Introductory (mathematical) statistics

Ümit Işlak

February 22, 2020

Contents

1 Probability measures

A process, real or hypothetical, in which the possible outcomes can be identified ahead of time is called an **experiment**. An **event** is a well-defined set of possible outcomes of the experiment. The collection of all possible outcomes of an experiment is called the **sample space** of the experiment, and is denoted by S.

In order to discuss probabilities in a given experiment, it is necessary to assign to each event A in the sample space S a number $\mathbb{P}(A)$ that indicates the probability that A will occur. In order to satisfy the mathematical definition of probability, the function $\mathbb{P}$ defined on events must satisfy three specific axioms:

Axioms of a probability measure

Axiom 1. For any event A, $\mathbb{P}(A) \geq 0$.

Axiom 2. $\mathbb{P}(S) = 1$ where S is the sample space.

Axiom 3. For every sequence of disjoint events $A_1, A_2, \ldots,$

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \qquad \text{(countable additivity).}$$

These axioms are known as *Kolmogorov's axioms*. The first two should be intuitively clear. One may wonder why only finite additivity is not enough in third axiom. This is rather technical, and we avoid going into details here.[1]

Definition 1.1 *A **probability measure**, or simply a **probability**, on a sample space S is a specification of numbers $\mathbb{P}(A)$ for all events A, that satisfy Kolmogorov's axioms.*

2 More properties of probability measures

Based on the axioms we introduced in previous section, we will now develop some theory. Below we denote the empty event by $\emptyset$.

Theorem 2.1 $\mathbb{P}(\emptyset) = 0$.

Proof: Let $A_i = \emptyset$ for $i = 1, 2, \ldots$. Then

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

[1] Yet, see Theorem 2.7 for an instance where this axiom is indeed useful.

which in particular implies that $\mathbb{P}(\emptyset) = 0$ since $\mathbb{P}(\emptyset) \geq 0$ by Axiom 1. □

Theorem 2.2 *For every* finite *sequence of n disjoint events $A_1, \ldots, A_n$,*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) \qquad \text{(finite additivity).}$$

Proof: Set $A_i = \emptyset$ for $i > n$. Then we have

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) &= \sum_{i=1}^{n} \mathbb{P}(A_i) + \sum_{i=n+1}^{\infty} \mathbb{P}(A_i) \\ &= \sum_{i=1}^{n} \mathbb{P}(A_i) + \sum_{i=n+1}^{\infty} 0 = \sum_{i=1}^{n} \mathbb{P}(A_i). \end{aligned}$$

□

Theorem 2.3 *For any event A,*

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

Proof: First note that $A \cup A^c = S$, and that A and A^c are disjoint. So, we can now use finite additivity and get

$$1 = \mathbb{P}(S) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c),$$

from which the result follows. □

Theorem 2.4 *For any two events A, B with $A \subseteq B$,*

$$\mathbb{P}(A) \leq \mathbb{P}(B).$$

Proof: We have

$$\mathbb{P}(B) = \mathbb{P}(A \cup (B \cap A^c)) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \geq \mathbb{P}(A),$$

because $\mathbb{P}(B \cap A^c) \geq 0$ by Axiom 1. □

Theorem 2.5 *For any event A,*

$$0 \leq \mathbb{P}(A) \leq 1.$$

Proof: First inequality is Axiom 1. For the second inequality, we note that $A \subseteq S$. So using Theorem 2.5 and Axiom 2, $\mathbb{P}(A) \leq \mathbb{P}(S) = 1$. □

Exercise 2.1 *Using the axioms of probability, and the properties we derived so far, prove the following:*

(i) For every two events A and B, we have

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B).$$

(ii) For every two events A and B, we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

In most real life problems, it will not not be easy (or maybe even not possible) to compute the exact value of a probability for unions of events. This takes us to our first inequality.

Theorem 2.6 *(Boole's inequality) For all events $A_1, \ldots, A_n$, we have*

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

(ii) $\mathbb{P}\left(\bigcap_{i=1}^n A_i^c\right) \geq 1 - \sum_{i=1}^n \mathbb{P}(A_i)$.

Proof: For the first part a simple proof can be given by using induction. Here we discuss another one which provides a useful technique that can be applied to several other problems. Let $B_1 = A_1$, and for $j \geq 2$, set

$$B_j = A_j - \bigcup_{k=1}^{j-1} A_k.$$

Then, (1) B_j's are disjoint, (2) $\bigcup_{i=1}^n A_i = \bigcup_{i=1}^n B_i$ and (3) $B_i \subseteq A_i$, $i = 1, \ldots, n$. These observations yield

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

(ii) follows easily by (i) once we consider the complementary event and use De Morgan's law. □

Here is an exercise for you which provides a variation of Boole's inequality.

Exercise 2.2 *(Bonferroni's inequalities) Let $A_1, \ldots, A_n$ be arbitrary events. Show that*

(a) $\mathbb{P}\big(\bigcup_{i=1}^n A_i\big) \geq \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j)$.

(b) $\mathbb{P}\big(\bigcup_{i=1}^n A_i\big) \leq \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k)$.

Next, we discuss an important property of probability measures[2] which is continuity under monotone sequences. This may look a little bit technical at first glance, but you will grasp its importance if you take an advanced course on statistics, probability theory or analysis.

[2] Note here that the countable additivity axiom is a necessary for the proof.

Theorem 2.7 *A sequence of events $A_1, A_2, \ldots,$ is said to be* ***monotone increasing*** *if $A_j \subset A_{j+1}$ for all $j \geq 1$. If $\{A_j\}_{j\geq 1}$ is a sequence of monotone increasing events, then*

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

Proof: Let $B_1 = A_1$, $B_2 = A_2 - A_1$, and more generally $B_k = A_k - \left(\bigcup_{i=1}^{k-1} A_i\right)$, $k \geq 2$. Now,

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

proving our claim. □

Exercise 2.3 *A sequence of events $A_1, A_2, \ldots,$ is said to be* ***monotone decreasing*** *if $A_{j+1} \subset A_j$ for all $j \geq 1$. Show that if $\{A_j\}_{j\geq 1}$ is a monotone decreasing sequence, then*

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

(Hint: Use a similar argument to the proof of Theorem 2.7.)

We lastly state the inclusion exclusion principle in this section.

Theorem 2.8 *For any n events $A_1, \ldots, A_n$,*

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

Exercise 2.4 *Prove the inclusion-exclusion principle.*

3 Conditional probability

Definition 3.1 *Suppose we learn that an event B has occured and we wish to find the probability of another event A. The new probability of A is called the **conditional probability** of A given that B has occured and is denoted by $\mathbb{P}(A \mid B)$.*

If $\mathbb{P}(B) > 0$, we compute this probability using the formula

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

$\mathbb{P}(A \mid B)$ is not defined if $\mathbb{P}(B) = 0$. [3]

We begin with two standard examples to keep in mind: tossing a coin and rolling dice.[4]

Example 3.1 *(Tossing coins) Suppose I toss a coin twice and I know that there is at least one head. What is the probability that both of them are heads?*

Solution: We may represent the sample space of this experiment $S = \{HH, HT, TH, TT\}$. The conditional probability in question then equals

$$\mathbb{P}(\text{both heads} \mid \text{ at least one head}) = \frac{\mathbb{P}(HH)}{\mathbb{P}(HH \text{ or } HT \text{ or } TH)} = \frac{1}{3}.$$

Note that the unconditional probability would be just 1/4. □

Remark 3.1 *The tossing coins example has a famous alternative interpretation. Suppose that a family has two children. If you know that one of them is a girl, then the probability that they have two daughters equals 1/3 - this will be an exercise for you.*

Exercise 3.1 *(Rolling dice) Suppose that two dice were rolled and the sum T was odd. Find the probability that T was less than 8.*

Remark 3.2 *Conditional probabilities behave just like the standard probability measures. For example, we have $\mathbb{P}(A^c \mid B) = 1 - \mathbb{P}(A \mid B)$. To see that this is indeed the case, observe*

$$\mathbb{P}(A^c \mid B) = \frac{\mathbb{P}(A^c \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B) - \mathbb{P}(A \cap B)}{\mathbb{P}(B)} = 1 - \mathbb{P}(A \mid B).$$

Other properties can be proven similarly.

[3] You should understand the intuition behind this definition by simply drawing a diagram, and considering the probabilities corresponding to outcomes in $A \cap B$ and in B.

[4] The reader is assumed to be familiar with very basic combinatorics; permutations, combinations, etc. In case you need a review, see the relevant appendix.

Letting A, B be events with $\mathbb{P}(B) > 0$, the definition of conditional probability is equivalent to

$$\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B).$$

This last equality is known as the **multiplication rule** for two events. Next result is its generalization to n many events.

Theorem 3.1 *(Multiplication rule) Letting $A_1, \ldots, A_n$ be events so that $\mathbb{P}(A_1 \cap \cdots \cap A_{n-1}) > 0$,*

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(A_n \mid A_1 \cap \cdots A_{n-1}).$$

Proof: We just use the definition of conditional probability multiple times to see that the right hand side equals

$$\mathbb{P}(A_1)\frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)}\frac{\mathbb{P}(A_1 \cap A_2 \cap A_3)}{\mathbb{P}(A_1 \cap A_2)} \cdots \frac{\mathbb{P}(A_1 \cap \cdots \cap A_n)}{\mathbb{P}(A_1 \cap \cdots \cap A_{n-1})}.$$

This is a telescoping product which is easily seen to equal the left hand side. □

Example 3.2 *Suppose that a box contains b blue, r red, and g green balls. We select four balls at random without replacement. Find the probability that the selected balls are red, green, blue, red, respectively.*

Solution: Let B_i, R_i, G_i be the events that we choose a blue, red, and a green ball, respectively in i^{th} selection, $i = 1, 2, 3, 4$. Then the required probability is

$$\begin{aligned}\mathbb{P}(R_1 \cap G_2 \cap B_3 \cap R_4) &= \mathbb{P}(R_1)\mathbb{P}(G_2 \mid R_1)\mathbb{P}(B_3 \mid R_1 \cap G_2)\mathbb{P}(R_4 \mid R_1 \cap G_2 \cap B_3) \\ &= \left(\frac{r}{r+b+g}\right)\left(\frac{g}{r+b+g-1}\right)\left(\frac{b}{r+b+g-2}\right)\left(\frac{r-1}{r+b+g-3}\right).\end{aligned}$$

□

Next, we discuss one of the most useful theorems of this course: The law of total probability. For this purpose first recall the definition of a partition.

Definition 3.2 *Let S be the sample space of some experiment, and consider k events $B_1, \ldots, B_k$ such that $B_1, \ldots, B_k$ are disjoint and $\bigcup_{i=1}^k B_i = S$. Then $B_1, \ldots, B_k$ is said to form a **partition** of S.*

Theorem 3.2 *(Law of total probability) Suppose that the events $B_1, \ldots, B_k$ form a partition of the sample space S and $\mathbb{P}(B_j) > 0$ for $j = 1, \ldots, k$. Then for every event A*

$$\mathbb{P}(A) = \sum_{j=1}^k \mathbb{P}(B_j)\mathbb{P}(A \mid B_j).$$

Exercise 3.2 *Prove Theorem 3.2.*

Example 3.3 *A box contains three coins with a head on each side, four coins with a tail on each side, and two fair coins. If one of these nine coins is selected at random, what is the probability that a head will be obtained?*

Solution: Let

- C_1 be the event that the chosen coin has heads on each side;
- C_2 be the event that the chosen coin has tails on each side;
- C_3 be the event that the chosen coin is fair;

Also let H be the event that a head is obtained. Then

$$\begin{aligned}\mathbb{P}(H) &= \mathbb{P}(C_1)\mathbb{P}(H \mid C_1) + \mathbb{P}(C_2)\mathbb{P}(H \mid C_2) + \mathbb{P}(C_3)\mathbb{P}(H \mid C_3) \\ &= \frac{3}{9} \times 1 + \frac{4}{9} \times 0 + \frac{2}{9} \times \frac{1}{2} = \frac{4}{9}.\end{aligned}$$

You could solve this problem in a much more straightforward way (How?), but our purpose here was to apply the law of total probability. □

Example 3.4 [5] *As a simplified model for weather forecasting, suppose that the weather (either wet or dry) tomorrow will be the same as the weather today with probability p. If the weather is dry on January 1, show that P_n, the probability that it will be dry n days later, satisfies*

$$P_n = (2p-1)P_{n-1} + (1-p), \quad n \geq 1$$

and

$$P_0 = 1.$$

Prove that

$$P_n = \frac{1}{2} + \frac{1}{2}(2p-1)^n, \quad n \geq 0.$$

Solution: We use the law of total probability to get

$$\begin{aligned}P_n = \mathbb{P}(n^{th} \text{ day dry}) &= \mathbb{P}(n^{th} \text{ day dry} \mid (n-1)^{st} \text{ day dry})\mathbb{P}((n-1)^{st} \text{ day dry}) \\ &\quad + \mathbb{P}(n^{th} \text{ day dry} \mid (n-1)^{st} \text{ day wet})\mathbb{P}((n-1)^{st} \text{ day wet}) \\ &= pP_{n-1} + (1-p)(1-P_{n-1}) \\ &= (2p-1)P_{n-1} + (1-p).\end{aligned}$$

Second claim follows from mathematical induction and is left for you[6]. □

[5] This is an example of a Markov chain (a special stochastic (i.e., random) process), an important topic in any field related to probability theory.

[6] In general, the formula will not be available to you. So you should be okay with finding solutions to recursions. A much useful technique for solving such recursions is the use of generating functions. I strongly suggest learning about it.

4 Independence

Definition 4.1 *Two events A and B are said to be (statistically)* ***independent*** *if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Remark 4.1 *A and B are independent if and only if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ (or, equivalently $\mathbb{P}(B \mid A) = \mathbb{P}(B)$). This explains the definition of independence: Getting information about B does not change the probability of A, and vice versa.*

Remark 4.2 *Note that we may have physically related events which are (statistically) independent. For such an example, consider the case where we roll a die, let A be the event that an even number is obtained and B be the event that one of the numbers $1, 2, 3,$ or 4 is obtained. Then A and B are obviously physically related, but it can be easily checked that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.*

Theorem 4.1 *If two events A and B are independent, then*
(i) A and B^c are also independent, and
(ii) A^c and B^c are also independent.

Proof: (i) We have

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A)(1 - \mathbb{P}(B)) = \mathbb{P}(A)\mathbb{P}(B^c),$$

showing that A and B^c are independent.
(ii) Follows from part i (Why?). □

Here is the generalization of independence to an arbitrary number of events.

Definition 4.2 *The k events $A_1, \ldots, A_k$ are* ***independent*** *if, for every subset $A_{i_1}, \ldots, A_{i_j}$ of j of these events ($j = 2, 3, \ldots, k$),*

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_j}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_j}).$$

Example 4.1 *When $k = 3$, the condition above reduces to having*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B),$$

$$\mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C),$$

$$\mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C),$$

and

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C),$$

Exercise 4.1 *How many equations do you need to check to see that the events $A_1, \ldots, A_k$ are independent for $k \geq 2$?*

Example 4.2 *Suppose that a machine produces a defective item with probability p $(0 < p < 1)$ and produces a non-defective item with probability $1 - p$. Assume that the productions are independent.*

(a) What is the probability that the first m, $m \geq 1$ items produced are defective?

(b) What is the probability that there is at most one defective items in first n, $n \geq 1$ productions?

(c) What is the probability that the first defective item will be the n^{th} one, $n \geq 1$?

(d) What is the probability that we will have a defective item at a finite time?

Solution: Let D_i, $i \geq 1$, be the event that the i^{th} produced item is defective. These are independent events by assumption. Also, we are given that $\mathbb{P}(D_i) = p$ for each i.

(a) Using independence $\mathbb{P}(D_1 \cap D_2 \cap \cdots \cap D_m) = \prod_{i=1}^m \mathbb{P}(D_i) = \prod_{i=1}^m p = p^m$.

(b) We will find the probability of the event that there is no defective item, and that there is exactly one defective item. Similar to part (a)

$$\mathbb{P}(\text{no defective}) = (1-p)^n.$$

Also, in order to have exactly one defective items there should be some $i = 1, 2, \ldots, n$ which corresponds to a defective item and all other productions should be non-defective. Noting that there are $\binom{n}{1} = n$ distinct ways of choosing i, the required probability is then

$$\mathbb{P}(\text{exactly one defective}) = np(1-p)^{n-1}.$$

Combining these observations we get

$$\mathbb{P}(\text{at most one defective}) = (1-p)^n + np(1-p)^{n-1}.$$

(c) We have

$$\mathbb{P}(\text{first defective at } n^{th} \text{ production}) = \mathbb{P}(D_1^c \cap \cdots \cap D_{n-1}^c \cap D_n) = \left(\prod_{i=1}^{n-1} (1-p)\right) p = (1-p)^{n-1} p.$$

(d) Let's consider the complementary event which is to have non-defective items at any time. We have

$$\mathbb{P}\left(\text{all non-defective}\right) = \mathbb{P}(D_1^c \cap D_2^c \cap \cdots) = \mathbb{P}\left(\bigcap_{i=1}^{\infty} D_i^c\right) = \lim_{n \to \infty} \mathbb{P}(\bigcap_{i=1}^{n} D_i^c).$$

(For a rigorous justification of the last step, adapt the result in Exercise 2.3 to our case.) Now,

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{n} D_i^c\right) = \lim_{n\to\infty} \prod_{i=1}^{n} \mathbb{P}(D_i^c) = \lim_{n\to\infty} \prod_{i=1}^{n} (1-p) = \lim_{n\to\infty} (1-p)^n = 0.$$

Therefore we will have a defective item with probability 1. Actually, it can be shown that there will be infinitely many defective items with probability 1. □

Example 4.3 *Independent trials that result in a success with probability p are successively performed until a total of r successes is obtained. Find the probability that exactly n trials are required.*

Solution: To have exactly n trials, the n^{th} trial should be a success. Among the previous $n-1$ trials, there should be exactly $r-1$ successes and there are $\binom{n-1}{r-1}$ many different ways to choose these successes. Since having a success and failure have probability p and $1-p$, respectively, and using the independence assumption we conclude that the probability that exactly n trials is given by

$$\binom{n-1}{r-1} p^r (1-p)^{n-r}.$$

□

Example 4.4 *Suppose that a family has exactly $n \geq 2$ children. Assume that the probability that any child will be a girl is $1/2$ and that all births are independent. Given that the family has at least one girl, determine the probability that the family has at least one boy.*

Solution: Let B and G be the number of boys and girls the family has. We are interested in the probability

$$p = \mathbb{P}(B \geq 1 \mid G \geq 1) = \frac{\mathbb{P}(B \geq 1, G \geq 1)}{\mathbb{P}(G \geq 1)}.$$

First observe

$$\mathbb{P}(G \geq 1) = 1 - \mathbb{P}(G = 0) = 1 - \mathbb{P}(\text{all boys}) = 1 - \frac{1}{2^n},$$

where the last equality here uses our independence assumption. Also

$$\mathbb{P}(B \geq 1, G \geq 1) = \mathbb{P}(G \geq 1) - \mathbb{P}(G \geq 1, B = 0) = 1 - \frac{1}{2^n} - \frac{1}{2^n} = 1 - \frac{1}{2^{n-1}}.$$

Hence,

$$p = \frac{1 - \frac{1}{2^{n-1}}}{1 - \frac{1}{2^n}} = \frac{2^n - 2}{2^n - 1}.$$

□

Remark 4.3 *In following sections we will also consider the independence of an infinite sequence $(A_n)_{n\geq 1}$ of events. By this, we mean that for any finite subset I of $\mathbb{N}$, the events in $\{A_i : i \in I\}$ are independent.*

There is a variant of the independence concept we defined which may sometimes help one to prove certain results under weaker assumptions. Though not much of an importance for this course, let's briefly go over it.

Definition 4.3 *The k events $A_1, \ldots, A_k$ are said to be **pairwise independent** if for any $i \neq j$ in $\{1, \ldots, k\}$,*

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j).$$

Clearly, for $k = 2$ independence and pairwise independence coincide. More generally:

Theorem 4.2 *Independence implies pairwise independence. The converse is not true in general.*

Proof: First claim is clear. For the second one, we provide an example. Suppose that a fair coin is tossed twice so that the sample space S is $\{HH, HT, TH, TT\}$. Let A be the event that we have H at the first toss, B the event that we have H at the second toss and C the event that both tosses are the same. Then one can easily check that A, B, C are pairwise independent, but not independent - check it. □

5 Bayes' theorem

Suppose that we are interested in which of several disjoint events $B_1, \ldots, B_k$ (that form a partition of the sample space) will occur and that we will get to observe some other event A. If $\mathbb{P}(A \mid B_i)$ is available for each i, then Bayes' theorem is a useful tool for computing the conditional probabilities of the B_i events given A. Here is the formal statement.

Theorem 5.1 *(Bayes' theorem) Let the events $B_1, \ldots, B_k$ form a partition of the sample space S such that $\mathbb{P}(B_j) > 0$ for each $j = 1, \ldots, k$, and let A be an event such that $\mathbb{P}(A) > 0$. Then, for $i = 1, \ldots, k$,*

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{k} \mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}. \quad (1)$$

Proof: We have

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{k} \mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}.$$

Here, the second equality followed by using the law of total probability. □

Example 5.1 *Consider three boxes, one with 1 black and 1 white marble, second one with 2 black and 1 white marble, and the last one with 2 black and 6 white balls. A box is selected at random, and a marble is drawn at random from the selected box.*

i. What is the probability that the marble is black?

ii. What is the probability that the first box was the one selected, given that the marble is white?

Solution: Let B, W be the events that the drawn marble is black and white, respectively. Let A_1, A_2, A_3 be the events that the first, second and third boxes are chosen, respectively.

(i.) $\mathbb{P}(B) = \mathbb{P}(B \mid A_1)\mathbb{P}(A_1) + \mathbb{P}(B \mid A_2)\mathbb{P}(A_2) + \mathbb{P}(B \mid A_3)\mathbb{P}(A_3) = \frac{1}{3}\frac{1}{2} + \frac{1}{3}\frac{2}{3} + \frac{1}{3}\frac{2}{8} = \frac{34}{72} = \frac{17}{36}$.

(ii.) $\mathbb{P}(A_1 \mid W) = \frac{\mathbb{P}(W|A_1)\mathbb{P}(A_1)}{\mathbb{P}(W)} = \frac{\mathbb{P}(W|A_1)\mathbb{P}(A_1)}{1-\mathbb{P}(B)} = \frac{4}{19}$. □

Example 5.2 *(Spam filter) Suppose that the probability that a "random" e-mail is spam is 80%*[7]*. Also assume that the probability that a spam e-mail contains the word* congratulations *with 10% and a non-spam e-mail with 5%. Find the probability that an e-mailed you received containing the word* congratulations *is a spam e-mail.*

Solution: Let S be the event that the e-mail is spam, and C be the event that it contains the word *congratulations*. We are interested in finding $\mathbb{P}(S \mid C)$. We have

$$\mathbb{P}(S \mid C) = \frac{\mathbb{P}(C \mid S)\mathbb{P}(S)}{\mathbb{P}(C)} = \frac{\mathbb{P}(C \mid S)\mathbb{P}(S)}{\mathbb{P}(C \mid S)\mathbb{P}(S) + \mathbb{P}(C \mid S^c)\mathbb{P}(S^c)} = \frac{(0.1)(0.8)}{(0.1)(0.8) + (0.05)(0.2)} = \frac{80}{81}.$$

□

Exercise 5.1 *(Monty Hall problem) Assume that a room is equipped with three doors. Behind two are goats, and behind the third is a shiny new car. You are asked to pick a door, and will win whatever is behind it. Let's say you pick door 1. Before the door is opened, however, someone who knows what's behind the doors (Monty Hall) opens one of the other two doors, revealing a goat, and asks you if you wish to change your selection to the third door (i.e., the door which neither you picked nor he opened). Does changing the door you selected increase the probability of getting the car? (Answer: Yes, it does. The initial probability of getting the car is* $1/3$*. Once Monty opens the door and we change our selection, this probability will increase.)*

Before continuing the probabilistic framework we stop for some statistical terminology in following few sections.

[7] The actual proportion of spam e-mails is around 88%!

6 Histograms

A histogram is a graphical representation of the distribution of numerical data, providing an estimate of a probability distribution of a quantitative (numeric) variable. It was first made use by Karl Pearson. To construct a histogram, we first "bin" the range of values, i.e., we divide the entire range of values into a series of intervals. Then we count how many values fall into each of these intervals.

Example 6.1 *Consider the following grade counts in a mass course at Boğaziçi University*[8]*:*

Grade	***Count***
0	*94*
1	*32*
1.5	*109*
2	*180*
2.5	*132*
3	*34*
3.5	*4*
4	*90*

,

Here we have 8 bins, and these correspond to 0 = F, 1 = DD,...,4 = A. The corresponding histogram in this case looks like

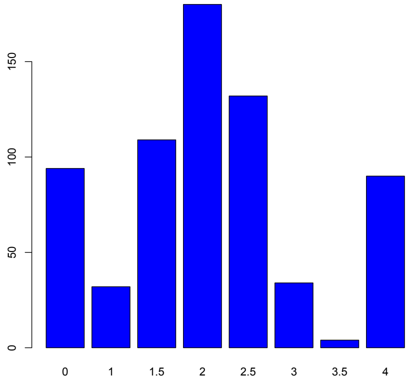


Figure 1: Grade counts

[8] Totally made up data

In general, we will be using histograms to approximate probability distributions. For our particular example case we could normalize the given histogram by dividing the counts for the individual grades by the total number of grades and could get a probability distribution.

□

Histograms are one of the most often encountered graphical tools in statistics, so you should get used to reading them[9]. In general, you can learn (or have a rough idea) about various features of the data distribution via histograms. Some of these are[10]:

- center of the data;
- spread of the data;
- skewness of the data;
- kurtosis of the data;
- presence of outliers;
- presence of multiple modes in the data.

7 Features of Data Distributions

In this section, we define various properties of data distributions. These help us to have some immediate observations about the data we have by just looking at the corresponding histogram, say.

1. The **center** of a data set is a number that represents the "middle" of the data. There are various ways to rigorously define a center; the sample mean, the sample median, trimmed mean, and so on[11].
2. A data distribution is said to be **symmetric** if the corresponding histogram is symmetric around its center.
3. The distribution is said to be **right-skewed** (left-skewed, resp.) if the right tail (left tail, resp.) seems to be stretched.
4. Another feature related to the data set is its kurtosis. This is a measure for how "peaked" the data is. If the distribution has a flat shape (no peak), then the data is said to be **platykurtic**. On the other hand, if the distribution has a steep peak, and then a heavy tail, it is called **leptokurtic**. The in between case is called **mesokurtic**.

[9] - and producing them, say, in R.

[10] Definitions of the terminology here will be introduced in next section

[11] We will learn about these in later sections.

5. The data is said to have **outliers** if there are data points that are far from the "typical" values of the data. Some reasons that may yield outliers; there can be a measurement error, or the outlier may not belong to data of interest, or maybe the data belongs to the population but it points out a deeper phenomenon.

6. **Spread** of a data set is a measure of variability of the data around its center. Again, there are multiple ways to define this., but intuitively data sets with small spread will be concentrated in a small strip containing the center.

7. A **clustering** in a data set is the clumping of data points around distinct values. In this case, the regions between different clusters are called **gaps**.

8. A **unimodal** distribution is a distribution with one clear peak. If the distribution has more than one peaks, then it is said to be **multimodal**.

8 Sampling statistics

Consider a data set $x_1, \ldots, x_n$ consisting of real numbers.

1. The **sample mean** of a data distribution is one measure of the center defined by
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

2. The **sample variance** is
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$
This measures the spread of the data around its center - which is the sample mean for this particular case.[12] s is called the **sample standard deviation**.

3. Given the data set $x_1, \ldots, x_n$, if we order them from smallest the largest so that
$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)},$$
then $x_{(1)}, \ldots, x_{(n)}$ are called the corresponding **order statistics**. In particular, $x_{(k)}$ is called the k^{th} **order statistic**, and approximately $100(k/n)\%$ of the observations fall to the left of $x_{(k)}$.

4. Based on order statistics, we define $x_{(1)}$ and $x_{(n)}$ to be the **sample min** and the **sample max** of the data distribution.

[12] We will explain why we use $n-1$ instead of n later.

Next, we will define the sample median and the sample trimmed mean. In order to motivate these definitions, let's note that sample mean is not robust to outliers. For example, if the data points are $\mathbf{x} = (1.5, 0.2, 1, 0, 9, 1.1, 1.3, 0.5, 2.7, 0.5, 0.3)$, then the sample mean is $11/10 = 1.1$. If we now replace only one of these data points, say 0.2 to 100.2, then the sample mean in this case would be 10.2 causing a significant change in the center. Such a significant change in the sample mean is explained in statistical terminology with the notion of robustness: Sample mean is not robust to outliers. We will now handle this issue by using the sample median or the sample trimmed mean as the center.

5 The **sample median** of the data set $x_1, \ldots, x_n$ is defined to be

$$med(x_1, \ldots, x_n) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2}, & \text{if } n \text{ is even.} \end{cases}$$

As an example, if we again consider $\mathbf{x} = (1.5, 0.2, 1, 1, 9, 1.1, 1.3, 0.5, 2.7, 0.5, 0.3)$, replacing 0.2 by 100.2 does not even change the sample median. For the reasons discussed previously we then say that the sample median is more robust than the sample mean. Another statistic that is more robust compared to the sample mean is the trimmed mean.

6 The **trimmed mean** is computed just as an ordinary mean except first a pre-specified percentage of the extremes is omitted. In particular, for the $m\%$ trimmed mean, the left-most (lowest) $m\%$ and right-most (highest) $m\%$ of the data are excluded; from the remaining observations the mean is found. For example, if the data we have is $\mathbf{x} = (1, 2, 3, 3, 4, 4, 4, 6, 8, 11)$, and if we are interested in the 20% trimmed mean, then we will discard 1, 2, 8, 11 and compute the mean of the remaining numbers. In this case, it turns out to be

$$\overline{x}_{.2} = \frac{3 + 3 + 4 + 4 + 4 + 6}{6} = 4.$$

Exercise 8.1 *Consider the data*

$$5,\ 4,\ 7,\ 6,\ 8,\ 10,\ 11,\ 0,\ 7,\ 18.$$

i. Find the sample mean and sample variance;

ii. Write the order statistics corresponding to this data set;

iii. Find the 20% trimmed mean;

iv. Find the sample median;

v. Explain with modifications of this data set why the trimmed mean and the median are more robust than the sample mean.

9 Review problems 1

Exercise 9.1 *Consider two events* A *and* B *so that* $\mathbb{P}(A) = 1/3$, $\mathbb{P}(B) = 1/2$. *Determine* $\mathbb{P}(B \cap A^c)$ *when*

i. A *and* B *are disjoint.*

ii. $A \subset B$.

iii. $\mathbb{P}(A \cap B) = \frac{1}{8}$.

Exercise 9.2 *If* $\mathbb{P}(A) = 1/3$ *and* $\mathbb{P}(B^c) = 1/4$, *can* A *and* B *be disjoint? Explain.*

Exercise 9.3 *Consider two events* A *and* B *with* $\mathbb{P}(A) = 0.4$ *and* $\mathbb{P}(B) = 0.7$. *Determine the maximum and minimum possible values of* $\mathbb{P}(A \cap B)$ *and the conditions under which each of these values is attained.*

Exercise 9.4 *Let* $\mathbb{P}(A) = 0.3$ *and* $\mathbb{P}(B) = 0.6$.
(a) Find $\mathbb{P}(A \cup B)$ *when* A *and* B *are independent.*
(b) Find $\mathbb{P}(A \mid B)$ *when* A *and* B *are disjoint.*

Exercise 9.5 *If the probability that student A will fail a certain statistics examination is 0.5, the probability that student B will fail the examination is 0.2, and the probability that both student A and student B will fail the examination is 0.1, what is the probability that at least one of these two students will fail the examination?*

Exercise 9.6 *Two students A and B are both registered for a certain course. Assume that student A attends class 80 percent of the time, student B attends class 60 percent of the time, and the absences of the two students are independent.*
a. What is the probability that at least one of the two students will be in class on a given day?
b. If at least one of the two students is in class on a given day, what is the probability that A is in class that day?

Exercise 9.7 *A box contains* r *red balls and* b *blue balls. One ball is selected at random and its color is observed. The ball is then returned to the box and* k *additional balls of the same color are also put into the box. A second ball is then selected at random, its color is observed, and it is returned to the box together with* k *additional balls of the same color. Each time another ball is selected, the process is repeated. If four balls are selected, what is the probability that the first three balls will be red and the fourth ball will be blue?*

Exercise 9.8 *Suppose that a fair n-sided die is rolled n independent times. A* ***match*** *occurs if side i is observed on the ith trial, $i = 1, 2, \ldots, n$.*

(a) Show that the probability of (at least) one match is

$$1 - \left(\frac{n-1}{n}\right)^n .$$

(b) Find the limit of this probability as n increases without bound.

Exercise 9.9 *Consider the birthdays of the students in a class of size consists of 365 days.*

(a) How many different ordered outcomes of birthdays are possible (for r students) allowing repetitions (with replacement)?

(b) The same as part (a) except requiring that all the students have different birthdays (without replacement)?

(c) If we can assume that each ordered outcome in part (a) has the same probability, what is the probability that at least two students have the same birthday?

(d) For what value of number of students is the probability in part (c) surprisingly high? (Use a calculator or a graphical software for this part.)

Exercise 9.10 *My phone rings 12 times each week, the calls being randomly distributed among the 7 days. What is the probability that I get at least one call each day?*

Exercise 9.11 *An insurance company has three types of customers - high risk, medium risk and low risk.* 20% *of its customers are high risk,* 30% *are medium risk, and* 50% *are low risk. Also, the probability that a customer has at least one accident in the current year is* 0.25 *for high risk,* 0.16 *for medium risk, and* 0.10 *for low risk.*

a. *Find the probability that a customer chosen at random will have at least one accident in the current year.*

b. *Find the probability that a customer is high risk, given that the person has had at least one accident during the current year.*

Exercise 9.12 *A pair of events A and B can not be simultaneously mutually exclusive*[13] *and independent. Prove that if $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, then:*

(a) If A and B are mutually exclusive, they cannot be independent.

(b) If A and B are independent, they cannot be mutually exclusive.

Exercise 9.13 *A fair die is cast until a 6 appears. What is the probability that it must be cast more than five times?*

[13] "Mutually exclusive" of A, B refers to disjointness of them.

Exercise 9.14 *The Smiths have two children. At least one of them is a boy. What is the probability that both children are boys? (See Gardner 1961 for a complete discussion of this problem.)*

Exercise 9.15 *Standardized tests provide an interesting application of probability theory. Suppose first that a test consists of 20 multiple-choice questions, each with 4 possible answers. If the student guesses on each question, then the taking of the exam can be modeled as a sequence of 20 independent events. Find the probability that the student gets at least 10 questions correct, given that he is guessing.*

Exercise 9.16 *Suppose that a fair coin is tossed independently until a head is obtained, and that this entire experiment is then performed independently a second time. What is the probability that the second experiment requires more tosses than the first experiment?*

Exercise 9.17 *Suppose that 5 percent of men and* $.25$ *percent of women are color blind. A colorblind person is chosen at random. What is the probability of this person being male? Assume that there are an equal number of males and females? What if the population consisted of twice as many males as females?*

Exercise 9.18 *You ask your neighbor to water a sickly plant while you are on vacation. Without water it will die with probability* $.8$*; with water it will die with probability* $.15$*. You are 90 percent certain that your neighbor will remember to water the plant.*

a. What is the probability that the plant will be alive when you return?

b. If it is dead, what is the probability your neighbor forgot to water it?

Exercise 9.19 *In an ESP experiment, one person chooses two different numbers between 1 and 10, inclusive. A second person in a different room then attempts to guess them. Suppose the second person correctly guesses the two numbers (Order is not important). What is the probability that this happens by pure chance?*

Exercise 9.20 *If* n *balls are placed at random into* n *cells, find the probability that exactly one cell remains empty.*

Exercise 9.21 *A man is given* n *keys, in random order, of which only one will unlock his door. He tries them successively (sampling without replacement). This procedure may require* $1, 2, \ldots, n$ *trials. Show that each of the* n *outcomes has probability* $1/n$*. (In other words,* $\mathbb{P}(\textit{man succeeds on } k\textit{th trial}) = 1/n$, $k = 1, 2, \ldots, n$*.)*

Exercise 9.22 *An employer is about to hire one new employee from a group of* N *candidates, whose future potential can be rated on a scale from 1 to* N*. The employer proceeds according to the following rules:*

a. *Each candidate is seen in succession (in random order) and a decision is made whether to hire the candidate.*

b. *Having rejected $m-1$ candidates $(m>1)$, the employer can hire the mth candidate only if the mth candidate is better than the previous $m-1$*

Suppose a candidate is hired on the ith trial What is the probability that the best candidate was hired?

Exercise 9.23 *In a town of $n+1$ inhabitants, a person tells a rumor to a second person, who in turn repeats it to a third person, etc. At each step the recipient of the rumor is chosen at random from the n people. Find the probability that the rumor will be told exactly r times*

a. *before returning to the originator;*

b. *without being repeated to any person.*

Exercise 9.24 *Two players, A and B, alternately and independently flip a coin and the first player to obtain a head wins. Assume player A flips first.*

a. *If the coin is fair, what is the probability that A wins?*

b. *Assume that $\mathbb{P}(head) = p$, not necessarily $1/2$. What is the probability that A wins?*

c. *Show that for all p, $0 < p < 1$, $\mathbb{P}(A\ wins) > 1/2$.*

Exercise 9.25 *(Coupon collector's problem) Suppose that each package of bubble gum contains the picture of a baseball player, that the pictures of r different players are used, that the picture of each player is equally likely to be placed in any given package of gum, and that pictures are placed in different packages independently of each other. Find the probability p that a person who buys n packages of gum $(n \geq r)$ will obtain a complete set of r different pictures.*

Exercise 9.26 *(*) Let $(A_\beta)_{\beta \in B}$ be a family of pairwise disjoint events. Show that if $\mathbb{P}(A_\beta) > 0$, each $\beta \in B$, then B must be countable.*

10 Discrete random variables

Definition 10.1 *We call a real-valued function that is defined on a sample space S a* ***random variable****.*

Example 10.1 *Consider an experiment in which a fair coin is tossed 10 times independently. In this experiment, the sample space S can be regarded as the set of outcomes consisting of the 2^{10} different sequences of heads and tails that are possible. Some random variables defined on this space are the number of heads we obtain, X, the number of tails we obtain, Y, or perhaps the difference between these two, $Z = X - Y$.*

Let's have closer look at X. In this case, the sample space can be given by

$$S = \{w_1, \ldots, w_{10} : w_j \in \{\textit{Head, Tail}\}, j = 1, \ldots, 10\}.$$

Then, letting

$$\mathbf{1}(A) = \begin{cases} 1, & \textit{if } A \textit{ occurs} \\ 0, & \textit{otherwise,} \end{cases}$$

then

$$X = \sum_{i=1}^{10} \mathbf{1}(\textit{ith one is head}) = \sum_{i=1}^{10} \mathbf{1}(w_i = \textit{head}).$$

Note here that the function $\mathbf{1}(\cdot)$ is called the ***indicator function*** *throughout these notes.* □

Example 10.2 *Suppose that I toss a die m many times. Letting X_i be the number of spots I obtain in i^{th} roll, $i = 1, 2, \ldots, m$, and setting $Z = X_1 + \cdots + X_m$, Z is a random variable on the sample space*

$$S = \{(i_1, i_2, \ldots, i_m) : i_j \in \{1, 2, 3, 4, 5, 6\}, \textit{for all } j = 1, \ldots, m\}.$$

□

Example 10.3 *We may consider the stock price tomorrow of a certain company as a random variable. Note that in this case the sample space is very difficult to describe. Fortunately, instead of understanding the sample space, we will just need to focus on the possible values of the random variable.* □

Definition 10.2 *Let X be a random variable. The* ***distribution*** *of X is the collection of all probabilities of the form $\mathbb{P}(X \in E)$ for all sets E of real numbers such that $\{X \in E\}$ is an event.*

Definition 10.3 *i. A random variable X is said to have a* ***discrete distribution*** *or that X is a* ***discrete random variable*** *if X can take only a finite number k of distinct values $x_1, \ldots, x_k$ or, at most, an infinite sequence of distinct values $x_1, x_2, \ldots$.*

ii. *If a random variable X has a discrete distribution, the* ***probability mass function*** *(abbreviated* ***pmf****) of X is defined as the function f such that for every real number x,*

$$f(x) = \mathbb{P}(X = x).$$

iii. *The closure[14] of the set $\{x : f(x) > 0\}$ is called the* ***support*** *of (the distribution of) X. We denote the support of X by $s(X)$*

Example 10.4 *Here are some examples of discrete random variables:*

1. *Number of students that are to attend Ümit's next Math 344 lecture.*
2. *Number of goals that are to be scored by Fenerbahçe in its next game.*
3. *Number of e-mails you get in a day.*
4. *Number of ants on our campus right now.*
5. *Number of times you get heads if you toss a coin n times.*
6. *Number of times you need to toss a fair coin until you get a head.*

Some comments

a. Example 6 is different than the other ones since the support in this case is countable. You may wait an arbitrarily large number of times in order to get your first head.

b. In some cases, you may model a data that seemingly can take only finitely many values using a probability distribution with denumerable support. For example, consider the number of ants on our campus. How would you model this? What would be the maximum number of ants if you decide to use a bounded support distribution? 10^7? Besides the difficulty in having an upper bound for this number, it is usually a lot more convenient (both from math and practical point of views) to study this problem with a distribution with denumerable support. This is due to skipping various mathematical technicalities by changing the probabilities only slightly (since the probabilities of extremes we include in denumerable setting will almost be negligible).

c. We usually write capital letters to denote random variables. For example, we could write $X =$ Number of students that are to attend Ümit's next Math 344 lecture, $Y =$ Number of times you get heads if you toss a coin n times, etc.

[14] Recall that the closure of $A \subset \mathbb{R}$ is the set A union its limit points.

Example 10.5 *Suppose that I roll an unbalanced die, and let X be the number of spots I obtain. Then some exemplary probabilities of interest could be*

$$\mathbb{P}(X \geq 3) = \frac{\mathbb{P}(\text{having } 3, 4, 5 \text{ or } 6 \text{ spots})}{6} = \frac{4}{6} = \frac{2}{3},$$

$$\mathbb{P}(X \text{ is even}) = \frac{\mathbb{P}(\text{having } 2, 4 \text{ or } 6 \text{ spots})}{6} = \frac{1}{2},$$

$$\mathbb{P}(X \leq 5) = \frac{\mathbb{P}(\text{having } 1, 2, 3, 4 \text{ or } 5 \text{ spots})}{6} = \frac{5}{6}.$$

Note that in the last case we could find $\mathbb{P}$(having at most five spots) by finding the probability of the complementary event {having exactly one spot} and subtracting that from 1. □

Example 10.6 *Suppose that I am rolling an unbalanced die twice, and let X_1, X_2 be the number of spots I obtain in first and second rolls.*

Find the following probabilities:
(a) $\mathbb{P}(X_1 > 3)$.
(b) $\mathbb{P}(X_1 + X_2 > 3)$.
(c) $\mathbb{P}(X_1 X_2 > 30)$.

Solution: The support of X_1 and X_2 are both the sets $\{1, 2, \ldots, 6\}$, and therefore when we look at the vector (X_1, X_2), it takes each value in $\{(i, j) : 1 \leq i, j \leq 6\}$ with equal probability, 1/36.

(a) Similar to our previous example, $\mathbb{P}(X_1 > 3) = 3/6 = 1/2$.

(b) We have

$$\mathbb{P}(X_1 + X_2 > 3) = 1 - \mathbb{P}(X_1 + X_2 = 2) - \mathbb{P}(X_1 + X_2 = 3) = 1 - \frac{1}{36} - \frac{2}{36} = \frac{11}{12}.$$

(c) Observe that

$$\mathbb{P}(X_1 X_2 > 30) = \mathbb{P}(X_1 = X_2 = 6) = \frac{1}{36}.$$

□

Now going back to pmfs, the following fact should be clear.

Fact: If $f(x)$ is the pmf of some random variable X, then

$$f(x) \geq 0, \quad \text{for all } x, \qquad \text{and} \qquad \sum_{x \in s(X)} f(x) = 1.$$

Example 10.7 *Suppose that the pmf of a discrete random variable X is given by:*

$$f(x) = \begin{cases} \frac{c}{3^x}, & \textit{if } x = 0, 1, 2, \dots, \\ 0, & \textit{otherwise} \end{cases}$$

(i.) Find c.
(ii.) Find $\mathbb{P}(X \geq 2)$.

Solution: (i.) Note that we should have

$$\sum_{x=0}^{\infty} \frac{c}{3^x} = 1.$$

Now since

$$\sum_{n=0}^{\infty} \frac{c}{3^x} = c \sum_{n=0}^{\infty} \frac{1}{3^x} = c \frac{1}{1 - 1/3} = \frac{3}{2} c,$$

we conclude that c should be $2/3$.

(ii.) $\mathbb{P}(X \geq 2) = 1 - f(0) - f(1)$. □

11 Expectation of discrete random variables

Definition 11.1 *Let X be a bounded*[15] *discrete random variable whose pmf is f. The **expectation** of X, denoted $\mathbb{E}[X]$, is defined by*

$$\mathbb{E}[X] = \sum_{x \in s(X)} x f(x),$$

where we recall $s(X)$ is the support of X.

Example 11.1 *Let X have the Bernoulli distribution with parameter p; that is,*

$$\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = p.$$

Then

$$\mathbb{E}[X] = 1 \times p + 0 \times (1 - p) = p.$$

□

Example 11.2 *Let X be the outcome of rolling a fair die. Find $\mathbb{E}[X]$.*

[15] X is bounded means that there exists some $M \in \mathbb{R}^+$ such that $|X(w)| \leq M$ for any $w \in S$.

Solution: We have

$$\mathbb{E}[X] = \sum_{x=1}^{6} x f(x) = \sum_{x=1}^{6} \frac{1}{6} x = \frac{1}{6}(1 + \cdots + 6) = \frac{7}{2}.$$

Note that

i 7/2 is just the midpoint of the line segment $[1, 6]$. This is natural if we consider the current problem in terms of finding the center of mass of certain unit masses placed at $x = 1, \ldots, 6$. Can you generalize this center of mass idea to expectations of more general random variables?

ii. This example in particular shows that the expectation of an integer valued random variable is not necessarily an integer. My students somehow find this confusing. □

Here is another example which shows that finding the expectation can get more involved.

Example 11.3 *(Binomial distribution) Let X be a binomial random variable with parameters n, p so that its pmf f is given by*

$$f(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Find $\mathbb{E}[X]$.

Solution: We have

$$\begin{aligned}\mathbb{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} &= \sum_{k=1}^{n} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^{n} k \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\ &= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\ &= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-k-1} \\ &= np,\end{aligned}$$

where the last step follows from the binomial theorem. □

Exercise 11.1 *Suppose that I have a coin whose head probability is p, and I toss it n many times. Let X be the number of heads I obtain. Check that the pmf of X is the same as the pmf of a binomial distribution with parameters n and p*[16].

How do we define $\mathbb{E}[X]$ when X is not bounded?

Definition 11.2 *Let X be a discrete random variable whose pmf is f. Suppose at least one of the following sums is finite:*

$$\sum_{positive\ x} xf(x), \qquad \sum_{negative\ x} xf(x).$$[17]

Then the ***expected value (or expectation or (distributional) mean)*** *of X is said to exist and is defined to be*

$$\mathbb{E}[X] = \sum_{x \in supp(X)} xf(x).$$

If both of the sums above are infinite, then expectation of X is said ***not to exist****.*

Example 11.4 *(A) Let X be a random variable with pmf*[18]

$$f(x) = \begin{cases} \frac{6}{\pi^2 x^2}, & if\ x = 1, 2, 3, \ldots, \\ 0, & otherwise. \end{cases}$$

Then we have

$$\sum_{positive\ x} xf(x) = \sum_{positive\ x} \frac{6}{\pi^2 x^2} x = \frac{6}{\pi^2} \sum_{positive\ x} \frac{1}{x} = \infty,$$

and

$$\sum_{negative\ x} xf(x) = 0.$$

So $\mathbb{E}[X] = \infty$.

(B) This time let Y be a random variable with pmf

$$f(x) = \begin{cases} \frac{3}{\pi^2 x^2}, & if\ x = \pm 1, \pm 2, \pm 3, \ldots, \\ 0, & otherwise. \end{cases},$$

[16] This is the standard example to keep in mind for the binomial distribution. We will learn more later.

[17] Here, the summation $\sum_{\text{positive } x}$ means that the summation is over all $x \in s(X)$ that are positive.

[18] Keep in mind that $\sum_{n=1}^{\infty} \frac{1}{n} = \infty$ and $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$.

In this case

$$\sum_{\text{positive } x} xf(x) = \sum_{\text{positive } x} \frac{3}{\pi^2 x^2} x = \frac{3}{\pi^2} \sum_{\text{positive } x} \frac{1}{x} = \infty,$$

and

$$\sum_{\text{negative } x} xf(x) = \sum_{\text{negative } x} \frac{3}{\pi^2 x^2} x = \frac{3}{\pi^2} \sum_{\text{negative } x} \frac{1}{x} = -\infty.$$

So $\mathbb{E}[Y]$ does not exist. □

12 Basic properties of expectation

In this section, we discuss some properties of expectation. Though the discussion will be given for the discrete case, everything here can be adapted to more general cases as we shall see later. Let's begin with the following fact:

Fact: If X is a random variable with pmf f, and r is a real valued function defined on real numbers, then we define the expectation of $Y = r(X)$ by

$$\mathbb{E}[Y] = \mathbb{E}[r(X))] = \sum_{x \in s(X)} r(x)f(x).$$

Example 12.1 *Let X be a random variable with pmf $f(x) = 1/5$ for $x \in \{-2, -1, 0, 1, 2\}$.*
(i.) Find $\mathbb{E}|X|$.
(ii.) Find $\mathbb{E}[X^2 - 2]$.

Solution: *(i.) We have $\mathbb{E}|X| = \frac{1}{5}(-2 - 1 + 0 + 1 + 2) = 0.$.*
(ii.) Using the fact above with $r(x) = x^2 - 2$, we obtain

$$\mathbb{E}[X^2 - 2] = \sum_{x=-2}^{2} (x^2 - 2)\frac{1}{5} = \frac{1}{5}((4 - 2) + (1 - 2) + (0 - 2) + (1 - 2) + (4 - 2)) = 0.$$

□

Proposition 12.1 *If X is a discrete random variable with pmf f and $Y = aX + b$, where a, b are constants, then $\mathbb{E}[Y] = a\mathbb{E}[X] + b$.*

Proof: We have

$$\mathbb{E}[Y] = \sum_{x \in s(X)} (ax + b)f(x) = a \sum_{x \in s(X)} xf(x) + b \sum_{x \in s(X)} f(x) = a\mathbb{E}[X] + b.$$

□

Using Proposition 12.1 with $a = 0$ yields:

Corollary 12.1 *If $Y = b$ with probability 1, then $\mathbb{E}[Y] = b$.*

Following generalization is fundamental.

Theorem 12.1 *(Linearity of expectation) Assume $\mathbb{E}[X_i]$ is finite for $i = 1, \ldots, n$. Then for any constants $a_1, \ldots, a_n, b$ we have*

$$\mathbb{E}[a_1 X_1 + \cdots + a_n X_n + b] = a_1 \mathbb{E}[X_1] + \cdots + a_n \mathbb{E}[X_n] + b.$$

The proof of Theorem 12.1 is omitted for now.

Example 12.2 *Suppose that three random variables X_1, X_2, X_3 form a random sample from a distribution for which the mean is 5. Determine the value of $\mathbb{E}[2X_1 - 3X_2 + X_3 - 4]$.*

Solution: We have $\mathbb{E}[2X_1 - 3X_2 + X_3 - 4] = 2\mathbb{E}[X_1] - 3\mathbb{E}[X_2] + \mathbb{E}[X_3] - 4 = -4$. □

The linearity of expectation can be extremely useful in counting problems:

Example 12.3 *(Matching problem) n men throw their hats and then they choose one hat uniformly at random. Let X be the number of men who find their own hats. Find $\mathbb{E}[X]$.*[19]

Solution: Let

$$X_i = \begin{cases} 1 & \text{if } i\text{th man finds his hat,} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$X = \sum_{i=1}^{n} X_i.$$

Using linearity of expectation, we obtain

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \mathbb{P}(X_i = 1) = \sum_{i=1}^{n} \frac{1}{n} = 1.$$

□

The idea used in previous example is quite general, and is known as the "Method of Indicators" in literature. A little bit more discussion on this method will be the content of next section.

[19] Finding $\mathbb{E}[X]$ without the use of linearity is really challenging. Try it yourself.

13 * Method of indicators

Method of indicators is a general technique to compute expectations in discrete setting by using the linearity property. This is especially useful when the underlying summands, which are Bernoulli random variables, are "dependent". For example, considering Example 12.3, the Bernoulli random variables $X_1, \ldots, X_n$ are dependent since, for example, $X_1 = \ldots = X_{n-1} = 1$ implies $X_n = 1$. This sort of dependence makes the pmf quite involved and without using method of indicators, it usually is quite challenging to find the expectation. Here we discuss two more examples.

Example 13.1 *We consider a problem that was posed and solved in the eighteenth century by Daniel Bernoulli. Suppose a jar contains $2N$ cards, two of them marked 1, two marked 2, two marked 3, and so on. Draw out m cards at random. What is the expected number of pairs that still remain in the jar? (Bernoulli proposed the above as a possible probabilistic model for determining the number of marriages that remain intact when there is a total of m deaths among N married couples.)*

Solution: Let

$$X_i = \begin{cases} 1 & \text{if } i\text{th pair remains in the jar,} \\ 0 & \text{otherwise.} \end{cases}$$

Then $X = \sum_{i=1}^{N} X_i$ is the number of pairs that remain in the jar. We have

$$\mathbb{E}[X] = \sum_{i=1}^{N} \mathbb{E}[X_i] = \sum_{i=1}^{N} \mathbb{P}(X_i = 1) = \sum_{i=1}^{N} \frac{\binom{2N-2}{m}}{\binom{2N}{m}} = N \frac{\binom{2N-2}{m}}{\binom{2N}{m}}.$$

(What are the conditions on m?) □

Example 13.2 *(Coupon collector problem) There are n types of coupons. Each newly obtained coupon is, independently, type i with probability p_i, $i = 1, \ldots, n$. Let X be the number of distinct types obtained in a collection of k coupons. Find the expected value of X.*

Solution: *Let*

$$X_i = \begin{cases} 1 & \textit{if ith coupon is in our sample,} \\ 0 & \textit{otherwise.} \end{cases}$$

Then $X = \sum_{i=1}^{n} X_i$, and

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \mathbb{P}(X_i = 1) = \sum_{i=1}^{n} (1 - \mathbb{P}(X_i = 0)) = \sum_{i=1}^{n} (1 - (1 - p_i)^k).$$

□

14 Independent random variables

Definition 14.1 *Two random variables X and Y are* ***independent*** *if for all $A, B \subset \mathbb{R}$ so that $\{X \in A\}$ and $\{Y \in B\}$ are events, we have*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B). \tag{2}$$

Example 14.1 *Suppose that X and Y denote the results of two coin tosses. We write 1 for a toss resulting a head, and 0 otherwise. Assume that*

$$\mathbb{P}(X = i, Y = j) = \frac{1}{4}, \quad i, j \in \{0, 1\}.$$

In this case

$$\mathbb{P}(X = 1) = \mathbb{P}(X = 1, Y = 0) + \mathbb{P}(X = 1, Y = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$

and

$$\mathbb{P}(X = 0) = 1 - \mathbb{P}(X = 1) = \frac{1}{2}.$$

Similarly,

$$\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = \frac{1}{2}.$$

From these, one can easily check that the condition in (13) is satisfied and so A and B are independent. □

Example 14.2 *Let X be a uniform random variable over $\{-1, 0, 1\}$, i.e. X takes the values $-1, 0, 1$ with equal porbabilities. Also set $Y = |X|$. It is intuitively clear that X and Y are dependent. To give a rigorous justification of this, let $A = (1/2, \infty)$ and $B = (-1/2, 1/2)$. Then*

$$\mathbb{P}(X \in A, Y \in B) = 0$$

but

$$\mathbb{P}(X \in A)\mathbb{P}(Y \in B) > 0,$$

and so X and Y can not be independent. □

In general, it is enough to find two subsets of the real line for which the condition in (13) is violated to show that X and Y are dependent. On the other hand, showing that the random variables are independent can be challenging since we need to check the condition for **all** intervals. Later in Section 25, we will learn a useful criteria, the factorization theorem, for proving independence of random variables.

The definition for independence of more than two random variables is as expected.

Definition 14.2 *The random variables $X_1, \ldots, X_n$ are said to be* ***independent*** *if for $A_1, \ldots, A_n \subset \mathbb{R}$ and for any $I \subset \{1, 2, \ldots, n\}$, we have*

$$\mathbb{P}\left(\bigcap_{i \in I} \{X_i \in A_i\}\right) = \prod_{i \in I} \mathbb{P}(X_i \in A_i).$$

We leave the theory for independent random variables for now, though, we shall see various related results in following sections.

15 Variance

The purpose of this section is to understand the variation a random variable has around its mean. In a certain sense, this measures the "randomness" of a random variable.

Definition 15.1 *Let X be a random variable with finite mean $\mu = \mathbb{E}[X]$. The* ***variance*** *of X, denoted by $Var(X)$, is defined by*

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

The ***standard deviation*** *of X is $\sqrt{Var(X)}$.*

Here, $|X(w) - \mu|^2$ is the deviation of our random variable when $w \in S$ occurs, and when we take the expectation we look at the average quadratic deviation around the mean.

Note that using linearity of expectation

$$Var(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

This provides an alternative expression for the variance.

Proposition 15.1 $Var(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$

Example 15.1 *Let X be a discrete random variable with pmf $f(x) = 1/5$ for $x \in \{-2, 0, 1, 3, 4\}$. Find $Var(X)$.*

Solution: It is easily seen that $\mathbb{E}[X] = 1.2$. Also

$$\mathbb{E}[X^2] = \frac{1}{5}(4 + 0 + 1 + 9 + 16) = 6.$$

Hence $Var(X) = 6 - (1.2)^2 = 4.56$ and $SD(X) \approx 2.135$.

Here, the distribution of X is said to be uniform over the set $\{-2, 0, 1, 3, 4\}$ since each result has equal probability. It can be shown that among all distributions over this set, uniform distribution has the maximal variance. This should be intuitively clear because the uniform distribution is the most "random" distribution over the given set. □

The next example is included in order to demonstrate that the variance and the standard deviation provide measures of spread around the mean.

Example 15.2 *Let the random variable X have pmf*

$$f(x) = \begin{cases} 0.5, & \textit{if } x = 0 \\ 0.499, & \textit{if } x = 1 \\ 0.001, & \textit{if } x = 10,000 \\ 0, & \textit{otherwise.} \end{cases}$$

Then, you can easily check that $Var(X) = 99,890.27$. If we remove the small mass at $10,000$ to get a new random variable Y with pmf

$$f(x) = \begin{cases} 0.5, & \textit{if } x = 0, 1, \\ 0, & \textit{otherwise,} \end{cases}$$

then the variance becomes 0.25. In other words, the spread around the mean diminished significantly after our small *change.* □

Let's next discuss some basic properties of the variance. The first two of these follow immediately from the definitions - check them.

Proposition 15.2 *For any random variable X, $Var(X) \geq 0$.*

This in particular says that $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ for any random variable X.

Proposition 15.3 *For a random variable X, $Var(X) = 0$ if and only if $\mathbb{P}(X = b) = 1$ for some constant b.*

When $Var(X) = 0$, we call X a **deterministic (random) variable**.

Proposition 15.4 *If X, Y are two random variables so that $Y = aX + b$ for some constants a and b, then*

$$Var(Y) = a^2 Var(X).$$

Proof: Letting $\mu = \mathbb{E}[X]$, we have

$$Var(Y) = \mathbb{E}[((aX+b)-(a\mu+b))^2] = a^2\mathbb{E}[(X-\mu)^2] = a^2Var(X).$$

□

Remark 15.1 *In particular, $Var(X+b) = Var(X)$ and $Var(X) = Var(-X)$ - translations and changes in sign do not change the variance (in contrast to the mean). Also, $Var(aX+b) = a^2Var(X)$.*

More generally, we have the following nice property for linear combinations of independent random variables.

Theorem 15.1 *If $X_1, \ldots, X_n$ are independent random variables, and if $a_1, \ldots, a_n, b$ are constants, then*

$$Var(a_1X_1 + \cdots + a_nX_n + b) = \sum_{i=1}^{n} a_i^2 Var(X_i).$$

We will prove Theorem 15.1 as a special case of Theorem 29.5 later.

Example 15.3 *Suppose X and Y are independent, $Var(X) = Var(Y) = 3$. Find $Var(X-Y)$ and $Var(2X-3Y+1)$.*

Solution: We have

$$Var(X-Y) = Var(X) + Var(Y) = 6$$

and

$$Var(2X-3Y+1) = 2^2Var(X) + (-3)^2Var(Y) = 39.$$

□

16 Uniform distribution

Definition 16.1 *A random variable X is said to have the* ***discrete uniform distribution*** *over the set $\mathcal{A} = \{a_1, \ldots, a_k\}$ if its pmf is given by*

$$f(x) = \mathbb{P}(X = x) = \begin{cases} \frac{1}{k}, & \textit{for } x \in \mathcal{A} \\ 0, & \textit{otherwise.} \end{cases}$$

In this case, we write $X \sim U(\mathcal{A})$[20].

[20] In general, we read $X \sim D$ as X is a random variable with distribution D.

Example 16.1 *If we roll a balanced die and let X be the outcome, then X is a uniform random variable on $\{1,2,3,4,5,6\}$, i.e. $X \sim U(\{1,2,3,4,5,6\})$.*

We may roll the same die n many times independently, and let $X_1,\dots,X_n$ be the corresponding outcomes. In this case defining $X=(X_1,\dots,X_n)$, $X \sim U(\{1,2,3,4,5,6\}^n)$. Note however that X is not real valued here, for that reason we call X a random vector. More discussion on random vectors (or, multivariate random variables) will be below. □

Example 16.2 *Let k be a positive integer and let $X \sim U(\{1,2,\dots,k\})$.*
(i) Find $\mathbb{E}[X]$;
(ii) Find $Var(X)$;

Solution. (i) We have

$$\mathbb{E}[X]=\sum_{i=1}^{k} i\frac{1}{k}=\frac{1}{k}\sum_{i=1}^{k} i=\frac{1}{k}\frac{k(k+1)}{2}=\frac{k+1}{2}.$$

(ii) First observe that

$$\mathbb{E}[X^2]=\sum_{i=1}^{k} i^2\frac{1}{k}=\frac{1}{k}\sum_{i=1}^{k} i^2=\frac{1}{k}\frac{k(k+1)(2k+1)}{6}=\frac{(k+1)(2k+1)}{6}.$$

So

$$Var(X)=\mathbb{E}[X^2]-(\mathbb{E}[X])^2=\frac{(k+1)(2k+1)}{6}-\left(\frac{k+1}{2}\right)^2=\frac{k^2-1}{12}.$$

□

Remark 16.1 *Note that we may also define a uniform distribution on the set S_n of all permutations with n, on all labeled trees with n vertices, on all graphs with n vertices where each vertex has degree at most three, and so on. Although we will not be interested in this type of situation for this course, if you take Math 345 then you will see uniform distribution appearing in various distinct contexts.*

17 Bernoulli and binomial distributions

Bernoulli random variables are used to model experiments having exactly two possible outcomes which are generically referred to as failure and success. Formally,

Definition 17.1 *A random variable X is said to have the* ***Bernoulli distribution*** *with parameter $p \in (0,1)$, denoted $X \sim BE(p)$, if its pmf is given by*

$$f(x \mid p)=\begin{cases} p^x(1-p)^{1-x}, & \textit{for } x \in \{0,1\} \\ 0, & \textit{otherwise.}\end{cases}$$

We have written $f(x \mid p)$ instead of $f(x)$ in previous definition. This is to emphasize that we have one and only one parameter, p, that determines the distribution. In general, modeling an experiment with two possible outcomes would only require an approximate value for p. Think about how you would find such approximation, there is no surprise here - more on this later.

Theorem 17.1 *Let $X \sim BE(p)$. Then*

1. $\mathbb{E}[X] = p$.
2. $Var(X) = p(1-p)$.
3. *The distribution of X^k is the same as the distribution of X for any $k \geq 1$.*

Exercise 17.1 *Prove Theorem 17.1.*

Next we move to a generalization of Bernoulli random variables which we already met in Example 11.3: Binomial distribution. Let's remember the example to keep in mind:

Example 17.1 *Consider an experiment where we toss a coin n times independently where for each trial obtaining a head has probability $p \in (0,1)$. Then, letting X be the number of heads (in other words, the number of successes) we have, it is easy to see that the pmf of X is given by $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, 1 \ldots, n$. Observe that this reduces to a Bernoulli random variable when $n = 1$.* □

Here is the formal definition of the binomial distribution.

Definition 17.2 *A random variable X is said to have the* ***binomial distribution*** *with parameters $n \in \mathbb{N}$ and $p \in (0,1)$, denoted $X \sim Bin(n,p)$, if its pmf is given by*

$$f(x \mid n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{1-x}, & \text{for } x = 0, 1, \ldots, n \\ 0, & \text{otherwise.} \end{cases}$$

The following fact is very important.

Fact: We can express binomial random variables as a sum of independent Bernoulli random variables. In particular, when $X \sim Bin(n,p)$, $X = \sum_{i=1}^n X_i$ where X_i's are independent Bernoulli random variables with parameter p.

Theorem 17.2 *Let $X \sim Bin(n,p)$. Then* *i.* $\mathbb{E}[X] = np$.

ii. $Var(X) = np(1-p)$.

Proof: (i) We have seen a proof of the first claim in Section 11 where we used direct computation. Let's give an alternative proof with the use of method of indicators. First, we know that we can write

$$X = \sum_{i=1}^{n} X_i,$$

where X_i's are independent Bernoulli random variables with parameter p. So using linearity of expectation, and the expectation for a Bernoulli random variable (which is p) we obtain

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = \sum_{i=1}^{n} p = np.$$

(ii) Since $X = \sum_{i=1}^{n} X_i$ where X_i's are independent Bernoulli random variables with parameter p, we have

$$Var(X) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} p(1-p) = np(1-p).$$

Here the second equality uses Theorem 15.1. □

18 Hypergeometric distribution

In Section 17, we discussed sums of independent Bernoulli random variables resulting in binomial distribution. This section will be similar except that the Bernoulli variables are dependent now.

Example 18.1 *Consider a box with R red and B blue balls. Select $0 \leq n \leq R+B$ balls randomly without replacement. Define the Bernoulli random variables $X_1, \ldots, X_n$ by*

$$X_i = \begin{cases} 1, & \text{if } i^{th} \text{ selected ball is red} \\ 0, & \text{otherwise.} \end{cases}$$

Then $X = \sum_{i=1}^{n} X_i$ is the number of red balls in our sample (Note that X_i's are dependent random variables) and the pmf of X is given by

$$f(x \mid R, B, n) = \mathbb{P}(X = x) = \frac{\binom{R}{x}\binom{B}{n-x}}{\binom{R+B}{n}} \quad \text{when} \quad 0 \leq x \leq R, 0 \leq n-x \leq B.$$

□

The distribution described in previous example is the hypergeometric distribution. It is important since sampling in real life occurs without replacement. For example, in a poll for elections, you would not like to double count a person as this would naturally cause a bias in your results.

Definition 18.1 *The pmf $f(x \mid R, B, n)$ described in previous example is the pmf of* ***hypergeometric distribution*** *with parameters R, B and n. If a random X has such distribution, we then write $X \sim HG(R, B, n)$.*

Theorem 18.1 *Let X be a random variable having the hypergeometric distribution with parameters R, B, n. Then*

i. $\mathbb{E}[X] = n\frac{R}{R+B}$.

ii. $Var(X) = \frac{nRB}{(R+B)^2}\frac{R+B-n}{R+B-1}$.

Proof: Label the red balls as $1, \ldots, R$ and for $i = 1, \ldots, R$, let $Y_i = 1$ if ith red ball is in the sample and $Y_i = 0$ otherwise. Then $X = \sum_{i=1}^R Y_i$. So

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^R Y_i\right] = \sum_{i=1}^R \mathbb{E}[Y_i] = R\mathbb{P}(Y_i = 1) = R\frac{\binom{R+B-1}{n-1}}{\binom{R+B}{n}} = n\frac{R}{R+B}.$$

We omit the computation of the variance which is similar, but lengthier. □

Exercise 18.1 *If $X \sim HG(R, B, n)$, show that*

$$Var(X) = \frac{nRB}{(R+B)^2}\frac{R+B-n}{R+B-1}.$$

Remark 18.1 *When R and B are very large, and n is comparably small it should be intuitively clear that the hypergeometric distribution will be close to the binomial distribution since it becomes less likely to resample an already sampled, say, ball. As a demontration you may observe that when X is hypergeometric with parameters R, B, n, where $R \sim pN$[21] and $B \sim (1-p)N$, with $p \in (0, 1)$, we would have*

$$\lim_{N\to\infty} \mathbb{E}[X] = \lim_{N\to\infty} n\frac{R}{N} = \lim_{N\to\infty} n\frac{R}{pN}\frac{pN}{N} = np.$$

That is, when N is very large, the probability of success is close to p and the Bernoullis we have behave almost independently. It is possible to make these statements precise, but we will not go into details here - though there will be a relevant problem in Review Problems.

[21] Here $R \sim pN$ means that $\lim_{n\to\infty} \frac{R}{pN} = 1$.

19 Geometric/Negative binomial distributions

We begin with the geometric distribution (negative binomial distribution will be a generalization). Suppose that independent trials, each having a success probability $p \in (0,1)$, are performed. Let X be the number of failures that occur before the first success. Then

$$\mathbb{P}(X = n) = (1-p)^n p, \qquad n = 0, 1, \ldots \tag{3}$$

Definition 19.1 *Any random variable X having pmf $f(n \mid p)$ as in (3) is said to have a **geometric distribution** with parameter p. In this case, we write $X \sim Geo(p)$.*

Remark 19.1 *(i) Note that $\sum_{n=0}^{\infty}(1-p)^n p = p\frac{1}{1-(1-p)} = 1$ by using the summation formula for geometric series[22], and so we have really defined a probability mass function.*

(ii) Some authors prefer to define the geometric distribution in a slightly different way. Namely, they call X a geometric random variable with parameter p if X is the number of trials required to obtain the first success. This is the same as our definition except that we have an additional +1 for the first success. So, the theory for these two definitions are almost identical.

Example 19.1 *An urn contains N white and M black balls. Balls are randomly selected one at a time, with replacement, and independently. Letting X be the number of white balls until we select a black ball;*

(i) What is the distribution of X?

(ii) Find $\mathbb{P}(X = n)$, $n \geq 0$.

(iii) Find $\mathbb{P}(X \geq n)$, $n \geq 1$.

Solution: (i) X is geometrically distributed with parameter $M/(N+M)$.
(ii) $\mathbb{P}(X = n) = \left(\frac{N}{N+M}\right)^n \frac{M}{N+M}$.
(iii) $\mathbb{P}(X \geq n) = \sum_{k=n}^{\infty}\left(\frac{N}{N+M}\right)^k \frac{M}{N+M}$. Simplifying this expression is left for you. You just need to play with geometric series. □

Next result gives the expectation and variance of the geometric distribution.

Theorem 19.1 *Assume that X is a geometric random variable with parameter p. Then*

(i) $\mathbb{E}[X] = \frac{1-p}{p}$.

[22] If $|r| < 1$, then $\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}$

(ii) $Var(X) = \frac{1-p}{p^2}$.

Proof: Let me just prove (i), the other one is similar. We have

$$\begin{aligned}\mathbb{E}[X] = \sum_{k=0}^{\infty} kp(1-p)^k = p(1-p)\sum_{k=1}^{\infty}(1-p)^{k-1} &= p(p-1)\sum_{k=1}^{\infty}\frac{d}{dp}(1-p)^k \\ &= p(p-1)\frac{d}{dp}\sum_{k=1}^{\infty}(1-p)^k \\ &= p(p-1)\frac{d}{dp}\left(\frac{1}{p}-1\right) \\ &= p(p-1)\frac{-1}{p^2} \\ &= \frac{1-p}{p}.\end{aligned}$$

□

Exercise 19.1 *Prove (ii) of Theorem 19.1.*

Theorem 19.2 *If X is a geometric random variable, then*

$$\mathbb{P}(X = n+k \mid X \geq n) = \mathbb{P}(X = k), \quad k, n \geq 0.$$

*(Note : This is called as the **memoryless property** of geometric distribution.)*

Proof: Since X is a geometric random variable, $\mathbb{P}(X = j) = p(1-p)^j$, $j \geq 0$ for some $p \in (0,1)$. We observe

$$\begin{aligned}\mathbb{P}(X = n+k \mid X \geq n) = \frac{\mathbb{P}(X = n+k, X \geq n)}{\mathbb{P}(X \geq n)} = \frac{\mathbb{P}(X = n+k)}{\mathbb{P}(X \geq n)} &= \frac{p(1-p)^{n+k}}{(1-p)^n} \\ &= p(1-p)^k \\ &= \mathbb{P}(X = k).\end{aligned}$$

□

Remark 19.2 *It is worth noting that geometric distribution is the only discrete non-negative distribution with the memoryless property. Its continuous version is the exponential distribution which we will see later.*

Geometric distribution has a natural generalization: Suppose that independent trials each with success probability p are performed. Let X be the number of failures that occur before the r^{th} success. Then

$$\mathbb{P}(X = k) = \binom{r+k-1}{k} p^r (1-p)^k, \qquad k = 0, 1, 2, \ldots \tag{4}$$

Definition 19.2 *A random variable X with pmf $f(k \mid r, p)$ as in (4) is said to have a **negative binomial distribution** with parameters r and p.*

The following theorem should at least be intuitively clear. Attempt a rigorous justification yourself.

Theorem 19.3 *If $X_1, \ldots, X_r$ are independent random variables each having geometric distribution with parameter p, then $X_1 + \cdots + X_r$ has the negative binomial distribution with parameters r and p.*

20 Moments

Definition 20.1 *For each random variable X and every positive integer k, $\mathbb{E}[X^k]$ is called the $\mathbf{k^{th}}$ **moment** of X provided that $\mathbb{E}[X^k]$ exists.*

Here are three basic observations:

- $\mathbb{E}[X]$ is the first moment of X.
- If $\mathbb{P}(a \leq X \leq b)$ for some $-\infty < a < b < \infty$, then all moments of X exist.
- X can be unbounded, but still all its moments may exist (e.g. consider the pmf $f(n \mid) = (1-p)^{n-1}p$, $n \geq 1$.).

Here is one nice result about the absolute moments - by k**th absolute moment** of X, we mean $\mathbb{E}|X|^k$.

Theorem 20.1 *If $\mathbb{E}|X|^k < \infty$ for some $k \in \mathbb{Z}^+$, then $\mathbb{E}|X|^j < \infty$ for every $j \in \mathbb{Z}^+$ with $j \leq k$.*

Proof: Let f be the pmf of X. Then we have

$$\begin{aligned}\mathbb{E}|X|^j = \sum_{x\in s(X)} |x|^j f(x)dx &= \sum_{x\in s(X):|x|\leq 1} |x|^j f(x)dx + \sum_{x\in s(X):|x|>1} |x|^j f(x)dx \\ &\leq \sum_{x\in s(X):|x|\leq 1} f(x)dx + \sum_{x\in s(X):|x|>1} |x|^k f(x)dx \\ &\leq \mathbb{P}(|X|\leq 1) + \mathbb{E}|X|^k < \infty.\end{aligned}$$

□

Exercise 20.1 *Let $k \in \mathbb{N}$. Find a random variable whose k^{th} moment exists but $(k+1)^{st}$ moment does not.*

Remark 20.1 *A related notion is the central moment: For a random variable X with finite expectation μ, the k**th central moment** is defined by $c_k = \mathbb{E}|X-\mu|^k$. In particular, variance is the second central moment.*

21 Special features of distributions

As we already know, one important distinction between distributions is based on whether the data is coming from a continuous or a discrete set. Below we discuss various features of distributions for the discrete case, but they can be extended to the continuous setting in a straightforward way.

1. A **mode** of a discrete probability distribution is a value at which the pmf takes its maximum value. Note that there can be multiple modes since the maximum value of a pmf can be attained at more than one value.

   If there is a single mode, the distribution function is called **unimodal**. If it has more modes it is "bimodal" (2), "trimodal" (3), etc., or more generally, **multimodal**. The binomial distribution can be seen to be unimodal when n is even and bimodal when n is odd.

2. An integer valued discrete random variable X is said to be **log-concave** if its pmf f satisfies
   $$f^2(i) \geq f(i-1)f(i+1), \qquad \text{for all } i \in \mathbb{Z}.$$
   Log-concavity is a fundamental property of distributions yielding various conclusions. One such exemplary result is that any log-concave this distribution is unimodal. It can be checked that binomial distribution is an example of log-concave distribution.

3. A distribution with the pmf $f(x)$ is said to be **symmetric** around $x = a$, if the function $f(x)$ is symmetric around $x = a$. When $a = 0$, we just say that the distribution, or the pmf is symmetric. Symmetric distributions satisfy several nice properties. For example, the odd moments of a symmetric pmf around $x = 0$ will be zero.

4. When the distribution is not symmetric, loosely speaking, the distribution is said to be **right-skewed** (left-skewed, resp.) if the right tail (left tail, resp.) seems to be stretched. Letting X be a sample from the underlying distribution, the rigorous definition is

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{SD(X)}\right)^3\right].$$

Larger γ_1 indicates a positive skewness[23] .

5. **Kurtosis** is a measure of the "tailedness" of the probability distribution of a real-valued distribution. It is defined as

$$Kurt(X) = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{SD(X)}\right)^4\right].$$

22 Review problems II

Exercise 22.1 *In a state lottery a non-negative three-digit integer is selected at random (this includes 000). If a player bets \$1 on a particular number and if that number is selected, the payoff is \$500 minus the \$1 paid for the ticket. Let X equal the payoff to the bettor, namely -\$1 or \$499, and find $\mathbb{E}[X]$.*

Exercise 22.2 *(Shifted geometric distribution) Let $p \in (0, 1)$. Let X be a random variable whose pmf is given by*

$$f(x) = \begin{cases} p(1-p)^{x-1}, & \textit{if } x = 1, 2, \dots \\ 0, & \textit{otherwise.} \end{cases}$$

Find (a) $\mathbb{E}[X]$ and (b) $Var(X)$. (Answer: (a) $1/p$ (b) $(1-p)/p^2$)

Exercise 22.3 *Let $\lambda > 0$. Let X be a random variable whose pmf is given by*

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

Find $\mathbb{E}[Y]$[24]. (Answer: λ)

[23] Let's note another measure of skewness. Pearson's mode skewness is defined by $\frac{mean-mode}{SD}$. Sketching a few distributions will provide you intuition towards this alternative definition.

[24] The distribution of Y will be called the Poisson distribution later on.

Exercise 22.4 *Suppose that $X_1, \ldots, X_n$ are n independent Bernoulli random variables with parameter p. Determine the conditional probability that $X_1 = 1$, given that $\sum_{i=1}^n X_i = k$, where $k \in \{1, \ldots, n\}$.*

Exercise 22.5 *Let $n > 1$ be a positive integer, and consider a die with n faces. Suppose that I roll this die $m \geq 1$ times, and let $X_1, \ldots, X_m$ be the number of spots in first,...,mth rolls, respectively.*

a. *Find $\mathbb{P}(X_1 + \cdots + X_m = nm)$.*

b. *Find $\mathbb{P}(X_1 + \cdots + X_m \neq m + 1)$.*

c. *Say that we have a* coincidence *at ith, $1 \leq i \leq m$, roll if $X_i = i$. Find an expression for the probability that we don't have any coincidences among these m rolls.*

Exercise 22.6 *Suppose X and Y are independent binomial random variables each having parameters n, p, and let $Z = X + Y$. Show that the conditional distribution of X given $Z = N$ is the hypergeometric distribution.*

Exercise 22.7 *Let U be the number of trials needed to get the first head and V be the number of trials needed to get two heads in repeated tosses of a fair coin. Are U and V independent random variables?*

Exercise 22.8 *Let $X_1, X_2, \ldots,$ be an infinite sequence of independent Bernoulli random variables where the probability of success is $1/3$ and the probability of failure is $2/3$. Let X be the number of failures required to obtain the first success.*

(i) *What is the distribution of X?*

(ii) *Find $\mathbb{P}(4 \leq X < 7)$.*

(iii) *Find the probability that X will be divisible by 3 and at the same time it will not be divisible by 7.*

(iv) *Let Y_1 and Y_2 be independent random variables with the same distribution as X. Find the conditional distribution of Y_1 given that $Y_1 + Y_2 = N$ where N is a given fixed positive integer.*

Exercise 22.9 *For each value of $p > 1$, let*

$$c(p) = \sum_{x=1}^{\infty} \frac{1}{x^p}.$$

Suppose that the random variable X has a discrete distribution with the following probability function:

$$f(x) = \frac{1}{c(p)x^p} \qquad \text{for} \quad x = 1, 2, \ldots.$$

(i) For each fixed positive integer n, determine the probability that X will be divisible by n.

(ii) Find the probability that X is divisible by a given prime number $q \geq 2$.

(iii) Find the probability that X is an odd integer.

Note: This exercise has various interesting connections to prime number theory which we do not go into details here[25].

Exercise 22.10 *Consider a group of n people. Assuming that their birthdays are independent of each other and that each 365 days are equally likely, find the expected number of pairs who have the same birthday.*

Exercise 22.11 *Let $X_1, X_2, \dots$ be a sequence of i.i.d. continuous random variables. We say that a record occurs at time n if $X_n > \max(X_1, \dots, X_{n-1})$. Let Y_n be the number of records by time n. Find $\mathbb{E}[Y_n]$.*

Exercise 22.12 *Five percent of computer parts produced by a certain supplier are defective.*
(a) What is the probability that a randomly and independently selected sample of 12 parts contains at least 3 defective ones?
(b) What is the expected number of defective parts among this sample of 12 parts?

Exercise 22.13 *Consider a sequence of independent coin flips, each of which has probability p of being heads. Define a random variable X as the length of the run (of either heads or tails) started by the first trial. (For example, $X = 3$ if either $TTTH$ or $HHHT$ is observed.) Find the distribution of X, and find $\mathbb{E}[X]$.*

Exercise 22.14 *(b) Let X be a discrete random variable whose range is the non-negative integers. Show that*

$$\mathbb{E}[X] = \sum_{n=0}^{\infty}(1 - F_X(n)),$$

where $F_X(n) = \mathbb{P}(X \leq n)$.

Exercise 22.15 *A couple decides to continue to have children until a daughter is born. What is the expected number of children of this couple?*

[25] Talk to me in case you are interested.

Exercise 22.16 *The flow of traffic at certain street corners can sometimes be modeled as a sequence of Bernoulli trials*[26] *by assuming that the probability of a car passing during any given second is a constant p and that there is no interaction between the passing of cars at different seconds. If we treat seconds as indivisible time units (trials), the Bernoulli model applies. Suppose a pedestrian can cross the street only if no car is to pass during the next 3 seconds. Find the probability that the pedestrian has to wait for exactly 4 seconds before starting to cross.*

Exercise 22.17 *A man with n keys wants to open his door and tries the keys at random. Exactly one key will open the door. Find the mean number of trials if*

(a) unsuccessful keys are not eliminated from further selections.

(b) unsuccessful keys are eliminated.

Exercise 22.18 *Let U_i, $i = 1, 2, \ldots$ be independent $U(0,1)$ random variables, and let X have distribution*

$$\mathbb{P}(X = x) = \frac{c}{x!}, \quad x = 1, 2, 3, \ldots$$

where $c = 1/(e-1)$. Find the distribution of $Z = \min\{U_1, \ldots, U_X\}$. (Hint: Note that the distribution of $Z \mid X = x$ is that of the first order statistic from a sample of size x.)

Exercise 22.19 *Prove that a log-concave pmf is unimodal.*

Exercise 22.20 *Prove that the binomial distribution is log-concave, and is therefore unimodal.*

23 First look at predictions: Minimizing MSE

Now suppose that X is a random variable whose value can be observed in some experiment, but that this value must be predicted before the observation can be made - e.g. the temperature tomorrow in Istanbul. There are various ways to do this. For example, if the pmf of X is as in following figure, one may decide to choose the mean, the median or the mode as a prediction.

[26] By a Bernoulli trial you should just understand that we have a random variable having the Bernoulli distribution.

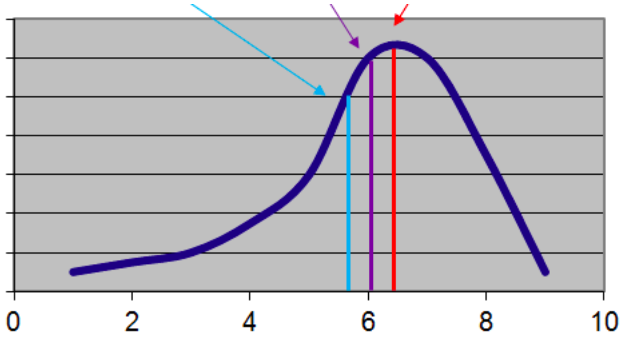


Figure 2: Red: Mode, Purple: Median, Blue: Mean

The basis for making the prediction in this section will be to select some $d \in \mathbb{R}$ for which the expected value of $(X-d)^2$ is minimized.

Definition 23.1 *The **mean square error** of a prediction d for a random variable X is defined by*

$$MSE(d;X) = \mathbb{E}[(X-d)^2].$$

When the underlying random variable is clear from the context we merely write $MSE(d)$ instead of $MSE(d;X)$.

Now for each w in the sample space S, one obtains a mean square error $MSE(d;X(w)) = (X(w)-d)^2$, and once we average this out over all sample space, we obtain the mean square error. Here is the main result of this section.

Theorem 23.1 *Let X be a random variable with finite mean. Then the prediction d that minimizes the mean squared error is given by*

$$d^* = \mathbb{E}[X].$$

That is, for any $d \in \mathbb{R}$

$$MSE(\mathbb{E}[X];X) \leq MSE(d;X).$$

Proof: Let $g(d) = MSE(d;X)$. Observe that

$$g(d) = \mathbb{E}[(X-d)^2] = \mathbb{E}[X^2 - 2dX + d^2] = d^2 - 2d\mathbb{E}[X] + \mathbb{E}[X^2],$$

where in the last step we used linearity of expectation. Therefore the problem reduces to minimizing a parabola. Noting $g'(d) = 2d - 2\mathbb{E}[X] = 0$ gives the critical point as $d = \mathbb{E}[X]$. Result now follows from the second derivative test since $g''(d) = 2 > 0$. □

24 Cumulative distribution functions

Definition 24.1 *The* ***cumulative distribution function*** *(cdf) F of a random variable X is defined as the function*

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

Example 24.1 1. *(Bernoulli distribution) Recall that X is a Bernoulli random variable if its pmf is given by $f(x) = p, x = 1$, $f(x) = 1 - p, x = 0$ and $f(x) = 0$ otherwise. It is then easy to see that the cdf of X is given by*

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - p, & \text{if } x \in [0, 1) \\ 1, & \text{if } x \geq 1. \end{cases}$$

2. *(Uniform distribution) Let X be uniformly distributed over $\{1, 2, 3, 4, 5, 6\}$. Then the pmf of X is given by*

$$f(x) = \frac{1}{6} \qquad x \in \{1, 2, 3, 4, 5, 6\}.$$

Using the pmf we can obtain the cdf easily as

$$F(x) = \frac{x(x+1)}{12}, \qquad x \in \{1, 2, 3, 4, 5, 6\}.$$

Sketch the pmf yourself.

3. *It is not always possible find a nice expression for the cdf. For example, if we consider a binomial random variable X with parameters n and p, the cdf will be a piecewise function with $n + 1$ pieces, and the sums corresponding to the probabilities for these pieces will not have closed forms.*

Theorem 24.1 *Let F be the cdf of some random variable X.*
(i) $F(x)$ is nondecreasing in x, that is, $F(x_1) \leq F(x_2)$ if $x_1 < x_2$.
(ii) Any cdf is a right-continuous function.
(iii) A cdf can have at most countably many discontinuties.
(iv) If X is integer valued and f is the pmf of X, then $f(n) = F(n) - F(n - 1)$, for any $n \in \mathbb{Z}$.

Proof: (i) First recall that if A and B are two events such that $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$. Using this with the events $\{X \leq x_1\} \subset \{X \leq x_2\}$, we have

$$F(x_1) = \mathbb{P}(X \leq x_1) \leq \mathbb{P}(X \leq x_2) = F(x_2),$$

proving the first part.

(ii) Let $x_0 \in \mathbb{R}$. We should show that $\lim_{h \searrow 0} F(x_0 + h) = F(x_0)$[27]. But since the sets $\{X \leq x_0 + h\}$ are decreasing as $h \searrow 0$, the result follows from Exercise 2.3.

(iii) Since F is a non-decreasing function it may only have jump discontinuities. At each jump we can pick a rational number and so the number of discontinuities can be at most as many as the rational numbers which is countable.

(iv) Follows from the definitions. □

25 Multivariate discrete distributions

In this section, we generalize the concept of the distribution of a random variable to the joint distribution of two and more random variables. We will do this focusing on just two random variables, more general case is similar.

Definition 25.1 *Let X and Y be random variables. The **joint distribution** (or the **bivariate distribution**) of X and Y is the collection of all probabilities of the form $\mathbb{P}((X,Y) \in E)$ for all sets E of pairs of real numbers such that $\{(X,Y) \in E\}$ is an event.*

The joint distribution of X and Y can be discrete, continuous or a mixture of these two. We focus on the discrete case for now.

Definition 25.2 *If there are only finitely or at most countably many different possible values (x,y) for the pair of random variables (X,Y), then we say that X and Y have a **discrete joint distribution**. The **joint probability mass function**, abbreviated the **joint pmf**, of X and Y is defined as the function f such that for every point (x,y) in the xy-plane,*

$$f(x,y) = \mathbb{P}(X = x, Y = y)^{28}.$$

Here are three immediate observations about discrete joint distributions:

- $f(x,y) \geq 0$ for any x, y. If X and Y have a discrete joint distribution and if (x,y) is not one of the possible values of the pair (X,Y), then $f(x,y) = 0$. Also, $\sum f(x,y) = 1$ when the summation is over all possible values of (X,Y).

- For each set E of countably many ordered pairs,

$$\mathbb{P}((X,Y) \in E) = \sum_{(x,y) \in E} f(x,y).$$

[27] Here the notation $\lim_{h \searrow 0}$ is used to emphasize that h approaches 0 from the right.

[28] Recall that we often write $\mathbb{P}(X = x, Y = y)$ in place of $\mathbb{P}(X = x \cap Y = y)$.

- If the individual distributions of X and Y are discrete, then their joint distribution is discrete as well (Why?).

Sometimes it is useful to consider the joint pmf as a matrix of probabilities - just like an Excel file.

Example 25.1 *In a certain suburban area, each household reported the number of cars and the number of television sets that they owned. Let X stand for the number of cars owned by a randomly selected household in this area and let Y stand for the number of television sets owned by that same randomly selected family. We assume that X takes only the values $1, 2,$ and 3; Y takes only the values $1, 2, 3,$ and 4; and the joint pmf f of X and Y is given as in following table:*

X \Y	1	2	3	4
1	0.1	0	0.1	0
2	0.3	0	0.1	0.2
3	0	0.2	0	0

Then, for example, if we are interested in the probability that a randomly chosen family has at least two cars and at least two TV sets, we have

$$\begin{aligned}\mathbb{P}(X \geq 2, Y \geq 2) &= f(2,2) + f(2,3) + f(2,4) + f(3,2) + f(3,3) + f(3,4) \\ &= 0 + 0.1 + 0.2 + 0.2 + 0 + 0 = 0.5.\end{aligned}$$

□

Exercise 25.1 *Consider the setting in previous example. Find the following probabilities.*
(i) $\mathbb{P}(X = 2, Y = 4)$;
(ii) $\mathbb{P}(X > 1, Y > 3)$;
(iii) $\mathbb{P}(X < 1, Y \geq 1)$;
(iv) $\mathbb{P}(X > 3, 2 \leq Y \leq 3)$.

Definition 25.3 *The **joint cumulative distribution function** (abbreviated **joint cdf**) of two random variables X and Y is defined by*

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y), \quad x, y \in \mathbb{R}.$$

Definition 25.4 *Let X and Y have a joint cdf F. Then the **marginal cdf** of X and the **marginal cdf** of Y are defined by $F_1(x) = \lim_{y\to\infty} F(x, y)$ and $F_2(y) = \lim_{x\to\infty} F(x, y)$, respectively. The pmfs associated with F_1 and F_2 are called the **marginal pmfs** of X and Y, and are denoted by f_1 and f_2, respectively.*[29]

[29] The definition looks complicated, but is not. Make it intuitively clear by thinking over it.

Here is an example on finding the marginal distribution[30] by using table of probabilities discussed in previous section.

Example 25.2 *(Table of probabilities) Recall the following probability table for car/TV set data for households in a certain area:*

X \Y	1	2	3	4
1	0.1	0	0.1	0
2	0.3	0	0.1	0.2
3	0	0.2	0	0

By the marginal pmf of X, we just mean the pmf of X. For this problem, support of X is $\{1,2,3\}$. For example,

$$\mathbb{P}(X=1) = f(1,1) + (1,2) + f(1,3) + f(1,4) = 0.1 + 0 + 0.1 + 0 = 0.2.$$

Finding the other probabilities similarly, the marginal pmf of X is given by

$$f_1(x) = \begin{cases} 0.2, & \textit{if } x = 1 \\ 0.6, & \textit{if } x = 2 \\ 0.2, & \textit{if } x = 2 \\ 0, & \textit{otherwise.} \end{cases}$$

Similarly, the marginal pmf of Y is given by

$$f_2(y) = \begin{cases} 0.4, & \textit{if } y = 1 \\ 0.2, & \textit{if } y = 2,3,4 \\ 0, & \textit{otherwise.} \end{cases}$$

□

More generally:

[30] By the marginal distribution, we just mean the marginal pmf.

Theorem 25.1 *If X and Y have a discrete joint distribution with joint pmf f, then the marginal pmf f_1 of X is*

$$f_1(x) = \sum_{y \in s(Y)} f(x,y).$$

Similarly, the marginal pmf f_2 of Y is given by

$$f_2(y) = \sum_{x \in S(X)} f(x,y).$$

Next, given a random pair (X,Y) with joint pmf $f(x,y)$ and a real-valued function $g : \mathbb{R}^2 \to \mathbb{R}$, how do we find $\mathbb{E}[g(X,Y)]$? This is no harder than the univariate case.

Let (X,Y) be a pair of random variables with joint pmf $f(x,y)$ and a real-valued function $g : \mathbb{R}^2 \to \mathbb{R}$. Then the expected value of $g(X,Y)$ is given by

$$\mathbb{E}[g(X,Y)] = \sum g(x,y) f(x,y)$$

where the summation is taken over all possible values of (X,Y) - you may call it the support of the random vector (X,Y). Here the summation is assumed to be absolutely convergent, otherwise we would say that the expectation does not exist.

Example 25.3 *Let (X,Y) be a pair of random variables with joint pmf*

$$f(x,y) = \begin{cases} \frac{1}{6}, & \textit{if } (x,y) = (1,2) \\ \frac{1}{3}, & \textit{if } (x,y) = (2,1) \\ \frac{1}{4}, & \textit{if } (x,y) = (1,0) \textit{ or } (x,y) = (0,1) \\ 0, & \textit{otherwise.} \end{cases}$$

Letting $g(x,y) = xy$, let's find the expected value of $g(X,Y)$. We have

$$\begin{aligned} \mathbb{E}[g(X,Y)] &= \frac{1}{6}(1 \times 2) + \frac{1}{3}(2 \times 1) + \frac{1}{4}(1 \times 0) + \frac{1}{4}(0 \times 1) \\ &= \frac{3}{2}. \end{aligned}$$

As already noted, generalization of results for bivariate random variables to more than two random variables is standard. In this case we would have, $n \geq 2$ random variables $X_1, X_2, \ldots, X_n$ and then their joint pmf would be given by

$$f(x_1, x_2, \ldots, x_n) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Other discussions above and below for the bivariate distributions can be generalized to this multivariate setting in more or less a straightforward way.

26 A few notes on independence

Recall that the random variables $X_1, \ldots, X_n$ are said to be independent if for any $A_1, \ldots, A_n \subset \mathbb{R}$ and for any $I \subset \{1, 2, \ldots, n\}$, we have

$$\mathbb{P}\left(\bigcap_{i \in I} \{X_i \in A_i\}\right) = \prod_{i \in I} \mathbb{P}(X_i \in A_i).$$

Again focusing on the $n = 2$ case here this reduces to checking whether $\mathbb{P}(X_1 \in A_1, X_2 \in A_2) = \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2)$ for any A_1, A_2 which is quite a difficult task, in general.

In our new setting there is a useful criteria to check independence, known as the factorization theorem.

Theorem 26.1 *(Factorization theorem) Suppose X and Y are random variables that have a joint pmf f.*
(i) Then X and Y are independent if and only if we can write

$$f(x, y) = h_1(x)h_2(y), \qquad -\infty < x, y < \infty$$

where h_1 is a function of x alone, and h_2 is a function of y alone.
(ii) Letting f_1, f_2 be the marginal pmfs of X and Y, respectively, an alternative criteria for independence is

$$f(x, y) = f_1(x)f_2(y), \quad \text{for all } x, y \in \mathbb{R}.$$

Proof of the factorization theorem will be given later in the continuous setting. Let's instead see two examples here.

Example 26.1 *Let the joint pmf of X and Y be given by*

$$f(x, y) = \frac{1}{4} \quad \text{for } (x, y) = (0, 0), (0, 2), (2, 0), (2, 2).$$

Then $f(x,y) = h_1(x)h_2(y)$ for any x,y where $h_1(x) = 1/2$ is a function of x only, and $h_2(y) = 1/2$ is a function of y only. Therefore X and Y are independent. □

Example 26.2 *Let $X \sim U(\{-1,0,1\})$, and $Y = X^2$. Call the joint pmf f. Then $f(-1,1) = 1/3$, $f_1(-1) = 1/3$ and $f_2(1) = 2/3$. So*

$$\frac{1}{3} = f(-1,1) \neq f_1(-1)f_2(1) = \frac{2}{9},$$

from which we conclude that X, Y are independent random variables. □

We conclude this section with a discussion on expectations of products of independent random variables. Let X, Y be independent random variables with joint pmf f, and corresponding marginal pmfs f_1 and f_2. Then we have

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{x,y} xyf(x,y) \\ &= \sum_{x,y} xyf_1(x)f_2(y) \\ &= \sum_{x} xf_1(x) \sum_{y} yf_1(y) \\ &= \mathbb{E}[X]\mathbb{E}[Y]. \end{aligned}$$

Following steps similar to this simple argument, you may prove the following.

Theorem 26.2 *(i) If $X_1, \ldots, X_n$ are independent random variables such that $\mathbb{E}[X_i]$ is finite for each i, then*

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

(ii) If $X_1, \ldots, X_n$ are independent random variables, $g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}$ are continuous functions, and if $\mathbb{E}[g_i(X_i)]$ is finite for each i,

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

The continuity assumption in part ii is not necessary, the result will hold as long as g_i's are "nice". We don't need these technicalities at the moment.

Example 26.3 *Suppose X_1, X_2, X_3 are independent random variables such that $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^2] = 1$ for $i = 1, 2, 3$. Find $\mathbb{E}[X_1^2(X_2 - 4X_3)^2]$.*

Solution: Since X_1 is independent of X_2 and X_3, X_1^2 is independent of $(X_2 - 4X_3)^2$. So

$$\mathbb{E}[X_1^2(X_2 - 4X_3)^2] = \mathbb{E}[X_1^2]\mathbb{E}[(X_2 - 4X_3)^2].$$

Manipulations now give

$$\mathbb{E}[X_1^2(X_2-4X_3)^2] = \mathbb{E}[X_1^2]\mathbb{E}[X_2^2-8X_2X_3+16X_3^2] = \mathbb{E}[X_1^2](\mathbb{E}[X_2^2]-8\mathbb{E}[X_2]\mathbb{E}[X_3]+16\mathbb{E}[X_3^2]) = 17.$$

□

27 Cauchy-Schwarz inequality

As we have already noted, it is often not possible to compute probabilities, and thereof expectations, exactly. Inequalities are of extreme use in such situations, besides their help on developing the underlying theory. In this section we will discuss one such inequality for random variables under convex functions. Recall that $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function if for any $x, y \in \mathbb{R}$ and for any $c \in (0, 1)$, the inequality

$$\phi(cx + (1 - c)y) \leq c\phi(x) + (1 - c)\phi(y).$$

At some elementary calculus course you learnt convex functions as ”concave up functions”.

The following fundamental inequality will be our first moment inequality.

Theorem 27.1 *(Cauchy-Schwarz inequality) Assume that X and Y are random variables for which $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$. Then*

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

Proof: For given $t \in \mathbb{R}$, set

$$f(t) = \mathbb{E}[(X - tY)^2] = \mathbb{E}[X^2] - 2t\mathbb{E}[XY] + t^2\mathbb{E}[Y^2].$$

Noting that $f(t) \geq 0$ for any t, the discriminant of the polynomial f should be non-positive which is to say that

$$4t^2(\mathbb{E}[XY])^2 - 4t^2\mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0.$$

This implies

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2],$$

completing the proof. □

Exercise 27.1 *When do we have equality in Cauchy-Schwarz inequality?*

Replacing X by $|X|$ and Y by $|Y|$ in previous proof, we could obtain the following stronger version of Cauchy-Schwarz inequality:

Theorem 27.2 *Assume that X and Y are random variables for which $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$. Then*

$$(\mathbb{E}|XY|)^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

28 * Jensen's inequality

Theorem 28.1 *(Jensen's inequality) Let X a real valued random variable whose expectation exists. Let ϕ be a convex function for which $\mathbb{E}[\phi(X)]$ exists. Then*

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

The equality holds if and only if $Y = cX$ for some constant c.

Proof: Let X be a real valued random variable for which $\mathbb{E}[X]$ exists, and ϕ be a convex function as in statement of the theorem. Then, for any x_0, the graph of ϕ lies entirely above its tangent line at the point x_0. Denoting the slope at x_0 by m, we then have

$$\phi(x) \geq \phi(x_0) + m(x - x_0), \qquad \text{for any } x \in \mathbb{R}.$$

By putting $x = X$ and $x_0 = \mathbb{E}[X]$, we get that:

$$\phi(X) \geq \phi(\mathbb{E}[X]) + m(X - \mathbb{E}[X]).$$

We can take the expected value of both sides and have:

$$\mathbb{E}[\phi(X)] \geq \mathbb{E}[\phi(\mathbb{E}[X]) + m(X - \mathbb{E}[X])].$$

By the linearity of expectation, we conclude that,

$$\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X]) + m(\mathbb{E}[X] - \mathbb{E}[X]) = \phi(\mathbb{E}[X]).$$

□

Exercise 28.1 *Check that the claim about having equality is true.*

Example 28.1 *Let X be a random variable taking values in $(0, \infty)$ with $\mathbb{E}|X| < \infty$. Here are some immediate consequences of Jensen's inequality.*

(i) $|\mathbb{E}[X]| \leq \mathbb{E}|X|$.

(ii) $(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$.

(iii) $\mathbb{E}[e^X] \leq e^{\mathbb{E}[X]}$.

(iv) $\mathbb{E}[\ln X] \geq \ln \mathbb{E}[X]$. *(Why did the inequality change direction?)*

Next, we discuss a very useful corollary to Jensen's inequality.

Theorem 28.2 *(Lyapounov's inequality) Assume that $\mathbb{E}|X|^t < \infty$ for some $t > 0$. Then for any $0 < s < t$, we have*

$$(\mathbb{E}|X|^s)^{1/s} \leq (\mathbb{E}|X|^t)^{1/t}.$$

Proof: Let $r = t/s > 1$. Then setting $Y = |X|^s$ and applying Jensen's inequality to $g(x) = |x|^r$, we obtain

$$|\mathbb{E}Y|^r \leq \mathbb{E}|Y|^r,$$

yielding

$$(\mathbb{E}|X|^s)^{t/s} \leq \mathbb{E}|X|^t.$$

Taking $1/t^{th}$ powers on both side concludes the proof. □

Corollary 28.1 *We have*

$$\mathbb{E}|X| \leq (\mathbb{E}|X|^2)^{1/2} \leq \cdots \leq (\mathbb{E}|X|^n)^{1/n}.$$

29 Covariance and correlation

The purpose of this section to measure the dependence of two random variables, in particular, by focusing on the linear dependence among them.

Definition 29.1 *Let X and Y have finite second moments. Let $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. The **covariance** of X and Y, denoted $Cov(X,Y)$, is defined to be*

$$Cov(X,Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

Note that when $X = Y$, we obtain $Cov(X,Y) = Var(X)$. So, covariance, in a certain sense, generalizes variance. Indeed, it will also help us to obtain very useful formulas for computing variances.

Next, using the linearity of expectation, we observe that

$$Cov(X,Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] = \mathbb{E}[XY] - \mu_X\mu_Y.$$

Let's record this latter comment as a separate formula.

Theorem 29.1 *For two random variables X and Y with finite second moments*

$$Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \tag{5}$$

We have the following relation between independence and the covariance being 0.

Theorem 29.2 *i.If X and Y are independent random variables, then $Cov(X,Y) = 0$. ii. Converse is not true in general.*

Proof: (i) Clear from (5).

(ii) Let's provide an example. Consider a random variable X with distribution

$$\mathbb{P}(X = -1) = \mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \frac{1}{3}.$$

Set $Y = X^2$. It can be easily shown that X and Y are dependent, but $Cov(X,Y) = 0$. □

Intuitively, you should understand that independence is a much stronger condition, and that it demands quadratic independence, cubic independence, exponential dependence, etc. along with the linear independence. A concept related to the covariance is the correlation of random variables.

Definition 29.2 *Let X and Y be random variables with finite variances σ_X^2 and σ_Y^2, respectively. Then the* ***correlation*** *of X,Y is defined to be*

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

X and Y are said to be ***positively correlated***, ***negatively correlated*** *and* ***uncorrelated*** *when $\rho(X,Y) > 0$, $\rho(X,Y) < 0$ and $\rho(X,Y) = 0$, respectively.*

Remark 29.1 *(Correlation is scale invariant) Consider two random variables X and Y for which $Cov(X,Y) = 1$. Then you can easily check that for $|c| \neq 0, 1$, $Cov(cX, cY) = c^2 Cov(X,Y) \neq Cov(X,Y)$ but $\rho(cX, cY) = \rho(X,Y)$ - check details. This is one main reason why we introduce the concept correlation besides the covariance. Multiplying two random quantities by the same constant should not change the amount of linear dependence among them, and the correlation does this for us, and therefore it provides a universal measure for linear relation.*

Theorem 29.3 *Let X, Y be random variables with finite variances. Then*

$$(Cov(X,Y))^2 \leq \sigma_X^2 \sigma_Y^2.$$

In particular,

$$-1 \leq \rho(X,Y) \leq 1.$$

Proof: Letting $U = X - \mu_X$ and $V = Y - \mu_Y$, Cauchy-Schwarz inequality says

$$(\mathbb{E}[UV])^2 \leq \mathbb{E}[U^2]\mathbb{E}[V^2]$$

from which the result follows immediately. □

Theorem 29.4 *Let X be a random variable with finite variance σ_X^2, and $Y = aX + b$, $a, b \in \mathbb{R}$, $a \neq 0$.*

(i) If $a > 0$, then $\rho(X,Y) = 1$

(ii) If $a > 0$, then $\rho(X,Y) = -1$.

Proof: For (i), we have

$$\begin{aligned}\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{Cov(X, aX+b)}{\sigma_X(a\sigma_X)} &= \frac{\mathbb{E}[(X-\mu_X)(aX+b-a\mu_X-b)]}{a\sigma_X^2} \\ &= \frac{a\mathbb{E}[(X-\mu_X)^2]}{a\sigma_X^2} = 1.\end{aligned}$$

(ii) is similar. □

Remark 29.2 *Let's note one more time that the covariance, and therefore the correlation, measures **only** linear relationship, and the last result is a good indicator for this.*

Here is one fundamental formula you should keep in mind - which says that the covariance is bilinear.

Theorem 29.5 *Let $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ be random variables so that $Cov(X_i, Y_j)$ is finite for any i, j, and suppose $a_1, \ldots, a_m, b_1, \ldots, b_n$ are constants. Then*

$$Cov\left(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j\right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j Cov(X_i, Y_j).$$

Proof: By using the definition of covariance and the linearity of expectation multiple times, we have

$$\begin{aligned}
Cov\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) &= \mathbb{E}\left[\sum_{i=1}^{m} a_i X_i \sum_{j=1}^{n} b_j Y_j\right] - \mathbb{E}\left[\sum_{i=1}^{m} a_i X_i\right] \mathbb{E}\left[\sum_{j=1}^{n} b_j Y_j\right] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \mathbb{E}\left[X_i Y_j\right] - \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \mathbb{E}\left[X_i\right]\mathbb{E}\left[Y_j\right] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j (\mathbb{E}[X_i Y_j] - \mathbb{E}[X_i]\mathbb{E}[Y_j]) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j Cov(X_i, Y_j),
\end{aligned}$$

which yields the required result. □

30 Variance formulas

Theorem 29.5 is very useful in obtaining formulas for variance of a sum of random variables.

Corollary 30.1 *If $X_1, \ldots, X_n$ are random variables with finite variances, then*

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j).$$

Proof: We have

$$\begin{aligned}
Var\left(\sum_{i=1}^{n} X_i\right) = Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} Cov(X_i, X_i) + \sum_{i\neq j} Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j).
\end{aligned}$$

Here the the first step follows from Theorem 29.5 and the last step uses the symmetry of covariance function. □

Let's collect some further immediate corollaries in following statement.

Corollary 30.2 *(i.) If X and Y are random variables with finite variances, and if a, b, c are constants, then*

$$Var(aX + bY + c) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y).$$

(ii.) If $X_1, \ldots, X_n$ are uncorrelated random variables with finite variances, then

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var\left(X_i\right).$$

Some examples are in order.

Example 30.1 *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $Var(X_1) = \sigma^2 < \infty$. Denoting the sample mean $\frac{\sum_{i=1}^n X_i}{n}$ by $\overline{X}_n$. Find $Var(\overline{X}_n)$.*

Solution: *We have*

$$Var(\overline{X}_n) = Var\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) \underbrace{=}_{independence} \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

□

Example 30.2 *Find the variance of a binomial random variable X with parameters $n \in \mathbb{N}$ and $p \in (0, 1)$.*

Solution: *Recall from Section 17 that we may write $X = \sum_{i=1}^n X_i$ where X_i's are independent Bernoulli random variables each having parameter p. Then, recalling that $Var(X_1) = p(1-p)$, we obtain*

$$Var(X) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var\left(X_i\right) = \sum_{i=1}^{n} p(1-p) = np(1-p).$$

□

Example 30.3 *Suppose that 20 men at a party throws his hat into the center of the room. The hats are first mixed up, and then each man randomly selects a hat. Let X_i be the event that i^{th} man finds his own hat for $i = 1, 2, \ldots, 20$, and X be the number of men who find their own hats.*

(i) Are the random variables $X_1, X_2, \ldots, X_{20}$ independent? Provide a mathematical explanation for your answer.

(ii) Find $\mathbb{E}[X]$ and $Var(X)$.

(iii) Let Y be the number of men who do not find their own hats. Find $\mathbb{E}[Y]$ and $Var(Y)$. (Hint: Express Y as a function of X and use part (ii).)

Solution : *(i) Note that $\mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1) = \frac{1}{20^2}$ and $\mathbb{P}(X_1 = 1, X_2 = 1) = \frac{1}{20(19)}$. Since these two are not equal, X_1 and X_2 are not independent, and therefore $X_1, \ldots, X_n$ are not independent.*[31]

(ii) Observe that $\mathbb{E}[X_i] = 1/20$, $\mathbb{E}[X_i^2] = 1/20$ and $Var(X_i) = 1/20 - (1/20)^2 = (1/20)(19/20)$. For $i \neq j$,

$$Cov(X_i, X_j) = \mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] = \frac{1}{20(19)} - \frac{1}{20^2}.$$

Hence

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{20} X_i\right] = \sum_{i}^{20} \mathbb{E}[X_i] = 20\mathbb{E}[X_1] = 20\frac{1}{20} = 1.$$

Next for the variance of X, observe that

$$Var(X) = \sum_{i=1}^{20} Var(X_i) + \sum_{i\neq j} Cov(X_i, X_j) = 20\frac{1}{20}\frac{19}{20} + 20(19)\left(\frac{1}{20(19)} - \frac{1}{20^2}\right) = 1.$$

(iii) Note that $Y = 20 - X$. So

$$\mathbb{E}[Y] = 20 - \mathbb{E}[X] = 19$$

and

$$Var(Y) = Var(20 - X) = Var(X) = 1.$$

□

Exercise 30.1 *Go over the above argument to see that if we had $n \geq 2$ people instead of 20, then we would still have $\mathbb{E}[X] = Var(X) = 1$.*

31 Markov and Chebyshev inequalities

Letting $X_1, X_2, \ldots$ be i.i.d. random variables with $Var(X_1) = \sigma^2 < \infty$, and denoting the sample mean $\frac{\sum_{i=1}^n X_i}{n}$ by $\overline{X}_n$, we already know from Example 30.1 that $Var(\overline{X}_n) = \frac{\sigma^2}{n}$. So $Var(\overline{X}_n) \longrightarrow 0$ as $n \to \infty$, or in other words, the variable $\overline{X}_n$ is almost like a constant for

[31] Maybe a better solution is: If $X_1 = X_2 = \cdots = X_{19} = 1$, then X_{20} is also necessarily 1.

large values of n. What is that constant and how fast does the convergence occur? To be able to answer this question, we first need some basic moment inequalities which we discuss next.

Theorem 31.1 *(Markov's inequality) Suppose that X is a non-negative random variable. Then for every real number $t > 0$, we have*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof: Let f be the pmf of X[32]. We have

$$\mathbb{E}[X] = \sum_{x \in s(X)} xf(x) = \sum_{x<t} xf(x) + \sum_{x \geq t} xf(x) \geq \sum_{x \geq t} xf(x) \geq \sum_{x \geq t} tf(x) = t\mathbb{P}(X \geq t).$$

Comparing the first and the last terms, we obtain

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t},$$

and we are done. □

Markov's inequality is helpful for large values of t. Let's now see an example.

Example 31.1 *(of hats and men) Consider the hats and men problem from Example 30.3 with n men and n hats. Let X be the number of men who find their own hats. For $t > 0$, Markov's inequality gives*

$$\mathbb{P}(X \geq t) \leq \frac{1}{t}.$$

For example, if $t = 50$, then

$$\mathbb{P}(X \geq 50) \leq 0.02.$$

Note that the result is independent of the number of men n. Even if we have a million of men and a million of hats, probability of at least 50 matches is at most 0.02. □

Next we discuss another fundamental inequality: Chebyshev's inequality.

[32] We prove the results for the discrete case, but they will also remain true for the continuous and mixed cases.

Theorem 31.2 *(Chebyshev's inequality) Let X be a random variable for which $Var(X) < \infty$. Then for any $t > 0$, we have*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{Var(X)}{t^2}.$$

Proof: Set $Y = (X - \mathbb{E}[X])^2 \geq 0$. By Markov's inequality,

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

□

Example 31.2 *(of hats and men) Again consider the hats and men problem from Example 30.3 with n men and n hats. Let X be the number of men who find their hats. Recall from Exercise 30.1 that $Var(X) = 1$ independent of n. Then, for any $t > 0$, Chebyshev's inequality gives*

$$\mathbb{P}(|X - 1| \geq t) \leq \frac{1}{t^2}.$$

When $t = 50$, this says

$$\mathbb{P}(|X - 1| \geq 50) \leq \frac{1}{2500}.$$

Observe that this is a much better estimate compared to the one we obtained via Markov's inequality. □

Here is another standard example for the Chebyshev's inequality.

Example 31.3 *Let X be a random variable with $\sigma^2 = Var(X) < \infty$. Then Chebyshev's inequality gives*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq 3\sigma) \leq \frac{\sigma^2}{9\sigma^2} = \frac{1}{9}.$$

So the probability that the random variable will deviate at least three standard deviations from the expected value is bounded by $1/9$. Such statements are frequently heard when you do basic data analysis. □

32 Weak law of large numbers

Let's first recall the following fact for the sample mean $\overline{X}_n = (X_1 + \cdots + X_n)/n$ of n random variables $X_1, \ldots, X_n$: If $X_1, \ldots, X_n$ are i.i.d. random variables with mean μ and variance σ^2, then

$$\mathbb{E}[\overline{X}_n] = \mu \qquad \text{and} \qquad Var(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Now we will get back to the question posed at the beginning of previous section: How close is $\overline{X}_n$ to μ for a given n? First, a definition.

Definition 32.1 *(Convergence in probability) A sequence of random variables $Z_1, Z_2, \ldots$ is said to* ***converge to*** *$b \in \mathbb{R}$* ***in probability*** *if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|Z_n - b| > \epsilon) = 0.$$

We write $Z_n \to_{\mathbb{P}} b$ when Z_n converges in probability to b.

Now we are ready to discuss our main result for this section: Weak law of large numbers. This fundamental result will show that as the sample size gets larger, the sample mean gets closer and closer to actual mean.

Theorem 32.1 *(Weak Law of Large Numbers - WLLN) Let $X_1, X_2, \ldots,$ be i.i.d. random variables each having mean μ and finite variance σ^2. Then*

$$\overline{X}_n \longrightarrow_{\mathbb{P}} \mu,$$

as $n \to \infty$.

Proof: Let $\epsilon > 0$. Using Chebyshev's inequality, we have

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq \frac{Var(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0, \quad \text{as } n \to \infty.$$

□

Weak law of large numbers, though very simple, is one of the most important results in probability theory. Of course it has a stronger version as well, but you will need to take a more advanced course for that. Here we state one other result whose proof is skipped - though interested reader may prove it with some effort.

Theorem 32.2 *Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. Assume that X_n is a sequence of random variables converging to $b \in \mathbb{R}$ in probability. Then $g(X_n)$ converges in probability to $g(b)$.*

Exercise 32.1 *The weak law of large numbers states that the successive arithmetic averages of a sequence of independent and identically distributed random variables converge ini probability to their common mean μ. What do the successive geometric averages converge to? That is, what does $\left(\prod_{i=1}^n X_i\right)^{1/n}$ converge to in probability?*

33 Convergence of histograms

We may use the law of large numbers to see that a histogram can be used as an approximation to a distribution.

Theorem 33.1 *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Let $c_1 < c_2$ be two constants. Define*

$$Y_i = \begin{cases} 1, & \text{if } c_1 \leq X_i < c_2 \\ 0, & \text{if not.} \end{cases}$$

So, $\sum_{i=1}^n Y_i$ counts the number of samples falling in the interval $[c_1, c_2)$. Then $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$, the proportion of $X_1, \ldots, X_n$ that lie in the interval $[c_1, c_2)$, satisfies

$$\overline{Y}_n \rightarrow_{\mathbb{P}} \mathbb{P}(c_1 \leq X_1 < c_2).$$

Proof: By construction, $Y_1, Y_2, \ldots$ are i.i.d. Bernoulli random variables with parameter $p = \mathbb{P}(c_1 \leq X_1 < c_2)$. Result now follows from the weak law of large numbers. □

So if we draw samples successively and form a corresponding histogram with the area of the bar over each subinterval being the proportion of our samples lying in the given subinterval, then the area of each bar converges in probability to the probability that a random variable with the underlying distribution lies in that subinterval. If the sample size is large, we would then expect the area of each bar to be close to the actual probabilities.

34 Discussion of Monte Carlo method

Mathematical methods that use random numbers for solving quantitative problems are commonly known as Monte Carlo methods. The name "Monte Carlo" was coined by Nicholas Constantine Metropolis (1915-1999) and was inspired by Stanslaw Ulam (1909-1986), because of the similarity of statistical simulations to games of chance, and because of the fame of Monte Carlo as a center for gambling and games of chance.

Here we discuss a very basic application of the method where the goal is to approximate the value of π. Consider a circle inscribed inside a square as in the following figure, where the radius of the circle is 1.

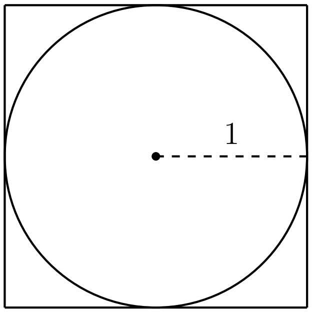


Figure 3: Circle inscribed in a square.

Clearly, we have

$$\frac{\text{Area of circle}}{\text{Area of square}} = \frac{\pi}{4}. \tag{6}$$

Here is how the Monte Carlo method provides an approximation for the value of π in such a setting: First, randomly select a large number of points within the square, and then determine how many of these points fall inside the circle.

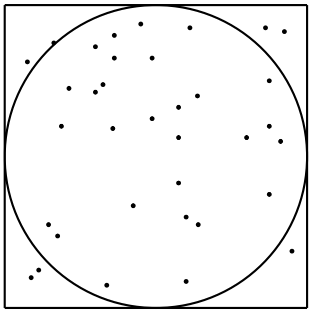


Figure 4: Randomly selected points inside a square with a circle in it.

It is intuitively clear that the ratio of number of points within the circle to the number points within the square approximates the ratio of area of the circle to the area of the square, and so provides an estimate for $\pi/4$. To put it another way, assuming that we have done the experiment by using N many random points in the unit square, and that M of these turned out to fall inside the circle, the relation in (6) suggests the value

$$\hat{\pi} \approx 4\frac{M}{N}. \tag{7}$$

as an approximation for π. How well does the approximation in (7) work? By using the weak law of large numbers we can easily show that as N tends to infinity, 4 times (number of points lying in the circle $/N$) will converge to π in probability - take it as an exercise to write this rigorously.

An important instance where the Monte Carlo method becomes very useful is when we need to approximate the value of an integral of some complicated (usually high dimensional) function. For a simple example, let us consider a non-negative bounded function f defined over an interval $[a, b]$ where we are interested in evaluating the value of $\int_a^b f(x)dx$. Let $M = \max_{x\in[a,b]} f(x)$, and define $g(x) = M$ for $x \in [a, b]$ and $g(x) = 0$, otherwise.

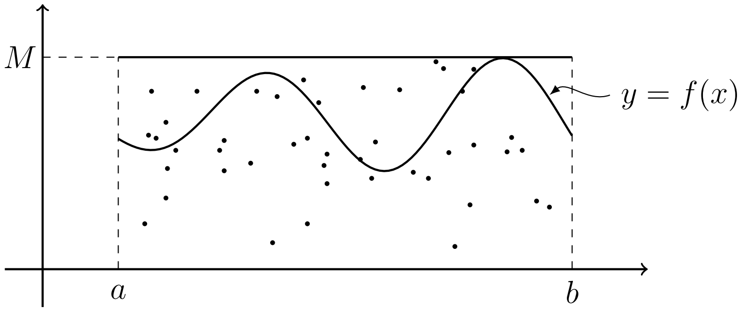


Figure 5: Random points in a bounded region (rectangle) containing the graph of a function.

Assume now that we choose N points over the region $[a, b] \times [0, M]$ uniformly at random, and let H be the number of sampled points that lie in the set $\{(x, y) : a \le x \le b, 0 \le y \le f(x)\}$. Then the Monte Carlo method suggests that for large N

$$\frac{N}{H} \approx \frac{\text{Area under } g}{\text{Area under } f}.$$

Since area under g is just $(b - a)M$, one concludes that

$$\int_a^b f(x)dx \approx \frac{H}{N}(b - a)M.$$

35 * Slightly more complicated example: Buffon's needle

The purpose of this section is to discuss a slightly more complicated experiment than the one described in Introduction for approximating the value of π. The experiment was first suggested by a French nobleman Georges Louis Leclerc, Comte de Buffon (1707-1788). In particular, Buffon poses the following question:

Problem. Assume that we have a needle which has length ℓ. We drop it on the $x - y$ plane in which we draw parallel lines having equal distance d. What is the probability that the needle intersects with one of the lines on the plane?

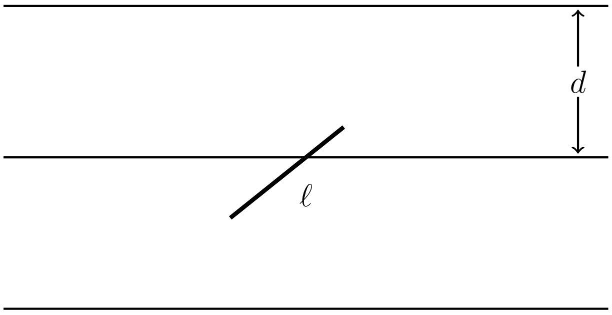


Figure 6: A needle thrown at random on $x - y$ plane with parallel lines

We will solve the problem in two complementary cases $\ell \leq d$ and $\ell > d$ seperately. Discussion below will make the necessity for this distinction clear.

First, suppose $\ell \leq d$. When a needle, which has length ℓ and a negligible thickness, is dropped on the x-y plane with parallel lines having distance d to their neighbor lines, the variables of interest for us will be (1) y: the distance from the midpoint of the needle to the nearest line and (2) θ: the angle which ranges between the line and the needle. In particular, since the needle is thrown at random, the values of y and θ are independent. Note that $0 \leq y \leq \frac{d}{2}$ and $0 \leq \theta \leq \pi$.

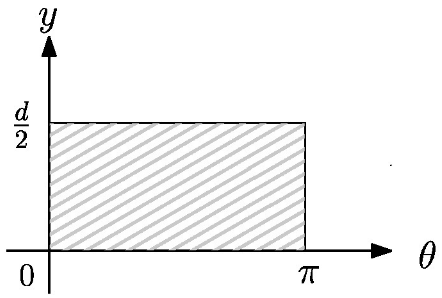


Figure 7: Possible values of (θ, y).

Next, a little thought should convince the reader that the needle will intersect a line if and only if

$$y \leq \frac{\ell}{2} \sin \theta, \tag{8}$$

and so the problem reduces to finding the ratio of the shaded area to the area of the outer rectangle in Figure 8.

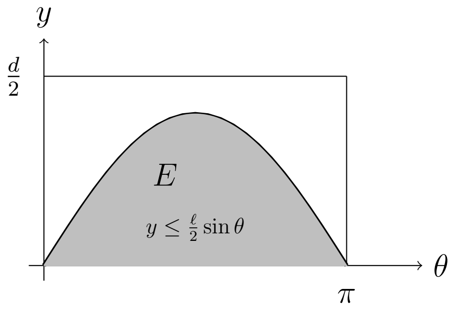


Figure 8: Shaded area represents all possible intersections of the needle with one of the lines.

Now we find the area of E as

$$\int_0^{\pi} \frac{\ell}{2}\sin\theta d\theta = \frac{\ell}{2}[-\cos(\pi) - (-\cos(0))] = \frac{\ell}{2}\cdot 2 = \ell$$

So

$$\text{Probability that needle intersects a line} = \frac{\text{Area}(E)}{\text{Area}(rectangle)} = \frac{\ell}{\frac{d}{2}\cdot\pi} = \frac{2}{\pi}\cdot\frac{\ell}{d}$$

solving the problem for $\ell \leq d$.

Next, let us consider the case $\ell > d$. This time the shaded area in following figure represents all cases with no intersections. So we would like to find the probability of (θ, y) coordinates that do not lie in the shaded area.

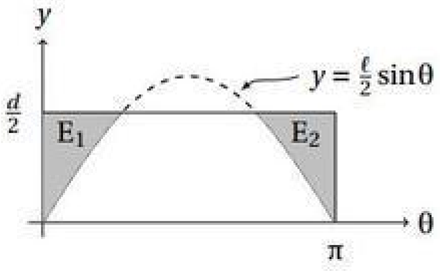


Figure 9: E_1 and E_2 represent all cases in which there is no intersection.

Denoting the probability of the non-shaded region by p, and considering the underlying symmetry, we have

$$\begin{aligned} p &= 2\int_0^{\pi/2}\int_0^{\min\{d/2,(\ell\sin\theta)/2\}} \frac{1}{(\pi d)/2} dy d\theta \\ &= 2\left[\int_0^{\arcsin(d/\ell)}\int_0^{(\ell\sin\theta)/2} \frac{2}{\pi d} dy d\theta + \int_{\arcsin(d/\ell)}^{\pi/2}\int_0^{d/2} \frac{2}{\pi d} dy d\theta\right] \\ &= \frac{4}{\pi d}\left[\int_0^{\arcsin(d/\ell)} \frac{\ell\sin\theta}{2} d\theta + \int_{\arcsin(d/\ell)}^{\pi/2} \frac{d}{2} d\theta\right] \\ &= \frac{4}{\pi d}\left[\frac{\ell}{2}\left(1-\cos(\arcsin(d/\ell))\right) + \frac{d}{2}(\pi/2 - \arcsin(d/\ell))\right] \\ &= 1 - \frac{2}{\pi}\arcsin(d/\ell) + \frac{2\ell}{\pi d} - \frac{2}{\pi d}\sqrt{\ell^2-d^2}, \end{aligned}$$

where the last step requires some elementary operations.

36 * A funny story about Monte Carlo: Lazzarini's experiment

Laplace was the first mathematician to realize that Buffon's problem can be used to approximate the value of π. Here we mainly discuss another subsequent work from the very early 20th century, by Mario Lazzarini. Lazzarini, in his paper in 1901, reported that he did an experiment with $\ell = 2.5$ cm, $d = 3$ cm, and that he threw 3408 needles, and that 1808 of these intersected a line. Such an experiment reveals

$$\hat{\pi} = \frac{2(2.5)}{3(3408/1808)} = 3.1415929...$$

as an approximation for $\pi = 3.1415926...$. Such an experiment looks suspicious to mathematicians of our own century as the experimental value is "too close" to the actual value π. Let us next discuss why that is the case?

1. Why 3408 and not a round number like 3500, and also similarly why 1808 and not 1800 instead? Such choice of numbers become even more interesting when we change 1808 to either 1807 or 1809, for which we respectively obtain 3.1433... and 3.1398... as estimates of π (not even three decimal digits are correct!).

2. Maybe the more gossipry fact is the following equation

$$\hat{\pi} = 2 \cdot \frac{2.5}{3} \cdot \frac{3408}{1808} = \frac{355}{113}$$

This ratio is a famous rational approximation for π discovered by Chinese mathematician Tsu Ch'ung-chih in the fifth century. This rational approximation was well-known to many mathematicians of 20th century, including probably Lazzarini.

3. Indeed one may go one step beyond and may question the probability of obtaining an estimate with a certain accuracy. For example, let us consider the case where we would like to be 95% confident that $|\pi - \hat{\pi}| < 0.5 \times 10^{-6}$. For such confidence one requires to throw approximately 134 trillion needles! We sketch a brief argument why that is the case - it is highly likely that you won't be able to follow what is going on here. In that case, come back to this at the end of the semester again.

   Denoting the number of trials we do by N, the number of intersections we obtain by H, and using the setting in Lazzarini's experiment, having a 95% confidence is equivalent to
   $$\mathbb{P}\left(\left|\pi - \frac{5N}{3H}\right| < 0.5 \times 10^{-6}\right) \geq 0.95.$$
   Now it can be easily shown that
   $$\left|\pi - \frac{5N}{3H}\right| < 0.5 \times 10^{-6} \qquad \text{if and only if} \qquad \left|\frac{5N}{3\pi} - H\right| < \frac{5N}{3\pi^2}(0.5) \times 10^{-6}.$$
   Noting that H is binomially distributed with parameters N and $p = 5/(3\pi)$, using the central limit theorem (which we will learn about in upcoming sections), one can easily conclude that we need
   $$\frac{\frac{5N}{3\pi^2}(0.5)10^{-6}}{\sqrt{Np(1-p)}} \approx 1.96$$
   in order to have
   $$\mathbb{P}\left(\left|\frac{3N}{5H} - H\right| < \frac{5N}{3\pi^2}(0.5) \times 10^{-6}\right).$$
   In other words, N should be approximately 134×10^{12}.

4. Let us finally include some simulation results we performed to support the argument in item 3:

N	$\overline{\pi}$
1	4
10	3.3333333333333335
10^2	2.7777777777777777
10^3	3.003003003003003
10^4	3.1501023783272957
10^5	3.146831141040972
10^6	3.1460341879535205.

37 Review problems III

Exercise 37.1 *Let $X_1, \ldots, X_n$ be independent random variables with finite mean μ and finite variance σ^2. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean of $X_1, \ldots, X_n$.*

i. Find the expectation of $\overline{X}$.

ii. Find the variance and standard deviation of $\overline{X}$.

iii. Show that $Cov(X_i - \overline{X}, \overline{X}) = 0$ for each $i = 1, 2, \ldots, n$.

Exercise 37.2 *The random pair (X, Y) has the distribution*

		X		
		1	2	3
Y	2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
	3	$\frac{1}{6}$	0	$\frac{1}{6}$
	4	0	$\frac{1}{3}$	0

a. Show that X and Y are dependent.

b. Give a probability table for random variables U and V that have the same marginals as X and Y but are independent.

Exercise 37.3 *Show that any random variable is uncorrelated with a constant.*

Exercise 37.4 *Let X and Y be independent random variables with means μ_X, μ_Y and variances σ_X^2, σ_Y^2. Find an expression for the correlation of XY and Y in terms of these means and variances.*

Exercise 37.5 *Let X_1, X_2, and X_3 be uncorrelated random variables, each with mean μ and variance σ^2. Find, in terms of μ and σ^2, $Cov(X_1+X_2, X_2+X_3)$ and $Cov(X_1+X_2, X_1-X_2)$.*

Exercise 37.6 *Calculate $\mathbb{P}(|X-\mu_X| \geq k\sigma_X)$ for $X \sim U(0,1)$ and $X \sim \exp(\lambda)$, and compare your answers to the bound from Chebyshev's inequality.*

Exercise 37.7 *Suppose $\overline{X}$ and S^2 are calculated from a random sample $X_1, \ldots, X_n$ drawn from a population with finite variance σ^2. We know that $\mathbb{E}[S^2] = \sigma^2$. Prove that $\mathbb{E}[S] \leq \sigma$, and if $\sigma^2 > 0$, then $\mathbb{E}[S] < \sigma$.*

Exercise 37.8 *Let X be a random variable. Show that if $\mathbb{E}[X^2] = 1$ and $\mathbb{E}[X^4] < \infty$, then*

$$\mathbb{E}[|X|] \geq \frac{1}{\sqrt{\mathbb{E}[X^4]}}.$$

(Hint: Use Cauchy-Schwarz inequality.)

Exercise 37.9 *Let $X_1, X_2, \ldots$ be a sequence of random variables that converges in probability to a constant a. Assume that $\mathbb{P}(X_i > 0) = 1$ for all i.*

a. Show that the sequence $Y_1, Y_2, \ldots$, defined by $Y_i = \sqrt{X_i}$ converges in probability to $\sqrt{a}$.

b. Show that, if $a > 0$, the sequence $Y_1, Y_2, \ldots$ defined by $Y_i = a/X_i$, converges in probability to 1.

c. Show that σ/S_n converges in probability to 1.

Exercise 37.10 *Suppose that it is known that the number of items produced in a factory during a week is a random variable with mean 50.*

i. What can be said about the probability that this week's production will exceed 75?

ii. If the variance of a week's production is known to equal 25, then what can be said about the probability that this week's production will be between 40 and 60?

Exercise 37.11 *Suppose that X is a nonnegative random variable such that $\mathbb{E}[X^k]$ exists. Prove that for $k > 0$ and $t > 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X^k]}{t^k}. \tag{9}$$

(Hint: Use the reasoning behind the proof of Chebyschev's inequality.) (Note that this is a generalization of Markov's inequality.)

Exercise 37.12 *Let g be strictly increasing and nonnegative. Show that*

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[g(|X|)]}{g(a)}, \qquad for \qquad a > 0.$$

Exercise 37.13 *Let X be binomial with parameters n and p with $p > 0$. Show that for $\lambda > 0$ and $\epsilon > 0$,*

$$\mathbb{P}(X - np \geq n\epsilon) \leq \mathbb{E}[\exp(\lambda(X - np - n\epsilon))].$$

(Hint: Exponential function is an increasing function.)

Exercise 37.14 *Let $X_1, \ldots, X_{10}$ be random variables with $\mathbb{E}[X_i] = 0$, $Var(X_i) = 1$ for each $i = 1, \ldots, 10$. Also assume that the correlation between each pair is $1/4$.*

i. *Find* $\mathbb{E}\left[\sum_{i=1}^{10}(iX_i + X_i^2)\right]$.

ii. *Find the variance and the standard deviation of* $X_1 + \cdots + X_{10}$.

Exercise 37.15 *There are* n *types of coupons. Each newly obtained coupon is, independently, type* i *with probability* p_i, $i = 1, \ldots, n$. *Let* X *be the number of distinct types obtained in a collection of* k *coupons. Find the expected value and variance of* X. *(Hint: For* $i = 1, \ldots, n$, *let* $X_i = 1$ *if* i^{th} *kind of coupon is in sample, and let* $X_i = 0$.*)*

Exercise 37.16 *Let* X *be the number of 1's and* Y *the number of 2's that occur in* n *rolls of a fair die (where we assume that the result of each roll is independent). Compute* $Cov(X, Y)$.

Exercise 37.17 *Let* $X_1, X_2, \ldots$ *be i.i.d. random variables having exponential distribution with parameter 1.*

i. *Find the distribution of* $\max\{X_1, \ldots, X_n\}$ *and* $\min\{X_1, \ldots, X_n\}$.

ii. *Find* $\mathbb{E}\left[\prod_{i=1}^n iX_i\right]$ *and* $\mathbb{E}\left[\prod_{i=1}^n iX_i^2\right]$.

iii. *Find the distribution of* $V_1 = e^{-X_1}$. *What is* $\mathbb{E}[V_1]$*?*

iv. *What can you say about* $Y_n = e^{-\frac{1}{n}\sum_{i=1}^n X_i}$ *for large values of* n*?*

Exercise 37.18 *A sequence of random variables* $X_1, X_2, \ldots$ *is said to* ***converge to*** b ***in quadratic mean*** *if*

$$\lim_{n\to\infty} \mathbb{E}[(X_n - b)^2] = 0. \tag{10}$$

Show that (10) is satisfied if and only if

$$\lim_{n\to\infty} \mathbb{E}[X_n] = b \qquad \text{and} \qquad \lim_{n\to\infty} Var(X_n) = 0.$$

Exercise 37.19 *Prove that if a sequence of random variables* $X_1, X_2, \ldots$ *converges to a constant* b *in quadratic mean, then the sequence also converges to* b *in probability.*

Exercise 37.20 *Let* $X_1, X_2, \ldots$ *be a sequence of random variables, and suppose that for* $n = 1, 2, 3, \ldots$, *the distribution of* X_n *is as follows:*

$$\mathbb{P}\left(X_n = \frac{1}{n}\right) = 1 - \frac{1}{n^2} \qquad \text{and} \qquad \mathbb{P}(X_n = n) = \frac{1}{n^2}.$$

a. Does there exist a constant c *to which the sequence converges in probability? b. Does there exist a constant* c *to which the sequence converges in quadratic mean?*

Exercise 37.21 *Let $X_1, X_2, \ldots$ be a sequence of random variables for which $\mathbb{E}[X_i] = 0$, $Var(X_i)/i \to 0$ as $i \to \infty$, and for $i \neq j$, $Cov(X_i, X_j) \leq \frac{1}{(j-i)^2}$. Let $S_n = \sum_{j=1}^n X_j$. Prove that S_n/n converges to 0 in probability and in quadratic mean.*

Exercise 37.22 *A sequence of mean-zero random variables $(X_n)_{n\in\mathbb{N}}$ is called weakly stationary if there is a function ϕ such that*

$$\mathbb{E}[X_i X_j] = \phi(j-i) < \infty \qquad \forall i, j.$$

Suppose that for some such sequence, we have $\phi(k) \to 0$ as $k \to \infty$. Show that the weak law of large numbers is valid, that is, $\frac{X_1+\cdots+X_n}{n}$ converges to 0 in probability.

Exercise 37.23 *How can we use Buffon's needle experiment to approximately find π?*

38 Likelihood estimation in discrete setting

Definition 38.1 *(In statistics terminology) A function of random variables is called a **statistic**.*

Example 38.1 *Let $X_1, \ldots, X_n$ be i.i.d. random variables[33]. Then some examples of statistics are:*

$$T_1 = \overline{X}_n, \qquad T_2 = S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2, \qquad T_3 = \left(\sum_{i=1}^n X_i^2\right)^{1/2}.$$

□

Noting that an arbitrary constant whose value characterizes a distribution is called a **parameter**, we have the following definition.

Definition 38.2 *A statistic $T(X_1, \ldots, X_n)$ used for estimating a parameter θ of the underlying distribution is called an **estimator**. The set of all possible values of θ is called the **parameter space**, and is denoted by Θ.*

Example 38.2 *Assuming that $X_1, \ldots, X_n$ are random samples from a Bernoulli distribution with parameter p, it should be intuitively clear that $\overline{X}$ is an estimator of p. In this case the parameter space, all possible values of p, is $\Theta = (0, 1)$- we exclude $p = 0, 1$ in order to avoid trivialities.* □

[33] You will often see "Let $X_1, \ldots, X_n$ is a **random sample**" instead of "Let $X_1, \ldots, X_n$ be i.i.d. random variables"

Definition 38.3 *If $X_1, \dots, X_n$ are i.i.d. samples from a distribution with pmf $f(x \mid \theta)$ where $\theta \in \Theta$, then*

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

*is called the **likelihood function**. The function*

$$\ell(\theta) = \ln L(\theta)$$

*is called the **log-likelihood function**.*

Given some samples $x_1, \dots, x_n$, $L(\theta) = L(\theta \mid x_1, \dots, x_n)$ is a measure of how likely to obtain these samples assuming that the underlying parameter is θ.

Definition 38.4 *If the function of the random sample $X_1, \dots, X_n$ that maximizes $L(\theta)$ is $T(X_1, \dots, X_n)$, then*

$$\hat{\theta} = T(X_1, \dots, X_n)$$

*is said to be a **maximum likelihood estimator** of θ. The corresponding observed value of this statistic, $T(x_1, \dots, x_n)$, is called a **maximum likelihood estimate**.*

Example 38.3 *(Bernoulli distribution) Assume that $X_1, \dots, X_n$ are i.i.d. samples from the Bernoulli distribution*

$$f(x \mid p) = p^x (1-p)^{1-x}, \qquad x = 0, 1, \qquad 0 < p < 1.$$

(a) Recalling that $\mathbb{E}[X_1] = p$, provide an intuitive guess for the estimator.

(b) Find the maximum likelihood estimator $\hat{p}$ of p.

(c) Assume that a sample of size 10 turns out to be $\mathbf{x} = (x_1, \dots, x_{10}) = (1, 0, 0, 0, 0, 1, 1, 1, 0, 0)$. Find the maximum likelihood estimate for p.

Solution: (a)The intuitive guess would be $\overline{X}_n$ - Why?

(b) The likelihood function and the log-likelihood functions are respectively given by

$$L(p \mid X_1, \dots, X_n) = \prod_{i=1}^{n} p^{X_i} (1-p)^{1-X_i},$$

and

$$\begin{aligned} \ell(p \mid X_1, \dots, X_n) = \ln L(p \mid X_1, \dots, X_n) &= \ln \left(\prod_{i=1}^{n} p^{X_i} (1-p)^{1-X_i} \right) \\ &= \sum_{i=1}^{n} X_i \ln p + \sum_{i=1}^{n} (1 - X_i) \ln(1-p). \end{aligned}$$

We would like to maximize L or ℓ[34] over p. In order to do so, we first note that ℓ is differentiable on $\Theta = (0,1)$, and the only possible critical points are therefore the solutions to

$$\ell'(p) = 0.$$

We find

$$\ell'(p) = \frac{1}{p}\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \frac{1-X_i}{1-p} = \frac{(1-p)\sum_{i=1}^{n} X_i - p\left(n - \sum_{i=1}^{n} X_i\right)}{p(1-p)}$$

and solving $\ell'(p) = 0$ provides

$$\hat{p}(X_1, \ldots, X_n) = \overline{X}_n$$

as a candidate for the maxima. To verify that the maximum is indeed attained at this point we check the second derivative

$$\ell''(p) = -\frac{1}{p^2}\sum_{i=1}^{n} X_i + \frac{1}{(1-p)^2}\sum_{i=1}^{n}(1-X_i),$$

and a simple computation shows

$$\ell''(\overline{X}_n) < 0.$$

Therefore via the second derivative test, we conclude that the maximum likelihood estimator of p is

$$\hat{p} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

(c) The estimate for p is given by

$$\overline{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{4}{10} = 0.4.$$

□

39 MLE examples

Example 39.1 *(Geometric distribution) Let $X_1, \ldots, X_n$ be a random sample from the geometric distribution with pmf $f(k \mid p) = p(1-p)^{k-1}$, $k \geq 1$, $p \in (0,1)$.*

(a) Recalling that $\mathbb{E}[X_1] = 1/p$, provide an intuitive guess for the estimator.

(b) Find the maximum likelihood estimator for p.

(c) Assume that we have the following 8 samples from the underlying distribution $\mathbf{x} = (2,4,1,9,5,1,2,8)$. Find the maximum likelihood estimate on the data you have.

[34] Note that the maximum is preserved under a strictly increasing function - still this will be an exercise for you.

Solution: (a)The intuitive guess would be $1/\overline{X}$ - Why?

(b) The likelihood and the log-likelihood functions are respectively given by

$$L(p \mid X_1,\dots,X_n)=\prod_{i=1}^{n} p(1-p)^{X_i-1}=p^n(1-p)^{\left(\sum_{i=1}^n X_i\right)-n},$$

and

$$\ell(p \mid X_1,\dots,X_n)=n\ln p+\left(\left(\sum_{i=1}^{n} X_i\right)-n\right)\ln(1-p).$$

Setting the first derivative $\ell'(p)$ to zero gives

$$\ell'(p)=\frac{n}{p}+\frac{n-\sum_{i=1}^n X_i}{1-p}=0$$

and solving this yields $\hat{p}=1/\overline{X}$ as the candidate for the MLE. Noting

$$\ell''(p)=-\frac{n}{p^2}-\frac{\sum_{i=1}^n X_i-n}{(1-p)^2}<0,$$

we have justified that this is indeed the case via the second derivative test.

(c) The maximum likelihood estimate is $1/\overline{x}=1/4$. □

Example 39.2 *(Uniform distribution) Let $X_1,\dots,X_n$ be a random sample from the pmf*

$$f(k)=\frac{1}{\theta},\qquad k=1,2,\dots,\theta,$$

where the positive integer θ is the unknown parameter. Find the maximum likelihood estimator for θ.

Solution: Note that since X_i's are uniformly distributed over $\{1,\dots,\theta\}$, θ has to be necessarily at least $\max\{X_1,\dots,X_n\}$. We will show that the MLE is indeed exactly the maximum of these samples[35]. To do so, first observe that the likelihood function is given by

$$\begin{aligned} L(\theta \mid X_1,\dots,X_n)=\prod_{i=1}^{n} f(X_k\mid\theta)=\prod_{i=1}^{n}\frac{1}{\theta}\mathbf{1}(1\le X_k\le\theta) &= \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbf{1}(1\le X_k\le\theta)\\ &= \frac{1}{\theta^n}\mathbf{1}(\max\{X_1,\dots,X_n\}\le\theta). \end{aligned}$$

The likelihood function is zero when $\theta<\max\{X_1,\dots,X_n\}$, is strictly positive for $\theta\ge\max\{X_1,\dots,X_n\}$, and is strictly decreasing again for $\theta\ge\max\{X_1,\dots,X_n\}$. From these observations we conclude that

$$\hat{\theta}=\max\{X_1,\dots,X_n\}$$

is indeed the MLE for θ. □

[35] Can you give an intuitive explanation for this?

Example 39.3 *Let X be a discrete random variable with the following pmf:*

$$f(x) = \begin{cases} \theta, & \text{if } x = 3, 5, \\ 1 - 2\theta, & \text{if } x = 7 \\ 0, & \text{otherwise}, \end{cases}$$

where $0 \leq \theta \leq 1$ is the unknown parameter. Assume that the 12 samples $(x_1, \ldots, x_{12}) = (5, 3, 3, 7, 3, 7, 5, 5, 5, 3, 3, 5)$ were taken from such a distribution independently. What is the maximum likelihood estimate of θ?

Solution. The likelihood function is

$$L(\theta) = \prod_{i=1}^{12} f(x_i \mid \theta) = \theta^{10}(1 - 2\theta)^2.$$

Taking ln on both sides, the log-likelihood function is then

$$\ell(\theta) = 10 \ln \theta + 2 \ln(1 - 2\theta).$$

Setting $\ell'(\theta) = 0$ yields $\theta = 5/12$, and the second derivative test justifies that the maximum likelihood estimate is $5/12$. □

40 Mark and recapture

Mark and recapture is a method commonly used in ecology to estimate an animal population in a certain region. As an example, we might be interested in finding the (approximate) number of a certain species of some fish in a lake. The method works as follows:

First, a portion of the population is captured, tagged, and released. Later, some other portion is captured and the number of tagged individuals within the sample is counted. Since the number of tagged individuals within the second sample should be proportional to the number of tagged individuals in the whole population, an estimate of the total population size can be obtained by dividing the number of tagged individuals by the proportion of tagged individuals in the second sample.

For a rigorous discussion, let's introduce the variables in mark and recapture. Let

- t be the number captured and tagged,
- k be the total number in the second capture
- r be the the number in the second capture that are tagged

- N be the total population.

Here t and k are set by the experimental design; r is a random number that may vary at each observation. The total population N is the unknown we would like to understand. The likelihood function for N is the hypergeometric distribution

$$L(N \mid r) = \frac{\binom{t}{r}\binom{N-t}{k-r}}{\binom{N}{k}}.$$

Question. Find the maximum likelihood estimator of N given that the number of recaptured individuals is r.

In order to approach this problem, we first observe that

$$\frac{L(N+1 \mid r)}{L(N \mid r)} = \frac{\binom{N+1-t}{k-r}/\binom{N+1}{k}}{\binom{N-t}{k-r}/\binom{N}{k}} = \frac{N-t+1}{N-t-k+r+1}\frac{N-k+1}{N+1}.$$

Some algebra shows that the right hand side is larger than 1 if and only if $N < \frac{tk}{r}$. So the sequence $L(N \mid r)$ increases when $N < \frac{tk}{r}$, and then starts decreasing. Therefore, the maximum likelihood estimator of N is $\lfloor \frac{tk}{r} \rfloor$, where $\lfloor \cdot \rfloor$ is the least integer function.

Let's give an example to make sure that the estimator we have found is what one would expect intuitively. Say again that we are interested in approximately counting the number of fish in a lake. For this purpose we capture and tag 500 of the fishes, and rthen elease them. Afterwards we capture 1000 fish and 100 of them turn out to be tagged. Intuitively, this means that we tagged 10% of all fishes in the lake initially and so the plausible guess for our problem would be 5000. This is also what we obtain using our maximum likelihood estimator.

41 Bias and consistency

In general, one will have several alternative estimators to estimate a certain parameter, and will need to choose one of them according to some criteria. In this section we discuss two important properties of estimators that might help: bias and consistency.

Definition 41.1 *(i) Suppose that we have a sequence of samples $X_1, X_2, \ldots$ from some common distribution*[36] *F having some parameter θ. An estimator $T = T(X_1, \ldots, X_n)$ is said*

[36] By saying common distribution F, we mean the unique distribution specified by a cumulative distribution function F.

to be an ***unbiased estimator*** *for θ if*

$$\mathbb{E}[T(X_1,\dots,X_n)] = \theta.$$

Otherwise, the estimator is said to be ***biased*** *and the* ***bias*** *is defined*

$$Bias(T;\theta) = \mathbb{E}[T(X_1,\dots,X_n)] - \theta.$$

(ii) If

$$\lim_{n\to\infty} \mathbb{E}[T(X_1,\dots,X_n)] = \theta,$$

then T is called an ***asymptotically unbiased estimator*** *of θ.*

(iii) $T = T(X_1,\dots,X_n)$ is said to be a ***consistent estimator*** *for θ if*

$$T(X_1,\dots,X_n) \to_{\mathbb{P}} \theta$$

as $n \to \infty$. That is $T(X_1,\dots,X_n)$ converges in probability to θ as n tends to infinity.

Example 41.1 *Let $X_1,\dots,X_n$ be i.i.d. random variables with finite mean $\mu = \mathbb{E}[X_1]$. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean.*

(i) Show that $\overline{X}_n$ is an unbiased estimator for μ.

(ii) If $Var(X_1) < \infty$, then show that $\overline{X}_n$ is a consistent estimator for μ.

Solution. (i) We have

$$\mathbb{E}[\overline{X}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^n \mu = \mu.$$

(ii) Follows from weak law of large numbers. □

Example 41.2 *(i.) For any real numbers $x_1,\dots,x_n$, show that*

$$\sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n x_i^2 - n\overline{x}^2.$$

(ii.) Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables for which $\mathbb{E}[X_1^4] < \infty$ and set $\sigma^2 = Var(X_1)$. Also define

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2.$$

Show that S_n^2 is an unbiased estimator for σ^2.

(iii) Show that

$$R_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$$

is a biased, but asymptotically unbiased estimator.

Solution. (i) exercise.

(ii) Using part (i) and linearity of expectation, observe that

$$\mathbb{E}[S_n^2] = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \frac{n}{n-1}\mathbb{E}[\overline{X}^2] = \frac{1}{n-1}\left(n(\mu^2+\sigma^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right) = \sigma^2.$$

(iii) We have

$$\mathbb{E}[R_n^2] = \frac{n}{n-1}\mathbb{E}[S_n^2].$$

Since S_n^2 is an unbiased estimator, this implies R_n^2 is biased.[37]

Noting that $\lim_{n\to\infty} n/(n-1) = 1$ and $\mathbb{E}[S_n^2] = \sigma^2$, we also see that R_n^2 is an asymptotically unbiased estimator. □

Example 41.3 *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with positive and finite variance, and with finite mean $\mu = \mathbb{E}[X_1]$. Let*

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

(i) Show that T is an unbiased estimator for μ.

(ii) Show that T is not a consistent estimator for μ.

(iii.) Conclude that we may have unbiased but inconsistent estimators. □

Solution. (i) Observe that

$$\mathbb{E}[T(X_1,\ldots,X_n)] = \mathbb{E}[X_1] = \mu,$$

and so T is an unbiased estimator.

(ii) X_1 clearly does not converge to μ in probability as $n \to \infty$.

(iii) Clear from first two parts. □

42 Poisson distribution and basic properties

Poisson distribution is in general used to model the number of events occurring in a fixed interval of time (and/or space) if these events occur with a known average rate and independently of the time since the last event.

Here are some examples from the literature where one often uses the Poisson distribution to model the underlying randomness.

[37] This is one reason why we have $n-1$ instead of n in definition of sample variance, but this is not the whole story.

1. The number of soldiers in the Prussian army killed accidentally hit by horse kicks in a day (Ladislaus Bortkiewicz, 1898[38]).
2. The number of e-mails you receive in a day.
3. The number of customers arriving at a coffee shop between 8:30am and 8:40am.
4. The number of typos in a given page in these lecture notes.
5. The number of births expected during the night in a hospital.
6. The number of suicides committed in Istanbul tomorrow.

Definition 42.1 *Let $\lambda > 0$. A random variable X has the* ***Poisson distribution*** *with parameter λ if its pmf is given by*

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

We use the notation $X \sim PO(\lambda)$ to say that X is Poisson random variable with parameter λ. It is easy to show that $f(x \mid \lambda)$ defines a pmf, just remember the Taylor expansion for e^{λ} (Check it.).

Theorem 42.1 *Let $X \sim PO(\lambda)$. Then*

i. $\mathbb{E}[X] = \lambda$.

ii. $Var(X) = \lambda$.

Proof: (i.) Observe

$$\mathbb{E}[X] = e^{-\lambda}\sum_{x=0}^{\infty}\frac{x\lambda^x}{x!} = e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = e^{-\lambda}\lambda e^{\lambda} = \lambda.$$

(ii.) The variance case is left as an exercise for you. (Hint: Similar to previous part. Start by computing $\mathbb{E}[X(X-1)]$ and keep taking the second derivative of a power series in mind.)

Remark 42.1 *Having the same expectation and variance is a useful property of Poisson distribution. When the sample mean and the sample variance tend to be close, one should keep the Poisson distribution in mind as a candidate distribution for modeling the underlying data.*

[38] First use of Poisson distribution in modeling a real life problem.

A nice property, whose proof is left for you, enjoyed by the Poisson distribution is:

Theorem 42.2 *If the random variables $X_1, \dots, X_k$ are independent Poisson random variables with parameters $\lambda_1, \dots, \lambda_k$, respectively, then $X_1 + \dots + X_k \sim PO(\lambda_1 + \dots + \lambda_k)$. (Hint: Showing it for $k = 2$ requires elementary probability theory. Then use induction.)*

Example 42.1 *Let X_1 and X_2 be the number of men and women arriving at a certain coffee shop in a given 10 minutes period. Assume X_1 and X_2 are independent[39], $X_1 \sim PO(3)$ and $X_2 \sim PO(4)$. Then $Z = X_1 + X_2$, number of all customers arriving during this 10 minutes period, has Poisson distribution with parameter 7.*

43 Poisson approximation

In this section we discuss the Poisson approximation to the binomial distribution. First recall the following fact from calculus.

Proposition 43.1 *If $a_n \to 0$, $c_n \to \infty$, and $a_n c_n \to b$ as $n \to \infty$, then $(1 + a_n)^{c_n} \to e^b$ as $n \to \infty$.*

Theorem 43.1 *For each integer n and each $p \in (0,1)$, let $f(x \mid n, p)$ be the binomial pmf with parameters n and p. Let $f(x \mid \lambda)$ denote the Poisson pmf with parameter $\lambda > 0$. Also let p_n be a sequence of real numbers between 0 and 1 such that $\lim_{n\to\infty} np_n = \lambda$. Then for any $x \in \mathbb{N}$,*

$$\lim_{n\to\infty} f(x \mid n, p_n) = f(x \mid \lambda).$$

Proof: We have

$$f(x \mid n, p_n) = \binom{n}{x} p_n^x (1 - p_n)^{n-x} = \frac{n(n-1)\dots(n-x+1)}{x!} p_n^x (1-p_n)^{n-x}.$$

Let $\lambda_n = np_n$ so that $\lim_{n\to\infty} \lambda_n = \lambda$. Then

$$f(x \mid n, p_n) = \frac{\lambda_n^x}{x!} \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-x+1}{n} \left(1 - \frac{\lambda_n}{n}\right)^{-x} \left(1 - \frac{\lambda_n}{n}\right)^{n} \longrightarrow \frac{\lambda^x e^{-\lambda}}{x!} = f(x \mid \lambda),$$

as $n \to \infty$, where for the last step we used Proposition 43.1. □

So, when p_n is small and np_n is around λ, probabilities related to a binomial random variable with parameters n and p_n can be approximated with a Poisson distributed random variable with parameter $\lambda = np_n$.

[39] Do you think that these would be independent?

Example 43.1 *Assume that $X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables each with parameter $1/n$, where $n \geq 1$. Theorem 43.1 says that the distribution of $\sum_{i=1}^n X_i$ is approximately Poisson with mean 1 when n is large. This is to say that binomial probabilities can be approximately found by using the corresponding Poisson probabilities where the parameter is 1. This is especially useful when the probabilities for the binomial case are complicated. As a specific example, if $n = 10^7$, it is quite involved to find the probabilities such as $\mathbb{P}\left(\sum_{k=1}^n X_k \leq 3\right)$ because of the factorial terms that will appear. However, these can be approximately found easily using the Poisson approximation.*

□

Example 43.2 *In this problem, we assume that 600,000 marriages took place in the state of Istanbul last year*[40]*. We are interested in the birthdays of these 1,200,000 people. Let's assume that births occur independently and that each birth is uniformly distributed over 365 days. (That is, ignore the possibility of someone having been born on February 29) Let X be the number of all couples among these 600,000 who were both born on June 22.*

(i) Find an exact expression for the probability $\mathbb{P}(X \leq 2)$.

(ii) Use Poisson distribution to approximate the probability you found in part ii.

Solution: *(i) Probability that any two specific people will have the same birthday is given by $p_0 = \frac{1}{365^2}$. Then X has binomial distribution with parameters $600,000$ and p_0. So*

$$\mathbb{P}(X \leq 2) = \sum_{k=0}^{2} \binom{600,000}{k} p_0^k (1 - p_0)^{600,000-k}.$$

(ii) The mean number of pairs having the same birthday is $\lambda_0 = (600,000)p_0$. Then the Poisson approximation of the exact probability is given by $\mathbb{P}(X \leq 2) \approx \sum_{k=0}^{2} \frac{\lambda_0^k e^{-\lambda_0}}{k!}$, □

Example 43.3 *Assume that each year there are $2,000,000$ Turkish people who receive scam emails promising them a huge amount of money. Assume also that each of these recipient of the scam emails, independently, becomes a victim with probability $1/1,000,000$.*

(i) Write down the exact probability that exactly 2 Turkish people will be victims by scam emails next year.

(ii) Compute the expected number of such scam cases in the next 2 years.

(iii) Write down a relevant approximate expression for the probability in part (i).

[40] According to TÜİK, there were 602,982 marriages and 131,830 divorces in 2015.

(iv) Write down an approximate expression for the probability that there will be no such scam cases during at least one of the next 2 years.

Solution : (i) The exact probability can be found by using the binomial pf as

$$\binom{2,000,000}{2}\left(\frac{1}{1,000,000}\right)^2\left(1-\frac{1}{1,000,000}\right)^{1,999,998}.$$

(ii) Expected number of such scam cases in the next 10 years is $(2,000,000)\frac{1}{1,000,000}2 = 4$.

(iii) We can use Poisson approximation with $\lambda = 2$. The probability in question will be approximately $2e^{-2}$.

(iv) This time $\lambda = 4$ since we have a period of 2 years[41]. So the required probability is

$$\begin{aligned} p &= \mathbb{P}(\text{no scams in either year 1 or year 2}) \\ &= \mathbb{P}(\text{no scams in year 1}) + \mathbb{P}(\text{no scams in year 2}) \\ &\quad -\mathbb{P}(\text{no scams in year 1 and in year 2}) \\ &\approx e^{-2} + e^{-2} - e^{-4} \\ &= 2e^{-2} - e^{-4}. \end{aligned}$$

□

Once we are given the limiting result in Theorem 43.1, a natural question is the quality of the approximation for a given $n \in \mathbb{N}$. Let's mention a classical result in this direction.

Theorem 43.2 *(A simple form of LeCam's inequality) Let $X_1, \dots, X_n$ be independent Bernoulli random variables with respective success probabilities $p_1, \dots, p_n$. Let N be a Poisson random variable with parameter $\lambda = \sum_{i=1}^n p_i$. Then, for any $A \subset \mathbb{R}$ we have*

$$\left|\mathbb{P}\left(\sum_{i=1}^n X_i \in A\right) - \mathbb{P}(N \in A)\right| \leq \sum_{i=1}^n p_i^2.$$

Indeed, one can generalize and improve the given inequality in various ways. For example, it is possible to derive Poisson approximation results for sums of dependent and non-identially distributed random variables. We do not go into details. Let's conclude the section with the following summary:

Rule of thumb: If n is large and p is small, then the distribution of a binomial random variable is approximately the same as the distribution of a Poisson with parameter np.

[41] What is the underlying assumption I am using here?

44 Distributional median

Definition 44.1 *For a given random variable X, a real number m is said to be a* ***median*** *(of the distribution) of X if both of the following conditions are satisfied:*

$$\mathbb{P}(X \leq m) \geq \frac{1}{2} \qquad \text{and} \qquad \mathbb{P}(X \geq m) \geq \frac{1}{2}.$$

Example 44.1 *(Median of a discrete distribution) Suppose X has distribution*

$$\mathbb{P}(X = 1) = 0.1, \quad \mathbb{P}(X = 2) = 0.2, \quad \mathbb{P}(X = 3) = 0.3, \quad \mathbb{P}(X = 4) = 0.4.$$

Then 3 is the unique median of X. □

Example 44.2 *(discrete case / median not unique) Suppose X has distribution*

$$\mathbb{P}(X = 1) = 0.1, \quad \mathbb{P}(X = 2) = 0.4, \quad \mathbb{P}(X = 3) = 0.3, \quad \mathbb{P}(X = 4) = 0.2.$$

Then any real number in $[2, 3]$ is a median of X. □

Theorem 44.1 *Let X be a random variable which takes values in an interval I of real numbers. Let r be a one-to-one function defined on the interval I. If m is a median of X, then $r(m)$ is a median of $r(X)$.*

Proof: Since r is one-to-one, it is either strictly increasing or strictly decreasing. Let's assume that the former is true, the other case being similar. Since r is one-to-one, it is invertible. We then have

$$\mathbb{P}(r(X) \geq r(m)) = \mathbb{P}((r^{-1} \circ r)(X) \geq (r^{-1} \circ r)(m)) = \mathbb{P}(X \geq m) \geq \frac{1}{2},$$

since r^{-1} is also a strictly increasing function when r is so. Similarly, $\mathbb{P}(r(X) \leq r(m)) \geq \frac{1}{2}$, and result follows. □

Remark 44.1 *The claim in Theorem 44.1 is not true if we replace the median by the mean. Here is a specific example for you to check. Let X be a Bernoulli random variable with parameter $p \in (0, 1)$ and $r(x) = x^3$. Show that $\mathbb{E}[r(X)] \neq r(\mathbb{E}[X])$.*[42]

Here are some remarks on the comparison of the mean and the median of a distribution.

[42] Exercise: For which X and for which r, we would have $\mathbb{E}[r(X)] \neq r(\mathbb{E}[X])$?

Median vs. mean

1. A finite median always exists, which is not true for the expectation.
2. We may have infinitely many medians of a distribution which is not the case for the mean.
3. Another property which is enjoyed by the median (and not by the mean) is: If r is a one-to-one function function and m is a median of X, then $r(m)$ is a median of $r(X)$.
4. If $Var(X) < \infty$, then we always have

   $$|Med(X) - \mathbb{E}[X]| \leq \sqrt{Var(X)}.$$

   So when the standard deviation of a random variable is small, expectation and median are close to each other. We will be proving this in next section.
5. When the underlying distribution is symmetric, mean and one of the medians will coincide.
6. When making predictions for a random variable, median minimizes the mean absolute error (which is to be defined in next section) whereas mean minimizes mean squared error.
7. The median and the mean may be considered as the center of the distribution according to different practice.
8. Sample mean and sample median are used to approximately find the distributional mean, and the distributional median, respectively.
9. The sample median is a more robust statistic compared to the sample mean.

Similar to the distributional median, we may also define distributional percentiles and distributional quartiles. This will be done in Section 46 where we will also discuss also discuss the estimators for these distributional properties. Before that we see a nice property enjoyed by the median.

45 Predictions II: The mean absolute error

Let X be a random variable and suppose that we would like to provide a prediction $d \in \mathbb{R}$ for X. Previously we have seen that the prediction that minimizes the mean squared error function $MSE(d) = \mathbb{E}[(X-d)^2]$ is $d^* = \mathbb{E}[X]$. Next we discuss minimizing another important error function.

Definition 45.1 *The* ***mean absolute error*** *of a prediction d for a random variable X is defined by*

$$MAE(d) = \mathbb{E}|X-d|.$$

Theorem 45.1 *If m is a median of a random variable X with support $\mathcal{S}$ and pmf f, then for any $d \in \mathbb{R}$,*

$$\mathbb{E}|X-m| \leq \mathbb{E}|X-d|.$$

Proof: Let's assume that $d > m$. The other case is similar - Still check it yourself for practice. We have

$$\begin{aligned}
\mathbb{E}[|X-d|-|X-m|] &= \sum_{x\in\mathcal{S}}(|x-d|-|x-m|)f(x) \\
&= \sum_{x\in\mathcal{S},\, x\leq m}(d-m)f(x) + \sum_{x\in\mathcal{S},\, m<x<d}(d+m-2x)f(x) \\
&\quad + \sum_{x\in\mathcal{S},\, x\geq d}(m-d)f(x) \\
&\geq \sum_{x\in\mathcal{S},\, x\leq m}(d-m)f(x) + \sum_{x\in\mathcal{S},\, m<x<d}(m-d)f(x) \\
&\quad + \sum_{x\in\mathcal{S},\, x\geq d}(m-d)f(x) \\
&= (d-m)(\mathbb{P}(X\leq m)-\mathbb{P}(X>m)) \\
&= (d-m)(\mathbb{P}(X\leq m)-1+\mathbb{P}(X\leq k)) \\
&= (d-m)\mathbb{P}(2\mathbb{P}(X\leq m)-1) \\
&\geq 0.
\end{aligned}$$

Here the first inequality follows since we have $d+m-2x \geq m-d$ when $m < x < d$. Therefore we obtain

$$\mathbb{E}|X-d| \geq \mathbb{E}|X-m|,$$

as required. □

So, the result to keep in mind is:

The prediction that **minimizes** the **mean absolute error** is a **median** of the underlying distribution.

We conclude this section with a related result.

Theorem 45.2 *If $Var(X) < \infty$, then we always have*

$$|Med(X) - \mathbb{E}[X]| \leq \sqrt{Var(X)}.$$

Proof: We have

$$\begin{aligned} |Med(X) - \mathbb{E}[X]| = |\mathbb{E}[Med(X) - X]| &\leq \mathbb{E}|Med(X) - X| \\ &\leq \mathbb{E}|X - \mathbb{E}[X]| \\ &\leq (\mathbb{E}|X - \mathbb{E}[X]|^2)^{1/2} \\ &= \sqrt{Var(X)}. \end{aligned}$$

□

46 Percentiles, and nearest rank method

We continue with a little bit more on learning descriptive statistics terminology. Our purpose is to extend the concept of the sample/distributional median to *percentiles*. We begin with data specific definitions - these are used to approximately find the corresponding distributional characteristics.

Definition 46.1 *Given N ordered values $x_1 < x_2 < \cdots < x_N$,* ***ordinal rank corresponding to p^{th} percentile****, $(p \in (0, 100))$ is*

$$n = \left\lceil \frac{p}{100} N \right\rceil,$$

where $\lceil \cdot \rceil$ is the ceiling function[43].

Example 46.1 *Assume that we have 5 samples. Then some examples of ordinal ranks are*

[43] i.e. $\lceil x \rceil$ the smallest integer at least as large as x. e.g. $\lceil 1.5 \rceil = 2$, $\lceil 1 \rceil = 1$, $\lceil -1.5 \rceil = -1$

Percentile	Ordinal rank
20^{th}	1
40^{th}	2
70^{th}	4

□

Definition 46.2 *For $p \in (0, 100)$, the* $\mathbf{p^{th}}$ ***(sample) percentile***[44] *is obtained by first computing the ordinal rank and then taking the value from the ordered list that corresponds to this rank.*

Example 46.2 *Consider $x_1 = 15$, $x_2 = 20$, $x_3 = 35$, $x_4 = 40$, $x_5 = 50$. Then, the ordinal rank corresponding to 20^{th} percentile is 1, and 20^{th} percentile is $x_1 = 15$. Similarly, 70^{th} percentile value is 40.* □

Example 46.3 *Assume that we are given the following data $3, 6, 7, 8, 8, 9, 10, 13, 15, 16, 20$ of size 11. Find the 25^{th} percentile value.*

Solution: Ordinal rank corresponding to 25^{th} percentile value is

$$\left\lceil \frac{25}{100} \times 11 \right\rceil = \lceil 2.75 \rceil = 3.$$

Then the 25^{th} percentile value is 7. □

Definition 46.3 *(i) 25^{th} and 75^{th} percentile values are called* ***sample first and third quartile values****.*

(ii) The difference between 75^{th} percentile and 25^{th} percentile is called the ***sample interquartile range****.*

(iii) The difference between the largest and the smallest sample values is called the ***sample range****.*

Note that some people call 50^{th} percentile as the sample median. We will avoid this to be consistent with our previous definition of the sample median, and will say 50^{th} percentile instead.

Next, let's briefly discuss the distributional percentiles and related quantities depending on our probabilistic model.

[44] You may see different definitions for the percentiles in other texts. We stick to this one in these notes.

Definition 46.4 *Let X be a random variable. The p**th (distributional) quantile** ($p \in [0,1]$) of the distribution of X is defined to be any number q_p such that*

$$\mathbb{P}(X \geq q_p) \geq 1 - p \quad \textit{and} \quad \mathbb{P}(X \leq q_p) \geq p.$$

Note that $q_{0.5}$ corresponds to our earlier definition for the distributional median. So, some of the properties stated for the median will also hold true percentiles (e.g. a distributional pth percentile always exists).

As in the case of sample percentiles, 25th and 75th distributional percentiles will be called the **first and the third distributional quartiles**, and the difference between them will be called the **distributional interquartile range**.

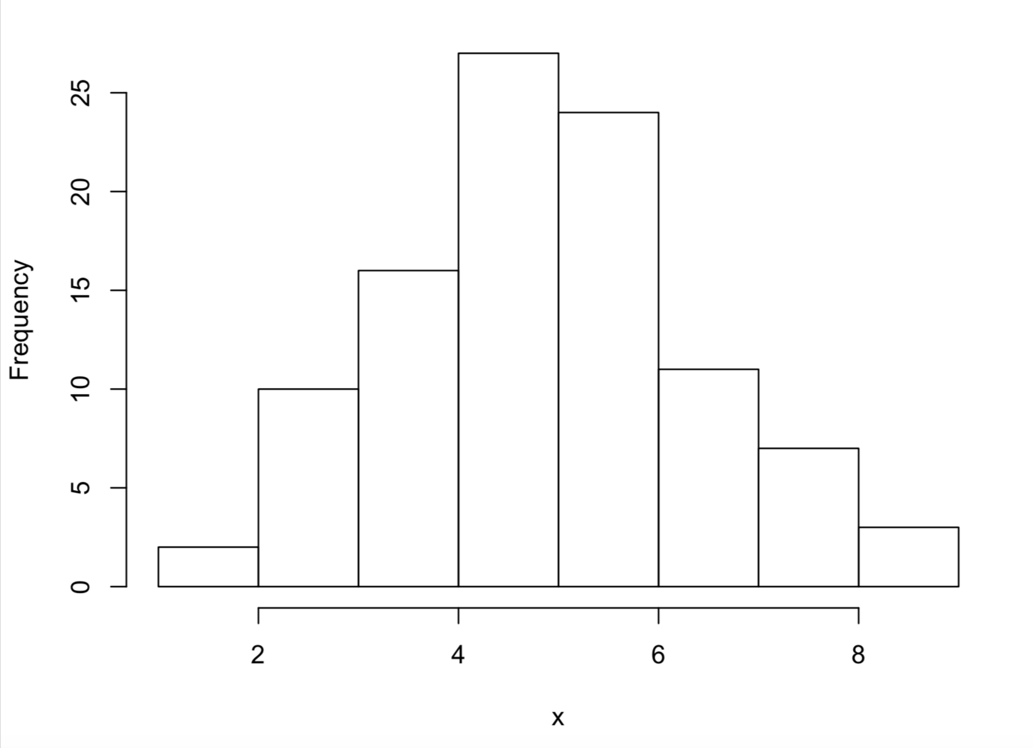
Example 46.4 *Let X be a uniform random variable over the set $\{1, 2, \ldots, 20\}$.*
(a) Find the first and third distributional quartiles of X.
(b) Find the distributional interquartile range of X.

Solution: (a) The first and the third distributional quartiles are given by $5, 15$, respectively.
(b) The distributional interquartile range $15 - 5 = 10$. □

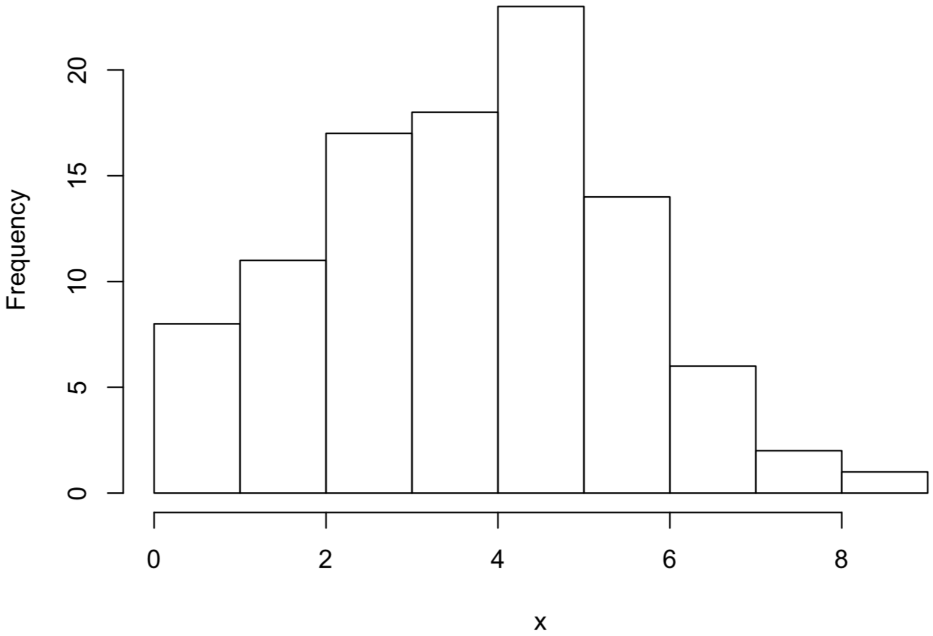
Once we have a random variable of interest, we estimate its distributional quartiles by using the sampling quartiles. One other use of quartiles will be seen in Section 69 where we will compare quartiles of two distributions in order to see whether they are indeed from the same distributions.

47 Model selection

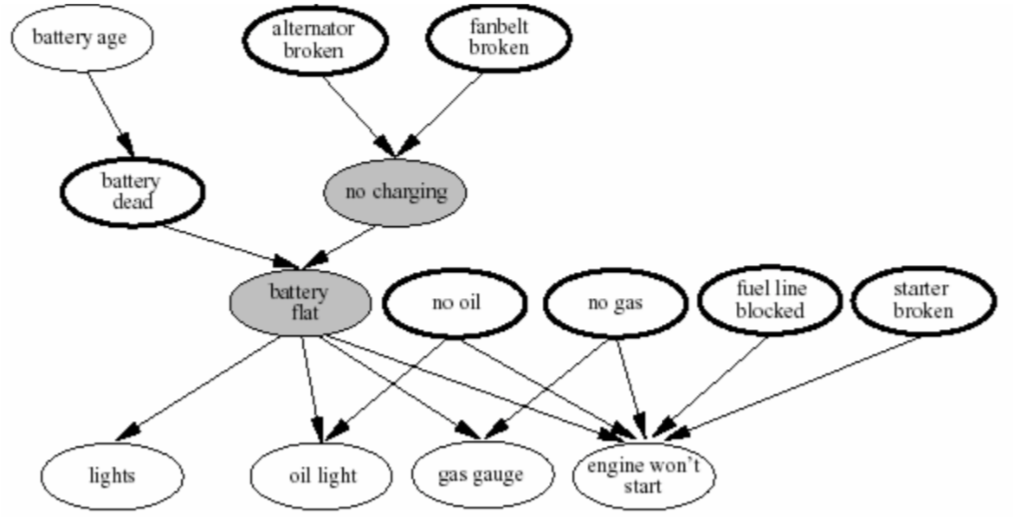
As mentioned earlier, these lecture notes is not intended for model selection. However, let's very briefly discuss that it is a fundamental part of statistics, and that it can get really complicated to choose a model in real life situations. So far you have learnt certain discrete distributions, binomial, Poisson, geometric, etc. Looking at the following data,



can you guess the distribution it comes from? I guess this should be fine with most of you. These were 100 samples from a binomial distribution with parameters 10 and 0.5. This was intuitively clear - but you can never be sure. How about the following one?



Now it got complicated. Is this one from binomial or Poisson or geometric (or some other)? Firstly, let me give you the answer: For this histogram I simulated 100 Poisson random variables with parameter 4. But by just looking at the data you have, you could very well propose a binomial distribution with certain parameters, or something like that. For that reason, while selecting your model you need to check various things in order to make sure that you are choosing right distributions. For this last particular example, for instance, we know for the Poisson distribution that the distributional mean and variance are the same. Then you could compute the sample mean and the sample variance, and check whether they are close to each other to support your proposal that the data comes from a Poisson distribution. It is not that simple of course, and model selection in general itself is a huge field. In real life problems, there are usually several variables, and you need to model how each of these affect each other probabilistically. This can get really tough. The following figure borrowed from Milos Hauskrecht's lecture notes represents a real life problem, diagnosing the car engine start problem, with a graphical model:



The relationship between all these variables is quite complex, and maybe there are even some missing arrows in this graphical model. In modeling, one main principle is Occam's razor which says that: "Choose the simplest possible model" - you will hear this often in detective series. But this brings a dilemma because real life problems are unfortunately not solvable if we keep the model "too simple".

In case you are interested in model selection issues, you will either need to visit the library for relevant books, or you will need to take courses in engineering, biostatistics, etc., where you really model a real life problem.

48 Review problems IV

Exercise 48.1 *Given a random sample $X_1, \ldots, X_n$ from a population with pdf $f(x \mid \theta)$, show that maximizing the likelihood function is equivalent to maximizing the log-likelihood function.*

Exercise 48.2 *Let $X_1, \ldots, X_n$ be i.i.d. with one of two pdfs. If $\theta = 0$, then*

$$f(x \mid \theta) = \begin{cases} 1, & \textit{if } 0 < x < 1 \\ 0, & \textit{otherwise,} \end{cases}$$

while if $\theta = 1$, then

$$f(x \mid \theta) = \begin{cases} 1/(2\sqrt{x}), & \textit{if } 0 < x < 1 \\ 0, & \textit{otherwise.} \end{cases}$$

Find the MLE of θ.

Exercise 48.3 *Let $f(x|\theta)$, $0 \leq \theta \leq 1$ be a pmf given by*

$$f(x|\theta) = \begin{cases} \frac{2\theta}{3}, & \text{if } x = 0 \\ \frac{\theta}{3}, & \text{if } x = 1 \\ \frac{2(1-\theta)}{3}, & \text{if } x = 2 \\ \frac{(1-\theta)}{3}, & \text{if } x = 3. \end{cases}$$

Assume that the 10 samples $(3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$ were taken from the distribution corresponding to f. Find the maximum likelihood estimate of θ. (Answer: $\theta = 1/2$.)

Exercise 48.4 *Let $X_1, \ldots, X_n$ be independent Poisson random variables with parameter λ.*

i. Find the conditional distribution of $X_1 + \cdots + X_n$ given that $X_1 = k$, $k \geq 0$.

ii. Find the conditional distribution of X_1 given that $X_1 + \cdots + X_n = k$, $k \geq 0$.

Exercise 48.5 *If $X \sim PO(\theta)$ and $Y \sim PO(\lambda)$, show that the distribution of $X|X + Y$ is binomial with success probability $\theta/(\theta + \lambda)$.*

Exercise 48.6 *An airline sells 200 tickets for a certain flight on an airplane that has only 198 seats because, on the average, 1 percent of purchasers of airline tickets do not appear for the departure of their flight. (a) Find an exact expression for the probability that every one who appears for the departure of this flight will have a seat. (b) Give an approximation for the probability you found in part (a) by using Poisson distribution*

Exercise 48.7 *Let X be a binomial random variable with parameters $2m$ and $1/2$ where m is some fixed positive integer.*

i. Explain how we can express X in terms of Bernoulli random variables.

ii. What prediction of X has the smallest MSE (mean squared error)? Explain.

iii. Show that the probability function of X is symmetric around the point $x = m$.

iv. What prediction of X has the smallest MAE (mean absolute error)? Explain.

Exercise 48.8 *In a given city it is assumed that the number of automobile accidents in a given year follows a Poisson distribution. In past years the average number of accidents per year was 15, and this year it was 10. Is it justified to claim that the accident rate has dropped?*

Exercise 48.9 *Let X be an observation from the pdf*

$$f(x \mid \theta) = \left(\frac{\theta}{2}\right)^{|x|} (1-\theta)^{1-|x|}, \quad x = -1, 0, 1; \quad 0 \leq \theta \leq 1.$$

a. Find the MLE of θ.
b. Define the estimator $T(X)$ by

$$T(X) = \begin{cases} 2, & \text{if } x = 1 \\ 0, & \text{otherwise.} \end{cases}$$

Show that $T(X)$ is an unbiased estimator of θ.
c. Find a better estimator than $T(X)$ and prove that it is better.

Exercise 48.10 *The hypergeometric distribution can be approximated by using the binomial or the Poisson distribution. Let X have the hypergeometric distribution*

$$\mathbb{P}(X = x \mid N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, x = 0, 1, \ldots, K.$$

i. Show that as $N \to \infty, M \to \infty$, and $M/N \to p$,

$$\mathbb{P}(X = x \mid N, M, K) \longrightarrow \binom{K}{x} p^x (1-p)^{K-x}, \quad x = 0, 1, \ldots, K.$$

(Hint: Stirling's formula will be useful.)

ii. Use the fact that the binomial can be approximated by the Poisson to show that if $N \to \infty$, $M \to \infty$, $K \to \infty$, $M/N \to 0$, and $KM/N \to \lambda$, then

$$\mathbb{P}(X = x \mid N, M, K) \longrightarrow \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots.$$

49 Continuous distributions

Definition 49.1 *We say that a random variable X has a **continuous distribution** (or that X is a **continuous random variable**) if there exists a non-negative function f, defined on the real line, such that for every interval of real numbers (bounded or unbounded), the probability that X takes a value in that interval is the integral of f over that interval.*

For example,

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx,$$

$$\mathbb{P}(X \geq a) = \int_a^\infty f(x)dx,$$

etc..

Definition 49.2 *If X has a continuous distribution, the function f described in Definition 49.1 is called the **probability density function** (abbreviated **pdf**) of X. The closure of the set*[45] *$\{x : f(x) > 0\}$ is called the **support** of (the distribution of) X.*

Analogous to the discrete case, every pdf must satisfy the following two requirements:

$$f(x) \geq 0, \quad \text{for all } x,$$

and

$$\int_{-\infty}^\infty f(x) = 1.$$

Also, let's note that the cumulative distribution function in continuous case is defined exactly as in the discrete case.

Here are some examples of continuous random variables:

Example 49.1 *i. **Uniform distribution**. A random variable with pdf $f(x) = 1/(b-a), a < b$ is called a uniform random variable on (a, b). Important special case : $a = 0, b = 1$.*

*ii. **Exponential distribution**. A random variable with pdf $f(x) = e^{-x}$, $x \geq 0$ is called an exponential random variable.*

*iii. **Cauchy distribution**. A random variable with pdf $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$, $x \in \mathbb{R}$ is called a Cauchy random variable.*

*iv. **Gumbel distribution**. A random variable with pdf $f(x) = e^{x-e^{-x}}$, $x \in \mathbb{R}$ is called a Gumbel random variable.* □

Example 49.2 *Let's exemplify the cumulative distribution functions in continuous setting by using the uniform distribution. Let X be a random variable uniformly distributed over*

[45] Closure of a set A is the smallest closed set containing A.

(a, b). Denoting the corresponding cdf by F, clearly, $F(x) = 0$ when $x \leq a$ and $F(x) = 1$ when $x \geq b$. If $a < x < b$, then

$$F(x) = \mathbb{P}(X \leq x) = \int_a^x f(x)dx = \int_a^x \frac{1}{b-a}dx = \frac{x-a}{b-a}.$$

Therefore, the cdf of X is

$$F(x) = \begin{cases} 0, & \text{if } x \leq a, \\ \frac{x-a}{b-a}, & \text{if } a < x < b, \\ 1 & \text{if } x \geq b. \end{cases}$$

□

Exercise 49.1 *Find the cumulative distribution functions of the random variables given in (ii) and (iii) in Example 49.1.*

Example 49.3 *Let n be a fixed non-negative integer. For some constant c, the random variable X has the pdf*

$$f(x) = \begin{cases} cx^n, & \text{if } 0 < x < 1 \\ 0, & \text{otherwise.} \end{cases}$$

(i) Find the value of c.

(ii) Find $\mathbb{P}(X \leq y)$ for $y \in \mathbb{R}$.

(iii) Find $\mathbb{P}(X > 1/2)$.

Solution: (i) We should have

$$1 = \int_0^1 cx^n dx = c\frac{1}{n+1}.$$

So $c = n + 1$.

(ii) Let $F(y) = \mathbb{P}(X \leq y)$. Clearly, $F(y) = 1$ for $y \geq 1$ and $F(y) = 0$ for $y \leq 0$. When $y \in (0, 1)$, we have

$$F(y) = \int_0^y f(x)dx = \int_0^y (n+1)x^n dx = y^{n+1}.$$

(iii) Using part (ii)

$$\mathbb{P}(X > 1/2) = 1 - F(1/2) = 1 - \frac{1}{2^{n+1}}.$$

Alternatively, you could just integrate the pdf on the proper region. □

Exercise 49.2 *Note that when $n = 0$ in previous example, we recover the uniform distribution on $(0,1)$. Of course, it is not possible to simulate a uniform random number on $(0,1)$ with a computer (Why?). What would your approach be to generate an approximately uniform random number on $(0,1)$ by using the discrete uniform distribution? What would be the problems you may face in this case?*

Some important remarks on continuous distributions:

Remark 49.1

1. ***(Points have zero mass)*** *Continuous distributions assign probability 0 to individual values. That is, if $\mathbb{P}$ is the probability measure corresponding to a continuous random variable with pdf f, then*

$$\mathbb{P}(X = x) = 0, \quad \textit{for any } x \in \mathbb{R}.$$

A heuristic argument showing that this is the case can be given by noting that for arbitrarily small $\epsilon > 0$,

$$\mathbb{P}(x - \epsilon \leq X \leq x + \epsilon) \approx 2\epsilon f(x) \approx 0.$$

2. ***(Non-uniqueness of the pdf)*** *Since $\mathbb{P}(X = x) = 0$ for every individual value x, each pdf can be changed at a finite number of points, or even at certain infinite sequences of points, without changing the value of the integral of the pdf over any subset A.*

3. ***(Density is not a probability)*** *A density function f itself is not a probability measure. Indeed f can even be unbounded. Next exercise provides such an example.*

Exercise 49.3 *Consider the function*

$$f(x) = \begin{cases} (2/3)x^{-1/3}, & \textit{if } 0 < x < 1 \\ 0, & \textit{otherwise.} \end{cases}$$

Show that f defines a pdf, but that it is unbounded. Conclude that a pdf itself is not a probability.

We now know that once we are given a continuous random variable with pdf f, we can find the corresponding cdf by integration. The next result explains the opposite direction. For the proof just recall the fundamental theorem of calculus.

Theorem 49.1 *Let X be a continuous random variable with pdf f and cdf F, that is*

$$F(x) = \int_{-\infty}^{x} f(y)dy.$$

Then

$$F'(x) = f(x), \quad x \in \mathbb{R}.$$

50 Expectations in continuous setting

Definition 50.1 *Let X be a continuous random variable whose pdf is f. Suppose at least one of the following integrals is finite:*

$$\int_0^\infty xf(x)dx, \qquad \int_{-\infty}^0 xf(x)dx. \tag{11}$$

*Then the **expectation** of X is said to exist, and is defined by*

$$\mathbb{E}[X] = \int_{-\infty}^\infty xf(x)dx.$$

*If both integrals in (11) are infinite, then we say that the **expectation of X does not exist.***

Example 50.1 *Let X be a uniform random variable over (a,b) so that its pdf is given by $f(x) = \frac{1}{b-a}$, $a < x < b$. Then*

$$\mathbb{E}[X] = \int_a^b x\frac{1}{b-a}dx = \frac{1}{b-a}\left(\frac{b^2-a^2}{2}\right) = \frac{a+b}{2}.$$

The result should be intuitively clear since we have found just the midpoint of a and b. As in discrete case you may try to relate this to the center of mass idea in physics. □

Example 50.2 *(1) Let X be a continuous random variable with pdf*

$$f(x) = \begin{cases} \frac{1}{x^2}, & \text{if } x \geq 1, \\ 0, & \text{otherwise} \end{cases}$$

Then $\mathbb{E}[X] = \int_1^\infty x\frac{1}{x^2}dx = \infty$

(2) Let X be a continuous random variable with pdf

$$f(x) = \begin{cases} \frac{1}{2x^2}, & \text{if } x \geq 1 \text{ or } x \leq -1, \\ 0, & \text{otherwise.} \end{cases}$$

Then the expectation of X does not exist since both expectations

$$\int_0^\infty xf(x)dx, \qquad \int_{-\infty}^0 xf(x)dx$$

are infinite. □

Remark 50.1 *As in case of discrete setting, if $Y = r(X)$ where X is a continuous random variable with pdf f, then the expectation of Y is given by*

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} r(x) f(x) dx.$$

Example 50.3 *Let X be a random variable whose pdf is given by $f(x) = e^{-x}$, $x \geq 0$. Find $\mathbb{E}[e^{tX}]$ where $t < 1$ is some given real number*[46]*.*

Solution: We have

$$\mathbb{E}[e^{tX}] = \int_0^{\infty} e^{tx} e^{-x} dx = \int_0^{\infty} e^{x(t-1)} dx = \frac{1}{t-1} e^{x(t-1)} \Big|_0^{\infty} = \frac{1}{1-t}.$$

□

51 Uniform distribution over an interval

We start by recalling uniform distribution over a given interval.

Definition 51.1 *A random variable X is said to be* ***uniformly distributed*** *over the interval (a, b)*[47]*, with $a < b$ if its pdf is given by*

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

When X is uniform over (a, b) we write $X \sim U(a, b)$.

Theorem 51.1 *Let $X \sim U(a, b)$. Then*

i $\mathbb{E}[X] = \frac{a+b}{2}$.

ii $Var(X) = \frac{(b-a)^2}{12}$.

Exercise 51.1 *Prove Theorem 51.1.*

[46] In general, for a given random variable X, the function $\phi(t) = \mathbb{E}[e^{tX}]$ is called the moment generating function. Although this is a topic of extreme importance in probability theory and statistics, we won't be able to go into details in these lecture notes.

[47] or, $[a, b]$ or $(a, b]$ or $[a, b)$

Uniform distribution is of extreme importance for various reasons. One instance that is to be discussed below is its use in random number generation. Once you can generate uniformly distributed random variables, you may reach at various other distributions after performing proper transformations on it.

Exercise 51.1 *Explain how we may define uniformly distributed random variables over unions of intervals.*

Exercise 51.2 *Let $a \in \mathbb{R}$. Can we have uniform distribution over the interval (a, ∞)? Why? Why not?*

52 * Pseudo-random number generators

Previously we mentioned that it is impossible to simulate a random variable having a continuous distribution due to finite memory. In this section we briefly discuss what is done in practice in such situations by focusing on pseudorandom number generators.

In general, pseudorandom number generators (PRNG) are algorithms for generating sequences of numbers that approximate the properties of random numbers. The sequence is not truly random in that it is completely determined by a relatively small set of initial values. To produce such random numbers, one uses some iterative scheme

$$x_{n+1} := f(x_n).$$

For instance, f can be chosen to be the logistic map,

$$f(x) = rx(1-x), \qquad r > 0$$

which is the archetypal example of chaos in 1 dimension. In this case we have the difference equation

$$x_{n+1} = rx_n(1-x_n)$$

and for some $x_0 \in (0,1)$, we may produce a "random sequence" in $(0,r)$.

As an alternative that is often used in practice, one may use the difference equation

$$x_{n+1} = (bx_n + r_{n+1}) \pmod{m}.$$

In this case the resulting random numbers over $(0,1)$ will be $u_n = x_n/m$. Here b is a given constant and in a conventional linear congruent PRNG, the *perturbation* r_{n+1} is also a constant.

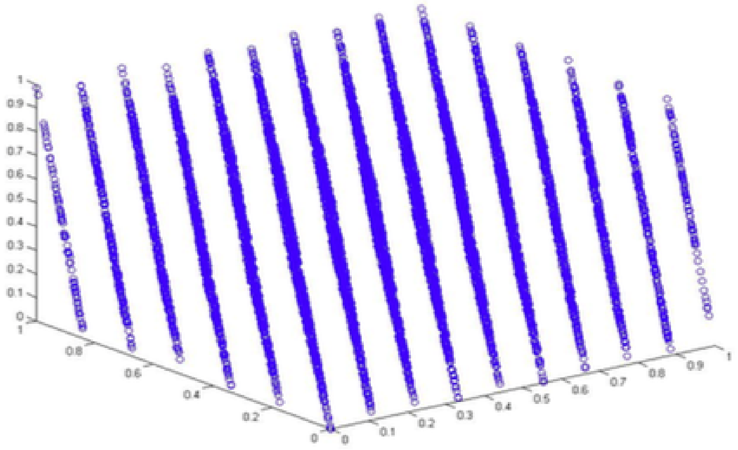
Let's here focus on a particular random number generator, RANDU, which was used in IBM machines in 1960's and let's show how erroneous it was. For this specific algorithm, the iterations are given by

$$x_{n+1} = 65539x_n \pmod{2^{31}} = (2^{16}+3)x_n \pmod{2^{31}}.$$

RANDU was used to generate random numbers in 3 dimensional unit cube as follows: Firstly generate $u_i \sim U(0,1)$ and then consider the vectors $\pi_1 = (u_1, u_2, u_3)$, $\pi_2 = (u_2, u_3, u_4)$ and so on. By uniformity of u_i's, π_i's were also considered to be uniform in the unit cube of $\mathbb{R}^3$. But the recursive definition of x_i's actually cause an enormous correlation within the three successive generated random numbers. In fact, we have

$$\begin{aligned} X_{n+2} = (2^{16}+3)^2 X_n \pmod{2^{31}} &= (2^{32} + 6 \cdot 2^{16} + 9) X_n \pmod{2^{31}} \\ &= (6 \cdot 2^{16} + 9) X_n \pmod{2^{31}} \\ &= (6(2^{16}+3) - 9) X_n \pmod{2^{31}} \\ &= (6(2^{16}+3) X_n - 9 X_n) \pmod{2^{31}} \\ &= (6 X_{n+1} - 9 X_n) \pmod{2^{31}}, \end{aligned}$$

causing the resulting π_i's to be concentrated on only 15 hyperplanes!



To see a discussion on more general cases of affine pseudo random number generators, we refer to [9].

53 A few notes on continuous distributions

The purpose of this section is to emphasize that various definitions/results we discussed for discrete distributions remain the same in continuous setting.

a. Expectation is linear: If $X_1, \ldots, X_n$ are continuous/discrete random variables, and if $c_1, \ldots, c_n$ are real numbers, then

$$\mathbb{E}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i \mathbb{E}\left[X_i\right].$$

The discussions for the variance, covariance and moments remain valid for continuous distributions. In particular, Cauchy-Schwarz inequality still holds true in non-negative setting.

b. All the variance and covariance formulas we obtained still hold. For example, when $X_1, \ldots, X_n$ are random variables with finite variances we still have, then

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j).$$

Or, when $X_1, \ldots, X_n$ are independent,

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i).$$

c. Markov inequality holds for a continuous random variable that is non-negative. That is, if X is a continuous non-negative random variable, then for any $t > 0$, we have

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Similarly, Chebyshev inequality, weak law of large numbers, etc. still stand for continuous random variables. It is an exercise foor you too adapt the proofs given in discrete setting to the continuous case. In general, all you need to will be replacing summations by integrals.

c. The definitions for (sample and distributional) median, quartiles, percentiles, etc. do not change. Let me show you two continuous examples for the median.

Example 53.1 *(i) (Median of a continuous distribution) Let X be a continuous random variable with pdf $f(x) = 4x^3, 0 < x < 1$. Then if m is a median we should have*

$$\frac{1}{2} = \int_0^m f(x)dx = \int_0^m 4x^3 dx = m^4$$

implying that the unique median of X is $1/(2^{1/4})$.

(ii) (Median of a continuous distribution / non-unique case) Suppose X has a continuous distribution with pdf

$$f(x) = \begin{cases} 1/2, & \textit{if } 0 \leq x \leq 1 \\ 1, & \textit{if } 2.5 \leq x \leq 3 \\ 0, & \textit{otherwise.} \end{cases}$$

Then it is easily seen that any real number in $[1, 2.5]$ is a median of X. □

d. The definition for independent random variables is the same in continuous setup. Two continuous random variables X and Y are *independent* if for any $A, B \subset \mathbb{R}$, we have

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

Below in Section 58 we will have a factorization theorem as in discrete setting which makes it easy to check that random variables are independent.

As before independent random variables will satisfy

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

The case for more than two random variables will follow the same lines as in our discussion on discrete random variables.

e. All other issues remain the same, so I do not go into more details here.

Let's give a brief discussion of above these items on a toy example.

Example 53.2 *Let $X_1, \ldots, X_n$ be independent uniformly distributed random variables on $[-1, 1]$. Then by Theorem 51.1 we know that $\mathbb{E}[X_1] = 0$ and $Var(X_1) = \frac{1}{3}$. By symmetry the median turns out to be 0. If we define $S_n = \sum_{i=1}^n X_i$, then*

$$\mathbb{E}[S_n] = \sum_{i=1}^n \mathbb{E}[X_i] = 0,$$

and since X_i's are also independent

$$Var(S_n) = Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i) = \sum_{i=1}^n \frac{1}{3} = \frac{n}{3}.$$

Note that we can not use Markov's inequality for S_n since it can be negative. But Chebyshev's inequality tells us that for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{|S_n| - \mathbb{E}[S_n]}{n} > \epsilon\right) = \mathbb{P}\left(\frac{|S_n|}{n} > \epsilon\right) \leq \frac{Var(S_n)}{n^2\epsilon^2} = \frac{1}{3n\epsilon^2}.$$

In particular, again for any $\epsilon > 0$

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| > \epsilon\right) \to 0, \qquad n \to \infty$$

for any given $\epsilon > 0$. That is,

$$S_n/n \to_{\mathbb{P}} 0$$

as $n \to \infty$. □

Now, let's continue with some other important special continuous distributions.

54 Exponential distribution

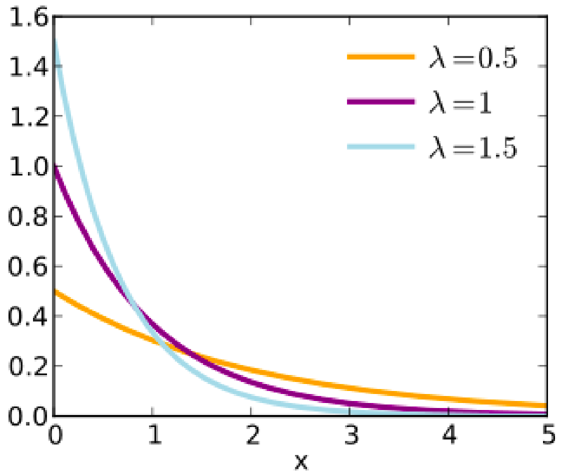
Exponential distribution is in general used to model waiting time for the occurrence of a certain event. This can be the service time at a bank cashier, or the distance required to see a deer when you travel on some country road. Of course there are certain assumptions that should be satisfied in order for a random quantity to obey the exponential distribution, but we skip these here. In case you are interested, please check *Poisson processes* online.

Definition 54.1 *Let $\lambda > 0$. A random variable X is said to have the* ***exponential distribution*** *with parameter λ if its pdf is given by*

$$f(x \mid \lambda) = \lambda e^{-\lambda x}, \qquad x > 0.$$

When X is exponential with parameter λ, we write $X \sim exp(\lambda)$.

The next figure borrowed from Wikipedia shows the exponential pdf for distinct values of λ.



Theorem 54.1 *If $X \sim exp(\lambda)$, then*

i. *The cdf of X is $F(x) = 1 - e^{-\lambda x}$ for $x > 0$ and $F(x) = 0$ for $x \leq 0$.*

ii. $\mathbb{E}[X] = \frac{1}{\lambda}$.

iii. $Var(X) = \frac{1}{\lambda^2}$.

Proof: (i) Observe that for $x > 0$,

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{0}^{x} \lambda e^{-\lambda y} dy = \lambda \left(-\frac{1}{\lambda}\right) e^{-\lambda y}\Big|_0^x = 1 - e^{-\lambda x}.$$

(ii) Using integration by parts with $u = x$ and $dv = e^{-\lambda x}dx$, we have

$$\int_0^\infty x\lambda e^{-\lambda x}dx = \lambda \left(x\frac{-1}{\lambda}e^{-\lambda x}\Big|_0^\infty - \int_0^\infty \frac{-1}{\lambda}e^{-\lambda x}dx \right) = \int_0^\infty e^{-\lambda x}dx = \frac{1}{\lambda}.$$

(iii) This part is similar to part (ii), and is left for you. You will just need to use integration by parts twice. □

Exponential distribution is the only continuous distribution with the memoryless property - recall that its counterpart in discrete case was the geometric distribution:

Theorem 54.2 ***(Memoryless property)*** *Let X be an exponential random variable with parameter λ and $t > 0$. Then for any $h > 0$,*

$$\mathbb{P}(X \geq t + h \mid X \geq t) = \mathbb{P}(X \geq h).$$

Proof: We know that the cdf of X is given by $F(x) = 1 - e^{-\lambda x}$, $x > 0$. So $\mathbb{P}(X \geq x) = 1 - F(x) = e^{-\lambda x}$, $x > 0$. Then

$$\mathbb{P}(X \geq t+h \mid X \geq t) = \frac{\mathbb{P}(X \geq t+h, X \geq t)}{\mathbb{P}(X \geq t)} = \frac{\mathbb{P}(X \geq t+h)}{\mathbb{P}(X \geq t)} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} = \mathbb{P}(X \geq h).$$

□

In words, if a probability distribution has the memoryless property the likelihood of something happening in the future has no relation to whether or not it has happened in the past. The history of the function is irrelevant to the future.

Showing that the memoryless property characterizes the exponential distribution is more challenging and it is omitted. Let's just see an example.

Example 54.1 *Suppose that duration of a couple's relationship is exponential with a mean of* 2 *weeks. Given that it lasted for 10 weeks so far, find the conditional probability that it will last at least one more week.*

Solution: Let T be the duration of the relationship. Since the mean is given to be 2, we see that T is an exponential random variable with parameter $1/2$. Then, using the memoryless property, and the cdf of exponential distribution we obtain

$$\mathbb{P}(T \geq 11 \mid T \geq 10) = \mathbb{P}(T \geq 10+1 \mid T \geq 10) = \mathbb{P}(T \geq 1) = 1 - \mathbb{P}(T < 1) = e^{-1/2}.$$

□

55 * Inversion method

Suppose that I would like to simulate from the distribution of a random variable X with cdf F. If F is a explicitly known, then this task can be done (at least theoretically) easily.

Theorem 55.1 *(Inversion method) Let F be the cdf of some random variable. If $U \sim U(0,1)$, then $F^{-}(U) \sim F$, where $F^{-}(u) = \inf\{x : F(x) \geq u\}$ is the generalized inverse of F. That is, $F^{-}(U)$ has the distribution whose cdf is given by F.*

Proof: Assume F is continuous and strictly increasing, and note that in this case $F^{-} = F^{-1}$, where F^{-1} is the standard inverse function of F. Then for $x \in \mathbb{R}$, we have

$$\mathbb{P}(F^{-}(U) \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(F(F^{-1}(U)) \leq F(x)) = \mathbb{P}(U \leq F(x)) = F(x),$$

concluding the proof when F is strictly increasing. The more general case requires playing with infimums[48], and I am leaving it as an exercise for those of you who are interested in. □

Example 55.1 *Explain how you can get samples from the exponential distribution with parameter $\lambda > 0$ by using the inversion method. (You may assume that you are able to generate uniformly distributed random variables over $(0,1)$[49].)*

Solution: Recall from previous section that the cdf of an exponential random variable X with parameter λ is given by

$$F(x) = \int_0^x \lambda e^{-\lambda y} dy = 1 - e^{-\lambda x}, \qquad x \geq 0.$$

Noting that F is strictly increasing on $[0, \infty)$, we may invert this function easily to get

$$F^{-1}(x) = -\frac{1}{\lambda} \ln(1-x).$$

Therefore using the inversion method the random variable

$$-\frac{1}{\lambda} \ln(1-U)$$

has the distribution of X when U is uniformly distributed over $(0,1)$. Now, in order to get i.i.d. samples from the distribution of X, we sample i.i.d. copies of U, say, $U_1, \ldots, U_n$, and then look at $-\frac{1}{\lambda} \ln(1-U_i)$, $i = 1, \ldots, n$[50]. □

[48] By the way, if you don't what infimum means just consider it as the minimum.

[49] In random number generation, uniform distribution over $(0,1)$ can be considered as the basis block. In general, we assume that we can generate such random numbers (via a pseudo-random number generator, or some other method) and then use them to generate variables with other distributions.

[50] Actually, $1 - U_i$ has the same distribution as U_i. So $-\frac{1}{\lambda} \ln U_i \sim exp(\lambda)$ as well

Exercise 55.1 *Explain how you can obtain samples from the Weibull distribution by using the inversion method, where the pdf of Weibull distribution is given by* $f(x \mid k, \lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$, $x \geq 0$, *and* $f(x) = 0$, *otherwise.*

56 Normal distribution

Normal distribution was initially used for observational measurement errors in $18^{th}/19^{th}$ century. It was later used to model physical attributes such as the height of a person (1835) and more complicated stuff such as the IQ of a person more recently. Thanks to the central limit theorem we will see in Section 65, the normal distribution turns out to be extremely important. This is perhaps reflected in 10 Deutsche Mark where they commemorate Gauss with the pdf of the normal distribution among all the work done by Gauss.

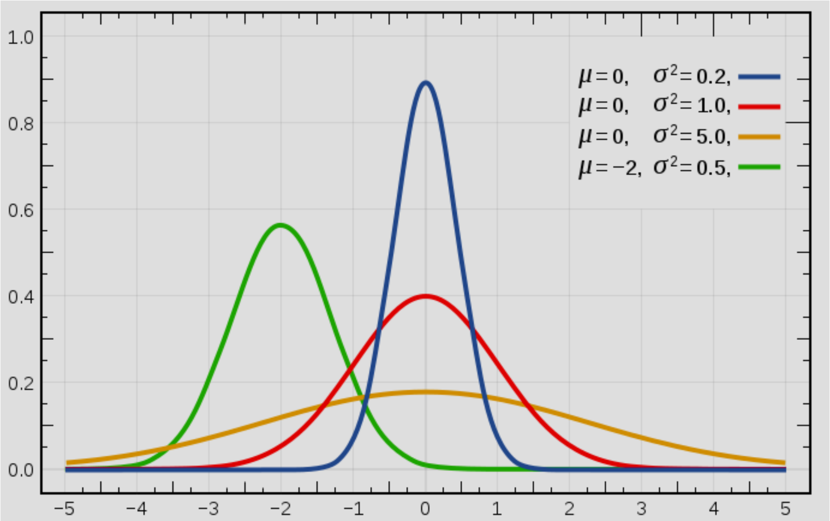
Definition 56.1 *A random variable* X *has the* ***normal distribution***[51] *with mean* $\mu \in (-\infty, \infty)$ *and variance* $\sigma^2 \in (0, \infty)$ *if* X *has a continuous distribution with pdf*

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \qquad -\infty < x < \infty. \tag{12}$$

We denote the distribution of such a random variable by $N(\mu, \sigma^2)$ *and write* $X \sim N(\mu, \sigma^2)$.

Next figure taken from Wikipedia demonstrates the graphs of the normal pdf for varying μ and σ^2 - note that they are all bell-shaped:

[51] Normal distribution is also known as **Gaussian distribution**.



Remark 56.1 *How do we check that (12) really defines a pdf? Here is a discussion for the case $\mu = 0$ and $\sigma = 1$. Clearly, f is non-negative and so we just need to show $\int_{-\infty}^{\infty} f(x)dx = 1$. To see that this is the case, let $I = \int_{-\infty}^{\infty} f(x)dx$. Then we have*

$$I^2 = \int_{-\infty}^{\infty} e^{-x^2/2}dx \int_{-\infty}^{\infty} e^{-y^2/2}dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}dxdy.$$

Changing to polar coordinates by letting $x = r\cos\theta$ and $y = r\sin\theta$, we see that

$$I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2} r dr d\theta.$$

Using u-substitution in last integral with $u = r^2/2$ gives $I^2 = 2\pi$, proving that f really defines a pdf. Check the details yourself.

$\mu = 0$, $\sigma^2 = 1$ is a special case worth a definition.

Definition 56.2 *X is said to be a **standard normal random variable**, denoted $X \sim N(0,1)$, if its pdf is given by*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad -\infty < x < \infty.$$

We denote the pdf and cdf of a standard normal by ϕ and Φ, respectively.

Remark 56.2 *For $x \in \mathbb{R}$*

$$\Phi(x) = \int_{-\infty}^{x} \phi(y)dy = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

It is a well-known fact that the right-most integral expression does not have a closed form[52]*. We read probabilities of the form $\Phi(x)$ from the z-table which can be found at the end of these lecture notes.*

For example, when Z is a standard normal random variable, we have

$$\mathbb{P}(>\leq 2.04) = 0.9793$$

and

$$\mathbb{P}(Z > 1) = 1 - \mathbb{P}(Z \leq 1) = 1 - 0.8413 = 0.1587.$$

Remark 56.3 *Due to symmetry reasons, for any $x \in \mathbb{R}$, we have*

$$\Phi(-x) = 1 - \Phi(x).$$

Exercise 56.1 *Let $Z \sim N(0,1)$. Read the following probabilities from the z-table:*

a. $\mathbb{P}(Z < 1.96)$*;*

b. $\mathbb{P}(-1.64 < Z < 1.64)$*;*

c. $\mathbb{P}(Z > -2.32)$*.*

The proof of the following result is left for you.

Theorem 56.1 *(Linear Tranformations) If $X \sim N(\mu, \sigma^2)$, and $Y = aX + b$, $a \neq 0$, then $Y \sim N(a\mu + b, a^2\sigma^2)$.*

The next standardization trick will be very useful.

Theorem 56.2 *Let $X \sim N(\mu, \sigma^2)$. Then*

$$\frac{X - \mu}{\sigma} \sim N(0,1).$$

The proof follows immediately from Theorem 56.1 with $a = 1/\sigma$ and $b = -\mu/\sigma$. Via this relation we can read probabilities related to non-standard normal random variables using the z-table. Here are two examples.

[52] A function is said not to have a closed form if it can not be expressed in terms of elementary functions, polynomials, exp, sin, etc..

Example 56.1 *Letting $X \sim N(5,4)$, find $\mathbb{P}(1 < X < 8)$.*

Solution: Denoting a standard normal random variable by Z and using Theorem 56.1, we have

$$\begin{aligned}\mathbb{P}(1 < X < 8) = \mathbb{P}\left(\frac{1-5}{2} < \frac{X-5}{2} < \frac{8-5}{2}\right) = \mathbb{P}(-2 < Z < 1.5) &= \Phi(1.5) - \Phi(-2)\\ &= \Phi(1.5) - (1-\Phi(2))\\ &= 0.9105.\end{aligned}$$

□

Example 56.2 *Assume that the IQ score of a current population has the normal distribution with parameters $\mu = 100$ and $\sigma^2 = 225$. Find the probability that a randomly chosen individual will have an IQ (i) more than 125; (ii) between 110 and 90.*

Solution: Let X be the IQ score of our randomly selected individual. Letting $Z \sim N(0,1)$ and Φ be the cdf of Z, we have

$$\mathbb{P}(X > 125) = \mathbb{P}\left(\frac{X-100}{15} > \frac{125-100}{15}\right) = \mathbb{P}(Z > 1.66) = 1-\Phi(1.66) = 1-0.9515 = 0.0485,$$

and

$$\begin{aligned}\mathbb{P}(110 > X > 90) \approx \mathbb{P}(0.66 > Z > -0.66) = \Phi(0.66) - \Phi(-0.66) &= \Phi(0.66) - (1-\Phi(0.66))\\ &= 2\Phi(0.66) - 1\\ &= 2(0.754) - 1 = 0.508.\end{aligned}$$

□

Next, we will look at what the parameters μ and σ^2 represent for the normal distribution.

Theorem 56.3 *Let $X \sim N(\mu, \sigma^2)$. Then:*
(i) $\mathbb{E}[X] = \mu$;
(ii) $Med(X) = \mu$;
(iii) $Var(X) = \sigma^2$.

Exercise 56.2 *Prove Theorem 56.3. (Hint: (i)-(ii) Use symmetry of the pdf around $x = \mu$. For (i), alternatively look at the integral $\mathbb{E}[X] = \int_{-\infty}^{\infty} x \frac{e^{-\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\sigma} dx$, and use u-substitution. For (iii), write $\mathbb{E}[X^2] = \int_{-\infty}^{\infty} xx \frac{e^{-\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\sigma} dx$, and do integration by parts.)*

Finally, we state the following result which extends Theorem 56.1 to sums of independent normal random variables. We are not able to give a proof of this result here, but afterwards we will see an important application on sample mean of i.i.d. normal random variables.

Theorem 56.4 *If $X_1, \ldots, X_n$ are independent, $X_i \sim N(\mu_i, \sigma_i^2)$, and if $a_1, \ldots, a_n$ are real numbers, then*

$$a_1X_1 + \cdots + a_nX_n \sim N(a_1\mu_1 + \cdots + a_n\mu_n, a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2).$$

Now, assuming that $X_1, \ldots, X_n$ are i.i.d. normal random variables with parameters μ and σ^2, Theorem 56.4 with $a_i = 1/n$, $i = 1, \ldots, n$ yields the following:

Theorem 56.5 *Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables. Then the sample mean is normally distributed with parameters μ and σ^2/n, i.e.,*

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

In particular,

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

57 Confidence intervals 0

The purpose of this short section is to give a first motivation for confidence intervals. Let $X_1, \ldots, X_n$ be i.i.d. normally distributed samples with an unknown mean μ and with a known variance σ^2. We already know that

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

provides a point estimate for μ, and that $\overline{X} \longrightarrow_{\mathbb{P}} \mu$ as $n \to \infty$. Our goal here is to strengthen this information by giving an interval estimator for μ. In particular, we would like to conclude that a certain *random* interval will contain the actual parameter μ with a predetermined probability $1 - \alpha$ in $(0, 1)$. Here is how we do it.

Since X_i's are i.i.d. normal random variables, we know from Theorem 56.5 that $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. Letting then z_β be the real number for which $\mathbb{P}(Z > z_\beta) = \beta$ for a standard normal random variable Z, and fixing some $\alpha \in (0,1)$, we have

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}\right) \\ &= \mathbb{P}\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right). \end{aligned}$$

This says that the random interval

$$\left[\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

contains the parameter μ with probability $1-\alpha$. We call it a $(1-\alpha)100\%$ confidence interval for the population mean μ.

Example 57.1 *Assume that the grades of students in a certain math course are normally distributed with an unknown mean μ and a known variance 144. Suppose that we will sample 36 people, record their grades, $X_1, \ldots, X_{36}$ and will construct a confidence interval based on this data.*

- *Taking $\alpha = 0.05$, our discussion above tells us that the random interval*

  $$\left[\overline{X}-z_{0.025}\frac{12}{6}, \overline{X}+z_{0.025}\frac{12}{6}\right]$$

  will contain the true parameter μ with probability .95.

- *Suppose that the sample mean for our specific sample turns out to be 60. Reading $z_{0.025} = 1.96$ from the z-table, the confidence interval for this specific sample is*

  $$\left[60-1.96\frac{12}{6}, 60+1.96\frac{12}{6}\right] \approx [56,64].$$

- *It is important you realize that the probability of μ being in $[56,64]$ is not 0.95. μ is just a fixed real number[53] and so it belongs to $[56,64]$ either with probability 0 or with probability 1.*

[53] μ being a fixed real number is part of the frequentist statistics. There is an alternative approach where the parameter μ is also random itself - check Bayesian statistics.

- *What we actually mean by a confidence interval is the following: If we form m many* 95% independent *confidence intervals via resampling, probability that μ will be in these confidence intervals is approximately* 95% *assuming that $\underline{m \text{ is large}}$. This statement can be made precise - via law of large numbers - but we skip the rigor here.* □

58 Multivariate distributions in continuous setting

Treatment of multivariate continuous distributions is similar to the discrete case. By the **joint pdf** of two continuous random variables X and Y, we mean a function $f : \mathbb{R}^2 \to \mathbb{R}^+$ so that for any $R \subset \mathbb{R}^2$ we have

$$\mathbb{P}((X,Y) \in R) = \int\int_R f(x,y)dxdy.$$

Note that we necessarily have

$$f(x,y) \geq 0, \quad \text{for all } x, y \in \mathbb{R}$$

and

$$\int\int_{\mathbb{R}^2} f(x,y)dxdy = 1.$$

In this setting **the marginal pdf** f_1 of X is

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy, \qquad -\infty < x < \infty,$$

and the marginal pdf f_2 of Y is

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)dx, \qquad -\infty < y < \infty.$$

Example 58.1 *Let X and Y have joint pdf*

$$f(x,y) = 2, \qquad x \in [0,1], y \in [0,1/2].$$

Find the marginal pdfs of X.

Solution: The marginal pdfs of X and Y

$$f_1(x) = \int_0^{1/2} f(x,y)dy = \int_0^{1/2} 2dy = 1, \qquad x \in [0,1],$$

and

$$f_2(y) = \int_0^1 2dy = \int_0^1 2dy = 2, \qquad y \in [0,2],$$

respectively. □

Example 58.2 *Let X and Y have joint pdf $f(x,y) = \frac{21}{4}x^2y,\ x^2 \leq y \leq 1$. Find the marginal pdf of X and Y.*

Solution: The marginal pdfs of X and Y are given by

$$f_1(x) = \int_{x^2}^{1} f(x,y)dy = \int_{x^2}^{1} \frac{21}{4}x^2ydy = \frac{21}{8}x^2(1-x^4), \qquad -1 \leq x \leq 1$$

and

$$f_2(y) = \int_{-\sqrt{y}}^{\sqrt{y}} f(x,y)dx = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{21}{4}x^2ydx = \frac{7}{2}y^{5/2}, \qquad 0 \leq y \leq 1.$$

□

Recall now the definition for independence of random variables.

Definition 58.1 *Two random variables X and Y are **independent** if for all $A, B \subset \mathbb{R}$*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B). \tag{13}$$

As in discrete setting, it is enough to find two subsets of the real line for which the condition in (13) is violated to show that X and Y are dependent. For proving independence, one useful tool is the following factorization theorem which we previously saw in the discrete setting.

Theorem 58.1 *(Factorization theorem) Suppose X and Y are random variables that have a joint pdf f. Then X and Y are independent if and only if*

$$f(x,y) = h_1(x)h_2(y), \qquad -\infty < x, y < \infty$$

where h_1 is a function of x alone, and h_2 is a function of y alone.

Proof: ($\Rightarrow$) Assume that X and Y are independent. Then for any $A, B \subset \mathbb{R}$, we have

$$\begin{aligned}\int_A\int_B f(x,y)dydx = \mathbb{P}(X \in A, Y \in B) \quad &\overbrace{=}^{\textit{independence}} \quad \mathbb{P}(X \in A)\mathbb{P}(Y \in B)\\ &= \quad \int_A f_1(x)dx \int_B f_2(y)dy\\ &= \quad \int_A\int_B f_1(x)f_2(y)dxdy.\end{aligned}$$

Since this is true for any A, B, we may conclude that $f(x,y) = f_1(x)f_2(y)$.

($\Leftarrow$) Assume $f(x,y) = h_1(x)h_2(y)$. Then

$$f_1(x) = \int_{-\infty}^{\infty} h_1(x)h_2(y)dy = c_1 h_1(x)$$

where $c_1 = \int_{-\infty}^{\infty} h_2(y)dy$. This gives $h_1(x) = \frac{f_1(x)}{c_1}$. Also,

$$f_2(y) = \int_{-\infty}^{\infty} h_1(x)h_2(y)dx = h_2(y)\int_{-\infty}^{\infty} h_1(x)dx = h_2(y)\int_{-\infty}^{\infty} \frac{f_1(x)}{c_1}dx = \frac{h_2(y)}{c_1},$$

yielding $h_2(y) = c_1 f_2(y)$.

So for any $A, B \subset \mathbb{R}$

$$\begin{aligned}\mathbb{P}(X \in A, Y \in B) = \int_A\int_B f(x,y)dydx &= \int_A\int_B h_1(x)h_2(y)dydx \\ &= \int_A\int_B \frac{f_1(x)}{c_1}c_1 f_2(y)dydx \\ &= \int_A f_1(x)dx \int_B f_2(y)dy \\ &= \mathbb{P}(X \in A)\mathbb{P}(Y \in B).\end{aligned}$$

□

Example 58.3 *Let X, Y have joint continuous distribution with pdf $f(x,y) = 1, 0 < x, y < 1$. Are X and Y independent?*

Solution: Yes, this follows immediately from Theorem 58.1 - just choose $h_1(x) = 1$, $h_2(y) = 1$ for $0 < x, y < 1$. □

Proof of Theorem 58.1 gives the following as a corollary.

Corollary 58.1 *Suppose X and Y are random variables that have a joint pdf f. Then X and Y are independent if and only if*

$$f(x,y) = f_1(x)f_2(y), \qquad -\infty < x, y < \infty$$

where f_1 and f_2 are the marginal pdfs of X and Y, respectively.

Example 58.4 *Suppose that the joint pdf of X and Y is given by*

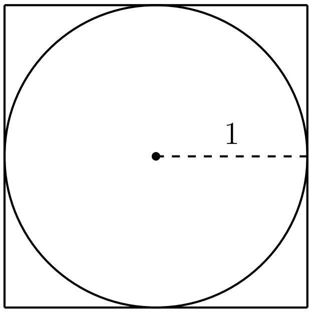
$$f(x,y) = kx^2y^2, \qquad x^2 + y^2 \leq 1.$$

Are X and Y independent?

Solution: Note that $f(0.9, 0.9) = 0$, but $f_1(0.9) \neq 0$ and $f_2(0.9) \neq 0$. So $f(0.9, 0.9) \neq f_1(0.9)f_2(0.9)$, and we conclude that X and Y are not independent. (The **important** point here is that, for independence we should have $f(x,y) = f_1(x)f_2(y)$ **for all** x, y.) □

59 * Von Neumann's rejection sampling

In this section, we discuss a slightly more complicated random number generation method. Let's begin with the idea behind accept/reject methods via a simple example. Assume that we want to generate a random point in the unit circle inscribed in a square whose corners are at $(1,1),(1,-1),(-1,1)$ and $(-1,-1)$:



Here is how the rejection sampling would work in this case:

1. Generate a candidate point (X,Y) where X,Y are independent $U(-1,1)$ random variables.
2. If $x^2+y^2 \leq 1$, then accept the sample.
3. Otherwise, go back to Step 1.

The idea is simple, right? Next, let's move on to Von Neumann's very nice observation. Assume that we would like to sample from a complicated distribution on $\mathbb{R}$ (in principle, and in practice, we may a space other than $\mathbb{R}$, but for the sake of simplicity, we assume so) with pmf/pdf f. Assume that g is some other pmf/pdf, which is easy to sample from, which satisfies

$$f(x) \leq Mg(x),$$

for all $x \in \mathbb{R}$ for some constant $M > 0$.

Here is the rejection sampling algorithm of Von Neumann:

Von Neumann's rejection sampling algorithm.

1. Sample X^* from g and U from $U(0,1)$, independently.
2. If $U \leq \frac{f(X^*)}{Mg(X^*)}$, then stop and return $X^{**} = X^*$ as your sample from f.
3. Otherwise, go back to step 1.

Theorem 59.1 *Von Neumann's rejection sampling algorithm produces samples X^{**} from the distribution with pmf/pdf f.*

Proof: We assume that we are in the continuous setting so that f is a pdf (In this case g is also necessarily a pdf). Observe that we have

$$\mathbb{P}(X^{**} \leq x) = \mathbb{P}\left(X^* \leq x \mid U \leq \frac{f(X^*)}{Mg(X^*)}\right) = \frac{\mathbb{P}\left(X^* \leq x, U \leq \frac{f(X^*)}{Mg(X^*)}\right)}{\mathbb{P}\left(U \leq \frac{f(X^*)}{Mg(X^*)}\right)}.$$

Now, we compute these two probabilities separately.

For the numerator

$$\begin{aligned}\mathbb{P}\left(X^* \leq x, U \leq \frac{f(X^*)}{Mg(X^*)}\right) &= \int_{-\infty}^{x}\int_{0}^{\frac{f(x^*)}{Mg(x^*)}} g(x^*)dudx^* \\ &= \int_{-\infty}^{x} \frac{f(x^*)}{M}dx^* \\ &= \frac{\mathbb{P}(X \leq x)}{M},\end{aligned}$$

where X has pdf f.

Also, for the denominator, we observe

$$\mathbb{P}\left(U \leq \frac{f(X^*)}{Mg(X^*)}\right) = \int_{-\infty}^{\infty}\int_{0}^{\frac{f(x^*)}{Mg(x^*)}} g(x^*)dudx^* = \int_{-\infty}^{\infty} \frac{f(x^*)}{M}dx^* = \frac{1}{M}.$$

Therefore, we conclude that

$$\mathbb{P}(X^{**} \leq x) = \frac{\frac{\mathbb{P}(X \leq x)}{M}}{\frac{1}{M}} = \mathbb{P}(X \leq x),$$

and we are done. □

Remark 59.1 *(i.) Note that the acceptance probability is given by*

$$\mathbb{P}\left(U \leq \frac{f(X^*)}{Mg(X^*)}\right) = \frac{1}{M}.$$

Naturally, we would like to maximize the acceptance probability, which amounts to choose M as small as possible. This choice will stem from an appropriate selection of the g function.

(ii.) For a more advanced application of the rejection sampling idea, you may check the Ziggurat algorithm.

Exercise 59.1 *Explain how you can use the standard Cauchy distribution (whose pdf is given by $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$, $-\infty < x < \infty$) as an envelope to generate a standard normal random number by using the accept/reject method. (Note that first of all you should be okay with generating random numbers according to Cauchy distribution. Consider the inversion method for this purpose.)*

60 Extremal statistics

In this section we are given an i.i.d. sample $X_1, \dots, X_n$ from some distribution with pdf f and cdf F. Our interest is in understanding the extremal statistics

$$\max\{X_1, \dots, X_n\}, \quad \text{and} \quad \min\{X_1, \dots, X_n\}.$$

Beginning with the max case, let's answer the following questions: Express the cdf, pdf and expectation of $\max\{X_1, \dots, X_n\}$ in terms of f and F. We begin with the cdf. For $x \in \mathbb{R}$ we have

$$\begin{aligned}\mathbb{P}(\max\{X_1, \dots, X_n\} \le x) = \mathbb{P}(X_1 \le x, \dots, X_n \le x) = \mathbb{P}\left(\bigcap_{i=1}^n \{X_i \le x\}\right) &= \prod_{i=1}^n \mathbb{P}(X_i \le x) \\ &= \prod_{i=1}^n F(x) \\ &= (F(x))^n .\end{aligned}$$

Taking the derivatives, the pdf of $\max\{X_1, \dots, X_n\}$ then turns out to be

$$f_{\max\{X_1,\dots,X_n\}}(x) = \frac{d}{dx}\left(\mathbb{P}(\max\{X_1, \dots, X_n\} \le x)\right) = nF^{n-1}(x)f(x),$$

and the expectation of the maximum can be found as

$$\mathbb{E}[\max\{X_1, \dots, X_n\}] = \int_{-\infty}^{\infty} xnF^{n-1}(x)f(x)dx.$$

The treatment for the minimum is similar. Just note that in this case $\min\{X_1, \dots, X_n\} \le x$ does not imply that each X_i is less or equal to x. Instead of this one needs to begin with the observation

$$\mathbb{P}(\min\{X_1, \dots, X_n\} \le x) = 1 - \mathbb{P}(\min\{X_1, \dots, X_n\} > x) = 1 - \mathbb{P}\left(\bigcap_{i=1}^n \{X_i > x\}\right).$$

Exercise 60.1 *In setting above derive the formulas for the cdf, pdf and the expectation of the minimum of n i.i.d. random samples $X_1, \dots, X_n$.*

Example 60.1 *Let $X_1, X_2, \ldots, X_n$ be independent random variables that are uniformly distributed over the interval $(0,1)$. Let $Y_n = \max\{X_1, X_2, \ldots, X_n\}$.*

(i) Find the cdf G_n of Y_n.

(ii) Find the pdf g_n of Y_n.

(iii) Find $\mathbb{E}[Y_n]$.

(iv) Find $\lim_{n\to\infty} \mathbb{E}[Y_n]$.

(v) For which values of n we will have $\mathbb{P}(Y_n \geq 0.99) \geq 0.95$? (Note : It is okay to leave your answer in terms of logarithms.)

Solution: (i) Let $y \in (0,1)$. Then

$$G_n(y) = \mathbb{P}(Y_n \leq y) = \mathbb{P}\left(\bigcap_{i=1}^{n}\{X_i \leq y\}\right) = (\mathbb{P}(X_i \leq y))^n = y^n.$$

Also, $G_n(y) = 0, y \leq 0$ and $G_n(y) = 1, y \geq 1$.

(ii) We take the derivative of G_n to obtain the pdf g_n as

$$g_n(y) = \begin{cases} ny^{n-1}, & \text{for } y \in (0,1) \\ 0, & \text{otherwise.} \end{cases}$$

(iii) We use the pdf to obtain

$$\mathbb{E}[Y_n] = \int_0^1 yny^{n-1}dy = \frac{n}{n+1}.$$

(iv) We take the limit in part (iii) to obtain

$$\lim_{n\to\infty} \mathbb{E}[Y_n] = \lim_{n\to\infty} \frac{n}{n+1} = 1$$

(v) Observe that

$$\mathbb{P}(Y_n \geq 0.99) = 1 - G_n(0.99) = 1 - (0.99)^n.$$

A simple manipulation shows that the required condition is

$$n \geq \frac{\log(0.05)}{\log(0.99)}.$$

□

Example 60.2 *If $X_1, \ldots, X_n$ form an i.i.d. sample from the exponential distribution with parameter λ. Show then that the random variable $Y = \min\{X_1, \ldots, X_n\}$ has exponential distribution with parameter $n\lambda$.*

Solution: Note that the minimum of $X_1, \ldots, X_n$ will be at least $x \in \mathbb{R}$ if and only if each of $X_1, \ldots, X_n$ is at least x. Using this and the i.i.d. assumption, we have

$$\begin{aligned}\mathbb{P}(Y \geq x) = \mathbb{P}(\min\{X_1, \ldots, X_n\} \geq x) &= \mathbb{P}(X_1 \geq x, \ldots, X_n \geq x) \\ &= \mathbb{P}(X_1 \geq x) \ldots \mathbb{P}(X_n \geq x) \\ &= e^{-n\lambda x}.\end{aligned}$$

So, $\mathbb{P}(Y \leq x) = 1 - e^{-n\lambda x}$ which is the cdf of an exponential with parameter $n\lambda$. Result follows. □

61 Review problems V

Exercise 61.1 *Let X be a continuous random variable with pdf f whose support is $\mathbb{R}$. Let $\{a_i\}_{i\geq 1}$ be a sequence of distinct real numbers. Let $A = \bigcup_{i=1}^{\infty}\{a_i\}$. Find $\mathbb{P}(A)$.*

Exercise 61.1 *Let X_1, X_2, X_3 be i.i.d. continuous random variables. Find*

(i) $\mathbb{P}(X_1 < X_2)$;

(ii) $\mathbb{P}(X_1 < X_2 < X_3)$;

(iii) $\mathbb{P}(X_1 < X_2, X_2 > X_3)$;

(iv) $\mathbb{P}(X_1 = X_2 = X_3)$;

(v) $\mathbb{P}(X_1 + X_2 = X_3)$.

Exercise 61.2 *For which value of c is the following function the pdf of a random variable:*

$$f(x) = ce^x(1 + e^x)^{-2}, \quad x \in \mathbb{R}?$$

Exercise 61.3 *(i) Prove that if f_1 and f_2 are density functions, and $0 \leq c \leq 1$, then $cf_1 + (1-c)f_2$ is also a density function.*

(ii) How can we generalize the statement in part (i) to k many density functions?

(iii) Again assume that f_1 and f_2 are density functions. Is $f_1 f_2$ a density function in general? (Answer: No.)

Exercise 61.4 *Prove that the following are cdfs.*

a. $\frac{1}{2} + \frac{1}{\pi}\arctan(x), x \in (-\infty, \infty)$;

b. $(1 + e^{-x})^{-1}, x \in (-\infty, \infty)$;

c. $e^{-e^{-x}}, x \in (-\infty, \infty)$*;*

Exercise 61.5 *Suppose that* $X \sim U(0,1)$*. Explain how we may use this random number to generate a random number that is uniformly distributed over* $(2,7)$*.*

Exercise 61.6 *Show that we can not have a uniform distribution over an unbounded set.*

Exercise 61.7 *Can you produce samples from the normal distribution by using the inversion method? (Answer: No. Question 2: What would be second best thing similar to inversion method you would try for generating approximately normal random variables?)*

Exercise 61.8 *Suppose that* $X_1, \ldots, X_n$ *are i.i.d. random variables, each of which has a continuous distribution with median* m*. Let* $Y_n = \max\{X_1, \ldots, X_n\}$*. Determine the value of* $\mathbb{P}(Y_n > m)$*.*

Exercise 61.2 *Let* X *and* Y *be jointly continuous functions whose joint pdf is given by*

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

(i) Find the constant c*.*
(ii) Find $\mathbb{P}(0 \leq X \leq 1/3, 1/3 \leq Y \leq 2/3)$*.*
(iii) Find $\mathbb{P}(Y = X^2)$*.*

Exercise 61.3 *Let* X *and* Y *be jointly continuous functions whose joint pdf is given by*

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

(i) Find the constant c*.*
(ii) Sketch the region where $f(x,y) > 0$*.*
(iii) Find the marginal pdfs f_1*,* f_2 *of* X *and* Y*.*
(iv) Find $\mathbb{P}(Y \leq X/2)$*.*
(v) Find $\mathbb{P}(Y \leq X/4 \mid Y \leq X/2)$*.*

Exercise 61.9 *An accident occurs at a point X that is uniformly distributed on a road of length L. At the time of the accident, an ambulance is at a location Y that is also uniformly distributed on the road. Assuming that X and Y are independent, find the expected distance between the ambulance and the point of the accident.*

Exercise 61.10 *Let X be a continuous random variable with pdf $f(x)$ and cdf $F(x)$. For a fixed number x_0, define the function*

$$g(x) = \begin{cases} \frac{f(x)}{1-F(x_0)}, & \text{if } x \geq x_0 \\ 0, & \text{if } x < 0. \end{cases}$$

Prove that $g(x)$ is a pdf assuming $F(x_0) < 1$.

Exercise 61.11 *In each of the following find the pdf of Y. Show that the pdf integrates to 1.*

(i) $f_X(x) = 42x^5(1-x)$, $0 < x < 1$; $Y = X^3$;

(ii) $f_X(x) = 7e^{-7x}$, $0 < x < \infty$; $Y = 4X + 3$;

(iii) $f_X(x) = 30x^2(1-x)^2$, $0 < x < 1$; $Y = X^2$.

Exercise 61.12 *Show that the given function is a cdf and find $F_X^{-1}(y)$:*

$$F_X(x) = \begin{cases} e^x/2, & \text{if } x < 0 \\ 1/2, & \text{if } 0 \leq x \leq 1 \\ 1 - (e^{1-x}/2), & \text{if } 1 \leq x. \end{cases}$$

Exercise 61.13 *Given the cdf $F(x) = \frac{x^3}{125}$ on $[0, 5]$, explain how you can get samples from the distribution corresponding to F.*

Exercise 61.14 *Compute $\mathbb{E}[X]$ and $Var(X)$ whose pdfs are as follows:*

(i) $f_X(x) = ax^{a-1}$, $0 < x < 1$, $a > 0$;

(iilç) $f_X(x) = \frac{3}{2}(x-1)^2$, $0 < x < 2$.

Exercise 61.15 *A certain river floods every year. Suppose that the low-water mark is set at 1 and the high-water mark Y has distribution function*

$$F_Y(y) = \mathbb{P}(Y \leq y) = 1 - \frac{1}{y^2}, \quad 1 \leq y < \infty.$$

(i) Verify that $F_Y(y)$ is a cdf.

(ii) Find $f_Y(y)$, the pdf of Y.

(iii) If the low-water mark is reset at zero and we use a unit of measurement which is $1/10$ of that given previously, the high-water mark becomes $Z = 10(Y-1)$. Find $F_Z(z)$.

Exercise 61.16 *Let $X \sim N(\mu, \sigma^2)$. Find values of μ and σ^2 such that $\mathbb{P}(|X| < 2) = \frac{1}{2}$. Prove or disprove that these values of μ and σ^2 are unique.*

Exercise 61.17 *Suppose that a random sample of 18 observations is drawn from the normal distribution with mean* 0 *and standard deviation 3, and that independently another random sample of 8 observations is drawn from the normal distribution with mean* 1 *and standard deviation 2. Let $\overline{X}$ and $\overline{Y}$ denote the sample means of the two samples.*

(i) Find $\mathbb{E}[\overline{X} - \overline{Y}]$ and $Var(\overline{X} - \overline{Y})$.

(ii) What is the distribution of $\overline{X} - \overline{Y}$? (Do not forget to include parameters of the distribution.)

(iii) Find $\mathbb{P}(\overline{X} - \overline{Y} \leq 1)$

(iv) Find the conditional probability $\mathbb{P}(\overline{X} - \overline{Y} \leq 1 \mid \overline{X} - \overline{Y} \leq 2)$.

(v) Find the conditional probability $\mathbb{P}(\overline{X} - \overline{Y} \leq 2 \mid \overline{X} - \overline{Y} \leq 1)$.

Exercise 61.18 *A random point (X, Y) is distributed uniformly on the square with vertices (1,1), (1, -1), (-1,1), (-1,-1). That is the joint pdf is $f(x, y) = \frac{1}{4}$ on the square. Find the following probabilities:*
a. $X^2 + Y^2 < 1$;
b. $2X - Y > 0$;
c. $|X + Y| < 2$.

Exercise 61.19 *Find $\mathbb{P}(X^2 < Y < X)$ if X and Y are jointly distributed with pdf*

$$f(x, y) = xy, \qquad 0 \leq x \leq 1, \quad 0 \leq y \leq 1.$$

Exercise 61.20 *Let X, Y, Z be i.i.d. random variables uniformly distributed over $(0, 1)$.*
a. Find $\mathbb{P}(X/Y \leq t)$ and $\mathbb{P}(XY \leq t)$. (Hint: Pictures will help)
b. Find $\mathbb{P}(XY/Z \leq t)$.

Exercise 61.21 *Find the pdf of $\prod_{i=1}^{n} X_i$, where the X_i's are independent $U(0, 1)$ random variables. (Hint: Try to calculate the cdf, and remember the relation between uniforms and exponentials.)*

Exercise 61.22 *Let $X_1, \ldots, X_n$ be a random sample from a $N(0, 1)$ population. Define*

$$Y_1 = \left| \frac{1}{n} \sum_{i=1}^{n} X_i \right|, \quad Y_2 = \frac{1}{n} \sum_{i=1}^{n} |X_i|.$$

Compute $\mathbb{E}[Y_1]$ and $\mathbb{E}[Y_2]$, and establish an inequality between them.

Exercise 61.23 *Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Find a function of S^2, the sample variance, say $g(S^2)$, that satisfies $\mathbb{E}[g(S^2)] = \sigma$. (Hint: Try $g(S^2) = c\sqrt{S^2}$, where c is a constant.)*

Exercise 61.24 *Let $X_1, X_2, \ldots, X_n$ be independent random variables that are uniformly distributed over the interval $(0, 1)$. Find $\mathbb{E}[\min\{X_1, \ldots, X_n\}]$.*

Exercise 61.25 *An urn contains 20 balls, numbered 1 through 20. 10 balls are independently withdrawn in sequence, each time replacing the ball selected previously. Let X be the minimum of the 10 chosen numbers.*

(i) Find the cdf of X.

(ii) Find $\mathbb{P}(X = k)$, for $k = 1, 2, ..., 20$.

Exercise 61.26 *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the uniform distribution over $\{1, 2, \ldots, M\}$ where M is an unknown parameter which is some positive integer greater or equal 2.*

(i) Prove that the likelihood estimator for M is $\hat{M} = \max\{X_1, \ldots, X_n\}$.

(ii) Give a heuristic argument that $\hat{M}$ is a biased estimator for M. If possible support your explanation by guessing whether $\mathbb{E}[\hat{M}] < M$ or $\mathbb{E}[\hat{M}] > M$, and explaining why.

(iii) Find the cdf of $\hat{M}$.

(iv) Make your explanation in part (ii) rigorous by finding an explicit value for the bias.

Exercise 61.27 *Suppose $\overline{X}$ and S^2 are calculated from a random sample $X_1, \ldots, X_n$ drawn from a population with finite variance σ^2. We know that $\mathbb{E}[S^2] = \sigma^2$. Prove that $\mathbb{E}[S] \leq \sigma$, and if $\sigma^2 > 0$, then $\mathbb{E}[S] < \sigma$.*

Exercise 61.28 *If a stick is broken at random into three pieces, what is the probability that the pieces can be put together in a triangle? (See Gardner 1961 for a complete discussion of this problem.)*

62 Likelihood estimation in continuous setting

The likelihood estimation in continuous setting follows similar lines to the discrete case. Letting $X_1, \ldots, X_n$ be i.i.d. samples from some distribution with pdf $f(x \mid \theta)$ where $\theta \in \Theta$, recall that

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

is the **likelihood function**, and

$$\ell(\theta) = \ln L(\theta)$$

is the **log-likelihood function**. Our goal is to maximize the likelihood $L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$ given that the samples we have are $x_1, \ldots, x_n$. We will now look at some examples in continuous setting.

Example 62.1 *Let*

$$f(x \mid \theta) = \theta x^{\theta - 1}, \quad 0 < x < 1, \quad \theta \in \Theta = (0, \infty).$$

Let $X_1, X_2, \ldots, X_n$ denote an i.i.d. sample of size n from the distribution with pdf f.

a. Sketch the pdf of X_1 for (i.) $\theta = 1/2$, (ii.) $\theta = 1$, and (iii.) $\theta = 2$.

b. Show that $\widehat{\theta} = -n / \ln\left(\Pi_{i=1}^{n} X_i\right)$ is the maximum likelihood estimator of θ.

c. Based on the following 10 observations find the maximum likelihood estimate for θ:

0.0256	0.3051	0.0278	0.8971	0.0739
0.3191	0.7379	0.3671	0.9763	0.0102

Solution. (a.) Left for you.
(b.) The likelihood function is given by

$$L(\theta) = L(\theta | X_1, \ldots, X_n) = \prod_{i=1}^{n} \theta X_i^{\theta - 1} = \theta^n \left(\prod_{i=1}^{n} X_i \right)^{\theta - 1}.$$

Taking ln on both sides we obtain the log-likelihood function as

$$\ell(\theta) = n \ln \theta + (\theta - 1) \ln \left(\prod_{i=1}^{n} X_i \right).$$

Now, solving for

$$\ell'(\theta) = \frac{n}{\theta} + \prod_{i=1}^{n} X_i = 0$$

gives

$$\widehat{\theta} = -n / \ln \left(\prod_{i=1}^{n} X_i \right)$$

as a candidate for the maximizer of the likelihood function. This is verified by using the second derivative test since $\ell'(\theta) = -\frac{n}{\theta^2} < 0$.

(c.) You just need to evaluate the likelihood function at the given data points with $n = 10$.

□

Example 62.2 *Let $X_1, \ldots, X_n$ be i.i.d. random samples from a gamma distribution with parameters α and β whose pdf is given by*

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1}, \quad x > 0$$

where

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

is the gamma function. We assume that the parameter α is known, and that we would like to estimate β. What is the maximum likelihood estimator of β?

Solution. The likelihood function is

$$L(\beta) = \prod_{i=1}^n f(X_i|\beta) = \prod_{i=1}^n \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta X_i} X_i^{\alpha-1} = \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n} e^{-\beta \sum_{i=1}^n X_i} \left(\prod_{i=1}^n X_i \right)^{\alpha-1}.$$

Then we obtain the log-likelihood as

$$\ell(\beta) = (n\alpha) \ln \beta - n \ln \Gamma(\alpha) - \beta \sum_{i=1}^n X_i + (\alpha - 1) \sum_{i=1}^n \ln(X_i).$$

Solving

$$\ell'(\beta) = \frac{n\alpha}{\beta} - \sum_{i=1}^n X_i = 0,$$

and using the second derivative test yield the maximum likelihood estimator as

$$\widehat{\beta} = \frac{\alpha n}{\sum_{i=1}^n X_i} = \frac{\alpha}{\overline{X}}.$$

□

Example 62.3 *Let $X_1, \ldots, X_n$ be a random sample (i.i.d.) from the distribution whose pdf is given by*

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

where θ is an unknown parameter. That is, X_i's are uniformly distributed over the unknown interval $[0, \theta]$.

(i) Find the maximum likelihood estimator of θ based on your sample.

(ii) Calculate the expected value of your estimator in part i. Is it an unbiased estimator? Is it asymptotically unbiased?

Solution. (i) Likelihood function is given by

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n}, & \text{if } \theta \geq \max\{X_1, \ldots, X_n\}, \\ 0, & \text{otherwise.} \end{cases}$$

Noting that $L'(\theta) < 0$ for $\theta \geq \max\{X_1, \ldots, X_n\}$, we should choose θ as small as possible. This yields the maximum likelihood estimator as

$$\widehat{\theta} = u(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\}.$$

(ii) First we find the cdf of $\widehat{\theta}$. For $x \in [0, \theta]$, we have

$$\mathbb{P}(\widehat{\theta} \leq x) = \mathbb{P}(X_1 \leq x, \ldots, X_n \leq x) = (\mathbb{P}(X_1 \leq x))^n = \left(\int_0^x \frac{1}{\theta} dy\right)^n = \left(\frac{x}{\theta}\right)^n.$$

Taking derivative, we find the pdf of $\widehat{\theta}$ as

$$g(x) = \frac{nx^{n-1}}{\theta^n}, \qquad 0 \leq x \leq \theta.$$

Then we compute

$$\mathbb{E}[\widehat{\theta}] = \int_0^\theta x \frac{nx^{n-1}}{\theta^n} dx = \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{n+1}\theta.$$

Since $\mathbb{E}[\widehat{\theta}] \neq \theta$, we conclude that $\widehat{\theta}$ is a biased estimator[54]. However,

$$\lim_{n\to\infty} u(X_1, \ldots, X_n) = \lim_{n\to\infty} \frac{n}{n+1}\theta = \theta,$$

and so $\widehat{\theta}$ is asymptotically unbiased. □

Example 62.4 *Let $X_1, X_2, \ldots, X_n$ be i.i.d. normal random variables with mean θ_1 and variance θ_2. We assume that we do not know neither θ_1 nor θ_2.*

(i) Find the maximum likelihood estimator $u_1(X_1, \ldots, X_n)$ of θ_1.

(ii) Find the maximum likelihood estimator $u_2(X_1, \ldots, X_n)$ of θ_2.

(iii) Is the estimator you found in part i unbiased?

(iv) Is the estimator you found in part ii unbiased?

Solution. (i) and (ii) The joint likelihood function for θ_1, θ_2 is given

$$L(\theta_1, \theta_2) = \prod_{i=1}^{n} f(X_i \mid \theta_1, \theta_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2}\frac{(X_i-\theta_1)^2}{\theta_2}} = \frac{1}{(2\pi)^{n/2}(\theta_2)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(X_i-\theta_1)^2}{\theta_2}}.$$

[54] Can you give an intuitive explanation for the bias here?

We then find the log-likelihood function as[55]

$$\ell(\theta_1, \theta_2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\theta_2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(X_i - \theta_1)^2}{\theta_2}.$$

The partial derivatives of ℓ are

$$\frac{\partial \ell}{\theta_1} = \sum_{i=1}^{n}\frac{X_i - \theta_1}{\theta_2},$$

and

$$\frac{\partial \ell}{\theta_2} = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(X_i - \theta_1)^2.$$

Setting $\frac{\partial \ell}{\theta_1} = 0$ and $\frac{\partial \ell}{\theta_2} = 0$, and using the vector form of second derivative test, we conclude that the maximum likelihood estimators are

$$\widehat{\theta_1} = \overline{X},$$

and

$$\widehat{\theta_2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

(iii) Yes, since $\mathbb{E}[\widehat{\theta}] = \mathbb{E}[\overline{X}] = \theta_1$.

(iv) No. To see this, recall that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an unbiased estimator of θ_2. Then, $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ can not be unbiased.

□

Remark 62.1 *Note that part (iv) in previous example explains the $1/(n-1)$ term instead of $1/n$ in definition of the sample variance.*

[55] Sometimes you will see people writing

$$\ell(\theta_1, \theta_2) \ \propto -\frac{n}{2}\ln\theta_2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(X_i - \theta_1)^2}{\theta_2}.$$

This is just to emphasize that the term $-\frac{n}{2}\ln(2\pi)$ can be regarded as a constant in this problem. For example, under the same setting if we knew the value of θ_2, then we could write

$$\ell(\theta_1) \ \propto \frac{1}{2}\sum_{i=1}^{n}\frac{(X_i - \theta_1)^2}{\theta_2}$$

since this time θ_2 is considered as a constant as well.

Example 62.5 *Find the maximum likelihood estimates for $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ if a random sample of size 15 from $N(\mu, \sigma^2)$ yielded the following values:*

31.5	36.9	33.8	30.1	33.9
35.2	29.6	34.4	30.5	34.2
31.6	36.7	35.8	34.5	32.7

Solution. Recall from the previous example that the maximum likelihood estimators for θ_1 and θ_2 are given by

$$\widehat{\theta_1} = \overline{X},$$

and

$$\widehat{\theta_2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Now you just need to substitute the numbers given in these estimators with $n = 15$. □

63 Mean squared error of estimators

Recall that previously we defined the mean square error of a prediction d for a random variable X to be $MSE(d) = \mathbb{E}|X - d|^2$. We have a similar notion for estimators.

Definition 63.1 *Let $T(X_1, \ldots, X_n)$ be an estimator for a parameter θ based on the random sample $X_1, \ldots, X_n$. Then the* ***mean squared error of the estimator*** *is defined by*

$$MSE(T; \theta) = \mathbb{E}|T(X_1, \ldots, X_n) - \theta|^2.$$

Remark 63.1 *As in case of predictions, we may also introduce mean absolute error or some other criteria. We will not go into details of these here.*

Example 63.1 *Let $X_1, \ldots, X_n$ be i.i.d. samples from a distribution with $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = Var(X_1) < \infty$. Find the mean squared error of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimator of μ.*

Solution: We have

$$MSE(\overline{X}; \mu) = \mathbb{E}[(\overline{X} - \mu)^2] = Var(\overline{X}) = \frac{\sigma^2}{n}.$$

□

Remark 63.2 *Here is a little bit discussion about the estimators for the variance. Recall that* $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ *was an unbiased estimator for* $\sigma^2 = Var(X_1)$ *and* $R_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{n-1}{n} S_n^2$ *was a biased one again for* σ^2. *That is one good thing about* S_n^2 *but of course it is not the end of story. For example, under a fourth moment assumption we may show that*

$$MSE(R_n^2; \sigma^2) < MSE(S_n^2; \sigma^2)$$

and so according to this criteria one would prefer R_n^2 *over* S_n^2. *In general, there are various properties of an estimator that needs to be considered and one selects the right one for the specific problem.*

The following result relates mean square error to the bias of the estimator by giving some insight to our discussion in previous remark.

Proposition 63.1 *We have*

$$MSE(T; \theta) = Var(T) + (Bias(T; \theta))^2.$$

Proof: We have

$$\begin{aligned} \mathbb{E}[(T-\theta)^2] &= \mathbb{E}[(T - \mathbb{E}[T] + \mathbb{E}[T] - \theta)^2] \\ &= \mathbb{E}[(T - \mathbb{E}[T])^2] + 2\mathbb{E}[(T - \mathbb{E}[T])(\mathbb{E}[T] - \theta)] + (\mathbb{E}[T] - \theta)^2. \end{aligned}$$

Noting

$$\mathbb{E}[(T - \mathbb{E}[T])^2] = Var(T),$$

$$2\mathbb{E}[(T - \mathbb{E}[T])(\mathbb{E}[T] - \theta)] = (\mathbb{E}[T] - \theta)2\mathbb{E}[(T - \mathbb{E}[T])] = 0$$

and

$$(\mathbb{E}[T] - \theta)^2 = (Bias(T; \theta))^2,$$

result follows. □

In words, the mean square error and the bias are inversely related. If one wishes to have a smaller mean square error, she has to sacrifice from the bias, and vice versa.

64 * Cramer-Rao lower bound

Suppose that $X_1, \ldots, X_n$ is a random sample from some distribution with cdf F and pdf f. Assume that the distribution relies on exactly one parameter θ, i.e., $F = F_\theta$ and $f = f_\theta$. Suppose further that the estimator $T(X_1, \ldots, X_n)$ is an unbiased estimator for θ:

$$\mathbb{E}[T(X_1, \ldots, X_n)] = \theta.$$

The Cramer-Rao inequality then gives a lower bound on the variance of T. Below we write f_θ, $\mathbb{E}_\theta$ and Var_θ in order to emphasize that the pdf, expectation or variance is computed for that specific θ.

Theorem 64.1 *(Cramer-Rao lower bound) Assume that $X_1, \ldots, X_n$ is a random sample from a common distribution with pdf f_θ, and that $T = T(X_1, \ldots, X_n)$ is an unbiased estimator for θ. Then*

$$Var(T(X_1, \ldots, X_n)) \geq \frac{1}{J(\theta)},$$

where

$$J(\theta) = \int_{-\infty}^{\infty} \left(\frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} \right)^2 f_\theta(x) dx = \mathbb{E}_\theta \left[\left(\frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} \right)^2 \right]$$

is the ***expected Fisher information****. We call $1/J(\theta)$ the* ***Cramer-Rao lower bound****.*

Remark 64.1 *Writing ℓ for the corresponding log-likelihood function, the expected Fisher information can also be written in the slightly more convenient form*

$$J(\theta) = \int_{-\infty}^{\infty} (\ell'(\theta \mid x))^2 f_\theta(x) dx = \mathbb{E}_\theta[(\ell'(\theta|X))^2].$$

Remark 64.2 *Obviously, an unbiased estimator with a small variance is preferred over another one with a large variance. If an unbiased estimator $T(X_1, \ldots, X_n)$ for θ satisfies*

$$Var(T(X_1, \ldots, X_n)) = \frac{1}{J(\theta)},$$

then T is said to be the ***best unbiased estimator*** *for θ.*

Proof: First observe that since $\mathbb{E}_\theta[T(X)] = \theta$, we have $\int_{-\infty}^{\infty} T(x) f_\theta(x) dx = \theta$. Also, since f_θ is a pdf, $\int_{-\infty}^{\infty} f_\theta(x) dx = 1$. Combining these, one arrives at

$$\int_{-\infty}^{\infty} (T(x) - \theta) f_\theta(x) dx = 0.$$

Taking derivatives on both sides with respect to θ - we avoid the technicalities here-, we get

$$\int_{-\infty}^{\infty} \left((T(x) - \theta) \frac{df_\theta(x)}{d\theta} - f_\theta(x) \right) dx = 0,$$

implying

$$\int_{-\infty}^{\infty}(T(x)-\theta)\frac{df_\theta(x)}{d\theta}dx = \int_{-\infty}^{\infty}(T(x)-\theta)\frac{\frac{df_\theta(x)}{d\theta}}{f_\theta(x)}f_\theta(x)dx = 1.$$

Using now Cauchy-Schwarz inequality gives

$$1 \leq \left(\int_{-\infty}^{\infty}(T(x)-\theta)^2 f_\theta(x)dx\right)\left(\int_{-\infty}^{\infty}\left(\frac{\frac{df_\theta(x)}{d\theta}}{f_\theta(x)}\right)^2 f_\theta(x)dx\right) = Var_\theta(T(X))J(\theta),$$

from which the result follows. □

65 Central limit theorem

In this section we discuss one of the most fundamental results of probability theory and statistics.

Theorem 65.1 *(Central limit theorem - CLT) Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and finite variance σ^2. Then for any fixed $x \in \mathbb{R}$*

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \leq x\right) = \Phi(x),$$

where $\Phi(x)$ is the cdf of a standard normal random variable.

In words, the central limit theorem says that for large n, the probabilities regarding $(\sum_{i=1}^n X_i - n\mu)/(\sqrt{n}\sigma)$ can be approximated by using the probabilities corresponding to a standard normal random variable. So for any $A \subset \mathbb{R}$, we have

$$\mathbb{P}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \in A\right) \approx \mathbb{P}(Z \in A),$$

where Z is standard normal.

The reason that this is considered as a **central** theorem is that there is no distributional assumption on X_i's except that they have finite second moments. X_i's could be binomial, exponential, and various other distributions we saw previously.

Here is a basic example on using CLT in applications.

Example 65.1 *If 10 fair dice are rolled independently, approximate the probability that the sums obtained will be between 30 and 40.*

Solution: Letting X_i be the number of spots we obtain in i^{th} roll for $i = 1, \ldots, 10$, the sum is given by $X = \sum_{i=1}^{10} X_i$. It is easy to see that $\mathbb{E}[X_i] = 7/2$ and $Var(X_i) = 35/12$ for each i - check these. So CLT gives

$$\begin{aligned}
\mathbb{P}\left(30 < X < 40\right) &= \mathbb{P}\left(\frac{30 - 10(7/2)}{\sqrt{10}\sqrt{35/12}} < \frac{X - 10(7/2)}{\sqrt{10}\sqrt{35/12}} < \frac{40 - 10(7/2)}{\sqrt{10}\sqrt{35/12}}\right) \\
&\approx \Phi(\sqrt{6/7}) - \Phi(\sqrt{-6/7}) \\
&\approx \Phi(0.925) - \Phi(-0.925) \\
&= 2\Phi(0.925) - 1 \\
&\approx 2(0.82) - 1 = 0.64,
\end{aligned}$$

where we read the value of $\Phi(0.925)$ from the z-table. We then conclude that the sum of these ten rolls will be somewhere between 30 and 40 is approximately 0.64. □

Of course one natural question here is the quality of the approximation we have. In particular, is the number of samples 10 enough to have a good normal approximation result. We will revisit this issue below when we discuss the Berry-Esseen inequality. Let's now see another example.

Example 65.2 *On each bet, a gambler loses 3 dollars with probability .8, or wins* 10 *dollars with probability .2. Assume that the bets are independent.*

(i) Use the central limit theorem to approximate the probability that the gambler will be winning after her first 36 bets.

(ii) Find an exact expression for the probability that the gambler will be winning after her first 36 bets.

Solution: Let X_i be the outcome of the i^{th} game so that $X_i \in \{-3, 10\}$. A quick computation gives $\mathbb{E}[X_1] = -0.4$ and $Var(X_1) = 27.04$ - check these yourself.

(i) Then the required probability is approximately

$$\mathbb{P}\left(\sum_{i=1}^{36} X_i > 0\right) \approx \mathbb{P}\left(Z > \frac{0 - (-0.4)36}{6\sqrt{27.04}}\right) = \mathbb{P}(Z > -0.46) = 0.677,$$

where Z is a standard normal random variable.

(ii) First note that $\sum_{i=1}^{36} X_i > 0$ if and only if she wins at least 9 games among these 36. So using binomial distribution, the exact answer is

$$\mathbb{P}(\text{at least 9 wins}) = \sum_{k=9}^{36} \binom{36}{k} (0.2)^k (0.8)^{36-k}.$$

Note here how easily we were able to obtain an approximate value for this *complicated* probability by using CLT. □

Lastly in this section we observe that

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} = \frac{\frac{\sum_{i=1}^n X_i - n\mu}{n}}{\frac{\sqrt{n}\sigma}{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

This provides an alternative formulation for the CLT in terms of the sample mean:

Corollary 65.1 *(CLT for the sample mean) Consider the setting in Theorem 65.1. Then for any $x \in \mathbb{R}$*

$$\mathbb{P}\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq x\right) \longrightarrow \Phi(x), \quad n \to \infty.$$

66 Some practical issues

In this section, we collect some practical info about the central limit theorem.

1. It is an important question whether the approximation provided by the CLT is good or not. In practical situations, people usually say that a sample of size at least 25 will provide a good approximation.

2. When the underlying distribution is symmetric, continuous and unimodal, the convergence to the normal in CLT will be faster.

3. Let's state one rigorous theorem related to convergence rates in CLT.

   Theorem 66.1 *(Berry-Esseen theorem) Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ, variance $\sigma^2 < \infty$ and a finite third moment. Then, letting Z be a standard normal,*

   $$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \leq x\right) - \mathbb{P}(Z \leq x) \right| \leq \frac{C}{\sqrt{n}}.$$

   Here C is a constant independent of n that depends on the third moment of X_1.[56]

   Proof of such theorems are unfortunately out of reach at this level.

[56] The distance on the left-hand side between probability measures is known to be the **Kolmogorov distance** between the given random variables.

4. **Continuity correction.** A continuity correction factor is used when you use a continuous function to approximate a discrete one. For instance, in our rolling dice example in previous section we approximated the distribution of the sum of 10 die rolls, which is discrete, via the normal distribution. When you use the normal distribution to approximate a discrete one, you are going to have to use a continuity correction factor. This is just adding or subtracting .5 to the discrete x-value depending on whether x is in upper bound or lower bound, respectively.

Example 66.1 *Suppose a fair coin is tossed 20 times and all tosses are independent. Approximate the probability that exactly 10 heads are obtained via using the CLT.*

Solution: Letting X be the number of heads we obtain in 20 trials. Then

$$\mathbb{P}(X = 10) = \mathbb{P}(9.5 \leq X \leq 10.5) \approx \mathbb{P}\left(\frac{-.5}{2.236} \leq Z \leq \frac{.5}{2.236}\right),$$

where Z is a standard normal random variable. Hence an approximation can be given by the value $\Phi(0.2236) - \Phi(-0.2236) \approx 0.177$.

Similarly if we were interested in approximating the probability $\mathbb{P}(5 \leq X \leq 14)$, then we would compute $\mathbb{P}(4.5 \leq X \leq 14.5)$ for the continuity correction. □

67 * A bit of fun: Stirling approximation with CLT and Poisson approximation

Let $X_1, X_2, ...$ be a sequence of independent Poisson random variables with mean 1. Set

$$S_n = \sum_{j=1}^{n} X_j.$$

We already know that S_n is a Poisson random variable with mean n, and so $\mathbb{E}[S_n] = n$ and $Var(S_n) = n$. Letting Z be a standard normal random variable, we have

$$\begin{aligned}
\mathbb{P}(S_n = n) = \mathbb{P}(n - 1 < S_n \leq n) &= \mathbb{P}\left(\frac{-1}{\sqrt{n}} < \frac{S_n - n}{\sqrt{n}} \leq 0\right) && (14)\\
&\approx \mathbb{P}\left(\frac{-1}{\sqrt{n}} < Z \leq 0\right) && (15)\\
&= \int_{-\frac{1}{\sqrt{n}}}^{0} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx && (16)\\
&\approx \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n}}. && (17)
\end{aligned}$$

On the other hand, since $\sum_{j=1}^{n} X_j$ is a Poisson random variable with parameter n,

$$\mathbb{P}(S_n = n) = \frac{e^{-n} n^n}{n!}. \tag{18}$$

By comparing the two expressions in (14) and (18) for $\mathbb{P}(S_n = n)$, we arrive at

$$\frac{e^{-n} n^n}{n!} \approx \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n}}.$$

This give us the Stirling's formula:

Stirling's formula:

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n, \qquad \text{as } n \to \infty.$$

Note here that $n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$ just means

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n} = 1$$

as before. In other words, for large n,

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

Here is a related - maybe challenging - exercise for you to think about.

Exercise 67.1 *Let $X_1, X_2, \ldots$ be a sequence of independent random variables where each X_i has exponential distribution with parameter 1.*

(i) Show that for any $x \in \mathbb{R}$

$$\mathbb{P}\left(\frac{\overline{X}_n - 1}{\frac{1}{\sqrt{n}}} \leq x\right) \longrightarrow \mathbb{P}(Z \leq x),$$

where Z is a standard normal random variable.

(ii) By using the result in part i., show that for any $x \in \mathbb{R}$

$$\lim_{n\to\infty} \frac{\frac{\sqrt{n}(n+\sqrt{n}x)^{n-1} e^{-(n+\sqrt{n}x)}}{(n-1)!}}{\frac{1}{\sqrt{2\pi}} e^{-x^2/2}} = 1.$$

(Hint : The density function of $Y = \sum_{i=1}^{n} X_i$ is $f(y) = \frac{y^{n-1} e^{-y}}{(n-1)!}$, $y > 0$. You can use this without proof.)

(iii) Use part ii. with $x = 0$ to prove Stirling's formula: $\lim_{n\to\infty} \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{n!} = 1$.

68 Confidence intervals for the mean I

In Section 57, we had our first discussion on interval estimation where we had normally distributed samples $X_1, \ldots, X_n$ with parameters μ and σ^2. Our result in that section was that the interval

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

is a $(1-\alpha)100\%$ confidence interval for the mean μ when σ is known. Our goal here is to understand the following question:

Question: We have i.i.d. samples $X_1, \ldots, X_n$ from some distribution, not necessarily normal, with mean μ and known variance σ^2. How can we obtain a $(1-\alpha)100\%$ confidence interval for the mean μ?

Observe that the only difference from the earlier setting is that the underlying distribution is not assumed to be normal this time. In order to handle this we will make use of the central limit theorem which tells us that the ratio $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is approximately standard normal (when n is large). Fixing $\alpha \in (0,1)$ and following exactly the same steps in Section 57 except the approximation due to CLT, we have

$$\begin{aligned} 1-\alpha &\approx \mathbb{P}\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}\right) \\ &= \mathbb{P}\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right). \end{aligned}$$

Therefore the random interval

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

is still a $(1-\alpha)100\%$ confidence interval for the mean, but this time an **approximate** one. Here the approximation will be *good* if the sample size is large enough so that the distribution of $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is close to a standard normal. In next section we will see a graphical tool to check closeness to normal distribution. Also remember our relevant discussions in Section 66.

Here is an example.

Example 68.1 *The average zinc concentration recovered from a random sample of zinc measurements in 36 different locations in a river is found to be 2.6 grams per milimeter.*

(i) Assuming that the population standard deviation is 0.03, find the 95% *and* 99% *confidence intervals for the mean zinc concentration in the river.*

(ii) How large a sample is required if we want to be 95% *confident that our estimate of* μ *is off by less than 0.05?*

Solution. (i) We have a sample size of 36. We assume that the measurements are i.i.d., but not necessarily normal. Noting that $z_{0.025} = 1.96$, the 95% confidence interval turns out to be

$$\left[2.6 - (1.96)\frac{0.3}{\sqrt{36}}, 2.6 + (1.96)\frac{0.3}{\sqrt{36}}\right] = [2.50, 2.70].$$

Similarly, reading $z_{0.005} = 2.575$ from the z-table, the 95% confidence interval is

$$\left[2.6 - (2.575)\frac{0.3}{\sqrt{36}}, 2.6 + (2.575)\frac{0.3}{\sqrt{36}}\right] = [2.47, 2.73].$$

99% confidence interval is longer as expected. Note that both of these confidence intervals are approximate since there is the underlying central limit theorem process going on.

(ii) When $\overline{x}$ is used as an estimator of μ, we are $(1-\alpha)100\%$ confident that the error $|\overline{x} - \mu|$ will not exceed $z_{\alpha/2}\sigma/\sqrt{n}$. We are now willing to find some n so that we are 95% confident that the error we make is less than 0.05. This will be the case if $z_{\alpha/2}\sigma/\sqrt{n} < 0.05$ which is equivalent to

$$n > \left(\frac{z_{\alpha/2}\sigma}{0.05}\right)^2.$$

In our case $\alpha = 0.5$ and $\sigma = 0.3$. Using these, we conclude that n should at least be 138.3 meaning that we should have at least 139 random samples.

□

69 QQ-plots

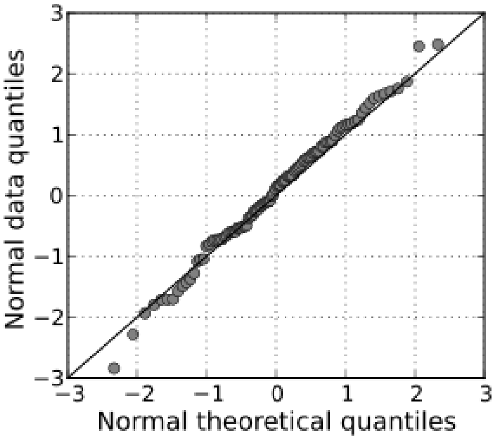
Let's begin by recalling that the p-th percentile of the distribution of a random variable X is defined to be a real number q_p that satisfies

$$\mathbb{P}(X \leq q_p) \geq p, \qquad \text{and} \qquad \mathbb{P}(X \geq q_p) \geq 1 - p.$$

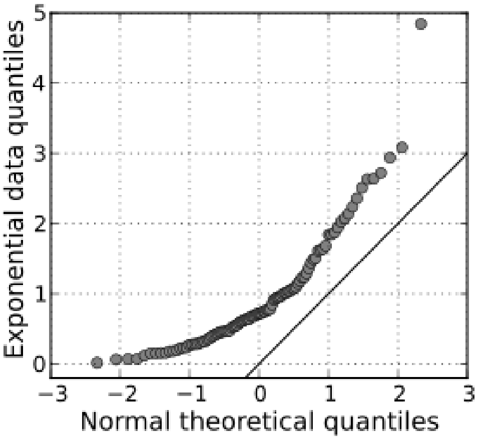
When $p = 0.25$, $p = 0.75$ and $p = 0.5$, the percentiles were called as the (distributional) first quantile, third quantile and median, respectively. Also, considering the sample versions, the p^{th} percentile was obtained by first computing the ordinal rank and then taking the value from the ordered list that corresponds to this rank. The sample first quantile, and so on were defined accordingly.

Sample quantiles can be used to compare two populations and to see whether they come from the same distribution or not. In particular, we may use the quantiles to decide (more correctly: to have an idea) whether a population is normally distributed or not. For the general case, we consider two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ from two populations. Assuming that their underlying distribution is common, the order statistics $X_{(1)}, \ldots, X_{(n)}$ and $Y_{(1)}, \ldots, Y_{(n)}$ would provide estimates for the same distributional quantile values.

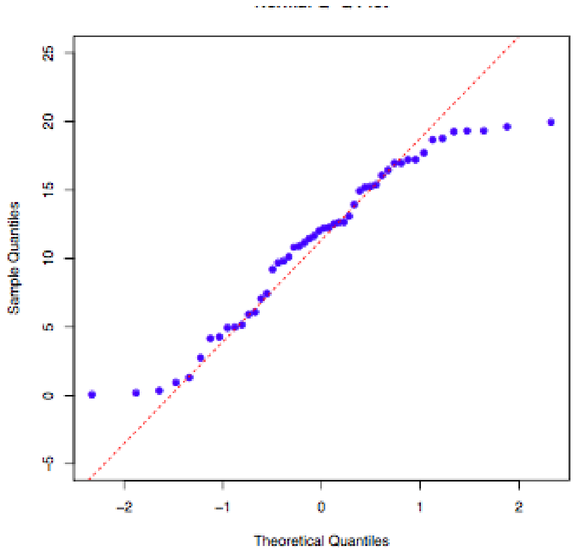
Now, the quantile-quantile plot, or QQ plot, is a graphical tool to compare the quantiles of two distributions. The following QQ plot borrowed from Wikipedia is comparing the sample quantiles of populations drawn from the standard normal distribution independently.



For a QQ-plot corresponding to two distinct populations, let's consider the case where the first population is generated from the standard normal distribution and the second one is generated from the exponential distribution with parameter 1 (Reference is again Wikipedia.)



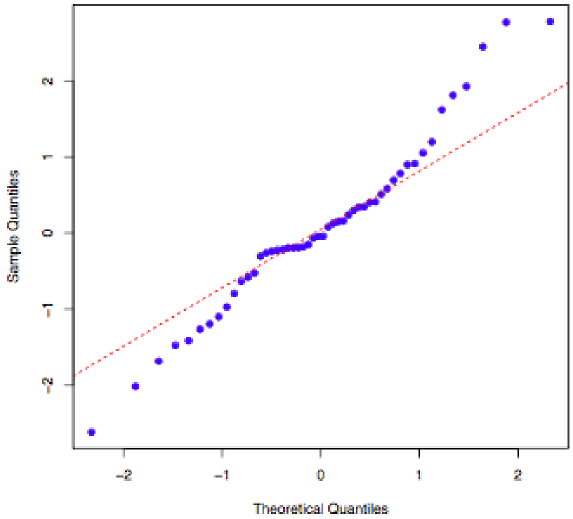
The plot here is non-linear. From this even if we did not know that the second population was drawn from the exponential distribution, we could easily see that its distribution is not the normal one. As already mentioned, this is one important use of the QQ-plots since the assumption that the underlying data is normal is crucial in various statistical arguments. As our last two examples, we will have a look at the QQ plots corresponding to quantiles of some sample distribution vs. the quantiles of the standard normal distribution[57]. The first one is



In this case the sampling distribution is quite different than the normal distribution. In particular, the sampling distribution's tails have less mass compared to the normal distribution.

The next plot is similar to the previous one with the relation in tails reversed. This time the sampling distribution has more mass in the tails.

[57] These figures are borrowed from http://people.reed.edu/~jones/Courses/P14.pdf.



Here is a summary for QQ plots with some rule of thumbs to keep in mind:

1. When the sample size is small we should not expect a good straight plot even for the normal vs. normal case.

2. If the two distributions agree after linearly transforming the values in one of the distributions, then the QQ plot follows some line, but not necessarily the line $y = x$.

3. If the general trend of the QQ plot is flatter than the line $y = x$, the distribution plotted on the horizontal axis is more dispersed than the distribution plotted on the vertical axis.

4. If the general trend of the QQ plot is steeper than the line $y = x$, the distribution plotted on the vertical axis is more dispersed than the distribution plotted on the horizontal axis. QQ plots are often arced, or "S" shaped, indicating that one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other.

5. In summary: QQ plots tell us differences in location, spread and shape of two distributions.

70 Review problems VI

Exercise 70.1 *Suppose that a pair of balanced dice are rolled 120 times, and let X denote the number of rolls on which the sum of the two numbers is 7. Use the central limit theorem to determine a value of k such that $\mathbb{P}(|X - 20| \le k)$ is approximately 0.95.*

Exercise 70.2 *If $\overline{X}_1$ and $\overline{X}_2$ are the means of two independent samples of size n from a population with variance σ^2, find a value for n so that $\mathbb{P}(|\overline{X}_1 - \overline{X}_2| < \sigma/5) \approx .99$. Justify your calculations.*

Exercise 70.3 *In 1000 tosses of a coin, 560 heads and 440 tails appear. Is it reasonable to assume that the coin is fair? Justify your answer.*

Exercise 70.4 *Suppose that people attending a party pour drinks from a bottle containing 63 ounces of a certain liquid. Suppose also that the expected size of each drink is 2 ounces, that the standard deviation of each drink is 1/2 ounce, and that all drinks are poured independently. Approximate the probability that the bottle will not be empty after 36 drinks have been poured.*

Exercise 70.5 *Let $X_1, \ldots, X_n$ be i.i.d. with pdf*

$$f(x \mid \theta) = \theta x^{\theta - 1}, \quad 0 \leq x \leq 1,\ 0 < \theta < \infty.$$

Find the MLE of θ, and show that its variance tends to zero as $n \to \infty$.

Exercise 70.6 *Let $X_1, \ldots, X_n$ be i.i.d. from a population with pmf*

$$\mathbb{P}_\theta(X = x) = \theta^x (1 - \theta)^{1-x}, \quad x = 0 \text{ or } 1,\ 0 \leq \theta \leq \frac{1}{2}.$$

Find the MLE of θ.

Exercise 70.7 *Let $X_1, \ldots, X_n$ be a random sample from a population with pdf*

$$f(x \mid \theta) = \frac{1}{2\theta}, \quad -\theta < x < \theta, \quad \theta > 0.$$

Find, if one exists, a best unbiased estimator of θ.

Exercise 70.8 *Let $X_1, \ldots, X_n$ be i.i.d. random variables with Bernoulli distribution with parameter p. Show that the variance of $\overline{X}$ attains the Cramer-Rao lower bound, and hence $\overline{X}$ is the best unbiased estimator of p.*

Exercise 70.9 *Let X_1, X_2, X_3 be a random sample of size three from a uniform distribution on $(\theta, 2\theta)$ where $\theta > 0$.*

(i) Find the MLE, $\hat{\theta}$ and find a constant k such that $\mathbb{E}[k\hat{\theta}] = \theta$.

(ii) Find the maximum likelihood estimate if the data you have is $1.29, 0.86$ and 1.33.

Exercise 70.10 *Let $X_1, \ldots, X_n$ be i.i.d. exponential random variables with parameter λ.*

(i) Find an unbiased estimator of λ based only on $Y = \min\{X_1, \ldots, X_n\}$.

(ii) Find an unbiased estimator better than the one in part a. Prove that it is better.

Exercise 70.11 *Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. random samples from the uniform distribution on $(0, \theta)$ where $\theta > 0$ is some unknown parameter. Some possible estimators for θ are*

1. $T_1 = 2\overline{X}$
2. $T_2 = \max\{X_1, \ldots, X_n\}$
3. $T_3 = \frac{n+1}{n} \max\{X_1, \ldots, X_n\}$
4. $T_4 = 2Median(X_1, \ldots, X_n)$

It should be intuitively clear that each of these estimators will be close to the value of θ when n is large. Analyze each of these estimators in terms of the properties we have learnt so far.

Exercise 70.12 *Suppose that the random variables $Y_1, \ldots, Y_n$ satisfy*

$$Y_i = \beta x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $x_1, \ldots, x_n$ are fixed constants, and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables with σ^2 unknown. Find the MLE of β, and show that it is an unbiased estimator of β.

Exercise 70.13 *Consider the setting in previous problem.*
(i) Show that $\sum Y_i / \sum x_i$ is an unbiased estimator of β.
(ii) Compute the variance of $\sum Y_i / \sum x_i$ and compare it to the variance of the MLE you have found in previous problem.

Exercise 70.14 *Consider again the setting two problems above.*
(i) Show that $(\sum (Y_i/x_i))$ is also an unbiased estimator of β.
(ii) Compute the variance of $(\sum (Y_i/x_i))$ and compare it with the variances in previous two problems.

71 Least squares method

Given some data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, what would be the best fitting line $y = \beta_0 + \beta_1 x$ for this data? In order to answer this question, we should first properly define what we mean by "best". Our goal here will be to minimize the function

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

and we will call the corresponding line to be the best fitting line. Note that there is no randomness here.

In order to minimize f, we set the partial derivatives to 0 to get

$$\frac{\partial f}{\partial \beta_0} = \sum_{i=1}^{n} 2(-1)(y_i - \beta_0 - \beta_1 x_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0,$$

and

$$\frac{\partial f}{\partial \beta_1} = \sum_{i=1}^{n} 2(-x_i)(y_i - \beta_0 - \beta_1 x_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0.$$

Solving these (and, say, by using a second derivative test) we obtain the best fitting line as

$$y = \beta_0 + \beta_1 x,$$

where

$$\beta_0 = \frac{\overline{y}\left(\sum_{i=1}^{n} x_i^2\right) - \overline{x} \sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2},$$

and

$$\beta_1 = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}.$$

This is nice, but it doesn't say anything whether a prediction based on x_{n+1} via this line is good or not.

72 Regression-motivating discussion

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

One variable, denoted x, is regarded as the *input, predictor, explanatory*, or *independent* variable. The other variable, denoted y, is regarded as the *output, response, outcome*, or *dependent* variable. Simple linear regression gets its adjective "simple", because it concerns the study of only one predictor variable. In contrast, multiple linear regression, which we will not study in this course, gets its adjective "multiple", because it concerns the study of two or more predictor variables. Below, X will be used for a predictor variable and Y for a response variable.

Our **probabilistic regression model** here will be the following:

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

where ϵ_i's are independent normal random variables with mean zero and variance σ^2.

Our eventual goal here is to find a prediction y for some given predictor value x. But this first requires the estimation of three parameters we have: α, β and σ^2. We will do this by using the likelihood method in the next section.

Though we will just focus on simple linear regression here, let's note that there are various other models which you may try to fit depending on the data you have. Two other examples could be:

- (Quadratic regression) $Y = \beta_0 + \beta_1 x + \beta_2 x^2$, and the goal is to estimate β_0, β_1 and β_2,
- (Exponential regression) $Y = \beta_0 e^{\beta_1 x}$, and the goal is to estimate β_0 and β_1.

One nice thing here to observe is that we can linearize certain non-linear models to obtain "approximate" predictions. For example, consider the exponential regression where

$$Y = \beta_0 e^{\beta_1 x},$$

with β_0 and β_1 certain constants that are to be estimated. In this case, taking ln on both sides we have

$$\ln Y = \ln \beta_0 + \beta_1 \ln x.$$

So, defining new variables $Y^* = \ln Y$, $x^* = \ln x$, $\alpha_0 = \ln \beta_0$ and $\alpha_1 = \beta_1$, we obtain the linear problem

$$Y^* = \alpha_0 + \alpha_1 x^*$$

which we will be able to understand. After this, one can try to turn the information gained about α_0 and α_1 to information on β_0 and β_1.

73 Simple linear regression

Recall that the simple linear regression model we have is:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i's are independent normal random variables with mean zero and variance σ^2. We are now ready to solve the following question.

Question: How do we find the best fitting line among all those infinitely many options?

In order to answer this, we will find the maximum likelihood estimators of β_0 and β_1. Assume that we are given the data points: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ (e.g., x_i is height of

ith person and y_i is the weight of ith person.). The joint likelihood function of the parameters $(\beta_0, \beta_1, \sigma^2)$ is

$$\begin{aligned} L(\beta_0, \beta_1, \sigma) &= \prod_{i=1}^{n} f_{Y_i}(y_i \mid \beta_0, \beta_1, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(\frac{-(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right), \end{aligned}$$

and the log-likelihood function as

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma} \right)^2 .$$

To maximize ℓ, we first find the critical points. We have

$$\frac{\partial \ell}{\partial \beta_0} = 0 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(-1)2(y_i - \beta_0 - \beta_1 x_i),$$

and $\partial\ell/\partial\beta_0 = 0$ implies

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i. \tag{19}$$

Next, we observe that

$$\frac{\partial \ell}{\partial \beta_1} = 0 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(-x_i)2(y_i - \beta_0 - \beta_1 x_i) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i y_i - \beta_0 x_i - \beta_1 x_i^2),$$

and $\partial\ell/\partial\beta_1 = 0$ implies

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{20}$$

Solving (19) and (20) for β_0 and β_1 gives the corresponding maximum likelihood estimators, respectively, as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n},$$

and

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

Conclusion:

Theorem 73.1 *The best fitting line by using the maximum likelihood estimators for β_0 and β_1 is given by*

$$\mu(x) = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)/n}{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2/n} \qquad \textit{and} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

Here, some elementary manipulations can be used to show that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)/n}{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2/n} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} =: \frac{S_{xy}}{S_{xx}}.$$

Note that S_{xx} is the sample variance of the input data $\{x_1, \ldots, x_n\}$. Also, $S_{xy} = \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$ is known to be the **sample covariance**.

Once we have our $\hat{\beta}_0, \hat{\beta}_1$,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

is the value corresponding to x_i according to our model. The quantity

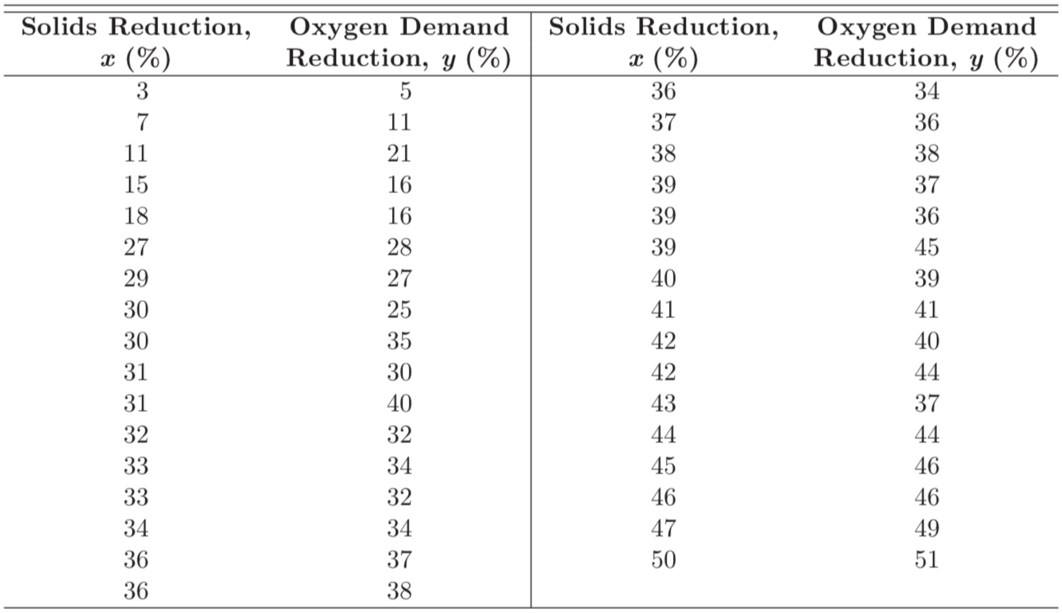
$$e_i = y_i - \hat{y}_i$$

is called the ith **residual** - or error. In general, when we have a good fit the set of residuals $\{e_1, \ldots, e_n\}$ will like look like uniformly distributed around the $y = 0$ line. Having the regression model, we may now also do **predictions** for $x \notin \{x_1, \ldots, x_n\}$; namely

$$\mu(x) = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Here is an example.

Example 73.1 *(Myers, Myers, Walpole) Consider the experimental data in following table, which were obtained from 33 samples of chemically treated waste in a study conducted at Virginia Tech. Readings on x, the percent reduction in total solids, and y, the percent reduction in chemical oxygen demand, were recorded.*



Solids Reduction, x (%)	**Oxygen Demand Reduction, y (%)**	**Solids Reduction, x (%)**	**Oxygen Demand Reduction, y (%)**
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

Figure 10: The table is borrowed from Myers, Myers, Walpole as well

For this example, we compute

$$\sum_{i=1}^{33} x_i = 1104, \quad \sum_{i=1}^{33} y_i = 1124, \quad \sum_{i=1}^{33} x_i y_i = 41,355, \quad \sum_{i=1}^{33} x_i^2 = 41,086.$$

Using these, we find the estimates for β_1 and β_0:

$$\hat{\beta_1} = \frac{(33)(41,355) - (1104)(1124)}{33(41,086) - (1104)^2} = 0.9036$$

and

$$\hat{\beta_0} = \frac{1124 - (0.9036)(1104)}{33} = 3.8296.$$

Therefore, the estimated regression line is given by

$$\hat{y} = 3.8296 + (0.9036)x.$$

For example, the predictions via the regression for $x = 3$ and $x = 50$ turn out to be approximately $\hat{y} = 6.5$ and $\hat{y} = 49$, respectively. □

74 Properties of likelihood estimators in simple linear regression

Consider the simple linear regression model of previous section. We have found

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

and

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

as the parameter estimates given the data points (x_i, y_i), $i = 1, \ldots, n$. Now we treat $\hat{\beta}_i$, $i = 0, 1$ as estimators by replacing y_i's by the random quantities Y_i. Our goal is to find unbiased estimators for β_0 and β_1. Let's begin with the latter.

Once we replace y_i by Y_i and $\overline{y}$ by $\overline{Y}$, we claim that $\hat{\beta}_1$ is an unbiased estimator for β_1. For this purpose observe first that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i - \sum_{i=1}^n (x_i - \overline{x})\overline{Y}}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

So

$$\hat{\beta}_1 = \sum_{i=1}^n c_i Y_i,$$

where

$$c_i = \frac{x_i - \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{x_i - \overline{x}}{s_{xx}}.$$

Then

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_1] = \sum_{i=1}^n c_i \mathbb{E}[Y_i] &= \sum_{i=1}^n \frac{x_i - \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \mathbb{E}[\beta_0 + \beta_1 x_i + \epsilon_i] \\
&= \sum_{i=1}^n \frac{x_i - \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} (\beta_0 + \beta_1 x_i) \\
&= \frac{\beta_0}{\sum_{i=1}^n (x_i - \overline{x})^2} \sum_{i=1}^n (x_i - \overline{x}) + \beta_1 \frac{\sum_{i=1}^n (x_i - \overline{x}) x_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= 0 + \beta_1 \frac{\sum_{i=1}^n (x_i - \overline{x})(x_i - \overline{x} + \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \beta_1 \frac{\sum_{i=1}^n (x_i - \overline{x})^2 + \overline{x} \sum_{i=1}^n (x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \beta_1 \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \beta_1.
\end{aligned}$$

So $\hat{\beta}_1$ is really an unbiased estimator for β_1. Also observe that

$$\begin{aligned} Var(\hat{\beta}_1) = Var\left(\frac{\sum_{i=1}^n (x_i-\overline{x})Y_i}{\sum_{i=1}^n (x_i-\overline{x})^2}\right) = \frac{\sum_{i=1}^n (x_i-\overline{x})^2 Var(Y_i)}{\left(\sum_{i=1}^n (x_i-\overline{x})^2\right)^2} &= \frac{Var(\beta_0+\beta_1 x_i+\epsilon_i)\sum_{i=1}^n (x_i-\overline{x})^2}{\left(\sum_{i=1}^n (x_i-\overline{x})^2\right)^2} \\ &= \frac{\sigma^2\sum_{i=1}^n (x_i-\overline{x})^2}{\left(\sum_{i=1}^n (x_i-\overline{x})^2\right)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_i-\overline{x})^2}. \end{aligned}$$

Let's summarize what we have proven so far.

Theorem 74.1 *Consider the simple linear regression model. Then $\hat{\beta}_1$ is an unbiased estimator for β_1,*

$$\mathbb{E}[\hat{\beta}_1] = \beta_1$$

and

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i-\overline{x})^2}.$$

How about $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{x}$. We have

$$\begin{aligned} \mathbb{E}[\hat{\beta}_0] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n Y_i\right] - \overline{x}\mathbb{E}[\hat{\beta}_1] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i] - \beta_1\overline{x} \\ &= \frac{1}{n}\sum_{i=1}^n (\beta_0+\beta_1 x_i) - \beta_1\overline{x} \\ &= \beta_0 + \beta_1\overline{x} - \beta_1\overline{x} \\ &= \beta_0. \end{aligned}$$

That is, $\hat{\beta}_0$ is an unbiased estimator for β_0 as well. I leave it as an exercise for you to show that

$$Var(\hat{\beta}_0) = \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\overline{x})^2}\sigma^2.$$

Here is a summary

- $\hat{\beta}_1$ is an unbiased estimator for β_1 and $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i-\overline{x})^2}$;
- $\hat{\beta}_0$ is an unbiased estimator for β_0 and $Var(\hat{\beta}_0) = \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\overline{x})^2}\sigma^2$;

Note that we haven't discussed the estimation of σ^2. Here is a related exercise.

Exercise 74.1 *Consider the setting in our setting in our simple linear regression model. Find the maximum likelihood estimator for* σ^2.

Now given this probabilistic setting, we may further find confidence intervals for the parameters, do hypothesis testing for them, etc. This will be briefly discussed in Section 92.

75 A story: Deterministic vs. stochastic models

In previous three sections, we had one deterministic and one stochastic version of line fitting. Is it always advantageous to use a stochastic model in our problems? Of course, the answer is no. The purpose of this section is to refer to you an excellent discussion about this where Tony Starfield advocates that people sometimes use stochastic models in some unnecessary cases making things complicated with no benefits [14]. However, he also gives an excellent story on his research where the right use of a statistical model can be crucial. I will quote and quote this story directly from the cited source. I encourage you to read both [14] and [15] completely.

" Many years ago I was working on a lion population model in South Africa, and we were simulating lions on an individual basis with a lot of social behavior, and that social behavior included what are called takeovers. You would perhaps have two or three pride lions in control of a pride, and you might have four or five younger nomadic lions wandering around looking for a pride to breed with. And we modeled whether or not they had a fight, and if they did have a fight, who won the fight. And the reason this was important in the model was because when young lions take over from older lions, they kill their cubs so that they can breed with the females as quickly as possible. So this was an important part of the population dynamics.

When we ran the model, some information that we had that we hadn't built into the model was that on average a pride would have - a group of males would have tenure with a pride for about two to three years before they were ousted. When we ran our model, we discovered that the pride males were in control for eight or nine years. So clearly something was wrong with the model.

We did a sensitivity analysis. We fiddled with the parameters. We couldn't get that number down. And then we realized something - we had modeled the sexing of lion cubs as a deterministic process. So, for example, if a litter of 10 cubs, or total number of 10 cubs were born in a particular year, we said five of them were male and five of them were female.

If we made that stochastic, then occasionally all 10 cubs would turn out to be male. Or occasionally you would have eight or nine cubs turning out to be male. Those occasions produced strong coalitions of young males, and those strong coalitions of young males were able more readily to displace older males in control of prides.

So by merely changing one little component of the model from deterministic to stochastic we saved the model and got realistic results. These are the things one has to learn by trial and error. "

Let me conclude this brief section with another quote again from [14]: "Remember our heuristic - was never build a stochastic model unless there is a real need to."

76 Chi-squared distribution

Definition 76.1 *X is said to have* ***chi-squared distribution with* r *degrees of freedom*** *if its pdf is given by*

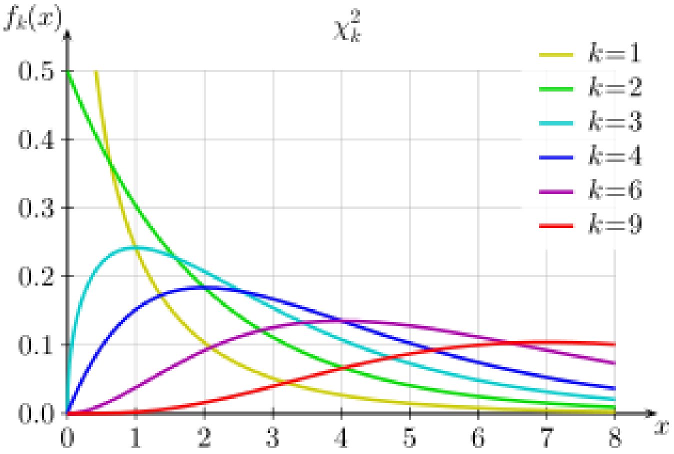
$$f(x|r) = \frac{1}{2^{r/2}\Gamma(r/2)} x^{r/2-1} e^{-x/2}, \qquad x > 0,$$

where

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

is the gamma function. When X has chi squared distribution with parameter r, we write $X \sim \chi^2(r)$.

The following figure borrowed from Wikipedia shows the pdf of chi-squared distribution for different values of the degrees of freedom - k in the figure.



As in the case of normal distribution, we read probabilities related to chi-squared distribution of the form

$$F(x) = \int_0^x \frac{1}{2^{r/2}\Gamma(r/2)} y^{r/2-1} e^{-y/2} dy$$

from the relevant table. You can find a copy of the χ^2-table at the end of these lecture notes.

We now state two useful facts about the chi-squared distribution without proof. The proofs are not hard, but they require some preliminaries and also are secondary for the purpose of this course.

Theorem 76.1 *If $U_1, \ldots, U_n$ are independent random variables, and $U_i \sim \chi^2(r_i)$, $i = 1, \ldots, n$, then*

$$U_1 + \cdots + U_n \sim \chi^2(r_1 + \cdots + r_n).$$

Theorem 76.2 *If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2(1)$.*

Now, letting $X_1, \ldots, X_n$ be independent random variables with $X_i \sim N(\mu_i, \sigma_i^2)$, we know that

$$\frac{X_i - \mu_i}{\sigma_i} \sim N(0,1).$$

Therefore combining previous two results, we obtain

Theorem 76.3 *If $X_1, \ldots, X_n$ be independent random variables with $X_i \sim N(\mu_i, \sigma_i^2)$, then*

$$\sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \sim \chi^2(n).$$

An immediate corollary is:

Corollary 76.1 *If $X_1, \ldots, X_n$ are i.i.d. normal random variables with parameters μ and σ^2, then*

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n).$$

Now let's assume that we don't know the value of μ, and we estimate it with $\overline{X}$. A natural question is the distribution of $\sum_{i=1}^n \left(\frac{X_i - \overline{X}}{\sigma} \right)^2$.

Theorem 76.4 *If $X_1, \ldots, X_n$ are i.i.d. normal random variables with parameters μ and σ^2, then*

$$\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 \sim \chi^2(n-1).$$

The proof of this result is again omitted

Note that the last statement says that "we lost" one degree of freedom by replacing a parameter by its estimate. In physics, the number of degrees of freedom is the minimum number of independent coordinates that can specify the position of the system completely. As an example, given that $\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 = c$, we need $(n-1)$-coordinates to determine all the coordinates.

As a final note in this section, recalling that $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$, we may observe that the conclusion of last theorem can be restated as

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1). \tag{21}$$

77 Confidence intervals for the variance

Setting of the section: We are given i.i.d. samples $X_1, \ldots, X_n$ from $N(\mu, \sigma^2)$, and we don't know the value of neither μ nor σ^2.

Question: How can we obtain a confidence interval for σ^2?

In order to answer this question first recall from (21) that

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1).$$

Also, for given $\beta \in (0,1)$, let $\chi^2_\beta(n-1) \in \mathbb{R}^+$ be so that a chi squared random variable with $n-1$ degrees of freedom has probability β for being at least $\chi^2_\beta(n-1)$.

Observe now that for $\alpha \in (0,1)$ we have

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1)\right) \\ &= \mathbb{P}\left(\frac{1}{(n-1)S^2}\chi^2_{1-\alpha/2}(n-1) < \frac{1}{\sigma^2} < \frac{1}{(n-1)S^2}\chi^2_{\alpha/2}(n-1)\right) \\ &= \mathbb{P}\left(\frac{1}{\chi^2_{\alpha/2}(n-1)}(n-1)S^2 < \sigma^2 < \frac{1}{\chi^2_{1-\alpha/2}(n-1)}(n-1)S^2\right). \end{aligned}$$

Therefore, we obtain the following result.

Theorem 77.1 *If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables, where σ is unknown, and $\alpha \in (0,1)$, then*

$$\left[\frac{1}{\chi^2_{\alpha/2}(n-1)}(n-1)S^2, \frac{1}{\chi^2_{1-\alpha/2}(n-1)}(n-1)S^2\right]$$

is a $(1-\alpha)100\%$ confidence interval for σ^2.

Remark 77.1 *(i) Note that the confidence interval of previous statement is exact. If the underlying random variables were not normal, then we could employ central limit theorem and obtain an approximate confidence interval.*

(ii) The same discussion provides a $(1-\alpha)100\%$ confidence interval for the population standard deviation σ as

$$\left[\sqrt{\frac{1}{\chi^2_{\alpha/2}(n-1)}(n-1)S^2}, \sqrt{\frac{1}{\chi^2_{1-\alpha/2}(n-1)}(n-1)S^2}\right]$$

Example 77.1 *In a study on cholesterol levels a sample of 12 men and women were chosen. The plasma cholesterol levels were as follows:*

$$6, 6.4, 7, 5.8, 6, 5.8, 5.9, 6.7, 6.1, 6.5, 6.3, 5.8.$$

Assume that these samples are independent and are drawn from a normal distribution. Find a 95% confidence interval for σ^2.

Solution: We have

$$S^2 = \frac{1}{11}\sum_{i=1}^{12}(x_i - \overline{x})^2 = 0.391.$$

Also from the χ^2-table we read

$$\chi^2_{0.025}(11) = 21.920, \qquad \text{and} \qquad \chi^2_{0.975}(11) = 3.816.$$

Hence the 95% confidence interval for σ^2 is given by

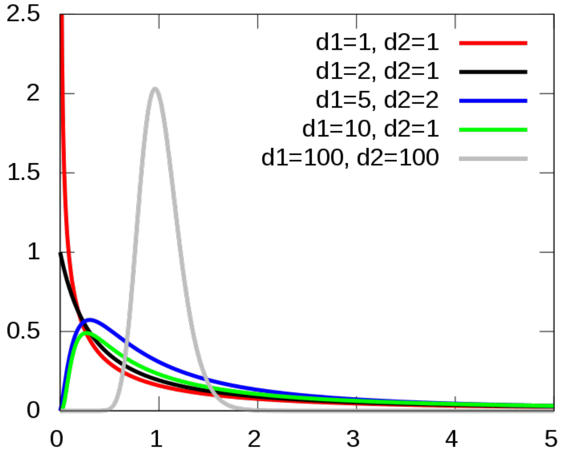
$$\left[\frac{1}{21.920}(11)(0.391), \frac{1}{3.816}(11)(0.391)\right] = [0.1966, 1.129].$$

□

78 F distribution and ratio of variances of two populations

Definition 78.1 *A random variable of the form $F = \frac{U_1/r_1}{U_2/r_2}$ is said to have **Fisher's F distribution** with parameters r_1 and r_2 if U_1, U_2 are independent and have respective distributions $\chi^2(r_1)$ and $\chi^2(r_2)$.*

The following figure shows the pdf of Fisher distribution for various choices of the parameters (Reference for the figure: Wikipedia).



When a random variable X has Fisher's distribution with parameters r_1 and r_2, we write $X \sim F(r_1, r_2)$. Probabilities related to F distribution are again read from the corresponding table which can be found at the end of the lecture notes[58].

[58] This one is different than the previous ones since we have two parameters. Make sure that you are okay with it.

Now, the purpose in this section is to understand whether two distinct populations have the same variation around their mean or not. For this, let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent random samples from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively. Then we claim that $\frac{S_X^2 \sigma_Y^2}{S_Y^2 \sigma_X^2}$ has F distribution with parameters $n-1$ and $m-1$. To see that this is the case observe

$$\frac{S_X^2 \sigma_Y^2}{S_Y^2 \sigma_X^2} = \frac{\left((n-1)\frac{S_X^2}{\sigma_X^2}/(n-1)\right)}{\left((m-1)\frac{S_Y^2}{\sigma_Y^2}/(m-1)\right)}.$$

So if we define $U_1 = (n-1)\frac{S_X^2}{\sigma_X^2} \sim \chi^2(n-1)$ and $U_2 = (m-1)\frac{S_Y^2}{\sigma_Y^2} \sim \chi^2(m-1)$, our claim follows from the definition of F distribution.

Now we will use this observation to construct a $(100)(1-\alpha)\%$ confidence interval for σ_X^2/σ_Y^2 where $\alpha \in (0,1)$ is some fixed given value. Using the notation $F_\beta(n-1, m-1)$ analogous to definitions for previous distributions, we have

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(F_{1-\alpha/2}(n-1,m-1) \leq \frac{S_X^2 \sigma_Y^2}{S_Y^2 \sigma_X^2} \leq F_{\alpha/2}(n-1,m-1)\right) \\ &= \mathbb{P}\left(F_{1-\alpha/2}(n-1,m-1)\frac{S_Y^2}{S_X^2} \leq \frac{\sigma_Y^2}{\sigma_X^2} \leq F_{\alpha/2}(n-1,m-1)\frac{S_Y^2}{S_X^2}\right) \\ &= \mathbb{P}\left(\frac{1}{F_{\alpha/2}(n-1,m-1)}\frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{1}{F_{1-\alpha/2}(n-1,m-1)}\frac{S_X^2}{S_Y^2}\right). \end{aligned}$$

Therefore, a $(1-\alpha)100\%$ confidence interval for the ratio of the population variances is given by

$$\left[\frac{1}{F_{\alpha/2}(n-1,m-1)}\frac{S_X^2}{S_Y^2}, \frac{1}{F_{1-\alpha/2}(n-1,m-1)}\frac{S_X^2}{S_Y^2}\right].$$

Let's record this as a theorem.

Theorem 78.1 *let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent i.i.d. samples from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively, where σ_X^2 and σ_Y^2 are unknown. Then a $(1-\alpha)100\%$ confidence interval for σ_X^2/σ_Y^2 is given by*

$$\left[\frac{1}{F_{\alpha/2}(n-1,m-1)}\frac{S_X^2}{S_Y^2}, \frac{1}{F_{1-\alpha/2}(n-1,m-1)}\frac{S_X^2}{S_Y^2}\right].$$

79 t-distribution

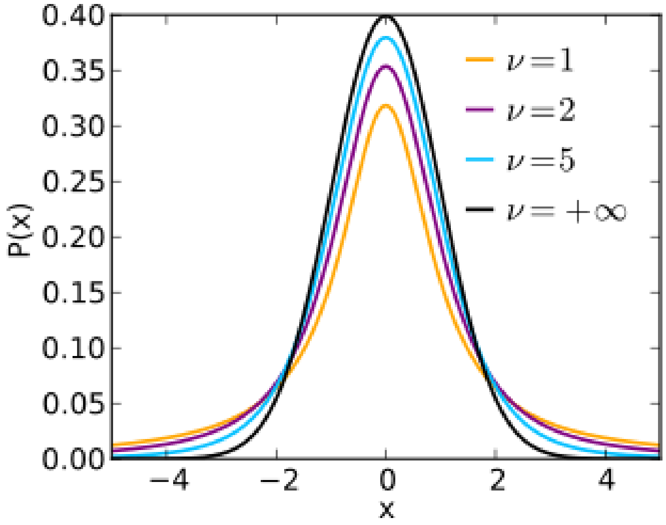
Letting $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables, previously we have seen that $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is a standard normal random variable when we know the value of σ^2. What if we also don't know the value of σ and estimate it with the sample standard deviation S?

Definition 79.1 *Let Z be a standard normal random variable, and U be a chi-squared random variable with r degrees of freedom. If Z and U are independent, then the distribution of*

$$T = \frac{Z}{\sqrt{U/r}}$$

is called the ***t-distribution with r degrees of freedom***[59]*. We write $T \sim t(r)$.*

Next figure borrowed from Wikipedia shows the pdf of t-distribution for different values of r - which is ν in the figure.



The following result provides an important instance where the t-distribution appears.

[59] t-distribution was first introduced by William Sealy Gosset who used the pseudonym *Student*. He has an interesting story, check the relevant Wikipedia page to learn more.

Theorem 79.1 *Assume that $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables. Then*

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1).$$

Proof: We already know that

1. $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$;
2. $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$.

We also take the result that $\overline{X}$ and S^2 are independent as granted, its proof is slightly technical. Once we have these,

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{(n-1)\frac{S^2}{\sigma^2}/n-1}} = \frac{Z}{\sqrt{U/(n-1)}},$$

where $Z \sim N(0,1)$ and $U \sim \chi^2(n-1)$ are independent. Therefore, by definition of the t-distribution, we may conclude $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$. □

We summarize some properties of the t-distribution that are to be kept in mind in next remark.

Remark 79.1 *(i) The support of the t-distribution is all real line, and the distribution is symmetric across $x = 0$.*

(ii) t-distribution is bell-shaped like the normal distribution.

(iii) t-distribution has heavier tails than the normal distribution. So, it is more likely to get extreme values compared to the normal distribution.

(iv) As in the case of standard normal distribution, one reads probabilities related to t-distribution from the relevant table which can be found at the end of these notes.

(v) When r is large, the distribution $t(r)$ is well approximated with the standard normal distribution.

(vi) The following table gives a general outline for approximation of t-distribution with the normal:

	Histogram looks like normal	Histogram does not look like normal
Small sample size ($n < 30$)	Good	Poor
Largel sample size ($n \geq 30$)	Good	Fair

80 Confidence intervals for the mean II

We now go back to the problem of finding confidence intervals for the mean. Let $X_1, \ldots, X_n$ be i.i.d. samples from $N(\mu, \sigma^2)$, where neither μ nor σ^2 is known. For given $\beta \in (0,1)$, let $t_\beta(r)$ be the real number for which one has $\mathbb{P}(T \geq t_\beta(r)) = \beta$ where T is a t-distributed random variable with r degrees of freedom.

In this setting, we already know that

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1),$$

where S is the sample standard deviation. Then, for $\alpha \in (0,1)$, we have

$$\begin{aligned} 1 - \alpha &= \mathbb{P}\left(-t_{\alpha/2}(n-1) \leq \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \leq t_{\alpha/2}(n-1)\right) \\ &= \cdots \\ &= \mathbb{P}\left(\overline{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right). \end{aligned}$$

Therefore we have the following conclusion.

Theorem 80.1 *Let $X_1, \ldots, X_n$ be i.i.d. samples from the $N(\mu, \sigma^2)$ distribution, where neither μ nor σ^2 is known. Then, for $\alpha \in (0,1)$, a $(1-\alpha)100\%$ confidence interval for μ is given by*

$$\left[\overline{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right].$$

Example 80.1 *A group of 10 foot surgery patients had a mean weight of 80 kilograms, and the sample standard deviation was 16 kilograms. Assuming that the weights are normally distributed, find a 95% confidence interval for the mean weight of all patients.*

Solution: In this question, $n = 10$, $\overline{x} = 80$, $\alpha = 0.05$, $t_{0.025}(9) = 2.262$, $s = 16$. Using these, the confidence interval is

$$\begin{aligned} \left[\overline{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right] &= \left[80 - (2.262)\frac{16}{\sqrt{10}}, 80 + (2.262)\frac{16}{\sqrt{10}}\right] \\ &\approx [72.85, 87.15]. \end{aligned}$$

□

81 Confidence intervals for the mean III

This is the last section concerning confidence intervals for the mean in univariate setting. We will just include some concluding discussions here.

D1 Whenever the underlying population is not assumed to be normal, we are employing the central limit theorem - in both z and t cases. So the confidence intervals we obtain result to be approximate ones. The goodness of approximation depends on various factors, but if the samples are independent and if you have a fair amount of samples, say 30, etc., then you will be doing okay in practice.

D2 Letting $X_1, \ldots, X_n$ be normally distributed random variables, $\overline{X}$ be their sample mean, and σ be the population standard deviation, we know that

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

is a $(1-\alpha)100\%$ confidence interval for the population mean μ. (The discussion to follow applies to the t-case as well.)

Question 1: Is this the only $(1-\alpha)100\%$ confidence interval we may find for μ?

Answer 1: No, indeed there are infinitely many of them. Let $0 \leq \gamma \leq \alpha$. Then we know that

$$\mathbb{P}\left(z_{1-(\alpha-\gamma)} \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\gamma}\right) = 1-\alpha.$$

Treating this as previously we may then conclude that for any $0 \leq \gamma \leq \alpha$, the interval

$$\left[\overline{X} - z_{\gamma}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-(\alpha-\gamma)}\frac{\sigma}{\sqrt{n}}\right]$$

is a $(1-\alpha)100\%$ confidence interval for μ. Note that when $\gamma = 0$, this gives us the one sided confidence interval

$$(-\infty, \overline{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}}).$$

How about the $\gamma = \alpha$ case?

Question 2: Then why do we usually work with $\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ among these infinitely many possible $(1-\alpha)100\%$ confidence intervals?

Answer 2: Because we would like to have the shortest possible interval for the given confidence level and $\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ has this property. This should be intuitively reasonable due to underlying symmetry. Can you give a rigorous proof?

Question 3: Should we have a shorter or longer confidence interval for more confidence?

Answer 3: More confidence requires a longer interval. No free lunch. At the end of the day, we could be 100% confident if we choose the confidence interval to be $(-\infty, \infty)$.

D3 Let's discuss one last time how we interprete confidence intervals. We compute confidence intervals so that a specified fraction of the intervals all computed form independent samples of the same size contain the population mean μ - justification of this can be done via law of large numbers. For example, when samples of the same size n are selected from a normal distribution and the confidence intervals are formed, approximately 95% of the intervals $[\overline{x} - 1.96\sigma/\sqrt{n}, \overline{x} + 1.96\sigma/\sqrt{n}]$ will contain the true parameter μ.

D4 In contrast to confidence intervals, we also have a concept called tolerance intervals which is more concerned about where the individual observations or measurements might fall. Let's define these latter intervals rigorously. For the normal distribution case with unknown mean μ and unknown standard deviation s, the **tolerance limits** are given by $\overline{x} - ks$ and $\overline{x} + ks$, where k is determined so that one can assert with $100(1-\gamma)\%$ confidence that the given limits contain at least the proportion $1-\alpha$ of the measurements.

D5 To conclude the discussion in D3 and D4: The confidence interval on the mean is not useful unless the data analyst is interested in the population mean. The tolerance interval is much more attentive to where individual observations fall; e.g. where the majority of the values will be.

82 Distribution free confidence intervals

The purpose of this section is to find confidence intervals for the **median** (or, more generally for percentiles) by making use of order statistics. The resulting intervals will be distribution free, meaning that there will be no model assumptions in forming them[60]. As a consequence,

[60] You may also see such intervals being called as "non-parametric" confidence intervals.

the resulting confidence intervals will tend to be longer compared to the ones where one has an idea about the underlying distribution. Again, no free lunch!

For a given sample, $X_1, \ldots, X_n$, let $X_{(1)} < \cdots < X_{(n)}$ be the corresponding order statistics. That is $X_{(1)}$ is the smallest, $X_{(2)}$ is the second smallest,...,$X_{(n)}$ is the largest among $X_1, \ldots, X_n$. Set $Y_j = X_{(j)}$ for $j = 1, \ldots, n$. So, for example

- $Y_n = \max\{X_1, \ldots, X_n\}$
- $Y_1 = \min\{X_1, \ldots, X_n\}$
- If n is odd, then $Y_{\frac{n+1}{2}} = Med(X_1, \ldots, X_n)$

We will explain the construction of distribution free confidence intervals with an example, and leave considering variations/generalizations to you.

Example 82.1 *Let $X_1, \ldots, X_5$ be i.i.d. samples from a continuous distribution, and let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ be the corresponding order statistics. Find a confidence interval for the median m of the distribution of X_1 based on the order statistics.*

Solution: Observe that

$$\begin{aligned}
\mathbb{P}(Y_1 < m < Y_5) &= \mathbb{P}(\text{at least one of } X_i\text{'s is larger than } m, \text{not all of } X_i\text{'s larger than } m) \\
&= 1 - \mathbb{P}(\text{none larger than } m) - \mathbb{P}(\text{all larger than } m) \\
&= 1 - \mathbb{P}\left(\bigcap_{j=1}^{n}\{X_j \leq m\}\right) - \mathbb{P}\left(\bigcap_{j=1}^{n}\{X_j > m\}\right) \\
&\geq 1 - \left(\frac{1}{2}\right)^5 - \left(\frac{1}{2}\right)^5 \\
&= 0.94.
\end{aligned}$$

Therefore, we may conclude that (Y_1, Y_5) is a 94% confidence interval for the median m. □

Generalizing the argument in previous example is an exercise for you.

Exercise 82.1 *Assume that $Y_1 < Y_2 < \cdots < Y_n$ is the order statistics of an i.i.d. sample $X_1, \ldots, X_n$ of size n from a continuous distribution.*

i. Show that

$$\mathbb{P}(Y_1 < m < Y_n) = 1 - \left(\frac{1}{2}\right)^n.$$

ii. Use the first part to construct a $(1-\alpha)100\%$ confidence interval for m.

Exercise 82.2 *Explain how we can use the reasoning in previous example to obtain confidence intervals for q_p, the distributional p^{th} percentile via order statistics.*

83 * Bayesian methods of estimation

In this deep but brief section we follow Myers/Myers/Walpole. The estimation techniques we discussed so far are all from the frequentist approach. For example, in deriving a 95% confidence interval for μ, we interpret the statement

$$\mathbb{P}(-1.96 < Y < 1.96) = 0.95$$

to mean that 95% of the time in repeated experiments Y falls between -1.96 and 1.96. Probabilities of this type are called objective probabilities. The Bayesian approach combines the sample information with any other available prior information.

Now consider the problem of finding a point estimate of the parameter θ for a population with pmf or pdf $f(x \mid \theta)$. The frequentist approach we had would here take a random sample and substitute the information provided into an appropriate estimator. For example, if the population is binomial with parameters n and p, our point estimate for p would be number of successes divided by n. In the Bayesian case, we suppose that some additional information is available for the parameter θ; namely, that it is known to be a random variable itself with a certain probability pmf or pdf $f(\theta)$. This distribution is called the **prior distribution**, with **prior mean** μ_0 and **prior variance** σ_0^2. The probabilities associated with this prior distribution are called **subjective probabilities** since they are considered to be a person's degree of belief in location of the parameter.

Bayesian techniques use the prior distribution $f(\theta)$ along with the joint distribution of the sample $f(x_1, \ldots, x_n \mid \theta)$ to compute the **posterior distribution** of θ: $f(\theta \mid x_1, \ldots, x_n)$. The posterior distribution consists information from both the prior distribution, and the objective sampling distribution. Then it expresses our belief in location of the parameter θ after having an observed sample.

Below $f(x_1, \ldots, x_n \mid \theta)$ is the joint probability pmf or pdf of the sample given θ. The joint distribution of $x_1, \ldots, x_n, \theta$ is then given by

$$f(x_1, \ldots, x_n, \theta) = f(x_1, \ldots, x_n \mid \theta) f(\theta),$$

from which the marginal distribution of the samples turn out to be

$$g(x_1, \ldots, x_n) = \sum_{\theta} f(x_1, \ldots, x_n, \theta)$$

in the discrete case and

$$g(x_1, \ldots, x_n) = \int_{-\infty}^{\infty} f(x_1, \ldots, x_n, \theta) d\theta$$

in the continuous case. In this setup, we may write the posterior pmf or pdf as

$$f(\theta \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n, \theta)}{g(x_1, \ldots, x_n)}.$$

Definition 83.1 *The mean of the posterior distribution $f(\theta \mid x_1, \ldots, x_n)$, denoted θ^*, is called the **Bayes estimate** of θ.*

Example 83.1 *We assume the the prior distribution for the proportion p of defective produced by a machine is*

p	0.1	0.2
$f(p)$	0.6	0.4

Find the Bayes estimate for the proportion of defectives being produced by this machine if a random sample of size 2 yields 1 defective.

Solution. Let X be the number of defectives in our sample. Then

$$f(x \mid p) = \binom{2}{x} p^x (1-p)^{2-x}, \quad x = 0, 1, 2.$$

If $p = 0.1$, the probability that the random sample of size 2 yields 1 defective is

$$f(1 \mid 0.1) = \binom{2}{1}(0.1)(0.9) = 0.18.$$

Similarly, when $p = 0.2$, we have

$$f(1 \mid 0.2) = \binom{2}{1}(0.2)(0.8) = 0.32$$

Using $f(x, p) = f(x \mid p) f(p)$, we obtain

$$f(1, 0.1) = 0.108 \quad \text{and} \quad f(1, 0.2) = 0.128.$$

From here we get

$$g(1) = \sum_p f(1, p) = 0.236.$$

Now we can get the posterior distribution for p by using $f(p \mid x) = f(x, p)/g(x)$. Indeed, the posterior probabilities turn out to be

$$f(0.1 \mid 1) = 0.458 \quad \text{and} \quad f(0.2 \mid 1) = 0.542.$$

Therefore the Bayesian estimate p^* of p is given by

$$p^* = (0.1)(0.458) + (0.2)(0.542) = 0.1542.$$

In this case, the Bayes estimate turned out to be much smaller than the one 1/2 we would obtain with a maximum likelihood approach - what is the maximum likelihood estimate? □

Exercise 83.1 *Repeat previous example if the prior distribution for p is the uniform distribution on $(0,1)$.*

Bayesian statistics is a whole field itself, and we will not be able to discuss much more here. The last thing I would like to mention is the definition for confidence intervals in Bayesian approach.

Definition 83.2 *The interval $a < \theta < b$ is called a $(1-\alpha)100\%$ Bayes interval for θ if*

$$\int_{\theta^*}^{b} f(\theta \mid x_1, \ldots, x_n) d\theta = \int_{a}^{\theta^*} f(\theta \mid x_1, \ldots, x_n) d\theta = \frac{1-\alpha}{2}.$$

84 Review problems VII

Exercise 84.1 *Let x and y equal the ACT scores in social science and natural science for a student who is applying for admission to a small liberal arts college. A sample of $n = 15$ such students yielded the following data.*

x	32	23	23	23	26	30	17	20	17	18	26	16	21	24	30
y	28	25	24	32	31	27	23	30	18	18	32	22	28	31	26.

(i) Calculate the least squares regression line for these data.
(ii) Plot the points and the least squares regression line on the same graph.
(iii) Find point estimates for α, β and σ^2.
(iv) Find 95% confidence intervals for α and σ^2 under the usual assumptions.

Exercise 84.2 *Midterm and final exam scores of 10 students in a statistics course are tabulated as shown.*

Midterm	70	74	80	84	80	67	70	64	74	82
Final	87	79	88	98	96	73	83	79	91	94.

(i) Calculate the least squares regression line for this data.

(ii) Plot the points and the least squares regression line on the same graph.

(iii) Find the value of $\widehat{\sigma^2}$.

Exercise 84.3 *A regression analysis relating test scores Y to training hours X produced the following fitted question: $\widehat{y} = 28 - 0.4x$.*
(i) What is the fitted value of the response variable corresponding to $x = 7$?

(ii) What is the residual corresponding to the data point with $x = 28$ and $y = 28$?

(iii) If x increases 2 units, how does $\widehat{y}$ change?

(iv) An additional test score is to be obtained for a new observation at $x = 7$. Would the test score for the new observation necessarily be 24.2? Explain.

Exercise 84.4 *Let $\overline{X}$, $\overline{Y}$, S_X^2 and S_Y^2 be the respective sample means and unbiased estimates of the variances using independent samples of sizes n and m from the normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, where μ_X, μ_Y, σ_X^2 and σ_Y^2 are unknown. If, however, $\sigma_X^2/\sigma_Y^2 = d$, a known constant,*

(i) What is the distribution of $\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sqrt{d\sigma_Y^2/n+\sigma_Y^2/m}}$?

(ii) What is the distribution of $\frac{(n-1)S_X^2}{d\sigma_Y^2} + \frac{(m-1)S_Y^2}{\sigma_Y^2}$?

(iiii) Using your results in first two parts, construct a random variable (not depending upon σ_Y^2) that has a t distribution and that can be used to construct a confidence interval for $\mu_X - \mu_Y$. (You may use the fact that the statistics in first two parts are independent without proof.)

(iv) By using the statistic in part (iii)., derive a formula for a 90% confidence interval for $\mu_X - \mu_Y$.

Exercise 84.5 *Let $X_1, ..., X_{20}$ be i.i.d. random variables whose expectation is μ_X and whose variance is $\sigma_X^2 = 1$. Also let $Y_1, ..., Y_{20}$ be i.i.d. random variables whose expectation is μ_Y and whose variance is $\sigma_Y^2 = 4$. We further assume that these two groups of observations are independent. Denote the corresponding sample means by $\overline{X}$ and $\overline{Y}$.*

(i) Find $Var(\overline{X} - \overline{Y})$.

(ii) What can we conclude about $\overline{X} - \overline{Y}$ by using the central limit theorem?

(iii) Construct a 95% confidence interval for $\mu_X - \mu_Y$ by using your observation in part (ii.). Provide all details.

(iv) Is the confidence interval you found in part iii an exact one? Why?

(v) Among the infinitely many possible choices of 95% confidence intervals, you have chosen the one in part iii. Is there a reason for that? Explain.

Exercise 84.6 *Find a $1 - \alpha$ confidence interval for θ, given Let $X_1, \ldots, X_n$ i.i.d. with pdf*

$$f(x \mid \theta) = 1, \quad \theta - \frac{1}{2} < x < \theta + \frac{1}{2}.$$

Exercise 84.7 *Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the normal distribution $N(\mu, \sigma^2)$. Calculate the expected length of (shortest) 95% confidence interval for μ assuming that $n = 5$, and the variance is* *(i) known,*

(ii) (for those that are really interested) unknown. Hint: To find $\mathbb{E}[S]$, first determine $\mathbb{E}[\sqrt{(n-1)S^2/\sigma^2}]$.

Exercise 84.8 *Assume that the prior distribution for the proportion p of drinks from a vending machine that overflow is*

p	0.05	0.10	0.15
$f(p)$	0.3	0.5	0.2

If 2 of next 9 drinks from this machine overflow, find
(a) the posterior distribution for the proportion p;
(b) the Bayes estimate p^* of p.

85 Motivating examples for hypothesis testing

Example 85.1 *What would be a reasonable method to test and to decide whether a particular coin is fair or not? Let p be the probability that the coin will show a head in a random toss. For this problem, the **null hypothesis** $\mathcal{H}_0$, (i.e., the hypothesis we would like to test whether it is true) and the **alternative hypothesis** $\mathcal{H}_1$ are respectively given by*

$$\mathcal{H}_0 : p = \frac{1}{2},$$

and

$$\mathcal{H}_1 : p \neq \frac{1}{2}.$$

An intuitively clear way to approach this problem would be to toss the coin N many times, and to conclude that the coin is not fair if there are too many or too few heads. For the sake of simplicity, assuming $N = 100$, the rejection region for the null hypothesis can be the set of integers falling in the interval $[0, 100] - [50 - x, 50 + x]$ where x is some non-negative integer. The question then is to decide about the value of x where the theory of statistics comes in.

One key point to make here is that whatever value we choose for $x \in (0, 50)$, there is always the chance to make a false conclusion. For example, assuming that the coin is fair, there is even some positive probability that we will have exactly 100 heads, though it is really small.

□

Example 85.2 *Statistical tests are used in law all the time. In this case, calling the person of interest Ümit, we have*

$$\mathcal{H}_0: \text{ Ümit is innocent,}$$

vs.

$$\mathcal{H}_1: \text{ Ümit is guilty.}$$

The null hypothesis here is being innocent since the stakes are very high in case of making a wrong decision. □

Example 85.3 *One other important application of hypothesis testing is in engineering problems. For example, the engineers working at a factory producing screws may test the following:*

$$\mathcal{H}_0 : \mu = 2cm,$$

vs.

$$\mathcal{H}_1 : \mu \neq 2cm,$$

where μ is the population mean of the lengths of screws produced at the factory. Here 2cm represents the ideal length of the screw. Note again that the null hypothesis is $\mu = 2cm$ since the stakes are very high if we conclude that the null hypothesis is not true (changing the machines may cost millions). So, usually the engineer will try to design a test which is inclined not to reject the null. □

Example 85.4 *Another engineering related statistical testing procedure is seen at casinos where certain machines are used to shuffle decks of cards. Assuming that the deck consists of 52 cards, and denoting the random permutation in S_{52} generated via the machine by ρ, the test can be formulated as*

$$\mathcal{H}_0 : \mathbb{P}(\rho = \pi) = \frac{1}{52!}, \qquad \text{for every } \pi \in S_{52},$$

vs. the alternative that there is bias towards certain permutations. There are various tests one may apply here, but we do not go into details. □

Example 85.5 *Finally, hypothesis testing has been of extreme use is in scientific experiments throughout the history. Some specific examples:*

i. the body temperature of a healthy human

ii. the height of a newly born baby in a certain population

iii. Mendel's experiments on the model of heredity.

These examples can be enriched easily. □

86 Hypothesis testing for the mean

In this section we would like to test whether the population mean μ of some quantity is exactly equal to some real number μ_0 or not. The setting is: We have i.i.d. samples from the normal distribution for which the mean is not known but the variance is known. The z-test in this case is as follows.

Two sided z-test for the mean:

Assume that $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$, where σ^2 is known.

Hypotheses

$\mathcal{H}_0 : \mu = \mu_0$
$\mathcal{H}_1 : \mu \neq \mu_0$

Test statistic

$$z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Decision rule

Reject $\mathcal{H}_0$ if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$, where $\alpha \in (0,1)$ is some predetermined real number (which is to be called the significance level below.)

Fail to reject $\mathcal{H}_0$ otherwise.

We will explain the reasoning behind the z-test after seeing an example.

Example 86.1 *The age of people in a certain population is normally distributed with variance* 25*. In order to understand the population mean, we take independent samples* $X_1, \ldots, X_{200}$ *and the sample mean turns out to be* 21*. Test the hypothesis that the population mean is* 18.9 *when* $\alpha = 0.05$

Solution: In this problem the hypotheses are

$$\mathcal{H}_0 : \mu = 18.9 \quad \text{and} \quad \mathcal{H}_1 : \mu \neq 18.9.$$

The test statistic takes the value

$$z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{21 - 18.9}{\frac{5}{\sqrt{200}}} = 5.939.$$

Noting that $z_{0.025} = 1.96$, and that $5.939 > 1.96$, we reject the null hypothesis. □

Let's now discuss why the method we use in z-test makes sense. Assuming that $\mathcal{H}_0$ is true, we know that $\frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ is standard normal. So if the test statistic $\frac{\overline{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ is too large or if it is too small, then this indicates a rare event, and suggests that there could be a problem with $\mathcal{H}_0$.

In setting of the test we defined above, the rarity is characterized in terms of the α value we decide prior to the test. Once we fix our α, we consider any value of the z-statistic that is larger than $z_{\alpha/2}$ or that is smaller than $-z_{\alpha/2}$ to be contradictory to our null hypothesis.

The reasoning in the two sided-test can also be used to form one-sided tests in cases where one is interested in just whether a parameter is just larger than (or smaller than) a certain threshold. We state the case for $\mathcal{H}_0 : \mu = \mu_0$ vs. $\mathcal{H}_1 : \mu > \mu_0$. Adapting it to $\mathcal{H}_1 : \mu < \mu_0$ is immediate and is left for you.

One sided z-test for the mean:

Assume that $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$, where σ^2 is known.

Hypotheses

$\mathcal{H}_0 : \mu = \mu_0$
$\mathcal{H}_1 : \mu > \mu_0$

Test statistic

$$z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Decision rule

Reject $\mathcal{H}_0$ if $z > z_\alpha$ where $\alpha \in (0,1)$ is some predetermined real number

Fail to reject $\mathcal{H}_0$ otherwise.

87 Some statistical terminology

Recall that in any statistical method we always have the chance of committing an error. For example, in case of legal prosecutions, one may judge a guilty person as innocent or an innocent person as guilty. We now define some terminology that classify the types of errors we may have.

(1) A **type I error** is rejecting the null hypothesis when $\mathcal{H}_0$ is actually true. The probability of committing a type I error is denoted by α and is called the **significance level**.
(2) A **type II error** is accepting $\mathcal{H}_0$ when it is indeed false. Probability of committing a type II error is denoted by β.
(3) The **power** of a statistical test is $1-\beta$, which is the probability of rejecting $\mathcal{H}_0$ when $\mathcal{H}_0$ is false.

Remark 87.1 *Note that the type II probability and the power of the test are actually functions of μ with $\mu \neq \mu_0$.*

Example 87.1 *Focusing on our law example where*

$$\mathcal{H}_0: \text{ Ümit is innocent,}$$

vs.

$$\mathcal{H}_1: \text{ Ümit is guilty,}$$

- *we do a type I error if we accuse Ümit although he is innocent.*
- *we do a type II error if we declare Ümit innocent although he is guilty.* □

Some comments.

(1) Minimizing type I error is equivalent to making the rejection region smaller, and minimizing the type II error is equivalent to enlarging the rejection region. So probabilities of these two types of errors are necessarily inversely related.

(2) In general, statistical tests are designed to minimize the probability of a type I error. This is related to the design of the tests since the null hypothesis is selected to be the one for which the stakes are very high.

(3) Typical levels of significance (α) in practice, i.e. probability of a type I error, are 0.01, 0.05 and 0.1.

Let's now see an example demonstrating what we learnt in this section so far.

Example 87.2 *Suppose that I find a coin and I would like to know whether the coin is fair or not. So I would like to test $H_0: p = 1/2$ vs. $H_1: p \neq 1/2$ where p is, say, the probability of a head. In order to decide about these two hypotheses, I toss the coin 16 times independently, and decide to reject H_0 if I obtain more than 10 heads or less than 6 heads.*

(i) Find an exact expression for the probability of a type I error.

(ii) Use central limit theorem to approximate the probability you found in part ii.

(iii) Assume now that we learn the true probability of having a head is $1/3$*. What is the probability of committing a Type II error under this information?*

(iv) What is the power of the test when probability of having a head is $1/3$*?*

Solution: (i) For $i = 1, \dots, 16$, let

$$X_i = \begin{cases} 1, & \text{if } i^{th} \text{ toss is head} \\ 0, & \text{if } i^{th} \text{ toss is tail.} \end{cases}$$

In particular, $\sum_{i=1}^{16} X_i$ is the total number of heads I obtain among these 16 trials which is binomially distributed with parameters 16 and 1/2 under the null hypothesis. Recalling that the type one error is rejecting the null hypothesis when it is true. In our case when $p = 1/2$, we have

$$\begin{aligned} \mathbb{P}(\text{type I error}) &= \mathbb{P}_{p=1/2}\left(\sum_{i=1}^{16} X_i < 6\right) + \mathbb{P}_{p=1/2}\left(\sum_{i=1}^{16} X_i > 10\right) \\ &= \sum_{i=0}^{5} \binom{16}{i} \frac{1}{2^{16}} + \sum_{i=11}^{16} \binom{16}{i} \frac{1}{2^{16}} \\ &= 2\frac{1}{2^{16}} \sum_{i=11}^{16} \binom{16}{i}. \end{aligned}$$

(ii) Letting $Z \sim N(0,1)$, we have

$$\mathbb{P}(\text{type I error}) = 2\mathbb{P}\left(\sum_{i=1}^{16} X_i > 10\right) \approx 2\mathbb{P}(Z > 1) = 2 - 2\mathbb{P}(Z < 1) \approx 0.32.$$

(iii) In order to commit a Type II error we should accept the null hypothesis although the true head probability is 1/3. We have

$$\beta(1/3) = \mathbb{P}_{p=1/3}(\text{type II error}) = \mathbb{P}_{p=1/3}\left(\left|\sum_{i=1}^{16} X_i - 8\right| \leq 2\right) = \sum_{i=6}^{10} \binom{16}{i} \left(\frac{1}{3}\right)^i \left(\frac{2}{3}\right)^{16-i}.$$

(iv) The power of the test is

$$1 - \beta(1/3) = 1 - \sum_{i=6}^{10} \binom{16}{i} \left(\frac{1}{3}\right)^i \left(\frac{2}{3}\right)^{16-i}.$$

□

Note that we did not have a given significance level in previous example, and so explaining our conclusions to other people requires more more definition.

Definition 87.1 *The **two-sided** **p-value** in a statistical test is the probability that the test statistic takes the value we observed or more extreme away from the null hypothesis if the null hypothesis is true.*

Example 87.3 *Recall the age of people in some community, Example 86.1. Here, we had found our z-value as 5.939 and so the p-value is given by*

$$p = \mathbb{P}(|Z| > 5.939) \approx 0.$$

□

Remark 87.2 *(i) When we are given some significance level α, we reject the null hypothesis if the p-value p is less than α.*

(ii) Some general guideline: If $p < 0.05$, then in general we have very strong evidence against $\mathcal{H}_0$, and if $p > 0.2$, then there is "no evidence" against $\mathcal{H}_0$. These are of course just very general suggestions, in general, you need to think about the p-values as problem dependent.

(iii) One sided tests. *If we want to test $\mathcal{H}_0 : \mu = \mu_0$ against $\mathcal{H}_1 : \mu > \mu_0$, then the p-value of the test statistic taking a value at least as large as observed value. The case where the alternative hypothesis is $\mathcal{H}_1 : \mu < \mu_0$ is similar.*

88 Test for the mean when σ is unknown

Two sided t-test for the mean: Assume that $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$, where σ^2 is known.

Hypotheses

$\mathcal{H}_0 : \mu = \mu_0$
$\mathcal{H}_1 : \mu \neq \mu_0$

Test statistic

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Decision rule

Reject $\mathcal{H}_0$ if $t > t_{\alpha/2}(n-1)$ or $t < -t_{\alpha/2}(n-1)$, where $\alpha \in (0,1)$ is the significance level.

Fail to reject $\mathcal{H}_0$ otherwise.

Example 88.1 *Assume that the IQ score of people in Turkey is normally distributed with parameters μ and σ^2. We would like to test*

$$\mathcal{H}_0 : \mu = 90 \qquad vs. \qquad \mathcal{H}_1 : \mu \neq 90.$$

For this purpose suppose that we sample 64 people randomly, and that the sample mean and the sample variance are respectively 94 and 36. Decide what your conclusion would be when the level of significance is 0.05%.

Solution. In this problem, $n = 64$, $\overline{x} = 94$, $s^2 = 36$ and $\alpha = 0.05$. One reads $t_{0.025}(63)$ from the t-table as 1.96. Also we compute the t-statistic

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{94 - 90}{\sqrt{36}/8} = 5.4.$$

Since $5.4 > t_{0.025}(63) = 1.96$, we reject the null hypothesis. □

One sided t-test for the mean:

Assume that $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$, where σ^2 is known.

Hypotheses

$\mathcal{H}_0 : \mu = \mu_0$
$\mathcal{H}_1 : \mu > \mu_0$

Test statistic

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Decision rule

Reject $\mathcal{H}_0$ if $t > t_\alpha$, where $\alpha \in (0,1)$ is the significance level.

Fail to reject $\mathcal{H}_0$ otherwise.

The case for $\mathcal{H}_1 : \mu < \mu_0$ is simlar.

In the rest of this section, let's discuss the relationship between confidence intervals and hypothesis testing by focusing on the t-test. In this case note that the statistic

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

played the key role. Indeed it turns out that the testing of $\mathcal{H}_0 : \mu = \mu_0$ vs. $\mathcal{H}_1 : \mu \neq \mu_0$ at a significance level α is equivalent to computing a $100(1-\alpha)\%$ confidence interval for the population mean μ when σ is unknown. If μ_0 is inside the confidence interval, then the hypothesis is not rejected. Written more rigorously, an observed value $\overline{x}$ results with a failure to reject $\mathcal{H}_0$ at significance level α implies that

$$-t_{\alpha/2}(n-1) \leq \frac{\overline{x}-\mu_0}{\frac{s}{\sqrt{n}}} \leq t_{\alpha/2}(n-1)$$

which is equivalent to

$$\overline{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} \leq \mu_0 \leq \overline{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}.$$

This equivalence to confidence intervals extends to differences between two means, variances, ratios of variances and so on. So, in terms of statistical inference, you should keep in mind that there is no reason to separate confidence intervals from hypothesis testing.

89 Comparison of two population means

Consider two independent random samples $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ of size n_1, n_2, respectively, that are drawn from two populations with means μ_1, μ_2 and variances σ_1^2, σ_2^2. Assume that n_1 and n_2 are both large enough so that the CLT applies. We then know that the random variable

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has the approximate standard normal distribution (Why?). Note here that

$$\mu_1 - \mu_2 = \mathbb{E}[\overline{X} - \overline{Y}] \quad \text{and} \quad \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = Var(\overline{X} - \overline{Y}).$$

Knowing the distribution of Z, it should then be straightforward for you to construct confidence intervals and do hypothesis testing for

$$\mathcal{H}_0 : \mu_1 - \mu_2 = \mu_0 \quad \text{vs.} \quad \mathcal{H}_1 : \mu_1 - \mu_2 \neq \mu_0$$

for some $\mu_0 \in \mathbb{R}$. A discussion of these will be included in the upcoming example. Before that, let's mention one case of particular importance in hypothesis testing. If we choose $\mu_0 = 0$ in above hypotheses, then the null and the alternative hypotheses can be rewritten as

$$\mathcal{H}_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad \mathcal{H}_1 : \mu_1 \neq \mu_2.$$

This is testing whether the two populations have the same mean or not, a very important question in practice; e.g. this can be used to test whether a certain drug helps the patients suffering from low iron in their blood or not. How?

Now, let's summarize the above discussion in an example.

Example 89.1 *Let $X_1, \dots, X_{25}$ be i.i.d. random variables whose expectation is μ_X and whose variance is $\sigma_X^2 = 1$. Also let $Y_1, \dots, Y_{25}$ be i.i.d. random variables whose expectation is μ_Y and whose variance is $\sigma_Y^2 = 9$. We further assume that these two groups of observations are independent. Denote the corresponding sample means by $\overline{X}$ and $\overline{Y}$.*

(i) Find $Var(\overline{Y} - \overline{X})$.

(ii) What can we conclude about $\overline{Y} - \overline{X}$ by using the central limit theorem? [61]

(iii) Construct a 95% confidence interval for $\mu_Y - \mu_X$ by using your observation in part (ii).

(iv) Suppose that we are interested in testing $H_0 : \mu_X = \mu_Y$ vs. $H_1 : \mu_Y > \mu_X$. Write a test statistic that can be used for this purpose. Also describe the rejection region for the test.

Solution: (i) $Var(\overline{Y} - \overline{X}) = Var(\overline{Y}) + Var(\overline{X}) = \frac{1}{25} + \frac{9}{25} = 0.4$.

(ii) CLT says that

$$\frac{\overline{Y} - \overline{X} - (\mu_Y - \mu_X)}{\sqrt{0.4}}$$

is approximately normal.

(iii) By previous item

$$\mathbb{P}\left(-z_{0.025} < \frac{\overline{Y} - \overline{X} - (\mu_Y - \mu_X)}{\sqrt{0.4}} < z_{0.025}\right)$$

is approximately 0.95. As in various previous examples we have seen in lecture we may then manipulate this expression to obtain the following approximate 95% confidence interval for $\mu_Y - \mu_X$:

$$\left[\overline{Y} - \overline{X} - 1.96\sqrt{0.4}, \overline{Y} - \overline{X} + 1.96\sqrt{0.4}\right].$$

(iv) Test statistic for this one-sided test is

$$z = \frac{\overline{Y} - \overline{X} - 0}{\sqrt{0.4}} = \frac{\overline{Y} - \overline{X}}{\sqrt{0.4}}.$$

Rejection region for a test of significance level α is $z > z_\alpha$. □

Everything we discussed for the standard univariate hypothesis testing and confidence intervals apply here as well. Thinking over the details is an exercise for you.

[61] Here, letting $Z_i = X_i - Y_i$, $\overline{X} - \overline{Y} = \overline{Z}$. This observation can be useful for you.

90 Tests for proportions

This time we are interested in testing population proportions. Here is an examplary problem where we may meet such a situation. Ümit Işlak is willling to run for the rectorate seat at Boğaziçi University. But before investing on this he wants to make sure that he has enough support to run a campaign. For example, he wants to test whether he has the support of 40% of the overall faculty or not. This hypothesis is formulated

$$\mathcal{H}_0 : p = 0.4 \quad \text{vs.} \quad \mathcal{H}_1 : p \neq 0.4.$$

Let's assume that actual value of the parameter is some p_0 which we do now know. In this case, denoting the number of samples by n (Here, the samples $X_1, \ldots, X_n$ are Bernoulli random variables), and the sample mean by $\overline{p}$, the z-statististic given by

$$z = \frac{\overline{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

will have an approximately normal distribution (why?) when $n\overline{p}$ is large enough; in practice, they say larger than 5. Based on this the test scheme for the proportion case is as follows:

Two sided proportions test: Assume that $X_1, \ldots, X_n$ are independent Bernoulli random variables with parameter p. Let $\overline{p}$ be the corresponding sample mean.

Hypotheses

$\mathcal{H}_0 : p = p_0$
$\mathcal{H}_1 : p \neq p_0$

Test statistic

$$z = \frac{\overline{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Decision rule

Reject $\mathcal{H}_0$ if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$ where $\alpha \in (0,1)$ is some predetermined real number.

Fail to reject $\mathcal{H}_0$ otherwise.

Example 90.1 *Let's continue our introductory example. Ümit was willing to be rectorate at Boğaziçi University and he was planning a hypothesis test in order to see whether his support is $p = 0.4$ or not. For this purpose he samples 129 faculty members randomly and learns about their opinion*[62]*. It turns out that 37 of them are willing to support Ümit in such an election. So here is how the hypothesis testing procedure:*

$$\mathcal{H}_0 : p = 0.4 \quad vs. \quad \mathcal{H}_1 : p \neq 0.4.$$

The sample size is quite large, so we may use a normal approximation. Noting that $\overline{p} = 37/129 = 0.287$, the z-value turns out to be

$$z = \frac{\overline{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.287 - 0.4}{\sqrt{\frac{0.4(0.6)}{129}}} = -2.62.$$

Checking the $z-$table, the p-value turns out to be around 0.0088*. For instance, if we take the significance level 0.05, it looks like it is impossible for Ümit to win the elections. How sorry he must be!* □

91 A first look at non-parametric tests

A statistical model is said to be a **parametric** one if the joint distribution of the observations involves certain unknown constants called parameters. One parametric model we have already seen is the normal model in which the observations are independently distributed with parameters μ and σ^2. In real life situations one will usually have models with too many parameters characterizing the distribution of interest.

A **non-parametric model** on the other hand is a model in which no assumption is made about the functional form of the joint distribution (e.g. normal distribution in previous paragraph). The only assumption made about the observations is that they are i.i.d. from some distribution. In particular, there are no parameters in a non-parametric model[63].

Here, I will discuss the idea behind non-parametric statistics by just focusing on one of the oldest such tests: Wilcoxon's rank sum test.

Example 91.1 *Suppose that we are given two groups each consisting of 4 people. Suppose that these eight people take a certain test and that we are interested in whether the grade distributions of the groups are the same. The test scores are given in the following table:*

[62] This is just a toy example, don't worry about the confounding factors

[63] There is also the **semi-parametric model** in which there are some parameters but very weak assumptions are made about the actual form of the distribution of the observations.

Group I	Group II
27	32
31	29
26	35
25	28.

Now assuming that the distribution for these two groups are the same, one would expect to have the corresponding rank[64] *sums close to each other. Next table contains the ranks of our eight samples and the rank sums of the two groups we have:*

	Group I	Group II
	3	7
	6	5
	2	8
	1	4
Rank sum	12	24.

As we see there is a significant difference between the rank sums of two populations which directs us to question whether the assumption of identical distribution really holds. It is possible to make this intuitive discussion rigorous by using some further statistical terminology, but I will skip that here referring you to more advanced books. □

92 A concluding discussion on simple linear regression

Recall that the simple linear regression model we had was:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i's are independent normal random variables with mean zero and variance σ^2. We are now ready to solve the following question. In Sections 73 and 74, we had shown that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

and

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

are unbiased estimators for β_1 and β_0, respectively.

[64] For example, the rank of 25 is 1, the rank of 26 is 2,..., the rank of 35 is 8.

Focusing on $\hat{\beta}_1$, we can indeed go further and show that the statistic

$$T = \frac{\hat{\beta}_1 - \beta_1}{\frac{S}{\sqrt{S_{xx}}}},$$

where $S = \sum_{i=1}^{n} \frac{(Y_i - \overline{Y_i})^2}{n-2}$ and $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$, is t distributed with $n-2$ degrees of freedom - we skip this derivation here. But once we have this distributional information, we may then obtain a confidence interval for β, do hypothesis testing for it, etc. The following should be clear by what we are given:

A $(1-\alpha)100\%$ confidence interval for β_1 is

$$\left[\hat{\beta}_1 - t_{\alpha/2}\frac{S}{\sqrt{S_{xx}}}, \hat{\beta}_1 + t_{\alpha/2}\frac{S}{\sqrt{S_{xx}}}\right].$$

The observation that $T = \frac{\hat{\beta}_1 - \beta_1}{\frac{S}{\sqrt{S_{xx}}}}$ is t-distributed also allows hypothesis testing for β_1, e.g. $\mathcal{H}_0 : \beta_1 = c$, vs. $\mathcal{H}_1 : \beta_1 \neq c$ for some $c \in \mathbb{R}$. By now, you should be comfortable with designing such a test. I would just like to emphasize the importance of $c = 0$ case in practice here. The truth of $\beta_1 = 0$ implies that the slope of the regression line is zero, but then this means that there is no direct linear influence of the input x to the output y, and that we just have the underlying noise[65].

As a last note on regression, I would like to mention that in practice you will often need to deal with multiple linear regression. In this case we have more than one independent variables, and the most general model looks like

$$Y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon_k,$$

where ϵ_k is the noise for the ith observation. Here the output (possibly) depends on $k \geq 1$ independent variables in contrast to simple linear regression where we had $k = 1$. A rigorous study of multiple linear regression requires approaching the problem with linear algebra, and we will not do it here in such a brief course. In case you are interested, you may just grab a textbook from the library on linear models to learn more (Or,of course you may take some other more specialized course - not necessarily from the math department - a course such as econometrics will involve lots of multiple regression).

[65] By a noise, we refer to the randomness resulting from the ϵ. terms.

93 Goodness of fit test

If k independent variables X_i are each normally distributed with mean μ_i and variance σ_i^2, we know that the the quantity

$$\chi^2 = \frac{(X_1-\mu_1)^2}{\sigma_1^2} + \frac{(X_2-\mu_2)^2}{\sigma_2^2} + \cdots + \frac{(X_k-\mu_k)^2}{\sigma_k^2}$$

is chi-square distributed with k degrees of freedom. Note that ideally, given the random fluctuations of the values of X_i about their mean values μ_i, each term in the sum will be of order unity. Hence, if we have chosen the μ_i and the σ_i correctly, we may expect that a calculated value of χ^2 will be approximately equal to k. If it is, then we may conclude that the data are well described by the values we have chosen for the μ_i, that is, by the hypothesized function. If a calculated value of χ^2 turns out to be much larger than k, and we have correctly estimated the values for the σ_i, we may possibly conclude that our data are not well described by our hypothesized set of the μ_i. This is the general idea of the χ^2 test. In what follows we provide the details of the procedure.

Let the possible outcomes $\mathcal{A}$ of an experiment be the union of k mutually disjoint sets $A_1, \ldots, A_k$. Further, let $p_i = \mathbb{P}(A_i)$ so that $p_1 + \cdots + p_k = 1$. In words, p_i is the probability that the outcome of the experiment is an element of A_i. Our purpose is to test the hypotheses:

$$\mathcal{H}_0 : p_1 = p_{10},\ p_2 = p_{20},\ \ldots, p_{k-1} = p_{k-1,0}, \quad (\text{and } \ p_k = 1 - p_{10} - \cdots - p_{k-1,0}),$$

vs. the alternative that $\mathcal{H}_0$ is not true, where $p_{10}, \ldots, p_{k-1,0}$ are some given numbers $(0,1)$ with $\sum_{j=1}^k p_{j0} = 1$. In order to test this we repeat the experiment n times and let X_i be the number of times outcome is in A_i. In particular, $\sum_{i=1}^k X_i = n$. Here is a fact which will be used without proof below:

Fact: Assuming $\mathcal{H}_0$ is true, the statistic

$$Q_{k-1} = \sum_{i=1}^k \frac{(X_i - np_{i0})^2}{np_{i0}}$$

is approximately chi square distributed with $k-1$ degrees of freedom.

Although we are not able to prove this result rigorously here, we may still discuss the special case $k=2$. In this case, the null hypothesis is $\mathcal{H}_0 : p_1 = p_{10}$ and $p_2 = p_{20} = 1 - p_{10}$. Noting the identity

$$(X_2 - np_{20})^2 = (n - X_1 - n(1-p_{10}))^2 = (X_1 - np_{10})^2,$$

we have

$$
\begin{aligned}
Q_1 &= \frac{(X_1 - np_{10})^2}{np_{10}} + \frac{(X_1 - np_{20})^2}{np_{20}} \\
&= \frac{p_{20}(X_1 - np_{10})^2 + p_{10}(X_1 - np_{10})^2}{np_{10}p_{20}} \\
&= \left(\frac{X_1 - np_{10}}{\sqrt{np_{10}(1 - p_{10})}} \right)^2 .
\end{aligned}
$$

Now noting that $\frac{X_1 - np_{10}}{\sqrt{np_{10}(1-p_{10})}}$ is approximately normal when n is large one concludes that Q_1 is approximately a chi square random variable with 1 degree of freedom. This is consistent what we have in the stated fact.

Now, we are ready to give the test procedure for the goodness of fit test.

Goodness of fit test:

Hypotheses

$\mathcal{H}_0 : p_1 = p_{10},\ p_2 = p_{20},\ \ldots, p_{k-1} = p_{k-1,0}$
$\mathcal{H}_1 : \exists i \in \{1, \ldots, k\}$ with $p_i \neq p_{i0}$

Test statistic

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_{i0})^2}{np_{i0}}$$

Decision rule

Reject $\mathcal{H}_0$ if $Q_{k-1} > \chi^2_\alpha(k-1)$, where $\alpha \in (0,1)$ is the significance level of the test determined by the experimenter.

Fail to reject $\mathcal{H}_0$ otherwise.

Example 93.1 *A number in* $\{1, 2, 3, 4, 5, 6\}$ *is to be chosen via rolling a balanced die. We would like to test*

$$\mathcal{H}_0 : \mathbb{P}(A_i) = p_{i0} = \frac{1}{6}, \qquad i = 1, \ldots, 6,$$

at significance level $\alpha = 0.05$. *Here* A_i *denotes the event that we obtain an* i *in a die roll. In order to test this we roll the die 60 times and the frequencies of* $A_1, \ldots, A_6$ *are* $13, 19, 11, 8, 5$ *and* 4. *What would be your conclusion?*

Solution: We have

$$\begin{aligned} Q_5 &= \sum_{i=1}^{6} \frac{\left(X_i - 60\frac{1}{6}\right)^2}{60\frac{1}{6}} \\ &= \frac{(13-10)^2}{10} + \frac{(19-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(4-10)^2}{10} \\ &= 15.6. \end{aligned}$$

From the χ^2 table read $\chi^2(5) = 11.1$ and since $15.6 > 11.1$ we reject the null hypothesis. □

94 χ^2-square test of independence

Chi-square test provides also a non-parametric statistical test to determine if the two or more classifications of the samples are independent or not. The chi square test of independence is very general, and can be used with variables measured on any type of scale, nominal, ordinal, interval, etc. The only limitation one has is that the sample sizes must be sufficiently large to ensure that the expected number of cases in each category is *five or more.*

Let's discuss some exemplary problems fitting in this framework.

Example 94.1

1. *Let X be an indicator variable showing whether a person believes afterlife or not, and let Y be the indicator that this particular person is happy. We may test whether X and Y are dependent on each other.*

2. *Let X_i, $i = 1, \ldots, 6$ be the number of goals scored by the Super Lig teams between the duration (in minutes) $[15(i-1), 15i]$. One may check whether the number of goals in distinct periods are independent or not.*

3. *Letting X be the variable holding the departments at Bogazici University (molecular biology, ee, math education, etc) taking the course Math 102, and letting Y be the indicator for getting a pass grade, we may test whether the department affects succeeding in Math 102 - dependence of X and Y.*

Chi square test for independence:

Setting: We have two variables X and Y; the former can take values $a_1, \ldots, a_k$ and the latter $b_1, \ldots, b_m$.

Total number of experiments is n.

O_{ij} = Number of observations in which $X = i$ and $Y = j$.

$p_i = \sum_{j=1}^{m} \frac{O_{ij}}{n}$, $i = 1, \ldots, k$;

$p_j = \sum_{i=1}^{k} \frac{O_{ij}}{n}$.

E_{ij} (= Expected number of occurences of $X = i$ and $Y = j$) $= np_ip_j$.

Hypotheses

$\mathcal{H}_0$: There is no dependence among the underlying variables X and Y.
$\mathcal{H}_1$: Variables of interest are dependent.

Test statistic

$$\chi^2_{(k-1)(m-1)} = \sum_{i=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Decision rule

Reject $\mathcal{H}_0$ if $Q_{(k-1)(m-1)} > \chi^2_\alpha(k-1)$, where $\alpha \in (0,1)$ is the significance level of the test determined by the experimenter.

Fail to reject $\mathcal{H}_0$ otherwise.

Exercise 94.1 *To test whether two distinct instruction methods affect teaching quality, 100 students are selected randomly and they are divided into two groups of equal sizes randomly. At the end of the semester the grade distribution (the grades can be A, B, C, D or F) for these two groups is given in the following table:*

	A	B	C	D	F	Total
Group I	8	13	16	10	3	50
Group II	4	9	14	16	7	50

Is there a convincing evidence that one of these instruction methods is superior over the other one?

Let's conclude the discussion with some general guidelines on assumptions made for using the chi square test for independence. Assumptions:

A. Independent and properly random samples;

B. All expected counts are greater than one;

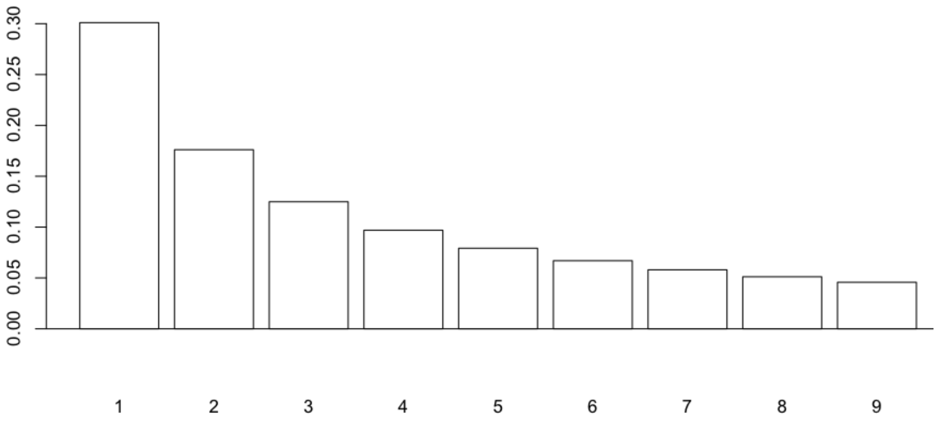
C. No more than 20% of cells with an expected count of less than five.

95 * Benford's law and hypothesis testing

Benford's law, also known as First-digit law or Significant-digit law, is an observation about the distribution of digits of numbers in a numerical data set. This law has a mysterious side since it appears in various real life settings such as the area of rivers, populations in cities, death rates etc. Moreover, there are many fields where the law has practical applications. Accounts payable data, estimations in general ledger, new combinations of selling prices are some fields where the Benford's law found applications previously. Benford's law (or, distribution) is defined rigorously with the following pmf

$$f(d) = Prob(D_1 = d) = \log_{10}(d+1) - \log_{10} d, \qquad d \in \{1, 2, \ldots, 9\}.$$

Here D_1 is a random variable following the Benford distribution. Its probability mass function looks like this:



Historically, Simon Newcomb (1835-1909) is the first scientist who realized that the first digit numbers have non-uniform distribution when he was looking at the books of logarithm table (1881). What he really noticed that is the first part of the book of logarithm table were more used than the rest, that is, scientists usually worked with the numbers having 1 as a first digit.

After around fifty years, Frank Benford, an engineer and a physicist, stated this unique distribution in his paper titled *The Law of Anomalous Numbers* in 1938. In this paper, he showed his analysis about first digits of numbers in different types of data sets such that scientific constants, population, street addresses and so on. By considering the first digits of numbers in data sets, Frank Benford observed that the numbers in first digit are not uniformly distributed on 1, . . . , 9. Actually first digits in the list of numbers was tend to show an exponential distribution; the probability of having a 1 in first digit is much more than the probability of having a 9. Moreover, a similar phenomenon arises when one considers the first two digits, or more generally the first k digits. As result of his observations, Benford established a general formula for these probabilities. Here we will have a brief discussion - just giving you the idea - of an application of Benford's law in tax frauds.

Before doing so let me show you some data in order to convince you there is really something interesting here. For example, the following figure shows the first digit distribution of the population of the 237 countries of the worlds as of 2010.

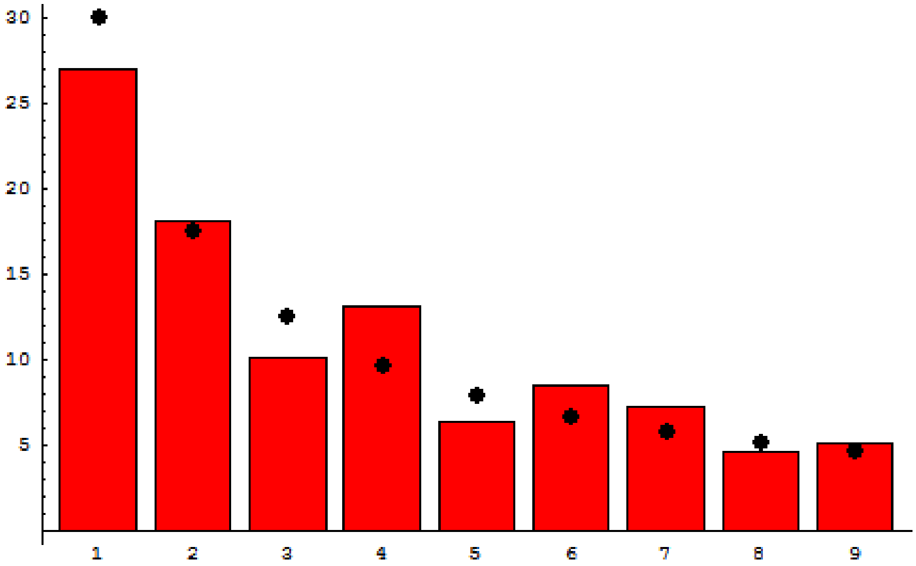


Figure 11: Black dots correspond to the distribution of Benford and the bars are the first digits of the population of the 237 countries of the world as of July 2010

That is not all, the following figure collects various data from different areas that seem to closely follow Benford's law when looked at the first digit:

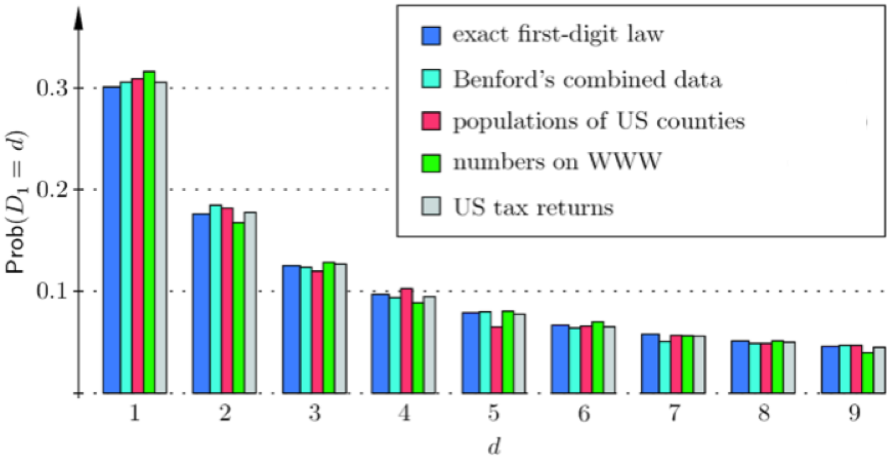
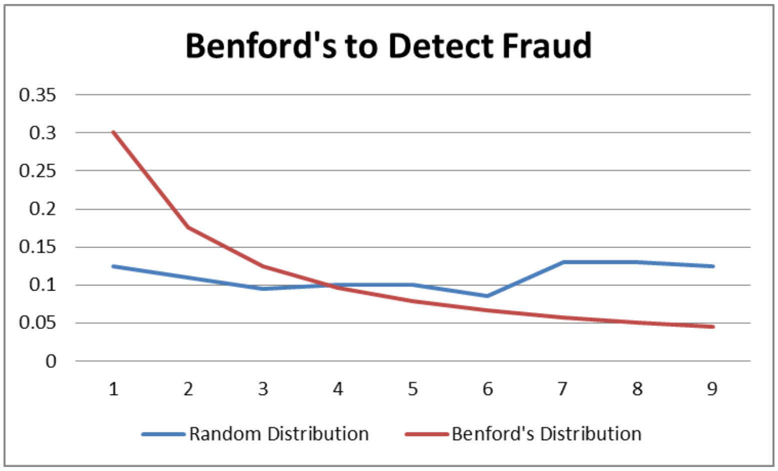


Figure 12: Various data following the Benford law

So how does this relate to tax frauds? The idea here is simple: if true data of a certain type is known to be close to Benford's Law, then chi-squared goodness-of-fit tests can be used as a simple test for data fabrication or falsification. Whether the tested data are close to Benford's Law or are not close proves nothing, but a poor fit raises the level of suspicion, at which time independent (non-Benford) tests or monitoring may be applied. For example, the following figure shows the Benford's distribution and some fraud going on via writing the first digits randomly.



As already mentioned, the detection is based on a chi-squared goodness of fit test. Let me note that there are various other statistical tests for fraud detection. We will not go into these here, but I encourage you to check this topic. It is really interesting.

96 One-way ANOVA

In this section, I will be mostly following Chapter 7 of Howard Seltman's book on experimental design and analysis [13]. One-way analysis of variance (abbreviated one-way ANOVA) is a technique that can be used to compare means[66] of two or more samples (using the F distribution). This technique can be used only for numerical response data, the "Y", usually one variable, and numerical or (usually) categorical input data, the "X", always one variable, hence "one-way".

The ANOVA is used to test the null hypothesis that samples in two or more groups are drawn from populations with the same mean values - typically when there are at least three groups[67].

Our example to follow here will be on understanding whether students' scores on a standardized test is a function of the teaching method they received. The independent variable represented the three different types of teaching methods:

1. lecture only;

2. hands-on

3. lecture and hands-on.

Denoting the actual mean scores of these three distinct methods by μ_1, μ_2, μ_3, respectively, we would like to test

$$\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3 \qquad \text{vs.} \qquad \mathcal{H}_1 : \exists i \neq j \text{ such that } \mu_i \neq \mu_j.$$

The data we have for these three groups are given as below

Method	Group size	Mean	Standard deviation
Lecture only	15	82.80	9.59
Hand-on only	15	88.53	8.73
Lecture and Hands-on	15	92.67	6.22
Total Group	45	88.00	9.09

In order to develop our test, we first need to have certain assumptions. Let's first assume that we have k groups. Further assumtions are as follows.

Assumptions in Anova

[66] Yes, "means" - and not the variances. The importance of variance will become clear below.

[67] When there are only two groups, it reduces to the t-test which we have seen earlier on.

(i) Subjects are chosen via a simple random sample.

(ii) The outcomes are assumed to be normally distributed.

(iii) The populations have the same variances $\sigma^2 > 0$.

A quick comment: ANOVA is somewhat robust, that is, results remain fairly trustworthy despite mild violations of these assumptions. In particular, as long as we have simple random samples from each group, assumptions in (ii) and (iii) are close enough to being true if you

A look at normal quantile plots for each group and, in each case, you see that the data points fall close to a line[68];

B compute the standard deviations for each group sample, and see that the ratio of the largest to the smallest group sample standard deviation is large[69].

So, under these assumptions on these k groups, we would like to test

$$\mathcal{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k \qquad \text{vs.} \qquad \mathcal{H}_1 : \exists i \neq j \text{ such that } \mu_i \neq \mu_j.$$

The statistic we will be using will be the F-statistic. The derivations here are fairly technical, so grasping the main ideas at a first look will be enough for our purposes.

First, some recall and some preliminaries. Recall that the sample variance of some given data $z_1, \ldots, z_m$ is $SS = \sum_{i=1}^{m}(z_i - \overline{z})^2$ and the degrees of freedom (df) will always be $m-1$ in our case when we focus on a particular group. Quantities of the form SS/df will be called **mean squares** or **MS**, in short.

The one way ANOVA requires dealing with two distinct MS values:

1. "mean square within groups", denoted MS_{within}, and
2. "mean square between groups", denoted MS_{between}. Our eventual test statistic will be

$$\frac{MS_{\text{between}}}{MS_{\text{within}}}.$$

So let's begin by understanding these two MS values.

Mean square within groups: MS_{within}

[68] Or you do some other normality test, and the data "seems" to be normal

[69] In practice, say that the ratio of the largest to the smallest group sample standard deviation is no more than two.

MS_{within} is an estimate for σ^2 that is independent of whether the null or the alternative hypotheses are true. The following figure shows a two-group experiment with 4 subjects in each group. If you have more groups, extension of the figure will be straightforward[70].

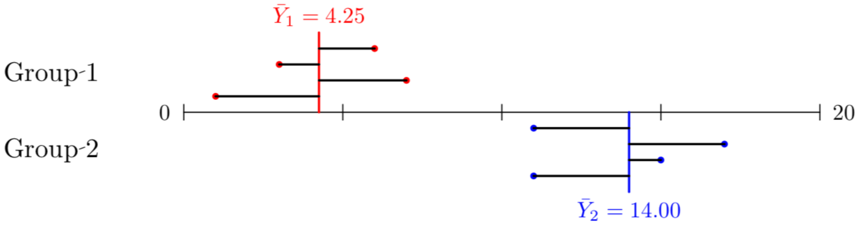


Figure 13: Deviations for within-group sum of squares

The deviation of subject j of group i is mathematically equal to $Y_{ij} - \overline{Y_i}$ where Y_{ij} is the jth observed value in group i and $\overline{Y_i}$ is the sample mean for group i.

Definition 96.1 *For an individual group i, we define*

$$SS_i = (Y_{ij} - \overline{Y_i})^2$$

and $df_i = n_i - 1$ where n_i is the size of group i. In this case, MS_{within} is defined to be

$$MS_{within} = SS_{within}/df_{within},$$

where

$$SS_{within} = \sum_{i=1}^{k} SS_i$$

and

$$df_{within} = \sum_{i=1}^{k} df_i = \sum_{i=1}^{k} (n_i - 1) = N - k$$

with $N = \sum_{i=1}^{k} n_i$.

Fact: MS_{within} is a good estimate of σ^2 whether $\mathcal{H}_0$ is true or not thanks to our model assumption that the populations have the same variances. The details of being ”good” are technical and we do not go into details of this here.

Next, it is time to discuss MS_{between}.

[70] Following two figures are borrowed from [13]

Mean square between groups: MS_{between}

This time consider the following figure between-group deviations:

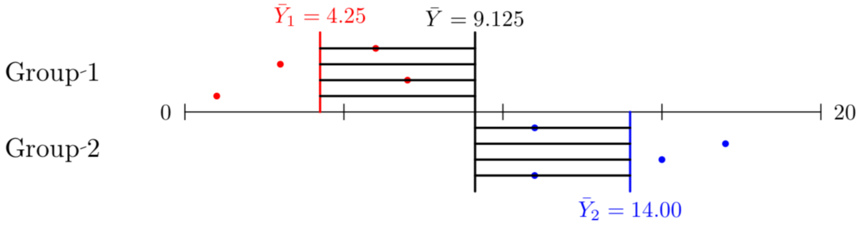


Figure 14: Deviations for between-group sum of squares

The middle vertical line is the sample average of all of the outcome values coming from all groups. For that reason, we call it the **grand mean**.

Definition 96.2 *We define the quantity $SS_{between}$ to be*

$$SS_{between} = \sum_{i=1}^{k} (\overline{Y_i} - \overline{\overline{Y}})^2, \tag{22}$$

where $\overline{\overline{Y}}$ is the grand mean and the other notations are as before. Then $MS_{between}$ is defined by

$$MS_{between} = \frac{SS_{between}}{k-1}.$$

Facts and notes: (1) MS_{between} is a "good" estimate of σ^2 only when the null hypothesis true since in this case we expect the group sample means to be close to each other and close to the grand mean.

(2) When you look at (22), since the k unique deviations add up to zero, we are free to choose only $k-1$ of them, and then the last one is determined by the others. This is the reason why $df_{\text{between}} = k - 1$ for one-way ANOVA.

Now we are ready to introduce the F-statistic we promised for testing whether the means of various populations are the same. We define F to be

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}.$$

Since the denominator is always (under $\mathcal{H}_0$ or $\mathcal{H}_1$) an estimator of σ^2 and the numerator is either an estimator of σ^2 (under $\mathcal{H}_0$) or is higher than the actual σ^2 (under $\mathcal{H}_1$), the (random) values of the F-statistic will tend to be around 1 when $\mathcal{H}_0$ is true and will be larger than 1 when the alternative is true. Let's summarize this:

> The F-statistic, $F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$, tends to be "much" larger if the alternative hypothesis is true.

Of course, we have a good understanding of the F-distribution, and so we will be able to statistically quantify the term "much" in previous box. This part is technical. Firstly, it is mathematically known (but not straightforward to prove) that the statistics MS_{between} and MS_{within} are independent of each other once we have our model assumptions. Therefore, knowing the degrees of freedom as well,

> Under the null hypothesis, the statistic
>
> $$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$
>
> has an F-distribution with parameters, $(k-1, N-k)$.

Just a small comment before we proceed to an example. Note that the F-statistic we have under the null hypothesis depends only on the number of groups and the number of subjects, and not on the values of the population variance and or the population group means. It is also nice to observe that the degrees of freedom are measures of "size" of the experiment, where bigger experiments (more groups or subjects) have bigger degrees of freedom.

Example 96.1 *Let's go back to our example at the beginning of this section. The goal was to understand whether students' scores on a standardized test depends on the teaching method they received. The independent variable represented the three different types of teaching methods: 1) lecture only; 2) hands-on only; and 3) lecture and hands-on. Denoting the actual mean scores of these three distinct methods by μ_1, μ_2, μ_3, respectively, we would like to test*

$$\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3 \qquad vs. \qquad \mathcal{H}_1 : \exists i \neq j \text{ such that } \mu_i \neq \mu_j.$$

The data we have for these three groups were given as follows:

Method	Group size	Mean	Standard deviation
Lecture only	15	82.80	9.59
Hand-on only	15	88.53	8.73
Lecture and Hands-on	15	92.67	6.22
Total Group	45	88.00	9.09

Further, let's assume that the three assumptions required for ANOVA hold in this problem[71]*. Doing the various necessary computations we then obtain the following result:*

Source	SS	degrees of freedom	MS	F	p-value
Between	736.53	2	368.27	5.34	.009
Within	2895.47	42	68.94		
Total	3632.00	44			

The p-value turns out to be $0.009 < 0.01$ *which is really small. At* 5% *significance level, you would reject the null hypothesis easily. But if you decide at a* 1% *significance level, then it would be good to repeat the experiment independently one more time.*

97 Review problems VIII

Exercise 97.1 *A suspect that vehemently denies her involvement in a terrorist plot is given a polygraph test to determine if she is lying.*

(a) Notice that the polygraph test is an hypothesis test. Formulate the null and alternative hypotheses. (Hint: recall that suspects are innocent until proven guilty.)

(b) Upon conclusion of the polygraph test the suspect is found to be telling the truth. However, DNA evidence received later positively confirms her involvement in the plot. What kind of error was made: type I or type II?

Exercise 97.2 *A random selection of 12 students at a university was asked, 'How much did you spend on textbooks this semester?' The responses (in dollars) are shown below.*

$$200, 175, 450, 300, 350, 250, 150, 200, 320, 370, 400, 250.$$

(a) Use the data to calculate a 95% confidence interval.

(b) What was the confidence interval calculated in part (a) estimating? Choose one option.

(i) The average amount spent by the 12 students on textbooks.

(ii) The average amount spent by all students at the university on textbooks.

(iii) The amount that a randomly selected student spends on textbooks.

(c) What assumptions were you making when you calculated the confidence interval in part a)? Do you think that these assumption are justified?

(d) An administrator at the university claims that the confidence interval calculated in part a) cannot be reliable since the sample was so small. Are they right? Explain.

[71] For this specific problem, the experimenters indeed checked the assumptions and did not see any flows in using ANOVA - http://oak.ucc.nau.edu/rh232/courses/EPS525/Handouts

Exercise 97.3 *Assume that the weight of cereal in a "10-ounce box" is* $N(\mu, \sigma^2)$*. To test* $H_0 : \mu = 10.1$ *against* $H_1 : \mu > 10.1$*, we take a random sample of size* $n = 16$ *and observe that* $\overline{x} = 10.4$ *and* $s = 0.4$*.*

(i.) Do we accept or reject H_0 *at the* 5% *significance level?*

(ii.) What is the approximate p-value of this test?

Exercise 97.4 *Kamil and Suphi are independently conducting statistical tests of* $H_0 : \mu = 15$ *versus* $H_1 : \mu \neq 15$*. The population variable is normally distributed and its standard deviation is not known. They each take a random sample of size 25. Kamil gets* $\overline{x} = 15.2$ *and* $s = 0.5$ *and Suphi gets* $\overline{x} = 15.1$ *and* $s = 0.2$*. Who has more compelling evidence for the alternative hypothesis?*

Exercise 97.5 *I toss a coin with a head probability* $p \in (0, 1)$ *16 times independently. For* $i = 1, \ldots, 16$*, let*

$$X_i = \begin{cases} 1, & \textit{if } i^{th} \textit{ toss is head} \\ 0, & \textit{if } i^{th} \textit{ toss is tail.} \end{cases}$$

In particular, $\sum_{i=1}^{16} X_i$ *is the total number of heads I obtain among these 16 trials.*

0. *What is the distribution of* $\sum_{i=1}^{16} X_i$*? Do not forget to include the parameters of this distribution.*

i. *Find an exact expression for the probability* $\mathbb{P}\left(\sum_{i=1}^{16} X_i > 10\right)$*.*

ii. *Consider the test* $H_0 : p = 1/2$ *vs.* $H_1 : p \neq 1/2$*. In order to compare these two hypotheses, I toss the coin 16 times independently, and decide to reject* H_0 *if I obtain more than 10 heads or less than 6 heads. Find an exact expression for the probability of a type I error.*

iii. *Use central limit theorem to approximate the probability you found in part ii.*

Exercise 97.6 *Suppose again that I find a coin, and decide to do a statistical test to see whether the coin is fair or not. That is, the hypotheses are given by*

$$\mathcal{H}_0 : p = \frac{1}{2},$$

and

$$\mathcal{H}_1 : p \neq \frac{1}{2}.$$

To do so I toss the coin 100 times and decide to reject the null hypothesis if the number of heads I obtain, say X*, satisfies* $|X - 50| > 15$*. What is the power of the test if the true parameter value is* $p = 0.6$ *Estimate the power of the test when* $p = 0.6$ *by using the central limit theorem.*

Exercise 97.7 *The mean number of credits taken by a sample of 9 statistics students was 16 and the sample standard deviation was 3. Assume that the number of credits taken by a student is normally distributed, and that it is independent of others.*

(i.) Construct a 90% *confidence interval for the mean number of credits taken by all statistics students.*

(ii.) Is the confidence interval in part i exact or approximate according to our model assumptions? Explain.

(iii.) Consider the null hypothesis $H_0 : \mu = 15.5$ *against the alternative* $H_1 : \mu \neq 15.5$*. Would you accept or reject* H_0 *with the data you have at* 90% *significance level?*

Exercise 97.8 *A point is to be selected from the unit interval* $\{x : 0 < x < 1\}$ *by a random process. Let* $A_1 = \{x : 0 < x \leq 1/4\}$*,* $A_2 = \{x : 1/4 < x \leq 1/2\}$*,* $A_3 = \{x : 1/2 < x \leq 3/4\}$*, and* $A_4 = \{x : 3/4 < x < 1\}$*. For* $i = 1, 2, 3, 4$*, suppose a certain hypothesis (*H_0*) assigns probabilities* p_{i0} *to these sets in accordance with* $p_{i0} = \int_{A_i} 2x dx$*,* $i = 1, 2, 3, 4$.

If the observed frequencies of the sets $A_i, i = 1, 2, 3, 4$ *in 80 independent experiments are 6,18,20 and 36, respectively, would* H_0 *be accepted at the* 2.5% *level of significance.*

Exercise 97.9 *A 1-pound bag of candy-coated chocolate-covered peanuts contained 224 pieces of candy colored brown, orange, green, and yellow. Test the null hypothesis that the machine filling these bags treats the four colors of candy equally likely; that is, test*

$$H_0 : p_B = p_O = p_G = p_Y = \frac{1}{4}.$$

The observed values were 42 brown, 64 orange, 53 green, and 65 yellow. You may select the significance level or give an approximate p-value.

Exercise 97.10 *Let* X *equal the number of female children in a three-child family. We shall use a chi-square goodness of fit statistic to test the null hypothesis that the distribution of* X *is* $Bin(3, 0.5)$.

(a) Define the test statistic and critical region using an $\alpha = 0.05$ *significance level.*

(b) Among students who were taking statistics, 52 came from families with 3 children. For these families, $x = 0, 1, 2,$ *and* 3 *for* $5, 17, 24,$ *and* 6 *families, respectively. Calculate the value of the test statistic and state your conclusion, considering how the sample was selected.*

Exercise 97.11 *In each of the following situations, compute the p-value of the observed data.*

a. For testing $H_0 : \theta \leq \frac{1}{2}$ *vs.* $H_1 : \theta > \frac{1}{2}$*, 7 successes are observed out of 10 Bernoulli trials.*

b. For testing $H_0 : \lambda \leq 1$ *vs.* $H_1 : \lambda > 1$*,* $X = 3$ *is observed, where* $X \sim PO(\lambda)$.

Exercise 97.12 *Let $X_1, \ldots, X_n$ be independent continuous random variables. Set m to be the median of X_1, and in particular note that m is unique by our assumption on continuity. Let $U = \min\{X_1, \ldots, X_n\}$ and $V = \max\{X_1, \ldots, X_n\}$.*

(0.) Find $\mathbb{P}(X_1 < m)$.

(i.) Find $\mathbb{P}(U < m)$.

(ii.) Show that $\mathbb{P}(U < m < V) = 1 - \left(\frac{1}{2}\right)^{n-1}$.

(iii.) Use your observation in part ii to construct a confidence interval for m which has at least $(1-\alpha)100\%$ confidence.

(iv.) In your derivation of (iii.), you had no model assumptions such as normality of the underlying distribution. Explain intuitively whether a $(1-\alpha)100\%$ confidence interval would be longer with or without model assumptions.

98 Learn more by solving problems

Exercise 98.1 *A cdf F is **stochastically greater** than a cdf F_Y if $F_X(t) \leq F_Y(y)$ for all t and $F_X(t) < F_Y(t)$ for some t. Prove that if $X \sim F_X$ and $Y \sim F_Y$, then*

$$\mathbb{P}(X > t) \geq \mathbb{P}(Y > t) \quad \textit{for every } t,$$

and

$$\mathbb{P}(X > t) > \mathbb{P}(Y > t) \quad \textit{for some } t,$$

Exercise 98.2 *A family of cdfs $\{F(x \mid \theta) : \theta \in \Theta\}$ is **stochastically increasing** in θ if $\theta_1 > \theta_2 \Rightarrow F(x \mid \theta_1)$ is stochastically greater than $F(x \mid \theta_2)$. Show that $N(\mu, \sigma^2)$ is stochastically increasing in μ for fixed σ^2.*

Exercise 98.3 *A **truncated discrete distribution** is one in which a particular class cannot be observed and is eliminated from the sample space. In particular, if X has range $0, 1, 2, \ldots$ and the 0 class cannot be observed (as is usually the case), the 0-truncated random variable X_T has pmf*

$$\mathbb{P}(X_T = x) = \frac{\mathbb{P}(X = x)}{\mathbb{P}(X > 0)}, \quad x = 1, 2, \ldots$$

Find the pmf, mean, and variance of the 0-truncated random variable starting from a Poisson random variable X with parameter λ.

Exercise 98.4 *Let the random variable X have the pdf*

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \quad 0 < x < \infty.$$

*Find the mean and variance of X. (The distribution of X is called the **folded normal distribution**.)*

Exercise 98.5 *Let Z be a random variable with pdf $f(z)$. Define z_α to be a number that satisfies this relationship:*

$$\alpha = \mathbb{P}(Z > z_\alpha) = \int_{z_\alpha}^{\infty} f(z)dz.$$

Show that if X is a random variable with pdf $(1/\sigma)f((x-\mu)/\sigma)$ and $x_\alpha = \sigma z_\alpha + \mu$, then $\mathbb{P}(X > x_\alpha) = \alpha$. (Thus if a table of z_α values were available, then values of x_α could be easily computed for any member of the location-scale family)

Exercise 98.6 *A large number, $N = mk$, of people are subject to a blood test. This can be administered in two ways.*

(i) Each person can be tested separately. In this case N tests are required.

(ii) The blood samples of k people can be pooled and analyzed together. If the test is negative, this one test suffices for k people. If the test is positive, each of the k persons must be tested separately, and, in all, $k+1$ tests are required for the k people.

Assume that the probability, p, that the test is positive is the same for all people and that the test results for different people are statistically independent.

(a) What is the probability that the test for a pooled sample of k people will be positive?

(b) Let X = number of blood tests necessary under plan (ii). Find $\mathbb{E}[X]$.

(c) In terms of minimizing the expected number of blood tests to be performed on the N people, which plan [(i) or (ii)] would be preferred if it is known that p is close to 0? Justify your answer using the expression derived in part b.

Exercise 98.7 *If the random variable X has pdf*

$$f(x) = \begin{cases} \frac{x-1}{2}, & \text{if } 1 < x < 3 \\ 0, & \text{otherwise,} \end{cases}$$

find a monotone function $u(x)$ such that the random variable $Y = u(X)$ has a uniform distribution over $(0,1)$.

Exercise 98.8 *Suppose that the pdf of a random variable X is an even function. Show that X and $-X$ are identically distributed.*

Exercise 98.9 *Let $f(x)$ be a pdf and let a be a number such that, for all $\epsilon > 0$, $f(a+\epsilon) = f(a-\epsilon)$. Such a pdf is said to be* ***symmetric*** *about the point a.*

a. *Give three examples of symmetric pdfs.*

b. *Show that if $X \sim f(x)$, symmetric, then the median of X is the number a.*

c. *Show that if $X \sim f(x)$, symmetric, and if $\mathbb{E}[X]$ exists, then $\mathbb{E}[X] = a$.*

d. *Show that $f(x) = e^{-x}$, $x \geq 0$ is not a symmetric pdf.*

e. *Show that for the pdf in part (d), the median is less than the mean.*

Exercise 98.10 *Let $f(x)$ be a pdf and let a be a number such that, if $a \geq x \geq y$ then $f(a) \geq f(x) \geq f(y)$ and, if $a \leq x \leq y$, then $f(a) \geq f(x) \geq f(y)$. Such a pdf is called **unimodal** with a mode equal to a.*

a. *Give an example of a unimodal pdf for which the mode is unique.*

b. *Give an example of a unimodal pdf for which the mode is not unique.*

c. *Show that if $f(x)$ is both symmetric and unimodal, then the point of symmetry is a mode.*

d. *Consider the pdf $f(x) = e^{-x}$, $x \geq 0$. Show that this pdf is unimodal. What is its mode?*

Exercise 98.11 *A distribution cannot be uniquely determined by a finite collection of moments. Let X have the standard normal distribution. Define a discrete random variable Y by*

$$\mathbb{P}(Y = \sqrt{3}) = \mathbb{P}(Y = -\sqrt{3}) = \frac{1}{6}, \qquad \mathbb{P}(Y = 0) = \frac{2}{3}.$$

Show that

$$\mathbb{E}[X^r] = \mathbb{E}[Y^r]$$

for $r = 1, 2, 3, 4, 5$.

Exercise 98.12 *Suppose the random variable T is the length of life of an object (possibly the lifetime of an electrical component or of a subject given a particular treatment). The **hazard function** $h_T(t)$ associated with the random variable T is defined by*

$$h_T(t) = \lim_{\delta \to 0} \frac{\mathbb{P}(t \leq T < t + \delta \mid T \geq t)}{\delta}.$$

h(t) d Thus, we can interpret $h_T(t)$ as the rate of change of the probability that the object survives a little past time t, given that the object survives to time t. Show that if T is a continuous random variable, then

$$h_T(t) = \frac{f_T(t)}{1 - f_T(t)} = -\frac{d}{dt} \ln(1 - F_T(t)).$$

Exercise 98.13 *If T is an exponential random variable with parameter β, show that its hazard function is $h_T(t) = 1/\beta$.*

Exercise 98.14 *Let $X_1, \dots, X_n$ be a random sample from a population with mean μ and variance σ^2.*

a. Show that the estimator $\sum_{i=1}^n a_i X_i$ is an unbiased estimator of μ.

*b. Among all unbiased estimators of this form (called **linear unbiased estimators**) find the one with the minimum variance, and compute its variance.*

Exercise 98.15 *Let X be a continuous, non-negative random variable. Show that*

$$\mathbb{E}[X] = \int_0^\infty (1 - F_X(x))dx$$

where $F_X(x)$ is the cdf of X.

Exercise 98.16 *Use previous problem to find the mean duration of certain telephone calls, where we assume that the duration, T, of a particular call can be described probabilistically by $\mathbb{P}(T > t) = ae^{-\lambda t} + (1-a)e^{-\mu t}$, where a, λ, μ are constants, $0 < a < 1$, $0 < \lambda, \mu < \infty$.*

Exercise 98.17 *The **Pareto distribution**, with parameters α and β, has pdf*

$$f(x) = \frac{\beta \alpha^\beta}{x^{\beta+1}}, \quad \alpha < x < \infty, \alpha > 0, \beta > 0.$$

(a) Verify that $f(x)$ is a pdf.

(b) Derive the mean and variance of this distribution.

(c) Prove that the variance does not exist if $\beta \leq 2$.

Exercise 98.18 *One of the earlier methods (not one of the better ones) of generating pseudo-random standard normal random variables from uniform random variables is to take $X = \sum_{i=1}^{12} U_i - 6$, where the U_i's are i.i.d. $U(0,1)$.*

(a) Justify the fact that X is approximately $N(0,1)$.

(b) Can you think of any obvious way in which the approximation fails?

(c) Show how good (or bad) the approximation is by comparing the first four moments.

Exercise 98.19 *An alternative form of estimation is accomplished through the **method of moments**. The method involves equating the population mean and variance to the corresponding sample mean $\overline{x}$ and sample variance s^2 and solving for the parameters, the results being the moment estimators. In the case of a single parameter, only the means are used. Give an argument that in the case of the Poisson distribution the maximum likelihood estimator and moment estimators are the same.*

A Further suggested reading

You now have the chance to do reading on various topics. Below I list some textbooks that are accessible to you.

Probability Theory

Sheldon, Ross. A first course in probability. Pearson Education India, 2002.

Rozanov, Yurii A. Probability theory: a concise course. Courier Corporation, 2013.

Walsh, John B. Knowing the odds: an introduction to probability. Vol. 139. American Mathematical Soc., 2012.

Capinski, Marek, and Tomasz Jerzy Zastawniak. Probability through problems. Springer Science & Business Media, 2013.

Mathematical finance and stochastic calculus

Shreve, Steven. Stochastic calculus for finance I: the binomial asset pricing model. Springer Science & Business Media, 2012.

Baxter, Martin, and Andrew Rennie. Financial calculus: an introduction to derivative pricing. Cambridge university press, 1996.

Capinski, Marek, and Tomasz Zastawniak. Mathematics for finance: an introduction to financial engineering. Springer, 2006.

Roman, Steven. Introduction to the Mathematics of Finance. Springer, 2004.

Stochastic processes

Ross, Sheldon M. Introduction to probability models. Academic press, 2014.

Brzezniak, Zdzislaw, and Tomasz Zastawniak. Basic stochastic processes: a course through exercises. Springer Science & Business Media, 2000.

Grimmett, Geoffrey, and David Stirzaker. Probability and random processes. Oxford university press, 2001.

Grimmett, Geoffrey, and David Stirzaker. One thousand exercises in probability. Oxford University Press, 2001.

Data analysis

Ramsey, Fred, and Daniel Schafer. The statistical sleuth: a course in methods of data analysis. Cengage Learning, 2012.

Ott, R. Lyman, and Micheal T. Longnecker. An introduction to statistical methods and data analysis. Nelson Education, 2015.

R: Statistical software

simpleR - Verzani, John, Using R for Introductory Statistics, CRC Press, 2014.

Dalgaard, Peter. Introductory statistics with R. Springer Science & Business Media, 2008.

Kerns, G. J., and G. A. Chang. "Introduction to Probability and Statistics Using R (IPSUR." 2010.

Phyton

Grus, Joel. Data science from scratch: first principles with python. " O'Reilly Media, Inc.", 2015.

VanderPlas, Jake. Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.", 2016.

Real analysis: If you are interested in theoretical statistics or theoretical probability, you will need to have a very good background in analysis. Following are a few suggested undergraduate level analysis books:

Rudin, Walter. Principles of mathematical analysis. Vol. 3. New York: McGraw-hill, 1964.

Kreyszig, Erwin. Introductory functional analysis with applications. Vol. 1. New York: wiley, 1978.

Capinski, Marek, and Peter E. Kopp. Measure, integral and probability. Springer Science & Business Media, 2013.

Other mixed reading

Advanced statistics: Casella, George, and Roger L. Berger. Statistical inference. Vol. 2. Pacific Grove, CA: Duxbury, 2002.

Randomized algorithms: Mitzenmacher, Michael, and Eli Upfal. Probability and computing: Randomized algorithms and probabilistic analysis. Cambridge university press, 2005.

Image analysis I: Barnsley, Michael Fielding. Superfractals. Cambridge University Press, 2006.

Image analysis II: Winkler, Gerhard. Image analysis, random fields and Markov chain Monte Carlo methods: a mathematical introduction. Vol. 27. Springer Science & Business Media, 2012.

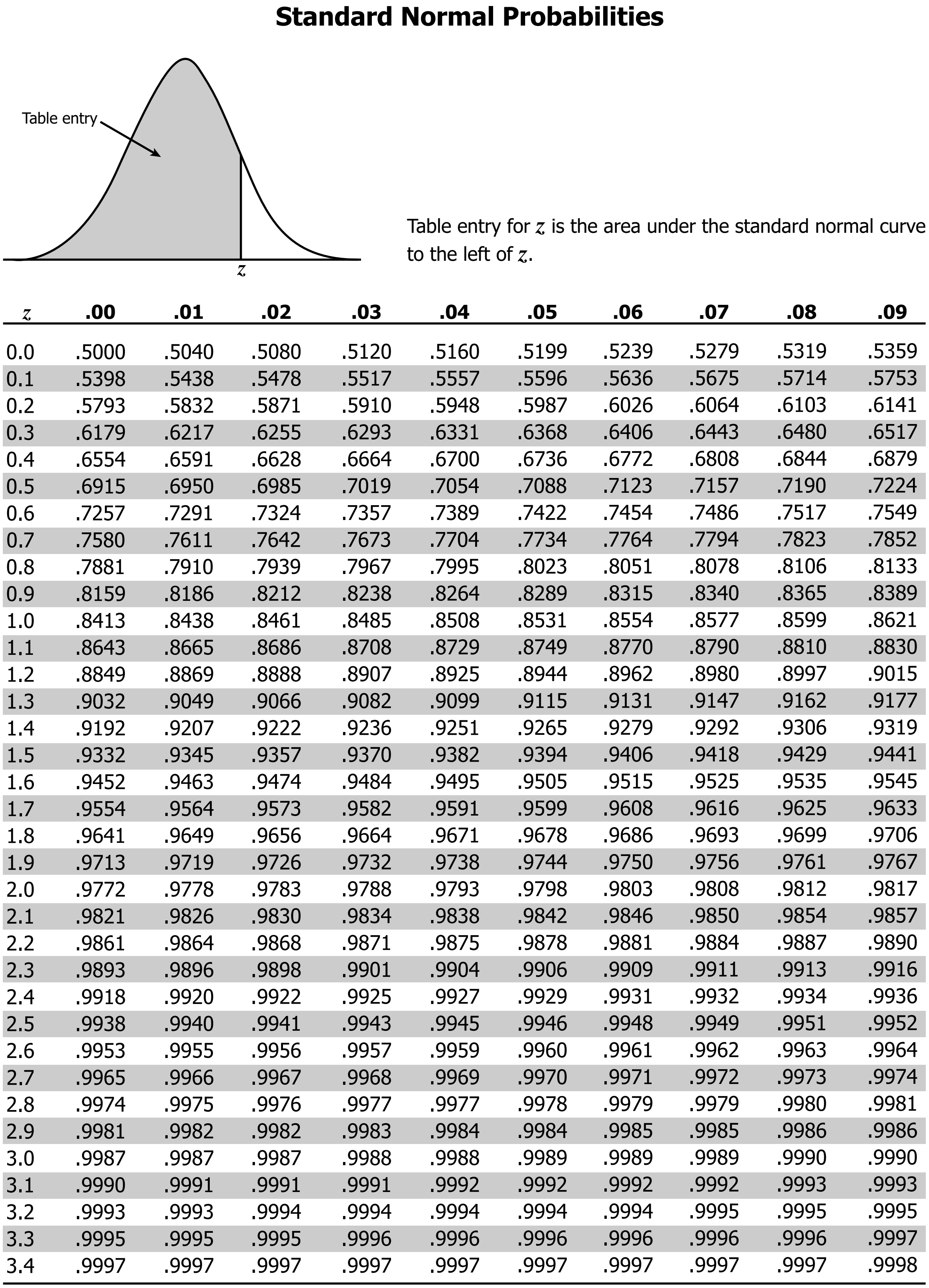
Simulation: Ross, Sheldon M. A course in simulation. Prentice Hall PTR, 1990.

Machine learning: Christopher, M. Bishop. PATTERN RECOGNITION AND MACHINE LEARNING. Springer-Verlag New York, 2016.

Monte Carlo methods: Robert, Christian P. Monte carlo methods. John Wiley & Sons, Ltd, 2004.

In case you need some specific suggetions, just talk to me.

B T



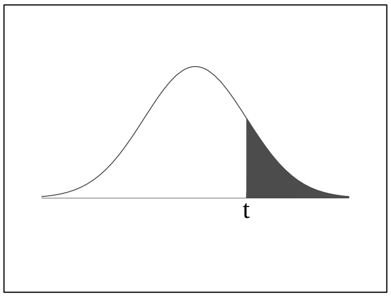
Standard Normal Probabilities

Table entry
z

Table entry for z is the area under the standard normal curve to the left of z.

z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.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

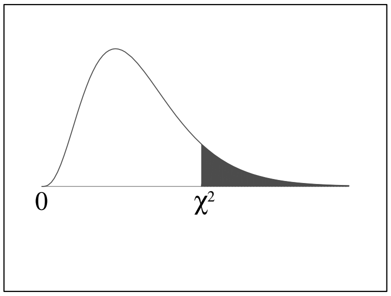
t-Distribution Table



The shaded area is equal to α for $t = t_\alpha$.

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
32	1.309	1.694	2.037	2.449	2.738
34	1.307	1.691	2.032	2.441	2.728
36	1.306	1.688	2.028	2.434	2.719
38	1.304	1.686	2.024	2.429	2.712
∞	1.282	1.645	1.960	2.326	2.576

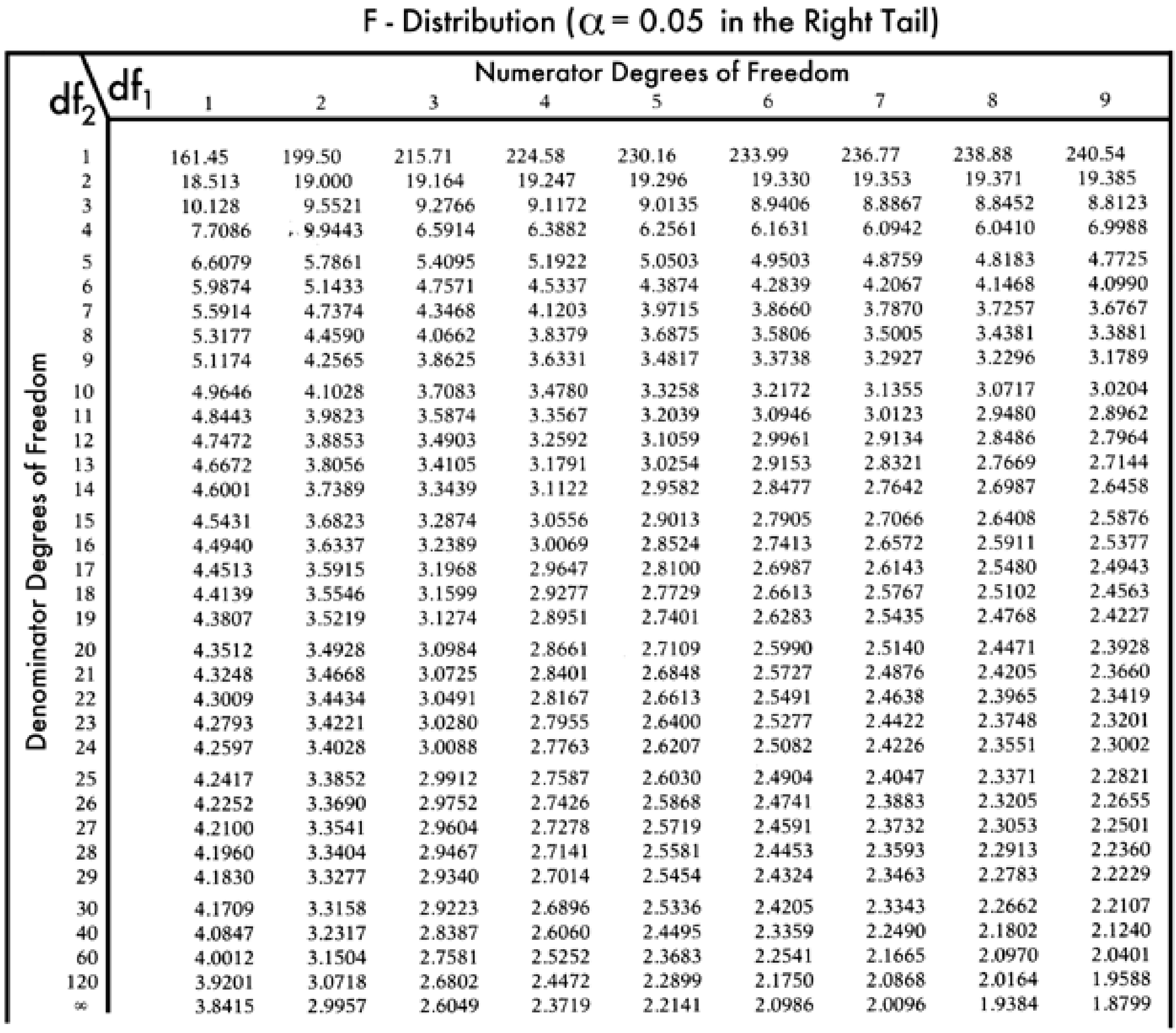
Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

F-Distribution Table



F - Distribution (α = 0.05 in the Right Tail)

Numerator Degrees of Freedom

Denominator Degrees of Freedom

df_2 \ df_1	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	7.7086	9.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799

C Some statistical quotes I like

All models are wrong, but some are useful. – **G. E. P. Box**

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem. – **J. Tukey**

I think it is much more interesting to live with uncertainty than to live with answers that might be wrong. **R. Feynman**

The best thing about being a statistician is that you get to play in everyone's backyard. – **J. Tukey**

It's easy to lie with statistics; it is easier to lie without them. – **F. Mosteller**

It is easy to lie with statistics. It is hard to tell the truth without statistics. - **A. Dunkels**

In God we trust. All others must bring data. – **W. E. Deming**

Statisticians, like artists, have the bad habit of falling in love with their models. – **G. Box**

Taking a model too seriously is really just another way of not taking it seriously at all. – **A. Gelman**

Statistics can be made to prove anything - even the truth. – **Author Unknown**

All life is an experiment. The more experiments you make, the better. – **R. W. Emerson**

The average human has one breast and one testicle. – **D. McHale**

I abhor averages. I like the individual case. A man may have six meals one day and none the next, making an average of three meals per day, but that is not a good way to live. – **L. D. Brandeis**

I could prove God statistically. Take the human body alone - the chances that all the functions of an individual would just happen is a statistical monstrosity. – **G. Gallup**

Maturity is the capacity to endure uncertainty. – **J. Finley**

"Natural selection is a mechanism for generating an exceedingly high degree of improbability." R. A. Fisher

---

9 out of ten dentists think the 10th dentist is an idiot. – **Author unknown**

Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital. – **Aaron Levenstein**

Statistics are like mini-skirts, shows a lot of things but does not show what is essentially. – **Alex Ferguson**

The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data. – **J. Tukey**

A big computer, a complex algorithm and a long time does not equal science. – **R. Gentleman**

All generalizations are false, including this one. – **M. Twain**

Statistical thinking will one day be as necessary a qualification for efficient citizenship as the ability to read and write. – **H. G. Wells**

With a large enough sample, any outrageous thing is likely to happen. **P. Diaconis and F. Mosteller**

Strange events permit themselves the luxury of occurring. – **C. Chan**

There are no routine statistical questions, only questionable statistical routines. – **D. R. Cox**

Say you were standing with one foot in the oven and one foot in an ice bucket. According to the percentage people, you should be perfectly comfortable. – **B. Bragan**

Statistics - A subject which most statisticians find difficult but in which nearly all physicians are expert. – **S. Senn**

Absence of evidence is not evidence of absence. **C. Sagan**

All we know about the world teaches us that the effects of A and B are always different—in some decimal place—for any A and B. Thus asking ”are the effects different?” is foolish. – **J. Tukey**

”... surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p?” – **R. L. Rosnow and R. Rosenthal**

Figures don’t lie, but liars do figure – **M. Twain**

Do not trust any statistics you did not fake yourself. – **W. Churchill**

I keep saying that the sexy job in the next 10 years will be statisticians. And I’m not kidding. – **H. Varian**

Those who ignore Statistics are condemned to reinvent it. – **B. Efron**

The death of one man is a tragedy. The death of millions is a statistic. – **J. Stalin**

There are three kinds of lies: lies, damned lies, and statistics. – **Author unknown**

To find out what happens when you change something, it is necessary to change it. – **Box, Hunter, and Hunter**

The greatest value of a picture is when it forces us to notice what we never expected to see. – **J. Tukey**

The statistician cannot evade the responsibility for understanding the process he applies or recommends. -– **Sir R. A. Fisher**

The Earth is round. $p < .05$ – **J. Cohen**

If I can’t picture it, I can’t understand it. – **A. Einstein**

An ecologist is a statistician who likes to be outside. – **Author unknown**

_ Anyone who considers arithmetical methods of producing random digits is, of course,

in a state of sin. – **J. V. Neumann**

The Central Limit Theorem is about the journey and the Strong Law of Large Numbers is about the destination. – **stats.SE user cardinal in a comment on this question**

Most of the time, when you get an amazing, counterintuitive result, it means you have screwed up the experiment – **M. Wigler**

If you think that statistics has nothing to say about what you do or how you could do it better, then you are either wrong or in need of a more interesting job. – **S. Senn**

---

efficiency = statistical efficiency x usage. – **J. Tukey**

Everybody knows that probability and statistics are the same thing, and statistics is nothing but correlation. Now the correlation is just the cosine of an angle, thus all is trivial. – **E. Artin**

Everybody is a Bayesian. It's just that some know it, and some don't. – **T. Raghunathan**

The best time to plan an experiment is after you've done it. – **R. A. Fisher**

People think that if you collect enormous amounts of data you are bound to get the right answer. You are not bound to get the right answer unless you are enormously smart. – **B. Efron**

Statistics is the grammar of science – **K. Pearson**

If you need statistics to prove it, it isn't true. – **Author unknown**

Facts speak louder than statistics. – **J. Streatfield**

The most important questions of life are, for the most part, really only problems of probability. – **P. Simon and M. de Laplace**

It is a capital mistake to theorize before one has data. – **Sir A. C. Doyle**

Absolute certainty is a privilege of uneducated minds-and fanatics. It is, for scientific folk, an unattainable ideal. –**C. J. Keyser**

The aim of science is to seek the simplest explanation of complex facts... Seek simplicity and distrust it. **A. N. Whitehead**

He uses statistics as a drunken man uses lampposts — for support rather than for illumination. – **A. Lang**

Statistics may be defined as ”a body of methods for making wise decisions in the face of uncertainty.” – **W. A. Wallis**

D Basic combinatorics review

We begin by recalling the basic counting principle which is also known as the multiplication rule.

Theorem D.1 *(Basic counting principle) Suppose that an experiment has k parts ($k \geq 2$), that the i^{th} part of the experiment can have n_i possible outcomes ($i = 1, \ldots, k$), and that all of the outcomes in each part can occur regardless of which specific outcomes have occurred in the other parts. Then the sample space S of the experiment will contain all vectors of the form $(u_1, \ldots, u_k)$, where u_i is one of the n_i possible outcomes of part i ($i = 1, \ldots, k$). The total number of these vectors in S will be equal to the product $n_1 n_2 \ldots n_k$.*

Here is an example:

Example D.1 *A well known nursery rhyme starts as follows:*

As I was going to St. Ives
I met a man with 7 wives.
Each wife had 7 sacks.
Each sack had 7 cats.
Each cat had 7 kittens.

How many kittens did the traveler meet?

Solution: Traveler met 7^4 kittens by using the basic counting principle.

The reader is assumed to know about the basics of permutations and combinations, so here I just include the definitions, the binomial theorem - you will see some problems in Combinatorics Review Problems section.

Definition D.1 *Suppose a set has n elements. Suppose an experiment consists of selecting k elements one at a time without replacement. Let each outcome consist of k elements in the order selected. Each such outcome is called a **permutation** of n elements taken k at a time. We denote the number of all distinct permutations by $P_{n,k}$.*

Definition D.2 *Given a set with n elements, each subset of size k is called a **combination** of n elements taken k at a time. The number of all such combinations is denoted by $C_{n,k}$.*

Remark D.1 *(1) Note that $P_{n,k} = C_{n,k} k!$. It is then immediate that*

$$C_{n,k} = \frac{n!}{k!(n-k)!}.$$

*(2) We will mostly use the notation $\binom{n}{k}$ instead of $C_{n,k}$. $\binom{n}{k}$ will be called a **binomial coefficient**.*

Theorem D.2 *(Binomial theorem) For any real numbers x, y, and any $n \in \mathbb{N}$*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Exercise D.1 *Prove the binomial theorem.*

The discussion on binomial coefficients can be generalized to count the number of ways to partition a finite set into more than two disjoint sets.

Example D.2 *Suppose that 20 members of an organization are to be divided into three committees A,B and C in such a way that A and B will have 8 members and C will have 4 members. Then, the number of different ways in which members can be assigned to these committees is given by*

$$\binom{20}{8}\binom{12}{8}\binom{4}{4}.$$

We now generalize the idea in previous example: Suppose that n distinct elements are to be divided into $k \geq 2$ different groups in such a way that, for $j = 1, \ldots, k$, the j^{th} group contains exactly n_j elements, and that $n_1 + \cdots + n_k = n$. Our goal is to determine the number of distinct ways of dividing these n into the k groups.

Observe that

n_1 elements in first group can be chosen in $\binom{n}{n_1}$ ways.
n_2 elements in second group can be chosen in $\binom{n-n_1}{n_2}$ ways.
$\cdots$
n_k elements in k^{th} group can be chosen in $\binom{n-n_1-\cdots-n_{k-1}}{n_k} = \binom{n_k}{n_k}$ ways.

Hence, the total number of different ways of dividing n elements into k groups is given by

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-\ldots-n_{k-1}}{n_k} = \cdots = \frac{n!}{n_1!\cdots n_k!}.$$

Definition D.3 *The number $\frac{n!}{n_1!\cdots n_k!}$, which we denote by $\binom{n}{n_1,\ldots,n_k}$, is called a* ***multinomial coefficient****.*

The following result generalizes the binomial theorem.

Theorem D.3 *(Multinomial theorem) For given real numbers $x_1, \dots, x_k$, and for each nonnegative integer n,*

$$(x_1 + \cdots + x_k)^n = \sum \binom{n}{n_1, \dots, n_k} x_1^{n_1} \dots x_k^{n_k},$$

where the summation extends over all possible combinations of nonnegative integers $n_1, \dots, n_k$ such that $n_1 + \cdots + n_k = n$.

problem D.1 *Prove the multinomial theorem.*

Example D.3 *Assume that a balanced die is rolled 12 times. What is the probability that all possible 6 outcomes will occur exactly twice?*

Solution: The required probability is

$$p = \frac{\binom{12}{2,\cdots,2}}{6^{12}} = \frac{12!}{(2!)^6 6^{12}} = 0.0034.$$

Example D.4 *If the letters s,s,s,t,t,t,i,i,a,c are arranged in a random order, what is the probability that they will spell the word "statistics"?*

Solution: There are $\binom{10}{3,3,2,1,1}$ different words one may write by using these letters. Thus, the required probability is $\frac{1}{\binom{10}{3,3,2,1,1}}$.

Example D.5 *A deck of 52 cards contains 13 hearts. Suppose that the deck is shuffled and distributed among A, B, C and D so that each player receives 13 cards. Find the probability p that player A receives six hearts, B receives four hearts, C receives two hearts and D receives one heart.*

Solution: The total number of ways in which 52 cards can be distributed among four players is $N = \binom{52}{13,13,13,13} = \frac{52!}{(13!)^4}$.

The number of ways in which the hearts can be distributed to players A, B, C and D so that the number of hearts they receive is 6,4,2 and 1, respectively, is $\binom{13}{6,4,2,1} = \frac{13!}{6!4!2!1!}$. Also the number of ways in which the other 39 cards can be distributed is $\binom{39}{7,9,11,12} = \frac{39!}{7!9!11!12!}$. Thus,

$$p = \frac{\frac{13!39!}{6!4!2!1!7!9!11!12!}}{\frac{52!}{(13!)^4}}.$$

Example D.6 *A deck of 52 cards contains four aces. If the cards are shuffled and distributed in a random manner to four players so that each player receives 13 cards, what is the probability that all four aces will be received by the same player?*

Solution 1: (Multinomial solution) The first approach is to consider that there are four players A, B, C and D. The probability that player A receives all four aces is given by

$$\frac{\binom{48}{9}\binom{39}{13}\binom{26}{13}\binom{13}{13}}{\binom{52}{13,13,13,13}}.$$

Since each player has the same probability to have all aces, and since the events are disjoint, the probability in question is given by

$$4\frac{\binom{48}{9}\binom{39}{13}\binom{26}{13}\binom{13}{13}}{\binom{52}{13,13,13,13}}.$$

Solution 2: (Binomial solution) Again let's first consider the probability of player A having all the aces. This time we consider only two players A and E, where player E receives 39 cards (representing cards of B, C and D). Then this time the probability that the player A receives all four aces is given by

$$\frac{\binom{48}{9}\binom{39}{39}}{\binom{52}{13}},$$

and following the reasoning in previous approach, the required probability is

$$4\frac{\binom{48}{9}\binom{39}{39}}{\binom{52}{13}}.$$

□

E Some combinatorics review problems

Exercise E.1 *Consider a group of 20 people. If everyone shakes hands with everyone else, how many handshakes take place? (Answer:* $\binom{20}{2}$*.)*

Exercise E.2 *How many vectors* $(x_1, ..., x_k)$ *are there for which each* x_i *is a positive integer such that* $1 \leq x_i \leq n$ *and* $x_1 < x_2 < \cdots < x_k$*?*

Exercise E.3 *Show that* $\sum_{k=0}^{n} \binom{n}{k} = 2^n$*.*

Exercise E.4 *(i) Show that* $\sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0$*. (Hint: Use binomial theorem.)*
(ii.) Show that

$$\sum_{k=0,\ k\ even}^{n} \binom{n}{k} = \sum_{k=0,\ k\ odd}^{n} \binom{n}{k}.$$

Exercise E.5 *How many subsets of size 4 of the set* $S = \{1, 2, \ldots, 20\}$ *contain at least one of the elements* $1, 2, 3, 4, 5$*?*

Exercise E.6 *How many different 7-place license plates are possible when 3 of the entries are letters and 4 are digits? (Assume that (i.) there are 26 letters, 10 digits, (ii) repetition of letters and numbers is allowed, (iii) there is no restriction on where the letters or numbers can be placed.)*

Exercise E.7 *How many different linear arrangements are there of the letters A, B, C, D, E, F for which*

a. A and B are next to each other?

b. A is before B?

c. A is before B and B is before C?

d. A is before B and C is before D?

e. A and B are next to each other and C and D are also next to each other?

f. E is not last in line?

Exercise E.8 *If 4 balanced dice are rolled, what is the probability that the number 5 and the number 6 will appear the same number of times?*

Exercise E.9 *A president, treasurer, and secretary, all different, are to be chosen from a club consisting of 10 people. How many different choices of officers are possible if* *i. there are no restrictions?*

ii. A and B will not serve together?

iii. C and D will serve together or not at all?

iv. E must be an officer?

v. F will serve only if he is president?

Exercise E.10 *Let n and k be two positive integers so that $n \geq k$. Determine the number of vectors $(x_1, ..., x_n)$, such that each x_i is either 0 or 1 and*

$$\sum_{i=1}^{n} x_i \geq k.$$

Exercise E.11 *In how many ways can 10 people be seated in a row if*

i. there are no restrictions on the seating arrangements;

ii. persons A and B will sit next to each other;

iii. persons A and B will not sit next to each other;

iv. there are 5 men and 5 women and no 2 men and no 2 women can sit next to each other;

v. there are 6 women and they must sit next to each other;

vi. there are 6 women and no two women can sit next to each other;

vii. there are 5 married couples and each couple must sit together?

Exercise E.12 *(Vandermonte's identity) Prove (by using a combinatorial argument) that*

$$\binom{n+m}{r} = \binom{n}{0}\binom{m}{r} + \binom{n}{1}\binom{m}{r-1} + \cdots + \binom{n}{r}\binom{m}{0}.$$

Hint : Consider a group of n men and m women. How many groups of size r are possible?

Exercise E.13 *Show that $\binom{2n}{n} = \sum_{k=0}^{n} \binom{n}{k}^2$. (Hint: Use Vandermonte's identity.)*

Exercise E.14 *How many distinct integer-valued vectors $(x_1, x_2, \ldots, x_r)$ exist that satisfy*

$$x_1 + x_2 + \cdots + x_r = n, \quad x_i > 0, \quad i = 1, ..., r?$$

Exercise E.15 *How many distinct integer-valued vectors* $(x_1, x_2, \ldots, x_r)$ *exist that satisfy*

$$x_1 + x_2 + \cdots + x_r = n, \quad x_i \geq 0, \quad i = 1, ..., r?$$

Exercise E.16 *Consider a function* $f(x_1, \ldots, x_n)$ *of* n *variables. How many different partial derivatives of order* r *does it possess? (You can assume that* f *is as smooth (in terms of differentiability) as you like.)*

Exercise E.17 *If 6 balls are thrown at random into 12 boxes, what is the probability that no box will receive more than one ball?*

Exercise E.18 *(Pairs of shoes) A closet has* $2n$ *pair of shoes where* $n \geq 1$ *is an even number. If* n *shoes are randomly selected, what is the probability that*

a. *there is no complete pair?*

b. *we have exactly* k *pairs?*

Exercise E.19 *A certain group has eight members. In January, three members are selected at random to serve on a committee. In February, four members are selected at random and independently of the first selection to serve on another committee. In March, five members are selected at random and independently of the previous two selections to serve on a third committee. Determine the probability that each of the eight members serves on at least one of the three committees.*

References

[1] Casella, George, and Roger L. Berger. Statistical inference. Vol. 2. Pacific Grove, CA: Duxbury, 2002.

[2] Dalgaard, Peter. Introductory statistics with R. Springer Science & Business Media, 2008.

[3] DeGroot, Morris H., and Mark J. Schervish. Probability and statistics. Pearson Education, 2012.

[4] Grimmett, Geoffrey, and David Stirzaker. Probability and random processes. Oxford university press, 2001.

[5] Grimmett, Geoffrey, and David Stirzaker. One thousand exercises in probability. Oxford University Press, 2001.

[6] Hogg, Robert V., and Allen T. Craig. Introduction to mathematical statistics.(5"" edition). Upper Saddle River, New Jersey: Prentice Hall, 1995.

[7] Hogg, Robert V., and Elliot A. Tanis. "A Brief Course in Mathematical Statistics." (2008).

[8] Kerns, G. J., and G. A. Chang. "Introduction to Probability and Statistics Using R (IPSUR." 2010.

[9] Marsaglia, George. "Random numbers fall mainly in the planes." Proceedings of the National Academy of Sciences 61.1 (1968): 25-28.

[10] Walpole, Ronald E., et al. Probability and statistics for engineers and scientists. Vol. 5. New York: Macmillan, 1993.

[11] Sheldon, Ross. A first course in probability. Pearson Education India, 2002.

[12] Rozanov, Yurii A. Probability theory: a concise course. Courier Corporation, 2013.

[13] Seltman, Howard J. "Experimental design and analysis." Online at: http://www. stat. cmu. edu/, hseltman/309/Book/Book. pdf (2012).

[14] Starfield, Tony. Discussion: Deterministic or Stochastic, recording, full version available at `http://www.uvm.edu/~tdonovan/modeling/Module5/05_DeterministicStochastic_transcript.pdf`, 2005.

[15] Starfield, Tony. Five Questions for Population Modeling, recording, full version available at `http://www.uvm.edu/~tdonovan/modeling/Module8/00_FiveQuestionsPop_transcript.pdf`, 2009.