GU4204: Statistical Inference

Bodhisattva Sen
Columbia University

February 27, 2020

Contents

1 Introduction

1.1 Statistical Inference: Motivation

Statistical inference is concerned with making *probabilistic statements* about *random variables* encountered in the analysis of data.

Examples: means, median, variances ...

Example 1.1. *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[1] *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 parameter θ. We may ask the following questions:

1. *Can we* **estimate** *θ from this data? If so, what is a reasonable estimator?*
2. *Can we quantify the uncertainty in the estimation procedure, i.e., can we construct* **confidence interval** *for θ?*

1.2 Recap: Some results from probability

Definition 1 (Sample mean)**.** *Suppose that $X_1, X_2, \ldots, X_n$ are n i.i.d r.v's 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 average) defined as*

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

Lemma 1.2. $\mathbb{E}(\bar{X}_n) = \mu$ *and* $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$.

Proof. Observe that

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

[1] X has an exponential distribution with (failure) rate $\theta > 0$, i.e., $X \sim \mathrm{Exp}(\theta)$, 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 (or expected value) of X is given by $\mathbb{E}(X) = \theta^{-1}$, and the variance of X is $\mathrm{Var}(X) = \theta^{-2}$.

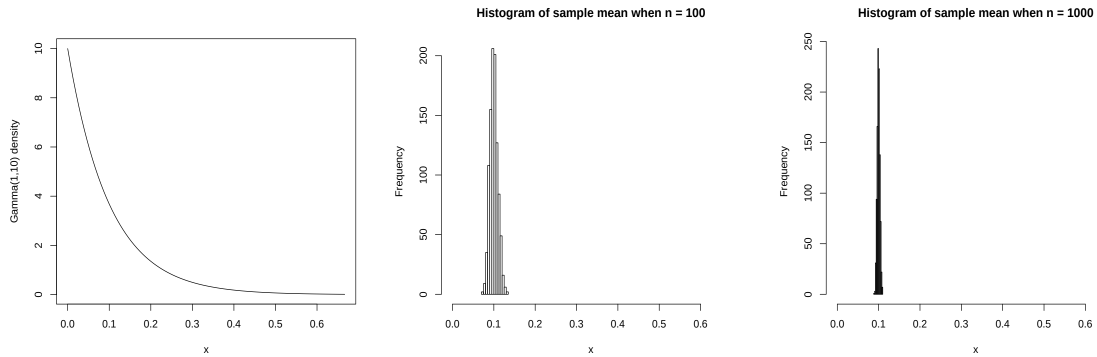


Figure 1: The plots illustrate the convergence (in probability) of the sample mean to the population mean.

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 1.3 (Weak law of large numbers)**.** *Suppose that $X_1, X_2, \ldots, X_n$ are n i.i.d r.v's 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 r.v's 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., $\bar{X}_n \stackrel{\mathbb{P}}{\to} 10^{-1}$ 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 the sequence of r.v's $\{Z_n\}_{n=1}^\infty$ converges to Z **in probability**, *and write*

$$Z_n \stackrel{\mathbb{P}}{\to} Z,$$

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 $\epsilon > 0$,

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

Definition 3 (Convergence in distribution)**.** *We say a sequence of r.v's $\{Z_n\}_{i=1}^n$ 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[2] *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 1.4 (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, for $X_1, X_2, \ldots$ are i.i.d with mean μ and variance $\sigma^2 < \infty$

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

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 is skewed and highly nor-normal (Exp(10)), 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 1.5. *If $Z_n \xrightarrow{\mathbb{P}} b$ and if $g(\cdot)$ is a function that is continuous at b, then*

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

[2] 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.

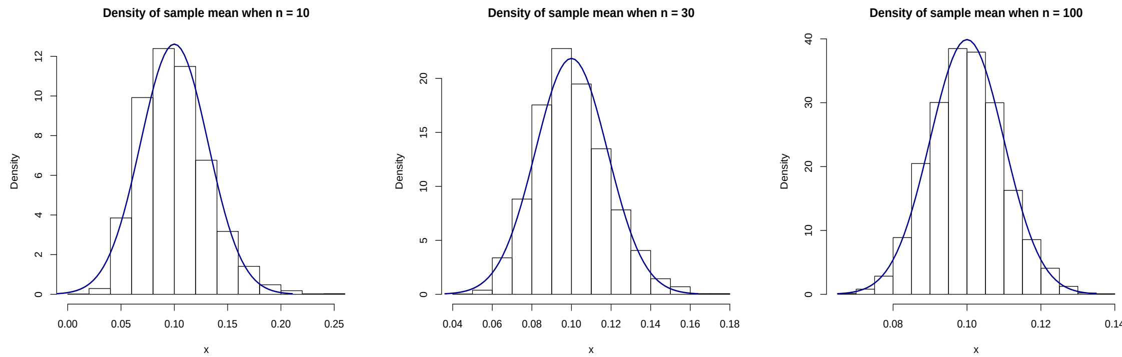


Figure 2: The plots illustrate the convergence (in distribution) of the sample mean to a normal distribution.

Theorem 1.6. *If $Z_n \xrightarrow{d} Z$ and if $g(\cdot)$ is a function that is continuous, then*

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

1.3 Back to Example 1.1

In the first example we have the following results:

- by the LLN, the sample mean $\bar{X}_n$ converges in probability to the expectation $1/\theta$ (failure rate), i.e.,
$$\bar{X}_n \xrightarrow{\mathbb{P}} \frac{1}{\theta};$$
- by the continuous mapping theorem (see Theorem 1.5) $\bar{X}_n^{-1}$ 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 $\mathrm{Var}(X_1) = \theta^{-2}$;
- But how does one find an approximation to the distribution of $\bar{X}_n^{-1}$?

1.4 Delta method

The first thing to note is that if $\{Z_n\}_{i=1}^n$ converges in distribution (or probability) to a constant θ, then $g(Z_n) \xrightarrow{d} g(\theta)$, for any continuous function $g(\cdot)$.

We can also "zoom in" to look at the asymptotic distribution (not just the limit point) of the sequence of r.v's $\{g(Z_n)\}_{i=1}^n$, whenever $g(\cdot)$ is sufficiently smooth.

Theorem 1.7. *Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of r.v's and let Z be a r.v 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

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

$\square$

In other words, limit distributions are passed through functions in a pretty simple way. This is called the **delta method** (I suppose because of the deltas and epsilons involved in this kind of limiting argument), and we'll be using it a lot.

The main application is when we've already proven a CLT for Z_n,

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

in which case

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

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?

1.5 Back to Example 1.1

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 $g(x) = \frac{1}{x}$; $g'(x) = -\frac{1}{x^2}$ (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: Estimation

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). *In a problem of statistical inference, a characteristic or combination of characteristics that* **determine the joint distribution** *for the random variables of interest is called a* **parameter** *of the distribution.*

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.*

Examples:

- 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).

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

Example 2.1. *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 $0 \le p \le 1$.

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**.

Examples:

- 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.*

Definition 9 (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 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 9 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.

Definition 10 (Consistent (in probability) estimator)**.** *A sequence of estimators $\widehat{\theta}_n$ that* **converges in probability** *to the unknown value of the parameter θ being estimated is called a* **consistent sequence of estimators***, i.e., $\widehat{\theta}_n$ is consistent if and only if for every $\epsilon > 0$,*

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

In this Chapter 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 11 (Method of moments estimator)**.** *Assume that $X_1, \ldots, X_n$ form a random sample 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*

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

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 ***sample moments*** *as*

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

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

The usual way of implementing the method of moments is to set up the k equations

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

and then solve for θ.

Example 2.2. *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: *Consider $\mu_1 = \mathbb{E}(X_1)$ and $\mu_2 = \mathbb{E}(X_1^2)$. Clearly, the parameter θ can be expressed as a function of the first two population moments, since*

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

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

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

where we have used the fact that $\hat{\mu}_2 = n^{-1} \sum_{j=1}^{n} X_j^2$.

Example 2.3. *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 and then solve for α and β. In this case, we get

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

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 $\hat{\theta}$.

Result: If M is continuous, then the fact that m_i converges in probability to μ_i, for every $i = 1, \ldots, k$, entails that

$$\hat{\theta} = M(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k) \xrightarrow{\mathbb{P}} M(\mu_1, \mu_2, \ldots, \mu_k) = \theta.$$

Thus MOM estimators are consistent under mild assumptions.

Proof. LLN: the sample moments converge in probability to the population moments $\mu_1(\theta), \ldots, \mu_k(\theta)$.

The generalization of the continuous mapping theorem (Theorem 6.2.5 in the book) 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. $\square$

Remark: 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 2.4. *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,

$$\hat{\theta}_{MOM} = \bar{X},$$

and

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

Example 2.5. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. Poisson(λ), $\lambda > 0$. Find the MOM estimator of λ.*

Solution: *We know that,*

$$\mathbb{E}(X_1) = \mu_1 = \lambda$$

and Var$(X_1) = \mu_2 - \mu_1^2 = \lambda$. Thus

$$\mu_2 = \lambda + \lambda^2.$$

Now, a MOM estimate of λ is clearly given by $\hat{\lambda} = \hat{\mu}_1 = \bar{X}$; thus a MOM estimate of $\mu_2 = \lambda^2 + \lambda$ is given by $\bar{X}^2 + \bar{X}$.

On the other hand, the obvious MOM estimate of $\hat{\mu}_2$ is $\hat{\mu}_2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2$. However these two estimates are not necessarily equal; in other words, it is not necessarily the case that $\bar{X}^2 + \bar{X} = (1/n)\sum_{j=1}^{n} X_j^2$.

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

Example 2.6. *Consider n systems with failure times $X_1, X_2, \dots, X_n$ 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

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

and the estimate based on the second moment is

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

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 $X_1, X_2, \dots, X_n$ 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 12 (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).

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 $0 \le \theta \le 1$. 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 $\theta \ge \max\limits_{i=1,\ldots,n} X_i$ 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,

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

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 (non-uniqueness of MLE)**.** *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. We want to find the MLE of θ. Show that 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 a random 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

3.1.1 Invariance

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(μ, σ^2), then the MLE of μ^2 is $\bar{X}_n^2$.

3.1.2 Consistency

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

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

3.2.1 Newton's Method

Let $f(x)$ be a real-valued function of a real variable, and suppose that we wish to solve the equation

$$f(x) = 0.$$

Let x_1 be an initial guess at the solution.

Newton's method replaces the initial guess with the updated guess

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

The rationale behind the Newton's method is: approximate the curve by a line tangent to the curve passing through the point $(x_1, f(x_1))$. The approximating line crosses the horizontal axis at the revised guess x_1. [Draw a figure!]

Typically, one replaces the initial guess with the revised guess and iterates Newton's method until the results stabilize (see e.g., http://en.wikipedia.org/wiki/Newton's_method).

3.2.2 The EM Algorithm

Read Section 7.6 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. This is one of the most

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,

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

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 $\mathbb{E}_\theta(T_n - \mathbb{E}_\theta(T_n)) = 0$. □

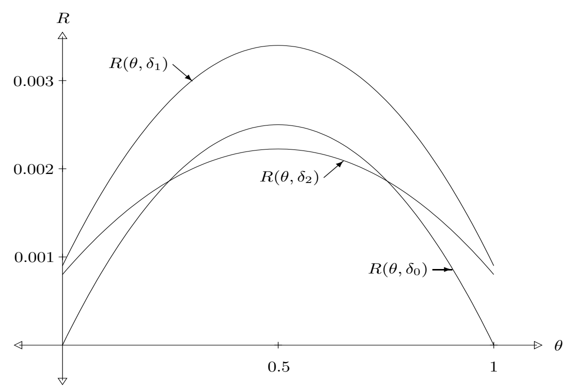


Figure 3: The plot shows the mean squared error for three estimators δ_1, δ_2 and δ_2. Here $R(\theta, \delta_i) = \mathbb{E}_\theta[(\delta_i(X) - \theta)^2]$ where $i = 0, 1, 2$.

The bias measures, on an average, by how much T_n overestimate or underestimate $g(\theta)$. 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.

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

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

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 $X \sim$ 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). Show that*

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

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):*

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

and

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

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.

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

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

Out of these $\hat{\sigma}^2$ is not unbiased for σ^2 but s^2 is. In fact s^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 $g(\theta)$. 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 4.3. *Consider $X_1, X_2, \ldots, X_n$ i.i.d Uniform($[0,\theta]$); $\theta > 0$. Here $\Omega = (0, \infty)$.*

A natural estimate of 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 $(1+n^{-1})X_{(n)}$.

Solution: We can show that

$$\begin{aligned}
\mathrm{MSE}(2\bar{X}_n, \theta) &= \frac{\theta^2}{3n} = \mathrm{Var}(2\bar{X}_n) \\
\mathrm{MSE}((1+n^{-1})X_{(n)}, \theta) &= \frac{\theta^2}{n(n+2)} = \mathrm{Var}((1+n^{-1})X_{(n)}) \\
\mathrm{MSE}(X_{(n)}, \theta) &= \frac{\theta^2}{n(n+2)} \cdot \frac{n^2}{(n+1)^2} + \frac{\theta^2}{(n+1)^2},
\end{aligned}$$

where in the last equality we have two terms — the variance and the squared bias.

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 rigorized at the end of this subsection.

Definition 13 (Sufficient statistic)**.** *Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution indexed by a parameter $\theta \in \Omega$. Let T be a statistic. Suppose that, for every $\theta \in \Omega$ and every possible value t of T, the conditional joint distribution of $X_1, X_2, \ldots, X_n$ given that $T = t$ (at θ) depends only on t but not on θ.*

That is, for each t, the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T = t$ is the same for all θ. Then we say that T is a sufficient statistic for the parameter 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 4.4. *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

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

But,

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

As

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

Also,

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

Hence,

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

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 4.5 (Factorization criterion)**.** *Let $X_1, X_2, \ldots, X_n$ form a random sample 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 $\theta \in \Omega$:

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

- *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.

Exercise: 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

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

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

Exercise: 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

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

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 the 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 and 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 4.6 (Rao-Blackwell theorem)**.** *For every value of* $\theta \in \Omega$*,*

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

The above result is proved in Theorem 7.9.1 of the text book (see deGroot and Schervish, Fourth Edition).

5 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.*

5.1 Prior distribution

Definition 14 (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**.

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

We will assume that if the prior distribution of Θ is continuous, then its p.d.f is called the prior p.d.f of Θ.

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.,

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

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

5.2 Posterior distribution

Definition 15 (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 5.1. *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 5.2 (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, $\Theta|X_1 = x_1, \ldots, X_n = x_n \sim Beta(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$.
$\square$

Example 5.3 (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 $Gamma(\alpha, \beta)$, where $\alpha, \beta > 0$.*

Then 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 common 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 are their observed values $\boldsymbol{x} = (x_1, \ldots, x_n)$, 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 5.4 (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 $\Theta \sim Gamma(\alpha, \beta)$, for known 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, $\Theta|\boldsymbol{X}_n = \boldsymbol{x} \sim Gamma(\alpha + n, \beta + \sum_{i=1}^n x_i)$.

5.3 Bayes Estimators

An estimator of a parameter is some function of the data that we hope is close to the parameter, i.e., $\hat{\theta} \approx \theta$.

Let $X_1, \ldots, X_n$ be data whose joint distribution is indexed by a parameter $\theta \in \Omega$.

Let $\delta(X_1, \ldots, X_n)$ be an estimator of θ.

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 $\xi(\cdot)$ is a prior p.d.f/p.m.f of $\theta \in \Omega$. Consider the problem of estimating θ without being able to observe the data. If the statistician chooses a particular estimate a, then her expected loss will be

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

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

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

$$\mathbb{E}[L(\theta, a)|\boldsymbol{x}] = \int_\Omega L(\theta, a)\xi(\theta|\boldsymbol{x})d\theta. \tag{2}$$

Hence, the statistician should now choose an estimate a for which the above expectation is minimum.

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 (2) is minimum. Then the function $\delta^*(\boldsymbol{X}_n)$ is called the **Bayes estimator** of θ.

Once $\boldsymbol{X}_n = \boldsymbol{x}$ is observed, $\delta^*(\boldsymbol{x})$ is called the Bayes estimate 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 5.5. *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*

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

Example 1: (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 $\Theta|X_1 = x_1, \ldots, X_n = x_n \sim$ Beta$(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$. Thus,

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

5.4 Sampling from a normal distribution

Theorem 5.6. *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

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

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 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,

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

Corollary 5.7. *Let $\theta \in \Omega \subset \mathbb{R}$. Suppose that the absolute error loss function is used. Then the Bayes estimator of θ $\delta^*(\mathbf{X}_n)$ equals the median of the posterior distribution of Θ.*

6 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 $T = \varphi(X_1, \ldots, X_n)$. Assume that $T \sim F_\theta$, where F_θ is the c.d.f of T (possibly dependent on θ).

The distribution of T (with θ fixed) is called the **sampling distribution** of T. Thus, the sampling distribution of T has c.d.f F_θ.

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

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

6.1 The gamma and the χ^2 distributions

6.1.1 The gamma distribution

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

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

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

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

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

The gamma distribution with parameters $\alpha > 0, \lambda > 0$ (denoted by $\text{Gamma}(\alpha, \lambda)$) 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.

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.

Exercise: Let X be a random variable following Gamma(α, λ). Then show that $Y = \lambda X$ (thus X is Y scaled by λ) follows the Gamma$(\alpha, 1)$ distribution. What is the distribution of $c\,X$ for some arbitrary positive constant c? You can use the change of variable theorem in one-dimension to work this out.

Reproductive Property of the gamma distribution:

Let $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i \sim$ Gamma(α_i, λ), for $i = 1, \ldots, n$. Then,

$$S_n := X_1 + X_2 + \ldots + X_n \;\sim\; \text{Gamma}\left(\sum_{i=1}^{n} \alpha_i, \lambda\right).$$

If X follows the Gamma(α, λ) distribution, the mean and variance of X can be explicitly expressed in terms of the parameters:

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

We outline the computation of a general moment $\mathbb{E}(X^k)$, where k is a positive integer. We have,

$$\begin{aligned}
\mathbb{E}(X^k) &= \int_0^\infty x^k \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1}\, dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{-\lambda x} x^{k+\alpha-1}\, dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\lambda^{\alpha+k}} \\
&= \frac{(\alpha+k-1)\cdots(\alpha)\,\Gamma(\alpha)}{\lambda^k\,\Gamma(\alpha)} \\
&= \frac{\Pi_{i=1}^k (\alpha+i-1)}{\lambda^k}\,.
\end{aligned}$$

The formulae for the mean and the variance should follow directly from the above computation. Note that in the above derivation, we have used the fact that

$$\int_0^\infty e^{-\lambda x} x^{k+\alpha-1}\, dx = \frac{\Gamma(\alpha+k)}{\lambda^{\alpha+k}}\,.$$

This is an immediate consequence of the fact that the gamma density with parameters $(\alpha + k, \lambda)$ integrates to 1.

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).

6.1.2 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_1^2).

The χ_1^2 distribution: Let $Z \sim N(0, 1)$. Then the distribution of $W := Z^2$ is called the χ_1^2 distribution, and W itself is called a χ_1^2 random variable.

Exercise: Show that W follows a Gamma$(1/2, 1/2)$ distribution. (You can do this by working out the density function of W from that of Z).

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

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

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

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

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

Thus, it follows that the sum of k i.i.d χ_1^2 random variables is a χ_k^2 random variable.

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 6.1. *If $X \sim \chi_m^2$ then $\mathbb{E}(X) = m$ and $\text{Var}(X) = 2m$.*

Theorem 6.2. *Suppose that $X_1, \ldots, X_k$ are independent and $X_i \sim \chi_{m_i}^2$ then the sum*

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

6.2 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.

You could think of the X_i's for example as a set of randomly sampled SAT scores from the entire population of SAT scores. Then μ is the average SAT score of the entire population and σ^2 is the variance of SAT scores in the entire population. We are interested in estimating μ and σ^2 based on the data. Note that SAT scores are actually discrete in nature — $N(\mu, \sigma^2)$ provides a good approximation to the actual population distribution. In other words, $N(\mu, \sigma^2)$ is the **model** that we use for the SAT scores.

In statistics as in any other science, models are meant to provide insightful approximations to the true underlying nature of reality.

Natural estimates of the mean and the variance are given by:

$$\hat{\mu} = \overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \,.$$

These are the *sample mean* and *sample variance* (biased version). In what follows, we will use a slightly different estimator of σ^2 than the one proposed above. We will use

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

One reason for using s^2 is that it has a natural interpretation as the multiple of a χ^2 random variable; further s^2 is an *unbiased estimator* of σ^2 whereas $\hat{\sigma}^2$ is not, i.e.,

$$\mathbb{E}(s^2) = \sigma^2 \ \text{ but } \ \mathbb{E}(\hat{\sigma}^2) \neq \sigma^2 \,.$$

For the sake of notational simplicity we will let S^2 denote the *residual sum of squares about the mean*, i.e., $S^2 := \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$.

Here is an interesting (and fairly profound) proposition.

Proposition 6.3. *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 are independent random variables. Moreover, when F is $N(\mu, \sigma^2)$, then*

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

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 are independent.

Proof. To this end, define new random variables $Y_1, Y_2, \ldots, Y_n$ where for each i,

$$Y_i = (X_i - \mu)/\sigma.$$

These are the *standardized versions* of the X_i's and are i.i.d. $N(0,1)$ random variables. Now, note that:

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

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 $(Y_1, Y_2, \ldots, Y_n)^\top$ and let P be an $n \times n$ orthogonal matrix with the first row of P (which has length n) being $(1/\sqrt{n}, 1/\sqrt{n}, \ldots, 1/\sqrt{n})$.

Recall that an orthogonal matrix satisfies

$$P^\top P = PP^\top = I$$

where I is the identity matrix.

Using standard linear algebra techniques it can be shown that such a P can always be constructed. Now define a new random vector

$$W = P\boldsymbol{Y}.$$

Then it can be established that the random vector $\boldsymbol{W} = (W_1, W_2, \ldots, W_n)^\top$ has the same distribution as $(Y_1, Y_2, \ldots, Y_n)^\top$; in other words, $W_1, W_2, \ldots, W_n$ are i.i.d $N(0,1)$ random variables.

Theorem 6.4. *Suppose that $Z_1, \ldots, Z_n$ are i.i.d $N(0,1)$. Suppose that A is an orthogonal matrix and*

$$\boldsymbol{V} = A\boldsymbol{Z}.$$

Then the random variables $V_1, \ldots, V_n$ are i.i.d $N(0,1)$. Also, $\sum_{i=1}^n V_i^2 = \sum_{i=1}^n Z_i^2$.

Note that

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

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

$$W_1 = Y_1/\sqrt{n} + Y_2/\sqrt{n} + \ldots + Y_n/\sqrt{n} = \sqrt{n}\,\overline{Y}\,.$$

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

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

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

Note that $\overline{Y} \sim N(0, 1/n)$. Deduce that $\overline{X}$ follows $N(\mu, \sigma^2/n)$. Since $\sum_{i=1}^n (Y_i - \overline{Y})^2 = W_2^2 + W_3^2 + \ldots + W_n^2$, it follows that

$$\frac{S^2}{\sigma^2} = \sum_{i=1}^n (Y_i - \overline{Y})^2 \sim \chi^2_{n-1}\,.$$

Thus,

$$s^2 = \frac{S^2}{n-1} \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}. \tag{3}$$

In the case $n = 2$, it is easy to check the details of the transformation leading from $\boldsymbol{Y}$ to $\boldsymbol{W}$. Set $\boldsymbol{W} = P\boldsymbol{Y}$ with

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

Thus $W_1 = (Y_1 + Y_2)/\sqrt{2}$ and $W_2 = (Y_1 - Y_2)/\sqrt{2}$.

Exercise: Use the change of variable theorem to deduce that W_1 and W_2 are i.i.d $N(0, 1)$.

Proof of Theorem 6.4: The joint p.d.f of $\boldsymbol{Z} = (Z_1, \ldots Z_n)$ is

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

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

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

Let $\boldsymbol{z} = A^{-1}\boldsymbol{v}$. Since A is orthogonal, $|\det A| = 1$ and $\boldsymbol{v}^\top\boldsymbol{v} = \sum_{i=1}^n v_i^2 = \boldsymbol{z}^\top\boldsymbol{z} = \sum_{i=1}^n z_i^2$. So,

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

Thus, $\boldsymbol{V}$ has the same joint p.d.f as $\boldsymbol{Z}$.

6.3 The t-distribution

Definition: Let $Z \sim N(0,1)$ and let $V \sim \chi^2_n$, 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).

Exercise: Let X be a random variable that is distributed symmetrically about 0, i.e., X and $-X$ have the same distribution function (and hence the same density function). If f denotes the density, show that it is an even function, i.e. $f(x) = f(-x)$ for all x.

Conversely, if the random variable X has a density function f that is even, then it is symmetrically distributed about 0, i.e $X =_d -X$.

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.

From 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; hence the quantiles of the t-distribution are well approximated by the quantiles of the normal 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 $K_1 + K_2 + \ldots + K_n$ where K_i's are i.i.d χ^2_1 random variables. But by the WLLN we know that

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

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

Theorem 6.5. *Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Let $\overline{X}_n$ denote the sample mean, and define $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$. Then*

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{s} \sim t_{n-1}.$$

7 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, $\theta \in \Omega \subset \mathbb{R}^k$ (that depends on a parameter θ).

Suppose that we want to estimate $g(\theta)$, a real-valued function of θ.

Let $A \leq B$ be two statistics that have 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 level (coefficient) $(1 - \alpha)$.

If the inequality "$\geq 1 - \alpha$" is an equality for all θ, the the CI is called *exact*.

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 $g(\theta) = \mu$.

Step 1: We want to construct $\Psi(X_1, X_2, \ldots, X_n, \mu)$ such that the distribution of this object is known to us.

How do we proceed here?

The usual way is to find some decent estimator of μ and combine it along with μ in some way to get a "pivot", i.e., a random variable whose distribution does not depend on θ.

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,

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

for every value of θ.

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

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

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

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

Thus our level $(1-\alpha)$ CI for μ is given by

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

Often a standard method of constructing CIs is the following *method of pivots* which we describe below.

(1) Construct a function Ψ using 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 is called a *pivot*.

(2) Let G denote the distribution function of the 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'th quantile of G. Thus,

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

(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) Vary θ across its domain and choose your level $1-\alpha$ confidence interval (set) as the set of all $g(\theta)$ such that the two inequalities in the above display are simultaneously satisfied.

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

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

will not work smoothly here. This certainly has a known ($N(0,1)$) distribution but involves the *nuisance parameter* σ making it difficult get a CI for μ directly.

However, one can replace σ by s, where s^2 is the natural estimate of σ^2 introduced before. So, set:

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

This only depends on the data and $g(\theta) = \mu$. We claim that this is indeed a pivot.

To see this write

$$\frac{\sqrt{n}(\overline{X} - \mu)}{s} = \frac{\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}}{\sqrt{s^2/\sigma^2}}\,.$$

The numerator on the extreme right of the above display follows $N(0,1)$ and the denominator is independent of the numerator and is the square root of a χ^2_{n-1} random variable over its degrees of freedom (from display (3)).

It follows from definition that $\Psi(\boldsymbol{X}_n, \mu) \sim t_{n-1}$ distribution.

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,

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

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

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

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

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.

Asymptotic pivots using the central limit theorem: 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

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

We are interested in constructing an approximate level $(1-\alpha)$ CI for μ.

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

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

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

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

As before, this translates to

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

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. Since we are dealing with large sample sizes, s is with very high probability close to σ and the interval

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

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

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

Thus

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

We want to find a level $(1-\alpha)$ CI (approximate) for θ.

Note that both mean and variance are unknown.

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.

See http://www.rossmanchance.com/applets/ConfSim.html and http://www.ruf.rice.edu/~lane/stat_sim/conf_interval/ for illustrations of confidence intervals.

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).

8 The (Cramer-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.

For most of the sequel, for notational simplicity, we will assume that the parameter space $\Omega \subset \mathbb{R}$. We define the **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 8.1 (Cramer-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 **Cramer-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 8.1: Let ρ_θ denote the correlation between $T(\boldsymbol{X}_n)$ and $\dot{\ell}_n(\boldsymbol{X}_n, \theta)$. Then $\rho_\theta^2 \leq 1$ which implies that

$$\mathrm{Cov}_\theta^2\left(T(\boldsymbol{X}_n), \dot{\ell}_n(\boldsymbol{X}_n, \theta)\right) \leq \mathrm{Var}_\theta(T(\boldsymbol{X}_n)) \cdot \mathrm{Var}_\theta(\dot{\ell}_n(\boldsymbol{X}_n, \theta)). \tag{4}$$

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,

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

since the $\dot{\ell}(X_i, \theta)$'s are i.i.d. Thus, we have $\mathbb{E}_\theta\left(\dot{\ell}(X_1, \theta)\right) = 0$. 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 *information* based on n observations. 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 (4) we get,

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

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

There is an alternative expression for the 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:

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

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 now look at some applications of the Cramer-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 information about θ based on n observations 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 Cramer-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 best unbiased estimator (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 information for V based on n observations 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 Cramer-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 best unbiased estimator of V in this model.

9 Large Sample Properties of the MLE

In this section 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 (A1, A2 and A3) 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 Cramer-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 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 $\overline{X}$ 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 best unbiased estimator 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 $\overline{X}$ and is unbiased for θ^2 is an MVUE (and indeed there is one such).

Verify that,

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

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).

Exercise: Verify, in the Bernoulli example above in this section, that

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

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,

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

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.

Proposition 9.1. *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 9.2 (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 9.1 with Proposition 9.2 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,

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

it follows that

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

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

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

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

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

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 $I(\hat{\theta})$ can be expected to be close to $I(\theta)$ with high probability and hence the following is also valid:

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

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

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

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 level $1-\alpha$ CI is then given by,

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

Exercise: Find explicitly

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

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:

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

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).

10 Hypothesis Testing

10.1 Principles of Hypothesis Testing

We are given data (say $X_1, \ldots, X_n$ i.i.d P_θ) from a model that is parametrized by θ. We consider a statistical problem involving θ whose value is unknown but must lie in a certain space Ω. We consider the testing problem

$$H_0 : \theta \in \Omega_0 \qquad \text{versus} \qquad H_1 : \theta \in \Omega_1, \tag{5}$$

where $\Omega_0 \cap \Omega_1 = \emptyset$ and $\Omega_0 \cup \Omega_1 = \Omega$.

Here the hypothesis H_0 is called the **null hypothesis** and H_1 is called the **alternative hypothesis**.

Question: Is there enough evidence in the data against the null hypothesis (in which case we reject it) or should we continue to stick to it?

Such questions arise very naturally in many different fields of application.

Definition 16 (One-sided and two-sided hypotheses)**.** *Let θ be a one-dimensional parameter.*

- *one-sided hypotheses*
  - $H_0 : \theta \leq \theta_0$*, and* $H_1 : \theta > \theta_0$*, or*
  - $H_0 : \theta \geq \theta_0$*, and* $H_1 : \theta < \theta_0$
- *two-sided hypotheses* $H_0 : \theta = \theta_0$*, and* $H_1 : \theta \neq \theta_0$*.*

H_0 is *simple* if Ω_0 is a set with only one point; otherwise, H_0 is *composite*.

Testing for a normal mean: Suppose that $X_1, X_2, \ldots, X_n$ is a sample from a $N(\mu, \sigma^2)$ distribution and let, initially, σ^2 be known.

We want to test the *null hypothesis* $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$.

Example: For concreteness, $X_1, X_2, \ldots, X_n$ could be the heights of n individuals in some tribal population. The distribution of heights in a (homogeneous) population is usually normal, so that a $N(\mu, \sigma^2)$ model is appropriate. If we have some a-priori reason to believe that the average height in this population is around 60 inches, we could postulate a null hypothesis of the form $H_0 : \mu = \mu_0 \equiv 60$; the alternative hypothesis is $H_1 : \mu \neq 60$.

10.2 Critical regions and test statistics

Consider a problem in which we wish to test the following hypotheses:

$$H_0 : \theta \in \Omega_0, \qquad \text{and} \qquad H_1 : \theta \in \Omega_1. \tag{6}$$

Question: How do we do the test?

The statistician must decide, after observing data, which of the hypothesis H_0 or H_1 appears to be true.

A procedure for deciding which hypothesis to choose is called a **test procedure** of simply a **test**. We will denote a test by δ.

Suppose we can observe a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ drawn from a distribution that involves the unknown parameter θ, e.g., suppose that $X_1, \ldots, X_n$ are i.i.d P_θ, $\theta \in \Omega$.

Let S denote the set of all possible values of the n-dimensional random vector $\mathbf{X}$.

We specify a test procedure by partitioning S into two subsets: $S = S_0 \cup S_1$, where

- the **rejection region** (sometimes also called the **critical region**) S_1 contains the values of $\mathbf{X}$ for which we will reject H_0, and
- the other subset S_0 (usually called the **acceptance region**) contains the values of $\mathbf{X}$ for which we will not reject H_0.

A test procedure is determined by specifying the critical region S_1 of the test.

In most hypothesis-testing problems, the critical region is defined in terms of a statistic, $T = \varphi(\mathbf{X})$.

Definition 17 (Test statistic/rejection region)**.** *Let* $\mathbf{X}$ *be a random sample from a distribution that depends on a parameter* θ*. Let* $T = \varphi(\mathbf{X})$ *be a statistic, and let* R *be a subset of the real line. Suppose that a test procedure is of the form:*

$$\textit{reject } H_0 \quad \textit{if} \quad T \in R.$$

Then we call T *a* **test statistic***, and we call* R *the rejection region of the test:*

$$S_1 = \{\mathbf{x} : \varphi(\mathbf{x}) \in R\}.$$

Typically, the rejection region for a test based on a test statistic T will be some fixed interval or the complement of some fixed interval.

If the test rejects H_0 when $T \geq c$, the rejection region is the interval $[c, \infty)$. Indeed, most of the tests can be written in the form:

$$\text{reject } H_0 \quad \text{if} \quad T \geq c.$$

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d $N(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ is unknown, and $\sigma > 0$ is assumed *known*.

Suppose that we want to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

Some of these procedures can be justified using formal paradigms. Under the null hypothesis the X_i's are i.i.d $N(\mu_0, \sigma^2)$ and the sample mean $\overline{X}$ follows $N(\mu_0, \sigma^2/n)$.

Thus, it is reasonable to take $T = \varphi(\mathbf{X}) = |\overline{X} - \mu_0|$.

Large deviations of the observed value of $\overline{X}$ from μ_0 would lead us to suspect that the null hypothesis might not be true.

Thus, a reasonable test can be to reject H_0 if $T = |\overline{X} - \mu_0| > c$, for some "large" constant c.

But how large is large? We will discuss this soon...

Associated with the test procedure δ are two different kinds of error that we can commit. These are called *Type 1 error* and *Type 2 error* (Draw the 2×2 table!).

Decision / State	Fail to reject H_0	Reject H_0
H_0 True	Correct	Type 1 error
H_1 True	Type 2 error	Correct

Table 1: Hypothesis test.

Type 1 error occurs if we reject the null hypothesis when actually H_0 is true.

Type 2 error occurs if we do not reject the null hypothesis when actually H_0 is false.

10.3 Power function and types of error

Let δ be a test procedure. If S_1 denotes the critical region of δ, then the **power function** of the test δ, $\pi(\theta|\delta)$, is defined by the relation

$$\pi(\theta|\delta) = \mathbb{P}_\theta(\mathbf{X} \in S_1) \quad \text{for} \quad \theta \in \Omega.$$

Thus, the power function $\pi(\theta|\delta)$ specifies for each possible value of θ, the *probability that δ will reject H_0*. If δ is described in terms of a test statistic T and rejection region R, the power function is

$$\pi(\theta|\delta) = \mathbb{P}_\theta(T \in R) \quad \text{for} \quad \theta \in \Omega.$$

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d Uniform$(0, \theta)$, where $\theta > 0$ is unknown.

Suppose that we are interested in the following hypotheses:

$$H_0 : 3 \leq \theta \leq 4, \qquad \text{versus} \qquad H_1 : \theta < 3, \text{ or } \theta > 4.$$

We know that the MLE of θ is $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

Note that $X_{(n)} < \theta$.

Suppose that we use a test δ given by the critical region

$$S_1 = \{\boldsymbol{x} \in \mathbb{R}^n : x_{(n)} \leq 2.9 \text{ or } x_{(n)} \geq 4\}.$$

Question: Find the power function $\pi(\theta|\delta)$?

Solution: The power function of δ is

$$\pi(\theta|\delta) = \mathbb{P}_\theta(X_{(n)} \leq 2.9 \text{ or } X_{(n)} > 4) = \mathbb{P}_\theta(X_{(n)} \leq 2.9) + \mathbb{P}_\theta(X_{(n)} \geq 4).$$

Case (i): Suppose that $\theta \leq 2.9$. Then

$$\pi(\theta|\delta) = \mathbb{P}_\theta(X_{(n)} \leq 2.9) = 1.$$

Case (ii): Suppose that $2.9 < \theta < 4$. Then

$$\pi(\theta|\delta) = \mathbb{P}_\theta(X_{(n)} \leq 2.9) = \left(\frac{2.9}{\theta}\right)^n.$$

Case (iii): Suppose that $\theta > 4$. Then

$$\pi(\theta|\delta) = \left(\frac{2.9}{\theta}\right)^n + \left[1 - \left(\frac{4}{\theta}\right)^n\right].$$

The ideal power function would be one for which

- $\pi(\theta|\delta) = 0$ for every value of $\theta \in \Omega_0$, and
- $\pi(\theta|\delta) = 1$ for every value of $\theta \in \Omega_1$.

If the power function of a test δ actually had these values, then regardless of the actual value of θ, δ would lead to the correct decision with probability 1.

In a practical problem, however, there would seldom exist any test procedure having this ideal power function.

- Type-I error: rejecting H_0 given that $\theta \in \Omega_0$. It occurs with probability $\pi(\theta|\delta)$.
- Type-II error: not rejecting H_0 given that $\theta \in \Omega_1$. It occurs with probability $1 - \pi(\theta|\delta)$.

Ideal goals: we would like the power function $\pi(\theta|\delta)$ to be **low** for values of $\theta \in \Omega_0$, and **high** for $\theta \in \Omega_1$.

Generally, these two goals work against each other. That is, if we choose δ to make $\pi(\theta|\delta)$ small for $\theta \in \Omega_0$, we will usually find that $\pi(\theta|\delta)$ is small for $\theta \in \Omega_1$ as well.

Examples:

- The test procedure δ_0 that never rejects H_0, regardless of what data are observed, will have $\pi(\theta|\delta_0) = 0$ for all $\theta \in \Omega_0$. However, for this procedure $\pi(\theta|\delta_0) = 0$ for all $\theta \in \Omega_1$ as well.
- Similarly, the test δ_1 that always rejects H_0 will have $\pi(\theta|\delta_1) = 1$ for all $\theta \in \Omega_1$, but it will also have $\pi(\theta|\delta_1) = 1$ for all $\theta \in \Omega_0$.

Hence, there is a need to strike an appropriate balance between the two goals of

low power in Ω_0 and *high power in* Ω_1.

1. The most popular method for striking a balance between the two goals is to choose a number $\alpha_0 \in (0, 1)$ and require that

   $$\pi(\theta|\delta) \leq \alpha_0, \quad \text{for all} \quad \theta \in \Omega_0. \tag{7}$$

   This α_0 will usually be a small positive fraction (historically .05 or .01) and will be called the **level of significance** or simply *level*.

   Then, among all tests that satisfy (7), the statistician seeks a test whose power function is as high as can be obtained for $\theta \in \Omega_1$.

2. Another method of balancing the probabilities of type I and type II errors is to minimize a linear combination of the different probabilities of error.

10.4 Significance level

Definition 18 (level/size)**.** *(of the test)*

- *A test that satisfies (7) is called a level α_0 test, and we say that the test has level of significance α_0.*
- *The size $\alpha(\delta)$ of a test δ is defined as follows:*

$$\alpha(\delta) = \sup_{\theta \in \Omega_0} \pi(\theta|\delta).$$

It follows from Definition 18 that:

- A test δ is a level α_0 test iff $\alpha(\delta) \leq \alpha_0$.
- If the null hypothesis is simple (that is, $H_0 : \theta = \theta_0$), then $\alpha(\delta) = \pi(\theta_0|\delta)$.

Making a test have a specific significance level

Suppose that we wish to test the hypotheses

$$H_0 : \theta \in \Omega_0, \qquad \text{versus} \qquad H_1 : \theta \in \Omega_1.$$

Let T be a test statistic, and suppose that our test will reject the null hypothesis if $T \geq c$, for some constant c. Suppose also that we desire our test to have the level of significance α_0. The power function of our test is $\pi(\theta|\delta) = \mathbb{P}_\theta(T \geq c)$, and we want that

$$\sup_{\theta \in \Omega_0} \mathbb{P}_\theta(T \geq c) \leq \alpha_0. \tag{8}$$

Remarks:

1. It is clear that the power function, and hence the left side of (8), are non-increasing functions of c.

   Hence, (8) will be satisfied for large values of c, but not for small values.

   If T has a continuous distribution, then it is usually simple to find an appropriate c.

2. Whenever we choose a test procedure, we should also examine the power function. If one has made a good choice, then the power function should generally be larger for $\theta \in \Omega_1$ than for $\theta \in \Omega_0$.

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d $N(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ is unknown, and $\sigma > 0$ is assumed *known*. We want to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

Suppose that the null hypothesis H_0 is true.

If the variance of the sample mean is, say, 100, a deviation of $\overline{X}$ from μ_0 by 15 is not really unusual.

On the other hand if the variance is 10, then a deviation of the sample mean from μ_0 by 15 is really sensational.

Thus the quantity $|\overline{X} - \mu_0|$ in itself is not sufficient to formulate a decision regarding rejection of the null hypothesis.

We need to adjust for the underlying variance. This is done by computing the so-called z-statistic,

$$Z := \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \equiv \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}$$

and rejecting the null hypothesis for large absolute values of this statistic.

Under the null hypothesis Z follows $N(0,1)$; thus an absolute Z-value of 3.5 is quite unlikely. Therefore if we observe an absolute Z-value of 3.5 we might rule in favor of the alternative hypothesis.

You can see now that we need a threshold value, or in other words a critical point such that if the Z-value exceeds that point we reject. Our test procedure δ then looks like,

$$\text{reject } H_0 \quad \text{ if } \quad \left| \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \right| > c_{n,\alpha_0}$$

where c_{n,α_0} is the *critical value* and will depend on α_0 which is the tolerance for the Type 1 error, i.e., the level that we set beforehand.

The quantity c_{n,α_0} is determined using the relation

$$\mathbb{P}_{\mu_0}\left(\left| \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \right| > c_{n,\alpha_0} \right) = \alpha_0.$$

Straightforward algebra then yields that

$$P_{\mu_0}\left(-c_{n,\alpha_0} \frac{\sigma}{\sqrt{n}} \le \overline{X} - \mu_0 \le c_{n,\alpha_0} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha_0,$$

whence we can choose $c_{n,\alpha_0} = z_{\alpha_0/2}$, the $\frac{\alpha_0}{2}$-th quantile of the $N(0,1)$ distribution.

The acceptance region $\mathcal{A}$ (or S_0) for the null hypothesis is therefore

$$\mathcal{A} = \left\{ \boldsymbol{x} = (x_1, x_2, \ldots, x_n) : \mu_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha_0/2} \le \overline{x} \le \mu_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha_0/2} \right\}.$$

So we accept whenever $\overline{X}$ lies in a certain window of μ_0, the postulated value under the null, and reject otherwise which is in accordance with intuition.

The length of the window is determined by the tolerance level α_0, the underlying variance σ^2 and of course the sample size n.

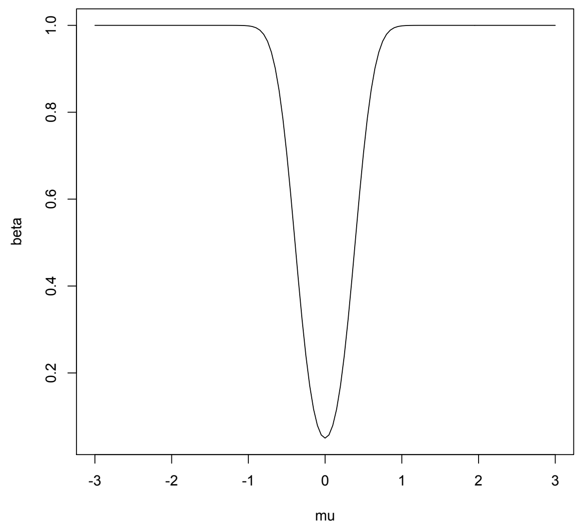


Figure 4: The power function $\pi(\mu|\delta)$ for $\mu_0 = 0$, $\sigma = 1$ and $n = 25$.

10.5 P-value

The p**-value** is the smallest level α_0 such that we would reject H_0 at level α_0 with the observed data.

For this reason, the p-value is also called the *observed level of significance*.

Example: If the observed value of Z was 2.78, and that the corresponding p-value = 0.0054. It is then said that the observed value of Z is just significant at the level of significance 0.0054.

Advantages:

1. No need to select beforehand an arbitrary level of significance α_0 at which to carry out the test.
2. When we learn that the observed value of Z was just significant at the level of significance 0.0054, we immediately know that H_0 would be rejected for every larger value of α_0 and would not be rejected for any smaller value.

10.6 Testing simple hypotheses: optimal tests

Let the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ come from a distribution for which the joint p.m.f/p.d.f is either $f_0(\mathbf{x})$ or $f_1(\mathbf{x})$. Let $\Omega = \{\theta_0, \theta_1\}$. Then,

- $\theta = \theta_0$ stands for the case in which the data have p.m.f/p.d.f $f_0(\mathbf{x})$,

- $\theta = \theta_1$ stands for the case in which the data have p.m.f/p.d.f $f_1(\mathbf{x})$.

We are then interested in testing the following simple hypotheses:

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

In this case, we have special notation for the probabilities of type I and type II errors:

$$\begin{aligned}\alpha(\delta) &= \mathbb{P}_{\theta_0}(\text{Rejecting } H_0),\\ \beta(\delta) &= \mathbb{P}_{\theta_1}(\text{Not rejecting } H_0).\end{aligned}$$

10.6.1 Minimizing the $\mathbb{P}$(Type-II error)

Suppose that the probability $\alpha(\delta)$ of an error of type I is not permitted to be greater than a specified level of significance, and it is desired to find a procedure δ for which $\beta(\delta)$ will be a minimum.

Theorem 10.1 (Neyman-Pearson lemma)**.** *Suppose that δ' is a test procedure that has the following form for some constant $k > 0$:*

- *H_0 is not rejected if $f_1(\mathbf{x}) < kf_0(\mathbf{x})$,*
- *H_0 is rejected if $f_1(\mathbf{x}) > kf_0(\mathbf{x})$, and*
- *H_0 can be either rejected or not if $f_1(\mathbf{x}) = kf_0(\mathbf{x})$.*

Let δ be another test procedure. Then,

$$\begin{aligned}&\textit{if} \quad \alpha(\delta) \le \alpha(\delta'), \quad \textit{then it follows that} \quad \beta(\delta) \ge \beta(\delta')\\ &\textit{if} \quad \alpha(\delta) < \alpha(\delta'), \quad \textit{then it follows that} \quad \beta(\delta) > \beta(\delta').\end{aligned}$$

Example: Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ is a random sample from the normal distribution with unknown mean θ and known variance 1. We are interested in testing:

$$H_0 : \theta = 0 \qquad \text{versus} \qquad H_1 : \theta = 1.$$

We want to find a test procedure for which $\beta(\delta)$ will be a minimum among all test procedures for which $\alpha(\delta) \le 0.05$.

We have,

$$f_0(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \qquad \text{and} \qquad f_1(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} (x_i - 1)^2\right].$$

After some algebra, the likelihood ratio $f_1(\mathbf{x})/f_0(\mathbf{x})$ can be written in the form

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \exp\left[n\left(\bar{x} - \frac{1}{2}\right)\right].$$

Thus, rejecting H_0 when the likelihood ratio is greater than a specified positive constant k is equivalent to rejecting H_0 when the sample mean $\bar{X}$ is greater than $k' := 1/2 + \log k/n$, another constant. Thus, we want to find, k' such that

$$\mathbb{P}_0(\bar{X} > k') = 0.05.$$

Now,

$$\begin{aligned} & \mathbb{P}_0(\bar{X} > k') = \mathbb{P}_0(\sqrt{n}\bar{X} > \sqrt{n}k') = \mathbb{P}_0(Z > \sqrt{n}k') = 0.05 \\ \Rightarrow \quad & \sqrt{n}k' = 1.645. \end{aligned}$$

10.7 Uniformly most powerful (UMP) tests

Let the null and/or alternative hypothesis be composite

- $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$, or
- $H_0 : \theta \geq \theta_0$ and $H_1 : \theta < \theta_0$

We suppose that Ω_0 and Ω_1 are disjoint subsets of Ω, and the hypotheses to be tested are

$$H_0 : \theta \in \Omega_0 \qquad \text{versus} \qquad H_1 : \theta \in \Omega_1. \tag{9}$$

- The subset Ω_1 contains at least two distinct values of θ, in which case the alternative hypothesis H_1 is composite.
- The null hypothesis H_0 may be either simple or composite.

We consider *only* procedures in which

$$\mathbb{P}_\theta(\text{Rejecting } H_0) \leq \alpha_0 \qquad \forall\, \theta \in \Omega_0.$$

that is

$$\pi(\theta|\delta) \leq \alpha_0 \qquad \forall\, \theta \in \Omega_0$$

or

$$\alpha(\delta) \leq \alpha_0. \tag{10}$$

Finally, among all test procedures that satisfy the requirement (10), we want to find one such that

- the probability of type II error is as small as possible for every $\theta \in \Omega_1$, or
- the value of $\pi(\theta|\delta)$ is as large as possible for every value of $\theta \in \Omega_1$.

There might be no single test procedure δ that maximizes the power function $\pi(\theta|\delta)$ simultaneously for every value of $\theta \in \Omega_1$.

In some problems, however, there will exist a test procedure that satisfies this criterion. Such a procedure, when it exists, is called a UMP test.

Definition 19 (Uniformly most powerful (UMP) test)**.** *A test procedure δ^* is a uniformly most powerful (UMP) test of the hypotheses (9) at the level of significance α_0 if*

$$\alpha(\delta^*) \le \alpha_0$$

and, for every other test procedure δ such that $\alpha(\delta) \le \alpha_0$, it is true that

$$\pi(\theta|\delta) \le \pi(\theta|\delta^*)$$

for every value of $\theta \in \Omega_1$.

Usually no test will uniformly most powerful against ALL alternatives, except in the special case of "monotone likelihood ratio" (MLR).

Example: Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which the mean μ (unknown) and the variance σ^2 (known). Consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Even in this simple example, there is no UMP test.

10.8 The t-test

10.8.1 Testing hypotheses about the mean with unknown variance

- Problem: testing hypotheses about the **mean** of a normal distribution when both the mean and the variance are unknown.
- The random variables $X_1, \ldots, X_n$ form a random sample from a normal distribution for which the mean μ and the variance σ^2 are unknown.
- The parameter space Ω in this problem comprises every two-dimensional vector (μ, σ^2), where $-\infty < \mu < \infty$ and $\sigma^2 > 0$.
- $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$
- Define

$$U_n = \frac{\bar{X}_n - \mu_0}{s_n/\sqrt{n}}, \tag{11}$$

where $s_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2}$.

- We reject H_0 if

$$|U_n| \geq T_{n-1}^{-1}\left(1 - \frac{\alpha_0}{2}\right),$$

  the $(1-\alpha_0/2)$-quantile of the t-distribution with $n-1$ degrees of freedom and U_n is defined in (11).

- p-values for t-tests: The p-value from the observed data and a specific test is the smallest α_0 such that we would reject the null hypothesis at level of significance α_0.

  Let u be the observed value of the statistic U_n. Thus the p-value of the test is

$$\mathbb{P}(|U_n| > |u|),$$

  where $U_n \sim T_{n-1}$, under H_0.

- The p-value is $2[1 - T_{n-1}(|u|)]$, where u be the observed value of the statistic U_n.

The Complete power function

Before we study the case when $\sigma > 0$ is unknown, let us go back to the case when σ is known.

Our test δ is "reject H_0 if $\left|\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma}\right| > z_{\alpha/2}$".

Thus we have,

$$\pi(\mu|\delta) = \mathbb{P}_\mu\left(\left|\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma}\right| > z_{\alpha/2}\right),$$

which is just,

$$\mathbb{P}_\mu\left(\left|\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} + \frac{\sqrt{n}(\mu-\mu_0)}{\sigma}\right| > z_{\alpha/2}\right).$$

But when μ is the population mean, $\sqrt{n}(\overline{X}-\mu)/\sigma$ is $N(0,1)$. If Z denotes a $N(0,1)$ variable then,

$$\begin{aligned}
\pi(\mu|\delta) &= \mathbb{P}_\mu\left(\left|Z + \frac{\sqrt{n}(\mu-\mu_0)}{\sigma}\right| > z_{\alpha/2}\right) \\
&= \mathbb{P}_\mu\left(Z + \frac{\sqrt{n}(\mu-\mu_0)}{\sigma} > z_{\alpha/2}\right) + \mathbb{P}\left(Z + \frac{\sqrt{n}(\mu-\mu_0)}{\sigma} < -z_{\alpha/2}\right) \\
&= 1 - \Phi\left(z_{\alpha/2} - \frac{\sqrt{n}(\mu-\mu_0)}{\sigma}\right) + \Phi\left(-z_{\alpha/2} - \frac{\sqrt{n}(\mu-\mu_0)}{\sigma}\right) \\
&= \Phi\left(-z_{\alpha/2} + \frac{\sqrt{n}(\mu-\mu_0)}{\sigma}\right) + \Phi\left(-z_{\alpha/2} - \frac{\sqrt{n}(\mu-\mu_0)}{\sigma}\right).
\end{aligned}$$

Check from the above calculations that $\pi(\mu_0|\delta) = \alpha$, the level of the test δ.

Notice that the test function δ depends on the value μ_0 under the null but it does not depend on any value in the alternative.

The power increases as the true value μ deviates further from μ_0.

It is easy to check that $\pi(\mu|\delta)$ diverges to 1 as μ diverges to ∞ or $-\infty$.

Moreover the power function is symmetric around μ_0. In other words, $\pi(\mu_0 + \Delta|\delta) = \pi(\mu_0 - \Delta|\delta)$ where $\Delta > 0$.

To see this, note that

$$\pi(\mu_0 + \Delta|\delta) = \Phi\left(-z_{\alpha/2} + \frac{\sqrt{n}\,\Delta}{\sigma}\right) + \Phi\left(-z_{\alpha/2} - \frac{\sqrt{n}\,\Delta}{\sigma}\right).$$

Check that you get the same expression for $\pi(\mu_0 - \Delta|\delta)$.

Exercise: What happens when $\sigma > 0$ is unknown?

We can rewrite U_n as

$$U_n = \frac{\sqrt{n}\big(\bar{X}_n - \mu_0\big)/\sigma}{s_n/\sigma},$$

- The numerator has the normal distribution with mean $\sqrt{n}(\mu - \mu_0)/\sigma$ and variance 1.
- The denominator is the square-root of a χ^2-random variable divided by its degrees of freedom, $n - 1$.
- When the mean of the numerator is not 0, U_n has a *non-central* t-distribution.

Definition 20 (Noncentral t-distributions)**.** *Let W and Y_m be independent random variables $W \sim \mathcal{N}(\psi, 1)$ and $Y \sim \chi^2_m$. Then the distribution of*

$$X := \frac{W}{\sqrt{Y_m/m}}$$

is called the **non-central** *t-distribution with m degrees of freedom and non-centrality parameter ψ. We define*

$$T_m(t|\psi) = \mathbb{P}(X \le t)$$

as the c.d.f of this distribution.

- The non-central t-distribution with m degrees of freedom and non-centrality parameter $\psi = 0$ is also the t-distribution with m degrees of freedom.

- The distribution of the statistic U_n in (11) is the non-central t-distribution with $n-1$ degrees of freedom and non-centrality parameter

$$\psi := \sqrt{n}\frac{(\mu - \mu_0)}{\sigma}.$$

- The power function of δ (see Figure 9.14) is

$$\pi(\mu, \sigma^2|\delta) = T_{n-1}(-c|\psi) + 1 - T_{n-1}(c|\psi),$$

where $c := T_{n-1}^{-1}(1 - \alpha_0/2)$.

Exercise: Prove this result.

10.8.2 One-sided alternatives

We consider testing the following hypotheses:

$$H_0 : \mu \leq \mu_0, \qquad \text{versus} \qquad H_1 : \mu > \mu_0. \tag{12}$$

- When $\mu = \mu_0$, $U_n \sim t_{n-1}$, regardless of the value of σ^2.
- The test rejects H_0 if

$$U_n \geq c,$$

where $c := T_{n-1}^{-1}(1 - \alpha_0)$ (the $(1-\alpha_0)$-quantile) of the t-distribution with $n-1$ degrees of freedom.

- $\pi(\mu, \sigma^2|\delta) = 1 - T_{n-1}(c|\psi)$.

Power function of the t-test

Let δ be the test that rejects H_0 in (12) if $U_n \geq c$.

The p-value for the hypotheses in (12) is $1 - T_{n-1}(u)$, where u is the observed value of the statistic U_n.

The power function $\pi(\mu, \sigma^2|\delta)$ has the following properties:

1. $\pi(\mu, \sigma^2|\delta) = \alpha_0$ when $\mu = \mu_0$,
2. $\pi(\mu, \sigma^2|\delta) < \alpha_0$ when $\mu < \mu_0$,
3. $\pi(\mu, \sigma^2|\delta) > \alpha_0$ when $\mu > \mu_0$,

4. $\pi(\mu, \sigma^2|\delta) \to 0$ as $\mu \to -\infty$,

5. $\pi(\mu, \sigma^2|\delta) \to 1$ as $\mu \to \infty$,

6. $\sup_{\theta\in\Omega_0} \pi(\theta|\delta) = \alpha_0$.

When we want to test

$$H_0 : \mu \geq \mu_0 \qquad \text{versus} \qquad H_1 : \mu < \mu_0. \tag{13}$$

the test rejects H_0 if $U_n \leq c$, where $c = T_{n-1}^{-1}(\alpha_0)$ (the α_0-quantile) of the t-distribution with $n-1$ degrees of freedom.

Power function of the t test

Let δ be the test that rejects H_0 in (13) if $U_n \leq c$.

The p-value for the hypotheses in (13) is $T_{n-1}(u)$. Observe that $\pi(\mu, \sigma^2|\delta) = T_{n-1}(c|\psi)$.

The power function $\pi(\mu, \sigma^2|\delta)$ has the following properties:

1. $\pi(\mu, \sigma^2|\delta) = \alpha_0$ when $\mu = \mu_0$,

2. $\pi(\mu, \sigma^2|\delta) > \alpha_0$ when $\mu < \mu_0$,

3. $\pi(\mu, \sigma^2|\delta) < \alpha_0$ when $\mu > \mu_0$,

4. $\pi(\mu, \sigma^2|\delta) \to 1$ as $\mu \to -\infty$,

5. $\pi(\mu, \sigma^2|\delta) \to 0$ as $\mu \to \infty$,

6. $\sup_{\theta\in\Omega_0} \pi(\theta|\delta) = \alpha_0$.

10.9 Comparing the means of two normal distributions (two-sample t test)

10.9.1 One-sided alternatives

Random samples are available from **two** normal distributions with common unknown variance σ^2, and it is desired to determine which distribution has the larger mean. Specifically,

- $\mathbf{X} = (X_1, \ldots, X_m)$ random sample of m observations from a normal distribution for which both the mean μ_1 and the variance σ^2 are unknown, and

- $\mathbf{Y} = (Y_1, \ldots, Y_n)$ form an independent random sample of n observations from another normal distribution for which both the mean μ_2 and the variance σ^2 are unknown.
- We shall assume that the variance σ^2 is the same for both distributions, even though the value of σ^2 is unknown.

If we are interested in testing hypotheses such as

$$H_0 : \mu_1 \le \mu_2 \qquad \text{versus} \qquad H_1 : \mu_1 > \mu_2, \tag{14}$$

We reject H_0 in (14) if the difference between the sample means is large. For all values of $\theta = (\mu_1, \mu_2, \sigma^2)$ such that $\mu_1 = \mu_2$, the test statistics

$$U_{m,n} = \frac{\sqrt{m+n-2}\big(\bar{X}_m - \bar{Y}_n\big)}{\sqrt{\big(\frac{1}{m} + \frac{1}{n}\big)\big(S_X^2 + S_Y^2\big)}}$$

follows the t-distribution with $m + n - 2$ degrees of freedom, where

$$S_X^2 = \sum_{i=1}^{m} \big(X_i - \bar{X}_m\big)^2, \quad \text{and} \quad S_Y^2 = \sum_{j=1}^{n} \big(Y_j - \bar{Y}_n\big)^2.$$

We reject H_0 if

$$U_{m,n} \ge T^{-1}_{m+n-2}(1 - \alpha_0).$$

The p-value for the hypotheses in (14) is $1 - T_{m+n-2}(u)$, where u is the observed value of the statistic $U_{m,n}$.

If we are interested in testing hypotheses such as

$$H_0 : \mu_1 \ge \mu_2 \qquad \text{versus} \qquad H_1 : \mu_1 < \mu_2, \tag{15}$$

we reject H_0 if

$$U_{m,n} \le -T^{-1}_{m+n-2}(1 - \alpha_0) = T^{-1}_{m+n-2}(\alpha_0).$$

The p-value for the hypotheses in (15) is $T_{m+n-2}(u)$, where u is the observed value of the statistic $U_{m,n}$.

10.9.2 Two-sided alternatives

If we are interested in testing hypotheses such as

$$H_0 : \mu_1 = \mu_2 \qquad \text{versus} \qquad H_1 : \mu_1 \ne \mu_2, \tag{16}$$

we reject H_0 if

$$|U_{m,n}| \geq T^{-1}_{m+n-2}(1 - \tfrac{\alpha_0}{2}).$$

The p-value for the hypotheses in (16) is $2[1 - T_{m+n-2}(|u|)]$, where u is the observed value of the statistic $U_{m,n}$.

The power function of the two-sided two-sample t test is based on the non-central t-distribution in the same way as was the power function of the one-sample two-sided t-test. The test δ that rejects H_0 when $|U_{m,n}| \geq c$ has power function

$$\pi(\mu_1, \mu_2, \sigma^2|\delta) = T_{m+n-2}(-c|\psi) + 1 - T_{m+n-2}(c|\psi),$$

where $T_{m+n-2}(\cdot|\psi)$ is the c.d.f of the non-central t-distribution with $m+n-2$ degrees of freedom and non-centrality parameter ψ given by

$$\psi = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)}}.$$

10.10 Comparing the variances of two normal distributions (F-test)

- $\mathbf{X} = (X_1, \ldots, X_m)$ random sample of m observations from a normal distribution for which both the mean μ_1 and the variance σ_1^2 are unknown, and
- $\mathbf{Y} = (Y_1, \ldots, Y_n)$ form an independent random sample of n observations from another normal distribution for which both the mean μ_2 and the variance σ_2^2 are unknown.

Suppose that we want to test the hypothesis of equality of the population variances, i.e., $H_0 : \sigma_1^2 = \sigma_2^2$.

Definition 21 (F-distribution)**.** *Let Y and W be independent random variables such that $Y \sim \chi^2_m$ and $W \sim \chi^2_n$. Then the distribution of*

$$X = \frac{Y/m}{W/n}$$

is called the F-distribution with m and n degrees of freedom.

The test statistic

$$V_{m,n}^* = \frac{\frac{S_X^2}{\sigma_1^2}/(m-1)}{\frac{S_Y^2}{\sigma_2^2}/(n-1)} = \frac{\sigma_2^2 S_X^2/(m-1)}{\sigma_1^2 S_Y^2/(n-1)}$$

follows the F-distribution with $m-1$ and $n-1$ degrees of freedom. In particular, if $\sigma_1^2 = \sigma_2^2$, then the distribution of

$$V_{m,n} = \frac{S_X^2/(m-1)}{S_Y^2/(n-1)}$$

is the F-distribution with $m-1$ and $n-1$ degrees of freedom.

Let ν be the observed value of the statistic $V_{m,n}$ below, and let $G_{m-1,n-1}(\cdot)$ be the c.d.f of the F-distribution with $m-1$ and $n-1$ degrees of freedom.

10.10.1 One-sided alternatives

If we are interested in testing hypotheses such as

$$H_0: \quad \sigma_1^2 \le \sigma_2^2 \qquad \text{versus} \qquad H_1: \quad \sigma_1^2 > \sigma_2^2, \tag{17}$$

we reject H_0 if

$$V_{m,n} \ge G_{m-1,n-1}^{-1}(1-\alpha_0).$$

The p-value for the hypotheses in (17) when $V_{m,n} = \nu$ is observed equals $1 - G_{m-1,n-1}(\nu)$.

10.10.2 Two-sided alternatives

If we are interested in testing hypotheses such as

$$H_0: \quad \sigma_1^2 = \sigma_2^2, \qquad \text{versus} \qquad H_1: \quad \sigma_1^2 \ne \sigma_2^2, \tag{18}$$

we reject H_0 if either $V_{m,n} \le c_1$ or $V_{m,n} \ge c_2$, where c_1 and c_2 are constants such that

$$\mathbb{P}(V_{m,n} \le c_1) + \mathbb{P}(V_{m,n} \ge c_2) = \alpha_0$$

when $\sigma_1^2 = \sigma_2^2$. The most convenient choice of c_1 and c_2 is the one that makes

$$\mathbb{P}(V_{m,n} \le c_1) = \mathbb{P}(V_{m,n} \ge c_2) = \frac{\alpha_0}{2},$$

that is,

$$c_1 = G_{m-1,n-1}^{-1}(\alpha_0/2) \quad \text{and} \quad c_2 = G_{m-1,n-1}^{-1}(1-\alpha_0/2).$$

10.11 Likelihood ratio test

A very popular form of hypothesis test is the **likelihood ratio test**.

Suppose that we want to test

$$H_0 : \theta \in \Omega_0, \quad \text{and} \quad H_1 : \theta \in \Omega_1. \tag{19}$$

In order to compare these two hypotheses, we might wish to see whether the likelihood function is higher on Ω_0 or on Ω_1.

The *likelihood ratio statistic* is defined as

$$\Lambda(\mathbf{X}) = \frac{\sup_{\theta \in \Omega_0} L_n(\theta, \mathbf{X})}{\sup_{\theta \in \Omega} L_n(\theta, \mathbf{X})}, \tag{20}$$

where $\Omega = \Omega_0 \cup \Omega_1$.

A likelihood ratio test of the hypotheses (19) rejects H_0 when

$$\Lambda(\mathbf{x}) \leq k,$$

for some constant k.

Interpretation: we reject H_0 if the likelihood function on Ω_0 is sufficiently small compared to the likelihood function on all of Ω.

Generally, k is to be chosen so that the test has a desired level α_0.

Exercise: Suppose that $\mathbf{X}_n = (X_1, \ldots, X_n)$ is a random sample from a normal distribution with unknown mean μ and known variance σ^2. We wish to test the hypotheses

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_a : \mu \neq \mu_0$$

at the level α_0. Show that the likelihood ratio test is equivalent to the z-test.

Exercise: Suppose that $X_1, \ldots, X_n$ from a normal distribution $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. We wish to test the hypotheses

$$H_0 : \sigma^2 = \sigma_0^2 \qquad \text{versus} \qquad H_a : \sigma^2 \neq \sigma_0^2$$

at the level α. Show that the likelihood ratio test is equivalent to the χ^2-test. [**Hint**: Show that $\Lambda(\mathbf{X}_n) = e^{n/2} \left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{n/2} \exp\left(-\frac{n}{2}\frac{\hat{\sigma}^2}{\sigma_0^2}\right)$ where $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Note that as $\Lambda(\mathbf{X}_n)$ is a function of $\frac{\hat{\sigma}^2}{\sigma_0^2}$, the inequality $\Lambda(\mathbf{X}_n) \leq c$ holds if and only if $\frac{\hat{\sigma}^2}{\sigma_0^2}$ is too big or too small; show this plotting the graph of $\log x - x$.]

Exercise: Suppose that $X_1, \ldots, X_n$ from a normal distribution $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. We wish to test the hypotheses

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_a : \mu \neq \mu_0$$

at the level α. Show that the likelihood ratio test is equivalent to the t-test [**Hint**: Show that $\Lambda(\mathbf{X}_n) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2}$ where $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ and $\hat{\sigma}_0^2 = n^{-1}\sum_{i=1}^n (X_i - \mu_0)^2$. Thus, $\Lambda(\mathbf{X}_n) \leq c \Leftrightarrow (\bar{X}_n - \mu_0)^2/s^2 \geq c'$, for a suitable c' where $s^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$.]

Theorem 10.2. *Let Ω be a open set of a p-dimensional space, and suppose that H_0 specifies that k coordinates of θ are equal to k specific values. Assume that H_0 is true and that the likelihood function satisfies the conditions needed to prove that the MLE is asymptotically normal and asymptotically efficient. Then, as $n \to \infty$,*

$$-2 \log \Lambda(\mathbf{X}) \xrightarrow{d} \chi_k^2.$$

Exercise: Let $X_1, \ldots, X_n$ be a random sample from the p.d.f

$$f_\theta(x) = e^{-(x-\theta)} \mathbf{1}_{[\theta,\infty)}(x).$$

Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, where θ_0 is a fixed value specified by the experimenter.

Show that the likelihood ratio test statistic is

$$\Lambda(\boldsymbol{X}) = \begin{cases} 1 & X_{(1)} \leq \theta_0 \\ e^{-n(X_{(1)}-\theta_0)} & X_{(1)} > \theta_0. \end{cases}$$

10.12 Equivalence of tests and confidence sets

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d $N(\mu, \sigma^2)$ where μ is unknown and σ^2 is known.

We now illustrate how the testing procedure ties up naturally with the CI construction problem.

Consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

First note that the acceptance region of the derived test δ can be written as:

$$S_0 = \mathcal{A}_{\mu_0} = \left\{ \boldsymbol{x} = (x_1, x_2, \ldots, x_n) : \overline{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha_0/2} \leq \mu_0 \leq \overline{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha_0/2} \right\}.$$

Now, consider a fixed data set $(X_1, X_2, \ldots, X_n)$ and based on this consider testing a family of null hypotheses:

$$\{H_{0,\tilde{\mu}} : \mu = \tilde{\mu} : \tilde{\mu} \in \mathbb{R}\}.$$

We can now ask the following question: Based on the observed data and the above testing procedure, *what values of $\tilde{\mu}$ would fail to be rejected by the level α_0 test*? This means that $\tilde{\mu}$ would have to fall in the interval

$$\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha_0/2} \leq \tilde{\mu} \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha_0/2} \, .$$

Thus, the set of $\tilde{\mu}$'s for which the null hypothesis would fail to be rejected by the level α_0 test is the set:

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha_0/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha_0/2}\right].$$

But this is precisely the level $1 - \alpha_0$ CI that we obtained before!

Thus, we obtain a level $1-\alpha_0$ CI for μ, the population mean, by *compiling all possible $\tilde{\mu}$'s for which the null hypothesis $H_{0,\tilde{\mu}} : \mu = \tilde{\mu}$ fails to be rejected by the level α_0 test.*

From hypothesis testing to CIs: Let $X_1, X_2, \ldots, X_n$ be i.i.d observations from some underlying distribution F_θ; here θ is a "parameter" indexing a family of distributions. The goal is to construct a CI for θ using hypothesis testing.

For each $\tilde{\theta}$ consider testing the null hypothesis $H_{0,\tilde{\theta}} : \theta = \tilde{\theta}$. Suppose, there exists a level α_0 test $\delta_{\tilde{\theta}}(\boldsymbol{X})$ for this problem with

$$\mathcal{A}_{\tilde{\theta}} = \{\boldsymbol{x} : T_{\tilde{\theta}}(\boldsymbol{x}) \leq c_{\alpha_0}\}$$

being the acceptance region of $\delta_{\tilde{\theta}}$ and

$$\mathbb{P}_{\tilde{\theta}}(\boldsymbol{X} \in \mathcal{A}_{\tilde{\theta}}) \geq 1 - \alpha_0.$$

Then a level $1 - \alpha$ confidence set for θ is:

$$\mathcal{S}(\boldsymbol{X}) = \{\tilde{\theta} : \boldsymbol{X} \in \mathcal{A}_{\tilde{\theta}}\}.$$

We need to verify that for any θ,

$$\mathbb{P}_\theta[\theta \in \mathcal{S}(\boldsymbol{X})] \geq 1 - \alpha.$$

But

$$\mathbb{P}_\theta(\theta \in \mathcal{S}(\boldsymbol{X})) = \mathbb{P}_\theta(\boldsymbol{X} \in \mathcal{A}_\theta) \geq 1 - \alpha_0.$$

Theorem 10.3. *For each $\theta_0 \in \Omega$, let $\mathcal{A}(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. For each $\boldsymbol{x} \in \mathcal{X}$ ($\mathcal{X}$ is the space of all data values), define a set $\mathcal{S}(\boldsymbol{x})$ in the parameter space by*

$$\mathcal{S}(\boldsymbol{x}) = \{\theta_0 : \boldsymbol{x} \in \mathcal{A}(\theta_0)\}.$$

Then the random set $\mathcal{S}(\boldsymbol{X})$ is a $1-\alpha$ confidence set. Conversely, let $\mathcal{S}(\boldsymbol{X})$ be a $1-\alpha$ confidence set. For any $\theta_0 \in \Omega$, define

$$\mathcal{A}(\theta_0) = \{\boldsymbol{x} : \theta_0 \in \mathcal{S}(\boldsymbol{x})\}.$$

Then $\mathcal{A}(\theta_0)$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$.

Proof. The first part is essentially done above!

For the second part, the type I error probability for the test of $H_0 : \theta = \theta_0$ with acceptance region $\mathcal{A}(\theta_0)$ is

$$\mathbb{P}_{\theta_0}(\boldsymbol{X} \notin \mathcal{A}_{\theta_0}) = \mathbb{P}_{\theta_0}[\theta_0 \notin \mathcal{S}(\boldsymbol{X})] \leq \alpha.$$

□

Remark: The more useful part of the theorem is the first part, i.e., given a level α test (which is usually easy to construct) we can get a confidence set by inverting the family of tests.

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d Exp(λ). We want to test $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$.

Find the LRT.

The acceptance region is given by

$$\mathcal{A}(\lambda_0) = \left\{\boldsymbol{x} : \left(\frac{\sum x_i}{\lambda_0}\right)^n e^{-\sum x_i/\lambda_0} \geq k^*\right\},$$

where k^* is a constant chosen to satisfy

$$\mathbb{P}_{\lambda_0}(\boldsymbol{X} \in \mathcal{A}(\lambda_0)) = 1 - \alpha.$$

Inverting this acceptance region gives the $1-\alpha$ confidence set

$$\mathcal{S}(\boldsymbol{x}) = \left\{\lambda : \left(\frac{\sum x_i}{\lambda}\right)^n e^{-\sum x_i/\lambda_0} \geq k^*\right\}.$$

This can be shown to be an interval in the parameter space.

11 Linear regression

- We are often interested in understanding the *relationship* between two or more variables.
- Want to model a functional relationship between a "predictor" (input, independent variable) and a "response" variable (output, dependent variable, etc.).
- But real world is noisy, no $f = ma$ (Force = mass $\times$ acceleration). We have observation noise, weak relationship, etc.

Examples:

- How is the *sales price* of a house related to its size, number of rooms and property tax?
- How does the probability of *surviving* a particular surgery change as a function of the patient's age and general health condition?
- How does the *weight* of an individual depend on his/her height?

11.1 Method of least squares

Suppose that we have n data points $(x_1, Y_1), \ldots, (x_n, Y_n)$. We want to predict Y given a value of x.

- Y_i is the value of the **response** variable for the i-th observation.
- x_i is the value of the **predictor** (covariate/explanatory variable) for the i-th observation.
- **Scatter plot:** Plot the data and try to visualize the relationship.
- Suppose that we think that Y is a **linear** function (actually here a more appropriate term is "affine") of x, i.e.,

  $$Y_i \approx \beta_0 + \beta_1 x_i,$$

  and we want to find the "best" such linear function.
- For the correct parameter values β_0 and β_1, the *deviation* of the observed values to its expected value, i.e.,

  $$Y_i - \beta_0 - \beta_1 x_i,$$

  should be *small*.

- We try to *minimize* the sum of the n squared deviations, i.e., we can try to minimize
$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2$$
as a function of b_0 and b_1. In other words, we want to minimize the sum of the squares of the vertical deviations of all the points from the line.

- The least squares estimators can be found by differentiating Q with respect to b_0 and b_1 and setting the partial derivatives equal to 0.

- Find b_0 and b_1 that solve:
$$\begin{aligned} \frac{\partial Q}{\partial b_0} &= -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i) = 0 \\ \frac{\partial Q}{\partial b_1} &= -2 \sum_{i=1}^{n} x_i (Y_i - b_0 - b_1 x_i) = 0. \end{aligned}$$

11.1.1 Normal equations

- The values of b_0 and b_1 that minimize Q are given by the solution to the *normal equations*:
$$\begin{aligned} \sum_{i=1}^{n} Y_i &= n b_0 + b_1 \sum_{i=1}^{n} x_i &(21) \\ \sum_{i=1}^{n} x_i Y_i &= b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2. &(22) \end{aligned}$$

- Solving the normal equations gives us the following point estimates:
$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, &(23) \\ b_0 &= \bar{Y} - b_1 \bar{x}, &(24) \end{aligned}$$
where $\bar{x} = \sum_{i=1}^{n} x_i / n$ and $\bar{Y} = \sum_{i=1}^{n} Y_i / n$.

In general, if we can parametrize the form of the functional dependence between Y and x in a linear fashion (linear in the parameters), then the method of least squares can be used to estimate the function. For example,
$$Y_i \approx \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$
is still linear in the parameters.

11.2 Simple linear regression

The model for **simple linear regression** can be stated as follows:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n.$$

- Observations: $\{(x_i, Y_i) : i = 1, \ldots, n\}$.
- β_0, β_1 and σ^2 are *unknown* parameters.
- ϵ_i is a (unobserved) **random error** term whose distribution is unspecified:

$$\mathbb{E}(\epsilon_i) = 0, \qquad \text{Var}(\epsilon_i) = \sigma^2, \qquad \text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \text{for } i \neq j.$$

- x_i's will be treated as known *constants*. Even if the x_i's are random, we condition on the predictors and want to understand the **conditional distribution** of Y given X.
- **Regression function**: **Conditional mean** on Y given x, i.e.,

$$m(x) := \mathbb{E}(Y|x) = \beta_0 + \beta_1 x.$$

- The regression function shows how the mean of Y changes as a *function* of x.
- $\mathbb{E}(Y_i) = \mathbb{E}(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i$
- $\text{Var}(Y_i) = \text{Var}(\beta_0 + \beta_1 x_i + \epsilon_i) = \text{Var}(\epsilon_i) = \sigma^2$.

11.2.1 Interpretation

- The slope β_1 has units "y-units per x-units".
  - For every 1 inch increase in height, the model predicts a β_1 *pounds increase* in the mean weight.
- The intercept term β_0 is not always meaningful.
- The model is *only valid* for values of the explanatory variable in the domain of the data.

11.2.2 Estimated regression function

- After formulating the model we use the observed data to *estimate* the *unknown* parameters.

- Three unknown parameters: β_0, β_1 and σ^2.

- We are interested in finding the estimates of these parameters that *best fit* the data.

- Question: *Best* in what sense?

- The **least squares** estimators of β_0 and β_1 are those values b_0 and b_1 that minimize:
$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2.$$

- Solving the normal equations gives us the following point estimates:
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{25}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \tag{26}$$
where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $\bar{Y} = \sum_{i=1}^{n} Y_i/n$.

- We estimate the regression function:
$$\mathbb{E}(Y) = \beta_0 + \beta_1 x$$
using
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

- The term
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad i = 1, \ldots, n,$$
is called the **fitted** or *predicted* value for the i-th observation, while Y_i is the observed value.

- The *residual*, denoted e_i, is the difference between the observed and the predicted value of Y_i, i.e.,
$$e_i = Y_i - \hat{Y}_i.$$

- The residuals show how far the individual data points fall from the regression function.

11.2.3 Properties

1. The sum of the residuals $\sum_{i=1}^{n} e_i$ is zero.

2. The sum of the squared residuals is a minimum.

3. The sum of the observed values equal the sum of the predicted values, i.e., $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$.

4. The following sums of weighted residuals are equal to zero:

$$\sum_{i=1}^{n} x_i e_i = 0 \qquad \sum_{i=1}^{n} e_i = 0.$$

5. The regression line always passes through the point $(\bar{x}, \bar{Y})$.

11.2.4 Estimation of σ^2

- Recall: $\sigma^2 = \mathrm{Var}(\epsilon_i)$.

- We might have used $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2}{n-1}$. But ϵ_i's are not *observed*!

- Idea: Use e_i's, i.e., $s^2 = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$.

- The divisor $n-2$ in s^2 is the number of **degrees of freedom** associated with the estimate.

- To obtain s^2, the two parameters β_0 and β_1 must first be estimated, which results in a loss of *two* degrees of freedom.

- Using $n-2$ makes s^2 an *unbiased* estimator of σ^2, i.e., $\mathbb{E}(s^2) = \sigma^2$.

11.2.5 Gauss-Markov theorem

The least squares estimators $\hat{\beta}_0, \hat{\beta}_1$ are **unbiased** (why?), i.e.,

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \qquad \mathbb{E}(\hat{\beta}_1) = \beta_1.$$

A *linear estimator* of β_j ($j = 0, 1$) is an estimator of the form

$$\tilde{\beta}_j = \sum_{i=1}^{n} c_i Y_i,$$

where the coefficients $c_1, \ldots, c_n$ are only allowed to depend on x_i.

Note that $\hat{\beta}_0, \hat{\beta}_1$ are linear estimators (show this!).

Result: No matter what the distribution of the error terms ϵ_i, the least squares method provides *unbiased* point estimates that have **minimum** variance among all **unbiased linear estimators**.

The Gauss-Markov theorem states that in a linear regression model in which the errors have **expectation zero** and are **uncorrelated** and have **equal variances**, the *best linear unbiased estimator* (BLUE) of the coefficients is given by the **ordinary least squares estimators**.

11.3 Normal simple linear regression

To perform *inference* we need to make assumptions regarding the distribution of ϵ_i.

We often assume that ϵ_i's are *normally* distributed.

The *normal error* version of the model for simple linear regression can be written:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

Here ϵ_i's are independent $N(0, \sigma^2)$, σ^2 unknown.

Hence, Y_i's are independent normal random variables with mean $\beta_0 + \beta_1 x_i$ and variance σ^2.

Picture?

11.3.1 Maximum likelihood estimation

When the probability distribution of Y_i is *specified*, the estimates can be obtained using the method of *maximum likelihood*.

This method chooses as estimates those values of the parameter that are most *consistent* with the observed data.

The *likelihood* is the *joint density* of the Y_i's viewed as a function of the unknown parameters, which we denote $L(\beta_0, \beta_1, \sigma^2)$.

Since the Y_i's are *independent* this is simply the *product* of the density of individual Y_i's.

We seek the values of β_0, β_1 and σ^2 that maximize $L(\beta_0, \beta_1, \sigma^2)$ for the given x and Y values in the sample.

According to our model:

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \qquad \text{for } i = 1, 2, \ldots, n.$$

The likelihood function for the n independent observations $Y_1, \dots, Y_n$ is given by

$$
\begin{aligned}
L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 x_i)^2\right\} && (27) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2\right\}.
\end{aligned}
$$

The value of $(\beta_0, \beta_1, \sigma^2)$ that maximizes the likelihood function are called *maximum likelihood estimates* (MLEs).

The MLE of β_0 and β_1 are *identical* to the ones obtained using the method of *least squares*, i.e.,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{S_x^2},$$

where $S_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

The MLE of σ^2: $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}$.

11.3.2 Inference

Our model describes the *linear* relationship between the two variables x and Y.

Different samples from the same population will produce different point estimates of β_0 and β_1.

Hence, $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables with sampling distributions that describe *what values* they can take and *how often* they take them.

Hypothesis tests about β_0 and β_1 can be constructed using these distributions.

The next step is to perform *inference*, including:

- Tests and confidence intervals for the *slope* and intercept.
- Confidence intervals for the *mean response*.
- *Prediction* intervals for new observations.

Theorem 11.1. *Under the assumptions of the normal linear model,*

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{S_x^2} & -\frac{\bar{x}}{S_x^2} \\ -\frac{\bar{x}}{S_x^2} & \frac{1}{S_x^2} \end{pmatrix}\right)$$

where $S_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Also, if $n \geq 3$, $\hat{\sigma}^2$ is independent of $(\hat{\beta}_0, \hat{\beta}_1)$ and $n\hat{\sigma}^2/\sigma^2$ has a χ^2-distribution with $n-2$ degrees of freedom.

Note that if the x_i's are random, the above theorem is still valid if we condition on the values of the predictor x_i's.

Exercise: Compute the variances and covariance of $\hat{\beta}_0, \hat{\beta}_1$.

11.3.3 Inference about β_1

We often want to perform tests about the *slope*:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0.$$

Under the null hypothesis there is *no linear relationship* between Y and x – the *means* of probability distributions of Y are equal at all levels of x, i.e., $\mathbb{E}(Y|x) = \beta_0$, for all x.

The *sampling distribution* of $\hat{\beta}_1$ is given by

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_x^2}\right).$$

Need to show that: $\hat{\beta}_1$ is normally distributed,

$$\mathbb{E}(\hat{\beta}_1) = \beta_1, \qquad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_x^2}.$$

Result: When $Z_1, \ldots, Z_k$ are *independent* normal random variables, the linear combination

$$a_1 Z_1 + \ldots + a_k Z_k$$

is also *normally* distributed.

Since $\hat{\beta}_1$ is a linear combination of the Y_i's and each Y_i is an *independent normally* distributed random variable, then $\hat{\beta}_1$ is also normally distributed.

We can write $\hat{\beta}_1 = \sum_{i=1}^n w_i Y_i$ where

$$w_i = \frac{x_i - \bar{x}}{S_x^2}, \qquad \text{for } i = 1, \ldots, n.$$

Thus,

$$\sum_{i=1}^n w_i = 0, \quad \sum_{i=1}^n x_i w_i = 1, \quad \sum_{i=1}^n w_i^2 = \frac{1}{S_x^2}.$$

- **Variance for the estimated slope:** There are *three* aspects of the scatter plot that affect the variance of the regression slope:

- The *spread* around the *regression line* (σ^2) – less scatter around the line means the slope will be more consistent from sample to sample.
- The *spread* of the *x values* ($\sum_{i=1}^n (x_i - \bar{x})^2/n$) – a large variance of x provides a more stable regression.
- The *sample size* n – having a larger sample size n, gives more consistent estimates.

- **Estimated variance:** When σ^2 is *unknown* we replace it with the

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

Plugging this into the equation for $\mathrm{Var}(\hat{\beta}_1)$ we get

$$se^2(\hat{\beta}_1) = \frac{\tilde{\sigma}^2}{S_x^2}.$$

Recall: *Standard error* $\mathrm{se}(\hat{\theta})$ of an estimator $\hat{\theta}$ is used to refer to an *estimate* of its *standard deviation*.

Result: For the normal error regression model:

$$\frac{SSE}{\sigma^2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2_{n-2},$$

and is *independent* of $\hat{\beta}_0$ and $\hat{\beta}_1$.

- (**Studentized statistic:**) Since $\hat{\beta}_1$ is *normally* distributed, the standardized statistic:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} \sim N(0, 1).$$

If we replace $\mathrm{Var}(\hat{\beta}_1)$ by its estimate we get the *studentized* statistic:

$$\frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)} \sim t_{n-2}.$$

Recall: Suppose that $Z \sim N(0,1)$ and $W \sim \chi^2_p$ where Z and W are independent. Then,

$$\frac{Z}{\sqrt{W/p}} \sim t_p,$$

the *t-distribution* with p *degrees of freedom*.

- **Hypothesis testing:** To test

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0$$

use the *test-statistic*

$$T = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}.$$

We reject H_0 when the observed value of $|T|$ i.e., $|t_{obs}|$, is *large*!

Thus, given *level* $(1-\alpha)$, we reject H_0 if

$$|t_{obs}| > t_{1-\alpha/2,n-2}$$

where $t_{1-\alpha/2,n-2}$ denotes the $(1-\alpha/2)$-quantile of the t_{n-2}-distribution, i.e.,

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

- **P-value:** p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

  The p-value depends on H_1 (one-sided/two-sided).

  In our case, we compute p-values using a t_{n-2}-distribution. Thus,

$$p\text{-value} = \mathbb{P}_{H_0}(|T| > |t_{obs}|).$$

  If we know the p-value then we can decide to accept/reject H_0 (versus H_1) at any given α.

- **Confidence interval:** A *confidence interval* (CI) is a kind of *interval estimator* of a population parameter and is used to indicate the reliability of an estimator.

  Using the sampling distribution of $\hat{\beta}_1$ we can make the following probability statement:

$$\begin{aligned}
\mathbb{P}\left(t_{\alpha/2,n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \leq t_{1-\alpha/2,n-2}\right) &= 1-\alpha \\
\mathbb{P}\left(\hat{\beta}_1 - t_{1-\alpha/2,n-2}\text{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 - t_{\alpha/2,n-2}\text{se}(\hat{\beta}_1)\right) &= 1-\alpha.
\end{aligned}$$

  Thus, a $(1-\alpha)$ confidence interval for β_1 is

$$\left[\hat{\beta}_1 - t_{1-\alpha/2,n-2} \cdot se(\hat{\beta}_1), \hat{\beta}_1 + t_{1-\alpha/2,n-2} \cdot se(\hat{\beta}_1)\right]$$

  as $t_{1-\alpha/2,n-2} = -t_{\alpha/2,n-2}$.

11.3.4 Sampling distribution of $\hat{\beta}_0$

The *sampling distribution* of $\hat{\beta}_0$ is

$$N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}\right)\right).$$

Verify at home using the same procedure as used for $\hat{\beta}_1$.

Hypothesis testing: In general, let c_0, c_1 and c_* be specified numbers, where at least one of c_0 and c_1 is nonzero. Suppose that we are interested in testing the following hypotheses:

$$H_0 : c_o\beta_0 + c_1\beta_1 = c_*, \qquad \text{versus} \qquad H_0 : c_o\beta_0 + c_1\beta_1 \neq c_*. \tag{28}$$

We should use a scalar multiple of

$$c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*$$

as the test statistic. Specifically, we use

$$U_{01} = \left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{S_x^2}\right]^{-1/2}\left(\frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*}{\tilde{\sigma}}\right),$$

where

$$\tilde{\sigma}^2 = \frac{S^2}{n-2}, \qquad S^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} e_i^2.$$

Note that $\tilde{\sigma}^2$ is an unbiased estimator of σ^2.

For each $\alpha \in (0, 1)$, a level α test of the hypothesis (28) is to reject H_0 if

$$|U_{01}| > T_{n-2}^{-1}\left(1 - \frac{\alpha}{2}\right).$$

The above result follows from the fact that $c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*$ is normally distributed with mean $c_0\beta_0 + c_1\beta_1 - c_*$ and variance

$$\begin{aligned}
\mathrm{Var}(c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*) &= c_0^2\mathrm{Var}(\hat{\beta}_0) + c_1^2\mathrm{Var}(\hat{\beta}_1) + 2c_0c_1\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\
&= c_0^2\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}\right) + c_1^2\sigma^2\frac{1}{S_x^2} - 2c_0c_1\frac{\sigma^2\bar{x}}{S_x^2} \\
&= \sigma^2\left[\frac{c_0^2}{n} + \frac{c_0^2\bar{x}^2}{S_x^2} - 2c_0c_1\frac{\bar{x}}{S_x^2} + c_1^2\frac{1}{S_x^2}\right] \\
&= \sigma^2\left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{S_x^2}\right].
\end{aligned}$$

Confidence interval: We can give a $1-\alpha$ confidence interval for the parameter $c_0\beta_0 + c_1\beta_1$ as

$$c_0\hat{\beta}_0 + c_1\hat{\beta}_1 \mp \tilde{\sigma}\left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{S_x^2}\right]^{1/2} T_{n-2}^{-1}\left(1 - \frac{\alpha}{2}\right).$$

11.3.5 Mean response

We often want to estimate the *mean* of the probability distribution of Y for some value of x.

- The *point estimator* of the mean response

  $$\mathbb{E}(Y|x_h) = \beta_0 + \beta_1 x_h$$

  when $x = x_h$ is given by

  $$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h.$$

  Need to:

  - Show that $\hat{Y}_h$ is *normally* distributed.
  - Find $\mathbb{E}(\hat{Y}_h)$.
  - Find $\mathrm{Var}(\hat{Y}_h)$.

- The sampling distribution of $\hat{Y}_h$ is given by

  $$\hat{Y}_h \sim N\left(\beta_0 + \beta_1 x_h, \sigma^2\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_x^2}\right)\right).$$

Normality:

Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are *linear combinations* of independent normal random variables Y_i.

Hence, $\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$ is also a linear combination of independent normally distributed random variables.

Thus, $\hat{Y}_h$ is also normally distributed.

Mean and variance of $\hat{Y}_h$:

Find the expected value of $\hat{Y}_h$:

$$\mathbb{E}(\hat{Y}_h) = \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_h) = \mathbb{E}(\hat{\beta}_0) + \mathbb{E}(\hat{\beta}_1) x_h = \beta_0 + \beta_1 x_h.$$

Note that $\hat{Y}_h = \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_h = \bar{Y} + \hat{\beta}_1 (x_h - \bar{x})$.

Note that $\hat{\beta}_1$ and $\bar{Y}$ are *uncorrelated*:

$$\mathrm{Cov}\left(\sum_{i=1}^{n} w_i Y_i, \sum_{i=1}^{n} \frac{1}{n} Y_i\right) = \sum_{i=1}^{n} \frac{w_i}{n}\sigma^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n} w_i = 0.$$

Therefore,

$$\begin{aligned} \mathrm{Var}(\hat{Y}_h) &= \mathrm{Var}(\bar{Y}) + (x_h - \bar{x})^2 \mathrm{Var}(\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + (x_h - \bar{x})^2 \frac{\sigma^2}{S_x^2}. \end{aligned}$$

When we do not know σ^2 we estimate it using $\tilde{\sigma}^2$. Thus, the *estimated variance* of $\hat{Y}_h$ is given by

$$\mathrm{se}^2(\hat{Y}_h) = \tilde{\sigma}^2 \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_x^2}\right).$$

The variance of $\hat{Y}_h$ is *smallest* when $x_h = \bar{x}$.

When $x_h = 0$, the variance of reduces to the variance of $\hat{\beta}_0$.

- The sampling distribution for the studentized statistic:

$$\frac{\hat{Y}_h - \mathbb{E}(\hat{Y}_h)}{\mathrm{se}(\hat{Y}_h)} \sim t_{n-2}.$$

All inference regarding $\mathbb{E}(\hat{Y}_h)$ are carried out using the t-distribution. A $(1-\alpha)$ CI for the *mean response* when $x = x_h$ is

$$\hat{Y}_h \mp t_{1-\alpha/2, n-2}\, \mathrm{se}(\hat{Y}_h).$$

11.3.6 Prediction interval

A CI for a *future* observation is called a *prediction interval*.

Consider the prediction of a new observation Y corresponding to a given level x of the predictor.

Suppose $x = x_h$ and the new observation is denoted $Y_{h(new)}$.

Note that $\mathbb{E}(\hat{Y}_h)$ is the *mean* of the distribution of $Y|X = x_h$.

$Y_{h(new)}$ represents the prediction of an *individual outcome* drawn from the distribution of $Y|X = x_h$, i.e.,

$$Y_{h(new)} = \beta_0 + \beta_1 x_h + \epsilon_{new},$$

where ϵ_{new} is independent of our data.

- The *point estimate* will be the *same* for both.

  However, the variance is *larger* when predicting an individual outcome due to the *additional variation* of an individual about the mean.

- When constructing prediction limits for $Y_{h(new)}$ we must take into consideration two sources of variation:

  - Variation in the *mean of* Y.
  - Variation around the mean.

- The *sampling* distribution of the studentized statistic:

$$\frac{Y_{h(new)} - \hat{Y}_h}{\text{se}(Y_{h(new)} - \hat{Y}_h)} \sim t_{n-2}.$$

  All inference regarding $Y_{h(new)}$ are carried out using the t-distribution:

$$\text{Var}(Y_{h(new)} - \hat{Y}_h) = \text{Var}(Y_{h(new)}) + \text{Var}(\hat{Y}_h) = \sigma^2 \left\{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_x^2}\right\}.$$

  Thus, $\text{se}_{pred} = \text{se}(Y_{h(new)} - \hat{Y}_h) = \tilde{\sigma}^2 \left\{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_x^2}\right\}$.

  Using this result, $(1 - \alpha)$ *prediction interval* for a new observation $Y_{h(new)}$ is

$$\hat{Y}_h \mp t_{1-\alpha/2,n-2} \; \text{se}_{pred}.$$

11.3.7 Inference about both β_0 and β_1 simultaneously

Suppose that β_0^* and β_1^* are given numbers and we are interested in testing the following hypothesis:

$$H_0 : \beta_0 = \beta_0^* \text{ and } \beta_1 = \beta_1^* \qquad \text{versus} \qquad H_1 : \text{at least one is different} \tag{29}$$

We shall derive the likelihood ratio test for (29).

The likelihood function (27), when maximized under the unconstrained space yields the MLEs $\hat{\beta}_1, \hat{\beta}_1, \hat{\sigma}^2$.

Under the constrained space, β_0 and β_1 are fixed at β_0^* and β_1^*, and so

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0^* - \beta_1^* x_i)^2.$$

The likelihood statistic reduces to

$$\Lambda(\mathbf{Y}, \mathbf{x}) = \frac{\sup_{\sigma^2} L(\beta_0^*, \beta_1^*, \sigma^2)}{\sup_{\beta_0,\beta_1,\sigma^2} L(\beta_0, \beta_1, \sigma^2)} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} = \left[\frac{\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^n (Y_i - \beta_0^* - \beta_1^* x_i)^2}\right]^{n/2}.$$

The LRT procedure specifies rejecting H_0 when

$$\Lambda(\mathbf{Y}, \mathbf{x}) \leq k,$$

for some k, chosen given the level condition.

Exercise: Show that

$$\sum_{i=1}^n (Y_i - \beta_0^* - \beta_1^* x_i)^2 = S^2 + Q^2,$$

where

$$\begin{aligned} S^2 &= \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ Q^2 &= n(\hat{\beta}_0 - \beta_0^*)^2 + \left(\sum_{i=1}^n x_i^2\right)(\hat{\beta}_1 - \beta_1^*)^2 + 2n\bar{x}(\hat{\beta}_0 - \beta_0^*)(\hat{\beta}_1 - \beta_1^*). \end{aligned}$$

Thus,

$$\Lambda(\mathbf{Y}, \mathbf{x}) = \left[\frac{S^2}{S^2 + Q^2}\right]^{n/2} = \left[1 + \frac{Q^2}{S^2}\right]^{-n/2}.$$

It can be seen that this is equivalent to rejecting H_0 when $Q^2/S^2 \geq k'$ which is equivalent to

$$U^2 := \frac{\frac{1}{2}Q^2}{\tilde{\sigma}^2} \geq \gamma.$$

Exercise: Show that, under H_0, $\frac{Q^2}{\sigma^2} \sim \chi_2^2$. Also show that Q^2 and S^2 are independent.

We know that $S^2/\sigma^2 \sim \chi_{n-2}^2$. Thus, under H_0,

$$U^2 \sim F_{2,n-2},$$

and thus $\gamma = F_{2,n-2}^{-1}(1 - \alpha)$.

12 Linear models with normal errors

12.1 Basic theory

This section concerns models for independent responses of the form

$$Y_i \sim N(\mu_i, \sigma^2), \qquad \text{where} \qquad \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$$

for some known vector of explanatory variables $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{ip})$ and *unknown* parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, where $p < n$.

This is the **linear model** and is usually written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(in vector notation) where

$$\mathbf{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X}_{n\times p} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix}, \quad \boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon}_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

Sometimes this is written in the more compact notation

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}),$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

It is usual to assume that the $n \times p$ matrix $\mathbf{X}$ has full rank p.

12.2 Maximum likelihood estimation

The log–likelihood (up to a constant term) for $(\boldsymbol{\beta}, \sigma^2)$ is

$$\begin{aligned}\ell(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2 \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(Y_i - \sum_{j=1}^p x_{ij}\beta_j\right)^2.\end{aligned}$$

An MLE $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ satisfies

$$\begin{aligned} 0 \quad &= \quad \frac{\partial}{\partial\beta_j}\ell(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^n x_{ij}(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}), \quad \text{for } j = 1, \ldots, p, \\ \text{i.e.,} \qquad &\sum_{i=1}^n x_{ij}\mathbf{x}_i^\top\hat{\boldsymbol{\beta}} = \sum_{i=1}^n x_{ij}y_i \qquad\qquad \text{for } j = 1, \ldots, p,\end{aligned}$$

so

$$(\mathbf{X}^\top\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{Y}.$$

Since $\mathbf{X}^\top\mathbf{X}$ is non-singular if $\mathbf{X}$ has rank p, we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

The **least squares estimator** of β minimizes

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Check that this estimator coincides with the MLE when the errors are normally distributed.

Thus the estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ may be justified even when the normality assumption is uncertain.

Theorem 12.1. *We have*

1.
$$\hat{\beta} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}), \tag{30}$$

2.
$$\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}})^2$$

   and that $\hat{\sigma}^2 \sim \frac{\sigma^2}{n}\chi^2_{n-p}$.

3. *Show that* $\hat{\boldsymbol{\beta}}$ *and* $\hat{\sigma}^2$ *are independent.*

Recall: Suppose that $\mathbf{U}$ is an n-dimensional random vector for which the mean vector $\mathbb{E}(\mathbf{U})$ and the covariance matrix $\mathrm{Cov}(\mathbf{U})$ exist. Suppose that $\mathbf{A}$ is a $q \times n$ matrix whose elements are constants. Let $\mathbf{V} = \mathbf{A}\mathbf{U}$. Then

$$\mathbb{E}(\mathbf{V}) = \mathbf{A}\mathbb{E}(\mathbf{U}) \qquad \text{and} \qquad \mathrm{Cov}(\mathbf{V}) = \mathbf{A}\mathrm{Cov}(\mathbf{U})\mathbf{A}^\top.$$

Proof of 1: The MLE of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$, and we have that the model can be written in vector notation as $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

Let $\mathbf{M} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ so that $\mathbf{M}\mathbf{Y} = \hat{\boldsymbol{\beta}}$. Therefore,

$$\mathbf{M}\mathbf{Y} \sim N_p(\mathbf{M}\mathbf{X}\boldsymbol{\beta}, \mathbf{M}(\sigma^2\mathbf{I})\mathbf{M}^\top).$$

We have that

$$\begin{aligned} \mathbf{M}\mathbf{X}\boldsymbol{\beta} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned} \qquad \text{and} \qquad \begin{aligned} \mathbf{M}\mathbf{M}^\top &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1} \end{aligned}$$

since $\mathbf{X}^\top\mathbf{X}$ is symmetric, and then so is it's inverse.

Therefore,

$$\hat{\boldsymbol{\beta}} = \mathbf{MY} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}).$$

These results can be used to obtain an exact $(1-\alpha)$-level confidence region for $\boldsymbol{\beta}$: the distribution of $\hat{\boldsymbol{\beta}}$ implies that

$$\frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p.$$

Let

$$\tilde{\sigma}^2 = \frac{1}{n-p}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \sim \frac{\sigma^2}{n-p}\chi^2_{n-p},$$

so that $\hat{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$ are still independent.

Then, letting $F_{p,n-p}(\alpha)$ denote the upper α-point of the $F_{p,n-p}$ distribution,

$$1-\alpha = \mathbb{P}_{\boldsymbol{\beta},\sigma^2}\left(\frac{\frac{1}{p}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\tilde{\sigma}^2} \leq F_{p,n-p}(\alpha)\right).$$

Thus,

$$\left\{\boldsymbol{\beta} \in \mathbb{R}^p : \frac{\frac{1}{p}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\tilde{\sigma}^2} \leq F_{p,n-p}(\alpha)\right\}$$

is a $(1-\alpha)$-level confidence set for $\boldsymbol{\beta}$.

12.2.1 Projections and orthogonality

The *fitted values* $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ under the model satisfy

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{Y} \equiv \mathbf{PY},$$

say, where $\mathbf{P}$ is an *orthogonal projection* matrix (i.e., $\mathbf{P} = \mathbf{P}^\top$ and $\mathbf{P^2} = \mathbf{P}$) onto the column space of $\mathbf{X}$.

Since $\mathbf{P^2} = \mathbf{P}$, all of the eigenvalues of $\mathbf{P}$ are either 0 or 1 (Why?).

Therefore,

$$\text{rank}(\mathbf{P}) = \text{tr}(\mathbf{P}) = \text{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top) = \text{tr}((\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{X}) = \text{tr}(\mathbf{I}_p) = p$$

by the *cyclic property* of the trace operation.

Some authors denote $\mathbf{P}$ by $\mathbf{H}$, and call it the **hat matrix** because it "puts the hat on $\mathbf{Y}$". In fact, $\mathbf{P}$ is an orthogonal projection. Note that in the standard linear model above we may express the **fitted** values

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

as $\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}$.

Example 12.2 (Problem 1)**.**

1. *Show that* $\mathbf{P}$ *represents an orthogonal projection.*
2. *Show that* $\mathbf{P}$ *and* $\mathbf{I} - \mathbf{P}$ *are positive semi-definite.*
3. *Show that* $\mathbf{I} - \mathbf{P}$ *has rank* $n - p$ *and* $\mathbf{P}$ *has rank* p.

Solution: To see that $\mathbf{P}$ represents a projection, notice that $\mathbf{X}^\top\mathbf{X}$ is symmetric, so its inverse is also, so

$$\mathbf{P}^\top = \{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\}^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{P}$$

and

$$\mathbf{P}^2 = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{P}.$$

To see that $\mathbf{P}$ is an orthogonal projection, we must show that $\mathbf{PY}$ and $\mathbf{Y} - \mathbf{PY}$ are orthogonal. But from the results above,

$$(\mathbf{PY})^\top(\mathbf{Y} - \mathbf{PY}) = \mathbf{Y}^\top\mathbf{P}^\top(\mathbf{Y} - \mathbf{PY}) = \mathbf{Y}^\top\mathbf{PY} - \mathbf{Y}^\top\mathbf{PY} = \mathbf{0}.$$

$\mathbf{I} - \mathbf{P}$ is positive semi-definite since

$$\mathbf{x}^\top(\mathbf{I} - \mathbf{P})\mathbf{x} = \mathbf{x}^\top(\mathbf{I} - \mathbf{P})^\top(\mathbf{I} - \mathbf{P})\mathbf{x} = \|\mathbf{x} - \mathbf{Px}\|^{\mathbf{2}} \geq \mathbf{0}.$$

Similarly, $\mathbf{P}$ is positive semi-definite.

Theorem 12.3 (Cochran's theorem). *Let* $\mathbf{Z} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, *and let* $\mathbf{A_1}, \ldots, \mathbf{A_k}$ *be* $n \times n$ *positive semi–definite matrices with rank*$(\mathbf{A}_i) = r_i$, *such that*

$$\|\mathbf{Z}\|^2 = \mathbf{Z}^\top\mathbf{A}_1\mathbf{Z} + \ldots + \mathbf{Z}^\top\mathbf{A}_k\mathbf{Z}.$$

If $r_1 + \cdots + r_k = n$, *then* $\mathbf{Z}^\top\mathbf{A}_1\mathbf{Z}, \ldots, \mathbf{Z}^\top\mathbf{A}_k\mathbf{Z}$ *are independent, and*

$$\frac{\mathbf{Z}^\top\mathbf{A}_i\mathbf{Z}}{\sigma^2} \sim \chi^2_{r_i}, \quad i = 1, \ldots, k.$$

Example 12.4 (Problem 2)**.** *In the standard linear model above, find the maximum likelihood estimator $\hat{\sigma}^2$ of σ^2, and use Cochran's theorem to find its distribution.*

Solution: Differentiating the log-likelihood

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

we see that an MLE $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ satisfies

$$0 = \frac{\partial \ell}{\partial \sigma^2}\bigg|_{(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2,$$

so

$$\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \equiv \frac{1}{n}\|\mathbf{Y} - \mathbf{P}\mathbf{Y}\|^2,$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top$. Observe that

$$\|\mathbf{Y} - \mathbf{P}\mathbf{Y}\|^{\mathbf{2}} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{P})^\top(\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y},$$

and from the previous question we know that $\mathbf{I} - \mathbf{P}$ and $\mathbf{P}$ are positive semi-definite and of rank $n - p$ and p, respectively. We cannot apply Cochran's theorem directly since $\mathbf{Y}$ does not have mean zero. However, $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ does have mean zero and

$$\begin{aligned}
&(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I} - \mathbf{P})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&\quad = \mathbf{Y}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y} - \mathbf{2}\boldsymbol{\beta}^\top\mathbf{X}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top)\mathbf{Y} + \boldsymbol{\beta}^\top\mathbf{X}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top)\mathbf{X}\boldsymbol{\beta} \\
&\quad = \mathbf{Y}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y}.
\end{aligned}$$

Since

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^{\mathbf{2}} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I} - \mathbf{P})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{P}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

we may therefore apply Cochran's theorem to deduce that

$$\mathbf{Y}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I} - \mathbf{P})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \sim \sigma^2\chi^2_{n-p},$$

and hence

$$\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{P}\mathbf{Y}\|^{\mathbf{2}} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I} - \mathbf{P})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \sim \frac{\sigma^2}{n}\chi^2_{n-p}.$$

12.2.2 Testing hypotheis

Suppose that we want to test

$$H_0 : \beta_j = \beta_j^* \qquad \text{versus} \qquad H_0 : \beta_j \neq \beta_j^*$$

for some $j \in \{1, \ldots, p\}$, where β_j^* is a fixed number. We know that

$$\hat{\beta}_j \sim N(\beta_j, \zeta_{jj}\sigma^2),$$

where $(\mathbf{X}^\top \mathbf{X})^{-1} = ((\zeta_{ij}))_{p \times p}$. Thus, we know that

$$T = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\tilde{\sigma}^2 \zeta_{jj}}} \sim t_{n-p} \text{ under } H_0,$$

where we have used Theorem 12.1.

12.3 Testing for a component of $\boldsymbol{\beta}$ – not included in the final exam

Now partition $\mathbf{X}$ and $\boldsymbol{\beta}$ as

$$\underbrace{\mathbf{X}}_{n\times p} = (\underbrace{\mathbf{X}_0}_{n\times p_0} \ \ \underbrace{\mathbf{X}_1}_{n\times(p-p_0)}) \qquad \text{and} \qquad \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \end{pmatrix} \begin{array}{l} \updownarrow p_0 \\ \updownarrow p-p_0 \end{array}.$$

Suppose that we are interested in testing

$$H_0 : \boldsymbol{\beta}_1 = 0, \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq 0.$$

Then, under H_0, the MLEs of $\boldsymbol{\beta}_0$ and σ^2 are

$$\hat{\hat{\boldsymbol{\beta}}}_0 = (\mathbf{X}_\mathbf{0}^\top \mathbf{X}_\mathbf{0})^{-\mathbf{1}} \mathbf{X}_\mathbf{0}^\top \mathbf{Y}, \qquad \hat{\hat{\sigma}}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}_0 \hat{\hat{\boldsymbol{\beta}}}_0\|^2.$$

$\hat{\hat{\beta}}_0$ and $\hat{\hat{\sigma}}^2$ are independent. The fitted values under H_0 are

$$\hat{\hat{\mathbf{Y}}} = \mathbf{X}_0 \hat{\hat{\boldsymbol{\beta}}}_0 = \mathbf{X}_0 (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top \mathbf{Y} = \mathbf{P}_0 \mathbf{Y}$$

where $P_0 = \mathbf{X}_0 (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top$ is an orthogonal projection matrix of rank p_0.

The likelihood ratio statistic is

$$\begin{aligned} -2\log\Lambda &= 2\left\{-\frac{n}{2}\log\left(\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2\right) - \frac{n}{2} + \frac{n}{2}\log\left(\|\mathbf{Y} - \mathbf{X}_0\hat{\hat{\boldsymbol{\beta}}}_0\|^2\right) + \frac{n}{2}\right\} \\ &= n\log\left(\frac{\|\mathbf{Y} - \mathbf{X}_0\hat{\hat{\boldsymbol{\beta}}}_0\|^2}{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}\right) = n\log\left(\frac{\|\mathbf{Y} - \mathbf{P}_\mathbf{0}\mathbf{Y}\|^2}{\|\mathbf{Y} - \mathbf{P}\mathbf{Y}\|^2}\right). \end{aligned}$$

We therefore reject H_0 if the ratio of the residual sum of squares under H_0 to the residual sum of squares under H_1 is large.

Rather than use Wilks' theorem to obtain the asymptotic "null distribution" of the test statistic [which anyway depends on unknown σ^2], we can work out the exact distribution in this case.

Since $(\mathbf{Y} - \mathbf{PY})^\top(\mathbf{PY} - \mathbf{P_0Y}) = \mathbf{0}$, Pythagorean theorem gives that

$$\|\mathbf{Y} - \mathbf{PY}\|^2 + \|\mathbf{PY} - \mathbf{P_0Y}\|^2 = \|\mathbf{Y} - \mathbf{P_0Y}\|^2. \tag{31}$$

Using (31),

$$\begin{aligned}\frac{\|\mathbf{Y} - \mathbf{P_0Y}\|^2}{\|\mathbf{Y} - \mathbf{PY}\|^2} &= \frac{\|\mathbf{Y} - \mathbf{PY}\|^2}{\|\mathbf{Y} - \mathbf{PY}\|^2} + \frac{\|\mathbf{PY} - \mathbf{P_0Y}\|^2}{\|\mathbf{Y} - \mathbf{PY}\|^2} \\ &= 1 + \frac{\|\mathbf{PY} - \mathbf{P_0Y}\|^2}{\|\mathbf{Y} - \mathbf{PY}\|^2.}\end{aligned}$$

Consider the decomposition:

$$\|\mathbf{Y}\|^2 = \|\mathbf{Y} - \mathbf{PY}\|^2 + \|\mathbf{PY} - \mathbf{P_0Y}\|^2 + \|\mathbf{P_0Y}\|^2$$

and a similar one for $\mathbf{Z} = \mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0$.

Under H_0, $\mathbf{Z} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. This allows the use of Cochran's theorem to ultimately conclude that $\|\mathbf{PY} - \mathbf{P_0Y}\|^2$ and $\|\mathbf{Y} - \mathbf{PY}\|^2$ are independent $\sigma^2\chi^2_{p-p_0}$ and $\sigma^2\chi^2_{n-p}$ random variables, respectively.

Example 12.5 (Problem 3)**.** *Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are partitioned as $\mathbf{X} = (\mathbf{X}_0 | \ \mathbf{X}_1)$ and $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_0^T | \, \boldsymbol{\beta}_1^T)$ respectively (where $\boldsymbol{\beta}_0$ has p_0 components and $\boldsymbol{\beta}_1$ has $p - p_0$ components).*

1. *Show that*
$$\|\mathbf{Y}\|^2 = \|\mathbf{P_0Y}\|^2 + \|(\mathbf{P} - \mathbf{P_0})\mathbf{Y}\|^2 + \|\mathbf{Y} - \mathbf{PY}\|^2.$$

2. *Recall that the likelihood ratio statistic for testing*
$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0} \qquad \textit{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}$$
is a strictly increasing function of $\|(\mathbf{P} - \mathbf{P_0})\mathbf{Y}\|^2 / \|\mathbf{Y} - \mathbf{PY}\|^2$.

   Use Cochran's theorem to find the joint distribution of $\|(\mathbf{P} - \mathbf{P}_0)\mathbf{Y}\|^2$ and $\|\mathbf{Y} - \mathbf{PY}\|^2$ under H_0. How would you perform the hypothesis test?

   [Hint: $rank(\boldsymbol{P}) = p$, and $rank(\mathbf{I} - \mathbf{P}) = n - p$. Similar arguments give that $rank(\mathbf{P_0}) = p_0$.

Solution: 1. Recall that since $(\mathbf{Y}-\mathbf{P}\mathbf{Y})^\top(\mathbf{P}\mathbf{Y}-\mathbf{P}_0\mathbf{Y})=0$ Pythagorean theorem gives that

$$\begin{aligned}
\|\mathbf{Y}-\mathbf{P}\mathbf{Y}\|^2+\|\mathbf{P}\mathbf{Y}-\mathbf{P}_0\mathbf{Y}\|^2 &= \|\mathbf{Y}-\mathbf{P}_0\mathbf{Y}\|^2\\
&= (\mathbf{Y}-\mathbf{P}_0\mathbf{Y})^\top(\mathbf{Y}-\mathbf{P}_0\mathbf{Y})\\
&= \mathbf{Y}^\top\mathbf{Y}-2\mathbf{Y}^\top\mathbf{P}_0\mathbf{Y}+\mathbf{Y}^\top\mathbf{P}_0^\top\mathbf{P}_0\mathbf{Y}\\
&= \mathbf{Y}^\top\mathbf{Y}-\mathbf{Y}^\top\mathbf{P}_0\mathbf{P}_0^\top\mathbf{Y}\\
&= \|\mathbf{Y}\|^2-\|\mathbf{P}_0\mathbf{Y}\|^2
\end{aligned}$$

giving that

$$\|\mathbf{Y}-\mathbf{P}\mathbf{Y}\|^2+\|\mathbf{P}\mathbf{Y}-\mathbf{P}_0\mathbf{Y}\|^2+\|\mathbf{P}_0\mathbf{Y}\|^2=\|\mathbf{Y}\|^2$$

as desired.

2. Under H_0, the response vector $\mathbf{Y}$ has mean $\mathbf{X}_0\boldsymbol{\beta}_0$, and so $\mathbf{Z}=\mathbf{Y}-\mathbf{X}_0\boldsymbol{\beta}_0$ satisfies

$$\begin{aligned}
\|\mathbf{Z}\|^2 &= \|\mathbf{Z}-\mathbf{P}\mathbf{Z}\|^2+\|\mathbf{P}\mathbf{Z}-\mathbf{P}_0\mathbf{Z}\|^2+\|\mathbf{P}_0\mathbf{Z}\|^2\\
&= \mathbf{Z}^\top\mathbf{Z}-2\mathbf{Z}^\top\mathbf{P}\mathbf{Z}+\mathbf{Z}^\top\mathbf{P}^\top\mathbf{P}\mathbf{Z}+\mathbf{Z}^\top(\mathbf{P}-\mathbf{P}_0)^\top(\mathbf{P}-\mathbf{P}_0)\mathbf{Z}+\mathbf{Z}^\top\mathbf{P}_0^\top\mathbf{P}_0\mathbf{Z}\\
&= \mathbf{Z}^\top(\mathbf{I}-\mathbf{P})\mathbf{Z}+\mathbf{Z}^\top(\mathbf{P}-\mathbf{P}_0)\mathbf{Z}+\mathbf{Z}^\top\mathbf{P}_0\mathbf{Z}.
\end{aligned}$$

But

$$\begin{aligned}
\mathbf{Z}^\top(\mathbf{P}-\mathbf{P}_0)\mathbf{Z} &= (\mathbf{Y}-\mathbf{X}_0\boldsymbol{\beta}_0)^\top(\mathbf{P}-\mathbf{P}_0)(\mathbf{Y}-\mathbf{X}_0\boldsymbol{\beta}_0)\\
&= \mathbf{Y}^\top(\mathbf{P}-\mathbf{P}_0)\mathbf{Y}-2\boldsymbol{\beta}_0^\top\mathbf{X}_0^\top(\mathbf{P}-\mathbf{P}_0)\mathbf{Y}+\boldsymbol{\beta}_0^\top\mathbf{X}_0^\top(\mathbf{P}-\mathbf{P}_0)\mathbf{X}_0\boldsymbol{\beta}_0.
\end{aligned}$$

Since $\mathbf{X}_0\boldsymbol{\beta}_0\in U_0$ and $(\mathbf{P}-\mathbf{P}_0)\mathbf{Y}\in U_0^\perp$, and U_0 and $U_0^\perp$ are mutually orthogonal, and moreover $\mathbf{P}\mathbf{X}_0\boldsymbol{\beta}_0=\mathbf{P}_0\mathbf{X}_0\boldsymbol{\beta}_0=\mathbf{X}_0\boldsymbol{\beta}_0$, this gives

$$\mathbf{Z}^\top(\mathbf{P}-\mathbf{P}_0)\mathbf{Z}=\mathbf{Y}^\top(\mathbf{P}-\mathbf{P}_0)\mathbf{Y},$$

Similarly,

$$\begin{aligned}
\mathbf{Z}^\top(\mathbf{I}-\mathbf{P})\mathbf{Z} &= (\mathbf{Y}-\mathbf{X}_0\boldsymbol{\beta}_0)^\top(\mathbf{I}-\mathbf{P})(\mathbf{Y}-\mathbf{X}_0\boldsymbol{\beta}_0)\\
&= \mathbf{Y}^\top(\mathbf{I}-\mathbf{P})\mathbf{Y}-2\boldsymbol{\beta}_0^\top\mathbf{X}_0^\top(\mathbf{I}-\mathbf{P})\mathbf{Y}+\boldsymbol{\beta}_0^\top\mathbf{X}_0^\top(\mathbf{I}-\mathbf{P})\mathbf{X}_0\boldsymbol{\beta}_0\\
&= \mathbf{Y}^\top(\mathbf{I}-\mathbf{P})\mathbf{Y},
\end{aligned}$$

since $\mathbf{X}_0\boldsymbol{\beta}_0\in U_0$ and $(\mathbf{I}-\mathbf{P})\mathbf{Y}\in U^\perp\subseteq U_0^\perp$, while $(\mathbf{I}-\mathbf{P})\mathbf{X}_0\boldsymbol{\beta}_0=\mathbf{X}_0\boldsymbol{\beta}_0-\mathbf{X}_0\boldsymbol{\beta}_0=0$. Since

$$\operatorname{rank}(\mathbf{I}-\mathbf{P})+\operatorname{rank}(\mathbf{P}-\mathbf{P}_0)+\operatorname{rank}(\mathbf{P}_0)=n-p+p-p_0+p_0=n$$

we may therefore apply Cochran's theorem to deduce that under H_0, $\|(\mathbf{P}-\mathbf{P}_0)\mathbf{Y}\|^2$ and $\|\mathbf{Y}-\mathbf{P}\mathbf{Y}\|^2$ are independent with

$$\|(\mathbf{P}-\mathbf{P}_0)\mathbf{Y}\|^2=\mathbf{Y}^\top(\mathbf{P}-\mathbf{P}_0)\mathbf{Y}=\mathbf{Z}^\top(\mathbf{P}-\mathbf{P}_0)\mathbf{Z}\sim\sigma^2\chi^2_{p-p_0},$$

and

$$\|(\mathbf{I}-\mathbf{P})\mathbf{Y}\|^2 = \mathbf{Y}^\top(\mathbf{I}-\mathbf{P})\mathbf{Y} = \mathbf{Z}^\top(\mathbf{I}-\mathbf{P})\mathbf{Z} \sim \sigma^2\chi^2_{n-p}.$$

It follows that under H_0,

$$F = \frac{\frac{1}{p-p_0}\|(\mathbf{P}-\mathbf{P}_0)\mathbf{Y}\|^2}{\frac{1}{n-p}\|(\mathbf{I}-\mathbf{P})\mathbf{Y}\|^2} \sim F_{p-p_0,n-p},$$

so we may reject H_0 if $F > F_{p-p_0,n-p}(\alpha)$, where $F_{p-p_0,n-p}(\alpha)$ is the upper α-point of the $F_{p-p_0,n-p}$ distribution.

Thus under H_0,

$$F = \frac{\frac{1}{p-p_0}\|\mathbf{PY}-\mathbf{P_0Y}\|^2}{\frac{1}{n-p}\|\mathbf{Y}-\mathbf{PY}\|^2} \sim F_{p-p_0,n-p}.$$

When $\mathbf{X}_0$ has one less column than $\mathbf{X}$, say column k, we can leverage the normality of the MLE $\hat{\beta}_k$ in (30) to perform a t-test based on the statistic

$$T = \frac{\hat{\beta}_k}{\sqrt{\tilde{\sigma}^2 \mathrm{diag}[(\mathbf{X}^\top\mathbf{X})^{-1}]_k}} \sim t_{n-p} \text{ under } H_0 \quad \text{[i.e., } \beta_k = 0\text{]}.$$

[This is what R uses, though the more general F–statistic can also be used in this case.]

The above theory also shows that under H_1, $\frac{1}{n-p}\|\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2$ is an unbiased estimator of σ^2. This is usually used in preference to the MLE, $\hat{\sigma}^2$.

Example 12.6. *1.* **Multiple linear regression***:*

For countries $i = 1,\ldots,n$, consider how the fertility rate Y_i (births per 1000 females in a particular year) depends on

- *the gross domestic product per capita x_{i1}*
- *and the percentage of urban dwellers x_{i2}.*

The model

$$\log Y_i = \beta_0 + \beta_1 \log x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad i = 1,\ldots,n$$

with $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$*, is of linear model form* $Y = X\beta + \varepsilon$ *with*

$$Y = \begin{pmatrix} \log Y_1 \\ \vdots \\ \log Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & \log x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & \log x_{n1} & x_{n2} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

On the original scale of the response, this model becomes

$$Y = \exp(\beta_0) \exp(\beta_1 \log x_1) \exp(\beta_2 x_2) \varepsilon$$

Notice how the possibility of transforming variables greatly increases the flexibility of the linear model. [But see how using a log response assumes that the errors enter multiplicatively.]

12.4 One-way analysis of variance (ANOVA)

Consider measuring yields of plants under a control condition and $J - 1$ different treatment conditions.

The explanatory variable (factor) has J levels, and the response variables at level j are $Y_{j1}, \ldots, Y_{jn_j}$. The model that the responses are independent with

$$Y_{jk} \sim N(\mu_j, \sigma^2), \qquad j = 1, \ldots, J; \;\; k = 1, \ldots, n_j$$

is of linear model form, with

$$Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{J1} \\ \vdots \\ Y_{Jn_J} \end{pmatrix} \qquad X = \left(\begin{array}{ccccc} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ & & \vdots & & \\ 0 & \cdots & \cdots & 0 & 1 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \end{array} \right) \begin{array}{l} \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} n_1 \\ \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} n_2 \\ \\ \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} n_J \end{array} \qquad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{pmatrix}.$$

An alternative parameterization, emphasizing the differences between treatments, is

$$Y_{jk} = \mu + \alpha_j + \varepsilon_{jk}, \qquad j = 1, \ldots, J; \;\; k = 1, \ldots, n_j$$

where

- μ is the baseline or mean effect
- α_j is the effect of the j^{th} treatment (or the control $j = 1$).

Notice that the parameter vector $(\mu, \alpha_1, \alpha_2, \ldots, \alpha_J)^\top$ is not **identifiable**, since replacing μ with $\mu + 10$ and α_j by $\alpha_j - 10$ gives the same model. Either a

- **corner point** constraint $\alpha_1 = 0$ is used to emphasise the differences from the control, or the
- **sum–to–zero** constraint $\sum_{j=1}^{J} n_j \alpha_j = 0$

can be used to make the model identifiable. R uses corner point constraints.

If $n_j = K$, say, for all j, the data are said to be **balanced**.

We are usually interested in comparing the null model

$$H_0 : Y_{jk} = \mu + \varepsilon_{jk}$$

with that given above, which we call H_1, i.e., we wish to test whether the treatment conditions have an effect on the plant yield:

$$H_0 : \alpha = 0, \text{ where } \alpha = (\alpha_1, \ldots, \alpha_J), \qquad \text{against} \qquad H_1 : \alpha \neq 0.$$

Check that the MLE fitted values are

$$\hat{Y}_{jk} = \bar{Y}_j \equiv \frac{1}{n_j} \sum_{k=1}^{n_j} Y_{jk}$$

under H_1, whatever parameterization is chosen, and are

$$\hat{\hat{Y}}_{jk} = \bar{Y} \equiv \frac{1}{n} \sum_{j=1}^{J} n_j \bar{Y}_j, \qquad \text{where } n = \sum_{j=1}^{J} n_j,$$

under H_0.

Theorem 12.7. *(Partitioning the sum of squares) We have*

$$SS_{total} = SS_{within} + SS_{between},$$

where

$$SS_{total} = \sum_{j=1}^{J} \sum_{k=1}^{n_j} (Y_{jk} - \bar{Y})^2, \qquad SS_{within} = \sum_{j=1}^{J} \sum_{k=1}^{n_j} (Y_{jk} - \bar{Y}_j)^2, \qquad SS_{between} = \sum_{j=1}^{J} n_j (\bar{Y}_j - \bar{Y})^2.$$

Furthermore, SS_{within} has $\sigma^2\chi^2$-distribution with $(n - J)$ degrees of freedom and is independent of $SS_{between}$. Also, under H_0, $SS_{between} \sim \sigma^2 \chi^2_{J-1}$.

Our linear model theory says that we should test H_0 by referring

$$F = \frac{\frac{1}{J-1}\sum_{j=1}^{J} n_j(\bar{Y}_j - \bar{Y})^2}{\frac{1}{n-J}\sum_{j=1}^{J}\sum_{k=1}^{n_j}(Y_{jk} - \bar{Y}_j)^2} \equiv \frac{\frac{1}{J-1}S_2}{\frac{1}{n-J}S_1}$$

to $F_{J-1,n-J}$, where S_1 is the "within groups" sum of squares and S_2 is the "between groups" sum of squares. We have the following ANOVA table.

Source of variation	Degrees of freedom	Sum of squares	F–statistic
Between groups	$J-1$	S_2	$F = \frac{\frac{1}{J-1}S_2}{\frac{1}{n-J}S_1}$
Within groups	$n-J$	S_1	
Total	$n-1$	$S_1 + S_2 = \sum_{j=1}^{J}\sum_{k=1}^{n_j}(Y_{jk} - \bar{Y})^2$	

13 Nonparametrics

13.1 The sample distribution function

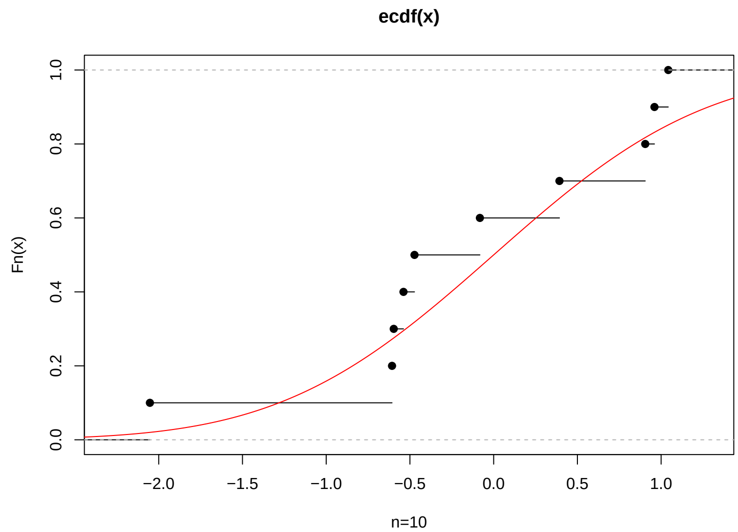
Let $X_1, \ldots, X_n$ be i.i.d F, where F is an unknown distribution function.

Question: We want to estimate F without assuming any specific parametric form for F.

Empirical distribution function (EDF): For each $x \in \mathbb{R}$, we define $F_n(x)$ as the proportion of observed values in the sample that are less than of equal to x, i.e.,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i).$$

The function F_n defined in this way is called the *sample/empirical distribution function.*



Idea: Note that

$$F(x) = \mathbb{P}(X \le x) = \mathbb{E}[I_{(-\infty,x]}(X)].$$

Thus, given a random sample, we can find an *unbiased* estimator of $F(x)$ by looking at the proportion of times, among the X_i's, we observe a value $\le x$.

By the WLLN, we know that

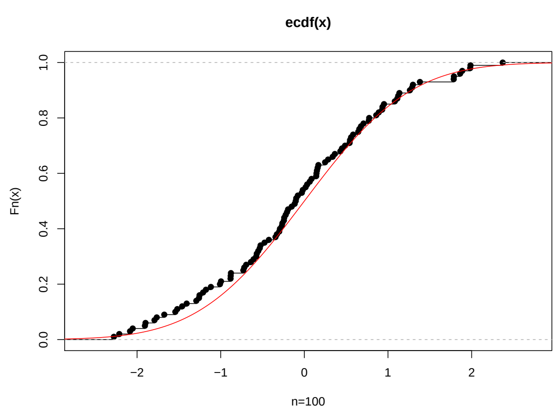
$$F_n(x) \xrightarrow{p} F(x), \qquad \text{for every } x \in \mathbb{R}.$$

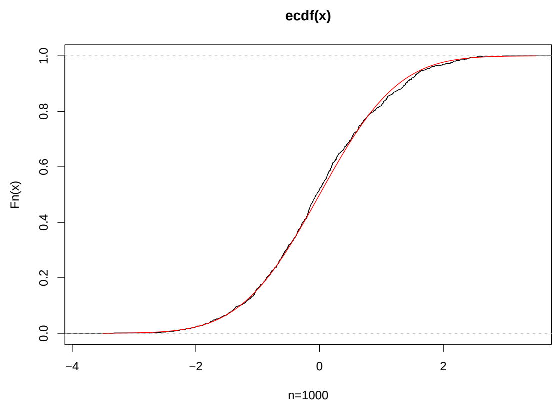
Theorem 13.1. Glivenko-Cantelli Theorem. *Let F_n be the sample c.d.f from an i.i.d sample $X_1, \ldots, X_n$ from the c.d.f F. Then,*

$$D_n := \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{p} 0.$$

By the CLT, we have

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x))), \qquad \text{for every } x \in \mathbb{R}.$$





As $F_n(x) \xrightarrow{p} F(x)$ for all $x \in \mathbb{R}$, we can also say that

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F_n(x)(1 - F_n(x))}} \xrightarrow{d} N(0,1), \qquad \text{for every } x \in \mathbb{R}.$$

Thus, an asymptotic $(1-\alpha)$ CI for $F(x)$ is

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

Likewise, we can also test the hypothesis $H_0 : F(x) = F_0(x)$ versus $H_1 : F(x) \neq F_0(x)$ for some known fixed c.d.f F_0, and $x \in \mathbb{R}$.

13.2 The Kolmogorov-Smirnov goodness-of-fit test

Suppose that we wish to test the simple null hypothesis that the unknown c.d.f F is actually a particular continuous c.d.f F^* against the alternative that the actual c.d.f is not F^*, i.e.,

$$H_0 : F(x) = F^*(x) \quad \text{for } x \in \mathbb{R}, \qquad H_0 : F(x) \neq F^*(x) \quad \text{for some } x \in \mathbb{R}.$$

This is a nonparametric ("infinite" dimensional) problem.

Let

$$D_n^* = \sup_{x \in \mathbb{R}} |F_n(x) - F^*(x)|.$$

D_n^* is the maximum difference between the sample c.d.f F_n and the hypothesized c.d.f F^*.

We should reject H_0 when

$$n^{1/2} D_n^* \geq c_\alpha.$$

This is called the **Kolmogorov-Smirnov** test.

How do we find c_α?

When H_0 is true, the distribution of D_n^* will have a certain distribution that is the same for every possible continuous c.d.f F. (Why?)

Note that, under H_0,

$$
\begin{aligned}
D_n^* &= \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n I(X_i \le x) - F^*(x) \right| \\
&= \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n I(F^*(X_i) \le F^*(x)) - F^*(x) \right| \\
&= \sup_{F^*(x) \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n I(U_i \le F^*(x)) - F^*(x) \right| = \sup_{t \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^n I(U_i \le t) - t \right| \\
&= \sup_{t \in [0,1]} \left| F_{n,U}(t) - t \right|,
\end{aligned}
$$

where $U_i := F^*(X_i) \sim \text{Uniform}(0,1)$ (i.i.d) and $F_{n,U}$ is the EDF of the U_i's. Thus, D_n^* is *distribution-free.*

Theorem 13.2. *(Distribution-free property) Under H_0, the distribution of D_n^* is the same for all continuous distribution functions F.*

We also have the following theorem.

Theorem 13.3. *Under H_0, as $n \to \infty$,*

$$n^{1/2} D_n^* \xrightarrow{d} H, \tag{32}$$

where H is a valid c.d.f.

In fact, the exact sampling distribution of the KS statistic, under H_0, can be approximated by *simulations*, i.e., we can draw *n data points from a Uniform(0,1) distribution and recompute the test statistic* multiple times.

13.2.1 The Kolmogorov-Smirnov test for two samples

Consider a problem in which a random sample of m observations $X_1, \ldots, X_m$ is taken from the unknown c.d.f F, and an independent random sample of n observations $Y_1, \ldots, Y_n$ is taken from another distribution with unknown c.d.f G.

It is desired to test the hypothesis that both these functions, F and G, are identical, without specifying their common form. Thus the hypotheses we want to test are:

$$H_0 : F(x) = G(x) \quad \text{for } x \in \mathbb{R}, \qquad H_0 : F(x) \neq G(x) \quad \text{for some } x \in \mathbb{R}.$$

We shall denote by F_m the EDF of the observed sample $X_1, \ldots, X_m$, and by G_n the EDF of the sample $Y_1, \ldots, Y_n$.

We consider the following statistic:

$$D_{m,n} = \sup_{x \in \mathbb{R}} |F_m(x) - G_n(x)|.$$

When H_0 holds, the sample EDFs F_m and G_n will tend to be close to each other. In fact, when H_0 is true, it follows from the Glivenko-Cantelli lemma that

$$D_{m,n} \xrightarrow{p} 0 \qquad \text{as } m, n \to \infty.$$

$D_{m,n}$ is also *distribution-free* (why?)

Theorem 13.4. *Under H_0,*

$$\left(\frac{mn}{m+n}\right)^{1/2} D_{m,n} \xrightarrow{d} H,$$

where H is a the same c.d.f as in (32).

A test procedure that rejects H_0 when

$$\left(\frac{mn}{m+n}\right)^{1/2} D_{m,n} \geq c_\alpha,$$

where c_α (is the $(1-\alpha)$-quantile of H) is an appropriate constant, is called a *Kolmogorov-Smirnov two sample test*.

Exercise: Show that this test statistic is also distribution-free under H_0. Thus, the critical of the test can be obtained via simulations.

13.3 Bootstrap

Example 1: Suppose that we model our data $\mathbf{X} = (X_1, \ldots, X_n)$ as coming from some distribution with c.d.f F having median θ.

Suppose that we are interested in using the sample median M as an estimator of θ.

We would like to estimate the MSE (mean squared error) of M (as an estimator of θ), i.e., we would like to estimate

$$\mathbb{E}[(M - \theta)^2].$$

We may also be interested in finding a confidence interval for θ.

Example 2: Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from a distribution F. We are interested in the distribution of the sample correlation coefficient:

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2\right]^{1/2}}.$$

We might be interested in the variance of R, or the bias of R, or the distribution of R as an estimator of the correlation ρ between X and Y.

Question: How do we get a handle on these problems?

How would we do it if an *oracle* told us F?

Bootstrap: The bootstrap is a method of replacing (plug-in) an unknown distribution function F with a known distribution in probability/expectation calculations.

If we have a sample of data from the distribution F, we first approximate F by $\hat{F}$ and then perform the desired calculation.

If $\hat{F}$ is a good approximation of F, then bootstrap can be successful.

13.3.1 Bootstrap in general

Let $\eta(\mathbf{X}, F)$ be a quantity of interest that possibly depends on both the distribution F and a sample $\mathbf{X}$ drawn from F.

In general, we might wish to estimate the mean or a quantile or some other probabilistic feature or the entire *distribution* of $\eta(\mathbf{X}, F)$.

The bootstrap estimates $\eta(\mathbf{X}, F)$ by $\eta(\mathbf{X}^*, \hat{F})$, where $\mathbf{X}^*$ is a random sample drawn from the distribution $\hat{F}$, where $\hat{F}$ is some distribution that we think is close to F.

How do we find the distribution of $\eta(\mathbf{X}^*, \hat{F})$?

In most cases, the distribution of $\eta(\mathbf{X}^*, \hat{F})$ is difficult to compute, but we can approximate it easily by simulation.

The bootstrap can be broken down in the following simple steps:

- Find a "good" estimator $\hat{F}$ of F.
- Draw a large number (say, v) of random samples $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(v)}$ from the distribution $\hat{F}$ and then compute $T^{(i)} = \eta(\mathbf{X}^{*(i)}, \hat{F})$, for $i = 1, \ldots, v$.
- Finally, compute the desired feature of $\eta(\mathbf{X}^*, \hat{F})$ using the sample c.d.f of the values $T^{(1)}, \ldots, T^{(v)}$.

13.3.2 Parametric bootstrap

Example 1: (Estimating the standard deviation of a statistic)

Suppose that $X_1, \ldots, X_n$ is random sample from $N(\mu, \sigma^2)$.

Suppose that we are interested in the parameter

$$\theta = \mathbb{P}(X \le c) = \Phi\left(\frac{c - \mu}{\sigma}\right),$$

where c is a given known constant.

What is the MLE of θ?

The MLE of θ is

$$\hat{\theta} = \Phi\left(\frac{c - \bar{X}}{\hat{\sigma}}\right).$$

Question: How do we calculate the standard deviation of $\hat{\theta}$? There is no easy closed form expression for this.

Solution: We can bootstrap!

Draw many (say v) bootstrap samples of size n from $N(\bar{X}, \hat{\sigma}^2)$. For the i-th sample we compute a sample average $\bar{X}^{*(i)}$, a sample standard deviation $\hat{\sigma}^{*(i)}$.

Finally, we compute

$$\hat{\theta}^{*(i)} = \Phi\left(\frac{c - \bar{X}^{*(i)}}{\hat{\sigma}^{*(i)}}\right).$$

We can estimate the mean of $\hat{\theta}$ by

$$\bar{\theta}^* = \frac{1}{v}\sum_{i=1}^{v} \hat{\theta}^{*(i)}.$$

The standard deviation of $\hat{\theta}$ can then be estimated by the sample standard deviation of the $\hat{\theta}^{*(i)}$ values, i.e.,

$$\left[\frac{1}{v}\sum_{i=1}^{v}(\hat{\theta}^{*(i)} - \bar{\theta}^*)^2\right]^{1/2}.$$

Example 2: (Comparing means when variances are unequal) Suppose that we have two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from two possibly different normal populations. Suppose that

$$X_1, \ldots, X_m \text{ are i.i.d } N(\mu_1, \sigma_1^2) \qquad \text{and} \qquad Y_1, \ldots, Y_n \text{ are i.i.d } N(\mu_2, \sigma_2^2).$$

Suppose that we want to test

$$H_0 : \mu_1 = \mu_2 \qquad \text{versus} \qquad H_1 : \mu_1 \neq \mu_2.$$

We can use the test statistic

$$U = \frac{(m+n-2)^{1/2}(\bar{X}_m - \bar{Y}_n)}{\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}(S_X^2 + S_Y^2)^{1/2}}.$$

Note that as $\sigma_1^2 \neq \sigma_2^2$, U does not necessarily follow a t-distribution.

How do we find the cut-off value of the test?

The parametric bootstrap can proceed as follows:

First choose a large number v, and for $i = 1, \ldots, v$, simulate $(\bar{X}_m^{*(i)}, \bar{Y}_n^{*(i)}, S_X^{2*(i)}, S_Y^{2*(i)})$, where all four random variables are independent with the following distributions:

- $\bar{X}_m^{*(i)} \sim N(0, \hat{\sigma}_1^2/m)$.
- $\bar{Y}_n^{*(i)} \sim N(0, \hat{\sigma}_2^2/n)$.
- $S_X^{2*(i)} \sim \hat{\sigma}_1^2\, \chi^2_{m-1}$.
- $S_Y^{2*(i)} \sim \hat{\sigma}_2^2\, \chi^2_{n-1}$.

Then we compute

$$U^{*(i)} = \frac{(m+n-2)^{1/2}(\bar{X}_m^{*(i)} - \bar{Y}_n^{*(i)})}{\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}(S_X^{2*(i)} + S_Y^{2*(i)})^{1/2}}$$

for each i.

We approximate the null distribution of U by the distribution of the $U^{*(i)}$'s.

Let c^* be the $\left(1 - \frac{\alpha}{2}\right)$-quantile of the distribution of $U^{*(i)}$'s. Thus we reject H_0 if

$$|U| > c^*.$$

13.3.3 The nonparametric bootstrap

Back to Example 1: Let $X_1, \ldots, X_n$ be a random sample from a distribution F.

Suppose that we want a CI for the median θ of F.

We can base a CI on the sample median M.

We want the distribution of $M - \theta$!

Let $\eta(\mathbf{X}, F) = M - \theta$.

We approximate the $\alpha/2$ and the $1 - \alpha/2$ quantiles of the distribution of $\eta(\mathbf{X}, F)$ by that of $\eta(\mathbf{X}^*, \hat{F})$.

We may choose $\hat{F} = F_n$, the empirical distribution function. Thus, our method can be broken in the following steps:

- Choose a large number v and simulate many samples $\mathbf{X}^{*(i)}$, for $i = 1, \ldots, n$, from F_n. This reduces to drawing **with replacement sampling** from $\mathbf{X}$.
- For each sample we compute the sample median $M^{*(i)}$ and then find the sample quantiles of $\{M^{*(i)} - M\}_{i=1}^{v}$.

Back to Example 2: Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from a distribution F. We are interested in the distribution of the sample correlation coefficient:

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]^{1/2}}.$$

We might be interested in the bias of R, i.e., $\eta(\mathbf{X}, \mathbf{Y}, F) = R - \rho$.

Let F_n be the discrete distribution that assigns probability $1/n$ to each of the n data points.

Thus, our method can be broken in the following steps:

- Choose a large number v and simulate many samples from F_n. This reduces to drawing **with replacement sampling** from the original paired data.
- For each sample we compute the sample correlation coefficient $R^{*(i)}$ and then find the sample quantiles of $\{T^{*(i)} = R^{*(i)} - R\}_{i=1}^{v}$.
- We estimate the mean of $R - \rho$ by the average $\frac{1}{n}\sum_{i=1}^{v} T^{*(i)}$.

14 Review

14.1 Statistics

- Estimation: Maximum likelihood estimation (MLE); large sample properties of the MLE; Information matrix; method of moments.
- Consistency of estimators; Mean squared error and its decomposition; unbiased estimation; minimum variance unbiased estimator; sufficiency.
- Bayes estimators: prior distribution; posterior distribution.
- Sampling distribution of an estimator; sampling from a normal distribution; t-distribution.

Exercise: Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution with mean 0 and unknown variance σ^2. Determine the asymptotic distribution of the statistic $T = \left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)^{-1}$.

Solution: We know that X_i^2's are i.i.d with mean $\mathbb{E}(X_1^2) = \sigma^2$ and $\text{Var}(X_1^2) = \mathbb{E}(X_1^4) - [\mathbb{E}(X_1^2)]^2 = 2\sigma^4$. Note that X_i^2's have a χ_1^2 distribution. Thus, by the CLT, we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \sigma^2\right) \xrightarrow{d} N(0, 2\sigma^4).$$

Let $g(x) = x^{-1}$. Thus, $g'(x) = -x^{-2}$. Therefore,

$$\sqrt{n}(T - \sigma^{-2}) \xrightarrow{d} N(0, 2\sigma^4 \cdot \sigma^{-8}).$$

Exercise: Consider i.i.d observations $X_1, \ldots, X_n$ where each X_i follows a normal distribution with mean and variance both equal to $1/\theta$, where $\theta > 0$. Thus,

$$f_\theta(x) = \frac{\sqrt{\theta}}{\sqrt{2\pi}} \exp\left[-\frac{(x - \theta^{-1})^2}{2\theta^{-1}}\right].$$

Show that the MLE is one of the solutions to the equation:

$$\theta^2 W - \theta - 1 = 0,$$

where $W = n^{-1}\sum_{i=1}^n X_i^2$. Determine which root it is and compute its approximate variance in large samples.

Solution: We have the log-likelihood (up to a constant) as

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

Therefore, the score equation is

$$\begin{aligned} & \frac{\partial \ell}{\partial \theta} = \frac{n}{2\theta} - \frac{1}{2}\sum_{i=1}^{n} X_i^2 + \frac{n}{2\theta^2} = 0 \\ \text{i.e.,} \quad & \frac{1}{2\theta} - \frac{1}{2}W + \frac{1}{2\theta^2} = 0 \\ \text{i.e.,} \quad & W\theta^2 - \theta - 1 = 0 \end{aligned}$$

The two roots are given by

$$\frac{1 \pm \sqrt{1+4W}}{2W}$$

and the admissible root is

$$\hat{\theta}_{MLE} = \frac{1+\sqrt{1+4W}}{2W}.$$

We know that

$$\hat{\theta}_{MLE} \sim N\left(\theta, \frac{1}{nI(\theta)}\right) \qquad \text{(approximately).}$$

Thus the approximate variance of $\hat{\theta}_{MLE}$ is $\frac{1}{nI(\theta)}$, where

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X_1)\right] = \frac{1}{2\theta^2} + \frac{1}{\theta^3}.$$

- Confidence intervals; Cramer-Rao information inequality.

Exercise: A biologist is interested in measuring the ratio of mean weight of animals of two species. However, the species are extremely rare and after much effort she succeeds in measuring the weight of one animal from the first species and one from the second. Let X_1 and X_2 denote these weights. It is assumed that $X_i \sim N(\theta_i, 1)$, for $i = 1, 2$. Interest lies in estimating θ_1/θ_2.

Compute the distribution of

$$h(X_1, X_2, \theta_1, \theta_2) = \frac{\theta_2 X_1 - \theta_1 X_2}{\sqrt{\theta_1^2 + \theta_2^2}}.$$

Is

$$\frac{X_1 - (\theta_1/\theta_2)X_2}{\sqrt{(\theta_1/\theta_2)^2 + 1}}$$

a pivot? Discuss how you can construct a confidence set for the ratio of mean weights.

Solution: Note that $\theta_2 X_1 - \theta_1 X_2 \sim N(0, \theta_1^2 + \theta_2^2)$ as

$$\mathbb{E}(\theta_2 X_1 - \theta_1 X_2) = \theta_2\theta_1 - \theta_1\theta_2 = 0$$

and

$$\mathrm{Var}(\theta_2 X_1 - \theta_1 X_2) = \mathrm{Var}(\theta_2 X_1) + \mathrm{Var}(\theta_1 X_2) = \theta_2^2 + \theta_1^2.$$

Thus,

$$h(X_1, X_2, \theta_1, \theta_2) = \frac{\theta_2 X_1 - \theta_1 X_2}{\sqrt{\theta_1^2 + \theta_2^2}} \sim N(0,1).$$

Now,

$$\frac{X_1 - (\theta_1/\theta_2)X_2}{\sqrt{(\theta_1/\theta_2)^2 + 1}} = \frac{\theta_2 X_1 - \theta_1 X_2}{\sqrt{\theta_1^2 + \theta_2^2}} \sim N(0,1)$$

and is thus indeed a pivot.

To get a confidence set for $\eta := \theta_1/\theta_2$, we know that

$$\mathbb{P}\left[-z_{\alpha/2} \leq \frac{X_1 - \eta X_2}{\sqrt{\eta^2+1}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

$$\text{i.e.,} \quad \mathbb{P}\left[\frac{|X_1 - \eta X_2|}{\sqrt{\eta^2+1}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

$$\text{i.e.,} \quad \mathbb{P}\left[(X_1 - \eta X_2)^2 - (\eta^2+1)^2 z_{\alpha/2} \leq 0\right] = 1 - \alpha.$$

Thus,

$$\{\eta : (X_1 - \eta X_2)^2 - (\eta^2+1)^2 z_{\alpha/2} \leq 0\}$$

gives a level $(1-\alpha)$ confidence set for η. This can be expressed explicitly in terms of the roots of the quadratic equation involved.

- Hypothesis testing: Null and the alternative hypothesis; rejection region; Type I and II errorx; power function; size (level) of a test; equivalence of tests and confidence sets; p-value; Neyman-Pearson lemma; uniformly most powerful test.
- t-test; F-test; likelihood ratio test
- Linear models: method of least squares; regression; Simple linear regression; inference on β_0 and β_1; mean response; prediction interval;
- General linear model; MLE; projection; one-way ANOVA

Exercise: Processors usually preserve cucumbers by fermenting them in a low-salt brine (6% to 9% sodium chloride) and then storing them in a high-salt brine until they are used by processors to produce various types of pickles. The high-salt brine is needed to retard softening of the pickles and to prevent freezing when

Weeks (X) in Storage at 72^o F	0	4	14	32	52
Firmness (Y) in pounds	19.8	16.5	12.8	8.1	7.5

they are stored outside in northern climates. Data showing the reduction in firmness of pickles stored over time in a low-salt brine (2% to 3%) are given in the following table.

(a) Fit a least-squares line to the data.

(b) Compute R^2 to evaluate the goodness of the fit to the data points?

(c) Use the least-squares line to estimate the mean firmness of pickles stored for 20 weeks.

(d) Determine the 95% CI for β_1.

(e) Test the null hypothesis that Y does not depend on X linearly.

Solution: (a) Fit a least-squares line to the data.

$$\begin{aligned}\widehat{\beta_1} &= \frac{S_{xy}}{S_{xx}} = \frac{-425.48}{1859.2} = -0.229 \\ \widehat{\beta_0} &= \overline{y} - \widehat{\beta_1}\overline{x} = 12.94 - (-0.229)(20.4) = 17.612 \\ \widehat{y} &= 17.612 - 0.229x.\end{aligned}$$

(b) Compute R^2 to evaluate the goodness of the fit to the data points?

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{15.4}{112.772} = 0.863.$$

(c) Use the least-squares line to estimate the mean firmness of pickles stored for 20 weeks.

$$\widehat{y}(20) = 17.612 - (0.229)(20) = 13.0$$

(d) Determine the 95% CI for β_1.

The 95% CI for β_1 is given by

$$\widehat{\beta_1} \pm t_{0.025,3} SE(\widehat{\beta_1}) \quad \text{where} \quad SE(\widehat{\beta_1}) = \frac{s}{\sqrt{S_{xx}}}.$$

We have $s = \sqrt{SSE/3} = \sqrt{(15.4)/3} = 2.266$, thus $SE(\widehat{\beta_1}) = \frac{2.66}{\sqrt{1859.2}} = 0.052$. Thus the 95% CI for β_1 is

$$-0.229 \pm (3.18)(0.052) = [-0.396, -0.062]$$

(e) Test the null hypothesis that Y does not depend on X linearly.

We test the hypothesis

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_a: \beta_1 \neq 0$$

with level at $\alpha = 0.05$. This can be tested with t-statistic

$$T = \frac{\widehat{\beta}_1}{SE(\widehat{\beta}_1)} \quad \text{and} \quad RR: |t| > t_{0.025,3} = 3.18.$$

The observed $t = \dfrac{-0.229}{0.052} = -4.404$, which is in the rejection region. Thus we reject the hypothesis that $\beta_1 = 0$. This means based on the data we reject H_0.

Exercise: A manager wishes to determine whether the mean times required to complete a certain task differ for the three levels of employee training. He randomly selected 10 employees with each of the three levels of training (Beginner, Intermediate and Advanced). Do the data provide sufficient evidence to indicate that the mean times required to complete a certain task differ for at least two of the three levels of training? The data is summarized in the following table. Use the level $\alpha = 0.05$.

	$\overline{x_i}$	s_i^2
Advanced	24.2	21.54
Intermediate	27.1	18.64
Beginner	30.2	17.76

Solution: Let α_i denote the mean effect of ith training level; advanced=1, intermediate=2 and beginner=3. We test the hypothesis

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 \quad \text{vs.} \quad H_a : \alpha_i \neq \alpha_j \quad \text{for some } i \text{ and } j$$

We have

$$\begin{aligned}
\overline{x_{1.}} &= 24.2 \qquad \overline{x}_{2.} = 27.1 \qquad \overline{x}_{3.} = 30.2 \\
\overline{x}_{..} &= \frac{1}{3}(24.2 + 27.1 + 30.2) = 27.17 \\
SSB &= 10\left((24.2 - 27.17)^2 + (27.1 - 27.17)^2 + (30.2 - 27.17)^2\right) = 180.1 \\
SSW &= 9(21.54 + 18.64 + 17.76) = 521.46
\end{aligned}$$

Thus we have the following ANOVA-table:

Source of variations	df	SS	MS	F
Treatments	2	180.1	90.03	4.67
Errors	27	521.46	19.31	
Total	29	683.52		

Since the observed $f = 4.67$ is in $RR : f > f_{0.05,2,27} = 3.35$, we reject the H_0. Thus the levels of training appear to have different effects on the mean times required to complete the task.

- The empirical distribution function; goodness-of-fit-tests; Kolmogorov-Smirnov tests.

Read the following sections from the text book for the final:

- Chapter 6 excluding 6.4
- Chapter 7 excluding 7.8, 7.9
- Chapter 8 excluding 8.6
- Chapter 9 excluding 9.3, 9.8, 9.9
- Chapter 11 excluding 11.4, 11.7, 11.8

Thank you!

Please complete course evaluations!

Questions?