STAT330S18 - Mathematical Statistics

Johnson Ng

May 20, 2018

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. 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.

1 Lecture 1 May 1st 2018

1.1 Introduction

Definition 1.1.1 (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 1.1.2 (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 1.1.3 (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 1.1.4 (Probability Measure)
*Suppose S is a sample space of a random experiment. Let $\mathscr{B} = \{A_1, A_2, ...\} \subseteq \mathbb{P}(S)$ 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 1.1.4 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 1.1.1 (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 1.1.4 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 1.1.4,*

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

[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 1.1.4, we would then have*

[7] Again, this is not hard to show

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

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 1.1.4, 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 1.1.4, $P(B \cap A^C) \geq 0$. *Hence from Equation (†), we have that*

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

as required. □

Definition 1.1.5 (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*

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

Definition 1.1.6 (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 1.1.2 (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 1.1.3 (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 1.1.4 (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 1.2.1 (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*

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

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 1.2.2 (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 1.2.1 (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 1.3.1 (Discrete Random Variable)

*An rv X is a **discrete random variable** when its image is finite or countably infinite, i.e. $X \in \{x_1, x_2, ...\}$. 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 1.3.1 (Properties of pmf)

With the notation from Definition 1.3.1, 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,

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

□

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

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

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

Solution

We have that

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

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**.

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

Exercise 1.3.2

Prove that the pdf of $X \sim (\mu)$ *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

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

show that

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

Solution

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

2 Lecture 2 May 03rd 2018

2.1 Continuous Random Variable

Definition 2.1.1 (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 2.1.1 (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 1.3.1.
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.

□

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

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

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 2.2.1 (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 $X \sim \text{Bin}(n, p)$ and its pmf is*

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

2.2.2 Geometric Distribution

Definition 2.2.2 (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 $X \sim \text{Geo}(p)$, and its pmf is*

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

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 2.2.3 (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 $X \sim \text{Poi}(\lambda)$, with $\lambda > 0$, and the pmf is*

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

2.3 Examples of Continuous RVs

2.3.1 Normal/Gaussian Distribution

Definition 2.3.1 (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 $X \sim \text{N}(\mu, \sigma^2)$, and its pdf is*

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

Definition 2.3.2 (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 $Z \sim \text{N}(0, 1)$ and its pdf is*

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

2.3.2 Uniform Distribution

Definition 2.3.3 (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 $X \sim \text{Unif}(a, b)$, and its pdf is*

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

2.3.3 Exponential Distribution

Definition 2.3.4 (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 $X \sim \text{Exp}(\theta)$, 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 2.3.5 (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:

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

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

Solution
Note that since

$$\text{Dom}\, X = \{0, 1, 2, 3, 4, ...\},$$

we have that

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

With that, note that

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

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

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

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*

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

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 $X \sim \text{Exp}(\frac{1}{2})$. So it is clear that X is a crv and since $Y = 1, 2,$ or 3, we have that Y is discrete. Now

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

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.

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

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:

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

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

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

2.4.4 A Formula for the Continuous Case

Theorem 2.4.1 (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

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

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. □

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

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

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 3.2.1 (Probability Integral Transformation)
*If X is a contiunous rv with cdf F, then $Y = F(X) \sim \text{Unif}(0,1)$. $Y = F(X)$ 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 $\text{Unif}(0,1)$ rv, i.e. the **standard uniform random variable**. Thus $Y \sim \text{Unif}(0,1)$.*

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

Note that the converse of Probability Integral Transformation is true:

Theorem 3.2.2 (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

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

□

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 3.2.2, 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 3.3.1 (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 3.3.2 (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 3.3.3 (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 $X \sim G(\mu,\sigma)$. Show that F_X belongs to a location-scale family of

distributions.

We know that if $\mu = 0$ and $\sigma = 1$, then $Y = \frac{X-\mu}{\sigma} \sim G(0,1)$, and we know that $G(0,1)$ 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 $X \in G(\mu, 2)$ 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 $G(0,2)$. 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

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

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

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

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

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

Therefore, F_X belongs to a scale family of distributions.

3.4 Expectations

3.4.1 Expectations

Definition 3.4.1 (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 3.4.2 (Expectation of A Continuous RV)
*If X is a continuous rv with pdf f and support set A, then the **expecta-***

tion *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 $X \sim \text{Poi}(\lambda)$. Calculate $E(X)$.

Solution
Note

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

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 4.1.1 (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,

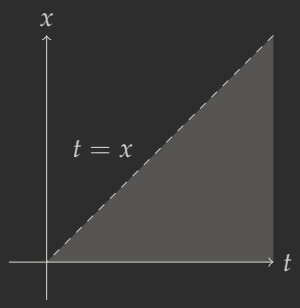
$$\int_0^\infty [1 - F(x)]\, dx = \int_0^\infty \int_x^\infty f(t)\, dt\, dx.$$

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.

□

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 $X \sim \text{Exp}(\theta)$. Use Theorem 4.1.1 to calculate $E(X)$.

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

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

Then

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

□

Theorem 4.1.2 (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 $X \sim \text{Unif}(0, \theta)$. 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 4.1.3 (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

$$\begin{aligned}
E\left[\sum_{i=1}^{n}(a_i g_i(X)+b_i)\right] &= \sum_{x\in A}\left[\sum_{i=1}^{n}(a_i g_i(x)+b_i)\right]f(x) \quad \because \textit{Theorem } 4.1.2 \\
&= \sum_{x\in A}\sum_{i=1}^{n}[a_i g_i(x)f(x)+b_i f(x)] \\
&= \sum_{i=1}^{n}\sum_{x\in A}[a_i g_i(x)f(x)+b_i f(x)] \qquad (*) \\
&= \sum_{i=1}^{n}\left[a_i\sum_{x\in A}g_i(x)f(x)+b_i\sum_{x\in A}f(x)\right] \\
&= \sum_{i=1}^{n}(a_i E[g_i(X)]+b_i)
\end{aligned}$$

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 4.1.1 (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 4.1.2 (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 4.1.4 (Variance of a Linear Function)
Suppose X is an rv with pf f and $a, b \in \mathbb{R}$. Then

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

Proof
Observe that

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

□

Example 4.1.3 (Example 2.22 (course notes - 2.6.10 (1)))
If $X \sim \text{Poi}(\theta)$, then $E[X^{(k)}] = \theta^k$ for $k = 1, 2, \dots$.

Solution

Note

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

So

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

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

Note that it is not necessarily true that

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

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 4.2.1 (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

$A = \{w \in S : X(w) \geq a\}$.

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

□

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

Theorem 4.2.2 (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

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

□

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 4.2.3 (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 4.2.2,

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

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 4.2.1, 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 4.2.3)?*

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

$$\begin{aligned}
&P(8000 < X < 12000)\\
&= P(-2000 < X - 10000 < 2000)\\
&= P(|X - 10000| < 2000) = 1 - P(|X - 10000| \geq 2000)\\
&\geq 1 - \frac{1}{(\sqrt{2000})^2} \quad \because \textit{Theorem 4.2.3} \wedge k = \frac{2000}{\sigma} = \sqrt{2000}\\
&= \frac{1999}{2000}
\end{aligned}$$

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 5.2.1 (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 5.2.2 (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 $X \sim \text{Poi}(\lambda)$. 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 5.2.1 (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*

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

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

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****.*

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

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

□

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.

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

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 5.2.1 (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 $X \sim (\theta_1, \theta_2)$. 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 5.2.1,

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

5.2.2 *Uniqueness of the MGF*

Theorem 5.2.2 (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 $X \sim (0,1)$. 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 6.1.1 (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 6.1.1 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 6.1.1 (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 1.1.1, 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 6.1.1.[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):

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

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

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

The joint cdf of (X_1, X_2) is the given F above. Verify that under this experi-

ment, F is indeed a cdf.

Solution

Note that conditions 3 and 4 of Proposition 6.1.1 are automatically satisfied by the definition of F.

incomplete example

Definition 6.1.2 (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 6.1.3 (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 6.1.2 (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. $\sum\limits_{(x,y)\in A} \sum f(x,y) = 1$
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 1.3.1

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 6.1.4 (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*

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

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

Definition 6.1.5 (Independence)
*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 6.1.1 (Independence by PF)
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, suppoes 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

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

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.

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

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, "indepdent" of each other. Indeed, since

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

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 Index