ST202: Probability, Distribution Theory, and 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.

Part I

Michaelmas Term

Chapter 2

Probability

2.1 Week 1: Lecture 1

2.1.1 A Pair of Dice

Tue 28 Sep 14:00

Example 2.1.1. Roll two dice. The probability sum is > 10. There are three favourable outcomes:

$$(5,6),(6,5),(6,6).$$

There are 36 total outcomes. Then the probability is $\frac{3}{36} = \frac{1}{12}$. ◇

Definition 2.1.2. The **sample space** Ω is the collection of every possible outcome. An **outcome** ω is an element of the sample space ($\omega \in \Omega$).

Definition 2.1.3. An **event** A is a set of possible outcomes in Ω ($A \subseteq \Omega$).

2.1.2 [a bit of] Measure Theory

Let ψ be a set and $\mathcal{G}$ be a collection of subsets of ψ. Note that if $A \in \mathcal{G}$, then $A \subseteq \psi$.

Definition 2.1.4. A **measure** is a function $m : \mathcal{G} \to R^+$ such that

i. $m(A) \geq 0$ for all $A \in \mathcal{G}$,

ii. $m(\varnothing) = 0$,

iii. if $A_1, A_2, \ldots \in \mathcal{G}$ are disjoint, then $m\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} m(A_i)$.

Definition 2.1.5. A set $\mathcal{G}$ is a σ-algebra on ψ if

i. $\varnothing \in \mathcal{G}$,

ii. if $A \in \mathcal{G}$ then $A^c \in \mathcal{G}$,

iii. if $A_1, A_2, A_3, \ldots \in \mathcal{G}$ then

$$\bigcup_{i=1}^{\infty} A_i = A_1 \cup A_2 \cup A_3 \cup \ldots \in \mathcal{G}.$$

Definition 2.1.6. Let ψ be a set, $\mathcal{G}$ a σ-algebra of ψ, and m a measure of $\mathcal{G}$. The space $(\psi, \mathcal{G})$ is a **measurable space.** The space $(\psi, \mathcal{G}, m)$ is a **measure space.**

Example 2.1.7. Let ψ be a set. The set $\{\varnothing, \psi\}$ is the smallest σ-algebra of ψ.

Suppose that $|\psi| > 1$. Let $A \subset \psi$. Then $\{\varnothing, A, A^c, \psi\}$ is the smallest non-trivial σ-algebra.

◇

Example 2.1.8. The σ-algebra $\mathcal{G} = \{A : A \subseteq \psi\} = \mathcal{P}(\psi)$ is the power set of ψ. Hence, if

$$\psi = \{\omega_1, \omega_2, \ldots, \omega_k\},$$

then $|\mathcal{G}| = 2^{|\psi|} = 2^k$. ◇

Example 2.1.9. Is $m(A) = |A|$, i.e., the number of elements of A, a well-defined measure?

i.) $m(A) \geq 0$? ✓

ii.) $m(\varnothing) = 0$? ✓

iii.) if $A_1, A_2, \ldots$ are disjoint,

$$m\left(\bigcup_{i=1}^{\infty} A_i\right) = \left|\bigcup_{i=1}^{\infty} A_i\right| = \sum_{i=1}^{\infty} |A_i| = \sum_{i=1}^{\infty} m\left(A_i\right). \quad \checkmark$$

◇

2.2 Week 1: Lecture 2

Wed 29 Sep 10:00

Let $(\psi, \mathcal{G})$ be a measurable space, with $\omega \in \psi$ and $A \in \mathcal{G}$. Consider

$$m(A) = \mathbf{1}_A(\omega) = \begin{cases} 1, \text{ if } w \in A \\ 0, \text{ if } w \notin A. \end{cases}$$

Exercise. Check that this is a measure! (on the problem set).

2.2.1 The Probability Measure

Definition 2.2.1. Consider the measurable space $(\Omega, \mathcal{F})$. Define $(\Omega, \mathcal{F}, P)$ as a **probability space.** The function P is a **probability measure** that satisfies $P(A) \in [0, 1]$ for all $A \in \mathcal{F}$ and $P(\Omega) = 1$.

Since P is a measure,

- $P(A) \geq 0$ for all $A \in \mathcal{F}$,
- $P(\varnothing) = 0$,
- $P(A \cup B) = P(A) + P(B)$ *if* $A \cap B = \varnothing$ *(mutually exclusive).*

In general, if $A, B \in \mathcal{F}$ do we have $A \cap B \in \mathcal{F}$? Observe that

$$(A \cap B)^c = A^c \cup B^c.$$

So yes! We do.

In general, if $A_1, A_2, A_3, \cdots \subseteq \Omega$ are mutually exclusive, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Basic Properties of Probability Measures

i. $P(A^c) = 1 - P(A)$

ii. If $A \subseteq B$, then $P(B \setminus A) = P(B) - P(A)$

iii. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Example proof (the remaining proofs are left as an exercise):

Proof. i. A, A^c are disjoint, and thus $A \cup A^c = \Omega$, but $P(A \cup A^c) = P(A) + P(A^c)$.

□

Corollary 2.2.2. If $A \subseteq B$, then $P(A) \leq P(B)$.

General Addition Rule:

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) = &\sum_{i=1}^{n} P(A_i) - \sum_{i,j=1,\ i<j}^{n} P(A_i \cap A_j) \\ &+ \sum_{i<j<k} P(A_i \cap A_j \cap A_k) \\ &- \dots \\ &+ (-1)^{n+1} P(A_1 \cap A_2 \cap \dots \cap A_n). \end{aligned}$$

2.2.2 More Properties of Probability Measures

Theorem 2.2.3 (Boole's Inequality)**.** If $(\Omega, \mathcal{F}, P)$ is a probability space and $A_1, A_2, A_3, \dots \in \mathcal{F}$, then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i).$$

Proof. Define

$$\begin{aligned} B_1 &= A_1 \\ B_2 &= A_2 \setminus B_1 \\ B_3 &= A_3 \setminus (B_1 \cup B_2) \\ &\vdots \\ B_i &= A_i \setminus (B_1 \cup \dots \cup B_{i-1}). \end{aligned}$$

Then $B_1, B_2, B_3, \dots \in \mathcal{F}$ (confirm this!) They are *disjoint,* and $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$. So,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i) \leq \sum_{i=1}^{\infty} P(A_i).$$

□

Proposition 2.2.4. If $A_1, A_2, A_3, \dots$ is an increasing sequence of sets $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$, then $\lim_{n\to\infty} P(A_n) = P\left(\bigcup_{i=1}^{\infty} A_i\right)$.

The following figure may come in handy:

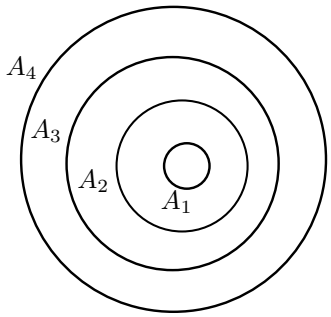


Figure 2.1: Visual representation of $A_1, A_2, \ldots$

Proof. Define

$$\begin{aligned} B_1 &= A_1 \\ B_2 &= A_2 \setminus A_1 \\ &\vdots \\ B_i &= A_i \setminus A_{i-1} \\ &\vdots \end{aligned}$$

Note that these events are mutually exclusive, and so $A_n = \bigcup_{i=1}^{n} B_i$. Moreover, $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$. Hence,

$$\begin{aligned} \lim_{n\to\infty} P(A_n) &= \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} B_i\right) \\ &= \lim \sum_{i=1}^{n} P(B_i) \\ &= P\left(\bigcup_{i=1}^{\infty} B_i\right) \\ &= P\left(\bigcup_{i=1}^{\infty} A_i\right). \end{aligned}$$

□

We will use $P(A) = \frac{|A|}{|\Omega|}$, for $A \in \mathcal{F}$, with assumptions that the each event is equally likely, and that the sample space is finite.

2.2.3 Sample Problems

1) **Lottery**: choose 6 numbers from $\{1, 2, \ldots, 59\}$. What is the probability of matching 6 numbers?

2) **Birthdays**: 100 people in this lecture. What is the probability that at least two share a birthday?

Note. Read how the multiplication rule applies to permutations and combinations.

2.3 Week 2: Lecture 1

2.3.1 Discrete Tools

Tue 4 Oct 14:00

Let A be an event in a σ-algebra $\mathcal{F}$, and let $P(A) = \frac{|A|}{|\Omega|}$ be a probability measure. Note that if we can break the experiment we are interested in into k subexperiments $\Omega_i \subseteq \Omega$, then the multiplication rule dictates

$$|\Omega| = |\Omega_1| \times |\Omega_2| \times \cdots |\Omega_k|.$$

Permutations

Definition 2.3.1. Take n distinct objects, and choose k of them to be put *in a specific order*. A **permutation** refers to one such ordering.

We can find the number of possible permutations of size k using the multiplication rule:

$$\begin{aligned}\underbrace{n}_{\text{1st choice}} \times \underbrace{(n-1)}_{\text{2nd}} \times \cdots \times \underbrace{(n-k+1)}_{k\text{th}} &= \frac{n(n-1)\cdots 1}{(n-k)(n-k-1)\cdots 1} \\ &= \frac{n!}{(n-k)!} \\ &= {}^nP_k.\end{aligned}$$

Combinations

Definition 2.3.2. Take n distinct objects, and choose k of them, but do *not* put them in order. A **combination** refers to one such group. The number of combinations of size k is represented as ${}^nC_k$

Note that a permutation can be represented as

$$\underbrace{\begin{pmatrix}\text{choose } k \text{ objects}\\ \text{out of } n\end{pmatrix}}_{{}^nC_k} \times \underbrace{\begin{pmatrix}\text{put these } k\\ \text{objects in order}\end{pmatrix}}_{k!}.$$

Hence,

$${}^nP_k = {}^n C_k \times k! \Rightarrow {}^n C_k = {}^n P_k = \frac{n!}{(n-k)!k!}.$$

Notation. We also denote combinations by $\binom{n}{k}$, also referred to as the **binomial coefficient.**

In general,

$$(a+b)^n = \sum^n \binom{n}{j} a^j b^{n-j}.$$

But why?

Remark 2.3.3. Take n objects, k of type I, and $n-k$ of type II. Put these n objects in order. How many possible ways of ordering them? We have $\binom{n}{k}$. Again, why? Think of it this way. Suppose there are n slots. We can put an object of type I or II in each slot. The order doesn't matter, so there are $\binom{n}{k}$ ways to choose k slots.

But how does this relate to the binomial coefficient? First note that

$$(a+b)^n = \underbrace{(a+b)(a+b)(a+b)\cdots(a+b)}_{n \text{ times}}$$

Now consider the form of each term of the polynomial:

$$a^j b^{n-j}$$

Each term can be thought of as one combination of slots. We "choose" a or b for each part of the product, and multiply them together to get a term of the form $a^k b^{n-k}$. By the above, $\binom{n}{k}$ is how many terms are constructed for each k.

Example 2.3.4. Let $\{1, 2, \ldots, 59\}$ be a set of numbers. Choose 6 without replacement. What is the probability that I match all 6 if I draw at random? Two ways to solve this:

1. Order Matters: suppose that we consider every permutation of drawings. Then
$$|\Omega| = {}^{59}P_6 = \frac{59!}{53!}.$$
Now let A be the event that we match all 6 numbers, regardless of order. Then $|A| = 6!$, and
$$P(A) = \frac{|A|}{|\Omega|} = \frac{6!\,53!}{59!}.$$

2. Order Doesn't Matter: suppose that we consider ever *combination* of drawings. Then
$$|\Omega| = {}^{59}C_6 = \frac{59!}{53!\,6!}.$$

Let A be the event that we match all 6 numbers. Then $|A| = 1$, because there is only one combination that fits the criteria of A. Then
$$P(A) = \frac{1}{{}^{59}C_6} = \frac{6!\,53!}{59!},$$
the same answer as before. ◇

2.3.2 Conditional Probability

Let $(\Omega,]\mathcal{F}, P)$ be a probability space. Let $B \in \mathcal{F}$ with $P(B) > 0$. Define a new probability measure P_B such that $A \in \mathcal{F}$, and
$$P_B(A) = P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Note that if $P(A) = \frac{|A|}{|\Omega|}$, then
$$P(A \mid B) = \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} = \frac{|A \cap B|}{|B|}.$$

So,
$$P(A^c \mid B) = 1 - P(A \mid B).$$

If $A_1, A_2, \cdots \in \mathcal{F}$, and $P(A_i) > 0$ for $P(A_i) > 0$ for $i = 1, 2, \ldots$, then

$$P(A_n \cap \cdots \cap A_1) = P(A_n \mid A_{n-1} \cap \cdots \cap A_1)P(A_{n-1} \mid A_{n-2} \cap \cdots A_1) \cdots P(A_1)$$

2.3.3 Bayes' Rule

Let $\mathcal{F}$ be a σ-algebra. For two events $A, B \in \mathcal{F}$,

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B \mid A)P(A)}{P(B)} \\ &= P(A)\frac{P(B \mid A)}{P(B)}. \end{aligned}$$

We refer to $P(A)$ as the **prior** and $P(B \mid A)$ as the Bayes factor.

2.3.4 The Law of Total Probability

Definition 2.3.5. A partition of Ω is a collection of events $\{B_1, B_2, \dots\}$ such that

i. $P(B_i) > 0$ for all i,

ii. $\bigcup_{i=1}^{\infty} = \Omega$ (collectively exhaustive),

iii. $B_i \cap B_j = \varnothing$ for all $i \neq j$ (pairwise mutually exclusive).

Theorem 2.3.6 (Law of Total Probability)**.** Let the set $\{B_1, B_2, \dots\}$ be a partition of Ω. Then for any $A \in \mathcal{F}$,

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) = \sum_{i=1}^{\infty} P(A \mid B_i) P(B_i).$$

2.4 Week 2: Lecture 2

Wed 6 Oct 10:00

Example 2.4.1. Suppose that 1.2% of live births lead to twins. Further suppose that $\frac{1}{3}$ are identical twins, and $\frac{2}{3}$ are fraternal. We can describe each of these events with the outcomes and their associated probabilities below:

$$\begin{aligned} &\frac{1}{3} \text{ identical} \qquad (\underset{\frac{1}{2}}{BB}, \underset{\frac{1}{2}}{GG}) \\ &\frac{2}{3} \text{ fraternal} \qquad (\underset{\frac{1}{4}}{BB}, \underset{\frac{1}{4}}{GG}, \underset{\frac{1}{4}}{BG}, \underset{\frac{1}{4}}{GB}). \end{aligned}$$

Define the events T, I, F, M as T : twins, I : identical twins, F : fraternal, M : twin boys. We now work out each of their associated probabilities.

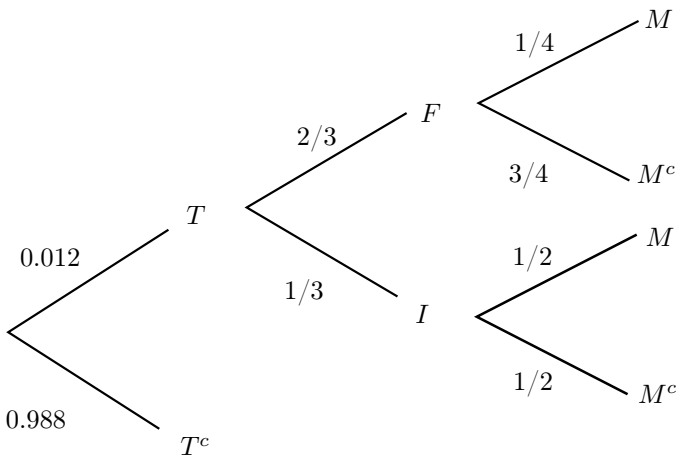


Figure 2.2: A probability tree representing this situation.

By multiplying along the paths of each event, we can obtain the probabilities of the events I, F, M, and $P(F \mid M)$.

$$\begin{aligned} P(I) &= P(I \mid T)P(T) = \frac{1}{3} \times 0.012 = 0.004 \\ P(F) &= P(F \mid T)P(T) = \frac{2}{3} \times 0.012 = 0.008 \\ P(M) &= \frac{1}{4} \times \frac{2}{3} \times 0.012 + \frac{1}{2} \times \frac{1}{3} \times 0.012 = 0.004 \\ P(F \mid M) &= \frac{P(M \mid F)P(F)}{P\,(M)} = \frac{\frac{1}{4} \times 0.008}{0.004} = \frac{1}{2}. \end{aligned}$$

◇

2.4.1 Independence

Let $A, B \in \Omega$. If A and B are independent, then

$$P(A|B) = P(A) \Rightarrow \frac{P(A \cap B)}{P(B)} = P(A),$$

which, in turn implies our definition of independence:

Definition 2.4.2. If A and B are **independent**, or $A \perp\!\!\!\perp B$, then $P(A \cap B) = P(A)P(B)$.

What if $B = \varnothing$? Then $P(B) = 0$, but also $P(A \cap B) = 0$. The definition also implies the following:

(i.) if $A \perp\!\!\!\perp B$ and $P(B) > 0$, then $P(A \mid B) = P(A)$

(ii.) if $A \perp\!\!\!\perp B$, then $A^c \perp\!\!\!\perp B$, $A \perp\!\!\!\perp B^c$, and $A^c \perp\!\!\!\perp B^c$.

Let $A_1, A_2, A_3, \ldots, A_n \in \mathcal{F}$. When do we say that these are independent?

Definition 2.4.3. (1) The set $\{A_1, \ldots, A_n\}$ are **pairwise independent** if

$$P(A_i \cap A_j) = P(A_i)P(A_j) \quad \text{for all } i \neq j$$

(2) The set $\{A_1, \ldots, A_n\}$ are (mutually) independent if any subset of at least two events are (mutually) independent.

Chapter 3

Random Variables & Univariate Distributions

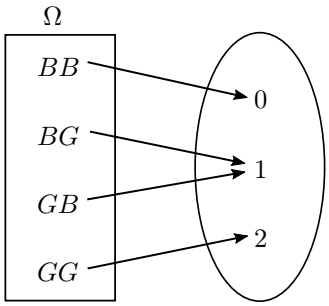
3.1 Week 2: Lecture 2 (continued)

3.1.1 The Random Variable

Wed 6 Oct 10:00

What is a random variable? Informally, it is a numerical quantity that takes different values with different probabilities. Its value is determined by the outcome of experiments.

Example 3.1.1. Consider the twin example from before. Let X represent the number of girls from a given birth. We can map each event to some value of X:



More formally we can say that X is a function, that is, $X : \Omega \to \mathbb{R}$, where

for $\omega \in \Omega$,$X(\omega) \in \mathbb{R}$. Then

$$\begin{aligned} P(X=1) &= P(\{\omega \in \Omega : X(\omega) = 1\} \\ &= P(\{BG, GB\}) = \frac{2}{4} = \frac{1}{2} \\ P(X>0) &= P(\{\omega \in \Omega : X(\omega) > 0\}) \\ &= P(\{BG, GB, GG\}) = \frac{3}{4}. \end{aligned}$$

◇

Definition 3.1.2. Let Ω be a sample space and E be a measurable space. For our purposes, let $E = \mathbb{R}$. A **random variable** is a function $X : \Omega \to E$ with the property that, if $A_x = \{\omega \in \Omega \ : \ X(\omega) \leq x\}$, then $A_x \in \mathcal{F}$ for all $x \in \mathbb{R}$.

3.2 Week 3: Lecture 1

3.2.1 Examples of Random Variables

Tue 12 Oct 14:00

Example 3.2.1. Let X be a random variable. For $x = 2$, we have $A_2 \in \mathcal{F}$, so we can write $P(A_2) = P(X \leq 2)$. ◇

Example 3.2.2. Suppose there is a family with two children. Let X = the number of girls. Then

$$\begin{aligned} P(A_0) &= P(\{BB\}) = \frac{1}{4} \\ P(A_1) &= P(\{BB, BG, GB\}) = \frac{3}{4} \\ P(A_{\frac{3}{2}}) &= P(A_1) = \frac{3}{4} \\ P(A_2) &= P(\Omega) = 1 \\ P(A_{-1}) &= P(\{\}) = 0 \\ P(A_\pi) &= P(\Omega) = 1. \end{aligned}$$

◇

3.2.2 The Cumulative Distribution Function

Definition 3.2.3. A random variable X is **positive** if $X(\omega) \geq 0$ for all $\omega \in \Omega$.

Definition 3.2.4. The **cumulative distribution function** (**CDF**) of a random variable X is the function $F_X : \mathbb{R} \to [0,1]$ given by $F_X(x) = P(\underbrace{X \leq x}_{A_x})$.

Example 3.2.5. Let X be a random variable and F_X be a valid CDF. Then

$$P(A_1) = P(X \leq 1) = F_X(1).$$

◇

Example 3.2.6. In our two child example,

$$F_X(0) = \frac{1}{4}, \quad F_X(1) = \frac{3}{4}, \quad F_X(2) = 1, \quad F_X(-1) = 0, \quad F_X\left(\frac{3}{2}\right) = \frac{3}{4}, \ldots$$

Moreover, note that the CDF in this case is a step function, as seen in the figure below.

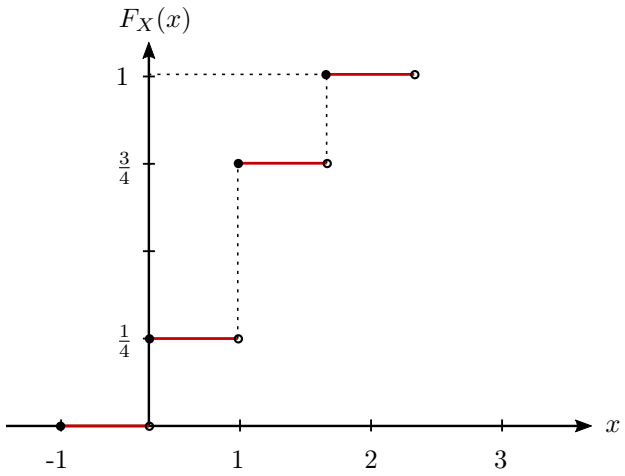


Figure 3.1: The Cumulative Distribution Function of the two child example.

◇

Definition 3.2.7. A function $g : \mathbb{R} \to \mathbb{R}$ is **right-continuous** if $g(x+) = g(x)$ for all $x \in \mathbb{R}$, where

$$g(x+) = \lim_{h \downarrow 0} g(x+h),$$
$$\text{and} \quad g(x-) = \lim_{h \downarrow 0} g(x-h).$$

Proposition 3.2.8. If F_X is a CDF, then

i. F_X is increasing, i.e., if $x < y$ then $F_X(x) \leq F_X(y)$.

ii. F_X is right-continuous, i.e. $F_X(x+) = F_X(x)$ for all $x \in \mathbb{R}$.

iii. $\lim_{x\to-\infty} = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.

Proof. i. If $x < y$, then $A_X \subseteq A_y$, so

$$F_X(x) = P(A_x) \leq P(A_y) = F_X(y).$$

ii. Take a decreasing sequence $\{x_n\}$ such that $x_n \downarrow x$ as $n \to \infty$ ($x_1 \geq x_2 \geq x_3 \geq \cdots$). We have

$$A_{x_1} \supseteq A_{X_2} \supseteq \cdots$$

and $A_x \supseteq A_{x_n}$. So

$$A_x = \bigcap_{n=1}^{\infty} A_{x_n}.$$

Then

$$\begin{aligned}
\lim_{x\to\infty} F_X(x_n) &= \lim_{n\to\infty} P(A_{x_n}) \\
&= P\left(\bigcap_{n\in\mathbb{N}} A_{x_n}\right) \\
&= P(A_x) \\
&= F_X(X) \\
\Rightarrow \lim_{h\downarrow 0} F_X(x+h) &= F_X(x).
\end{aligned}$$

iii. In M&P textbook.

□

Some basic properties of CDFs

Observe that

- $P(X > x) = 1 - P(X \leq x) = 1 - F_X(x)$
- $P(x < X \leq y) = F_X(y) - F_X(x)$
- $P(X < x) = F_X(x-)$
- $P(X = x) = F_X(x) - F_X(x-)$.

3.3 Week 3: Lecture 2

3.3.1 Types of Random Variables

Wed 13 Oct 10:00

Some examples of random variable types:

- Discrete: hurricanes (0,1,2,3,...)
- Continuous: javelin throw distance
- Continuous model for discrete situation: average salary
- Neither discrete nor continuous: queuing time
- Neither discrete nor continuous for discrete situation: yearly income

These types are not as clear cut as we may believe.

Definition 3.3.1. The **support** of a non-negative function $g : \mathbb{R} \to [0, \infty)$ is the subset of $\mathbb{R}$ where g is *strictly* positive.

Notation.

(i) $X \sim \text{Poisson}(\lambda)$, $X \sim N(\lambda, \sigma^2)$

(ii) $X \sim F_x$ (a CDF)

Example 3.3.2. Recall the prior two child example. Our discrete random variable was in the form of a step function. ◇

Definition 3.3.3. X is a discrete random variable if and only if it takes values in $\{x_1, x_2, x_3, \dots\} \subset \mathbb{R}$.

Definition 3.3.4. The probability mass function (PMF) of a discrete random variable X is the function $f_x : \mathbb{R} \to [0, 1]$ where $f_X(x) = P(X = x)$.

In our example: $f_X(0) = 1/4,\ f_X(1) = 1/2,\ f_X(2) = 1/4\ f_X(x) = 0$ for all other x. Hence, $\{0, 1, 2\}$ is the support.

(i) $f_X(x) = F_X(x) - F_X(x-)$

(ii) $F_X(x) = \sum_{u \in \mathbb{R}, u \leq x} f_X(u)$ i.e. $P(X \leq x) = \sum_{u : u \leq x} P(X = u)$

Proposition 3.3.5. For valid PMF $f_X(x)$ and valid CDF $F_X(x)$,

(i.) $f_X(x) = F_X(x) = F_X(x-)$, or

$$P(X = x) = P(X \leq x) - P(X < x).$$

(ii.) $F_X(x) = \sum_{u \in \mathbb{R},\ u \leq x} f_x(u)$, or

$$P(X \leq x) = \sum_{u : u \leq x} P(X = u).$$

3.3.2 Some Distributions

Binomial Distribution

We obtain a binomial distribution with the following:

- Repeat experiment n times.
- Each time, declare one of two outcomes: success or failure.
- Every trial is independent.
- $P(\text{“success"}) = p$ every repetition.

Define X as the number of successes. Let $X \sim \text{Bin}(n, p)$, where n is the number of trials and p is the probability of success. The PMF of X is

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots n.$$

and $f_X(x) = 0$ otherwise. Check if f_X is a valid PMF:

$$\sum_{x=0}^{n} f_X(x) = (p + (1-p))^n = 1^n = 1. \checkmark$$

Example 3.3.6. For our previous example, where X is the number of girls, $X \sim \text{Bin}\left(2, \frac{1}{2}\right)$. ◇

Bernoulli Distribution

$X \sim \text{Bernoulli}(p)$ is the same as $\text{Bin}(1, p)$.

Geometric Distribution

Same (Bernoulli) setup as Binomial, but:

- the number of trials is not fixed

- we repeat the experiment until first success

Let Y : number of trials required. Then $Y \sim \text{Geo}(p)$.

$$f_Y(y) = P(Y = y) = (1-p)^{y-1}p, \quad \text{for } y = 1, 2, \dots$$

Check the validity of f_X:

$$\sum_{y=1}^{\infty} f_Y(y) = \frac{p}{1-(1-p)} = \frac{p}{p} = 1.$$

Sometimes: Y^* : the number of failures before first success. Note that $Y^* = Y - 1$.

Negative Binomial Distribution

Same setup as Geometric, but we stop when we obtain the rth success for some given $r \in \mathbb{Z}^+$.

Let X : number of trials required to obtain r successes. Then $X \sim \text{NegBin}(r, p)$, and

$$f_X(x) = p^r(1-p)^{x-r} \quad \text{for } x = r, r+1, r+2, \dots$$

A more common iteration of the Negative Binomial Distribution:

- X^* : number of failures before r successes. (support is $\{0, 1, 2, \dots\}$).
- $X^* = X - r$

Note that $\text{NegBin}(1, p)$ is the same as $\text{Geo}(p)$.

3.4 Week 4: Lecture 1

3.4.1 A Distribution of Emails

Tue 19 Oct 14:00

Example 3.4.1. Suppose that we want to create a probability distribution of how many emails are sent in each point of time on the LSE server between 10AM and 11AM. There aren't exactly any Bernoulli trials here by default, so we need to create them. Split the 1 hour period into 60 one minute intervals. If an email is sent during a given interval, we count that as a success. Call this random variable X_{60}. Note that $X_{60} \sim \text{Bin}(60, p)$. Here's the problem: this model will systematically undercount the number of emails, since there are a maximum of 60 trials, and each trial counts for only 1. Now increase the number of Bernoulli trials, and assume that the probability remains constant over any given time interval. Hence, if we have 120 30-second intervals, then $X_{120} \sim \text{Bin}(120, \frac{p}{2})$.

$\diamond$

What if we continue increasing the number of trials? Or, what is X_∞, i.e. $\lim \text{Bin}(n,p)$ as $n \to \infty$, $p \to 0$ with np remaining constant?

$$\begin{aligned}
f_X(x) &= \lim_{n\to\infty,\ p\to 0,\ np=\lambda} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \frac{\lambda^x}{n^x} \left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-x+1)}{n\times n\cdots\times n} \left(1-\frac{\lambda}{n}\right)^{n} \left(1-\frac{\lambda}{n}\right)^{-x} \frac{\lambda^x}{x!} \\
&= 1 \times e^{-\lambda} \times 1 \times \frac{\lambda^x}{x!} \\
&= \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots
\end{aligned}$$

which is the PMF of the Poisson distribution. Hence, $X \sim \text{Poisson}(\lambda)$.

3.4.2 Discrete Uniform Distribution

Definition 3.4.2. A discrete random variable X is **uniformly distributed** if it has PMF

$$f_X(x) = \begin{cases} \frac{1}{n} & \text{for } x \in \{x_1, x_2, \ldots, x_n\} \\ 0 & \text{otherwise.} \end{cases}$$

3.4.3 Continuous Random Variables

Note that a discrete CDF is a step function. A continuous CDF, however, is continuous everywhere.

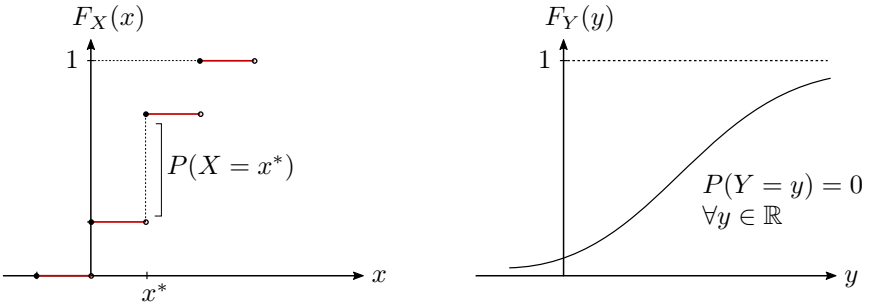


Figure 3.2: Discrete vs. Continuous CDF

Definition 3.4.3. A random variable X is **continuous** if its CDF can be written as

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\ du$$

for some integrable real-valued function f_X.

Definition 3.4.4. We write $f_X(x)$ to denote the **probability density function** (PDF) of X.

Proposition 3.4.5. For all $x \in \mathbb{R}$,

i. $f_X(x) = \frac{d}{dx} F_X(x) = F_X'(x)$

ii. $f_X(x) \geq 0$ for all $x \in \mathbb{R}$

iii. $\int_{\mathbb{R}} f_X(x)\ dx = 1$.

iv. Let $a, b \in \mathbb{R}$, with $a < X \leq b$. Then

$$\begin{aligned} F_X(b) - F_X(a) &= P(a < X \leq b) \\ &= \int_a^b f_X(u)\ du. \end{aligned}$$

v. For any $B \subseteq \mathbb{R}$,

$$P(X \in B) = \int_B f_X(x)\ dx.$$

Example 3.4.6. A continuous random variable X is uniformly distributed for parameters a, b if

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

More compactly, $X \sim \text{Unif}[a, b]$. We can say

$$P(c \leq X \leq c + h) = \frac{h}{b-a}.$$

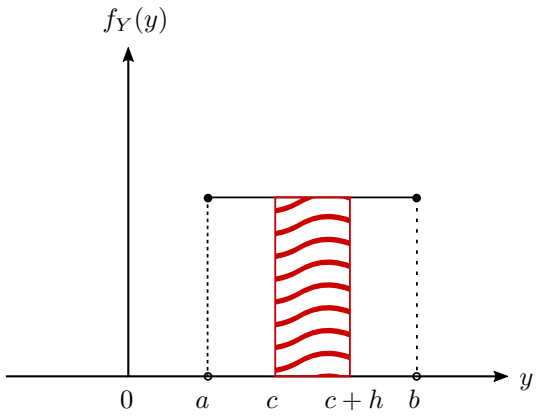


Figure 3.3: The continuous Uniform Distribution with parameters a, b.

◇

Example 3.4.7. Let X be the number of email arrivals in an hour, and suppose $X \sim \text{Poisson}(\lambda)$. Note that we can scale this, with $X(t) = \text{Poisson}(t\lambda.)$ Let Y be the time of the first arrival. Note that

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= 1 - P(Y > y) \\ &= 1 - P(X(y) = 0) \\ &= 1 - \frac{e^{-\lambda y}(\lambda y)^0}{0!} \\ &= 1 - e^{\lambda y}, \ y \geq 0. \end{aligned}$$

Differentiate $F_Y(y)$ to find the density function:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \frac{d}{dy}(1 - e^{-\lambda y}) \\ &= \lambda^{-\lambda y}, \quad y \geq 0; \end{aligned}$$

which is the PDF of the exponential distribution with rate λ. Hence, $Y \sim \text{Exp}(\lambda)$. ◇

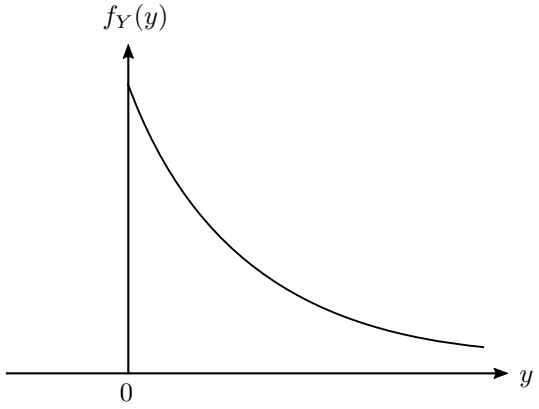


Figure 3.4: A simple graphical representation of the PDF of the Exponential Distribution.

3.5 Week 4: Lecture 2

3.5.1 Some Continuous Distributions

Wed 20 Sep 10:00

Exponential Distribution

An Exponential Distribution can be described with either a rate parameter or a scale parameter.

Definition 3.5.1. We say $X \sim \text{Exp}(\lambda)$ for **rate parameter** λ if $f_x(x) = \lambda e^{-\lambda x}$ for $x > 0,\ \lambda > 0$. We say $X \sim \text{Exp}(\theta)$ for **scale parameter** θ if $f_x(x) = 1/\theta e^{-x/\theta}$, for $x > 0$.

Note that $\theta = \frac{1}{\lambda}$.

Normal Distributions

Definition 3.5.2. We say that X is **normally distributed**, or $X \sim N(\mu, \sigma^2)$ for mean μ and variance σ^2 if

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } x \in \mathbb{R}.$$

Example 3.5.3. The *standard normal distribution* is Normal$(0, 1)$. ◇

Some properties of the normal distribution:

- If $X \sim N(\mu, \sigma^2)$, then $\frac{x-\mu}{\sigma} \sim N(0, 1)$.
- If $Z \sim N(0, 1)$, then $\mu + \sigma Z \sigma N(\mu, \sigma^2)$.

Remark 3.5.4. The normal CDF has no closed form. It can be written as an infinite sum, but it cannot be written in a finite number of operations.

Remark 3.5.5. If $Z \sim N(0,1)$ we write $\Phi(z)$ for $F_Z(z)$. The Φ function is the CDF for the standard normal.

Gamma Distribution

Definition 3.5.6. The PDF of the **Gamma Distribution** is

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0, \quad 0 \text{ otherwise}$$

We denote the gamma distribution as Gamma(α, λ), where α is the shape parameter, and λ is the rate parameter.

Definition 3.5.7. The **gamma function** is as follows

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

If $k \in \mathbb{Z}^+$, then $\Gamma(k) = (k-1)!$, so Gamma$(1, \lambda)$ is Exp(λ).

Remark 3.5.8. Beware! The scale parameter θ is commonly in use too, with $\theta = \frac{1}{\lambda}$.

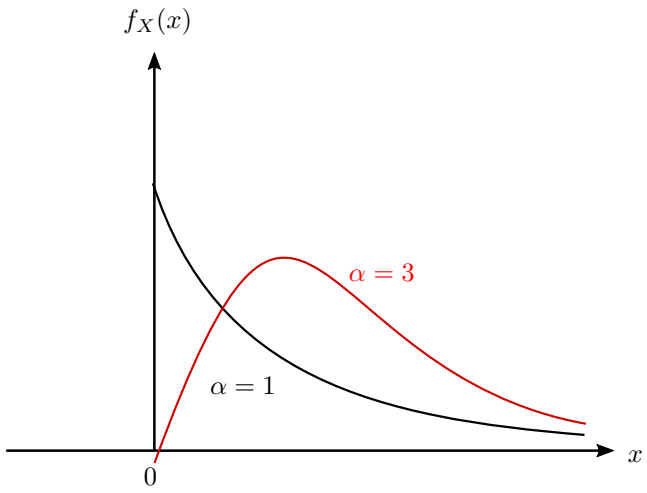


Figure 3.5: PDF of the Gamma Distribution for $\alpha = 1$ and $\alpha = 3$

3.5.2 Expectation, Variance, and Moments

We can now characterize some properties of random variables:

- mode$((X) = \arg\max(f_X(x))$

- median$(X) = m$, where $F_X(m) = 1/2$.
- mean(X)? See below:

Definition 3.5.9. The mean, or **expected value** of X is

$$\mu = \mathbb{E}(x) = \begin{cases} \sum_x x f_x(x), & X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_X(x)\ dx, & X \text{ is continuous.} \end{cases}$$

Note. We generally ask that

$$\sum_x |x| f_X(x) < \infty.$$

or for continuous random variables,

$$\int_{\mathbb{R}} |x| f_X(x)\ dx < \infty.$$

Example 3.5.10. Let $X \sim \text{Uniform}[a, b]$. Then

$$\begin{aligned} \mathbb{E}(X) &= \int_{\mathbb{R}} x f_X(x)\ dx \\ &= \int_a^b x \frac{1}{b-a}\ dx \\ &= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{(b-a)(b+a)}{2(b-a)} \\ &= \frac{a+b}{2} \end{aligned}$$

◇

Remark 3.5.11. Note that

$$\mathbb{E}(g(x)) = \begin{cases} \sum_x g(x) f_X(x) & \text{(discrete)} \\ \int_{\mathbb{R}} g(x) f_X(x)\ dx & \text{(continuous),} \end{cases}$$

as long as

$$\sum_x |g(x)| f_X(x) < \infty,$$

and similarly when X is continuous.

Example 3.5.12. For a random variable and $a_0, a_1, a_2, \cdots \in \mathbb{R}$,

$$\mathbb{E}(a_0 + a_1 X + a_2 X^2 + \cdots) = a_0 + a_1 \mathbb{E}(X) + a_2 \mathbb{E}(X^2) + \cdots$$

◇

Remark 3.5.13. A quick aside:

$$\int (e^x + 2\sin(x))\ dx = \int e^x\ dx + 2\int \sin x\ dx$$

Definition 3.5.14. The **variance** (σ^2) of a function is

$$\mathrm{Var}(x) = \mathbb{E}[(x - \mathbb{E}(x))^2].$$

Observe that

$$\sigma^2 = \begin{cases} \sum_x (x-\mu)^2 f_X(x) & \text{(discrete)} \\ \int_{\mathbb{R}} (x-\mu)^2 f_X(x)\ dx & \text{(continuous).} \end{cases}$$

Some properties of variance:

i. $\mathrm{Var}(X) \geq 0$,

ii. $\mathrm{Var}(a_0 + a_1 X) = a_1^2 \mathrm{Var}(X)$ (prove this!),

iii. $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ (this too!).

Definition 3.5.15. The standard deviation σ of a random variable X is $\sqrt{\mathrm{Var}(X)}$.

3.6 Week 5: Lecture 1

3.6.1 Markov Inequality

Tue 26 Oct 14:00

Theorem 3.6.1 (Markov Inequality)**.** If Y is a positive random variable, and $\mathbb{E}(Y) < \infty$, then

$$P(Y \geq a) \leq \frac{\mathbb{E}(Y)}{a},$$

for any $a > 0$.

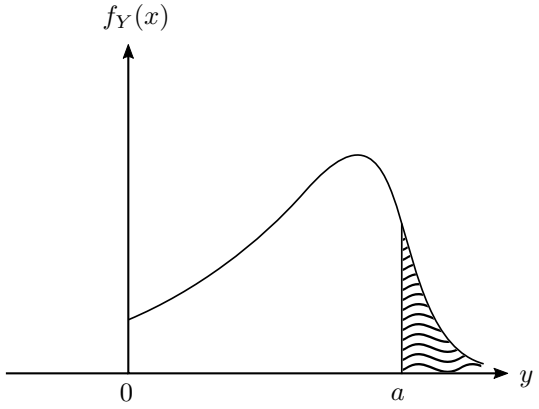


Figure 3.6: The Markov Inequality shows that the shaded area (the survival function of X evaluated at a) is always less than or equal to $\frac{\mathbb{E}(Y)}{a}$.

Proof. Observe that

$$\begin{aligned} P(Y \geq a) &= \int_a^\infty f_Y(y)\ dy \\ &\leq \int_a^\infty f_Y(y)\ dy \\ &= \frac{1}{a}\int_a^\infty y f_Y(y)\ dy \\ &\leq \frac{1}{a}\int_0^\infty y f_Y(y)\ dy \\ &= \frac{1}{a}\mathbb{E}(Y). \end{aligned}$$

□

Example 3.6.2. Let Y be the random variable representing a person's lifespan. Say that $\mathbb{E}(Y) = 80$. Note that

$$P(Y \geq 160) \leq \frac{80}{160} = \frac{1}{2}.$$

◇

Theorem 3.6.3 (Chebyshev Inequality)**.** If X is a random variable with $\text{Var}(X) \leq \infty$, then

$$P(|X - \mathbb{E}(x)| \geq a) \leq \frac{\text{Var}(x)}{a^2},$$

for any $a > 0$.

Proof. Let $Y = \text{Var}(X)$. Observe that

$$\begin{aligned} P(|X - \mathbb{E}(X)| \geq a) &= P((X - \mathbb{E}(X))^2 \geq a^2) \\ &\leq \frac{\mathbb{E}(Y)}{a^2} \\ &= \frac{\mathbb{E}[(X - \mathbb{E}(X))^2]}{a^2} \\ &= \frac{\text{Var}(X)}{a^2}. \end{aligned}$$

Alternatively,

$$\begin{aligned} P(X \geq \mu + \lambda\sigma \text{ or } X \leq \mu - \lambda\sigma) &= P\left(\left|\frac{x-\mu}{\sigma} \geq \lambda\right|\right) \\ &= P\left(|x - \mu \geq \lambda\sigma\right) \\ &\leq \frac{\sigma^2}{\lambda^2\sigma^2} \\ &= \frac{1}{\lambda^2}. \end{aligned}$$

□

Example 3.6.4. Let $X \sim \text{Normal}(\mu, \sigma^2)$. Then

$$P\left(\left|\frac{x-\mu}{\sigma}\right| \geq 2\right) \leq \frac{1}{4}.$$

Note that the exact probability is ≈ 0.05. ◇

Definition 3.6.5. A function $g : \mathbb{R} \to \mathbb{R}$ is **convex** if for any $a \in \mathbb{R}$, we can find λ such that

$$g(x) \geq g(a) + \lambda(x - a) \quad \text{for all } x \in \mathbb{R}.$$

A **concave** function is the same principle, but with

$$g(x) \leq g(a) + \lambda(x - a) \quad \text{for all } x \in \mathbb{R}.$$

3.6.2 Jensen Inequality

Theorem 3.6.6 (Jensen Inequality)**.** If X is a random variable (with $\mathbb{E}(x)$ defined) and $g : \mathbb{R} \to \mathbb{R}$ is convex (with $\mathbb{E}(g(x)) < \infty$), then

$$\mathbb{E}(g(x)) \geq g(\mathbb{E}(x)).$$

Proof. Using the definition of a convex function with $a = \mathbb{E}(X)$, we have

$$\begin{aligned}
\mathbb{E}(g(X)) &= \int_{\mathbb{R}} g(x) f_X(x)\ dx \\
&\geq \int_{\mathbb{R}} \left[g(\mathbb{E}(X)) + \lambda(x - \mathbb{E}(X)) \right] f_X(x)\ dx \\
&= g(\mathbb{E}(X)) \int_{\mathbb{R}} f_X(x)\ dx + \lambda \int_{\mathbb{R}} (x - \mathbb{E}(X)) f_X(x)\ dx \\
&= g(\mathbb{E}(X)) + \lambda \mathbb{E}(X - \mathbb{E}(X)) \\
&= g(\mathbb{E}(X)).
\end{aligned}$$

□

If $h : \mathbb{R} \to \mathbb{R}$ is concave, then $\mathbb{E}(h(X)) \leq h(\mathbb{E}(X))$.

Example 3.6.7. A special case:

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

◇

Example 3.6.8. Note that

$$\mathbb{E}(X^2) \geq (\mathbb{E}(X))^2.$$

◇

Example 3.6.9. If $Y > 0$,

$$\mathbb{E}\left(\frac{1}{Y}\right) \geq \frac{1}{\mathbb{E}(Y)}.$$

◇

3.6.3 Moments

Definition 3.6.10. The rth moment of a random variable X is

$$\mu'_r = \mathbb{E}(X^r), \quad \text{for } r = 1, 2, 3, \ldots$$

Definition 3.6.11. The rth central moment of X is

$$\mu_r = \mathbb{E}[(X - \mathbb{E}(X))^r], \quad \text{for } r = 1, 2, 3, \ldots$$

Example 3.6.12. Some moments:

$$\begin{aligned}
&\mu'_1 = \mathbb{E}(X), \\
&\mu_1 = 0, \\
&\mu_2 = \mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\
&\Rightarrow \mu_2 = \mu'_2 - (\mu'_1)^2.
\end{aligned}$$

◇

Example 3.6.13. Let $X \sim \text{Exp}(\lambda)$. Then

$$\begin{aligned}
\mu_r' &= \mathbb{E}(X^r) \\
&= \int_{\mathbb{R}} x^r f_X(x)\ dx \\
&= \int_0^\infty x^r \lambda e^{-\lambda x}\ dx \\
&= \int_0^\infty x^r \frac{d}{dx}(-e^{-\lambda x})\ dx \\
&= \left[x^r(-e^{-\lambda x})\right]_0^\infty - \int_0^\infty r x^{r-1}(-e^{-\lambda x})\ dx \\
&= \frac{r}{\lambda}\int_0^\infty x^{r-1}\lambda e^{-\lambda x}\ dx \\
&= \frac{r}{\lambda}\mu_{r-1}'.
\end{aligned}$$

Observe that

$$\begin{aligned}
\mu_r' &= \frac{r}{\lambda}\mu_{r-1}' \\
&= \frac{r}{\lambda}\frac{r-1}{\lambda}\mu_{r-2}' \\
&= \dots \\
&= \frac{r}{\lambda}\frac{r-1}{\lambda}\cdots\frac{1}{\lambda}\mu_0' \\
&= \frac{r!}{\lambda^r}.
\end{aligned}$$

So $\mathbb{E}(X) = \frac{1}{\lambda}$, $\mathbb{E}(X^2) = \frac{2}{\lambda^2}$, and so on. Further note that

$$\text{Var}(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

◇

3.7 Week 5: Lecture 2

Wed 28 Oct 10:00

3.7.1 Moment-Generating Function

Definition 3.7.1. The **moment-generating function** (MGF) of a random variable X is a function $M_X : \mathbb{R} \to \mathbb{R}_0^+$ given by

$$M_X(t) = \mathbb{E}(e^{tX}) = \begin{cases} \sum_x e^{tx} f_X(x) & \text{(discrete)} \\ \int_{\mathbb{R}} e^{tx} f_X(x)\, dx & \text{(continuous)}, \end{cases}$$

where we require that $M_X(t) < \infty$ for all $t \in [-h, h]$ for some $h > 0$ (a neighborhood of 0).

Remark 3.7.2. Note that

$$e^y = 1 + y + \frac{y^2}{2!} + \frac{y^3}{3!} + \cdots = \sum_{j=0}^{\infty} \frac{y^j}{j!}.$$

And so

$$\begin{aligned} M_X(t) &= \mathbb{E}(e^{tX}) \\ &= \mathbb{E}\left(1 + tX + \frac{(tX)^2}{2!} + \cdots\right) \\ &= \mathbb{E}\left[\sum_{j=0}^{\infty} \frac{(tX)^j}{j!}\right] \\ &= \sum_{j=0}^{\infty} \frac{t^j}{j!} \mathbb{E}(X^j) \\ &= 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \frac{t^3}{3!}\mu_3' + \cdots \end{aligned}$$

The coefficient of t^r is $\frac{\mu_r'}{r!} = \frac{\mathbb{E}(X^r)}{r!}$.

Proposition 3.7.3. The rth derivative of $M_X(t)$ at $t = 0$ is μ_r'.

Proof.

$$\begin{aligned} M_X^{(r)}(t) &= \frac{d^r}{dt^r} \mu_X(t) \\ &= \frac{d^r}{dt^r}\left(1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \frac{t^3}{3!}\mu_3' + \cdots\right) \\ &= \mu_r' + t\mu_{r+1}' + \frac{t^2}{2!}\mu_{r+2}' + \cdots \end{aligned}$$

This implies

$$\mu_X^{(r)}(0) = \mu_r' = \mathbb{E}(X^r).$$

□

Proposition 3.7.4. If X, Y are random variables and we can find $h > 0$ such that $M_X(t) = M_Y(t)$ for all $|t| < h$, i.e., $t \in (-h, h)$, then

$$F_X(x) = F_Y(x) \quad \text{for all } x \in \mathbb{R}.$$

Proof. Omitted. □

Example 3.7.5. Let $X \sim \text{Poisson}(\lambda)$. Observe that

$$\begin{aligned}
M_X(t) &= \mathbb{E}(e^{tX}) \\
&= \sum_x e^{tx} f_X(x) \\
&= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=0}^{\infty} \frac{e^{-\lambda}(\lambda e^t)^x}{x!} \\
&= e^{\lambda e^t} e^{-\lambda} \sum_{x=0}^{\infty} \frac{e^{-\lambda e^t}(\lambda e^t)^x}{x!} \\
&= e^{\lambda(e^t-1)} \quad \text{for } t \in \mathbb{R} \\
&= \exp(\lambda(e^t - 1)) \\
&= \exp(\lambda(e^t - 1)) \\
&= \exp(\lambda(1 + t + \frac{t^2}{2} + \frac{t^3}{6} + \cdots - 1 \\
&= 1 + \lambda(t + \frac{t^2}{2} + \cdots) + \frac{\lambda^2(t + \frac{t^2}{2} + \cdots)^2}{2} + \cdots \\
&= 1 + \lambda t + \frac{\lambda t^2}{2} + \frac{\lambda^2 t^2}{2} + \cdots \\
&= 1 + \lambda t + \frac{\lambda t^2}{2} + (\lambda + \lambda^2)\frac{t^2}{2} + \cdots .
\end{aligned}$$

From this, $\mathbb{E}(X) = \lambda$, and $\mathbb{E}(X^2) = \lambda + \lambda^2$. Moreover,

$$\text{Var}(X) = \lambda + \lambda^2 - \lambda^2 = \lambda.$$

Or,

$$M_X' = \exp(\lambda(e^t - 1))\lambda e^t \Rightarrow \mu_1' = M_X'(0) = \lambda.$$

◇

Example 3.7.6. Let $Y \sim \Gamma(\alpha, \lambda)$. Then

$$\begin{aligned} M_Y(t) &= \mathbb{E}(e^{tY}) \\ &= \int_0^\infty e^{tY} \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}\, dy \\ &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha} \int_0^\infty \frac{(\lambda-t)^\alpha}{\Gamma(\alpha) y^{\alpha-1} e^{-(\lambda-t)y}}\, dy \\ &= \left(\frac{\lambda}{\lambda-t}\right)^\alpha \\ &= \left(1-\frac{t}{\lambda}\right)^{-\alpha}, \quad \text{for } |t| < \lambda. \end{aligned}$$

◇

Negative Binomial Expansion

$$\begin{aligned} M_Y(t) &= \left(1-\frac{t}{\lambda}\right)^{-\alpha} \\ &= \sum_{j=0}^{\infty} \binom{j+\alpha-1}{\alpha-1} \left(\frac{t}{\lambda}\right)^j \end{aligned}$$

So, for example, the coefficient of $\frac{t^j}{j!}$ is $\frac{(j+\alpha-1)!}{(\alpha-1)!}\lambda^{-j}$. Then

$$\begin{aligned} \mathbb{E}(Y) &= \frac{(1+\alpha-1)!}{\alpha-1)!}\lambda^{-1} \\ &= \frac{\alpha!}{(\alpha-1)!}\frac{1}{\lambda} \\ &= \frac{\alpha}{\lambda}. \end{aligned}$$

3.7.2 Cumulant-Generating function

Definition 3.7.7. The **cumulant-generating function** (CGF) of a random variable X is $K_X(t) \ln(M_X(t))$.

We can write

$$K_X(t) = \kappa_1 t + \frac{\kappa_2}{2!}t^2 + \frac{\kappa_3}{3!}t^3 + \cdots$$

The rth cumulant, κ_r, is the coefficient of $\frac{t^r}{r!}$ in the power series expansion of $K_X(t)$ about 0.

Example 3.7.8. Let $X \sim \text{Poisson}(\lambda)$. Then

$$\begin{aligned} K_X(t) &= \ln M_X(t) \\ &= \ln(\exp(\lambda(e^t - 1))) \\ &= \lambda(e^t - 1) \\ &= \lambda t + \lambda\frac{t^2}{2} + \lambda\frac{t^3}{3!} + \cdots . \end{aligned}$$

So, $\kappa_1 = \kappa_2 = \kappa_3 = \cdots = \lambda$. ◇

Proposition 3.7.9. If X is a random variable, then

i. $\kappa_1 = \mu_1' = \mathbb{E}(X)$

ii. $\kappa_2 = \mu_2' - (\mu_1')^2 = \mu_2 = \mathrm{Var}(X)$

iii. $\kappa_3 = \mu_3 = \mathbb{E}[(X - \mathbb{E}(X))^3]$.

Proof.

i. Observe that

$$\begin{aligned} K_X(t) &= \ln(M_X(t)) \\ \Rightarrow K_X'(t) &= \frac{1}{M_X(t)} M_X'(t) \\ \Rightarrow \kappa_1 = K_X'(0) &= \frac{1}{M_X(0)} M_X'(0) = \mu_1. \end{aligned}$$

ii.

$$\begin{aligned} K_X''(t) &= \left(\frac{M_X'(t)}{M_X(t)}\right)' = \frac{M_X''(t)M_X(t) - (M_X'(t))^2}{(M_X(t))^2} \\ \Rightarrow \kappa_2 &= K_X''(0) = \mu_2' - (\mu_1')^2. \end{aligned}$$

iii. Left as an exercise.

□

3.8 Week 6: Reading Week

:)

3.9 Week 7: Lecture 1

Tue 9 Nov 14:00

3.9.1 Functions of Random Variables

Let X be a random variable, and $g : \mathbb{R} \to \mathbb{R}$ be a well-behaved function. We're interested in

$$Y = g(X), \qquad \mathbb{E}(g(X))$$

When we first encountered functions of random variables, we started with the CDF, and we worked from there. But observe that

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(g(X) \leq y) \neq P(X \leq g^{-1}(y)). \end{aligned}$$

Hence, this doesn't work, e.g., for $g(x) = x^2$.

Definition 3.9.1. If $B \subseteq \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, the **inverse image** of B is defined as

$$g^{-1}(B) = \{x \in \mathbb{R} : \ g(x) \in B\}.$$

Example 3.9.2. If $g(x) = x^2$,

$$\begin{aligned} g^{-1}(\{4\}) &= \{-2, 2\} \\ g^{-1}([0, 1]) &= [-1, 1] \end{aligned}$$

◇

Example 3.9.3. For a random variable Y and some set B,

$$\begin{aligned} P(Y \in B) &= P(g(X) \in B) \\ &= P(\{\omega \in \Omega : \ g(X(\omega)) \in B\}) \\ &= P(\{\omega \in \Omega : \ X(\omega) \in g^{-1}(B)\}) \\ &= P(X \in g^{-1}(B)) \end{aligned}$$

◇

Remark 3.9.4. Note that

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(Y \in (-\infty, y]) \\ &= P(X \in g^{-1}((-\infty, y]) \\ &= \begin{cases} \displaystyle\sum_{x : g(x) \leq y} & \text{(discrete)} \\ \displaystyle\int_{x : g(x) \leq y} f_X(x) \ dx & \text{(continuous).} \end{cases} \end{aligned}$$

Further,

$$f_Y(y) = \ldots = \sum_{x:g(x)=y} f_X(x).$$

Example 3.9.5. Let X be a continuous random variable and $Y = g(X) = X^2$. For $y \geq 0$:

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(X^2 \leq y) \\
&= P(-\sqrt{y} \leq X \leq \sqrt{y} \\
&= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\
\Rightarrow f_Y(y) &= \frac{d}{dy} F_Y(y) \\
&= \begin{cases} \frac{1}{2\sqrt{y}}(f_X(\sqrt{y}) + f_X(-\sqrt{y}), & y \geq 0 \\ 0, & y < 0. \end{cases}
\end{aligned}$$

If $X \sim \text{Normal}(0, 1)$, then

$$\begin{aligned}
f_y(y) &= \frac{1}{2\sqrt{y}} \left(\frac{1}{\sqrt{2\pi}} e^{\frac{-(\sqrt{y})^2}{2}} + \ldots \right) \\
&= \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}}, \quad y \geq 0. \\
&= \frac{(1/2)^2}{\sqrt{\pi}} y^{\frac{1}{2}-1} e^{-\frac{1}{2}y}.
\end{aligned}$$

Note that $\sqrt{\pi} = \Gamma(\frac{1}{2})$. Hence, $Y \sim \Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$. ◇

Monotonicity

Definition 3.9.6. A function is **monotone** if it is strictly increasing or strictly decreasing.

Remark 3.9.7. If a function is monotone increasing, then

$$y \in (c, d) \iff x \in (a, b), .$$

and hence, $g^{-1}((c, d)) = (a, b)$.

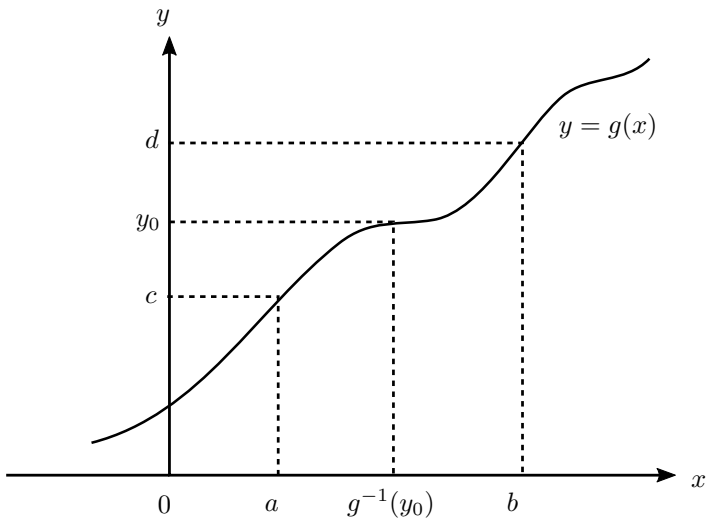


Figure 3.7: A monotone increasing function

Similarly, if a function is monotone decreasing, then

$$y \in (c, d) \iff x \in (a, b), .$$

and hence, $g^{-1}((c, d)) = (a, b)$

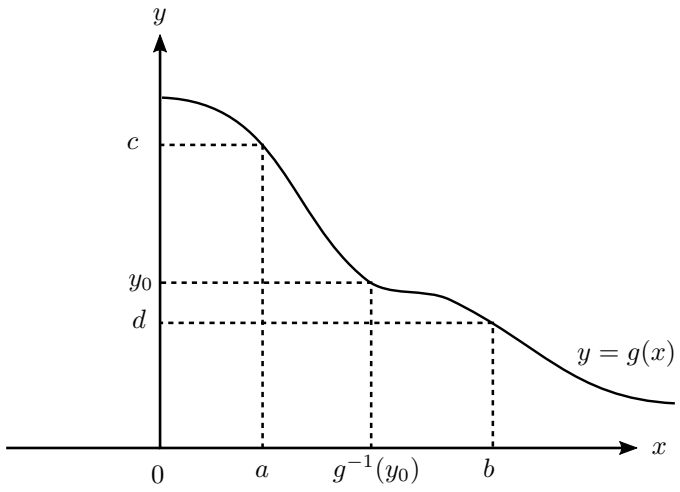


Figure 3.8: A monotone decreasing function

In general, if g is monotone (increasing or decreasing),

$$g^{-1}((-\infty, y]) = \{x \in \mathbb{R} : \ g(x) \in (-\infty, y]\}$$
$$= \begin{cases} (-\infty, g^{-1}(y)], & g \text{ is increasing} \\ [g^{-1}(y), \infty), & g \text{ is increasing.} \end{cases}$$

$$F_Y(y) = P(X \in g^{-1}((-\infty, y]))$$
$$= \begin{cases} P(X \in (-\infty, g^{-1}(y)]), & g \uparrow \\ P(X \in (g^{-1}(y), \infty]), & g \downarrow \end{cases}$$
$$= \begin{cases} F_X(g^{-1}(y)) & g \uparrow \\ 1 - F_X(g^{-1}(y)-), & g \downarrow \end{cases}$$

Note. $F_X(x-) = \lim_{h \downarrow 0} F_X(x-h) = P(X < x)$.

Remark 3.9.8. If X is continuous, then

$$f_Y(y) = \begin{cases} \dfrac{d}{dy} F_X(g^{-1}(y)), & g \ \uparrow \\ \dfrac{d}{dy}(1 - F_X(g^{-1}(y))), & g \ \downarrow \end{cases}$$
$$= \begin{cases} \left(\dfrac{d}{dy} g^{-1}(y)\right) f_X(g^{-1}(y)), & g \ \uparrow \\ \left(-\dfrac{d}{dy} g^{-1}(y)\right) f_X(g^{-1}(y)), & g \ \downarrow \end{cases}$$
$$= f_X(g^{-1}(y)) \left|\frac{d}{dy} g^{-1}(y)\right|, \quad g \ \uparrow \text{ or } \downarrow .$$

Example 3.9.9. Let $Y = e^X$. Note that

$$g(x) = e^x \iff g^{-1}(y) = \log y,$$

so

$$f_Y(y) = f_X(g^{-1}(y)) \left|\frac{d}{dy} g^{-1}(y)\right|$$
$$= f_X(\log y) \left|\frac{1}{y}\right|$$
$$= f_X(\log y) \frac{1}{y}, \quad y \geq 0.$$

If we define $y = g(x)$, $x = g^{-1}(y)$, we can write

$$\boxed{f_Y(y) = f_X(x) \left|\frac{dx}{dy}\right|}$$

$\diamond$

3.10 Week 7: Lecture 2

3.10.1 Location-scale transformation

Wed 10 Nov 10:00

Let Y be a continuous random variable, and $Y = \mu + \sigma X$, with $\sigma > 0$. Then

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|.$$

Note that

$$y = \mu + \sigma x \iff x = \frac{y-\mu}{\sigma},$$

so

$$f_Y(y) = f_X\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma}.$$

What about the MGF/CGF?

$$\begin{aligned} M_Y(t) &= \mathbb{E}(e^{tY}) = \mathbb{E}(e^{t(\mu+\sigma x)}) \\ &= \mathbb{E}(e^{t\mu}e^{t\mu x}) = e^{t\mu}M_X(t\sigma) \\ &\Rightarrow K_Y(y) = \ln M_Y(t) = t\mu + K_X(t\sigma) \\ &= t\mu + t\sigma K_{X,1} + \frac{(t\sigma)^2}{2}K_{X,2} + \frac{(t\sigma)^3}{3!}K_{X,3}\cdots \end{aligned}$$

$$\begin{aligned} K_{Y,1}(t) &= \mu + \sigma K_{X,1}(t) \\ K_{Y,r}(t) &= \sigma^r K_{X,r}(t), \quad \text{for } r = 2,3,4,\ldots \end{aligned}$$

3.10.2 Sequences of Random Variables & Convergence

Definition 3.10.1. A sequence **converges** $(x_n) \to x$ if, for all $\varepsilon > 0$ there exists some $N \in \mathbb{N}$ such that $|x_n - x| < \varepsilon$ for all $n \geq N$.

Say we have a sequence of random variables (X_n). What does it mean to say that (X_n) "converges"?

Convergence in...

Definition 3.10.2. We say that a sequence of random variable (X_n) converges in

- **probability**, if $\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1$, then $X_n \to^P X$.
- **distribution**, if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all $x \in \mathbb{R}$, then $X_n \to^d X$. Convergence in distribution is a *milder* form of convergence than Probability.
- **mean square**, if $\mathbb{E}[(X_n - X)^2] = 0$ then $X_n \to^{m.s} X$. Convergence in mean square is *stronger* than convergence in probability.

Remark 3.10.3. Note that

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) \leq \frac{\mathbb{E}[(X_n - X)^2]}{\varepsilon^2} \to 0.$$

So

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) \to 0, \quad X_n \to^P X.$$

And thus,

$$\boxed{\text{convergence in m.s.} \Rightarrow \text{c. in probability} \Rightarrow \text{c. in distribution}}.$$

Convergence almost surely

Definition 3.10.4. We say that X_n converges to X **almost surely** if

$$P\left(\lim_{n\to\infty} |X_n - X| < \varepsilon\right) = 1.$$

More compactly, we say $X_n \to^{\text{a.s.}} X$.

Remark 3.10.5. Alternatively, if

$$A = \{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\},$$

then we want $P(A) = 1$. Now consider A^c. There exists $\varepsilon > 0$ where for every n we can find $m \geq n$ with $|X_m(\omega) - X(\omega)| > \varepsilon$. Equivalently: There are infinitely many m with $|X_m(\omega) - X(\omega)| > \varepsilon$.

If

$$A_n = |X_n - 0| > \varepsilon$$

for some $\varepsilon \in \mathbb{R}$, then $P(\text{finitely many } A_n \text{ occur}) = 1$, i.e. $X_n \to^{\text{a.s.}} 0$. Or,

"There's going to be a last one" - Milt Mavrakakis.

Remark 3.10.6. Note that convergence...

$$\boxed{\text{Almost Surely} \Rightarrow \text{in Probability} \Rightarrow \text{in Distribution}}$$

and, again, that convergence in

$$\boxed{\text{Mean Square} \Rightarrow \text{in Probability} \Rightarrow \text{in Distribution.}}$$

3.10.3 The Borel-Cantelli Lemmas

Definition 3.10.7. The **limit superior** is defined as

$$A^c = \limsup_{n\to\infty} E_n = \bigcap_{n\in\mathbb{N}} \left(\bigcup_{m=n}^{\infty} E_m \right).$$

Note that $\bigcup_{m=n}^{\infty} E_m$ occurs when at least one E_m $(m \geq n)$ occurs.

Theorem 3.10.8 (First Borel-Cantelli Lemma)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and $E_1, E_2, E_3, \ldots \in \mathcal{F}$ with $\sum_{n\in\mathbb{N}} P(E_n) < \infty$ then $P(\limsup_{n\to\infty} E_n) = 0$.

Proof. Observe that

$$\begin{aligned} P(\limsup_{n\to\infty} E_n) &= P\left(\bigcap_{n=1}^{\infty} \left(\bigcup_{m=n}^{\infty} E_m \right) \right) \\ &= P\left(\bigcap_{n=1}^{\infty} \right) \\ &= \lim_{n\to\infty} P(B_n) \\ &= \lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} \right) \\ &\leq \lim_{n\to\infty} \sum_{m=n}^{\infty} P(E_m). \end{aligned}$$

Example 3.10.9. Define

$$S_{n-1} = \sum_{m=1}^{n-1} P(E_m) = \lim_{n\to\infty} (S_\infty - S_{n-1}) = S_\infty - S_\infty = 0$$

as long as $S_\infty < \infty$.

For a coin, the probability of tails is $P(E_m) = 1/2^m$. Then

$$\sum_{m=1}^{\infty} P(E_m) = \sum_{m=1}^{\infty} 1/2^m = 1 < \infty$$

so $P(\text{"infinitely many tails"}) = 0$. ◇

□

We can show that $X_n \to^{\text{a.s.}} X$ by showing that

$$\sum_{n\in\mathbb{N}} P(|X_n - X| > \varepsilon)$$

converges.

Theorem 3.10.10 (Second-Borel-Cantelli Lemma)**.** Suppose that $E_1, E_2, E_3, \ldots$ are mutually independent and

$$\sum_{n\in\mathbb{N}} P(E_n) = \infty.$$

Then $P(\limsup_{n\to\infty} E_n) = 1$.

Proof. Omitted. See here if you are still curious. □

Chapter 4

Multivariate Distributions

4.1 Week 8: Lecture 1

4.1.1 Joint CDFs and PDFs

Tue 16 Nov 14:00

Recall. Note that

$$F_X : \mathbb{R} \to [0,1] \quad F_{X_1,\dots,X_n} : \mathbb{R}^n \to [0,1].$$

Definition 4.1.1. The **joint cumulative distribution function** of $X_1, \dots, X_n$ is the function

$$F_{X_1,\dots,X_n}(x_1,\dots,x_n) = P(X_1 \le x_1, X_2 \le x_2,\ \dots,\ X_n \le x_n).$$

Note that the commas in the last expression indicate $\cap$.

Bivariate CDFs

Notation. We write a bivariate CDF as

$$F_{X,Y}(x,y) = P(X \le x, Y \le y).$$

Note that

$$\begin{aligned} P(x_1 < X \le x_2,\ y_1 < Y \le y_2) = \quad & F_{X,Y}(x_2,y_2) - F_{X,Y}(x_1,y_2) \\ & -F_{X,Y}(x_2,y_1) + F_{X,Y}(x_1,y_1). \end{aligned}$$

Moreover,

$$F_{X,Y}(-\infty, y) = 0 = F_{X,Y}(x, -\infty)$$

Similarly,

$$F_{X,Y}(\infty,\infty) = 1.$$

Lastly,

$$\begin{aligned} F_{X,Y}(x,\infty) &= \lim_{y\to\infty} F_{X,Y}(x,y) \\ &= P(X \le x, Y \le \infty) \\ &= P(X \le x) \\ &= F_X(x), \end{aligned}$$

which is defined as the marginal CDF of X. Naturally,

$$\lim_{x\to\infty} F_{X,Y}(x,y) = F_Y(y).$$

If X, Y are both discrete, the joint PMF is $f_{X,Y}(x,y) = P(X = x, Y = y)$. So

$$F_{X,Y}(x,y) = \sum_{u\le x}\sum_{v\le y} f_{X,Y}(u,v).$$

Example 4.1.2. Draw 2 cards from a deck of 52 cards. Let X : number of kings drawn, and Y : the number of aces drawn. Note that

$$f_{X,Y}(0,0) = \frac{44}{52}\frac{43}{51} \approx 0.713.$$

We can represent the probabilities of each event using an array:

$x \downarrow y \rightarrow$	0	1	2	$f_X(x)$
0	0.713	0.133	0.004	0.850
1	0.133	0.012	0	0.145
2	0.004	0	0	0.004
$f_Y(y)$	0.85	0.145	0.004	1

Figure 4.1: An array representing the probabilities for values of x and y.

It follows that

$$\sum_x \sum_y f_{X,Y}(x,y) = 1.$$

and that

$$f_X(x) = \sum_y f_{X,Y}(x,y).$$

◇

Definition 4.1.3. Random variables X, Y are jointly continuous if

$$F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(x,y)(u,v)\, du\, dv$$

for all $x, y \in \mathbb{R}$.

So

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

Now, we have

$$\int_{\mathbb{R}^2} f_{X,Y}(x,y)\ dx\ dy = 1,$$

and

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\ dy, \qquad f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)\ dx,$$

and

$$P((X,Y) \in B) = \int \int_B f_{X,Y}(x,y)\ dx\ dy.$$

Remark 4.1.4. Aside: Note that

$$f_X(x) = \sum_y f_{X,Y}(x,y),$$

so

$$f_X(0) = f_{X,Y}(0,0) + f_{X,Y}(0,1) + f_{X,Y}(0,2) + \cdots.$$

4.2 Week 8: Lecture 2

Wed 17 Nov 10:00

Note. Sometimes when you have jointly continuous random variables, you need to be careful about the support.

4.2.1 Bivariate Density

Example 4.2.1 (Bivariate Density)**.**

$$f_{X,Y}(x,y) = \begin{cases} 8xy, & 0 < x < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

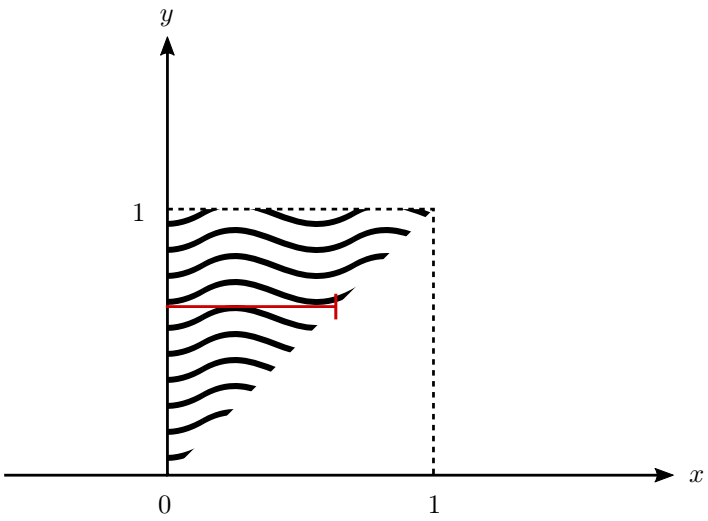


Figure 4.2: The support of Example 4.2.1. The support of a bivariate PDF is an interval in $\mathbb{R}^2$. Finding the limits of integration for each axis can be tricky.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dxdy = \int_0^1\int_0^y 8xy\ dx\,dy = \int_0^1\int_x^1 8xy\ dy\,dx.$$

Note that

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\ dy = \int_x^1 8xy\ dy.$$

◇

$$\mathbb{E}(g(x)) = \begin{cases} \sum_x g(x)f_X(x) & \text{(discrete)} \\ \int_{-\infty}^{\infty} g(x)f_X(x)\ dx & \text{(continuous).} \end{cases}$$

4.2.2 Multiple Random Variables

Recall. If X is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is a well-behaved function, then

Example 4.2.2. $X_1, X_2, \ldots, X_n$: Daily max temperatures. Say $n = 365$. You might want to take the average:

$$\frac{X_1, X_2, \ldots, X_n}{365}.$$

Or the maximum, or the median, etc. These are all functions $g : \mathbb{R} \to \mathbb{R}^n$. If $X_1, X_2, \ldots, X_n$ are random variables, and $g : \mathbb{R}^n \to \mathbb{R}$ is a well-behaved

function.

$$\mathbb{E}(g(X_1, X_2, \ldots, X_n)) =$$
$$\begin{cases} \sum_{x_1} \cdots \sum_{x_n} g(x_1, \ldots, x_n) f(x_1, \ldots x_n) & \text{(discrete)} \\ \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} g(x_1, \ldots, x_n) f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\ dx_1 \cdots dx_n & \text{(continuous).} \end{cases}$$

◇

In previous example:

$$\begin{aligned} \mathbb{E}(X + 2Y) &= \int_{\mathbb{R}^2} x f_{X,Y}(x, y)\ dx\, dy \\ &= \int_0^1 \int_0^y (x + 2y) 8xy\ dx\, dy. \end{aligned}$$

But we can split them!

$$\begin{aligned} \mathbb{E}(X + 2Y) &= \int_{\mathbb{R}^2} x f_{X,Y}(x, y)\ dx\, dy \\ &= 2 \int_{\mathbb{R}^2} y f_{X,Y}(x, y)\ dx\, dy. \end{aligned}$$

4.2.3 Covariance and Correlation

Definition 4.2.3. Let X, Y be random variables. The **covariance** of X and Y is defined as

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}(x))(Y - \mathbb{E}(Y))] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

Some helpful properties of covariance:

- $\mathrm{Cov}(aX, aY) = ab\mathrm{Cov}(X, Y)$
- $\mathrm{Cov}(X + c,\ Y + d) = \mathrm{Cov}(X, Y)$
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(X + Y,\ U + V) = \mathrm{Cov}(X, U) + \mathrm{Cov}(X, V) + \mathrm{Cov}(Y, U) + \mathrm{Cov}(Y, V)$

Definition 4.2.4. Let X, Y be random variables. The **correlation coefficient** of X and Y is

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \rho,$$

with $-1 \leq \rho \leq 1$.

4.3 Week 9: Lecture 1

Tue 23 Nov 14:00

Proposition 4.3.1. Let X, Y be random variables. Then

$$-1 \leq \operatorname{Corr}(X, Y) \leq 1,$$

Moreover, $|\operatorname{Corr}(X, Y)| = 1$ iff $Y = rX + k$, for constants $r \neq 0$ and k.

Proof. Define $Z = Y - rX$, where $r \in \mathbb{R}$. Observe that

$$\begin{aligned} 0 &\leq \operatorname{Var}(Z) \\ &= \operatorname{Var}(Y - rX) \\ &= \operatorname{Var}(Y) + \operatorname{Var}(-rX) + 2\operatorname{Cov}(Y, -rX) \\ &= \operatorname{Var}(Y) + r^2 \operatorname{Var}(X) - 2r \operatorname{Cov}(X, Y). \end{aligned}$$

Let $h(r) = \operatorname{Var}(Y) + r^2 \operatorname{Var}(X) - 2r \operatorname{Cov}(X, Y)$. Note that $h(r)$ is a quadratic equation. Let Δ be the discriminant of $h(r)$. Then

$$\begin{aligned} \Delta &= b^2 - 4ac \\ &= (-2\operatorname{Cov}(X, Y))^2 - 4\operatorname{Var}(X)\operatorname{Var}(Y) \\ &= 4(\operatorname{Cov}(X, Y)^2 - \operatorname{Var}(X)\operatorname{Var}(Y)). \end{aligned}$$

Since $0 \leq h(r)$, $h(r)$ has at most one root. Then $\Delta \leq 0$, and hence

$$\operatorname{Cov}(X, Y)^2 \leq \operatorname{Var}(X)\operatorname{Var}(Y).$$

Thus,

$$\left(\frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} \right)^2 \leq 1,$$

which implies that $-1 \leq \operatorname{Corr}(X, Y) \leq 1$. If $\Delta = 0$, or $\operatorname{Corr}(X, Y)^2 = 1$, then $h(r)$ has a double root, i.e., $h(r^*) = 0$ for some $r^* \in \mathbb{R}$. Moreover,

$$h(r^*) = 0 \iff \operatorname{Var}(Y - r^* X) = 0,$$

so

$$Y - r^* X = k \iff Y = r^* X + k.$$

We can show that $r^* = -\frac{b}{2a} = \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}$.

Now suppose that $Y = rX + k$. Then

$$\begin{aligned}\operatorname{Cov}(X, Y) &= \operatorname{Cov}(X, rX + k) \\ &= r \operatorname{Cov}(X, X) \\ &= r \operatorname{Var}(X) \\ &= \operatorname{Var}(Y) \\ &= \operatorname{Var}(rX + k) \\ &= r^2 \operatorname{Var}(X).\end{aligned}$$

So

$$\begin{aligned}\operatorname{Corr}(X, Y) &= \frac{r \operatorname{Var}(X)}{\sqrt{\operatorname{Var}(X) r^2 \operatorname{Var}(X)}} \\ &= \frac{r}{\sqrt{r^2}} \\ &= \frac{r}{|r|} \\ &= \begin{cases} 1, & \text{if } r > 0 \\ -1, & \text{if } r < 0. \end{cases}\end{aligned}$$

□

4.3.1 Joint Moments

Definition 4.3.2. If X, Y are random variables, the $(r, s)^{\text{th}}$ **joint moment** of X and Y is

$$\mu'_{r,s} = \mathbb{E}(X^r Y^s).$$

Definition 4.3.3. The $(r, s)^{\text{th}}$ **joint central moment** is

$$\mu_{r,s} = \mathbb{E}[(X - \mathbb{E}(X))^r (Y - \mathbb{E}(Y))^s].$$

Example 4.3.4. Note that

$$\begin{aligned}\mu'_{1,0} &= \mathbb{E}(X) \\ \mu'_{r,0} &= \mathbb{E}(X^r) \\ \mu'_{0,3} &= \mathbb{E}(Y^3).\end{aligned}$$

◇

Example 4.3.5. Note that

$$\operatorname{Corr}(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X) \operatorname{Var}(Y)}} = \frac{\mu_{1,1}}{\sqrt{\mu_{2,0}\mu_{0,2}}}.$$

◇

Example 4.3.6. Let

$$f_{X,Y}(x,y) = \begin{cases} x+y, & 0 \le x,\, y \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} \mu'_{r,s} &= \mathbb{E}(X^r Y^s) \\ &= \int_{\mathbb{R}^2} x^r y^s f_{X,Y}(x,y)\; dx\, dy \\ &= \int_0^1 \int_0^1 x^r y^s (x+y)\; dx\, dy \\ &= \int_0^1 \int_0^1 (x^{r+1} y^s + x^r y^{s+1})\; dx\, dy \\ &= \dots \end{aligned}$$

◇

4.3.2 Joint MGFs

Definition 4.3.7. The **joint MGF** of X and Y is

$$\begin{aligned} M_{X,Y}(t,u) &= \mathbb{E}(e^{tX+uY}) \\ &= \mathbb{E}(e^{tX} e^{uY}) \\ &= \mathbb{E}\left[\left(\sum_{i \in \mathbb{N}_0} \frac{(tX)^i}{i!} \right) \left(\sum_{j \in \mathbb{N}_0} \frac{(uY)^j}{j!} \right) \right] \\ &= \mathbb{E}\left[\sum_{i \in \mathbb{N}_0} \sum_{j \in \mathbb{N}_0} X^i Y^j \frac{t^i u^j}{i!j!} \right] \\ &= \sum_{i \in \mathbb{N}_0} \sum_{j \in \mathbb{N}_0} \mathbb{E}(X^i Y^j) \frac{t^i u^j}{i!j!}. \end{aligned}$$

Note that $\mathbb{E}(X^i Y^j) = \mu_{i,j}$.

The $(r,s)^{\text{th}}$ joint moment of X, Y is the coefficient of $\frac{t^r u^s}{r!s!}$ in the power series expansion of $M_{X,Y}(t,u)$. Moreover,

$$\begin{aligned} M^{(r,s)}_{X,Y}(0,0) &= \frac{\partial^{r+s}}{\partial t^r \partial u^s} M_{X,Y}(t,u) \Big|_{t=0,\, u=0} \\ &= \mu'_{r,s} \\ &= \mathbb{E}(X^r Y^s). \end{aligned}$$

4.3.3 Joint CGFs

Definition 4.3.8. Define

$$K_{X,Y}(t,u) = \log M_{X,Y}(t,u).$$

Then $K_{X,Y}(t,u)$ is the **joint cumulant generating function** of X and Y. Let

$$K_{X,Y}(t,u) = \sum_{i\in\mathbb{N}_0}\sum_{j\in\mathbb{N}_0} \kappa_{i,j}\frac{t^i u^j}{i!j!}.$$

Then $\kappa_{i,j}$ is the $(i,j)^{\text{th}}$ **joint cumulant**.

Example 4.3.9. Let X, Y be random variables. Then $\kappa_{1,1} = \text{Cov}(X,Y)$. ◇

Proof. Observe that

$$\begin{aligned} &M_{X,Y}(t,u) = 1 + \mu'_{1,0}t + \mu'_{0,1}u + \mu'_{1,1}tu + \cdots \\ &\Rightarrow K_{X,Y}(t,u) = \log M_{X,Y}(t,u). \end{aligned}$$

This implies that

$$\begin{aligned} \frac{\partial}{\partial t}K_{X,Y}(t,u) &= \frac{\frac{\partial}{\partial t}M_{X,Y}(t,u)}{M_{X,Y}(t,u)} = \frac{\mu'_{1,0} + \mu'_{1,1}u + \cdots}{M_{X,Y}(t,u)} \\ \Rightarrow \frac{\partial^2}{\partial u\,\partial t}K_{X,Y}(t,u) &= \frac{\frac{\partial}{\partial t}M_{X,Y}(t,u)}{M_{X,Y}(t,u)} \\ &= \frac{\mu'_{1,0} + \mu'_{1,1}u + \cdots}{M_{X,Y}(t,u)} - \frac{(\mu'_{1,0} + \mu'_{1,1}u + \cdots)(\mu'_{0,1} + \cdots)}{(M_{X,Y}(t,u))^2} \end{aligned}$$

Thus,

$$\begin{aligned} \kappa_{1,1} = K^{(1,1)}_{X,Y}(0,0) &= \mu'_{1,1} - \mu'_{1,0}\mu'_{0,1} \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\ &= \text{Cov}(X,Y). \end{aligned}$$

□

By Example 4.3.9, we can write

$$\text{Corr}(X,Y) = \frac{\kappa_{1,1}}{\sqrt{\kappa_{2,0}\kappa_{0,2}}}.$$

4.4 Week 9: Lecture 2

4.4.1 Independent Random Variables

Wed 24 Nov 10:00

Definition 4.4.1. Two random variables X and Y are independent ($X \perp\!\!\!\perp Y$) iff $\{X \leq x\}$ and $\{Y \leq y\}$ are independent events for all $x, y \in \mathbb{R}$, i.e.:

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y).$$

If X, Y are independent and jointly continuous, then

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

If (X, Y) are independent and discrete, then

$$\begin{aligned} f_{X,Y}(x,y) &= P(X = x, Y = y) \\ &= P(X = x)P(Y = y) \\ &= f_X(x)f_Y(y). \end{aligned}$$

Let X, Y be jointly continuous. If $X \perp\!\!\!\perp Y$ then

$$\begin{aligned} \mathbb{E}(X,Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)\ dx\ dy \\ &= \int_{-\infty}^{\infty} xf_X(x)\ dx \int_{-\infty}^{\infty} yf_Y(y)\ dy \\ &= \mathbb{E}(XY) \\ &= \mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

Hence, $X \perp\!\!\!\perp Y \Rightarrow X, Y$ are uncorrelated, i.e., $\mathrm{Cov}(X,Y) = 0$.

Proposition 4.4.2. If $X \perp\!\!\!\perp Y$ and $g, h : \mathbb{R} \to \mathbb{R}$ are well-behaved functions, then $g(X) \perp\!\!\!\perp h(Y)$ and $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$.

Proof. Omitted. Left as an exercise. □

Example 4.4.3. For random variables X, Y with $X \perp\!\!\!\perp Y$,

$$\begin{aligned} M_{X,Y}(t,u) &= \mathbb{E}(e^{tx}e^{uY}) \\ &= \mathbb{E}(e^{tx})\mathbb{E}(e^{uY}) \\ &= M_X(t)M_Y(u), \end{aligned}$$

and thus, $K_{X,Y} = K_X(t) + K_Y(t)$.

◇

Example 4.4.4. Let X, Y be continuous random variables with joint density

$$f_{X,Y}(x,y) = \begin{cases} x+y, & 0 < x,\, y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Note that

$$f_X(x) = \ldots = x + 1/2, \quad 0 < x < 1$$
$$f_Y(y) = \ldots = y + 1/2, \quad 0 < y < 1,$$

and thus,

$$f_{X,Y}(x,y) \neq f_X(x) f_Y(y),$$

so $X \not\perp\!\!\!\perp Y$. ◇

Example 4.4.5. Let

$$f_{X,Y}(x,y) = \begin{cases} kxy, & 0 < x < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Two functions that don't have the same support cannot be the same function. Hence, $X \not\perp\!\!\!\perp Y$ because of the support. ◇

Notation. We write that $X_1, X_2, \ldots, X_n$ are independent iff $\{X_1 \leq x_1\}, \ldots, \{X_n \leq x_n\}$ are *mutually independent*. Hence,

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{u} F_{X_i}(x_i).$$

Also

$$\mathbb{E}(X_1 X_2 \cdots X_n) = \mathbb{E}(X_1) \cdots \mathbb{E}(X_n).$$

4.4.2 Random Vectors & Random Matrices

Definition 4.4.6. We say that $\mathbf{X}$ is a **random vector** if

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

for random variables (X_i). We say that $\mathbf{W}$ is a **random matrix** if

$$\mathbf{W} = \begin{pmatrix} W_{1,1} & \cdots & W_{1,n} \\ \vdots & \ddots & \cdots \\ W_{m,1} & \cdots & W_{m,n} \end{pmatrix}$$

for random variables $(W_{i,j})$.

Let $\mathbf{X} = (X_1, \ldots, X_n)^T, \quad \mathbf{x} = (x_1, \ldots, x_n)^T$. So

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X_1,\ldots X_n}(x_1, \ldots, x_n).$$

And similarly for $f_{\mathbf{X}}(\mathbf{x})$ and $M_{\mathbf{x}}(\mathbf{t})$. The expectation of a random vector X is given by

$$\mathbb{E}(\mathbf{X}) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix},$$

and the expectation of a random matrix $\mathbf{W}$ is given by

$$\mathbb{E}(\mathbf{W}) = \begin{pmatrix} \mathbb{E}(W_{1,1}) & \ldots & \mathbb{E}(W_{1,n}) \\ \vdots & \ddots & \vdots \\ \mathbb{E}(W_{m,1}) & \ldots & \mathbb{E}(W_{m,n}) \end{pmatrix}$$

What is the variance of a random vector?

Recall. For a random variable X,

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}_n} \mathbf{g}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x}.$$

Then

$$\begin{aligned} \mathrm{Var}(\mathbf{X}) &= \mathbb{E}[(\mathbf{X} - \mathbb{E}(\mathbf{X})(\mathbf{X} - \mathbb{E}(\mathbf{X}))^T] \\ &= \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \mathrm{Cov}(X_1, X_3) \ldots \\ \mathrm{Cov}(X_2, X_1) & \ddots & \vdots \\ \vdots & \ldots & \mathrm{Var}(X_n) \end{pmatrix}. \end{aligned}$$

Note that this is a symmetric $n \times n$ matrix. If $X_1, \ldots X_n$ are independent & identically distributes (IID), or $F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^n f_{X_1}(x_i)$, then $\mathrm{Var}(\mathbf{X}) = \sigma^2 \mathbf{I}_n$ where $\sigma^2 = \mathrm{Var}(\mathbf{X}_1)$.

Definition 4.4.7. An $n \times n$ matrix $\mathbf{A}$ is **positive semidefinite** (or **non-negative definite**) if, for any $\mathbf{b} \in \mathbb{R}^n$ if it holds that

$$\mathbf{b}^T \mathbf{A} \mathbf{b} \geq 0.$$

Let $\mathbf{X}$ be an $n \times 1$ random vector and let $\mathbf{b} \in \mathbb{R}^n$ (vector of constants). Then

$$\begin{aligned} 0 \leq \mathrm{Var}(\underbrace{\mathbf{b}^T \mathbf{X}}_{n \text{ scalar } (1\times 1)}) &= \mathbb{E}[(\mathbf{b}^T \mathbf{X} - \mathbb{E}(\mathbf{b}^T \mathbf{X}))(\ldots)^T] \\ &= \mathbb{E}[\mathbf{b}^T (\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X})^T \mathbf{b}] \\ &= \mathbf{b}^T \mathbb{E}[(\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))^T] \mathbf{b} \\ &= \mathrm{Var}(\mathbf{X}). \end{aligned}$$

We often write $\mathrm{Var}(\mathbf{X}) \geq 0$ in place of writing that the variance matrix is positive semidefinite.

4.4.3 Transformations of Random Variables

Recall. Univariate case: Let X be a random variable, and $Y = g(x)$ where $g : \mathbb{R} \to \mathbb{R}$ is monotonic. Then

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|,$$

where $x = g^{-1}(y)$.

Remark 4.4.8. We now want to transform (U, V) into (X, Y). We have

$$X = g_1(U, V)$$
$$Y = g_2(U, V),$$

where $(X, Y) = \mathbf{g}(U, V)$. Assume that the transformation is a bijective function. The inverse is $(U, V) = \mathbf{h}(X, Y) = \mathbf{g}^{-1}(X, Y)$. Then

$$f_{X,Y}(x, y) = f_{U,V}(u, v)|J_{\mathbf{h}}(x, y)|,$$

where $J_{\mathbf{h}}(x, y)$ is the Jacobian of $\mathbf{h}$.

4.5 Week 10: Lecture 1

4.5.1 Sums of Random Variables

Tue 30 Nov 14:00

Recall. For random variables X, Y,

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

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

$$\mathbb{E}[(X + Y)^r] = \sum_{j=0}^{r} \binom{r}{j} \mathbb{E}(X^j Y^{r-j}) = \sum_{j=0}^{r} \binom{r}{j} \mu'_{j,r-j}.$$

Proposition 4.5.1. If $Z = X + Y$, then

$$f_Z(z) = \begin{cases} \sum_u f_{X,Y}(u, z - u), & \text{(discrete)}, \\ \int_{\mathbb{R}} f_{X,Y}(u, z - u)du & \text{(continuous)}. \end{cases}$$

Proof. For the discrete case, note that

$$\begin{aligned} f_Z(z) &= P(Z = z) \\ &= P(X + Y = z) \\ &= \sum_u P(X = u,\ Y = z - u) \\ &= \sum_u f_{X,Y}(u, z - u). \end{aligned}$$

By the Law of Total Probability,

$$\{X + Y = Z\} = \bigcup_u \{X = u,\ Y = Z - U\}.$$

For the continuous case, note that

$$Z = X + Y,\ U = X \iff X = U,\ Y = Z - U.$$

Let

$$(Z, U) = \boldsymbol{\partial}(X, Y), \quad (X, Y) = \mathbf{h}(U, Z).$$

Then

$$\begin{aligned} J_{\mathbf{h}}(x, y) &= \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial z} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial z} \end{vmatrix} \\ &= \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix}. \end{aligned}$$

Then

$$f_{U,Z}(u, z) = f_{X,Y}(u, z - u)|J_{\mathbf{h}}| = 1,$$

which implies that

$$f_Z(z) = \int_{\mathbb{R}} f_{U,Z}(u, z)\ du = \int_{\mathbb{R}} f_{X,Y}(u, z - u)\ du$$

□

Definition 4.5.2. Let f and g be functions. The **convolution** of f and g is

$$\int_{\mathbb{R}} f(x)g(y - x)\ dy.$$

Notation. If f and g are functions, their convolution is denoted by $f * g$.

Remark 4.5.3. If $X \perp\!\!\!\perp Y$, then

$$f_Z(z) = \begin{cases} \sum_u f_X(u) f_Y(z-u) & \text{(discrete)}, \\ \int_{\mathbb{R}} f_X(u) f_Y(z-u)\ du & \text{(continuous)}. \end{cases}$$

Hence,

$$f_Z = f_X * f_Y = f_Y * f_X.$$

Assume $X \perp\!\!\!\perp Y$. To work out the distribution of $Z = X + Y$, either work out the convolution of f_X, f_Y, or use their MGFs/CGFs:

$$M_Z(t) = M_X(t) M_Y(t) \iff K_Z(t) = K_X(t) + K_Y(t).$$

Example 4.5.4.

$$X \sim N(\mu_X, \sigma_x^2),\ Y \sim N(\mu_Y(\sigma_Y^2), \quad \text{with} \quad X \perp\!\!\!\perp Y,\ Z = X + Y$$

Recall that $K_x(t) = \mu_X t + \sigma_X^2 \frac{t^2}{2}$, so

$$\begin{aligned} K_Z(t) &= K_X(t) + K_Y(t) \\ &= \mu_x t + \sigma_x^2 \frac{t^2}{2} + \mu_Y t + \sigma_Y^2 \frac{t^2}{2} \\ &= (\mu_X + \mu_Y) t + (\sigma_X^2 + \sigma_Y^2) \frac{t^2}{2} \\ \Rightarrow Z &\sim \text{Normal}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2). \end{aligned}$$

◇

Example 4.5.5. Let

$$X \sim \text{Exp}(\lambda),\ Y \sim \text{Exp}(\theta), \qquad X \perp\!\!\!\perp Y,\ Z = X + Y$$

Observe that

$$\begin{aligned} f_Z(z) &= \int_{\mathbb{R}} f_X(u) f_Y(z-u)\ du \\ &= \int_0^z \lambda e^{-\lambda u} \theta e^{-\lambda(z-u)}\ du \\ &= \lambda\theta e^{-\theta z} \left[-\frac{1}{\lambda - \theta} e^{-(\lambda\theta)u} \right]_0^z \\ &= \frac{\lambda\theta}{\lambda - \theta} e^{-\theta z} (1 - e^{-(\lambda-\theta)z} \\ &= \frac{\lambda\theta(e^{-\theta z} - e^{-\lambda z})}{\lambda - \theta}, \quad \text{for } z > 0,\ \lambda \neq 0 \end{aligned}$$

◇

Example 4.5.6. Consider

$$X_1, X_2, \ldots, X_n.$$

Let $S = \sum_{i=1}^{n} X_i$. Suppose that (X_i) are mutually independent. Then

$$f_S = f_{X_1} * f_{X_2} * \ldots f_{X_n}, \quad M_S(t) = \prod_{i=1}^{n} M_{X_i}(t).$$

◇

If $X_1, \ldots, X_n$ are IID (identically distributed), then

$$M_S(t) = \prod_{i=1}^{n} M_{X_i}(t) = (M_{X_1}(t))^n,$$

which implies that $K_S(t) = nK_{X_1}(t)$.

Example 4.5.7. If $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, then

$$M_S(t) = (M_{X_1}(t))^n = (1 - p + pe^t)^n,$$

which implies that $S \sim \text{Bin}(n, p)$. ◇

4.5.2 Multivariate Normal Distributions

Bivariate Normal Distribution

How can we derive a bivariate normal distribution? Starting point: take $U, V \sim \text{Normal}(0, 1)$, with $U \perp\!\!\!\perp V$. Then

$$f_{U,V}(u, v) = f_U(u) f_V(v) = \frac{1}{2\pi} e^{-(u^2+v^2)/2}, \quad u, v \in \mathbb{R}.$$

Moreover,

$$M_{U,V}(s, t) = e^{(s^2+t^2)/2}, \quad s, t \in \mathbb{R}$$

Let $U, V \sim^i N(0, 1)$. Define

$$X = U, Y = \rho U + \sqrt{1 - \rho^2} V, \quad \text{where } |\rho| \leq 1.$$

Some quick properties of the bivariate standard normal:

(1) $X \sim N(0, 1)$ by definition. Moreover, Y is normal, as it is a sum of independent Normals:

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

and thus,

$$\begin{aligned} \text{Var}(Y) &= \rho^2 \text{Var}(U) + (\sqrt{1 - \rho^2})^2 \text{Var}(V) \\ &= \rho^2 + 1 - \rho^2 = 1, \end{aligned}$$

which implies that $Y \sim \text{Normal}(0, 1)$.

(2) $\mathrm{Corr}(X,Y) = \rho$. Observe that

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{Cov}(U, \rho U + \sqrt{1-\rho^2} \\ &= \mathrm{Cov}(U,\rho U) + \mathrm{Cov}(U, \sqrt{1-\rho^2}V) \\ &= \rho\,\mathrm{Cov}(U,U) \\ &= \rho.\end{aligned}$$

Thus,

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \rho$$

(3) Any linear combination of X and Y is normally distributed:

$$\begin{aligned}aX + bY + c &= aU + b(\rho + \sqrt{1-\rho^2}V) + c \\ &= (a+b\rho)u + b\sqrt{1-\rho^2}V + c.\end{aligned}$$

which is Normal, as $U \perp\!\!\!\perp V$.

Example 4.5.8. Let $U, V \sim^i \mathrm{Normal}(0,1)$, and

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

We have

$$\begin{aligned}f_{X,Y}(x,y) &= \ldots (\text{try this!}) \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right), \quad x,y \in \mathbb{R} \\ &= f_{U,V}(u,v)|J_{\mathbf{u}}(x,y)|, \quad x,y \in \mathbb{R}.\end{aligned}$$

◇

Example 4.5.9. Let $U, V \sim^i \mathrm{Normal}(0,1)$, and

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

Then

$$K_{X,Y}(t,u) = \frac{1}{2}(s^2 + 2\rho st + t^2).$$

◇

Proof. Try this! □

Finally, to obtain the bivariate normal from X and Y, we take

$$X^* = \mu_X + \sigma_X X, \quad Y^* = \mu_Y + \sigma_Y Y.$$

Then $X^* \sim N(\mu_X, \sigma_X)$, and $Y^* \sim N(\mu_Y, \sigma_Y)$, and $\mathrm{Corr}(X^*, Y^*) = \rho$.

Chapter 5

Conditional Distributions

5.1 Week 10: Lecture 2

5.1.1 Another Deck of Cards

Wed 1 Dec 10:00

Example 5.1.1. Draw 2 cards from full deck. Define Y : # of aces, X : # of kings (see Figure 4.1).

$$\begin{aligned} P(\text{one Ace} \mid \text{one King}) &= P(Y = 1 \mid X = 1) \\ &= \frac{P(Y = 1, X = 1)}{P(X = 1)} \\ &= \frac{f_{X,Y}(1, 1)}{f_X(1)}. \end{aligned}$$

◇

5.1.2 Conditional Mass and Density

In general,

$$P(Y = y \mid X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

Definition 5.1.2. The **conditional probability mass function** of Y given $X = x$ is

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

Question: Does it sum to 1?

$$\begin{aligned}\sum_y f_{Y|X}(y \mid X) &= \sum_y \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \frac{1}{f_X(x)} \sum_y f_{X,Y}(x,y) \\ &= \frac{f_X(x)}{f_X(x)} \\ &= 1. \quad \checkmark\end{aligned}$$

Definition 5.1.3. The **conditional cumulative distribution function** of Y given $X = x$ is

$$F_{Y|X}(y \mid x) = \sum_{u \le y} f_{Y|X}(u \mid x)$$

Definition 5.1.4. If X, Y are jointly continuous, we define the **conditional probability density function** of Y given $X = x$ as

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

Example 5.1.5. Let X, Y be jointly continuous random variables with

$$f_{X,Y}(x,y) = \begin{cases} 8xy, & 0 < x < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Recall that

$$f_X(x) = 4x(1 - x^2), \quad 0 < x < 1.$$

Then

$$f_{Y|X}(y \mid x) = \frac{8xy}{4x(1-x^2)} = \frac{2y}{1-x^2}, \quad x < y < 1.$$

Furthermore,

$$\begin{aligned}F_{Y|X}(y \mid x) &= \int_{-\infty}^{y} f_{Y|X}(u \mid x)\, du \\ &= \int_x^y \frac{2u}{1-x^2}\, du \\ &= \frac{y^2 - x^2}{1-x^2}, \quad x < y < 1.\end{aligned}$$

Plug in $y = x$ to check if this is plausible.

Recall. $P(A \cap B \cap C) = P(A \mid B \cap C)P(B \mid C)P(C)$. Similarly,

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

and

$$f_{X,Y,Z}(x,y,z) = f_{Z|X,Y}(z \mid x,y) f_{Y|X}(y \mid x) f_X(x).$$

◇

A Simple Model

Example 5.1.6. X : # of hurricanes formed, Y : # of hurricanes making landfall

Suppose that $X \sim \text{Poisson}(\lambda)$ and $(Y \mid X = x) \sim \text{Bin}(x,p)$. Then

$$f_{X,Y}(x,y) = f_{Y|X}(y \mid x) f_X(x) = \binom{x}{y} p^y (1-p)^{x-y} \frac{e^{-\lambda}\lambda^x}{x!}.$$

supported by $x, y = 0, 1, 2, \ldots$, $y \leq x$. Then

$$\begin{aligned} f_Y(y) &= \sum_x f_{X,Y}(x,y) \\ &= \sum_{x=y}^{\infty} \frac{x!}{y!(x-y)!} p^y (1-p)^{x-y} \frac{e^{-\lambda}\lambda x}{x!} \\ &= \frac{e^{-\lambda}p^y}{y!} \sum_{x=y}^{\infty} \frac{(1-p)^{x-y}\lambda^x}{(x-y)!}. \end{aligned}$$

Let $z = x - y$. Then

$$\begin{aligned} f_Y(y) &= \frac{e^{-\lambda}p^y}{y!}\lambda^y \sum_{z=0}^{\infty} \frac{(1-p)^z \lambda^z}{z!} \\ &= \frac{e^{-\lambda}(\lambda p^y)}{y!} e^{\lambda(1-p)} \\ &= \frac{e^{-\lambda p}(\lambda p)^y}{y!} \\ &= \frac{e^{-\lambda p}(\lambda p)^y}{y!}, \quad y = 0, 1, 2, \ldots \end{aligned}$$

This implies that $Y \sim \text{Poisson}(\lambda p)$. ◇

In general, if X is discrete and Y is continuous,

$$\underbrace{f_{X,Y}(x,y)}_{\text{joint mass / density}} = \underbrace{f_{Y|X}(y \mid x)}_{\text{conditional density}} \times \underbrace{f_X(x)}_{\text{marginal mass}}$$

Moreover,

$$\int_{\mathbb{R}} \sum_x f_{X,Y}(x,y)\, dy = 1.$$

Insurance Example

Example 5.1.7. Define Z : total value of claims, Y : # of claims submitted, and X : average # of claims. Suppose that

$$X \sim \Gamma(\alpha, \lambda), \quad (Y \mid X = x) \sim \text{Poisson}(x).$$

Then

$$(Z \mid Y = y) \sim \text{ some continuous model.}$$

◇

5.2 Week 11: Lecture 1

5.2.1 Conditional Expectation

Tue 7 Dec 14:00

Definition 5.2.1. The **conditional expectation** of Y given X is $\mathbb{E}(Y \mid X) = \psi(X)$.

Example 5.2.2 (Hurricanes)**.**

$$(Y \mid X = x) \sim \text{Bin}(x, p),$$

so $\mathbb{E}(Y \mid X = x) = xp \Rightarrow \mathbb{E}(Xp)$.

Important difference: $\mathbb{E}(Y \mid X)$ gives a random variable, $\mathbb{E}(Y \mid X = x)$ gives $\psi(x)$. ◇

5.2.2 Law of Iterated Expecations

Proposition 5.2.3. For random variables X and Y, we have

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

Proof.

$$\begin{aligned}
\mathbb{E}[\mathbb{E}(Y \mid X)] &= \mathbb{E}[\psi(X)] \\
&= \int_{\mathbb{R}} \psi(x) f_X(x)\ dx \\
&= \int_{\mathbb{R}} \mathbb{E}(Y \mid X = x) f_X(x)\ dx \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y f_{Y|X}(y \mid x)\ dy \right) f_X(x)\ dx \\
&= \int_{\mathbb{R}^2} y f_{Y|X}(y \mid x) f_X(x)\ dy\ dx \\
&= \int_{\mathbb{R}^2} y f_{X,Y}(x, y)\ dy\ \ dx \\
&= \mathbb{E}(Y).
\end{aligned}$$

□

Example 5.2.4 (More Hurricanes)**.** $\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y \mid X)] = \mathbb{E}(Xp) = \lambda p$. Then $X \sim \text{Poisson}(\lambda)$, $(Y \mid X = x) \sim \text{Bin}(x, p)$. ◇

Note that the Law of Iterated Expectations is conceptually similar to The Law of Total Probability, which states

$$P(A) = \sum_{i \in \mathbb{N}} P(A \mid B_i) P(B_i).$$

Example 5.2.5. Let X, Y be random variables with joint density

$$f_{X,Y}(x, y) = \begin{cases} xe^{-xy}e^{-x}, & x, y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Find $\mathbb{E}(Y \mid X)$:

$$\begin{aligned}
\int_{\mathbb{R}} f_{X,Y}(x, y)\ dy &= \int_0^\infty xe^{-xy}e^{-x}\ dy \\
&= [e^{-xy}e^{-x}]_{y=0}^{y \to \infty} \\
&= e^{-x}, \quad x > 0.
\end{aligned}$$

Hence, $X \sim \text{Exp}(1)$. It is often helpful to write this explicitly. Now,

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{xe^{-xy}e^{-x}}{e^{-x}} = xe^{-xy}.$$

Hence, $(Y \mid X = x) \sim \text{Exp}(x)$. This implies that

$$\mathbb{E}(Y \mid X = x) = \frac{1}{x} \quad \text{and} \quad \mathbb{E}(Y \mid X) = \frac{1}{X}.$$

◇

If $g : \mathbb{R} \to \mathbb{R}$ is a well-behaved function, and we define

$$h(x) = \mathbb{E}[g(Y) \mid X = x] = \begin{cases} \sum_y g(y) f_{Y|X}(y \mid x) & \text{(discrete)} \\ \int_{\mathbb{R}} g(y) f_{Y|X}(y \mid x)\, dy & \text{(continuous)}, \end{cases}$$

then the conditional expectation of $g(Y)$ given X is $\mathbb{E}(g(Y) \mid X) = h(X)$.

5.2.3 Properties of Conditional Expectation

For any two random variables X and Y,

- $\mathbb{E}((aX + b) \mid Y) = a\mathbb{E}(X \mid Y) + b$,
- $E(XY \mid X) = X\mathbb{E}(Y \mid X)$.
- Think about: $\mathbb{E}(XY \mid X = x) = \mathbb{E}(xY \mid X = x) = x\mathbb{E}(Y \mid X = x)$.
- $\mathbb{E}[\mathbb{E}(X \mid Y)Y \mid X] = \mathbb{E}(Y \mid X)\mathbb{E}(Y \mid X) = \mathbb{E}(Y \mid X)^2$, since $\mathbb{E}(Y \mid X)$ is a function of X.

Definition 5.2.6. The rth **conditional moment** of Y given X is $\mathbb{E}(Y^r \mid X)$, and the rth conditional central moment is

$$\mathbb{E}[(Y - \mathbb{E}(Y \mid X))^r \mid X].$$

Example 5.2.7 (Conditional Variance)**.** Let X, Y be random variables. Then

$$\begin{aligned} \mathrm{Var}(Y \mid X) &= \mathbb{E}[(Y - \mathbb{E}(Y \mid X))^2 | X] \\ &= \mathbb{E}(Y^2 \mid X) - \mathbb{E}(Y \mid X)^2. \end{aligned}$$

◇

Proof. Prove this! □

5.2.4 Law of Iterated Variance

Proposition 5.2.8. Let X and Y be random variables. Then

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

Proof. We have

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \\ &= \mathbb{E}[\mathbb{E}(Y^2 \mid X)] - (\mathbb{E}[\mathbb{E}(Y \mid X)])^2 \\ &= \mathbb{E}[\mathrm{Var}(Y \mid X) + \underbrace{\mathbb{E}(Y \mid X)^2}_{\psi(X)^2}] - \mathbb{E}[\underbrace{\mathbb{E}(Y \mid X)}_{\psi(X)}]^2 \\ &= \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathbb{E}[\psi(X)^2] - (\mathbb{E}(\psi(X)))^2 \\ &= \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}[\mathbb{E}(Y \mid X)].\end{aligned}$$

□

Hurricanes Again

Example 5.2.9. Let $(Y \mid X = x) \sim \mathrm{Bin}(x, p)$, and $X \sim \mathrm{Poisson}(\lambda)$. Then

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}[\mathbb{E}(Y \mid X)] \\ &= \mathbb{E}(Xp(1-p)) + \mathrm{Var}(Xp) \\ &= \lambda p(1-p) + \lambda p^2 \\ &= \lambda p.\end{aligned}$$

◇

5.3 Week 11: Lecture 2

5.3.1 Conditional Moment Generating Function

Wed 8 Dec 10:00

Definition 5.3.1. If

$$M_{Y|X}(u \mid v) = \mathbb{E}[e^{uY} | X = x] = \phi(u, x),$$

then the **conditional moment generating function** is $\phi(u, X)$. Hence,

$$M_{Y|X}(u \mid X) = \mathbb{E}[e^{uY} \mid X].$$

Observe that by the Law of Iterated Expecations,

$$M_Y(u) = \mathbb{E}[M_{Y|X}(u \mid x)] = \mathbb{E}(e^{uY}).$$

Example 5.3.2. Let

$$X \sim \mathrm{Poisson}(\lambda), \quad (Y \mid X = x) \sim \mathrm{Bin}(x, p).$$

Then

$$M_{Y|X}(y \mid x) = (1 - p + pe^u) \Rightarrow M_{Y|X}(u \mid X) = (1 - p + pe^u)^X.$$

So

$$\begin{aligned} M_Y(u) &= \mathbb{E}[M_{Y|X}(u \mid X)] \\ &= \mathbb{E}[(1-p+pe^u)^X] \\ &= \mathbb{E}[e^{X\ln(1-p+pe^u)}] \\ &= M_X(\ln(1-p+pe^u)) \\ &= \exp(\lambda(e^{\ln(1-p+pe^u)}-1) \\ &= e^{\lambda p(e^u-1)}, \end{aligned}$$

so $Y \sim \text{Poisson}(\lambda p)$.

Remark 5.3.3. Aside: $M_X(t) = e^{\lambda(e^t-1)}$.

Thus,

$$\begin{aligned} M_{X,Y}(t,u) &= \mathbb{E}[e^{tX}e^{uY}] \\ &= \mathbb{E}[\mathbb{E}[e^{tX}e^{uY} \mid X]] \\ &= \mathbb{E}[e^{tX}\mathbb{E}[e^{uY} \mid X]] \\ &= \mathbb{E}[e^{tX}M_{Y|X}(u \mid X)]. \end{aligned}$$

◇

5.3.2 Some Practical Applications

Example 5.3.4 (Height)**.** Suppose that you know the mean height and variance of the male and female population. Let

$$\begin{aligned} X &: \text{ height of a student,} \\ W &: \text{ male or female} \quad (\text{male} = 0,\ \text{female} = 1). \end{aligned}$$

Then

$$\begin{aligned} (X \mid W = 1) &\sim \text{Normal}(\mu_W, \sigma_W^2) \\ (X \mid W = 0) &\sim \text{Normal}(\mu_M, \sigma_M^2) \\ W &\sim \text{Bernoulli}(p). \end{aligned}$$

Moreover,

$$\begin{aligned} f_X(x) &= \sum_w \underbrace{f_{X|W}(x \mid w) f_W(w)}_{f_{X,W}(x,w)} \\ &= p(f_{X|W}(x \mid 1) + (1-p) f_{X|W}(x \mid 0). \end{aligned}$$

Note that

$$f_{X|W}(x \mid 1) = \frac{1}{\sqrt{2\pi\sigma^2 w}} e^{-\frac{(x-\mu_w)^2}{2\sigma^2 w}}.$$

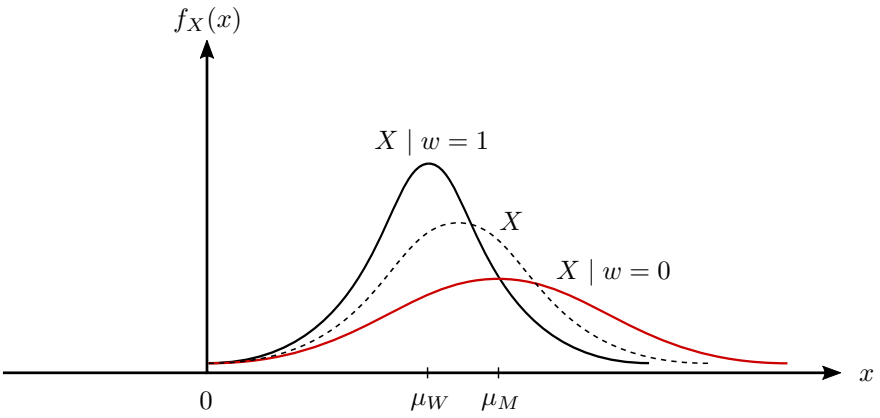


Figure 5.1: The distribution of X from Example 5.3.4 is somewhere in between the distributions of $X \mid w = 1$ and $X \mid w = 0$.

◇

Example 5.3.5 (Household Insurance)**.**

$\text{Exp}(\lambda)$:	amount claimed each year
Normal $\sim \text{Geo}(p)$:	years policy is held
Y :	total amount claimed

Then

$$Y = \sum_{i=1}^{N} X_i, \quad \text{a random sum.}$$

We assume that N is independent of X_i. This is a basic assumption that may or may not be reasonable, depending on the situation. We further assume that $N \geq 0$, with $Y = 0$ if $N = 0$. Then

$$\begin{aligned} \mathbb{E}(Y \mid N = n) &= \mathbb{E}\left(\sum_{i=1}^{n} X_i)\right) \\ &= n\mathbb{E}(X_1). \end{aligned}$$

So

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}[\mathbb{E}(Y \mid N)] \\ &= \mathbb{E}[N\mathbb{E}(X_1)] \\ &= \mathbb{E}(X_1)\mathbb{E}(N). \end{aligned}$$

Moreover,

$$\begin{aligned}\operatorname{Var}(Y \mid N = n) &= \operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= n \operatorname{Var}(X_1).\end{aligned}$$

Now, how do we iterate variances? We use the Law of Iterated Variance:

$$\begin{aligned}\operatorname{Var}(Y) &= \mathbb{E}[\operatorname{Var}(Y \mid N)] + \operatorname{Var}(\mathbb{E}(Y \mid N)] \\ &= \mathbb{E}[N \operatorname{Var}(X_1)] + \operatorname{Var}(N\mathbb{E}(X_1)) \\ &= \operatorname{Var}(X_1)\mathbb{E}(N) + \mathbb{E}(X_1)^2 \operatorname{Var}(N).\end{aligned}$$

Moreover,

$$\begin{aligned}M_{Y|N}(u \mid n) &= \mathbb{E}(e^{uY} \mid N = n) \\ &= \mathbb{E}\left(e^{u\sum_{i=1}^{n} X_i}\right) \\ &= (M_{X_1}(u))^n.\end{aligned}$$

Finally,

$$\begin{aligned}M_Y(u) &= \mathbb{E}[N_{Y|N}(u \mid N)] \\ &= \mathbb{E}[(M_{X_1}(u))^N] \\ &= \mathbb{E}[\exp(N \ln M_{X_1}(u)] \\ &= M_n(\log M_{X_1}(u)).\end{aligned}$$

This implies

$$K_Y(u) = K_N(K_{X_1}(u)).$$

Back to the insurance example. Note that $X_1, X_2, \ldots \sim \operatorname{Exp}(\lambda)$, and $N \sim \operatorname{Geo}(p)$. We have

$$\mathbb{E}(Y) = \mathbb{E}(N)\mathbb{E}(X_1) = \frac{1}{p}\frac{1}{\lambda} = \frac{1}{\lambda p}.$$

Then

$$\begin{aligned}M_Y(u) &= M_N(\ln(M_{X_1}(u)) \\ &= M_N(\ln(1 - \frac{u}{\lambda}^{-1}) \\ &= \left(1 - \frac{1}{p} + \frac{1}{p}\left(1 - \frac{u}{\lambda}\right)\right)^{-1} \\ &= \left(1 - \frac{1}{p} + \frac{1}{p} - \frac{u}{\lambda p}\right)^{-1} \\ &= \left(1 - \frac{u}{\lambda p}\right),\end{aligned}$$

so $Y \sim \operatorname{Exp}(\lambda p)$. ◇

Part II

Lent Term

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 $S_n = \bar{Y}_n$. 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:

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

What is the sample equivalent of moments?

Definition 7.2.2. Let

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

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

Example 7.2.3. We have

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

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 Normal(μ, σ^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 $Z_1, \ldots, Z_n \sim$ 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,

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

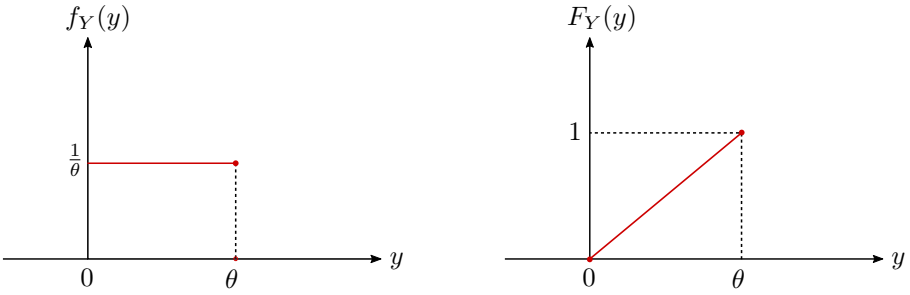


Figure 7.1: The PDF and CDF of the Unif$[0, \theta]$ distribution, respectively.

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 $F_{Y_{(n)}}(y) = \left(\frac{y}{\theta}\right)^n$.

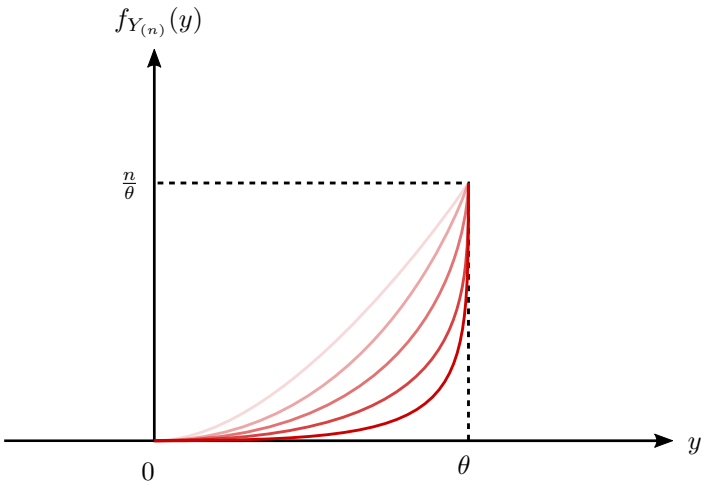


Figure 7.2: The PDF of $Y_{(n)}$ in Example 7.4.1. As n increases, the degree of the polynomial also increases. This, in turn, increases the probability that $Y_{(n)}$ takes a value closer to θ.

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

The CDF of $Y_{(i)}$

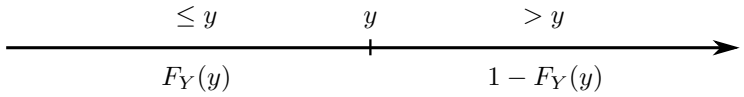


Figure 7.3: For a given random variable Y, the probability that the observed value is less than y is $F_Y(y)$, and the probability that it is greater than y is $1 - F_Y(y)$.

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:

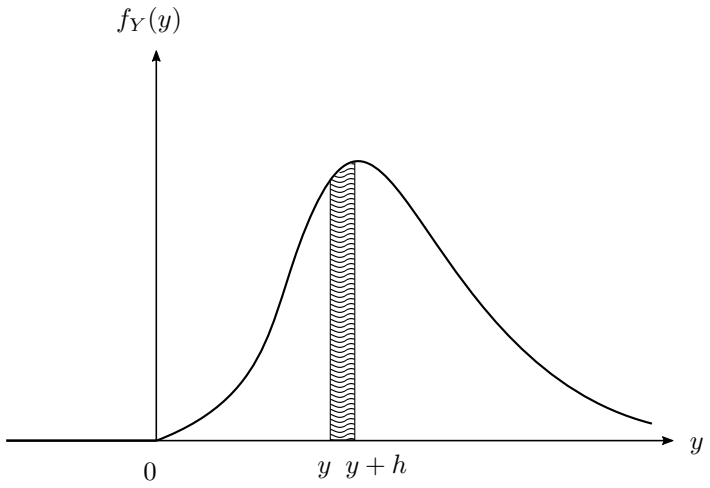


Figure 7.4: As h approaches 0, the shape under the distribution approaches a rectangle.

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 $y + h$?

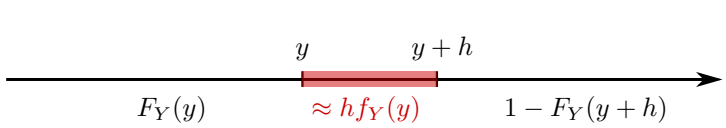


Figure 7.5: The probability that a value falls between y and $y + h$ is approximately equal to $hf_Y(y)$ for small values of h.

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

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

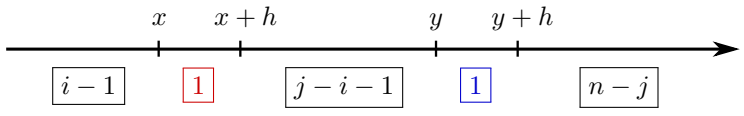
◇

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.

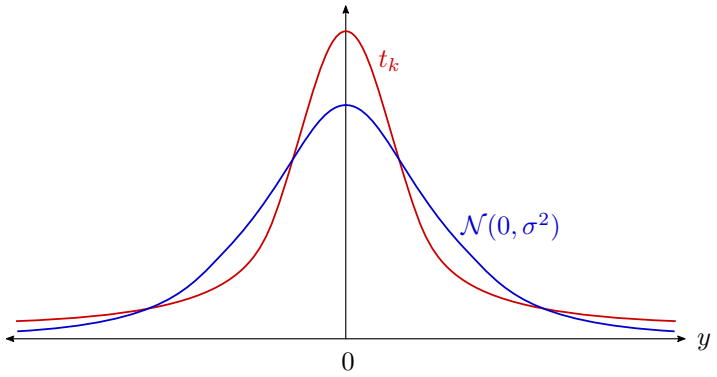


Figure 8.1: A t-distribution compared with a Normal distribution. The t-distribution has clearly fatter tails.

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

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

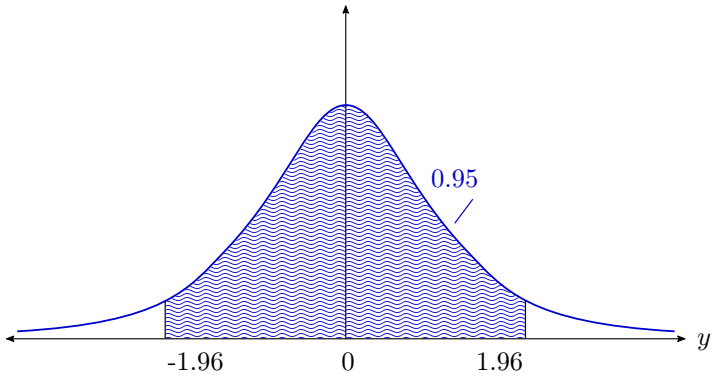


Figure 8.2: A 95% confidence interval of the standard normal.

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

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

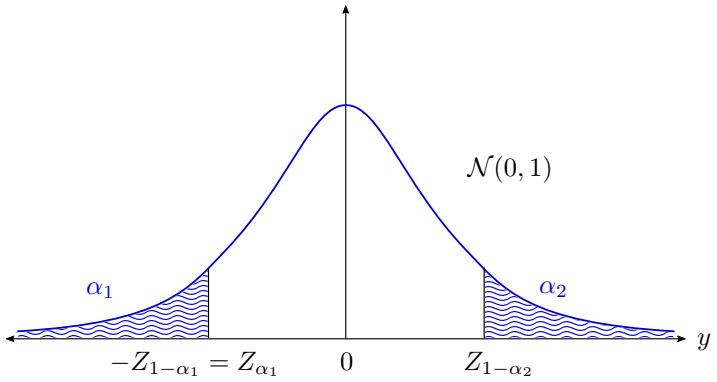


Figure 8.3: A more general case for a standard normal pivotal.

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 $W = g(\mathbf{Y}, \theta) = a(\mathbf{Y}) + b(\mathbf{Y})\theta$.

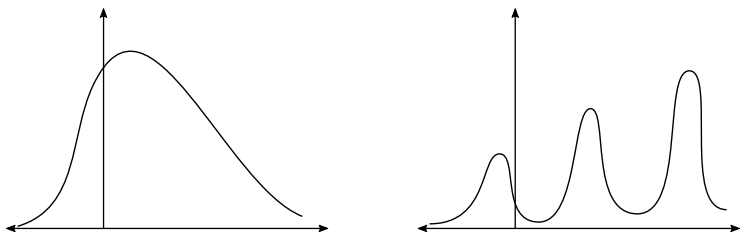


Figure 8.4: A unimodal (left) distribution vs. a multimodal distribution.

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:

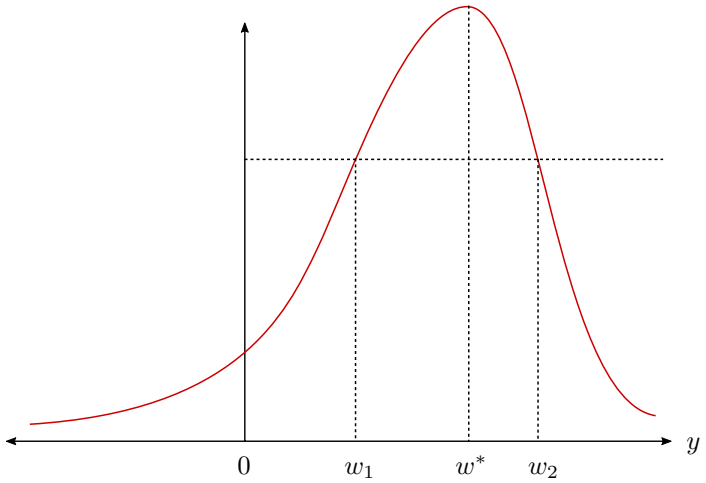


Figure 8.5: Suppose that the area encompassed by w_1 and w_2 corresponds to the desired value of α, and w^* is the mode. If we shift w_1 to the left, then it follows that we will have to shift w_2 to the left as well. However, we shift w_2 *less* than w_1, because the area under the distribution curve 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:

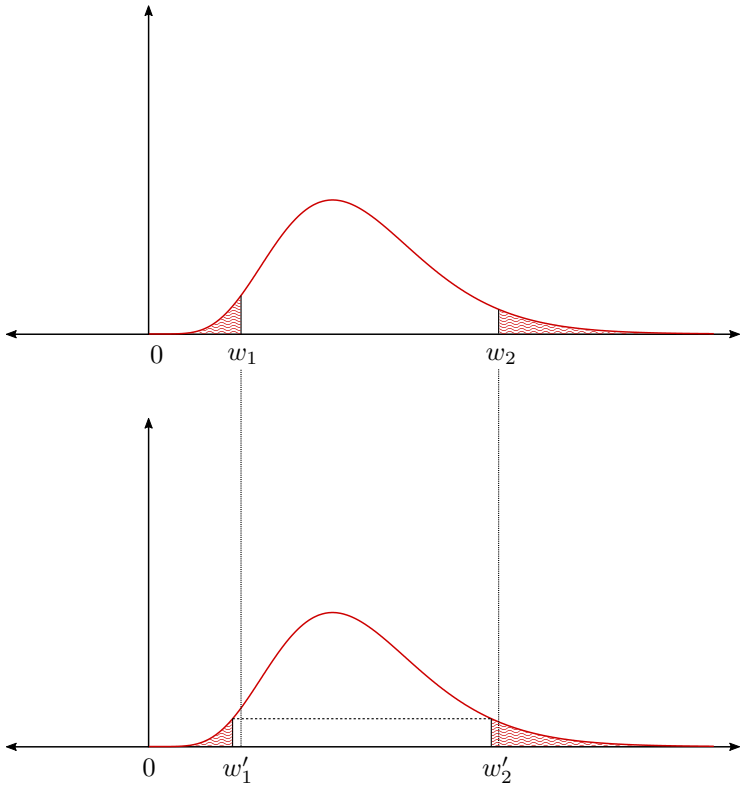


Figure 8.6: Equal-tail confidence interval (top) vs. optimal equal-density confidence interval for a Gamma distribution. *Note: Gamma curve taken from Wikimedia Commons.*

◇

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 $\mu < \mu_0$,

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

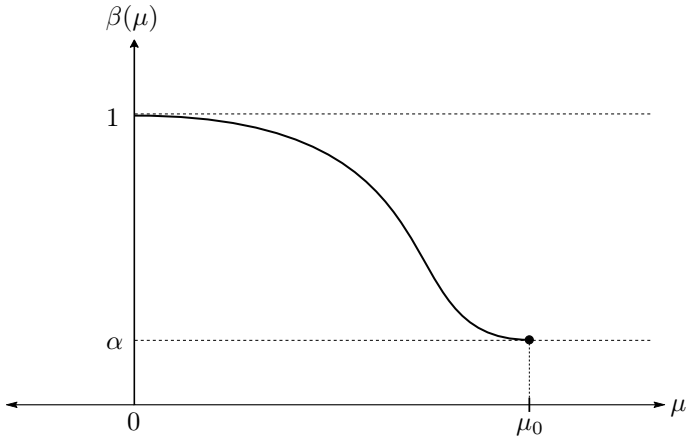


Figure 8.7: The power function in Example 8.4.9.

◇

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:

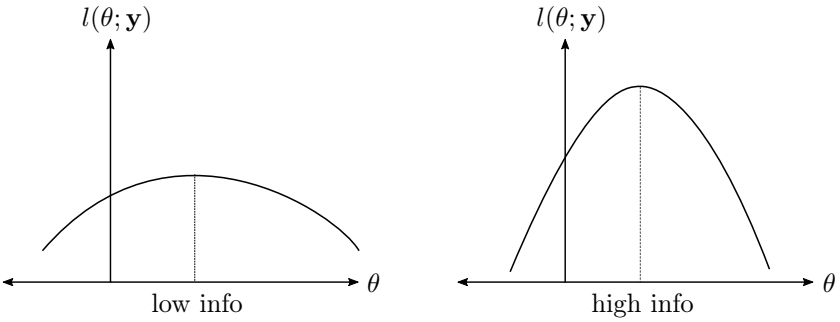


Figure 9.1: Two log-likelihood functions. The function that is relatively flat (left) doesn't tell us much about θ since its likelihood function does not differ much, and therefore has low information. The likelihood of different values of θ differs significantly on the right function, and so we say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

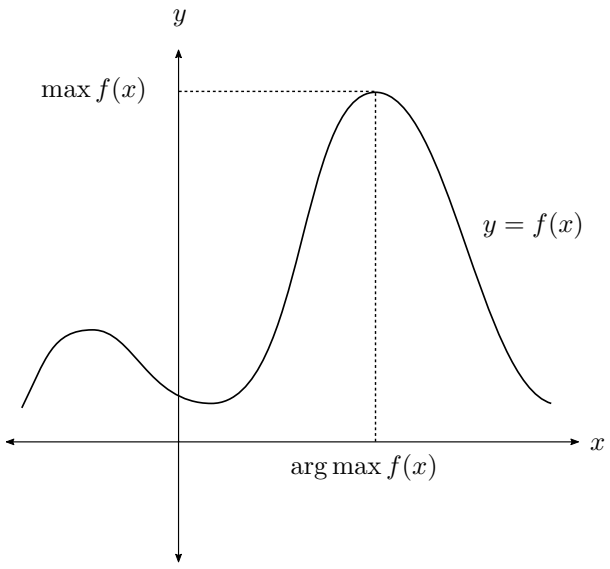


Figure 9.2: For a function $y = f(x)$, the $\arg\max f(x)$ is the x value which outputs $\max f(x)$.

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.

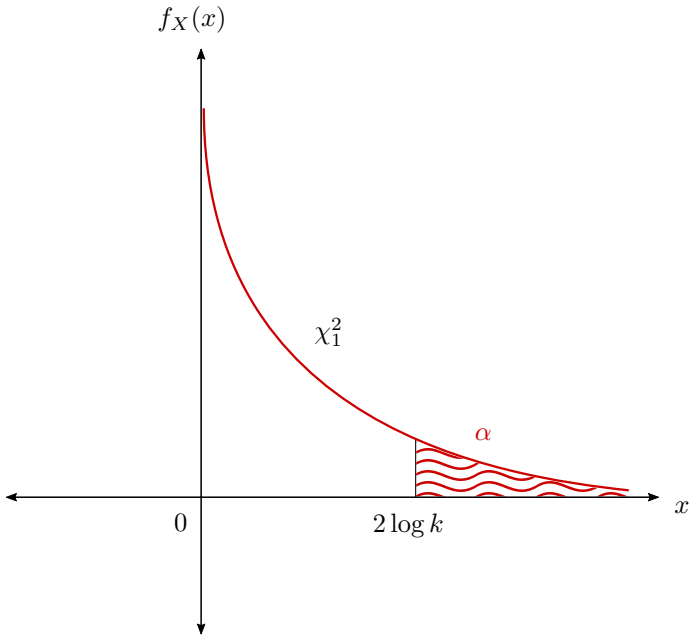


Figure 9.3: The rejection region of the likelihood-ratio test in Example 9.3.6, with $X \sim \chi^2_1$.

◇

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 $Y_1, \dots, Y_n \sim F_Y(y; \theta)$ 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 $n\bar{y} = \sum_{i=1}^{n} y_i$.

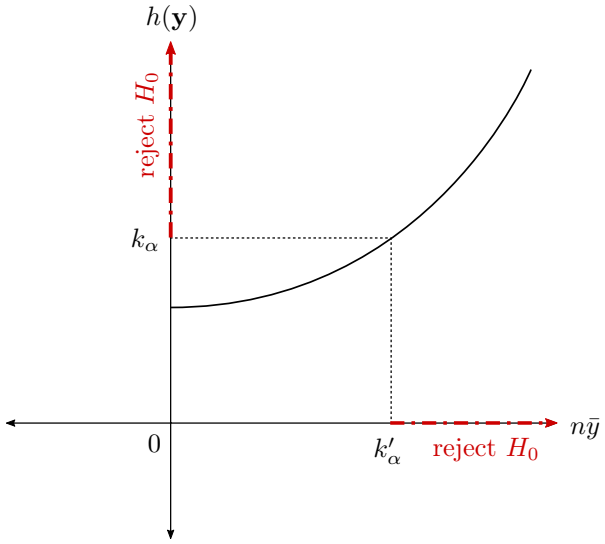


Figure 10.1: Intuition behind the critical region .

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.

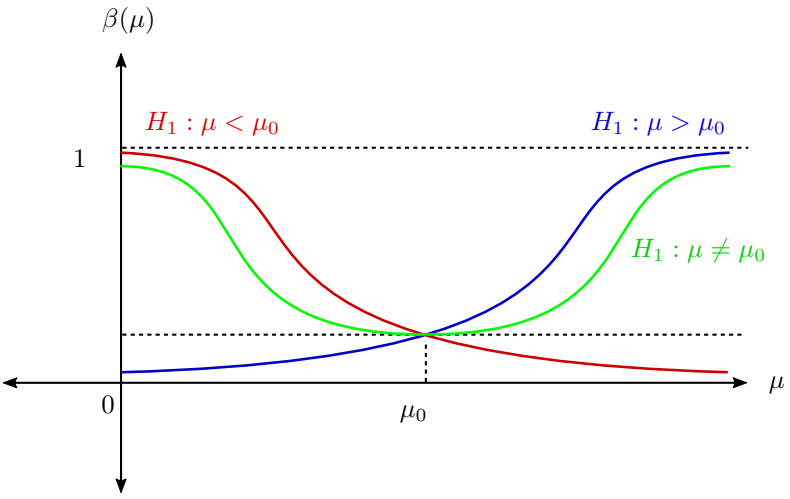


Figure 10.2: The alternative hypotheses $H_1 : \mu < \mu_0$ and $H_1 : \mu > \mu_0$ are more powerful than $H_1 : \mu \neq \mu_0$ at values less than and greater than μ_0, respectively.

◇

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.

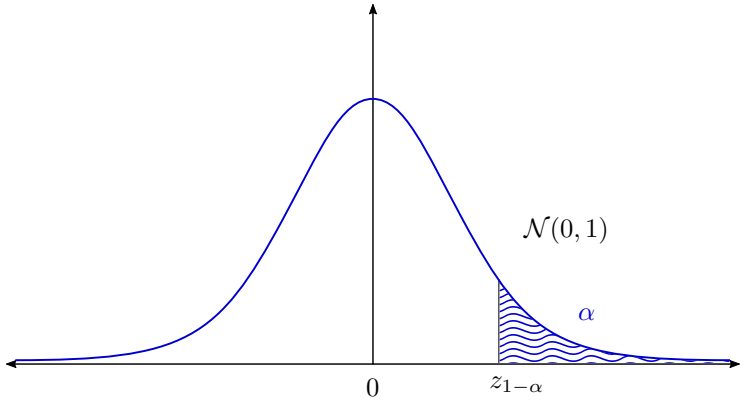


Figure 10.3: The rejection region of Z in Example 10.4.6.

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 $\mathrm{Var}(\varepsilon_i) = 1$.

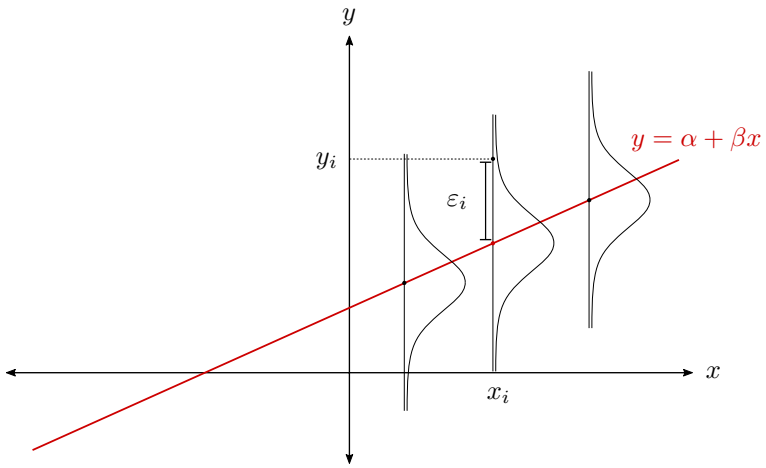


Figure 6.1: We can imagine a density curve at each point on the fitted model that determines y_i. The distance from the mean is the error ε_i.

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

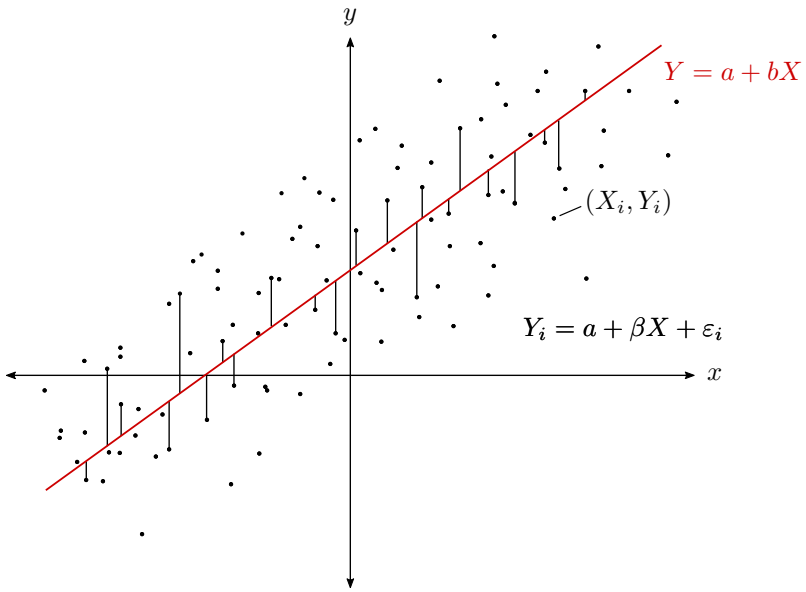


Figure 6.2: A simple fitted linear model. The dispersed vertical lines represent the size of the error.

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,

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

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

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

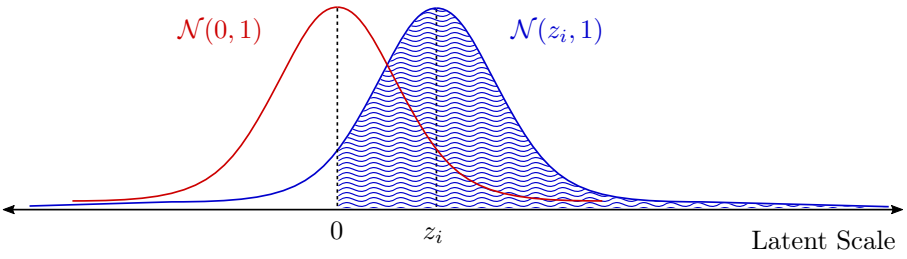


Figure 6.3: We can visualize the probit method through a latent scale depending on z_i. If individual i falls to the right of the standard normal, our model predicts that they default. The probability that an individual defaults is Normal$(z_i, 1)$ distributed, with this example showing a positive value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