}{\sqrt{n}}. \end{cases}$$

For example, suppose $\theta_0 = 5$, $\theta_1 = 6$, $\sigma_0 = 1$, $\alpha = 0.05$, $n = 4$ and $x = \{5.1, 5.5, 4.9, 5.3\}$ so that $\bar{x}_4 = 5.2$. Using tables or software, $z_{0.05} = 1.645$ and $z = 0.4 < 1.645$, so x is not in the rejection region (the threshold for $\bar{x}_4$ is $\theta_0 + z_\alpha \frac{\sigma_0}{\sqrt{n}} = 5.82$). We thus do not reject H_0 at the 5% level. Note this does not mean we *accept* H_0, just that we do not have sufficient reason to reject it.

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} Exp(\theta)$ and consider testing*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta = \theta_1,$$

where $\theta_1 > \theta_0$. The likelihood ratio equals

$$\Lambda(x) = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} = \frac{\prod_{i=1}^n (\theta_1 e^{-\theta_1 x_i})}{\prod_{i=1}^n (\theta_0 e^{-\theta_0 x_i})} = \left(\frac{\theta_1}{\theta_0}\right)^n e^{(\theta_0 - \theta_1)\sum_{i=1}^n x_i}.$$

Using the Neyman-Pearson Lemma, we reject H_0 if $\Lambda(x) \geq k$. But $\Lambda(x)$ is a decreasing function of $\bar{x}_n$ so $\Lambda(x) \geq k \iff \bar{x}_n \leq c$ for some c determined by α. But $n\bar{X}_n = \sum_{i=1}^n X_i \sim \Gamma(n, \theta)$, so we reject H_0 if

$$P_{\theta_0}(\bar{X}_n \leq c_\alpha) = P\left(\Gamma(n, \theta_0) \leq n c_\alpha\right) = \alpha,$$

where nc_α is chosen based on the quantiles of gamma distribution.

In fact, since $\sum_{i=1}^n X_i \sim \Gamma\left(n, \theta_0\right)$ under H_0, we have $2\theta_0 \sum_{i=1}^n X_i \sim \chi^2_{2n}$ under H_0. Hence,

$$\alpha = P_{\theta_0}\left(\sum_{i=1}^n x_i \leq n c_\alpha\right) = P_{\theta_0}\left(2\theta_0 \sum_{i=1}^n X_i \leq 2\theta_0 n c_\alpha\right) = P_{\theta_0}(\chi^2_{2n} \leq 2\theta_0 n c_\alpha).$$

We thus take $2\theta_0 n c_\alpha = q_{2n}(1-\alpha)$, where $P(\chi^2_{2n} > q_{2n}(1-\alpha)) = 1 - \alpha$. Therefore, the UMP test rejects H_0 if $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \leq \frac{q_{2n}(1-\alpha)}{2\theta_0 n}$, or in other words,

$$\phi(x) = \begin{cases} 1 & \text{if } \bar{X}_n \leq \dfrac{q_{2n}(1-\alpha)}{2\theta_0 n}, \\ 0 & \text{if } \bar{X}_n > \dfrac{q_{2n}(1-\alpha)}{2\theta_0 n}. \end{cases}$$

6.2.2 UMP tests for one-sided hypotheses

Suppose $X \sim f_\theta$, where $\theta \in \Theta \subseteq \mathbb{R}$, and consider the <u>one-sided</u> hypotheses

$$H_0 : \theta \leq \theta_0, \qquad H_1 : \theta > \theta_0,$$

where θ_0 is a known value. It is possible to extend the Neyman-Pearson lemma to this setting if the family $\{f_\theta : \theta \in \Theta\}$ has a property known as monotone likelihood ratio.

Definition. *A family of distributions $\{f_\theta(x) : \theta \in \Theta\}$ is said to have <u>monotone likelihood ratio</u> if there exists a function $T(x)$ such that for any $\theta_2 > \theta_1$, the ratio $\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}$ is a non-decreasing function of $T(x)$.*

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, \sigma_0^2)$, where σ_0^2 is known. We showed previously that for any $\theta_2 > \theta_1$,*

$$\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} = \exp\left\{\frac{\theta_2 - \theta_1}{\sigma_0^2}\sum_{i=1}^n x_i + \frac{n(\theta_1^2 - \theta_2^2)}{2\sigma_0^2}\right\}.$$

This is an increasing function of $T(x) = \sum_{i=1}^n x_i$ and hence the family $\{N(\theta, \sigma_0^2) : \theta \in \mathbb{R}, \sigma_0^2 \text{ known}\}$ has monotone likelihood ratio in $T(x)$.

Lemma 6.2. *Let $X \sim f_\theta(x)$ belong to the one-parameter exponential family*

$$f_\theta(x) = \exp\{c(\theta)T(x) - d(\theta) + S(x)\}.$$

If $c(\theta)$ is a non-decreasing function of θ, then $\{f_\theta(x) : \theta \in \Theta\}$ has monotone likelihood ratio in $T(x)$.

Proof. This follows directly from the definition of a monotone likelihood ratio. □

The following theorem shows how to obtain UMP tests for one-sided hypotheses.

Theorem 6.1 (Karlin-Rubin theorem)**.** *Suppose $X \sim f_\theta(x)$ and consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. If $\{f_\theta(x) : \theta \in \Theta\}$ has monotone likelihood ratio in a statistic $T(x)$, then the UMP test at level α is*

$$\phi(x) = \mathbb{1}\{x : T(x) \geq k\} = \begin{cases} 1 & \text{if } T(x) \geq k, \\ 0 & \text{if } T(x) < k, \end{cases}$$

for k such that $P_{\theta_0}(T(X) \geq k) = \alpha$.

Proof. Due to the monotone likelihood ratio property, the power function of the test $\theta \mapsto \pi_\phi(\theta) = P_\theta((T(X) > k)$ is non-decreasing (exercise). Thus $\sup_{\theta \leq \theta_0} \pi_\phi(\theta) \leq \pi_\phi(\theta_0) = \alpha$ and this is a level α test. Now fix $\theta' > \theta_0$ and consider testing the simple hypotheses

$$H_0' : \theta = \theta_0, \qquad H_1' : \theta = \theta'. \tag{6.1}$$

By the Neyman-Pearson lemma, the level α UMP test for (6.1) rejects H_0' if $\frac{f_{\theta'}(x)}{f_{\theta_0}(x)} \geq k'$, or equivalently if $T(x) \geq k$ due to the monotone property, i.e. the test ϕ. Therefore, $\pi_{\phi'}(\theta') \leq \pi_\phi(\theta')$ for any other level α test ϕ', i.e. satisfying $\pi_{\phi'}(\theta_0) \leq \alpha$. However, any level α test of H_0 satisfies $\pi_{\phi^*}(\theta_0) \leq \sup_{\theta \leq \theta_0} \pi_{\phi^*}(\theta) \leq \alpha$ and so $\pi_\phi(\theta') \geq \pi_{\phi^*}(\theta')$. Since $\theta' > \theta_0$ was arbitrary, ϕ is the level α UMP test for H_0 versus H_1. □

Remark. *For testing $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$, the UMP test at level α is similarly*

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) \leq k, \\ 0 & \text{if } T(x) > k, \end{cases}$$

where $\alpha = P_{\theta_0}(T(X) \leq k)$. The proof is the same after interchanging the inequality directions.

Remark. *It is crucial in the Karlin-Rubin theorem that the family $\{f_\theta(x) : \theta \in \Theta\}$ has monotone likelihood ratio in a statistic $T(x)$, which is often a sufficient statistic. Fortunately, many common families, including exponential families, location families $\{f_\theta(x) = f(x - \theta) : \theta \in \Theta\}$, scale families $\{f_\theta(x) = \frac{1}{\theta} f(\frac{1}{\theta}) : \theta \in \Theta\}$ and location-scale families $\{f_{\theta_1,\theta_2}(x) = \frac{1}{\theta_2} f(\frac{x-\theta_1}{\theta_2}) : (\theta_1, \theta_2) \in \Theta\}$ have monotone likelihood ratio.*

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, \sigma_0^2)$, where σ_0^2 is known, and consider testing*

$$H_0 : \theta \leq \theta_0, \qquad H_1 : \theta > \theta_0.$$

We already showed that $\{N(\theta, \sigma_0^2) : \theta \in \mathbb{R}, \sigma_0^2 \text{ known}\}$ has monotone likelihood ratio in $T(x) = \sum_{i=1}^n x_i$, so by Karlin-Rubin, the UMP test at level α is

$$\phi(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n x_i \geq k, \\ 0 & \text{if } \sum_{i=1}^n x_i < k, \end{cases}$$

where k is chosen so that $\alpha = P_{\theta_0}(\sum_{i=1}^n X_i \geq k)$. To obtain k, note that $\sum_{i=1}^n X_i \sim N(n\theta_0, n\sigma_0^2)$ at $\theta = \theta_0$, so

$$\alpha = P_{\theta_0}\left(\sum_{i=1}^n X_i \geq k\right) = P\left(Z \geq \frac{k - n\theta_0}{\sqrt{n}\sigma_0}\right)$$

and thus $k = n\theta_0 + z_\alpha \sqrt{n}\sigma_0$ for $P(N(0,1) \geq z_\alpha) = \alpha$. The UMP test at level α thus rejects H_0 if $\sum_{i=1}^n X_i \geq n\theta_0 + z_\alpha \sqrt{n}\sigma_0$, or $\bar{X}_n \geq \theta_0 + z_\alpha \frac{\sigma_0}{\sqrt{n}}$.

Example. *In the previous example, we can similarly show that for testing $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$, the UMP test at level α rejects H_0 if $\bar{X}_n < \theta_0 - z_\alpha \frac{\sigma_0}{\sqrt{n}}$*

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ and consider testing*

$$H_0 : \sigma^2 \leq \sigma_0^2, \qquad H_1 : \sigma^2 > \sigma_0^2.$$

It is easy to show that the UMP test at level α is

$$\phi(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n x_i^2 \geq k, \\ 0 & \text{if } \sum_{i=1}^n x_i^2 < k, \end{cases}$$

where k is chosen so that $P_{\sigma_0^2}(\sum_{i=1}^n X_i^2 \geq k) = \alpha$.

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Bernoulli(θ), where $0 < \theta < 1$, and consider testing*

$$H_0 : \theta \leq \tfrac{1}{2}, \qquad H_1 : \theta > \tfrac{1}{2}.$$

One can check this family has monotone likelihood ratio in $T(x) = \sum_{i=1}^n x_i$, so by Karlin-Rubin, the UMP test at level α is

$$\phi(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n x_i \geq k, \\ 0 & \text{if } \sum_{i=1}^n x_i < k, \end{cases}$$

where k is chosen so that $P_{1/2}(\sum_{i=1}^n X_i \geq k) = P(\text{Binomial}(n, 1/2) \geq k) = \alpha$. Note that because the Binomial distribution is discrete, it is <u>not</u> possible to obtain k for all sizes α.

Remark. *In both the Neyman-Pearson lemma and Karlin-Rubin theorem, when $f_\theta(x)$ is a discrete distribution, it is <u>not always</u> possible to find k so that the size of the given tests is exactly α. For instance, in the previous example with $n = 4$ and $\theta = 1/2$,*

t	0	1	2	3	4
$P_{1/2}(\sum_{i=1}^4 X_i = t)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

and so there is no k satisfying $P(\text{Binomial}(4, 1/2) \geq k) = 0.05$ exactly.

To overcome this issue, the UMP test in the Neyman-Pearson lemma can be modified as follows

$$\phi(x) = \begin{cases} 1 & \text{if } f_{\theta_1}(x) > k f_{\theta_0}(x) \\ \gamma & \text{if } f_{\theta_1}(x) = k f_{\theta_0}(x) \\ 0 & \text{if } f_{\theta_1}(x) < k f_{\theta_0}(x) \end{cases}$$

where $0 < \gamma < 1$ and k are chosen so that $E_{\theta_0}[\phi(X)] = \alpha$, or *if this is not exactly possible* then

$$P_{\theta_0}(f_{\theta_1}(x) > k f_{\theta_0}(x)) + \gamma P_{\theta_0}(f_{\theta_1}(x) = k f_{\theta_0}(x)) = \alpha.$$

As written, this is <u>not</u> formally a test $\phi : \mathcal{X} \to \{0, 1\}$ (binary function), but the interpretation is that if we observe x such that $f_{\theta_1}(x) = k f_{\theta_0}(x)$, then with probability γ we reject H_0 and probability $1 - \gamma$ we do not. This is known as a <u>randomized test</u> and the above is a convenient way of writing this.

Similarly, the UMP test at level α in Karlin-Rubin can be modified as

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > k \\ \gamma & \text{if } T(x) = k \\ 0 & \text{if } T(x) < k \end{cases}$$

where $0 < \gamma < 1$ and k are chosen so that

$$P_{\theta_0}(T(X) > k) + \gamma P_{\theta_0}(T(X) = k) = \alpha.$$

Example. *Suppose* $X_1, \ldots, X_{10} \overset{\text{i.i.d}}{\sim}$ *Poisson*(λ) *and consider testing*

$$H_0 : \lambda \leq 0.1, \qquad H_1 : \lambda > 0.1.$$

Let $T(x) = \sum_{i=1}^{10} x_i$*. Using Karlin-Rubin and that the Poisson distribution is discrete, the UMP test at level* $\alpha = 0.05$ *is*

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > k \\ \gamma & \text{if } T(x) = k \\ 0 & \text{if } T(x) < k \end{cases}$$

where γ *and* k *are chosen so that*

$$P_{0.1}(T(x) > k) + \gamma P_{0.1}(T(x) = k) = 0.05.$$

Since $T \sim$ *Poisson*(10λ) *under* P_λ*, we have* $P_{0.1}(T > 2) = 0.080$ *and* $P_{0.1}(T > 3) = 0.019$*, so we take* $k = 3$ *(the largest* k *such that* $P_{0.1}(T > k) \leq \alpha$*). Substituting* $k = 3$ *into the last display and solving for* γ *yields* $\gamma = 0.506$*. Therefore, the UMP test at level* $\alpha = 0.05$ *is*

$$\phi(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{10} x_i > 3 \\ 0.506 & \text{if } \sum_{i=1}^{10} x_i = 3 \\ 0 & \text{if } \sum_{i=1}^{10} x_i < 3 \end{cases}$$

An issue with randomized tests is that different conclusions can be obtained using the same data.

Remark. *There is no general method to obtain UMP tests for the two-sided hypotheses*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta \neq \theta_0,$$

where θ_0 *is a known value, without making further restrictions on the class of tests.*

6.3 Likelihood ratio tests

We now consider likelihood ratio tests for more general situations. So far we have considered disjoint hypotheses Θ_0, Θ_1, but sometimes it is easier to fit the likelihood of $\Theta_1 = \Theta$ rather than $\Theta_1 = \Theta \backslash \Theta_0$, since we can then use maximum likelihood estimation.

Definition. *Let* $X \sim f_\theta(x)$*, where* $\theta \in \Theta$*. The* likelihood ratio test statistic *for testing* $H_0 : \theta \in \Theta_0$ *against* $H_1 : \theta \in \Theta_1$ *is defined as*

$$\Lambda(x) = \frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)} = \frac{L(\hat{\theta}_{ML})}{L(\hat{\theta}_0)},$$

where $\hat{\theta}_0$ *and* $\hat{\theta}_{ML}$ *are the MLEs for* θ *under the models* Θ_0 *and* Θ*, respectively.*

Note that $\Lambda(x) \geq 1$. The larger the likelihood ratio test statistic, the greater the evidence against H_0. A likelihood ratio test (LRT) at level α rejects H_0 if $\Lambda(x) \geq k$, where $k \geq 1$ is chosen so that

$$\sup_{\theta \in \Theta_0} P_\theta(\Lambda(X) \geq k) = \alpha.$$

To compute k, and thus the critical/rejection region $R = \{x : \Lambda(x) \geq k\}$, we need the distribution of the likelihood ratio statistic $\Lambda(x)$ under H_0.

Example. *Suppose* $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, 1)$ *and consider testing*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta \neq \theta_0.$$

Here $\Theta_0 = \{\theta_0\}$ *and* $\Theta = \mathbb{R}$. *Recall that the MLE in the full model* $\Theta = \mathbb{R}$ *is* $\hat{\theta}_{ML} = \bar{X}_n$. *Since* $\sup_{\theta \in \Theta_0} L(\theta) = L(\theta_0)$,

$$\Lambda(x) = \frac{L(\hat{\theta}_{ML})}{L(\theta_0)} = \frac{(2\pi)^{-\frac{n}{2}} \exp\{-\frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2\}}{(2\pi)^{-\frac{n}{2}} \exp\{-\frac{1}{2}\sum_{i=1}^n (x_i - \theta_0)^2\}}.$$

So we reject H_0 *if* $\Lambda(x) \geq k$ *is large. Taking logarithms and simplifying the expression,*

$$2\log\Lambda(x) = \left\{\sum_{i=1}^n (x_i - \theta_0)^2 - \sum_{i=1}^n (x_i - \bar{x}_n)^2\right\} = n(\bar{x}_n - \theta_0)^2.$$

Since $t \mapsto 2\log t$ *is an increasing function, we have* $\Lambda(x) \geq k \iff 2\log\Lambda(x) \geq 2\log k$. *Thus we pick* k *such that the type I error satisfies*

$$\alpha = P_{\theta_0}(\Lambda(X) \geq k) = P_{\theta_0}\left(n(\bar{X}_n - \theta_0)^2 \geq 2\log k\right) = P_{\theta_0}\left(\sqrt{n}|\bar{X}_n - \theta_0| \geq \sqrt{2\log k}\right).$$

But under H_0, $Z = \sqrt{n}(\bar{X}_n - \theta_0) \sim N(0,1)$. *Using the symmetry of* Z, *if* $P(Z > z_{\alpha/2}) = \alpha/2$,

$$P(|Z| > z_{\alpha/2}) = 2P(Z > z_{\alpha/2}) = \alpha$$

and so we set $\sqrt{2\log k} = z_{\alpha/2}$. *Hence the LRT of size* α *rejects* H_0 *if* $2\log\Lambda(x) \geq 2\log k = z_{\alpha/2}^2$, *or equivalently if* $|\bar{x}_n - \theta_0| \geq z_{\alpha/2}/\sqrt{n}$.

Alternatively, since $n(\bar{X}_n - \theta_0)^2 \sim \chi_1^2$ *under* H_0, *we reject* H_0 *if*

$$n(\bar{X}_n - \theta_0)^2 > q_1(\alpha),$$

where $P(\chi_1^2 > q_1(\alpha)) = \alpha$ *(one can check that indeed* $q_1(\alpha) = z_{\alpha/2}^2$*). Note that this is a two-tailed test, i.e. we reject* H_0 *both for high and low values of* $\bar{x}_n$.

Example. *Suppose* $X_1, \ldots, X_m \overset{\text{i.i.d}}{\sim} Exp(\theta_1)$ *and* $Y_1, \ldots, Y_n \overset{\text{i.i.d}}{\sim} Exp(\theta_2)$ *are independent and consider*

$$H_0 : \theta_1 = \theta_2, \qquad H_1 : \theta_1 \neq \theta_2.$$

Under H_1, *the MLEs for* (θ_1, θ_2) *are just the MLEs of the individual samples* (X_i) *and* (Y_i),

$$\hat{\theta}_{1,ML} = \frac{1}{\bar{X}_m}, \qquad \hat{\theta}_{2,ML} = \frac{1}{\bar{Y}_n}.$$

Under $H_0 : \theta_1 = \theta_2$, *the MLE of* $\theta_1 = \theta_2$ *is*

$$\hat{\theta}_0 = \frac{m+n}{m\bar{X}_m + n\bar{Y}_n}.$$

After some computations, the likelihood ratio test statistic can be shown to equal

$$\Lambda(x,y) = \left(\frac{m}{m+n} + \frac{n}{m+n}\frac{\bar{Y}_n}{\bar{X}_m}\right)^m \left(\frac{n}{m+n} + \frac{m}{m+n}\frac{\bar{X}_m}{\bar{Y}_n}\right)^n,$$

and the likelihood ratio test at level α *rejects* H_0 *if* $\Lambda(x,y) \geq k$. *To compute* k *such that* $P_{H_0}(\Lambda(X,Y) \geq k) = \alpha$, *we require the distribution of* $\Lambda(X,Y)$ *under* $H_0 : \theta_1 = \theta_2$. *Recall that for* $U \sim \Gamma(\alpha_U, \beta_U)$ *and* $V \sim \Gamma(\alpha_V, \beta_V)$ *independent,* $\frac{\alpha_V \beta_U U}{\alpha_U \beta_V V} \sim F_{2\alpha_U, 2\alpha_V}$. *This implies that under under* $H_0 : \theta_1 = \theta_2$,

$$\bar{X}_m/\bar{Y}_n \sim F_{2m,2n},$$

from which we can compute k.

Remark. *If $T(X)$ is a sufficient statistic for θ and $\Lambda^*(t)$ and $\Lambda(x)$ are the likelihood ratio statistics based on $T(X)$ and X, respectively, then*

$$\Lambda^*(T(x)) = \Lambda(x), \qquad \forall x \in \mathcal{X},$$

i.e. the likelihood ratios are identical as functions of x. Therefore, the LRT can equivalently be conducted based on the sufficient statistic $T(X)$. Indeed, using the factorization criterion,

$$\begin{aligned}\Lambda(x) = \frac{\sup_{\theta\in\Theta} L(\theta)}{\sup_{\theta\in\Theta_0} L(\theta)} = \frac{\sup_{\theta\in\Theta} f_\theta(x)}{\sup_{\theta\in\Theta_0} f_\theta(x)} &= \frac{\sup_{\theta\in\Theta} g(T(x),\theta)h(x)}{\sup_{\theta\in\Theta_0} g(T(x),\theta)h(x)} \\ &= \frac{\sup_{\theta\in\Theta} g(T(x),\theta)}{\sup_{\theta\in\Theta_0} g(T(x),\theta)} \\ &= \frac{\sup_{\theta\in\Theta} L^*_{T(x)}(\theta)}{\sup_{\theta\in\Theta_0} L^*_{T(x)}(\theta)} = \lambda^*(T(x)).\end{aligned}$$

6.3.1 Asymptotic distribution of the LR statistic

To conduct a likelihood ratio test, we need the distribution of the likelihood ratio (LR) statistic $\Lambda(X)$ under H_0, which we could explicitly derive for some specific examples above. The next theorem allows us to use likelihood ratio tests even when we cannot find the exact null distribution by providing the *asymptotic* null distribution.

Theorem 6.2 (Wilks' theorem). *Let $\{f_\theta : \theta \in \Theta\}$ be a statistical model satisfying Assumption 3.1, except $\Theta \subseteq \mathbb{R}^p$ for possibly $p \geq 1$. Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} f_\theta$ and consider the testing problem $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Then under H_0, as $n \to \infty$,*

$$2\log\Lambda(X) \to^d \chi^2_p.$$

Proof. We prove this only for $p = 1$. Since the MLE $\hat\theta$ is the maximizer of the log-likelihood $l_n(\theta) = \sum_{i=1}^n \log f_\theta(X_i)$, we have $l_n'(\hat\theta) = 0$. Using a second-order Taylor expansion about $\hat\theta$,

$$l_n(\theta_0) = l_n(\hat\theta) + l_n'(\hat\theta)(\theta_0 - \hat\theta) + \frac{1}{2!}l_n''(\xi)(\theta_0 - \hat\theta)^2 = l_n(\hat\theta) + \frac{1}{2!}l_n''(\xi)(\theta_0 - \hat\theta)^2$$

for some ξ between θ_0 and $\hat\theta$. Substituting this into the definition of $\Lambda(x)$:

$$2\log\Lambda(x) = 2\log\frac{L_n(\hat\theta)}{L_n(\theta_0)} = 2l_n(\hat\theta) - 2_n(\theta_0) = -l_n''(\xi)(\theta_0 - \hat\theta)^2.$$

Under Assumption 3.1 and $H_0 : \theta = \theta_0$, Theorem 3.2 implies $\sqrt{n}(\hat\theta - \theta_0) \to^d N\left(0, \frac{1}{I_{X_1}(\theta_0)}\right)$, and hence by the continuous mapping theorem,

$$n(\hat\theta - \theta_0)^2 \to^d \frac{\chi^2_1}{I_{X_1}(\theta_0)}.$$

Turning to $-l_n''(\xi)$, note that $|\xi - \theta_0| \leq |\hat\theta - \theta_0|$ and $\hat\theta \to^P \theta_0$ (Theorem 3.1, consistency of the MLE), so that $\xi \to^P \theta_0$. Under $H_0 : \theta = \theta_0$, by the weak law of large numbers and the continuous mapping theorem,

$$-\frac{1}{n}l_n''(\xi) = -\frac{1}{n}\sum_{i=1}^n l_1''(\xi; X_i) \to^P -E_{\theta_0}[l_1''(\theta_0; X_1)] = I_{X_1}(\theta_0).$$

Therefore, by Slutsky's theorem,

$$2\log\Lambda(x) = -\frac{1}{n}l_n''(\xi)n(\theta_0 - \hat\theta)^2 \to^d I_{X_1}(\theta_0)\frac{\chi^2_1}{I_{X_1}(\theta_0)} = \chi^2_1.$$

□

Remark. *The LRT of (asymptotic) size α thus rejects H_0 if $2\log\Lambda(x) > k_\alpha$, where k_α satisfies $P(\chi^2_p > k_\alpha) = \alpha$.*

Remark. *Wilks' theorem extends to composite null hypotheses. If $H_0 : \theta \in \Theta_0$, then under H_0,*

$$2\log\Lambda(X) \to^d \chi^2_r,$$

where $r = |\Theta| - |\Theta_0|$ and $|\Theta_0|$, $|\Theta|$ are the number of free parameters specified by Θ_0, Θ, respectively (roughly the dimension of Θ_0, Θ).

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Poisson(θ) and consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Then, since $\hat{\theta}_{ML} = \bar{X}_n$,*

$$\Lambda(x) = \frac{L(\hat{\theta}_{ML})}{L(\hat{\theta}_0)} = \frac{e^{-n\bar{x}_n}\bar{x}_n^{\sum_{i=1}^n x_i}}{e^{-n\theta_0}\theta_0^{\sum_{i=1}^n x_i}} = e^{n(\theta_0 - \bar{x}_n)}\left(\frac{\bar{x}_n}{\theta_0}\right)^{\sum_{i=1}^n x_i}.$$

Therefore,

$$2\log\Lambda(x) = 2n(\theta_0 - \bar{x}_n) + 2n\bar{x}_n \log\left(\frac{\bar{x}_n}{\theta_0}\right).$$

The asymptotic likelihood ratio test at level α rejects H_0 if $2\log\Lambda(x) \geq \chi^2_1(\alpha)$.

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} f_\theta(x)$, where $f_\theta(x)$ is a multinomial distribution with $f_\theta(x) = P_\theta(X = k) = p_k$, $k = 1, \ldots, 5$, and $\sum_{k=1}^5 p_k = 1$. Consider testing*

$$H_0 : p_1 = p_2 = p_3 \text{ and } p_4 = p_5, \qquad H_1 : p_1, \ldots, p_5 \text{ unconstrained.}$$

The full parameter space (H_1) has 4 free parameters $p_1, \ldots, p_4$, since $p_5 = 1 - \sum_{i=1}^4 p_i$. There is only one free parameter under H_0, since for $q \in [0, 1]$ we could write $p_1 = p_2 = p_3 = q$ and $p_4 = p_5 = (1 - 3q)/2$. The asymptotic distribution of $2\log\Lambda(X)$ is thus $\chi^2_{4-1} = \chi^2_3$.

Remark. *When a UMP test exists, the likelihood ratio test is often UMP. Even when UMP tests do not exist, the likelihood ratio test often works.*

6.4 Confidence Intervals

Definition. *Let $X \sim f_\theta$. For $0 < \alpha < 1$, a set $C = C(X)$ is called a $100(1-\alpha)\%$ confidence set (or interval if $p = 1$) for θ if*

$$P_\theta(\theta \in C(X)) = 1 - \alpha$$

(or $\geq 1 - \alpha$) for all $\theta \in \Theta$. The probability $1 - \alpha$ is called the coverage.

Observe that $C = C(X)$ is a set-valued statistic of the data X. It is important to note that having observed some data x and calculated a 95% confidence set $C(x)$, we *cannot* say that θ has a 95% chance of being within the set. The parameter θ is a fixed value and is either in the set or not, and hence we cannot assign probabilities to this event. We can instead interpret this in terms of repeated sampling. If we calculate $C(x)$ for a large number of samples $X = x$, then approximately $100(1 - \alpha)\%$ of them will cover (contain) the true value of θ. This distinction is illustrated in the following example.

Example. *Suppose $X_1, X_2 \overset{\text{i.i.d}}{\sim} U(\theta - 1/2, \theta + 1/2)$ and we want to construct a 50% confidence interval for θ. We know that each X_i is equally likely to be less than θ or greater than θ. So there is a 50% chance we get one observation on each side:*

$$P_\theta(\min(X_1, X_2) \leq \theta \leq \max(X_1, X_2)) = 0.5.$$

Thus $[\min(X_1, X_2), \max(X_1, X_2)]$ *is a 50% confidence interval for* θ.

But suppose during the experiment we obtain $|x_1 - x_2| \geq 1/2$, *e.g.* $x_1 = 0.2$ *and* $x_2 = 0.9$. *Then we know that, in this particular case,* θ *must lie in* $[\min(X_1, X_2), \max(X_1, X_2)]$. *In this instance, we are 100% sure this 50% confidence interval contains the true* θ. *This is why we should not say "there is a* $100(1-\alpha)\%$ *chance that* θ *lies in here". The confidence interval says "if we keep making these intervals,* $100(1-\alpha)\%$ *of them will contain* θ". *After we have calculated a particular confidence interval, the probability that that particular interval contains* θ *is not* $100(1-\alpha)\%$.

We discuss two ways to construct confidence sets. The first involves a *pivotal quantity*.

Definition. *A random variable* $Q(X, \theta)$ *is a pivotal quantity if its distribution does not depend on the parameter* θ.

Such quantities can be used to construct confidence intervals:

1. Find a pivotal quantity $Q(X, \theta)$ such that the P_θ-distribution of $Q(X, \theta)$ does not depend on θ.
2. Write down a probability statement of the form $P_\theta(a \leq Q(X, \theta) \leq b) = 1 - \alpha$.
3. Rearrange the inequalities inside $P_\theta(\dots)$ to find the interval.

Example. *Suppose* $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, 1)$ *and we want to construct a 95% confidence interval for* μ. *We know* $\bar{X}_n \sim N(\mu, 1/n)$, *so* $Q(X, \mu) = \sqrt{n}(\bar{X}_n - \mu) \sim N(0, 1)$ *is pivotal quantity for* μ. *For* $Z \sim N(0, 1)$, *we can find* a, b *such that*

$$P(a \leq Z \leq b) = P_\mu(a \leq \sqrt{n}(\bar{X}_n - \mu) \leq b) = 1 - \alpha$$

for all $\mu \in \mathbb{R}$. *Rearranging the inequalities in the last equation,*

$$P_\mu\left(\bar{X}_n - \frac{b}{\sqrt{n}} \leq \mu \leq \bar{X}_n - \frac{a}{\sqrt{n}}\right) = 1 - \alpha.$$

There are many possible choices for a *and* b. *For the* $N(0, 1)$ *density, the shortest such interval is of the form* $[-b, b]$, *and we thus take* $b = -a = z_{\alpha/2}$, *where* $P(N(0, 1) \geq z_{\alpha/2}) = \alpha/2$ *(e.g.* $z_{0.025} = 1.96$ *for* $\alpha = 0.05$*). A 95% confidence interval for* μ *is then* $[\bar{X}_n - 1.96/\sqrt{n}, \bar{X}_n + 1.96/\sqrt{n}]$. *We could also take other values of* a, b, *such as* $a = -\infty$ *and* $b = 1.64$, *but we usually want the shortest interval.*

Note this can be considered as a random interval, since it depends on the random quantity $\bar{X}_n$. *If instead* $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, *where* σ^2 *is unknown, then a pivotal quantity for* μ *is* $Q(X, \mu) = \frac{\bar{X}_n - \mu}{(S/\sqrt{n})} \sim t_{n-1}$, *a t-distribution.*

Notice in the last example that it is the endpoints of the confidence interval that are random quantities, while μ is a fixed constant that we want to find out.

Example. *One can construct asymptotic confidence intervals using asymptotically pivotal quantities. Suppose* $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} f_\theta$. *By the asymptotic normality of the MLE (Theorem 3.2),*

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \to^d N(0, I_{X_1}^{-1}(\theta_0)).$$

This gives the following (asymptotic) $100(1-\alpha)\%$ *confidence interval for* θ*:*

$$\left[\hat{\theta}_{ML} - z_{\frac{\alpha}{2}} \frac{1}{\sqrt{nI(\theta_0)}}, \hat{\theta}_{ML} + z_{\frac{\alpha}{2}} \frac{1}{\sqrt{nI(\theta_0)}}\right].$$

Since the true unknown parameter θ_0 *appears in the variance, we must replace it by a good estimator like* $\hat{\theta}_{ML}$, *giving*

$$\left[\hat{\theta}_{ML} - z_{\frac{\alpha}{2}} \frac{1}{\sqrt{nI(\hat{\theta}_{ML})}}, \hat{\theta}_{ML} + z_{\frac{\alpha}{2}} \frac{1}{\sqrt{nI(\hat{\theta}_{ML})}}\right].$$

Confidence intervals or sets can also be obtained by inverting hypothesis tests, and vice-versa.

Definition. *The acceptance region A of a test is the complement of the critical/rejection region R.*

This is the set of observations $x \in \mathcal{X}$ for which we do *not* reject H_0 ("acceptance" is used for historic reasons).

Theorem 6.3. *For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. Then the set*

$$C(X) = \{\theta : X \in A(\theta)\}$$

is a $100(1-\alpha)\%$ confidence set for θ. Conversely, let $C(X)$ be a $100(1-\alpha)\%$ confidence set for θ. Then

$$A(\theta_0) = \{X : \theta_0 \in C(X)\}$$

is the acceptance region for a level α test of $H_0 : \theta = \theta_0$.

Proof. Note that $\theta_0 \in C(X) \iff X \in A(\theta_0)$. For the first statement, since the test has size α,

$$P(\text{accept } H_0 | H_0 \text{ is true}) = P_{\theta_0}(X \in A(\theta_0)) = 1 - \alpha.$$

Thus, $P_{\theta_0}(\theta_0 \in C(X)) = 1 - \alpha$ and so $C(X)$ is a $100(1-\alpha)\%$ confidence set.

For the converse, since $C(X)$ is a $100(1-\alpha)\%$ confidence set, we have $P_{\theta_0}(\theta_0 \in C(X)) = 1 - \alpha$. Thus

$$P_{\theta_0}(X \in A(\theta_0)) = P_{\theta_0}(\theta_0 \in C(X)) = 1 - \alpha,$$

i.e. $A(\theta_0)$ is the acceptance region of a level α test. □

Intuitively, this says that "confidence intervals" and "hypothesis acceptance/rejection" are equivalent. After observing X, we can produce a 95% confidence interval (a, b). Then to test the hypothesis $H_0 : \theta = \theta_0$, we simply need to check whether $\theta_0 \in (a, b)$. On the other hand, if we have a test for $H_0 : \theta = \theta_0$, then a confidence interval is all θ_0's such that we would accept $H_0 : \theta = \theta_0$.

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, 1)$ and we want a $100(1-\alpha)\%$ confidence interval for μ. We previously constructed a level α likelihood ratio test for $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ that rejects H_0 if $\sqrt{n}|\bar{x}_n - \mu_0| > z_{\alpha/2}$, where $P(N(0,1) > z_{\alpha/2}) = \alpha/2$. Thus the acceptance region is*

$$\begin{aligned} A(\mu_0) &= \left\{(x_1, \ldots, x_n) : \sqrt{n}|\bar{x}_n - \mu_0| \leq z_{\alpha/2}\right\} \\ &= \left\{(x_1, \ldots, x_n) : \bar{x}_n - z_{\alpha/2}/\sqrt{n} \leq \mu_0 \leq \bar{x}_n + z_{\alpha/2}/\sqrt{n}\right\}. \end{aligned}$$

A $100(1-\alpha)\%$ confidence interval is then

$$\begin{aligned} C(X) = \{\mu : x \in A(\mu)\} &= \{\mu : \bar{x}_n - z_{\alpha/2}/\sqrt{n} \leq \mu \leq \bar{x}_n + z_{\alpha/2}/\sqrt{n} \\ &= [\bar{x}_n - z_{\alpha/2}/\sqrt{n}, \bar{x}_n + z_{\alpha/2}/\sqrt{n}] \end{aligned}$$

as we derived before.

Example. *Returning to the last example, we can also construct a one-sided confidence interval for μ. For testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$, we constructed using the Karlin-Rubin theorem a (UMP) level α test where we reject H_0 if $\bar{X}_n \geq \mu_0 + z_\alpha/\sqrt{n}$. Thus the acceptance region is*

$$A(\mu_0) = \left\{(x_1, \ldots, x_n) : \bar{x}_n \leq \mu_0 + z_\alpha/\sqrt{n}\right\} = \left\{(x_1, \ldots, x_n) : \bar{x}_n - z_\alpha/\sqrt{n} \leq \mu_0\right\}.$$

A $100(1-\alpha)\%$ confidence interval is then

$$C(X) = \{\mu : x \in A(\mu)\} = [\bar{x}_n - z_\alpha/\sqrt{n}, \infty).$$

-End of Statistical Theory-