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.

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:

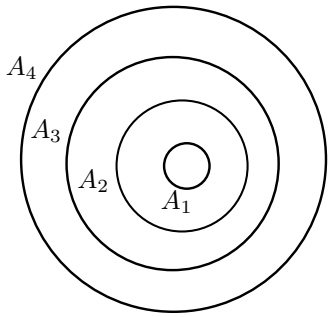


Figure 2.1: Visual representation of $A_1, A_2, \ldots$

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:

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

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.

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

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

$\diamond$

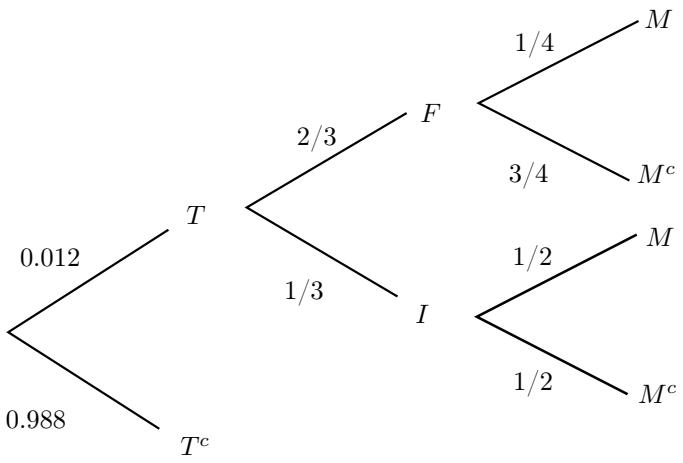


Figure 2.2: A probability tree representing this situation.

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

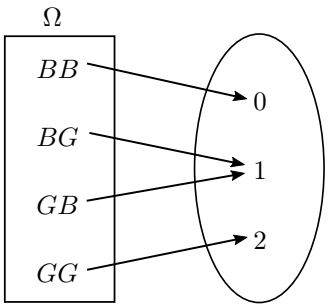
3.1 Week 2: Lecture 2 (continued)

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. ◇

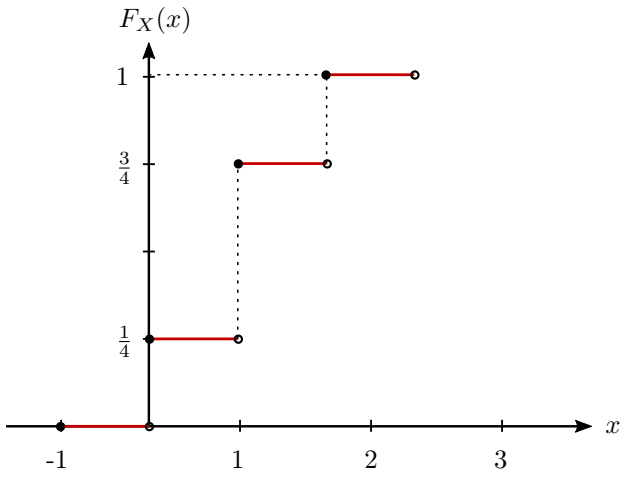


Figure 3.1: The Cumulative Distribution Function of the two child example.

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.

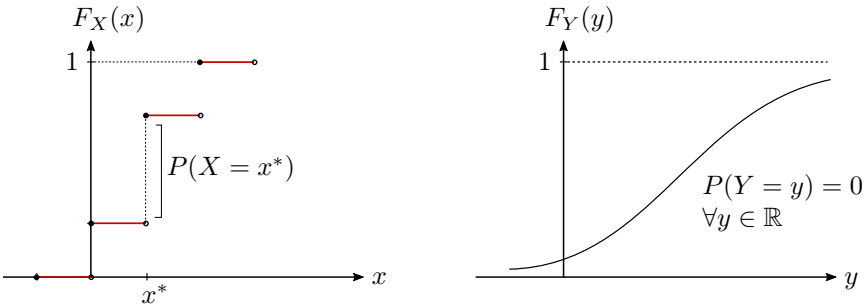


Figure 3.2: Discrete vs. Continuous CDF

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

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

◇

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

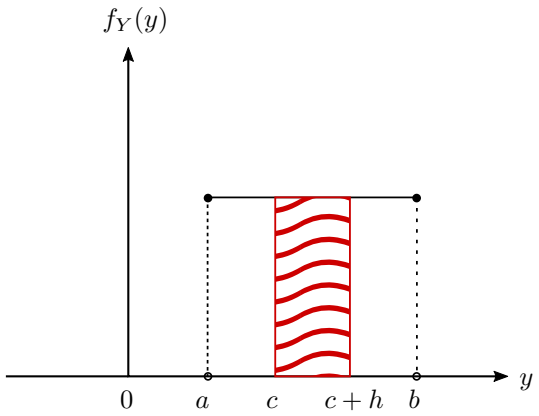


Figure 3.3: The continuous Uniform Distribution with parameters a, b.

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

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, $Y \sim \text{Exp}(\lambda)$. ◇

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.

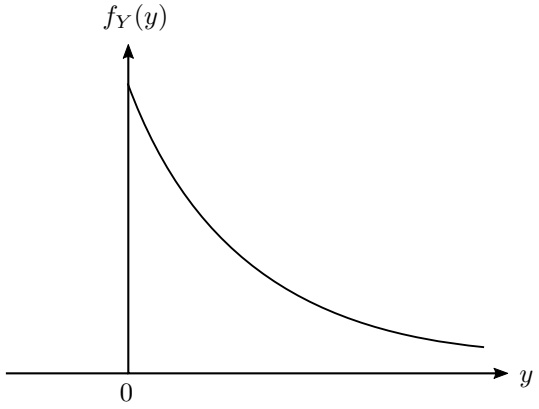


Figure 3.4: A simple graphical representation of the PDF of the Exponential Distribution.

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

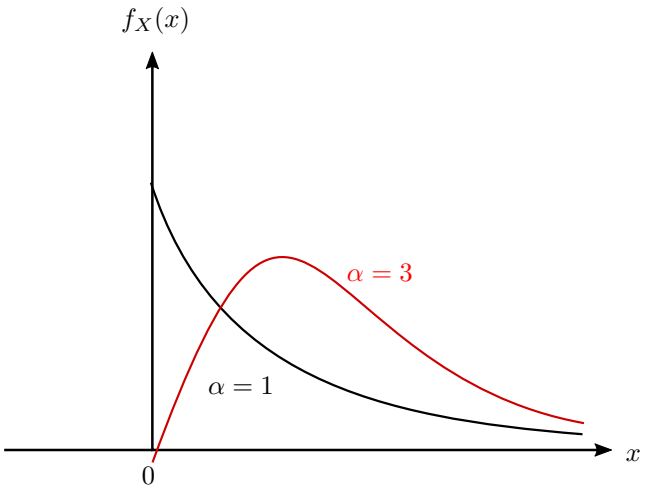


Figure 3.5: PDF of the Gamma Distribution for $\alpha = 1$ and $\alpha = 3$

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 $a > 0$.

Proof. Observe that

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

□

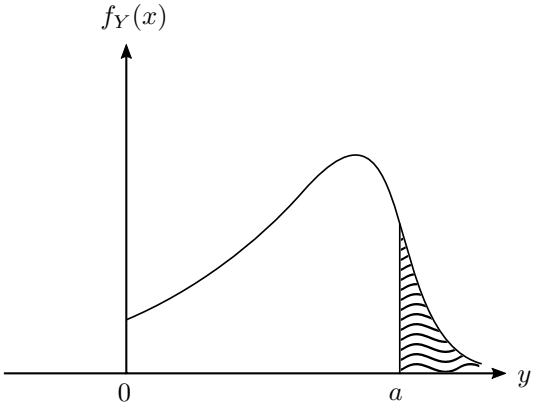


Figure 3.6: The Markov Inequality shows that the shaded area (the survival function of X evaluated at a) is always less than or equal to $\frac{\mathbb{E}(Y)}{a}$.

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

Similarly, if a function is monotone decreasing, then

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

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

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

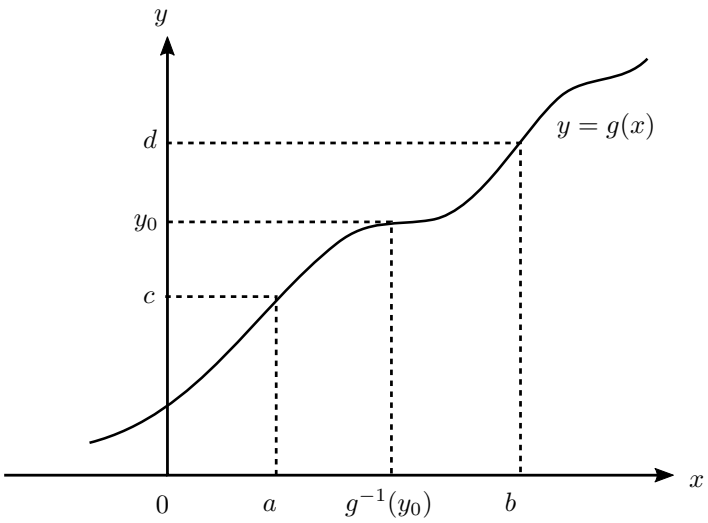


Figure 3.7: A monotone increasing function

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

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

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

Remark 3.9.8. If X is continuous, then

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

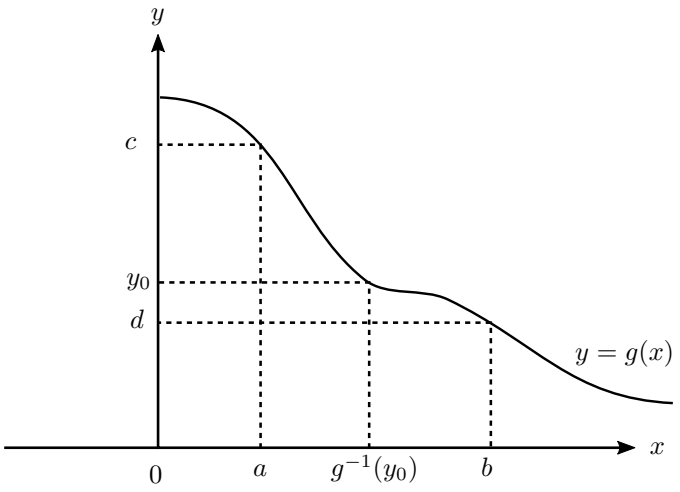


Figure 3.8: A monotone decreasing function

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

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

so

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

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

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

◇

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)**.**

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

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

Note that

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

◇

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

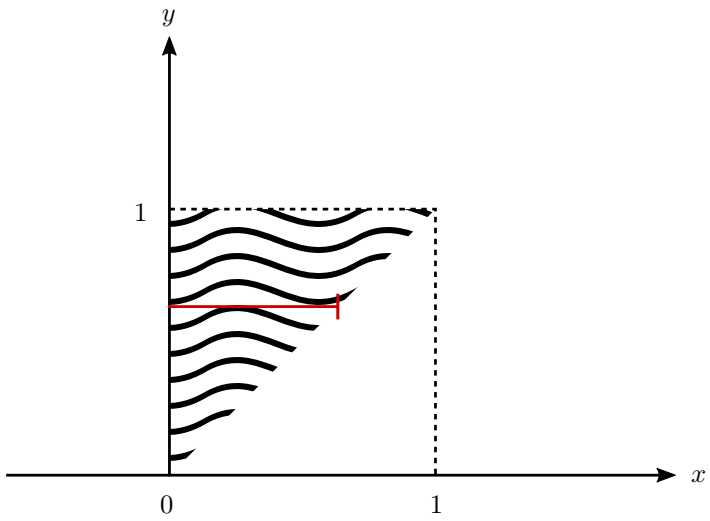


Figure 4.2: The support of Example 4.2.1. The support of a bivariate PDF is an interval in $\mathbb{R}^2$. Finding the limits of integration for each axis can be tricky.

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 $-1 \le \rho \le 1$.

4.3 Week 9: Lecture 1

Tue 23 Nov 14:00

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

$$-1 \le \mathrm{Corr}(X,Y) \le 1,$$

Moreover, $|\mathrm{Corr}(X,Y)| = 1$ 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,

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

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 $\kappa_{1,1} = \mathrm{Cov}(X,Y)$. ◇

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,

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

□

By Example 4.3.9, we can write

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

4.4 Week 9: Lecture 2

Wed 24 Nov 10:00

4.4.1 Independent Random Variables

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

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

◇

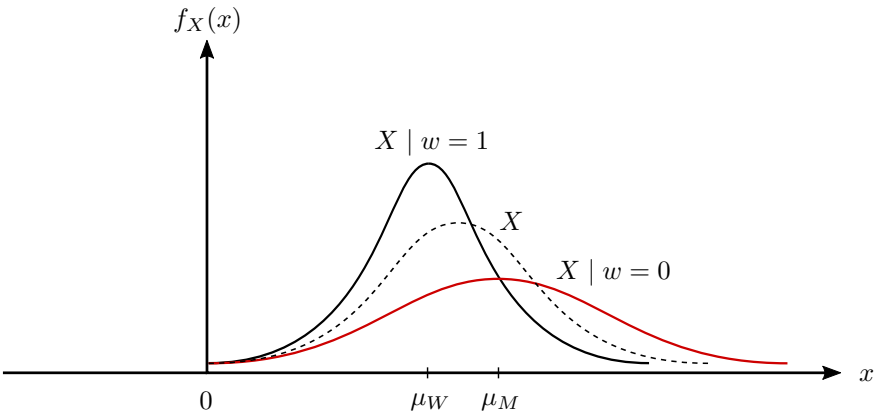


Figure 5.1: The distribution of X from Example 5.3.4 is somewhere in between the distributions of $X \mid w = 1$ and $X \mid w = 0$.

Example 5.3.5 (Household Insurance)**.**

$\text{Exp}(\lambda)$:	amount claimed each year
$\text{Normal} \sim \text{Geo}(p)$:	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 $Y \sim \operatorname{Exp}(\lambda p)$. ◇