GR5204: Statistical Inference*

Johannes Wiesel
Columbia University

February 23, 2023

Contents

*adapted from Prof. Bodhisettva Sen's and Prof. Thibault Vatter's notes

1 Recap of Probability Theory

Definition 1 (Sample mean)**.** *Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with (unknown) mean $\mu \in \mathbb{R}$ (i.e., $\mathbb{E}(X_1) = \mu$) and variance $\sigma^2 < \infty$. A natural "estimator" of μ is the* ***sample mean*** *(or* ***sample average****) defined as*

$$\bar{X}_n := \frac{1}{n}(X_1 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Lemma 1.1. *$\mathbb{E}(\bar{X}_n) = \mu$ and $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$.*

Proof. Observe that

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

Also,

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

□

Theorem 1.2 (Weak law of large numbers)**.** *Suppose that $X_1, X_2, \ldots, X_n$ are n i.i.d. random variables with finite mean μ. Then for any $\epsilon > 0$, we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right| > \epsilon\right) \to 0 \qquad \text{as } n \to \infty.$$

This says that if we take the sample average of n i.i.d random variables the sample average will be close to the true population average. Figure 1 illustrates the result: The left panel shows the density of the data generating distribution (in this example we took $X_1, \ldots, X_n$ i.i.d. Exp(10)); the middle and right panels show the distribution (histogram obtained from 1000 replicates) of $\bar{X}_n$ for $n = 100$ and $n = 1000$, respectively. We see that as the sample size increases, the distribution of the sample mean concentrates around $\mathbb{E}(X_1) = 1/10$ (i.e., $\bar{X}_n \xrightarrow{\mathbb{P}} 10^{-1}$ as $n \to \infty$).

Definition 2 (Convergence in probability)**.** *In the above, we say that the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to the true (population) mean.*

More generally, we say that a sequence of random variables $\{Z_n\}_{n=1}^{\infty}$ converges to Z **in probability***, and write*

$$Z_n \xrightarrow{\mathbb{P}} Z,$$

if for every $\epsilon > 0$,

$$\mathbb{P}(|Z_n - Z| > \epsilon) \to 0 \qquad \text{as } n \to \infty.$$

This is equivalent to saying that for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|Z_n - Z| \leq \epsilon) = 1.$$

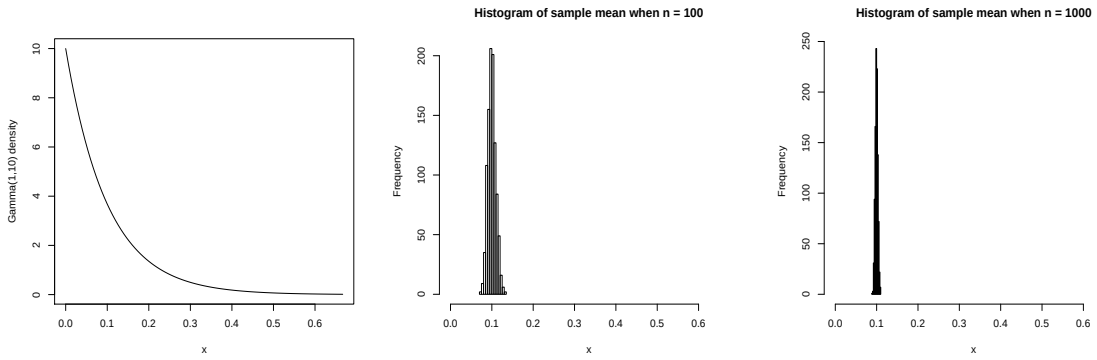

Figure 1: The plots illustrate the convergence (in probability) of the sample mean to the population mean.

Definition 3 (Convergence in distribution)**.** *We say a sequence of random variables* $\{Z_n\}_{i=1}^n$ *with c.d.f's* $F_n(\cdot)$ **converges in distribution** *to* F *if*

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

for all u *such that* F *is continuous*[1] *at* u *(here* F *is itself a c.d.f).*

The second fundamental result in probability theory, after the law of large numbers (LLN), is the Central limit theorem (CLT), stated below. The CLT gives us the approximate (asymptotic) distribution of $\bar{X}_n$

Theorem 1.3 (Central limit theorem)**.** *If* $X_1, X_2, \ldots$ *are i.i.d with mean zero and variance* 1*, then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} N(0,1),$$

where $N(0,1)$ *is the standard normal distribution. More generally, the usual rescaling tell us that if* $X_1, X_2, \ldots$ *are i.i.d with mean* μ *and variance* $\sigma^2 < \infty$*, then*

$$\sqrt{n}(\bar{X}_n - \mu) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} N(0, \sigma^2).$$

[1]Explain why do we need to restrict our attention to continuity points of F. (Hint: think of the following sequence of distributions: $F_n(u) = I(u \geq 1/n)$, where the "indicator" function of a set A is one if $x \in A$ and zero otherwise.) It's worth emphasizing that convergence in distribution — because it only looks at the c.d.f. — is in fact **weaker** than convergence in probability. For example, if p_X is symmetric, then the sequence $X, -X, X, -X, \ldots$ trivially converges in distribution to X, but obviously doesn't converge in probability. Also, if $U \sim \text{Unif}(0,1)$, then the sequence

$$U, 1-U, U, 1-U, \ldots$$

converge in distribution to a uniform distribution. But obviously they do not converge in probability.

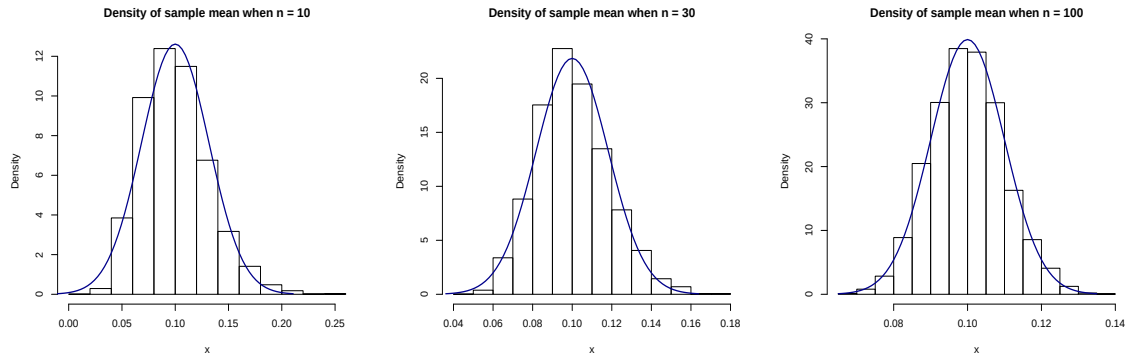

Figure 2: The plots illustrate the convergence (in distribution) of the sample mean to a normal distribution.

The following plots illustrate the CLT: The left, center and right panels of Figure 2 show the (scaled) histograms of $\bar{X}_n$ when $n = 10, 30$ and 100, respectively (as before, in this example we took $X_1, \ldots, X_n$ i.i.d. Exp(10); the histograms are obtained from 5000 independent replicates). We also overplot the normal density with mean 0.1 and variance $10^{-1}/\sqrt{n}$. The remarkable agreement between the two densities illustrates the power of the CLT. Observe that the original distribution of the X_i's, Exp(10), is skewed and highly non-normal, but even for $n = 10$, the distribution of $\bar{X}_{10}$ is quite close to being normal.

Another class of useful results we will use very much in this course go by the name "continuous mapping theorem". Here are two such results.

Theorem 1.4. *If $Z_n \xrightarrow{\mathbb{P}} b$ and if $g(\cdot)$ is a function that is continuous at b, then*

$$g(Z_n) \xrightarrow{\mathbb{P}} g(b).$$

Theorem 1.5. *If $Z_n \xrightarrow{d} Z$ and if $g(\cdot)$ is a function that is continuous, then*

$$g(Z_n) \xrightarrow{d} g(Z).$$

The last result that we need from probability theory—and this may be new to you—is the so-called **delta method**. It allows us to find the asymptotic distribution of a *continuous transformation* of the rescaled sample mean. But let's first state the abstract result.

Theorem 1.6. *Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of random variables and let Z be a random variable with a continuous c.d.f F^*. Let $\theta \in \mathbb{R}$, and let $a_1, a_2, \ldots,$ be a sequence such that $a_n \to \infty$. Suppose that*

$$a_n(Z_n - \theta) \xrightarrow{d} F^*.$$

Let $g(\cdot)$ be a function with a continuous derivative such that $g'(\theta) \neq 0$. Then

$$a_n \frac{g(Z_n) - g(\theta)}{g'(\theta)} \xrightarrow{d} F^*.$$

Proof. We will only give an outline of the proof (think $a_n = n^{1/2}$, if Z_n as the sample mean). As $a_n \to \infty$, Z_n must get close to θ with high probability as $n \to \infty$. As $g(\cdot)$ is continuous, $g(Z_n)$ will be close to $g(\theta)$ with high probability. Let's say $g(\cdot)$ has a Taylor expansion around θ, i.e.,

$$g(Z_n) \approx g(\theta) + g'(\theta)(Z_n - \theta),$$

where we have ignored all terms involving $(Z_n - \theta)^2$ and higher powers. Then if

$$a_n(Z_n - \theta) \xrightarrow{d} Z,$$

for some limit distribution F^* and a sequence of constants $a_n \to \infty$, then

$$a_n \frac{g(Z_n) - g(\theta)}{g'(\theta)} \approx a_n(Z_n - \theta) \xrightarrow{d} F^*.$$

□

In other words, limit distributions are passed through functions in a pretty simple way. We'll be using the delta method a lot. The main application is when we've already proven a CLT for Z_n, that is, when

$$\frac{\sqrt{n}(Z_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

Then, by Theorem 1.6,

$$\sqrt{n}(g(Z_n) - g(\mu)) \xrightarrow{d} N(0, \sigma^2(g'(\mu))^2).$$

Exercise 1: Assume $n^{1/2}Z_n \xrightarrow{d} N(0, 1)$. What is the asymptotic distribution of

1. $g(Z_n) = (Z_n - 1)^2$?
2. What about $g(Z_n) = Z_n^2$? Does anything go wrong when applying the delta method in this case? Can you fix this problem?

Let's illustrate the theory through an example.

Example 1.7. A company sells a certain kind of electronic component. The company is interested in knowing about *how long* a component is likely to last on average. They can collect data on many such components that have been used under typical conditions. They choose to use the family of *exponential* distributions[2] to model the length of time (in years) from when a component is put into service until it fails.

The company believes that, if they knew the failure rate θ, then $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ would be n i.i.d random variables having the exponential distribution with rate θ. Based on these hypotheses, we have the following results:

[2] X has an exponential distribution with rate $\theta > 0$ (in symbols, $X \sim \text{Exp}(\theta)$), if the p.d.f of X is given by

$$f_\theta(x) = \theta e^{-\theta x} \mathbf{1}_{[0,\infty)}(x), \quad \text{for } x \in \mathbb{R}.$$

The mean of X is given by $\mathbb{E}(X) = \theta^{-1}$, and the variance of X is $\text{Var}(X) = \theta^{-2}$.

- by the LLN, the sample mean $\bar{X}_n$ converges in probability to the expectation $1/\theta$, that is,
$$\bar{X}_n \xrightarrow{\mathbb{P}} \frac{1}{\theta};$$
- by the continuous mapping theorem (see Theorem 1.4) $\bar{X}_n^{-1}$ converges in probability to θ, i.e.,
$$\bar{X}_n^{-1} \xrightarrow{\mathbb{P}} \theta;$$
- by the CLT, we know that
$$\sqrt{n}(\bar{X}_n - \theta^{-1}) \xrightarrow{d} N(0, \theta^{-2})$$
where $\mathrm{Var}(X_1) = \theta^{-2}$;
- By the delta method, we can show that
$$\sqrt{n}(\bar{X}_n^{-1} - \theta) \xrightarrow{d} N(0, (\theta^2)^2\theta^{-2}),$$
where we have considered $g(x) = \frac{1}{x}$ with $g'(x) = -\frac{1}{x^2}$ (observe that g is continuous on $(0, \infty)$). Note that the variance of X_1 is $\mathrm{Var}(X_1) = \theta^{-2}$.

2 Statistical Inference: Introduction

2.1 Statistical model

Definition 4 (Statistical model)**.** *A statistical model is*

- *an identification of random variables of interest,*
- *a specification of a joint distribution or a family of possible joint distributions for the observable random variables,*
- *the identification of any parameters of those distributions that are assumed unknown,*
- *(Bayesian approach, if desired) a specification for a (joint) distribution for the unknown parameter(s).*

Definition 5 (Statistical Inference)**.** *Statistical inference is a procedure that produces a probabilistic statement about some or all parts of a statistical model.*

Definition 6 (Parameter space)**.** *The set* Ω *of* **all possible values of a parameter** θ *or of a vector of parameters* $\theta = (\theta_1, \ldots, \theta_k)$ *is called the parameter space.*

Example 2.1.

- The family of *binomial* distributions has parameters n and p.
- The family of *normal* distributions is parameterized by the mean μ and variance σ^2 of each distribution (so $\theta = (\mu, \sigma^2)$ can be considered a pair of parameters, and $\Omega = \mathbb{R} \times \mathbb{R}^+$).
- The family of *exponential* distributions is parameterized by the rate parameter θ (the failure rate must be positive: Ω will be the set of all positive numbers).

Note: The parameter space Ω must contain all possible values of the parameters in a given problem.

Example 2.2. Suppose that n patients are going to be given a treatment for a condition and that we will observe for each patient whether or not they recover from the condition.

For each patient $i = 1, 2, \ldots$, let $X_i = 1$ if patient i recovers, and let $X_i = 0$ if not. As a collection of possible distributions for $X_1, X_2, \ldots$, we could choose to say that the X_i's are i.i.d having the Bernoulli distribution with parameter p, for $0 \leq p \leq 1$.

In this case, the parameter p is known to lie in the closed interval $[0, 1]$, and this interval could be taken as the parameter space. Notice also that by the LLN, p is the limit as $n \to \infty$ of the proportion of the first n patients who recover.

Definition 7 (Statistic)**.** *Suppose that the observable random variables of interest are $X_1, \ldots, X_n$. Let φ be a real-valued function of n real variables. Then the random variable $T = \varphi(X_1, \ldots, X_n)$ is called a* **statistic***.*

Example 2.3.

- the sample mean $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$;
- the maximum $X_{(n)}$ of the values $X_1, \ldots, X_n$;
- the sample variance $s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ of the values $X_1, \ldots, X_n$.

Definition 8 (Estimator/Estimate)**.** *Let $X_1, \ldots, X_n$ be observable data whose joint distribution is indexed by a parameter θ taking values in a subset Ω of the real line. An* **estimator** *$\widehat{\theta}_n$ of the parameter θ is a real-valued function $\widehat{\theta}_n = \varphi(X_1, \ldots, X_n)$. If $\{X_1 = x_1, \ldots, X_n = x_n\}$ is observed, then $\varphi(x_1, \ldots, x_n)$ is called the* **estimate** *of θ.*

More generally, let $X_1, \ldots, X_n$ be observable data whose joint distribution is indexed by a parameter θ taking values in a subset Ω of d-dimensional space, i.e., $\Omega \subset \mathbb{R}^d$. Let $h : \Omega \to \mathbb{R}^d$, be a function from Ω into d-dimensional space. Define $\psi = h(\theta)$. An ***estimator*** *of ψ is a function $g(X_1, \ldots, X_n)$ that takes values in d-dimensional space. If $\{X_1 = x_1, \ldots, X_n = x_n\}$ are observed, then $g(x_1, \ldots, x_n)$ is called the* ***estimate*** *of ψ.*

When h in Definition 8 is the identity function $h(\theta) = \theta$, then $\psi = \theta$ and we are estimating the original parameter θ. When $g(\theta)$ is one coordinate of θ, then the ψ that we are estimating is just that one coordinate.

Notice that an estimator need not be a "good" one. In fact, any transformation of the observations $X_1, \ldots, X_n$ is an estimator by Definition 8. For example, if θ is the unknown mean of distribution,

$$\widehat{\theta}_n = X_1 + \sum_{i=4}^{n} X_i - e^{X_2 + X_3}$$

is formally an estimator for θ (but probably really bad one). So we need some sort of criteria to evaluate how good an estimator is. Here is one criterion that is often used.

Definition 9 (Consistent estimator)**.** *A sequence of estimators $\widehat{\theta}_n$ is said to be* ***consistent*** *for the unknown parameter θ if $\widehat{\theta}_n$ converges in probability to θ, that is, if for every $\epsilon > 0$,*

$$\mathbb{P}(|\widehat{\theta}_n - \theta| > \epsilon) \to 0, \qquad as \;\; n \to \infty.$$

In the following, we shall discuss three types of estimators:

- **Method of moments** estimators,
- **Maximum likelihood** estimators, and
- **Bayes** estimators.

2.2 Method of moments estimators

The *method of moments* (MOM) is an intuitive method for estimating parameters when other, more attractive, methods may be too difficult (to implement/compute).

Definition 10 (Method of moments estimator)**.** *Assume that $X_1, \ldots, X_n$ are observations from a distribution that is indexed by a k-dimensional parameter θ and that has at least k finite moments. For $j = 1, \ldots, k$, let*

$$\mu_j(\theta) := \mathbb{E}_\theta(X_1^j)$$

be the ***jth moment*** *of $X_1, \ldots, X_n$. Suppose that the function*

$$\mu(\theta) = (\mu_1(\theta), \ldots, \mu_k(\theta))$$

is a one-to-one function of θ. Let $M(\mu_1, \ldots, \mu_k)$ denote the inverse function, that is, for all θ,

$$\theta = M(\mu_1, \ldots, \mu_k).$$

Define the ***jth sample moment*** *as*

$$\hat{\mu}_j := \frac{1}{n} \sum_{i=1}^{n} X_i^j \quad for \;\; j = 1, \ldots, k.$$

The method of moments estimator of θ is $M(\hat{\mu}_1, \ldots, \hat{\mu}_k)$.

Equivalently, the method of moments estimators can be obtained by setting up k equations

$$\hat{\mu}_j = \mu_j(\theta), \qquad \text{for} \;\; j = 1, \ldots, k,$$

and then solving for θ.

Theorem 2.4 (Consistency of the MOM estimator)**.** *Suppose that $X_1, X_2, \ldots$ are i.i.d with a distribution indexed by a k-dimensional parameter vector θ. If the first k moments of that distribution exist and are finite for all θ and the inverse function M in Definition (10) is continuous, then the sequence of MOM estimators based on $X_1, X_2, \ldots$ is consistent for θ.*

Proof. By the LLN, the sample moments converge in probability to the population moments $\mu_1(\theta), \ldots, \mu_k(\theta)$. A generalization of the continuous mapping theorem to functions of k variables implies that $M(\cdot)$ evaluated at the sample moments converges in probability to θ, i.e., the MOM estimator converges in probability to θ. □

Example 2.5. Let $X_1, X_2, \ldots, X_n$ be from a $N(\mu, \sigma^2)$ distribution. Thus $\theta = (\mu, \sigma^2)$. What is the MOM estimator of θ?

Solution: Because $\mu_1 = \mathbb{E}(X_1) = \mu$ and $\mu_2 = \mathbb{E}(X_1^2) = \sigma^2 + \mu^2$, it is easy to express the unknown parameters μ and σ^2 in terms of μ_1 and μ_2:

$$\mu = \mu_1, \qquad \sigma^2 = \mu_2 - \mu^2 = \mu_2 - \mu_1^2.$$

To get MOM estimates of μ and σ^2 we are going to plug in the sample moments. Thus

$$\hat{\mu} = \hat{\mu}_1 = \bar{X}\,,$$

and

$$\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2 \; - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

where we have used the fact that

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

Example 2.6. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d Gamma(α, β), $\alpha, \beta > 0$. Thus, $\theta = (\alpha, \beta) \in \Omega := \mathbb{R}_+ \times \mathbb{R}_+$. The first two moments of this distribution are:

$$\mu_1(\theta) = \frac{\alpha}{\beta}, \qquad \mu_2(\theta) = \frac{\alpha(\alpha+1)}{\beta^2},$$

which implies that

$$\alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2}, \qquad \beta = \frac{\mu_1}{\mu_2 - \mu_1^2}.$$

The MOM says that we replace the right-hand sides of these equations by the *sample moments*. In this case, we get

$$\hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}, \qquad \hat{\beta} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2}.$$

Remark: MOM can thus be thought of as "plug-in" estimates; to get an estimate $\hat{\theta}$ of $\theta = M(\mu_1, \mu_2, \ldots, \mu_k)$, we plug-in estimates of the μ_i's, which are the $\hat{\mu}_i$'s, to get $\hat{\theta}$.

In general, we might be interested in estimating $\Psi(\theta)$ where $\Psi(\theta)$ is some (known) function of θ; in such a case, the MOM estimate of $\Psi(\theta)$ is $\Psi(\hat{\theta})$ where $\hat{\theta}$ is the MOM estimate of θ.

Example 2.7. Let $X_1, X_2, \ldots, X_n$ be the indicators of n Bernoulli trials with success probability θ. We are going to find a MOM estimator of θ.

Solution: Note that θ is the probability of success and satisifes,

$$\theta = \mathbb{E}(X_1) \ , \ \theta = \mathbb{E}(X_1^2) \, .$$

Thus we can get MOMs of θ based on both the first and the second moments. Thus,

$$\hat{\theta} = \bar{X}$$

or

$$\hat{\theta} = \frac{1}{n} \sum_{j=1}^{n} X_j^2 = \frac{1}{n} \sum_{j=1}^{n} X_j = \bar{X}.$$

Here, the MOM estimate based on the second moment μ_2 coincides with the MOM estimate based on μ_1 However, this is not necessarily the case; the MOM estimate of a certain parameter *may not be unique* as illustrated by the following example.

Example 2.8. Let $X_1, X_2, \ldots, X_n$ be i.i.d. Poisson(λ), $\lambda > 0$. Find the MOM estimator of $\theta := \lambda + \lambda^2$.

Solution: On the one hand, because $\mu_1 = \mathbb{E}(X_1) = \lambda$, the MOM estimate of θ based on the first moment is

$$\hat{\theta} = \hat{\mu}_1 + \hat{\mu}_1^2 = \bar{X} + \bar{X}^2.$$

On the other hand, $\mu_2 = \mathbb{E}(X_1^2) = \mathrm{Var}(X_1) + \mathbb{E}(X_1)^2 = \lambda + \lambda^2 = \theta$, so the MOM estimate of θ based on the second moment is

$$\hat{\theta} = \frac{1}{n} \sum_{j=1}^{n} X_j^2.$$

However, these two estimates are not necessarily equal; in other words, it is not necessarily the case that $\bar{X}^2 + \bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j^2$.

This illustrates one of the disadvantages of MOM estimates—they may not be uniquely defined.

Example 2.9. Consider n systems with failure times $X_1, X_2, \ldots, X_n$ assumed to be i.i.d Exp(λ), $\lambda > 0$. Find the MOM estimators of λ.

Solution: It is not difficult to show that

$$\mathbb{E}(X_1) = \frac{1}{\lambda}\,,\ \mathbb{E}(X_1^2) = \frac{2}{\lambda^2}.$$

Therefore

$$\lambda = \frac{1}{\mu_1} = \sqrt{\frac{2}{\mu_2}}\,.$$

The above equations lead to two different MOM estimators for λ; the estimate based on the first moment is

$$\hat{\lambda} = \frac{1}{\hat{\mu}_1}\,,$$

and the estimate based on the second moment is

$$\hat{\lambda} = \sqrt{\frac{2}{\hat{\mu}_2}}\,.$$

Once again, note the non-uniqueness of the estimates.

We finish up this section by some key observations about method of moments estimates.

(i) The MOM principle generally leads to procedures that are easy to compute and which are therefore valuable as preliminary estimates.

(ii) For large sample sizes, these estimates are likely to be close to the value being estimated (consistency).

(iii) The prime disadvantage is that they do not provide a unique estimate and this has been illustrated before with examples.

3 Method of Maximum Likelihood

As before, we have i.i.d observations $X_1, X_2, \ldots, X_n$ with common probability density (or mass function) $f(x, \theta)$, where $\theta \in \Omega \subseteq \mathbb{R}^k$ is a Euclidean parameter indexing the class of distributions being considered.

The goal is to estimate θ or some $\Psi(\theta)$ where Ψ is some known function of θ.

Definition 11 (Likelihood function)**.** *The likelihood function for the sample* $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ *is*

$$L_n(\theta) \equiv L_n(\theta, \boldsymbol{X}_n) := \prod_{i=1}^{n} f(X_i, \theta).$$

This is simply the joint density (or mass function) but we now think of this as a function of θ *for a fixed* $\boldsymbol{X}_n$*; namely the* $\boldsymbol{X}_n$ *that is realized.*

Intuition: Suppose for the moment that X_i's are discrete, so that f is actually a p.m.f. Then $L_n(\theta)$ is exactly the probability that the observed data is realized or "happens".

We now seek to obtain that $\theta \in \Omega$ for which $L_n(\theta)$ is maximized. Call this $\hat{\theta}_n$ (assume that it exists). Thus $\hat{\theta}_n$ is that value of the parameter that maximizes the likelihood function, or in other words, makes the observed data most likely.

It makes sense to pick $\hat{\theta}_n$ as a guess for θ.

When the X_i's are continuous and $f(x, \theta)$ is in fact a density we do the same thing – maximize the likelihood function as before and prescribe the maximizer as an estimate of θ.

For obvious reasons, $\hat{\theta}_n$ is called an **maximum likelihood estimate** (MLE).

Remarks:

- Note that $\hat{\theta}_n$ is itself a deterministic function of $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ and is therefore a random variable. Of course there is nothing that guarantees that $\hat{\theta}_n$ is unique, even if it exists.
- Sometimes, in the case of multiple maximizers, we choose one which is more desirable according to some "sensible" criterion.

Example 3.1. Suppose that $X_1, \ldots, X_n$ are i.i.d Poisson(θ), $\theta > 0$. Find the MLE of θ.

Solution: In this case, it is easy to see that

$$L_n(\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\, \theta^{X_i}}{X_i\,!} = C(\boldsymbol{X}_n)\, e^{-n\,\theta}\, \theta^{\sum_{i=1}^{n} X_i}.$$

To maximize this expression, we set

$$\frac{\partial}{\partial \theta} \log L_n(\theta) = 0\,.$$

This yields that

$$\frac{\partial}{\partial \theta}\left[-n\,\theta + \left(\sum_{i=1}^{n} X_i\right) \log\,\theta\right] = 0;$$

i.e.,

$$-n + \frac{\sum_{i=1}^{n} X_i}{\theta} = 0\,,$$

showing that

$$\hat{\theta}_n = \bar{X}\,.$$

It can be checked (by computing the second derivative at $\hat{\theta}_n$) that the stationary point indeed gives (a unique) maximum (or by noting that the log-likelihood is a (strictly) concave function).

Exercise 2: Let $X_1, X_2, \ldots, X_n$ be i.i.d Ber(θ) where $0 \leq \theta \leq 1$. What is the MLE of θ?

Example 3.2. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d Uniform($[0, \theta]$) random variables, where $\theta > 0$. We want to obtain the MLE of θ.

Solution: The likelihood function is given by,

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} \frac{1}{\theta} I_{[0,\theta]}(X_i) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} I_{[X_i,\infty)}(\theta) \\ &= \frac{1}{\theta^n} I_{[\max_{i=1,\ldots,n} X_i,\infty)}(\theta). \end{aligned}$$

It is then clear that $L_n(\theta)$ is constant and equals $1/\theta^n$ for $\theta \geq \max\limits_{i=1,\ldots,n} X_i$ and is 0 otherwise. By plotting the graph of this function, you can see that

$$\hat{\theta}_n = \max_{i=1,\ldots,n} X_i.$$

Here, differentiation will not help you to get the MLE because the likelihood function is not differentiable at the point where it hits the maximum.

Example 3.3. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d $N(\mu, \sigma^2)$. We want to find the MLEs of the mean μ and the variance σ^2.

Solution: We write down the likelihood function first. This is,

$$L_n(\mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right).$$

It is easy to see that,

$$\begin{aligned}\log L_n(\mu, \sigma^2) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 + \text{constant} \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 - \frac{n}{2\sigma^2}(\bar{X}_n - \mu)^2.\end{aligned}$$

To maximize the above expression w.r.t μ and σ^2 we proceed as follows. For any (μ, σ^2) we have,

$$\log L_n(\mu, \sigma^2) \leq \log L_n(\bar{X}_n, \sigma^2),$$

showing that we can choose $\hat{\mu}_{MLE} = \bar{X}_n$.

It then remains to maximize $\log L_n(\bar{X}_n, \sigma^2)$ with respect to σ^2 to find $\hat{\sigma}^2_{MLE}$.

Now,

$$\log L_n(\bar{X}_n, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Differentiating the left-side w.r.t σ^2 gives,

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} n\,\hat{\sigma}^2 = 0\,,$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. The above equation leads to,

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

The fact that this actually gives a global maximizer follows from the fact that the second derivative at $\hat{\sigma}^2$ is negative.

Note that, once again, the MOM estimates coincide with the MLEs.

Exercise 3: We now tweak the above situation a bit. Suppose now that we restrict the parameter space, so that μ has to be non-negative, i.e., $\mu \geq 0$.

Thus we seek to maximize $\log L_n(\mu, \sigma^2)$ but subject to the constraint that $\mu \geq 0$ and $\sigma^2 > 0$. Find the MLEs in this scenario.

Example 3.4 (MLEs might not be unique)**.** Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[\theta, \theta+1]$, where $\theta \in \mathbb{R}$ is unknown. Show that the MLE of θ is not unique.

Solution: The likelihood has the form

$$L_n(\theta) = \prod_{i=1}^{n} I_{[\theta,\theta+1]}(X_i).$$

The condition that $\theta \leq X_i$, for all $i = 1, \ldots, n$, is equivalent to the condition that $\theta \leq \min\{X_1, \ldots, X_n\} = X_{(1)}$. Similarly, the condition that $X_i \leq \theta + 1$, for all $i = 1, \ldots, n$, is equivalent to the condition that $\theta \geq \max\{X_1, \ldots, X_n\} - 1 = X_{(n)} - 1$. Thus the likelihood can be written as

$$L_n(\theta) = I_{[X_{(n)}-1, X_{(1)}]}(\theta).$$

Hence it is possible to select as an MLE any value of θ in the interval $[X_{(n)} - 1, X_{(1)}]$, and thus the MLE is not unique.

Example 3.5 (MLEs might not exist)**.** Consider a random variable X that can come with equal probability either from a $N(0, 1)$ or from $N(\mu, \sigma^2)$, where both μ and σ are unknown.

Thus, the p.d.f. $f(\cdot, \mu, \sigma^2)$ of X is given by

$$f(x, \mu, \sigma^2) = \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} e^{-x^2/2} + \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}\right].$$

Suppose now that $X_1, \ldots, X_n$ form an i.i.d sample from this distribution. As usual, the likelihood function

$$L_n(\mu, \sigma^2) = \prod_{i=1}^{n} f(X_i, \mu, \sigma^2).$$

We want to find the MLE of $\theta = (\mu, \sigma^2)$.

Let X_k denote one of the observed values. Note that

$$\max_{\mu\in\mathbb{R}, \sigma^2>0} L_n(\mu, \sigma^2) \geq L_n(X_k, \sigma^2) \geq \frac{1}{2^n}\left[\frac{1}{\sqrt{2\pi}\sigma}\right] \prod_{i\neq k} \frac{1}{\sqrt{2\pi}} e^{-X_i^2/2}.$$

Thus, if we let $\mu = X_k$ and let $\sigma^2 \to 0$ then the factor $f(X_k, \mu, \sigma^2)$ will grow large without bound, while each factor $f(X_i, \mu, \sigma^2)$, for $i \neq k$, will approach the value

$$\frac{1}{2\sqrt{2\pi}} e^{-X_i^2/2}.$$

Hence, when $\mu = X_k$ and $\sigma^2 \to 0$, we find that $L_n(\mu, \sigma^2) \to \infty$.

Note that 0 is not a permissible estimate of σ^2, because we know in advance that $\sigma > 0$. Since the likelihood function can be made arbitrarily large by choosing $\mu = X_k$ and choosing σ^2 arbitrarily close to 0, it follows that the MLE *does not exist*.

3.1 Properties of MLEs

Theorem 3.6 (Invariance property of MLEs)**.** *If $\widehat{\theta}_n$ is the MLE of θ and if Ψ is any function, then $\Psi\big(\widehat{\theta}_n\big)$ is the MLE of $\Psi(\theta)$.*

See Theorem 7.6.2 and Example 7.6.3 in the text book.

Thus if $X_1, \ldots, X_n$ be i.i.d $N(\mu, \sigma^2)$, then the MLE of σ is $\sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}$.

Consider an estimation problem in which a random sample is to be taken from a distribution involving a parameter θ. Then, under certain conditions, which are typically satisfied in practical problems, the sequence of MLEs is *consistent*, i.e.,

$$\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta, \qquad \text{as } n \to \infty.$$

3.2 Computational methods for approximating MLEs

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d from a Gamma distribution for which the p.d.f is as follows:

$$f(x, \alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \qquad \text{for } x > 0.$$

The likelihood function is

$$L_n(\alpha) = \frac{1}{\Gamma(\alpha)^n} \left(\prod_{i=1}^n X_i \right)^{\alpha-1} e^{-\sum_{i=1}^n X_i},$$

and thus the log-likelihood is

$$\ell_n(\alpha) \equiv \log L_n(\alpha) = -n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^n \log(X_i) - \sum_{i=1}^n X_i,$$

The MLE of α will be the value of α that satisfies the equation

$$\begin{aligned} \frac{\partial}{\partial \alpha} \ell_n(\alpha) &= -n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^n \log(X_i) = 0 \\ \text{i.e.,} \quad \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} &= \frac{1}{n} \sum_{i=1}^n \log(X_i). \end{aligned}$$

Solving this equation for α cannot be done analytically and requires computational methods to obtain an approximate solutions.

Read pages 428–430 and pages 434–439 of the text-book. I will cover this later, if time permits.

4 Principles of estimation

Setup: Our data $X_1, X_2, \ldots X_n$ are i.i.d observations from the distribution P_θ where $\theta \in \Omega$, the parameter space (Ω is assumed to be the k-dimensional Euclidean space). We assume identifiability of the parameter, i.e. $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$.

Estimation problem: Consider now, the problem of estimating $g(\theta)$ where g is some function of θ.

In many cases $g(\theta) = \theta$ itself.

Generally $g(\theta)$ will describe some important aspect of the distribution P_θ.

Our estimator of $g(\theta)$ will be some function of our observed data $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$.

In general there will be several different estimators of $g(\theta)$ which may all seem reasonable from different perspectives — the question then becomes one of finding the most optimal one.

This requires an objective **measure of performance** of an estimator.

If T_n estimates $g(\theta)$ a criterion that naturally suggests itself is the distance of T_n from $g(\theta)$. Good estimators are those for which $|T_n - g(\theta)|$ is generally small.

Since T_n is a random variable no deterministic statement can be made about the *absolute deviation*; however what we can expect of a good estimator is a high chance of remaining close to $g(\theta)$.

Also as n, the sample size, increases we get hold of more information and hence expect to be able to do a better job of estimating $g(\theta)$.

These notions when coupled together give rise to the **consistency** requirement for a sequence of estimators T_n; as n increases, T_n ought to converge in probability to $g(\theta)$ (under the probability distribution P_θ). In other words, for any $\epsilon > 0$,

$$\mathbb{P}_\theta \left(|T_n - g(\theta)| > \epsilon \right) \to 0.$$

The above is clearly a *large sample property*; what it says is that with probability increasing to 1 (as the sample size grows), T_n estimates $g(\theta)$ to any pre-determined level of accuracy.

However, the consistency condition alone, does not tell us anything about how well we are performing for any particular sample size, or the rate at which the above probability is going to 0.

4.1 Mean squared error

Question: For a fixed sample size n, how do we measure the performance of an

estimator T_n?

A way out of this difficulty is to obtain an average measure of the error, or in other words, average out $|T_n - g(\theta)|$ over all possible realizations of T_n.

The resulting quantity is then still a function of θ but no longer random. It is called the **mean absolute error** and can be written compactly (using acronym) as:

$$\text{MAD} := \mathbb{E}_\theta \left[|T_n - g(\theta)| \right].$$

However, it is more common to avoid absolute deviations and work with the square of the deviation, integrated out as before over the distribution of T_n. This is called the **mean squared error** (MSE) and is defined as

$$\text{MSE}(T_n, g(\theta)) := \mathbb{E}_\theta \left[(T_n - g(\theta))^2 \right]. \tag{1}$$

Of course, this is meaningful, only if the above quantity is finite for all θ. Good estimators are those for which the MSE is generally not too high, whatever be the value of θ.

There is a standard decomposition of the MSE that helps us understand its components.

Theorem 4.1. *For any estimator T_n of $g(\theta)$, we have*

$$\text{MSE}(T_n, g(\theta)) = \text{Var}_\theta(T_n) + b(T_n, g(\theta))^2,$$

where $b(T_n, g(\theta)) = \mathbb{E}_\theta(T_n) - g(\theta)$ is the **bias** *of T_n as an estimator of $g(\theta)$.*

Proof. We have,

$$\begin{aligned}
\text{MSE}(T_n, g(\theta)) &= \mathbb{E}_\theta \left[(T_n - g(\theta))^2 \right] \\
&= \mathbb{E}_\theta \left[(T_n - \mathbb{E}_\theta(T_n) + \mathbb{E}_\theta(T_n) - g(\theta))^2 \right] \\
&= \mathbb{E}_\theta \left[(T_n - \mathbb{E}_\theta(T_n))^2 \right] + (\mathbb{E}_\theta(T_n) - g(\theta))^2 \\
&\qquad + 2\, \mathbb{E}_\theta[(T_n - \mathbb{E}_\theta(T_n))(\mathbb{E}_\theta(T_n) - g(\theta))] \\
&= \text{Var}_\theta(T_n) + b(T_n, g(\theta))^2,
\end{aligned}$$

where

$$b(T_n, g(\theta)) := \mathbb{E}_\theta(T_n) - g(\theta)$$

is the **bias** of T_n as an estimator of $g(\theta)$.

The cross product term in the above display vanishes since $\mathbb{E}_\theta(T_n) - g(\theta)$ is a constant and $\mathbb{E}_\theta(T_n - \mathbb{E}_\theta(T_n)) = 0$. □

The bias measures, on an average, by how much T_n overestimates or underestimates $g(\theta)$. If we think of the expectation $\mathbb{E}_\theta(T_n)$ as the center of the distribution of T_n, then the bias measures by how much the *center deviates from the target*.

The variance of T_n, of course, measures how closely T_n is clustered around its center. Ideally one would like to minimize both simultaneously, but unfortunately this is rarely possible.

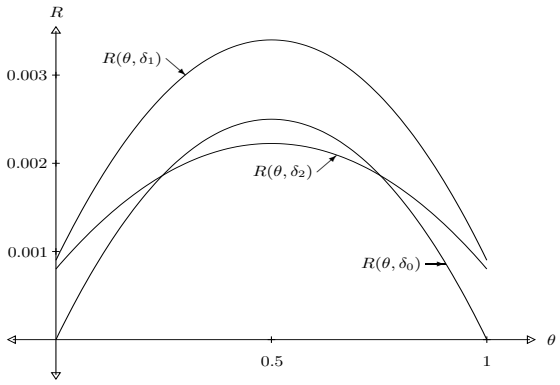

Figure 3: The plot shows the mean squared error for three estimators δ_1, δ_2 and δ_2. Here $R(\theta, \delta_i) = \mathbb{E}_\theta[(\delta_i(X) - \theta)^2]$ where $i = 0, 1, 2$.

4.2 Comparing estimators

Two estimators T_n and S_n can be compared on the basis of their MSEs. Under parameter value θ, T_n dominates S_n as an estimator if

$$\text{MSE}(T_n, \theta) \leq \text{MSE}(S_n, \theta) \qquad \text{for all } \theta \in \Omega.$$

In this situation we say that S_n is *inadmissible* in the presence of T_n.

The use of the term "inadmissible" hardly needs explanation. If, for all possible values of the parameter, we incur less error using T_n instead of S_n as an estimate of $g(\theta)$, then clearly there is no point in considering S_n as an estimator at all.

Continuing along this line of thought, is there an estimate that improves all others? In other words, is there an estimator that makes every other estimator inadmissible? The answer is **no**, except in certain pathological situations.

Example 4.2. Suppose that $X \sim \text{Binomial}(100, \theta)$, where $\theta \in [0, 1]$. The goal is to estimate the unknown parameter θ. A natural estimator of θ in this problem is $\delta_0(X) = X/100$ (which is also the MLE and the method of moments estimator). Then

$$R(\theta, \delta_0) := \text{MSE}(\delta_0(X), \theta) = \frac{\theta(1-\theta)}{100}, \quad \text{for } \theta \in [0, 1].$$

The MSE of $\delta_0(X)$ as a function of θ is given in Figure 3.

We can also consider two other estimators in this problem: $\delta_1(X) = (X+3)/100$ and $\delta_2(X) = (X+3)/106$. Figure 3 shows the MSEs of δ_1 and δ_2, which can be shown to be (show this):

$$R(\theta, \delta_1) := \text{MSE}(\delta_1(X), \theta) = \frac{9 + 100\theta(1-\theta)}{100^2}, \quad \text{for } \theta \in [0, 1],$$

and

$$R(\theta, \delta_2) := \text{MSE}(\delta_2(X), \theta) = \frac{(9 - 8\theta)(1 + 8\theta)}{106^2}, \quad \text{for } \theta \in [0, 1].$$

Looking at the plot, δ_0 and δ_2 are both better than δ_1, but the comparison between δ_0 and δ_2 is ambiguous. When θ is near $1/2$, δ_2 is the preferable estimator, but if θ is near 0 or 1, δ_0 is preferable. If θ were known, we could choose between δ_0 and δ_2. However, if θ were known, there would be no need to estimate its value.

As we have noted before, it is generally not possible to find a universally best estimator. One way to try to construct optimal estimators is to restrict oneself to a subclass of estimators and try to find the best possible estimator in this subclass. One arrives at subclasses of estimators by constraining them to meet some desirable requirements. One such requirement is that of *unbiasedness*. Below, we provide a formal definition.

4.3 Unbiased estimators

An estimator T_n of $g(\theta)$ is said to be *unbiased* if $\mathbb{E}_\theta(T_n) = g(\theta)$ for all possible values of θ; i.e.,

$$b(T_n, g(\theta)) = 0 \qquad \text{for all } \theta \in \Omega.$$

Thus, unbiased estimators, on an average, hit the target, for all parameter values. This seems to be a reasonable constraint to impose on an estimator and indeed produces meaningful estimates in a variety of situations.

Lemma 4.3. *Let $X_1, \ldots, X_n$ be iid with mean μ and finite variance σ^2. Then*

$$\mathbb{E}[\bar{X}_n] = \mu, \qquad \mathbb{E}[s_n^2] = \sigma^2.$$

In other words, sample mean and sample variance are unbiased estimators for μ and σ^2, respectively.

Proof. Exercise! In particular, make sure you see in the calculation why we have to define $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ (with division by $n-1$) to obtain an unbiased estimator. □

Note that for an unbiased estimator T_n, the MSE under θ is simply the variance of T_n under θ.

In a large class of models, it is possible to find an unbiased estimator of $g(\theta)$ that has the smallest possible variance among all possible unbiased estimators. Such an estimate is called an **minimum variance unbiased estimator** (MVUE). Here is a formal definition.

MVUE: We call S_n an MVUE of $g(\theta)$ if

$$(i)\ \ \mathbb{E}_\theta(S_n) = g(\theta) \qquad \text{for all } \theta \in \Omega$$

and (ii) if T_n is an unbiased estimate of $g(\theta)$, then $\mathrm{Var}_\theta(S_n) \leq \mathrm{Var}_\theta(T_n)$.

Here are a few examples to illustrate some of the various concepts discussed above.

(a) Consider $X_1, \ldots, X_n$ i.i.d $N(\mu, \sigma^2)$.

A natural unbiased estimator of $g_1(\theta) = \mu$ is $\bar{X}_n$, the sample mean. It is also consistent for μ by the WLLN. It can be shown that this is also the MVUE of μ.

In other words, *any* other unbiased estimate of μ will have a larger variance than $\bar{X}_n$. Recall that the variance of $\bar{X}_n$ is simply σ^2/n.

Consider now, the estimation of σ^2. Two estimates of this that we have considered in the past are

$$(i)\ \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{and} \qquad (ii)\ s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\,.$$

Out of these $\hat{\sigma}_n^2$ is not unbiased for σ^2 but s_n^2 is. In fact s_n^2 is the MVUE of σ^2.

(b) Let $X_1, X_2, \ldots, X_n$ be i.i.d from some underlying density function or mass function $f(x, \theta)$. Let $g(\theta) = \mathbb{E}_\theta(X_1)$.

Then the sample mean $\bar{X}_n$ is always an unbiased estimate of $g(\theta)$ (see Lemma 4.3). Whether it is MVUE or not depends on the underlying structure of the model.

(c) Suppose that $X_1, X_2, \ldots, X_n$ be i.i.d Ber(θ). It can be shown that $\bar{X}_n$ is the MVUE of θ.

Now define $g(\theta) = \theta/(1-\theta)$. This is a quantity of interest because it is precisely the odds in favor of Heads. It can be shown that there is *no unbiased estimator* of $g(\theta)$ in this model (**Why?**).

However an intuitively appealing estimate of $g(\theta)$ is $T_n \equiv \bar{X}_n/(1 - \bar{X}_n)$. It is *not unbiased* for $g(\theta)$; however it does converge in probability to $g(\theta)$.

This example illustrates an important point — unbiased estimators may not always exist. Hence imposing unbiasedness as a constraint may not be meaningful in all situations.

(d) Unbiased estimators are not always better than biased estimators.

Remember, it is the MSE that gauges the performance of the estimator and a biased estimator may actually outperform an unbiased one owing to a significantly smaller variance.

Example 4.4. Consider $X_1, X_2, \ldots, X_n$ i.i.d Uniform$([0, \theta])$ with $\theta > 0$. Here $\Omega = (0, \infty)$. The MLE for θ is the maximum of the X_i's, which we denote by $X_{(n)}$. Another estimate of θ is obtained by observing that $\bar{X}_n$ is an unbiased estimate of $\theta/2$, the common mean of the X_i's; hence $2\,\bar{X}_n$ is an unbiased estimate of θ. Show that $X_{(n)}$ in the sense of MSE outperforms $2\bar{X}_n$ by an order of magnitude. The best unbiased estimator (MVUE) of θ is $(1 + n^{-1})X_{(n)}$.

4.4 Sufficient Statistics

In some problems, there may not be any MLE, or there may be more than one. Even when an MLE is unique, it may not be a suitable estimator (as in the Unif$(0, \theta)$ example, where the MLE always underestimates the value of θ).

In such problems, the search for a good estimator must be extended beyond the methods that have been introduced thus far.

In this section, we shall define the concept of a **sufficient statistic**, which can be used to simplify the search for a good estimator in many problems.

Suppose that in a specific estimation problem, two statisticians A and B must estimate the value of the parameter θ.

Statistician A can observe the values of the observations $X_1, X_2, \ldots, X_n$ in a random sample, and statistician B cannot observe the individual values of $X_1, X_2, \ldots, X_n$ but can learn the value of a certain statistic $T = \varphi(X_1, \ldots, X_n)$.

In this case, statistician A can choose any function of the observations $X_1, X_2, \ldots, X_n$ as an estimator of θ (including a function of T). But statistician B can use only a function of T. Hence, it follows that A will generally be able to find a better estimator than will B.

In some problems, however, B will be able to do just as well as A. In such a problem, *the single function $T = \varphi(X_1, \ldots, X_n)$ will in some sense summarize all the information contained in the random sample about θ*, and knowledge of the individual values of $X_1, \ldots, X_n$ will be irrelevant in the search for a good estimator of θ. A statistic T having this property is called a **sufficient statistic**.

A statistic is **sufficient** with respect to a statistical model P_θ and its associated unknown parameter θ if it provides "all" the information on θ; e.g., if "no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter". This intuition will be made rigorous at the end of this subsection.

Definition 12 (Sufficient statistic)**.** *Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution indexed by a parameter $\theta \in \Omega$. A statistic T is called a **sufficient statistic** for the parameter θ if the following holds: For all possible values of t and no matter what the true value of $\theta \in \Omega$ is, the conditional joint distribution of $X_1, X_2, \ldots, X_n$ given that $T = t$ does not depend on θ.*

So, if T is sufficient, and one observed only T instead of $(X_1, \ldots, X_n)$, one could, at least in principle, simulate random variables $(X_1', \ldots, X_n')$ with the same joint distribution.

In this sense, T is sufficient for obtaining as much information about θ as one could get from $(X_1, \ldots, X_n)$.

Example 4.5. Suppose that $X_1, \ldots, X_n$ are i.i.d Poisson(θ), where $\theta > 0$. Show that $T = \sum_{i=1}^n X_i$ is sufficient. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$.

Note that

$$\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x} | T(\boldsymbol{X}) = t) = \frac{\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x}, T(\boldsymbol{X}) = t)}{\mathbb{P}_\theta(T(\boldsymbol{X}) = t)}.$$

But,

$$\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x}, T(\boldsymbol{X}) = t) = \begin{cases} 0 & T(\boldsymbol{x}) \neq t \\ \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) & T(\boldsymbol{x}) = t. \end{cases}$$

Because

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \frac{e^{-n\theta}\theta^{T(\boldsymbol{x})}}{\prod_{i=1}^n x_i!}$$

and

$$\mathbb{P}_\theta(T(\boldsymbol{X}) = t) = \frac{e^{-n\theta}(n\theta)^t}{t!},$$

we have that

$$\frac{\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}_\theta(T(\boldsymbol{X}) = t} = \frac{t!}{\prod_{i=1}^n x_i! n^t},$$

which does not depend on θ. So $T = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.

Other sufficient statistics are: $T = 3.7\sum_{i=1}^n X_i$, $T = (\sum_{i=1}^n X_i, X_4)$, and $T = (X_1, \ldots, X_n)$.

We shall now present a simple method for finding a sufficient statistic that can be applied in many problems.

Theorem 4.6 (Factorization criterion)**.** *Let $X_1, X_2, \ldots, X_n$ be iid from either a continuous distribution or a discrete distribution for which the p.d.f or the p.m.f is $f(x, \theta)$, where the value of θ is unknown and belongs to a given parameter space Ω. A statistic $T = r(X_1, X_2, \ldots, X_n)$ is a sufficient statistic for θ if and only if the joint p.d.f or the joint p.m.f $f_n(\boldsymbol{x}, \theta)$ of $(X_1, X_2, \ldots, X_n)$ can be factored as follows for all values of $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and all values of $\theta \in \Omega$:*

$$f_n(\boldsymbol{x}, \theta) = u(\boldsymbol{x})\nu(r(\boldsymbol{x}), \theta),$$

where

- *u and ν are both non-negative,*
- *the function u may depend on $\boldsymbol{x}$ but does not depend on θ,*
- *the function ν will depend on θ but depends on the observed value $\boldsymbol{x}$ only through the value of the statistic $r(\boldsymbol{x})$.*

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d Poi(θ), $\theta > 0$. Thus, for every non-negative integers $x_1, \ldots, x_n$, the joint p.m.f $f_n(\mathbf{x}, \theta)$ of $(X_1, \ldots, X_n)$ is

$$f_n(\mathbf{x}, \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{1}{\prod_{i=1}^{n} x_i!} e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}.$$

Thus, we can take $u(\boldsymbol{x}) = 1/(\prod_{i=1}^{n} x_i!)$, $r(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$, $\nu(t, \theta) = e^{-n\theta}\theta^t$. It follows that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d Gamma(α, β), $\alpha, \beta > 0$, where α is known, and β is unknown. The joint p.d.f is

$$f_n(\boldsymbol{x}, \beta) = \Big\{ [\Gamma(\alpha)]^n \big(\underbrace{\prod_{i=1}^{n} x_i}_{u(\boldsymbol{x})} \big)^{\alpha - 1} \Big\}^{-1} \times \Big\{ \underbrace{\beta^{n\alpha} \exp(-\beta t)}_{\nu(t,\beta)} \Big\}, \quad \text{where } t = \sum_{i=1}^{n} x_i.$$

The sufficient statistics is $T_n = \sum_{i=1}^{n} X_i$.

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d Gamma(α, β), $\alpha, \beta > 0$, where α is unknown, and β is known.

The joint p.d.f in this exercise is the same as that given in the previous exercise. However, since the unknown parameter is now α instead of β, the appropriate factorization is now

$$f_n(\boldsymbol{x}, \alpha) = \Big\{ \underbrace{\exp\big(-\beta \sum_{i=1}^{n} x_i \big)}_{u(\boldsymbol{x})} \Big\} \times \Big\{ \underbrace{\frac{\beta^{n\alpha}}{[\Gamma(\alpha)]^n} t^{\alpha - 1}}_{\nu(t,\alpha)} \Big\}, \quad \text{where } t = \prod_{i=1}^{n} x_i.$$

The sufficient statistics is $T_n = \prod_{i=1}^{n} X_i$.

Exercise: Suppose that $X_1, \ldots, X_n$ are i.i.d Unif($[0, \theta]$), $\theta > 0$ is the unknown parameter. Show that $T = \max\{X_1, \ldots, X_n\}$ is a sufficient statistic.

Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from a distribution for which the p.d.f or p.m.f. is $f(\cdot|\theta)$, where the parameter θ must belong to some parameter space Ω. Let $\boldsymbol{T}$ be a sufficient statistic for θ in this problem.

We show how to improve upon an estimator that is not a function of a sufficient statistic by using an estimator that is a function of a sufficient statistic. Let $\delta(\boldsymbol{X})$ be an estimator of $g(\theta)$. We define the estimator $\delta_0(T)$ by the following conditional expectation:

$$\delta_0(\boldsymbol{T}) = \mathbb{E}_\theta[\delta(\boldsymbol{X})|\boldsymbol{T}].$$

Since $\boldsymbol{T}$ is a sufficient statistic, the conditional expectation of the function $\delta(\boldsymbol{X})$ will be the same for every value of $\theta \in \Omega$. It follows that the conditional expectation above will depend on the value of $\boldsymbol{T}$ but will not actually depend on the value of θ. In other words, the function $\delta_0(\boldsymbol{T})$ is indeed an estimator of $g(\theta)$ because it depends only on the observations $\boldsymbol{X}$ and does not depend on the unknown value of θ.

We can now state the following theorem, which was established independently by D. Blackwell and C. R. Rao in the late 1940s.

Theorem 4.7 (Rao–Blackwell theorem)**.** *For every value of* $\theta \in \Omega$*,*

$$\mathrm{MSE}(\delta_0(\boldsymbol{T}), g(\theta)) \leq \mathrm{MSE}(\delta(\boldsymbol{X}), g(\theta)).$$

The above result is proved in Theorem 7.9.1 of the textbook.

5 The sampling distribution of a statistic

A **statistic** is a function of the data, and hence is itself a random variable with a distribution. This distribution is called its **sampling distribution**. It tells us what values the statistic is likely to assume and how likely is it to take these values. Formally, suppose that $X_1, \ldots, X_n$ are i.i.d with p.d.f/p.m.f $f_\theta(\cdot)$, where $\theta \in \Omega \subset \mathbb{R}^k$. Let T be a statistic, i.e., suppose that $T = \varphi(X_1, \ldots, X_n)$. The distribution of T (with θ fixed) is called the **sampling distribution** of T.

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d $N(\mu, \sigma^2)$. Then we know that

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

5.1 The gamma and the χ^2 distributions

5.1.1 The gamma distribution

The gamma function is a real-valued non-negative function defined on $(0, \infty)$ in the following manner

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx \;\;,\;\; \alpha > 0\,.$$

The Gamma function enjoys some nice properties. Two of these are listed below:

$$(a)\; \Gamma(\alpha+1) = \alpha\,\Gamma(\alpha) \;\;,\;\; (b)\; \Gamma(n) = (n-1)! \;\; (n \;\text{ integer})\,.$$

Property (b) is an easy consequence of Property (a). Start off with $\Gamma(n)$ and use Property (a) recursively along with the fact that $\Gamma(1) = 1$. Another important fact is that $\Gamma(1/2) = \sqrt{\pi}$.

Definition 13. *The **gamma distribution** with parameters $\alpha > 0, \lambda > 0$ (denoted by Gamma(α, λ)) is defined through the following density function:*

$$f(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\, e^{-\lambda x}\, x^{\alpha-1} I_{(0,\infty)}(x).$$

*The first parameter α is called the **shape** parameter and the second parameter λ is called the **scale** parameter.*

By definition, the Gamma$(1, \lambda)$ is nothing else but the exponential distribution with rate λ. For fixed λ the shape parameter regulates the shape of the gamma density. Here is a simple exercise that justifies the term "scale parameter" for λ.

The following properties can be proved using techniques from probability theory (see Chapter 5.7 of textbook):

- **Scaling property:** Let X be a random variable following Gamma(α, λ). Then

$$c\,X \sim \text{Gamma}(\alpha, \tfrac{\lambda}{c}).$$

- **Reproductive property:** Let $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i \sim \text{Gamma}(\alpha_i, \lambda)$, for $i = 1, \ldots, n$. Then,

$$\sum_{i=1}^{n} X_i \;\sim\; \text{Gamma}\left(\sum_{i=1}^{n} \alpha_i, \lambda\right).$$

- **Moments:** If X follows the Gamma(α, λ) distribution, then

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

  In general, if k is a positive integer,

$$\mathbb{E}(X^k) = \frac{\Pi_{i=1}^{k}(\alpha + i - 1)}{\lambda^k}.$$

Exercise: Here is an exercise that should follow from the discussion above. Let $S_n \sim \text{Gamma}(n, \lambda)$, where $\lambda > 0$. Show that for large n, the distribution of S_n is well approximated by a normal distribution (with parameters that you need to identify).

5.1.2 The Chi-squared distribution

We now introduce an important family of distributions, called the chi-squared family. To do so, we first define the **chi-squared distribution** with 1 degree of freedom (for brevity, we call it "chi-squared one" and write it as χ^2_1).

The χ^2_1 distribution: Let $Z \sim N(0, 1)$. Then the distribution of $W := Z^2$ is called the χ^2_1 distribution, and W itself is called a χ^2_1 random variable.

Exercise: Show that W follows a Gamma$(1/2, 1/2)$ distribution. You can do this by relating the c.d.f. of W to that of Z and differentiation.

For any integer $d > 0$ we can now define the χ^2_d distribution (chi-squared d distribution, or equivalently, the chi-squared distribution with d degrees of freedom).

The χ^2_d distribution: Let $Z_1, Z_2, \ldots, Z_d$ be i.i.d $N(0, 1)$ random variables. Then the distribution of

$$W_d := Z_1^2 + Z_2^2 + \cdots + Z_d^2$$

is called the χ^2_d distribution and W_d itself is called a χ^2_d random variable.

Exercise: Using the reproductive property of the Gamma distribution, show that $W_d \sim \text{Gamma}(d/2, 1/2)$.

Exercise: Let Z_1, Z_2, Z_3 be i.i.d $N(0,1)$ random variables. Consider the vector (Z_1, Z_2, Z_3) as a random point in 3-dimensional space. Let R be the length of the radius vector connecting this point to the origin. Find the density functions of (a) R and (b) R^2.

Theorem 5.1. *If $X \sim \chi^2_m$, then $\mathbb{E}(X) = m$ and $\mathrm{Var}(X) = 2m$.*

Theorem 5.2. *Suppose that $X_1, \ldots, X_k$ are independent and $X_i \sim \chi^2_{m_i}$ then the sum*

$$X_1 + \cdots + X_k \sim \chi^2_{\sum_{i=1}^k m_i}.$$

In particular, the sum of k i.i.d χ^2_1 random variables is a χ^2_k random variable.

5.2 Sampling from a normal population

Let $X_1, X_2, \ldots, X_n$ be i.i.d $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$, $\sigma > 0$ are unknown. We have seen that the MLE *and* MOM estimators of the mean and the variance are given by

$$\hat{\mu}_n = \overline{X}_n \quad \text{and} \quad \hat{\sigma}^2_n = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X}_n)^2 .$$

Because $\hat{\sigma}^2_n$ is biased, we will use a slightly different estimator of σ^2 than the one proposed above. We will use the sample variance

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

Besides being an unbiased estimator (see Lemma 4.3), it turns out that s^2_n has a nice interpretation as the multiple of a χ^2 random variable. Here is an interesting (and fairly profound) proposition.

Proposition 5.3. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d sample from some distribution F with mean μ and variance σ^2. Then F is the $N(\mu, \sigma^2)$ distribution if and only if for all n, $\overline{X}_n$ and s^2_n are independent random variables. Moreover, when F is $N(\mu, \sigma^2)$, then*

$$\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n}), \quad \text{and} \quad s^2 \sim \frac{\sigma^2}{n-1} \chi^2_{n-1}.$$

The "if" part is the profound part. It says that the independence of the natural estimates of the mean and the variance for any sample size forces the underlying distribution to be normal. We will sketch a proof of the "only if" part, i.e., we will assume that F is $N(\mu, \sigma^2)$ and show that $\overline{X}_n$ and s^2_n are independent.

Proof. Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Define the *standardized versions* of the X_i's as

$$Y_i = \frac{X_i - \mu}{\sigma}.$$

These are i.i.d. $N(0,1)$ random variables. Now, note that:

$$\overline{X} = \overline{Y}\,\sigma + \mu \quad \text{and} \quad s^2 = \frac{\sigma^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1}.$$

From the above display, we see that it suffices to show the independence of $\overline{Y}$ and $\sum_{i=1}^n (Y_i - \overline{Y})^2$.

The way this proceeds is outlined below: Let $\boldsymbol{Y}$ denote the $n \times 1$ column vector $(Y_1, Y_2, \ldots, Y_n)^T$ and let A be an $n \times n$ orthogonal matrix with the first row of A (which has length n) being $(1/\sqrt{n}, 1/\sqrt{n}, \ldots, 1/\sqrt{n})$.

Recall that an orthogonal matrix satisfies

$$A^\top A = A A^\top = I$$

where I is the identity matrix. Using standard linear algebra techniques it can be shown that such a A can always be constructed. For instance, in the case $n = 2$, we have

$$A = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

Exercise: Show that this A is orthogonal.

Now define a new random vector

$$\boldsymbol{W} = A\boldsymbol{Y}$$

and use the following result:

Theorem 5.4. *If $Y_1, \ldots, Y_n$ are i.i.d $N(0,1)$ and A is an orthogonal matrix and*

$$\boldsymbol{W} = A\boldsymbol{Y},$$

then the random variables $W_1, \ldots, W_n$ are i.i.d $N(0,1)$.

Proof of Theorem 5.4. The joint p.d.f of $\boldsymbol{Y} = (Y_1, \ldots Y_n)$ is

$$f_n(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n y_i^2\right), \qquad \text{for } \boldsymbol{y} \in \mathbb{R}^n.$$

Note that as $\boldsymbol{Y} \mapsto A\boldsymbol{Y}$ is a linear transformation. The joint p.d.f of $\boldsymbol{W} = A\boldsymbol{Y}$ is

$$g_n(\boldsymbol{w}) = \frac{1}{|\det A|} f_n(A^{-1}\boldsymbol{w}), \qquad \text{for } \boldsymbol{w} \in \mathbb{R}^n.$$

Let $\boldsymbol{y} = A^{-1}\boldsymbol{w}$. Since A is orthogonal, $|\det A| = 1$ and $\boldsymbol{w}^\top \boldsymbol{w} = \sum_{i=1}^n w_i^2 = \boldsymbol{y}^\top \boldsymbol{y} = \sum_{i=1}^n y_i^2$. So,

$$g_n(\boldsymbol{w}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n w_i^2\right), \qquad \text{for } \boldsymbol{w} \in \mathbb{R}^n.$$

Thus, $\boldsymbol{W}$ has the same joint p.d.f as $\boldsymbol{Y}$. □

Exercise: Compute W if $n = 2$ and

$$A = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

Returning to the proof of Proposition 5.3, note that

$$\boldsymbol{W}^\top \boldsymbol{W} = (A\boldsymbol{Y})^\top A\boldsymbol{Y} = \boldsymbol{Y}^\top A^\top A\boldsymbol{Y} = \boldsymbol{Y}^\top \boldsymbol{Y}$$

by the orthogonality of A—in other words, $\sum_{i=1}^n W_i^2 = \sum_{i=1}^n Y_i^2$. Also,

$$W_1 = Y_1/\sqrt{n} + Y_2/\sqrt{n} + \cdots + Y_n/\sqrt{n} = \sqrt{n}\,\bar{Y}_n.$$

Note that W_1 is independent of $W_2^2 + W_3^2 + \cdots + W_n^2$. But

$$\sum_{i=2}^n W_i^2 = \sum_{i=1}^n W_i^2 - W_1^2 = \sum_{i=1}^n Y_i^2 - n\,\bar{Y}_n^2 = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2.$$

It therefore follows that $\sqrt{n}\,\bar{Y}_n$ and $\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ are independent – which implies that $\bar{Y}_n$ and $\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ are independent.

Finally, we prove the distributional properties of $\bar{X}_n$ and s_n^2. Note that $\bar{Y}_n \sim N(0, 1/n)$. Deduce that $\bar{X}_n$ follows $N(\mu, \sigma^2/n)$. Since $\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 = W_2^2 + W_3^2 + \cdots + W_n^2$, it follows that

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 \sim \chi^2_{n-1}.$$

Thus,

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}. \tag{2}$$

□

5.3 The t-distribution

Definition 14. *Let $Z \sim N(0,1)$ and let $V \sim \chi^2_n$ be independent of each other. Then,*

$$T = \frac{Z}{\sqrt{V/n}}$$

is said to follow the **t-distribution** *on n degrees of freedom. We write $T \sim t_n$.*

The density of the t-distribution is derived in the text book (see Chapter 8.4). With a little bit of patience, you can also work it out, using the change of variable theorem appropriately (I won't go into the computational details here).

Here are some important facts about the t-distribution. Let $T \sim t_n$.

(a) T and $-T$ have the same distribution. Thus, the distribution of T is symmetric about 0 and it has an even density function.

Indeed, by definition,

$$-T = \frac{-Z}{\sqrt{V/n}} = \frac{\tilde{Z}}{\sqrt{V/n}},$$

where $\tilde{Z} \equiv -Z$ follows $N(0,1)$, and is independent of V where V follows χ^2_n. Thus, by definition, $-T$ also follows the t-distribution on n degrees of freedom.

(b) As $n \to \infty$, the t_n distribution converges to the $N(0,1)$ distribution.

This follows from the law of large numbers. Consider the term V/n in the denominator of T for large n. As V follows χ^2_n it has the same distribution as $K_1 + K_2 + \cdots + K_n$ where K_i's are i.i.d χ^2_1 random variables. But by the WLLN we know that

$$\frac{K_1 + K_2 + \cdots + K_n}{n} \xrightarrow{\mathbb{P}} \mathbb{E}(K_1) = 1 \text{ (check!)}.$$

Thus V/n converges in probability to 1; hence the denominator in T converges in probability to 1 and T converges in distribution to Z, where Z is $N(0,1)$.

Theorem 5.5. *Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Then*

$$\frac{\overline{X}_n - \mu}{\sqrt{s_n^2/n}} \sim t_{n-1}.$$

6 Confidence intervals

Confidence intervals (CIs) provide a method of quantifying uncertainty to an estimator $\hat{\theta}$ when we wish to estimate an unknown parameter θ. We want to find an interval (A, B) that we think has high probability of containing θ.

Definition: Suppose that $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ is a random sample from a distribution P_θ, $\theta \in \Omega$. Suppose that we want to estimate $g(\theta)$, a real-valued function of θ. If $A \leq B$ are two statistics with the property that for all values of θ,

$$\mathbb{P}_\theta(A \leq g(\theta) \leq B) \geq 1 - \alpha,$$

where $\alpha \in (0, 1)$, then the random interval (A, B) is called a **confidence interval** for $g(\theta)$ with **(confidence) level** $1 - \alpha$. If the inequality "$\geq 1 - \alpha$" is an equality for all θ, the CI is called **exact**.

6.1 Construction of confidence interval using a pivot

How do we construct confidence intervals? One approach is to use a so-called pivot.

Definition: A random variable $\Psi(X_1, X_2, \ldots, X_n, g(\theta))$ is called a **pivot** for $g(\theta)$ if its distribution is independent of θ.

Example 1: Find a level $(1 - \alpha)$ CI for μ from data $X_1, X_2, \ldots, X_n$ which are i.i.d. $N(\mu, \sigma^2)$ where σ is **known**. Here $\theta = \mu$ and $g(\theta) = \mu$.

The most intuitive estimator of μ here is the sample mean $\overline{X}_n$. We know that

$$\overline{X}_n \sim N(\mu, \sigma^2/n).$$

The standardized version of the sample mean follows $N(0, 1)$ and can therefore act as a pivot. In other words, construct,

$$\Psi(\boldsymbol{X}_n, \mu) = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}\,(\overline{X}_n - \mu)}{\sigma} \sim N(0, 1)$$

for every value of μ.

With z_β denoting the **upper** β**-quantile** of $N(0, 1)$ (i.e., $\mathbb{P}(Z > z_\beta) = \beta$ where Z follows $N(0, 1)$) we can write:

$$\mathbb{P}_\mu\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

From the above display we can find limits for μ such that the above inequalities are simultaneously satisfied. On doing the algebra, we get:

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

Thus

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

is an exact CI for μ with level $1 - \alpha$.

The general **method of pivots** works as follows:

(1) Construct a function Ψ of the data $\boldsymbol{X}_n$ and $g(\theta)$, say $\Psi(\boldsymbol{X}_n, g(\theta))$, such that the distribution of this random variable under parameter value θ *does not depend on* θ and is known. Such a Ψ gives a pivot for $g(\theta)$.

(2) Let G denote the distribution function of this pivot. The idea now is to get a range of plausible values of the pivot. The level of confidence $1 - \alpha$ is to be used to get the appropriate range.

This can be done in a variety of ways but the following is standard. Denote by $q(G; \beta)$ the β**-quantile** of G, i.e.,

$$\mathbb{P}_\theta[\Psi(\boldsymbol{X}_n, g(\theta)) \leq q(G; \beta)] = \beta.$$

(3) Choose $0 \leq \beta_1, \beta_2 \leq \alpha$ such that $\beta_1 + \beta_2 = \alpha$. Then,

$$\mathbb{P}_\theta[q(G; \beta_1) \leq \Psi(\boldsymbol{X}_n, g(\theta)) \leq q(G; 1 - \beta_2)] = 1 - \beta_2 - \beta_1 = 1 - \alpha\,.$$

(4) Solve the inequalities $q(G; \beta_1) \leq \Psi(\boldsymbol{X}_n, g(\theta)) \leq q(G; 1 - \beta_2)$ for θ to obtain a confidence interval for $g(\theta)$.

Example 2: The data are the same as in Example 1 but now σ^2 is no longer known. Thus, the parameter of unknowns $\theta = (\mu, \sigma^2)$ and we are interested in finding a CI for $g(\theta) = \mu$.

Clearly, setting

$$\Psi(\boldsymbol{X}_n, \mu) = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

will not work smoothly here. This certainly has a known ($N(0, 1)$) distribution but involves the *nuisance parameter* σ (an unknown parameter that we are not primarily interested in).

However, one can replace σ by s_n, where s_n^2 is the sample variance. So, set:

$$\Psi(\boldsymbol{X}_n, \mu) = \frac{\overline{X}_n - \mu}{\sqrt{s_n^2/n}}\,.$$

This only depends on the data and $g(\theta) = \mu$.

This is indeed a pivot: By Theorem 5.5, the new $\Psi(\boldsymbol{X}_n, \mu)$ has a t_{n-1} distribution (which is independent of μ). Thus, G here is the t_{n-1} distribution and we can choose

the quantiles to be $q(t_{n-1}; \alpha/2)$ and $q(t_{n-1}; 1 - \alpha/2)$. By symmetry of the t_{n-1} distribution about 0, we have, $q(t_{n-1}; \alpha/2) = -q(t_{n-1}; 1 - \alpha/2)$. It follows that,

$$\mathbb{P}_{\mu,\sigma^2}\left[-q(t_{n-1}; 1 - \alpha/2) \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} \leq q(t_{n-1}; 1 - \alpha/2)\right] = 1 - \alpha\,.$$

As with Example 1, direct algebraic manipulations show that this is the same as the statement:

$$\mathbb{P}_{\mu,\sigma^2}\left[\overline{X}_n - \frac{s_n}{\sqrt{n}}\, q(t_{n-1}; 1 - \alpha/2) \leq \mu \leq \overline{X}_n + \frac{s_n}{\sqrt{n}}\, q(t_{n-1}; 1 - \alpha/2)\right] = 1 - \alpha\,.$$

This gives a level $1 - \alpha$ confidence set for μ.

Remark: Both z_α and $q(t_n, \alpha)$ can be found from the table on p. 860 of the textbook. For example,

$$z_{0.05} = q(N(0,1), 1 - 0.05) = q(t_\infty, 0.95) = 1.645, \qquad q(t_{16}, 0.99) = 2.583.$$

Food for thought: In each of the above examples there are innumerable ways of decomposing α as $\beta_1 + \beta_2$. It turns out that when α is split equally the level $1 - \alpha$ CIs obtained in Examples 1 and 2 are the shortest.

What are desirable properties of confidence sets? On one hand, we require high levels of confidence; in other words, we would like α to be as small as possible. On the other hand we would like our CIs to be shortest possible. Unfortunately, we cannot simultaneously make the confidence levels of our CIs go up and the lengths of our CIs go down.

In Example 1, the length of the level $(1 - \alpha)$ CI is

$$2\sigma \frac{z_{\alpha/2}}{\sqrt{n}}.$$

As we reduce α (for higher confidence), $z_{\alpha/2}$ increases, making the CI wider.

However, we can reduce the length of our CI for a fixed α by increasing the sample size. If my sample size is 4 times yours, I will end up with a CI which has the same level as yours but has half the length of your CI.

Can we hope to get absolute confidence, i.e. $\alpha = 0$? That is too much of an ask. When $\alpha = 0$, $z_{\alpha/2} = \infty$ and the CIs for μ are infinitely large. The same can be verified for Example 2.

6.2 Asymptotic confidence intervals

The CLT allows us to construct an *approximate pivot* for large sample sizes for estimating the population mean μ for any underlying distribution F.

Let $X_1, X_2, \ldots, X_n$ be i.i.d observations from some common distribution F and let

$$\mathbb{E}(X_1) = \mu \qquad \text{and} \qquad \text{Var}(X_1) = \sigma^2.$$

We are interested in constructing an approximate level $(1-\alpha)$ CI for μ, *assuming that σ is known.*

By the CLT we have $\overline{X}_n \sim_{appx} N(\mu, \sigma^2/n)$ for large n; in other words,

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim_{appx} N(0,1)\,.$$

If σ is known the above quantity is an approximate pivot and following Example 1, we can therefore write,

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

As before, this translates to

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

This gives an approximate level $(1-\alpha)$ CI for μ when σ is known. The approximation will improve as the sample size n increases. Note that the true coverage of the above CI may be different from $1-\alpha$ and can depend heavily on the nature of F and the sample size n.

Realistically, however, σ is unknown and is replaced by s_n. Since we are dealing with large sample sizes, s_n is with very high probability close to σ and the interval

$$\left(\overline{X}_n - \frac{s_n}{\sqrt{n}}\,z_{\alpha/2}\;,\;\overline{X}_n + \frac{s_n}{\sqrt{n}}\,z_{\alpha/2}\right),$$

still remains an approximate level $(1-\alpha)$ CI.

If σ^2 is unknown but can be expressed in terms of the unknown parameter μ, there is a better approach than just using s_n.

Exercise: Suppose $X_1, X_2, \ldots, X_n$ are i.i.d Bernoulli(θ). The sample size n is large.

Thus

$$\mathbb{E}(X_1) = \theta \qquad \text{and} \qquad \text{Var}(X_1) = \theta(1-\theta).$$

We want to find an approximate CI for θ at level $1-\alpha$. Note that both mean and variance are unknown but $\sigma^2 = \theta(1-\theta)$ is a function of θ.

Show that if $\hat{\theta}$ is natural estimate of θ obtained by computing the sample proportion of 1's, then

$$\left[\hat{\theta} - \sqrt{\frac{\hat{\theta}\,(1-\hat{\theta})}{n-1}}\, z_{\alpha/2}\ ,\ \hat{\theta} + \sqrt{\frac{\hat{\theta}\,(1-\hat{\theta})}{n-1}}\, z_{\alpha/2}\right]$$

is an approximate level $(1-\alpha)$ CI for θ.

Interpretation of confidence intervals: Let (A, B) be a coefficient γ confidence interval for a parameter θ. Let (a, b) be the observed value of the interval.

It is NOT correct to say that "θ lies in the interval (a, b) with *probability* γ".

It is true that "θ will lie in the random intervals having endpoints $A(X_1, \ldots, X_n)$ and $B(X_1, \ldots, X_n)$ with probability γ".

After observing the specific values $A(X_1, \ldots, X_n) = a$ and $B(X_1, \ldots, X_n) = b$, it is not possible to assign a probability to the event that θ lies in the specific interval (a, b) without regarding θ as a random variable.

We usually say that there is *confidence* γ that θ lies in the interval (a, b).

$$\frac{\sqrt{n}\,(\hat{\theta} - \theta)}{\sqrt{\theta\,(1-\theta)}} \sim_{appx} N(0, 1)\,,$$

so that

$$P_\theta\left[-z_{\alpha/2} \le \frac{\sqrt{n}\,(\hat{\theta} - \theta)}{\sqrt{\theta\,(1-\theta)}} \le z_{\alpha/2}\right] =_{appx} 1 - \alpha\,.$$

An approximate level $(1-\alpha)$ CI can now be obtained by solving for all θ for which the above inequalities are both satisfied. This amounts to solving a quadratic and will yield a different C.I. than the one proposed in the Exercise. You should try to work out what this gives you.

7 The Cramér–Rao Information Inequality

We saw in the last lecture that for a variety of different models one could differentiate the log-likelihood function with respect to the parameter θ and set this equal to 0 to obtain the MLE of θ.

In these examples, the log-likelihood as a function of θ is strictly concave (looks like an inverted bowl) and hence solving for the stationary point gives us the unique maximizer of the log-likelihood.

We start this section by introducing some notation. Let X be a random variable with p.d.f $f(\cdot, \theta)$, where $\theta \in \Omega$, and

$$\ell(x, \theta) = \log f(x, \theta) \qquad \text{and} \qquad \dot{\ell}(x, \theta) = \frac{\partial}{\partial \theta} \ell(x, \theta).$$

As before, $\boldsymbol{X}_n$ denotes the vector $(X_1, X_2, \ldots, X_n)$ and $\boldsymbol{x}$ denotes a particular value $(x_1, x_2, \ldots, x_n)$ assumed by the random vector $\boldsymbol{X}_n$.

We denote by $f_n(\boldsymbol{x}, \theta)$ the value of the density of $\boldsymbol{X}_n$ at the point $\boldsymbol{x}$. Then,

$$f_n(\boldsymbol{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

Thus,

$$L_n(\theta, \boldsymbol{X}_n) = \prod_{i=1}^{n} f(X_i, \theta) = f_n(\boldsymbol{X}_n, \theta)$$

and

$$\ell_n(\boldsymbol{X}_n, \theta) = \log L_n(\theta, \boldsymbol{X}_n) = \sum_{i=1}^{n} \ell(X_i, \theta).$$

Differentiating with respect to θ yields

$$\dot{\ell}_n(\boldsymbol{X}_n, \theta) = \frac{\partial}{\partial \theta} \log f_n(\boldsymbol{X}_n, \theta) = \sum_{i=1}^{n} \dot{\ell}(X_i, \theta).$$

We call $\dot{\ell}(x, \theta)$ the **score function** and

$$\dot{\ell}_n(\boldsymbol{X}_n, \theta) = 0$$

the **score equation**. If differentiation is permissible for the purpose of obtaining the MLE, then $\hat{\theta}_n$, the MLE, solves the equation

$$\dot{\ell}_n(\boldsymbol{X}_n, \theta) \equiv \sum_{i=1}^{n} \dot{\ell}(X_i, \theta) = 0.$$

In this section, our first goal is to find a (nontrivial) **lower bound** on the **variance of unbiased estimators** of $g(\theta)$ where $g : \Omega \to \mathbb{R}$ is some differentiable function.

If we can indeed find such a bound (albeit under some regularity conditions) and there is an unbiased estimator of $g(\theta)$ that attains this lower bound, we can conclude that it is the MVUE of $g(\theta)$.

We now impose the following restrictions (regularity conditions) on the model.

(A.1) The set $A_\theta = \{x : f(x, \theta) > 0\}$ actually does NOT depend on θ and is subsequently denoted by A.

(A.2) If $W(\boldsymbol{X}_n)$ is a statistic such that $\mathbb{E}_\theta(|W(\boldsymbol{X}_n)|) < \infty$ for all θ, then,

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta[W(\boldsymbol{X}_n)] = \frac{\partial}{\partial\theta}\int_{A^n} W(\boldsymbol{x})\, f_n(\boldsymbol{x}, \theta)\, d\boldsymbol{x} = \int_{A^n} W(\boldsymbol{x})\, \frac{\partial}{\partial\theta} f_n(\boldsymbol{x}, \theta) d\boldsymbol{x}.$$

(A.3) The quantity $\frac{\partial}{\partial\theta}\log f(x, \theta)$ exists for all $x \in A$ and all $\theta \in \Omega$ as a well-defined finite quantity.

The first condition says that the set of possible values of the data vector on which the distribution of $\boldsymbol{X}_n$ is supported does not vary with θ; this therefore rules out families of distribution like the uniform.

The second assumption is a "smoothness assumption" on the family of densities and is generally happily satisfied for most parametric models we encounter in statistics.

There are various types of simple sufficient conditions that one can impose on $f(x, \theta)$ to make the interchange of integration and differentiation possible — we shall however not bother about these for the moment.

7.1 Information

For most of the sequel, for notational simplicity, we will assume that the parameter space $\Omega \subset \mathbb{R}$. We define the **Fisher information** about the parameter θ in the model, namely $I(\theta)$, by

$$I(\theta) := \mathbb{E}_\theta[\dot{\ell}^2(X, \theta)],$$

provided it exists as a finite quantity for every $\theta \in \Omega$.

We then have the following theorem.

Theorem 7.1 (Cramér–Rao inequality)**.** *All notation being as above, if $T(\boldsymbol{X}_n)$ is an unbiased estimator of $g(\theta)$, then*

$$\mathrm{Var}_\theta(T(\boldsymbol{X}_n)) \geq \frac{[g'(\theta)]^2}{nI(\theta)},$$

provided assumptions A.1, A.2 and A.3 hold, and $I(\theta)$ exists and is finite for all θ.

The above inequality is the celebrated **Cramér–Rao inequality** (or the information inequality) and is one of the most well-known inequalities in statistics and has important ramifications in even more advanced forms of inference.

Notice that if we take $g(\theta) = \theta$ then $n^{-1}I(\theta)^{-1}$ gives us a lower bound on the variance of unbiased estimators of θ in the model.

If $I(\theta)$ is small, the lower bound is large, so unbiased estimators are doing a poor job in general—in other words, the data is not that informative about θ (within the context of unbiased estimation).

On the other hand, if $I(\theta)$ is big, the lower bound is small, and so if we have a best unbiased estimator of θ that actually attains this lower bound, we are doing a good job. That is why $I(\theta)$ is referred to as the information about θ.

Proof of Theorem 7.1: By the Cauchy–Schwarz inequality (see Theorem 4.6.3 in the textbook),

$$\mathrm{Cov}_\theta^2\left(T(\boldsymbol{X}_n), \dot{\ell}_n(\boldsymbol{X}_n, \theta)\right) \leq \mathrm{Var}_\theta(T(\boldsymbol{X}_n))\mathrm{Var}_\theta(\dot{\ell}_n(\boldsymbol{X}_n, \theta)). \tag{3}$$

As

$$1 = \int f_n(\boldsymbol{x}, \theta)\, d\boldsymbol{x}, \quad \text{for all } \theta \in \Omega,$$

on differentiating both sides of the above identity with respect to θ and using (A.2) with $W(\boldsymbol{x}) \equiv 1$ we obtain,

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta} f_n(\boldsymbol{x}, \theta) d\boldsymbol{x} = \int \left(\frac{\partial}{\partial\theta} f_n(\boldsymbol{x}, \theta)\right) \frac{1}{f_n(\boldsymbol{x}, \theta)} f_n(\boldsymbol{x}, \theta) d\boldsymbol{x} \\ &= \int \left(\frac{\partial}{\partial\theta} \log f_n(\boldsymbol{x}, \theta)\right) f_n(\boldsymbol{x}, \theta) d\boldsymbol{x}. \end{aligned}$$

The last expression in the above display is precisely $\mathbb{E}_\theta[\dot{\ell}_n(\boldsymbol{X}_n, \theta)]$ which therefore is equal to 0. Note that

$$\mathbb{E}_\theta[\dot{\ell}_n(\boldsymbol{X}_n, \theta)] = \mathbb{E}_\theta\left[\sum_{i=1}^n \dot{\ell}(X_i, \theta)\right] = n\mathbb{E}_\theta[\dot{\ell}(X, \theta)],$$

since the $\dot{\ell}(X_i, \theta)$'s are i.i.d. Thus, we have $\mathbb{E}_\theta[\dot{\ell}(X_1, \theta)] = 0$. This implies that

$$I(\theta) = \mathrm{Var}_\theta(\dot{\ell}(X, \theta)).$$

Further, let $I_n(\theta) := \mathbb{E}_\theta[\dot{\ell}_n^2(\boldsymbol{X}_n, \theta)]$. Then

$$\begin{aligned} I_n(\theta) &= \mathrm{Var}_\theta(\dot{\ell}_n(\boldsymbol{X}_n, \theta)) = \mathrm{Var}_\theta\left(\sum_{i=1}^n \dot{\ell}(X_i, \theta)\right) \\ &= \sum_{i=1}^n \mathrm{Var}_\theta(\dot{\ell}(X_i, \theta)) = nI(\theta). \end{aligned}$$

We will refer to $I_n(\theta)$ as the **Fisher information in the sample $\boldsymbol{X}_n$**. Since $\mathbb{E}_\theta[\dot{\ell}_n(\boldsymbol{X}_n, \theta)] = 0$, it follows that

$$\begin{aligned}
\mathrm{Cov}_\theta\left(T(\boldsymbol{X}_n), \dot{\ell}_n(\boldsymbol{X}_n, \theta)\right) &= \int T(\boldsymbol{x})\dot{\ell}_n(\boldsymbol{x}, \theta) f_n(\boldsymbol{x}, \theta) d\boldsymbol{x} \\
&= \int T(\boldsymbol{x}) \left(\frac{\partial}{\partial\theta} f_n(\boldsymbol{x}, \theta)\right) d\boldsymbol{x} \\
&= \frac{\partial}{\partial\theta} \int T(\boldsymbol{x}) f_n(\boldsymbol{x}, \theta) d\boldsymbol{x} \qquad \text{(by (A.2))} \\
&= \frac{\partial}{\partial\theta} g(\theta) = g'(\theta).
\end{aligned}$$

Using the above in conjunction in (3) we get,

$$[g'(\theta)]^2 \leq \mathrm{Var}_\theta(T(\boldsymbol{X}_n))\ I_n(\theta)$$

which is equivalent to what we set out to prove. □

There is an alternative expression for the Fisher information $I(\theta)$ in terms of the second derivative of the log-likelihood with respect to θ. If

$$\ddot{\ell}(x, \theta) := \frac{\partial^2}{\partial\theta^2} \log f(x, \theta)$$

exists for all $x \in A$ and for all $\theta \in \Theta$, then we have the following identity:

$$I(\theta) = \mathbb{E}_\theta[\dot{\ell}(X, \theta)^2] = -\mathbb{E}_\theta[\ddot{\ell}(X, \theta)],$$

provided we can differentiate twice under the integral sign; more concretely, if

$$\int \frac{\partial^2}{\partial\theta^2} f(x, \theta) dx = \frac{\partial^2}{\partial\theta^2} \int f(x, \theta) dx = 0 \quad (\star).$$

To prove the above identity, first note that,

$$\dot{\ell}(x, \theta) = \frac{1}{f(x, \theta)} \left[\frac{\partial}{\partial\theta} f(x, \theta)\right].$$

Now,

$$\begin{aligned}
\ddot{\ell}(x, \theta) &= \frac{\partial}{\partial\theta}\left(\dot{\ell}(x, \theta)\right) = \frac{\partial}{\partial\theta}\left(\frac{1}{f(x, \theta)} \frac{\partial}{\partial\theta} f(x, \theta)\right) \\
&= \frac{\partial^2}{\partial\theta^2} f(x, \theta) \frac{1}{f(x, \theta)} - \frac{1}{f^2(x, \theta)} \left(\frac{\partial}{\partial\theta} f(x, \theta)\right)^2 \\
&= \frac{\partial^2}{\partial\theta^2} f(x, \theta) \frac{1}{f(x, \theta)} - \dot{\ell}(x, \theta)^2.
\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}_\theta[\ddot{\ell}(X,\theta)] &= \int \ddot{\ell}(x,\theta) f(x,\theta)dx \\ &= \int \frac{\partial^2}{\partial\theta^2} f(x,\theta)dx - \mathbb{E}_\theta[\dot{\ell}^2(X,\theta)] \\ &= 0 - \mathbb{E}_\theta[\dot{\ell}^2(X,\theta)],\end{aligned}$$

where the first term on the right side vanishes by virtue of $(\star)$. This establishes the desired equality. It follows that,

$$I_n(\theta) = \mathbb{E}_\theta[-\ddot{\ell}_n(\boldsymbol{X}_n,\theta)],$$

where $\ddot{\ell}_n(\boldsymbol{X}_n,\theta)$ is the second partial derivative of $\ell_n(\boldsymbol{X}_n,\theta)$ with respect to θ. To see this, note that,

$$\ddot{\ell}_n(\boldsymbol{X}_n,\theta) = \frac{\partial^2}{\partial\theta^2}\left(\sum_{i=1}^n \ell(X_i,\theta)\right) = \sum_{i=1}^n \ddot{\ell}(X_i,\theta),$$

so that

$$\mathbb{E}_\theta[\ddot{\ell}_n(\boldsymbol{X}_n,\theta)] = \sum_{i=1}^n \mathbb{E}_\theta[\ddot{\ell}(X_i,\theta)] = n\,\mathbb{E}_\theta[\ddot{\ell}(X,\theta)] = -n\,I(\theta).$$

We have just established the following result.

Theorem 7.2. *If (A.1)–(A.3) as well as $(\star)$ hold, then*

$$I(\theta) = -\mathbb{E}_\theta[\ddot{\ell}(X,\theta)] = \mathrm{Var}_\theta(\dot{\ell}(X,\theta)).$$

7.2 Examples

We now look at some applications of the Cramér–Rao inequality.

Example 1: Let $X_1, X_2, \ldots, X_n$ be i.i.d Pois(θ), $\theta > 0$. Then

$$\mathbb{E}_\theta(X_1) = \theta \qquad \text{and} \qquad \mathrm{Var}_\theta(X_1) = \theta.$$

Let us first write down the likelihood of the data. We have,

$$f_n(\boldsymbol{x},\theta) = \prod_{i=1}^n \frac{e^{-\theta}\,\theta^{x_i}}{x_i!} = e^{-n\,\theta}\,\theta^{\sum_{i=1}^n x_i}\left(\prod_{i=1}^n x_i!\right)^{-1}.$$

Thus,

$$\begin{aligned}\ell_n(\boldsymbol{x},\theta) &= -n\theta + \log\theta\left(\sum_{i=1}^n x_i\right) - \log\prod_{i=1}^n x_i! \\ \dot{\ell}_n(\boldsymbol{x},\theta) &= -n + \frac{1}{\theta}\sum_{i=1}^n x_i.\end{aligned}$$

Thus the Fisher information about θ in the sample $X_1, \ldots, X_n$ is given by

$$I_n(\theta) = \mathrm{Var}_\theta\left(-n + \frac{1}{\theta}\sum_{i=1}^n X_i\right) = \frac{1}{\theta^2}\mathrm{Var}_\theta\left(\sum_{i=1}^n X_i\right) = \frac{n\theta}{\theta^2} = \frac{n}{\theta}.$$

The assumptions needed for the Cramé–Rao inequality to hold are all satisfied for this model, and it follows that for any unbiased estimator $T(\boldsymbol{X}_n)$ of $g(\theta) = \theta$ we have,

$$\mathrm{Var}_\theta(T(\boldsymbol{X}_n)) \geq \frac{1}{I_n(\theta)} = \frac{\theta}{n}.$$

Since $\overline{X}_n$ is unbiased for θ and has variance θ/n we conclude that $\overline{X}_n$ is the MVUE of θ.

Example 2: Let $X_1, X_2, \ldots, X_n$ be i.i.d $N(0, V)$. Consider once again, the joint density of the n observations:

$$f_n(\boldsymbol{x}, V) = \frac{1}{(2\pi V)^{n/2}}\exp\left(-\frac{1}{2V}\sum_{i=1}^n x_i^2\right).$$

Now,

$$\begin{aligned}\dot{\ell}_n(\boldsymbol{x}, V) &= \frac{\partial}{\partial V}\left(-\frac{n}{2}\log 2\pi - \frac{n}{2}\log V - \frac{1}{2V}\sum_{i=1}^n x_i^2\right)\\ &= -\frac{n}{2V} + \frac{1}{2V^2}\sum_{i=1}^n x_i^2.\end{aligned}$$

Differentiating yet again we obtain,

$$\ddot{\ell}_n(\boldsymbol{x}, V) = \frac{n}{2V^2} - \frac{1}{V^3}\sum_{i=1}^n x_i^2.$$

Then, the Fisher information for V based on $X_1, \ldots, X_n$ is

$$I_n(V) = -\mathbb{E}_V\left(\frac{n}{2V^2} - \frac{1}{V^3}\sum_{i=1}^n X_i^2\right) = \frac{n}{2\,V^2} + \frac{1}{V^3}\,nV = \frac{n}{2\,V^2}\,.$$

Now consider the problem of estimating $g(V) = V$. For any unbiased estimator $S(\boldsymbol{X}_n)$ of V, the Cramér–Rao inequality tells us that

$$\mathrm{Var}_V\left(S(\boldsymbol{X}_n)\right) \geq I_n(V)^{-1} = \frac{2V^2}{n}.$$

Consider, $\sum_{i=1}^n X_i^2/n$ as an estimator of V. This is clearly unbiased for V and the variance is given by,

$$\mathrm{Var}_V\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) = \frac{1}{n}\mathrm{Var}_V(X_1^2) = \frac{V^2}{n}\mathrm{Var}_V\left(\frac{X_1^2}{V}\right) = \frac{2V^2}{n},$$

since $X_1^2/V \sim \chi_1^2$ which has variance 2. It follows that $\sum X_i^2/n$ is the MVUE of V.

7.3 Large sample properties of the MLE

In this subsection we study some of the large sample properties of the MLE in standard parametric models and how these can be used to construct confidence sets for θ or a function of θ. We will see in this section that in the long run MLEs are the best possible estimators in a variety of different models.

We will stick to models satisfying the restrictions (A.1)–(A.3) imposed in the last section. Hence our results will not apply to the uniform distribution (or ones similar to the uniform).

Let us throw our minds back to the Cramér–Rao inequality. When does an unbiased estimator $T(\boldsymbol{X}_n)$ of $g(\theta)$ attain the bound given by this inequality? This requires:

$$\mathrm{Var}_\theta(T(\boldsymbol{X}_n)) = \frac{(g'(\theta))^2}{n\, I(\theta)}\,.$$

But this is equivalent to the assertion that the correlation between $T(\boldsymbol{X}_n)$ and $\dot{\ell}_n(\boldsymbol{X}_n, \theta)$ is equal to 1 or -1.

This means that $\dot{\ell}_n(\boldsymbol{X}_n, \theta)$ can be expressed as a *linear function* of $T(\boldsymbol{X}_n)$.

In fact, this is a necessary and sufficient condition for the information bound to be attained by the variance of $T(\boldsymbol{X}_n)$.

It turns out that this is generally difficult to achieve. Thus, there will be many different functions of θ, for which best unbiased estimators will exist but whose variance will not hit the information bound. The example below will illustrate this point.

Example: Let $X_1, X_2, \ldots, X_n$ be i.i.d Ber(θ). We have,

$$f(x, \theta) = \theta^x (1-\theta)^{1-x} \qquad \text{for } x = 0, 1.$$

Thus,

$$\ell(x, \theta) = x \log \theta + (1-x) \log(1-\theta)\,,$$

$$\dot{\ell}(x, \theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

and

$$\ddot{\ell}(x, \theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}\,.$$

Thus,

$$\dot{\ell}_n(\boldsymbol{X}_n, \theta) = \sum_{i=1}^n \dot{\ell}(X_i, \theta) = \frac{\sum_{i=1}^n X_i}{\theta} - \frac{n - \sum_{i=1}^n X_i}{1-\theta}\,.$$

Recall that the MLE solves $\dot{\ell}_n(\boldsymbol{X}_n, \theta) = 0$.

Check that in this situation, this gives you precisely $\overline{X}_n$ as your MLE.

Let us compute the Fisher information $I(\theta)$. We have,

$$I(\theta) = -\mathbb{E}_\theta[\ddot{\ell}(X_1, \theta)] = \mathbb{E}_\theta\left(\frac{X_1}{\theta^2} + \frac{1 - X_1}{(1-\theta)^2}\right) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta\,(1-\theta)}.$$

Thus,

$$I_n(\theta) = nI(\theta) = \frac{n}{\theta(1-\theta)}.$$

Consider unbiased estimation of $\Psi(\theta) = \theta$ based on $\boldsymbol{X}_n$. Let $T(\boldsymbol{X}_n)$ be an unbiased estimator of θ. Then, by the information inequality,

$$\mathrm{Var}_\theta(T(\boldsymbol{X}_n)) \geq \frac{\theta(1-\theta)}{n}.$$

Note that the variance of $\overline{X}_n$ is precisely $\theta(1-\theta)/n$, so that it is the MVUE of θ. Note that

$$\dot{\ell}_n(\boldsymbol{X}_n, \theta) = \frac{n\overline{X}}{\theta} - \frac{n(1-\overline{X})}{1-\theta} = \left(\frac{n}{\theta} + \frac{n}{1-\theta}\right)\overline{X} - \frac{n}{1-\theta}.$$

Thus, $\overline{X}_n$ is indeed linear in $\dot{\ell}_n(\boldsymbol{X}_n, \theta)$.

Consider now estimating a different function of θ, say $g(\theta) = \theta^2$.

This is the probability of getting two consecutive heads. Suppose we try to find an unbiased estimator of this parameter.

Then $S(\boldsymbol{X}_n) = X_1X_2$ is an unbiased estimator ($\mathbb{E}_\theta(X_1X_2) = \mathbb{E}_\theta(X_1)\mathbb{E}_\theta(X_2) = \theta^2$), but then so is X_iX_j for any $i \neq j$.

We can find the MVUE of θ^2 in this model by using techniques beyond the scope of this course—it can be shown that any estimator $T(\boldsymbol{X}_n)$ that can be written as a function of $\overline{X}_n$ and is unbiased for θ^2 is an MVUE (and indeed there is one such).

Verify that,

$$T^*(\boldsymbol{X}_n) = \frac{n\,\overline{X}_n^2 - \overline{X}_n}{n-1}$$

is unbiased for θ^2 and is therefore an (in fact *the*) MVUE.

However, the variance of $T^*(\boldsymbol{X}_n)$ does not attain the information bound for estimating $g(\theta)$ which is $4\theta^3(1-\theta)/n$ (Exercise). This can be checked by direct (somewhat tedious) computation or by noting that $T^*(\boldsymbol{X}_n)$ is not a linear function of $\dot{\ell}_n(\boldsymbol{X}_n, \theta)$.

The question then is whether we can propose an estimator of θ^2 that does achieve the bound, at least approximately, in the long run.

It turns out that this is actually possible. Since the MLE of θ is $\overline{X}$, the MLE of $g(\theta)$ is proposed as the plug-in value $g(\overline{X}) = \overline{X}^2$.

This is *not an unbiased estimator of* $g(\theta)$ in finite samples, but has excellent behavior in the long run. In fact,

$$\sqrt{n}\,(g(\overline{X}_n) - g(\theta)) \to_d N(0, 4\theta^3(1-\theta)).$$

Exercise: Prove the last statement.

Thus for large values of n, $g(\overline{X})$ behaves approximately like a normal random variable with mean $g(\theta)$ and variance $4\theta^3(1-\theta)/n$.

In this sense, $g(\overline{X}_n)$ is *asymptotically (in the long run) unbiased* and *asymptotically efficient* (in the sense that it has minimum variance).

Here is an important proposition that establishes the limiting behavior of the MLE.

Theorem 7.3. *If $\hat{\theta}_n$ is the MLE of θ obtained by solving*

$$\sum_{i=1}^{n} \dot{\ell}(X_i, \theta) = 0,$$

then the following representation for the MLE is valid:

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\theta)^{-1} \dot{\ell}(X_i, \theta) + r_n,$$

where r_n converges to 0 in probability. It follows by a direct application of the CLT that,

$$\sqrt{n}(\hat{\theta}_n - \theta) \to_d N(0, I(\theta)^{-1}).$$

The above result shows MLE $\hat{\theta}$ is (asymptotically) the best possible estimator: Not only does its long term distribution center around θ, the quantity of interest, its distribution is also less spread out than that of any "reasonable" estimator of θ. If S_n is a "reasonable" estimator of θ, with

$$\sqrt{n}\,(S_n - \theta) \to_d N(0, \xi^2(\theta))\,,$$

then $\xi^2(\theta) \geq I(\theta)^{-1}$.

Recall the delta method.

Proposition 7.4 (Delta method)**.** *Suppose T_n is an estimator of θ (based on i.i.d observations, $X_1, X_2, \ldots, X_n$ from P_θ) that satisfies:*

$$\sqrt{n}(T_n - \theta) \to_d N(0, \sigma^2(\theta)).$$

Here $\sigma^2(\theta)$ is the limiting variance and depends on the underlying parameter θ. Then, for a continuously differentiable function h such that $h'(g(\theta)) \neq 0$, we have:

$$\sqrt{n}\,(g(T_n) - g(\theta)) \to_d N(0, (g'(\theta))^2\,\sigma^2(\theta))\,.$$

We can now deduce the limiting behavior of the MLE of $g(\theta)$ given by $g(\hat{\theta}_n)$ for any smooth function g such that $g'(\theta) \neq 0$.

Combining Proposition 7.3 with Proposition 7.4 yields (take $T_n = \hat{\theta}_n$)

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \to_d N(0, g'(\theta)^2\, I(\theta)^{-1}).$$

Thus, for large n,

$$g(\hat{\theta}_n) \sim_{appx} N(g(\theta), g'(\theta)^2\, (n\, I(\theta))^{-1})\,.$$

Thus $g(\hat{\theta}_n)$ is asymptotically unbiased for $g(\theta)$ (unbiased in the long run) and its variance is approximately the information bound for unbiased estimators of $g(\theta)$.

Constructing confidence sets for θ: Suppose that, for simplicity, θ takes values in a subset of $\mathbb{R}$. Since,

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

it follows that

$$\sqrt{n\, I(\theta)}(\hat{\theta}_n - \theta) \to_d N(0, 1).$$

Thus, the left side acts as an *approximate pivot* for θ. We have,

$$\mathbb{P}_\theta \left(-z_{\alpha/2} \leq \sqrt{nI(\theta)}(\hat{\theta}_n - \theta) \leq z_{\alpha/2} \right) \approx 1 - \alpha.$$

An approximate level $1 - \alpha$ confidence set for θ is obtained as

$$\left\{ \theta : -z_{\alpha/2} \leq \sqrt{nI(\theta)}(\hat{\theta}_n - \theta) \leq z_{\alpha/2} \right\}.$$

To find the above confidence set, one needs to solve for all values of θ satisfying the inequalities in the above display; this can however be a potentially complicated exercise depending on the functional form for $I(\theta)$.

However, if the sample size n is large, $I(\hat{\theta}_n)$ can be expected to be close to $I(\theta)$ with high probability and hence the following is also valid:

$$P_\theta \left[-z_{\alpha/2} \leq \sqrt{n\, I(\hat{\theta})}(\hat{\theta}_n - \theta) \leq z_{\alpha/2} \right] \approx 1 - \alpha. \quad (\star\star)$$

This immediately gives an approximate level $1 - \alpha$ CI for θ as:

$$\left[\hat{\theta}_n - \frac{1}{\sqrt{nI(\hat{\theta}_n)}}\, z_{\alpha/2}, \hat{\theta}_n + \frac{1}{\sqrt{nI(\hat{\theta}_n)}} z_{\alpha/2} \right].$$

Let's see what this implies for the Bernoulli example discussed above. Recall that $I(\theta) = (\theta\,(1-\theta))^{-1}$ and $\hat{\theta} = \overline{X}$. The approximate $(1-\alpha)$-CI is then given by,

$$\left[\overline{X}_n - \sqrt{\frac{\overline{X}_n\,(1 - \overline{X}_n)}{n}}\, z_{\alpha/2}, \overline{X}_n + \sqrt{\frac{\overline{X}_n\,(1 - \overline{X}_n)}{n}}\, z_{\alpha/2} \right].$$

Exercise: Find explicitly

$$\left\{\theta : -z_{\alpha/2} \leq \sqrt{n\, I(\theta)}(\hat{\theta}_n - \theta) \leq z_{\alpha/2}\right\}$$

in the following cases (a) $X_1, X_2, \ldots, X_n$ are i.i.d Bernoulli(θ). (b) $X_1, X_2, \ldots, X_n$ are i.i.d Pois(θ).

You will see that this involves solving for the roots of a quadratic equation. As in the Bernoulli example, one can also get an approximate CI for θ in the Poisson setting on using ($\star\star$). Verify that this yields the following level $1-\alpha$ CI for θ:

$$\left[\overline{X}_n - \sqrt{\frac{\overline{X}_n}{n}}\, z_{\alpha/2}, \overline{X}_n + \sqrt{\frac{\overline{X}_n}{n}}\, z_{\alpha/2}\right].$$

The recipe ($\star\star$) is somewhat unsatisfactory because it involves one more level of approximation in that $I(\theta)$ is replaced by $I(\hat{\theta})$ (note that there is already one level of approximation in that the pivots being considered are only approximately $N(0,1)$ by the CLT).

8 Bayesian paradigm

Frequentist versus Bayesian statistics:

Frequentist:

- Data are a repeatable random sample — there is a frequency.
- *Parameters are fixed.*
- Underlying parameters remain constant during this repeatable process.

Bayesian:

- Parameters are unknown and described probabilistically.
- *Analysis is done conditioning on the observed data; i.e., data is treated as fixed.*

8.1 Prior distribution

Definition 15 (Prior distribution)**.** *Suppose that one has a statistical model with parameter θ. If one treats θ as random, then the distribution that one assigns to θ before observing the data is called its* **prior distribution**. *Its pdf/pmf is called the* ***prior pdf/pmf*** *of θ.*

Thus, now θ is random and will be denoted by Θ (note the change of notation).

Example: Let Θ denote the probability of obtaining a head when a certain coin is tossed.

- Case 1: Suppose that it is known that the coin either is fair or has a head on each side. Then Θ only takes two values, namely $1/2$ and 1. If the prior probability that the coin is fair is 0.8, then the prior p.m.f of Θ is $\xi(1/2) = 0.8$ and $\xi(1) = 0.2$.
- Case 2: Suppose that Θ can take any value between $(0, 1)$ with a prior distribution given by a Beta distribution with parameters $(1, 1)$.

Suppose that the observable data $X_1, X_2, \ldots, X_n$ are modeled as random sample from a distribution indexed by θ. Suppose $f(\cdot|\theta)$ denote the p.m.f/p.d.f of a single random variable under the distribution indexed by θ.

When we treat the unknown parameter Θ as random, then the joint distribution of the observable random variables (i.e., data) indexed by θ is understood as the **conditional distribution** of the data given $\Theta = \theta$.

Thus, in general we will have $X_1, \ldots, X_n | \Theta = \theta$ are i.i.d with p.d.f/p.m.f $f(\cdot|\theta)$, and that $\Theta \sim \xi$, i.e.,

$$f_n(\boldsymbol{x}|\theta) = f(x_1|\theta) \cdots f(x_n|\theta),$$

where f_n is the joint conditional distribution of $\boldsymbol{X} = (X_1, \ldots, X_n)$ given $\Theta = \theta$.

8.2 Posterior distribution

Definition 16 (Posterior distribution)**.** *Consider a statistical inference problem with parameter θ and random variables $X_1, \ldots, X_n$ to be observed. The conditional distribution of Θ given $X_1, \ldots, X_n$ is called the* **posterior distribution** *of θ. The conditional p.m.f/p.d.f of Θ given $X_1 = x_1, \ldots, X_n = x_n$ is called the* ***posterior p.m.f/p.d.f*** *of θ and is usually denoted by $\xi(\cdot|x_1, \ldots, x_n)$.*

Theorem 8.1. *Suppose that the n random variables $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f/p.m.f is $f(\cdot|\theta)$. Suppose also that the value of the parameter θ is unknown and the prior p.d.f/p.m.f of θ is $\xi(\cdot)$. Then the posterior p.d.f/p.m.f of θ is*

$$\xi(\theta|\boldsymbol{x}) = \frac{f(x_1|\theta) \cdots f(x_n|\theta)\xi(\theta)}{g_n(\boldsymbol{x})}, \qquad \text{for } \theta \in \Omega,$$

where g_n is the marginal joint p.d.f/p.m.f of $X_1, \ldots, X_n$.

Example 8.2 (Sampling from a Bernoulli distribution)**.** Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with mean $\theta > 0$, where $0 < \theta < 1$ is unknown. Suppose that the prior distribution of Θ is Beta(α, β), where $\alpha, \beta > 0$.

Then the posterior distribution of Θ given $X_i = x_i$, for $i = 1, \ldots, n$, is Beta$(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$.

Proof. The joint p.m.f of the data is

$$f_n(\boldsymbol{x}|\theta) = f(x_1|\theta) \cdots f(x_n|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Therefore the posterior density of $\Theta | X_1 = x_1, \ldots, X_n = x_n$ is given by

$$\begin{aligned} \xi(\theta|\boldsymbol{x}) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ &= \theta^{\sum_{i=1}^n x_i + \alpha - 1}(1-\theta)^{\beta+n-\sum_{i=1}^n x_i - 1}, \end{aligned}$$

for $\theta \in (0,1)$. Thus, $\Theta | X_1 = x_1, \ldots, X_n = x_n \sim \text{Beta}(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$. □

Example 8.3 (Sampling from a Poisson distribution)**.** Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with mean $\theta > 0$, where θ is unknown. Suppose that the prior distribution of Θ is Gamma(α, β), where $\alpha, \beta > 0$. Show that the posterior distribution of Θ given $X_i = x_i$, for $i = 1, \ldots, n$, is Gamma$(\alpha + \sum_{i=1}^n x_i, \beta + n)$.

Definition: Let $X_1, X_2, \ldots,$ be conditionally i.i.d given $\Theta = \theta$ with p.m.f/p.d.f $f(\cdot|\theta)$, where $\theta \in \Omega$. Let Ψ be a family of possible distributions over the parameter space Ω. Suppose that no matter which prior distribution ξ we choose from Ψ, no matter how many observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ we observe, and no matter what their observed values $\boldsymbol{x} = (x_1, \ldots, x_n)$ are, the posterior distribution $\xi(\cdot|\boldsymbol{x})$ is a member of Ψ. Then Ψ is called a *conjugate family of prior distributions* for samples from the distributions $f(\cdot|\theta)$.

Example 8.4 (Sampling from an Exponential distribution)**.** Suppose that the distribution of the lifetime of fluorescent tubes of a certain type is the exponential distribution with parameter θ. Suppose that $X_1, \ldots, X_n$ is a random sample of lamps of this type. Also suppose that $\Theta \sim$ Gamma(α, β). Then

$$f_n(\boldsymbol{x}|\theta) = \prod_{i=1}^n \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^n x_i}.$$

Then the posterior distribution of Θ given the data is

$$\xi(\theta|\boldsymbol{x}) \propto \theta^n e^{-\theta \sum_{i=1}^n x_i} \cdot \theta^{\alpha-1} e^{-\beta\theta} = \theta^{n+\alpha-1} e^{-(\beta + \sum_{i=1}^n x_i)\theta}.$$

Therefore, $\Theta|\boldsymbol{X}_n = \boldsymbol{x} \sim$ Gamma$(\alpha + n, \beta + \sum_{i=1}^n x_i)$.

8.3 Bayes Estimators

Definition: A *loss function* is a real-valued function of two variables, $L(\theta, a)$, where $\theta \in \Omega$ and $a \in \mathbb{R}$.

The interpretation is that the statistician loses $L(\theta, a)$ if the parameter equals θ and the estimate equals a.

Example: (Squared error loss) $L(\theta, a) = (\theta - a)^2$.

(Absolute error loss) $L(\theta, a) = |\theta - a|$.

Suppose that $\Theta \sim \xi(\cdot)$ is a p.d.f/p.m.f. Consider the problem of estimating Θ without being able to observe the data. If the statistician chooses a particular estimate a, then their expected loss will be

$$\mathbb{E}[L(\Theta, a)] = \int L(\theta, a)\xi(\theta)d\theta.$$

It is sensible that the statistician wishes to choose an estimate a for which the expected loss is *minimal*.

Definition: Suppose now that the statistician can observe the value $\boldsymbol{x}$ of data $\boldsymbol{X}_n$, and let $\xi(\cdot|\boldsymbol{x})$ denote the posterior p.d.f/p.m.f of $\theta \in \Omega$. For each estimate a that the statistician might use, their expected loss in this case will be

$$\mathbb{E}[L(\theta, a)|\boldsymbol{x}] = \int_{\Omega} L(\theta, a)\xi(\theta|\boldsymbol{x})d\theta. \tag{4}$$

For each possible value $\boldsymbol{x}$ of $\boldsymbol{X}_n$, let $\delta^*(\boldsymbol{x})$ denote a value of the estimate a for which the expected loss (4) is minimum. Then $\delta^*(\boldsymbol{x})$ is called the **Bayes estimate** of θ. Plugging in $\boldsymbol{X}_n$ instead of $\boldsymbol{x}$, we obtain $\delta^*(\boldsymbol{X}_n)$, which is called the **Bayes estimator** of θ.

Thus, a Bayes estimator is an estimator that is chosen to minimize the *posterior mean* of some measure of how far the estimator is from the parameter.

Corollary 8.5. *Let $\theta \in \Omega \subset \mathbb{R}$. Suppose that the squared error loss function is used and the posterior mean of Θ, i.e., $\mathbb{E}(\Theta|\boldsymbol{X}_n)$, is finite. Then the Bayes estimator of θ is*

$$\delta^*(\boldsymbol{X}_n) = \mathbb{E}(\Theta|\boldsymbol{X}_n).$$

Example 8.6 (Bernoulli distribution with Beta prior)**.** Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with mean $\theta > 0$, where $0 < \theta < 1$ is unknown. Suppose that the prior distribution of Θ is Beta(α, β), where $\alpha, \beta > 0$. Recall that $\Theta|X_1 = x_1, \ldots, X_n = x_n \sim \text{Beta}(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$. Thus,

$$\delta^*(\boldsymbol{X}_n) = \frac{\alpha + \sum_{i=1}^n X_i}{\alpha + \beta + n}.$$

8.4 Sampling from a normal distribution

Theorem 8.7. *Suppose that $X_1, \ldots, X_n$ form a random sample from $N(\theta, \sigma^2)$, where θ is unknown and the value of the variance $\sigma^2 > 0$ is known. Suppose that $\Theta \sim N(\mu_0, v_0^2)$. Then*

$$\Theta|X_1 = x_1, \ldots, X_n = x_n \sim N(\mu_1, v_1^2),$$

where

$$\mu_1 = \frac{\sigma^2\mu_0 + nv_0^2\bar{x}_n}{\sigma^2 + nv_0^2} \qquad \textit{and} \qquad v_1^2 = \frac{\sigma^2 v_0^2}{\sigma^2 + nv_0^2}.$$

Proof. The joint density has the form

$$f_n(\boldsymbol{x}|\theta) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right].$$

The method of completing the squares tells us that

$$\sum_{i=1}^{n}(x_i-\theta)^2 = n(\theta-\bar{x}_n)^2 + \sum_{i=1}^{n}(x_i-\bar{x}_n)^2.$$

Thus, by omitting the factor that involves $x_1,\ldots,x_n$ but does not depend on θ, we may rewrite $f_n(\boldsymbol{x}|\theta)$ as

$$f_n(\boldsymbol{x}|\theta) \propto \exp\left[-\frac{n}{2\sigma^2}(\theta-\bar{x}_n)^2\right].$$

Since the prior density has the form

$$\xi(\theta) \propto \exp\left[-\frac{1}{2v_0^2}(\theta-\mu_0)^2\right],$$

it follows that the posterior p.d.f $\xi(\theta|\boldsymbol{x})$ satisfies

$$\xi(\theta|\boldsymbol{x}) \propto \exp\left[-\frac{n}{2\sigma^2}(\theta-\bar{x}_n)^2 - \frac{1}{2v_0^2}(\theta-\mu_0)^2\right].$$

Completing the squares again establishes the following identity:

$$\frac{n}{\sigma^2}(\theta-\bar{x}_n)^2 + \frac{1}{v_0^2}(\theta-\mu_0)^2 = \frac{1}{v_1^2}(\theta-\mu_1)^2 + \frac{n}{\sigma^2+nv_0^2}(\bar{x}_n-\mu_0)^2.$$

The last term on the right side does not involve on θ. Thus,

$$\xi(\theta|\boldsymbol{x}) \propto \exp\left[-\frac{1}{2v_1^2}(\theta-\mu_1)^2\right].$$

□

Thus, the Bayes estimator (under the squared error loss) in this problem is

$$\delta^*(\boldsymbol{X}_n) = \frac{\sigma^2\mu_0 + nv_0^2\bar{X}_n}{\sigma^2+nv_0^2}.$$