STAT330S18 - Mathematical Statistics

Classnotes for Spring 2018

by

Johnson Ng

BMath (Hons), Pure Mathematics major, Actuarial Science Minor

University of Waterloo

Table of Contents

List of Definitions

List of Theorems

Foreword

The proofs in this set of notes will be more rigourous compared to the expectations of the course (at least, for the course this term). If you are not the author and is interested in reading the notes, you may skip the proofs should you have little interest in them. The rigour is required almost exclusively for the author himself, for his own practice, and because he transferred his STAT230 course from a class that is clean of proofs.

Also, many of the common mathematical notations will be heavily used both in the author's notes and proofs. The author cannot guarantee that his proofs are absolutely correct, but he tries, to the best of his abilities to minimize and assure of that. Should you be suspicious about the proofs, or should you notice errorneous ones, please do inform the author at `https://github.com/japorized/TeX_notes/issues`.

You are also warned that the author is rather bummed with how the course is presented in the term that he is/was taking it, and so there may be sarcastic language (towards the lectures) mixed in his notes.

1 Lecture 1 May 1st 2018

1.1 Introduction

Definition 1 (Sample Space)

*A **sample space**, S of a random experiment is the set of all possible outcomes of the experiment.*

Example 1.1.1

The following are some random experiments and their sample space.

- *Flipping a coin*
  $S = \{H, T\}$ *where H denotes head and T tail.*
- *Rolling a 6-faced dice twice*
  $S = \{(x, y) : x, y \in \mathbb{N},\ 1 \leq x, y \leq 6\}$
- *Measuring a patient's height*
  $S = R^+ = \{x \in \mathbb{R} : x \geq 0\}$

Definition 2 (σ-field)

*Let S be a sample space. The collection of sets $\mathscr{B} \subseteq \mathbb{P}(S)$[1], is called a σ-field (or **σ-algebra**) on S if:*

1. $\emptyset \in \mathscr{B}$ and $S \in \mathscr{B}$;
2. $\forall A \in \mathscr{B} \quad A^C \in \mathscr{B}$;[2] and
3. $\forall n \in \mathbb{N} \quad \forall \{A_j\}_{j=1}^n \subseteq \mathscr{B} \quad \cup_{j=1}^n A_j \in \mathscr{B}$.

[1] The **power set** of S, $\mathbb{P}(S)$, is defined as the set that contains all subsets of S.

[2] We shall denote the compliment of a set by a superscript C in this set of notes. The supplemental notes provided in the class uses an overhead bar, e.g. $\overline{A}$, while lecture notes will use A^C and A' interchangably.

Definition 3 (Measurable Space)

*Given that S is a non-empty set, and $\mathscr{B}$ is a σ-field, $(S, \mathscr{B})$ is a **measurable space**.*[3]

[3] A measurable space is a basic object in **measure theory**.

Example 1.1.2

Consider $S = \{1, 2, 3, 4\}$. Check if $\mathscr{B} = \{\emptyset, \{1, 2, 3, 4\}, \{1, 2\}, \{3, 4\}\}$ is a σ-field on S.

1. *It is clear that $\emptyset, S \in \mathscr{B}$.*
2. *Note that $S^C = \emptyset$ and $\{1, 2\}^C = \{3, 4\}$.*
3. *Note that the largest possible result of any countable union of the elements of $\mathscr{B}$ is $\{1, 2, 3, 4\}$, which is an element of $\mathscr{B}$.*

Because $(S, \mathscr{B})$ is a measurable space, we can define a measure on it.

Definition 4 (Probability Measure)

*Suppose S is a sample space of a random experiment. Let $\mathscr{B} = \{A_1, A_2, \ldots\} \subseteq \mathbb{P}(S)$ be the σ-field on S. The **probability set function** (or **probability measure**), $P : \mathscr{B} \to [0, 1]$, is a function that satisfies the following:*[4]

[4] These conditions are also known as **Kolmogorov Axioms**, or **probability axioms**.

- $\forall A \in \mathscr{B} \;\; P(A) \geq 0$;
- $P(S) = 1$;
- $\forall \{A_j\}_{j=1}^{\infty} \subseteq \mathscr{B} \;\; \forall i \neq j \in \mathbb{N} \;\; A_i \cap A_j = \emptyset \implies$

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j) \tag{1.1}$$

*$(S, \mathscr{B}, P)$ is called a **probability space**.*

Example 1.1.3

Consider flipping a coin where $S = \{H, T\}$. Let P be defined as follows

$$P(\{H\}) = \frac{1}{3} \quad P(\{T\}) = \frac{2}{3} \quad P(\emptyset) = 0 \quad P(S) = 1$$

Conditions 1 and 2 of Definition 4 are met. Notice that

$$P(\{H\} \cup \{T\}) = P(S) = 1 \text{ and } P(\{H\}) + P(\{T\}) = \frac{1}{3} + \frac{2}{3} = 1.$$

Hence condition 3 is also fulfilled.

Proposition 1 (Properties of Probability Set Functions)

Let P be a probability set function and A, B be any set in $\mathscr{B}$. Prove the following:[5]

1. $P(A^C) = 1 - P(A)$
2. $P(\emptyset) = 0$
3. $P(A) \leq 1$
4. $P(A \cap B^C) = P(A) - P(A \cap B)$
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
6. $A \subseteq B \implies P(A) \leq P(B)$

[5] Many among these properties illustrate that the probability is indeed a **measure**.

Exercise 1.1.1
Prove that $A \subseteq B \iff B^C \subseteq A^C$.

Proof

Let S be the sample space for P.

1. *Note that*
   $A \in \mathscr{B} \implies A \in \mathbb{P}(S) \iff A \subseteq S$
   $A \in \mathscr{B} \iff A^C \in \mathscr{B} \implies A^C \subseteq S$*. Also, since A^C is the complement of A, it is clear that $S = A \cup A^C$.*

   $$\therefore P(S) = 1 \iff P(A \cup A^C) = 1 \overset{1}{\iff} P(A) + P(A^C) = 1$$

   where 1 is by condition 3 in Definition 4 since $A \cap A^C = \emptyset$ by definition of a complement of a set.

2. *Note that $S \cup \emptyset = S$ and $S \cap \emptyset = \emptyset$. Using a similar argument as above,*

   $$1 = P(S) = P(S \cup \emptyset) = P(S) + P(\emptyset) \implies P(\emptyset) = 0$$

3. *By 1 from above,* $P(A) = 1 - P(A^C)$. *Since* $0 \leq P(A^C) \leq 1$, *we have that* $P(A)$ *is at most* 1, *as required.*

4. *Note that* $A = (A \cap B) \cup (A \cap B^C)$. *Clearly,* $(A \cap B) \cap (A \cap B^C) = \emptyset$.[6] *Hence by condition 3 in Definition 4,*

$$P(A) = P(A \cap B) + P(A \cap B^C)$$

[6] This is an easy proof using the basic way of proving membership.

5. *Consider* $P(A \cup B) + P(A \cap B)$. *By definition,*

$$A \cup B = (A \cap B^C) \cup (A \cap B) \cup (A^C \cap B)$$

where each of the sets in brackets are disjoint from each other[7]. *By condition 3 of Definition 4, we would then have*

$$\begin{aligned} &P(A \cup B) + P(A \cap B) \\ &= P(A \cap B^C) + P(A \cap B) + P(A^C \cap B) + P(A \cap B) \\ &= 2P(A \cap B) + P(A) - P(A \cap B) + P(B) - P(A \cap B) \quad \text{by } 4 \\ &= P(A) + P(B) \end{aligned}$$

[7] Again, this is not hard to show

6. *Note that* $B = B \cap S = B \cap (A^C \cup A) = (B \cap A^C) \cup A$. *Clearly,* $A \cap (B \cap A^C) \neq \emptyset$. *By condition 3 in Definition 4, we thus have that*

$$P(B) = P(B \cap A^C) + P(A). \tag{†}$$

Suppose $A \subsetneq B$. *Then* $B \cap A^C \neq \emptyset$. *I shall make the claim that* $B \cap A^C \in \mathscr{B}$. *Since* $A \subseteq B$ *we have that*

$$\begin{aligned} a \in (B \cap A^C) &\iff a \in B \wedge a \in A^C \\ &\iff a \in B \wedge a \notin A \\ &\iff a \in (B \setminus A). \end{aligned}$$

But $B \setminus A$ *is a subset of* B *from the above steps*[8]. *Therefore,* $(B \cap A^C) \subseteq B \in \mathscr{B}$ *as required.*

[8] This is rather obvious from the steps, since $\forall a \in (B \cap A^C)$, $a \in B$.

With that done, by condition 1 in Definition 4, $P(B \cap A^C) \geq 0$. *Hence from Equation (†), we have that*

$$\begin{aligned} P(B) &= P(B \cap A^C) + P(A) \\ &\geq P(A) \end{aligned}$$

as required. $\square$

Definition 5 (Conditional Probability)

*Suppose S is a sample space of a random experiment, and $A, B \subseteq S$. The **conditional probability of A given B** is given by*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \textit{provided } P(B) > 0. \tag{1.2}$$

Definition 6 (Independent Events)

*Suppose S is a sample space of a random experiment, and $A, B \subseteq S$. A and B are said to be **independent of each other** if*

$$P(A \cap B) = P(A)P(B)$$

Proposition 2 (Boole's Inequality)

If $\{A_j\}_{j=1}^{\infty}$ is a sequence of events, then

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) \leq \sum_{j=1}^{\infty} P(A_j)$$

Proof

Proof shall be provided later

Proposition 3 (Bonferroni's Inequality)

If $\{A_j\}_{j=1}^{k}$ is a set of events where $k \in \mathbb{N}$, then

$$P\left(\bigcap_{j=1}^{k} A_j\right) \geq 1 - \sum_{j=1}^{k} P(A_j^C)$$

Proof

Proof shall be provided later

Proposition 4 (Continuity Property)

If $A_1 \subset A_2 \subset \ldots$ is a sequence where $A = \cup_{i=1}^n A_i$, then

$$\lim_{n\to\infty} P\left(\bigcup_{i=1}^n A_i\right) = P(A)$$

Proof

Proof shall be provided later

1.2 Random Variable

Definition 7 (Random Variable)

*In a given probability space $(S, \mathscr{B}, P)$, the function $X : S \to \mathbb{R}$ is called a **random variable**[9] if*

$$P(X \leq x) = P(\{\omega \in S : X(\omega) \leq x\}) \tag{1.3}$$

is defined for all $x \in \mathbb{R}$[10].

[9] We shall use rv as shorthand for random variable in this set of notes.

[10] $X \leq x$ is an abbreviation for $\{\omega \in S : X(\omega) \leq x\} \in \mathscr{B}$.

Example 1.2.1

In a coin flip experiment, we have that $S = \{H, T\}$ where $\mathbb{P}(S) = \{\emptyset, S, \{H\}, \{T\}\}$. Define X : the number of heads in a flip, i.e.

$$X(\{H\}) = 1 \text{ and } X(\{T\}) = 0$$

To prove why X is a random variable given this definition, notice that

$$\begin{aligned} x < 0 &\implies P(X \leq x) = P(\{\omega \in S : X(\omega) < 0\}) = P(\emptyset) = 0 \\ x \geq 1 &\implies P(X \leq x) = P(\{\omega \in S : X(\omega) \leq x\}) = P(\{H, T\}) \\ &\qquad = P(\{H\}) + P(\{T\}) = 1 \text{ by Independence} \\ 0 \leq x < 1 &\implies P(X \leq x) = P(\{\omega \in S : X(\omega) \leq x\}) = P(T) \geq 0 \end{aligned}$$

which shows that P is deiined for all $x \in \mathbb{R}$. Hence X is a random variable.

Definition 8 (Cumulative Distribution Function)

The ***cumulative distribution function (c.d.f)*** *of a random variable X is defined as*

$$\forall x \in \mathbb{R} \quad F(x) = P(X \leq x)$$

Note

Notice *that $F(x)$ is defined for* ***all*** *real numbers, and since it is a probability, we have $0 \leq F(x) \leq 1$.*

Proposition 5 (Properties of the cdf)

1. $\forall x_1 < x_2 \in \mathbb{R} \quad F(x_1) \leq F(x_2)$
2. $\lim_{x \to -\infty} = 0 \wedge \lim_{x \to \infty} = 1$
3. $\lim_{x \to a^+} F(x) = F(a)$ [11]
4. $\forall a < b \in \mathbb{R} \quad P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$
5. $P(X = b) = F(b) - \lim_{a \to b^-} F(a)$ [12]

[11] F is a **right-continuous** function.

[12] This is also called **the magnitude of the jump**.

Proof

Proof shall be provided later

Note

The definition and properties of the cdf hold for the rv X regardless of whether S is discrete (finite or countable) or not.

1.3 Discrete Random Variable

Definition 9 (Discrete Random Variable)

An rv X is a ***discrete random variable*** *when its image is finite or countably infinite, i.e.* $X \in \{x_1, x_2, \ldots\}$*. The function*

$$\forall x \in \mathbb{R} \quad f(x) := P(X = x) = F(x) - \lim_{\varepsilon \to 0^+} F(x - \varepsilon)$$

is its probability function, commonly known as the ***probability mass function*** *(pmf). The set* $A := \{x : f(x) > 0\}$ *is called the* ***support set*** *of X, and*

$$\sum_{x \in A} f(x) = \sum_{i=1}^{\infty} f(x_i) = 1. \tag{1.4}$$

Proposition 6 (Properties of pmf)

With the notation from Definition 9*, prove that*

1. $\forall x \in \mathbb{R} \quad f(x) \geq 0$
2. $\sum_{x \in A} f(x) = 1$

Proof

1. *This result follows from the fact that* f *is a pdf, a probability, i.e.* $\forall x \in R$, $f(x) = 0$ *is* $x \notin S$ *where* S *is the sample space, and* $0 \leq f(x) \leq 1$ *if* $x \in S$.
2. *Since* $A = \{x : f(x) > 0\}$*, we know that*

$$\sum_{x \in A} f(x) > 0.$$

If we consider all the elements of A, we have that the events $(X = x_i)$, for $x_i \in A$, constitutes the entire sample space. Therefore,

$$\sum_{x \in A} f(x) = \sum_{x \in A} P(X = x) = P(S) = 1.$$

$\square$

Exercise 1.3.1

Consider an urn containing r red marbles and b black marbles. Find the pmf of the rv for the following:

1. *X = number of red balls in n selections without replacement.*
2. *X = number of red balls in n selections with replacement.*
3. *X = number of black balls selected before obtaining the first red ball if sampling is done with replacement.*
4. *X = number of black balls selected before obtaining the kth red ball if sampling is done with replacement.*

Solution

1. *Let $d = \max\{n, r+b\}$. The desired pmf is therefore the pmf from the hypergeometric distribution*

$$\forall x \in \mathbb{Z}^+_{\leq r} \quad f(x) = \frac{\binom{r}{x}\binom{b}{d-x}}{\binom{r+b}{d}}.$$

2. $\forall x \in \mathbb{Z}^+ \quad f(x) = \binom{n}{x}\left(\frac{r}{r+b}\right)^x\left(\frac{b}{r+b}\right)^{n-x}$*, which is the pmf of the binomial distribution.*
3. $\forall x \in \mathbb{Z}^+ \quad f(x) = \left(\frac{b}{r+b}\right)^x\left(\frac{r}{r+b}\right)$
4. $\forall x \in \mathbb{Z}^+ \quad f(x) = \binom{x+k-1}{k-1}\left(\frac{b}{r+b}\right)^x\left(\frac{r}{r+b}\right)^k$

Example 1.3.1

Consider the function

$$f(x) = \begin{cases} \frac{C\mu^x}{x!} & x \in \mathbb{Z}^+, \mu > 0 \\ 0 & \textit{otherwise} \end{cases}$$

Find C such that $f(x)$ is a pmf for the rv X.

Solution

We have that

$$\begin{aligned} 1 &= \sum_{x \in \mathbb{Z}^+} \frac{C\mu^x}{x!} \\ &= C \sum_{x \in \mathbb{Z}^+} \frac{\mu^x}{x!} \\ &= Ce^\mu \end{aligned}$$

Thus $C = e^{-\mu}$.

This gives us that $\forall x \in \mathbb{Z}^+, f(x) = \frac{e^{-\mu}\mu^x}{x!}$, and this is, of course, the pmf of the **Poisson distribution**.

Exercise 1.3.2

Prove that the pdf of $X \sim \mathrm{Poi}(\mu)$ *sums to* 1 *over all of its values.*

Solution

$$\begin{aligned} \sum_{x \in \mathbb{N}} \frac{\mu^x e^{-\mu}}{x!} &= e^{-\mu} \sum_{x \in \mathbb{N}} \frac{\mu^x}{x!} \\ &= e^{-\mu} e^{\mu} \quad \because \sum_{x \in \mathbb{N}}^{\infty} \frac{k^x}{x!} = e^k \\ &= 1 \end{aligned}$$

Exercise 1.3.3

If X is a random variable with pmf

$$f(x) = \frac{-(1-p)^x}{x \log p}, \quad x = 1, 2, \ldots;\ 0 < p < 1,$$

show that

$$\sum_{x \in \mathbb{N}} f(x) = 1$$

Solution

$$\begin{aligned} \sum_{x \in \mathbb{N}} \frac{-(1-p)^x}{x \log p} &= -\frac{1}{\log p} \sum_{x \in \mathbb{N}} \frac{(-1)^x (p-1)^x}{x} \\ &= -\frac{1}{\log p} \underbrace{\left[-(p-1) + \frac{(p-1)^2}{2} - \frac{(p-1)^3}{3} + \ldots \right]}_{\text{Taylor expansion of } -\log p} \\ &= 1 \end{aligned}$$

2 Lecture 2 May 03rd 2018

2.1 Continuous Random Variable

Definition 10 (Continuous Random Variable)

*Suppose X is an rv with cdf F. If F is a continuous function for all $x \in \mathbb{R}$ and F is differentiable except possibly at countably many points, then X is a **continuous rv**. The probability function, or more commonly known as the **probability density function** (pdf), of X is $f(x) = F'(x)$ wherever F is differentiable on x and 0 otherwise.*

*The set $A = \{x : f(x) > 0\}$ is called the **support set** of X and*

$$\int_{x \in A} f(x)\, dx = 1$$

Proposition 7 (Properties of pdf)

Let X be a random variable and f be its pdf.

1. $\forall x \in \mathbb{R} \quad f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$
3. $f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \to 0} \frac{P(x \leq X \leq x+h)}{h}$ *(if the limit exists)*
4. $\forall x \in \mathbb{R} \quad F(x) = \int_{-\infty}^{x} f(t)\, dt$
5. $P(a < X \leq b) = \int_a^b f(x)\, dx = F(b) - F(a)$
6. $P(X = b) = F(b) - \lim_{a \to b^-} F(a) = F(b) - F(b) = 0$

Proof

1. *The argument of this proof is similar to that provided in* Proposition 6.
2. *Same as above, except that the support set can now have complete intervals.*
3. *The first equation follows from the first principles of Calculus. The second equation follows by method of calculation using the cdf.*
4. $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt.$
5. *This follows immediately from the above property.*
6. *The first part of the equation is a way to interpret the above property. The limit equates to $F(b)$ since F is continuous.*

$\square$

Example 2.1.1

Consider the function

$$f(x) = \begin{cases} \frac{\theta}{x^{\theta+1}} & x \geq 1 \\ 0 & x < 1 \end{cases}$$

For what values of θ is f a pdf?

Solution

If f is a pdf, then $\theta \geq 0$. In fact, $\theta \neq 0$; otherwise f would be equivalently 0 for all $x \in \mathbb{R}$, which would imply that $\int_{\mathbb{R}} f = 0$, which is impossible. It remains to check if $\theta > 0$ is a safe choice. Now

$$\int_1^\infty \frac{\theta}{x^{\theta+1}}\, dx = -\frac{1}{x^\theta}\bigg|_1^\infty = 1$$

Note that the above integral is valid because $\frac{1}{x^{\theta+1}} \leq \frac{1}{x}$. Therefore the choice of $\theta > 0$ is safe.

2.2 Examples of Discrete RVs

2.2.1 Binomial Distribution

Definition 11 (Binomial RV)

Consider X to be the number of successes in a sequence of n experiments where

1. *experiments are* ***independent****;*
2. *the outcome of each experient is a* ***binary*** *(e.g. success or failure); and*
3. *has the* ***probability of success,*** p *for each singular experiment.*

X is called a ***Binomial*** *rv, and we write $X \sim \mathrm{Bin}(n, p)$ and its pmf is*

$$P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, ..., n \\ 0 & \textit{otherwise} \end{cases}$$

2.2.2 Geometric Distribution

Definition 12 (Geometric RV)

Consider a sequence of independent success/failure (binary) experiments, each of which has a success probability of p. Let X be the ***number of failures*** *before the* ***first success*** *is reached. We call X a* ***Geometric*** *rv, and we write $X \sim \mathrm{Geo}(p)$, and its pmf is*

$$P(X = x) = \begin{cases} (1-p)^x p & x = 0, 1, 2, ..., n \\ 0 & \textit{otherwise} \end{cases}$$

Note

Some authors would define the Geometric rv as:

> *Let X be the number of experiments until the first success.*

But that really is just a play of words.

2.2.3 Poisson Distribution

Definition 13 (Poisson RV)

Suppose X is defined to be the number of occurrences of an event in a given time period. If the process on which the events occur satisfies the following:

1. *The number of occurrences in non-overlapping intervals are independent of each other;*
2. *The probability of the occurrence of an event in a short interval of length h is proportional to h;*
3. *For sufficiently short time periods of length h, the probability of 2 or more events occurring in the interval is negligible, i.e. almost zero;*

*then X is a **Poisson** rv, and we write $X \sim \mathrm{Poi}(\lambda)$, with $\lambda > 0$, and the pmf is*

$$P(X = x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x = 0, 1, \ldots \\ 0 & \text{otherwise} \end{cases}$$

2.3 Examples of Continuous RVs

2.3.1 Normal/Gaussian Distribution

Definition 14 (Normal / Gaussian RV)

*The **Normal/Gaussian** Distribution is a continuous probability distribution that is symmetric about the mean, showing that data around* *the mean is more frequent than data far from the mean. If X is a* ***Normal/Gaussian*** *rv, we write* $X \sim \mathrm{N}(\mu, \sigma^2)$*, and its pdf is*

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

Definition 15 (Standard Normal Distribution)

The ***Standard Normal Distribution*** *is the simplest case of a Normal Distribution. An rv Z is called the* ***Standard Normal*** *rv if* $\mu = 0$ *and* $\sigma = 1$*. We write* $Z \sim \mathrm{N}(0, 1)$ *and its pdf is*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \textit{for } x \in \mathbb{R}.$$

2.3.2 Uniform Distribution

Definition 16 (Uniform RV)

If X represents the result of drawing a real number from an interval (a, b)*, with* $a < b$*, such that all numbers in between are equally likely to be chosen, then X is called a* ***Uniform*** *rv, and we write* $X \sim \mathrm{Unif}(a, b)$*, and its pdf is*

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in (a, b) \\ 0 & \textit{otherwise} \end{cases}$$

2.3.3 Exponential Distribution

Definition 17 (Exponential RV)

Let X show the time between two consecutive events in a ***Poisson process****, i.e. the 3 conditions in Poisson Distribution are satisfied. Then X is* *called an* ***Exponential*** *rv, and we write* $X \sim \mathrm{Exp}(\theta)$*, where* $\theta > 0$*, with its pdf*

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

2.3.4 Gamma Distribution

Definition 18 (Gamma RV)

Let X *be the sum of* n *independent* ***Exponential*** *rvs with some fixed* θ*. Then* X *is called a* ***Gamma*** *rv, in which we write* $X \sim \Gamma(n, \theta)$*, and its pdf is*

$$f(x) = \begin{cases} \frac{x^{n-1} e^{-\frac{x}{\theta}}}{\Gamma(n)\theta^n} & x > 0 \wedge \theta, n > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\Gamma(n) = \int_0^\infty e^{-y} y^{n-1} \, dy = (n-1)!$*, where the last equality is true when* n *is an integer.*

Note
The Gamma Distribution is usually used for when we are looking for the probability of the occurrence of the n*-th event in the desired waiting time.*

2.4 Functions of Random Variables

Consider the rv X with pdf/pmf f and cdf F. Given $Y = h(X)$ where h is some real-valued function, we are interested in finding the pdf/pmf of Y.

The following are some possible scenarios:

1. X and Y are both discrete;

2. X is continuous and Y is discrete;

3. X and Y are both continuous

We may also define $Y = h(X)$ for a continuous rv X such that Y is **neither discrete nor continuous** (e.g. discrete for some values of X while continuous for others).

2.4.1 *Discrete X and Discrete Y*

If X and $Y = h(X)$ are both discrete, we can derive $P(Y = y)$ by mapping values in Y onto their corresponding value through h, i.e.

$$P(Y = y) = \sum_{\{x : h(x) = y\}} P(X = x)$$

Exercise 2.4.1

Let X have the following probability function:

$$f_X(x) = \begin{cases} \frac{e^{-1}}{x!} & x = 0, 1, 2, \ldots \\ 0 & \textit{otherwise} \end{cases}$$

Find the pmf of $Y = (X - 1)^2$.

Solution

Note that since

$$\operatorname{Dom} X = \{0, 1, 2, 3, 4, \ldots\},$$

we have that

$$\operatorname{Dom} Y = \{1, 0, 1, 4, 9, \ldots\}.$$

With that, note that

$$\begin{aligned} P(Y = 0) &= P(X = 1) = \frac{e^{-1}}{1!} \\ P(Y = 1) &= P(X = 0 \textit{ or } 2) = P(X = 0) + P(X = 2) \\ &= \frac{e^{-1}}{0!} + \frac{e^{-1}}{2!} = e^{-1}\left(1 + \frac{1}{2}\right) = \frac{3}{2}e^{-1} \\ P(Y = 4) &= P(X = 3) = \frac{e^{-1}}{3!} \\ P(Y = 9) &= P(X = 4) = \frac{e^{-1}}{4!}. \end{aligned}$$

Therefore, the pmf of $Y = (X-1)^2$ *is*

$$P(Y=y) = \begin{cases} e^{-1} & y=0 \\ \frac{3}{2}e^{-1} & y=1 \\ \frac{e^{-1}}{(1+\sqrt{y})!} & y=4,9,16,\ldots \\ 0 & \textit{otherwise} \end{cases}$$

2.4.2 Continuous X and Discrete Y

If X is continuous and Y is discrete, we can use the method that we have used in the previous subsection, and replace Σ by the integral sign $\int$, i.e. define $A := \{x : h(x) = y\}$ such that we have

$$P(Y=y) = \int_A f(x)\,dx$$

Example 2.4.1 (Example 2.9)

Suppose X *is a random variable with the following probability function*

$$f_X(x) = \begin{cases} 2e^{2x} & x > 0 \\ 0 & \textit{otherwise} \end{cases}.$$

Suppose $Y = h(X)$ *is defined as follows:*

$$Y = \begin{cases} 1 & X < 1 \\ 2 & 1 \leq X \leq 2 \\ 3 & X > 2 \end{cases}$$

Find the probability function of Y.

Solution

Note that $X \sim \mathrm{Exp}(\frac{1}{2})$. *So it is clear that* X *is a crv and since* $Y = 1, 2,$ *or*

3, we have that Y is discrete. Now

$$\begin{aligned}
P(Y=1) = P(X<1) &= \int_0^1 2e^{-2x}\,dx \\
&= -e^{-2x}\Big|_0^1 = 1-e^{-2} \\
P(Y=2) = P(1 \le X \le 2) &= \int_1^2 2e^{-2x}\,dx \\
&= -e^{-2x}\Big|_1^2 = e^{-2}-e^{-4} \\
P(Y=3) = P(X>2) &= \int_2^\infty 2e^{-2x}\,dx \\
&= -e^{-2x}\Big|_2^\infty = e^{-4}
\end{aligned}$$

Thus the pmf is

$$P(Y=y) = \begin{cases} 1-e^{-2} & Y=1 \\ e^{-2}-e^{-4} & Y=2 \\ e^{-4} & Y=3 \end{cases}$$

2.4.3 Continuous X and Continuous Y

If X and $Y = h(X)$ are both continous, start with the definition of the cdf of Y, i.e.

$$F_Y(y) = P(Y \le y) = P(h(X) \le y)$$

solve the inequality for X, and then obtain the cdf of Y. We will then only need to differentiate the cdf wrt y to get the pdf that we desire.

Example 2.4.2 (Example 2.10)

Let X have the following pdf:

$$f_X(x) = \begin{cases} 2e^{-2x} & x \ge 0 \\ 0 & \textit{otherwise} \end{cases}$$

Find the pdf of $Y = \sqrt{X}$.

Solution

We have that the range of values where $f_Y(y) \leq 0$ *is* $y \geq 0$. *Now*

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(\sqrt{X} \leq y) = P(X \leq y^2) \\ &= \int_0^{y^2} 2e^{-2x}\, dx \\ &= -e^{-2x}\Big|_0^{y^2} = 1 - e^{-2y^2} \end{aligned}$$

Therefore, the pdf of Y *is*

$$f_Y(y) = \begin{cases} \frac{d}{dy} 1 - e^{-2y^2} = 4ye^{-2y^2} & y \leq 0 \\ 0 & \textit{otherwise} \end{cases}.$$

2.4.4 A Formula for the Continuous Case

Theorem 8 (One-to-One Transformation of a Random Variable)

Suppose X *is a continuous random variable with pdf* f_X *and support set* $A = \{x : f_X(x) > 0\}$ *and* $Y = h(X)$ *where* h *is a real-valued function. Let* f_Y *be the pdf of the rv* Y *and let* $B = \{y : f_Y(y) > 0\}$. *If* h *is a one-to-one function from* A *to* B *and if* h' *is continuous, then*

$$f_Y(y) = f\left(h^{-1}(y)\right) \cdot \left| \frac{d}{dy} h^{-1}(y) \right|, \quad y \in B$$

Proof

Note that since h *is one-to-one, it is monotonous. Suppose* h *is increasing. Then* h^{-1} *is also an increasing function. Note that the cdf of* Y *is*

$$F_Y(y) = P(Y \leq y) = P(X \leq h^{-1}(y)) = F_X\left(h^{-1}(y)\right).$$

Then the cdf of Y *is*

$$f_Y(y) = \frac{d}{dy} F_X\left(h^{-1}(y)\right) = f_X\left(h^{-1}(y)\right) \cdot \frac{d}{dy} h^{-1}(y)$$

If h is decreasing, then so is its inverse. Thus

$$F_Y(y) = P(Y \leq y) = P(X \geq h^{-1}(y)) = 1 - F_X(h^{-1}(y))$$

Thus the cdf of Y is

$$f_Y(y) = \frac{d}{dy}(1 - F_X(h^{-1}(y))) = -f_X(h^{-1}(y)) \cdot \frac{d}{dy}h^{-1}(y).$$

Note that the pdf of Y is indeed positive since h^{-1} is decreasing.

Combining the two, we have that

$$f_Y(y) = f_X(h^{-1}(y)) \cdot \left| \frac{d}{dy}h^{-1}(y) \right|,$$

as required. $\square$

3 Lecture 3 May 08th 2018

3.1 Functions of Random Variables (Continued)

3.1.1 Special Cases

Example 3.1.1

Recall Example 2.4.1. Suppose X is a rv with the following probability function

$$f_X(x) = \begin{cases} 2e^{-2x} & x > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Define $Y = h(X)$ as follows:

$$Y = \begin{cases} 1 & X < 1 \\ X & 1 \leq X \leq 2 \\ 3 & X > 2 \end{cases}$$

Find the cdf of Y.

Solution

Solution is given differently in the 2 sections. I am not happy with either solutions because some things don't add up. My opinion is that the definition of Y is badly given, along with a badly phrased question. As a result, there are more ways than one to interpret an already confusing information, and thus we have ourselves one hell of a mess.

3.2 Probability Integral Transformation

Theorem 9 (Probability Integral Transformation)

If X is a contiuous rv with cdf F, then $Y = F(X) \sim \mathrm{Unif}(0,1)$. $Y = F(X)$ is called the ***probability integral transformation****.*

Note

The distribution of $Y = F(X)$ can be proven.

Proof

Let X be a continuous rv and $Y = F(X)$. Since $F(X)$ is one-to-one and increasing (i.e. monotonous), there exists $F^{-1}(Y)$ that is a real-valued and increasing function. Then

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

Note that $F_Y(y) = y$ is the cdf of a $\mathrm{Unif}(0,1)$ rv, i.e. the ***standard uniform random variable****. Thus $Y \sim \mathrm{Unif}(0,1)$.*

Note

This theorem essentially states that any rv from a continuous distribution can be transformed into a standard uniform distribution.

Example 3.2.1 (Example 2.11)

Suppose $X \sim \mathrm{Exp}(01)$. We know that $F_X(x) = 1 - e^{-10x}$ for all $x \in \mathbb{R}a$. By Probability Integral Transformation, we have that $Y = F_X(X) = 1 - e^{-10X} \sim \mathrm{Unif}(0,1)$.

Note that the converse of Probability Integral Transformation is true:

Theorem 10 (Converse of Probability Integral Transformation)

Suppose X is a continuous rv with cdf F such that F^{-1} exists. If $U \sim \text{Unif}(0,1)$, we have that $Y = F^{-1}(U) \sim X$.

Proof

Note that

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(F^{-1}(U) \leq y) \\ &= P(U \leq F_X(y)) = F_X(y). \end{aligned}$$

$\square$

Example 3.2.2 (Example 2.12)

Suppose $X \sim \text{Unif}(0,1)$. Find a transformation T such that $T(X) \sim \exp(\theta)$.

Solution

Let $Y = T(X) \sim \text{Exp}(\theta)$. Note that

$$F_Y(y) = 1 - e^{-\frac{y}{\theta}}, \quad y > 0$$

Observe that since

$$x = 1 - e^{-\frac{y}{\theta}} \implies y = -\theta \ln(1-x)$$

we have that

$$F_Y^{-1}(X) = -\theta \ln(1-X).$$

By Converse of Probability Integral Transformation 10, we have that $T = F_Y^{-1}$.

3.3 *Location-Scale Families*

When we look into methods for constructing confidence intervals for an unknown parameter θ. If the parameter θ is either a *scale parameter* or *location parameter*, then a confidence interval is easier to construct.

Definition 19 (Location Parameter and Family)

Suppose X is a continuous rv with pdf $f(x;\mu)$, where μ is a parameter of the distribution of X. Let $F_0(x) = F_X(x;\mu = 0)$, where F_X is the cdf of X, and $f_0(x) = f(x;\mu = 0)$. The parameter μ is called a ***location paramter*** *of the distribution if*

$$F_X(x;\mu) = F_0(x-\mu), \quad \mu \in \mathbb{R}$$

or equivalently,

$$f(x;\mu) = f_0(x-\mu), \quad \mu \in \mathbb{R}.$$

We say that F belongs to a ***location family*** *of distributions.*

Definition 20 (Scale Parameter and Family)

Suppose X is a continuous rv with pdf $f(x;\theta)$, where θ is a parameter of the distribution of X. Let $F_1(x) = F_X(x;\theta = 1)$, where F_X is the cdf of X, and $f_1(x) = (x;\theta = 1)$. The parameter θ is called a ***scale parameter*** *of the distribution if*

$$F_X(x;\theta) = F_1(\frac{x}{\theta}). \quad \theta > 0$$

or equivalently,

$$f(x;\theta) = \frac{1}{\theta} f_0(\frac{x}{\theta}), \quad \theta > 0.$$

We say that F belongs to a ***scale family*** *of distributions.*

Definition 21 (Location-Scale Family)

Suppose X is an rv with cdf $F(x;\mu,\sigma)$ where $\mu \in \mathbb{R}$ and $\sigma > 0$ are the parameters of the distribution. Let $Y = \frac{X-\mu}{\sigma}$. If the distribution of Y does not depend on μ and/or σ, then F is said to belong to a ***location-scale family*** *of distributions, with* ***location parameter*** *μ and* ***scale parameter*** *σ. In other words, F belongs to a location-scale family of distributions if*

$$F(x;\mu,\theta) = F_0\left(\frac{x-\mu}{\theta}\right),$$

where $F_0(x) = F(x;\mu = 0, \theta = 1)$, *or equivalently,*

$$f(x;\mu,\theta) = \frac{1}{\theta} f_0\left(\frac{x-\mu}{\theta}\right),$$

where $f_0(x) = f(x;\mu = 0, \theta = 1)$.

Example 3.3.1 (Example 2.13)

Consider $X \sim \mathrm{G}(\mu,\sigma)$. *Show that* F_X *belongs to a location-scale family of distributions.*

We know that if $\mu = 0$ *and* $\sigma = 1$, *then* $Y = \frac{X-\mu}{\sigma} \sim \mathrm{G}(0,1)$, *and we know that* $\mathrm{G}(0,1)$ *has no dependence on unknowns* μ *and* σ. *Therefore,* F_X *belongs to the location-scale family of distributions, with location parameter* μ *and scale parameter* σ.

Another solution is to show that one of the equations in the definition is fulfilled. Observe that

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

So if we set $\mu = 0$ *and* $\sigma = 1$ *to get* f_0, *we have that*

$$f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Now, note that

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

Let $y = \frac{x-\mu}{\sigma}$, *and we have ourselves*

$$f(x) = \frac{1}{\sigma}\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} = \frac{1}{\sigma} f_0(\frac{x-\mu}{\sigma})$$

Example 3.3.2 (Example 2.14)

Consider $X \in \mathrm{G}(\mu,2)$ *where* $\mu = E(X)$. *Show that* μ *is a location parameter.*

We can use a similar approach as before and define $Y = X - \mu$ *which follows* $\mathrm{G}(0,2)$. *It is clear that we then have that* F_X, *the cdf of* X, *belongs to a location family of distributions.*

Example 3.3.3 (Example 2.15)

Consider $X \sim \text{Exp}(\theta)$. Show that F_X belongs to a scale family of distributions and find the scale parameter.

Note that

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & x > 0 \\ 0 & \textit{otherwise} \end{cases}$$

Let $Y = \frac{X}{\theta}$. Then

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(\frac{X}{\theta} \leq y) \\ &= P(X \leq \theta y) = \int_0^{\theta y} \frac{1}{\theta} e^{-\frac{x}{\theta}} \, dx \\ &= -e^{-\frac{x}{\theta}} \Big|_0^{\theta} y = 1 - e^{-y} \end{aligned}$$

and we have

$$f_Y(y) = \begin{cases} e^{-y} & y > 0 \\ 0 & \textit{otherwise} \end{cases}$$

Note that if we set $\sigma = 1$ to get f_1, we have

$$f_1(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & \textit{otherwise} \end{cases} .$$

Therefore, F_X belongs to a scale family of distributions.

3.4 Expectations

3.4.1 Expectations

Definition 22 (Expectation of A Discrete RV)

*If X is a discrete rv with pmf f and support set A, then the **expectation** of X, or the **expected value** of X is defined by*

$$E(X) = \sum_{x \in A} x f(x) \tag{3.1}$$

provided that the sum converges absolutely, i.e.

$$E(|X|) = \sum_{x \in A} |x| f(x) < \infty.$$

If $E(|X|)$ does not converge, then we say that $E(X)$ does not exist.

Definition 23 (Expectation of A Continuous RV)

*If X is a continuous rv with pdf f and support set A, then the **expectation** of X, or the **expected value** of X is defined by*

$$E(X) = \int_{x \in A} x f(x) \tag{3.2}$$

provided that the integral converges absolutely, i.e.

$$E(|X|) = \int_{x \in A} |x| f(x) < \infty.$$

If $E(|X|)$ does not converge, then we say that $E(X)$ does not exist.

Example 3.4.1 (Example 2.16)

Suppose $X \sim \mathrm{Poi}(\lambda)$. Calculate $E(X)$.

Solution

Note

$$f(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x = 0, 1, 2, \ldots \\ 0 & \textit{otherwise} \end{cases}.$$

Then

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} \\ &= 0 + \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} \\ &= e^{-\lambda}\lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= e^{-\lambda}\lambda e^{\lambda} = \lambda \end{aligned}$$

Example 3.4.2 (Example 2.18)

Suppose X is an rv with

$$f(x) = \begin{cases} \frac{1}{x^2} & 1 < x < \infty \\ 0 & \text{otherwise} \end{cases}.$$

Calculate $E(X)$.

Solution

Observe that $x \cdot \frac{1}{x^2} = \frac{1}{x}$ and the antiderivative of $\frac{1}{x}$ is $\ln x$, which would need to be evaluated at $\ln \infty$. Thus, we should instead immediately check if the integral converges absolutely.

$$\begin{aligned} E(|X|) &= \int_1^\infty |x| \frac{1}{x^2} \, dx \\ &= \int_1^\infty |x| \frac{1}{|x| \, |x|} \, dx \\ &= \int_1^\infty \frac{1}{|x|} \, dx \\ &= \int_1^\infty \frac{1}{x} \, dx, \end{aligned}$$

and we notice that the integral would not converge. Therefore, $E(X)$ does not exist.

4 Lecture 4 May 10th 2018

4.1 Expectations (Continued)

4.1.1 Expectations (Continued)

Theorem 11 (Expectation from the cdf)

Suppose X is a non-negative continuous rv with cdf F, and $E(X) < \infty$. Then

$$E(X) = \int_0^\infty [1 - F(x)]\, dx = \int_0^\infty P(X \geq x)\, dx \tag{4.1}$$

If X is a discrete rv with cdf F, and $E(X) < \infty$, then

$$E(X) = \sum_{x=0}^{\infty} [1 - F(x)] = \sum_{x=0}^{\infty} P(X \geq x) \tag{4.2}$$

Proof

Note that for a continuous rv X, we have

$$1 - F(x) = P(X \geq x) = \int_x^\infty f(t)\, dt$$

Therefore,

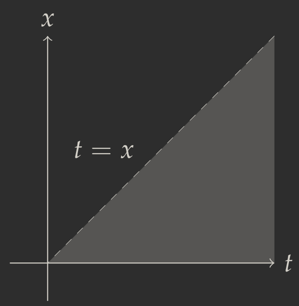
$$\int_0^\infty [1 - F(x)]\, dx = \int_0^\infty \int_x^\infty f(t)\, dt\, dx.$$

Since $1 - F(x)$ is a finite value, so is $\int_0^\infty f(t)\, dt$, and thus we can apply ***Fubini's Theorem***[1]*:*

$$\int_0^\infty [1 - F(x)]\, dx = \int_0^\infty \int_x^\infty f(t)\, dt\, dx = \int_0^\infty \int_0^t f(t)\, dx\, dt$$

Note that the limits of the integral utilizes the following figure:

[1] Condition for Fubini's Theorem to hold is that the integrand of the double integral must be absolutely convergent. See Wikipedia.



With that, note that

$$\int_0^t f(t)\,dx = xf(t)\Big|_0^t = tf(t)$$

Since t is just a dummy variable, we can indeed let $t = x$, and thus we have

$$\int_0^\infty [1 - F(x)]\,dx = \int_0^\infty xf(x)\,dx = E(X)$$

as required.

$\square$

Work on the discrete case as an exercise.

Exercise 4.1.1
For a non-negative discrete rv X with cdf F and $E(X) < \infty$, prove that

$$E(X) = \sum_{x=0}^{\infty} [1 - F(x)]$$

Example 4.1.1 (Example 2.20)

Suppose $X \sim \mathrm{Exp}(\theta)$. Use Theorem 11 to calculate $E(X)$.

Solution

Note that X is a non-negative rv. The cdf of $X \,\mathrm{Im}\, \mathrm{Exp}(\theta)$ is

$$F_X(x) = 1 - e^{-\frac{x}{\theta}}.$$

Then

$$\begin{aligned} E(X) &= \int_0^\infty 1 - F_X(x)\,dx = \int_0^\infty e^{-\frac{x}{\theta}}\,dx \\ &= -\theta e^{-\frac{x}{\theta}}\Big|_0^\infty = \theta \end{aligned}$$

$\square$

Theorem 12 (Expected Value of a Function of X)

Suppose $h(x)$ is a real-valued function.

If X is a discrete rv with pmf f and support set A, then

$$E[h(x)] = \sum_{x \in A} h(x) f(x) \tag{4.3}$$

provided that the sum converges absolutely.

If X is a continuous rv with pdf f, then

$$E[h(x)] = \int_{-\infty}^{\infty} h(x) f(x) \, dx \tag{4.4}$$

provided that the integral converges absolutely.

The proof is, unfortunately, not trivial. One would have to look into Lesbesgue integrals (or at the very least, Riemann-Stieltjes integrals) in order to prove this statement. This "theorem" is also called **The Law of the Unconscious Statistician** [Reference - Wikipedia]. An idea of the proof is given on Math SE.

Example 4.1.2

Suppose $X \sim \mathrm{Unif}(0, \theta)$. Calculate $E(X^2)$.

Solution

$$E(X^2) = \int_0^\theta \frac{x^2}{\theta} \, dx = \frac{1}{\theta} \frac{x^3}{3} \Big|_{x=0}^{\theta} = \frac{\theta^2}{3}$$

Exercise 4.1.2

Find the pdf of $Y = X^2$ and find $E(Y)$ by evaluating $\int_{-\infty}^{\infty} y f_Y(y) \, dy$

Theorem 13 (Linearity of Expectation)

Suppose X is an rv with pf f. Let $a_i, b_i \in \mathbb{R}$, for $i = 1, ..., n$, be constants, and $g_i(x)$, for $i = 1, ..., n$, are real-valued functions. Then

This theorem essentially states that the expectation is a linear operator.

$$E\left[\sum_{i=1}^n (a_i g_i(X) + b_i)\right] = \sum_{i=1}^n (a_i E[g_i(X)] + b_i) \tag{4.5}$$

provided that $E[g_i(X)] < \infty$ for $i = 1, ..., n$.

Proof

Suppose X is a discrete rv with support set A. Then

$$\begin{aligned}
E\left[\sum_{i=1}^{n}(a_i g_i(X)+b_i)\right] &= \sum_{x\in A}\left[\sum_{i=1}^{n}(a_i g_i(x)+b_i)\right] f(x) \qquad \because 1 \text{ Theorem 12} \\
&= \sum_{x\in A}\sum_{i=1}^{n}[a_i g_i(x) f(x) + b_i f(x)] \\
&= \sum_{i=1}^{n}\sum_{x\in A}[a_i g_i(x) f(x) + b_i f(x)] \qquad (*) \\
&= \sum_{i=1}^{n}\left[a_i \sum_{x\in A} g_i(x) f(x) + b_i \sum_{x\in A} f(x)\right] \\
&= \sum_{i=1}^{n}(a_i E[g_i(X)] + b_i)
\end{aligned}$$

where note that $(*)$ is valid because a_i, b_i are constants, and $g_i(x), f(x)$ are finite real-valued functions.

Note

In general, $E(g(X)) \neq g(E(X))$ unless if g is a linear function. For example, for $a, b \in \mathbb{R}$, we have

$$E(aX+b) = aE(X)+b$$

4.1.2 Moments and Variance

Since these concepts were introduced in STAT230 and were given little treatment in the lecture, we shall only cover over them briefly.

Definition 24 (Variance)

*The expectation tof the squared deviation of an rv from its mean is called the **variance**, i.e. for an rv X with mean $\mu = E(X)$,*

$$\sigma^2 = \mathrm{Var}(X) = E[(X-\mu)^2] = E(X^2) - E(X)^2$$

Definition 25 (Moments)

Let X be an rv with mean μ.

*The k^{th} **moment about the origin** is defined as:*

$$E(X^k)$$

*The k^{th} **moment about the mean** is defined as:*

$$E[(X-\mu)^k]$$

*The k^{th} **factorial moment** is defined as:*

$$E[X^{(k)}] = E[X(X-1)\dots(X-k+1)] = E\left[\frac{X!}{(X-k)!}\right]$$

Theorem 14 (Variance of a Linear Function)

Suppose X is an rv with pf f and $a, b \in \mathbb{R}$. Then

$$\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$$

Proof

Observe that

$$\begin{aligned}
\mathrm{Var}(aX+b) &= E[(aX+b)^2] - E(aX+b)^2 \\
&= E[a^2X^2 + 2abX + b^2] - (aE(X)+b)^2 \\
&= a^2E(X^2) + 2abE(X) + b^2 - (a^2E(X)^2 + 2abE(X) + b^2) \\
&= a^2E(X^2) - a^2E(X)^2 = a^2\,\mathrm{Var}(X)
\end{aligned}$$

□

Example 4.1.3 (Example 2.22 (course notes - 2.6.10 (1)))

If $X \sim \mathrm{Poi}(\theta)$, then $E[X^{(k)}] = \theta^k$ for $k = 1, 2, \dots$.

Solution

Note

$$f_X(x) = \begin{cases} \frac{e^{-\theta}\theta^x}{x!} & x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

So

$$\begin{aligned} E[X^{(k)}] &= E(X(X-1)(X-2)\ldots(X-k+1)) \\ &= \sum_{x=0}^{\infty} x(x-1)(x-2)\ldots(x-k+1)\frac{e^{-\theta}\theta^x}{x!} \\ &= 0 + \sum_{x=k}^{\infty} x(x-1)(x-2)\ldots(x-k+1)\frac{e^{-\theta}\theta^x}{x!} \qquad (*) \\ &= \sum_{x=k}^{\infty} \frac{x!}{(x-k)!}\frac{e^{-\theta}\theta^x}{x!} \qquad \because x(x-1)\ldots(x-k+1) = \frac{x!}{(x-k)!} \\ &= e^{-\theta}\theta^k \sum_{x=k}^{\infty} \frac{\theta^{x-k}}{(x-k)!} \\ &= e^{-\theta}\theta^k \sum_{y=0}^{\infty} \frac{\theta^y}{y!} \qquad \text{let } y = x-k \\ &= e^{-\theta}\theta^k e^{\theta} = \theta^k \end{aligned}$$

where for $(*)$ *we have that* $\sum_{x=0}^{k-1} x(x-1)\ldots(x-k+1)A = 0$ *for any* $A \in \mathbb{R}$.

Note that it is not necessarily true that

$$x(x-1)\ldots(x-k+1) = \frac{x!}{(x-k)!}$$

for $0 \leq x \leq k-1$. And so we can only say that the equality is true for $x \geq k$, and hence we have the approach that we use in $(*)$.

4.2 *Inequalities*

4.2.1 *Markov/Chebyshev Style Inequalities*

Theorem 15 (Markov's Inequality)

If X is a non-negative rv and $a > 0$*, then the probability that X is no less than a is no greater than the expectation of X divided by a, i.e.*

$$P(X \geq a) \leq \frac{E(X)}{a} \tag{4.6}$$

Proof

We shall prove for the discrete case. Suppose X is a non-negative discrete rv with pf f. Let $A \subset S$, where S is the sample space, such that $A = \{w \in S : X(w) \geq a\}$.

$$\begin{aligned}
E(X) &= \sum_{x \in S} x f(x) \\
&= \sum_{x \in A} x f(x) + \sum_{x \notin A} x f(x) \\
&\geq \sum_{x \in A} x f(x) \quad \because \sum_{x \notin A} x f(x) \geq 0 \\
&\geq \sum_{x \in A} a f(x) \\
&= a \sum_{x \in A} f(x) = a \cdot P(A) \\
&= a \cdot P(\{w \in S : X(w) \geq a\}) = aP(X \geq a).
\end{aligned}$$

□

Exercise 4.2.1
Prove Markov's Inequality for a continuous rv.

Theorem 16 (Markov's Inequality 2)

If X is a non-negative rv and $a, k > 0$, then the probability that X is no less than a is no greater than the expectation of X divided by a, i.e.

$$P(|X| \geq a) \leq \frac{E(|X|^k)}{a^k} \tag{4.7}$$

Proof

We shall, again, prove for the discrete case. Suppose X is a non-negative discrete rv with pf f. $A := \{w \in S : |X(w)| \geq a\} \subseteq S$. Then

$$\begin{aligned}
E(|X|^k) &= \sum_{x \in S} |x|^k f(x) \\
&= \sum_{x \in A} |x|^k f(x) + \sum_{x \notin A} |x|^k f(x) \\
&\geq \sum_{x \in A} |x|^k f(x) \geq \sum_{x \in A} a f(x) \\
&= a^k P(A) = a^k P(|X| \geq a).
\end{aligned}$$

□

Question: Can we write

$P(\{w \in S : |X(w)| \geq a\}) = P(|X| \geq a)$?

Exercise 4.2.2
Prove for the continuous case.

Theorem 17 (Chebyshev's Inequality)

Suppose X is an rv with finite mean μ and finite variance σ^2. Then for any $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{4.8}$$

Proof

By Theorem 16,

$$P(|X - \mu| \geq k\sigma) \leq \frac{E(|X - \mu|^2)}{(k\sigma)^2} = \frac{1}{k^2}$$

since $E(|X - \mu|^2) = \mathrm{Var}(X) = \sigma^2$. □

Example 4.2.1 (Example 2.23)

A post office handles, on average, 10000 letters a day. What can be said about the probability that it will handle at least 15000 letters tomorrow?

Solution

$X :=$ number of letters handled in a day. Note that by its definition, X is a non-negative discrete rv. Then, using Theorem 15, since $E(X) = 10000$

$$P(X \geq 15000) \leq \frac{10000}{15000} = \frac{2}{3}.$$

Thus, we know that there is less than two-third of chance that the post office will handle more than 15000 tomorrow.

5 Lecture 5 May 15th 2018

5.1 Inequalities (Continued)

5.1.1 Markov/Chebyshev Style Inequalities (Continued)

Example 5.1.1 (Example 2.24)

A post office handles 10000 *letters per day with a variance of* 2000 *letters. What can be said about the probability that this post office handles between* 8000 *and* 12000 *letters tomorrow? What about the probability that more than* 15000 *letters come in (use* ☕ *Theorem 17)?*

1. *Probability that this post office handles between* 8000 *and* 12000 *letters tomorrow:*

$$\begin{aligned}
&P(8000 < X < 12000) \\
&= P(-2000 < X - 10000 < 2000) \\
&= P(|X - 10000| < 2000) = 1 - P(|X - 10000| \geq 2000) \\
&\geq 1 - \frac{1}{(\sqrt{2000})^2} \qquad \because 1 \text{ Theorem 17} \wedge k = \frac{2000}{\sigma} = \sqrt{2000} \\
&= \frac{1999}{2000}
\end{aligned}$$

2. *Probability that more than 15000 letters come in:*

$$\begin{aligned}
P(X > 15000) &= P(X - 10000 > 15000 - 10000) \\
&= P(X - 10000 > 5000) \\
&\leq P(X - 10000 > 5000) + P(X - 10000 < 5000) \\
&\leq P(|X - 10000| > 5000) \\
&\leq \frac{1}{\left(\frac{5000}{\sqrt{2000}}\right)^2} = \frac{2000}{5000^2}
\end{aligned}$$

5.2 Moment Generating Function

Moment generating functions are important because they uniquely define the distribution of an rv.

Definition 26 (Moment Generating Function)

If X is an rv, then $M_X(t) = E(e^{tx})$ is called the ***moment generating function*** *(mgf) of X provided this expectation exists for all $t \in (-h, h)$ for some $h > 0$.*

Note

When determining the mgf of an rv, the values of t for which the expectation exists must always be stated. The range of t where the expectation is defined is "essentially" the ***radius of convergence****.*

Exercise 5.2.1 (Example 2.25 (2.9.2 (1) of the course notes))

Find the mgf of $X \sim \Gamma(\alpha, \beta)$. Make sure you specify the domain on which the mgf is defined.

Solution

Note that the pdf of the Gamma distribution is:

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha - 1} e^{-\frac{X}{\beta}} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Therefore

$$
\begin{aligned}
M_X(t) = E(e^{tx}) &= \int_0^\infty e^{tx} \frac{1}{\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}\, dx \\
&= \frac{1}{\beta^\alpha} \int_0^\infty \frac{1}{\Gamma(\alpha)} x^\alpha e^{-x\left(\frac{1}{\beta}-t\right)}\, dx \\
&= \frac{\left(\frac{\beta}{1-t\beta}\right)^\alpha}{\beta^\alpha} \underbrace{\int_0^\infty \frac{1}{\left(\frac{\beta}{1-t\beta}\right)^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\frac{\beta}{1-t\beta}}}\, dx}_{\text{sum over all values for pdf of } \Gamma(\alpha, \frac{\beta}{1-t\beta}=1)} \quad \text{for } \frac{1}{\beta} - t > 0 \\
&= (1-t\beta)^{-\alpha} \qquad \text{for } t < \frac{1}{\beta}
\end{aligned}
$$

Definition 27 (Indicator Function)

The function $\mathbb{1}_A$ is called the ***indicator function*** *of the set A, i.e.*

$$
\mathbb{1}_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A^C \text{ occurs} \end{cases} \tag{5.1}
$$

Example 5.2.1

The pdf

$$
f(x) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}
$$

can be represented as

$$
f(x) = \frac{1}{\theta} \mathbb{1}_{\{0 \leq x \leq \theta\}}
$$

Example 5.2.2 (Example 2.26)

Find the mgf of $X \sim \mathrm{Poi}(\lambda)$. Make sure you specify the domain on which the mgf is defined.

Solution

Note that the pmf of X is

$$
f_X(x) = \frac{e^{-\lambda} \lambda^x}{x!} \mathbb{1}_{\{0,1,2,\ldots\}}
$$

The mgf is thus

$$\begin{aligned}
M_X(t) = E(e^{tX}) &= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t \lambda)^x}{x!} = e^{-\lambda} e^{e^t \lambda} \\
&= e^{\lambda(e^t - 1)} \qquad \forall t \in \mathbb{R}
\end{aligned}$$

Proposition 18 (Properties of the MGF)

Suppose X is an rv. Then

1. $M_X(0) = 1$
2. *Suppose the derivatives $M_X^{(k)}(t)$, for $k = 1, 2, \ldots$, exists for $t \in (-h, h)$ for some $h > 0$, then the* ***Maclaurin Series***[1] *of $M_X(t)$ is*

$$M_X(t) = \sum_{k=0}^{\infty} \frac{M_X^{(k)}(t)\Big|_{t=0}}{k!} t^k$$

3. *If the mgf exists, then the k^{th} moment of X is:*

$$E(X^k) = \frac{d^k M_X(t)}{dt^k}\Big|_{t=0}$$

4. *Putting 2 and 3 together, we have*

$$M_X(t) = \sum_{k=0}^{\infty} \frac{E(X^k)}{k!} t^k$$

The final item shows why $M_X(t)$ is called the ***moment generating function****.*

[1] The Maclaurin series is the Taylor expansion around 0.

Proof

1. $M_X(t)\Big|_{t=0} = E(e^{tX})\Big|_{t=0} = E(e^0) = 1$
2. *This is simply a result of using the Maclaurin series.*

3. *Note that*

$$\begin{aligned} E(e^{tX}) &= E\left[1 + tX + \frac{1}{2}(tX)^2 + \frac{1}{3!}(tX)^3 + \dots\right] \\ &= 1 + tE(X) + \frac{t^2}{2}E(X^2) + \frac{t^3}{3!}E(X^3) + \dots . \end{aligned}$$

So

$$\frac{d^k}{dt^k}E(e^{tX})\Big|_{t=0} = \frac{k!}{k!}E(X^k) + \underbrace{\frac{k! \cdot t}{(k+1)!}E(X^{k+1}) + \dots\Big|_{t=0}}_{=0 \text{ when } t=0} = E(X^k)$$

$\square$

Example 5.2.3 (Example 2.27)

A discrete random variable X has the pmf

$$f(x) = \left(\frac{1}{2}\right)^{x+1} \mathbb{1}_{\{0,1,2,\dots\}}$$

Derive the mgf of X and use it calculate its mean and variance.

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{\infty} e^{tx}\left(\frac{1}{2}\right)^{x+1} \\ &= \frac{1}{2} \cdot \sum_{x=0}^{\infty}\left(\frac{e^t}{2}\right)^x \\ &= \frac{1}{2} \cdot \frac{1}{1 - \frac{e^t}{2}} \quad \text{for } \left|\frac{e^t}{2}\right| < 1 \text{ or } t < \ln 2 \\ &= \frac{1}{2 - e^t} \end{aligned}$$

To get the first two moments,

$$\begin{aligned} E(X) &= \frac{d}{dt}M_X(t)\Big|_{t=0} \\ &= \frac{e^t}{(2-e^t)^2}\Big|_{t=0}^{=} 1 \\ E(X^2) &= \frac{d^2}{dt^2}M_X(t)\Big|_{t=0} \\ &= \frac{e^t}{(2-e^t)^2} + \frac{2e^t}{(2-e^t)^3}\Big|_{t=0} \\ &= 1 + 2 = 3 \end{aligned}$$

Thus we have that the expected value and variance are

$$\begin{aligned} E(X) &= 1 \\ \mathrm{Var}(X) &= E(X^2) - E(X)^2 = 3 - 1 = 2 \end{aligned}$$

respectively.

5.2.1 *MGF of a Linear Transformation*

Theorem 19 (MGF of a Linear Transformation)

Suppose the rv X has an mgf $M_X(t)$ defined for $t \in (-h, h)$ for some $h > 0$. Let $Y = aX + b$, where $a, b \in \mathbb{R}$ and $a \neq 0$. Then the mgf of Y is

$$M_Y(t) = e^{bt} M_X(at), \quad |t| \leq \frac{h}{|a|}. \tag{5.2}$$

Proof

Observe that

$$M_Y(t) = E(e^{tY}) = E(e^{t(aX+b)}) = E(e^{atX} e^{tb}) = e^{bt} M_X(at).$$

The range of t is

$$|at| < h \overset{a \neq 0}{\iff} |t| < \frac{h}{|a|}$$

Example 5.2.4 (Example 2.28)

Consider $X \sim \mathrm{Unif}(\theta_1, \theta_2)$. Find the mgf of $Y = 5X + 3$.

Solution

Note that

$$
\begin{aligned}
M_X(t) &= \int_{\theta_1}^{\theta_2} \frac{e^{tx}}{\theta_2 - \theta_1}\,dx \\
&= \begin{cases} \frac{e^{tx}}{t(\theta_2-\theta_1)}\Big|_{\theta_1}^{\theta_2} & t \neq 0 \\ 1 & t = 0 \end{cases} \\
&= \begin{cases} \frac{e^{t\theta_2}-e^{t\theta_1}}{t(\theta_2-\theta_1)} & t \neq 0 \\ 1 & t = 0 \end{cases}
\end{aligned}
$$

Thus by Theorem 19,

$$
M_Y(t) = e^{3t} M_X(5t) = \begin{cases} e^{3t}\frac{e^{5t\theta_2}-e^{5t\theta_1}}{5t(\theta_2-\theta_1)} & t \neq 0 \\ 1 & t = 0 \end{cases}
$$

5.2.2 Uniqueness of the MGF

Theorem 20 (Uniqueness of the MGF)

Suppose the rv X has mgf $M_X(t)$ and the rv Y has mgf $M_Y(t)$. Suppose also that $M_X(y) = M_Y(t)$ for all $t \in (-h, h)$ for some $h > 0$. Then X and Y have the same distribution, that is, $\forall s \in \mathbb{R}$,

$$
P(X \leq s) = F_X(s) = F_Y(s) = P(Y \leq s)
$$

Proof

The proof of this theorem is not trivial. See this comment on Math SE for information. It appears that the 2nd bullet point points to a material that I might be able to understand. If I can find that material, and understand it, I may change this proof section to become my own notes.

Example 5.2.5 (Example 2.29)

Suppose $X \sim \mathrm{Unif}(0,1)$*. Define* $Y = -2\log X$*, and use the mgf method to show that* $Y \sim \chi_2^2$*.*
(Hint: Find mgf of χ_2 *and show that* Y *has the same mgf)*

Solution
Let $Z = \chi_2^2$*. The pdf of* Z *is therefore*

$$f_Z(z) = \frac{1}{2}e^{-\frac{z}{2}}\mathbb{1}_{\{z>0\}}.$$

Then

$$\begin{aligned}
M_Z(t) = E(e^{tZ}) &= \int_0^\infty e^{tz}\frac{1}{2}e^{-\frac{z}{2}}\,dz \\
&= \frac{1}{2}\int_0^\infty e^{(t-\frac{1}{2})z}\,dz \\
&= \begin{cases} \frac{1}{2}\frac{1}{t-\frac{1}{2}}e^{(t-\frac{1}{2})z}\Big|_{z=0}^{\infty} & t \neq \frac{1}{2} \\ \infty & t = \frac{1}{2} \end{cases} \\
&= \frac{1}{2t-1} \qquad t \neq \frac{1}{2}
\end{aligned}$$

6 Lecture 6 May 17th 2018

6.1 Joint Distributions

6.1.1 Introduction to Joint Distributions

Note (Motivation)

Most studies collect information for multiple variables per subject rather than just one variable. Because these variables may interfere/interact with each other and hence give us results that may not be fully reliant on a single variable, it is in our interest to study the interaction of these variables.

To start off with the basics, we will first look at the bivariate case of a joint distribution.

6.1.2 Joint and Marginal CDFs

Definition 28 (Joint CDF)

Suppose X and Y are rvs defined on a sample space S. The ***joint cdf*** *of X and Y is given by*

$$\forall (x,y) \in \mathbb{R}^2 \qquad F(x,y) = P(X \leq x, Y \leq y).$$

Note

- *Depending on whether X and Y are both discrete or both continuous, we can derive the joint pmf or joint pdf of (X, Y), respectively.*
- *Definition 28 only concerns two variables (a bivariate case), but we can certainly extend the idea to a k-dimensional joint cdf for the rvs $X_1, X_2, ..., X_k$ as $\forall (x_1, x_2, ..., x_k) \in \mathbb{R}^k$,*

$$F(x_1, x_2, ..., x_k) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_k \leq x_k).$$

Proposition 21 (Properties of Joint CDF)

Suppose X, Y are rvs, either both continuous or discrete, and has a joint cdf F. Then

1. *F is non-decreasing in x for fixed y.*
2. *F is non-decreasing in y for fixed x.*
3. *$\lim_{x \to -\infty} F(x, y) = 0$ and $\lim_{y \to -infty} F(x, y) = 0$.*
4. *$\lim_{(x,y) \to (-\infty, -\infty)} F(x, y) = 0$ and $\lim_{(x,y) \to (\infty, \infty)} F(x, y) = 1$*

Proof

1. *Suppose not, i.e. that we have instead that F is decreasing for x. Then for $x_1 < x_2 \in \mathbb{R}$, we would have*

$$\begin{aligned} &F(x_1, y) > F(x_2, y) \\ &\implies P(X \leq x_1, Y \leq y) > P(X \leq x_2, Y \leq y) \end{aligned}$$

In other words,

$$\begin{aligned} &P(\{(w, v) : (w, v) \in S,\ X(w) \leq x_1,\ Y(v) \leq y\}) \\ &> P(\{(w, v) : (w, v) \in S,\ X(w) \leq x_2,\ Y(v) \leq y\}) \end{aligned}$$

However, note that for fixed y, since $x_1 < x_2$, we must have that

$$\begin{aligned} &\{(w, v) \in S : X(w) \leq x_1, Y(v) \leq y\} \\ &\subseteq \{(w, v) \in S : X(w) \leq x_2, Y(v) \leq y\}. \end{aligned}$$

By Proposition 1, we have that

$$\begin{aligned}&P(\{(w,v) : (w,v) \in S,\ X(w) \leq x_1,\ Y(v) \leq y\})\\ &\leq P(\{(w,v) : (w,v) \in S,\ X(w) \leq x_2,\ Y(v) \leq y\}).\end{aligned}$$

This is clearly a contradiction.

2. *The proof for this statement is similar to the above.*

3. *Note that*

$$\begin{aligned}\lim_{x\to-\infty} F(x,y) &= \lim_{x\to-\infty} P(X \leq x, Y \leq y)\\ &= P(X \leq -\infty, Y \leq y)\\ &= P([X \leq -\infty] \cap [Y \leq y])\\ &= P(\emptyset \cup [Y \leq y]) = P(\emptyset) = 0\end{aligned}$$

   The proof for the case where $y \to -\infty$ is similar.

4. *This is simply a consequence of 3.*

Note

We say that F is a joint cdf if it satisfies all the conditions in Proposition 21.[1]

[1] Many literature actually claims this, and it does look like it will be assumed so for this class.

Example 6.1.1 (Example 3.1)

Consider the following joint cdf of two rvs (X_1, X_2):

$$F(x_1, x_2) = \begin{cases} 0 & x_1 < 0 \lor x_2 < 0 \\ 0.49 & 0 \leq x_1 < 1 \land 0 \leq x_2 < 1 \\ 0.7 & 0 \leq x_1 < 1 \land x_2 > 1 \\ 0.7 & x_1 \geq 1 \land 0 \leq x_2 < 1 \\ 1 & x_1 \geq 1 \land x_2 \geq 1 \end{cases}$$

Flipping an unfair coin with $P(\{H\}) = 0.3$ twice independently, we define *for* $i = 1, 2$

$$X_i = \begin{cases} 1 & \textit{if the } i^{th} \textit{ flip is heads} \\ 0 & \textit{otherwise} \end{cases}$$

The joint cdf of (X_1, X_2) *is the given* F *above. Verify that under this experiment,* F *is indeed a cdf.*

Solution

Note that conditions 3 and 4 of Proposition 21 are automatically satisfied by the definition of F.

incomplete example

Definition 29 (Marginal CDF)

For the rvs X, Y *with joint cdf* F*, the* ***marginal cdf*** *of* X *is*

$$F_X(x) = P(X \leq x) = \lim_{y \to \infty} F(x, y) = F(x, \infty) \quad \forall x \in \mathbb{R}$$

and the ***marginal cdf*** *of* Y *is*

$$F_Y(y) = P(Y \leq y) = \lim_{x \to \infty} F(x, y) = F(\infty, y) \quad \forall y \in \mathbb{R}$$

Note that the marginal cdf is defined for both discrete and continuous cases.

Example 6.1.2

Based on Example 6.1.1, derive $F_{X_i}(x_i)$ *for* $i = 1, 2$.

Solution

$$\begin{aligned} F_{X_1}(x_1) &= \lim_{x_2 \to \infty} F(x_1, x_2) \\ &= \begin{cases} 0 & x_1 < 0 \\ 0.7 & 0 \leq x_1 < 1 \\ 1 & x_1 \geq 1 \end{cases} \end{aligned}$$

The solution for $F_{X_2}(x_2)$ *is similar.*

6.1.3 Joint Discrete RVs

Definition 30 (Joint Discrete RV)

*Suppose X amd Y are rvs defined on a sample space S. If S is discrete then X and Y are discrete rvs. The **joint pmf** of X and Y is given by*

$$\forall (x,y) \in \mathbb{R}^2 \quad f(x,y) = P(X = x, Y = y).$$

*The set $A = \{(x,y) : f(x,y) > 0\}$ is called the **support set** of (X,Y).*

Proposition 22 (Properties of Joint PMF)

Suppose X, Y are discrete rvs with joint pmf f and support set A. Then

1. $\forall (x,y) \in \mathbb{R}^2 \qquad f(x,y) \geq 0$
2. $\sum_{(x,y) \in A} \sum f(x,y) = 1$
3. $\forall R \subset \mathbb{R}^2$,
$$P[(X,Y) \in R] = \sum_{(x,y) \in R} \sum f(x,y)$$

The proof is analogous to the univariate case as seen in Proposition 6

Example 6.1.3 (Example 3.2)

Consider the following joint pmf where the numbers inside the table show $P(X = x, Y = y)$. Find c. Then, calculate $P(X + Y \leq 2)$.

	x = -2	x = 0	x = 2
y = 0	0.05	0.1	0.15
y = 1	0.07	0.11	c
y = 2	0.02	0.25	0.05

Solution

Since the sum of all the probabilities must be 1, thus

$$c = 1 - 0.05 - 0.07 - 0.02 - \ldots - 0.15 - 0.05 = 0.2.$$

Notice that the only cases where $X + Y > 2$ *is when*

- $X = 2, Y = 1$; *and*
- $X = 2, Y = 2$.

Thus

$$\begin{aligned} P(X + Y \leq 2) &= 1 - P(X = 2, Y = 1) - P(X = 2, Y = 2) \\ &= 1 - 0.2 - 0.05 = 0.75 \end{aligned}$$

Example 6.1.4 (Example 3.3)

A small college has 90 *male and* 30 *female professors. An ad hoc committee of* 5 *is selected at random to write the vision and mission of the college. Let* X *and* Y *be the number of men and women in this committee, respectively. Derive the joint distribution of* (X, Y).

Solution

Observe that the support set of this distribution is

$$A = \{(x, y) : x + y = 5, x, y = 0, 1, 2, 3, 4, 5\}.$$

We have that the distribution is

$$P(X = x, Y = y) = \begin{cases} \frac{\binom{90}{x}\binom{30}{y}}{\binom{120}{5}} & \substack{x,y=0,1,2,3,4,5 \\ x+y=5} \\ 0 & \text{otherwise} \end{cases}$$

Definition 31 (Marginal Distribution - Discrete Case)

Suppose X *and* Y *are discrete rvs with joint pf* f. *Then the* ***marginal pf*** *of* X *is*

$$\forall x \in \mathbb{R}^2 \quad f_X(x) = P(X = x) = \sum_{y \in \mathbb{R}} f(x, y),$$

and the ***marginal pf*** *of* Y *is*

$$\forall y \in \mathbb{R}^2 \quad f_Y(y) = P(Y = Y) = \sum_{x \in \mathbb{R}} f(x, y).$$

Example 6.1.5 (Example 3.4)

Consider the joint pmf from Example 6.1.3. Find the marginal distributions, *i.e. marginal pmfs of X and Y.*

	x = -2	x = 0	x = 2
y = 0	0.05	0.1	0.15
y = 1	0.07	0.11	0.2
y = 2	0.02	0.25	0.05

Solution

Using the definition, we have that

$$f_X(x) = \sum_{y \in \mathbb{R}} f(x,y) = \begin{cases} 0.14 & x = -2 \\ 0.46 & x = 0 \\ 0.40 & x = 2 \end{cases}$$

and

$$f_Y(y) = \sum_{x \in \mathbb{R}} f(x,y) = \begin{cases} 0.3 & y = 0 \\ 0.38 & y = 1 \\ 0.32 & y = 2 \end{cases}$$

Example 6.1.6 (Example 3.5)

Suppose that a penny and a nickel are each tossed 10 *times so that every pair of sequences of tosses (n tosses in each sequence) is equally likely to occur. Let X be the number of heads obtained with the penny, and Y be the number of heads obtained with the nickel. It can be shown that (show it!) the joint pmf of X and Y is as follows.*

$$P(X = x, Y = y) = \begin{cases} \binom{10}{x}\binom{10}{y}\left(\frac{1}{2}\right)^{20} & x, y = 0, ..., 10 \\ 0 & \textit{otherwise} \end{cases}$$

Solution

Note that the support set of X and Y are the same, i.e.

$$A_X = A_Y = \{0, 1, ..., 10\}.$$

We may assume that the penny and the nickel are fair coins, i.e. if we let p_x *and* p_y *be the probability of getting a head for a penny and nickel, respectively, then* $p_x = p_y = \frac{1}{2}$. *Since there are* 10 *ways to get* x *heads with the penny, and similarly so for the nickel, we have that*

$$P(X = x, Y = y) = \begin{cases} \binom{10}{x}\binom{10}{y}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^{10} & x, y = 0, 1, ..., 10 \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} \binom{10}{x}\binom{10}{y}\left(\frac{1}{2}\right)^{20} & x, y = 0, 1, ..., 10 \\ 0 & \text{otherwise} \end{cases}$$

as required.

Note

It is interesting to observe that the two rvs in the last example have seemingly no relationship with one another in terms of the experiment conducted, since they do not affect each other. This leads us to introducing the next concept.

6.1.4 Independence of Discrete RVs

Definition 32 (Independence of Discrete RVs)

Two rvs X and Y with joint cdf F are said to be ***independent*** *if and only if*

$$\forall x, y \in \mathbb{R} \quad F(x, y) = F_X(x)F_Y(y)$$

Theorem 23 (Independence by PF)

I am not certain as to why this is presented as a theorem that repeats the definition. As so, the prove for the 2nd equation will not be shown.

Suppose X and Y are rvs with joint cdf F, joint pf f, marginal cdf F_X and F_Y respectively, and marginal pf f_X and f_Y respectively. Also, suppose that $A_X = \{x : f_X(x) > 0\}$ is the support set of X and $A_Y = \{y : f_Y(y) > 0\}$ is the support set of Y. Then X and Y are independent rvs if and only if either

$$\forall (x,y) \in A_X \times A_Y \quad f(x,y) = f_X(x) f_Y(y)$$

holds, or

$$\forall x, y \in \mathbb{R} \quad F(x,y) = F_X(x) F_Y(y)$$

Proof

*The ($\implies$) direction is simply a result of **Clairaut's Theorem**[2]. While the ($\impliedby$) direction is a direct result of applying double integrals.* □

[2] Work needs to be done to show that our statement actually satisfies the condition for Clairaut's Theorem to apply. Clairaut's Theorem states that:

Theorem 24 (Clairaut's Theorem)
If (x_0, y_0) is a point in the domain of a function f with

- *f is defined on all points in an open disk centered at (x_0, y_0);*
- *the first partial derivatives, f_{xy} and f_{yx} are all continuous for all points in the open disk.*

Then $f_{xy}(x_0, y_0) = f_{yx}(x_0, y_0)$.

Example 6.1.7 (Example 3.6)

Suppose X and Y are discrete rvs with joint pf

$$f(x,y) = \frac{\theta^{x+y} e^{-2\theta}}{x! y!} \mathbb{1}_{\{x,y=0,1,\ldots\}}.$$

Are X and Y independent of each other?

Solution

Note that we may write f as

$$f(x,y) = \left(\frac{\theta^x e^{-\theta}}{x!} \cdot \frac{\theta^y e^{-\theta}}{y!} \right) \mathbb{1}_{\{x,y=0,1,\ldots\}}$$

and so this suggests that we can indeed break down f into two parts, each only affected by x and y respectively, "indenpdent" of each other. Indeed, *since*

$$\begin{aligned} f_X(x) &= \sum_{y=0}^{\infty} \frac{\theta^{x+y} e^{-\theta}}{x!y!} \mathbb{1}_{\{x,y=0,1,\ldots\}} \\ &= \sum_{y=0}^{\infty} \left(\frac{\theta^x e^{-\theta}}{x!} \cdot \frac{\theta^y e^{-\theta}}{y!} \right) \mathbb{1}_{\{x=0,1,\ldots\}} \\ &= \frac{\theta^x e^{-\theta}}{x!} \underbrace{\sum_{y=0}^{\infty} \frac{\theta^y e^{-\theta}}{y!}}_{\textit{sum of pmf of } \mathrm{Poi}(\theta)=1} \\ &= \frac{\theta^x e^{-\theta}}{x!} \end{aligned}$$

Similarly, we can obtain

$$f_Y(y) = \frac{\theta^y e^{-\theta}}{y!}$$

Multiplying $f_X(x)$ and $f_Y(y)$ together, we indeed get back to the original joint pmf.

7 Lecture 7 May 24th 2018

7.1 Joint Distributions (Continued)

7.1.1 Independence of Discrete RVs (Continued)

Example 7.1.1 (Example 3.7)

Consider the joint pmf below from Example 6.1.3. Are X and Y independent? Prove or disprove.

	x = -2	x = 0	x = 2	$P(Y = y)$
y = 0	0.05	0.1	0.15	0.3
y = 1	0.07	0.11	0.2	0.38
y = 2	0.02	0.25	0.05	0.32
$P(X = x)$	0.14	0.46	0.4	

Solution

Note that

$$P(X = -2, Y = 0) = 0.5 \text{ but}$$
$$P(X = -2)P(Y = 0) = 0.14 \cdot 0.3 = 0.042 \neq 0.5.$$

Thus X and Y are not independent.

7.1.2 Joint Continuous RVs

Definition 33 (Joint Continuous RVs)

Two random variables X and Y are said to be ***jointly continuous*** *if there exists a function $f(x, y)$ such that the joint cdf of X and Y can be written as*

$$\forall (x, y) \in \mathbb{R}^2 \quad F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(t_1, t_2) d_{t_2} d_{t_1}.$$

The function f is called the ***joint density function*** *of X and Y. It follows from the above defintiion that when the second partial derivative exists, we have*

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

The set $\{(x, y) : f(x, y) > 0\}$ is called the ***support set*** *of (X, Y).*

Note (Convention)

Define $f(x, y) = 0$ when $\frac{\partial^2}{\partial x \partial y} F(x, y)$ does not exist.

Example 7.1.2 (Example 3.8)

Suppose X and Y have joint pdf $f(x, y) = \mathbb{1}_{\{0<x,y<1\}} = \mathbb{1}_{\{0<x<1, 0<y<1\}}$. Calculate the joint cdf of X and Y.

Solution

$$F(x, y) = \begin{cases} 0 & x \leq 0, \vee y \leq 0 \\ \int_0^x \int_0^y 1 \, ds \, dt = xy & 0 < x < 1 \wedge 0 < y < 1 \\ \int_0^1 \int_0^y 1 \, ds \, dt = y & x \geq 1 \wedge 0 < y < 1 \\ \int_0^x \int_0^1 1 \, ds \, dt = x & 0 < x < 1 \wedge y \geq 1 \\ \int_0^1 \int_0^1 1 \, ds \, dt = 1 & x \geq 1 \wedge y \geq 1 \end{cases}$$

Proposition 25 (Properties of Joint PDF)

1. $\forall (x, y) \in \mathbb{R}^2 \quad f(x, y) \geq 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = 1$

3. $\forall B \subset \mathbb{R}^2$,

$$P[(X,Y) \in B] = \int\int_{(x,y)\in B} f(x,y)\,dx\,dy$$

Proof

Example 7.1.3 (Example 3.9)

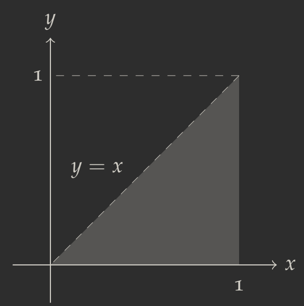
Suppose that $f(x,y) = Kxy \cdot \mathbb{1}_{\{0<xy<1\}}$ for some constant $K > 0$. Find K so that f is a valid joint pdf. If X and Y have the joint density f, calculate $P(X > Y)$.

Solution

Note that

$$1 = \int_0^1 \int_0^1 Kxy\,dx\,dy = \frac{K}{4}.$$

Thus $K = 4$. To solve the next part, observe that for $X > Y$, we have the diagram to the right to show the support set of the joint distribution. The shaded region is the support set. We then have



$$\begin{aligned} P(X > Y) &= \int_0^1 \int_0^x 4xy\,dy\,dx = \int_0^1 2xy^2\Big|_0^x\,dx \\ &= \int_0^1 2x^3\,dx = \frac{1}{2}x^3\Big|_0^1 = \frac{1}{2} \end{aligned}$$

Example 7.1.4 (Example 3.10)

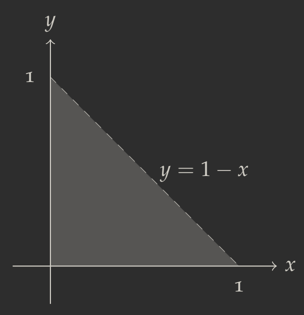
Suppose that

$$f(x,y) = \begin{cases} Cxy & 0 < x,y < 1,\ x+y < 1 \\ 0 & \textit{otherwise} \end{cases}$$

Find C so that $f(x,y)$ is a valid joint probability density function, and calculate $P(Y^2 < X)$.

Solution

Note that the diagram on the right shows the support set of (X,Y). To find



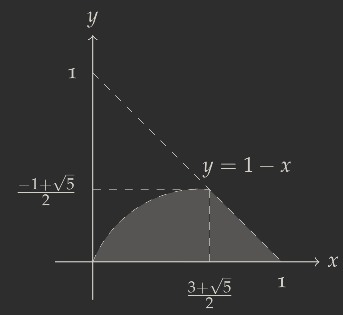
C,

$$\begin{aligned} 1 &= \int_0^1 \int_0^{1-x} Cxy\,dy\,dx = \int_0^1 \frac{C}{2} xy^2 \Big|_0^{1-x} dx \\ &= C \int_0^1 \frac{1}{2} x(x^2 - 2x + 1)\,dx = C \int_0^1 \frac{1}{2}(x^3 - 2x^2 + x)\,dx \\ &= C \left(\frac{1}{8}x^4 - \frac{1}{3}x^3 + \frac{1}{4}x^2 \right) \Big|_0^1 = C \left(\frac{3}{24} - \frac{8}{24} + \frac{6}{24} \right) = \frac{C}{24}. \end{aligned}$$

And so $C = 24$.

To calculate $P(Y^2 < X)$, note the diagram to the right. Then

$$\begin{aligned}
P(Y^2 < X) &= \int_0^{\frac{3+\sqrt{5}}{2}} \int_0^{\sqrt{x}} 24xy \, dy \, dx + \int_{\frac{3+\sqrt{5}}{2}}^1 \int_0^{1-x} 24xy \, dy \, dx \\
&= \int_0^{\frac{3+\sqrt{5}}{2}} 12xy^2 \Big|_0^{\sqrt{x}} dx + \int_{\frac{3+\sqrt{5}}{2}}^1 12xy^2 \Big|_0^{1-x} dx \\
&= 4x^3 \Big|_0^{\frac{3+\sqrt{5}}{2}} + \int_{\frac{3+\sqrt{5}}{2}}^1 24(x^3 - 2x^2 + x) \, dx \\
&= 4\left(\frac{3+\sqrt{5}}{2}\right)^3 + 24\left[\frac{1}{4}x^4 - \frac{2}{3}x^3 + \frac{1}{2}x^2\right]\Big|_{\frac{3+\sqrt{5}}{2}}^1 = \dots
\end{aligned}$$



Solve for $y = 1 - x$ and $y^2 = x$ to get the intersection.

We shall not proceed to get the final solution since it is a messy process and the result is not important.

7.1.3 *Marginal Distribution (Continuous)*

Definition 34 (Marginal PDF)

Suppose X and Y are continuous rvs with joint pdf f. Then the marginal pdf of X is given by

$$\forall x \in \mathbb{R} \quad f_X(x) = \int_{-\infty}^{\infty} f \, dy,$$

and the marginal pdf of Y is

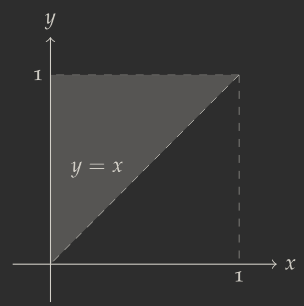
$$\forall y \in \mathbb{R} \quad f_Y(y) = \int_{-\infty}^{\infty} f \, dx.$$

Example 7.1.5 (Example 3.11)

Suppose X and Y have joint pdf $f(x,y) = K(x+y)\mathbb{1}_{0 \leq x < y \leq 1}$ for some constant K. Find K. Then, calculate the marginal density of X.

Solution

A diagram showing the region of the support set is on the right.



To get K,

$$\begin{aligned} 1 &= \int_0^1 \int_x^1 K(x+y)\,dy\,dx = \int_0^1 \left(Kxy + \frac{1}{2}Ky^2 \right) \Big|_x^1 \,dx \\ &= \int_0^1 Kx + \frac{K}{2} - Kx^2 - \frac{1}{2}Kx^2 \,dx \\ &= \frac{K}{2}\left(x^2 + x - x^3\right)\Big|_0^1 = \frac{K}{2} \end{aligned}$$

Thus $K = 2$.

To get the marginal density of X, note that our joint pdf is now the following:

$$f(x,y) = 2(x+y)\mathbb{1}_{\{0 \leq x < y \leq 1\}}$$

Thus

$$\int_x^1 2(x+y)\,dy = 2xy + y^2\Big|_x^1 = 2x + 1 - 3x^2$$

And hence

$$f_X(x) = \begin{cases} -3x^2 + 2x + 1 & 0 \leq x \leq 1 \\ 0 & \textit{otherwise} \end{cases}$$

7.1.4 *Independence of Continuous RVs*

Definition 35 (Independence of Continuous RVs)

Two random variables X and Y with joint cdf F and joint pdf f are independent iff

$$\forall x, y \in \mathbb{R} \quad F(x,y) = F_X(x)F_Y(y)$$

or[1]

$$\forall x, y \in \mathbb{R} \quad f(x,y) = f_X(x)f_Y(y).$$

[1] It's really an "AND"

Note

A necessary, but insufficient, condition for X and Y to be independent is that

$$\operatorname{supp}(X,Y) = \operatorname{supp}(X) \times \operatorname{supp}(Y)$$

Example 7.1.6 (Example 3.12)

Are random variables X and Y introduced in Example 7.1.5 independent? Explain.

Solution

Recall that the pdf was given as

$$f(x, y) = 2(x + y)\mathbb{1}_{\{1 \leq x < y \leq 1\}}.$$

We derived the marginal pdf of X in the earlier example:

$$f_X(x) = (-3x^2 + 2x + 1)\mathbb{1}_{\{0 \leq x \leq 1\}}.$$

To get the marginal pdf of Y, note

$$\int_0^y 2(x + y)\, dx = x^2 + 2xy \Big|_0^y = 3y^2.$$

Thus

$$f_Y(y) = \begin{cases} 3y^2 & 0 \leq y \leq 1 \\ 0 & \textit{otherwise.} \end{cases}$$

Note that

$$f_X(x) f_Y(y) = -9x^2y^2 + 6xy^2 + 3y^2 \quad 0 \leq x < y \leq 1$$

which is not equal to f. Thus, X and Y are not independent.

8 *Lecture 8 May 29th 2018*

8.1 *Joint Distributions (Continued 2)*

8.1.1 *Independence of Continuous RVs (Continued)*

Example 8.1.1 (Example 3.12 (3.4.8 course note))

Suppose X and Y are continuous with joint pdf

$$f(x,y) = \frac{3}{2}y(1-x^2)\mathbb{1}_{\{-1\leq x\leq 1\}}\mathbb{1}_{\{0\leq y\leq 1\}}$$

Are X and Y independent?

Solution

The marginal pdf of X is

$$\begin{aligned} f_X(x) &= \int_0^1 \frac{3}{2}y(1-x^2)\,dy = \frac{3}{4}y^2(1-x^2)\Big|_0^1 \\ &= \begin{cases} \frac{3}{4}(1-x^2) & -1\leq x\leq 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

The marginal pdf of Y is

$$\begin{aligned} f_Y(x) &= \int_{-1}^1 \frac{3}{2}y(1-x^2)\,dx = \frac{3}{2}y\left(x-\frac{1}{3}x^3\right)\Big|_{-1}^1 \\ &= \begin{cases} 2y & 0\leq y\leq 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Clearly, we have

$$f_X(x)f_Y(y) = \frac{3}{2}y(1-x^2) = f(x,y) \quad -1 \leq x \leq 1, 0 \leq y \leq 1.$$

Thus X and Y are independent.

Theorem 26 (Factorization Theorem for Independence)

Suppose X and Y are rvs with joint pf f, and marginal pf f_X and f_Y, respectively. Suppose also that

$$\begin{aligned} A &= \{(x,y) : f(x,y) > 0\} \text{ is the support set of } (X,Y) \\ A_X &= \{x : f_X(x) > 0\} \text{ is the support set of } X \text{, and} \\ A_Y &= \{y : f_Y(y) > 0\} \text{ is the support set of } Y \end{aligned}$$

Then X and Y are independent rvs iff $A = A_X \times A_Y$ and there exist non-negative functions g and h such that

$$f(x,y) = g(x)h(y)$$

for all $(x,y) \in A_X \times A_Y$.

Proof

The $\implies$ direction is straightforward: Since X and Y are independent, we have that $f = f_X f_Y$, and so clearly, $A = A_X \times A_Y$ and so $\forall (x,y) \in A = A_x \times A_Y$, we have that f_X and f_Y are non-negative.

For the $\impliedby$ direction, note that

$$\begin{aligned} f_Y(y) &= \int_{x \in A_X} g(x)h(y)\,dx = h(y)\int_{x\in A_X} g(x)\,dx \\ f_X(x) &= \int_{y\in A_Y} g(x)h(y)\,dy = g(x)\int_{y \in A_Y} h(x)\,dy. \end{aligned}$$

Thus,

$$\begin{aligned} f_X(x)f_Y(y) &= g(x)h(y)\int_{x\in A_X} g(x)\,dx \int_{y \in A_Y} h(y)\,dy \\ &= g(x)h(x)\int_{x \in A_X}\int_{y\in A_Y} g(x)h(y)\,dy\,dx = g(x)h(y) \end{aligned}$$

where line 2 is by ***linearity of integration****. Thus $f(x,y) = f_X(x)f_Y(y)$. Thus X and Y are independent.* $\square$

Note

1. *If Theorem 26 holds, then f_X will be proportional to g and f_Y will be proportional to h. Clearly so, since*

$$g(x) \cdot h(y) = f_X(x) f_Y(y)$$
$$g(x) \propto f_X(x) \wedge h(y) \propto f_Y(y)$$

2. *The definitions and theorems can be easily extended to the random vector $(X_1, X_2, ..., X_n)$. Indeed, if we apply mathematical induction on the proof above, we will be able to get our desired result.*[1]

[1] I wonder if this statement is equivalent to the Fisher-Neyman Factorization Theorem.

8.1.2 Conditional Distributions

Definition 36 (Conditional Distributions)

Suppose X and Y are rvs with joint pf f, and marginal pfs f_X and f_Y, respectively. Suppose also that $A = \{(x, y) : f(x, y) > 0\}$. The ***conditional pf*** *of X given $Y = y$ is given by*

$$f_X(x|y) = \frac{f(x, y)}{f_Y(y)}$$

for $(x, y) \in A$ provided that $f_Y(y) \neq 0$. The ***conditional pf*** *of Y given $X = x$ is given by*

$$f_Y(y|x) = \frac{f(x, y)}{f_X(x)}$$

for $(x, y) \in A$ provided that $f_X(x) \neq 0$.

Remark

If X and Y are discrete rvs then

$$f_X(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x, y)}{f_Y(y)}$$

and

$$\sum_x f_X(x|y) = \sum_x \frac{f(x, y)}{f_Y(y)} = \frac{1}{f_Y(y)} \sum_x f(x, y) = \frac{f_Y(y)}{f_Y(y)} = 1,$$

and similarly so for $f_Y(y|x)$. Similarly, if X and Y are both continuous rvs, then

$$\int_{-\infty}^{\infty} f_X(x|y)\,dx = \int_{-\infty}^{\infty} \frac{f(x,y)}{f_Y(y)}\,dx = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f(x,y)\,dx = \frac{f_Y(y)}{f_Y(y)} = 1$$

Now consider if X is a continuous rv such that $f_X(x) \neq P(X = x)$ and $P(X = x) = 0$ for all x. Then to justify the definition of the conditional pdf of Y given $X = x$, when X and Y are both continuous rvs, we consider $P(Y \leq y|X = x)$ using the limit approach:

$$\begin{aligned}
P(Y \leq y|X = x) &= \lim_{h\to 0} P(Y \leq y|x \leq X \leq x+h) \\
&= \lim_{h\to 0} \frac{\int_x^{x+h} \int_{-\infty}^{y} f(u,v)\,dv\,du}{\int_x^{x+h} f_X(u)\,du} \\
&= \lim_{h\to 0} \frac{\frac{d}{dh}\int_x^{x+h} \int_{-\infty}^{y} f(u,v)\,dv\,du}{\frac{d}{dh}\int_x^{x+h} f_X(u)\,du} \quad \text{by \textbf{L'Hôpital's Rule}} \\
&= \lim_{h\to 0} \frac{\int_{-\infty}^{y} \frac{d}{dh}\int_x^{x+h} f(u,v)\,du\,dv}{\frac{d}{dh}\int_x^{x+h} f_X(u)\,du} \quad (1) \\
&= \lim_{h\to 0} \frac{\int_{-\infty}^{y} f(x+h,v)\,dv}{f_X(x+h)} \quad (2) \\
&= \frac{\int_{-\infty}^{y} f(x,v)\,dv}{f_X(x)}
\end{aligned}$$

*where (1) is by assuming that the integrands are all convergent so that we may interchange the integral signs and the differential operator, and (2) by the **Fundamental Theorem of Calculus**. If we differentiate the last line with respect to y, by the Fundamental Theorem of Calculus, we have*

$$\frac{d}{dy} P(Y \leq y|X = x) = \frac{f(x,y)}{f_X(x)}$$

which justifies the using of our definition

$$f_Y(y|x) = \frac{f(x,y)}{f_X(x)}.$$

Example 8.1.2

A fair coin is flipped 10 times independently.

1. *What is the distribution of Y, the number of heads in 10 flips?*
2. *Suppose the first 4 flips have all landed on tails. What is the distribution of Y given this information?*

Solution

1. *Clearly, we know that* $Y \sim \text{Bin}\left(10, \frac{1}{2}\right)$.

2. *Since each flip is independent of each other and the first four flips have already been determined, the range of values for* Y *changes from* $\{0, ..., 10\}$ *to* $\{0, ..., 6\}$. *Since the experiment is still essentially the same, we have that*
$$Y \mid \textit{first 4 flips are tails} \sim \text{Bin}\left(6, \frac{1}{2}\right).$$

Example 8.1.3 (Example 3.13)

Consider the experiment carried out in Example 8.1.2. Let

$$\begin{aligned} X &:= \textit{ number of heads in the first 4 flips} \\ Y &:= \textit{ number of heads in 10 flips} \end{aligned}$$

Derive the conditional distribution of Y *given that the first 4 flips landed on heads, i.e. derive the distribution for* $Y|X = 4$.

Solution

Let W *be the number of heads in the last 6 flips. Then* W *has the same distribution as in part 2 of our earlier example. Also,* $X \sim \text{Bin}\left(4, \frac{1}{2}\right)$ *Clearly,* $Y = X + W$. *We proceed to derive the joint pf of* X *and* Y:

$$\begin{aligned} P(X = x, Y = y) &= P(X = x, X + W = y) = P(X = x, W = y - x) \\ &= P(X = x)P(W = y - x) \quad \textit{by } \textbf{\textit{Independence}} \\ &= \binom{4}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{4-x} \cdot \binom{6}{y-x}\left(\frac{1}{2}\right)^{y-x}\left(\frac{1}{2}\right)^{6-y+x} \\ &= \binom{4}{x}\binom{6}{y-x}\left(\frac{1}{2}\right)^{10} \end{aligned}$$

Then

$$\begin{aligned} P(Y|X = 4) &= \frac{P(X = 4, Y = y)}{P(X = 4)} \\ &= \frac{\binom{4}{4}\binom{6}{y-4}\left(\frac{1}{2}\right)^{10}}{\binom{4}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0} \\ &= \binom{6}{y-4}\left(\frac{1}{2}\right)^6 \quad y \in \{4, 5, ..., 10\}. \end{aligned}$$

We may also re-label the conditional distribution to have

$$\binom{6}{y^*}\left(\frac{1}{2}\right)^6 \quad y^* \in \{0, ..., 6\}$$

Example 8.1.4 (Example 3.14)

From Example 7.1.5, we had that

$$f(x, y) = 2(x + y)\mathbb{1}_{\{0 \leq x < y \leq 1\}}$$

and the marginal density of X is

$$f_X(x) = (2x - 3x^2 + 1)\mathbb{1}_{\{0 \leq x < 1\}}.$$

Derive the conditional distribution of $Y|X = \frac{1}{2}$.

Solution

Observe that

$$f(y|X = \frac{1}{2}) = \frac{f\left(\frac{1}{2}, y\right)}{f_X\left(\frac{1}{2}\right)} = \frac{2\left(\frac{1}{2} + y\right)}{2\left(\frac{1}{2}\right) - 3\left(\frac{1}{2}\right)^2 + 1} = \frac{8}{13}(1 + 2y)$$

for $\frac{1}{2} < y \leq 1$.

Proposition 27 (Properties of Conditional Distributions)

Let X and Y be rvs. If both X and Y are discrete, then

- $\sum_x f(x|y) = 1$;
- $F(x|y) = \sum_{\{w : w \leq x\}} f(w|y)$; *and*
- $f(x|y) = F(x|y) - F(x^-|y)$.

If X and Y are both continuous, then

- $\int_x f(x|y)\, dx = 1$;
- $F(x|y) = \int_{-\infty}^{x} f(t|y)\, dt$; *and*
- $f(x|y) = \frac{\partial}{\partial x} F(x|y)$

Exercise 8.1.1
Prove Proposition 27.

Theorem 28 (Product Rule)

Suppose X and Y are rvs with joint pf f, marginal pfs $f_X(x)$ and $f_Y(y)$ respectively, and conditional pfs $f_X(x|y)$ and $f_Y(y|x)$ respectively. Then

$$f(x, y) = f_X(x|y) f_Y(y) = f_Y(y|x) f_X(x).$$

Proof

Notice once and for all that by rearranging the definition of conditional distribution

$$f_X(x|y)f_Y(y) = f(x,y) = f_Y(y|x)f_X(x)$$

□

Proposition 29 (Independence from Conditionality)

Suppose X and Y are rvs with marginal pfs $f_X(x)$ and $f_Y(y)$ respectively, and conditional pfs $f_X(x|y)$ and $f_Y(y|x)$ respectively. Let $A_X = \{x : f_X(x) > 0\}$ and $A_Y = \{y : f_Y(y) > 0\}$. X and Y are independent rvs iff either of the following holds:

$$\forall x \in A_X \quad f_X(x|y) = f_X(x)$$

or

$$\forall y \in A_Y \quad f_Y(y|x) = f_Y(y).$$

Proof

Suppose that X and Y are independent rvs. Then

$$f(x,y) = f_X(x)f_Y(y).$$

Then

$$\begin{aligned} f_X(x|y) &= \frac{f(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x) \\ f_Y(y|x) &= \frac{f(x,y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_Y(y). \end{aligned}$$

for $x \in A_X$ and $y \in A_Y$.

Suppose WLOG that $f_X(x|y) = f_X(x)$. Thus by Theorem 28

$$f(x,y) = f_X(x|y)f_Y(y) = f_X(x)f_Y(y) \text{ for } x \in A_X,\ y \in A_Y.$$

□

Example 8.1.5 (Example 3.15)

In a game of chance, a random number is generated from $P \sim \text{Beta}(\alpha, \beta)$. Given $P = p$, a coin with $P(\{H\}) = p$ is flipped independently n times, where the player is rewarded the same amount of dollars as the number of heads in n. Calculate the probability that a random player earns at least $1 in this game.

Solution

Let X be the number of heads that appear in n flips, which equates to the total amount of $1 earned. Then

$$X \mid P = p \sim \text{Bin}(n, p)$$

However, note that

$$P(X \geq 1) = P(\textit{earn at least \$1}) = 1 - P(\textit{earn nothing}) = 1 - P(X = 0)$$

To get $P(X = x)$, we need to do the following: note that the support set of P is from 0 to 1, then

$$\begin{aligned}
Pr(X = x) &= \int_0^1 Pr(X = x, P = p)\, dp \\
&= \int_0^1 Pr(X = x \mid P = p) Pr(P = p)\, dp \qquad \textit{by } 1 \;\; \textit{Theorem } 28 \\
&= \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \frac{p^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\, dp \\
&= \binom{n}{x} \frac{1}{B(\alpha,\beta)} \int_0^1 p^{\alpha+x-1}(1-p)^{\beta+n-x-1}\, dp \\
&= \binom{n}{x} \frac{B(\alpha+x, \beta+n-x)}{B(\alpha,\beta)} \int_0^1 \underbrace{\frac{p^{\alpha+x-1}(1-p)^{\beta+n-x-1}}{B(\alpha+x,\beta+n-x)}}_{\textit{pdf of } \text{Beta}(\alpha+x,\beta+n-x)}\, dp \\
&= \binom{n}{x} \frac{B(\alpha+x, \beta+n-x)}{B(\alpha,\beta)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
P(X \geq 1) = 1 - P(X = 0) &= 1 - \binom{n}{0} \frac{\Gamma(\alpha)\Gamma(\beta+n)}{\Gamma(\alpha+\beta+n)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= 1 - \frac{\Gamma(\alpha+\beta)\Gamma(\beta+n)}{\Gamma(\alpha+\beta+n)\Gamma(\beta)}
\end{aligned}$$

Definition 37 (Beta Distribution)

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

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

where

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

8.1.3 Joint Expectations

Definition 38 (Joint Expectation)

Suppose $h(x, y)$ is a real-valued function. If X and Y are discrete rvs with joint pf f and support A, then

$$E[h(x,y)] = \sum_{(x,y)\in A}\sum h(x,y)f(x,y).$$

provided that the joint sum converges absolutely.

If X and Y are continuous rvs with joint pf f, then

$$E[h(x,y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f(x,y)\,dx\,dy$$

provided that the joint integral converges absolutely.

*This is also known as the **Law of the Unconscious Statistician** for two rvs.*

Example 8.1.6 (Example 3.16)

Consider X and Y with the following joint probability distribution. Calculate $E(XY)$.

	x = -2	x = 0	x = 2
y = 0	0.05	0.1	0.15
y = 1	0.07	0.1	0.2
y = 2	0.02	0.25	0.05

Solution

$$\begin{aligned} E(XY) &= \sum_x \sum_y xyf(x,y) \\ &= -2(1)(0.07) - 2(2)(0.02) + 2(1)(0.2) + 2(2)(0.05) \\ &= 0.38 \end{aligned}$$

9 Lecture 9 May 31st 2018

9.1 Joint Distributions (Continued 3)

9.1.1 Joint Expectations (Continued)

Theorem 30 (Linearity of Expectation in Bivariate Case)

Suppose X and Y are two rvs with joint pf f, a_i, b_i, for $i = 1, ..., n$, are constants, and $g_i(x, y)$, for $i = 1, ..., n$, are real-valued functions. Then

$$E\left[\sum_{i=1}^{n}(a_i g_i(X, Y) + b_i)\right] = \sum_{i=1}^{n}(a_i E[g_i(X, Y)]) + \sum_{i=1}^{n} b_i$$

provided that $E[g_i(X, Y)]$ is finite for $i = 1, ..., n$.

Proof

This is simply an extension of Theorem 13.

Theorem 31 (Implication of Independence on Joint Expectation)

If X and Y are independent rvs with joint pf f, and $g(x)$ and $h(y)$ are real valued funcctions, then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

Proof

We shall prove for the discrete case and leave the continuous case for future exercises.

Observe that

$$\begin{aligned} E[g(X)h(Y)] &= \sum_x \sum_y g(x)h(y)f(x,y) \quad \because \text{ Definition 38} \\ &= \sum_x \sum_y g(x)h(y)f_X(x)f_Y(y) \\ &= \sum_x g(x)f_X(x) \sum_y h(y)f_Y(y) \\ &= E[g(X)]E[h(Y)] \end{aligned}$$

where $f_X(x)$ and $f_Y(y)$ are the marginal pfs of X and Y respectively. □

Exercise 9.1.1
Prove Theorem 31 for the continuous case.

We may repeatedly apply the above proof for n rvs through induction and get the following result.

Theorem 32 (Generalized Implication of Independence on Joint Expectation)

If $X_1, X_2, \ldots, X_n$, for some $n \in \mathbb{N}$, are independent rvs and $h_1, h_2, \ldots, h_n$ are real valued functions, then

$$E\left[\prod_{i=1}^n h_i(X_i)\right] = \prod_{i=1}^n E[h_i(X_i)].$$

9.1.2 Covariance

Independence of two rvs X and Y implies that knowledge of the value of X does not provide any information whatsoever about the distribution of Y. Essentially, we can say that there is no "relationship" between X and Y. In statistics, **linear relationships** are often the subject of interest. The strength of a linear relationship is related to **covariance** and measured by the **correlation coefficient**, usually denoted by ρ.

It can be shown that when X and Y have no linear relationship iff their covariance is 0.

On a related thought, does covariance relate to independence? If so, how?

Definition 39 (Covariance)

*The **covariance** of rvs X and Y is given by*

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y. \tag{9.1}$$

where $\mu_X = E[X]$ *and* $\mu_Y = E[Y]$.

If $\mathrm{Cov}(X,Y) = 0$, *then X and Y are called **uncorrelated** rvs.*

Note

Note that the 2nd and 3rd term are equivalent in Equation (9.1) since

$$\begin{aligned} E[(X-\mu_X)(Y-\mu_Y)] &= E[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] \\ &= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X\mu_Y \because 1\ \textit{Theorem } 30 \\ &= E[XY] - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y = E[XY] - \mu_X\mu_Y \end{aligned}$$

From here, it is easy to see from Theorem 31, that since $E[XY] = E[X]E[Y] = \mu_X\mu_Y$, *we have that the independence of X from Y will imply that* $\mathrm{Cov}(X,Y) = 0$.

*However, the converse of the above is **not true**.*

Example 9.1.1

Source: Stats SE

Let X be an rv that it is -1 *or* 1 *with probability* 0.5. *Then let Y be an rv such that* $Y = 0$ *if* $X = -1$, *and Y is randomly* -1 *or* 1 *with probability* 0.5 *if* $X = 1$.

Clearly X and Y are highly dependent (since knowing Y allows me to *perfectly know X). They both have zero mean:*

$$E[X] = -1\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = 0$$
$$E[Y] = -1\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = 0$$

and

$$\begin{aligned} E[XY] &= (-1)\cdot 0 P(X=-1, Y=0) + 1(-1)\cdot P(X=1, Y=-1) \\ &\quad + 1(1)P(X=1, Y=1) \\ &= -\frac{1}{4} + \frac{1}{4} = 0 \end{aligned}$$

Thus $\mathrm{Cov}(X, Y) = 0$

Or more generally, take any distribution $P(X)$ *and any* $P(Y|X)$ *such that* $P(Y = a|X) = P(Y = -a|X)$ *for all X (i.e., a joint distribution that is symmetric around the x axis), and you will always have zero covariance. But you will have non-independence whenever* $P(Y|X) \neq P(Y)$*, i.e., the conditionals are not all equal to the marginal, and vice versa for symmetry around the y axis.*

Note

- *If* $\mathrm{Cov}(X, Y) = 0$*, then X and Y are called* ***uncorrelated*** *rvs.*
- *By definition,* $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$*, since*

$$\mathrm{Cov}(X, X) = E[(X - \mu_X)^2] = \mathrm{Var}(X)$$

Example 9.1.2 (Example 3.17)

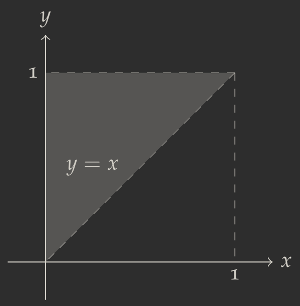
Consider the rvs X and Y with the joint pdf

$$f(x, y) = \begin{cases} 2 & 0 \leq x \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate $\mathrm{Cov}(X, Y)$.

Solution

Observe the diagram of the support set of X and Y to our right.



Then we can calculate

$$
\begin{aligned}
E[XY] &= \int_0^1 \int_x^1 2xy \, dy \, dx = \int_0^1 xy^2 \Big|_x^1 dx \\
&= \int_0^1 x - x^3 \, dx = \frac{1}{2} - \frac{1}{4} = \frac{1}{4} \\
f_X(x) &= \int_x^1 2 \, dy = \begin{cases} 2 - 2x & 0 \leq x \leq 1 \\ 0 & \textit{otherwise} \end{cases} \\
E[X] &= \int_0^1 x(2 - 2x) \, dx = 1 - \frac{2}{3} = \frac{1}{3} \\
f_Y(y) &= \int_0^y 2 \, dx = \begin{cases} 2y & 0 \leq y \leq 1 \\ 0 & \textit{otherwise} \end{cases} \\
E[Y] &= \int_0^1 2y^2 \, dy = \frac{2}{3}
\end{aligned}
$$

Thus

$$
\mathrm{Cov}(X, Y) = \frac{1}{4} - \frac{1}{3}\left(\frac{2}{3}\right) = \frac{1}{36}
$$

We observe that the covariance is positive. This implies a ***positive linear relationship****. However, we cannot tell from this value the strength of the relationship between X and Y.*

Example 9.1.3 (Example 3.18)

Consider rvs X and Y with the joint pf

$$
f(x, y) = \frac{xy}{7} \mathbb{1}_{\{y=1,2\}} \mathbb{1}_{\{x=0,\ldots,y\}}.
$$

Calculate $\mathrm{Cov}(X, Y)$.

Solution

The following table captures all the probabilities that can be found from the given pf

	$x = 0$	$x = 1$	$x = 2$
$y = 1$	0	$\frac{1}{7}$	0
$y = 2$	0	$\frac{2}{7}$	$\frac{4}{7}$

Observe that we thus have

$$f_X(x) = \begin{cases} \frac{3}{7} & x = 1 \\ \\ \frac{4}{7} & x = 2 \end{cases}$$

$$E[X] = \frac{3}{7} + 2\left(\frac{4}{7}\right) = \frac{11}{7}$$

$$f_Y(y) = \begin{cases} \frac{1}{7} & y = 1 \\ \\ \frac{6}{7} & y = 2 \end{cases}$$

$$E[Y] = \frac{1}{7} + 2\left(\frac{6}{7}\right) = \frac{13}{7}$$

Also,

$$E[XY] = \frac{1}{7} + 2\left(\frac{2}{7}\right) + 4\left(\frac{4}{7}\right) = 3$$

Therefore,

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = 3 - \frac{11}{7}\frac{13}{7} = \frac{4}{49}$$

Theorem 33 (Variance of Linear Combinations)

Suppose X and Y are rvs and a, b, c are real constants. Then

$$\mathrm{Var}(aX + bY + c) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y)$$

Proof

Let $E[aX + bY + c] = \mu$. Observe that

$$\begin{aligned}
&\mathrm{Var}[aX + bY + c] \\
&= E\left[[(aX + bY + c) - \mu]^2\right] \\
&= E\left[(aX + bY + c)^2 - 2\mu(aX + bY + c) + \mu^2\right] \\
&= E\left[a^2X^2 + abXY + acX + abXY + b^2Y^2 + bcY + c^2\right. \\
&\qquad \left. -2\mu(aX + bY + c) + \mu^2\right] \\
&= a^2E[X^2] + 2abE[XY] + acE[X] + b^2E[Y^2] + bcE[Y] + c^2 - \mu^2
\end{aligned}$$

Note that

$$\begin{aligned}
\mu^2 &= E\left[aX + bY + c\right]^2 \\
&= (aE[X] + bE[Y] + c)^2 \\
&= a^2E[X]^2 + b^2E[Y]^2 + 2abE[X]E[Y] + acE[X] + bcE[Y] + c^2
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
&\operatorname{Var}[aX + bY + c] \\
&= a^2E[X^2] + a^2E[X]^2 + b^2E[Y^2] - b^2E[Y]^2 + 2abE[XY] - 2abE[X]E[Y] \\
&= a^2\left(E[X^2] - E[X]^2\right) + b^2\left(E[Y^2] - E[Y]^2\right) + 2ab\left(E[XY] - E[X]E[Y]\right) \\
&= a^2\operatorname{Var}(X) + b^2\operatorname{Var}Y + 2ab\operatorname{Cov}(X, Y)
\end{aligned}$$

as required. □

By applying Theorem 33 repeatedly, we have the following generalized theorem.

Theorem 34 (Generalized Variance of Linear Combinations)

Suppose $X_1, X_2, \ldots, X_n$ *are rvs with* $\operatorname{Var}(X_i) = \sigma_i^2$, *and* $a_1, a_2, \ldots, a_n$ *are real constants. Then*

$$\operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2\sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j \operatorname{Cov}(X_i, X_j)$$

Note that the prove the above, we have to also use the fact that

$$\operatorname{Cov}(X_i, X_j) = \operatorname{Cov}(X_j, X_i)$$

Note

Note that in Theorem 34, if the rvs are independent rvs, then $\operatorname{Cov}(X_i, X_j) = 0$ *for* $i \neq j$, *thus wiping off the 2nd term in the equation, leaving us with*

$$\operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2\sigma_i^2$$

Example 9.1.4 (Example 3.19)

To build a ship engine piece, suppose two pole-shaped components A and B are attached at one end to each other to make one long pole-shaped component C. Suppose the length of part A is an rv with a mean of 3 inches and a variance of 0.25 inch^2. Similarly, the length of component B is an rv with a mean of 25 inches and a variance of 0.5 inch^2.

Find the mean and the variance of the length of part C if

1. *the lengths of A and B are independent;*
2. *the covariance between lengths of A and B is -0.3 inch^2.*

✎ **Solution**

Note that we are given that $C = A + B$

1. *We have that*

$$E(C) = E(A+B) = E(A) + E(B) = 3 + 25 = 28.$$

For variance, since A and B are independent, $\mathrm{Cov}(A, B) = 0$, thus

$$\begin{aligned} \mathrm{Var}(C) &= \mathrm{Var}(A+B) = \mathrm{Var}(A) + \mathrm{Var}(B) + 2\,\mathrm{Cov}(A, B) \\ &= 0.25 + 0.5 + 0 = 0.75 \end{aligned}$$

2. *Since C is a linear equation, the covariance does not affect the expectation and thus we still have*

$$E(C) = 28.$$

Now, given that $\mathrm{Cov}(A, B) = -0.3$, we have

$$\mathrm{Var}(C) = \mathrm{Var}(A) + \mathrm{Var}(B) + 2\,\mathrm{Cov}(A, B) = 0.75 - 0.6 = 0.15.$$

9.1.3 *Correlation*

The **covariance** is a real number which depends on the units of measurement of X and Y. The information part of a covariance is its **sign**, unless if it is used as the context.

To use the covariance as the context, and to quantitatively measure the strength of a linear relationship, which we have discussed and desired before, we use the **correlation coefficient**.

Definition 40 (Correlation Coefficient)

*The **correlation coefficient** of rvs X and Y is given by*

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where $\sigma_X = \sqrt{\mathrm{Var}(X)}$ *and* $\sigma_Y = \sqrt{\mathrm{Var}(Y)}$.

Note that this is the definition of the **Pearson Correlation Coefficient**. There are other correlation coefficients but we will be using only Pearson, at it seems.

Proposition 35 (Properties of the Correlation Coefficient)

Let X and Y be rvs, and $\rho(X, Y)$ the correlation coefficient of X and Y. Then

1. $|\rho(X, Y)| \leq 1$;
2. (***perfect positive linear relationship***)
   $\rho(X, Y) = 1 \iff Y = aX + b$ *for some* $a > 0$;
3. (***perfect inverse linear relationship***)
   $\rho(X, Y) = -1 \iff Y = aX + b$ *for some* $a < 0$.

Proof

1. *This is somewhat beyond the scope of what we can cover now but we shall use this result presented on Wikipedia:*

$$|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}.$$

Then given the formula of $\rho(X, Y)$, the proof is complete:

$$|\rho(X, Y)| = \left| \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \right| \leq \frac{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = 1$$

Proving 2 and 3 is currently outside of my abilities. Refer to this Math SE Q&A for a hint on how to prove this statement.

Example 9.1.5 (Example 3.20)

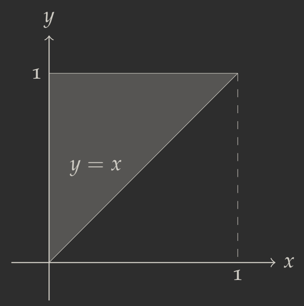
Consider rvs X and Y with the joint pdf

$$f(x,y) = \begin{cases} 2 & 0 \leq x \leq y \leq 1 \\ 0 & \textit{otherwise} \end{cases}$$

Calculate $\rho(X,Y)$.

Solution

The diagram to the right is an illustration of the region of support for X and Y. *We now calculate the following values:*



$$\begin{aligned}
E(XY) &= \int_0^1 \int_x^1 2xy \, dy \, dx = \int_0^1 x - x^3 \, dx = \frac{1}{4} \\
f_X(x) &= \int_x^1 2 \, dy = 2 - 2x \quad 0 \leq x \leq 1 \\
f_Y(y) &= \int_0^y 2 \, dx = 2y \quad 0 \leq y \leq 1 \\
E(X) &= \int_0^1 2x - 2x^2 \, dx = \left(x^2 - \frac{2}{3}x^3 \right) \Big|_0^1 = \frac{1}{3} \\
E(X^2) &= \int_0^1 2x^2 - 2x^3 \, dx = \left(\frac{2}{3}x^3 - \frac{1}{2}x^4 \right) \Big|_0^1 = \frac{1}{6} \\
E(Y) &= \int_0^1 2y^2 \, dy = \frac{2}{3} \\
E(Y^2) &= \int_0^1 2y^3 \, dy = \frac{1}{2} \\
\mathrm{Var}(X) &= E(X^2) - E(X)^2 = \frac{1}{6} - \left(\frac{1}{3} \right)^2 = \frac{1}{18} \\
\mathrm{Var}(Y) &= E(Y^2) - E(Y)^2 = \frac{1}{2} - \left(\frac{2}{3} \right)^2 = \frac{1}{18}
\end{aligned}$$

Therefore we have that

$$\begin{gathered}
\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{4} - \frac{2}{9} = \frac{1}{36} \\
\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)} = \frac{1}{18}
\end{gathered}$$

and so

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{\frac{1}{36}}{\frac{1}{18}} = \frac{1}{2}$$

9.1.4 Conditional Expectation

Definition 41 (Conditional Expectation)

Let G be a real-valued function. The ***conditional expectation*** *of $g(Y)$ given $X = x$, denoted as $g(Y) \mid (X = x)$ is given by*

$$E\left[g(Y) \mid x\right] = \sum_y g(y) f_Y(y \mid x)$$

if $Y|(X = x)$ is a discrete rv and

$$E\left[g(Y) \mid x\right] = \int_y g(y) f_Y(y \mid x)$$

if $Y|(X = x)$ is a continuous rv. This definition only holds provided that the sum and the integral converges absolutely. The conditional expectation of $h(X)$ given $Y = y$, where h is a real-valued function, is defined in a similar manner.

We also call $E[Y \mid X = x]$ the ***conditional mean****, which may be denoted as $E(Y \mid x)$, and $\mathrm{Var}(Y \mid X = x)$ the* ***conditional variance****, which may be denoted as $\mathrm{Var}(Y \mid x)$.*

Note

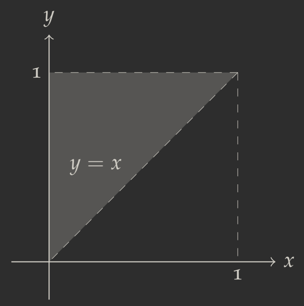
Note that there is also the notation $E(Y \mid X)$, which is an rv, and hence different from $E(Y \mid x)$.

Example 9.1.6 (Example 3.21)

Consider $f(x, y) = 8xy\mathbb{1}_{\{0<x<y<1\}}$. Calculate the conditional mean and the conditional variance of $X \mid \left(Y = \frac{1}{2}\right)$.

Solution

The diagram to the right illustrates the region of support for X and Y. To derive the conditional distribution, we first need $f_Y(y)$.



$$f_Y(y) = \int_0^y 8xy \, dx = 4x^2 y \Big|_0^y = 4y^3$$

Thus

$$f_X\left(X \mid Y = \frac{1}{2}\right) = \frac{f\left(x, \frac{1}{2}\right)}{f_Y\left(\frac{1}{2}\right)} = \frac{4x}{\frac{1}{2}} = 8x \quad 0 < x < \frac{1}{2}$$

Therefore the conditional mean is

$$E\left[X \mid \frac{1}{2}\right] = \int_0^{\frac{1}{2}} 8x^2\,dx = \frac{8}{3} \cdot \frac{1}{2^3} = \frac{1}{3},$$

the second conditional moment is

$$E\left[X^2 \mid \frac{1}{2}\right] = \int_0^{\frac{1}{2}} 8x^3\,dx = 2 \cdot \frac{1}{2^4} = \frac{1}{8},$$

and so the conditional variance is

$$\mathrm{Var}\left(X^2 \mid \frac{1}{2}\right) = E\left[X^2 \mid \frac{1}{2}\right] + E\left[X \mid \frac{1}{2}\right]^2 = \frac{1}{8} - \frac{1}{9} = \frac{1}{72}$$

10 Lecture 10 Jun 05th 2018

10.1 Joint Distribution (Continued 4)

10.1.1 Conditional Expectation (Continued)

Example 10.1.1 (Example 3.22)

Given the joint distribution below, calculate $\mathrm{Var}(X \mid Y = 1)$ *and compare it to* $\mathrm{Var}(X)$.

	$x = -2$	$x = 0$	$x = 2$	$P(Y = y)$
$y = 0$	0.05	0.1	0.15	0.3
$y = 1$	0.07	0.11	0.2	0.38
$y = 2$	0.02	0.25	0.05	0.32
$P(X = x)$	0.14	0.46	0.4	

Solution

Note that

$$\begin{aligned}\mathrm{Var}(X) &= E(X^2) - E(X)^2 \\ &= 4 \cdot 0.14 + 4 \cdot 0.4 - (-2 \cdot 0.14 + 2 \cdot 0.4)^2 = 1.8896\end{aligned}$$

To get $\mathrm{Var}(X \mid Y = 1)$*, we first need*

$$f(X \mid Y = 1) = \frac{P(X = x, Y = 1)}{P(Y = 1)} = \begin{cases} \frac{0.07}{0.38} = 0.1842 & x = -2 \\ \frac{0.11}{0.38} = 0.2895 & x = 0 \\ \frac{0.2}{0.38} = 0.5263 & x = 2 \end{cases}$$

Thus

$$\begin{aligned} \mathrm{Var}(X \mid Y = 1) &= E[X^2 | Y = 1] - E[X \mid Y = 1]^2 \\ &= \frac{14}{19} + \frac{40}{19} - (\frac{13}{19})^2 = \frac{857}{361} = 2.3740 \end{aligned}$$

Proposition 36 (Independence on Conditional Expectation)

If X and Y are independent rvs then $E[g(Y) \mid x] = E[g(Y)]$ *and* $E[h(X) \mid y] = E[h(X)]$.

Proof

We shall prove one of the above for the other will follow a similar argument. Also, we shall prove the continuous case and leave the discrete case as an exercise.

Exercise 10.1.1
Prove the discrete case for Proposition 36.

Observe that

$$\begin{aligned} E[g(Y) \mid X = x] &= \int_y g(y) \frac{f(x,y)}{f_X(x)} \, dy \\ &= \int_y g(y) \frac{f_X(x) f_Y(y)}{f_X(x)} \, dy \quad \because \text{ independence} \\ &= \int_y g(y) f_Y(y) \, dy = E[g(Y)] \end{aligned}$$

□

Theorem 37 (Law of Total Expectation)

Suppose X and Y are rvs, then

$$E\left(E[g(Y) \mid X]\right) = E[g(Y)]$$

If g is the identity function, we have $E(E[Y \mid X]) = E(Y)$.

Proof

We shall prove for the discrete case and leave the continuous case as an exercise. Observe that

$$
\begin{aligned}
E[g(Y) \mid X] &= \sum_y [g(y) \cdot P(Y = y \mid X)] \\
E[E[g(Y) \mid X]] &= \sum_x \left[\sum_y [g(y) \cdot P(Y = y \mid X)] \right] P(X = x) \\
&= \sum_x \sum_y g(y) \cdot P(X = x, Y = y) \\
&= \sum_y g(y) \sum_x P(X = x, Y = y) \\
&= \sum_y [g(y) \cdot P(Y = y)] = E[g(Y)]
\end{aligned}
$$

□

Exercise 10.1.2
Prove the continuous case for *Theorem 37.*

Example 10.1.2 (Example 3.23 - A Classical Example)

This is a very classical example to illustrate the power of the **Law of Total Expectation**.

A man is lost in a mine, and 3 paths are in front of him. If he takes path 1, after 3 hours, he will be back at his current place. If he takes path 2, the time to get out of the mine (in hours) follows an $\mathrm{Exp}(1)$ distriburion. If he takes the 3rd path, he will be back to his current place after 2 hours. Suppose that the man cannot recognize which path he took every time he comes back to the original spot (after going through either path 1 or 3), and so he randomly chooses a path every time he comes back to this original spot. What is the expected time that he will take to get out of the mine?

Solution

Let X an rv that represents the path number, i.e. $X = 1, 2$ or 3, and let Y represent the total time that the man takes to exit the mine. We are given that

$$P(X = 1) = P(X = 2) = P(X = 3) = \frac{1}{3}.$$

We are also given that

$$\begin{aligned} E[Y \mid X = 1] &= 3 + E[Y] \\ E[Y \mid X = 2] &= 1 \quad \because Y \mid (X = 2) \sim \operatorname{Exp}(1) \\ E[Y \mid X = 3] &= 2 + E[Y]. \end{aligned}$$

Therefore, to get the expected time, by Theorem 37,

$$\begin{aligned} E[Y] &= E[E[Y \mid X]] \\ &= \frac{1}{3} \cdot E[Y \mid X = 1] + \frac{1}{3} \cdot E[Y \mid X = 2] + \frac{1}{3} \cdot E[Y \mid X = 3] \\ &= \frac{1}{3} \cdot (3 + E[Y]) + \frac{1}{3} \cdot 1 + \frac{1}{3}(2 + E[Y]) \\ &= 2 + \frac{2}{3} E[Y] \end{aligned}$$

and hence

$$E[Y] = 6$$

Theorem 38 (Law of Total Variance)

Suppose X and Y are rvs. Then

$$\operatorname{Var}(Y) = E[\operatorname{Var}(Y \mid X)] + \operatorname{Var}[E(Y \mid X)]$$

Proof

Note that

$$\begin{aligned} \operatorname{Var}(Y|X) &= E(Y^2|X) - E(Y|X)^2 \\ E[\operatorname{Var}(Y|X)] &= E[E(Y^2|X) - E(Y|X)^2] \\ &= E[Y^2] - E[E(Y|X)^2] \\ &= E[Y^2] - \left[\operatorname{Var}(E(Y|X)) + E\left(E(Y|X)\right)^2\right] \\ &= \operatorname{Var}(Y) - \operatorname{Var}(E(Y|X)) \end{aligned}$$

By rearranging the above, we get

$$\operatorname{Var}(Y) = E[\operatorname{Var}(Y|X)] + \operatorname{Var}[E(Y|X)]$$

$\square$

Example 10.1.3 (Example 3.24 (Course Note 3.7.11))

Suppose $P \sim \text{Unif}(0, 0.1)$ *and* $Y \mid P = p \sim \text{Bin}(10, p)$. *Find* $E(Y)$ *and* $\text{Var}(Y)$.

Solution

$$\begin{aligned}
E[Y] &= E[E[Y|P]] = E[10P] = 10E[P] \\
&= 10 \cdot \int_0^{0.1} \frac{p}{0.1}\, dp = \frac{10}{0.2} p^2 \Big|_0^{0.1} = 0.05 \\
\textit{Note: } E[P^2] &= \int_0^{0.1} \frac{p^2}{0.1}\, dp = \frac{p^3}{0.3}\Big|_0^{0.1} = \frac{0.001}{0.3} = \frac{1}{300} \\
\text{Var}(Y) &= E[\text{Var}(V|P)] + \text{Var}[E[Y|P]] \\
&= E[10P(1-P)] + \text{Var}[10P] \\
&= 10E[P] - 10E[P^2] + 100\,\text{Var}(P) \\
&= 0.05 - \frac{1}{30} + 100\left[E[P^2] - E[P]^2\right] \\
&= \frac{1}{60} + 100\left[\frac{1}{300} - 0.05^2\right] = \frac{1}{60} + 100\left[\frac{1}{300} - \frac{1}{400}\right] \\
&= \frac{1}{60} + \frac{1}{12} = \frac{1}{10}
\end{aligned}$$

10.1.2 *Joint Moment Generating Functions*

Definition 42 (Joint Moment Generating Functions)

The ***joint moment generating function*** *of two rvs X and Y is defined as*

$$M(t_1, t_2) = E\left(e^{t_1 X + t_2 Y}\right)$$

if the expectation exists for all $t_1 \in (-h_1, h_1)$ *and* $t_2 \in (-h_2, h_2)$ *for some* $h_1, h_2 > 0$.

More generally, if $X_1, X_2, ..., X_n$ *are rvs, then*

$$M(t_1, t_2, ..., t_n) = E\left[\exp\left(\sum_{i=1}^n t_i X_i\right)\right]$$

is called the ***joint mgf*** *of* $X_1, X_2, ..., X_n$ *is the expectation exists for all* $t_i \in (-h_i, h_i)$ *for some* $h_i > 0$, *where* $i = 1, ..., n$.

Definition 43 (Joint Moments and Marginal MGF)

Given the joint mgf $M(t_1, t_2)$, we can calculate the joint moments. In particular,

$$E\left(X^j Y^k\right) = \frac{\partial^{j+k}}{\partial t_i^j \partial t_2^k} M(t_1, t_2)\Big|_{(t_1,t_2)=(0,0)}$$

If $M(t_1, t_2)$ exists for all $t_1 \in (-h_1, h_1)$ and $t_2 \in (-h_2, h_2)$ for some $h_1, h_2 > 0$, then the mdf of X is given by

$$M_X(t) = E\left(e^{tX}\right) = M(t, 0) \quad t \in (-h_1, h_1)$$

and the mgf of Y is given by

$$M_Y(t) = E\left(e^{tY}\right) = M(0, t) \quad t \in (-h_2, h_2).$$

Example 10.1.4 (Example 3.25)

Given the joint distribution below, calculate the joint mgf $M(t_1, t_2)$, the first joint moment, $E[XY]$, from the joint mgf, and the marginal mgf of X and that of Y.

	x = -1	x = 1
y = 1	0.5	0.3
y = 2	0.1	0.1

Solution

Since all probabilities are provided,

$$
\begin{aligned}
M(t_1, t_2) &= E\left(e^{t_1 X + t_2 Y}\right) = \sum_x \sum_y e^{t_1 x + t_2 y} P(X = x, Y = y) \\
&= 0.5e^{-t_1+t_2} + 0.3e^{t_1+t_2} + 0.1e^{-t_1+2t_2} + 0.1e^{t_1+2t_2} \\
E(XY) &= \frac{\partial^2}{\partial t_1 \partial t_2} M(t_1, t_2)\Big|_{t_1=0,\, t_2=0} \\
&= \frac{\partial}{\partial t_1}\left[0.5e^{-t_1+t_2} + 0.3e^{t_1+t_2} + 0.2e^{-t_1+2t_2} + 0.2e^{t_1+2t_2}\right]\Big|_{t_1=0, t_2=0} \\
&= -0.5e^{-t_1+t_2} + 0.3e^{t_1+t_2} - 0.2e^{-t_1+2t_2} + 0.2e^{t_1+2t_2}\Big|_{t_1=0, t_2=0} \\
&= -0.2 \\
M_X(t_1) &= M(t_1, 0) = 0.5e^{-t_1} + 0.3e^{t_1} + 0.1e^{-t_1} + 0.1^{t_1} \\
&= 0.6e^{-t_1} + 0.4e^{t_1} \\
M_Y(t_2) &= M(0, t_2) = 0.5e^{t_2} + 0.3e^{t_2} + 0.1e^{2t_2} + 0.1e^{2t_2} \\
&= 0.8e^{t_2} + 0.2e^{2t_2}
\end{aligned}
$$

Proposition 39 (Independence on Joint MGF)

Suppose X and Y are rvs with joint mgf $M(t_1, t_2)$ which exists $\forall t_1 \in (-h_1, h_1)$, $t_2 \in (-h_2, h_2)$, for some $h_1, h_2 > 0$. Then X and Y are independent rvs iff

$$\forall t_1 \in (-h_1, h_1),\ t_2 \in (-h_2, h_2) \quad M(t_1, t_2) = M_X(t_1) M_Y(t_2)$$

where $M_X(t_1) = M(t_1, 0)$ and $M_Y(t_2) = M(0, t_2)$.

Proof

to be proven later

Example 10.1.5 (Example 3.26 (Course Note 3.8.5))

Suppose X and Y are continous rvs with joint pdf

$$f(x, y) = e^{-y} \quad 0 < x < y < \infty$$

Find the joint mdf of X and Y. Are X and Y independent rvs? What is the *marginal mgf of X and Y?*

Solution

$$\begin{aligned}
M(t_1, t_2) = E[e^{t_1 X + t_2 Y}] &= \int_0^\infty \int_0^y e^{t_1 x + t_2 y} e^{-y} \, dx \, dy \\
&= \int_0^\infty \frac{1}{t_1} e^{t_1 x + t_2 y - y} \Big|_0^y \, dy \\
&= \int_0^\infty \frac{1}{t_1} \left[e^{y(t_2 - 1)} - e^{y(t_1 + t_2 - 1)} \right] dy \\
&= \frac{1}{t_1} \left[\frac{1}{t_2 - 1} e^{y(t_2-1)} - \frac{1}{t_1 + t_2 - 1)} e^{y(t_1 + t_2 - 1)} \right] \Big|_0^\infty \\
&= \frac{1}{t_1} \left[\frac{t_1}{(t_2 - 1)(t_1 + t_2 - 1)} \right] \\
&= \frac{1}{(t_2 - 1)(t_1 + t_2 - 1)} \quad t_2 < 1 \wedge t_1 + t_2 < 1 \\
M_X(t_1) = M(t_1, 0) &= \frac{1}{t_1 - 1} \quad t_1 < 1 \\
M_Y(t_2) = M(0, t_2) &= \frac{1}{(t_2 - 1)^2} \quad t_2 < 1
\end{aligned}$$

Observe that

$$M_X(t_1) M_Y(t_2) = \frac{1}{(t_1 - 1)(t_2 - 1)^2} \neq M(t_1, t_2)$$

and so by Proposition 39, X and Y are not independent.

Example 10.1.6 (Example 3.27)

Investigate the independence of X and Y in Example 10.1.4 using the mgf method.

Solution

We had that

$$\begin{aligned}
M_X(t_1) &= 0.6^{-t_1} + 0.4e^{t_1} \quad t_1 \in \mathbb{R} \\
M_Y(t_2) &= 0.8e^{t_2} + 0.2e^{2t_2} \quad t_2 \in \mathbb{R}.
\end{aligned}$$

Since

$$M_X\left(\frac{1}{2}\right) M_Y\left(\frac{1}{2}\right) \neq M\left(\frac{1}{2}, \frac{1}{2}\right)$$

we have that X and Y are not independent.

11 Lecture 11 Jun 07th 2018

11.0.1 Working with Multivariate Cases

Almost everything that has been introduced above can be extended to cases where we have more than just 2 rvs. For example:

Definition 44 (k-variate CDF)

The ***k-variate CDF****, $k > 2$, rvs $X_1, ..., X_k$ is defined as*

$$F(x_1, ..., x_k) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_k \leq x_k).$$

In the continuous case, we may write

$$f(x_1, ..., x_k) = \frac{\partial^k}{\partial x_1 \ldots \partial x_k} F(x_1, ..., x_k).$$

Definition 45 (k-variate Support Set)

The support set of the distribution for $X_1, X_2, ..., X_k$ is

$$\{(x_1, ..., x_k) : f(x_1, ..., x_k) > 0\}$$

We also have the following:

Proposition 40 (Law of Total Probability - Multivariate)

If $X_1, ..., X_k$ are continuous rvs, then

$$\int_{x_1} \dots \int_{x_k} f(x_1, ..., x_k) \, dx_1 \dots dx_k = 1.$$

Should they by discrete, then

$$\sum_{x_1} \dots \sum_{x_k} f(x_1, ..., x_k) = 1$$

Definition 46 (k-Variate Marginal Distribution)

To get the marginal distribution of a subset of m variables from $X_1, ..., X_k$ $(1 \leq m \leq k)$, we will sum or integrate over the other ones if they are discrete or continuous, respectively. For example,

$$f(x_1, x_2, x_3) = \int_{x_4} \dots \int_{x_k} f(x_1, ..., x_k) \, dx_4 \dots dx_k$$

Definition 47 (k-Variate Joint MGF)

The joint mgf of $X_1, ..., X_k$ is defined as

$$M(t_1, t_2, ..., t_k) = E\left(e^{t_1 X_1 + ... + t_k X_k}\right)$$

Proposition 41 (Independence for Multivariate Cases)

If $X_1, ..., X_k$ are independent, then

$$f(x_1, ..., x_k) = \prod_{i=1}^{k} f_{X_i}(x_i) \qquad F(x_1, ..., x_k) = \prod_{i=1}^{k} F_{X_i}(x_i)$$
$$M(t_1, ..., t_k) = \prod_{i=1}^{k} M_{X_i}(t_i)$$

THERE ARE many different examples of multivariate distributions. We shall discuss two:

- Multinomial Distribution

- Multivariate Normal Distribution

The **multinomial distribution** is an extension of the binomial distribution to cases where there are more categories than two results. For a multinomial distribution, we have that

- the experiment involves n trials, each with k categories
- the outcome of trials are independent of each other
- the probability of each category, p_i, remains the same across n trials
- $X = (X_1, ..., X_k) \sim \mathrm{Mult}(n, p_1, ..., p_k)$ counts the number of elements in each category among the n trials.

Definition 48 (Mutlinomial Distribution)

Suppose $X_1, ..., X_k$ are discrete rvs with joint pf

$$f(x_1, ..., x_k) = \frac{n!}{x_1!x_2! \dots x_k!x_{k+1}!} p_1^{x_1} p_2^{x_2} \dots P_k^{x_k} p_{k+1}^{x_{k+1}}$$

where $x_i = 0, ..., n,$, $x_{k+1} = n - \sum_{i=1}^k x_i$, $0 < p_i < 1$, $p_{k+1} = 1 - \sum_{i=1}^k p_i$, for $i = 1, ..., k+1$.

Under these conditions, $(X_1, ..., X_k)$ is said to have a multinomial distribution, and we write $(X_1, ..., X_k) \sim \mathrm{Mult}(n, p_1, ..., p_k)$.

Note

Observe that $\mathrm{Bin}(n, p) = \mathrm{Mult}(n, p, p)$.

Proposition 42 (Properties of Multinomial Distribution)

Suppose $(X_1, ..., X_k) \sim \mathrm{Mult}(n, p_1, ..., p_k)$, then

1. *$\forall (t_1, ..., t_k) \in \mathbb{R}^k$, the random vector $(X_1, ..., X_k)$ has joint mgf*

$$M(t_1, ..., t_k) = E\left(e^{t_1 X_1 + \dots t_k X_k}\right) = (p_1 e^{t_1} + \dots + p_k e^{t_k} + p_{k+1})^n$$

2. *Any subset of $X_1, ..., X_{k+1}$ also has a multinomial distribution. In particular, $X_i \sim \mathrm{Bin}(n, p_i)$, $i = 1, ..., k+1$.*

3. *If* $T = X_i + X_j$, *for* $i \neq j$, *then* $T \sim \mathrm{Bin}(n, p_i + p_j)$
4. $\mathrm{Cov}(X_i, X_j) = -np_ip_j$, *for* $i \neq j$
5. *The conditional distribution of any subset of* $(X_1, ..., X_{k+1})$ *given the rest of the coordinates is a multinomial distribution. In particular, the conditional pf of* X_i *given* $X_j = x_j$, $i \neq j$, *is*

$$X_i \mid X_j = x_j \sim \mathrm{Bin}\left(n - x_j, \frac{p_i}{1 - p_j}\right)$$

6. *The conditional distribution of* X_i *given* $T = X_i + X_j = t$, *for* $i \neq j$, *is*

$$X_i \mid X_i + X_j = t \sim \mathrm{Bin}\left(t, \frac{p_i}{p_i + p_j}\right)$$

We shall look at the bivariate normal distribution so that it is easier to be explained. The same idea can be extended to a multivariate Normal distribution.

Definition 49 (Bivariate Normal Distribution)

Let X_1 *and* X_2 *be rvs with joint pdf*

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi \left|\Sigma\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right\} \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} \exp\left\{-\frac{1}{2(1 - \rho^2)}\left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2}\right.\right. \\ &\qquad \left.\left. + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2}\right]\right\} \end{aligned}$$

where $(x_1, x_2) \in \mathbb{R}^2$ *and*

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

where Σ *is a nonsingular matrix. Then* $X = (X_1, X_2)^T$ *is said to have a bivariate normal distribution, and we write* $X \sim \mathrm{BVN}(\mu, \Sigma)$.

Proposition 43 (Properties of Bivariate Normal Distribution)

Suppose $X \sim \mathrm{BVN}(\mu, \Sigma)$, *where*

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

1. *X has a joint mgf of*

   $$M(t_1, t_2) = E[\exp\left(t^T X\right)] = E\left(e^{t_1 X_1 + t_2 X_2}\right) = \exp\left(\mu^T t + \frac{1}{2} t^T \Sigma t\right)$$

   $\forall (t_1, t_2) \in \mathbb{R}^2$.
2. $X_1 \sim N(\mu_1, \sigma_1^2)$ *and* $X_2 \sim N(\mu_2, \sigma_2^2)$
3. $\mathrm{Cov}(X_1, X_2) = \rho\sigma_1\sigma_2$ *and* $\mathrm{Corr}(X_1, X_2) = \rho$ *where* $|rho| \leq 1$
4. X_1 *and* X_2 *are independent rvs iff* $\rho = 0$
5. $c = (c_1, c_2)^T$ *is a nonzero vector of constants* $\implies$

   $$c^T X = \sum_{i=1}^{2} c_i X_i \sim N\left(c^T \mu, c^T \Sigma c\right).$$

6. *If A is a* 2×2 *nonsingular matrix and b is a* 2×1 *vector, then* $Y = AX + b \sim \mathrm{BVN}\left(A\mu + b, A\Sigma A^T\right)$.
7. $$X_2 \mid X_1 = x_1 \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1 - \rho^2)\right)$$
   $$X_1 \mid X_2 = x_2 \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1 - \rho^2)\right)$$
8. $(X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi^2(2)$

Definition 50 (Random Sample (IID))

Rvs $X_1, \ldots, X_2$ *are said to form a* ***simple random sample*** *or are said to be* ***independent and identically distributed*** *(IID) if* $X_1, \ldots, X_n$ *are independent, and* $f_{X_i} = f_{X_j}, \forall i \neq j$.

Example 11.0.1 (Example 3.28)

Let $X_1, \ldots, X_{10}$ be a random sample of standard normal distribution. Let Y_1 denote the number of these variables that are between -1 and 1, let Y_2 denote the number that have absolute value between 1 and 2, and let Y_3 denote the number that have absolute value larger than 2. Calculate:

1. $P(Y_1 \leq 2)$
2. $E[Y_2 \mid Y_1 = 5]$

12 Lecture 12 Jun 12th 2018

12.1 Functions of Random Variables

12.1.1 Transformation of Two or More Random Variables

In earlier lectures we discussed about basic transformations from one random variable to another, for example, from a continuous rv X to $Y = g(X)$. In particular, we methods were presented:

- The direct method, i.e. $P(Y \leq y) = P(g(X) \leq y)$, and taking the derivative of $P(Y \leq y)$ with respect to y.

- Using the mgf of Y, and then translate it as the mgf of X.

Note (Recall)

In Section 2.4.4, we used the following idea to obtain the result that we desire: for rvs X and $Y = g(X)$ where g is some continuous and injective function

$$
\begin{aligned}
P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) \\
f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = (g^{-1}(y))' f_X(g^{-1}(y))
\end{aligned}
$$

In this chapter, we will now study the case where we have more than one rv involved. In particular, let X and Y be two continuous rvs with joint pdf $f(x, y)$. Our questions are:

1. What is the distribution of $U = h_1(X, Y)$?
2. What is the joint distribution of $U = h_1(X, Y)$ and $V = h_2(X, Y)$?

To answer the first question, we can actually still employ the direct method:

Example 12.1.1 (Example 4.1 (Course Notes 4.1.1))

Suppose X and Y are continuous rvs with joint pdf

$$f(x,y) = 3y\mathbb{1}_{0 \leq x \leq y \leq 1}$$

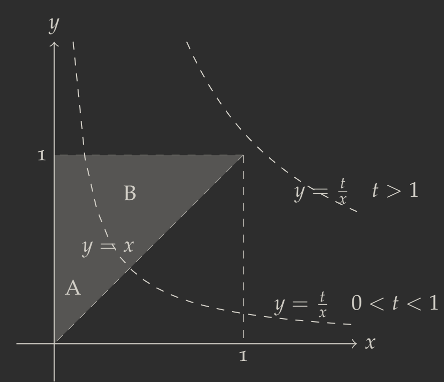
Find the pdf of $T = XY$.

Solution

First, note that[1]

$$P(T \leq t) = P(XY \leq t) = P\left(Y \leq \frac{t}{X}\right)$$

[1] I wonder if the last step is actually valid. The support of X definitely includes 0, so the division would not make sense with $\frac{t}{X}$. We can, however, still make sense of the event in the 2nd term, which we would have $P(0 \leq t)$. Should we be concerned about $X = 0$, or can we neglect that single point given that X is a continuous rv?



The diagram to the right shows us the support of the joint probability. We observe that if $t \leq 0$, then $P(T \leq t) = 0$, and if $t \geq 1$, then $P(T \leq t) = 1$. Now if $0 < t < 1$, the region that we are looking for is the shaded region with the label A, and so we consider

$$\begin{aligned}
P(T \leq t) &= 1 - P(B) = 1 - \iint_B f(x,y)\,dx\,dy \\
&\overset{(1)}{=} 1 - \int_{\sqrt{t}}^{1} \int_{\frac{t}{y}}^{\sqrt{t}} f(x,y)\,dx\,dy - \int_{\sqrt{t}}^{1} \int_{\sqrt{t}}^{y} f(x,y)\,dx\,dy \\
&\overset{(2)}{=} 1 - \int_{\sqrt{t}}^{1} \int_{\frac{t}{y}}^{y} f(x,y)\,dx\,dy \\
&= 1 - \int_{\sqrt{t}}^{1} \int_{\frac{t}{y}}^{y} 3y\,dx\,dy \\
&= 1 - \int_{\sqrt{t}}^{1} 3y\left(y - \frac{t}{y}\right) dy \quad \because \textbf{\textit{FTC}} \\
&= 1 - \frac{\sqrt{t}}{1}\left(3y^2 - 3t\right) dy \\
&= 1 - \left[y^3 - 3ty\right]\Big|_{\sqrt{t}}^{1} = 1 - (1 - 3t - \sqrt{t}^3 + 3t\sqrt{t}) \\
&= 3t - 2t\sqrt{t} \text{ for } 0 < t < 1
\end{aligned}$$

where for step (1), we broke B into two parts, in particular at $x = \sqrt{t}$ where $y = x$ and $y = \frac{t}{x}$ coincide, and step (2) is true by linearity of integration.

With that, i.e. with the CDF of T, we can then obtain

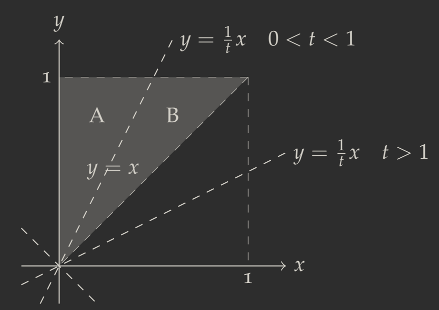
$$f_T(t) = \begin{cases} 3 - 3\sqrt{t} & 0 < t < 1 \\ 0 & \textit{otherwise} \end{cases}$$

Example 12.1.2 (Example 4.2 (Course Note 4.1.2))

Using the info in Example 12.1.1, find the pdf of $T = \frac{X}{Y}$.

Solution

The diagram to the right shows the support of X, Y and the function $t = \frac{x}{y}$. We observe that if $t = 0$, then $x = 0$, and we would have the line on the axis, and so $P(T \leq t) = 0$. If $t < 0$, we would have $y = mx$ where $m = \frac{1}{t} < 0$, which, regardless of what $t < 0$ is, will not interact with the support of X and Y. So for $t < 0$, $P(T \leq t) = 0$. Now if $t > 0$, we have



$$P(T \leq t) = P\left(\frac{X}{Y} \leq t\right) = P\left(Y \geq \frac{1}{t}X\right).$$

Consider the case where $t \geq 1$, we have that the event would still cover the entire support set of X and Y, and so $P(T \leq t) = 1$ for $t \geq 1$. With that, the only remaining case is when $0 < t < 1$. In this case,

$$P(T \leq t) = \int_0^1 \int_0^{ty} 3y \, dx \, dy = \int_0^1 3ty^2 \, dy = t$$

and so the pdf of T is

$$f_T(t) = \begin{cases} 1 & 0 < t < 1 \\ 0 & \textit{otherwise} \end{cases}$$

Example 12.1.3 (Example 4.3 (Course Note 4.1.3) - Order Statistics)

Suppose $X_1, ..., X_n$ are IID samples, each from a continuous distribution, and with pdf f and cdf F. Find the pdf of

1. $T = \min(X_1, ..., X_n) = X_{(1)}$
2. $Y = \max(X_1, ..., X_n) = X_{(n)}$

Solution

1. *For T, we have that its cdf is*[2]

[2] The use the **Law of Total Probability** here so that we can use the following argument: "the *smallest* rv is larger than t, and so rest of the rvs must be the same."

$$\begin{aligned} P(T \leq t) &= 1 - P(T > t) = 1 - P(\min(X_1, ..., X_n) > t) \\ &= 1 - P(X_1 > t, X_2 > t, ..., X_n > t) \\ &= 1 - \prod_{i=1}^{n} P(X_i > t) \qquad \because \textit{independence} \\ &= 1 - \prod_{i=1}^{n} P(X_1 > t) \qquad \because \textit{identical distribution} \\ &= 1 - P(X_1 > t)^n = 1 - [1 - F_{X_1}(t)]^n \end{aligned}$$

and so its pdf is

$$\begin{aligned} f_T(t) &= -\frac{d}{dt}[1 - F_{X_1}(t)]^n = -n(-F_{X_1}(t))'[1 - F_{X_1}(t)]^{n-1} \\ &= n f_{X_1}(t)[1 - F_{X_1}(t)]^{n-1}. \end{aligned}$$

Since T relies entirely on X_1 (due to IID), and since we did not have to condition on the values of t, we have that

$$\operatorname{supp}(T) = \operatorname{supp}(X_1) = \operatorname{supp}(X_i) \quad \textit{for } i = 1, ..., n.$$

2. *For Y, we have that its cdf is*[3]

$$\begin{aligned} P(Y \leq t) &= P(\max(X_1, ..., X_n) \leq y) = P(X_1 \leq y, ..., X_n \leq y) \\ &= \prod_{i=1}^{n} P(X_i \leq y) \quad \because \textit{independence} \\ &= \prod_{i=1}^{n} P(X_1 \leq y) \quad \because \textit{identical distribution} \\ &= P(X_1 \leq t)^n = F_{X_1}(t)^n \end{aligned}$$

and therefore its pdf is

$$f_Y(y) = \frac{d}{dy} F_{X_1}(y)^n = n f_{X_1}(y) F_{X_1}(y)^{n-1}.$$

[3] This time, we do not have to employ the **Law of Total Probability**, because we simply have that "the *largest* rv is smaller than t, and so must the rest of the rvs."

Exercise 12.1.1

From Example 12.1.3, find the joint distribution of $X_{(1)}$ and $X_{(n)}$.

12.1.2 One-to-One Bivariate Transformations

Definition 51 (One-to-One Bivariate Transformation)

Let X and Y be rvs, and $R_{XY} = \mathrm{supp}[(X,Y)] \in \mathbb{R}^2$. We define

$$U = h_1(X,Y) \quad V = h_2(X,Y)$$
$$S : R_{XY} \to \mathbb{R}^2 \text{ by } (x,y) \mapsto (h_1(x,y), h_2(x,y))$$

The mapping S is called a ***one-to-one mapping*** *if and only if* [4] $\forall (u,v) \in R_{UV}$, $\exists!(x,y) \in R_{XY}$, $\exists w_1, w_2$ *that are functions such that*

$$x = w_1(u,v) \quad y = w_2(u,v)$$

i.e. $\exists S^{-1} : R_{UV} \to R_{XY}$ *such that* $(u,v) \mapsto (x,y)$. *The* ***Jacobian*** *of the transformation* S^{-1} *is*

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \left| \frac{\partial(u,v)}{\partial(x,y)} \right|^{-1}$$

where $\frac{\partial(u,v)}{\partial(x,y)}$ *is the Jacobian of the transformation* S.

[4] There is nothing magnificent about this definition, since this is simply the definition of a one-to-one function.

Theorem 44 (One-to-One Bivariate Transformations)

This is a generalization of Theorem 8.

Let X and Y be continuous rvs with joint pdf f_{XY} and let $R_{XY} = \{(x,y) : f(x,y) > 0\}$ be the support set of (X,Y), and R_{UV} be the support set of (U,V). Suppose the transformation $S : R_{XY} \to R_{UV}$ defined by

$$U = h_1(X,Y) \quad V = h_2(X,Y)$$

is a one-to-one transformation, with inverse transformation

$$X = w_1(U,V) \quad Y = w_2(U,V).$$

Then $g(u,v)$, the joint pdf of U and V, is given by

$$\forall (u,v) \in R_{UV} \quad g(u,v) = f(w_1(u,v), w_2(u,v)) \left| \frac{\partial(x,y)}{\partial(u,v)} \right|.$$

Proof

Let the inverse transformation be labelled as $S^{-1} : R_{UV} \supset B \to A \subset$

R_{XY}. Then

$$\iint_B g(u,v)\,dv\,du = P[(U,V) \in B] = P[(X,Y) \in A]$$
$$= \iint_A f(x,y)\,dx\,dy$$
$$= \iint_B f(w_1(u,v), w_2(u,v)) \left| \frac{\partial(x,y)}{\partial(u,v)} \right| du\,dv$$

where the last step is by the Change of Variables Theorem. And so by comparing integrands, we have

$$\forall (u,v) \in R_{UV} \quad g(u,v) = f(w_1(u,v), w_2(u,v)) \left| \frac{\partial(x,y)}{\partial(u,v)} \right|$$

as required. □

Proposition 45 (Properties of the Jacobian)

Given the setup in Definition 51, we have that

1. *if S is a linear transformation, i.e. $\exists a_1, b_1, c_1, a_2, b_2, c_2 \in \mathbb{R}$ such that $u(x,y) = a_1x + b_1y + c_1$ and $v(x,y) = a_2x + b_2y + c_2$, then the Jacobian is a constant;*
2. *if S is a one-to-one transformation, then $\left| \frac{\partial(x,y)}{\partial(u,v)} \right| \neq 0$*

Proof

1. We have

$$\frac{\partial u}{\partial x} = a_1 \qquad \frac{\partial u}{\partial y} = b_1$$
$$\frac{\partial v}{\partial x} = a_2 \qquad \frac{\partial v}{\partial y} = b_2$$

and so

$$|J| = \frac{\partial(u,v)}{\partial(x,y)} = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1b_2 - a_2b_1$$

which is a constant, as required.

2. ***I have no idea how to prove this.***

Example 12.1.4 (Example 4.4 (Course Notes 4.2.4))

Suppose $X \sim \mathrm{Gam}(a, 1)$ and $Y \sim \mathrm{Gam}(b, 1)$ independently. Find the joint pdf of $U = X + Y$ and $V = \frac{X}{X+Y}$. Show that $U \sim \mathrm{Gam}(a+b, 1)$ and $V \sim \mathrm{Beta}(a, b)$, independently. Find $E(V)$.

Solution

Given $U = X + Y$ and $V = \frac{X}{X+Y}$, rearranging variables, we have

$$X = UV \text{ and } Y = U(1 - V)$$

In order for U to have a Gamma distribution, we need U to be non-negative, which we do since both X and Y have Gamma distributions. Note that the transformation is indeed one to one, since $\forall (u_1, v_1), (u_2, v_2) \in R_{UV}$ with

$$\begin{aligned} u_1 &= x_1 + y_1 \quad v_1 = \frac{x_1}{x_1 + y_1} \\ u_2 &= x_2 + y_2 \quad v_2 = \frac{x_2}{x_2 + y_2}, \end{aligned}$$

we have that, if we let ϕ denote the transformation,

$$\begin{aligned} &\phi(u_1, v_1) = \phi(u_2, v_2) \\ \implies &\left(x_1 + y_1, \frac{x_1}{x_1 + y_1} \right) = \left(x_2 + y_2, \frac{x_2}{x_2 + y_2} \right) \end{aligned}$$

which then

$$x_1 + y_1 = x_2 + y_2 \tag{12.1}$$

$$\begin{aligned} \frac{x_1}{x_2 + y_2} &= \frac{x_2}{x_2 + y_2} \\ \overset{\text{Equation (12.1)}}{\implies} x_1 &= x_2 \\ \overset{\text{Equation (12.1)}}{\implies} y_1 &= y_2. \end{aligned}$$

We shall now get the Jacobian so that we may use Theorem 44, so that we may consequently get the distributions for U and V.

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ 1 - v & -u \end{vmatrix} = -vu - u(1 - v) = -u$$

$$|J| = u$$

By ☕ Theorem 44, and since X and Y are independent, we have

$$
\begin{aligned}
f_{U,V}(u,v) = f_{X,Y}(x,y) \cdot |J| &= f_X(x) f_Y(y) \cdot |J| \\
&= \frac{x^{a-1}e^{-x}}{\Gamma(a)} \frac{y^{b-1}e^{-y}}{\Gamma(b)} \cdot |J| \\
&= \frac{e^{-a}e^{-b}}{\Gamma(a)\Gamma(b)} (uv)^{a-1} u^{b-1} (1-v)^{b-1} u \\
&= \underbrace{\frac{u^{a+b-1}e^{-(a+b)}}{\Gamma(a+b)}}_{\textit{pdf of } \mathrm{Gam}(a+b,1)} \cdot \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} v^{a-1}(1-v)^{b-1}}_{\textit{pdf of } \mathrm{Beta}(a,b)}
\end{aligned}
$$

We have already shown that U is a non-negative rv, and so $U \sim \mathrm{Gam}(a+b,1)$ as required. Note that for V, we have $X + Y > X > 0 \implies 1 > \frac{X}{X+Y} > 0 \quad \because X + Y \neq 0$ and so $0 < V < 1$. Therefore $V \sim \mathrm{Beta}(a,b)$.

With that, we can look for $E[V]$.

$$
\begin{aligned}
E[V] &= \int_0^1 v \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} v^{a-1}(1-v)^{b-1} \, dv \\
&= \frac{\Gamma(a+b)\Gamma(a+1)}{\Gamma(a)\Gamma(a+b+1)} \int_0^1 \underbrace{\frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b)} v^a (1-v)^{b-1}}_{\textit{pdf of } \mathrm{Beta}(a+1,b)} \, dv \\
&= \frac{a\Gamma(a)\Gamma(a+b)}{(a+b)\Gamma(a)\Gamma(a+b)} = \frac{a}{a+b}
\end{aligned}
$$

13 Lecture 13 Jun 14th 2018

13.1 Functions of Random Variables (Continued)

13.1.1 One to One Bivariate Transformations (Continued)

Example 13.1.1 (Example 4.5 (Course Notes 4.2.8))

Suppose X and Y are continuous rvs with joint pdf

$$f(x,y) = e^{-x-y}\mathbb{1}_{\{0<x,y<\infty\}}.$$

Let $U = X + Y$ and $V = X - Y$. Find the joint pdf of U and V. (Note: Be sure to specify the support of (U,V).) Then, find the marginal pdf of U and V respectively.

Solution

Note that because $0 < x, y < \infty$,

$$\begin{aligned} u = x + y &\implies 0 < u < \infty \\ v = x - y &\implies -\infty < v < \infty \end{aligned}$$

The Jacobian is

$$\left|\frac{\partial(u,v)}{\partial(x,y)}\right| = \left\| \begin{matrix} 1 & 1 \\ 1 & -1 \end{matrix} \right\| = |-2| = 2$$

and so

$$\left|\frac{\partial(x,y)}{\partial(u,v)}\right| = \left|\frac{\partial(u,v)}{\partial(x,y)}\right|^{-1} = \frac{1}{2}.$$

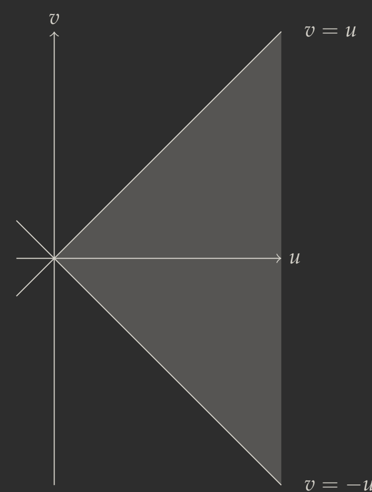
Note that

$$X = \frac{U-V}{2} \quad Y = \frac{U-V}{2}.$$

and so

$$x = \frac{u+v}{2} > 0 \implies u > -v$$
$$y = \frac{u-v}{2} > 0 \implies u > v.$$

The diagram to the right shows the support of U, V. With that,



$$\begin{aligned} g(u,v) &= f\left(\frac{u+v}{2}, \frac{u-v}{2}\right) \cdot \frac{1}{2} \\ &= \frac{1}{2} e^{-\left(\frac{u+v}{2}\right) - \left(\frac{u-v}{2}\right)} = \frac{1}{2} e^{-u}. \end{aligned}$$

With that, the marginal pdf of V is[1]

$$\begin{aligned} v < 0 &\implies f_V(v) = \int_{-v}^{\infty} \frac{1}{2} e^{-u} \, du = \frac{1}{2} e^{v} \\ v \geq 0 &\implies f_V(v) = \int_{v}^{\infty} \frac{1}{2} e^{-v} \, du = \frac{1}{2} e^{-v}. \end{aligned}$$

For the marignal pdf of U, we have

$$f_U(u) = \int_{-u}^{u} \frac{1}{2} e^{-u} \, dv = ue^{-u} \quad \text{for } 0 < u \leq \infty.$$

[1] V is also called the Double Exponential Distribution.

13.1.2 Moment Generating Function Method

This method is particularly useful in finding distributions of sums of independent rvs.

Theorem 46 (Sums of MGF)

Suppose $X_1, ..., X_n$ are independent rvs and X_i has mgf $M_i(t)$ which exists for $t \in (-h, h)$ for some $H > 0$. The mgf of $Y = \sum_{i=1}^{n} X_i$ is given by

$$M_Y(t) = \prod_{i=1}^{n} M_i(t)$$

for $t \in (-h, h)$.

Proof

Observe that

$$\begin{aligned} M_Y(t) &= E\left[e^{tY}\right] = E\left[e^{t\sum_{i=1}^n X_i}\right] = E\left[\prod_{i=1}^n e^{tX_i}\right] \\ &= \prod_{i=1}^n E\left[e^{tX_i}\right] = \prod_{i=1}^n M_i(t) \end{aligned}$$

and $t \in (-h, h)$ is preserved. □

Note

1. *If X_i's are IID rvs each with mgf $M(t)$ then Y has mgf*

$$M_Y(t) = [M(t)]^n \quad \text{for } t \in (-h, h).$$

2. *Used in conjunction with the Uniqueness Theorem for mgfs, this theorem can be used to find the distribution of Y.*

Exercise 13.1.1

Show the following results:

1. *If $X \sim \mathrm{Gam}(\alpha, \beta)$, where $\alpha \in \mathbb{N}$, then $\frac{2X}{\beta} \sim \chi^2(2\alpha)$.*

2. *If $X_i \sim \mathrm{Gam}(\alpha_i, \beta)$, $i = 1, ..., n$ independently, then*

$$\sum_{i=1}^n X_i \sim \mathrm{Gam}\left(\sum_{i=1}^n \alpha_i,\ \beta\right).$$

3. *If $X_i \sim \mathrm{Gam}(1, \beta) = \mathrm{Exp}(\beta)$, $i = 1, ..., n$ independently, then*

$$\sum_{i=1}^n X_i \sim \mathrm{Gam}(n,\ \beta).$$

4. *If $X_i \sim \mathrm{Gam}\left(\frac{k_i}{2}, 2\right) = \chi^2(k_i)$, $i = 1, ..., n$ independently, then*

$$\sum_{i=1}^n X_i \sim \chi^2\left(\sum_{i=1}^n k_i\right).$$

5. *If* $X_i \sim \mathrm{N}(\mu, \sigma^2)$*,* $i = 1, ..., n$ *independently, then*

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n).$$

6. *If* $X_i \sim \mathrm{Poi}(\mu_i)$*,* $i = 1, ..., n$ *independently, then*

$$\sum_{i=1}^{n} X_i \sim \mathrm{Poi}\left(\sum_{i=1}^{n} \mu_i \right).$$

7. *If* $X_i \sim \mathrm{Bin}(n_i, p)$*,* $i = 1, ..., n$ *independently, then*

$$\sum_{i=1}^{n} X_i \sim \mathrm{Bin}\left(\sum_{i=1}^{n} n_i, p \right).$$

8. *If* $X_i \sim \mathrm{NB}(k_i, p)$*,* $i = 1, ..., n$ *indepdently, then*

$$\sum_{i=1}^{n} X_i \sim \mathrm{NB}\left(\sum_{i=1}^{n} k_i, p \right).$$

Solution

1. *Using the mgf method,*

$$\begin{aligned}
M_Y(t) &= E\left[e^{tY}\right] = E\left[e^{t \cdot \frac{2x}{\beta}}\right] = \int_0^\infty e^{\frac{2tx}{\beta}} \frac{1}{\beta^\alpha \Gamma(a)} x^{\alpha-1} e^{-\frac{x}{\beta}} \, dx \\
&= \frac{1}{\beta^\alpha} \int_0^\infty \frac{1}{\Gamma(a)} x^{\alpha-1} e^{-\frac{x(1-2t)}{\beta}} \, dx \\
&\overset{(*)}{=} \frac{1}{\beta^\alpha} \left(\frac{\beta}{1-2t} \right)^\alpha \int_0^\infty \underbrace{\frac{1}{\left(\frac{\beta}{1-2t}\right)\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\left(\frac{\beta}{1-2t}\right)}}}_{\text{pdf of } \mathrm{Gam}\left(\alpha, \frac{\beta}{1-2t}\right)} \, dx \\
&= \left(\frac{1}{1-2t} \right)^\alpha = (1-2t)^\alpha
\end{aligned}$$

where in $(*)$ *we note that* $1 - 2t > 0$ *and so* $t > \frac{1}{2}$*. Observe that the mgf of* Y *is the mgf of* $\chi^2(2\alpha)$ *and so by the* *Theorem 20,* $Y = \frac{2X}{\beta} \sim \chi^2(2\alpha)$*.*

2. *Using the mgf method, let* $Y = \sum_{i=1}^{n} X_i$

$$\begin{aligned}
M_Y(t) &= E\left[e^{tY}\right] = E\left[\prod_{i=1}^{n} e^{tX_i}\right] = \prod_{i=1}^{n} E[e^{tX_i}] \\
&= \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} (1-\beta t)^{-\alpha_1} = (1-\beta t)^{-\sum_{i=1}^{n} \alpha_1}
\end{aligned}$$

and we observe that the last term is the mgf of $\operatorname{Gam}\left(\sum_{i=1}^n \alpha_i, \beta\right)$ *and so the result follows.*

3. *From* (2) *is* $\alpha_i = 1$ *for* $i = 1, ..., n$, *then* $\sum_{i=1}^n \alpha_1 = \sum_{i=1}^n 1 = n$. *The result follows.*

4. *By* (2), *we have* $\sum_{i=1}^n X_i \sim \operatorname{Gam}\left(\sum_{i=1}^n \frac{k_1}{2}, 2\right)$. *Then by* (1), *we have that*

$$\sum_{i=1}^n X_i \sim \chi^2\left(\sum_{i=1}^n k_i\right)$$

as required.

5. *We know that* $Y_i = \frac{X_i - \mu}{\sigma} \sim Z(0,1)$. *Let* $W_i = Y_i^2$. *Then*

$$\begin{aligned}
M_{W_i}(t) = E\left[e^{tY_i}\right] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} e^{ty^2}\, dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2(1-2t)^{-1}}\right\} dy \\
&= (1-2t)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}(1-2t)^{-\frac{1}{2}}} \exp\left\{\frac{y^2}{2(1-2t)^{-1}}\right\}}_{\text{pdf of } \mathrm{N}\left(0,(1-2t)^{-\frac{1}{2}}\right)} dy \\
&= (1-2t)^{-\frac{1}{2}},
\end{aligned}$$

which is the mgf of $\chi^2(1)$. *Let* $W = \sum_{i=1}^n W_i$. *Then*

$$M_W(t) = \prod_{i=1}^n (1-2t)^{-\frac{1}{2}} = (1-2t)^{-\frac{n}{2}}$$

which is the mgf of $\chi^2(n)$.

6. *Let* $Y = \sum_{i=1}^n X_i$. *Then*

$$M_Y(t) = \prod_{i=1}^n e^{\mu_i(e^t - 1)} = \exp\left[(e^t - 1)\sum_{i=1}^n \mu_i\right]$$

is the mgf of $\operatorname{Poi}\left(\sum_{i=1}^n \mu_i\right)$.

7. *Let* $Y = \sum_{i=1}^n X_i$. *Then*

$$M_Y(t) = \prod_{i=1}^n (pe^t + q)^{n_i} = (pe^t + q)^{\sum_{i=1}^n n_i}$$

is the mgf of $\operatorname{Bin}\left(\sum_{i=1}^n n_i, p\right)$.

8. *Again, a similar approach: let* $Y = \sum_{i=1}^{n} X_i$. *Then*

$$M_Y(t) = \prod_{i=1}^{n} \left(\frac{1-p}{1-pe^t} \right)^{k_i} = \left(\frac{1-p}{1-pe^t} \right)^{\sum_{i=1}^{n} k_i}$$

is the mgf of $\mathrm{NB}\left(\sum_{i=1}^{n} k_i, p\right)$.

14 Lecture 14 Jun 19th 2018

14.1 Functions of Random Variables (Continued 2)

14.1.1 Moment Generating Function Method (Continued)

Theorem 47 (Gaussian Distribution)

If $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, with $i = 1, ..., n$ independently, then

$$\sum_{i=1}^{n} a_i X_i \sim \mathrm{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

Proof

Let $Y_i = a_i X_i$. Then by Theorem 19, we have

$$M_{Y_i}(t) = M_{X_i}(a_i t) = \exp\left[\mu_i a_i t + \frac{\sigma_i^2 a_i^2 t^2}{2}\right],$$

which implies that $Y_i \sim \mathrm{N}(a_i \mu_i, a_i^2 \sigma_i^2)$ by Theorem 20. Then let $Y = \sum_{i=1}^{n} Y_i$. Then

$$M_Y(t) = \prod_{i=1}^{n} \exp\left[\mu_i a_i t + \frac{\sigma_i^2 a_i^2 t_2}{2}\right] = \exp\left[t \sum_{i=1}^{n} a_i \mu_i + \frac{t^2}{2} \sum_{i=1}^{n} a_i^2 \sigma_i^2\right],$$

which implies that $Y \sim \mathrm{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$. □

Theorem 48 (Properties of the Gaussian Distribution)

Asumme $X_1, ..., X_n \sim \mathrm{N}(\mu, \sigma_2)$, *independently, where* $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, *and* $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$. *Then*

1. $\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
2. (**Cochran's Theorem**) $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(n-1)$.
3. (**t-test**) $\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$.

Proof

1. *Let* $Y_i = \frac{X_i}{n}$. *Theorem 47 implies that* $Y_i \sim \mathrm{N}\left(\frac{\mu}{n}, \frac{\sigma^2}{n^2}\right)$. *Let* $Y = \sum_{i=1}^{n} Y_i$. *Then Theorem 47 implies that* $Y \sim \mathrm{N}\left(\sum_{i=1}^{n}\frac{\mu}{n}, \sum_{i=1}^{n}\frac{\sigma^2}{n^2}\right) = \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$. □
2. *The proof of Cochran's Theorem is beyond the scope of this course, for it uses knowledge from linear algebra and involving Fourier Transforms.* [*Reference - Wikipedia*]
3. *The proof of this statement is non-trivial:* [*Reference - Wikipedia*] [*Reference - Math SE*]

Theorem 49 (F Distribution)

Suppose $X_1, ..., X_n$ *is a random sample from the* $\mathrm{N}(\mu_1, \sigma_1^2)$ *distribution and independently* $Y_1, ..., Y_m$ *is a random sample from the* $\mathrm{N}(\mu_2, \sigma_2^2)$ *distribution. Let*

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ and } S_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2.$$

Then

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1).$$

Proof

This is, once again, an incomplete proof. Rigor shall be appended where it is due or they shall be stated with reference, or if it is at a level that I cannot understand and that I cannot find a reference, I shall state that as well.

Note that the definition of the F distribution is as follows:

$$F_{\gamma_1,\gamma_2} = \frac{\chi^2(\gamma_1)/\gamma_1}{\chi^2(\gamma_2)/\gamma_2}$$

where $\chi^2(\gamma_1)$ *and* $\chi^2(\gamma_2)$ *are independent.*

Now we are given that $X_1, ..., X_n$ *and* $Y_1, ..., Y_m$ *are independent of one another. It can likely be shown, using induction, that the following two are independent of one another:*

$$\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma_1} \right)^2 \sim \chi^2(n-1)$$

$$\sum_{j=1}^{m} \left(\frac{Y_j - \overline{Y}}{\sigma_2} \right)^2 \sim \chi^2(m-1).$$

Then

$$\frac{\sum_{i=1}^{n} \left(\frac{X_1 - \overline{X}}{\sigma_1} \right)^2}{\sum_{j=1}^{m} \left(\frac{Y_j - \overline{Y}}{\sigma_2} \right)^2} = \frac{(n-1)S_1^2/\sigma_1^2}{(m-1)S_2^2/\sigma_2^2}.$$

With that, we have that

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\frac{(n-1)S_1^2}{\sigma_1^2}/(n-1)}{\frac{(m-1)S_2^2}{\sigma_2^2}/(m-1)}.$$

The result follows.

15 Lecture 15 Jun 21st 2018

15.1 Limiting or Asymptotic Distributions

15.1.1 Convergence in Distribution

Definition 52 (Convergence in Distribution)

Let $X_1, X-2, ..., X_n, ...$ be a sequence of rvs such that X_i has cdf F_i, for $i \in \mathbb{N}$. Let X be an rv with cdf F. We say that X_i ***converges in distribution*** *to X and we write*

$$X_n \xrightarrow{D} X$$

if

$$\lim_{i \to \infty} F_i(x) = F(x)$$

at all points x at which F is continuous. We call F the ***limiting*** *or* ***asymptotic distribution*** *of X_i.*

The following theorem, that we shall not prove, will be useful for the remainder of the notes.

Theorem 50 (Taylor Series with Lagrange's Remainder)

Suppose that $f : [a, b] \to \mathbb{R}$ is infinitely differentiable, and $x \in [a, b]$. Then $\forall x \in [a, b]$ and $\forall k \in \mathbb{N}$,

$$f(x) = \sum_{i=0}^{k} \frac{f^{(i)}(c)(x-c)^i}{i!} + \frac{f^{(k+1)}(\zeta_X)(x-c)^{k+1}}{(k+1)!}$$

for some $\zeta_X \in [c, x]$.

The proof for the above involves the use of the **Mean Value Theorem for Integrals**.

Theorem 51 (Generalized Limit Definition of e)

If $b, c \in \mathbb{R}$ *are constants and* $\lim_{n\to\infty} \psi(n) = 0$*, then*

$$\lim_{n\to\infty} \left[1 + \frac{b}{n} + \frac{\psi(n)}{n}\right]^{cn} = e^{bc}.$$

Proof

The above equation can be rewritten as

$$\lim_{n\to\infty} e^{cn \log\left(1 + \frac{b}{n} + \frac{\psi(n)}{n}\right)} = e^{bc}$$

and so it suffices to prove that

$$\lim_{n\to\infty} cn \log\left(1 + \frac{b}{n} + \frac{\psi(n)}{n}\right) = bc$$

Note that the Taylor Expansion with Lagrange's Remainders (Theorem 50) of $\log(1+x)$*, where we pick* $c = 0$ *and* $k = 1$ *for convenience,*

$$\log(1+x) = \frac{\log(1)x^0}{0!} + \frac{\left(\frac{1}{1-0}\right)x^1}{1!} + \frac{-\frac{1}{(1+\zeta_x)^2}x^2}{2!} = x - \frac{x^2}{2(1+\zeta_x)}$$

where $\zeta \in [0, x]$*. Then*

$$cn \log\left[1 + \frac{b}{n} + \frac{\psi(n)}{n}\right] = cb + c\psi(n) - \frac{c(b+\psi(n))^2}{2n(1+\zeta)^2}$$

where $\zeta \in \left[0, \frac{b+\psi(n)}{n}\right]$*. Note that by* ***L'Hôpital's Rule****, we have that*

$$\lim_{n\to\infty} \frac{b+\psi(n)}{n} = \lim_{n\to\infty} \frac{\psi'(n)}{1} = 0$$

and so the possible highest value for ζ *goes to* 0 *as* $n \to \infty$*. Then*

$$\lim_{n\to\infty} \frac{c(b+\psi(n))^2}{2n(1+\zeta)^2} = 0.$$

As a result, we have

$$\lim_{n\to\infty} cn \log\left(1+\frac{b}{n}+\frac{\psi(n)}{n}\right) = bc$$

as required. $\square$

Corollary 52 (Limit definition of e)

If $b, c \in \mathbb{R}$ are constants, then

$$\lim_{n\to\infty}\left(1+\frac{b}{n}\right)^{cn} = e^{bc}$$

Example 15.1.1 (Example 5.1 (Course Notes 5.1.4))

Let $X_i \sim \mathrm{Exp}(1)$, where $i \in \mathbb{N}$, independently so. Consider the sequence of rvs $Y_1, Y_2, ..., Y_n, ...$, where $Y_n = \max(X_1, ..., X_n) - \log n$. Find the limiting distribution of Y_n.

Solution

Firstly, observe that to find the support set of Y_n, note

$$\begin{aligned}0 &< x_1, ..., x_n < \infty \\ 0 &< \max(x_1, ..., x_n) < \infty \\ -\log n &< \max(x_1, ..., x_n) - \log n < \infty\end{aligned}$$

so

$$\mathrm{supp}(Y) = (-\log n, \infty)$$

Now the pf of Y_n is

$$\begin{aligned}F_n(y) &= P(Y_n \leq y) = P(\max(X_1, ..., X_n) - \log n \leq y) \\ &= P(\max(X_1, ..., X_n) \leq y + \log n) \\ &= \prod_{i=1}^n P(X_1 \leq y + \log n) \quad \because X_1, ..., X_n \text{ are IID} \\ &= \prod_{i=1}^n \left[1 - e^{-y-\log n}\right] = \prod_{i=1}^n \left[1 - \frac{e^-y}{n}\right] = \left[1 - \frac{e^-y}{n}\right]^n.\end{aligned}$$

Thus the limiting distribution of Y_i is

$$\lim_{n\to\infty} F_n(y) = \lim_{n\to\infty}\left[1-\frac{e^-y}{n}\right]^n = e^{-e^{-y}} \text{ for } y \in (-\log n, \infty).$$

Example 15.1.2 (Example 5.2 (Course Notes 5.1.5))

Let $X_i \sim \text{Unif}(0, \theta)$, $i \in \mathbb{N}$, independently so. Consider the sequence of random variables $Y_1, Y_2, ..., Y_n, ...$, where $Y_n = \max(X_1, ..., X_n)$. Find the limiting distribution of Y_n.

Solution

Clearly, support of $Y_n = (0, \theta)$. Note that

$$f_{X_i}(x) = \frac{1}{\theta} \mathbb{1}_{0 < x < \theta}.$$

Now since X_i are IID,

$$F_n(y) = P(Y_n \leq y) = \prod_{i=1}^{n} P(X_1 \leq y) = \prod_{i=1}^{n} \frac{y}{\theta} = \left(\frac{y}{\theta}\right)^n \quad y \in (0, \theta)$$

Then the limiting distribution is

$$\lim_{n \to \infty} F_n(y) = \lim_{n \to \infty} \left(\frac{y}{\theta}\right)^n = \begin{cases} 0 & y < \theta \\ 1 & y \geq \theta \end{cases}$$

Define Y to have a distribution such that $P(Y = \theta) = 1$. Then

$$Y_n \xrightarrow{D} Y.$$

Definition 53 (Degenerate Distribution)

A function F is the cdf of a degenerate distribution at value $y = c$ if

$$F(y) = \begin{cases} 0 & y < c \\ 1 & y \geq c \end{cases}.$$

In other words, F is the CDF of a discrete distribution where

$$P(Y = y) = \begin{cases} 1 & y = c \\ 0 & \text{otherwise} \end{cases}$$

We have that the earlier example gives us that the limiting distribution is a degenerate distribution.

15.1.2 Convergence in Probability

Definition 54 (Convergence in Probability)

A sequence of rvs $X_1, X_2, \ldots, X_n, \ldots$ ***converges in probability*** *to an rv X if, for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

or equivalently

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

We write

$$X_n \xrightarrow{P} X.$$

Example 15.1.3 (Example 5.3)

Consider $X \sim \mathrm{Bernoulli}(0.3)$. Define the sequence $X_n = \left(1 + \frac{1}{n}\right) X$, $n \in \mathbb{N}$. Show that $X_n \xrightarrow{P} X$.

Solution

Note that

$$P(X = x) = \begin{cases} 0.3 & x = 1 \\ 0.7 & x = 0 \end{cases}$$

Since $X_n = \left(1 + \frac{1}{n}\right) X$, we have

$$X_1 = 2X,\ X_2 = \frac{3}{2}X,\ X_3 = \frac{4}{3}X,\ \ldots .$$

Note that $|X_n - X| = \left|\left(1 - \frac{1}{n}\right) X - X\right| = \left|\frac{1}{n} X\right| = \frac{1}{n} X$, *so*

$$P(|X_n - X| < \varepsilon) = P\left(\frac{1}{n} X < \varepsilon\right) = P(X < n\varepsilon)$$
$$= \begin{cases} P(X = 0) = 0.7 & n < \frac{1}{\varepsilon} \\ P(X = 1) + P(X = 0) = 1 & n \geq \frac{1}{\varepsilon} \end{cases}.$$

Therefore

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

and so

$$X_n \xrightarrow{P} X.$$

We shall look into the following proposition in the next lecture.

Proposition

Suppose $\forall n \in \mathbb{N}$, $\{X_n\}$ *is a sequence of rvs. Then*

$$X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$$

where X *is an rv.*

---

16 Lecture 16 Jun 26th 2018

16.1 Limiting or Asymptotic Distributions (Continued)

16.1.1 Convegence in Probability (Continued)

Before proving the proposition introduced in last class, we require the following lemma.

Lemma 53

Let X, Y be rvs, $a \in \mathbb{R}$, and $\varepsilon > 0$. Then

$$P(Y \leq a) \leq P(X \leq a + \varepsilon) + P(|Y - X| > \varepsilon)$$

Proof

Note that

$$\begin{aligned} P(Y \leq a) &= P(Y \leq a, X \leq a + \varepsilon) + P(Y \leq a, X > a + \varepsilon) && (16.1) \\ &\leq P(X \leq a + \varepsilon) + P(Y - X \leq a - X, a - X < -\varepsilon) && (16.2) \\ &\leq P(X \leq a + \varepsilon) + P(Y - X < -\varepsilon) && (16.3) \\ &\leq P(X \leq a + \varepsilon) + P(Y - X < -\varepsilon) + P(Y - X > \varepsilon) && (16.4) \\ &= P(X \leq a + \varepsilon) + P(|Y - X| > \varepsilon) \end{aligned}$$

where Equation (16.1) is by ***Law of Total Probability****, Equation (16.2) and Equation (16.3) are by the fact that for non-empty sets A and B, $P(A \cap B) \leq P(A)$, and Equation (16.4) is because $P(Y - X > \varepsilon) \geq 0$.* $\square$

Proposition 54 (Convergence in Probability Implies Convergence in Distribution)

This only states for the case of scalar random variables.

Suppose $\forall n \in \mathbb{N}$, $\{X_n\}$ is a sequence of rvs. Then

$$X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$$

where X is an rv.

Proof

By Lemma 53, we have that[1]

$$P(X_n \leq a) \leq P(X \leq a + \varepsilon) + P(|X_n - X| > \varepsilon)$$
$$P(X \leq a - \varepsilon) \leq P(X_n \leq a) + P(|X_n - X| > \varepsilon)$$

[1] We choose to use Lemma 53 to have $P(X_n \leq a)$ in two inequalities. Note that our goal is to show that

$$\lim_{n\to\infty} P(X_n \leq a) = P(X \leq a).$$

We already know that $\varepsilon \to 0$, so we should use that to our advantage in the case of X. Also, in hindsight, it is clear why we choose this method as the **Squeeze Theorem** is suitable to help us reach our conclusion.

Then

$$\begin{aligned} &P(X \leq a - \varepsilon) - P(|X_n - X| > \varepsilon) \\ &\leq P(X_n \leq a) \\ &\leq P(X \leq a + \varepsilon) + P(|X_n - X| > \varepsilon) \end{aligned}$$

As $n \to \infty$, by assumption that $X_n \xrightarrow{P} X$, we have that $P(|X_n - X| > \varepsilon) \to 0$, and so

$$P(X \leq a - \varepsilon) \leq \lim_{n\to\infty} P(X_n \leq a) \leq P(X \leq a + \varepsilon)$$

Now as $\varepsilon \to 0^+$, note that we must have the cdf of X be continuous on a by assumption, and so

$$P(X \leq a) \leq \lim_{n\to\infty} P(X_n \leq a) \leq P(X \leq a)$$

*and so by the **Squeeze Theorem**,*

$$\lim_{n\to\infty} P(X_n \leq a) = P(X \leq a)$$

which then

$$X_n \xrightarrow{D} X$$

as required. $\square$

Definition 55 (Convergence in Probability to a Constant)

A sequence of rvs $\{X_i\}_{i \in \mathbb{N}}$ converges in probability to a constant $b \in \mathbb{R}$ if $\forall \varepsilon > 0$,

$$\lim_{n \to \infty} P(|X_n - b| \geq \varepsilon) = 0$$

or equivalently

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

We write

$$X_n \xrightarrow{P} b.$$

Example 16.1.1 (Example 5.4 - Weak Law of Large Numbers)

The Weak Law of Large Numbers is also known as **Bernoulli's Theorem**.

Let $X_1, ..., X_n, ...$ be a sequence of IID rvs, each having mean μ and variance σ^2. Let

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then $\overline{X}_n \xrightarrow{P} \mu$.

Solution

Observe that

$$E\left[\overline{X}_n\right] = E\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{n\mu}{n} = \mu,$$

Also, we have that

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

Then by Theorem 16, we have that $\forall \varepsilon > 0$,

$$P(\left|\overline{X}_n - \mu\right| \geq \varepsilon) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Therefore, we have

$$0 \leq P(\left|\overline{X}_n - \mu\right| \geq \varepsilon) \leq \min\left(1, \frac{\sigma^2}{n\varepsilon^2}\right)$$

As $n \to \infty$, by ***Squeeze Theorem****, we have that*

$$\lim_{n \to \infty} P(\left|\overline{X}_n - \mu\right| \geq \varepsilon) = 0$$

i.e. $\overline{X}_n \xrightarrow{P} \mu$.

Note that the converse of Proposition 54 is not generally true. **(Example required)** However, with an additional condition, the converse becomes true.

Proposition 55 (Partial Converse of Proposition 54)

Suppose $\forall n \in \mathbb{N}$, $\{X_n\}$ *is a sequence of rvs, each with cdf* F_n. *If*

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P(X_n \leq x) = \begin{cases} 0 & x < c \\ \\ 1 & x > c \end{cases},$$

then

$$X_n \xrightarrow{P} c.$$

Proof

Note that for $\varepsilon > 0$, *we have*

$$\begin{aligned} P(|X_n - c| > \varepsilon) &= P(X_n - c < -\varepsilon) + P(X_n - c > \varepsilon) \\ &= P(X_n < c - \varepsilon) + P(X_n > c + \varepsilon) \\ &= 1 - P(X_n < c + \varepsilon) + P(X_n < c - \varepsilon) \end{aligned}$$

By assumption, we have that

$$\begin{aligned} &\lim_{n\to\infty} P(X_n \leq c + \varepsilon) = 1 \\ &\lim_{n\to\infty} P(X_n \leq c - \varepsilon) = 0 \end{aligned}$$

So

$$\lim_{n\to\infty} P(|X_n - c| > \varepsilon) = 1 - 1 - 0 = 0$$

and so by definition, we have $X_n \xrightarrow{P} c$. $\square$

Definition 56 (Double Parameter Exponential Distribution)

We say that an rv $X \sim \mathrm{Exp}(\lambda, \theta)$ *when* X *has pdf*

$$f(x) = e^{-\frac{x-\theta}{\lambda}} \text{ for } x \in (\theta, \infty)$$

where λ is the scale parameter, and θ the location parameter.

Example 16.1.2 (Example 5.5 (Course Notes 5.2.5))

Let $X_i \sim \mathrm{Exp}(1, \theta)$, $i = 1, 2, \ldots$, independently. Consider the sequence of rvs $Y_1, Y_2, \ldots$ where $Y_n = \min(X_1, X_2, \ldots, X_n)$, $n = 1, 2, \ldots$. Show that $Y_n \xrightarrow{P} \theta$.

Solution

Since we want to show that $Y_n \xrightarrow{P} \theta$, and θ is a constant, we can use *Proposition 55. Thus we need to show that*

$$\lim_{n\to\infty} F_{Y_n}(y) = \lim_{n\to\infty} P(Y_n \leq y) = \begin{cases} 0 & y < \theta \\ 1 & y > \theta \end{cases}$$

Since Y_n is defined as the minimum of n of the first X_i rvs, we need to use the ***Law of Total Probability*** *in order to be able to make sense of the order statistics, i.e.*

$$P(Y_n \leq y) = 1 - P(Y_n > y)$$

Now

$$\begin{aligned} P(Y_n > y) &= \prod_{i=1}^{n} P(X_i > y) = \prod_{i=1}^{n} \left[\int_y^\infty e^{-(x-\theta)}\, dx \right] \\ &= \prod_{i=1}^{n} \left[e^{-(y-\theta)} \right] = e^{-n(y-\theta)} \end{aligned}$$

Thus

$$\begin{aligned} \lim_{n\to\infty} P(Y_n \leq y) &= \begin{cases} 1 - \lim_{n\to\infty} P(Y_n > y) & y > \theta \\ 0 & y \leq \theta \end{cases} \\ &= \begin{cases} 1 - \lim_{n\to\infty} e^{-n(y-\theta)} = 1 & y > \theta \\ 0 & y \leq \theta \end{cases} \end{aligned}$$

since if $y < \theta$

$$P(Y_n \leq y) = P(\min(X_1, X_2, \ldots, X_n) \leq y) = 0.$$

Note that if $y = \theta$, then

$$e^{-n(y-\theta)} = 1.$$

The proof is complete with Proposition 55.

16.1.2 Limit Theorems

Proposition 56 (Convergence in Distribution and MGF)

Let $X_1, X_2, ..., X_n, ...$ be a sequence of rvs such that X_n has mgf $M_n(t)$. Let X be an rv with mgf $M(t)$.

$$X_n \xrightarrow{D} X \iff \left[\exists h > 0 \; \forall t \in (-h, h) \quad \lim_{n\to\infty} M_n(t) = M(t)\right]$$

Proof

Note that

$$\lim_{n\to\infty} M_n(t) = \lim_{n\to\infty} \int_{\mathrm{supp}(X_n)} e^{tx} \frac{d}{dx} F_{X_n}(x)\, dx$$

and

$$M(t) = \int_{\mathrm{supp}(X)} e^{tx} \frac{d}{dx} F_X(x)\, dx$$

The result follows assuming that the integral converges[2]. □

[2] This allows us to "swap" the limit, the integral sign, and the differential operator.

Example 16.1.3

Consider the sequence $Y_1, Y_2, ...,$ where $Y_i \sim \mathrm{Bin}\left(n, \frac{\mu}{n}\right)$, for $i = 1, 2, ...$ Find the limiting distribution of Y_n.

Solution (Example 5.6)

Note that $Y_n \sim \mathrm{Bin}\left(n, \frac{\mu}{n}\right) \implies$

$$M_{Y_n}(t) = \left(\frac{\mu}{n}e^t + 1 - \frac{\mu}{n}\right)^n = \left(1 + \frac{\mu\left(e^t - 1\right)}{n}\right)^n.$$

So

$$\lim_{n\to\infty} M_{Y_n}(t) = \lim_{n\to\infty}\left(1 + \frac{\mu\left(e^t - 1\right)}{n}\right)^n = e^{\mu\left(e^t - 1\right)}$$

which is the mgf of a Poisson Distribution. Thus by Proposition 56, we have that

$$Y_n \xrightarrow{D} Y \sim \text{Poi}(\mu).$$

Example 16.1.4 (Example 5.7)

Let $Y_1, Y_2, \ldots \sim \text{G}(\mu, \sigma)$. Then

$$\frac{\overline{Y}_n - \mu}{\sigma / \sqrt{n}} \sim \text{G}(0, 1)$$

where $\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

Solution

Note that by Theorem 19, we have that

$$M_{\overline{Y}_n}(t) = \prod_{i=1}^{n} M_{Y_i}\left(\frac{t}{n}\right) = \prod_{i=1}^{n} e^{\frac{\mu t}{n} + \frac{(\sigma t)^2}{2n^2}} = e^{\mu t + \frac{\sigma^2 t^2}{n}}.$$

which is the mgf of $\text{G}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. *Let* $X_n = \frac{\overline{Y}_n - \mu}{\sigma / \sqrt{n}}$.

Then again, by Theorem 19, we have

$$\begin{aligned} M_{X_n}(t) &= e^{-\frac{\mu t}{\sigma/\sqrt{n}}} M_{\overline{Y}_n}(t) \left(\frac{t}{\sigma/\sqrt{n}}\right) \\ &= e^{-\frac{\mu t}{\sigma/\sqrt{n}}} e^{\frac{\mu t}{\sigma/\sqrt{n}} + \frac{\sigma^2/n}{2}\left(\frac{t}{\sigma/\sqrt{n}}\right)^2} = e^{\frac{t^2}{2}} \end{aligned}$$

which is the mgf of $\text{G}(0, 1)$.

Example 16.1.5 (Example 5.8)

Let $Y_n \sim \text{Gam}(n, 2)$. Find the limiting distribution of

$$W_N = \frac{1}{2\sqrt{n}}(Y_n - 2n).$$

Solution

By Theorem 19, we have

$$\begin{aligned}
M_{W_n}(t) &= e^{-t\sqrt{n}} M_{Y_n}(t)\left(\frac{t}{2\sqrt{n}}\right) = e^{-t\sqrt{n}}\left(1 - 2\left(\frac{t}{2\sqrt{n}}\right)\right)^{-n} \\
&= \left[e^{\frac{t}{\sqrt{n}}}\left(1 - \frac{t}{\sqrt{n}}\right)\right]^{-n} = \left[\left(1 - \frac{t}{\sqrt{n}}\right)\sum_{k=0}^{\infty}\frac{(t/\sqrt{n})^k}{k!}\right]^{-n} \\
&= \left[\left(1 - \frac{t}{\sqrt{n}}\right)\left(1 + \frac{t}{\sqrt{n}} + \frac{t^2}{2n} + \frac{t^3}{3\sqrt{n^3}} + \ldots\right)\right]^{-n} \\
&= \left(1 - \frac{t}{\sqrt{n}} + \frac{t}{\sqrt{n}} - \frac{t^2}{n} + \frac{t^2}{2n} - \frac{t^3}{2\sqrt{n^3}} + \frac{t^3}{3\sqrt{n^3}} - \frac{t^4}{3n^2} + \ldots\right)^{-n} \\
&= \left(1 - \frac{t^2}{2n} - \underbrace{\frac{t^3}{6\sqrt{n^3}} - \frac{t^4}{3n^2} + \ldots}_{\to 0 \text{ as } n \to \infty}\right)^{-n}
\end{aligned}$$

So we have, by Theorem 51,

$$\lim_{n\to\infty} M_{W_n}(t) = \lim_{n\to\infty}\left(1 - \frac{t^2/2}{n} - \frac{t^3}{6\sqrt{n^3}} - \frac{t^4}{3n^2} + \ldots\right)^{-n} = e^{-(-\frac{t^2}{2})} = e^{\frac{t^2}{2}},$$

which is the mgf of $\mathrm{G}(0,1)$*. So we know that*

$$W_n \xrightarrow{D} W \sim \mathrm{G}(0,1)$$

for some limiting distribution W that has a $\mathrm{G}(0,1)$ *distribution.*

Theorem 57 (Central Limit Theorem)

Suppose $X_1, X_2, \ldots$ is a sequence of IID rvs with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$, for $i = 1, 2, \ldots$. Then

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} Z \sim \mathrm{N}(0,1)$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

We actually also need the condition that all moments of $\frac{X_i - \mu}{\sigma}$ exists and is bounded.

Proof

Since the first and second moment exist, the mgf must be well-defined, and is at least of class C^2 [3]. We will use Proposition 56 and

[3] You should remember this from your calculus courses. Otherwise, [Reference - Wiki].

Theorem 50 in this proof. Firstly note that

$$E\left[\sum_{i=1}^{n} X_i\right] = n\mu \text{ and } \operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2$$

Let $\Sigma X_n = \sum_{i=1}^n X_i$ *and* $Z = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$. *Note that*

$$Z = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} = \frac{n\overline{X}_n - \mu n}{\sigma\sqrt{n}} = \frac{\Sigma X_n - n\mu}{\sigma\sqrt{n}} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma\sqrt{n}}.$$

Let $Y_i = \frac{X_i - \mu}{\sigma}$ *so that*

$$Z = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} Y_i.$$

Now the mgf of Z, by Theorem 19 *and* Proposition 39*, is*

$$M_Z(t) = \prod_{i=1}^{n} M_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)^n. \tag{16.5}$$

Note that

$$E(Y_i) = E\left(\frac{X_i - \mu}{\sigma}\right) = \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} = 0$$
$$\operatorname{Var}(Y_i) = \operatorname{Var}\left(\frac{X_i - \mu}{\sigma}\right) = \frac{1}{\sigma^2}\operatorname{Var}(X_i) = 1.$$

So by Theorem 50 *up to the second derivative, we have*

$$M_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + \frac{M_{Y_i}^{(3)}(\zeta)\left(\frac{t}{\sqrt{n}}\right)^3}{3!}$$

where $\zeta \in \left[0, \frac{t}{\sqrt{n}}\right]$. *Note that the last term in the above equation goes to* 0 *faster than the second term as* $n \to \infty$, *since we would have* $n^{\frac{3}{2}}$ *in the denominator, which is larger than* n *in the second term. With that, continuing with* Equation (16.5) *and now taking the limit as* $n \to \infty$, *we have*

$$\lim_{n\to\infty} M_Z(t) = \lim_{n\to\infty}\left[1 + \frac{t^2}{2n} + o\left(\frac{t}{n}\right)\right]^n$$

where $o\left(\frac{t}{n}\right)$ *represents the last term*[4]. *Then by* Theorem 51*, we have*

[4] This is a commonly used notation called the Little-o Notation.

$$\lim_{n\to\infty} M_Z(t) = e^{\frac{t^2}{2}}$$

which is the mgf of $\mathrm{N}(0,1)$. *This completes the proof.* □

Example 16.1.6 (Example 5.9)

Revisit Example 16.1.5 using the ☕ Central Limit Theorem. Given

$$W_n = \frac{Y - 2n}{2\sqrt{n}}$$

where $Y \sim \mathrm{Gam}(n, 2)$, find sequences a_n and b_n such that

$$a_n(Y - b_n) \xrightarrow{D} Z \sim G(0, 1).$$

This involves the knowledge that the sum of n exponential distributions with mean μ results in a gamma distribution with parameter n and μ. Since this has never been proven in this set of notes, it shall be proven here in this solution.

Solution

Let $X_1, X_2, \ldots$ be a sequence of IID rvs with distribution $\mathrm{Exp}(2)$. Note that we then have $E(X_i) = 2$ and $\mathrm{Var}(X_i) = 4$. Let $Y_n = \sum_{i=1}^n X_i$. Note that

$$M_{Y_n}(t) = [M_{X_i}(t)]^n = (1 - 2t)^{-n}$$

which is the mgf of $\mathrm{Gam}(n, 2)$. By CLT, we have that

$$\frac{\frac{1}{n}Y_n - 2}{2/\sqrt{n}} \xrightarrow{D} \mathrm{G}(0, 1).$$

Note that

$$\frac{\frac{1}{n}Y_n - 2}{2/\sqrt{n}} = \frac{\sqrt{n}}{2n}(Y_n - 2n)$$

and so

$$a_n = \frac{1}{2}\sqrt{n} \text{ and } b_n = 2n$$

Example 16.1.7 (Example 5.10 (Course Notes 5.3.6))

Suppose $Y_n \sim \chi^2(n)$, $n = 1, 2, \ldots$. Consider the sequence of rvs $Z_1, Z_2, \ldots$ where $Z_n = (Y_n - n)/\sqrt{2n}$. Show that

$$Z_n = \frac{Y_n - n}{\sqrt{2n}} \xrightarrow{D} Z \sim G(0, 1).$$

Once again our class pulls the card of "not explaining or showing stuff that they are supposed to". We will be using the fact that a sum of n Chi-Squared Distirbutions, each with degree of freedom k, will result in a gamma distribution with parameter $\frac{nk}{2}$ and 2. This will be proven in this exercise.

Solution

Let $X_1, X_2, \ldots \sim \chi^2(1)$ be IID rvs, and so $E[X_i] = 1$ and $\mathrm{Var}(X_i) = 2$ for $1 \leq i \leq n$, and let $Y_n = \sum_{i=1}^n X_i$. Note that

$$M_{Y_n}(t) = [M_{X_i}(t)]^n = (1 - 2t)^{-\frac{n}{2}}$$

is the mgf of $\chi^2(n)$. By CLT, we have that

$$\frac{\frac{1}{n}\sum_{i=1}^n X_i - 1}{\sqrt{2}/\sqrt{2}} \xrightarrow{D} Z \sim \mathrm{G}(0, 1).$$

Note that

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - 1}{\sqrt{2}/\sqrt{n}} = \frac{Y_n - n}{\sqrt{2n}/\sqrt{n}} = \frac{Y_n - n}{\sqrt{2n}}.$$

This completes our example.

Example 16.1.8 (Example 5.11 (Course Notes 5.3.7))

Suppose $Y_n \sim \mathrm{Bin}(n, p)$, $n = 1, 2, \ldots$. *Consider the sequence of rvs* $Z_1, Z_2, \ldots$, *where* $Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}}$. *Show that*

$$Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}} \xrightarrow{D} Z \sim \mathrm{N}(0, 1).$$

Solution

Let $X_1, X_2, \ldots$ *be a sequence of IID Bernoulli trials with success probability* $0 \leq p \leq 1$. *Note that for* $i \in \mathbb{N}$, *we have that* $E[X_i] = p$ *and* $\mathrm{Var}(X_i) = p(1-p)$. *Now let* $Y_n = \sum_{i=1}^{n} X_i$. *We note that this satisfies our assumption, i.e.* $Y_n \sim \mathrm{Bin}(n, p)$ *since*

$$M_{Y_n}(t) = [M_{X_1}]^n = (pe^t + 1 - p)^n$$

is the mgf of $\mathrm{Bin}(n, p)$. *By CLT, we have that*

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - p}{\sqrt{p(1-p)}/\sqrt{n}} \xrightarrow{D} Z \sim \mathrm{N}(0, 1).$$

Note that

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - p}{\sqrt{p(1-p)/\sqrt{n}}} = \frac{Y_n - np}{\sqrt{np(1-p)}}$$

and so this completes our proof.

Proposition 58 (Other Limit Theorems)

1. $X_n \xrightarrow{P} a \land g$ *continuous at* $x = a \implies g(X_n) \xrightarrow{P} g(a)$.

2. $X_n \xrightarrow{P} a \land Y_n \xrightarrow{P} b \land g$ *continous at* $(x, y) = (a, b) \implies g(X_n, Y_n) \xrightarrow{P} g(a, b)$.

3. (***Slutsky's Theorem***) $X_n \xrightarrow{D} X \land Y_n \xrightarrow{P} b \land g$ *continous at* $(x, y) = (x, b)$ *for all* $x \in \mathrm{supp}(X) \implies g(X_n, Y_n) \xrightarrow{D} g(X, b)$.

4. (***Continuous Mapping Theorem***) $X_n \xrightarrow{D} X \land g$ *continous at all* $x \in \mathrm{supp}(X) \implies g(X_n) \xrightarrow{D} g(X)$.

Proof

1. *Since g is continuous at a, we have that*

$$\forall \varepsilon > 0 \, \exists \delta > 0 \, \forall x \in \text{supp}(X) \, (|x - a| < d \implies |g(x) - g(a)| < \varepsilon)$$

It follows that

$$P[|g(X_n) - g(a)| < \varepsilon] \geq P[|X_n - a| < \delta]$$

since $P(B) \geq P(A)$ whenever $A \subset B$ [5]. *Since $X_n \xrightarrow{P} a$, we have that for any $\delta > 0$, we have*

$$1 \geq \lim_{n \to \infty} P[|g(X_n) - g(a)| < \varepsilon] \geq \lim_{n \to \infty} P[|X_n - a| < \delta] = 1.$$

Thus

$$\lim_{n \to \infty} P[|g(X_n - g(a)| < \varepsilon] = 1$$

and so

$$g(X_n) \xrightarrow{P} g(a)$$

as required.

2. *This is simply a more general case than* (1). □

3. *See this Wikipedia article for a proof. Requires knowledge of measure theory (in particular, convergence of measures).*

4. *See this Wikipedia article for a proof. Requires knowledge of measure theory (in particular, convergence of measures).*

[5] ~~While I do not have a rigorous proof of this for our case here, it is a sensible result seeing that δ depends on ε.~~ My hunch was right: thinking about the two probability measures using the definition of a random variable, we see that the $w \in S$ that works for $|X_n - a| < \delta$ will definitely work for $|g(X_n) - g(a)| < \varepsilon$, but the converse is not necessarily true. Therefore, the events covered in the left term is a larger set that contains the event on the right.

Example 16.1.9

If $X_n \xrightarrow{P} 10$ and $Y_n \xrightarrow{P} 2$, then $X_n Y_n \xrightarrow{P} 20$ since $g(x, y) = xy$ is continuous at $(10, 2)$.

If $Z_n \xrightarrow{D} Z \sim G(0, 1)$ and $a_n \xrightarrow{P} a$ where a is a constant, then $a_n Z_n \xrightarrow{D} aZ \sim (0, a)$ since $g(a, z) = az$ is continous at (a, z) for all $z \in \text{supp}(Z)$..

Example 16.1.10 (Example 5.12 (Course Notes 5.3.10))

If $X_n \xrightarrow{P} a > 0$, $Y_n \xrightarrow{P} b \neq 0$ and $Z_n \xrightarrow{D} Z \sim G(0, 1)$, the find the limiting distributions of each of the following:

1. $\sqrt{X_n}$

2. $X_n + Y_n$

3. $X_n Z_n$

4. $\frac{1}{Z_n}$

Solution

1. *Since $a > 0$, we have the function $g(x) = \sqrt{x}$ is continuous on a, and so by Proposition 58, we have that $\sqrt{X_n} \xrightarrow{P} \sqrt{a}$.*

2. *Since the function $g(x, y) = x + y$ is continous on the real number 2-tuple (a, b), we have that $X_n + Y_n \xrightarrow{P} a + b$.*

3. *Since the function $g(x, z) = xz$ is continuous on the real number 2-tuple (x, z) for all $x \in \operatorname{supp}(X_n)$ and $z \in \operatorname{supp}(Z_n)$, by* ***Slutsky's Theorem****, we have $X_n Z_n \xrightarrow{D} aZ \sim \mathrm{G}(0, a)$.*

4. *Note that the function $g(z) = \frac{1}{z}$ is continuous for all $z \in \operatorname{supp}(X)$ except at $z = 0$. But since $Z \sim \mathrm{G}(0, 1)$ is a continuous distribution, at a single point $z = 0$, $P(Z = 0) = 0$, and in a distribution it is of negligible value[6]. Therefore, we have that*

$$\frac{1}{Z_n} \xrightarrow{D} \frac{1}{Z}$$

[6] This is a painful thing to write down for me...

Example 16.1.11 (Example 5.13 (Course Notes 5.3.11))

Suppose $X_1, X_2, \ldots \sim \mathrm{Poi}(\mu)$ are IID rvs. Define $Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\overline{X}_n}}$. Find the limiting distribution of Z_n.

Solution

Firstly, note that the parameter for a Poisson distribution is positive. Note that by CLT, we have

$$\frac{\overline{X}_n - \mu}{\sqrt{\mu}/\sqrt{n}} \sim Z \sim \mathrm{G}(0, 1)$$

Note that since $\sqrt{m}$ is a constant, we have that $\sqrt{m} \xrightarrow{P} \sqrt{m}$ is trivially true. Thus by our limit theorems, we have

$$\sqrt{n}(\overline{X}_n - \mu) \sim \sqrt{m}Z \sim \mathrm{G}(0, \sqrt{m}).$$

Now by the Weak Law of Large Numbers, we have

$$\overline{X}_n \xrightarrow{P} \mu.$$

Since a function $g(x) = \sqrt{x}$ is continuous for $x = \mu > 0$, we have that

$$\sqrt{\overline{X}_n} \xrightarrow{P} \sqrt{m}.$$

Then by the ***Continuous Mapping Theorem****, we have that*

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\overline{X}_n}} \sim \frac{\sqrt{\mu}Z}{\sqrt{\mu}} \sim \mathrm{G}(0,1).$$

Theorem 59 (Generalized δ-Method)

Let $X_1, X_2, \ldots$ be a sequence of rvs such that

$$n^b(X_n - a) \xrightarrow{D} X$$

for some $b > 0$. Suppose the function $g(x)$ is differentiable at a and $g'(a) \neq 0$. Then

$$n^b[g(X_n) - g(a)] \xrightarrow{D} g'(a)X.$$

Proof

Using the ***Mean Value Theorem****, we have that*

$$g(X_n) = g(a) + g'(c)(X_n - a) \tag{16.6}$$

for some c in between X_n and a. Since $X_n \xrightarrow{P} a$ [7]*, and, WLOG, $X_n < c < a$, $c \xrightarrow{P} a$. Then by the* ***Continuous Mapping Theorem****, we have $g(c) \xrightarrow{P} g(a)$.*

[7] **I have no idea why...**

Now by rearranging Equation (16.6) and multiplying both sides by n^b, we get

$$n^b[g(X_n) - g(a)] = g'(c)n^b[X_n - a] \xrightarrow{D} g'(a)X$$

by ***Slutsky's Theorem****.* □

Corollary 60 (δ-Method)

Let $X_1, X_2, \ldots$ be a sequence of IID rvs with mean μ and variance σ^2. Suppose the function $g(x)$ is differentiable at μ and $g'(\mu) \neq 0$. Then

$$\sqrt{n}[g(\overline{X}_n) - g(\mu)] \xrightarrow{D} Z \sim N\left(0, [g'(\mu)]^2\sigma^2\right)$$

Proof

Note that by the same working as in Example 16.1.11 for $\sqrt{n}(\overline{X}_n - \mu)$, we have that[8]

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} Y \sim \mathrm{N}(0, \sigma^2).$$

Then by Theorem 59, we have that

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{D} g'(\mu)Y \sim \mathrm{N}\left(0, [g'(\mu)]^2\sigma^2\right)$$

as required. □

[8] This is actually stated as an assumption on the δ-Method Wikipedia article - [Reference].

Example 16.1.12 (Example 5.14)

Let $X_1, X_2, \ldots \sim \mathrm{Geo}(p)$ is a sequence of IID rvs, each with support $\mathrm{supp}(X_i) = \{0, 1, 2, \ldots\}$. Derive the limiting distribution of

$$W_n = \sqrt{n}\left(\frac{1}{\overline{X}_n} - \frac{p}{1-p}\right)$$

Solution

Note that by CLT, we have that

$$\frac{\overline{X}_n - \frac{1-p}{p}}{\sqrt{\frac{1-p}{p^2}} / \sqrt{n}} \xrightarrow{D} Z \sim \mathrm{N}(0,1)$$

Then using Proposition 58, we have

$$\sqrt{n}\left(\overline{X}_n - \frac{1-p}{p}\right) \xrightarrow{D} \sqrt{\frac{1-p}{p^2}} Z \sim N\left(0, \frac{1-p}{p^2}\right)$$

***We took** $g(x) = \frac{1}{x}$ and the given solution in class somehow circumvents the fact that $x = 0$ is a case.*

Example 16.1.13 (Example 5.15)

Suppose that $X_1, X_2, \ldots \sim \mathrm{Exp}(\theta)$ is a sequence of IID rvs. Find constants a_n and b_n such that

$$W_n = b_n(\overline{X}_n^2 - a_n)$$

has a non-denegerate limiting distribution.

Solution

By CLT, we have

$$\frac{\overline{X}_n - \theta}{\theta/\sqrt{n}} \xrightarrow{D} Z \sim \mathrm{N}(0,1)$$

which then by Proposition 58, we have

$$\sqrt{n}(\overline{X}_n - \theta) \xrightarrow{D} \theta Z \sim \mathrm{N}(0, \theta^2).$$

Let $g(x) = x^2$*. Then* g *is continous on any* $x \in \operatorname{supp}(X_n)$ *and on* θ*. Then using Corollary 60, we have*

$$\sqrt{n}(\overline{X}_n^2 - \theta^2) \sim \theta^2 Z \sim \mathrm{N}(0, 4\theta^2).$$

So we just need to pick $b_n = \sqrt{n}$ *and* $a_n = \theta^2$*.*

17 Lecture 17 Jun 28th 2018

17.1 Estimation

SUPPOSE $X_1, ..., X_n \sim f(x; \theta)$ is an IID sequence of rvs, where $f(x; \theta)$ is the pf of the X_i's. The joint distribution of $X_1, ..., X_n$ is

$$\prod_{i=1}^{n} f(x_i; \theta)$$

where the unknown parameter θ can either be a scalar in Ω, where Ω is the parameter space or the set of possible values of θ, or a vector, i.e. $\theta = \begin{pmatrix} \theta_1 & \theta_2 & \ldots & \theta_n \end{pmatrix}^T$.

We are interested in making inferences about the unknown parameter θ, i.e. we want to find **estimators** (point and interval) of θ and we want to test our hypothesis about θ.

Before proceeding the the rest of this chapter (actual chapter, ahem...), we require the following definitions.

Definition 57 (Statistic)

*A **statistic**, $T = T(X) = T(X_1, ..., X_n)$, is a function of the data which does not depend on any unknown parameter(s).*

Example 17.1.1

Suppose $X_1, ..., X_n$ is a random sample from a distribution with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, where μ and σ^2 are unknown. The sample mean $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and the sample variance $S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$ are statistics, *while* $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ *is not a statistic.*

Definition 58 (Estimators and Estimates)

A statistic $T = T(X) = T(X_1, ..., X_n)$ *that is used to estimate* $\tau(\theta)$*, a function of* θ*, is called an* ***estimator*** *of* $\tau(\theta)$*, and an observed value of the statistic* $t = t(x) = t(x_1, ..., x_n)$ *is called an* ***estimate*** *of* $\tau(\theta)$*.*

Example 17.1.2

Suppose $X_1, ..., X_n$ *are IID rvs with* $E[X_i] = \mu$*. The rv* $\overline{X}$ *is an estimator of* μ*. For a given set of observations,* $x_1, ..., x_n$*, the number* $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *is an estimate of* μ*.*

There are certain "properties" of an estimator that we look for, in particular, for an estimator, $\tilde{\theta}$, of an unknown parameter θ, we want that the estimator

- is not biased, aka an **unbiased estimator**, i.e. $E[\tilde{\theta}]$;
- has small variance;
- is consistent, i.e. $\tilde{\theta} \xrightarrow{P} \theta$.

17.1.1 *Maximum Likelihood Estimation*

Suppose X is a discrete rv with pf $P(X = x; \theta) = f(x; \theta)$, $\theta \in \Omega$ where the scalar parameter θ is unknown. Suppose x is an observed value of the rv X. Then the probability of observing this value is

$$P(X = x; \theta) = f(x; \theta).$$

With the observed value of x substituted into $f(x; \theta)$, we have a function of the parameter θ, referred to as the **likelihood function**, and denoted $L(\theta)$. In the absense of any other information, it seems logical (temptingly so) that we should estimate the parameter θ using a value that is "the most compatible" with the data. E.g., we might choose the value of θ of which it maximizes the probability of the observed data, or equivalently, the value of θ which maximizes the likelihood function $L(\theta)$.

But first, let us formally state the definition of a likelihood function.

Definition 59 (Likelihood function)

*Suppose X is an rv with pf $f(x;\theta)$, where $\theta \in \Omega$ is a scalar. If x is the observed data, then the **likelihood function** for θ based on x is*

$$L(\theta) = P(X = x; \theta) = f(x;\theta) \quad \text{for } \theta \in \Omega.$$

*Similarly so, suppose $X_1, ..., X_n$ is a random sample from a distribution, each with pf $f(x;\theta)$, and let $x_1, ..., x_n$ be the observed data. Then the **likelihood function** for θ based on $x_1, ..., x_n$ is*

$$L(\theta) = P(X_1 = x_1, ..., X_n = x_n; \theta) = \prod_{i=1}^{n} f(x_i;\theta) \quad \text{for } \theta \in \Omega.$$

Note

As discussed earlier, a natural estimate of θ is the value which maximizes the probability of the observed sample, and we denote this notion as

$$\hat{\theta} = \arg\max_{\theta \in \Omega} L(\theta)$$

*and call $\hat{\theta}$ the **maximum likelihood estimate** (ML estimate). In practice, it is often convenient to work, instead, with the natural logarithm of the likelihood function, in which we call the **Log-likelihood**:*

$$\ell(\theta) = \ln L(\theta).$$

Note that the maximum likelihood estimate for both the likelihood function and the log-likelihood are the same since log *is a strictly increasing function.*

Example 17.1.3 (Example 6.1)

Consider flipping a coin repeatedly, where for $i \in \mathbb{N}$

$$X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ flip lands on heads} \\ 0 & \text{otherwise} \end{cases}$$

Based on 4 independent flips, we are given that $X_1, X_2, X_3, X_4 \sim \mathrm{Bernoulli}(p)$ are IID rvs. The sample has been observed as $x_1 = 1$, $x_2 = 1$ $x_3 = 0$, $x_4 = 1$. Write the probability of this sample as a function of p. Then, compute the likelihood function and find the ML estimate.

Solution

Since the X_i's are IID rvs, we have

$$\begin{aligned} f(x_1, x_2, x_3, x_4; p) &= \prod_{i=1}^{4} P(X_i = x_i; p) \\ &= \prod_{i=1}^{4} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{4} x_i}(1-p)^{\sum_{i=1}^{4}(1-x_i)}. \end{aligned}$$

Note that the likelihood function for p under the observed values is therefore

$$L(p) = p^3(1-p).$$

To find the ML estimate, we do

$$0 = \frac{dL(p)}{dp} = 3p^2 - 4p^3 = p^2(3-4p) \tag{17.1}$$

and we observe that $p = 0$ or $\frac{3}{4}$. Clearly, $p = \frac{3}{4}$ maximizes the likelihood function.

Note

We have been *talking almost solely about the discrete case, so what about the continuous case? We have that $P(X = x) = 0$, and so we consider a small neighbourhood of radius $\delta > 0$. Then for a small $\delta > 0$ around any point $x \in \mathrm{supp}(X)$, we have that*

$$P(x - \delta < X < x + \delta; \theta) = \int_{x-\delta}^{x+\delta} f(t;\theta)\, dt \approx 2\delta \cdot f(x;\theta). \tag{17.2}$$

And so for an observed value x, since δ is fixed in Equation (17.2), we have that the value of θ that maximizes $f(t;\theta)$ also maximizes $2\delta \cdot f(t;\theta)$.

Warning

We shall clarify the following notations: we use

- $\tilde{\theta}$ *as the estimator, which is an rv; and*
- $\hat{\theta}$ *as an estimate, which is a fixed value.*

Example 17.1.4 (Example 6.2 (Course Notes 6.2.4))

Recall the coin flip example in Example 17.1.3. Suppose $X_1, ..., X_n \sim \mathrm{Bernoulli}(p)$ *is a sequence of IID rvs. Calculate the ML estimate of* p.

Solution

Since $X_1, ..., X_n$ *are IID, let* $\vec{x} = (x_1 \; x_2 \; \dots \; x_n)$, *then the joint distribution is*

$$\begin{aligned} f(\vec{x}; p) &= \prod_{i=1}^n f(x_i; p) \\ &= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n (1-x_i)} \end{aligned}$$

To get an ML estimator, we shall use the log-likelihood:

$$\ell(p) = \left(\sum_{i=1}^n x_i \right) \ln p + \left(n - \sum_{i=1}^n x_i \right) \ln(1-p)$$

and so to get the ML estimator

$$\begin{aligned} 0 = \frac{\partial}{\partial p} \ell(p) \bigg|_{p=\hat{p}} &= \left(\sum_{i=1}^n x_i \right) \left(\frac{1}{\hat{p}} \right) - \left(n - \sum_{i=1}^n x_i \right) \left(\frac{1}{1-\hat{p}} \right) \\ &= \frac{\Sigma x_i}{\hat{p}} - \frac{n - \Sigma x_i}{1-\hat{p}} = \frac{(1-\hat{p})\Sigma x_i - n\hat{p} + \hat{p}\Sigma x_i}{\hat{p}(1-\hat{p})} \\ &= \frac{\Sigma x_i - n\hat{p}}{\hat{p}(1-\hat{p})} \end{aligned}$$

Note that when looking for an **ML estimate**, we should also check for the case when $\frac{\partial^2}{\partial p^2}\ell(p)\Big|_{p=\hat{p}} < 0$ to ensure maximality (instead of minimality).

where we represent $\sum_{i=1}^n x_i$ *by* Σx_i *for sanity. Thus we have that*

$$n\hat{p} = \sum_{i=1}^n x_i \implies \hat{p} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

Note that the **ML estimator** in this case would be represented as

$$\tilde{p} = \bar{X}.$$

Example 17.1.5 (Example 6.3 (Course Notes 6.2.5))

Suppose we have the data $\vec{x} = (x_1 \; x_2 \; \dots \; x_n)$ *for the sequence of IID rvs* $X_1, X_2, ..., X_n \sim \mathrm{Poi}(\theta)$. *Find the likelihood function, log-likelihood, the ML* *estimate $\hat{\theta}$, and the ML estimator $\tilde{\theta}$.*

Solution

Since each of the X_i's are IID, we have that their joint pmf, and in particular, their likelihood function, is

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

Then the log-likelihood is

$$\ell(\theta) = -n\theta + \left(\sum_{i=1}^{n} x_i \right) \log\theta - \sum_{i=1}^{n} \log(x_i!).$$

To get the ML estimate, note that

$$0 = \frac{\partial}{\partial\theta}\ell(\theta)\Big|_{\theta=\hat{\theta}} = -n + \frac{1}{\hat{\theta}} \sum_{i=1}^{n} x_i$$

and so

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}.$$

Consequently, we have that the ML estimator is

$$\tilde{\theta} = \overline{X}.$$

18 Lecture 18 Jul 3rd 2018

18.1 Estimation (Continued)

18.1.1 Maximum Likelihood Estimation (Continued)

Definition 60 (Score Function)

*The **score function** is defined as*

$$S(\theta) = S(\theta; x) = \frac{d}{d\theta}\ell(\theta) = \frac{d}{d\theta}\ln L(\theta) \quad \theta \in \Omega.$$

Note

Notice that to find the ML estimate, we usually set the score function to zero and solve for $S(\theta) = 0$.

Definition 61 (Information Function)

*The **information function** is defined as*

$$I(\theta) = I(\theta; x) = -\frac{d^2}{d\theta^2}\ell(\theta) = -\frac{d^2}{d\theta^2}\ln L(\theta) \quad \theta \in \Omega$$

*If $\hat{\theta}$ is the ML estimate of θ, then $I(\hat{\theta})$ is called the **observed information**.*

Example 18.1.1

Going back to our example in Example 17.1.4, we had that the likelihood function was

$$L(p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}$$

the log-likelihood was

$$\ell(p) = \left(\sum_{i=1}^{n} x_i \right) \ln p + \left(n - \sum_{i=1}^{n} x_i \right) \ln(1-p).$$

So the score function is

$$S(p) = \frac{d}{dp} \ell(p) = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \left(n - \sum_{i=1}^{n} x_i \right)$$

Consequently, the information function is

$$\begin{aligned} I(p) &= -\frac{d^2}{dp^2} \ell(p) = -\frac{d}{dp} S(p) \\ &= \frac{1}{p^2} \sum_{i=1}^{n} x_i + \frac{1}{(1-p)^2} \left(n - \sum_{i=1}^{n} x_i \right) \end{aligned}$$

Note

Recall from the calculus knowledge of the 2nd derivative being the "curvature" of the original function, and so $I(\theta)$ is a function about the curvature of the log-likelihood. In particular, we have that $I(\hat{\theta})$ tells us about the convacity of the log-likelihood function at the ML estimate.

Note that the information function is a function that has two variables: the unknown parameter θ, and the data $\vec{X} = (X_1 \dots X_n)$.

In a later lecture, we shall see how the observed information $I(\hat{\theta})$ can be used to construct approximate confidence intervals for the unknown parameter θ.

Definition 62 (Fisher Information)

*If θ is a scalar, then the **expected information**, or **Fisher information** (function) is given by*

$$J(\theta) = E[I(\theta; \vec{X})] = E\left[-\frac{\partial^2}{\partial \theta^2} \ell(\theta; \vec{X}) \right] \quad \theta \in \Omega$$

Note

Just to take away the layers of definitions and compare the Fisher information with our pf for the rv(s), if $X_1, ..., X_n$ is a random sample (i.e. IID rvs), each with pf $f(x;\theta)$, then

$$J(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2}\ell(\theta;\vec{X})\right] = nE\left[-\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\right]$$

where $\vec{X} = (X_1 \ldots X_n)$.

Example 18.1.2 (Example 6.4 (Course Notes 6.2.10))

Suppose $X_1, ..., X_n \sim \text{Bernoulli}(p)$ is a sequence of IID rvs. We have showed in Example 17.1.4 that the ML estimator of p is $\tilde{p} = \overline{X}$. Calculate the Fisher information and compare it with the variance of the ML estimator of p.

Solution

Recall that from Example 18.1.1, we had

$$\begin{aligned}
\ell(p) &= \left(\sum_{i=1}^n x_i\right)\ln p + \left(n - \sum_{i=1}^n x_i\right)\ln(1-p) \\
S(p) &= \frac{1}{p}\sum_{i=1}^n x_i - \frac{1}{1-p}\left(n - \sum_{i=1}^n x_i\right) \\
I(p) &= \frac{1}{p^2}\sum_{i=1}^n x_i + \frac{1}{(1-p)^2}\left(n - \sum_{i=1}^n x_i\right),
\end{aligned}$$

So the Fisher information is

$$\begin{aligned}
E[I(p)] &= E\left[\frac{1}{p^2}\sum_{i=1}^n X_i + \frac{1}{(1-p)^2}\left(n - \sum_{i=1}^n X_i\right)\right] \\
&= \frac{1}{p^2}\sum_{i=1}^n E[X_i] + \frac{1}{(1-p)^2}\left(n - \sum_{i=1}^n E[X_i]\right) \\
&= \frac{1}{p^2}(np) + \frac{1}{(1-p)^2}(n-np) \\
&= \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)}.
\end{aligned}$$

On the other hand, note that the variance of the ML estimator is

$$\text{Var}(\tilde{p}) = \text{Var}(\overline{X}) = \frac{1}{n}\text{Var}(X_i) = \frac{p(1-p)}{n}.$$

Example 18.1.3 (Example 6.5 (Course Notes 6.2.10))

Suppose $X_1, ..., X_n \sim \mathrm{Poi}(\theta)$ is a sequence of IID rvs. We showed in Example 17.1.5 the ML estimator of θ is $\tilde{\theta} = \overline{X}$. Calculate the Fisher information and compare it with the variance of the ML estimator of θ.

Solution

We had the that log-likelihood is

$$\ell(\theta) = -n\theta + \left(\sum_{i=1}^{n} x_i\right) \log\theta - \sum_{i=1}^{n} \log(x_i!).$$

So the score function is

$$S(\theta) = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i,$$

and so the information function is

$$I(\theta; x) = \frac{1}{\theta^2}\sum_{i=1}^{n} x_i.$$

Then the Fisher information is

$$E[I(\theta; \vec{X})] = \frac{n}{\theta^2} E[X_i] = \frac{n}{\theta}.$$

The variance of the ML estimator of θ is

$$\mathrm{Var}(\tilde{\theta}) = \frac{1}{n}\mathrm{Var}(X_i) = \frac{\theta}{n}.$$

Note

We observe from the above two examples that

$$J(\theta) = \frac{1}{\mathrm{Var}\,\tilde{\theta}}$$

where $\tilde{\theta} = \overline{X}$. It is tempting to verify if this is the case in general, but there does not seem to be reliable sources that points to this case (at least, online searches have yet to yield me results). The lecture claims that this is only true for certain families of distributions.

Example 18.1.4 (Example 6.6 (Course Notes 6.2.12))

Suppose $X_1, ..., X_n$ is a random sample, each from a distribution with pdf

$$f(x;\theta) = \theta x^{\theta - 1} \quad 0 \leq x \leq 1, \ \theta > 0.$$

Find the score function, the ML estimator of θ, the information function and the observed information.

Solution

Since $X_1, ..., X_n$ are IID, we have

$$L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta - 1}.$$

So the log-likelihood is

$$\ell(\theta) = n \ln \theta + (\theta - 1) \ln \left(\prod_{i=1}^{n} x_i \right) = n \ln \theta + (\theta - 1) \sum_{i=1}^{n} \ln x_i.$$

The score function is therefore

$$S(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i,$$

and so the ML estimator of θ is

$$\tilde{\theta} = -\frac{n}{\sum\limits_{i=1}^{n} \ln x_i}.$$

From the score function, the information function is

$$I(\theta) = \frac{n}{\theta^2},$$

and so the observed information is

$$I(\hat{\theta}) = \frac{n}{\left(-\frac{n}{\sum_{i=1}^{n} \ln x_i} \right)^2} = \frac{1}{n} \left(\sum_{i=1}^{n} \ln x_i \right)^2$$

Note

In the above example, note that the Fisher information is

$$J(\theta) = E[I(\theta; \vec{X})] = \frac{n}{\theta^2}.$$

Example 18.1.5 (Example 6.7 (Course Notes 6.2.13))

Suppose $X_1, ..., X_n \sim \mathrm{Unif}(0, \theta)$ *is a random sample. Find the ML estimator of* θ*.*

✎ Solution

Note that $f_{X_i}(x_i; \theta) = \frac{1}{\theta}$ *for* $0 \leq x_i \leq \theta$*. So the likelihood function is*

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{\{0 \leq x_i \leq \theta\}} \\ &= \theta^{-n} \mathbb{1}_{\{0 \leq x_1 \leq \theta, 0 \leq x_2 \leq \theta, ..., 0 \leq x_n \leq \theta\}} \\ &= \theta^{-n} \mathbb{1}_{\{0 \leq x_{(n)} \leq \theta\}} \end{aligned}$$

where we use the order statistics notation to simplify the equation. Notice that if we take the derivative of the likelihood function from this point and try to get the ML estimate/estimator, we would run into the following equation:

$$-\frac{n}{\theta^{n+1}} = 0$$

in which we would have trouble getting the MLE.

Example 18.1.6 (Example 6.8)

Suppose $X_1, ..., X_n$ *is a random sample, each from the* $\mathrm{Unif}(\theta, \theta + 1)$ *distribution. Find the ML estimator of* θ*.*

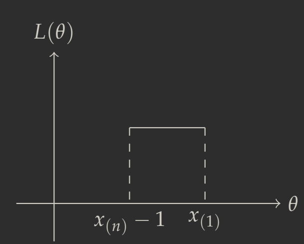
✎ Solution

This time, we have that the pdf for each X_i *is*

$$f_{X_i}(x_i; \theta) = \mathbb{1}_{\{\theta \leq x_i \leq \theta + 1\}}.$$

Then the likelihood function is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x_i; \theta) = \mathbb{1}_{\{\theta \leq x_1 \leq \theta + 1, ..., \leq x_n \leq \theta + 1\}} \\ &= \mathbb{1}_{\{\theta \leq x_{(1)}\}} \mathbb{1}_{\{x_{(n)} \leq \theta + 1\}} \\ &= \mathbb{1}_{\{\theta \leq x_{(1)}\}} \mathbb{1}_{\{x_{(n)} - 1 \leq \theta\}} = \mathbb{1}_{\{x_{(n)} - 1 \leq \theta \leq x_{(1)}\}} \end{aligned}$$



where we, once again, use the order statistics notation. Consequently, we have the graph of θ *versus* Ł(θ) *on the right.*

Thus we observe that the MLE is not unique, since it can be any number between $x_{(n)} - 1$ *and* $x_{(1)}$*.*

Theorem 61 (Invariance Property of the MLE)

Suppose $\tau = h(\theta)$ is an injective function of θ. Suppose also that $\hat{\theta}$ is the ML estimator of θ. Then $\hat{\tau} = h(\theta)$ is the ML estimator of τ.

Warning

We are short on certain tools to actually prove this theorem.

Example 18.1.7 (Example 6.9)

Suppose $X_1, ..., X_n \sim f(x;\theta) = \theta x^{\theta-1}\mathbb{1}_{\{0<x<1\}}$, where $\theta > 0$. Find the MLE of the median of the distribution.

Solution

Recall from Example 18.1.4 that we had

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \ln x_i}.$$

Let m be the median. The goal is to use Theorem 61. ***Something feels off...***

19 Appendix

19.1 Commonly Used Distributions

Distribution	pf	Mean	Var	mgf
Binomial Distribution : $X \sim \text{Bin}(n, p)$				
$x \in \mathbb{N} \cup \{0\}$ $n \in \mathbb{N}$ $0 < p < 1$	$\binom{n}{x} p^x (1-p)^{n-x}$	np	$np(1-p)$	$(pe^t + 1 - p)^n$
Geometric Distribution : $X \sim \text{Geo}(p)$				
$x \in \mathbb{N} \cup \{0\}$ $0 < p < 1$	$p(1-p)^x$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$	$\frac{p}{1-(1-p)e^t}$ $t < -\log(1-p)$
Poisson Distribution : $X \sim \text{Poi}(\mu)$				
$x \in \mathbb{N} \cup \{0\}$ $\mu > 0$	$\frac{e^{-\mu}\mu^x}{x!}$	μ	μ	$e^{\mu(e^t-1)}$
Uniform Distribution : $X \sim \text{Unif}(a, b)$				
$a \leq x \leq b$ $a < b \in \mathbb{R}$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)^t}$ $t \neq 0$
Normal Distribution : $X \sim \text{N}(\mu, \sigma)$				
$x \in \mathbb{R}$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	$e^{\mu t + \frac{\sigma^2 t^2}{2}}$

$\mu \in \mathbb{R}$ $\sigma > 0$				
Gamma Distribution $: X \sim \Gamma(\alpha, \beta)$				
$x \in \mathbb{R}_{\geq 0}$ $\alpha > 0$ $\beta > 0$	$\frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$	$\alpha\beta$	$\alpha\beta^2$	$(1-\beta t)^{-\alpha}$ $t < \frac{1}{\beta}$
Exponential Distribution $: X \sim \text{Exp}(\theta)$				
$x \in \mathbb{R}_{>0}$ $\theta > 0$	$\frac{1}{\theta} e^{-\frac{x}{\theta}}$	θ	θ^2	$(1-\theta t)^{-1}$ $t < \frac{1}{\theta}$
Negative Binomial Distribution $: X \sim \text{NB}(r, p)$				
$x \in \mathbb{N} \cup \{0\}$ $r > 0$ $p \in (0, 1)$	$\binom{x+r-1}{x} \cdot (1-p)^r p^x$	$\frac{pr}{1-p}$	$\frac{pr}{(1-p)^2}$	$\left(\frac{1-p}{1-pe^t}\right)^r$ $t < -\log p$
Beta Distribution $: X \sim \text{Beta}(\alpha, \beta)$				
$x \in [0, 1] or$ $x \in (0, 1)$ $\alpha > 0$ $\beta > 0$	$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(a+b)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r}\right) \frac{t^k}{k!}$
Chi-Squared Distribution $: X \sim \chi^2(k)$				
$k \in \mathbb{N}$ if k = 1 $x \in (0, \infty)$ otherwise $x \in [0, \infty]$	$\frac{1}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$	k	2k	$(1-2t)^{-\frac{k}{2}}$ $t < \frac{1}{2}$

20 Index