Lecture notes for STA 130A

1 Probability

Probability theory is concerned with outcomes that occur randomly in an experiment. For example, if we toss a fair coin, then each side of the coin is equally likely to come up. The coin toss is an experiment and heads and tails are outcomes. These outcomes are random because we cannot predict with 100% accuracy if heads or tails comes up before the coin is tossed. Instead, we can predict that if we repeat the experiment many times, then about 50% of the times we will get heads (and other 50% of the times we will get tails). We say that tails occurs with probability 1/2 and heads occurs also with probability 1/2. In this section, we learn how to make this intuitive discussion mathematically rigorous.

1.1 Sample space

Definition 1.1 (Sample space)**.** A sample space is the set of all possible outcomes of an experiment. The sample space is denoted by Ω and an element of Ω (an outcome) is denoted by ω.

Example 1.2. We consider a few experiments and corresponding sample spaces.

- In the coin toss experiment, $\Omega = \{Heads, Tails\}$.
- Commuter passes through three intersections. At each one, she can stop (s) or continue (c).
$$\Omega = \{ccc, ccs, css, csc, sss, ssc, scc, scs\}.$$
- How many days will it rain in 2019?
$$\Omega = \{0, 1, 2, \dots, 365\}.$$
- Heights
$$\Omega = \{t : t \geq 0\}.$$

Some outcomes are more interesting than others. Usually we are only interested in certain subsets of Ω. These are *events*. In probability theory, sets and events are the same thing.

- If we let A denote the event that the commuter stops at the first light, then
$$A = \{sss, ssc, scc, scs\}.$$

- What is the event that there are more than 10 days of rain in 2019?

$$A = \{11, 12, \ldots, 365\}.$$

- NBA basketball players' heights

$$A = \{t : t \geq 5 \ ft \ 3 \ in \ (\text{Muggsy Bogues} : 1987 - 2001)\}$$

1.2 Basic set theory

If A is a subset of Ω, then we write $A \subset \Omega$. The *union* of two events A and B is the set of outcomes ω belonging to A or B

$$A \cup B = \{\omega : \omega \in A, \text{ or } \omega \in B\}.$$

The *intersection* of two events is the set of outcomes ω belonging to A and B

$$A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}.$$

The *complement* of an event A is the set of outcomes not belonging to A, denoted

$$A^c = \{\omega : \omega \notin A\}.$$

If A and B are events, then

$$A \setminus B = \{\omega : \omega \in A \text{ and } \omega \notin B\}.$$

The empty set is the set that has no elements. It is denoted as $\emptyset$. For instance, if two sets A and B have no elements in common, then

$$A \cap B = \emptyset.$$

For example, if A is the event it rains more than 10 days in 2019, and B is the event it rains less than 5 days. You can also check that for any set A, we have $A \cap A^c = \emptyset$. When two events have no outcomes in common (empty intersection), then they are *disjoint*.

The following are some laws of set theory.

- Commutative law:

$$A \cup B = B \cup A$$

$$A \cap B = B \cap A$$

- Associative law:

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

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

- Distributive laws:

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

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

- DeMorgan's laws:

$$\Big(\bigcup_{i=1}^{n} A_i\Big)^c = \bigcap_{i=1}^{n} A_i^c.$$

  and

$$\Big(\bigcap_{i=1}^{n} A_i\Big)^c = \bigcup_{i=1}^{n} A_i^c.$$

1.3 Probability measure

So far, we are able to talk about outcomes and events. But how do we talk about the probability of an event happening? Intuitively, we can first assign a probability to each outcome and then calculate the probability of an event by summing up the probabilities of all outcomes contained in that event.

Thus, a *probability measure* $\mathbb{P}$ on Ω is a function that takes in a subset $A \subset \Omega$ as input and returns a non-negative real number between zero and one. It satisfies the following axioms:

1. $\mathbb{P}(\Omega) = 1$
2. If $A \subset \Omega$, then $0 \leq \mathbb{P}(A) \leq 1$.
3. If A_1 and A_2 are disjoint, then

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2).$$

   More generally, if $A_1, A_2, \ldots,$ are disjoint, then

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Example 1.3. Suppose a coin is thrown twice. Then,

$$\Omega = \{hh, ht, th, tt\}.$$

Suppose each outcome in Ω is equally likely. Write $A_1 = \{hh\}$, $A_2 = \{ht\}$, $A_3 = \{th\}$, and $A_4 = \{tt\}$. Hence, our assumption is $P(A_1) = P(A_2) = P(A_3) = P(A_4)$. Note that this requires all probabilities are equal to 1/4, since these events are disjoint

$$1 = \mathbb{P}(\Omega) = P(A_1) + P(A_2) + P(A_3) + P(A_4) = 4P(A_1).$$

Hence $P(A_1) = 1/4$ and also $P(A_i) = 1/4$ for $i = 1, \ldots, 4$.

We can use this result to calculate other probabilities. Let C be the event that heads comes up on the first toss or on the second toss, i.e.

$$C = \{ht\} \cup \{hh\} \cup \{th\}.$$

By disjointness $P(C) = \mathbb{P}(\{ht\} \cup \{hh\} \cup \{th\}) = P(\{ht\}) + P(\{hh\}) + P(\{th\}) = 3/4$.

We list some properties of a probability measure. They are direct consequences of the axioms.

- $P(A^c) = 1 - P(A)$.
  Proof. Write $\Omega = A \cup A^c$. Then $1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$.
- $P(\emptyset) = 0$.
  Proof. Note that $\emptyset = \emptyset \cup \emptyset$, and $\emptyset \cap \emptyset = \emptyset$. Then

  $$\mathbb{P}(\emptyset) = \mathbb{P}(\emptyset \cup \emptyset) = P(\emptyset) + P(\emptyset) = 2P(\emptyset)$$

  .
  Hence $2P(\emptyset) = P(\emptyset)$.
- If $A \subset B$, then $P(A) \leq P(B)$
  Proof. Write $B = C \cup A$, where we define $C = B \setminus A$. Note that $C \cap A = \emptyset$. Hence

  $$P(B) = P(C \cup A) = P(C) + P(A) \geq P(A),$$

  since $\mathbb{P}(C) \geq 0$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
  Proof: Try on your own; or read the book.

1.4 Computing probability: counting method

Suppose the space of outcomes is finite with N possible outcomes

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}.$$

When all outcomes occur with equal probability (i.e. each outcome $\omega_i \in \Omega$ has probability $\mathbb{P}(\omega_i) = 1/N$), computing probabilities of an event can be reduced to counting the number of outcomes the event contains. Indeed, if A contains k outcomes (we say A can occur in k ways) then we sum up probabilities of those k elements (each equal $1/N$) to get $\mathbb{P}(A) = k/N$. Putting it in another way,

$$\mathbb{P}(A) = \frac{\text{number of ways } A \text{ can occur}}{\text{total number of outcomes}}.$$

We have used the axiom 3 of probability measure for the simple case when all outcomes occur with the same probability. In general, if $\mathbb{P}(\omega_i) = p_i$ then $\mathbb{P}(A)$ is the sum of all p_i such that ω_i is contained in A. That is,

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i.$$

We've seen that being able to count the number of ways an event can occur is very important in calculating probabilities. The following principle helps us to count possibilities.

Multiplication principle. Suppose there are two experiments to be performed. If the first experiment has m outcomes, and if for each outcome of the first experiment, the second experiment has n outcomes, then there are $m \cdot n$ possible outcomes for the two experiments.

Example. A certain type of shoe comes in 12 sizes and 18 colors. The number of versions of the shoe is $12 \times 18 = 216$.

Extended multiplication principle. If there are k experiments, and the first has n_1 possible outcomes, the second has n_2 possible outcomes,..., and the kth has n_k possible outcomes, then there are a total of

$$n_1 \times n_2 \times \cdots \times n_k$$

possible outcomes for the k experiments.

Example. How many different 8 bit words are there? Consider a word of the form

$$b_1 b_2 b_3 b_4 b_5 b_6 b_7 b_8$$

Each $b_i = 0$ or 1 for $i = 1, \ldots, 8$. Hence, there are $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 2^8 = 256$ words.

Example. A DNA molecule is a sequence of nucleotides, denoted by four possible letters A, T, C, G. The number of possible sequences of length one million is 4^{10^6}.

1.4.1 Permutations

Definition. A permutation is an ordered arrangement of things. For example, $(1, 2, 3)$ and $(2, 3, 1)$ are different permutations of the numbers 1,2, and 3.

Example 1.4. How many passwords can be generated using the letters E, A, S, Y?

Lemma 1.5. *There are $n \cdot (n-1) \cdot (n-2) \cdots 1 = n!$ permutations of n distinct items.*

Proof. Let's think about the process of forming a permutation as drawing labeled items out of a bag. In the first step, we draw one item and set it aside. We have n ways of doing this.

In the second step, we drawn another item and set it aside. But now, we have $n-1$ ways of doing this. We continue until there is only 1 item left.

Hence, the extended multiplication principle says that the number of possible permutations is

$$n \cdot (n-1) \cdot (n-2) \cdots 1 = n!$$

and the lemma is proved. □

Remark (zero factorial). By convention, we set $0! = 1$.

1.4.2 Sampling without replacement

When we proved that there are $n!$ permutations of n items, we used the idea of "sampling without replacement", i.e. once an item is picked, it is no longer available for later use.

Note that we use all n available items to form a permutation of n item. How many ways can we pick/sample r items without replacement from the set of n item? (where $1 \leq r \leq n$ and the ordering of the sample matters). Answer: $n \cdot (n-1) \cdot (n-(r-1))$.

1.4.3 Sampling with replacement

Suppose we have a bag of n items. How many ways can we sample r items with replacement (where $r \leq n$, and the ordering of the sample matters)? Answer: $\underbrace{n \cdot n \cdots n}_{r\text{times}} = n^r$.

1.4.4 Combinations

Definition 1.6. A combination is an arrangement of items where order doesn't matter. For instance, $\{1, 2, 3\} = \{2, 3, 1\}$ are the same combination of the numbers 1, 2, 3.

Example 1.7. Three teams play against one another exactly once in a tournament. How many matches are there? What if the order matters, for example home and away games?

Lemma 1.8. *There are $\frac{n!}{(n-r)!r!}$ combinations of r items from a set of n items.*

Proof. Let x be the number of combinations. For each unordered sample, we can make $r!$ ordered samples. Also, every ordered sample can be obtained in this way. Therefore $x \times r!$ is the number of r items sampled without replacement from n items. Therefore

$$x = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!}$$

and the lemma is proved. □

Remark. The number $\frac{n!}{(n-r)!r!}$is denoted by a special symbol $\binom{n}{r}$ – called the binomial coefficient.

Example There are 18 professors in the statistics department, and 3 need to be chosen for a special committee on admitting masters students. How many possible committees are there? Answer: $\binom{18}{3} = 816$.

1.4.5 Multinomial coefficient

The binomial coefficient $\binom{n}{k}$ tells us how many ways we can form a subset of k items of a set of n items. How many ways can we form two subsets from the set of n items, one contains k items and the other contains $n-k$ items? The answer is still $\binom{n}{k}$ because once a subset of k items is chosen, we only have one choice for the other subsets of $n-k$ items.

What about forming multiple subsets from a set of n items?

Example 1.9. Suppose you run a business of 9 employees, and you need to assign 2 employees to work in the morning shift, 3 employees to work in the afternoon, and 4 employees to work at night.

How many ways can this be done? There is a formula that answers this question known as the multinomial coefficient, but we will just derive it in this special case.

Let's look at the morning shift first. How many ways can we assign 2 of the 9 workers to the morning shift? Answer: $\binom{9}{2}$.

Now, of the 7 workers that remain, how many ways can we allocate 3 of them to the afternoon? Answer: $\binom{(9-2)}{3}$.

Finally, of the 4 workers that remain, how many ways can we allocate 4 of them to the night shift? Answer: $\binom{(9-2-3)}{4} = \binom{4}{4} = 1$.

Now, using the multiplication principle, the total number of allocations is

$$\binom{9}{2} \times \binom{(9-2)}{3} \times \binom{(9-2-3)}{4} = \frac{9!}{2!3!4!}$$

General formula. In general, if we have n items, and we want to allocate them to r subsets of sizes $n_1, n_2, \ldots, n_r$ such that $n_1 + n_2 + \cdots n_r = n$, then the number of ways this can be done is equal to

$$\frac{n!}{n_1!n_2!\cdots n_r!} = \binom{n}{n_1, n_2, \cdots n_r} = \text{ the multinomial coefficient.}$$

Example 1.10. How many distinct 10-letter sequences can be spelled from the letters S,T,A,T,I,S,T,I,C,S?

Let's think about it in the following way. We have 10 positions to allocate to the possible letters S,T,A,I,C.

How many are in the "S group"? Ans: 3
How many are in the "T group"? Ans: 3
How many are in the "A group"? Ans: 1
How many are in the "I group"? Ans: 2
How many are in the "C group"? Ans 1

So, the multinomial coefficient tells us the number of distinct 10 letter sequences is

$$\binom{10}{3, 3, 1, 2, 1}.$$

2 Conditional probability

Conditional probability refers to the probability of an event, given the information that some other event has already occurred.

Example 2.1. Your friend rolls a dice without showing it to you. He then tells you that the value is at least 3 (assuming that he tells the truth). Find the probability that the value is 5.

Note that if the additional information is not provided then the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $\mathbb{P}(D = 5) = 1/6$. With the information that D is at least 3, we can narrow down the set of possible outcomes to $\{3, 4, 5, 6\}$. Therefore the probability that the outcome takes value 5 is $1/4$.

We can think of the additional information (the value is at least 5) as an event that has occurred already. This event is $E = \{3, 4, 5, 6\}$, which can be thought of as a new sample space. We use the following notation to denote the conditional probability that $D = 5$ given that E has occurred (or just given E for short):

$$\mathbb{P}(D = 5|E) = 1/4.$$

Let D_1 and D_2 be values rolled on dice 1 and dice 2. What is the probability that $D_1 = 2$ given that $D_1 + D_2 \leq 5$?

We denote this probability by $P(D_1 = 2|D_1 + D_2 \leq 5)$. By drawing a table, it is easy to see that the event that $D_1 + D_2 \leq 5$ contains 10 possible outcomes. Among them, three outcomes have

$D_1 = 2$. Thus

$$\mathbb{P}(D_1 = 2 | D_1 + D_2 \leq 5) = \frac{3}{10} = \frac{3/36}{10/36} = \frac{\mathbb{P}(D_1 = 2 \textbf{ and } D_1 + D_2 \leq 5)}{\mathbb{P}(D_1 + D_2 \leq 5)}.$$

This motivates the following definition.

Definition 2.2. Let A and B be two events with $P(B) \neq 0$. Then, the conditional probability of A given B is defined to be

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

Multiplication law. If A and B are events with $P(B) \neq 0$, then

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

Example 2.3. An urn contains three red balls and one blue ball. Two balls are subsequently drawn without replacement from the urn. What is the probability that both of them are red?

The probability that the first ball is red is 3/4. Given that the first ball is red, there are 2 red balls and one blue ball left in the urn, and the probability that the second ball is red is 2/3. Therefore the probability that both of them are red is $3/4 \times 2/3 = 1/2$ (why do we multiply those probabilities?).

Let's make the argument above rigorous using the multiplication law. Denote

$$R_1 = \{\text{red ball drawn on first trial}, \qquad R_2 = \{\text{red ball drawn on second trial}\}.$$

Then

$$P(R_1) = \frac{3}{4}, \qquad P(R_2|R_1) = \frac{2}{3}.$$

Therefore by the multiplication law we have

$$P(R_1 \cap R_2) = P(R_1)P(R_2|R_1) = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}.$$

Law of total probability. Let $B_1, \ldots, B_n$ be some events such that $\cup_{i=1}^n B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for all $i \neq j$. Then for any event A we have

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

Example 2.4. Suppose occupations are grouped into Upper Middle and Lower. The following table of conditional probabilities is calculated from observations:

	U_2	M_2	L_2
U_1	0.45	0.48	0.07
M_1	0.05	0.7	0.25
L_1	0.01	0.5	0.49

Here, U_1 is the event that father's occupation is in Upper, U_2 is the event that son's occupation is in Upper, etc. For example, $P(U_2|U_1) = 0.45$.

We are also given the information that for the father's generation, the occupation probabilities are:

$$P(U) = 0.1, \quad P(M) = 0.4 \quad P(L) = 0.5.$$

Using this data, what is the probability that a son in the next generation will have an occupation in U?

Use law of total probability, we get

$$\begin{aligned} P(U_2) &= P(U_2|U_1)P(U_1) + P(U_2|M_1)P(M_1) + P(U_2|L_1)P(L_1) \\ &= 0.45 \times 0.10 + 0.05 \times 0.4 + 0.01 \times 0.5 \\ &= 0.07. \end{aligned}$$

What if we know the son's occupation and we want to calculate the probability that the father's occupation is in Upper? For example, what is $P(U_1|U_2)$?

From the multiplication law we have

$$P(U_1|U_2)P(U_2) = P(U_1 \cap U_2) = P(U_2|U_1)P(U_1).$$

Therefore we get

$$\begin{aligned} \mathbb{P}(U_1|U_2) &= \frac{\mathbb{P}(U_2|U_1)\mathbb{P}(U_1)}{\mathbb{P}(U_2)} \\ &= \frac{\mathbb{P}(U_2|U_1)\mathbb{P}(U_1)}{\mathbb{P}(U_2|U_1)\mathbb{P}(U_1) + \mathbb{P}(U_2|M_1)\mathbb{P}(M_1) + \mathbb{P}(U_2|L_1)\mathbb{P}(L_1)} \end{aligned}$$

This is a special case of the following theorem

Theorem 2.5 (Bayes Theorem I)**.** *For any events A and B with $P(B) > 0$ and $P(A) > 0$,*

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

Theorem 2.6 (Bayes Theorem II)**.** *Let A and $B_1, \ldots, B_n$ be events with $P(A) > 0$ and $P(B_i) > 0$ for all $i = 1, \ldots, n$. Also assume the $B_i \cap B_j = \emptyset$ for $i \neq j$ and $\Omega = \cup_{i=1}^{n} B_i$.*

Then, for any $j = 1, \dots, n$,

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}.$$

Example 2.7 (False positive rates of lie detector tests)**.** If the lie detector test is positive, what is the chance that the person is actually telling the truth? Define

$$\begin{aligned} L &= \{\text{subject is lying}\}, \\ T &= \{\text{subject is telling truth}\}, \\ D_+ &= \{\text{lie detector test is positive}\}, \\ D_- &= \{\text{lie detector test is negative}\}. \end{aligned}$$

According to research on lie detector tests

$$P(D_+|L) = 0.88, \quad \text{(which also gives } P(D_-|L) = 0.12\text{)}$$

$$P(D_-|T) = 0.86, \quad \text{(which also gives } P(D_+|T) = 0.14\text{)}$$

Let's suppose that the person is very likely to be honest, i.e. we assume

$$P(T) = 0.99 \text{ (which also gives } P(L) = 0.01\text{)}.$$

From the data and the assumption, we can calculate $P(T|D_+)$ using Bayes theorem:

$$\begin{aligned} P(T|D_+) &= \frac{P(D_+|T)P(T)}{P(D_+|T)P(T) + P(D_+|L)P(L)} \\ &= \frac{0.14 \times 0.99}{0.14 \times 0.99 + 0.88 \times 0.01} \\ &= 0.94. \end{aligned}$$

Is this good news or bad news?

3 Independent events

Intuitively what does it mean when we say that two events are independent? Roughly speaking, two events are independent if knowing that one event occurred gives no information about the other event. For example, if a fair coin is tossed twice the event that the first toss is heads is independent of the event that the second toss is heads

One way of formalizing the idea that B gives no information about A is to say that $P(A|B) = P(A)$. It is equivalent to the following definition.

Definition 3.1 (Independent events)**.** Two events A and B are independent if

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

We often write $A \perp B$ to denote that A and B are independent.

Example 3.2. Sometimes it's intuitively obvious that events are independent. Sometimes it's not. Suppose a single card is drawn from a deck of 52 cards. Let

$$A = \{\text{the card is an ace}\}, \qquad D = \{\text{the card is a diamond}\}.$$

If we know that the card is a diamond, does this give any information that it's an ace? or vice versa?

We can compute $P(A \cap D) = 1/52$. (Only one diamond ace.) Also, we know $P(A) = 4/52$ and $P(D) = 1/4$, so indeed $P(A)P(D) = 1/52 = P(A \cap D)$.

What about more than two events? We say that A, B, C are pairwise independent if any pair are independent. However, even if A, B, C are pairwise independent, it's still possible for $A \cap B$ to give information about C.

Example 3.3. A fair coin is tossed twice. Define

$$\begin{aligned} A &= \{\text{ heads on first toss}\}, \\ B &= \{\text{ heads on second toss}\}, \\ C &= \{\text{exactly one head is tossed}\}. \end{aligned}$$

It is simple to check that A, B, C are pairwise independent. However, we can check that

$$P(C|A \cap B) \neq P(C).$$

First note that

$$P(C|A \cap B) = \frac{P(C \cap A \cap B)}{P(A \cap B)} = 0$$

since A, B and C cannot all occur at the same time. But on other hand, $P(C) = 1/2$, so indeed $P(C|A \cap B) \neq P(C)$. This example motivates the following definition

Definition 3.4 (Independence for several events)**.** We say that a collection of events are mutually independent (or just independent) if for any sub-collection $A_1, \ldots, A_n$, we have

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n).$$

Example 3.5. Suppose that if you make contact with a diseased person, the chance that you will become infected is 1/10. If you make contact 10 times, what is the probability that you don't get infected?

Let

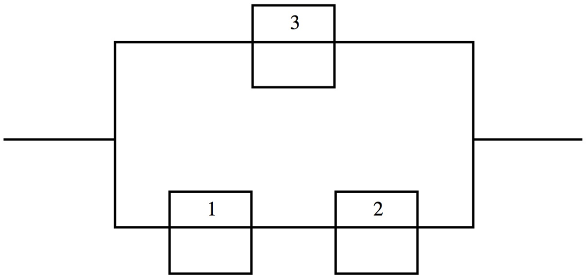
$$C_i = \{ \text{ infection does not occur on } i\text{th trial}\}.$$

Then, the probability no infection occurs in 10 trials is

$$P(C_1 \cap \cdots \cap C_{10}) = (1 - \frac{1}{10})^{10} = 0.35$$

Hence, the probability that you will get infected after 10 contacts is 0.65.

Example 3.6. Suppose there are three independent switches in a circuit.



Let $A_i = \{i\text{th switch works}\}$ and suppose $P(A_i) = p$ for all i. Denote by F the event that current flows through the circuit. What is $\mathbb{P}(F)$?

Note that $F = A_3 \cup (A_1 \cap A_2)$ and for general A and B we have

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

Applying this to F gives

$$P(F) = P(A_1) + P(A_2 \cap A_3) - P(A_1 \cap A_2 \cap A_3) = p + p^2 - p^3.$$

4 Discrete random variables

Random variables are just random numbers that come from an experiment.

Consider the experiment of flipping a coin three times:

$$\Omega = \{hhh, hht, htt, hth, ttt, tth, thh, tht\}.$$

There are lots of numbers that describe what is happening in this experiment.

- total number of heads
- total number of tails
- the number of heads minus the number of tails.

All of these numbers are random variables. How do make sense of this mathematically?

Definition. A random variable is a function $X : \Omega \to \mathbb{R}$. In other words, X takes in a point $\omega \in \Omega$ and returns a real number.

So for example, if X is the total number of heads, then

$$\begin{aligned} X(hhh) &= 3 \\ X(hht) &= 2 \\ X(htt) &= 1 \\ X(hth) &= 2 \\ X(ttt) &= 0 \\ X(tth) &= 1 \\ X(thh) &= 2 \\ X(tht) &= 1. \end{aligned} \tag{4.1}$$

Definition. If the set of possible values that X can take is finite (or countable), then X is discrete.

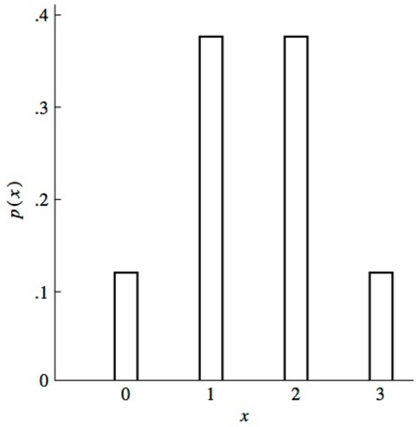
The last example was a discrete r.v.

Definition (probability mass function). If a discrete random variable X can take values $x_1, x_2, \ldots,$ then the probability mass function (pmf) (or frequency function) of X is the function defined by

$$p(x_i) := P(X = x_i).$$

Example.

$$\begin{aligned} p(0) &= P(X = 0) = \frac{1}{8} \\ p(1) &= P(X = 1) = \frac{3}{8} \\ p(2) &= P(X = 2) = \frac{3}{8} \\ p(3) &= P(X = 3) = \frac{1}{8} \end{aligned} \tag{4.2}$$



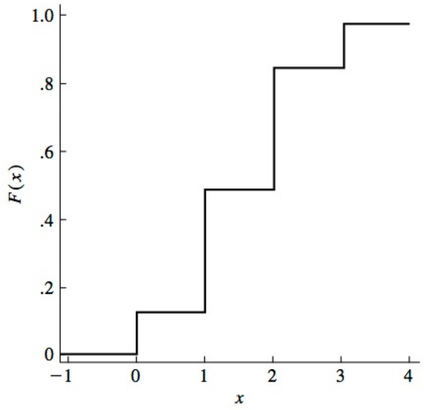
Remark. If we list the total set of outcomes as $\{x_1, \ldots, x_N\}$, then the pmf has to satsify

$$\sum_{i=1}^{N} p(x_i) = 1$$

This is also true if the number of outcomes is infinite.

Definition (cumulative distribution function). If X is a random variable, its cumulative distribution function is defined by

$$F(x) := P(X \leq x) \qquad \text{for any } -\infty < x < \infty$$



Comments (properties). $F(x)$ is increasing: If $x \leq x'$, then $F(x) \leq F(x')$.

As $x \to -\infty$, $F(x) \to 0$. As $x \to \infty$, $F(x) \to 1$.

Definition (independent random variables). Suppose X and Y are rvs with possible values $x_1, x_2, \ldots$, and $y_1, y_2, \ldots$. Then, X and Y are independent iff

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j) \qquad \text{for all} \quad i, j.$$

Let Z be a third random variable with possible values $z_1, z_2, \ldots$. Then, X, Y, and Z are independent if

$$P(X = x_i, Y = y_j, Z = z_k) = P(X = x_i)P(Y = y_j)P(Z = z_k) \qquad \text{for all} \quad i, j, k.$$

4.1 Bernoulli rv

A bernoulli rv is a "coin flip" variable that can be either 0 or 1, with probability q or $1 - q$.

$$p(1) = p, \quad \text{and } p(0) = 1 - p.$$

$$p(x) = p^x(1-p)^{1-x} \qquad \text{if x=0,1 and 0 otherwise}$$

Although Bernoulli rv are simple, they are important, because they occur as *indicators* of events.

Defnition The indicator function of an event $I_A(\omega) = 1$ if $\omega \in A$ and 0 otherwise.

4.2 Binomial Distribution

Suppose we have a sum of independent Bernoulli rv. i.e. How many heads do we toss in n coin flips? This number is a binomial rv.

If X is a binomial rv arising from n coin flips, then what are the possible values of X? ans: $0, 1, \ldots, n$

What is the pmf of a binomial rv? For each $k = 0, \ldots, n$, we need to calculate $p(k) = P(X = k)$. How many ways are there that X can equal k? ans $\binom{n}{k}$. Consider $k = 2$ in the previous example. Then,

$$\{X = 2\} = \{hht, hth, thh\},$$

and there are $\binom{3}{2}$ ways this can happen.

In general we can write

$$\{X = k\} = A_1 \cup A_2 \cup \cdots \cup A_m$$

e.g. $A_1 = \{hht\}$, $A_2 = \{hth\}$, etc, where $m = \binom{n}{k}$.

The A_i are disjoint and equiprobable, Let $P(A_i) = c$ for some c.

Then,

$$P(X = k) = \sum_{i=1}^{m} P(A_i) = m \cdot c = \binom{n}{k} \cdot c.$$

We just need to calculate c. What is the probability of a string of coin tosses involving k heads and $n - k$ heads? Ans $p^k(1-p)^{n-k}$. Hence,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Example (genetic disease). Suppose a couple has a 0.25 chance of passing on a disease to a child, and that the couple plans to have four children. Suppose that the births are independent.

Let X be the total number of children who inherit the disease. Let's derive the pmf for X.

$$p(k) = \binom{4}{k} 0.25^k 0.75^{4-k}.$$

$$p(0) = 0.315, \quad p(1) = 0.422 \quad p(2) = 0.211 \quad p(3) = 0.047 \quad p(4) = 0.004.$$

4.3 Geometric

Suppose that we flip a coin (with heads prob p) until we get heads.

Let X be the number of flips it takes to get heads.

$$p(k) = (1-p)^{k-1} p.$$

4.4 Negative binomial

Suppose we flip a coin (with heads prob p) until we get r heads.

Let X be the number of flips it takes to get r heads.

To compute $p(k)$, all outcomes with $X = k$ must end with a heads, and other otherwise a sequence of $k - 1$ arbitrary coin flips containing r heads.

Hence,

$$P(X = k) = \binom{k-1}{r-1} p^{r-1}(1-p)^{k-r} \times p$$

4.5 Hypergeometric

Suppose we have an urn containing n balls, with r black balls and $n - r$ white balls.

We draw m balls from the urn without replacement.

Let X denote the number of black balls that are drawn.

The total number of subsets that can be drawn is $\binom{n}{m}$. To have exactly k black balls in the subset of size m two things have to happen: we have to draw a subset of size k from a total of r black balls and we have to draw $m - k$ white balls from $n - r$ total white balls. Hence

$$P(X = k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}.$$

4.6 Poisson

Suppose you flip a coin many times (n) where the probability of heads is small p.

Let X be the total number of heads.

Examples Fires

Computer failures

Number of men in an army regiment who die of horse kicking.

Counting diseased cells in a specimine

Also assume that $pn = \lambda$ (you can think of typical number of heads)

When p is small and n is large, it is possible to simplify the binomial pmf to get

$$P(X = k) \approx \frac{\lambda^k}{k!} e^{-\lambda}.$$

Is the right side actually a pmf?

Define $p(k) := \frac{\lambda^k}{k!} e^{-\lambda}$.

Can we check that

$$\sum_{k=0}^{\infty} p(k) = 1?$$

$$\sum_{k=0}^{\infty} p(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

5 Continuous random variables

A continuous random variable is one that can take on a "continuum" of values, say any value in an interval $[a, b]$, or any real number.

Example. How long did it take me to drive to work today (time is continuous).

What is the height of a person drawn from a population? (length is continuous).

Continuous rv are described completely by their density function:

Definition (density function) If X is a continuous random variable, the for any interval $[a, b]$, we have

$$P(a < X < b) = \int_a^b f(x) dx.$$

The density function plays the role for continuous rv that pmf does for discrete rv.

Example. Uniform random variable in interval $[0, 1]$.

Remark (relation between cdf and density)

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u) du.$$

How to get density from cdf?

$$f(x) = \frac{d}{dx}F(x) = F'(x).$$

Also, we have

$$P(a \leq X \leq b) = \int_a^b f(x)dx = \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx = F(b) - F(a).$$

Let X be a uniform variable between a and b. Let's derive its cdf.

$$F(x) = \int_{-\infty}^x f(u)du = \int_{-\infty}^a 0du + \int_a^x \frac{1}{b-a}du = \frac{x-a}{b-a}.$$

Note. If $x \leq a$, then $F(x) = 0$, and if $x \geq b$, then $F(x) = 1$.

Remark. In some sense, a cdf, a density, a random variable, and a probability measure are all the same thing. They are all just different ways of representing each other. If I give you a density function, that also specifies an rv. If I give you an rv, that also specifies a cdf, etc.

Remark. The probability that a continuous rv is exactly equal to a single real number is 0. If X is the amount of time it takes to drive to work, there is a small chance it could be between 1000 seconds and 1010 seconds. However, if narrow the range to between 1000 seconds and 1000.0000001 seconds, the chance becomes extremely small. Hence, if we make an arbitrarily small neighborhood, the probability goes to 0.

More formally

$$P(X = c) = \int_c^c f(x)dx = 0.$$

To see why, consider the probability that $X \in [x_0 - \delta/2, x_0 + \delta/2]$ for some small $\delta > 0$. Then, if f is continuous at x, we get

$$P(x_0 - \delta/2 < X < x_0 + \delta/2) = \int_{x_0-\delta/2}^{x_0+\delta/2} f(x)dx \approx f(x_0) \cdot \delta.$$

As we take $\delta \to 0$, the right side goes to 0.

Quantile. Suppose 100 students take an exam, and we rank the scores. If Bob does at least as well as 75 students but not better than 25 students, then Bob's score is the 0.75 quantile.

More generally, for a random variable, the pth quantile of an rv X with a continuous density is the number x_p such that

$$F(x_p) = P(X \leq x_p) = p.$$

PICTURE
What is the 0.5 quantile? Special name: median.

What are the 0.25 and 0.75 quantiles called: lower and upper quartiles.

Example. Uniform random variable in interval $[a, b]$.

In addition to uniform rv, we will also introduce the exponential rv, as well as Gaussian rv (next time)

5.1 Exponential random variables

What is the story for an exponential rv? This rv is the waiting time until something happens. e.g. How long until a light bulb goes out, or how long a human will live.

If X represents a waiting time, then X should be non-negative. Hence the density for X only is positive on the positive x axis.

PICTURE

Draw curves for two lambdas.

Density:

$$f(x) = \lambda e^{-\lambda x}, x \geq 0,$$

and

$$f(x) = 0 \quad \text{otherwise}$$

Let's calculate the cdf. Suppose $x \geq 0$. Then,

$$F(x) = \int_{-\infty}^{x} f(u)du = \int_{0}^{x} \lambda e^{-\lambda u} du = -e^{-\lambda u}\Big|_{0}^{x} = -e^{-\lambda x} + 1.$$

Let's calculate the median $x_{1/2}$. ie let's solve the equation $F(x_{1/2}) = 0.5$ for $x_{1/2}$.

Hence,

$$-e^{-\lambda x_{1/2}} + 1 = 0.5 \Longrightarrow e^{-\lambda x_{1/2}} = 1/2 \Longrightarrow -\lambda x_{1/2} = \log(1/2).$$

Which implies the median of an exponential variable is.

$$x_{1/2} = \frac{\log(2)}{\lambda}.$$

Remark. As lambda gets big what happens? Median gets small.

Memoryless property. Fix a number $s > 0$, and suppose T is the time it takes for a light bulb to go out (with exponentially distributed lifetime). What is the probability that it takes longer than s for the lightbulb to go out?

Ans:

$$P(T > s) = 1 - P(T \leq s) = e^{-\lambda s}.$$

If the lightbulb lasts longer than a time s, what is the probability that it will last an additional time t?

Ans:

$$\begin{aligned}
P(T > t + s | T > s) &= \frac{P(T > t + s \text{ and } T > t)}{P(T > s)} \\
&= \frac{P(T > t + s)}{P(T > s)} \\
&= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} \\
&= e^{-\lambda t}.
\end{aligned} \tag{5.1}$$

What this says is that for an an exponential rv, the additional waiting time does not depend on how long you have already waited. Good for light bulbs but for some other things.

Remark. Not all waiting times are exponentially distributed.

Human lifetimes are not well modeled by an exponential rv. Are a 20 year old and an 80 year old equally likely to live an additional 10 years? No.

Suppose you wait a week for someone to respond to your email. Is the additional waiting time likely to be the same as if you had just sent the email? No.

5.2 The Normal Distribution (Gaussian rv)

The story: The normal distribution arises from a sum of many independent random variables. We area all familiar with the idea of white noise.

Examples. A room full of people talking. Standing on the sidewalk next to a busy street. In all these cases, the white noise comes from many independent sources of noise being added together. The cumulative effect can be thought of as a Gaussian rv.

Another example is in terms of "complex" phenomena where many independent factors lead to the outcome. For example, heights in a population, or the movement of particles.

Density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/2\sigma^2).$$

If X is an rv with density f, we write $X \sim N(\mu, \sigma^2)$.

when $\mu = 0$ and $\sigma = 1$, we usually write

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

Also, we write

$$\Phi(x) = \int_{-\infty}^{x} \phi(u) du.$$

[Pictures varying μ and σ.]

Examples measuring height and weight.

6 Functions of a Random Variable

Suppose you have a quantity that is a linear transformation. For instance, suppose we have daily temperatures measured in Fahrenheit and we think of temperature as a Gaussian rv.

What is the distribution of the temperature $\tilde{Y}$ measured in Celsius?

Unit conversion says there are constants a and b such that

$$Y = aX + b.$$

If X has cdf F_X, can we calculate the cdf F_Y of Y?

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(aX + b \leq y) \\ &= P(X \leq \frac{y-b}{a}) \\ &= F_X(\frac{y-b}{a}) \end{aligned} \tag{6.1}$$

This also shows that we can get the density function f_Y of Y from the density f_X of X.

Take derivatives

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \frac{d}{dy} F_X(\frac{y-b}{a}) \\ &= \frac{1}{a} f_X(\frac{y-b}{a}) \end{aligned} \tag{6.2}$$

Let's use this to check that if X is Gaussian, then a linear transformation of X is Gausian.

Proposition. If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$ for constants a and b, then

$$Y \sim N(a\mu + b, a^2\sigma^2).$$

Proof. By assumption

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

Now, let's use our formula

$$\begin{aligned} f_Y(y) &= \frac{1}{a} f_X(\frac{y-b}{a}) \\ &= \frac{1}{a\sigma\sqrt{2\pi}} \exp\Big(-\frac{((y-b)-a\mu)^2}{2a^2\sigma^2} \Big) \end{aligned} \tag{6.3}$$

This shows that f_Y is a Gaussian density with mean $b + a\mu$ and variance $a^2\sigma^2$.✓

Example (non-linear transformation) Say X is the square of another random variable Z,

$$X = Z^2.$$

Let F_Z denote the cdf of Z. What is the cdf of X?

$$\begin{aligned} F_X(x) &= P(X \le x) && (6.4)\\ &= P(-\sqrt{x} \le Z \le \sqrt{x}) && (6.5)\\ &= F_Z(\sqrt{x}) - F_Z(-\sqrt{x}) && (6.6) \end{aligned}$$

Hence, the density f_X is given by

$$\begin{aligned} f_X(x) &= F'_X(x) && (6.7)\\ &= \frac{1}{2}x^{-1/2} f_Z(\sqrt{x}) + \frac{1}{2}x^{-1/2} f_Z(-\sqrt{x}) && (6.8)\\ & && (6.9) \end{aligned}$$

6.1 General non-linear transformation

Let X be a continuous rv with density f_X, and let

$$Y = g(X)$$

where g is a differentiable and strictly increasing function on an interval I. Suppose that $f(x) = 0$ if x is not in I. Then,

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

Here, the inverse function satisfies $g^{-1}(y) = x$ if and only if $y = g(x)$.

Example 6.1. Verify this formula for the linear transformation. Next, Let U be a uniform random variable. Find the density function of $V = U^{-\alpha}$ for $\alpha > 0$. Compare the rates of decrease of the tails of the densities as a function of α. Does the comparison make sense intuitively?

For $Y = g(X) = aX + b$ with $a > 0$ we have $g^{-1}(y) = (y - b)/a$ and

$$\frac{d}{dy} g^{-1}(y) = \frac{1}{a}.$$

Therefore

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a}.$$

Next, let $V = g(U) = U^{-\alpha}$ where U is uniform random variable on $[0, 1]$. Then $f_U(u) = I_{[0,1]}(u)$ and $g^{-1}(v) = v^{-1/\alpha}$. Also,

$$\frac{d}{dv}g^{-1}(v) = \frac{-v^{-1/\alpha+1}}{\alpha}.$$

Therefore

$$f_V(v) = I_{[0,1]}(v^{-1/\alpha}) \cdot \frac{v^{-1/\alpha+1}}{\alpha}.$$

Note that

$$I_{[0,1]}(v^{-1/\alpha}) = \begin{cases} 1, & \text{if } v \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Hence

$$f_V(v) = I_{[1,\infty)}(v) \cdot \frac{v^{-1/\alpha+1}}{\alpha}.$$

6.2 How to generate random variables?

Suppose that we can generate uniform[0,1] variables on a computer. We can use this to generate many other kinds of rv.

Let U be uniform on [0,1] and F be the cdf of a random variable we want to generate. Define X to be the random variable

$$X = F^{-1}(U).$$

Then, the cdf of X is F.

Proof. Recall that for a uniform[0,1] variable U we have $P(U \leq y) = y$ for any y. Now we use this to calculate

$$F_X(x) = \mathbb{P}(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

Hence $F_X(x) = F(x)$.

Remark. The opposite direction is also true. If we start with a random variable X with cdf F and define $Y = F(X)$ then Y is uniform on [0,1].

7 Joint Distributions

Joint distributions describe the relationships between two or more random variables. They allow us to quantify *dependence* between random variables. Examples:
- Height and weight
- Demographic variables: age, income, zipcode,

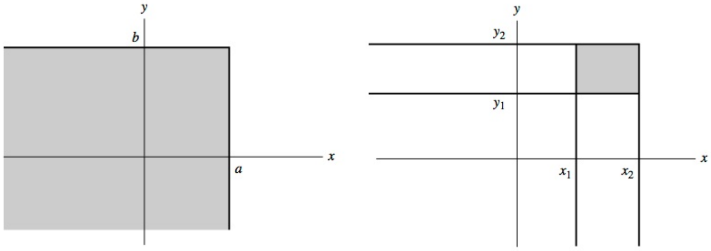
Joint cdf The joint behavior of two rv's X and Y is determined by their joint cumulative distribution function

$$F(x,y) := P(X \leq x, Y \leq y).$$

We can get the probabilities of lots of other events from knowing $F(x,y)$, e.g.

$$P(x_1 \leq X \leq x_2,\ \ y_1 \leq Y \leq y_2) = F(x_2,y_2) - F(x_2,y_1) - F(x_1,y_2) + F(x_1,y_1). \tag{7.1}$$

This works regardless of whether the rvs are discrete or continuous.



Suppose there are n variables $X_1, \ldots, X_n$ rather than just 2. Then, the joint cdf is defined by

$$F(x_1, \ldots, x_n) := P(X_1 \leq x_1, \ldots, X_n \leq x_n). \tag{7.2}$$

7.1 Joint distributions for discrete rvs

Suppose X and Y are discrete rvs on the sample space; i.e. when an experiment occurs, we observe a value for both X and Y.

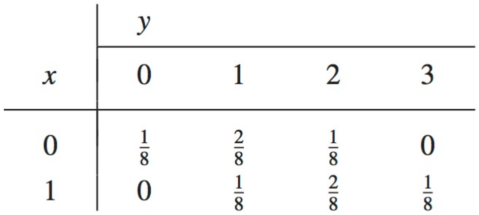
Let the possible values of X and Y be denoted by $x_1, x_2, \ldots,$ and the possible values of Y as $y_1, y_2, \ldots$. Then, for any pair x_i, y_j, the joint frequency function is

$$p(x_i, y_j) = P(X = x_i, Y = y_j). \tag{7.3}$$

Example. Suppose a coin is tossed three times. Let X be 1 if heads occurs on the first toss and 0 if tails. Let Y be the total number of heads in the three tosses.
The total sample space is

$$\Omega = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}. \tag{7.4}$$

For example $p(0,2) = P(X = 0, Y = 2) = 1/8$.



	y			
x	0	1	2	3
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

How to get the frequency functions of X and Y from the joint pmf.

$$\begin{aligned} p_Y(0) &= P(Y=0) \\ &= P(Y=0, X=0) + P(Y=0, X=1) \\ &= 1/8 + 0 \\ &= 1/8 \end{aligned} \tag{7.5}$$

$$\begin{aligned} p_Y(1) &= P(Y=1) \\ &= P(Y=1, X=0) + P(Y=1, X=1) \\ &= 2/8 + 1/8 \\ &= 3/8 \end{aligned} \tag{7.6}$$

To find the frequency function of Y we simply sum down the appropriate column of the table. For this reason, p_Y is called the marginal frequency function of Y.

Similarly summing across the rows gives the marginal frequency function of X,

$$p_X(x) = \sum_i p(x, y_i). \tag{7.7}$$

Multinomial distribution. Suppose you have n balls and you can throw the balls into r bins. The probability that a ball lands in the ith bin is p_i. Also suppose the throws are independent.

Let N_i be the number of balls in the ith bin. The joint distribution $(N_1, \ldots, N_r)$ is called the multinomial distribution.

Let's calculate $p(n_1, \ldots, n_r) = P(N_1 = n_1, \ldots, N_r = n_r)$.

Suppose $n = 4$ and $r = 3$. What is the probability that first two balls are in bin one, third ball in bin two and last ball in bin three? Answer: $p_1^2 p_2 p_3$.

How many ways can we get two balls in bin one (not necessarily the first two balls), one ball in bin two, and one ball in bin three? i.e. $n_1 = 2, n_2 = 1, n_3 = 1$? Answer: $\binom{4}{211}$.

So, in general

$$p(n_1, \ldots, n_r) = \binom{n}{n_1 n_2 \cdots n_r} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}.$$

7.2 Joint distributions for continuous rv

What is the joint density function of X and Y? The density function of (X, Y) allow us to calculate the probability that (X, Y) falls into a region A in the plane:

$$P((X,Y) \in A) = \iint\limits_{(x,y)\in A} f(x,y)\ dxdy$$

Let's recall the one-variable situation. The double integral is volume under the curve (don't worry if you haven't seen double integrals before).

Let's consider a rectangular region A such that $(X, Y) \in A$ iff $X \leq x$ and $Y \leq y$. Then,

$$P((X,Y) \in A) = \int_{-\infty}^{x} \Big(\int_{-\infty}^{y} f(u,v)dv \Big) du. \tag{7.8}$$

But we also know that

$$P((X,Y) \in A) = P(X \leq x, Y \leq y) = F(x,y). \tag{7.9}$$

so

$$F(x,y) = \int_{-\infty}^{x} \Big(\int_{-\infty}^{y} f(u,v)dv \Big) du.$$

This shows that we can represent the joint density of (X, Y) as

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

Recall the one-variable case. If F_X is the cdf of X, we know that the density function of X is

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

Another way to interpret the joint density function:

Let's look at the probability that (X, Y) lies in a small rectangle $[x, x + \Delta_x] \times [y, y + \Delta_y]$:

$$\begin{aligned} P(x \leq X \leq x+\Delta_x, y \leq Y \leq y+\Delta_y) &= \int_{x}^{x+\Delta_x} \Big(\int_{y}^{y+\Delta_y} f(u,v)dv \Big) du \\ &\approx f(x,y)\Delta_x\Delta_y. \end{aligned} \tag{7.10}$$

Example. Consider the density function

$$f(x,y) = \frac{12}{7}(x^2 + xy), \qquad \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1.$$

Let's look at the region A of points (x, y) such that $x > y$.

Then, if $(X, Y) \sim f$, we have

$$P((X,Y) \in A) = P(X > Y) = \frac{12}{7} \int_0^1 \Big(\int_0^x (u^2 + uv) dv \Big) du = \frac{9}{14}.$$

Now let's show how we can get the marginal cdf of X from the joint density.

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= P(X \le x, Y \le \infty) \\ &= \lim_{y \to \infty} F(x,y) \\ &= \int_{-\infty}^{x} \Big(\int_{-\infty}^{\infty} f(u,v) du \Big) dv \end{aligned} \tag{7.11}$$

Let's get the marginal density function of x from the joint density:

$$\begin{aligned} f_X(x) &= F_X'(x) \\ &= \frac{d}{dx} \int_{-\infty}^{x} \Big(\int_{-\infty}^{\infty} f(u,v) dv \Big) du \\ &= \int_{-\infty}^{\infty} f(x,v) dv \end{aligned} \tag{7.12}$$

Example Go back to previous example

$$f_X(x) = \frac{12}{7} \int_0^1 (x^2 + xy) dy = \frac{12}{7} \Big(x^2 y + x \frac{y^2}{2} \Big|_{y=0}^{1} \Big) = \frac{12}{7} \Big(x^2 + \frac{x}{2} \Big)$$

You can do a similar calculation to get $f_Y(y) = \frac{12}{7}(\frac{1}{3} + y/2)$.

7.3 Independent random variables

Definition. An arbitrary collection of (continuous or discrete) random variables $X_1, \ldots, X_n$ are independent if their joint cdf factors into the product of their marginal cdfs

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n).$$

for all $x_1, \ldots, x_n$.

Remark. For discrete random variables, the above definition is equivalent to saying that the joint pmf factors as $p(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n)$ for all $x_1, \ldots, x_n$. For continuous random variables,it is equivalent to saying that the joint density function factors as $f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$ for all $x_1, \ldots, x_n$.

Example. Suppose we are transmitting packets in a network. If two packets arrive at approximately the same time, then they "collide" (which is bad). Specifically, if packets arrive within τ seconds of eachother, then they collide. Suppose the arrival times of two packets are independent rv, and are each distributed uniformly on the interval $[0, t_0]$.

Recall than the uniform density on the interval $[0, t_0]$ looks like $f(t) = 1/t_0$.

By independence, the joint density of T_1, T_2 is

$$f(t_1, t_2) = \begin{cases} 1/t_0^2 \text{ if } (t_1, t_2) \in [0, t_0] \times [0, t_0] \\ 0 \text{ otherwise} \end{cases}$$

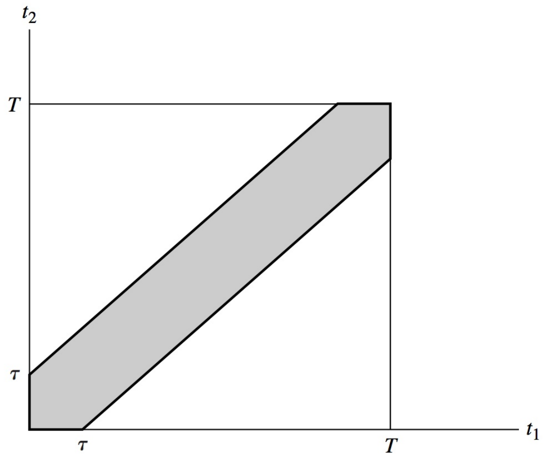
Let's look at the event of collision. We must have

$$t_2 \leq t_1 + \tau,$$

and

$$t_2 \geq t_1 - \tau.$$

It's the shaded region.



Let S denote the shaded region.

We want to calculate the probability $P((T_1,T_2) \in \mathrm{S})$.

which is

$$P(collision) = P(|T_1 - T_2| \leq \tau) = \int\int_{(t_1,t_2)\in S} f(t_1,t_2)dt_1dt_2$$

But this is simple because the density is constant $1/t_0^2$so

$$P(collision) = P(|T_1 - T_2| \leq \tau) = \frac{1}{t_0^2}\int\int_{(t_1,t_2)\in S} 1dt_1dt_2 = \frac{1}{t_0^2}\mathrm{area}(S).$$

But area$(S) = t_0^2$ - two triangles $= t_0^2 - (t_0 - \tau)^2$. So altogether

$$P(collision) = \frac{t_0^2 - (t_0-\tau)^2}{t_0^2}.$$

Remark. If X and Y are (continuous or discrete) rv, then for any two sets A and B we have

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

Functions of independent rv are independent. Suppose X and Y are independent rv (continuous or discrete). Define

$$Z = g(X) \quad \text{and} \quad W = h(Y).$$

for some functions g and h. Then, Z and W are independent.

8 Conditional Distributions

8.1 Conditional distributions for discrete rv

Suppose X and Y are rv. that take values (respectively) $x_1, x_2, \ldots,$ and $y_1, y_2, \ldots.$

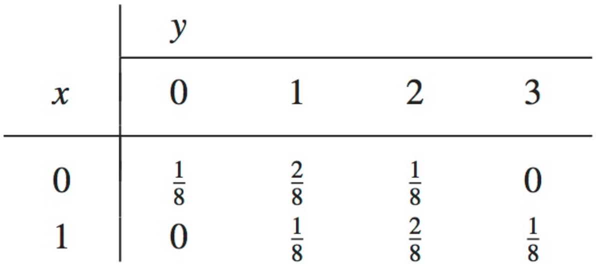
Consider a fixed y_j and suppose $p_Y(y_j) > 0$.

The conditional probability of $X = x_i$ given $Y = y_j$ is

$$\begin{aligned} p_{X|Y}(x_i|y_j) := P(X = x_i|Y = y_j) &= \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} \qquad (8.1) \\ &= \frac{p_{XY}(x_i, y_j)}{p_Y(y_j)} \qquad (8.2) \end{aligned}$$

What happens when X and Y are independent? Then,

$$p_{X|Y}(x|y) = p_X(x).$$



	y			
x	0	1	2	3
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

Example.

$$P_{X|Y}(0|1) = \frac{2/8}{3/8} = 2/3.$$

$$P_{X|Y}(1|1) = \frac{1/8}{3/8} = 1/3.$$

Relation between joint pmf and conditional pmf

$$p_{XY}(x, y) = p_{X|Y}(x|y)p_Y(y).$$

Now, marginalization tells us that

$$p_X(x) = \sum_y p_{XY}(x, y) = \sum_y p_{X|Y}(x|y)p_Y(y).$$

Divide and conquer.

Example. Suppose a particle counter is imperfect. It only detects an incoming particle with probability p. Suppose also that a unit of time, the number of incoming particles is a random number N that has a poisson distribution with parameter λ.

Let X denote the number of particles that are counted. We want pmf for X,

$$
\begin{aligned}
P(X=k) &= \sum_{n=1}^{\infty} P(X=k|N=n)P(N=n) \\
&= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \frac{\lambda^n e^{-\lambda}}{n!} \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda} \sum_{n=k}^{\infty} \lambda^{n-k} \frac{(1-p)^{n-k}}{(n-k)!} \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j (1-p)^j}{j!} \qquad j := n-k \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda} e^{\lambda(1-p)} \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda p} \qquad \text{Poisson pmf with param } \lambda p
\end{aligned}
\tag{8.3}
$$

8.2 Conditional distributions for continuous rv

Suppose we have two continuous rv X and Y with densities f_X and f_Y.

The conditional density of Y given that X lies in a strip of width dx satisfies

$$P(y \leq Y \leq y+dy \mid x \leq X \leq x+dx) \approx f_{Y|X}(y|x)dy$$

Let's derive a formula for $f_{Y|X)}(y|x)$,

$$P(y \leq Y \leq y+dy \mid x \leq X \leq x+dx) \approx \frac{f_{XY}(x,y)dxdy}{f_X(x)dx} = \frac{f_{XY}(x,y)}{f_X(x)}dy.$$

Comparing, we conclude that

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}.$$

Get joint density from $f_{Y|X}$ and f_X.

$$f_{XY}(x,y) = f_{Y|X}(y|x)f_X(x).$$

Get marginal f_Y from conditional $f_{Y|X}$ and marginal f_X.

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x) dx.$$

This is the continuous version of the divide and conquer formula we showed earlier for the discrete case.

9 Expected value of random variable

Expectation for discrete random variables. If X is a random variable taking values $x_1, x_2, \ldots$ with pmf $p_X(x)$, then

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i p_X(x_i)$$

Interpretation of expected value. If we flip many independent coints $X_1, \ldots, X_n$ each with probability of heads p, and we look at $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, then for large values of n we have

$$\bar{X} \approx p = \mathbb{E}[X_1].$$

More generally, if X_i are independent and they all have the same distribution, then

$$\bar{X} \approx \mathbb{E}[X_1],$$

provided that $\mathbb{E}[X_1]$ exists.

Example. (Geometric random variable) Suppose that items produced in a plant are independently defective with probability p. Items are inspected one by one until a defective item is found. On average, how many inspected until a defective item is found?

Let X denote the number of inspected items until a defective is found. X is a geometric rv and

$$P(X = k) = (1-p)^{k-1} p = q^{k-1} p$$

where $q := 1 - p$.

Therefore,

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k P(X = k) = \sum_{k=1}^{\infty} k \Big(q^{k-1} p \Big) = p \sum_{k=1}^{\infty} k q^{k-1}$$

Pull a rabbit out of the hat. Consider the function $f_k(q) = q^k$. Note that

$$\frac{d}{dq} f(q) = kq^{k-1}.$$

Hence

$$\mathbb{E}[X] = p \sum_{k=1}^{\infty} \frac{d}{dq} f_k(q) = p \frac{d}{dq} \Big(\sum_{k=1}^{\infty} f_k(q) \Big)$$

But note that we have a geometric series

$$\sum_{k=1}^{\infty} f_k(q) = \sum_{k=1}^{\infty} q^k = \frac{q}{1-q}.$$

So,

$$\mathbb{E}[X] = p \frac{d}{dq} \Big(\frac{q}{1-q} \Big) = p \frac{1}{(1-q)^2} = p \frac{1}{p^2} = \frac{1}{p}.$$

Definition (Expectation for continuous random variable) Suppose X is a continuous random variable with density function $f(x)$. Then, we define

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx,$$

provided that $\int |x| f(x) dx < \infty$. If the integral diverges, then the expectation is undefined.

Example (Exponential random variable). Suppose X is exponential with parameter λ. Think of X as a waiting time. The density function is

$$f(x) = \lambda e^{-\lambda x}, \text{for } x \geq 0 \text{ and } 0 \text{ otherise}$$

Hence,

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} xf(x)dx \\
&= \int_{0}^{\infty} x \cdot \lambda e^{-\lambda x} dx \qquad \text{consider } g(x) = x, \quad h(x) = e^{-\lambda x} \\
&= \int_{0}^{\infty} g(x)[-h'(x)]dx \qquad \text{use} \quad (gh)' = g'h + gh' \\
&= -\int_{0}^{\infty} (g(x)h(x))'dx + \int_{0}^{\infty} g'(x)h(x)dx \\
&= -g(x)h(x)|_0^\infty - \int_{0}^{\infty} h(x)dx \\
&= 0 + \int_{0}^{\infty} e^{-\lambda x}dx \\
&= -\frac{1}{\lambda} e^{-\lambda x}|_0^\infty \\
&= \frac{-e^{-\infty}}{\lambda} - (\frac{-e^0}{\lambda}) \\
&= \frac{1}{\lambda}
\end{aligned} \tag{9.1}$$

9.1 Markov's inequality

The existence of $\mathbb{E}[X]$ implies that the probability of large values has to be fairly small.

Markov's inequality Suppose $X \geq 0$ is a non-negative rv and $\mathbb{E}[X]$ exists. Then, for any $t \geq 0$,

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

Proof. Let's do the discrete case. Suppose $X \in \{x_1, x_2, \dots\}$. Then,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_i x_i p(x_i) \\
&= \sum_{i:x_i<t} x_i p(x_i) + \sum_{i:x_i\geq t} x_i p(x_i) \\
&\geq 0 + \sum_{i:x_i\geq t} x_i p(x_i) \\
&\geq \sum_{i:x_i\geq t} t p(x_i) \\
&= t \sum_{i:x_i\geq t} p(x_i) \\
&= tP(X \geq t)
\end{aligned}$$

Which is the same as what we wanted.

9.2 Expectation of functions of random variable

Often we need to calculate $\mathbb{E}[g(X)]$ where g is some function on an random variable.

If X is discrete one way of calculating this would be to consider an random variable $Y = g(X)$, find $p_Y(y)$ using $p_X(x)$, and then calculate $\mathbb{E}[Y] = \sum_y y p_Y(y)$. However, calculating the pmf of Y can be tricky. We would like to find an easier way to do this.

Theorem (Discrete case) Suppose $Y = g(X)$. If X is discrete with pmf $p_X(x)$, then

$$\mathbb{E}[Y] = \sum_x g(x) p_X(x),$$

provided that $\sum_x |g(x)|p(x) < \infty$.

(Continuous case) If X is continuous with density function $f_X(x)$, then

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} g(x) f_X(x) dx,$$

provided that $\int_{-\infty}^{\infty} |g(x)| f_X(x) dx < \infty$.

9.3 Expectation for several variables (joint distributions)

Suppose $Y = g(X_1, \ldots, X_n)$ for jointly distributed random variables $X_1, \ldots, X_n$.

If the X_i are discrete with joint pmf $p(x_1, \ldots, x_n)$ then,

$$\mathbb{E}[Y] = \sum_{x_1,\ldots,x_n} g(x_1, \ldots, x_n) p(x_1, \ldots, x_n),$$

provided that $\sum_{x_1,\ldots,x_n} |g(x_1, \ldots, x_n)| p(x_1, \ldots, x_n) < \infty$.

If the X_i are continuous with joint density $f(x_1, \ldots, x_n)$, then

$$\mathbb{E}[Y] = \int \int \cdots \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$

Example. Compute $\mathbb{E}[e^X]$ in two ways when $X \sim$ Uniform[0,1]; using density and using expectation of function.

Recall that $f_X(x) = 1\{0 \leq x \leq 1\}$.

Let's do the new way of calculating $\mathbb{E}[e^X]$ first:

$$\mathbb{E}[e^X] = \int_0^1 e^x f_X(x) dx = \int_0^1 e^x dx = e^x \Big|_0^1 = e - 1.$$

Now let's do the old way. Define $Y = e^X$. Let's find $f_Y(y)$, with $1 \leq y \leq e$. Then,

$$F_Y(y) = P(Y \leq y) = P(e^X \leq y) = P(X \leq \log(y)) = F_X(\log(y)) \underbrace{=}_{Recalluniformcdf=identity} \log(y).$$

So, we take the derivative

$$f_Y(y) = F'_Y(y) = \frac{1}{y} \text{ when } 1 \leq y \leq e \text{ and } 0 \text{ otherwise.}$$

Finally

$$\mathbb{E}[e^X] = \mathbb{E}[Y] = \int_1^e y f_Y(y) dy = \int_1^e y(\frac{1}{y}) dy = \int_1^e 1 dy = e - 1 \checkmark$$

9.4 Expectation and independence

Suppose X and Y are independent random variables. Then,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Also let g and h be fixed functions. Then,

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)].$$

Do you see why the second follows from the first?

9.5 Linearity of expectation

Suppose $Y = a + bX$. Then,

$$\mathbb{E}[Y] = a + b\mathbb{E}[X].$$

More generally, if $Y = a + \sum_{i=1}^{n} b_i X_i$, then,

$$\mathbb{E}[Y] = a + \sum_{i=1}^{n} b_i \mathbb{E}[X_i].$$

Example (Binomial rv) .

Suppose $X_1, \dots, X_n$ are independent Bernoulli variables with head-probability p. Also let $Y = \sum_{i=1}^{n} X_i$. Then, Y is a binomial(n, p) rv.

We can evaluate the mean of a binomial by using linearity of expectation

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

Example (Coupon Collector) Suppose you want to collect n different kinds of baseball cards. Suppose you can buy the cards one at a time, but you don't know what card will be inside the wrapper. Each of the n cards are equally likely to be purchased.

How many cards do you have to buy to get a complete set? Call this number X.

Let $X_1 = 1$.

Let X_2 be the number of extra cards you have to buy to get one different from the first.

Let X_3 be the number of extra cards you have to buy to get one different from the first 2.

Let X_r be the number of extra cards you have to buy to get one different from the first r.

Then we can write

$$X = \sum_{r=1}^{n} X_r.$$

By linearity of expectation, we have

$$\mathbb{E}[X] = \sum_{r=1}^{n} \mathbb{E}[X_r],$$

So, it's enough to calculate $\mathbb{E}[X_r]$.

Note that X_2 is a geometric rv where the probability of heads is $\frac{n-1}{n}$.

X_3 is a geometric rv where the probability of heads is $\frac{n-2}{n}$.

X_r is a geometric rv where the probability of heads is $\frac{n-(r-1)}{n}$.

Recall from our calculation of expectation of geomtric rv that if p is the probability of heads this $1/p$. Hence

$$\mathbb{E}[X_r] = \frac{n}{n-(r-1)},$$

so

$$\mathbb{E}[X] = \sum_{r=1}^{n} \frac{n}{n-(r-1)} = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1}.$$

so

$$\mathbb{E}[X] = n \sum_{r=1}^{n} \frac{1}{r} \approx n \int_{1}^{n} \frac{1}{x} dx = n \log(n).$$

10 Variance

The expectation of a random variable gives us a 1-number summary of a typical value of an rv.

A more refined question is to ask how much does an rv fluctuate around its mean.

Suppose you're crossing a river on foot, and you want to know if it's safe. Someone tells you that the average depth of the river is only 3 feet. Are you confident it will be safe to cross.

What if there is a small part of the river that is 10 feet? This is not reflected in the average, since the deep part maybe very narrow. The variance helps to get more insight into these types of situations by telling us how much variation there is around the mean value.

Definition If X is an rv with expected value $\mathbb{E}[X]$, the variance of X is

$$\text{var}(X) = \mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big],$$

provided that that the expectation of $(X - \mathbb{E}[X])^2$ exists.

Note: The standard deviation of X is simply the square root of its variance. We abbreviate this as $\text{sd}(X)$.

It is customary to write the variance of an rv as $\sigma^2 = \text{var}(X)$. In this notation, the std deviation of X is simply σ.

If X is discrete with pmf $p_X(x)$, and expected value $\mu = \mathbb{E}[X]$, then the variance of X may be written as

$$\text{var}(X) = \sum_x (x - \mu)^2 p_X(x).$$

If X is continuous with density function $f_X(x)$ and mean $\mu = \mathbb{E}[X]$, then the variance may be written as

$$\text{var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx.$$

Example (Variance of coin flip). Suppose $X = 1$ with probability p and $X = 0$ with probability $1 - p$. What is the variance of X? Intuitively which choice of p should give the highest variance?

$$
\begin{aligned}
\mathrm{var}(X) &= \mathbb{E}((X-\mathbb{E}[X])^2) \\
&= \sum_x (x-\mu)^2 p_X(x) \\
&= (1-\mu)^2 \cdot p + (0-\mu)^2(1-p) \\
&= (1-p)^2 \cdot p + (0-p)^2(1-p) \\
&= (1-p)^2 \cdot p + p^2(1-p) \\
&= (1-2p+p^2)\cdot p + p^2 - p^3 \\
&= (p-2p^2+p^3) + (p^2-p^3) \\
&= p - p^2 \\
&= p(1-p)
\end{aligned}
\tag{10.1}
$$

Let's look at this as a function of p. When p is very close to 0, we know the coin is almost always going to be tails – the variance is small. When p is very close to 1, we know the coin is almost always going to be heads – the variance is small.

Also, we can see that the variance is the highest when $p = 1/2$; i.e. a fair coin has the highest variance.

Variance under linear transformation. (Theorem) Suppose we define a random variable $Y = bX + a$ where a and b are constants. Then,

$$\mathrm{var}(Y) = b^2\,\mathrm{var}(X).$$

Example Suppose X is temperature in Fahrenheit and Y is temperature in celsius. How much is the variance affected by changing units? The variance is not affected by translation; only scaling.

Proof.

$$
\begin{aligned}
\mathrm{var}(Y) &= \mathbb{E}[(Y-\mathbb{E}[Y])^2] \\
&= \mathbb{E}[(a+bX-(a+b\mathbb{E}[X]))^2] \\
&= \mathbb{E}[(bX-b\mathbb{E}[X])^2] \\
&= \mathbb{E}[b^2(X-\mathbb{E}[X])^2] \\
&= b^2\mathbb{E}[(X-\mathbb{E}[X])^2] \\
&= b^2\,\mathrm{var}(X)\checkmark
\end{aligned}
\tag{10.2}
$$

Other formula for variance

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

Proof.

$$\begin{aligned} \operatorname{var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}\Big[(X^2 - 2X \cdot \mathbb{E}[X] + (\mathbb{E}[X])^2\Big] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X] \cdot \mathbb{E}[X] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \checkmark \end{aligned} \tag{10.3}$$

Now calculate variance of coin flip this way:

$$\operatorname{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2.$$

Let's calculate the mean and variance of a fair die; i.e. if $X \in \{1, 2, 3, 4, 5, 6\}$ with equal probability.

Then,

$$\mathbb{E}[X] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{1}{6}21 = \frac{7}{2} = 3.5.$$

$$\operatorname{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{6}\Big(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) - 3.5^2 = \frac{35}{12}.$$

That implies $sd(X) = \sqrt{\operatorname{var}(X)} \approx 1.71$.

Variance under linear transformation. (Theorem) Suppose we define a random variable

$$Y = bX + a,$$

where a and b are constants. Then,

$$\operatorname{var}(Y) = b^2 \operatorname{var}(X).$$

Example Suppose X is temperature in Fahrenheit and Y is temperature in celsius. How much is the variance affected by changing units? The variance is not affected by translation; only scaling.

Proof.

$$\begin{aligned}\operatorname{var}(Y) &= \mathbb{E}[(Y-\mathbb{E}[Y])^2] \\ &= \mathbb{E}[(a+bX-(a+b\mathbb{E}[X]))^2] \\ &= \mathbb{E}[(bX-b\mathbb{E}[X])^2] \\ &= \mathbb{E}[b^2(X-\mathbb{E}[X])^2] \\ &= b^2\mathbb{E}[(X-\mathbb{E}[X])^2] \\ &= b^2 \operatorname{var}(X)\checkmark\end{aligned} \tag{10.4}$$

11 Covariance

Definition (Covariance). Let $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. Then we define

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

Note: $\operatorname{Cov}(X,X) = \operatorname{var}(X)$. Also, $\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X)$.

Covariance Formula for discrete rv Let $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. Let $p_{X,Y}(x,y)$ be the joint pmf of X and Y. Then,

$$\operatorname{Cov}(X,Y) = \sum_{(x,y)} (x-\mu_X)\cdot(y-\mu_Y)p_{X,Y}(x,y)$$

Covariance Formula for continuous rv Let $f_{X,Y}(x,y)$ be the joint density for X and Y.

$$\operatorname{Cov}(X,Y) = \int\int (x-\mu_X)\cdot(y-\mu_Y)f_{X,Y}(x,y)dxdy.$$

Other formula for covariance

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

Proof.

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}[XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[\mu_Y X] - \mathbb{E}[\mu_X Y] + \mathbb{E}[\mu_X\mu_Y] \\ &= \mathbb{E}[XY] - \mu_Y\mathbb{E}[X] - \mu_X\mathbb{E}[Y] + \mu_X\mu_Y \\ &= \mathbb{E}[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y \\ &= \mathbb{E}[XY] - \mu_X\mu_Y \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned} \tag{11.1}$$

11.1 Covariance under shifting, scaling, and addition

1. (Shifting by a constant a)

$$\mathrm{Cov}(a+X,Y) = \mathrm{Cov}(X,Y).$$

$$\begin{aligned}\mathrm{Cov}(a+X,Y) &= \mathbb{E}[(a+X-\mathbb{E}[a+X])\cdot(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[(a+X-a-\mathbb{E}[X])\cdot(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[(X-\mathbb{E}[X])\cdot(Y-\mathbb{E}[Y])] = \mathrm{Cov}(X,Y)\checkmark\end{aligned} \tag{11.2}$$

2. Scaling by constants a and b,

$$\mathrm{Cov}(aX,bY) = ab\mathrm{Cov}(X,Y).$$

3. Adding random variables

$$\mathrm{Cov}(X,Y+Z) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,Z).$$

If we combine 1,2,3 we conclude that for all random variables W, X, Y,Z, and all constants a, b, c, d that

$$\mathrm{Cov}(aW+bX, cY+dZ) = ac\mathrm{Cov}(W,Y) + bc\mathrm{Cov}(X,Y) + ad\mathrm{Cov}(W,Z) + bd\mathrm{Cov}(X,Z).$$

For a set of n variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, let $U = X_1 + \cdots X_n$ and $V = Y_1 + \cdots Y_n$. Then, we have Based on our rules for covariance, the following relation always holds

$$\mathrm{Cov}(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i,Y_j) \tag{11.3}$$

11.2 Uncorrelated random variables.

Two random variables X and Y are said to be uncorrelated if $\mathrm{Cov}(X,Y)=0$.

Also, we say $X_1,\dots,X_n$ are uncorrelated if $\mathrm{Cov}(X_i,X_j)=0$ for all $i\neq j$.

Claim. If X and Y are independent, then they are also uncorrelated.

Proof. Let $U=X-\mathbb{E}[X]$ and $V=Y-\mathbb{E}[Y]$.

Then,

$$\mathrm{Cov}(X,Y)=\mathbb{E}[(X-\mathbb{E}[X])\cdot(Y-\mathbb{E}[Y])]=E[U\cdot V]=\mathbb{E}[U]\mathbb{E}[V]=0\cdot 0=0\checkmark.$$

Remark. If X and Y are uncorrelated, this does not imply they are independent. (i.e. it's possible to come up with examples where X and Y are uncorrelated by not independent.

Theorem. If $X_1,\dots,X_n$ are uncorrelated. Then,

$$\mathrm{var}(X_1+\cdots X_n)=\mathrm{var}(X_1)+\cdots\mathrm{var}(X_n).$$

Proof.

$$\begin{aligned}\mathrm{var}(X_1+\cdots X_n)&=\mathrm{Cov}(X_1+\cdots X_n,X_1+\cdots X_n)\\&=\sum_{i=1}^n\sum_{j=1}^n\mathrm{Cov}(X_i,X_j)\\&=\sum_{i=1}^n\mathrm{Cov}(X_i,X_i)\\&=\sum_{i=1}^n\mathrm{var}(X_i)\end{aligned}\tag{11.4}$$

Where the third line follows since we assume the variables are uncorrelated, then $\mathrm{Cov}(X_i,X_j)=0$ if $i\neq j$.

11.3 Correlation

$$cor(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}.$$

Always between -1 and +1. Compare with covariance. The magnitude of covariance can be very large even if the variables have small correlation.

Correlation is not affected by standardizing the random variables, but covariance is.

12 Conditional expectation

$$\mathbb{E}[Y|X=x] = \sum_y yp(y|x).$$

$$\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} yf_{Y|X}(y|x)dy.$$

$$\mathbb{E}[h(Y)|X=x] = \int_{-\infty}^{\infty} h(y)f_{Y|X}(y|x)dy.$$

Remark (Conditional expectation as an rv). There are different ways we can think about conditional expectation.

For any fixed x there is a well defined number $\mathbb{E}[Y|X=x]$. This means that we can define a function $g:\mathbb{R}\to\mathbb{R}$ such that

$$g(x) := \mathbb{E}[Y|X=x]$$

e.g. $g(5) = \mathbb{E}[Y|X=5]$.

Now, we are allowed to plug anything we want into g, so let's plug in a random choice of x, say $g(X)$.

Now, in terms of the definition, we would have

$$g(X) = \mathbb{E}[Y|X=X],$$

but this is a little bit strange looking, and so it's more common to write

$$\mathbb{E}[Y|X],$$

but the real meaning of this symbol is $g(X)$.

12.1 Tower property

Theorem For any two random variables X and Y we have

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]].$$

This is the same as

$$\mathbb{E}[Y] = \mathbb{E}[g(X)]$$

Proof. (Discrete case)

To prove the theorem, we must show that

$$\mathbb{E}[Y] = \mathbb{E}[g(X)],$$

which is the same as

$$\mathbb{E}[Y] = \sum_x \mathbb{E}[Y|X = x]p_X(x)](*)$$

To show this, recall $\mathbb{E}[Y|X = x] = \sum_y yp_{Y|X}(y|x)$. Let's calculate RHS of (*).

$$\begin{aligned}\mathbb{E}[g(X)] &= \sum_x g(x)p_X(x)\\ &= \sum_x \mathbb{E}[Y|X = x]p_X(x)\\ &= \sum_x \Big(\sum_y yp_{Y|X}(y|x)\Big)p_X(x)\\ &= \sum_y \sum_x yp_{Y|X}(y|x)p_X(x)\\ &= \sum_y y\sum_x p_{Y|X}(y|x)p_X(x)\\ &= \sum_y y\sum_x p_{X,Y}(x,y)\\ &= \sum_y yp_Y(y)\\ &= \mathbb{E}[Y]\checkmark.\end{aligned} \tag{12.1}$$

Example Suppose you are an insurance company in a town where there is an earthquake. People who own homes are going to come to you and say, I paid my insurance premiums, and my house was destroyed, so you need to buy me a new house. (That's called an insurance claim.) So, as the insurance company, there are two things for you to consider.

One is, how big are most of the insurance claims? Let's call these $X_1, X_2, \ldots,$. Another question is, how many of these claims are there going to be? Let's call that number N. What kind of variable might you use to model N? (Poisson).

So, we are interested in the size of the random variable

$$T = \sum_{i=1}^{N} X_i.$$

We want to calculate $\mathbb{E}[T]$. However, we have to be careful because the number of things we are summing is also random.

Using conditional expectation can solve this problem. Let's assume that N and the X_i are independent. Also let $\mathbb{E}[X_i] = \mu$ for all $i = 1, \ldots, n$.

$$\mathbb{E}[T] = \mathbb{E}[\mathbb{E}[T|N]] = \mathbb{E}\Big[\mathbb{E}\Big[\sum_{i=1}^{N} X_i \Big| N\Big]\Big].$$

Now, since $g(n) = \mathbb{E}\Big[\sum_{i=1}^{N} X_i \Big| N = n\Big] = n\mu$, that implies

$$g(N) = \mathbb{E}[T|N] = N\mu.$$

Hence,

$$\mathbb{E}[T] == \mathbb{E}[\mathbb{E}[T|N]] = \mathbb{E}[g(N)] = \mathbb{E}[N \cdot \mu] = \mu \cdot \mathbb{E}[N].$$

12.2 Law of total variance

Let X and Y be random variables and suppose $\mathrm{var}(Y)$ exists.

Define the conditional variance of Y given $X = x$ to be

$$\mathrm{var}(Y|X = x) = \mathbb{E}[Y^2|X = x] - (\mathbb{E}[Y|X = x])^2.$$

So if we define the function $\varphi(x) := \mathrm{var}(Y|X = x)$, then we define

$$\mathrm{var}(Y|X) := \varphi(X).$$

Law of total variance

$$\mathrm{var}(Y) = \mathrm{var}(\mathbb{E}[Y|X]) + \mathbb{E}[\mathrm{var}(Y|X)].$$

Remark. We are free to choose any variable X we want. Often, we choose X as something convenient to condition on.

Application, calculate $\operatorname{var}(T)$. Let's again look at the sum

$$T = X_1 + \cdots X_N.$$

We assume

- N is random, independent of the X_i, and $\operatorname{var}(N)$ exists.
- X_i are independent
- $\operatorname{var}(X_i) = \sigma^2$
- $\mathbb{E}[X_i] = \mu$ for all i.

Given these assumptions, we want to calculate $\operatorname{var}(T)$. By the formula

$$\operatorname{var}(T) = \underbrace{\mathbb{E}[\operatorname{var}(T|N)]}_{a} + \underbrace{\operatorname{var}(\mathbb{E}[T|N])}_{b}.$$

Let's calculate b first. We know that $\mathbb{E}[T|N] = \mu N$, and so

$$b = \operatorname{var}(\mathbb{E}[T|N]) = \operatorname{var}(\mu N) = \mu^2 \operatorname{var}(N).$$

Now let's calculate a. Note that if $N = n$, then since the X_i and N are all independent,

$$\varphi(n) = \operatorname{var}(T|N = n) = \operatorname{var}(X_1 + \cdots + X_n) = \operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n) = n\sigma^2.$$

So, $\varphi(n) = n\sigma^2$, and so

$$\varphi(N) = \operatorname{var}(T|N) = \sigma^2 N.$$

This means

$$a = \mathbb{E}[\operatorname{var}(T|N)] = \mathbb{E}[\sigma^2 N] = \sigma^2 \mathbb{E}[N].$$

Altogether

$$\operatorname{var}(T) = a + b = \mu^2 \operatorname{var}(N) + \sigma^2 \mathbb{E}[N].$$

Concrete instance. Let's analyze health care costs for the university. Let's suppose some number N of people on the university health plan need surgery next year.

What might be a reasonable way to model N?
How about Poisson with mean $\lambda = n \cdot p = 40000 \cdot (1/100) = 400$

Next, for each person that has a surgery, how much will this cost? Call this X_i and say it has mean $\mu = \mathbb{E}[X_i] = 20,000$ and sd $\sigma = sd(X_i) = 10,000$, i.e. $\sigma^2 = 10^8$.

The university wants to know $\mathbb{E}[T]$ and $\sqrt{\text{var}(T)}$.

From Tower Property we get

$$\mathbb{E}[T] = \mu\mathbb{E}[N] = (20,000) \cdot 400 = 8 \cdot 10^4 \cdot 10^2 = 8Million$$

Then, we get, noting for Poisson, $\mathbb{E}[N] = \text{var}(N) = \lambda = 400$

$$\text{var}(T) = \mu^2 \,\text{var}(N) + \mathbb{E}[N]\sigma^2 = (20,000)^2 \cdot 400 + 400 \cdot 10^8 = 2 \cdot 10^{11}$$

so

$$\sqrt{\text{var}(T)} \approx 450,000.$$

Hence, the insurance company might guess that the actual cost in 2018 would be between $\mathbb{E}[T]-3\sigma$ and $\mathbb{E}[T] + 3\sigma$, or

$$8Million \pm 3 \cdot 450k$$

13 Moment generating functions

The mgf is a function "attached" to a random variable that packages all of the information related to the rv.

Definition of mgf

$$M_X(t) = \sum_x e^{tx}p(x)$$

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx.$$

$$M_X(t) = \mathbb{E}[e^{tX}].$$

Theorem. (Mgf determines a distribution) Suppose two random variables X and Y have the same mgf on an interval around 0, i.e.

$$M_X(t) = M_Y(t) \quad \text{for all} - \varepsilon < t < \varepsilon.$$

for some small ε. The, X and Y have the same distribution, i.e. for any number s,

$$\mathbb{P}(X \leq s) = \mathbb{P}(Y \leq s).$$

Theorem (Moments from the mgf.) We can calculate all of the moments of a random variable by taking derivatives of the mgf.

$$M_X'(0) = \mathbb{E}[X].$$

and

$$M_X''(0) = \mathbb{E}[X^2].$$

and in general

$$M_X^{(r)}(0) = \mathbb{E}[X^r],$$

for all $r \geq 1$.

Remark. What does the notation mean? Note that the derivative of $M_X(t)$ is another function, say $g(t) = M_X'(t)$. When we plug the value $t = 0$ into g, we write this as $M_X'(0) = g(0)$.

Example for Poisson.

$$\begin{aligned} M(t) &= \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} \\ &= e^{-\lambda} e^{\lambda e^t} \\ &= e^{\lambda(e^t - 1)} \end{aligned} \tag{13.1}$$

$$M'(t) = e^{\lambda(e^t-1)} \cdot \lambda e^t.$$
$$M'(0) = e^{\lambda(1-1)} \cdot \lambda e^0 = \lambda = \mathbb{E}[X]\checkmark$$

$$M''(t) = e^{\lambda(e^t-1)} \cdot (\lambda e^t)^2 + e^{\lambda(e^t-1)} \cdot \lambda e^t.$$
$$M''(0) = 1 \cdot \lambda^2 + 1 \cdot \lambda = \lambda^2 + \lambda.$$

So,

$$\mathbb{E}[X^2] = \lambda^2 + \lambda.$$

So,

$$\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X]^2) = \lambda^2 + \lambda - (\lambda^2) = \lambda\checkmark.$$

13.1 MGF under linear function

Consider

$$Y = a + bX.$$

Then,

$$M_Y(t) = e^{at} M_X(bt).$$

Proof.

$$M_Y(t) = \mathbb{E}[e^{t(a+bX)}]\mathbb{E}[e^{at} \cdot e^{tbX}] = e^{at}\mathbb{E}[e^{tbX}] = e^{at} M_X(bt)\checkmark.$$

13.2 MGF for independent sums

If X and Y are independent, and $Z = X + Y$, then

$$M_Z(t) = M_Y(t) M_X(t).$$

This is very useful because it can allow us to find the distribution of the sum of two independent rv.

Let's show that the sum of X, poisson λ_1 and Y, poisson λ_2 is Poisson $\lambda_1 + \lambda_2$.

$$Z = X + Y.$$

$$M_Z(t) = M_X(t)M_Y(t) = e^{\lambda_1(e^t-1)} \cdot e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}.\checkmark$$

This also implies that the sum of n independent poison with parameters $\lambda_1, \ldots, \lambda_n$ is Poisson with parameter $\lambda_1 + \cdots + \lambda_n$.

14 Limit theorems

This section is about understanding sums of independent random variables. We will use the notation

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

Earlier in the course, I claimed to you that if the X_i's are independent, and have $E[X_i] = \mu$ for all $i = 1, 2, ...$, then when n is large,

$$\bar{X}_n \approx \mu.$$

This fact is called the law of large numbers. Our goal is to make the $\approx$ more precise.

First, we need to review some useful inequalities.

14.1 Markov and Chebyshev

Markov's inequality For any non-negative rv Y, and any $t > 0$, we have

$$\mathbb{P}(Y > t) \leq \frac{\mathbb{E}[Y]}{t}.$$

We proved this earlier in the course.

Chebyshev's inequality. Let W be any rv with finite variance. Then for any $t > 0$,

$$\mathbb{P}(|W - \mathbb{E}[W]| > t) \leq \frac{\mathrm{var}(W)}{t^2}.$$

Remark. This is say that when t is large, the random variable W is likely to be in the interval

$$E[W] - t \leq W \leq E[W] + t.$$

Proof. Consider the non-negative random variable $Y = (W - E[W])^2$, and let s and t be two numbers such that $s = t^2$. Then Markov's inequality says

$$P(|W-E[W]| > t) = P((W-E[W])^2 > t^2) = P(Y > s) \leq \frac{\mathbb{E}[Y]}{s} = \frac{1}{t^2}\mathbb{E}[(W-E[W])^2] = \frac{\text{var}(W)}{t^2}.\checkmark$$

Example. Let's consider the choice $t = 3\sigma$ in Chebyshev. Then, for any rv that has a variance, we have

$$\mathbb{P}(|X - \mu| > 3\sigma) \leq \frac{\sigma^2}{9\sigma^2} = \frac{1}{9}.$$

Hence, no matter what distribution X has – as long as its variance is finite, the chance that it is more than 3 sd from the mean is only 1/9.

14.2 Review: Convergence of sequences of real numbers

Before we can speak of the law of large numbers, we need to recall what it means for a sequence of real numbers to converge.

Let $a_1, a_2, \ldots$ be a sequence of real numbers. (Also written $\{a_n\}$.)

Definition. We say $a_n \to a$ as $n \to \infty$ if for any $\delta > 0$, there is a number N_δ, such that for all $n \geq N_\delta$, we have that

$$|a_n - a| < \delta.$$

14.3 The law of large numbers (LLN)

Theorem. Let $X_1, X_2, \ldots$ be a sequence of independent random variables with the same expectation $E[X_i] = \mu$ and variance $var(X_i) = \sigma^2$, for all $i = 1, 2, \ldots$. Then, for any $\varepsilon > 0$, we have

$$P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2 n} \to 0 \quad \text{as} \quad n \to \infty.$$

Remarks. Let's pause and think about what this statement means.

If we consider the sequence of numbers $a_n := P(|\bar{X}_n - \mu| > \varepsilon)$, the law of large numbers is saying that $a_n \to 0$ as $n \to \infty$.

What does the LLN mean?

Let's think about the event $\{|\bar{X}_n - \mu| > \varepsilon\}$. When this event happens, it means that $\bar{X}_n$ is not close to $\mu = E[X_1]$. So as n becomes large, the LLN is saying that this event becomes very unlikely.

In other words, if we set $\varepsilon = 0.01$, and we think of $X_1, X_2, \ldots$ as a sequence of fair coins, where $\mathbb{E}[X_i] = \mu = 1/2$, then the LLN says that if n is large enough, then the probability

$$P(|\bar{X}_n - 0.5| > 0.01) = P(\bar{X}_n > 0.51 \quad \text{or} \quad \bar{X} < 0.49)$$

becomes very small.

Proof of LLN. From a homework problem, we know that

$$E[\bar{X}] = \mu, \quad var(X) = \sigma^2/n.$$

For any fixed $\varepsilon > 0$, using Chebyshev's inequality we have

$$P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{var(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n} \to 0 \quad \text{as } n \to \infty.$$

Remark. The bound in the LLN provides a way to choose n. For example, if we want our estimate $\bar{X}_n$ to be within $\varepsilon = 0.01$ from the true mean μ with probability at least 0.99, then we should choose n so that

$$\frac{\sigma^2}{\varepsilon^2 n} = \frac{\sigma^2}{(0.01)^2 n} \leq 1 - 0.99 = 0.01.$$

This implies

$$n \geq \frac{\sigma^2}{(0.01)^2 \times 0.01} = 10^6 \times \sigma^2.$$

If X_i are Bernoulli random variables with success probability $p = 0.1$, then $\sigma^2 = p(1-p) = 0.09$, so to estimate $p = 1/2$, we need a sample size $n = 0.09 \times 10^6 = 90000$. Since we don't know p in practice (we are estimating it), we use the fact that $p(1-p) \leq 1/4$ for all p and use a slightly larger (conservative) sample size $n = 10^6/4$.

14.4 CLT

For a general random variable X, it's often of interest to approximately know $F(x) = P(X \leq x)$, even if we can't compute it exactly.

The CLT allows us to approximate cdf's when X is a sum of independent rv.

Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p) rv, and let

$$S_n = X_1 + \cdots + X_n.$$

Let's look at

$$F_(s) = P(S_n \leq s) = \sum_{k=0}^{\lfloor s \rfloor} P(S = k) = \sum_{k=0}^{\lfloor s \rfloor} \binom{n}{k} p^k (1-p)^{n-k}.$$

This is an ugly function of s. Want something simpler.

Specifically, we want to find an approximation for situations where the number of terms n is large.

That is, we want to understand the cdf of S_n as $n \to \infty$.

However, we have to be careful about how we do this because as $n \to \infty$,

$$\mathbb{E}[S_n] = np \to \infty,$$

and

$$\operatorname{var}(S_n) = np(1-p) \to \infty.$$

Instead, we will standardize S_n and then study an approximation for large n.

Standardization means adjusting S_n so that it has mean 0 and variance 1.

Define

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}.$$

Note, $\mathbb{E}[Z_n] = 0$ and $\operatorname{var}(Z_n) = 1$.

Hence, even as n gets huge, the mean and variance of Z_n don't change. Also, if we have the cdf of Z_n we can get the cdf of S_n since

$$
\begin{aligned}
F_{S_n}(s) &= P(S_n \leq s) && (14.1)\\
&= P(\frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{s - np}{\sqrt{np(1-p)}}) && (14.2)\\
&= P(Z_n \leq \frac{s - np}{\sqrt{np(1-p)}}) && (14.3)\\
&= F_{Z_n}(\frac{s - np}{\sqrt{np(1-p)}}). && (14.4)\\
& && (14.5)
\end{aligned}
$$

The idea is to come up with a nice formula for F_n say

$$F_{Z_n} \approx G$$

and then calculate properties about S_n by using the approximation

$$F_{S_n}(s) \approx G(\frac{s - np}{\sqrt{np(1-p)}}).$$

Goal Approximate cdfs for variables that look like Z_n.

Defn: Convergence in distribution

Suppose Z is a random variable that has cdf G.

We say a sequence of variables $W_1, W_2, \ldots, W_n, \ldots$ converges in distribution to Z if

$$F_{W_n}(z) \to G(z),$$

at every point z where the function G is continuous.

CLT background Recall that $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2.

When $\mu = 0$ and $\sigma^2 = 1$ we call this standard normal

The cdf for a standard normal distribution is Φ i.e.

$$P(Z \leq z) = \Phi(z),$$

where Z is a standard normal rv.

The CLT Let $X_1, X_2, \ldots,$ be i.i.d rv with $\mathbb{E}[X_1] = \mu$ and $\text{var}(X_1) = \sigma^2$. Also put

$$Z_n = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{1}{\sigma^2} \times \sqrt{n}(\bar{X} - \mu).$$

Then, Z_n convergences in distribution to a standard normal distribution $N(0, 1)$, i.e.

$$P(Z_n \leq z) \to \Phi(z),$$

as $n \to \infty$.

Example 70 coin flips with $p = 0.5$

$$S = X_1 + \cdots + X_{70}.$$

Use CLT to estimate $P(S \leq 37)$.

$$Z_{70} = \frac{S/70 - 0.5}{\sqrt{0.5 \cdot 0.5/70}}.$$

$$\begin{aligned}
P(S \leq 37) &= P\left(\frac{S/70 - 0.5}{\sqrt{0.5 \cdot 0.5/70}} \leq \frac{37/70 - 0.5}{\sqrt{0.5 \cdot 0.5/70}}\right) && (14.6)\\
&= P\left(Z_{70} \leq \frac{37/70 - 0.5}{\sqrt{0.5 \cdot 0.5/70}}\right) && (14.7)\\
&\approx \Phi\left(\frac{37/70 - 0.5}{\sqrt{0.5 \cdot 0.5/70}}\right) && (14.8)\\
&= 0.684 && (14.9)
\end{aligned}$$

R code: pbinom(37,70,0.5) = 0.7247937 and pnorm $\left(\frac{37/70-0.5}{\sqrt{0.5\cdot0.5/70}}\right)$ = 0.6837074

Continuity theorem

Suppose F_n are a sequence of cdfs with corresponding mgfs M_n. Also, let F be another cdf with mgf M

If $M_n(t) \to M(t)$ as $n \to \infty$ for all t in an interval around 0, then

$$F_n(x) \to F(x),$$

at all points where $F(x)$ is continuous.

Example: Poisson MGF expansion Let X_n be Poisson(λ_n) with $\lambda_n \to \infty$ as $n \to \infty$.

Want to show X_n is approximately normal when it is standardized and n is large.

$$Z_n = \frac{X_n - \mathbb{E}[X_n]}{\sqrt{\text{var}(X_N)}} = \frac{X_n - \lambda_n}{\sqrt{\lambda_n}}.$$

$$\begin{aligned}
M_{Z_n}(t) &= \mathbb{E}[e^{tZ_n}] \\
&= \mathbb{E}[e^{-t\sqrt{\lambda_n} + \frac{t}{\sqrt{\lambda_n}} X_n}] \\
&= e^{-t\sqrt{\lambda_n}} \cdot M_{X_n}(\frac{t}{\sqrt{\lambda_n}}) \\
&= \exp(-t\sqrt{\lambda_n}) \cdot \exp(\lambda_n(e^{\frac{t}{\sqrt{\lambda_n}}} - 1)) \\
&= \exp(-t\sqrt{\lambda_n} + \lambda_n(e^{\frac{t}{\sqrt{\lambda_n}}} - 1)
\end{aligned}$$

It's enough to show that $\log M_{Z_n}(t)$ converges to $\log M_Z(t)$ where $M_Z(t)$ is the mgf of a standard normal. Note that

$$M_Z(t) = \exp(t^2/2),$$

so

$$\log M_Z(t) = \frac{t^2}{2}.$$

Taking log of M_{Z_n} we get

$$\begin{aligned}
\log M_{Z_n}(t) &= -t\sqrt{\lambda_n} + \lambda_n(e^{\frac{t}{\sqrt{\lambda_n}}} - 1) \\
&= -t\sqrt{\lambda_n} + \lambda_n(1 + \tfrac{t}{\sqrt{\lambda}} + \tfrac{1}{2!}(\tfrac{t}{\sqrt{\lambda}})^2 + \tfrac{1}{3!}(\tfrac{t}{\sqrt{\lambda}})^3 + \cdots - 1) \\
&= \frac{t^2}{2} + \frac{1}{3!}\frac{t^3}{\lambda_n^{3/2}} + \cdots
\end{aligned}$$

As $n \to \infty$ we get

$$\log M_{Z_n}(t) \to \log M_Z(t).$$

as desired

Example Let's look at a Poisson $\lambda = 900$ variable X

Want to know

$$\begin{aligned}
P(X > 950) &= P(\frac{X - 900}{\sqrt{900}} > \frac{950 - 900}{\sqrt{900}}) && (14.10)\\
&\approx P(Z > \frac{950 - 900}{\sqrt{900}}) && (14.11)\\
&\approx 1 - \Phi(\frac{950 - 900}{\sqrt{900}}) && (14.12)\\
&= 0.04779. && (14.13)
\end{aligned}$$

The exact answer is 0.04712

Proof of CLT. For simplicity, we assume that $\sigma^2 = 1$ and $\mu = 0$. According to the Continuity Theorem, we only need to show that mgf of $Z_n = S_n/\sqrt{n}$ converges to that of standard normal random variable Φ. Since Z_n is a sum of independent and identically distributed random variables, we have

$$M_{Z_n}(t) = [M(t/\sqrt{n})]^n,$$

where $M(t)$ is the mgf of X_i. Note that $t/\sqrt{n} \to 0$ as $n \to \infty$, so we can use Taylor's expansion for M around zero:

$$M(s) = M(0) + M'(0)s + M''(0)s^2/2 + R(s),$$

where $R(s)$ is the remainder that goes to zero faster than s^2: $R(s)/s^2 \to 0$ as $s \to 0$. Since $M(0) = \mathbb{E}[e^{0 \cdot X_i}] = 1$ and $M'(0) = \mathbb{E}[X_i] = 0$, $M''(0) = \mathbb{E}[X_i^2] = 1$ by the Theorem about moments from the mgf, using $s = t/\sqrt{n}$, we have

$$M(t/\sqrt{n}) = 1 + (t/\sqrt{n})^2/2 + R(t/\sqrt{n}) = 1 + \frac{t^2}{2n} + R(t/\sqrt{n}),$$

where

$$\frac{R(t/\sqrt{n})}{(t/\sqrt{n}))^2} = \frac{nR(t/\sqrt{n})}{t^2} \to 0 \quad \text{as } n \to \infty.$$

Recall from calculus that if $a_n \to a$ then

$$\left(1 + \frac{a_n}{n}\right)^n \to e^a.$$

In this case, we will use

$$a_n = t^2/2 + nR(t/\sqrt{n}) \to t^2/2 = a,$$

and with that choice

$$M(t/\sqrt{n}) = 1 + a_n/n.$$

Therefore

$$M_{Z_n}(t) = [M(t/\sqrt{n})]^n = \Big(1 + \frac{a_n}{n}\Big)^n \to e^a = e^{t^2/2} = \Phi(t),$$

and the CLT is proved.

Confidence intervals. By CLT, if n is large then for any v we have

$$P\Big(\frac{\bar{X}_n - \mu}{\sigma_{\bar{X}_n}} \le v\Big) \approx \Phi(v).$$

Also, if $u, v > 0$ are fixed numbers then,

$$P\Big(-u \le \frac{\bar{X}_n - \mu}{\sigma_{\bar{X}_n}} \le v\Big) \approx \Phi(v) - \Phi(-u).$$

By rearranging, we get

$$P\Big(\bar{X}_n - v\sigma_{\bar{X}} \le \mu \le \bar{X}_n + u\sigma_{\bar{X}_n}\Big) \approx \Phi(v) - \Phi(-u). \quad (*)$$

Quantiles. Recall that $z(\alpha)$ is defined to be the number that satisfies $\Phi(z(\alpha)) = 1 - \alpha$. By symmetry it follows that $\Phi(-z(\alpha)) = \alpha$

Consider choosing $u = z(\alpha/2)$ and $v = z(\alpha/2)$ in (*). Then

$$\begin{aligned} P\Big(\bar{X}_n - z(\tfrac{\alpha}{2})\sigma_{\bar{X}} \le \mu \le \bar{X}_n + z(\tfrac{\alpha}{2})\sigma_{\bar{X}_n}\Big) &\approx \Phi(z(\tfrac{\alpha}{2}) - \Phi(-z(\tfrac{\alpha}{2}) \\ &= 1 - \frac{\alpha}{2} - \tfrac{\alpha}{2} \\ &= 1 - \alpha. \end{aligned}$$

Therefore if $\sigma_{\bar{X}_n}$ is known to us then the approximate $(1-\alpha)$-confidence interval for μ is

$$\big[\bar{X}_n - z(\tfrac{\alpha}{2})\sigma_{\bar{X}}, \bar{X}_n + z(\tfrac{\alpha}{2})\sigma_{\bar{X}_n}\big].$$

The numbers $\sigma_{\bar{X}_n}$ are unknown. We often can't use this in practice because the numbers $\sigma_{\bar{X}_n}$ are not known. Instead we must estimate these numbers.

Fact. Let $s^2 : \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Also, put

$$s_{\bar{X}}^2 := \frac{s^2}{n}\Big(1 - \frac{n}{N}\Big).$$

Then, $s_{\bar{X}}^2$ is an unbiased estimator for $\sigma_{\bar{X}}^2$, i.e.

$$E[s_{\bar{X}}^2] = \sigma_{\bar{X}}^2.$$

Practical confidence interval for μ Using $s_{\bar{X}}$ instead of $\sigma_{\bar{X}}$, we get for large n,

$$\mathbb{P}\Big(\mu \in [\bar{X}_n - z(\tfrac{\alpha}{2})s_{\bar{X}}\ ,\ \bar{X}_n + z(\tfrac{\alpha}{2})s_{\bar{X}_n}]\Big) \approx 1 - \alpha.$$

Practical confidence interval for p in dichotomous case If x_i are 0 or 1, then we want to get a confidence interval for p – the proportion of 1's in the population.

In this case we write

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

and

$$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}\Big(1 - \tfrac{n}{N}\Big),$$

is an unbiased estimate for $\sigma_{\hat{p}}^2$.

Example There is a population of 8000 people who own condominiums. We want to know what proportion p of these people plan to sell their condo within the next year.

We collect a simple random sample of size $n = 100$. Of the 100 people we contact, 12 of them say they plan to sell in the next year.

So

$$\hat{p} = 0.12.$$

and

$$s_{\hat{p}} = \sqrt{\frac{0.12(1-0.12)}{99}}\sqrt{1 - \tfrac{100}{8000}} = 0.03.$$

So, if we construct the interval with $\alpha = 0.05$, and $z(0.025) = 1.96$, then

$$[0.12 - 1.96 \cdot 0.03\ ,\ 0.12 + 1.96 \cdot 0.03],$$

$$[0.061\ ,\ 0.179]$$

is our 95% confidence interval for p.

Interpretation. If we were to repeat the experiment, the number we would get would be slightly different. However, if we repeated the experiment many times, the interval would contain p about 95% of the time.

15 The method of maximum likelihood

Suppose $X_1, \ldots, X_n$ have a joint density function $f(x_1, x_2, \ldots, x_n|\theta)$ where θ is an unknown parameter. If we observe $X_1 = x_1, ..., X_n = x_n$, how can we estimate θ?

With the x_i held fixed, we can view $f(x_1, \ldots, x_n|\theta)$ as a function of the θ, i.e.

$$\text{lik}(\theta) = f(x_1, \ldots, x_n|\theta).$$

If the samples $X_1, \ldots, X_n$ are i.i.d., then the likelihood function factors

$$\text{lik}(\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta).$$

Example If we have an i.i.d. sample of size n from Poisson(λ), then λ plays the role of θ and we have

$$\text{lik}(\lambda) = e^{-\lambda}\tfrac{\lambda^{x_1}}{x_1!}\, e^{-\lambda}\tfrac{\lambda^{x_2}}{x_2!}\cdots e^{-\lambda}\tfrac{\lambda^{x_n}}{x_n!}$$

The maximum likelihood principle Which choice of θ fits the data the best? What if $lik(\theta)$ is very small? That means the data we observed was unlikely. However, we usually don't observe unlikely things. Usually, we observe the things that are most likely to occur. This motivates the idea that we should choose the value of θ that makes $lik(\theta)$ the largest, i.e.

$$\hat{\lambda} = \text{argmax}_{\lambda\geq 0}\text{lik}(\lambda).$$

Example (Poisson) For an i.i.d. sample from Poisson(λ),

$$\begin{aligned}
l(\lambda) &= \log f(X_1|\lambda) + \cdots \log f(X_n|\lambda) = \\
&= \log\left(e^{-\lambda}\tfrac{\lambda^{X_1}}{X_1!}\right) + \cdots + \log\left(e^{-\lambda}\tfrac{\lambda^{X_n}}{X_n!}\right) \\
&= \Big[-\lambda + X_1\log(\lambda) - \log(X_1!)\Big] + \cdots + \Big[-\lambda + X_n\log(\lambda) - \log(X_n!)\Big] \\
&= -n\lambda + \log(\lambda)\Big(X_1 + \cdots + X_n\Big) - \Big(\log(X_1!) + \cdots + \log(X_n!)\Big)
\end{aligned}$$

How to maximize? Take derivative and set equal to 0.

Differentiating,

$$l'(\lambda) = -n + \frac{X_1 + \cdots + X_n}{\lambda} = 0.$$

Which implies

$$\hat{\lambda}_{mle} = \frac{1}{n}\Big(X_1 + \cdots + X_n\Big) = \bar{X}.$$

Question When does the equation $l'(\hat{\lambda}) = 0$ imply $\hat{\lambda}$ maximizes $l(\lambda)$? You need to check that $l''(\hat{\lambda}) < 0$.

In the previous example, $l''(\lambda) = -\frac{1}{\lambda^2}(X_1 + \cdots + X_n) < 0$ which does hold when $X_1 + \cdots + X_n > 0$.

Example: Normal The density function for $N(\mu, \tau)$ is $f(x|\mu, \tau) = \frac{1}{\sqrt{2\pi\tau}} \exp(-\frac{(x-\mu)^2}{2\tau})$. Here we write τ instead of σ^2 because that will make our calculations easier.

The log likelihood function for an i.i.d. sample $X_1, \ldots, X_n$ is

$$l(\mu, \sigma^2) = \sum_{i=1}^{n} \log f(X_i|\mu, \tau)$$

sample from $N(\mu, \sigma^2)$ is

$$f(X_1, \ldots, X_n|\mu, \tau) = \Big(\frac{1}{\sqrt{2\pi\tau}}\Big)^n \exp(-\frac{\sum_{i=1}^n (X_i - \mu)^2}{2\tau}).$$

The loglikelihood function is

$$l(\mu, \tau) = -\frac{n}{2}\log(\tau) - \frac{n}{2}\log(2\pi) - \frac{1}{2\tau}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Now we need to solve the equations

$$\frac{\partial l(\mu, \tau)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial l(\mu, \tau)}{\partial \tau} = 0.$$

Solving the first equation for μ,

$$\frac{\partial l(\mu, \tau)}{\partial \mu} = -\frac{1}{\tau}\sum_{i=1}^{n}(X_i - \mu) = 0,$$

i.e.

$$\Big(\sum_{i=1}^{n} X_i\Big) - n\mu = 0,$$

which gives

$$\hat{\mu}_{mle} = \bar{X}.$$

Now let's solve the equation for τ

$$\frac{\partial l(\mu,\tau)}{\partial \tau} = -\frac{n}{2\tau} + \frac{1}{\tau^2}\sum_{i=1}^{n}(X_i - \mu)^2 = 0.$$

Solving for τ we get

$$\tau = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Now, remember that since we're solving the equations simultaneously, we plug in $\mu = \hat{\mu}_{mle} = \bar{X}$, we get

$$\hat{\tau}_{mle} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Example (Uniform) Consider a uniform distribution on the interval $[0,\theta]$; i.e. we don't know how long the interval is.

The density function is given by

$$f(x|\theta) = \frac{1}{\theta}1\{0 \le x \le \theta\}.$$

Hence, for an i.i.d. sample $X_1, \ldots, X_n \sim$ uniform$[0,\theta]$, we have

$$\begin{aligned} lik(\theta) &= f(X_1|\theta)\cdots f(X_n|\theta) && (15.1)\\ &= \Big(\frac{1}{\theta}1\{0 \le X_1 \le \theta\}\Big)\cdots\Big(\frac{1}{\theta}1\{0 \le X_n \le \theta\}\Big) && (15.2)\\ &= \Big(\frac{1}{\theta}\Big)^n \cdot 1\{0 \le X_{(1)} \le X_{(n)} \le \theta\}. && (15.3) \end{aligned}$$

To make lik(θ) as big as possible, we have to make θ as small as possible; we see that this occurs when $\theta = X_{(n)}$, i.e.

$$\hat{\theta}_{mle} = X_{(n)}.$$

Example (Angular distribution) Consider the density

$$f(x|\alpha) = \frac{1 + \alpha x}{2} \qquad -1 \leq x \leq 1,$$

and the parameter lies in $\alpha \in [-1, 1]$.

The log likelihood function is

$$l(\alpha) = \sum_{i=1}^{n} \log\Big(1 + \alpha X_i\Big) - \log(2).$$

Setting the derivative to 0 gives

$$l'(\alpha) = \sum_{i=1}^{n} \frac{X_i}{1 + \alpha X_i} = 0$$

This equation must be solved numerically.