ST202.2: Statistical Inference

Tay Meshkinyar

Dr. Milt Mavrakakis

The London School of Economics and Political Science

2021-2022

Contents

Chapter 1

Introduction

Welcome to my transcribed set of lecture notes for ST202: Probability, Distribution Theory, and Inference. This document uses an edited version of the theme used in Gilles Castel's differential geometry notes. Much of the workflow used to write these notes was ported from his lightning-fast, elegant setup on Linux. Check out his github here, as well as his personal website. You can also find the most up to date version of these notes here. This chapter serves mainly for theme consistency, and to match the numbering of the course textbook. The course thus begins with Chapter 2.

Chapter 7

Sample Moments and Quantiles

Chapter 7 is the first topic covered this term, and the bulk of Chapter 6 is the *last* topic covered this term. Chapter 6 is thus at the *end* of this document. I had to compromise a little between chronological consistency and consistency with the textbook, so unfortunately this half of the course might be a little weird as a reference.

7.1 Week 12: Lecture 1

7.1.1 First bit of Ch. 6

Tue 18 Jan 14:00

First, we start with some data:

$$y_1, y_2, \ldots, y_n \quad (\mathbf{y}) \quad \text{with } y_i \in \mathbb{R}$$

This is our *observed sample*. We can think of these as single realizations of our sample:

$$Y_1, Y_2, \ldots, Y_n \quad (\mathbf{Y}).$$

We let $\Theta = (\theta_1, \ldots, \theta_r)^T$ be our parameters.

Definition 7.1.1. A **random sample** is a set of IID random variables $\{Y_1, \ldots, Y_n\}$ such that

$$Y_1, \ldots, Y_n \sim F_Y,$$

for some distribution F_Y. An **observed sample**, denoted $y_1, \ldots, y_n$, is a set of possible values for each random variable.

Note. When we take a random sample, we do so *without* replacement.

7.1.2 Sample Moments

Definition 7.1.2. Let Y be a random variable with moment and central moment μ_r', μ_r and MGF $M_Y(t)$. The **sample mean**, given some random sample $Y_1, \ldots, Y_n$, is

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Some properties of $\bar{Y}$:

(i) $\mathbb{E}(\bar{Y}) = \mu$, the population mean:

$$\mathbb{E}(\bar{Y}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i) = \frac{1}{n}n\mu = \mu.$$

(ii) Note that $\mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{n}$, where σ^2 is the population variance (as long as $\sigma^2 < \infty$. Then

$$\begin{aligned}\mathrm{Var}(\bar{Y}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i)\\ &= \frac{1}{n^2}n\sigma^2\\ &= \frac{\sigma^2}{n}.\end{aligned}$$

7.1.3 The Central Limit Theorem

Theorem 7.1.3 (Central Limit Theorem)**.** Given a random sample $Y_1, \ldots, Y_n$ with $\mathbb{E}(Y_1) = \mu$, $\mathrm{Var}(Y_1) = \sigma^2 < \infty$, and $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$,

$$\frac{\bar{Y}_n - \mu}{\sqrt{\sigma^2/n}} \xrightarrow{d} \mathrm{Normal}(0,1),$$

as $n \to \infty$.

Proof. Let

$$Z_n = \frac{\bar{Y}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}},$$

and note that

$$Z_n = \frac{\bar{Y}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{n\bar{Y}_n - n\mu}{\sqrt{n\sigma^2}}.$$

Let $S_n = n\bar{Y}_n$. Note that

$$\begin{aligned} M_{Z_n} &= \mathbb{E}(e^{tZ_n}) \\ &= \mathbb{E}\left[\exp\left(t\frac{S_n - n\mu}{\sqrt{n\sigma^2}}\right)\right] \\ &= \mathbb{E}\left[\exp\left(\frac{t}{\sqrt{n\sigma^2}}S_n\right)\right]\exp\left(-\frac{n\mu t}{\sqrt{n\sigma^2}}\right) \\ &= M_{S_n}\left(\frac{t}{\sqrt{n\sigma^2}}\right)\exp\left(\frac{-\sqrt{n}\mu t}{\sigma}\right) \\ &= \left[M_{Y_1}\left(\frac{t}{\sqrt{n\sigma^2}}\right)\right]^n \exp\left(\frac{-\sqrt{n}\mu t}{\sigma}\right). \end{aligned}$$

The last equality is justified through the IID property of random samples. Now observe that

$$\begin{aligned} K_{Z_n}(t) &= nK_{Y_1}\left(\frac{t}{\sqrt{n\sigma^2}}\right) - \frac{\mu t\sqrt{n}}{\sigma} \\ &= n\left(\mu\frac{t}{\sqrt{n\sigma^2}} + \frac{\sigma^2}{2}\left(\frac{t}{\sqrt{n\sigma^2}}\right) + \left(\begin{matrix}\text{terms in} \\ \left(\frac{1}{n}\right)^{3/2} \text{ and higher}\end{matrix}\right)\right) - \frac{\mu t\sqrt{n}}{\sigma} \\ &= \frac{\mu t\sqrt{n}}{\sigma} + \frac{t^2}{2} + \left(\begin{matrix}\text{terms in} \\ \left(\frac{1}{n}\right)^{1/2} \text{ and higher}\end{matrix}\right) - \frac{\mu t\sqrt{n}}{\sigma} \\ &= \frac{t^2}{2} + \left(\begin{matrix}\text{terms in} \\ \left(\frac{1}{n}\right)^{1/2} \text{ and higher}\end{matrix}\right) \end{aligned}$$

So $K_{Z_n}(t) \to \frac{t^2}{2}$ as $n \to \infty$, i.e. the CGF of Normal$(0, 1)$. Since the CGF of a distribution characterizes that distribution,

$$Z_n \to \text{Normal}(0, 1)$$

as $n \to \infty$. □

7.2 Week 12: Lecture 2

7.2.1 More on Sample Moments

Thu 20 Jan 10:00

Recall. Let $Y_1, \ldots, Y_n \sim F_Y$. Then

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Remark 7.2.1. If you're interested in something without having access to the underlying distribution of the population, then you use the sample average. Sample averages converge to their population equivalent as the sample size increases.

Recall. Moments and central moments:

$$\mu_1' = \mathbb{E}(Y)$$
$$\mu_r' = \mathbb{E}(Y^r)$$
$$\mu_r = \mathbb{E}[(Y - \mathbb{E}(Y))^r].$$

What is the sample equivalent of moments?

Definition 7.2.2. Let

$$m_r' = \frac{1}{n}\sum_{i=1}^{n} Y_i^r$$
$$m_r = \frac{1}{n}\sum_{i=1}^{m} (Y_i - \bar{Y})^2.$$

Then m_r' is the rth **sample moment** and m_r is the rth sample central moment.

Example 7.2.3. We have

$$m_1' = \bar{Y}$$
$$m_2 = \frac{n-1}{n} s^2,$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

the sample variance. ◇

7.2.2 New Tricks, New Properties

Let $Y_1, \ldots, Y_n \sim^i F_Y$. Define

$$Z_i = Y_i - \bar{Y}, \quad \text{for } i = 1, \ldots n.$$

We have

$$\begin{aligned} m_r^{(Z)} &= \frac{1}{n}\sum_{i=1}^{n} (Z_i - \bar{Z})^r \\ &= \frac{1}{n}\sum_{i=1}^{n} (Y_i - \bar{Y})^r \\ &= m_r^{(Y)}. \end{aligned}$$

Moreover,

$$\begin{aligned}\bar{Z} &= \frac{1}{n}\sum_{i=1}^{n} Z_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}(Y_i) - n\bar{Y}\right) \\ &= 0.\end{aligned}$$

But

$$\begin{aligned}\mathbb{E}(Z_i) &= \mathbb{E}(Y_i - \bar{Y}) \\ &= \mathbb{E}(Y_i) - \mathbb{E}(\bar{Y}) \\ &= \mu_Y - \mu_Y \\ &= 0.\end{aligned}$$

7.2.3 Sample Variance

Observe that

$$\begin{aligned}Y_i - \bar{Y} &= Y_i - \frac{1}{n}\sum_{j=1}^{n} Y_j \\ &= Y_i - \frac{1}{n}Y_i - \frac{1}{n}\sum_{j=1, j\neq i}^{n} Y_j \\ &= \underbrace{\frac{n-1}{n}Y_i}_{V_i} - \underbrace{\frac{1}{n}\sum_{j=1, j\neq i}^{n} Y_j}_{W_i}\end{aligned}$$

Notice that $V_i \perp\!\!\!\perp W_i$, $Y_i \perp\!\!\!\perp W_i$, and $V_1, \ldots, V_n$ are independent.

Alternative Expressions

We have

$$\begin{aligned}s^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}_n^2\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{n} Y_i(Y_i - \bar{Y}).\end{aligned}$$

So,

$$\begin{aligned}
\mathbb{E}(s^2) &= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n} Y_i(Y_i - \bar{Y})\right) \\
&= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n} Y_i(V_i - W_i)\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbb{E}(Y_iV_i) - \underbrace{\mathbb{E}(Y_iW_i)}_{=0}\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left(\frac{n-1}{n}Y_i^2\right) \\
&= \frac{1}{n-1}n\frac{n-1}{n}\mathbb{E}(Y_1^2) \\
&= \mathrm{Var}(Y_1) \\
&= \sigma^2.
\end{aligned}$$

Now, what is $\mathrm{Cov}(\bar{Y}, s^2)$?

$$\begin{aligned}
\mathrm{Cov}(\bar{Y}, s^2) &= \mathbb{E}(\bar{Y}s^2) - \underbrace{\mathbb{E}(\bar{Y})}_{=0}\mathbb{E}(s^2) \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)\left(\frac{1}{n-1}\sum_{j=1}^{n} Y_j(V_j - W_j)\right)\right] \\
&= \frac{1}{n(n-1)}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(Y_iY_jV_j - Y_iY_jW_j)\right].
\end{aligned}$$

Whenever $i \neq j$ in the expression $Y_iY_jV_j$, we have $Y_iY_jV_j = 0$, since $Y_i \perp\!\!\!\perp Y_j$ and $Y_i \perp\!\!\!\perp V_j$. Moreover, we have $Y_iY_jW_j = 0$ where $i \neq j$. What about when $i = j$?

$$\begin{aligned}
\mathrm{Cov}(\bar{Y}, s^2) &= \frac{1}{n(n-1)}\mathbb{E}\left[\sum_{i=1}^{n}\left(Y_i^2V_i - \underbrace{Y_i^2W_i}_{=\mathbb{E}(Y_i^2 - W_i^2)=0}\right)\right] \\
&= \frac{1}{n(n-1)}\mathbb{E}\left[\sum_{i=1}^{n}\left(\frac{n-1}{n}Y_i^3\right)\right] \\
&= \frac{n-1}{n^2(n-1)}\sum_{i=1}^{n}\mathbb{E}(Y_i^3) \\
&= \frac{1}{n^2}n\mu_3' \\
&= \frac{1}{n^2}n\mu_3 \\
&= \frac{\mu_3}{n}.
\end{aligned}$$

The penultimate equality results from Y having 0 mean. Note that $\mu_3 = 0$ for any symmetrical distribution.

7.2.4 Joint Sample Moments

Definition 7.2.4. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample. The **joint sample moment** and **joint central sample moment** are

$$m'_{r,s} = \frac{1}{n}\sum_{i=1}^{n} X_i^r Y_i^s$$
$$m_{r,s} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^r (Y_i - \bar{Y})^s,$$

respectively.

Example 7.2.5. Note that

$$\begin{aligned} m_{1,1} &= \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \frac{n-1}{n} C_{X,Y}, \end{aligned}$$

where $C_{X,Y}$ is the sample covariance. ◇

Definition 7.2.6. With a random sample defined as in Definition 7.2.4, the **sample correlation** is

$$r_{X,Y} = \frac{C_{X,Y}}{\sqrt{s_X^2 s_Y^2}}.$$

Remark 7.2.7. Note that $|r_{X,Y}| \leq 1$. We have $|r_{X,Y}| = 1$ only when

$$Y_i = \alpha + \beta X_i, \quad i = 1, \ldots n$$

for some $\alpha, \beta \in \mathbb{R}$, $(\beta \neq 0)$. Further note that $r_{X,Y} = 1$ when $\beta > 0$, and $r_{X,Y} = -1$ when $\beta < 0$.

7.3 Week 13: Lecture 1

7.3.1 A Normal Sample

Tue 25 Jan 14:00

Proposition 7.3.1. Let $Y_1, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2)$. Then

(i) $\bar{Y} \perp\!\!\!\perp (Y_j - \bar{Y})$ for all $j = 1, \ldots, n$

(ii) $\bar{Y} \perp\!\!\!\perp S^2$.

Proof.

(i) Since we can express any linear combination of $\bar{Y}$ and $(Y_j - \bar{Y})$ as a linear combination of $Y_1, \ldots, Y_n$, $\bar{Y}$ and $(Y_j - \bar{Y})$ are jointly normally distributed. Moreover, they are uncorrelated:

$$\begin{aligned}
\operatorname{Cov}(\bar{Y}, Y_j - \bar{Y}) &= \operatorname{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i,\ Y_j - \frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\
&= \operatorname{Cov}(\bar{Y}, Y_j) - \operatorname{Cov}(\bar{Y}, \bar{Y}) \\
&= \operatorname{Cov}\left(\frac{1}{n}Y_j,\ Y_j\right) - \operatorname{Var}(\bar{Y}) \\
&= \frac{1}{n}\operatorname{Var}(Y_j) - \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} \\
&= 0.
\end{aligned}$$

Hence, $\bar{Y}$ and $\text{Normal}(\mu, \sigma^2)$ are jointly normal and uncorrelated. Thus, they are independent.

(ii) Note that $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$ is independent of $\bar{Y}$.

□

7.3.2 The χ^2 Distribution

Recall. Let $X_1, \ldots, X_n \sim F_x$. Then

$$\begin{aligned}
M_{\bar{X}}(t) &= \mathbb{E}(e^{t\bar{X}} \\
&= \mathbb{E}\left(e^{\frac{t}{\lceil} n}\sum_{i=1}^{n} X_i\right) \\
&= \mathbb{E}(e^{\frac{t}{n}}X_1) \cdots \mathbb{E}(e^{\frac{t}{n}}X_n \\
&= \left(M_X\left(\frac{t}{n}\right)\right)^n
\end{aligned}$$

Proposition 7.3.2. Let $Z_1, \ldots, Z_n \sim \text{Normal}(0, 1)$. Then

(i) $\bar{Z} \sim \text{Normal}\left(0, \frac{1}{n}\right)$

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

Remark 7.3.3. Note that χ^2_k, where k is the degrees of freedom, is the same

as $\Gamma\left(\frac{k}{2}, \frac{1}{2}\right)$. So, if $U \sim \chi^2_k$, then

$$M_n(t) = (1-2t)^{-\frac{k}{2}},$$

which is the MGF of the distribution of $\sum_{i=1}^k Z_i^2$.

Proof.

(i) Note that

$$\begin{aligned} M_{\bar{Z}} &= \left(M_{Z_1}\left(\frac{t}{n}\right)\right)^n \\ &= \left(e^{\frac{\left(\frac{t}{n}\right)^2}{2}}\right) \\ &= e^{\frac{t^2}{2n}}, \end{aligned}$$

which is the MGF of Normal $\left(0, \frac{1}{n}\right)$.

(ii) Observe that

$$\underbrace{\sum_{i=1}^n Z_i^2}_{\sim\chi^2_n} = \underbrace{\sum_{i=1}^n (Z_i - \bar{Z})^2}_{(n-1)s^2} + \underbrace{n\bar{Z}^2}_{\left(\frac{\bar{Z}-0}{\sqrt{1/n}}\right)^2 \sim \chi^2_1}.$$

Observe that the MGF of the left hand side is $(1-2t)^{-\frac{n}{2}}$, and the MGF of the right hand side is

$$M_{(n-1)s^2}(t) \times (1-2t)^{-\frac{1}{2}}.$$

This is because the two terms are independent random variables. This implies that

$$M_{(n-1)s^2}(t) = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-\frac{n-1}{2}},$$

so

$$(n-1)s^2 \sim \chi^2_{n-1}.$$

□

We can extend these properties to the general normal.

Proposition 7.3.4. Let $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$. Then

(i) $\bar{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$

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

Proof. We can set $X_i = \mu + \sigma Z_i$, where $Z_i \sim \text{Normal}(0, 1)$. The rest of the proof follows. □

7.3.3 Sample quantiles and order statistics

Definition 7.3.5. The α-**quantile** q_α is the smallest value such that

$$F_Y(q_\alpha) = \alpha.$$

Example 7.3.6. The median is $q_{0.5}$, and $q_{0.25}$ is the lower quartile. ◇

Notation. If $Y_1, \ldots, Y_n$ is a random sample. Then

$$Y_{(1)} = \min\{Y_1, \ldots, Y_n\} \vdots Y_{(n)} = \max\{Y_1, \ldots, Y_n\}.$$

Definition 7.3.7. We say that $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ are the **order statistics** of the sample.

Remark 7.3.8. Notice that $\mathbb{E}(Y_{(n)}) > \mathbb{E}(Y)$. Order statistics do *not* have the same distribution as the population they came from:

$$\begin{aligned} F(Y_{(n)}) &= P(Y_{(n)} \leq y) \\ &= P(Y_1 \leq y, \ldots, Y_n \leq y) \\ &= P(Y_1 \leq y) \cdots P(Y_n \leq y) \\ &= (F_Y(y))^n. \end{aligned}$$

Similarly, $F_{Y_{(1)}}(y) = 1 - (1 - F_Y(y))^n$. How do we find the PDF/PMF?

Continuous Case

$$\begin{aligned} f_{Y_{(n)}}(y) &= \frac{d}{dy} F_{Y_{(n)}}(y) \\ &= \frac{d}{dy} (F_Y(y))^n \\ &= n(F_Y(y))^{n-1} f_Y(y). \end{aligned}$$

Moreover,

$$\begin{aligned} f_{Y_{(1)}}(y) &= \frac{d}{dy} (1 - (1 - F_Y(y))^n) \\ &= n(1 - F_Y(y))^{n-1} f_Y(y). \end{aligned}$$

Discrete Case

If the support is $\{a_1, a_2, \dots\}$, then

$$f_{Y_{(n)}}(y) = \begin{cases} (F_Y(a_k))^n - (F_Y(a_{k-1}))^n, & \text{if } y = \alpha_k, \\ 0, & \text{otherwise.} \end{cases}$$

7.3.4 Sample quantiles

Notation. For $a \in \mathbb{R}$, $\{a\}$ is equal to a rounded to the nearest integer.

Definition 7.3.9. The **sample α-quantile** is defined as

$$Q_\alpha = \begin{cases} Y_{(\{n\alpha\})}, & \frac{1}{2n} < \alpha < \frac{1}{2} \\ Y_{(n+1-\{n(1-\alpha)\})}, & \frac{1}{2} < \alpha < 1 - \frac{1}{2n}. \end{cases}$$

Try applying this definition, for, say, $n = 5$ and various α.

Example 7.3.10. The sample median is

$$Q_{0.5} = \begin{cases} Y_{\left(\frac{n+1}{2}\right)}, & \text{if } n \text{ is odd} \\ \dfrac{Y_{\left(\frac{n}{2}\right)} + Y_{\left(\frac{n}{2}+1\right)}}{2}, & \text{if } n \text{ is even.} \end{cases}$$

◇

7.4 Week 13: Lecture 2

7.4.1 More on Order Statistics

Thu 27 Jan 10:00

Example 7.4.1. Let $Y_1, \dots, Y_n \sim \text{Unif}[0, \theta]$. Then $f_Y(y) = \frac{1}{\theta}$, for $0 \le y \le \theta$. Moreover,

$$F_Y(y) = \begin{cases} 0, & y < 0, \\ \frac{y}{\theta}, & 0 \le y \le \theta \\ 1, & y > \theta. \end{cases}$$

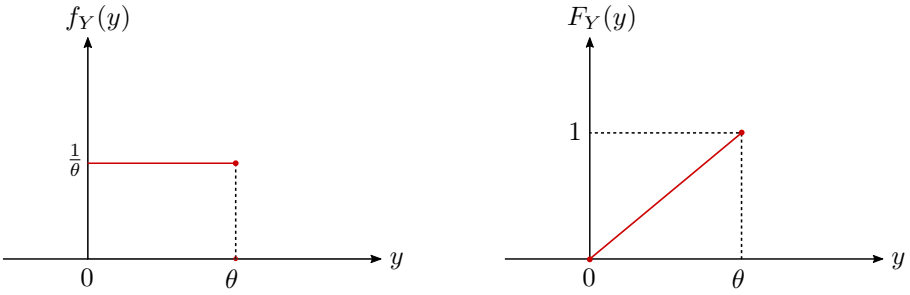


Figure 7.1: The PDF and CDF of the Unif$[0, \theta]$ distribution, respectively.

So

$$\begin{aligned} f_{Y_{(n)}}(y) &= n(F_Y(y))^{n-1} f_Y(y) \\ &= n\left(\frac{y}{\theta}\right)^{n-1} \frac{1}{\theta} \\ &= \frac{n y^{n-1}}{\theta^n} \end{aligned}$$

Note. Alternatively, we could differentiate $F_{Y_{(n)}}(y) = \left(\frac{y}{\theta}\right)^n$.

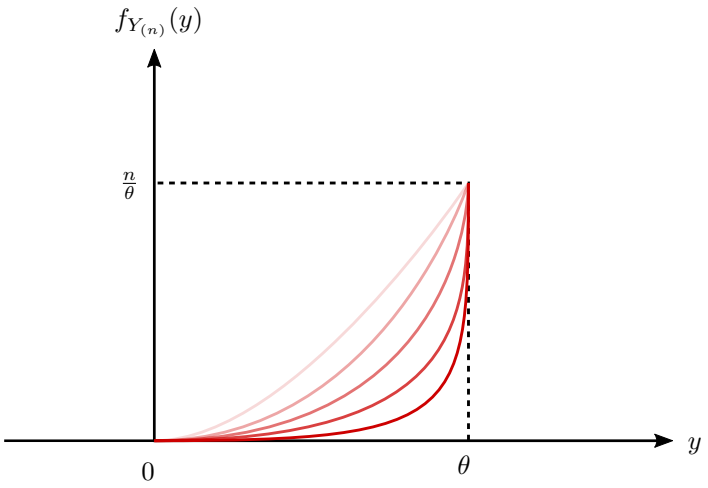


Figure 7.2: The PDF of $Y_{(n)}$ in Example 7.4.1. As n increases, the degree of the polynomial also increases. This, in turn, increases the probability that $Y_{(n)}$ takes a value closer to θ.

But what about $Y_{(i)}$?

The CDF of $Y_{(i)}$

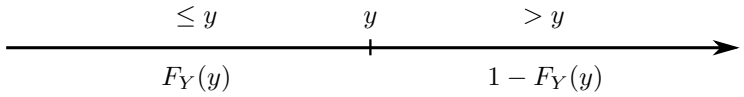


Figure 7.3: For a given random variable Y, the probability that the observed value is less than y is $F_Y(y)$, and the probability that it is greater than y is $1 - F_Y(y)$.

Using Figure 7.3, we have

$$\begin{aligned} F_{Y_{(i)}}(y) &= P(Y_{(i)} \leq y) \\ &= P(\text{"At least } i \text{ observations are } \leq y\text{"}) \\ &= \sum_{j=i}^{n} P(\text{"Exactly } j \text{ observations are } \leq y\text{"}) \\ &= \sum_{j=i}^{n} \binom{n}{j} (F_Y(y))^j (1 - F_Y(y))^{n-j} \end{aligned}$$

(try it for $i = 1$ and $i = n$).

◇

PDF of $Y_{(i)}$ (continuous case)

Note that

$$\begin{aligned} f_Y(y) &= \lim_{h \downarrow 0} \frac{F_Y(y+h) - F_Y(y)}{h} \\ &= \lim_{h \downarrow 0} \frac{P(y < Y \leq y + h)}{h} \\ &\Rightarrow P(y < Y \leq y + h) \approx h f_Y(y) \quad \text{(if } h \text{ is small)}. \end{aligned}$$

What does this tell us? Refer to Figure 7.4:

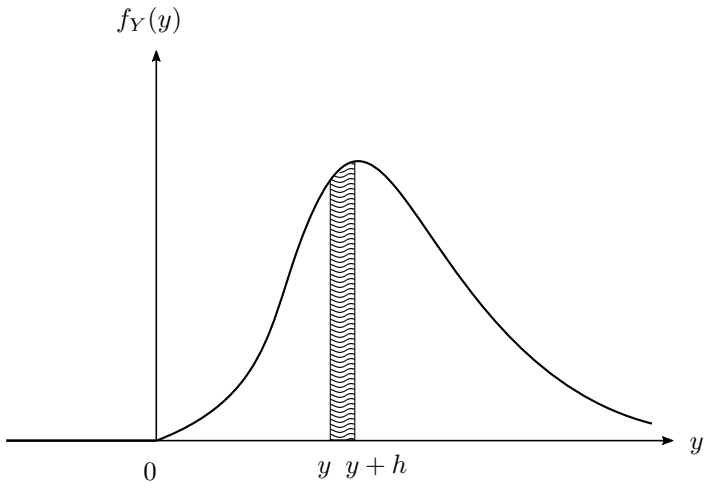


Figure 7.4: As h approaches 0, the shape under the distribution approaches a rectangle.

More rigorous explanation? Note that

$$P(y < Y \leq y + h) = hf_Y(y) + o(h).$$

As h approaches 0, $o(h)$ converges to 0. Now, what is the probability that y takes a value between y and $y + h$?

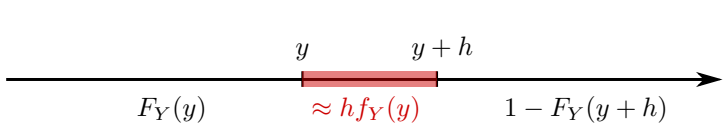


Figure 7.5: The probability that a value falls between y and $y + h$ is approximately equal to $hf_Y(y)$ for small values of h.

Note that if i falls between y and $y+h$, we must have exactly $i-1$ observations to the left of it. Now, note that an observation can fall in each of the three intervals. We know $i - 1$ observations are in the interval $(-\infty, y]$ and that for very small h, only one observation falls within $[y, y + h]$.

Finally, $n - i$ observations fall within $(y + h, \infty)$. Using the multinomial coefficient, we can write down the density function of the ith order statistic directly:

$$\begin{aligned} f_{Y_{(i)}}(y) &= \lim_{h \downarrow 0} \frac{P(y < Y_{(i)} \leq y + h)}{h} \\ &= \lim_{h \downarrow 0} \frac{\frac{n!}{(i-1)!1!(n-i)!}(F_Y(y))^{i-1} h f_Y(y)(1 - F_Y(y+h))^{n-i}}{h} \\ &= \frac{n!}{(i-1)!(n-i)!}(F_Y(y))^{i-1} f_Y(y)(1 - F_Y(y))^{n-i}. \end{aligned}$$

7.4.2 The Beta Distribution

Example 7.4.2 (Beta Distribution)**.** Let $Y_1, \ldots, Y_n \sim \text{Unif}[0,1]$, $f_Y(y) = 1$, $0 \leq y \leq 1$ and

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ y, & 0 \leq y \leq 1 \\ 1, & y > 1 \end{cases}$$

Note that

$$f_{Y_{(i)}}(y) = \frac{n!}{(i-1)!(n-i)!} y^{i-1}(1-y)^{n-i}, \quad 0 \leq y \leq 1.$$

If $X \sim \text{Beta}(\alpha, \beta)$,

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1.$$

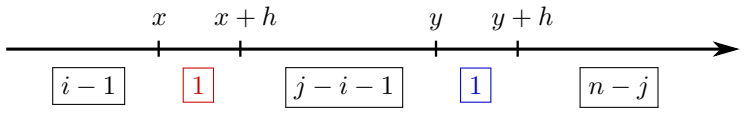
◇

Note that

$$\text{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Joint PDF of Two Order Statistics

The approach we used to find density functions directly can be used to find joint densities, i.e., the joint PDF of $Y_{(i)}$ and $Y_{(j)}$, $(i \neq j)$.



We have

$$\begin{aligned} f_{Y_{(i)},Y_{(j)}}(x,y) = &\frac{n!}{(i-1)!(j-i-1)!(n-j)!} \times (F_Y(x))^{i-1} f_Y(x) \\ &\times (F_Y(y) - F_Y(x))^{j-i-1} \times f_Y(y)(1 - F_Y(y))^{n-j}. \end{aligned}$$

Chapter 8

Estimation, Testing, and Prediction

8.1 Week 14: Lecture 1

In this lecture, we introduce three topics that we will focus on for the rest of the term: point estimation, interval estimation, and hypothesis testing. Tue 1 Feb 14:00

8.1.1 A Few Questions

What's one of the first things you do when you first get a dataset? Well, you find summary statistics, namely, the means. The sample mean is an example of a statistic.

Definition 8.1.1. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be a random sample. A **statistic** is a function of the *sample* and *known constants* alone.

Let $\mathbf{u} = \mathbf{h}(\mathbf{Y})$ be a statistic. Is it a random variable? Yes, it is a function of the sample mean, which itself is a random variable. In practice, what we observe is not a random variable, because we plug in the observed values for $\mathbf{Y}$.

Definition 8.1.2. We denote an **observed statistic** as $\mathbf{u} = \mathbf{h}(\mathbf{y})$.

Example 8.1.3 (Sample Mean)**.** Note that

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is a statistic, but

$$\frac{Y_1 + Y_2^3 + e^{Y_3}}{Y_4}$$

is also. We are interested in statistics that give us a meaningful value, generally some dimension reduction. ◇

Definition 8.1.4. The distribution of a statistic **U** is known as a **sampling distribution.**

In practice, what we observe is not the statistic itself, but the observed statistic. We do not see the distribution unless we repeat data collection. The sampling distribution of a statistic is going to depend on three things: the function of the statistic, the population distribution, and the sample size.

8.1.2 Pivotals

Example 8.1.5. Let $Y_1, \ldots, Y_n \sim \text{Normal}(\mu, 1)$. Then

$$\frac{\bar{Y} - \mu}{\sqrt{1/n}} = \sqrt{n}(\bar{Y} - \mu) \sim \text{Normal}(0, 1).$$

Note that $\sqrt{n}(\bar{Y} - \mu)$ is not a statistic, because the expression includes μ, a parameter of the underlying distribution. ◇

Definition 8.1.6. The quantity $g(\bar{Y}, \theta)$, where $\bar{Y}$ is a random sample and θ is a parameter, is a **pivotal** if its distribution does not depend on θ (or any unknown parameters).

Example 8.1.7. Let $Y_1, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2)$. We know

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$

and thus it is a pivotal. ◇

Remark 8.1.8. While pivotals are functions of the sample, they are not always statistics, as seen in Example 8.1.5. In fact, they more often aren't.

The t-distribution

Definition 8.1.9. Let $Z \sim \text{Normal}(0, 1)$, $V \sim \chi^2_K$, $Z \perp\!\!\!\perp V$. Then

$$\frac{z}{\sqrt{v/k}} \sim t_k,$$

which is the ***t*-distribution** with k degrees of freedom.

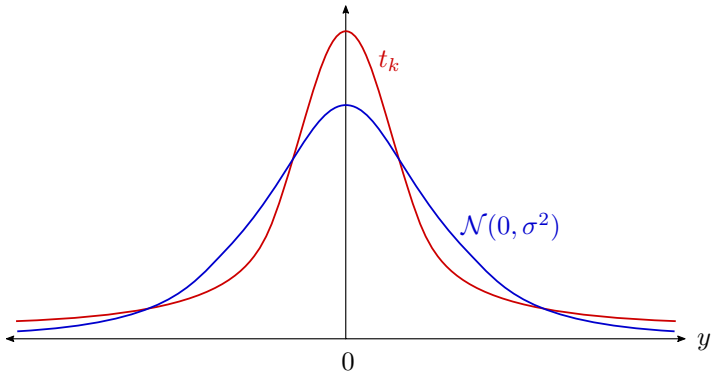


Figure 8.1: A t-distribution compared with a Normal distribution. The t-distribution has clearly fatter tails.

The t-distribution is another bell curve distribution, but its different because it has heavier tails for lower values of k, and thus slimmer peaks. In general, if you are sampling from the standard normal, you're probably going to be moderately far from the mean. But if you sample from the t-distribution, you're probably going to be closer to the mean than the standard normal. Occasionally, however, you will be at the extremes.

Why is the t-distribution helpful?

Example 8.1.10. Let $Y_1, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2)$. Then

$$\frac{Y - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1), \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}, \quad Y \perp\!\!\!\perp s^2.$$

This implies that

$$\frac{\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{n-1}}} = \boxed{\frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1}}$$

a pivotal quantity $g(\mathbf{Y}, \mu)$. ◇

Obtaining Pivotals

There are no standard steps in obtaining a pivotal. Using some properties from previous distributions, however, can be helpful.

Example 8.1.11. Let $Y_1, \ldots, Y_n \sim \text{Exp}(\lambda)$. Note that

$$\sum_{i=1}^{n} Y_i \sim \Gamma(n, \lambda).$$

Moreover,

$$kY_1 \sim \text{Exp}(\lambda/k),$$

so

$$M_{Y_1}(t) = \left(1 - \frac{t}{\lambda}\right)^{-1},$$

and

$$\begin{aligned} M_{kY_1}(t) &= \mathbb{E}(e^{tkY_1}) \\ &= M_{Y_1}(tk) \\ &= \left(1 - \frac{tk}{\lambda}\right)^{-1}. \end{aligned}$$

◇

8.1.3 Point Estimation

Definition 8.1.12. Any scalar statistic U can be considered as a **point estimator** for a parameter θ. Moreover,

$$U = h(\mathbf{Y}) \text{ is a point estimator}$$
$$u = h(\mathbf{y}) \text{ is a } \textbf{point estimate}.$$

An estimator is a random variable. In practice, we obtain *estimates.*

Bias

Definition 8.1.13. We define **bias** as

$$\underset{\theta}{\text{Bias}}(U) = \mathbb{E}_\theta(U) - \theta$$

Note that

$$\mathbb{E}_\theta(U) = \begin{cases} \sum u f_U(u;\theta) & \text{(discrete)} \\ \displaystyle\int_{\mathbb{R}} u f_U(u;\theta)\, du & \text{(continuous)} \end{cases}$$

Moreover, U is *unbiased* for θ if $\text{Bias}_\theta(U) = 0$.

All else being equal, we want no bias. But in practice, we also want low variance, because an estimator with high variance isn't very useful.

Definition 8.1.14. The **mean squared error** of an estimator is

$$\text{MSE}[\theta](U) = \mathbb{E}_\theta[(U-\theta)^2] = \left(\underset{\theta}{\text{Bias}}(U)\right)^2 + \text{Var}(\theta).$$

Exercise. Prove the second equality of this definition.

8.2 Week 14: Lecture 2

8.2.1 Estimator Convergence

Thu 3 Feb 10:00

Example 8.2.1. Let $Y_1, \dots, Y_n \sim F_Y(\cdot; \boldsymbol{\theta})$ be a random sample. We now expand our definition of an estimator. Suppose that the underlying distribution is Normal(μ, σ^2) distributed:

$$\boldsymbol{\theta} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \quad \text{example estimator: } \begin{pmatrix} \bar{Y} \\ s^2 \end{pmatrix}$$

◇

We denote the estimator $U(\mathbf{Y}) = \hat{\theta}$. Unfortunately we also denote the *estimate* $\hat{\theta} = U(\mathbf{y})$.

Recall. The definition of MSE is

$$\begin{aligned} \underset{\theta}{\mathrm{MSE}}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\ &= \left(\underset{\theta}{\mathrm{Bias}}(\hat{\theta})\right)^2 + \mathrm{Var}(\hat{\theta}). \end{aligned}$$

Example 8.2.2. Let

$$Y_1, \dots, Y_n \sim F_Y, \quad \mathbb{E}(Y_1) = \mu, \ \mathrm{Var}(Y_1) = \sigma^2.$$

Take $\hat{\mu} = \bar{Y}$. Then

$$\underset{\mu}{\mathrm{MSE}}(\hat{\mu}) = \left(\underset{\mu}{\mathrm{Bias}}(\hat{\mu})\right)^2 + \mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{n}.$$

Note that $\lim_{n\to\infty} \frac{\sigma^2}{n} = 0$. This is the single most important property of an estimator. ◇

Note that we've only talked about convergence in terms of sequences. We can consider $\hat{\theta} = \hat{\theta}(\mathbf{Y})$ as a *sequence* of estimators:

$$\hat{\theta}_1 = \hat{\theta}(Y_1), \ \hat{\theta}_2 = \hat{\theta}(Y_1, Y_2), \dots, \ \hat{\theta}_n = \hat{\theta}(Y_1, \dots, Y_n)$$

Remark 8.2.3. Convergence in distribution to a *constant* is equivalent to convergence in probability.

Definition 8.2.4. An estimator $\hat{\theta}$ is a **consistent** estimator of θ if $\hat{\theta} \to^P \theta$ as $n \to \infty$. Alternatively,

$$P(|\hat{\theta} - \theta| < \varepsilon) \to 1 \text{ as } n \to \infty.$$

Remark 8.2.5. Basically, consistency implies that however small you want ε, you can achieve this by increasing the sample size.

But is it enough for an estimator to be consistent? No. This fact alone doesn't tell us the rate of convergence. But it is an essential first property.

Example 8.2.6. If $\hat{\theta} \to^{\text{m.s}} \theta$, then $\hat{\theta}$ is a mean-square consistent estimator. $\diamond$

Note that mean square consistency is easy to show; bias and variance should converge to 0 for a mean square consistent estimator.

Recall.

$$\bar{Y} \to^{P} \mu \quad \text{is the weak law of large numbers}$$

We can rephrase this as "$\bar{Y}$ is a consistent estimator of μ." Moreover,

$$\boxed{\bar{Y} \to^{\text{a.s.}} \mu \quad \text{is the strong law of large numbers.}}$$

The strong law of large numbers implies the weak law, but does not imply mean-square consistency.

Remark 8.2.7. In principle, we always want to pick the estimator that minimizes the MSE. In practice, the minimum MSE estimator typically has no closed form. What's the next best thing? Choose an unbiased estimator, so the MSE only depends on $\text{Var}(\hat{\theta})$. Among all estimators, the best is the one that minimizes $\text{Var}(\hat{\theta})$.

Ultimately, we need a guarantee that our estimator gets better with more data.

8.2.2 The Method of Moments (Moment Matching)

So how do we find estimators for unknown quantities? Suppose we have a random sample $Y_1, \ldots, Y_n \sim F_Y(\,\cdot\,; \boldsymbol{\theta})$, where

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}$$

Which "moments" are we talking about? Well, we can either be looking at the moments of the probability distribution, or the sample moments:

$$\begin{aligned} m_1' &= \bar{Y}, & \mu_1' &= \mathbb{E}(Y) \\ &\vdots & &\vdots \\ m_r' &= \frac{1}{n}\sum_{i=1}^{n} Y_i^r, & &\underbrace{\mu_r' = \mathbb{E}(Y^r)}_{\text{functions of } \boldsymbol{\theta}}. \end{aligned}$$

Sample moments are statistics, as they are functions of random variables. Population moments are constants. For $\Gamma(\alpha, \lambda)$, we have

$$\mu_1' = \frac{\alpha}{\lambda}, \quad \mu_2' = \frac{\alpha^2}{\lambda^2} + \frac{\alpha}{\lambda^2}.$$

These are functions of the parameters. So, how do we use the sample to estimate the unknown parameters? We take the first sample moment and set it equal to the first population moment, and the second and so on.

Example 8.2.8. Take $Y_1, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2)$. Note that

$$\begin{aligned} m_1' &= \bar{Y}, & \mu_1' &= \mu \\ m_2' &= \frac{1}{n}\sum_{i=1}^{n} Y_i^2, & \mu_2' &= \mathbb{E}(Y^2) = \mu^2 + \sigma^2. \end{aligned}$$

Now, set $\hat{\mu} = \bar{Y}$ and $\hat{\mu}^2 + \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$. We have

$$\hat{\mu} = \bar{Y}, \qquad \hat{\sigma}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - n\bar{Y}^2\right) = m_2 = \frac{n-1}{n}s^2.$$

◇

In practice, we are relying on a method to maximize a very complicated equation. Methods are sensitive to starting values. The faster they converge, the more sensitive they are.

Example 8.2.9. Let $Y_1, \ldots, Y_n \sim \text{Bin}(r, p)$. Then

$$\begin{aligned} \mu_1' = rp, \qquad \mu_2' &= rp(1-p) + (rp^2) \\ &= \mu_2 + (\mu_1')^2. \end{aligned}$$

We can set $\hat{\mu}_2 = m_2$. We have

$$\bar{Y} = \hat{r}\hat{p}.\frac{1}{n}\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right) = \hat{r}\hat{p}(1-\hat{p}).$$

So $\hat{p} = \frac{\bar{Y}}{\hat{r}}$, which implies

$$\begin{aligned} &\hat{r}\frac{\bar{Y}}{\hat{r}}\left(1 - \frac{\bar{Y}}{\hat{r}}\right) = \bar{Y}\left(1 - \frac{\bar{Y}}{\hat{r}}\right) = m_2 \\ &\Rightarrow 1 - \frac{\bar{Y}}{\hat{r}} = \frac{m_2}{\bar{Y}}. \\ &\Rightarrow \frac{\bar{Y}}{\hat{r}} = 1 - \frac{m_2}{\bar{Y}} \\ &\Rightarrow \hat{r} = \frac{\bar{Y}}{1 - \frac{m_2}{\bar{Y}}} = \frac{\bar{Y}^2}{\bar{Y} - m_2}. \end{aligned}$$

◇

8.2.3 Interval Estimation

An interval estimator of θ is a *random interval* of the form (U_1, U_2), where U_1, U_2 are *statistics* with the property $U_1 \leq U_2$. With an interval estimator, we want a range of values that contains θ.

Definition 8.2.10. The **coverage probability** is

$$P(U_1 \leq \theta \leq U_2).$$

Definition 8.2.11. The **confidence coefficient** is

$$\inf_{\theta} P(U_1 \leq \theta \leq U_2)$$

Is it enough to have a high confidence coefficient? No, we also care about the expected length of the interval.

Definition 8.2.12. The **expected length** is

$$\mathbb{E}(U_2 - U_1).$$

8.3 Week 15: Lecture 1

8.3.1 More Interval Estimation

Tue 8 Feb 14:00

Definition 8.3.1. Let (U_1, U_2) be an interval estimator, with $U_1 \leq U_2$. Then (u_1, u_2) is an interval estimate, where $u_1 = u_1(\mathbf{y})$.

Example 8.3.2. Let $Y_1, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. A pivotal quantity for μ is

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1).$$

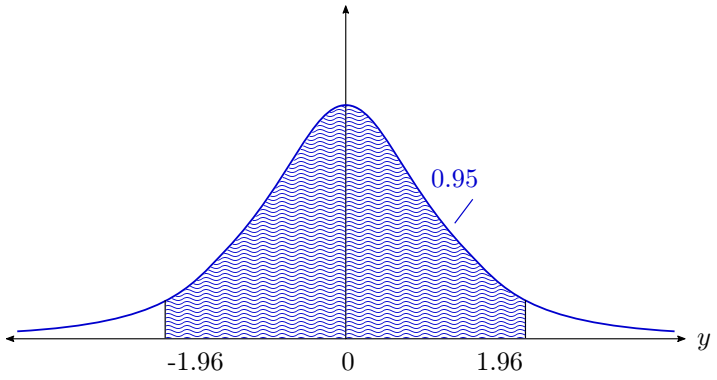


Figure 8.2: A 95% confidence interval of the standard normal.

Recall that a pivotal quantity relies only on known parameters. How do we turn a pivotal into a confidence interval? Since our pivotal is standard normal, we can utilize the z-score at the limits of a 95% confidence interval:

$$0.95 = P\left(-1.96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < 1.96\right).$$

After rearranging, the solution is

$$P\left(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Then

$$\left(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is an interval estimator for μ, or, a 95% confidence interval. ◇

In general, we take $\alpha_1 + \alpha_2 = \alpha$, with

$$1 - \alpha = P\left(Z_{\alpha_1} < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < Z_{1-\alpha_2}\right).$$

The confidence interval is thus

$$\left(\bar{Y} - Z_{1-\alpha 1}\frac{\sigma}{\sqrt{n}}, \bar{Y} + Z_{1-\alpha 2}\frac{\sigma}{\sqrt{n}}\right).$$

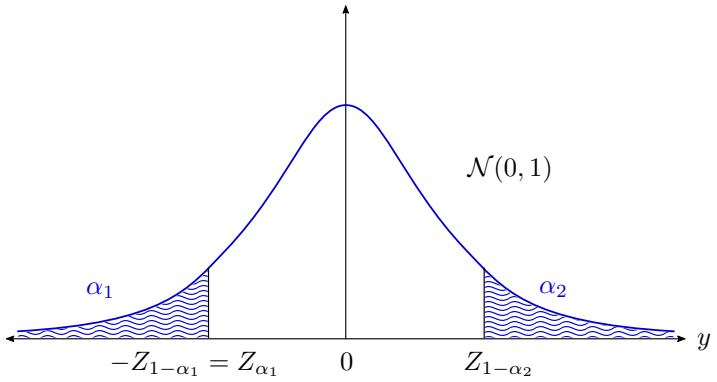


Figure 8.3: A more general case for a standard normal pivotal.

8.3.2 Some Pivotal Assumptions

Remark 8.3.3. For the purposes of this course, we assume that our pivotals have the following properties:

- The distribution of our pivotal, W, is unimodal.
- Moreover, W is continuous and linear in θ, with $W = g(\mathbf{Y}, \theta) = a(\mathbf{Y}) + b(\mathbf{Y})\theta$.

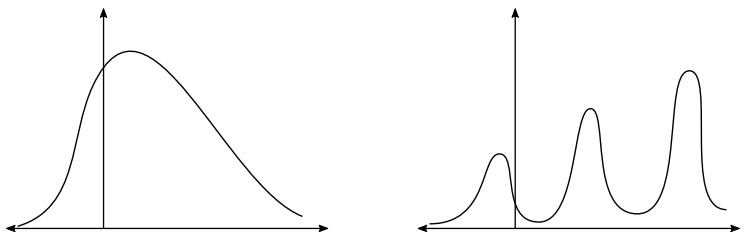


Figure 8.4: A unimodal (left) distribution vs. a multimodal distribution.

Thus,

$$\begin{aligned} 1 - \alpha &= P(w_1 < W < w_2) \\ &= P(w_1 < a(\mathbf{Y}) + b(\mathbf{Y})\theta < w_2) \\ &= P\left(\frac{w_1 - a(\mathbf{Y})}{b(\mathbf{Y})} < \theta < \frac{w_2 + a(\mathbf{Y})}{b(\mathbf{Y})}\right) \end{aligned}$$

The length of the CI is $\frac{w_2 - w_1}{b(\mathbf{Y})}$. The *optimal* CI is where w_1 and w_2 have the same density. Why? No formal proof here; a geometric argument is presented in Figure 8.5:

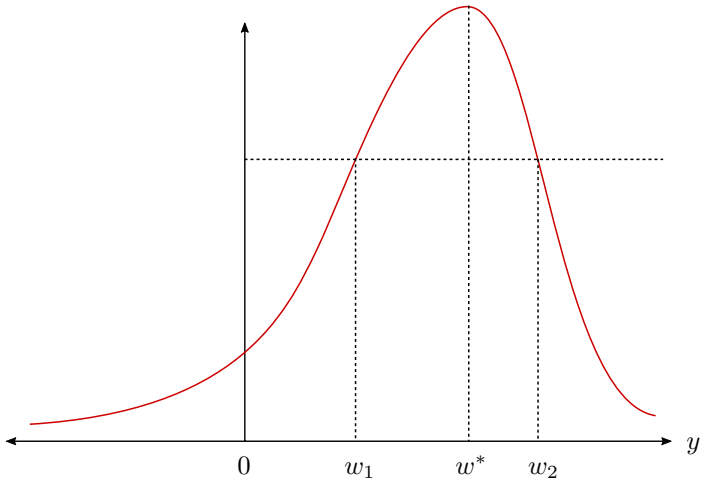


Figure 8.5: Suppose that the area encompassed by w_1 and w_2 corresponds to the desired value of α, and w^* is the mode. If we shift w_1 to the left, then it follows that we will have to shift w_2 to the left as well. However, we shift w_2 *less* than w_1, because the area under the distribution curve is greater to the left of w_2 than w_1. If we shift w_1 to the right, it follows that we shift w_2 to the right by a greater amount to maintain the same value of α. Both of these actions therefore *increase* the length of the confidence interval. A similar argument holds for w_2. Hence, the optimal interval is w_1, w_2.

Example 8.3.4. Let $Y_1, \ldots, Y_n \sim \text{Exp}(\lambda)$. Then

$$\begin{aligned} &\lambda Y_1, \lambda Y_2, \ldots, \lambda Y_n \sim \text{Exp}(1) \\ &\Rightarrow \sum_{i=1}^{n} \lambda Y_i \sim \Gamma(n, 1) \\ &2\sum_{i=1}^{n} \lambda Y_i = 2n\lambda \bar{Y} \sim \Gamma\left(n, \frac{1}{2}\right) \end{aligned}$$

Choose w_1, w_2 so that

$$f_W(w_1) = f_W(w_2).$$

In practice, w_1, w_2 are difficult to find, and have no closed form solution. The equal-tail CI is a good approximation:

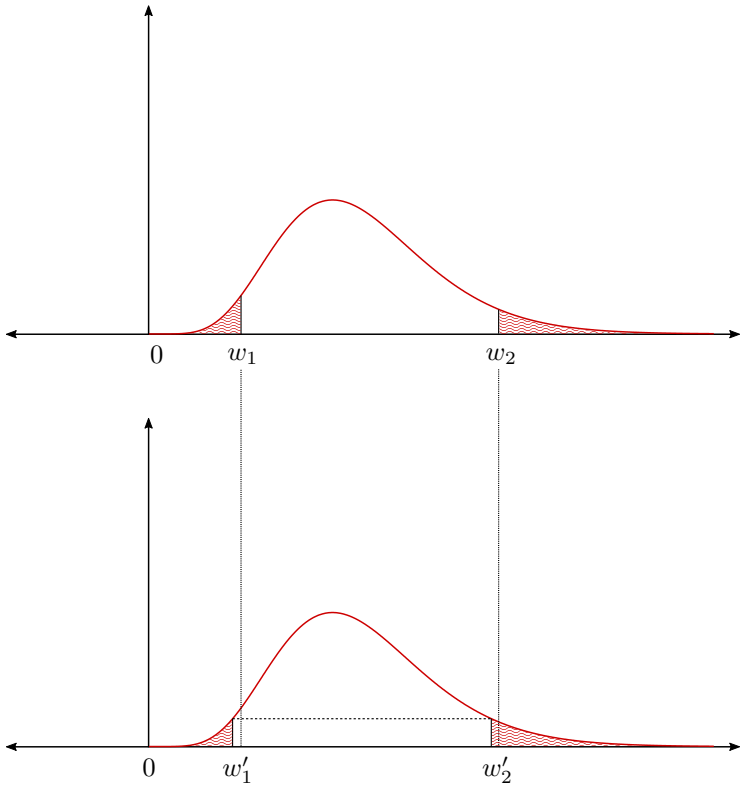


Figure 8.6: Equal-tail confidence interval (top) vs. optimal equal-density confidence interval for a Gamma distribution. *Note: Gamma curve taken from Wikimedia Commons.*

◇

Example 8.3.5. Let $Y_1, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2)$. We have

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Then

$$1 - \alpha = P\left(-t_{n-1,\ 1-\frac{\alpha}{2}} \leq \frac{\bar{Y} - \mu}{s/\sqrt{n}} \leq t_{n-1,\ 1-\frac{\alpha}{2}}\right)$$

So $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\bar{Y} - t_{n-1,\ 1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{Y} + t_{n-1,\ 1-\frac{\alpha}{2}}\right).$$

◇

8.4 Week 15: Lecture 2

8.4.1 Hypothesis Testing

Thu 10 Feb 10:00

Definition 8.4.1. We denote the **null hypothesis** as $H_0 : \theta \in \Theta_0$, where $\Theta_0 \subset \Theta$, the **parameter space.**

We can have either a *simple* or a *composite* null hypothesis. Some examples:

$$\begin{aligned}
&\Theta_0 = \{\theta_0\}, && (H_0 : \theta = \theta_0) && \text{(simple)}\\
&\Theta_0 = [a, b], && (H_0 : \theta = \theta_0 a \leq \theta \leq b) && \text{(composite)}\\
&\Theta_0 = [a, \infty), && (H_0 : \theta = \theta \geq a) && \text{(composite).}
\end{aligned}$$

Remark 8.4.2. The general form of a hypothesis test is

null:	$H_0 : \theta \in \Theta_0$	$\Theta_0 \cap \Theta_1 = \varnothing, \quad \Theta_0, \Theta_1 \subset \Theta$
alternative:	$H_1 : \theta \in \Theta_1$	(but we don't require $\Theta_0 \cup \Theta_1 = \Theta$)

Definition 8.4.3. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$. The **decision** rule is

$$\phi(\mathbf{y}) = \begin{cases} 1, & \text{when } H_0 \text{ is rejected} \\ 0, & \text{when } H_0 \text{ is not rejected.} \end{cases}$$

Definition 8.4.4. The **rejection** or **critical** region is defined as

$$C = \{\mathbf{y} : \phi(\mathbf{y}) = 1\}.$$

Hence, we can write

$$\phi(\mathbf{y}) = \begin{cases} 1, & \text{if } \mathbf{y} \in C \\ 0, & \text{if } \mathbf{y} \notin C. \end{cases}$$

Definition 8.4.5. We define **Type I error** as rejecting a true H_0, and **Type II** error as not rejecting a false H_0.

Definition 8.4.6. Let $H_0 : \theta \in \Theta_0$. A test has **significance level** α if

$$\sup_{\theta \in \Theta_0} P_\theta(\text{reject } H_0) \leq \alpha,$$

and **size** α if

$$\sup_{\theta \in \Theta_0} P_\theta(\text{reject } H_0) = \alpha.$$

If $H_0 : \theta = \theta_0$, then

$$\begin{aligned}
\text{size} &= P_\theta(\text{reject } H_0) \\
&= P(\text{type I error})
\end{aligned}$$

If something has size α, it has significance level α. It is often difficult to find a test of size α, so we set an upper bound instead.

8.4.2 Power Function

Definition 8.4.7. For $\theta \in \Theta_1$, the **power function** is

$$\beta(\theta) = P_\theta(H_0 \text{ rejected}).$$

Definition 8.4.8. The **power** of a specific $\theta_1 \in \Theta_1$ is

$$\begin{aligned}\beta(\theta_1) &= P_{\theta_1}(H_0 \text{ rejected}) \\ &= 1 - P_{\theta_1}(H_0 \text{ not rejected}) \\ &= 1 - P_{\theta_1}(\text{type II error}).\end{aligned}$$

When we talk about a study being underpowered, often we do not see the effect of the treatment.

For $\theta_0 \in \Theta_0$ we have

$$\beta(\theta_0) = P_{\theta_0}(H_0 \text{ rejected}) = P_{\theta_0}(\text{type I error})$$

So, if

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha,$$

the test has size α.

Example 8.4.9. Let $Y_1, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2)$, with σ^2 known. We have

$$\begin{aligned}H_0 &: \ \mu = \mu_0 \quad \text{simple} \\ H_1 &: \ \mu < \mu_0 \quad \text{composite},\end{aligned}$$

with critical region

$$C = \left\{ \mathbf{y} : \bar{y} < \mu_0 - Z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right\}.$$

We want

$$\begin{aligned}\beta(\mu) &= P_\mu(\text{reject } H_0) \\ &= P_\mu(\mathbf{Y} \in C) \\ &= P_\mu\left(\bar{Y} < \mu_0 - Z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right)\end{aligned}$$

If $\mu = \mu_0$,

$$\begin{aligned}\beta(\mu_0) &= P_{\mu_0}\left(\bar{Y} < \mu_0 - Z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right) \\ &= P_{\mu_0}\left(\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} < -Z_{1-\alpha} \right) \\ &= \alpha,\end{aligned}$$

the size. If $\mu < \mu_0$,

$$\begin{aligned}\beta(\mu) &= P_\mu\left(\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} < \frac{\mu_0-\mu}{\sigma/\sqrt{n}} - Z_{1-\alpha}\right)\\ &= \Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} + Z_\alpha\right)\\ &> \alpha.\end{aligned}$$

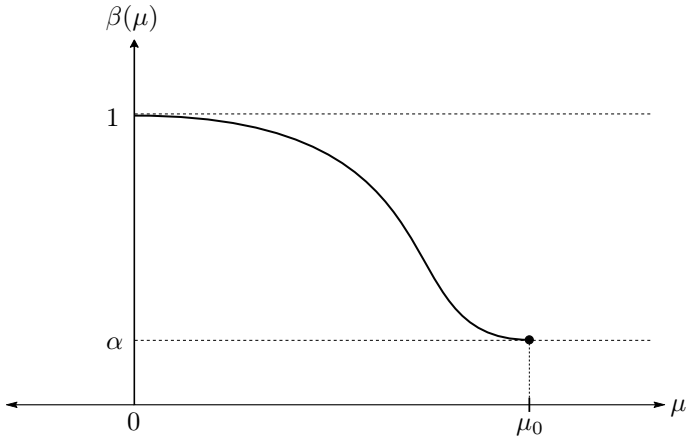


Figure 8.7: The power function in Example 8.4.9.

◇

Chapter 9

Likelihood-based Inference

9.1 Week 16: Lecture 1

9.1.1 Likelihood

Tue 15 Mar 14:00

Definition 9.1.1. Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, and $f_{\mathbf{Y}}(\mathbf{y}; \theta)$. The **likelihood** or **likelihood function** is

$$L_{\mathbf{Y}}(\theta; \mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}; \theta).$$

The **log-likelihood** function is

$$\ell_{\mathbf{Y}}(\theta; \mathbf{y}) = \log L_{\mathbf{Y}}(\theta; \mathbf{y}).$$

Note that

$$\begin{aligned} L_{\mathbf{Y}}(\theta; \mathbf{y}) &= f_{\mathbf{Y}}(\mathbf{y}; \theta) \\ &= \prod_{i=1}^{n} f_{Y_1}(y_i; \theta) \\ &= \prod_{i=1}^{n} L_Y(\theta; y_i) \end{aligned}$$

Then

$$\ell_{\mathbf{Y}}(\theta; y) = \sum_{i=1}^{n} \ell l_Y(\theta; y_i).$$

Remember, we can have $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^T$, so

$$L_{\mathbf{Y}}(\boldsymbol{\theta}; \mathbf{y}) = L_{\mathbf{Y}}(\theta_1, \ldots, \theta_r; \mathbf{y}).$$

Example 9.1.2. Let $Y_1, \ldots, Y_n \sim \text{Exp}(\lambda)$. Then

$$\begin{aligned} L(\lambda_i; \mathbf{y}) &= \prod_{i=1}^{n} f_Y(y; \lambda) \\ &= \prod_{i=1}^{n} \lambda e^{-\lambda y_i} \\ &= \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i} \\ &= \lambda^n e^{-\lambda n \bar{y}}. \end{aligned}$$

This implies

$$\ell(\lambda; \mathbf{y}) = n \log \lambda - n\bar{y}\lambda, \quad \lambda > 0.$$

◇

Example 9.1.3. Let $Y_1, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2)$, with $\boldsymbol{\theta} = (\mu, \sigma^2)^T$. Then we have

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{y}) &= L(\mu, \sigma^2; \mathbf{y}) \\ &= \prod_{i=1}^{n} f_Y(y_i; \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right) \\ &= (2\pi\sigma^2)^{\frac{-n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right), \quad \mu \in \mathbb{R},\ \sigma^2 > 0. \end{aligned}$$

This implies that

$$\ell(\mu, \sigma^2; y) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2.$$

◇

Remark 9.1.4. So, what exactly are we going to do with this likelihood function? How do we interpret it, how does this let us accomplish anything? Quick aside: maximizing the likelihood also maximizes the log-likelihood, because both are increasing functions. Moreover, why do we like the log-likelihood? Because the likelihood is a product, while the log-likelihood is a sum, and you'd rather differentiate sums than products.

9.1.2 The Score Function

Definition 9.1.5. We define the **score function** as

$$s_{\mathbf{Y}}(\theta; \mathbf{y}) = \frac{\partial}{\partial \theta} \ell_{\mathbf{Y}}(\theta; \mathbf{y}).$$

By the chain rule,

$$\begin{aligned} s_{\mathbf{Y}}(\theta;\mathbf{y}) &= \frac{\partial}{\partial\theta}\log L_{\mathbf{Y}}(\theta;\mathbf{y}) \\ &= \frac{1}{L_{\mathbf{Y}}(\theta;\mathbf{y})}\frac{\partial}{\partial\theta}L_{\mathbf{Y}}(\theta;\mathbf{y}). \end{aligned}$$

Now, what is $\mathbb{E}(s_{\mathbf{Y}}(\theta;\mathbf{Y}))$?

$$\begin{aligned} \mathbb{E}(s_{\mathbf{Y}}(\theta;\mathbf{Y})) &= \int_{\mathbb{R}^n} s_{\mathbf{Y}}(\theta;\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y};\theta)\ d\mathbf{y} \\ &= \int_{\mathbb{R}^n} \frac{\frac{\partial}{\partial\theta}L_{\mathbf{Y}}(\theta;\mathbf{y})}{L_{\mathbf{Y}}(\theta;\mathbf{y})} f_{\mathbf{Y}}(\mathbf{y};\theta)\ d\mathbf{y} \\ &= \int_{\mathbb{R}^n} \frac{\frac{\partial}{\partial\theta}L_{\mathbf{Y}}(\theta;\mathbf{y})}{L_{\mathbf{Y}}(\theta;\mathbf{y})} L_{\mathbf{Y}}(\theta;\mathbf{y})\ d\mathbf{y} \\ &= \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta}L_{\mathbf{Y}}(\theta;\mathbf{y})\ d\mathbf{y} \\ &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} L_{\mathbf{Y}}(\theta;\mathbf{y})\ d\mathbf{y} \\ &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} f_{\mathbf{Y}}(\mathbf{y};\theta)\ d\mathbf{y} \\ &= \frac{\partial}{\partial\theta}[1] \\ &= 0. \end{aligned}$$

Remember this!

9.1.3 Fisher Information

Remark 9.1.6. In general, log-likelihood functions tend to be concave.

We generally want to have a log-likelihood. Consider the following (crude) figure:

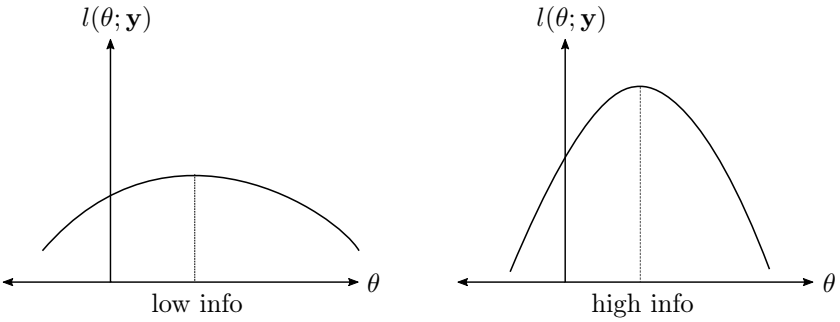


Figure 9.1: Two log-likelihood functions. The function that is relatively flat (left) doesn't tell us much about θ since its likelihood function does not differ much, and therefore has low information. The likelihood of different values of θ differs significantly on the right function, and so we say it has higher information.

Definition 9.1.7. The **Fisher information** is

$$\mathcal{I}_{\mathbf{Y}}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ell_{\mathbf{Y}}(\theta; \mathbf{y})\right)^2\right].$$

9.1.4 Properties of Information

Note that

$$\begin{aligned}\mathcal{I}_{\mathbf{Y}}(\theta) &= \mathbb{E}[(s_{\mathbf{Y}}(\theta; \mathbf{y}))^2] \\ &= \mathrm{Var}(s_{\mathbf{Y}}(\theta; \mathbf{Y})),\end{aligned}$$

since $\mathbb{E}(s_{\mathbf{Y}}(\theta; \mathbf{Y})) = 0$.

Proposition 9.1.8. For a random sample $\mathbf{Y}$ and parameter θ,

$$\begin{aligned}\mathcal{I}_{\mathbf{Y}}(\theta) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} l_{\mathbf{Y}}(\theta; \mathbf{y})\right] \\ &= -\mathbb{E}\left[\frac{\partial}{\partial\theta} s_{\mathbf{Y}}(\theta; \mathbf{y})\right].\end{aligned}$$

Proof. Observe that

$$\begin{aligned}
\frac{\partial}{\partial\theta}\mathbb{E}[s(\theta;\mathbf{Y})] &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} s(\theta;\mathbf{Y})f_Y(\mathbf{y};\theta)\ d\mathbf{y} \\
&= \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta}\left(s_{\mathbf{Y}}(\theta;\mathbf{y})f_Y(\mathbf{y};\theta)\right)\ d\mathbf{y} \\
&= \int_{\mathbb{R}^n} \left(\frac{\partial}{\partial\theta}s(\theta;\mathbf{y})\right) f_{\mathbf{Y}}(\mathbf{y};\theta) + s(\theta;\mathbf{y})\left(\frac{\partial}{\partial\theta}f_{\mathbf{Y}}(\mathbf{y};\theta)\right)\ d\mathbf{y} \\
&= \int_{\mathbb{R}^n} \left(\frac{\partial}{\partial\theta}s(\theta;\mathbf{y})\right) f_{\mathbf{Y}}(\mathbf{y};\theta)\ d\mathbf{y} + \int_{\mathbb{R}^n} s(\theta;\mathbf{y})\left(\frac{\partial}{\partial\theta}f_{\mathbf{Y}}(\mathbf{y};\theta)\right)\ d\mathbf{y} \\
&= \mathbb{E}\left[\frac{\partial}{\partial\theta}s(\theta;\mathbf{y})\right] + \int_{\mathbb{R}^n} \frac{\frac{\partial}{\partial\theta}L_{\mathbf{Y}}(\theta;\mathbf{y})}{L_{\mathbf{Y}}(\theta;\mathbf{y})}\left(\frac{\partial}{\partial\theta}L_{\mathbf{Y}}(\theta;\mathbf{y})\right)\ d\mathbf{y} \\
&= \mathbb{E}\left[\frac{\partial}{\partial\theta}s(\theta;\mathbf{y})\right] + \int_{\mathbb{R}^n} \frac{\left(\frac{\partial}{\partial\theta}L_{\mathbf{Y}}(\theta;\mathbf{y})\right)^2}{L_{\mathbf{Y}}(\theta;\mathbf{y})}\ d\mathbf{y} \\
&= \mathbb{E}\left[\frac{\partial}{\partial\theta}s(\theta;\mathbf{y})\right] + \int_{\mathbb{R}^n} \left(\frac{\partial}{\partial\theta}l_{\mathbf{Y}}(\theta;\mathbf{y})\right)^2 f_{\mathbf{Y}}(\theta;\mathbf{y})\ d\mathbf{y} \\
&= \mathbb{E}\left[\frac{\partial}{\partial\theta}s(\theta;\mathbf{y})\right] + \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}l_{\mathbf{Y}}(\theta;\mathbf{y})\right)^2\right].
\end{aligned}$$

Since $\frac{\partial}{\partial\theta}\mathbb{E}[s(\theta;\mathbf{Y})] = 0$,

$$\mathcal{I}_{\mathbf{Y}}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}l_{\mathbf{Y}}(\theta;\mathbf{y})\right)^2\right] = -\mathbb{E}\left[\frac{\partial}{\partial\theta}s(\theta;\mathbf{y})\right].$$

□

Example 9.1.9. Applying this to $\text{Exp}(\lambda)$, we have

$$\ell(\lambda;\mathbf{y}) = n\log\lambda - n\bar{y}\lambda,$$

which implies that

$$\begin{aligned}
s(\lambda;\mathbf{y}) &= \frac{\partial}{\partial\lambda}\ell(\lambda;\mathbf{y}) \\
&= \frac{n}{\lambda} - n\bar{y}.
\end{aligned}$$

Now we check the expectation:

$$\begin{aligned}
\mathbb{E}(s(\lambda;\mathbf{Y})) &= \mathbb{E}\left(\frac{n}{\lambda} - n\bar{Y}\right) \\
&= \frac{n}{\lambda} - \frac{n}{\lambda} \\
&= 0.
\end{aligned}$$

Now we have

$$\begin{aligned}\mathcal{I}(\lambda) &= \text{Var}(s_{\mathbf{Y}}(\lambda, \mathbf{Y})) \\ &= \text{Var}\left(\frac{n}{\lambda} - n\bar{Y}\right) \\ &= (-n)^2 \,\text{Var}(\bar{Y}) \\ &= \frac{\frac{1}{\lambda^2}}{n} \\ &= \frac{n}{\lambda^2}.\end{aligned}$$

Or,

$$\begin{aligned}\mathcal{I}(\lambda) &= -\mathbb{E}\left[\frac{\partial}{\partial\lambda} s_{\mathbf{Y}}(\lambda; \mathbf{Y})\right] \\ &= -\mathbb{E}\left[-\frac{n}{\lambda^2}\right] \\ &= \frac{n}{\lambda^2}.\end{aligned}$$

◇

9.2 Week 16: Lecture 2

9.2.1 Vector Parameter Extension

Thu 17 Feb 10:00

Recall that if $\mathbf{Y}$ is a random sample,

$$\begin{aligned}L_{\mathbf{Y}}(\theta, \mathbf{y}) &= \prod_{i=1}^{n} L_Y(\theta; y_i) \\ \Rightarrow l_{\mathbf{Y}}(\theta, \mathbf{y}) &= \sum_{i=1}^{n} l_Y(\theta; y_i) \\ \Rightarrow s_{\mathbf{Y}}(\theta, \mathbf{y}) &= \sum_{i=1}^{n} s_Y(\theta; y_i) \\ \Rightarrow \mathcal{I}_{\mathbf{Y}}(\theta) &= nI_Y(\theta).\end{aligned}$$

Example 9.2.1. Let $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ be a random sample. Then

$$\begin{aligned}L_Y(\mu; y_1) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_1-\mu}{\sigma}\right)^2} \\ \Rightarrow l_Y(\mu; y_1) &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_1-\mu)^2}{2\sigma^2} \\ \Rightarrow s_Y(\mu; y_1) &= \frac{2(y_1-\mu)}{2\sigma^2} = \frac{y_1-\mu}{\sigma^2} \\ \Rightarrow \mathcal{I}_Y(\mu) &= -\mathbb{E}\left[\frac{\partial}{\partial\mu} s_Y(\mu; Y_1)\right] = -\mathbb{E}\left[-\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2} \\ \Rightarrow \mathcal{I}_{\mathbf{Y}}(\mu) &= \frac{n}{\sigma^2}\end{aligned}$$

◇

Definition 9.2.2. A **normalizing constant** ensures that the CDF equals 1.

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^T$, then

$$\nabla_{\boldsymbol{\theta}} l_{\mathbf{Y}}(\theta; \mathbf{y}) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} l_{\mathbf{Y}}(\theta; \mathbf{y}) \\ \vdots \\ \frac{\partial}{\partial \theta_r} l_{\mathbf{Y}}(\theta; \mathbf{y}) \end{pmatrix}$$

is the same vector. Moreover,

$$I\mathbf{Y}(\boldsymbol{\theta}) = \mathbb{E}[s_{\mathbf{Y}}(\boldsymbol{\theta}; \mathbf{Y}) s_{\mathbf{Y}}(\boldsymbol{\theta}; \mathbf{Y})^t]$$

is the $r \times r$ information matrix. As in the scalar case, $\mathbb{E}[s_{\mathbf{Y}}(\boldsymbol{\theta}; \mathbf{Y})] = 0$, so

$$\begin{aligned} \mathcal{I}_{\mathbf{Y}}(\boldsymbol{\theta}) &= \mathrm{Var}(s_{\mathbf{Y}}(\boldsymbol{\theta}; \mathbf{Y})) \\ &= -\mathbb{E}[\nabla_{\boldsymbol{\theta}^T} s_{\mathbf{Y}}(\boldsymbol{\theta}; \mathbf{Y})] \end{aligned}$$

9.2.2 Maximum Likelihood Estimation

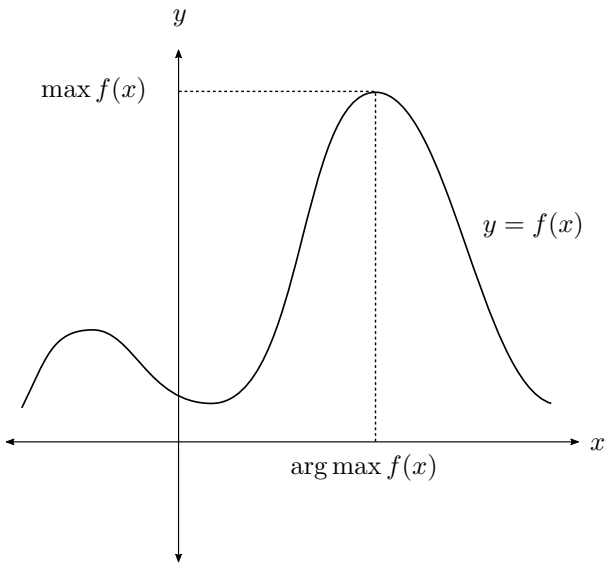


Figure 9.2: For a function $y = f(x)$, the $\arg\max f(x)$ is the x value which outputs $\max f(x)$.

Definition 9.2.3. Let $Y_1, \ldots, Y_n \sim f_Y(y;\theta)$ be a random sample. The **maximum-likelihood estimate** of θ is

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} L(\theta;\mathbf{y}) \\ &= \arg\max_{\theta} \ell(\theta;y).\end{aligned}$$

If $\hat{\theta}(\mathbf{y})$ is the maximum-likelihood estimate, then $\hat{\theta}(\mathbf{Y})$ is the **maximum-likelihood estimator.**

Example 9.2.4. If $\hat{\mu}(\mathbf{y}) = \bar{y}$, then $\hat{\mu}(\mathbf{Y}) = \bar{Y}$. ◇

Note. Be careful, MLE can mean either maximum-likelihood *estimate* or maximum-likelihood *estimator*. A bit inconvenient.

If $\ell(\theta;\mathbf{y})$ is differentiable, we find $\hat{\theta}$ by taking

$$\left.\frac{\partial\,\ell(\theta;y)}{\partial\theta}\right|_{\theta=\hat{\theta}} = 0,$$

since $s(\hat{\theta};\mathbf{y}) = 0$.

The maximum likelihood estimate $\hat{\theta}$ is consistent, i.e., $\hat{\theta} \to^p \theta$ as $n \to \infty$. But it gets even better than this!

Proposition 9.2.5. Let $\mathbf{Y}$ be a random sample with parameter θ. Then

$$\sqrt{n}(\hat{\theta} - \theta) \to^d \text{Normal}(0, \mathcal{I}_{\mathbf{Y}}^{-1}(\theta)), \quad \text{as } n \to \infty.$$

Proof. This proof is quite involved. See Proposition 9.3.4 in the textbook for a sketch proof, and see Section 5.2 here for a rigorous proof. □

Remark 9.2.6. Note that

1. This is a *general result* about the error of estimation.
2. It puts a distribution on the error of estimation.
3. This result shows that MLEs are asymptotically unbiased.
4. The accuracy of the MLE is the inverse of the information matrix.

Note. The quantity $\sqrt{n}(\hat{\theta} - \theta)$ is *not* a pivotal, but it is functionally close to one.

In principle, if you have a pivotal, then you can build a confidence interval. This property shows that for large samples, we have an approximate pivotal in the MLE.

Week 17: Reading Week

9.3 Week 18: Lecture 1

9.3.1 More on MLEs

Tue 1 Mar 14:00

Example 9.3.1. Let $Y_1, \ldots, Y_n \sim \text{Exp}(\lambda)$. Then

$$L(\lambda; \mathbf{y}) = \prod_{i=1}^{n} \lambda e^{0\lambda y_i} = \lambda^n e^{-\lambda n \mathbf{y}}$$
$$\Rightarrow \ell(\lambda; \mathbf{y}) = n \log \lambda - n\mathbf{y}\lambda$$
$$\Rightarrow s(\lambda; \mathbf{y}) = \frac{n}{\lambda} - n\mathbf{y}..$$

Setting this final expression equal to 0 yields

$$\frac{n}{\hat{\lambda}} - n\mathbf{y} = 0$$
$$\Rightarrow \hat{\lambda} = \frac{1}{\mathbf{y}} \qquad \text{(estimate)} \Rightarrow \hat{\lambda} = \frac{1}{\mathbf{Y}} \quad \text{(estimator)}.$$

Take

$$\frac{\partial^2}{\partial \lambda^2} \ell(\hat{\lambda}; \mathbf{y}) = -\frac{n}{\lambda^2},$$

which is negative. Hence, we know that $\hat{\lambda}$ is indeed a maximum. ◇

Remark 9.3.2. Given $Y_1, \ldots, Y_n \sim F_Y(y; \theta)$, and $\hat{\theta}$ as the MLE of θ,

$$\hat{\theta} \to^d \text{Normal}(\theta, \mathcal{I}_{\mathbf{Y}}(\theta)^{-1}) \quad \text{as } n \to \infty..$$

Since $\mathcal{I}_{\mathbf{Y}}(\theta) = n\mathcal{I}_Y(\theta)$, we have

$$\mathcal{I}_{\mathbf{Y}}(\theta)^{-1} = \frac{1}{n\mathcal{I}_Y(\theta)}.$$

If you have a large sample, we can use this limiting distribution as an approximation, and thus construct a pivotal and therefore a confidence interval.

Example 9.3.3. In the Exp(λ) case, we have

$$\hat{\lambda} = \frac{1}{\overline{Y}}, \quad \mathcal{I}_{\mathbf{Y}} = \frac{n}{\lambda^2},$$

so $\hat{\lambda} \to^d \text{Normal}\left(\lambda, \frac{\lambda^2}{n}\right)$. In this example we use the rate parameter λ. There are, however, two parameters for the exponential. Recall that the scale parameter is $\theta = \frac{1}{\lambda}$. Now is it the case that $\hat{\theta} = \frac{1}{\hat{\lambda}}$? Yes, since a particular value of λ maximizes the likelihood, that same value maximizes the likelihood of θ. Thus, likelihood is transformation-invariant. This holds for both 1-1 and many-to-one transformations. ◇

Example 9.3.4 (Odds)**.** Let $Y_1, \ldots, Y_n \sim \text{Bernoulli}(p)$. We often prefer to work with **odds:** $\frac{p}{1-p}$. What is the MLE of odds? The MLE of p is $\hat{p} = \bar{Y}$. So the MLE of odds is

$$\frac{\hat{p}}{1-\hat{p}} = \frac{\bar{Y}}{1-\bar{Y}}.$$

◇

Note. Check out the section on induced likelihood.

9.3.2 Likelihood-ratio test

Definition 9.3.5. Let

$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1,$$

i.e., we have $\Theta_0 \cup \Theta_1 = \Theta$. The **likelihood ratio test statistic** is

$$r(\mathbf{Y}) = \frac{\sup_{\theta \in \Theta} L(\theta; \mathbf{Y})}{\sup_{\theta \in \Theta_0} L(\theta; \mathbf{Y})} = \frac{L(\hat{\theta}; \mathbf{Y})}{L(\hat{\theta}_0; \mathbf{Y})},$$

where $L(\hat{\theta}_0; \mathbf{Y})$ is the **constrained MLE.**

The likelihood ratio test tells us to reject H_0 for large values of $r(\mathbf{Y})$. We find some value k then reject H_0 when $r(\mathbf{Y}) > k$. How do we find k? We set

$$P_{H_0}(r(\mathbf{Y}) > k) = \alpha.$$

Example 9.3.6. Let $Y_1, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2)$, with σ^2 known. Let

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

Note that

$$L(\mu; \mathbf{y}) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]}$$

and

$$\begin{aligned} r(\mathbf{Y}) &= \frac{L(\hat{\mu}; \mathbf{Y})}{L(\mu_0; \mathbf{Y})} \\ &= \frac{(2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(n-1)s^2} e^{-\frac{n(\bar{Y}-\bar{Y})}{2\sigma^2}}}{(2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(n-1)s^2} e^{-\frac{n(\bar{Y}-\mu_0)}{2\sigma^2}}} \\ &= e^{\frac{n}{2\sigma^2}(\bar{Y}-\mu_0)^2} \end{aligned}$$

This is an increasing function of $|\bar{Y} - \mu_0|$, i.e., $r(\bar{Y})$ is large when $\bar{Y}$ is far from

μ_0. Then

$$\begin{aligned}\alpha &= P_{H_0}(r(\mathbf{Y}) > k) \\ &= P_{\mu_0}\left(e^{\frac{n}{2\sigma^2}(\bar{Y}-\mu_0)^2} > k\right) \\ &= P_{\mu_0}\left(\frac{n(\bar{Y}-\mu_0)^2}{\sigma^2} > 2\log k\right).\end{aligned}$$

Since $\frac{n(\bar{Y}-\mu_0)^2}{\sigma^2} \sim^{H_0} \chi^2_1$, we can find a value of k.

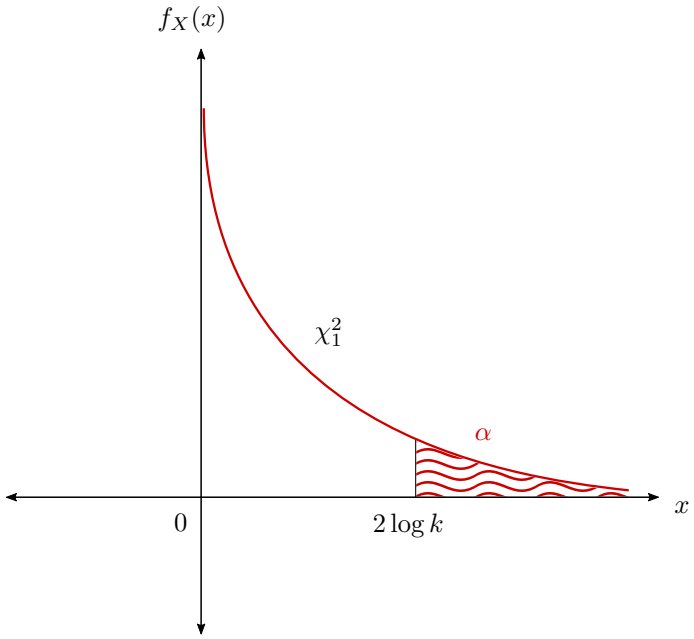


Figure 9.3: The rejection region of the likelihood-ratio test in Example 9.3.6, with $X \sim \chi^2_1$.

◇

Chapter 10

Inferential Theory

10.1 Week 18: Lecture 2

10.1.1 Sufficiency

Thu 3 Mar 10:00

When you summarize information into a statistic, you are also throwing away some information included within the sample. Moreover, not all statistics are equally useful. What we're trying to do with sufficiency is to show if something is a useful summary, and if it throws away too much information.

Definition 10.1.1. Let $Y_1, \ldots, Y_n \sim F_Y(y; \theta)$ be a random sample. Consider the statistic $U = h(\mathbf{Y})$. We say U is **sufficient** for θ if the conditional distribution of $\mathbf{Y} \mid U$ does not depend on θ.

Note. Both $\mathbf{Y}$ and U are both sample statistics.

First, consider the discrete case:

Discrete Case

Let $f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y} \mid \mathbf{u})$. If $\mathbf{Y} = \mathbf{y}$, then $h(\mathbf{Y}) = h(\mathbf{y})$, which implies that $\mathbf{U} = \mathbf{u}$. We can say that $\mathbf{Y} = \mathbf{y}$ is a *subset* of $\mathbf{U} = \mathbf{u}$. Note that we can express

$$\begin{aligned} f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y} \mid \mathbf{u}) &= \frac{f_{\mathbf{Y},\mathbf{U}}(\mathbf{y}, \mathbf{u})}{f_{\mathbf{U}}} \\ &= \frac{P(\mathbf{Y} = \mathbf{y}, \mathbf{U} = \mathbf{u})}{P(\mathbf{U} = \mathbf{u})} \\ \Rightarrow f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y} \mid \mathbf{u}) &= \begin{cases} \frac{f_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{U}(\mathbf{u})}}, & \text{if } \mathbf{u} = h(\mathbf{y}) \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Example 10.1.2. Let $Y_1, \ldots, Y_n \sim$ Bernoulli(p) and consider $U = \sum_{i=1}^n Y_i \sim \text{Bin}(n, p)$. Then

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \prod_{i=1}^n f_Y(y_i) \\ &= \prod_{i=1}^n p^{y_i}(1-p)^{1-y_i} \\ &= p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i} \\ &= p^u(1-p)^{n-u}, \end{aligned}$$

where $u = \sum_{i=1}^n y_i$. So

$$\begin{aligned} f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y} \mid u) &= \frac{f_{\mathbf{Y}}(\mathbf{y})}{f_U(u)} \\ &= \frac{p^u(1-p)^{n-u}}{\binom{n}{u}p^u(1-p)^{n-u}} \\ &= \frac{1}{\binom{n}{u}}, \end{aligned}$$

which does not depend on p. This implies that $U = \sum_{i=1}^n Y_i$ is a sufficient statistic for p. Hence, $\mathbf{y}$ is a vector of 0s and 1s with $\binom{n}{u}$ possibilities. Moreover, this tells us nothing about p. ◇

Remark 10.1.3. The key idea of sufficiency is that knowing the raw data tells us nothing about the parameter. Estimation of unknown parameters should be based on sufficient statistics. We will see that there is a strong link between sufficiency and MLEs.

10.1.2 Finding Sufficient Statistics

Theorem 10.1.4 (Factorisation Criterion)**.** If we can express the joint mass/density of $\mathbf{Y}$ in the form

$$f_{\mathbf{Y}}(\mathbf{y}; \theta) = L(\theta; \mathbf{y}) = b(\theta, h(\mathbf{y}))c(\mathbf{y}),$$

then $\mathbf{U} = h(\mathbf{Y})$ is sufficient for θ.

Proof. If $\mathbf{U}$ is sufficient for θ, then $f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y} \mid \mathbf{u}) = \frac{f_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{U}}(\mathbf{u})}$ does not depend on θ. This implies that

$$f_{\mathbf{Y}}(\mathbf{y}) = \underbrace{f_{\mathbf{U}}(\mathbf{u})}_{b(\theta, h(\mathbf{y}))} \underbrace{f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y} \mid \mathbf{u})}_{c(\mathbf{y})}.$$

The "if" direction for the discrete case is covered in Theorem 10.1.14, page

330 in the course textbook. The full proof for both cases is very involved, and can be found here. □

Example 10.1.5 (Bernoulli Case)**.** Let $Y_1, \dots, Y_n \sim \text{Bernoulli}(p)$. Then

$$L(p; \mathbf{y}) = \underbrace{p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i}}_{b\left(p, \sum_{i=1}^n y_i\right)},$$

with $c(\mathbf{y}) = 1$. This implies that $\sum_{i=1}^n Y_i$ is sufficient for p, Note that this isn't unique, as we can instead write this in terms of $\bar{Y}$. ◇

Example 10.1.6 (Poisson Case)**.** We have

$$L(\lambda; \mathbf{y}) = \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \underbrace{e^{-n\lambda}\lambda^{n\bar{y}}}_{b(\lambda, \bar{y})} \underbrace{\frac{1}{\prod_{i=1}^n y_i!}}_{c(\mathbf{y})},$$

which implies that $\bar{Y}$ is sufficient for λ. ◇

MLEs are always functions of sufficient statistics. The part that is not a function of the sufficient statistic plays no role in finding the MLE.

10.2 Week 19: Lecture 1

10.2.1 More on Sufficient Statistics

Tue 8 Mar 14:00

Example 10.2.1. Let $Y_1, \dots, Y_n \sim \text{Normal}(\mu, \sigma^2)$. Then

$$\begin{aligned}
\mathcal{L}(\mu, \sigma^2; \mathbf{y}) &= (2\pi\sigma^2)^{-\frac{n}{2}} e^{\frac{-\sum_{i=1}^n (y_i - \mu)}{2\sigma^2}} \\
&= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}} \left(\sum y_i^2 - 2\mu \sum y_i + \mu^2 \right) \\
&= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}} \left(\sum (y_i - \bar{y})^2 + n(\bar{y} - \mu) \right) \\
&= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} e^{-\frac{(n-1)s^2}{2\sigma^2}} e^{-\frac{n}{2\sigma^2}(y - \bar{\mu})^2}.
\end{aligned}$$

◇

There are multi dimensional sufficient statistics. Grouping these differently gives different sufficient statistics. Which one is preferable?

Let $Y_1, \dots, Y_n \sim F_Y(y; \theta)$. Then $\mathbf{Y}$ is a sufficient statistic!

Suppose that $U = \sum_{i=1}^n Y_i$ is sufficient for θ. Let $\mathbf{V} = (Y_1, \sum_{i=2}^n Y_i)$. Then $\mathbf{V}$ is sufficient since you can figure out U from $\mathbf{V}$. But it doesn't make sense to use V, since it is two dimensional. Notice that $U = g(\mathbf{V})$. What does this tell us?

Suppose that V is a statistic. If U is a function of V alone (i.e., $U = g(V)$), then

Proposition 10.2.2. i. U is a statistic;

ii. if U is sufficient for θ, then V is also sufficient;

iii. if V is not sufficient, then U is not sufficient;

iv. if V is sufficient and g is injective, then U is also sufficient.

Proof. Exercise 10.1 in the course textbook. □

Definition 10.2.3. A statistic U is a **minimal sufficient statistic** if, for any other sufficient statistic V, U is a function of V.

A minimal sufficient statistic has the lowest number of dimensions among all sufficient statistics.

Proposition 10.2.4. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ with mass/density $f_{\mathbf{Y}}(\mathbf{y}; \theta)$. If we can find a function h such that:

$$h(\mathbf{y}) = h(\mathbf{x}) \iff \frac{f_{\mathbf{Y}}(\mathbf{y}; \theta)}{f_{\mathbf{Y}}(\mathbf{x}; \theta)} = k(\mathbf{y}, \mathbf{x})$$

with $\mathbf{x} = (x_1, \ldots, x_n)^T$ and where k does not depend on θ, then $h(\mathbf{Y})$ is a minimal sufficient statistic for θ.

Proof. Omitted. □

Example 10.2.5. Let $Y_1, \ldots Y_n \sim \text{Normal}(\mu, \sigma^2)$. Note that

$$\begin{aligned}
\frac{f_{\mathbf{Y}}(\mathbf{y}; \mu, \sigma^2)}{f_{\mathbf{Y}}(\mathbf{x}; \mu, \sigma^2)} &= \frac{(2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left(\sum_i y_i^2 - 2\mu \sum_i y_i + \mu^2\right)}}{(2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - 2\mu \sum_i x_i + \mu^2\right)}} \\
&= \exp\left(-\frac{1}{2\sigma^2}\left[\left(\sum_i y_i^2 - \sum_i x_i^2\right) - 2\left(\sum_i y_i - \sum_i x_i\right)\mu\right]\right) \\
&= k(\mathbf{y}, \mathbf{x}) \\
&\iff \left(\sum_i y_i, \sum_i y_i^2\right) = \left(\sum_i x_i, \sum_i x_i^2\right).
\end{aligned}$$

Then $\left(\sum_i Y_i, \sum_i Y_i^2\right)$ is a minimal sufficient statistic for $(\mu, \sigma^2)^T$. ◇

Proposition 10.2.6. If $\mathbf{Y}$ is a random sample, $\mathbf{U}$ is a statistic, and $\mathbf{S}$ is a sufficient statistic for θ, then $\mathbf{T} =: \mathbb{E}(\mathbf{U} \mid \mathbf{S})$ is a *statistic,* i.e., its value does not depend on θ.

Proof.

$$\begin{aligned}\mathbb{E}(\mathbf{U} \mid \mathbf{S} = \mathbf{s} = \mathbb{E}(h(\mathbf{Y}) \mid \mathbf{S} = \mathbf{s}) \\ &= \int_{\mathbb{R}^n} h(\mathbf{y}) \underbrace{f_{\mathbf{Y}|\mathbf{S}}(\mathbf{y} \mid \mathbf{s})}_{\text{dnd on } \theta}\ d\mathbf{y},\end{aligned}$$

i.e., it does not contain θ. □

The next lecture introduces the Rao-Blackwell Theorem.

10.3 Week 19: Lecture 2

10.3.1 The Rao-Blackwell Theorem

Thu 10 Mar 10:00

Theorem 10.3.1 (Rao-Blackwell Theorem)**.** Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be a random sample with parameter $\boldsymbol{\theta}$, estimator $\mathbf{U}$ (of $\boldsymbol{\theta}$), sufficient statistic $\mathbf{S}$. Then $\mathbf{T} = \mathbb{E}(\mathbf{U} \mid \mathbf{S})$ is an estimator with

$$\underset{\boldsymbol{\theta}}{\text{MSE}}(\mathbf{T}) \leq \underset{\boldsymbol{\theta}}{\text{MSE}}(\mathbf{U}).$$

Proof. Observe that

$$\begin{aligned}\underset{\boldsymbol{\theta}}{\text{MSE}}(\mathbf{U}) &= \mathbb{E}[(\mathbf{U} - \boldsymbol{\theta})^2] \\ &= \mathbb{E}[\mathbb{E}[(\mathbf{U} - \boldsymbol{\theta})^2 \mid \mathbf{S}]] \\ &\geq \mathbb{E}[[\mathbb{E}((\mathbf{U} - \boldsymbol{\theta}) \mid \mathbf{S})]^2] \\ &= \mathbb{E}[(\mathbb{E}(\mathbf{U} \mid \mathbf{S}) - \boldsymbol{\theta})^2] \\ &= \mathbb{E}[(\mathbf{T} - \boldsymbol{\theta})^2] \\ &= \underset{\boldsymbol{\theta}}{\text{MSE}}(\mathbf{T}).\end{aligned}$$

□

Also notice that

$$\mathbb{E}(\mathbf{T}) = \mathbb{E}[\mathbb{E}(\mathbf{U} \mid \mathbf{S})] = \mathbb{E}(\mathbf{U}).$$

For equality, we need

$$\mathbb{E}[(\mathbf{U} - \boldsymbol{\theta})^2 \mid \mathbf{S}] = [\mathbb{E}(\mathbf{U} - \boldsymbol{\theta}) \mid \mathbf{S}]^2 \iff \text{Var}(\mathbf{U} - \boldsymbol{\theta} \mid \mathbf{S}) = 0,$$

so $\mathbf{U}$ is a function of $\mathbf{S}$.

10.3.2 Cramer-Rao lower bound

What is the best unbiased estimator? The one with the lowest variance: the minimum-variance unbiased estimator (MVUE).

Theorem 10.3.2 (Cramer-Rao Bound)**.** Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be a random sample with distribution $f_Y(y;\theta)$. If $U = h(\mathbf{Y})$ is an *unbiased* estimator of $g(\theta)$, then:

$$\text{Var}(U) \geq \frac{(\frac{d}{d\theta} g(\theta))^2}{\mathcal{I}_{\mathbf{Y}}(\theta)}.$$

Special case: if $g(\theta) = \theta$, then $\text{Var}(U) \geq \frac{1}{\mathcal{I}_{\mathbf{Y}}(\theta)}$.

An unbiased estimator that is fully-efficient attains this lower bound. MLEs are asymptotically unbiased, asymptotically normal, and asymptotically fully-efficient.

Proof. Recall that $\mathbb{E}(s(\theta;\mathbf{Y})) = 0$, $\text{Var}(s(\theta,\mathbf{Y})) = \mathcal{I}_{\mathbf{Y}}(\theta)$. If $U = h(\mathbf{Y})$ is unbiased for $g(\theta)$, then

$$\begin{aligned} g(\theta) &= \mathbb{E}(U) \\ &= \int_{\mathbb{R}^n} h(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y};\theta)\ d\mathbf{y}, \end{aligned}$$

which implies that

$$\begin{aligned} \frac{d}{d\theta} g(\theta) &= \int_{\mathbb{R}^n} h(\mathbf{y}) \frac{d}{d\theta} f_{\mathbf{Y}}(\mathbf{y};\theta)\ d\mathbf{y} \\ &= \int_{\mathbb{R}^n} h(\mathbf{y}) s(\theta;\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y};\theta)\ d\mathbf{y} \\ &= \mathbb{E}[h(\mathbf{Y}) s(\theta;\mathbf{Y})] \\ &= \text{Cov}(h(\mathbf{Y}), s(\theta;\mathbf{Y})), \end{aligned}$$

which implies that

$$\begin{aligned} \left(\frac{d}{d\theta} g(\theta)\right)^2 &= (\text{Cov}(h(\mathbf{Y}), s(\theta;\mathbf{Y})))^2 \\ &\leq \text{Var}(h(\mathbf{Y}))\, \text{Var}(s(\theta;\mathbf{Y})) \\ &= \text{Var}(h(\mathbf{Y})) \mathcal{I}_{\mathbf{Y}}(\theta) \\ &\Rightarrow \text{Var}(h(\mathbf{Y})) \geq \frac{\left(\frac{d}{d\theta} g(\theta)\right)^2}{\mathcal{I}_{\mathbf{Y}}(\theta)}. \end{aligned}$$

□

When do we have equality? We need

$$\begin{aligned} &\text{Cov}(h(\mathbf{Y}), s(\theta;\mathbf{Y}))^2 = \text{Var}(h(\mathbf{Y}))\, \text{Var}(s(\theta;\mathbf{Y})) \\ \iff &\text{Corr}(h(\mathbf{Y}), s(\theta;\mathbf{Y}))^2 = 1, \end{aligned}$$

so $s(\theta;\mathbf{Y}) = b(\theta)U + a(\theta)$, *but*

$$\mathbb{E}(s(\theta;\mathbf{Y})) = 0 \Rightarrow b(\theta)\mathbb{E}(U) + a(\theta) = 0,$$

which in turn implies that $a(\theta) = -b(\theta)g(\theta)$, and we can write

$$s(\theta; \mathbf{Y}) = b(\theta)[U - g(\theta)].$$

Then U is the MLE.

So, if we can write the score function as

$$s(\theta; \mathbf{Y}) = b(\theta)[U - g(\theta)],$$

where $b(\theta)$ is a function of θ *only*, the quantity of interest is $g(\theta)$, and U is our estimator of $g(\theta)$, we can deduce:

(1) $U = h(\mathbf{Y})$ is the MLE of $g(\theta)$.

(2) U is unbiased for $g(\theta)$.

(3) U attains the Cramer-Rao lower bound (CRLB).

(4) U is the MVUE.

If we *can't*, then no unbiased estimator can attain the CRLB.

Example 10.3.3 (Normal Case)**.** Let $Y_1, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2)$, where σ^2 is known. Then

$$\begin{aligned} L(\mu; \mathbf{y}) &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + n(\bar{Y} - \mu)^2\right) \\ &\propto \exp\left(-\frac{n}{2\sigma^2}(\bar{Y} - \mu)^2\right) \\ &\Rightarrow \ell(\mu; \mathbf{y}) = -\frac{n}{2\sigma^2}(\bar{Y} - \mu)^2 + C \\ &\Rightarrow s(\mu; \mathbf{Y}) = \frac{n}{\sigma^2}(\bar{Y} - \mu) \end{aligned}$$

which is in the form $b(\mu)[U - g(\mu)]$, so $\hat{\mu}$ is the MLE, it is unbiased, it attains the CRLB, and it is the MVUE. So

$$\text{Var}(\bar{Y}) = \frac{1}{\mathcal{I}_{\mathbf{Y}}(\mu)} = \frac{1}{n/\sigma^2} = \frac{\sigma^2}{n}.$$

This is an example of a case where the CRLB is attainable. ◇

Proposition 10.3.4. Let $\mathbf{Y}$ be a random sample. Let $\phi_S(\mathbf{Y})$ be a decision function, c_s be the critical region, and $\beta_S(\theta)$ be the power function. Then

$$\beta_S(\theta) = \mathbb{E}_\theta[\phi_S(\mathbf{Y})]$$

Proof.

$$\begin{aligned}
\mathbb{E}_\theta[\phi_S(\mathbf{Y})] &= \int_{\mathbb{R}^n} \phi_S(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y};\theta)\ d\mathbf{y} \\
&= \int_{C_S} \phi_S(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y};\theta)\ d\mathbf{y} \\
&= \int_{C_S} \phi_S(\mathbf{y})\ f(\mathbf{Y})(\mathbf{y};\theta)\ d\mathbf{y} \\
&= \int_{C_S} f_{\mathbf{Y}}(\mathbf{y};\theta) \\
&= P_\theta(\mathbf{Y} \in C_S) \\
&= P_\theta(\text{reject } H_0) \\
&= \beta_S(\theta).
\end{aligned}$$

□

Remark 10.3.5. Let $H_0 : \theta = \theta_0$, and $H_1 : \theta = \theta_1$. Then the size is $\beta(\theta_0)$, and the power is $\beta(\theta_1)$. We say that T is a most powerful test (MPT) if $\beta_T(\theta_1) \geq \beta_S(\theta_1)$ for all tests S such that $\beta_T(\theta_0) = \beta_S(\theta_0)$.

10.3.3 Neyman-Pearson Lemma

Lemma 10.3.6 (Neyman-Pearson Lemma)**.** For testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, the MPT of size α is the one which rejects H_0 if

$$\frac{L_{\mathbf{Y}}(\theta_1;\mathbf{y})}{L_{\mathbf{Y}}(\theta_0;\mathbf{y})} > k_\alpha \iff L_{\mathbf{Y}}(\theta_1;\mathbf{y}) - k_\alpha L_{\mathbf{Y}}(\theta_0;\mathbf{y}) > 0.$$

Critical Region: $C_T = \{\mathbf{y} \in \mathbb{R}^n : L_{\mathbf{Y}}(\theta_1;\mathbf{y}) - k_\alpha L_{\mathbf{Y}}(\theta_0;\mathbf{y}) > 0\}$

Proof. Observe that

$$\begin{aligned}
\beta_S(\theta_1) - k_\alpha \beta_S(\theta_0) &= \int_{R^n} \phi_S(\mathbf{y})[f_{\mathbf{Y}}(\mathbf{y};\theta_1) - k_\alpha(f_{\mathbf{Y}}(\mathbf{y};\theta_0))]\, d\mathbf{y} \\
&\leq \int_{C_T} \phi_S(\mathbf{y})[f_{\mathbf{Y}}(\mathbf{y};\theta_1) - k_\alpha f_{\mathbf{Y}}(\mathbf{y};\theta_0)]\, d\mathbf{y} \\
&\leq \int_{C_T} \phi_T(\mathbf{y})[f_{\mathbf{Y}}(\mathbf{y};\theta_1) - k_\alpha f_{\mathbf{Y}}(\mathbf{y};\theta_0)]\, d\mathbf{y} \\
&= \int_{\mathbb{R}^n} \phi_T(\mathbf{y})[f_{\mathbf{Y}}(\mathbf{y};\theta_1) - k_\alpha f_{\mathbf{Y}}(\theta_0)]\, d\mathbf{y} \\
&= \beta_T(\theta_1) - k_\alpha \beta_T(\theta_0).
\end{aligned}$$

But $\beta_S(\theta_0) = \beta_T(\theta_0) = \alpha$, so we have proved that $\beta_S(\theta_1) \leq \beta_T(\theta_1)$, as required. □

Example 10.3.7. Let $Y_1, \ldots, Y_n \sim \text{Bernoulli}(p)$, with $H_0 : p = p_0$ and $H_1 : p = p_1$. Then

$$L(p; \mathbf{y}) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{n\bar{y}}(1-p)^{n-n\bar{y}}.$$

Now utilize the Neyman-Pearson Lemma, and take the test statistic:

$$\begin{aligned} h(\mathbf{y}) &= \frac{L(p_1; \mathbf{y})}{L(p_0; \mathbf{y})} \\ &= \frac{p_1^{n\bar{y}}(1-p_1)^{n(1-\bar{y})}}{p_0^{n\bar{y}}(1-p_0)^{n(1-\bar{y}}} \\ &= \underbrace{\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^{n\bar{y}}}_{>1} \left(\frac{1-p_1}{1-p_0}\right)^{n}. \end{aligned}$$

◇

Suppose that $p_1 > p_0$. Then $h(\mathbf{y})$ is increasing in $\mathbf{y}$, or equivalently, in $n\bar{y} = \sum_{i=1}^{n} y_i$.

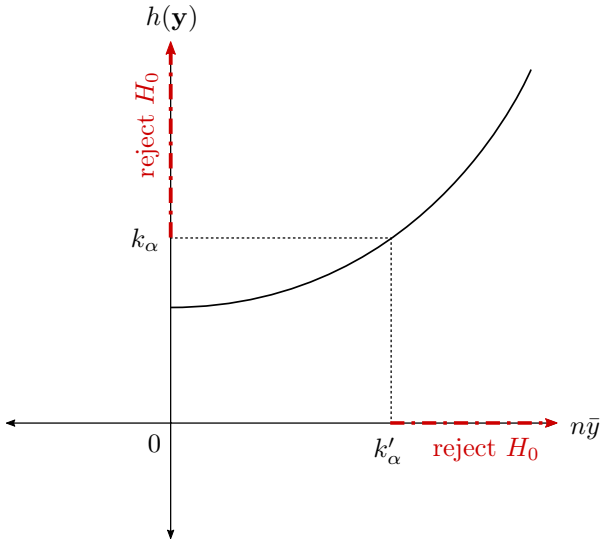


Figure 10.1: Intuition behind the critical region .

The fact that the function is monotonic increasing shows that $h(\mathbf{y}) > k_\alpha$ is equivalent to $\bar{y} > k'_\alpha$. Then

$$\text{Reject } H_0 \text{ if } h(\mathbf{Y}) > k_\alpha \iff \text{Reject } H_0 \text{ if } n\bar{Y} > k'_\alpha.$$

We want $P_{\theta_0}(n\bar{Y} > k'_\alpha) = \alpha$, which we can find because $\bar{Y} \sim^{H_0} \text{Bin}(n, p_0)$. Note that if $p_0 > p_1$, then $h(\mathbf{y})$ is a monotonic decreasing function, and a similar argument holds.

In general, if $h(\mathbf{Y}) = \frac{L(\theta_1;\mathbf{Y})}{L(\theta_0;\mathbf{Y})}$ is increasing in some sufficient statistic $T(\mathbf{Y})$, then:

$$\text{Reject } H_0 \text{ if } h(\mathbf{Y}) > k_\alpha \iff \text{Reject } H_0 \text{ if } T(\mathbf{Y}) > k'_\alpha.$$

If $h(\mathbf{Y})$ is decreasing in $T(\mathbf{Y})$, this becomes

$$\text{Reject } H_0 \text{ if } h(\mathbf{Y}) > k_\alpha \iff \text{Reject } H_0 \text{ if } T(\mathbf{Y}) < k'_\alpha.$$

10.4 Week 20: Lecture 2

10.4.1 Uniformly Most Powerful Tests

Tue 15 Mar 14:00

Example 10.4.1. Let $Y_1, \ldots, Y_n \sim \text{Normal}(\mu, 1)$ (i.e. variance is known) Let $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1$. We have

$$\begin{aligned} L(\mu;\mathbf{y}) &= \prod_{i=1}^{n} \frac{1}{2\pi} e^{\frac{-(y_i-\mu)^2}{2}} \\ &\propto e^{-\frac{1}{2}\sum(y_i^2 - 2\mu_i + \mu^2)} \\ &\propto e^{\mu n \bar{y}} e^{\frac{-n\mu^2}{2}}. \end{aligned}$$

Our test statistic is

$$\begin{aligned} h(\mathbf{y}) = \frac{L(\mu_1;\mathbf{y})}{L(\mu_0;\mathbf{y})} &= \frac{e^{\mu_1 n\bar{y}} e^{-n\frac{\mu_1^2}{2}}}{e^{\mu_0 n\bar{y}} e^{-n\frac{\mu_0^2}{2}}} \\ &= e^{\frac{n}{2}(\mu_0^2 - \mu_1^2)} e^{n(\mu_1-\mu_0)\bar{y}}. \end{aligned}$$

Then if $\mu_1 > \mu_0$, $h(\mathbf{y})$ is increasing in $\bar{y}$, and we reject H_0 when $\bar{Y} > k_\alpha$. To find k_α, we set: $P_{H_0}(\mathbf{Y} > k_\alpha) = \alpha$. If we change μ_1 to another value larger than μ_0, then the form of the test does not change, and the critical region doesn't change either, because k_α depends on μ_0, not μ_1. Any size calculation is performed under the null hypothesis, assuming that the null hypothesis is true. The power of the test assumes the alternative. Notice that this MPT is the same for any $\mu_1 > \mu_0$. This is an important property. ◇

Definition 10.4.2. Suppose we want to test $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$. If we take any $\theta \in \Theta_1$, and obtain the same MPT, then that MPT is the is the **Uniformly MPT (UMPT).**

Now, what if the alternative were two-sided? Is the MPT always the same? No, because for values greater than μ_0, we reject for $\bar{y} > k$, but for $\mu_1 < \mu_0$ we reject for $\bar{y} < k$. If you had a two sided alternative, you could not come up with

a UMPT because you do not obtain the same MPT for every special case. The form of the test changes for different parameter values.

Example 10.4.3. In these cases, a UMPT exists:

$$H_1 : \mu < \mu_0$$
$$H_1 : \mu > \mu_0.$$

But in this case, a UMPT does not exist:

$$H_1 : \mu \neq \mu_0.$$

Remark 10.4.4. Aside: how would you test $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$? We could use the Likelihood-ratio test.

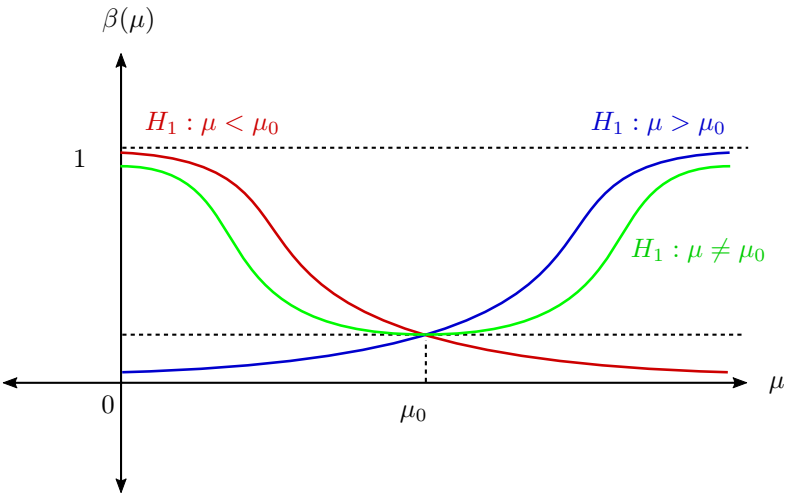


Figure 10.2: The alternative hypotheses $H_1 : \mu < \mu_0$ and $H_1 : \mu > \mu_0$ are more powerful than $H_1 : \mu \neq \mu_0$ at values less than and greater than μ_0, respectively.

◇

Remark 10.4.5. It's not always possible to find a UMPT for a one-sided test, but it is more often the case.

Note that the most powerful test is on the side of the alternative that you care about. It is not always the case that we can find a UMPT for a one-sided alternative but not for a two-sided case, but it *is* often true.

Example 10.4.6. Now, what if we had:

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0.$$

Firstly, if we were to test the simple null vs. the alternative, we would get a UMPT of size α. For $H_0^* : \mu = \mu_0 \quad \text{vs.} \quad H_1$, we have the UMPT which rejects H_0^* when $\bar{Y} > k_\alpha$, where

$$\begin{aligned}
\alpha &= P_{H_0}(\bar{Y} > k_\alpha) \\
&= P_{H_0}\left(\frac{\bar{Y} - \mu_0}{\sqrt{\frac{1}{n}}} > \frac{k_\alpha - \mu_0}{\sqrt{\frac{1}{n}}}\right) \\
&= P(Z > \sqrt{n}(k_\alpha - \mu_0)).
\end{aligned}$$

So $k_\alpha = \mu_0 + z_\alpha\sqrt{\frac{1}{n}}$. We use μ_0 because this is a size calculation. Note that under H_0, the first term is Normal$(0, 1)$ distributed. Moreover, Z is in the right tail of the standard normal.

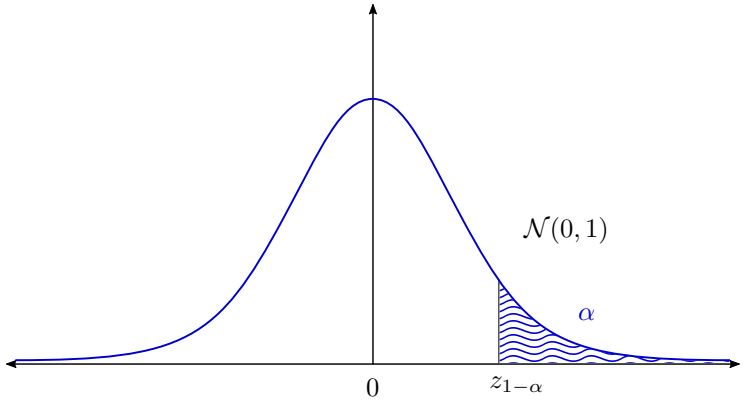


Figure 10.3: The rejection region of Z in Example 10.4.6.

Now what if $\mu = \mu_0'$, where $\mu_0' < \mu_0$? The probability of Type I error would be

$$\begin{aligned}
P_{\mu_0'}(\bar{Y} > k_\alpha) &= P_{\mu_0'}\left(\bar{Y} > \mu_0 + z_{1-\alpha}\sqrt{\frac{1}{n}}\right) \\
&= P_{\mu_0'}\left(\frac{\bar{Y} - \mu_0'}{\sqrt{\frac{1}{n}}} > \frac{\mu_0 - \mu_0' + z_{1-\alpha}\sqrt{\frac{1}{n}}}{\sqrt{\frac{1}{n}}}\right) \\
&= P(Z > z_{1-\alpha} + \sqrt{n}(\mu_0 - \mu_0')) \\
&< \alpha.
\end{aligned}$$

Note that the inequality holds because $z_{1-\alpha} + \sqrt{n}(\mu_0 - \mu_0') > z_{1-\alpha}$. The probability of type I error for any $\mu_0' < \mu_0$ is less than α. The worst case scenario for this test is therefore α. In fact, α is the *supremum* of the size of the test. Hence, the UMPT has

$$\sup_{\mu \leq \mu_0} P_\mu(\text{reject } H_0) = \alpha.$$

So, it has size α. ◇

Recall. In general, the size of a test is the supremum of the type I error.

10.4.2 LRT and nuisance parameters

Example 10.4.7. Let $X_1 \dots X_n \sim \text{Poisson}(\lambda)$ and $Y_1, \dots Y_n \sim \text{Poisson}(\mu)$, with

$$H_0 : \lambda = \mu,$$
$$H_1 : \lambda \neq \mu.$$

We have an alternative parameterisation: $\mu = \lambda + \psi$. Our parameter of interest is ψ, so the test becomes

$$H_0 : \psi = 0,$$
$$H_1 : \psi \neq 0.$$

Now, λ is a *nuisance* parameter. It is an added layer of complexity that we have to work around, but is not the parameter of interest. When we set up LRTs, we are expressing the hypotheses in terms of the parameter of interest.

Remark 10.4.8. If we had a third sample, then we would have two parameters of interest. The advantage of the above formulation is that it is easy to write it for an arbitrary number of samples.

The test statistic for LRT:

$$\begin{aligned} h(\mathbf{Y}) &= \frac{\sup_\theta(\theta; \mathbf{Y})}{\sup_{\theta \in \Theta_0} L(\theta; \mathbf{Y})} \\ &= \frac{L(\hat{\theta}; \mathbf{Y})}{L(\hat{\theta_0}; \mathbf{Y})}. \end{aligned}$$

If $\theta = \binom{\psi}{\delta}$, then $\hat{\theta} = \binom{\hat{\psi}}{\hat{\lambda}}$and $\hat{\theta}_0 = \binom{0}{\hat{\lambda}_0}$. Under both the null and the alternative, you have to estimate the nuisance parameters. For a null hypothesis of $\psi_1 = \psi_2 = \dots = \psi_k = 0$, however, we only have to estimate the distribution of the parameter that each is equal to, in this case λ. The unconstrained model therefore requires that we estimate $k - 1$ additional parameters. Moreover, the asymptotic distribution for $2\log(h(\mathbf{Y}))$ is χ^2_d, where d is the dimension of $\boldsymbol{\psi}$. ◇

Remark 10.4.9. For the LHR test statistic, we assume that we can solve both for the null and the alternative. There are situations where this is not feasible. There exist alternatives to the LHR in the score and Wald test.

Chapter 6

Statistical Models

6.1 Week 21: Lecture 1

Tue 22 Mar 14:00

Let X, Y be variables, where Y is the response variable (dependent variable, outcome), and X is the explanatory variable (independent variable, covariate, treatment). We have data

$$\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}.$$

A simple linear model for this can be written as

$$Y_i = \alpha + \beta X_i + \varepsilon_i.$$

for $i = 1, 2, \ldots n$. What can we say about the noise/error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$?

(i) $\mathbb{E}(\varepsilon_i) = 0$,

(ii) $\mathrm{Var}(\varepsilon_i) = \sigma^2$,

(iii) The sequence $\varepsilon_1, \ldots, \varepsilon_n$ are IID.

Every statistical model is going to have a signal (randomness we can quantify), and noise (randomness that is irreducible). We can write

$$Y_i \mid X_i = x_i = \alpha + \beta x_i + \varepsilon_i.$$

So $\mathbb{E}(Y_i \mid X_i = \alpha + \beta \mathbb{E}(x_i)$, and

$$\begin{aligned} \mathrm{Var}(Y_i \mid X_i = x_i) &= \mathbb{E}[(\alpha + \beta x_i + \varepsilon_i - \alpha - \beta x_i)^2] \\ &= \mathbb{E}[(\varepsilon_i - \mathbb{E}(\varepsilon))^2] \\ &= \sigma^2. \end{aligned}$$

The simple linear model has 3 parameters that we need to fully quantify the model: α; β; and σ^2, the variance of the error terms.

Remark 6.1.1. In M&P, we use $Y_i = \alpha + \beta X_i + \sigma\varepsilon_i$, where $\mathbb{E}(\varepsilon_i) = 0$, and $\mathrm{Var}(\varepsilon_i) = 1$.

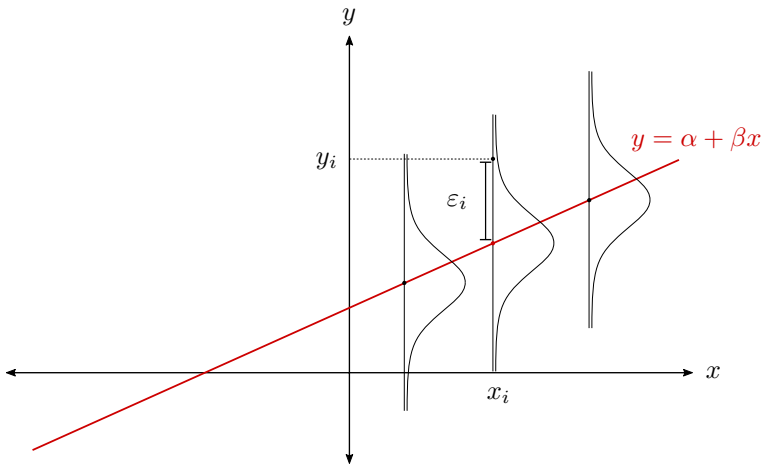


Figure 6.1: We can imagine a density curve at each point on the fitted model that determines y_i. The distance from the mean is the error ε_i.

In practice, however, we don't have data that is produced based on a line. We instead have a scatter plot:

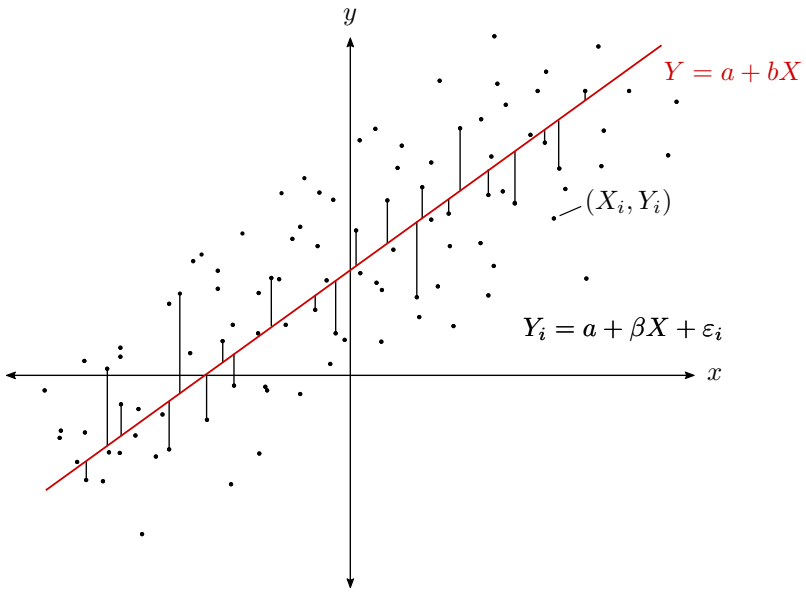


Figure 6.2: A simple fitted linear model. The dispersed vertical lines represent the size of the error.

So, what exactly are the optimal values of a, b? By what measure are some

lines a better fit than other lines, and what is a line of best fit? We can define our linear model as $\hat{Y}_i = a + bX_i$, so take $Y_i - \hat{Y}_i$ (the residual). We minimize the residuals by minimizing

$$S(a,b) = \sum_{n=1}^{n} (Y_i - \hat{Y}_i)^2,$$

the sum of squares of the residuals.

Remark 6.1.2. The advantage of taking the square of the residuals vs. the absolute value is that $(Y_i - \hat{Y}_i)^2$ is differentiable, while $|Y_i - \hat{Y}_i|$ is not. Taking absolute values in an L_1 regression *is* more robust, and allows us to better quantify the residuals, but it is not mathematically convenient.

We then have

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{(a,b)} S(a,b).$$

Definition 6.1.3. This approach is known as **OLS**, or **(ordinary least squares).**

6.1.1 Multiple Linear Regression

Suppose we have p explanatory variables:

$$\{(X_{1,1}, X_{1,2}, \ldots, X_{1,p}, Y_1), \ldots, (X_{n,1}, \ldots, X_{n,p}, Y_n)\},$$

where $X_{i,j}$ is the ith observation, and the jth variable. We have

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Notation. We can write this in matrix notation as

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{\substack{\mathbf{Y} \\ (n\times 1)}} = \underbrace{\begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{pmatrix}}_{\substack{\mathbf{X} \\ (n\times(p+1))}} \underbrace{\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}}_{\substack{\boldsymbol{\beta} \\ ((p+1)\times 1)}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\substack{\boldsymbol{\varepsilon} \\ (n\times 1)}}.$$

Or more concisely as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the same model in matrix notation. Moreover,

$$\mathrm{Var}(\boldsymbol{\varepsilon}) = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}_n.$$

If we take $x_i^T = \begin{pmatrix} 1 & x_{i,1} & \cdots & x_{i,p} \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$, we can write

$$\hat{Y}_i(\mathbf{b}) = b_0 + b_1 + x_{i,1} + \ldots + b_p x_{i,p} = \mathbf{x}_i^T \mathbf{b},$$

and so the sum of squares is:

$$\begin{aligned} S(\mathbf{b}) &= \sum_{n=1}^{n} (Y_i - \hat{Y}_i) \\ &= \sum_{n=1}^{n} (Y_i - \mathbf{x}_i^T \mathbf{b})^2 \\ &= (\mathbf{Y} - \mathbf{X}\mathbf{b})^T (\mathbf{Y} - \mathbf{X}\mathbf{b}). \end{aligned}$$

and the least-squares estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b}} S(\mathbf{b})$.

Remark 6.1.4. If $\mathbf{A}$ is an $r \times c$ constant matrix and $\mathbf{w}$ is a $c \times 1$ vector, then:

$$\frac{\partial}{\partial \mathbf{A}\mathbf{w}} \mathbf{w} = \mathbf{A}, \quad \frac{\partial \mathbf{w}^T \mathbf{A}}{\partial \mathbf{w}} = \mathbf{A}^T, \quad \frac{\partial \mathbf{w}^T \mathbf{A} \mathbf{w}}{\partial \mathbf{w}} = \mathbf{w}^T (\mathbf{A} + \mathbf{A}^T).$$

So:

$$\begin{aligned} \frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} &= \frac{\partial}{\partial \mathbf{b}} \left[(\mathbf{Y} - \mathbf{X}\mathbf{b})^T (\mathbf{Y} - \mathbf{X}\mathbf{b}) \right] \\ &= \frac{\partial}{\partial \mathbf{b}} \left[\mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{X}\mathbf{b} - \mathbf{b}^T \mathbf{X}^T \mathbf{Y} + \mathbf{b}^T \mathbf{X}^T \mathbf{X}\mathbf{b} \right] \\ &= 0 - \mathbf{Y}^T \mathbf{X} - \mathbf{Y}^T \mathbf{X} + \mathbf{b}^T (\mathbf{X}^T \mathbf{X} + \mathbf{X}^T \mathbf{X}) \\ &= -2\mathbf{Y}^T \mathbf{X} + 2\mathbf{b}^T \mathbf{X}^T \mathbf{X}. \end{aligned}$$

We set this final expression to 0, which implies

$$\begin{aligned} &\mathbf{b}^T \mathbf{X}^T \mathbf{X} = \mathbf{Y}^T \mathbf{X} \\ &\Rightarrow \mathbf{X}^T \mathbf{X}\mathbf{b} = \mathbf{X}^T \mathbf{Y} \\ &\Rightarrow \mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \end{aligned}$$

We found $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$. This is the solution for an arbitrary number of random variables. So:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \underbrace{\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T}_{\text{the "hat" matrix}} \mathbf{Y} = \mathbf{H}\mathbf{Y}.$$

Definition 6.1.5. We can say that $\hat{\boldsymbol{\beta}}$ is a linear estimator of $\mathbf{Y}$.

Note. This proof is not examinable.

6.2 Week 21: Lecture 2

6.2.1 The Hat and the Annihilator

Thu 24 Mar 2022

Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ be a linear model. Recall that $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$, and $\text{Var}(\boldsymbol{\varepsilon}) - \sigma^2 \boldsymbol{I}_n$. Note that

$$S(b) = (\mathbf{Y} - \hat{\mathbf{Y}})^T(\mathbf{Y} - \hat{\mathbf{Y}}) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Then $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ minimizes $S(b)$, and

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} = \underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\mathbf{H}\text{, the hat matrix.}}\mathbf{Y} = \mathbf{H}\mathbf{Y}.$$

The residuals are

$$\begin{aligned}
\hat{\boldsymbol{\varepsilon}} &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{Y} - \hat{\mathbf{Y}} \\
&= \mathbf{Y} - \mathbf{H}\mathbf{Y} \\
&= \underbrace{(\mathbf{I}_n - \mathbf{H})}_{\mathbf{A}\text{, the annihilator matrix}} \times \mathbf{Y} \\
&= \mathbf{A}\mathbf{Y}.
\end{aligned}$$

Note that $\mathbf{H}$ and $\mathbf{A}$ are symmetric and idempotent.

Definition 6.2.1. An **idempotent** matrix $\mathbf{A}$ satisfies the property $\mathbf{A} = \mathbf{A}^2$.

6.2.2 Linear Models I

$$\mathbf{1}_n = \underset{(n\times 1)}{\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1, \end{pmatrix}}, \quad \mathbf{e}_1 = \underset{(p+1)\times 1}{\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}.$$

Moreover,

$$\mathbf{X} = \underset{n\times(p+1)}{\begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots X_{n,p} & \end{pmatrix}}.$$

Then

$$\mathbf{1}^T\mathbf{Y} = \sum_{i=1}^{n} Y_i, \quad \mathbf{X}\mathbf{e}_1 = 1.$$

Moreover,

$$\begin{aligned}\mathbf{1}^T\hat{\mathbf{Y}} &= (\mathbf{X}\mathbf{e}_1)^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{e}_1^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{1}^T\mathbf{Y}.\end{aligned}$$

Or,

$$\sum_{i=1}^{n}\hat{Y}_i = \sum_{i=1}^{n} Y_i \iff \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right) = 0,$$

i.e., the residuals sum to zero. Moreover,

$$\begin{aligned}\frac{1}{n}\mathbf{1}^T\mathbf{X} &= \frac{1}{n}\left(n\sum_{i=1}^{n} X_{i,1} \cdots \sum_{i=1}^{n} X_{i,p}\right) \\ &= (1 \times \bar{X}_{\cdot 1} \ldots \bar{X}_{\cdot p}) \\ &= \bar{\mathbf{X}}^T.\end{aligned}$$

Then

$$\begin{aligned}\hat{\mathbf{Y}} &= \bar{\mathbf{X}}\hat{\boldsymbol{\beta}} \\ &= \frac{1}{n}\mathbf{1}^T\bar{\mathbf{X}}\hat{\boldsymbol{\beta}} \\ &= \frac{1}{n}\mathbf{1}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \frac{1}{n}\mathbf{1}^T\hat{\mathbf{Y}} \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i \\ &= \bar{Y}.\end{aligned}$$

This shows that the hyperplane passes through the vector $\left(\bar{\mathbf{X}}, \mathbf{Y}\right)$.

Properties

Some properties:

1. $\mathbb{E}(\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ (unbiased).
2. $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$.

3. $\mathbb{E}(\hat{\boldsymbol{\varepsilon}}) = 0$.

To estimate σ^2, take

$$\sum_{i=1}^{n} \hat{\varepsilon_i^2} = \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} = S(\hat{\boldsymbol{\beta}}).$$

This has $n - p - 1$ degrees of freedom, so

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

Example 6.2.2. If $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \text{Normal}(0, \sigma^2 \mathbf{I}_n)$, then $\mathbf{Y} \sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$. Also:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \quad \text{is Normal.}$$

because for a multivariate normal distribution, any linear transformation of the variable is also normal. Moreover,

$$\mathbf{Y} = \hat{\mathbf{HY}} \quad \text{is Normal.}$$

and

$$\hat{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbb{H})\mathbf{Y}) \quad \text{is Normal.}$$

and

$$\hat{\boldsymbol{\varepsilon}} = (I_n - \mathbf{H})\mathbf{Y} \quad \text{is Normal.}$$

◇

Now, what if Y_i is Bernoulli distributed?

Definition 6.2.3. The **mean function** is

$$\mu(\mathbf{x}_i) = \mathbb{E}(Y_i \mid \mathbf{X}_i = \mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Definition 6.2.4. Let $g(\mu(\mathbf{x}_i)) = \mathbf{x}_i^T\boldsymbol{\beta}$, where g is the **link function**. Then $\mu(\mathbf{x}_i) = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$.

Note. The link function is the inverse of the mean function.

Example 6.2.5 (Logistic Regression)**.** Let $Y_i \mid \mathbf{X}_i = \mathbf{x}_i \sim \text{Bernoulli}(p(\mathbf{x}_i))$, where

$$p(\mathbf{x}_i) = \mu(\mathbf{x}_i) = \mathbb{E}(Y_i \mid \mathbf{X}_i = \mathbf{x}_i).$$

◇

Note that

$$\underbrace{\log\left(\frac{p_i}{1-p_i}\right)}_{\in\mathbb{R}} = \mathbf{x}_i^T\boldsymbol{\beta},$$

so

$$\mu(\mathbf{x}_i^T\boldsymbol{\beta}) = \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}.$$

6.3 Week 22: Lecture 1

6.3.1 Linear Models II

Tue 29 Mar 14:00

Consider the following situation. We have (X_i, Y_i), with $i = 1, \ldots, n$. Moreover,

$$X_i = \begin{cases} 0, & \text{if } Y_i \text{ is from population A,} \\ 1, & \text{if } Y_i \text{ is from population B.} \end{cases}.$$

Now, we have a simple linear regression model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

Then

$$\mathbb{E}(Y_i \mid X_i = 0) = \alpha,$$
$$\mathbb{E}(Y_i \mid X_i = 1) = \alpha + \beta.$$

Note that both of the populations still have the same variance σ^2. The parameter β determines whether these populations are the same or different.

What if we have populations $0, 1, 2, \ldots, k-1$? Observe:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_{i,k-1} + \varepsilon_i.$$

and

$$X_{i,j} = \begin{cases} 0, & \text{if } Y_i \text{ is not from population } j \\ 1, & \text{if } Y_i \text{ is from population } j. \end{cases}$$

If we define

$$\mu_j = \mathbb{E}(Y_i \mid X_{i,j} = 1) = \beta_0 + \beta_j, \quad \text{for } j = 1, \ldots, k-1$$

then $\mu_0 = \beta_0$, the baseline. Note that β_j tells us how much the j^{th} group differs from the baseline.

Remark 6.3.1. When you treat something as a categorical variable, the different values it can take are now categories. And if you have a categorical value that can take k different values you can represent this using k populations and the $k-1$ indicator variables. If you attempt to fit this in your statistical analysis software of your choice, you will typically get coefficients for $k-1$ of the different groups.

Definition 6.3.2. We can use this model to test whether the groups are different from each other. This application is known as **1-way ANOVA** (**ANalysis Of VAriance**).

In ANOVA, the key quantities we are comparing are

- $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$: the total sum of squares (or total variability in data)
- $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$: the residual sum of squares (the within-group variability),

where $\hat{Y}_i = \hat{\mu}_j$ if Y_i comes from population j.

Remark 6.3.3. In ANOVA, we compare how much smaller the residual sum of squares is than the total sum of squares. If we find the RSS is very small, we reduce the within-group variability a lot, and hence within-group variability is smaller than between group variability. Thus, there would be evidence that this group alignment is statistically significant. If there isn't a big difference, we would conclude that the groups aren't very different. This is known as the *F-test*.

6.3.2 Logistic Regression

In a logistic regression model, $Y_i \mid \mathbf{X}_i = x_i \sim \text{Bernoulli}(p_i)$, where

$$p_i = \mu(\mathbf{x}_i) = \mathbb{E}(Y_i \mid \mathbf{X}_i = \mathbf{x}_i).$$

Definition 6.3.4. The standard choice of link function for a logistic regression is the **logit** function, or the **log-odds**. The logit function is defined as

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T\beta \iff p_i = \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}.$$

Remark 6.3.5. Least squares does not work here because, for the linear model, we assumed that $\text{Var}(\varepsilon_i) = \text{Var}(\varepsilon_j)$ for all $i, j \in \{1, \ldots, n\}$. With a logistic regression, this constant-variance assumption breaks down. If Y_i is Bernoulli distributed with probability p_i, the different Y_i values have different variance. We can however, fit them using maximum-likelihood:

$$\begin{aligned} f(y_i \mid x_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \left(\frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{y_i}\left(1-\frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{1-y_i}. \end{aligned}$$

Two choices of link functions for a logistic regression are

- logit: $\mathbf{x}_i^T\beta = \log\left(\frac{p_i}{1-p_i}\right)$

- probit: $\mathbf{x}_i^T \boldsymbol{\beta} = \Phi^{-1}(p_i)$,

where Φ is the CDF of the Standard Normal. Note that $p_i = \frac{1}{2}$ implies

$$\log\left(\frac{p_i}{1-p_i} = 0\right)$$
$$\Phi^{-1}(p_i) = 0.$$

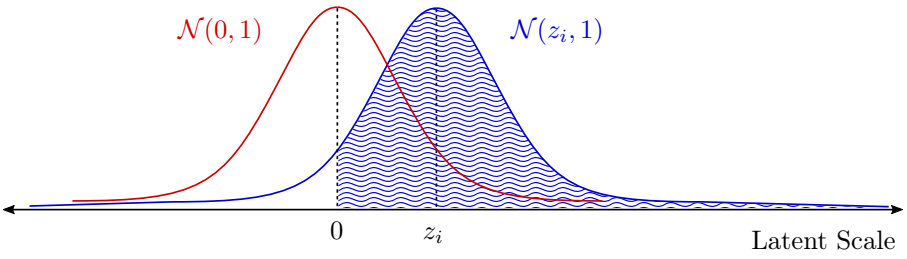


Figure 6.3: We can visualize the probit method through a latent scale depending on z_i. If individual i falls to the right of the standard normal, our model predicts that they default. The probability that an individual defaults is Normal$(z_i, 1)$ distributed, with this example showing a positive value of z_i.

Let $z_i = \mathbf{x}_i^T \boldsymbol{\beta}$. Then

$$P(Y_i = 1 \mid \mathbf{x}_i = \mathbf{x}_i) = p_i = \Phi(\mathbf{x}_i^T \boldsymbol{\beta}) = \Phi(z_i).$$

The probability of default is the shaded area from 0 to ∞.

Definition 6.3.6. A **latent scale** is a scale derived from unobserved variables that are inferred from a mathematical model.

We do not observe whether a person is a safer investment, i.e., where they are on the latent scale. We *do* observe whether they default on a loan or not. Every individual is drawn on a normal distribution centered at a mean somewhere on a latent scale.

Note. Lecture 2 was a revision lecture, and covered no new content.

Conclusion

Any issues with the lecture notes can be reported on the git repository, by either submitting a pull request or an issue. I am happy to fix any typos or inaccuracies with the content. In addition, feel free to edit my work, just keep my name on it if you're going to publish it somewhere else. The figures can be edited with Inkscape, the software I used to create them. When editing the figures, make sure to save to pdf, and choose the option that exports the text directly to LaTeX. I hope these notes helped!