Class Notes for Applied Probability and Statistics

Mark Siggers

ver .2022/06/10

These notes are for a upper undergraduate or graduate one semester course on Probability and Statistics, and are largely based on Hogg, McKean, and Craig's 'Introduction to Mathematical Statistics' (International Seventh edition) which we refer to as [1], or as 'the text'. They cover much of the material in the first four chapters of the text.

Section numbering follows the text, and problem numbers often refer to the text. Problems within the notes are usually quite easy, checking that we know definitions, and making simple observations that we will use later. It is important to look also at the problems from the text.

1 Probability and Distributions

1.1 Introduction

Probabililty theory in its pure form has much of the flavour of real analysis. In this course on applied probability theory, we try to avoid this formality not by sacrificing rigour, but by avoiding exceptional cases. We generally assume things to be nice. Our goal is to get an introduction to how probablility can be applied, both in statistics and in mathematics.

Probability theory for statistics is concerned with *random experiments* - experiments that can be repeated several times, under the same conditions, and have different outcomes. They are characterised by the fact that we cannot predict the outcome of an individual experiment but can predict the frequency of the outcome over many repetitions of the experiment.

For example, an experiment might consist of tossing a coin. We cannot predict whether the outcome will be heads or tails, but if we toss the same coin 100 times, we are all going to guess that the outcome will be heads 50 times.

Would we bet on it though? Would you take the following bet? You pay ₩1000 and toss a coin 100 times. If the outcome is heads exactly 50 times, you win ₩2000.

Probably not. Would you take the bet if you win in the case that the outcome is heads between 40 and 60 times? This is the kind of question that probability theory lets us address.

In statistics we will not be really be interested in probability that a coin comes up heads, but perhaps we will be interested in the probability that a given person in a population tests positive for some disease. Our goal will be to look at the data of an experiment, estimate such a parameter, and then give some measure of how good our estimate is.

On the other hand, the application of probability to mathematics is usually a way of counting structures, and through this, determining properties that *most* of the structures have, or showing that a structure with a given property must exist.

For example, by constructing a graph by randomly adding an edge between any two vertices with some given probability, and then calculating the probabilities that the graph has small cycles or large independent sets, we show the existence of graphs of large girth and large chromatic number.

The obvious commonality in these applications is the notion of something happening randomly. This brings us to *random variables*, which are the starting point of our course.

1.2 Set Theory

We will generally consider sets of points in $\mathbb{R}$ or $\mathbb{R}^n$, such as

$$C = \{(x, y) | x \in \mathbb{R}, y = 2x\}.$$

The union and intersection of sets are denoted standardly by such notation as $C_1 \cup C_2$, $\cup_{i=1}^{k} C_i$, $\cup_{i=1}^{\infty} C_i$, $C_1 \cap C_2$, $\cap_{i=1}^{k} C_i$, and $\cap_{i=1}^{\infty} C_i$.

The empty set, or null set is often (but not always) denoted in [1] by ϕ, but I will use the more standard $\emptyset$.

Subsets are denoted $C_1 \subset C_2$ and may be equal. A set C is usually assumed to be a subset of and underlying *universe* $\mathscr{C}$, and the complement of a set C is defined as

$$C^c = \mathscr{C} - C = \{x \in \mathscr{C} \mid x \notin C\}.$$

(The notation $\overline{X}$ will play a different role.)

Recall the following basic set laws.

i) $C \cup C^c = \mathscr{C}$.

ii) $C \cap C^c = \emptyset$.

iii) $C \cup \mathscr{C} = \mathscr{C}$.

iv) $C \cap \mathscr{C} = C$.

v) $(C_1 \cup C_2)^c = C_1^c \cap C_2^c$ (Demorgan).

vi) $(C_1 \cap C_2)^c = C_1^c \cup C_2^c$.

Definition 1.2.1. The powerset $\mathscr{P}(\mathscr{C})$ of a set $\mathscr{C}$ is the set of all sets in $\mathscr{C}$. A *set function* on $\mathscr{C}$ is a function from $\mathcal{P}(\mathscr{C})$ to the set $\mathbb{R}^{\text{ex}} := \mathbb{R} \cup \{\pm\infty\}$ of extended reals.

Note

In a more thorough treatment of probability theory, we would define a set function as a function from a σ-algebra on $\mathscr{C}$ to the exended reals. This is a subset of the powerset $\mathscr{P}(\mathscr{C})$ that is closed uner complements, and countable intersections and unions.

Example 1.2.2. The cardinality map $|\cdot|$, acting like:

$$|\{1, 3, 5, 6, 7\}| = 5,$$

is a set function on any finite set $\mathscr{C}$. Its range is the natural numbers $\mathbb{N} \subset \mathbb{R}$.

Example 1.2.3. The area/volume function Vol is a set function on $\mathbb{R}^2$.

- $\text{Vol}(\{(x, y) \mid x, y \in [0, 1]\}) = 1$.
- $\text{Vol}(\{(x, y) \mid |(x, y) - 0| \leq 1\}) = 2\pi$.
- $\text{Vol}(\{(x, y) \mid y = 2x\}) = 0$.
- $\text{Vol}(\{(x, y) \mid x \geq 0\}) = \infty$.

The area function is really just the integral $\text{Vol}(S) = \oiint_S 1 \, \mathrm{d}\vec{x}$. More generally for any function $f : \mathbb{R}^n \to \mathbb{R}$ we can define a set function Vol_f on $\mathbb{R}^n$ by

$$\text{Vol}_f(C) = \oiint_C f(\vec{x}) \, \mathrm{d}\vec{x}.$$

Example 1.2.4. Let $\text{Vol} = \text{Vol}_{e^{-x}}$ then for $C = [0, \infty) \in \mathbb{R}$, we have:

$$\begin{aligned} \text{Vol}(C) &= \int_0^\infty e^{-x} \, \mathrm{d}x = \lim_{N \to \infty} \int_0^N e^{-x} \, \mathrm{d}x \\ &= \lim_{N \to \infty} (-e^{-x}) \, |_{x=0}^N \\ &= \lim_{N \to \infty} (-e^{-N} + e^0) = 0 + 1 = 1 \end{aligned}$$

The *support* $\mathrm{Supp}(f)$ of a function $f : \mathscr{C} \to \mathbb{R}$ is the subset of $\mathscr{C}$ on which f is non-zero. Often we artificially extend a function to a more convenient universe by defining it to be 0 where it wasn't defined. To refer back to properties of the original function, we then talk of its support. For example, a function defined on the integers can be considered a real function, but with countable support.

Now, it is clear that for any $f : \mathbb{R}^n \to \mathbb{R}$ with finite or countable support, $\mathrm{Vol}_f = 0$. In such cases, we will consider a discrete analogue $\mathrm{Sum}_f(C) = \sum_{x \in C} f(x)$.

Example 1.2.5. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = (1/2)^x$ for $x \in \mathbb{Z}^+$ (positive integers) and $f(x) = 0$ otherwise. Then

$$\mathrm{Sum}_f(\{x \in \mathbb{N} \mid x < 3\}) = 1/2 + 1/4 = 3/4$$

and

$$\mathrm{Sum}_f(\mathbb{R}) = Q_f(\mathbb{Z}^+) = 1/2 + 1/4 + \cdots = 1.$$

This is our first introduction to what will become a consistant theme in the course: concepts and definitions will frequently have *continuous* and *discrete* versions.

Problems from the Text

Section 1.2: 1,5,6,8,11,14,16

1.3 The Probability Set Function

Definition 1.3.1. A *probablility set function* or *distribution* on a set $\mathscr{C}$ is a set function P of $\mathscr{C}$ such that

i) $P(C) \geq 0$ for all $C \subset \mathscr{C}$.

ii) $P(\mathscr{C}) = 1$.

iii) P is countably additive: for a family $\{C_n\}_{n \in \mathbb{N}}$ of pairwise disjoint sets $C_n \subset \mathscr{C}$,

$$P(\cup_{n=1}^{\infty} C_n) = \sum_{n=1}^{\infty} P(C_n).$$

We are now ready to define the basic setup that we will assume throughout the course.

Definition 1.3.2. A *(random) experiment* or a *probability space* consists of a set $\mathscr{C}$, and a probability set function P of $\mathscr{C}$. We call $\mathscr{C}$ the *sample space*, and subsets of $\mathscr{C}$ *events*. Elements $x \in \mathscr{C}$ are called *outcomes*, or sometimes, viewed as singleton sets in $\mathscr{C}$, *elementary events*.

Often a probability function, expecially for finite (discrete) $\mathscr{C}$, is defined additively by its value on elementary events.

Example 1.3.3. Tossing a coin is a random experiment with two outcomes: $\mathscr{C} = \{H, T\}$. Setting $P(\{H\}) = 1/2$ defines P as a probablility function on $\mathscr{C}$ via the axioms of a probablitiy set function:

$$P(\{T\}) = P(\mathscr{C} - \{H\}) = P(\mathscr{C}) - P(\{H\}) = 1 - 1/2 = 1/2.$$

For elementary events like $\{H\}$, we will write $P(H)$ for $P(\{H\})$.

Example 1.3.4. Tossing two identical coins is a random experiment with three possible outcomes: $\mathscr{C} = \{HH, HT, TT\}$. The probability function is defined by $P(HH) = P(TT) = 1/4$ and $P(HT) = 1/2$. The event $C = \{HT, HH\}$ has probability $P(C) = 3/4$.

Given a random experiment, we often define events non-formally: the event $C = \{HT, HH\}$ can be described as the event that 'at least one head is tossed'. We would say 'The probability that at least on head is tossed is 3/4.' and write $P($ at least on head is tossed $) = 3/4$.

Problem 1.3.5. Tossing two non-identical coins is an experiment with four possible outcomes: $\mathscr{C} = \{HH, HT, TH, TT\}$. What is the probability that at least one head is tossed?

Problem 1.3.6. Let $\mathscr{C}$ be the set of 36 possible outcomes

$$\{(i, j) \mid i, j \in [6]\}$$

when two different dice are rolled. What is the probablility of the following events (assuming that each outcome is equally likely)?

i) $i + j = 7$

ii) $i + j$ is even

iii) $i > j$.

Example 1.3.7. A p-coin is a coin that when tossed, shows heads with probabilty p and shows tails with probability $1 - p$. Tossing a p-coin is a random experiment with $\mathscr{C} = \{H, T\}$ such that $P(H) = p$ and $P(T) = 1 - p$. (A $\frac{1}{2}$-coin is called a fair coin.)

Unless otherwise stated, when we talk about events, it is always assumed that they are events of a sample space $\mathscr{C}$ with a probabilitiy set function P.

Theorem 1.3.8. *For events C, C_1 and C_2, the following hold.*

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

ii) $P(\emptyset) = 0$.

iii) $C_1 \subset C_2 \Rightarrow P(C_1) \leq P(C_2)$.

iv) $0 \leq P(C) \leq 1$.

v) $P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2)$.

Proof. All of these are pretty easy using the additivity of P. □

Now, the above theorem immediately yields the following which is known as **Bonferroni's Inequality**.

$$P(C_1 \cap C_2) \geq P(C_1) + P(C_2) - 1. \tag{1}$$

A sequence of events $\{C_n\}$ is *non-decreasing* if $C_n \subset C_{n+1}$ for each n. A sequence $\{D_n\}$ is *non-increasing* if $D_n \supset D_{n+1}$. In this case we often write $\lim_{n\to\infty} C_i$ for $\cup_{n=1}^{\infty} C_i$ and $\lim_{n\to\infty} D_i$ for $\cap_{n=1}^{\infty} D_i$.

Given a non-decreasing sequence $\{C_n\}$ of events, if we let $R_{n+1} = C_{n+1} - C_n$ for each n, then the events R_n are pairwise disjoint, and so by the additivity of P we have that

$$\begin{aligned} P(\lim_{n\to\infty} C_n) &= P(\cup_{n=1}^{\infty} C_n) = P(\cup_{n=1}^{\infty} R_n) = \sum_{n=1}^{\infty} P(R_n) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} P(R_i) = \lim_{n\to\infty} P(C_i). \end{aligned}$$

That is, we can interchange P and the limit. We have essentially shown the 'non-decreasing' part of the following.

Theorem 1.3.9. *Let $\{C_n\}$ be a non-decreasing or a non-increasing sequence of events. Then*

$$\lim_{n\to\infty} P(C_n) = P(\lim_{n\to\infty} C_n).$$

Problem 1.3.10. Prove the above theorem for a non-increasing sequence of events.

Using $C_n' = C_n - \cup_{i=1}^{n-1} C_i$ instead of R_n in the proof of the above theorem, we get the following.

Theorem 1.3.11 (Boole's Inequality)**.** *Let $\{C_n\}$ be a sequence of events. Then*

$$P(\cup_{n=1}^{\infty} C_n) \leq \sum_{n=1}^{\infty} P(C_n).$$

Example 1.3.12. In a experiment, you flip a coin until you get two consecutive heads or two consecutive tails. The sample space is

$$\mathscr{C} = \{HH, TT, HTT, THH, HTHH, THTT, HTHTT, THTHH, \dots\}.$$

What is the probabilitly that the experiment ends with an H?

Letting C_i be the event that we finish with two heads in at most i flips, we get that $P(C_2) = P(\{HH\}) = 1/4$, $P(C_3) = P(\{HH, THH\}) = 1/4+1/8$, and in general that $P(C_n) = 1/4+1/8+\dots 1/2^n$. By Theorem 1.3.9, the probability $C = \cup_{i=2}^{\infty} C_i$ that the experiment ends in two heads is

$$P(C) = P(\cup_{n=2}^{\infty} C_n) = \lim_{n\to\infty}(\sum_{i=2}^{n} 1/2^i) = \sum_{i=2}^{\infty} 1/2^i = 1/2.$$

There is another easy way to do the above example. Assuming that the experiment ends, it is easy to see, by symmetry, that it is equally likely to end with H or with T, so with probability $1/2$ it ends with H. This uses conditional probability, which we will see next section, but still it must be shown that the experiment ends. Or more precisely, it must be shown that the probability that the experiment ends is 1.

Problems from the Text

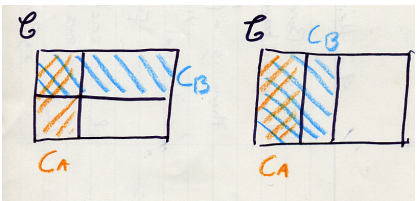
Section 1.3: 1,3,5,8,10,13,15,20

1.4 Conditional Probability

In an experiment with some event C_A of probability $1/3$, and another event C_B of probability $1/2$, does the knowledge that C_A occurs affect the probability that C_B occurs?

It can.

In the following picture let C_A be the event that a randomly placed dot in $\mathscr{C}$ is placed in the orange region and C_B be the event that a randomly placed dot in $\mathscr{C}$ is placed in the blue region.



In the first picture, knowing that event C_A happened doesn't affect the probability of event C_B. In the second picture, the fact that C_A has occured implies that event C_B **definitely** occurs.

In the first picture, the events C_A and C_B are *independent* and in the second picture they are not. Let's give this a mathematical definition.

1.4.1 Conditional probability and independence

Definition 1.4.1. For events C_1 and C_2, the *conditional probability of* C_1 *given* C_2 is

$$P(C_1 \mid C_2) = \frac{P(C_1 \cap C_2)}{P(C_2)}.$$

The events C_1 and C_2 are *independent* if $P(C_1 \mid C_2) = P(C_1)$.

Notice that if two events C_1 and C_2 are independent then we have that $P(C_1) = P(C_1 \mid C_2) = \frac{P(C_1 \cap C_2)}{P(C_2)}$ and so

$$P(C_1 \cap C_2) = P(C_1) \cdot P(C_2). \tag{2}$$

Indeed, this is an alternate definition of the independence of events, and because of this, independence is sometimes called multiplicity.

Example 1.4.2. In the experiment $\mathscr{C} = \{(i,j) \mid i,j \in [6]\}$ where we roll two independent dice. We define the events $C_1 : i \leq 3$, $C_2 : j \leq 3$, and $C_3 : i+j = 8$. Intuitively, we feel that the events C_1 and C_2 should be independent, while C_3 should depend on either of them. Indeed, we see, among other things that $P(C_1) = P(C_2) = 1/2$, $P(C_3) = 5/36$, $P(C_1 \cap C_2) = 9/36 = 1/4$, and

$$P(C_3 \cap C_1) = \frac{|\{(2,6),(3,5)\}|}{36} = 1/18.$$

This gives the conditional probabilities,

- $P(C_1 \mid C_2) = P(C_1 \cap C_2)/P(C_2) = (9/36)/(3/36) = 1/2 = P(C_1)$,
- $P(C_2 \mid C_1) = P(C_2 \cap C_1)/P(C_1) = (1/6)/(1/3) = 1/2 = P(C_2)$, and
- $P(C_3 \mid C_1) = P(C_3 \cap C_1)/P(C_1) = (2/36)/(3/6) = 1/9 \neq 5/36 = P(C_3)$.

We conclude that C_1 is independent of C_2 and C_2 is independent of C_1, but C_3 is not independent of C_1.

Notice that C_1 and C_2 were independent of each other. This should be expected, as it is clear from (2) that independence is a symmetric relationship. Here are some other obvious facts.

i) $P(C_1 \mid C_1) = 1$.

ii) $P(C_1 \mid C_2) = P(C_1 \cap C_2 \mid C_2)$.

iii) For fixed C_2 the function $P(\cdot \mid C_2) : \mathscr{P}(\mathscr{C}) \to \mathbb{R} : C_1 \mapsto P(C_1 \mid C_2)$ is a probability set function.

iv) $P(C_1 \cap C_2) = P(C_2)P(C_1 \mid C_2) = P(C_2)P(C_2 \mid C_1)$.

This last fact can be extended to more events

$$\begin{aligned} P(C_1 \cap C_2 \cap C_3) &= P(C_1 \cap C_2) \cdot P(C_3 \mid C_1 \cap C_2) \\ &= P(C_1) \cdot P(C_2 \mid C_1) \cdot P(C_3 \mid C_1 \cap C_2) \end{aligned}$$

and used as a way to calculate the probability of an intersection of events.

Example 1.4.3. There is a bucket 10 different coloured jelly-beans. In an experiment you reach your hand into the bucket and pull out 3 jellybeans unseen. The probability of the that we pull blue, green, and red, is $1/\binom{10}{3}$. But we can compute this another way. The probability $P(C_b)$ that one of the chosen jellybeans is blue is $P(C_b) = \binom{9}{2}/\binom{10}{3}$, the probability that one of the other two is green is $P(C_g \mid C_b) = 8/\binom{9}{2}$ and the probability that the final one is red is $P(C_r \mid C_g \cap C_b) = 1/8$.

This checks out, as

$$\frac{\binom{9}{2}}{\binom{10}{3}} \cdot \frac{8}{\binom{9}{2}} \cdot \frac{1}{8} = \frac{1}{\binom{10}{3}}.$$

We have been using the notion of independence implicitly in some of our examples. In the experiment when we tossed two coins, we said the probability of the outcome, say HH, was $P(HH) = 1/4$. We were assuming that the outcome of the second toss was independent of the outcome of the first. In this case we say that the two tosses, or experiments, are *independent*.

We also assumed independence of the two dice rolls in the two dice experiment.

1.4.2 Bayes Theorem

Let the events $C_1, \ldots, C_n$ be a partition of the sample space $\mathscr{C}$; that is, assume that

i) C_i and C_j are independent for $i \neq j \in [n]$, and

ii) $\bigcup C_i = \mathscr{C}$.

The outcome of an experiment on $\mathscr{C}$ must be in exactly one of the C_1, and so for any event C we have that

$$P(C) = \sum_{i=1}^{n} P(C \cap C_i) = \sum_{i=1}^{n} P(C \mid C_i) \cdot P(C_i). \tag{3}$$

Example 1.4.4. Consider the following experiment:

i) Flip a 1/3-coin A.

ii) Event C_1 is the event that A shows heads, in this event, flip a 1/3-coin; event C_2 is the event that A shows tails, in this event, flip a 1/2 coin.

iii) C_H is the event that the second flip is a head.

Now it is easy to compute $P(C_H|C_1) = 1/3$, and say,

$$P(C_H) = P(C_H|C_1)\cdot P(C_1)+P(C_H|C_2)\cdot P(C_2) = (1/3\cdot 1/3)+(2/3\cdot 1/2) = 4/9.$$

But what is $P(C_1 \mid C_H)$? Intuitively, we see that in the computation of $P(C_H)$, 1/9 of the 4/9 came from the case when C_1 held. So $P(C_1 \mid C_H) = 1/4$.

This is exactly what Bayes Theorem says.

Theorem 1.4.5 (Bayes Theorem)**.** *Let the events $C_1, \ldots, C_n \in \mathcal{B}$ be a partition of the sample space $\mathscr{C}$, and $C \in \mathcal{B}$. Then for any $j \in [n]$,*

$$P(C_j \mid C) = \frac{P(C \cap C_j)}{\sum_{i=1}^n P(C \cap C_i)} = \frac{P(C_j)P(C \mid C_j)}{\sum_{i=1}^n P(C_i)P(C \mid C_i)}.$$

Proof. Indeed,

$$P(C_j \mid C) = \frac{P(C \cap C_j)}{P(C)} = \frac{P(C \mid C_j)P(C_j)}{P(C)}.$$

Putting (3) in the bottom of the right-hand side gives the identity. □

Example 1.4.6. Plants 1, 2 and 3 produce respectively 10%, 50%, and 40% of the lightbulbs produced by a lightbulb company. Lightbulbs made in these plants are defective with probabilities .01, .03, and .04 respectively. What is the probability that a randomly chosen defective lightbulb was produced in plant 1?

By Bayes Theorem the probabilitly is

$$\frac{.10 * .01}{(.10 * .01) + (.50 * .03) + (.40 * .04)} = \frac{1}{32}.$$

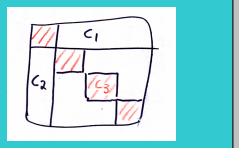
1.4.3 Mutual Independence

Definition 1.4.7. Events $C_1, \ldots, C_n$ are *(pairwise) independent* if for all $i \neq j$, C_i and C_j are independent: $P(C_i \cap C_j) = P(C_i) \cdot P(C_j)$. They are *mutually independent* if for all $S \subset [n]$,

$$P(\bigcap_{i\in S} C_i) = \prod_{i\in S} P(C_i).$$

Problem 1.4.8. Show that a family of pairwise independent events need not be mutually independent.

Hint



Problem 1.4.9. Show that if $C_1, \ldots, C_n$ are mutually independent, then so are

i) $C_1 \cup C_2$ and C_3, or

ii) $C_1^c \cap C_2$ and C_3.

Problems from the Text

Section 1.4: 6,8,11,18,23,30,34

1.5 Random Variables

Definition 1.5.1. A *random variable* or *RV* is a real function

$$X : \mathscr{D} \to \mathbb{R}$$

on the sample space $\mathscr{D}$ of some experiment. The image $\mathscr{C} = X(\mathscr{D})$ is called the *space* of X.

Example 1.5.2. In an experiment, we flip 100 fair coins. So the sample space is $\mathscr{D} = \{H, T\}^{100}$. Let X be the random variable such that counts the number of heads in an outcome of $\mathscr{D}$. Then $\mathscr{C} = X(\mathscr{D}) = \{0, 1, 2, \ldots, 100\}$.

The random variable is used to define a new probablity space on $\mathscr{C} = \{0, 1, 2, \ldots, 100\}$, which is usually a bit easier to work with than the original probability space on $\mathscr{D} = \{H, T\}^{100}$. Indeed, by the construction of the experient we can compute the probabilty of any event.

Problem 1.5.3. In the above example what are the following probabilities?

i) $P(X = 3)$

ii) $P(2 \leq X \leq 30)$

iii) $P(2 < X < 30)$

This allows us to defined the cumulative distribution function, which we will use in the next section to define the a probablility set function for the new probability space.

Definition 1.5.4. The *cumulative distribution function* or *cdf* of a random variable X is the function $F_x : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = P(X \leq x).$$

Example 1.5.5. Continuing the above example, we have that $F_X(0) = \left(\frac{1}{2}\right)^{100} = F_X(100)$, that $F_X(1) = F_X(0) + \binom{100}{1}\left(\frac{1}{2}\right)^{100}$ and that in general, for $x \in \mathscr{C}$,

$$F_X(x) = \left(\frac{1}{2}\right)^{100} \sum_{i=0}^{x} \binom{100}{i}.$$

Observe also that we have such values as $F_X(-3) = 0$, $F_X(499) = 1$, and $F_X(2.3) = F_X(2)$.

Problem 1.5.6. What is the cdf of a random variable X that counts the number of heads in an experiment in which 100 p-coins are tossed?

Problem 1.5.7. In the two dice experiment with the sample space $\mathscr{D} = \{(x, y) \mid x, y \in [6]\}$ of 36 equally likely outcomes, let $X : \mathscr{C} \to \mathbb{R}$ be the random variable defined by $X((x, y)) = x + y$. Find $F_X(4)$.

Example 1.5.8. Let X be the identity on the sample space $\mathscr{C} = [0, 1]$ in which each outcome is equally likely. Then $F_X(x) = P(X \leq x) = x$, for $x \in [0, 1]$. (Also $F_X(x) = 0$ if $x < 0$ and $F_X(x) = 1$ if $x > 1$.)

A random variable X is *discrete* if its sample space $\mathscr{C}$ is finite or countable. Examples 1.5.2 and 1.5.5 have discrete RVs while the RV in Example 1.5.8 is not discrete. The treatment of discrete and non-discrete RVs is a little different, and we will consider them separately in the next two sections. But before we do this, we make a couple more observations about the cdf.

Theorem 1.5.9. *Let F be the cdf of an RV. Then*

i) $a < b \Rightarrow F(a) \leq F(b)$,

ii) $\lim_{x \to -\infty} F(x) = 0$,

iii) $\lim_{x \to \infty} F(x) = 1$, *and*

iv) $\lim_{x \to a^+} F(x) = F(a)$.

Problem 1.5.10. Prove the above theorem.

Problem 1.5.11. Show that $\lim_{x \to a^-} F(x) = F(a)$ need not be true.

Problem 1.5.12. For an event $B \subset \mathscr{D}$, the *indicator random variable* $I_B : \mathscr{D} \to [0, 1]$, is defined by

$$I_B(x) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{otherwise.} \end{cases}$$

Show that $P(I_B = 1) = P(B)$.

1.6 Discrete Random Variables

Recall that a random variable X is discrete if it has a countable sample space $\mathscr{C}$. For such a variable one can talk of the probability of a given outcome $x \in \mathscr{C}$.

Definition 1.6.1. For a discrete random variable X, the *probability mass function* or *pmf* of X is the function $p_X : \mathbb{R} \to [0,1]$ defined by

$$p_X(x) = P(X = x).$$

Example 1.6.2. Let X be the random variable that counts the number of flips of a fair coin you make until one coin shows up heads. The sample space $\mathscr{C}$ of X is the positive integers. Then we have, for example, $p_X(1) = 1/2$, $p_X(2) = 1/4$, $p_X(n) = 1/2^n$, and $p_X(1/2) = 0$.

Problem 1.6.3. Show that for the pmf p of a discrete RV X, $\sum_{x \in \mathscr{C}} p(x) = 1$.

Example 1.6.4. Let X be the random variable from Problem 1.5.6 that counts the number of heads showing up when we a p-coin 100 times. Then

$$p_X(37) = \binom{100}{37} p^{37} q^{63} = F_X(37) - F_X(36).$$

This exhibits a fundamental relationship between the cdf F_X and the pmf p_X of a discrete RV:

$$F_X(x) = \sum_{i \in \mathscr{C}, i \leq x} p_X(i).$$

Problem 1.6.5. Prove this. Show that it need not hold for non-discrete RVs. (Hint: What is p_X when X is the RV from Example 1.5.8?)

The problem above exhibits the main difference between discrete and non-discrete random variables. The pmf may be trivial for non-discrete RVs. In the next section we will define an analogue of the pmf for certain nice non-discrete RVs. Before we do this though, we talk about transformations of random variables.

Problems from the Text

Section 1.5: 2,3,8,9

1.6.1 Transformations of Discrete RV

Often we will define one random variable as a function of another.

Example 1.6.6. Let X be an RV with space $\mathscr{C} = \{\pm 1, \pm 2, \ldots, \pm 5\}$, and $p_X(x) = 1/10$ for all $x \in \mathscr{C}$

Let $Y = X^2$. Then Y is an RV with space $\mathscr{D} = \{1, 4, 9, 16, 25\}$. For each $y \in \mathscr{D}$ we have that

$$\begin{aligned} p_Y(y) &= P(Y = y) \\ &= P(X^2 = y) \\ &= P(X \in \{\pm\sqrt{y}\}) \\ &= p_X(-\sqrt{y}) + p_X(\sqrt{y}) \end{aligned}$$

So $p_Y(y) = 1/5$ for all $y \in \mathscr{D}$.

It is easy to see that in general, if $Y = g(X)$ then

$$p_Y(y) = \sum_{x \in g^{-1}(y)} p_X(x).$$

If g is one-to-one, this simplifies to

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

Problem 1.6.7. Show that if $Y = g(X)$ for monotone strictly increasing g then $F_y(g(x)) = F_x(x)$. What can we say if g is decreasing?

Problems from the Text

Section 1.6: 1,2,3,5,7,10

1.7 Continuous Random Variables

Recall that if an RV X is not discrete, its pmf $p_X(x) = P(X = x) = \lim_{\varepsilon \to 0}(F_X(x) - F_X(x - \varepsilon))$ may be identically 0. Indeed this is the situation we want.

Definition 1.7.1. A random variable X is *continuous* if its cdf F_X is a continuous function on $\mathbb{R}$.

For such functions p_X is identically 0.

Definition 1.7.2. A random variable X is *absolutely continuous* if

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

for some function f_X. The function f_X is called the *probability density function* or *pdf* of X.

We will never consider RVs that are continuous but not absolutely continuous (though they exist); so **any time we say an RV X is continuous, we will assume that it has a pdf f_X**. It then follows by the fundamental theorem of calculus that

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

Though $f_X(x)$ is not uniquely defined, it is uniquely defined except on a set of zero measure, and so, taking the 'most continuous' version of it, we call it 'the pdf' as though it were unique.

The pdf of a continuous RV is the analogue of the pmf of a discrete RV.

Problem 1.7.3. Show that for a continuous RV X,

$$P(a < X < b) = \int_a^b f_X(t)\, dt.$$

Example 1.7.4. Recall the RV X from Example 1.5.8 that had cdf $F_X(x) = x$ for all $x \in [0, 1]$. Its pdf is the derivative

$$f_x(x) = F_X'(x) = \frac{d}{dx}(x) = 1.$$

We say that an RV X (or its sample space $\mathscr{C}$) has a *uniform distribution* if the pdf (or pmf) is constant on its support. The RV X from the above example is said to have the *standard* uniform distribution. This is denoted $X \sim \text{Unif}([0, 1])$. The RV X from Example 1.6.6 is a uniformly distributed discrete random variable.

Note

Given a sample space $\mathscr{C}$, we sometimes say that an event is 'chosen at random' to mean we an event is chosen according to a uniform distribution.

Problem 1.7.5. Find the pdf of the uniformly distributed random variable $X \sim \text{Unif}([-1, 1])$ on the space $\mathscr{C} = [-1, 1]$.

Problem 1.7.6. Show that for a uniformly distributed space $\mathscr{C} \subset \mathbb{R}^n$ the probability of an event C is

$$P(C) = \frac{\iiint_C 1\, dx}{\iiint_{\mathscr{C}} 1\, dx}.$$

That is, show that the probability of an event is proportional to its volume.

Example 1.7.7. Let a point (x, y) be chosen randomly from $\mathscr{C} = \{(x, y) \mid x^2 + y^2 < 1\}$ and let X be its distance from $(0, 0)$. By definition $\mathscr{C}$ is uniformly

distributed, but $\mathscr{C}$ is not the space of X. (The space of X is $[0,1]$.) We observe that X is not uniformly distributed.

Indeed, its cdf is $F_X(t) = P(X \leq t)$ which is the area of the event $\{(x,y) \mid x^2 + y^2 \leq t\}$ over the area of $\mathscr{C}$ (which is π.) So

$$F_X(x) = 1/\pi \int_0^x 2\pi t \, dt = \int_0^x 2t \, dt = x^2$$

for $x \in [0,1]$. And so $f_X(x) = \frac{d}{dx}x^2 = 2x$, which is not constant.

1.7.1 Transformations of Continuous RVs

Now let $Y = X^2$ be a transformation of X from the above example. So $Y = g(X)$ where the function $g(x) = x^2$ is monotone strictly increasing on the space $(0,1]$. It is tempting to follow the discrete case and say that the pdf of f is

$$f_Y(y) = f_X(g^{-1}(y)) = 2\sqrt{y}.$$

But this is not true! Indeed, the cdf of Y is

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = F_X(\sqrt{y}) = \sqrt{y}^2 = y,$$

and so the pdf is $f_Y(y) = \frac{d}{dy}y = 1$.

Of course! A transformation is just a change of variables from calculus. In general, differentiating the above equation $F_Y(y) = F_X(g^{-1}(y))$ with respect to y, we get, by the chain rule,

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

Indeed, where $f_X(x) = 2x$ and $y = g(x) = x^2$, so $x = g^{-1}(y) = \sqrt{y}$, we found:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \frac{dx}{dy} = 2\sqrt{y} \cdot \frac{d}{dy}\sqrt{y} = 2\sqrt{y} \cdot \frac{1}{2\sqrt{y}} = 1.$$

We have proved the following theorem (in the case that g is monotone increasing).

Theorem 1.7.8. *Let X be a continuous RV with pdf $f_X(x)$ and let $Y = g(X)$ where g is one-to-one and differentiable on the support of X. Then the pdf of Y is*

$$f_Y(y) = f_X(g^{-1}(y)) \cdot |J|$$

where $J = \frac{d}{dy}g^{-1}(y)$, for y in the support $\{g(x) \mid x \in \operatorname{Supp} X\}$ of Y.

Problem 1.7.9. Where in the proof are we using that fact that g is monontone increasing?

The value $J = \frac{d}{dy} g^{-1}(y)$ is called the *Jacobian* of the transformation g. In other books it may be called the Jacobian of g^{-1}.

Example 1.7.10. Where $X \sim \text{Unif}((0,1))$ let $Y = -2\log X$. So Y has support $(0,\infty)$. The transformation $h : X \to Y : x \mapsto -2\log x$ is one-to-one with inverse $h^{-1}(y) = e^{-y/2}$, so it has Jacobian

$$J = \frac{d}{dy} e^{-y/2} = -\frac{1}{2} e^{-y/2}.$$

The pdf of Y is thus

$$f_Y(y) = f_X(e^{-y/2}) \cdot |J| = 1 \cdot \frac{1}{2} e^{-y/2} = \frac{1}{2} e^{-y/2}$$

on the support of Y.

Problems from the Text

Section 1.7: 1,2,5,6,8,9,10

1.8 Expectation of a Random Variable

Definition 1.8.1. For a random variable X, the *expected value* or *expectation* of X is

$$E(X) = \sum_{x \in \mathscr{C}} x p_X(x)$$

or

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

depending on whether X is discrete or continuous.

Note

Technically, being an integral or possibly infinite sum, the expectation need not always exist for an RV. And indeed, we should insist that the sum/integral is absolutely convergent so that the expectation is independent of an ordering of the sample space. But this is not an issue for all RVs that we consider. In theorems dealing with expectation, we will implicitly assume sufficiently strong convergence.

Example 1.8.2. The expected value when you roll a die is $(1+2+\cdots+6)/6 = 3.5$.

Example 1.8.3. Let X be the distance from $(0,0)$ of a randomly chosen point in the unit circle $S = \{(x,y) \mid x^2 + y^2 \leq 1\}$. What is $E(X)$?

Using that the pdf of X is $f(x) = 2x$ (from Example 1.7.7), we get that

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot 2x \, dx = \int_0^1 x \cdot 2x \, dx \\ &= 2\int_0^1 x^2 \, dx = 2\left(1/3x^3\right)_0^1 = 2/3 \end{aligned}$$

If $Y = g(X)$ is a transformation of an X then it is very believable that

$$E(Y) = E(g(X)) = \sum_{\mathscr{C}_x} g(x)p_X(x).$$

There are issues of convergence to deal with, of course, as the 'obvious' proof

$$\begin{aligned} E(Y) &= \sum_{\mathscr{C}_y} y p_Y(y) = \sum_{\mathscr{C}_y} y \sum_{g(x)=y} p_X(x) \\ &= \sum_{\mathscr{C}_y} \sum_{g(x)=y} y p_X(x) = \sum_{\mathscr{C}_y} \sum_{g(x)=y} g(x) p_X(x) \\ &= \sum_{\mathscr{C}_x} g(x) p_X(x). \end{aligned}$$

requires reordering sums. We will always assume though that $\sum g(x)p_X(x)$ converges absolutely, so this proof holds.

The following tool is immediate from the above using the linearity of sums and integrals.

Theorem 1.8.4 (Linearity of Expectation)**.** *If* $Y = k_1 g_1(X) + k_2 g_2(X)$ *then* $E(Y) = k_1 E(g_1(X)) + k_2 E(g_2(X))$.

Problem 1.8.5. Prove Theorem 1.8.4.

Problem 1.8.6. Where $Y = X^2$ for X from Example 1.8.3, what is $E(Y)$? (Make a guess before you compute it. What should it be?)

Problem 1.8.7. Let I_C be the indicator variable (see Problem 1.5.12) for an event $C \in \mathscr{C}$. Show that $E(I_C) = P(A)$.

Problem 1.8.8. Let X count the number of heads that show up when n independent p-coins are flipped. Find $E(X)$.

Problem 1.8.9. Let v be a randomly chosen vertex in $G_{n,p}$. What is the expected degree $E(\deg(v))$ of v.

Problems from the Text

Section 1.8: 3,4,6,7,8,9,11

1.9 Mean Variance and Moments

Given a random variable X, we will be interested in the expected value of various functions of X. Certain ones get special names and notation. The expected value of X is also called the *mean* $\mu = E(X)$ of X. The *variance* of X is

$$\sigma^2 = \mathrm{Var}(X) = E\left((X-\mu)^2\right).$$

Expanding the square in the expression on the right, and using the linearity of expectation, we get that

$$\begin{aligned} \sigma^2 &= E\left(X^2 - 2\mu X + \mu^2\right) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

The positive square root σ of the variance is called the *standard deviation* of X.

Example 1.9.1. Let X have pdf $f(x) = \frac{1}{2}(x+1)$ for $x \in [-1,1]$. Find μ and σ^2.

We get

$$\begin{aligned} \mu = E(X) &= \int_{-1}^{1} x f(X)\, dx = \frac{1}{2}\int_{-1}^{1} x^2 + x\, dx \\ &= \frac{1}{2}\left(\frac{1}{3}(1+1)\right) = \frac{1}{3}, \end{aligned}$$

and

$$\begin{aligned} \sigma^2 = E(X^2) - \mu^2 &= \frac{1}{2}\int_{-1}^{1} x^3 + x^2\, dx - \frac{1}{9} \\ &= \frac{1}{2}\left(\frac{1}{4}(1+1) + \frac{1}{3}(1+1)\right) - \frac{1}{9} \\ &= 17/36. \end{aligned}$$

Problem 1.9.2. In terms of $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$, what is $\mathrm{Var}(X-Y)$?

1.9.1 The moment generating function

The mean $\mu_X = E(X)$ has yet another name. It is also called the *first moment* of X. And the value $E(X^2)$, used in the compuation of the variance, is called the *second moment* of X. In general $E(X^n)$ is the n^{th} *moment* of X. The 0^{th} moment is $E(1) = 1$.

Letting the moments be the coefficients of an exponential generating function:

$$M_X(t) = E(1) + tE(X) + t^2\frac{E(X^2)}{2!} + t^3\frac{E(X^3)}{3!} + \dots,$$

we get by the linearity of expectation that

$$M_X(t) = E(1 + (tX) + \frac{(tX)^2}{2!} + \dots) = E(e^{tX}).$$

So we have that the n^{th} moment $E(X^n)$ of X is also denoted $M_X^{[n]}(t)$.

From the cdf of an RV, one can compute the moments, and so find the moment generating function. On the other hand, we know from an analysis class, that the power series of a function expanded on an open interval around a point, is uniqely defined, so the moment generating function uniquely defines the moments of a distribution. The following theorem takes this one step further and asserts that from the moments, we can recover the cdf. The proof of this is beyond our scope, (and beyond the scope of the text).

Theorem 1.9.3. *If $M_X(t) = M_Y(t)$ on some open interval around t, then $F_X(z) = F_Y(z)$ for all z.*

This function $M_X(t) = E(e^{tx})$ is called the *moment generating function* or *mgf* of X. As the taylor expansion of a function about a point does not always converge, not all RVs necessarily have mgfs, and indeed the text gives examples of RVs for which the mgf does not exist. But the mgf does exist for many RVs that we will consider, and it will become a useful tool.

Problems from the Text

Section 1.9: 1,2,3,5,6,18,23

1.10 Important Inequalities and Bounds

We finish the chapter with some basic inequalities.

1.10.1 Markov's Inequality

Theorem 1.10.1 (Markov's Inequality)**.** *For any non-negative RV X and any constant c:*

$$P(X \geq c) \leq E(X)/c.$$

More generally, for any RV X, any non-negative function u of X, and any constant c:

$$P(u(X) \geq c) \leq E(u(X))/c.$$

Proof. The first statement is simply a special case of the second, so we prove just the second. We prove it in the case that X is continuous. The proof in the discrete case is essentially the same.

Let $A = \{x \mid u(x) \geq c\}$. (Recall that A^c is its complement.) Then

$$\begin{aligned} E(u(X)) &= \int_{-\infty}^{\infty} u(x) f_X(x)\, dx \\ &= \int_A u(x) f_X(x)\, dx + \int_{A^c} u(x) f_X(x)\, dx \\ &\geq \int_A u(x) f_X(x)\, dx \\ &\geq c \int_A f_X(x)\, dx = cP(x \in A) = cP(u(x) \geq c). \end{aligned}$$

The inequality follows. □

Markov's inequality is crude. Indeed if $X \sim \text{Unif}([0,4])$ then $E(X) = 2$ and taking $c = 1$ the inequality says $P(X \geq 1) \leq 2$. We could certainly give a better bound. However, the inequality is incredibly useful due to its universality.

1.10.2 Chebyshev's Inequality

Corollary 1.10.2 (Chebyshev's Inequality)**.** *Let X be an RV, then for every $\varepsilon, k > 0$, the following, clearly equivalent statements, all hold.*

i) $P(|X - \mu| \geq k\sigma) \leq 1/k^2$

ii) $P(|X - \mu| < k\sigma) > 1 - 1/k^2$

iii) $P(|X - \mu| < \varepsilon) > 1 - \sigma^2/\varepsilon^2$

Proof. Applying Markov wih $u(x) = (X - \mu)^2$ and $c = k^2\sigma^2$ gives

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

□

In Problem 1.9.3 of the text, you were asked to find $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$ for an RV X with pdf $f(x) = 6x(1-x)$. Compare this with the quick bound we can now get without even computing μ and σ:

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) > 1 - 1/4 = 3/4$$

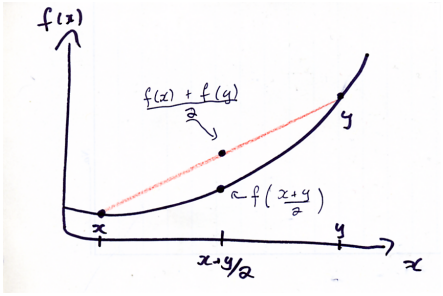
1.10.3 Jensen's Inequality

You have probably seen the next inequality several times, and proved it in a linear algebra class. It won't hurt to see it again. We state it without proof.

Definition 1.10.3. A function f is *convex* on an interval $I = [a, b]$ if for all $x, y \in I$ and all $n > 1$,

$$f(x/n + y\frac{n-1}{n}) < f(x)/n + f(y)\frac{n-1}{n}.$$

The following picture for the case $n = 2$ show that this definition of convexity agrees with the definition for a continuous function that it is convex if the second derivative is positive.



Theorem 1.10.4 (Jensen's Inequality)**.** *If f is convex, then*

$$f\left(\frac{x_1 + x_2 + \ldots x_n}{n}\right) \leq \frac{f(x_1) + f(x_2) + \cdots + f(x_n)}{n}.$$

For any RV X, this means that

$$f(E(X)) \leq E(f(X)).$$

Example 1.10.5. The function x^2 is convex so $E(X)^2 < E(X^2)$. Thus $\mathrm{Var}(X) = E(X^2) - E(X)^2$ is non-negative.

Problem 1.10.6. Sometimes the mean of a set of numbers $\{x_1, \ldots, x_n\}$ is called the *arithmetic mean* $AM = \frac{1}{n}\sum x_i$, distinguishing it from the *geometric mean* $GM = (\prod x_i)^{1/n}$ and the *harmonic mean* $HM = (\frac{1}{n}\sum \frac{1}{x_i})^{-1}$.

Using that $-\log x$ is convex, use Jensen's Inequality to show that for any set $\{x_1, \ldots, x_n\}$ of positive numbers, $HM \leq GM \leq AM$.

Problems from the Text

Section 1.10: 2,3,4,6

2 Multivariate Distributions

In Example 1.5.2 we considered the experiment of tossing 100 p-coins and let Y be the random variable counting the number of heads. The experiment can be viewed as a set of 100 random variables $X_1, \ldots, X_n$ each having the pmf of a p-coin. In this context we can view Y as a function $Y = \sum_{i=1}^{100} X_i$ of the multivariate distribution $\mathbf{X} = (X_1, \ldots, X_{100})$. Let's go into more detail.

Definition 2.0.1. A *random vector* $\mathbf{X} = (X_1, \ldots, X_n)$ is a set of RVs on a sample space $\mathscr{D}$. The *space* of $\mathbf{X}$ is

$$\mathscr{C} = \{(X_1(c), X_2(c), \ldots, X_n(c)) \mid c \in \mathscr{C}\}.$$

Example 2.0.2. Where $\mathscr{D}$ are people in a sample population, $\mathbf{X} = (\text{Height}, \text{Weight}, \text{Age})$ is a random vector.

In the above example, the component RVs in our random vector are dependent. Often, we define a random vector such that the component vectors are independent.

Example 2.0.3. The random graph $G_{n,p}$ can be viewed as a random vector consisting of $\binom{n}{2}$ independent RVs X_e each with the distribution of a p-coin, one for each possible edge e on the vertices $[n]$.

2.1 Distributions of Two Random Variables

We extend many of our definitions for RVs to random vectors. For most definitions the extension from two variables to arbitrarily many is trivial. For those that it isn't, we will revisit them later for more than two variables.

Definition 2.1.1. The *(joint) cdf* of $\mathbf{X} = (X_1, X_2)$ is

$$F_{\mathbf{X}}(\boldsymbol{x}) = F_{X_1,X_2}(x_1, x_2) = P\left((X_1 \leq x_1) \text{ and } (X_2 \leq x_2)\right).$$

The *(joint) pmf* for discrete $\mathbf{X}$ is

$$p_{\mathbf{X}}(\boldsymbol{x}) = p_{X_1,X_2}(x_1, x_2) = P\left((X_1 = x_1) \text{ and } (X_2 = x_2)\right).$$

The *(joint) pdf* for continuous $\mathbf{X}$ is a function $f_{\mathbf{X}}$ such that

$$F_{\mathbf{X}}(\boldsymbol{x}) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{\mathbf{X}}(t_1, t_2)\, dt_1\, dt_2,$$

so almost everywhere we have

$$f_{\mathbf{X}}(\boldsymbol{x}) = \frac{\partial^2 F_{\mathbf{X}}(x_1, x_2)}{\partial x_1 \partial x_2}.$$

Example 2.1.2. Let $f_{\boldsymbol{x}}(x_1, x_2) = 6x_1^2 x_2$ for $x_i \in (0,1)$ be the joint pdf of $\mathbf{X} = (X_1, X_2)$. Then $P(1/4 < X_1 \leq 3/4, 0 < x_2 < 2)$ is

$$\int_{\frac{1}{4}}^{\frac{3}{4}} \int_0^1 6x_1^2 x_2 \, \mathrm{dx}_2 \, \mathrm{dx}_1 \quad = \quad \int_{\frac{1}{4}}^{\frac{3}{4}} 3x_1^2 \, \mathrm{dx}_1 = [x_1^3]_{\frac{1}{4}}^{\frac{3}{4}} = 13/32$$

From the joint pmf of a random vector, we can isolate the *(marginal) pmf* of any one component RV as follows,

$$p_{X_1}(x_1) = \sum_{x_2} p_{\mathbf{X}}(x_1, x_2),$$

where $\sum_{x_2}$ is over all x_2 such that $(x_1, x_2) \in \mathscr{C}$.

The marginal pdf of a component of a continous random vector is defined analogously.

Problem 2.1.3. Find the marginal pdf $f_{X_1}(x_1)$ of X_1 for the joint distribution $f_{\mathbf{X}}(x_1, x_2) = 6x_1^2 x_2$ from the above example.

We can talk of the *expected value of a random vector*. It is simply the vector

$$E(\mathbf{X}) = (E(X_1), \ldots, E(X_n))$$

of expected values of its components. To find the expected value of a random vector one must find the expected value of the components. To find $E(X_1)$, or more generally $E(g(X_1))$, we can get the marginal distribution of X_1, and then find the expected value as in the previous chapter. Or we can find it directly:

Example 2.1.4. Where $f_{\mathbf{X}}(x_1, x_2) = 6x_1^2 x_2$, the expected value of X_1^2 is

$$\begin{aligned} E(X_1^2) &= \int_0^1 \int_0^1 x_1^2 \cdot 6x_1^2 x_2 \, dx_2 \, dx_1 \\ &= \frac{1}{2} \int_0^1 6x_1^4 \, dx_1 = \frac{6}{10} \end{aligned}$$

Problem 2.1.5. Find $E(X_1^2)$ in the above example by using the marginal distribution of X_1 which you found in Problem 2.1.3.

The following generalisation of Theorem 1.8.4 to random vectors is a key tool in saying anything of substance in statistical inference or with the probabilistic method.

Theorem 2.1.6 (Additivity of Expectation)**.** *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector and $k_1, \ldots, k_n$ be real numbers. Then*

$$E(\sum k_i X_i) = \sum k_i E(X_i).$$

Proof. We do the continuous case for a vector of two RVs:

$$\begin{aligned} E(k_1X_1 + k_2X_2) &= \int_{\mathbb{R}}\int_{\mathbb{R}} (k_1x_1 + k_2x_2) f_{\mathbf{X}}(x_1, x_2)\,\mathrm{dx}_2\,\mathrm{dx}_1 \\ &= k_1 \int\int x_1 f_{\mathbf{X}}(x_1, x_2)\,\mathrm{dx}_2\,\mathrm{dx}_1 + k_2 \int\int x_2 f_{\mathbf{X}}(x_1, x_2)\,\mathrm{dx}_2\,\mathrm{dx}_1 \\ &= k_1E(X_1) + k_2E(X_2) \end{aligned}$$

□

Definition 2.1.7. The mgf of $\mathbf{X} = (X_1, X_2)$ is

$$M_{\mathbf{X}}(\boldsymbol{t}) = E(e^{\boldsymbol{t}\cdot\boldsymbol{x}}) = E(e^{t_1x_1+t_2x_2})$$

Observe that $M_{X_1}(t)$ is recovered as $M_{\mathbf{X}}(t, 0)$.

Problems from the Text

Section 2.1: 1,2,3,6,7,9,12

2.2 Transformations of Bivariate RV

Assume that $\mathbf{X}$ is a random vector and Y is some function $Y = g(\mathbf{X})$ of $\mathbf{X}$. Given the joint distribution of a the RV $\mathbf{X}$ will want to find the distribution of Y. As in the univariate case, we can go through the cdf, or go directly, using, in the continuous case, a transformation multiplier.

As before, the discrete case is more straight-forward.

2.2.1 Discrete Case

In the one variable case when we had $Y = g(X)$ for one-to-one (and increasing) g we observed that $p_Y(y) = p_X(g^{-1}(y))$. Now, when $Y = g(X_1, X_2)$, g is very rarely one-to-one This causes little problem in the discrete case, the main difficulty is finding the set $g^{-1}(y)$.

Example 2.2.1. Let $\mathbf{X} = (X_1, X_2)$ have pmf

$$p_{\mathbf{X}}(x_1, x_2) = \frac{\mu_1^{x_1}\mu_2^{x_2}e^{-\mu_1}e^{-\mu_2}}{x_1!x_2!}, \qquad x_i \in \mathbb{N},$$

and $Y = X_1 + X_2$.

Now $p_Y(y) = \sum_S p_{\mathbf{X}}(x_1, x_2)$ where the sum is over

$$S = \{(x_1, x_2) \in \mathbb{N}^2 \mid x_1 + x_2 = y\} = \{(x_1, y - x_1) \in \mathbb{N}^2 \mid 0 \le x_1 \le y\}.$$

Computing, we get

$$\begin{aligned} p_Y(y) &= \sum_{x_1=0}^{y} \frac{\mu_1^{x_1}\mu_2^{y-x_1}e^{-\mu_1}e^{-\mu_2}}{x_1!(y-x_1)!} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{y!}\sum_{x_1=0}^{y}\frac{y!}{x_1!(y-x_1)!}\mu_1^{y-x_1}\mu_2^{x_1} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{y!}\sum_{x_1=0}^{y}\binom{y}{x_1}\mu_1^{y-x_1}\mu_2^{x_1} \\ &= \frac{(\mu_1+\mu_2)^y e^{-(\mu_1+\mu_2)}}{y!} \end{aligned}$$

where the last line uses the recognition of the expansion of the binomial $(\mu_1 + \mu_2)^y$.

2.2.2 Continuous Case

Assume now that $\mathbf{X} = (X_1, X_2)$ is a random vector of continuous variables, and $Y = g(X_1, X_2)$.To get the pdf of Y from the joint pdf of $\mathbf{X}$ we can go through the cdf.

Example 2.2.2. Let $\mathbf{X} = (X_1, X_2)$ have the uniform distribution on the unit square $D = \{(x_1, x_2) \mid 0 \le x_i \le 1\}$; so the pdf $f_{\mathbf{X}}$ is 1 on D and 0 elsewhere. Let $Y = X_1 + X_2$.

The cdf of Y is $F_Y(y) = \iint_S \mathrm{dx}_1\,\mathrm{dx}_2$ where S consists of the set of pairs (x_1, x_2) such that $x_1 + x_2 \le y$. This breaks into the cases

$$F_Y(y) = \begin{cases} \int_0^y \int_0^{y-x_1} \mathrm{dx}_2\,\mathrm{dx}_1 = \frac{y^2}{2} & 0 \le y \le 1 \\ 1 - \int_{y-1}^1 \int_{y-x_1}^1 \mathrm{dx}_2\,\mathrm{dx}_1 = 1 - \frac{(2-y)^2}{2} & 1 \le y \le 2 \end{cases}$$

Differentiating, we thus get that $f_Y(y) = y$ when $0 \le y \le 1$ and $f_Y(y) = 2-y$ when $1 \le y \le 2$.

This is fine, but for more complicated transformations the regions S can become quite complicated.

The method of transformations can remove some of this complication. Replacing the transformation $Y = g(X_1, X_2)$ with a one-to-one transformation $(Y_1, Y_2) = \boldsymbol{u}(X_1, X_2)$, usually by replacing Y with Y_1 and choosing Y_2 strategically, we get a reverse transformation $(X_1, X_2) = \boldsymbol{w}(Y_1, Y_2)$, and get that $f_{\mathbf{Y}}(\boldsymbol{y}) = f_{\mathbf{X}}(\boldsymbol{x})|J|$ for the Jacobian

$$J = \frac{\partial x_1}{\partial y_1}\frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2}\frac{\partial x_2}{\partial y_1}.$$

Let's repeat the above example using this method, and then recall the proof of this result from calculus.

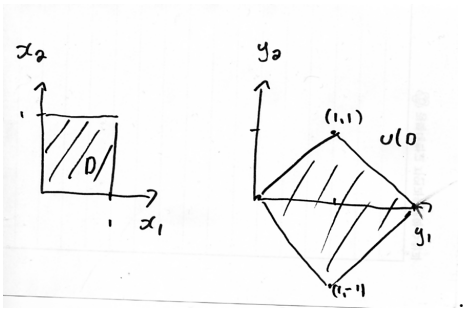
Example 2.2.3. Where $\mathbf{X} = (X_1, X_2)$ has the uniform distribution on the unit square cube D let

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \boldsymbol{u} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} X_1 + X_2 \\ X_1 - X_2 \end{bmatrix}.$$

This has inverse transformation

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \boldsymbol{w} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \frac{Y_1+Y_2}{2} \\ \frac{Y_1-Y_2}{2} \end{bmatrix}.$$

The transformation $\boldsymbol{u}$ takes the sample space D of $\mathbf{X}$ to the sample space $\boldsymbol{u}(D)$ shown below.



Computing

$$J = \left(\frac{1}{2}\right)\left(-\frac{1}{2}\right) - \frac{1}{2}\frac{1}{2} = -\frac{1}{2},$$

we get that

$$f_{\mathbf{Y}}(\boldsymbol{y}) = f_{\mathbf{X}}(\boldsymbol{w}(\boldsymbol{y})) \cdot |J| = 1 \cdot \left|-\frac{1}{2}\right| = \frac{1}{2}$$

on $\boldsymbol{u}(D)$ and 0 elsewhere.

To get the marginal pdf f_{Y_1} we then integrate with respect to Y_2. When $y_1 \in [0, 1]$, $f_{\mathbf{Y}}(y_1, y_2)$ is 1 for $y_2 \in [-y_1, y_1]$, so

$$f_{Y_1}(y_1) = \int_{-y_1}^{y_1} \frac{1}{2}\, dy_2 = y_1$$

and (as $\boldsymbol{u}(D)$ is symmetric about $y_1 = 1$,) when $y_1 \in [1, 2]$, $f_{\mathbf{Y}}(y_1, y_2) = f_{\mathbf{Y}}(2 - y_1, y_2)$ so $f_{Y_1}(y_1) = 2 - y_1$.

Problem 2.2.4. We might write the transformation $\boldsymbol{u}$ in the above example as

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

calling the square matrix in the middle U. Write the inverse transformation $\boldsymbol{w}$ as

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = W \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

for some square matrix W. What do you notice about U and W?

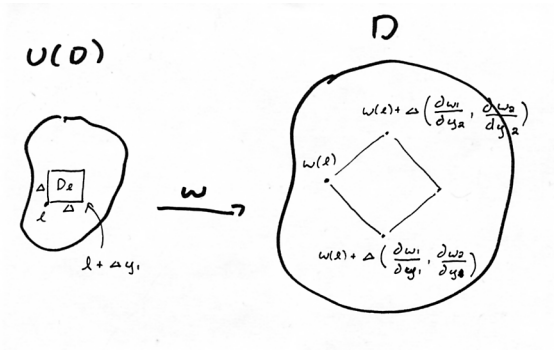
2.2.3 Recalling Jacobian from Calculus

To see that $f_{\mathbf{Y}}(\boldsymbol{y}) = f_{\mathbf{X}}(\boldsymbol{x})|J|$ we should see that verify that this gives us

$$\oiint_{\boldsymbol{u}(R)} f_{\mathbf{Y}}(y_1, y_2)\, dy_1\, dy_2 = \oiint_R f_{\mathbf{X}}(x_1, x_2)\, \mathrm{dx}_1\, \mathrm{dx}_2,$$

for every region R.

Recall that by definition, $\oiint_{\boldsymbol{u}(R)} f_{\mathbf{Y}}(y_1, y_2)\, dy_1\, dy_2$ is the limit, as Δ gets small, of the sum $\sum_L f_{\mathbf{Y}}(\ell)\Delta^2$ where L consists of all points ℓ of a Δ-lattice on $\boldsymbol{u}(R)$. To get the value $f_{\mathbf{Y}}(\ell)\Delta^2$ in terms of $f_{\mathbf{X}}$ we must consider what the transformation w does to the Δ-lattice.



(Note, the w_i in the picture should be x_i.)

It maps a Δ-square D_ℓ with corner ℓ to the (approximate) parallelogram between the vectors $\Delta(\frac{\partial x_1}{\partial y_1}, \frac{\partial x_2}{\partial y_1})$ and $\Delta(\frac{\partial x_1}{\partial y_2}, \frac{\partial x_2}{\partial y_2})$, so has area approximately

$$A = \Delta^2 \left| \begin{matrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} \end{matrix} \right|.$$

Taking limits in this argument gives the Jacobian formula.

Problems from the Text

Section 2.2: 1,3,5,6

2.3 Conditional Distributions

We have talked about the conditional probability $P(X \in A \mid X \in B)$ of an event A given an event B. When we have two RVs on a space, it is natural to consider the conditional probability for elementary events such as $P(X = x \mid Y = y)$.

Definition 2.3.1. Let (X, Y) be a random vector. The *conditional pmf or pdf* of X, conditioned on Y, is

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \qquad \text{or} \qquad f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Note

We may use shortcut notation such as $p_{1|2}$ for $p_{X_1|X_2}$.

More generally, we might ask about the distribution of X when we fix $Y = y$.

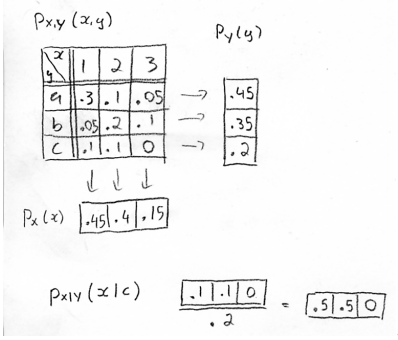
Definition 2.3.2. For fixed y the function

$$p_{X|y} : x \mapsto p_{X|Y}(x|y)$$

is itself the pmf of a random variable, the *conditional random variable* (or *conditional distribution*) which we denote $X|y$.

Being an RV, we can compute its mean and variance.

Example 2.3.3. Let (X, Y) have the joint distribution $p_{X,Y}$ shown.



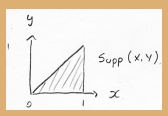
The conditional pmf $p_{X|Y}(x|y)$ restricts to the y^{th} row, scales it by $1/p_Y(y)$ (so that it sums to 1) and then returns the x value.

Compare this to the marginal pmf p_X which x to the sum of the x^{th} column, and the marginal pmf p_Y which takes y to the sum of the y^{th} row. So $p_x(3) = .15$ and $p_Y(b) = .35$.

Note

The notation $p_{X|Y}$ vs $p_{X|y}$ can get confusing. The conditional pmf $p_{X|Y}$ is a function of two variables, x and y, which we usually write as $x|y$, to parallel the indexing. The function $p_{X|y}$ is the pmf of the distribution $X|y$. We think of y as being fixed, so the argument of the function is usually written as x. Use this heuristic aid: if the letter in the index is uppercase, the argument has a corresponding lower case value.

Note



Example 2.3.4. Let (X, Y) have the pdf $f_{X,Y}(x, y) = 6y$ on its support $0 \leq y \leq x \leq 1$. We find the mean μ of $X|y$.

First, by definition $\mu = E(X|y) = \int_y^1 x \cdot f_{X|y}(x)\,\mathrm{dx}$, so we need to find $f_{X|y}(x) = f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$. Now the marginal pdf of y is

$$f_Y(y) = \int_y^1 f_{X,Y}(x, y)\,\mathrm{dx} = \int_y^1 6y\,\mathrm{dx} = 6y(1 - y),$$

so

$$f_{X|y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{6y}{6y(1 - y)} = \frac{1}{1 - y},$$

and so

$$E(X|y) = \frac{1}{1 - y}\int_y^1 x\,\mathrm{dx} = \frac{1}{1 - y}\frac{1}{2}(1 - y^2) = \frac{1 + y}{2}.$$

Problem 2.3.5. Show that the variance of $X|y$ above is $\sigma^2 = \frac{y^2-2y+1}{12}$.

Problems from the Text

Section 2.3: 1,2,3,5,7

2.5 Independent Random Variables

Definition 2.5.1. Random variables X and Y, with joint pdf f_{XY} and marginal pdfs f_X and f_Y, are *independent* if

$$f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$$

(holds with probability 1). They are *dependent* otherwise.

Problem 2.5.2. Show that if X and Y are independent, then $E(X|y) = E(X)$.

There are several equivalent definitions of independence.

Theorem 2.5.3. *The following are equivalent for RVs X and Y.*

i) X and Y are independent

ii) $f_{XY}(x,y) = g(x)h(y)$ (almost everywhere) for some non-negative functions g and h

iii) The cdfs satisty $F_{XY}(x,y) = F_X(x)F_Y(y)$ for all x and y.

iv) For all intervals S_X and $S_Y \subset \mathbb{R}$,

$$P(X \in S_X, Y \in S_Y) = P(X \in S_X)P(Y \in S_Y).$$

Proof. That i) impies ii) is immediate from the definition. We first show ii) implies i). Assuming ii), the marginal pdfs are

$$f_X(x) = \int_{\mathbb{R}} g(x)h(y)\,\mathrm{d}y = g(x)\int_{\mathbb{R}} h(y)\,\mathrm{d}y = c_1 g(x)$$

for some constant c_1 and $f_Y(y) = c_2 h(y)$ for some constant c_2. As

$$\begin{aligned} 1 &= \int_{\mathbb{R}}\int_{\mathbb{R}} f_{XY}(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int_{\mathbb{R}}\int_{\mathbb{R}} g(x)h(y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{\mathbb{R}} g(x)\,\mathrm{d}x \int_{\mathbb{R}} h(y)\,\mathrm{d}y = c_1 c_2 \end{aligned}$$

we get that $c_1 c_2 = 1$. So

$$f_{XY}(x,y) = g(x)h(y) = \frac{f_X(x)f_Y(y)}{c_1 c_2} = f_X(x)f_Y(y).$$

For i) implies iii):

$$\begin{aligned} F_{XY}(x,y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(s,t)\,\mathrm{d}t\,\mathrm{d}s = \int\int f_X(s)f_Y(t)\,\mathrm{d}t\,\mathrm{d}s \\ &= \int f_X(s)\,\mathrm{d}s \int f_Y(t)\,\mathrm{d}t = F_X(x)F_Y(y) \end{aligned}$$

For iii) implies i):

$$\begin{aligned} f_{XY}(x,y) &= \frac{\partial^2 F_{XY}(x,y)}{\partial y \partial x} = \frac{\partial^2}{\partial y \partial x} F_X(x)F_Y(y) \\ &= f_X(x)\frac{\partial}{\partial y}F_Y(y) = f_X(x)f_Y(y) \end{aligned}$$

The proof of the equivalence of iii) and iv) is just as straight forward, so we skip it. □

Theorem 2.5.4. *If X and Y are independent, then $E(XY) = E(X)E(Y)$.*

Proof.

$$\begin{aligned} E(XY) &= \int\int xyf_{XY}(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int xyf_X(x)f_Y(y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int xf_X(x)\,\mathrm{d}x \int yf_Y(y)\,\mathrm{d}y = E(X)E(Y) \end{aligned}$$

□

All the proofs are basically the same: if X and Y are indepentent, then we can seperate our double sums or integrals. With essentially the same proof as we used above one can do the following problems.

Problem 2.5.5. Show that if X and Y are independent and $u(X)$ and $w(Y)$ are transformations of X and Y respectively, then $E(u(X)w(Y)) = E(u(X))E(w(Y))$.

The *joint mgf* of (X,Y) is $M_{(X,Y)}(t_X,t_Y) = E(e^{t_X X + t_Y Y})$. It is a function of two variables; observe that $M_{(X,Y)}(t,0) = M_X(t)$.

Problem 2.5.6. Show that X and Y are independent if and only if

$$M_{(X,Y)}(t_X,t_Y) = M_X(t_X)M_Y(t_Y).$$

Problems from the Text

Section 2.5: 1,3,4,5,8,9,12

2.4 The Correlation Coefficients

The following is done for continuous variables, which are assumed to be nice, (ie., all necessary expectations are assumed to exist). It holds though for discrete variables as well.

For random variables X and Y with means μ_X and μ_Y respectively, the *covariance* is $\mathrm{Cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y))$.

Problem 2.4.1. Show that $\mathrm{Cov}(X,Y) = E(XY) - \mu_X\mu_Y$.

Observe that if $X = Y$ then $\mathrm{Cov}(X,Y) = E(X^2) - E(X)^2 = \mathrm{Var}(X)$; so the covariance can be seen as a generalisation of the variance.

Problem 2.4.2. Show that if X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$. Show if $E(X|y)$ is an increasing (decreasing) function of y then $\mathrm{Cov}(X,Y) > 0$ ($\mathrm{Cov}(X,Y) < 0$).

The magnitude of $\mathrm{Cov}(X,Y)$ is hard to interpret, but the normalised version, the *correlation coefficient*

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

has the property that $-1 \leq \rho \leq 1$. If $X = Y$ then $\rho = 1$ and if $X = -Y$ then $\rho = -1$. So $|\rho|$ is a measure of how closely X and Y are related.

Problems from the Text

Section 2.4: 1,3,4,10

2.6 Extension to more Random Variables

Let $\mathbf{X} = (X_1, \ldots X_n)$ be an n dimensional random vector. Its joint cdf is

$$F_{\mathbf{X}}(\boldsymbol{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots X_n \leq x_n)$$

and its joint pdf is a function $f_{\mathbf{X}}$ such that

$$F_{\mathbf{X}}(\boldsymbol{y}) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_n} f_{\mathbf{X}}(\boldsymbol{x})\, \mathrm{d}x_n \ldots \mathrm{d}x_1.$$

The conditional pdfs are

$$f_{\mathbf{X}|X_i}(\boldsymbol{x}|x_i) = \frac{f_{\mathbf{X}}(\boldsymbol{x})}{f_{X_i}(x_i)}.$$

The variables $X_1, \ldots X_n$ are *mutully independent* if

$$f_{\mathbf{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

(with probability 1.)

In this case

$$E(\prod u_i(X_i)) = \prod E(u_i(X_i))$$

for any transformations u_i of X_i. In particular:

Theorem 2.6.1. *Let $T = \sum_{i=1}^{n} k_i X_i$ where $X_1, \ldots, X_n$ are mutually independent RVs having respective mgfs $M_1(t), \ldots, M_n(t)$. The RV T has mgf*

$$M_T(t) = \prod M_i(k_i t).$$

Proof.

$$\begin{aligned} M_T(t) &= E(e^{tT}) = E(e^{t\sum k_i X_i}) = E(\prod e^{tk_i X_i}) \\ &= \prod E(e^{tk_i X_i}) = \prod M_i(k_i t) \end{aligned}$$

□

A vector of RVs is *independent identically distributed* or *iid* if the components are mutually independent and all have the same pdfs. An n-dimensional iid random vector of variables all having the same pdf as a RV X is an *random sample of distribution* X ; it has n *tests*, or n *samples*, or simply has *size* n. Often we implicitly assume n is the size of the sample.

Corollary 2.6.2. *If* $\mathbf{X}$ *is a random sample of distribution* X *then* $M_{\sum X_i}(t) = (M_{X_1}(t))^n$.

Problems from the Text

Section 2.6: 1,2(a),3

2.7 Transformations for more Variables

This is mostly the same as Section 2.2 so we skip it, except for noting what the jacobian looks like for more variables.

For n-dimensional random vectors $\mathbf{X}$ and $\mathbf{Y}$ a transformation $\boldsymbol{u} : \mathbf{X} \to \mathbf{Y}$ is described by $Y_i = u_i(X_1, \ldots X_n)$ for $i = 1, \ldots, n$ and its inverse is described by $X_i = w_i(Y_1, \ldots, Y_n)$.

The jacobian is the determinant

$$\begin{vmatrix} \frac{\partial w_1}{\partial y_1} & \cdots & \frac{\partial w_1}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial w_n}{\partial y_1} & \cdots & \frac{\partial w_n}{\partial y_n} \end{vmatrix}$$

2.8 Linear Combinations of Random Variables

Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors. By the linearity of expectation

$$E(\sum a_i X_i) = \sum a_i E(X_i).$$

Further

$$\begin{aligned} \operatorname{Cov}(\sum a_i X_i, \sum b_i Y_i) &= E\left((\sum a_i X_i - \sum a_i E(X_i))(\sum b_i Y_i - \sum b_i E(Y_i))\right) \\ &= E(\sum\sum a_i b_j X_i Y_j - \sum\sum a_i b_j E(X_i) Y_j + \ldots \\ &= \sum\sum a_i b_j E[X_i Y_j - X_i E(Y_j) - E(X_i) Y_j + E(X_i) E(Y_j)] \\ &= \sum\sum a_i b_j E[(X_i - E(X_i))(Y_j - E(Y_j))] \\ &= \sum\sum a_i b_j \operatorname{Cov}(X_i, Y_j) \end{aligned}$$

In the case that $\mathbf{X} = \mathbf{Y}$ this gives that

$$\operatorname{Var}(\sum a_i X_i) = \sum\sum a_i a_j \operatorname{Cov}(X_i, X_j) = \sum a_i^2 \operatorname{Var}(X_i)$$

where the last inequality uses that the non-diagonal terms are 0 by the independence of the variables.

If $\mathbf{X}$ is a random sample of a distribution X having mean μ and variance σ^2, then the *sample mean* is

$$\overline{X} = \frac{\sum X_i}{n}.$$

It has expected value

$$E(\overline{X}) = E\left(\frac{1}{n}\sum X_i\right) = \frac{nE(X)}{n} = E(X)$$

and variance

$$\operatorname{Var}(\overline{X}) = \frac{1}{n^2}\sum \operatorname{Var}(X_i) = \frac{\sigma^2}{n}.$$

The *sample variance* is

$$S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1} = \frac{\sum X_i^2 - n\overline{X}^2}{n-1}$$

We get the second equality above as follows.

$$\begin{aligned}\sum (X_i - \overline{X})^2 &= \sum (X_i^2 - 2X_i\overline{X} + \overline{X}^2) \\ &= \sum X_i^2 - 2\overline{X}\sum X_i + n\overline{X}^2 \\ &= \sum X_i^2 - 2n\overline{X}^2 + n\overline{X}^2 = \sum X_i^2 - n\overline{X}^2\end{aligned}$$

The sample variance is a random variable. It has expected value

$$\begin{aligned}E(S^2) &= \frac{1}{n-1}\left(\sum E(X_i^2) - nE(\overline{X}^2)\right) \\ &= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n(\mu^2 + \frac{\sigma^2}{n})\right) = \sigma^2\end{aligned}$$

Problems from the Text

Section 2.8: 2,3,10

3 Some Special Distributions

3.1 Binomial and Related Distributions

We have given a name to only one distribution so far: the uniform distribution whose pdf or pmf is constant on its support. There are several other distributions that occur repeatedly in mathematics and statistics. One of the most basic is the Binomial Distribution, which we build from the following distribution, which we will recognise as the distribution of outcomes when tossing a p-coin.

3.1.1 The Bernoulli Distribution

Bernoulli $X \sim b(1,p)$

$p_X(x)$	$\begin{cases} p & \text{if } x = 1 \\ 1-p & \text{otherwise.} \end{cases}$
μ	p
σ^2	pq
$M_X(t)$	$pe^t + q$

Definition 3.1.1. An RV X has a *Bernoulli distribution*, or is a *Bernoulli RV*, if its support is $\{0,1\}$. Its pmf is

$$f(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{otherwise,} \end{cases}$$

for some probability $p \in [0,1]$.

If X is a Bernoulli distribution with probability p, then its mean is $\mu = p$ and its variance is $\sigma^2 = p(1-p) = pq$.

Indeed, $\mu = p \cdot 1 + (p-1) \cdot 0 = p$, and $E(X^2) = p \cdot 1^2 + (p-1) \cdot 0^2 = p$, so $\sigma^2 = E(X^2) - \mu^2 = p - p^2 = p(1-p)$.

The Bernoulli distribution of probabilitly p is thus more often called the *Bernoulli distribution of mean p*.

3.1.2 The Binomial Distribution

Binomial $X \sim b(n,p)$

$p_X(x)$	$\binom{n}{x} p^x (1-p)^{n-x}$
μ	np
σ^2	npq
$M_X(t)$	$(pe^t + q)^n$

Often a probabilitly space consists of n independent Bernoulli spaces. When tossing 100 p-coins, the only random variable of any interest is that which counts the number of times the outcome is 'heads'. This is the Binomial distribution.

Definition 3.1.2. A random variable Y has the *Binomial distribution* $b(n,p)$ if

$$Y = \sum_{i=1}^{n} X_i$$

for a family $\{X_i\}_{i\in[n]}$ of iid Bernoulli RVs with mean p.

Clearly the pmf of $Y \sim b(n,p)$ is

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

on its support $y = 0, 1, \ldots, n$. By the linearity of expectation, its mean is

$$\mu = E(Y) = \sum_{i=1}^{n} E(X_i) = \sum p = np.$$

Problem 3.1.3. Show that the variance of $Y \sim b(n,p)$ is $\sigma^2 = np(1-p)$.

The following is a special case of the 'Law of Large Numbers' which is covered in Chapter 5 of the text.

Example 3.1.4. If $Y \sim b(n,p)$, then Y/n can be viewed as the 'rate of success' of the trials $X_1, \ldots, X_n$ making up Y. Clearly $E(Y/n) = E(Y)/n = np/n = p$, and one can show that $\mathrm{Var}(Y/n) = \frac{p}{n}(1-p)$. So by Chebyshev,

$$P(|Y/n - p| \geq \varepsilon) \leq \frac{\mathrm{Var}(Y/n)}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \to 0.$$

This means that the rate of success of the trials X_i is more and more concentrated around p as n gets bigger. The take away is that to get a good estimate of p, one can take the average of several samples of the distribution. The more samples we take, the more likely the estimate is close to the actual value.

The mgf of $X \sim b(n,p)$ is

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{n} e^{xt} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x} \\ &= (pe^t + q)^n. \end{aligned}$$

Problem 3.1.5. Use the mgf to find μ and σ^2 of $X \sim b(n,p)$.

Problem 3.1.6. Show that if $X_i \sim b(n_i, p)$ for $i = 1, \ldots, d$, and $X_1, X_2, \ldots, X_d$ are pairwise independent, then $Y = \sum_{i=1}^{d} X_i$ has distribution $b(\sum_{i=1}^{d} n_i, p)$.

We do not do much with the rest of the distributions in this section. We simply define them so that we have seen them.

3.1.3 The geometric and negative binomial distributions

For the binomial distribution $b(n,p)$ we conducted n independent Bernoulli trials with mean p and counted the number of successes. For the Geometic distribution Y, we conduct Bernoulli trials with mean p until there is a success. We let Y count the number of failures.

Formally, the *geometric RV with parameter* p is the RV with pmf

$$p(y) = (1-p)^y \cdot p.$$

More generally the *negative binomial RV with parameters* p *and* r is the RV that counts the number of failures that occur, when conducting Bernoulli trials $b(1,p)$, until the r^{th} success. It has pmf

$$p(y) = \binom{y+r-1}{r-1} p^r (1-p)^y.$$

3.1.4 The Hypergeometric Distribution

Hypergeometric

$p_X(x)$	$\frac{\binom{N-D}{n-x}\binom{D}{x}}{\binom{N}{n}}$
μ	$n\frac{D}{N}$
σ^2	$n\frac{D}{N}\frac{N-D}{N}\frac{N-n}{N-1}$

In a lot of N items, D are defective. We choose n items. The RV X that counts the number of chosen items that are defective is a *hypergeometric* distribution. Its pmf is

$$p(x) = \frac{\binom{N-D}{n-x}\binom{D}{x}}{\binom{N}{n}}.$$

The expected value of X is

$$E(X) = \sum_{x=0}^{n} xp(x) = \sum \binom{N-D}{n-x}\binom{D}{x}\binom{N}{n}^{-1} x.$$

Using that $b\binom{a}{b} = a\binom{a-1}{b-1}$ and so

$$\binom{a}{b} = \frac{a}{b}\binom{a-1}{b-1}$$

this becomes

$$\begin{aligned} E(X) &= \sum x \binom{(N-1)-(D-1)}{(n-1)-(x-1)} \binom{D-1}{x-1} \binom{N-1}{n-1}^{-1} \frac{D}{x}\frac{n}{N} \\ &= \frac{Dn}{N} \sum \binom{(N-1)-(D-1)}{(n-1)-(x-1)} \binom{D-1}{x-1} \binom{N-1}{n-1}^{-1} \\ &= \frac{Dn}{N} \sum p'(x) \end{aligned}$$

where $p'(x)$ is the pmf of a hypergeometric distribution with parameters $N-1, D-1$ and $n-1$. So $E(X) = \frac{Dn}{N}$.

Problem 3.1.7. Show that $\text{Var}(X) = n\frac{D}{N}\frac{N-D}{N}\frac{N-n}{N-1}$. (Hint: Compute $E(X(X-1))$ using the trick in the above calculation; use this to compute $E(X^2)$.)

Problems from the Text

Section 3.1: 3,4,5,6,11,14,15,18,23

3.2 The Poisson Distribution

Poisson $X \sim \text{pois}(\mu)$

$p_X(x)$	$\frac{e^{-\mu}\mu^x}{x!}$
μ	μ
σ^2	μ
$M_X(t)$	$e^{\mu(e^t-1)}$

A *Poisson process* is a random process with a recurring event such that

i) The probability of an occurence of the event in an interval of time is proportianal to the length of the interval.

ii) The probability of more than one occurence in a small interval is negligible.

iii) The probability of occurences in disjoint intervals is independent.

The random variable X that counts the number of occurences in an interval of length one has a *Poisson distribution*. The pmf of a poisson distribution is

$$p(x) = \frac{\mu^x e^{-\mu}}{x!}$$

for $x = 0, 1, 2, \ldots$, and for some $\mu > 0$. We write $X \sim \text{pois}(\mu)$.

First, let's verify that this is indeed a pmf. Putting $y = \mu$ in the taylor expansion $e^y = \sum_{x=0}^{\infty} \frac{y^x}{x!}$ of e^y about 0, we get

$$\sum_{x=0}^{\infty} p(x) = \sum \frac{\mu^x e^{-\mu}}{x!} = e^{-\mu} \sum \frac{\mu^x}{x!} = e^{-\mu} e^{\mu} = 1,$$

as needed.

Now let's verify that $p(x)$ arises from the given properties of a Poisson process.

Given a Poisson process, let $g_x(t)$ be the probabililty that there are x occurences in an interval of length t. So $p(x) = g_x(1)$. We find an expression for $g_x(t)$ using the above three properties, and then see that taking $t = 1$ we get $\frac{\mu^x e^{-\mu}}{x!}$ thus deriving $p(x)$.

The above three properties yield the following properties of $g_x(t)$.

i) $g_1(t) = \mu t$ for some $\mu > 0$.

ii) $g_0(h) + g_1(h) \to 1$ as $h \to 0$, and for $i \geq 2$, $g_i(h) \to 0$.

iii) $g_x(t+h) = \sum_{i=1}^{x} g_i(t) \cdot g_{x-i}(h) \to g_{x-1}(t)\mu h + g_x(t)(1 - \mu h)$ as $h \to 0$.

For iii) we use i) and ii) along with the third property of a poisson process.

From this we get

$$\begin{aligned} g_x(t)' &= \lim_{h \to 0} \frac{g_x(t+h) - g_x(t)}{h} \\ &= \lim_{h \to 0} \frac{g_{x-1}(t)\mu h + g_x(t)(1 - \mu h) - g_x(t)}{h} \\ &= \lim_{h \to 0} \frac{g_{x-1}(t)\mu h - g_x(t)\mu h}{h} \\ &= \mu g_{x-1}(t) - \mu g_x(t) \end{aligned}$$

When $x = 0$ this becomes the differential equation $g_0'(t) = -\mu g_0(t)$. As $(e^{-\mu t})' = -\mu e^{-\mu t}$, $g_0(t) = e^{-\mu t}$ is the solution with $g_0(0) = 1$.

Inductively, we can solve

$$g_x'(t) = \mu g_{x-1}(t) - \mu g_x(t) = \frac{\mu(\mu t)^{x-1} e^{-\mu t}}{(x-1)!} - \mu g_x(t)$$

to get $g_x(t) = \frac{(\mu t)^x e^{-\mu t}}{x!}$. Thus $p(x) = g_x(1)$, as required.

Now, the mgf of $X \sim \text{pois}(\mu)$ is

$$\begin{aligned} M(t) &= E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} \frac{\mu^x e^{-\mu}}{x!} \\ &= e^{-\mu} \sum_{x=0}^{\infty} \frac{(\mu e^t)^x}{x!} = e^{-\mu} e^{\mu e^t} = e^{\mu(e^t-1)} \end{aligned}$$

Problem 3.2.1. Show that $E(X) = M'(0) = \mu$, and that $\sigma^2 = M''(0) - \mu^2 = \mu$.

There is no nice closed form for the cdf X, but we can compute it easily enough for small values.

Example 3.2.2. Let $X \sim \text{pois}(2)$. Then

$$\begin{aligned} P(1 \le X) &= 1 - P(X = 0) = 1 - p_X(0) \\ &= 1 - 2^0 e^{-2}/1 = 1 - e^{-2} \approx .865 \end{aligned}$$

For common values of μ and small values of x, values of the cdf $F_X(x)$ of $X \sim \text{pois}(\mu)$ are listed in a table in the back of the text.

The time interval for a Poisson process is easily scaled: if one expects μ occurences in an hour, then one expects 24μ occurences in a day. The same process can be described by the RV $H \sim \text{pois}(\mu)$ counting occurences per hour, or by the RV $D \sim \text{pois}(24\mu)$ counting occurences per day.

Example 3.2.3. Let X, counting the number of fish bites in a 1 minute interval, be a Poisson process with mean .2 bites per minute. We want to calculate the probability that we get at least 1 bite in a 10 minute interval. The RV Y counting the number of bites in a 10 minute inverval is Poisson with mean $\mu = .2 * 10 = 2$. So by Example 3.2.2, the probability is $P(Y \ge 1) \approx .865$.

The following will also be useful.

Theorem 3.2.4. *If $X_1, \ldots, X_n$ are independent RVs with $X_i \sim \text{pois}(\mu_i)$, then*

$$Y = \sum X_i \sim \text{pois}(\sum \mu_i).$$

Proof.

$$M_Y(t) = \prod M_{X_i}(t) = \prod e^{\mu_i e^t - 1} = e^{(\sum \mu_i)(e^t - 1)}.$$

□

Problem 3.2.5. Find the probability P_n that a vertex v in $G_{n,p}$ has degree d. Show that (for fixed d) $P_n \to p_X(d)$ as $n \to \infty$ where $X \sim \text{pois}(np)$.

Problems from the Text

Section 3.2: 1,3,5,8,10,12

3.3 The exponential and related distributions

We define the exponential distribution in detail, and the more general Γ distribution in less detail. We then use the Γ distribution to define the χ^2 distribution,

3.3.1 The exponential distribution and waiting time

Exponential $X \sim \Gamma(1, \mu)$

$f_X(x)$	$e^{-x/\mu}/\mu$
μ	μ
σ^2	μ^2
$M_X(t)$	$(1-\mu t)^{-1}$

Given a poisson process, the time one waits until the first occurence of the process is the *waiting time*. What is the waiting time of a poisson process with mean τ?

Let X be the RV $X \sim \text{pois}(\tau)$ that count the occurences of the process in unit time. The exponential distribution $Y = \Gamma(1, \frac{1}{\tau})$ measures the waiting time of the process. This Y is continuous RV with support $(0, \infty)$. (The choice of notation for the distribution will make sense after a couple of calculations.)

To calculate its distribution functions, it is useful to consider a scaled poisson RV $X_I \sim \text{pois}(\tau I)$ which counts the number of occurences of the above poisson process in a time interval of length I. This has pmf

$$p_I(x) = \frac{(\tau I)^x e^{-\tau I}}{x!}.$$

Now, we have

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= \text{Prob(At least one occurence of the poisson process by time } y). \\ &= P(X_y \ge 1) \\ &= 1 - p_y(0) = 1 - \frac{(\tau y)^0 e^{-\tau y}}{0!} = 1 - e^{-\tau y} \end{aligned}$$

Differentiating the cdf, we get $f_Y(y) = \tau e^{-\tau y}$.

Now to find the mean μ of Y we compute

$$\mu = E(Y) = \int_0^\infty y\tau e^{-\tau y}\,\mathrm{dy}$$

Using the method of integration by parts with $u = y$ and $v' = \tau e^{-\tau y}$ we get that the antiderivative is

$$\int y\tau e^{-\tau y}\,\mathrm{dy} = -ye^{-\tau y} - \int -e^{-\tau y}\,\mathrm{dy} = -\frac{e^{-\tau y}}{y} - \frac{1}{\tau}e^{-\tau y}.$$

When $y = 0$ this is $-\frac{1}{\tau}$, and as $y \to \infty$ it goes to 0, so we get that $\mu = \frac{1}{\tau}$. Thus $Y = \Gamma(1, \mu)$, and in terms of μ the pdf is $f_Y(y) = \frac{e^{-y/\mu}}{\mu}$.

We now compute the moment generating function of $Y \sim \Gamma(1, \mu)$.

$$M_Y(t) = E(e^{ty}) = \int_0^\infty e^{ty} e^{-y/\mu}/\mu \, \mathrm{d}y = \int_0^\infty \frac{1}{\mu} e^{-\frac{y}{\mu}(1-\mu t)} \, \mathrm{d}y$$

Now letting $x = y(1-\mu t)$ we differentiate with respect to y to get the differential $\mathrm{d}x = (1 - \mu t) \, \mathrm{d}y$. When $y = 0$ we have $x = 0$ and when $y \to \infty$, taking t small enough that $0 < (1 - \mu t) < 1$ we have that $x \to \infty$, and so substitiuting in x the above integral is

$$\frac{1}{1-\mu t} \int_0^\infty \frac{1}{\mu} e^{-x/\mu} \, \mathrm{d}x = \frac{1}{1-\mu t} [-e^{-x/\mu}]_0^\infty = \frac{1}{1-\mu t}.$$

Recognising this as the geometric series

$$M_Y(t) = \frac{1}{1-\mu t} = 1 + \mu t + (\mu t)^2 + (\mu t)^3 + \ldots ...$$

you should easily be able to do the following.

Problem 3.3.1. Show that the varience σ^2 of $Y \sim \Gamma(1, \mu)$ is μ^2.

3.3.2 The Gamma distribution

Gamma $X \sim \Gamma(\alpha, \beta)$

$f_X(x)$	$\frac{x^{\alpha-1} e^{-(x/\beta)}}{\Gamma(\alpha)\beta^\alpha}$
μ	$\alpha\beta$
σ^2	$\alpha\beta^2$
$M_X(t)$	$(1-\beta t)^{-\alpha}$

It is non-trivial, but true, that the Γ function

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} \, \mathrm{d}y$$

exists for all $\alpha > 0$.

One can show that for integers $\alpha > 1$ that $\Gamma(\alpha) = (\alpha - 1)!$. So $\Gamma(2) = 1$ and $\Gamma(3) = 2 \cdot 1$.

Using the change of variables $y = x/\beta$ we get that $\mathrm{d}y = \frac{1}{\beta} \, \mathrm{d}x$, and so

$$\Gamma(\alpha) = \int_0^\infty \left(\frac{x}{\beta}\right)^{\alpha-1} \frac{e^{-(x/\beta)}}{\beta} \, \mathrm{d}x = \int_0^\infty \frac{x^{\alpha-1} e^{-(x/\beta)}}{\beta^\alpha} \, \mathrm{d}x.$$

Dividing through by $\Gamma(\alpha)$ we get that $f(x) = \frac{x^{\alpha-1} e^{-(x/\beta)}}{\Gamma(\alpha)\beta^\alpha}$ on $(0, \infty)$ is the pdf of a random variable X which we call the gamma distribution $\Gamma(\alpha, \beta)$.

Note that by taking $\alpha = 1$ we get the pdf of the exponential distribution $\Gamma(1, \beta)$. This explains this notation for the exponential distribution.

Problem 3.3.2. Show that the mgf of $X \sim \Gamma(\alpha, \beta)$ is $M_X(t) = (1 - \beta t)^{-\alpha}$. Hint: use the change of variable $y = \frac{x(1-t\beta)}{\beta}$ in the integral you get.

From this we get that $X \sim \Gamma(\alpha, \beta)$ has mean

$$\mu = M'_X(0) = \frac{\alpha\beta}{(1 - \beta \cdot 0)^{\alpha+1}} = \alpha\beta,$$

and variance

$$\sigma^2 = M''_X(0) - \mu^2 = \cdots = \alpha\beta^2.$$

The following additivity of Gamma distributions will be useful.

Problem 3.3.3. Let $X_i \sim \Gamma(\alpha_i, \beta)$ and $Y = \sum X_i$. Using the mgf, show that $Y \sim \Gamma(\sum \alpha_i, \beta)$.

3.3.3 The χ^2 distribution

The *chi-squared* distribution with r degrees of freedom is $\chi^2(r) = \Gamma(r/2, 2)$.

The following is immediate from Problem 3.3.3, and will be used in analysing ANOVA.

Problem 3.3.4. Show that if $X_i \sim \chi^2(r_i)$ for each i, and $Y = \sum X_i$, then $Y \sim \chi^2(\sum r_i)$.

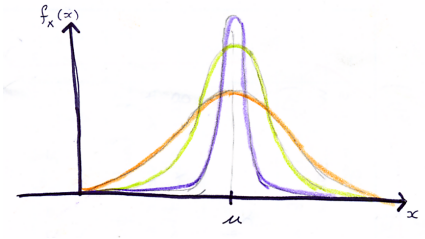
Problems from the Text

Section 3.3: 1,2, 6,9,15

3.4 The Normal distribution

Normal $X \sim N(\mu, \sigma^2)$

$f_X(x)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
$M_X(t)$	$e^{\mu t + \frac{1}{2}\sigma^2 t^2}$



The normal distribution will be our most important distribution for statistical inference: not all populations are normally distributed, but the means of large samples from any population are approximately normal. This is a very useful tool.

A continuous RV Z has the *standard normal distribution* $N(0,1)$ if its pdf is

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

To see that this is a pdf observe that

$$\begin{aligned}\left(\int_{\mathbb{R}} f_Z(z)\,\mathrm{d}z\right)^2 &= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{-z^2/2}e^{-y^2/2}\,\mathrm{d}y\,\mathrm{d}z = \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{-\left(\frac{z^2+y^2}{2}\right)}\,\mathrm{d}y\,\mathrm{d}z \\ &= \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} e^{-\left(\frac{r^2}{2}\right)} r\,\mathrm{d}r\,\mathrm{d}\theta = \frac{1}{2\pi}\int_0^{2\pi} 1\,\mathrm{d}\theta \\ &= 1\end{aligned}$$

In the second last line, use the substitutition $u = r^2$ to get that $\int_0^\infty e^{-\left(\frac{r^2}{2}\right)} r\,\mathrm{d}r = 1$. Since $f_Z(z)$ is positve, this give that $\int f_Z = 1$, as needed.

Problem 3.4.1. Show that $M_Z(t) = e^{t^2/2}$. The integral is easy once you notice that you can seperate the t from the z using the simple identity

$$tz - \frac{1}{2}z^2 = \frac{1}{2}(2zt - z^2) = \frac{1}{2}(t^2 - z^2 + 2zt - t^2) = \frac{1}{2}t^2 - \frac{1}{2}(z-t)^2.$$

It follows that Z has mean $\mu = 0$ and $E(Z^2) = 1$ so it has variance $\sigma^2 = 1$. Where $Z \sim N(0,1)$, the RV $X = \mu + \sigma Z$ has *normal distribution* $N(\mu, \sigma^2)$. Clearly $E(X) = \mu$ and

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

The pdf of X is

$$f_X(x) = f_Z\left(\frac{x-\mu}{\sigma}\right)\left|\frac{\mathrm{d}z}{\mathrm{d}x}\right| = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

and the mgf is

$$M_X(t) = E(e^{Xt}) = E(e^{(\mu+\sigma z)t}) = e^{\mu t} E(e^{t\sigma z}) = e^{\mu t + \frac{1}{2}(\sigma t)^2}.$$

There is no good closed form for the cdf of $N(\mu, \sigma^2)$. Again, if you find yourself without a good calculator, you can refer to the tables at the back of the book where the cdf of $Z \sim N(0,1)$ is computed. As the distribution is symmetric, it is only computed for $z > 0$. For the cdf of $N(\mu, \sigma^2)$, one can

simply transform to X. Where $X \sim N(\mu, \sigma^2)$, we have that $X = \mu + \sigma Z$, (and so $Z = \frac{X-\mu}{\sigma}$), so

$$F_X(x) = P(X < x) = P(Z < \frac{x-\mu}{\sigma}).$$

We can look this up as $\Phi(\frac{x-\mu}{\sigma})$, in the back of the book.

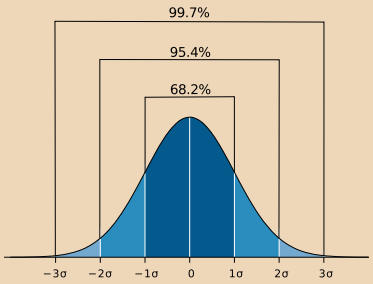
Example 3.4.2. Where $X \sim N(2, 16)$, we find $P(1 < x < 3)$ by

$$\begin{aligned} P(1 < X < 4) &= P(\frac{1-2}{4} < Z < \frac{4-2}{4}) = P(\frac{-1}{4} < Z < \frac{1}{2}) \\ &= \Phi(1/2) - \Phi(-1/4) = \Phi(1/2) - (1 - \Phi(1/4)) \\ &= \Phi(1/2) + \Phi(1/4) - 1 \approx .6915 + .5987 - 1 = .2902 \end{aligned}$$

Problem 3.4.3. The class grades on the next test will have distribution $X \sim N(73, 10)$. What is the probability that you will get an A (over %85)? What is the probabilitiy that you will fail (under %50)? Does it matter how many students are in the class? Does it matter if you do the homework?

Note

A nice quick description of the normal distribution is the rule of $68-95-99.7$ which says that 68% of the distribution is within σ (a standard deviation) of the mean, 95% is within 2σ and 99.7% of the distribution is within 3σ of the mean.



Theorem 3.4.4. *If $X_i = N(\mu_i, \sigma_i^2)$ are mutually independent then $Y = \sum a_i X_i$ is $N(\sum a_i \mu_i, \sum a_i^2 \sigma_i^2)$.*

Proof. Indeed, for each i we have that

$$a_i X_i = a_i(\mu_i + \sigma_i Z) = a_i \mu_i + a_i \sigma_i Z$$

is $N(a_i \mu_i, a_i^2 \sigma_i^2)$, so as mgfs are multiplicative for independent distributions, we have

$$M_Y(t) = \prod M_{a_i X_i}(t) = \prod e^{t a_i \mu_i + t^2 a_i^2 \mu_i^2/2} = e^{t \sum a_i \mu_i + t^2/2 \sum a_i^2 \mu_i^2}.$$

□

Corollary 3.4.5. *If $X_1, \ldots, X_n$ is a random sample of $X \sim N(\mu, \sigma^2)$ then $\overline{X} \sim N(\mu, \sigma^2/n)$.*

Problems from the Text

Section 3.4: 1,2,4,5,6,10,13,16,19,28

4 Statistical Inference

4.1 Sampling and statistics

The game in statistical inference is that we have some distribution X and by taking a sample from the distribution, we want to determine what X is.

For example, the height of people in a population has some distribution X. We measure 100 random people from the population, and from this decide whether X is $N(150, 10)$ or $\Gamma(160, 2)$ or something else.

At the end of this subsection, we outline how a histogram can be used to help us determine if X normal or gamma or from some other family, but usually, we will assume that we know what family of distributions X is from, and so only try to determine its parameters: μ and σ, or α and β, etc. A typical problem is as follows.

Example 4.1.1. The X-gene, (which, according to Wikipedia, allows that person to naturally develop superhuman powers and abilities), is known (suspected) to occur in any given person with probability p.

Our task is to find p.

To do this, we take a sample of n people, and test them for the X-gene.

Say we got the following 5 sample points:

$$x_1 = 1, \text{ and } x_2 = x_3 = x_4 = x_5 = 0$$

Here 1 denotes the presence of the gene, and 0 its absence. How should we interpret this data?

The sample yields us an estimate of the parameter p. Indeed, from the above data, most of us would estimate that $p = 1/5 = .2$. This is the sample mean. Usually when estimating the mean μ of a distribution, it is pretty clear that the sample mean is the best estimate. But for other parameters, the variance of a normal distribution, for example, it is not alway clear what the best estimate is. In later chapters, we will spend a lot of time trying to find 'best estimates' of parameters. In this chapter, we will just except the estimators we are given, and use them in two ways.

1) Although most of us would estimate that $p = .2$, we do not really believe this. It is unlikely. But perhaps we believe that p is p is in the interval $(.15, .25)$. This is more likely. How likely? This interval is a *confidence interval* for our parameter. We will calculate the probability that p is in this interval, conditioned on the fact that our sample yielded a mean of .2. This will be our confidence that p is in this interval.

2) In *hypothesis testing*, we make a hypothesis such as '$p < .18$' and then based

on our sample data, decide if the hypothesis is reasonable, or unreasonable. In the above test, the 5 data points don't give very strong evidence to dismiss the hypothesis $p < .18$. However, they would have given strong evidence to dismiss a hypothesis such as $p > .8$.

Our setup for most of the rest of the course is that X has a distribution $f_X(x;\theta)$ which depends on some unknown parameter, or vector of parameters, θ. We will use a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ to make inferences about θ.

For example, we might say that $X \sim b(1,p)$ for some unknown p. It has distribution $f_X(x;p) = p, (1-p)$ depending on whether x is 1 or 0. We might say that $X \sim N(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$. What is $f_X(x;(\mu,\sigma^2))$?

Definition 4.1.2. A function $T = T(X_1, \ldots, X_n)$ is called a *statistic* of the sample. If T is used to estimate θ, then T is a *point estimator* for θ. It is *unbiased* if

$$\theta = E(T(X_1, \ldots, X_n)).$$

Based on a sample $\mathbf{X}$, there can be many estimators for a given parameter. Generally we want an unbiased one, but there can be many of these too.

Example 4.1.3. All of $\overline{\mathbf{X}}, X_1$ and $\frac{X_1+2X_2}{2}$ can be estimators of the mean μ of X. Both $\overline{\mathbf{X}}$ and X_1 are unbiased, but

$$E(\frac{X_1+2X_2}{2}) = \frac{3}{2}\mu,$$

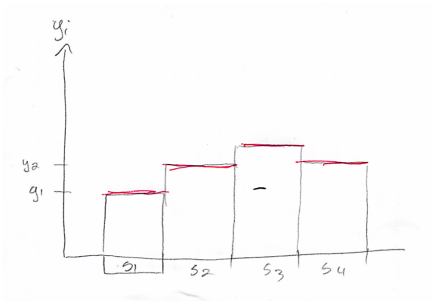
so this has bias.

Problem 4.1.4. Show that the sample variance $S^2 = \frac{\sum_{i=1}^{n}(X_i-\overline{\mathbf{X}})^2}{n-1}$ is an unbiased estimator of the variance σ^2.

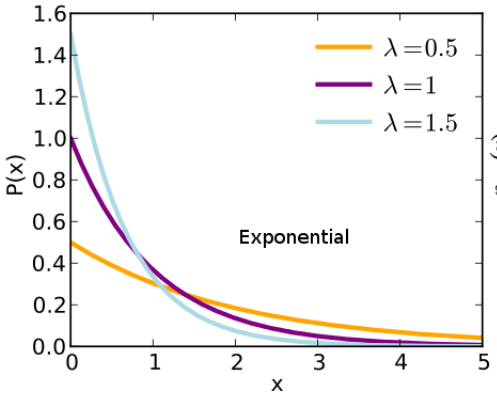
Both $\overline{\mathbf{X}}$ and X_1 are unbiased, but $\overline{\mathbf{X}}$ seems better as its variance is smaller. Is it the best estimator? This is a complicated question that we consider more closely in Chapters 6,7 and 8.

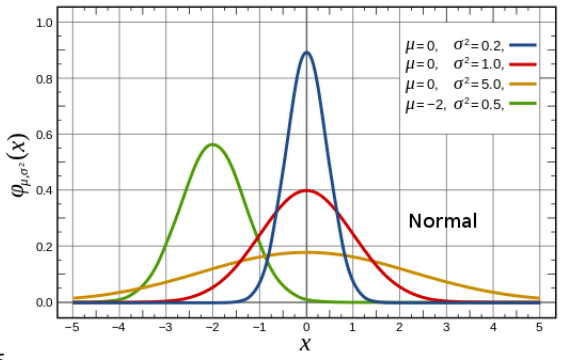
To finish off this section we offer a first practical solution to a practical problem of statistics that we mostly ignore for the rest of the course. (We will address it again in Section 4.5, about order statistics, but only there.) For most of our testing, we will assume that the population X fits in a parametrised family of distributions. However sometimes, we have no reason to assume that the distribution is normal, or exponential, or what have you. In this case, we can use a histogram as an unbiased estimator of the pdf f_X.

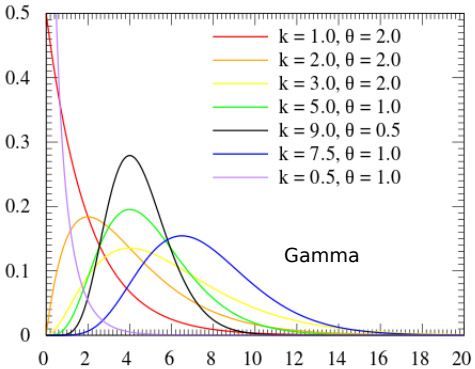
Example 4.1.5. Let $\boldsymbol{x}$ be a realisation of a random sample of X. For a partition of the sample space into (equal) parts $S_1, \ldots S_d$. Let y_i count the number of sample points in $\boldsymbol{x}$ that lie in the part S_i. A *histogram*, as below, is an approximation of f_x. It is an unbiased estimator of the corresponding discretisation of f_x.

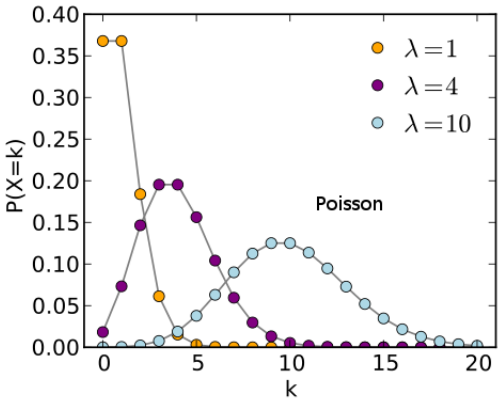


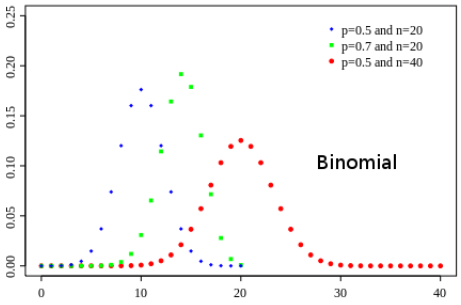
Here are pdf/pmfs of our common distributions. Which does the above most closely resemble? (All these pictures are from wikipedia.)

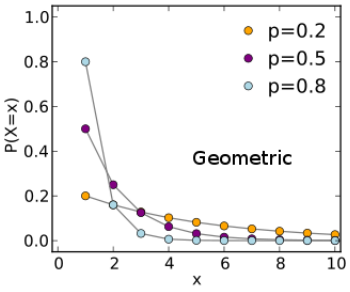












Problems from the Text

Section 4.1: 1 (a,c), 5

4.2 Confidence Intervals

(This is the material from Sections 4.2 and 4.3, mixed up a bit.)

For Example 4.1.1, instead of asserting that $p = .2$, which is almost surely wrong we want to say that we are %90 sure that $p \in [.1, .3]$.

Definition 4.2.1. Let $\mathbf{X}$ be a random sample of X and let $\theta_L = \theta_L(\mathbf{X})$ and $\theta_R = \theta_R(\mathbf{X})$ be statistics. The interval (θ_L, θ_R) is a $(1-\alpha)$ *confidence interval* for a parameter θ if

$$P(\theta_L < \theta < \theta_R) = 1 - \alpha.$$

As innocent as this definition looks, it is not so straightforward. The parameter θ is fixed, but we do not know it. It is not immediatly clear how to compute

$$P(\theta_L(\mathbf{X}) < \theta < \theta_R(\mathbf{X})),$$

let alone to find $\theta_L(\mathbf{X})$ and $\theta_R(\mathbf{X})$?

But $\theta_L(\mathbf{X})$ and $\theta_R(\mathbf{X})$ will, of course, depend on an estimator $T = T(\mathbf{X})$ of θ. It makes sense that $\theta_L(\mathbf{X}) < T(\mathbf{X}) < \theta_R(\mathbf{X})$. Now, let $\theta_L(\mathbf{X})$ be the smallest value of θ for which

$$F_T(T(\mathbf{X}); \theta) < 1 - \alpha/2.$$

So $\theta < \theta_L$ implies $F_T(T(\mathbf{X}); \theta) \geq 1 - \alpha/2$. Moreover, we have by definition that

$$P(F_T(T(\mathbf{X}); \theta) \geq 1 - \alpha/2) = \alpha/2;$$

so

$$P(\theta < \theta_L(\mathbf{X})) = \alpha/2.$$

Similarily, taking θ_R as the greatest value for which $F_T(T(\mathbf{X}); \theta) > \alpha/2$, we get

$$P(\theta_R(\mathbf{X}) < \theta) = \alpha/2$$

and so these values of θ_L and θ_R will do to defined a $(1-\alpha)$-confidence interval for θ.

This is not a full proof, but it can be made rigourous. Changing some inequalities to equalities, and so making a compensating small adjustment for discrete distributions we record it as the following.

Theorem 4.2.2. *Let $\mathbf{X}$ be a random sample of X having distribution $f_X(x; \theta)$ for some θ, and let $T = T(\mathbf{X})$ be an estimator of θ. Let $T^-(\mathbf{X})$ be the greatest value in the support of T less than $T(\mathbf{X})$. Where θ_L is the minimum value of θ for which $F_T(T^-(\mathbf{X}); \theta) = 1 - \alpha/2$ and θ_R is the greatest value of θ for which $F_T(T(\mathbf{X}); \theta) = \alpha/2$, the interval (θ_L, θ_R) is a $(1-\alpha)$-confidence interval for θ.*

Notice for an $(1-\alpha)$ confidence interval one need not split the α into $\alpha/2$ and $\alpha/2$. For symmetric distributions this is a good choice, but for other distributions this is not always efficient (meaning that it is the shortest $(1-\alpha)$

confidence interval). This is covered in later chapters, but we will (mostly?) ignore it, and split α into $\alpha/2$ and $\alpha/2$.

This is a general result but depending on the family of distributions that X belongs to, the way we compute θ_L and θ_R can vary.

4.2.1 Confidence Intervals for Discrete Distributions

We start with an example for a discrete distribution.

Example 4.2.3. Let $\mathbf{X}$ be a 30 point random sample of $X \sim b(1, p)$, and use the sample mean $\overline{\mathbf{X}}$ to estimate p. Say we get a realisation

$$\overline{\boldsymbol{x}} = .6.$$

To get a 90%-confidence interval for p, we compute values we need to compute values p_L and p_R such that

$$P(p_L < p < p_R) = .90.$$

Using Theorem 4.2.2 we can take p_L as the least p such that $F_{\overline{\mathbf{X}}}(T^-; p) = .05$. What is T^-? Well, T is our estimator $\overline{\mathbf{X}} = .6 = 18/30$; so the greatest value less than this in our support is $T^- = 17/30$. So

$$F_{\overline{\mathbf{X}}}(T^-; p) = \sum_{i=0}^{17} \binom{30}{i} p^i (1-p)^{30-i}.$$

This, as a function of p is a bit hard to invert, so we break out a computer and use a low machinery attack. Plugging in $p = .4$ we get a value of 0.978. This is to high, so we try $p = .45$, getting 0.92. We try $p_L = .425$, etc. We settle on $p_L = .434$, for which we get .95, as needed. Similarily $p_R \approx .75$ gives that

$$.05 = \sum_{i=0}^{18} \binom{30}{i} p_R^i (1-p_R)^{30-i}.$$

So $(.434, .75)$ is a %90 confidence interval for p.

Notice that our estimator $T = \overline{\mathbf{X}}$ is a simple transformation $T = S/30$ of the RV $S = \sum X_i$, and in computing $F_{\overline{\mathbf{X}}}(T^-; p)$, we sum values of $p_{\overline{\mathbf{X}}}$ that we get from values of p_S via this simple transformation. Our calculation found a confidence interval $(30(.434), 30(.75))$ for the mean $30p$. This often happens for a mean, and sometimes it is easier to think about it by making this explicit. We do this in the following example.

Example 4.2.4. Let $\mathbf{X}$ be a 20 point random sample of $X \sim \text{pois}(\mu)$. Assuming a realisation of $\overline{\mathbf{X}}$ yields a point estimate 10 of μ, find a %95 confidence interval for μ.

Instead of $\overline{\mathbf{X}}$ we use the estimator $T = \sum X_i = 20\overline{\mathbf{X}}$ of the mean $\theta = 20\mu$ of the distribution $T \sim \text{pois}(\theta)$. Our point estimate is $t = 200$ so the value t^- of T^- we are using in Theorem 4.2.2 is 199. We compute θ_L such that

$$.975 = \sum_{y=0}^{199} p_T(y;\theta_L) = \sum_{y=0}^{199} e^{-\theta_L}\frac{(\theta_L)^y}{y!},$$

and θ_R such that

$$.025 = \sum_{y=0}^{200} p_T(y;\theta_L) = \sum_{y=0}^{200} e^{-\theta_L}\frac{(\theta_L)^y}{y!}.$$

Problems from the Text

Section 4.3: 3 (The problem is not clear on this fact, but they are asking for a symmetric %20-confidence interval.)

We will come back to the discrete case. This nonsense about T^- can often be avoid using the CLT. For this we first look at the continuous case of confidence intervals.

4.2.2 Confidence intervals for μ

The first continuous confidence interval we look at is one for the mean μ of a normal distribution $X \sim N(\mu, \sigma^2)$.

As the distribution is symmetric around μ, one can get a 90%-confidence interval for μ by finding a such that $P(\overline{\mathbf{X}} - a \le \mu \le \overline{\mathbf{X}} + a) = .90$. This is a such that

$$P(-a \cdot \sqrt{n}/\sigma \le \frac{\overline{\mathbf{X}} - \mu}{\sigma/\sqrt{n}} \le a \cdot \sqrt{n}/\sigma).$$

As $\frac{\overline{\mathbf{X}}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, we could look this up in our tables, if we knew σ^2. But we don't know it, so the best we can do is estimate it, which we do with

$$S^2 = \frac{\sum(\overline{\mathbf{X}} - X_i)^2}{n-1}.$$

We must then find a such that

$$P(-a \cdot \sqrt{n}/S \le \frac{\overline{\mathbf{X}} - \mu}{S/\sqrt{n}} \le a \cdot \sqrt{n}/S).$$

Assuming that $T_{n-1} = \frac{\overline{\mathbf{X}}-\mu}{S/\sqrt{n}}$ is $N(0,1)$ is not so bad, (it will give you close to a $(1-\alpha)$-confidence interval) but it isn't exact, and so we get better estimates

by calculating its cdf exactly. This is done in tables in the back of the book. T_{n-1} is called the *Student's T distribution with $n-1$ degrees of freedom.*

The t-value $t_{.05} = t_{.05,19}$, which we can look up in the back of the book, is the value such that

$$P(T_{19} > t_{.05}) = .05.$$

(Note that the table is indexed by the degree of freedom $r = n-1$ and the value So

$$(\overline{\mathbf{X}} - t_{.05}(S/\sqrt{n}), \overline{\mathbf{X}} + t_{.05}(S/\sqrt{n}))$$

is a %90 confidence interval for μ.

Example 4.2.5. The height of people in a population is $X \sim N(\mu, \sigma^2)$ for some μ and σ^2. We sample 20 people, and get that $\overline{\mathbf{X}} = 168cm$ and that $S^2 = 40$. Let's find a 95% confidence interval for μ.

Checking that $t_{.025} = t_{.025,19} = 2.093$, the %95 confidence interval for μ is

$$(168 - 2.093(40/20)^{1/2}, 168 + 2.093(40/20)^{1/2}).$$

Now this is nice if X is normal, but what if it is something else? Well, we pretend that it is normal.

We will prove the following in Chapter 5.

Theorem 4.2.6 (Central Limit Theorem (CLT))**.** *Let $X_1, \ldots, X_n$ be a random sample of any distribution X with mean μ and variance σ^2. The cdf F_{W_n} of*

$$W_n = \frac{\overline{\mathbf{X}} - \mu}{\sigma/\sqrt{n}}$$

limits pointwise to the cdf Φ of $N(0,1)$ as n goes to ∞.

In Chapter 5, the above Theorem will be the statement that W_n *converges in distribution* to $N(0,1)$.

It will follow from the ideas in Chapter 5 that the same also holds for

$$T_{n-1} = \frac{\overline{\mathbf{X}} - \mu}{S/\sqrt{n}}.$$

where we replace σ^2 with the sample variance S^2. In fact, we will be able to replace σ with any *consistant* estimator of σ. We will use this fact (with comment but without proof) in this chapter, but will define and prove it in Chapter 5.

With this theorem, we can get a pretty nice approximate confidence interval for the mean of any distribution. Where $z_{\alpha/2}$ is the value such that

$$1 - \alpha/2 = \Phi(z_{\alpha/2}) = P(N(0,1) < z_{\alpha/2}),$$

we have that

$$\begin{aligned} 1-\alpha &\approx P(-z_{\alpha/2} < T < z_{\alpha/2}) \\ &= P(\overline{\mathbf{X}} - z_{\alpha/2}(S/\sqrt{n}) < \mu < \overline{\mathbf{X}} + z_{\alpha/2}(S/\sqrt{n})) \end{aligned}$$

gives an approximate $1-\alpha$ confidence interval for μ. As n gets bigger, the approximation gets better. This is called a *large sample confidence interval*.

Traditionally, one would use a t-value when the sample is size 30 or less, and the variance is unknown. If the sample size is larger than 30 or the variance is known, one uses the z-value. For a Bernoulli or Poisson distribution, the variance is a simple function of the mean ($\sigma^2 = p(1-p)$ or $\sigma^2 - \mu$ respectively) so if one estimates p with $\overline{x}$ it is reasonable to assume that $\sigma^2 = \overline{x}(1-\overline{x})$. In this case one uses the z-value in place of the t-value.

Problem 4.2.7. Using the CLT to approximate $b(30, p)$ with a normal distribution, find an approximate %90 confidence interval for an estimate $\overline{x} = .6$ of p. Compare with Example 4.2.3.

4.2.3 Other Confidence Intervals

In the following example we want to address the difference between two random variables.

Example 4.2.8. Let X and Y, having means μ_X and μ_Y respectivley, measure the occurrence of cancer among people taking drugs 1 and 2 respectively. To show that drug 1 is effective, we want to show that the difference of means

$$\Delta = \mu_X - \mu_Y$$

is positive, or large.

Taking 10 point samples $\mathbf{X} = (X_1, \ldots, X_{10})$ and $\mathbf{Y} = (Y_1, \ldots, Y_{10})$, of the two populations, we get that $\overline{\mathbf{X}} = 4.2$, $S_X^2 = 49$, $\overline{\mathbf{Y}} = 3.4$, and $S_Y^2 = 32$. It seems clear that $\overline{\Delta} = \overline{\mathbf{X}} - \overline{\mathbf{Y}} = 4.2 - 3.4 = .8$ is an estimator for Δ. Indeed, by linearity of expectation, we have that $E(\overline{\Delta}) = E(\overline{\mathbf{X}}) - E(\overline{\mathbf{Y}}) = \mu_x - \mu_y = \Delta$, so it is an unbiased estimator. But how do we get a %90 confidence interval for Δ?

Letting $\Delta_i = X_i - Y_i$ for each i we view the random vector $(\Delta_1, \ldots, \Delta_{10})$ as a 10 point sample of the distribution $X - Y$. In Exercise 1.9.2 we showed that this has variance $\sigma^2 = \sigma_X^2 + \sigma_Y^2$.

By the CLT we have that

$$W = \frac{\overline{\Delta} - \Delta}{\sqrt{(\sigma_X^2 + \sigma_Y^2)/10}} \to N(0,1).$$

If we could show that the sample variance S^2 of our sample points Δ_i was equal to $S_X^2 + S_Y^2$ then we could apply the CLT directly. However this is not true.

Problem 4.2.9. Show that S^2 is not generally equal to $S_X^2 + S_Y^2$.

We will be able to show, however, that S_X^2 and S_Y^2 are consistant estimators of σ_X^2 and σ_Y^2, and so that $\sqrt{S_X^2 + S_Y^2}$ is a consistant estimator of σ, so by the discussion following the CLT we get that

$$\frac{\overline{\Delta} - \Delta}{\sqrt{(S_X^2 + S_Y^2)/10}} \to N(0,1).$$

This allows us to compute an approximate %90 confidence interval for $\mu_X - \mu_Y$:

$$\left[\overline{\Delta} - z_{.05}\sqrt{(S_X^2 + S_Y^2)/10}\,,\, \overline{\Delta} + z_{.05}\sqrt{(S_X^2 + S_Y^2)/10}\right]$$

or approximately $[.8 - 4.68, .8 + 4.68] = [-3.88, 5.48]$

Though the proof is different when $\mathbf{X}$ and $\mathbf{Y}$ have different lengths, the result is essentially the same.

Example 4.2.10. Let $\mathbf{X}$ be a 10 point sample of $X \sim N(\mu_X, \sigma_X^2)$ and $\mathbf{Y}$ be a 7 point sample of $Y \sim N(\mu_Y, \sigma_Y^2)$. Assume that $\overline{\mathbf{X}} = 4.2$, $S_X^2 = 49$, $\overline{\mathbf{Y}} = 3.4$, and $S_Y^2 = 37$.

A %90 confidence interval for $\Delta = \mu_X - \mu_Y$ is

$$[.8 - z_{.05}\sqrt{49/10 + 37/7}, .8 + z_{.05}\sqrt{49/10 + 37/7}] \approx [-4.24, 5.84].$$

Compare this with Example 4.2.4 of the text (and the discussion before it) where they assume that $\sigma_X^2 = \sigma_Y^2 =: \sigma^2$. In doing this, the get that

$$\frac{\overline{\Delta} - (\mu_X - \mu_Y)}{\sigma\sqrt{1/10 + 1/7}} \sim N(0,1)$$

exactly, without using the CLT. Using a *pooled estimator* $S_p^2 = \frac{(10-1)S_X^2 + (7-1)S_Y^2}{10-1+7-1}$ of the variance, then then go on to show (with some work) that

$$T = \frac{\overline{\Delta} - (\mu_X - \mu_Y)}{S_p\sqrt{1/10 + 1/7}}$$

is a t-distribution with $n-2$ degrees of freedom. They use this to get a confidence interval $[-4.81, 6.41]$ that is slightly bigger than ours. Theirs, for the assumption $\sigma_X^2 = \sigma_Y^2$, is better though, as it is exact, whereas ours is just approximate. And indeed, ours is pretty far off. Really we shouldn't be using the CLT for n as small as 17.

Problems from the Text

Section 4.2: 1,2,3,5,6,8,10,12,17,21,22

The last discussion above will help with question 12.

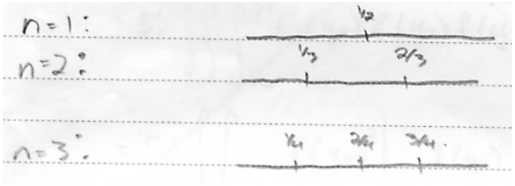
Section 4.2.2. of the text will help with question 22.

4.4 Order Statistics

We have seen how to estimate μ and σ^2 from a sample $\mathbf{X}$ of a distribution X. When X is normal or poisson or binomial, this is all the information we need. But as we mentioned before, we might not always know what kind of family of distributions X belongs to. In this case, we might like some more information from our sample.

Definition 4.4.1. Let $X_1, \ldots, X_n$ be a random sample of a continuous random variable X. Where $\{Y_1, \ldots, Y_n\} = \{X_1, \ldots, X_n\}$ and $Y_1 < Y_2 < \cdots < Y_n$, the parameters $Y_1, \ldots Y_n$ are the *order statistics* of $X_1, \ldots, X_n$.

Example 4.4.2. If X has a pdf $f_X(x) = 1$ on $[0, 1]$ then it is reasonable to assume the order statistics of a sample will tend to spread out evenly, falling something like:



More generally we expect $Y_1, Y_2, \ldots, Y_n$ to fall at around $F^{-1}(1/(n+1)), F^{-1}(2/(n+1)), \ldots, F^{-1}(n/(n+1))$. Indeed this is true. Let's say this with some notation.

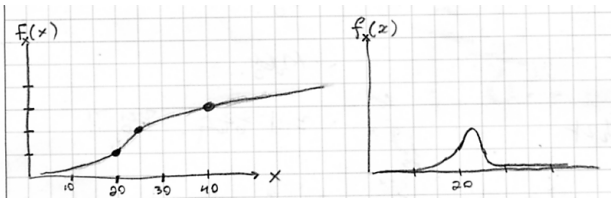
Definition 4.4.3. The p^{th} *quantile* or $100p^{th}$ *percentile* of a distribution is the value ξ_p such that

$$P(X < \xi_p) = p.$$

That is, if F is the cdf of X, then $\xi_p = F^{-1}(p)$.

Knowing several quantiles can give us a good feel for the distribution.

Example 4.4.4. If we know $(\xi_{.25}, \xi_{.50}, \xi_{75}) = (20, 25, 40)$ for a distribution X then we can sketch the cdf and pdf maybe as follows.



With the notion of a quantile defined, the statement that we expect Y_i to be $F_X^{-1}(i/n+1)$ becomes the precise statement

$$E(Y_i) = \xi_{\frac{1}{n+1}}$$

which requires proof.

To show this, we should first observe the following.

Fact 4.4.5. *Where Y_i is the i^{th} order statistic of an n-point random sample of X, the pdf (pmf) of Y_I is*

$$f_{Y_i}(y) = \frac{F_X^{i-1}(y) f_X(y)(1-F_X(y))^{n-i} n!}{(n-i)!(i-1)!}.$$

Proof. For this to occur in a sample $(X_1, \ldots, X_n)$ we need that $i-1$ of the sample points are less than y, one is equal to it, and $n-i$ are greater. There are $\binom{n}{i} \cdot i = \frac{n!}{(n-i)!(i-1)!}$ ways to choose the i points that will be less than or equal to y, and the one that will equal y; and the probablility for such a tuple that it will satisfy the required conditions is

$$F_X^{i-1}(y) f_X(y)(1-F_X(y))^{n-i}.$$

So the probability $f_{Y_i}(y)$ that a random sample orders to order statistics satisfying thess conditions is the product of these numbers, as needed. □

We show this, skipping some of the messier details.

Problem 4.4.6. Observe that whereas the joint pdf of $X_1, \ldots, X_n$ is

$$f_{\mathbf{X}}(\boldsymbol{x}) = \prod f_X(x_i)$$

the joint pdf of $Y_1, \ldots, Y_n$ is

$$f_{\mathbf{Y}}(\boldsymbol{y}) = g(y_1, \ldots, y_n) = \begin{cases} n! \prod f_X(y_i) & \text{if } y_1 < y_2 < \cdots < y_n \\ 0 & \text{otherwise.} \end{cases}$$

Re-derive the above formula for $f_{Y_i}(y)$ by finding the appropiate marginal distribuition of $f_{\mathbf{Y}}(\boldsymbol{y})$:

$$g_{Y_i}(y_i) = \int_{y_{n-1}}^{\infty} \int_{y_{n-2}}^{\infty} \cdots \int_{y_i}^{\infty} \int_{-\infty}^{y_i} \cdots \int_{-\infty}^{y_3} \int_{-\infty}^{y_2} n! \prod f_X(y_i) \, \widetilde{\mathrm{d}\boldsymbol{y}}$$

where $\widetilde{\mathrm{d}\boldsymbol{y}}$ stands for $\mathrm{d}y_1 \, \mathrm{d}y_2 \ldots \mathrm{d}y_{i-1} \, \mathrm{d}y_{i+1} \ldots \mathrm{d}y_n$. Hint: first observe that by the chain rule $\int F(x)^{\alpha-1} f(x) \, \mathrm{d}x = F(x)^{\alpha}/\alpha$, and $\int (1-F(x))^{\alpha-1} f(x) \, \mathrm{d}x = (1-F(x))^{\alpha}/\alpha$.

First we show that this, we need the pdf of Y_i, which we can get as a marginal pdf from the joint pdf of $\mathbf{Y}$.

Whereas the joint pdf of $X_1, \ldots, X_n$ is

$$f_{\mathbf{X}}(\boldsymbol{x}) = \prod f_X(x_i)$$

the joint pdf of $Y_1, \ldots, Y_n$ is

$$f_{\mathbf{Y}}(\boldsymbol{y}) = f_{\mathbf{Y}}(y_1, \ldots, y_n) = \begin{cases} n! \prod f_X(y_i) & \text{if } y_1 < y_2 < \cdots < y_n \\ 0 & \text{otherwise.} \end{cases}$$

Now we can show that the marginal pdf of Y_i is

$$f_{Y_i}(y_i) = \frac{n!}{(i-1)!(n-i)!} F_X(y_i)^{i-1} \cdot (1 - F_X(y_i))^{n-i} f_X(y_i).$$

We do just to a representive example.

Example 4.4.7. The marginal pdf of Y_3, when $n = 4$ is (where we write F and f for F_X and f_X,)

$$\begin{aligned} f_{Y_3}(y_3) &= 4! \int_{y_3}^{\infty} \int_{-\infty}^{y_3} \int_{-\infty}^{y_2} f(y_1) f(y_2) f(y_3) f(y_4) \, \mathrm{d}y_1 \, \mathrm{d}y_2 \, \mathrm{d}y_4 \\ &= 4! \int \int_{-\infty}^{y_3} (F(y_2) - 0) f(y_2) f(y_3) f(y_4) \, \mathrm{d}y_2 \, \mathrm{d}y_4 \\ &= 4! \int_{y_3}^{\infty} \frac{1}{2} (F^2(y_3) - 0) f(y_3) f(y_4) \, \mathrm{d}y_4 \\ &= \frac{4!}{2} F^2(y_3) f(y_3) \int_{y_3}^{\infty} f(y_4) \, \mathrm{d}y_4 \\ &= \frac{4!}{2} F^2(y_3)(1 - F(y_3)) f(y_3) \end{aligned}$$

Now to show that $E(Y_i) = \xi_{i/n+1}$ we should compute $\int_{-\infty}^{\infty} y f_{Y_i}(y) \, \mathrm{d}y$. The text suggests rather showing

$$E(F_X(Y_i)) = \int_{\mathbb{R}} F_X(y_i) g(y_i) \, \mathrm{d}y_i = \frac{i}{n+1} = E(\xi_{i/n+1}).$$

From which it follows that $E(Y_i) = \xi_{i/n+1}$. But even with this trick, the non-trivial middle equality here takes a fair bit of work, and so skip it.

Having estimators for the quantiles, and a distribution of these estimators, a sample now gives us a confidence intervals for the quantiles.

Example 4.4.8 (Confidence interval for quantiles)**.** Let $\mathbf{Y}$ be the order statistics of an n-point sample of X. For $i < j$, the value $P(Y_i < \xi_p < Y_j)$ is the probability that i to $j-1$ of the n trial observations are less than ξ_p. As the probability that any given trial is less than ξ_p is p, we have that

$$P(Y_i < \xi_p < Y_j) = \sum_{w=i}^{j-1} \binom{n}{w} p^w (1-p)^{n-w} =: (1 - \alpha).$$

Thus (Y_i, Y_j) is an $(1-\alpha)$-confidence interval for the p^{th} quantile.

Problem 4.4.9. Taking $n = 30$, $i = 12$ and $j = 18$ in the above example compute, (or approximate using the CLT) the confidence of the confidence interval $[Y_{12}, Y_{18}]$ for Q_2.

As they estimate the quantiles, the order statistics give us a good picture of the distribution X. Indeed, they can be used as a pretty good test to see if X is normal.

Example 4.4.10 (Quantile-quantile plot)**.** Let $Y_1, \ldots, Y_n$ be the order statistics of a sample $\mathbf{X}$, and let $\xi_{1/(n+1)}, \ldots, \xi_{n/(n+1)}$ be quantiles of the standard normal distribution Z. Plot the points $(Y_i, \xi_{i/(n+1)})$. If the plot is linear, it suggests that X is normal.

But the set of all the order statistics is a lot of data, the goal in statistics is to give some nice summarising measures.

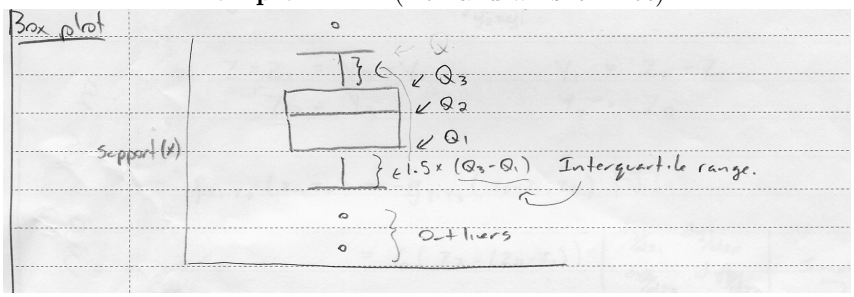
A couple of more concise statistics of a sample are the following.

Definition 4.4.11. Given the order statistics $\mathbf{Y}$ of an n point random sample $\mathbf{X}$, the statistic

- $Q_1 = Y_{(n+1)/4}$, $Q_2 = Y_{2(n+1)/4}$, and $Q_3 = Y_{3(n+1)/4}$ are the quartiles of the sample;
- $Y_n - Y_1$ is the *range*; and
- $(Y_1 + Y_n)/2$ is the *midrange*.

These suggest another (further than the histogram) graphical representation of sample data.

Example 4.4.12 (Box and whisker Plot)**.**



As we know the distributions of the order statistics, we can get the pdf of these secondary order statistics as well.

Example 4.4.13. Let Y_1, Y_2, Y_3 be the order statistics of a 3-point sample of $X \sim \text{Unif}([0, 1])$. Let $Z = Y_3 - Y_1$ be the range of the sample. To find the

pdf of Z we first find the joint pdf of Y_1 and Y_3, and then use the method of transformations.

(Before we continue, what do you think the expected value $E(Z)$ is?)

First, recalling that for $X \sim \text{Unif}([0,1])$ we have $f(x) = 1$ for all x in the support, we compute

$$f_{Y_1,Y_3}(y_1, y_3) = \int_{y_1}^{y_3} 3! f(y_1) f(y_2) f(y_3)\, \mathrm{d}y_2 = \int_{y_1}^{y_3} 3!\, \mathrm{d}y_2 = 6(y_3 - y_1).$$

Letting $Z_1 = Z = Y_3 - Y_1$ and $Z_2 = Y_3$ we get a one-to-one transformation with inverse $Y_1 = Z_2 - Z_1$ and $Y_3 = Z_2$, for $0 < Z_1 < Z_2 < 1$. So

$$\begin{aligned} f_{Z_1,Z_2}(z_1, z_2) &= f_{Y_1,Y_3}(z_2 - z_1, z_2)|J| \\ &= 6(z_2 - (z_2 - z_1)) \begin{vmatrix} \frac{\partial y_1}{\partial z_1} & \frac{\partial y_1}{\partial z_2} \\ \frac{\partial y_2}{\partial z_1} & \frac{\partial y_2}{\partial z_2} \end{vmatrix} \\ &= 6z_1 \begin{vmatrix} -1 & 1 \\ 0 & 1 \end{vmatrix} = 6z_1 \end{aligned}$$

Taking a marginal pdf, we get

$$f_{Z_1}(z_1) = \int_{z_1}^{1} 6z_1\, \mathrm{d}z_2 = 6z_1(1 - z_1).$$

Problem 4.4.14. Find the expected value $E(Z)$ of the range Z in the above example. Does this agree with your intuition?

Problems from the Text

Section 4.4: 3,5,8,11,13,24,27

4.5 Introduction to Hypothesis Testing

In a *hypothesis test*, we use a random sample $\mathbf{X}$ of an RV X to decide the truth of a hypothesis about a parameter θ of the distribution of X.

Example 4.5.1. Standard chickens produce an average of 6 eggs a week. I have modified chickens, and I want to determine if the produce more eggs than standard chickens. Where X is the distribution of 'eggs per week' of a modified chicken, I take a 5-point sample $\mathbf{X}$ of X to test the alternate hypothesis that $\mu_X > 6$ against the null hypothesis that $\mu_X = 6$.

If I get an outcome of $\overline{\boldsymbol{x}} = 6.7$, I have some evidence to accept the alternate hypothesis. Is this evidence significant enough? Should I have taken a bigger sample to make it more significant? Is this difference of .7 eggs per day big enough that I care about it?

Note

'Hypotheses' is plural, 'hypothesis' is singular.

We introduce the notation to address such questions. Let θ be an unknown parameter of a distribution X. A *(simple) hypothesis test* consists of a *null hypothesis*

$$H_0 : \theta = \theta_0$$

and an *alternate hypothesis*

$$H_1 : \theta \geq \theta_0 \qquad (\text{ or } H_1 : \theta \leq \theta_0);$$

Note

Sometimes 'rejecting H_1' is called 'accepting H_0', but some people feel this is inaccurate. We will see why in a bit.

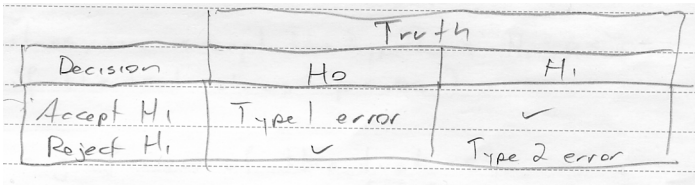
and a random sample $\mathbf{X}$ of X based on which we either accept H_1 or reject it. These alternate hypotheses are for what is called a *one-sided test*, in the next section we will look at a two-sided test with a alternate hypothesis $H_1 : \theta \neq \theta_0$. Which alternate hypothesis we use depends on what we expect to happen, and what we want to show. (In later chapters we will consider a more general hypothesis test with hypotheses $H_i : \theta \in \omega_i$ for some partition $\omega_0 \cup \omega_1$ of the support Ω of θ.)

The null hypothesis is so named because it is usually used for the case that something we are testing has had no effect.

The *critical region*, C, is the set of values of our sample $\mathbf{X}$ under which we accept H_1. For our sample $\mathbf{X}$ of the egg yield we will estimate μ with $\overline{\mathbf{X}}$ so our critical region be the set of values of $\overline{\mathbf{X}}$ under which we would conclude that H_1 is true. It will be something like

$$C = (7, \infty).$$

One might think that it should be $C = (6, \infty)$, but in this case our test will often give false results. There are four possible outcomes of a test.



The *size* or *significance* of C (or of the test) is probability of *Type 1 error* or a *false positive*:

$$\alpha = P_{\theta_0}(\mathbf{X} \in C).$$

The *power* γ of the test is the probability of a *correct positive*. It is a function of a specific value of θ in C:

$$\gamma(\theta) = P_{\theta}(\mathbf{X} \in C).$$

So $1 - \gamma(\theta)$ is the probability of a *Type II error* or a *false negative* when the actual value of the parameter θ is θ.

We want a test with low significance and high power. But this is a trade off. If $C = (6, \infty)$ in the previous example, then when $\mu = 6$ our sample will be in C half the time. Our significance is .50, which is no good at all. So to have a decent significance, our critical region must not contains values close to 6. So we say $C = (7, \infty)$. Our signficance is better, but then for a value 7 we have power $\gamma(7) = .50$. Maybe we will only have decent power for values of $\mu > 8$. When defining our critical region, we have to first decide what values of μ we care to prove the null hypothesis for.

Let's do a numerical example. We us a discrete distribution and calculate the significance and power of various critical regions.

Example 4.5.2. Assume that $p_0 = .05$ of all people exposed to coronavirus will contract Covid. So exposed people contract the disease according to a binomial distribution with probability p_0. We expect that if we vaccinate people, this will be reduced. We expose 100 vaccinated people to the coronavirus and find that k contract Covid disease. (So $\overline{x} = k/100$ estimates the probability p that a vaccinated person exposed to the virus contracts the disease.)

We want to test the alternate hypothesis

$$H_1 : p < .05$$

against the null hypothesis

$$H_0 : p = .05.$$

Some obvious critical regions that we might consider are, for $i = 1, 2, 3, 4$,

$$C_i = \{\mathbf{X} \mid \overline{x} \leq i/100\}.$$

The significance of this test when $i = 4$ is

$$\alpha = P(\overline{x} \leq .04 \mid p = .05) = \sum_{i=0}^{4} \binom{100}{i} (.05)^i (.95)^{100-i} \approx .43.$$

It's power, for say $p = .02$ is

$$\gamma(.02) = P(\overline{x} \leq .04 \mid p = .02) = \sum_{i=0}^{4} \binom{100}{i} (.02)^i (.98)^{100-i} \approx .95.$$

Figure 1 shows significance and power values for all of these critical regions.

To have a reasonable significance, say of .05 or below, we don't have very good power. To have good power for even a value as low as $p = .01$ we need to take a critical region as big as C_3, which has an okay, but not great, significance.

C_i	α	$\gamma(.005)$	$\gamma(.01)$	$\gamma(.02)$	$\gamma(.04)$
C_1	.01	.91	.37	.13	.02
C_2	.04	.99	.74	.40	.09
C_3	.12	.99	.92	.68	.23
C_4	.43	.99	.98	.86	.43

Figure 1: Significance and Power calculations for various critical regions

Well, maybe people will not buy a vaccine that only reduces the probability of contracting Covid from .05 to a .02, so maybe we only care about the power for $p = .01$ or $p = .005$. Then maybe this test is okay. But if we do care about $p = .02$ what should we do? We have to use a bigger sample.

To set up a test, we generally do the following.

i) Choose a level α of significance, and a level γ of power.

ii) Choose the values in the alternate hypothese H_1 that we want this power for.

iii) Choose n and C so that the test has significance α and power γ for the chosen values in H_1.

In the next example we do this, and set up a test.

Example 4.5.3. For an RV X with an unknown normal distribution $N(\mu, \sigma^2)$, we want to test the hypotheses

$$H_0 : \mu = 5 \qquad \text{vs.} \qquad H_1 : \mu > 5$$

We expect a mean of about $\mu = 10$ and a variance of about $\sigma^2 = 4$.

A good level of significance is always $\alpha = .05$, and a good power is .95. As we expect a mean of about 10, perhaps we want our level of power for any value of $\mu > 8$. As the power function $\gamma(\mu)$ is clearly increasing in μ for $\mu > 5$, it will be enough to setup the test so that $\gamma(8) \geq .95$.

Figure 2 shows the distribution of the sample mean $\overline{\mathbf{X}}$ according to the null hypothesis, and in the case $\mu = 8$ of the alternate hypothesis, the lowest value for which we want power of .95.

When we set our critical region as the region to the right of the bold black bar in the middle (at $\mu = 6.5$) the significance is the probability of falling in the red region when our mean is 5. The power for $\mu = 8$ is the probability of falling in the blue region (which includes the red region).

To have a power of $\gamma(8) = .95$ and significance .05 must choose n large enough that the red region has probability .05 and the blue has probability .95.

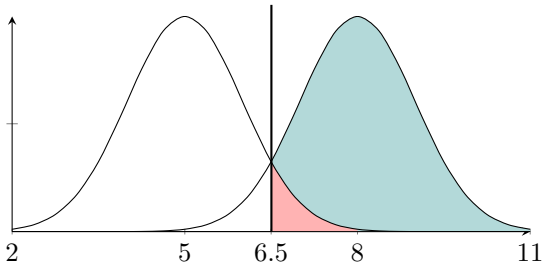


Figure 2: Distributions for $H_0 : \mu = 5$ and $H_1 : \mu > 5$ for value $\mu = 8$

With $\mu = 6.5$ these two thing happen at the same time, so it is enough to find n so that the red region has probabilty is .05.

We expect a value of $\sigma^2 = 4$, and so we approximate this as

$$.95 = P(\overline{\mathbf{X}} \leq 6.5 \mid \mu = 5) = P(\frac{\overline{\mathbf{X}} - \mu}{2/\sqrt{n}} \leq \frac{6.5 - 5}{2/\sqrt{n}}) = P(Z \leq .75\sqrt{n}).$$

We know this z-value: $z_{.95} = 1.65$. Taking $n \geq 5$ we get $.75\sqrt{n} \geq 1.65$, so this is approximately the value of n that we need.

Problem 4.5.4. Taking the approximation $\sigma^2 = 4$ as an approximation of the sample variance S^2. With this, use a t-tables instead of z-tables to approximate the n that you will need to get a significance of .05.

Well, we should maybe bump n up a bit to be safe, particularly when we consider that we do not really know σ, so when we compute the actual significance later, we will need to use a T distribution rather than the normal distribution Z. Lets take $n = 10$. Taking $n = 100$ would be safer, but could also be more expensive. This is a practical issue in statistical testing.

So! We run a test with 10 samples points, and use a critical region of $C = (6.5, \infty)$. Say we get a sample mean of $\boldsymbol{x} = 7.5$ and a sample variance of $S^2 = 5$, slightly higher than the 4 we were expecting.

The significance of our test is

$$\begin{aligned} P(\overline{\mathbf{X}} > 6.5 \mid \mu = 5) &= P(\overline{\mathbf{X}} - \mu > 1.5) \\ &= P(\frac{\overline{\mathbf{X}} - \mu}{S/\sqrt{10}} > 1.5/\sqrt{5/10} \approx 2.12) \end{aligned}$$

Checking the t-values we see that $t_{9,.95} = 1.833$ and $t_{9,.975} = 2.228$. Our 2.12 falls between these, so we estimate this probability at about $\alpha = .035$.

The power for the value $\mu = 8$ is

$$P(\overline{\mathbf{X}} < 6.5 \mid \mu = 8) = \cdots \approx .035$$

Problems from the Text

Section 4.5: 3,4,8,11,12

4.6 Two-sided Tests and p-values

As there can be a difference of opinion on what an appropriate significance is a test can be done without an explicit significance or confidence interval. Instead of accepting or rejecting H_1 we simple return the minimum possible value of α at which H_1 would have been accepted. This is the *p-value* of the test. It is then up to the reader of the results to decide if they find them significant.

Definition 4.6.1. Formally, the *p-value* p is defined as the probability, assuming H_0, that the test would yield an outcome at least as extreme as the observed outcome.

Now 'at least as extreme' allows room for interpretation, and will depend on the hypotheses, but it is usually quite clear what it should mean.

Say our hypotheses are

$$H_0 : \mu = \mu_0 \qquad \text{vs.} \qquad H_1 : \mu > \mu_0$$

and our test yields a sample mean of $\overline{\boldsymbol{x}}$. The p-value is

$$P(\overline{\mathbf{X}} \geq \overline{\boldsymbol{x}} \mid \mu = \mu_0).$$

Example 4.6.2. The weight, in ounces, of cereal in a 10-ounce box is $N(\mu, \sigma^2)$. To test

$$H_0 : \mu = 10.1 \qquad \text{vs.} \qquad H_1 : \mu > 10.1$$

we take a sample of size $n = 16$ and find that $\overline{\boldsymbol{x}} = 10.4$ and $S = 0.4$.

Let's find the p-value of the test.

$$\begin{aligned} P(\overline{\mathbf{X}} \geq 10.4, \mu = 10.1) &= P\left(\frac{\overline{\mathbf{X}} - 10.1}{.4/\sqrt{n}} > \frac{.3}{.4/\sqrt{n}}\right) \\ &= P\left(T(15) > 3\right) \end{aligned}$$

Looking up in the T-tables we get that $t_{.0043,15} \approx 3$, so our p-value, this probability, is about $p = .0043$. That is a fairly significant test.

When we were testing the effect of a vaccination, we were only interested in the possibility that in reduces the probability of contracting a disease, so we used a one-sided test. In some situations, we may be interested in an effect that can act up or down. In such situations, we would opt for a two-sided test.

Example 4.6.3. The height of people in a population, in centimeters, is $H \sim N(\mu_0 = 175, \sigma_0^2)$. We suspect that there is a relationship between the time of year someone is born, and their height. So we sample 30 people born in April an measure their heights to get a sample of the distribution $X \sim N(\mu, \sigma^2)$ of heights of people born in April.

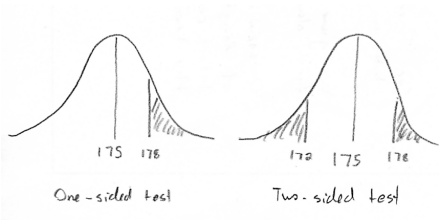
Our hypotheses are

$$H_0 : \mu = \mu_0 = 175 \qquad \text{vs.} \qquad H_1 : \mu \neq \mu_0.$$

H_1 is a *two-sided hypothesis*.

What is the p-value if our sample yields $\overline{\mathbf{X}} = 178$ and $S^2 = 120$?

In a one-sided test we would be calculating the probability of falling in the this region on the right of the following figure. However, in a two sided test, the probability of getting this outcome or one more extreme is the probability of falling in the region on the left.



The p-value of this is

$$1 - P(172 < \overline{\mathbf{X}} < 178) = 1 - P(-1.5 < \frac{\overline{\mathbf{X}} - 175}{\sqrt{120/30}} < 1.5) \approx 2 \cdot .068 = .136.$$

Problems from the Text

Section 4.6: 4,6

4.7 Chi-squared Tests

A chi-squared test is used to test a hypothesis of the following type.

Where $\omega_1, \ldots, \omega_k$ is a partition of the the support Ω of an RV X, and for each i we have

$$p_i = P(X \in \omega_i)$$

consider the null hypothesis

$$H_0 : p_i = p_{i,0} \text{ for all } i \in [k],$$

and the alternate nypothesis

$$H_0 : p_i \neq p_{i,0} \text{ for some } i \in [k].$$

We take an n point sample of X, and for $i = 1, \ldots, k$ let X_i be the RV that counts the number of sample points in ω_i. This is an estimator of the parameter $n \cdot p_i$.

The RV

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_{i,0})^2}{np_{i,0}}$$

can be shown to be a chi-squared distribution with $k-1$ degrees of freedom. Clearly, it has low expected value if H_0 is true. So we will accept H_0 if it is low, and reject it if H_1 is high. To decide what 'low' and 'high' are we have to look at the chi-squared distribution.

Example 4.7.1. To test if a die is fair, we roll it 60 times and let X_i be the number of times that i shows up. Our results are

$$X_1 = 13,\ X_2 = 19,\ X_3 = 11,\ X_4 = 8,\ X_5 = 5,\ X_6 = 4.$$

To test

$$H_0 : p_i = 1/6 \qquad \forall i \qquad \text{vs.} \qquad H_1 : p_i \neq 1/6 \qquad \exists i$$

we compute

$$Q_5 = \sum_{i=1}^{6} \frac{(X_i - 10)^2}{10} = 15.6.$$

Looking at table for the cdf of a chi-squared distribution, we find the $P(Q_5 \geq 15.086) = .01$. So our test has a p-value of less than .01. We would accept the hypothesis that the die is biased at a significance of .01.

Problems from the Text

Section 4.7: 3

4.8 The Monte-Carlo Method

The generation of random samples of a given distribution is called the Monte-Carlo method, and has several uses. The problem of generating random (or randon-like) samples of any distribution is a fairly deep computational problem, but all modern computer languages have pretty good generators of the standard uniform distribution Unif($[0, 1]$).

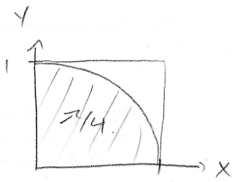
Problem 4.8.1. Generate a Unif$([a,b])$ distribution using a Unif$([0,1])$ generator.

We look at a couple of 'non-statistics' applications of a random Unif$([0,1])$ generator, and then use it to generate other distributions.

Example 4.8.2 (Estimation of π)**.** Using a random Unif$([0,1])$ generator, generate n random pairs $(X_i, Y_i) \in [0,1] \times [0,1]$. Let Z be the RV that counts the number of pairs (X_i, Y_i) that satisy

$$X_i^2 + Y_i^2 < 1.$$

All pairs fall inside a region of area one, and Z counts the number of pairs that fall inside the shown region of area of $\pi/4$.



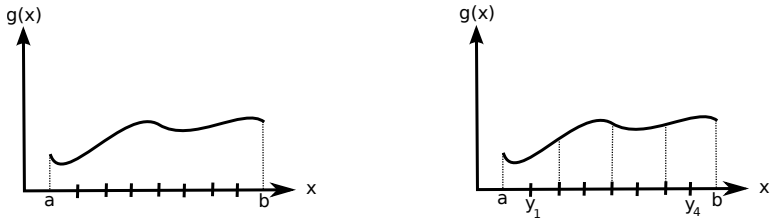
So $4Z/n$ is an unbiased estimator of $4(\pi/4) = \pi$.

Example 4.8.3 (Monte Carlo Integration)**.** Say we want to evaluate

$$\int_a^b g(x)\,\mathrm{d}x$$

for some $g(x)$ but we cannot find its antiderivative.

We want to find the area under the curve in the first picture below. The order statistics $y_1, \ldots, y_4$ of a 4 point statistic are expected to fall as in teh second picture, so we can use them to approximate the area with the Reimann sum $\sum g(y_i)\frac{b-a}{n}$.



Doing this, as n gets big, we get

$$\int_a^b g(x)\,\mathrm{d}x = (b-a)\int_a^b \frac{g(x)}{b-a}\,\mathrm{d}x = (b-a)\int_a^b f_X(x)g(x)\,\mathrm{d}x = (b-a)E(g(X))$$

using that the distribution f_X of $X \sim \text{Unif}((a,b))$ is $1/(b-a)$. Using a generator of $\text{Unif}((a,b))$ we generate n samples X_i of X and let

$$I = \frac{\sum g(X_i)}{n}.$$

This has expected value $E(g(X))$ so $(b-a)I$ has expected value $\int_a^b g(x)\,\mathrm{d}x$.

Now we look at generating samples from other distributions, using a generator of $U \sim \text{Unif}([0,1])$. If it is a distribution X whose cdf F_X we can invert, the obvious technique is to generate samples U_i of U and then return $X_i = F_X^{-1}(U_i)$.

Example 4.8.4 (Generating distributions whose cdf we can invert)**.** The exponential distribution $X \sim \Gamma(1, 1/\mu)$ with mean $\mu = 2$ has cdf

$$F(X) = 1 - e^{-2x} \qquad x \geq 0.$$

To generate a random sample of X we generate a random sample of U:

```
sage: U = [random() for i in range(4)]; U
[0.09742247457107367,
 0.5705002816351732,
 0.949933512893782,
 0.44883965991368235]
```

and then plug it into $F^{-1}(u) = -\frac{1}{2}\ln(1-u)$:

```
sage: [(-1/2)*ln(1-u) for u in U]
[0.05125034585604324,
 0.4225670966430177,
 1.4972017073768573,
 0.29786475690962677]
```

Example 4.8.5 (Generating Poisson process and a Poisson RV)**.** We want to generate a Poisson process that produces a mean of μ occurence per unit time. The exponential RV $X \sim \Gamma(1, 1/\mu)$ counts the waiting time until the first occurence of such a process.

So from our generated values of X we get a Poisson process whose first occurence is at time .051, whose second is at time $.051 + .422 = .473$, whose third is at time $.473 + 1.497 = 1.970$, etc.

As the Poisson RV $Y \sim \text{pois}(\mu)$ counts the number of occurences of this Poisson process in unit time, we have just generated a sample value 2 of Y.

Now we can generate distributions whose cdf we can invert, and we have also seen how to generate a Poisson distribution whose cdf would be difficult to invert. The text shows how to generate two independent points of a a normal distribution from to independent points of U. We skip this, and give one final technique that will work for all distributions we would like to generate.

Example 4.8.6 (The Accept-Reject Generation Algorithm)**.** We generate a distribution W with pdf f

Simple Version . Take two uniform distributions: $X \sim \text{Unif}([a,b])$ and $Y \sim \text{Unif}([0,M])$ where $[a,b]$ is the support of f and M is (greater than) its maximum value.

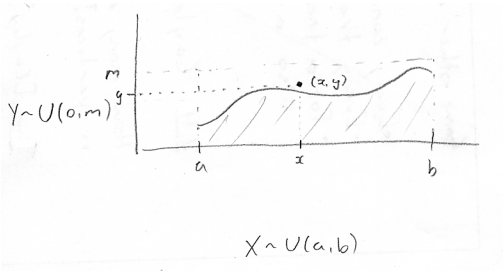
Generate a sample of W as follows.

i) Generate samples x of X and y of Y.

ii) If $f(x) < y$ return x, and stop.

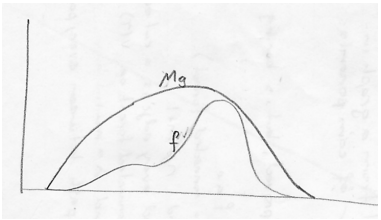
iii) Otherwise, throw out samples and go back to (i).

What we are doing is randomly generating a point in the rectangle below, and returning its x value if it falls under the pdf of W.



This works: if $f(x) = 2f(x')$ then x and x' are equally likely to be generated from X, but x is twice as likely to be accepted by the algorithm. So it is twice as likley to be returned.

Now, if we are rejecting a lot of points, the algorithm may take a while, so we can replace X and Y with another distribution that is closer to W.



More General Version

Let X be a distribution with pdf g for which we can generate samples. Let M be such that $f(x) < Mg(x)$ for all values of x. Let $Y \sim \text{Unif}([0,1])$.

Generate a sample of W as follows.

i) Generate samples x of X and y of Y.

ii) If $f(x)/g(x) < y$ return x, and stop.

iii) Otherwise, throw out samples and go back to (i).

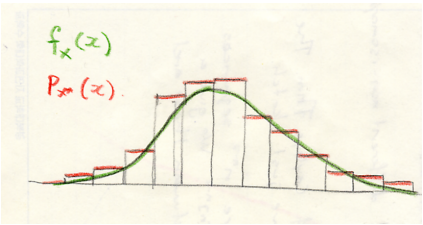
Problems from the Text

Section 4.8: 4,6,18

4.9 Bootstrap Procedures

Bootstrapping is a clever tool for getting around situations in which mathematical analysis is tricky.

The idea is that a random sample $\mathbf{X}$ of an RV X, gives a good approximation of the distribution of X. Indeed, let X^* be a randomly chosen element of $\{X_1, \ldots, X_n\}$. The distribution (pmf) p_{X^*} of X^* is the function defined by scaling the histogram of the sample $\mathbf{X}$. This is called the *empirical distribution*; it is a close approximation of the distribution (pdf) f_X of X.



It follows that an n-point *resample*; that is, a set $\mathbf{X}^*$ of n points chosen iid from $\mathbf{X}$ (with replacement) will have a similar distribution to $\mathbf{X}$. So statistics of a resample should be similar to statistics of a sample.

Example 4.9.1 (Bootstrap confidence intervals)**.** To find a 95% confidence interval $[\theta_L, \theta_R]$ for a parameter θ of a distribution X based on a random sample $\mathbf{X}$ the main ideas was to find the distribution $F_{\hat{\theta}}$ of an estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$ and set

$$[\theta_L, \theta_R] = [F_{\hat{\theta}}^{-1}(.025), F_{\hat{\theta}}^{-1}(.975)].$$

That is we are finding the 'spread' of the distribution of the estimator $\hat{\theta}$.

When the parameter we are interested in is the mean μ then this is easy, the estimator $\overline{\mathbf{X}}$ (upto normalisation) has a $\hat{\theta}$ distribution if X is normal or otherwise an approximate Z distribution by the CLT.

For other parameters, however, the distribution of an estimator may be difficult to determine.

Assume $\hat{\theta} = \hat{\theta}(\mathbf{X})$ is an unbiased estimator for a parameter θ of the distribution X. We want to find a 95% confidence interval for θ.

One way to do it would be to take N different n point random samples of X, and get a point estimate $\hat{\theta}$ for each. Ordering the estimates

$$\hat{\theta}_1 < \hat{\theta}_2 < \cdots < \hat{\theta}_N,$$

a 95% confidence interval for θ is $[\hat{\theta}_{.025\cdot N}, \hat{\theta}_{.975\cdot N}]$. The problem, though, is that taking a lot of samples can become expensive. So we take resamples instead.

For $j = 1, \ldots, N$ let $\mathbf{X}_j^*$ be an *n-point resample* of $\boldsymbol{x}$. Let $\hat{\theta}_j = \hat{\theta}(\mathbf{X}_j^*)$. Ordering the $\hat{\theta}_j$ as

$$\hat{\theta}_{(1)} \leq \hat{\theta}_{(2)} \leq \ldots \hat{\theta}_{(N)},$$

the 95% *percentile bootstrap* confidence interval for θ is

$$[\hat{\theta}_{(.025\cdot N)}, \hat{\theta}_{(.975\cdot N)}].$$

The proof that percentile bootstrap confidence interval is valid is beyond our scope. We finish with two examples of using bootstrapping for hypothesis testing.

Example 4.9.2. Let X and Y be RV such that $(X - \mu_x) \sim (Y - \mu_Y)$. We want to test the hypotheses

$$H_0 : \mu_x = \mu_y \qquad \text{vs.} \qquad H_1 : \mu_x \neq \mu_y$$

about whether the two X and Y have the different mean.

For samples $\mathbf{X}$ and $\mathbf{Y}$ of sizes m and n respectively, we let $\Delta = \overline{Y} - \overline{X}$. If $|\Delta| > c$ for some critical c we will accept H_1, and if $|\Delta| < c$ we will reject it.

What should the critical value c be for a test of significance $\alpha = .05$? In Section , we looked at finding the distribution of Δ. Using bootstrapping, we simply pool our samples into a population $Z = \{x_1, \ldots, x_m\} \cup \{y_1, \ldots, y_n\}$. Under the null hypothesis $H_0 : \mu_x = \mu_y$, Z is just a sample from the common distribution $X \sim Y$.

Taking N m-point resamples X_i^* of Z and N n-point resamples Y_i^* of Z, we let $\Delta_i^* = Y_i^* - X_i^*$. Ordering the Δ_i^*

$$|\Delta_{(1)}^*| \leq |\Delta_{(2)}^*| \leq \cdots \leq |\Delta_{(N)}^*|,$$

let $c = |\Delta_{(.95\cdot N)}^*|$.

The basic idea we are using is that the emperical distribution based on a sample has a similar distribution to the original RV. This is fine when using it to estimate variance, as we have been doing, but the emperical distribution is not exactly the original distribution. We would expect it to have a different mean than the original distribution. In the next example we see how we may have to account for this.

Example 4.9.3. For an RV X we want to test the hypotheses,

$$H_0 : \mu = \mu_0 \qquad \text{vs.} \qquad H_1 : \mu \neq \mu_0$$

.

For a random sample $\mathbf{X}$, we will accept H_1 if $\overline{\mathbf{X}} > \mu_0 + c$. We want to use bootstrapping to find the p-value of the test; that is, to find the probability that a sample is at least c greater than μ. The distribution of $\mathbf{X}^*$ is similar to that of $\mathbf{X}$ but clearly it has mean $\overline{\mathbf{X}}$ whereas $\mathbf{X}$ has mean μ. So when we resample, we find the the probability that a *resample* is at least c greater than $\overline{\mathbf{X}}$.

Letting μ_i^* be the means of N resamples, we have that p is the percentage of the μ_i^* that satisfy $\mu_i^* > \overline{\mathbf{X}} + c$.

Problems from the Text

Section 4.9: 1

5 Tables

You will be given the following two pages on tests.

Bernoulli $X \sim b(1,p)$

$p_X(x)$	$\begin{cases} p & \text{if } x = 1 \\ 1-p & \text{otherwise.} \end{cases}$
μ	p
σ^2	pq
$M_X(t)$	$pe^t + q$

Binomial $X \sim b(n,p)$

$p_X(x)$	$\binom{n}{x} p^x (1-p)^{n-x}$
μ	np
σ^2	npq
$M_X(t)$	$(pe^t + q)^n$

Hypergeometric

$p_X(x)$	$\frac{\binom{N-D}{n-x}\binom{D}{x}}{\binom{N}{n}}$
μ	$n\frac{D}{N}$
σ^2	$n\frac{D}{N}\frac{N-D}{N}\frac{N-n}{N-1}$

Poisson $X \sim \text{pois}(\mu)$

$p_X(x)$	$\frac{e^{-\mu}\mu^x}{x!}$
μ	μ
σ^2	μ
$M_X(t)$	$e^{\mu(e^t-1)}$

Gamma $X \sim \Gamma(\alpha, \beta)$

$f_X(x)$	$\frac{x^{\alpha-1}e^{-(x/\beta)}}{\Gamma(\alpha)\beta^\alpha}$
μ	$\alpha\beta$
σ^2	$\alpha\beta^2$
$M_X(t)$	$(1-\beta t)^{-\alpha}$

Exponential $X \sim \Gamma(1, \mu)$

$f_X(x)$	$e^{-x/\mu}/\mu$
μ	μ
σ^2	μ^2
$M_X(t)$	$(1-\mu t)^{-1}$

Normal $X \sim N(\mu, \sigma^2)$

$f_X(x)$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
$M_X(t)$	$e^{\mu t+\frac{1}{2}\sigma^2 t^2}$

Table I
Poisson Distribution

The following table presents selected Poisson distributions. The probabilities tabled are

$$P(X \le x) = \sum_{w=0}^{x} e^{-m} \frac{m^w}{w!},$$

for the values of m selected.

	$m = E(X)$										
x	0.5	1.0	1.5	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
0	0.607	0.368	0.223	0.135	0.050	0.018	0.007	0.002	0.001	0.000	0.000	0.000
1	0.910	0.736	0.558	0.406	0.199	0.092	0.040	0.017	0.007	0.003	0.001	0.000
2	0.986	0.920	0.809	0.677	0.423	0.238	0.125	0.062	0.030	0.014	0.006	0.003
3	0.998	0.981	0.934	0.857	0.647	0.433	0.265	0.151	0.082	0.042	0.021	0.010
4	1.000	0.996	0.981	0.947	0.815	0.629	0.440	0.285	0.173	0.100	0.055	0.029
5	1.000	0.999	0.996	0.983	0.916	0.785	0.616	0.446	0.301	0.191	0.116	0.067
6	1.000	1.000	0.999	0.995	0.966	0.889	0.762	0.606	0.450	0.313	0.207	0.130
7	1.000	1.000	1.000	0.999	0.988	0.949	0.867	0.744	0.599	0.453	0.324	0.220
8	1.000	1.000	1.000	1.000	0.996	0.979	0.932	0.847	0.729	0.593	0.456	0.333
9	1.000	1.000	1.000	1.000	0.999	0.992	0.968	0.916	0.830	0.717	0.587	0.458
10	1.000	1.000	1.000	1.000	1.000	0.997	0.986	0.957	0.901	0.816	0.706	0.583
11	1.000	1.000	1.000	1.000	1.000	0.999	0.995	0.980	0.947	0.888	0.803	0.697
12	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.991	0.973	0.936	0.876	0.792
13	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.996	0.987	0.966	0.926	0.864
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.994	0.983	0.959	0.917
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.992	0.978	0.951
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.996	0.989	0.973
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.995	0.986
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.993
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.997
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998
21	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
22	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table II
Chi-square Distribution

The following table presents selected quantiles of chi-square distribution; i.e, the values x such that

$$P(X \le x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1} e^{-w/2}\, dw,$$

for selected degrees of freedom r.

	$P(X \le x)$							
r	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Table III
Normal Distribution

The following table presents the standard normal distribution. The probabilities tabled are

$$P(X \le x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw.$$

Note that only the probabilities for $x \ge 0$ are tabled. To obtain the probabilities for $x < 0$, use the identity $\Phi(-x) = 1 - \Phi(x)$.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998

Table IV
t-Distribution

The following table presents selected quantiles of the t-distribution; i.e, the values x such that

$$P(X \le x) = \int_{-\infty}^{x} \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\Gamma(r/2)(1+w^2/r)^{(r+1)/2}}\, dw$$

for selected degrees of freedom r. The last row gives the standard normal quantiles.

	$P(X \le x)$					
r	0.900	0.950	0.975	0.990	0.995	0.999
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
∞	1.282	1.645	1.960	2.326	2.576	3.090

References

[1] Hogg, McKean, Craig, *Introduction to Mathematical Statistics (Seventh Edition; International Edition)*. Pearson Education, Inc. (1995, Seventh Ed. 2013).