Introduction
Class 1, 18.05
Jeremy Orloff and Jonathan Bloom

1 Probability vs. Statistics

In this introduction we will preview what we will be studying in 18.05. Don't worry if many of the terms are unfamiliar, they will be explained as the course proceeds.

Probability and statistics are deeply connected because all statistical statements are at bottom statements about probability. Despite this the two sometimes feel like very different subjects. Probability is logically self-contained; there are a few rules and answers all follow logically from the rules, though computations can be tricky. In statistics we apply probability to draw conclusions from data. This can be messy and usually involves as much art as science.

Probability example
You have a fair coin (equal probability of heads or tails). You will toss it 100 times. What is the probability of 60 or more heads? There is only one answer (about 0.028444) and we will learn how to compute it.

Statistics example
You have a coin of unknown provenance. To investigate whether it is fair you toss it 100 times and count the number of heads. Let's say you count 60 heads. Your job as a statistician is to draw a conclusion (inference) from this data. There are many ways to proceed, both in terms of the form the conclusion takes and the probability computations used to justify the conclusion. In fact, different statisticians might draw different conclusions.

Note that in the first example the random process is fully known (probability of heads = 0.5). The objective is to find the probability of a certain outcome (at least 60 heads) arising from the random process. In the second example, the outcome is known (60 heads) and the objective is to illuminate the unknown random process (the probability of heads).

2 Frequentist vs. Bayesian Interpretations

There are two prominent and sometimes conflicting schools of statistics: Bayesian and frequentist. Their approaches are rooted in differing interpretations of the meaning of probability.

Frequentists say that probability measures the frequency of various outcomes of an experiment. For example, saying a fair coin has a 50% probability of heads means that if we toss it many times then we expect about half the tosses to land heads.

Bayesians say that probability is an abstract concept that measures a state of knowledge or a degree of belief in a given proposition. In practice Bayesians do not assign a single value for the probability of a coin coming up heads. Rather they consider a range of values each with its own probability of being true.

In 18.05 we will study and compare these approaches. The frequentist approach has long

been dominant in fields like biology, medicine, public health and social sciences. The Bayesian approach has enjoyed a resurgence in the era of powerful computers and big data. It is especially useful when incorporating new data into an existing statistical model, for example, when training a speech or face recognition system. Today, statisticians are creating powerful tools by using both approaches in complementary ways.

3 Applications, Toy Models, and Simulation

Probability and statistics are used widely in the physical sciences, engineering, medicine, the social sciences, the life sciences, economics and computer science. The list of applications is essentially endless: tests of one medical treatment against another (or a placebo), measures of genetic linkage, the search for elementary particles, machine learning for vision or speech, gambling probabilities and strategies, climate modeling, economic forecasting, epidemiology, marketing, googling…We will draw on examples from many of these fields during this course.

Given so many exciting applications, you may wonder why we will spend so much time thinking about toy models like coins and dice. By understanding these thoroughly we will develop a good feel for the simple essence inside many complex real-world problems. In fact, the modest coin is a realistic model for any situations with two possible outcomes: success or failure of a treatment, an airplane engine, a bet, or even a class.

Sometimes a problem is so complicated that the best way to understand it is through computer simulation. Here we use software to run *virtual* experiments many times in order to estimate probabilities. In this class we will use R for simulation as well as computation and visualization. Don't worry if you're new to R; we will teach you all you need to know.

Counting and Sets
Class 1, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the definitions and notation for sets, intersection, union, complement.
2. Be able to visualize set operations using Venn diagrams.
3. Understand how counting is used computing probabilities.
4. Be able to use the rule of product, inclusion-exclusion principle, permutations and combinations to count the elements in a set.

2 Counting

2.1 Motivating questions

Example 1. A coin is *fair* if it comes up heads or tails with equal probability. You flip a fair coin three times. What is the probability that exactly one of the flips results in a head?

Solution: With three flips, we can easily list the eight possible outcomes:

$$\{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}$$

Three of these outcomes have exactly one head:

$$\{TTH, THT, HTT\}$$

Since all outcomes are equally probable, we have

$$P(\text{1 head in 3 flips}) = \frac{\text{number of outcomes with 1 head}}{\text{total number of outcomes}} = \frac{3}{8}.$$

Think: Would listing the outcomes be practical with 10 flips?

A deck of 52 cards has 13 ranks (2, 3, ... , 9, 10, J, Q, K, A) and 4 suits ($\heartsuit, \spadesuit, \diamondsuit, \clubsuit,$). A poker hand consists of 5 cards. A *one-pair* hand consists of two cards having one rank and three cards having three other ranks, e.g., $\{2\heartsuit, 2\spadesuit, 5\heartsuit, 8\clubsuit, \text{K}\diamondsuit\}$

Test your intuition: the probability of a one-pair hand is:

(a) less than 5%
(b) between 5% and 10%
(c) between 10% and 20%
(d) between 20% and 40%
(e) greater than 40%

At this point we can only guess the probability. One of our goals is to learn how to compute it exactly. To start, we note that since every set of five cards is equally probable, we can compute the probability of a one-pair hand as

$$P(\text{one-pair}) = \frac{\text{number of one-pair hands}}{\text{total number of hands}}$$

So, to find the exact probability, we need to count the number of elements in each of these sets. And we have to be clever about it, because there are too many elements to simply list them all. We will come back to this problem after we have learned some counting techniques.

Several times already we have noted that all the possible outcomes were equally probable and used this to find a probability by counting. Let's state this carefully in the following principle.

Principle: Suppose there are n possible outcomes for an experiment and each is equally probable. If there are k desirable outcomes then the probability of a desirable outcome is k/n. Of course we could replace the word desirable by any other descriptor: undesirable, funny, interesting, remunerative, …

Concept question: Can you think of a scenario where the possible outcomes are not equally probable?

Here's one scenario: on an exam you can get any score from 0 to 100. That's 101 different possible outcomes. Is the probability you get less than 50 equal to 50/101?

2.2 Sets and notation

Our goal is to learn techniques for counting the number of elements of a set, so we start with a brief review of sets. (If this is new to you, please come to office hours).

2.2.1 Definitions

A **set** S is a collection of elements. We use the following notation.

Element: We write $x \in S$ to mean the element x is in the set S.

Subset: We say the set A is a subset of S if all of its elements are in S. We write this as $A \subset S$.

Complement:: The complement of A in S is the set of elements of S that are **not** in A. We write this as A^c or $S - A$.

Union: The union of A and B is the set of all elements in A **or** B (or both). We write this as $A \cup B$.

Intersection: The intersection of A and B is the set of all elements in both A **and** B. We write this as $A \cap B$.

Empty set: The empty set is the set with no elements. We denote it $\emptyset$.

Disjoint: A and B are **disjoint** if they have no common elements. That is, if $A \cap B = \emptyset$.

Difference: The difference of A and B is the set of elements in A that are not in B. We write this as $A - B$.

Let's illustrate these operations with a simple example.

Example 2. Start with a set of 10 animals

$$S = \{\text{Antelope, Bee, Cat, Dog, Elephant, Frog, Gnat, Hyena, Iguana, Jaguar}\}.$$

Consider two subsets:

$M =$ the animal is a mammal $= \{\text{Antelope, Cat, Dog, Elephant, Hyena, Jaguar}\}$
$W =$ the animal lives in the wild $= \{\text{Antelope, Bee, Elephant, Frog, Gnat, Hyena, Iguana, Jaguar}\}$.

Our goal here is to look at different set operations.

Intersection: $M \cap W$ contains all wild mammals: $M \cap W = \{\text{Antelope, Elephant, Hyena, Jaguar}\}$.
Union: $M \cup W$ contains all animals that are mammals or wild (or both).
$M \cup W = \{\text{Antelope, Bee, Cat, Dog, Elephant, Frog, Gnat, Hyena, Iguana, Jaguar}\}$.
Complement: M^c means everything that is *not* in M, i.e. not a mammal. $M^c = \{\text{Bee, Frog, Gnat, Iguana}\}$.
Difference: $M - W$ means everything that's in M and not in W. So, $M - W = \{\text{Cat, Dog}\}$.

There are often many ways to get the same set, e.g. $M^c = S - M$, $M - W = M \cap W^c$.

The relationship between union, intersection, and complement is given by DeMorgan's laws:

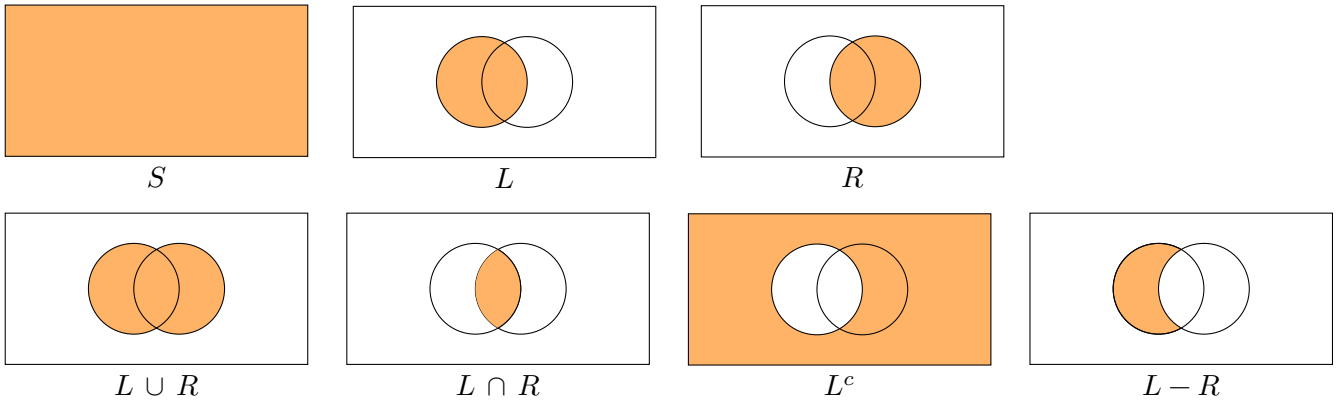
$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

In words the first law says everything not in (A or B) is the same set as everything that's (not in A) and (not in B). The second law is similar.

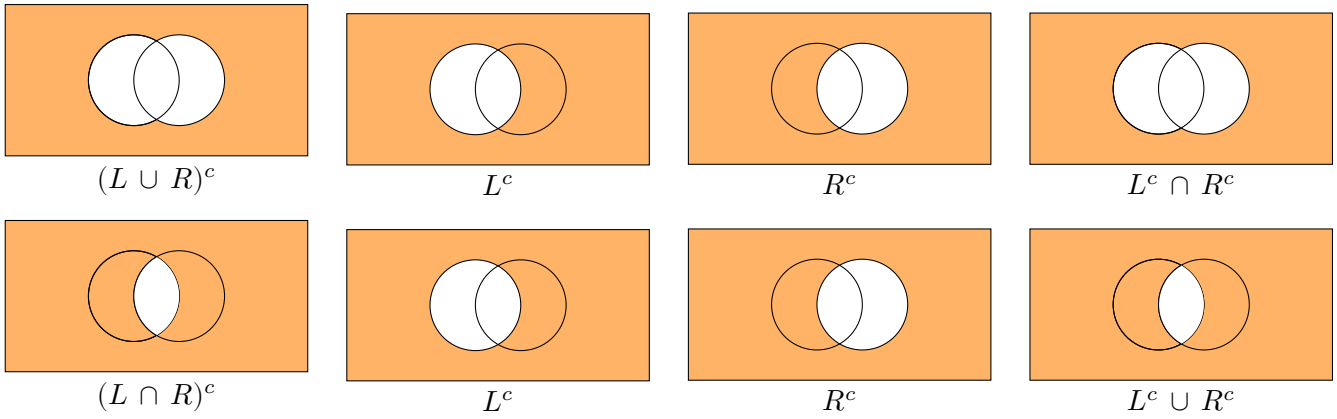
2.2.2 Venn Diagrams

Venn diagrams offer an easy way to visualize set operations.

In all the figures S is the region inside the large rectangle, L is the region inside the left circle and R is the region inside the right circle. The shaded region shows the set indicated underneath each figure.



Proof of DeMorgan's Laws



Example 3. Verify DeMorgan's laws for the subsets $A = \{1, 2, 3\}$ and $B = \{3, 4\}$ of the set $S = \{1, 2, 3, 4, 5\}$.

Solution: For each law we just work through both sides of the equation and show they are the same.

1. $(A \cup B)^c = A^c \cap B^c$:
Right hand side: $A \cup B = \{1, 2, 3, 4\} \Rightarrow (A \cup B)^c = \{5\}$.
Left hand side: $A^c = \{4, 5\}$, $B^c = \{1, 2, 5\} \Rightarrow A^c \cap B^c = \{5\}$.
The two sides are equal. QED

2. $(A \cap B)^c = A^c \cup B^c$:
Right hand side: $A \cap B = \{3\} \Rightarrow (A \cap B)^c = \{1, 2, 4, 5\}$.
Left hand side: $A^c = \{4, 5\}$, $B^c = \{1, 2, 5\} \Rightarrow A^c \cup B^c = \{1, 2, 4, 5\}$.
The two sides are equal. QED

Think: Draw and label a Venn diagram with A the set of Brain and Cognitive Science majors and B the set of sophomores. Shade the region illustrating the first law. Can you express the first law in this case as a non-technical English sentence?

2.2.3 Products of sets

The product of sets S and T is the set of ordered pairs:

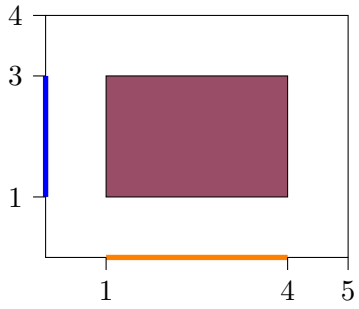
$$S \times T = \{(s, t) \mid s \in S, t \in T\}.$$

In words the right-hand side reads "the set of ordered pairs (s, t) such that s is in S and t is in T.

The following diagrams show two examples of the set product.

×	1	2	3	4
1	(1,1)	(1,2)	(1,3)	(1,4)
2	(2,1)	(2,2)	(2,3)	(2,4)
3	(3,1)	(3,2)	(3,3)	(3,4)

$\{1,2,3\} \times \{1,2,3,4\}$



$[1,4] \times [1,3] \subset [0,5] \times [0,4]$

The right-hand figure also illustrates that if $A \subset S$ and $B \subset T$ then $A \times B \subset S \times T$.

2.3 Counting

If S is finite, we use $|S|$ or $\#S$ to denote the number of elements of S.

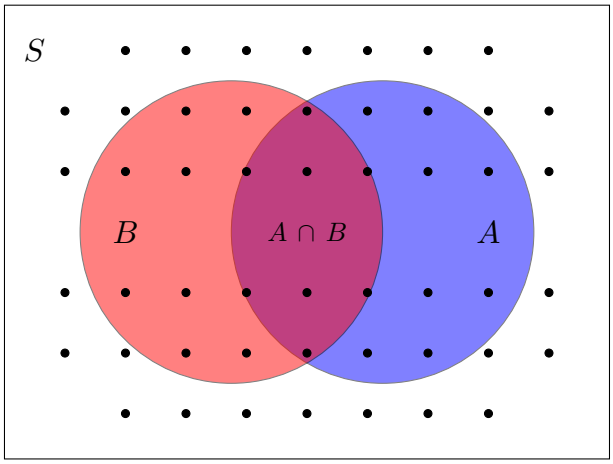
Two useful counting principles are the *inclusion-exclusion principle* and the *rule of product*.

2.3.1 Inclusion-exclusion principle

The inclusion-exclusion principle says

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

We can illustrate this with a Venn diagram. S is all the dots, A is the dots in the blue circle, and B is the dots in the red circle.



$|A|$ is the number of dots in A and likewise for the other sets. The figure shows that $|A|+|B|$ *double*-counts $|A \cap B|$, which is why $|A \cap B|$ is subtracted off in the inclusion-exclusion formula.

Example 4. In a band of singers and guitarists, seven people sing, four play the guitar, and two do both. How big is the band?

Solution: Let S be the set singers and G be the set guitar players. The inclusion-exclusion principle says

$$\text{size of band} = |S \cup G| = |S| + |G| - |S \cap G| = 7 + 4 - 2 = 9.$$

2.3.2 Rule of Product

The Rule of Product says:

> If there are n ways to perform action 1 and then by m ways to perform action 2, then there are $n \cdot m$ ways to perform action 1 followed by action 2.

We will also call this the multiplication rule.

Example 5. If you have 3 shirts and 4 pants then you can make $3 \cdot 4 = 12$ outfits.

Think: An extremely important point is that the rule of product holds even if the ways to perform action 2 depend on action 1, as long as the *number* of ways to perform action 2 is independent of action 1. To illustrate this:

Example 6. There are 5 competitors in the 100m final at the Olympics. In how many ways can the gold, silver, and bronze medals be awarded?

Solution: There are 5 ways to award the gold. Once that is awarded there are 4 ways to award the silver and then 3 ways to award the bronze: answer $5 \cdot 4 \cdot 3 = 60$ ways.

Note that the choice of gold medalist affects who can win the silver, but the number of possible silver medalists is always four.

2.4 Permutations and combinations

2.4.1 Permutations

A permutation of a set is a particular ordering of its elements. For example, the set $\{a, b, c\}$ has six permutations: $abc, acb, bac, bca, cab, cba$. We found the number of permutations by listing them all. We could also have found the number of permutations by using the rule of product. That is, there are 3 ways to pick the first element, then 2 ways for the second, and 1 for the third. This gives a total of $3 \cdot 2 \cdot 1 = 6$ permutations.

In general, the rule of product tells us that the number of permutations of a set of k elements is

$$k! = k \cdot (k-1) \cdots 3 \cdot 2 \cdot 1.$$

We also talk about the permutations of k things out of a set of n things. We show what this means with an example.

Example 7. List all the permutations of 3 elements out of the set $\{a, b, c, d\}$.

Solution: This is a longer list,

abc	acb	bac	bca	cab	cba
abd	adb	bad	bda	dab	dba
acd	adc	cad	cda	dac	dca
bcd	bdc	cbd	cdb	dbc	dcb

Note that *abc* and *acb* count as distinct permutations. That is, for permutations the order matters.

There are 24 permutations. Note that the rule of product would have told us there are $4 \cdot 3 \cdot 2 = 24$ permutations without bothering to list them all.

2.4.2 Combinations

In contrast to permutations, in combinations order does *not* matter: permutations are lists and combinations are sets. We show what we mean with an example

Example 8. List all the combinations of 3 elements out of the set $\{a, b, c, d\}$.

Solution: Such a combination is a collection of 3 elements without regard to order. So, *abc* and *cab* both represent the same combination. We can list all the combinations by listing all the subsets of exactly 3 elements.

$$\{a, b, c\} \quad \{a, b, d\} \quad \{a, c, d\} \quad \{b, c, d\}$$

There are only 4 combinations. Contrast this with the 24 permutations in the previous example. The factor of 6 comes because every combination of 3 things can be written in 6 different orders.

2.4.3 Formulas

We'll use the following notations.
${}_nP_k$ = number of permutations (lists) of k distinct elements from a set of size n
${}_nC_k = \binom{n}{k}$ = number of combinations (subsets) of k elements from a set of size n
We emphasize that by the number of combinations of k elements we mean the number of subsets of size k.

These have the following notation and formulas:

$$\text{Permutations:} \quad {}_nP_k = \frac{n!}{(n-k)!} = n(n-1)\cdots(n-k+1)$$

$$\text{Combinations:} \quad {}_nC_k = \frac{n!}{k!(n-k)!} = \frac{{}_nP_k}{k!}$$

The notation ${}_nC_k$ is read "n choose k". The formula for ${}_nP_k$ follows from the rule of product. It also implies the formula for ${}_nC_k$ because a subset of size k can be ordered in $k!$ ways.

We can illustrate the relation between permutations and combinations by lining up the results of the previous two examples.

abc	*acb*	*bac*	*bca*	*cab*	*cba*	$\{a, b, c\}$
abd	*adb*	*bad*	*bda*	*dab*	*dba*	$\{a, b, d\}$
acd	*adc*	*cad*	*cda*	*dac*	*dca*	$\{a, c, d\}$
bcd	*bdc*	*cbd*	*cdb*	*dbc*	*dcb*	$\{b, c, d\}$
	Permutations: ${}_4P_3$					Combinations: ${}_4C_3$

Notice that each row in the permutations list consists of all 3! permutations of the corresponding set in the combinations list.

2.4.4 Examples

Example 9. Count the following:

(i) The number of ways to choose 2 out of 4 things (order does not matter).
(ii) The number of ways to list 2 out of 4 things.
(iii) The number of ways to choose 3 out of 10 things.

Solution: (i) This is asking for combinations: $\binom{4}{2} = \frac{4!}{2!\,2!} = 6$.

(ii) This is asking for permuations: ${}_4P_2 = \frac{4!}{2!} = 12$.

(iii) This is asking for combinations: $\binom{10}{3} = \frac{10!}{3!\,7!} = \frac{10\cdot 9\cdot 8}{3\cdot 2\cdot 1} = 120$.

Example 10. (i) Count the number of ways to get 3 heads in a sequence of 10 flips of a coin.
(ii) If the coin is fair, what is the probability of exactly 3 heads in 10 flips?

Solution: (i) This asks for the number sequences of 10 flips (heads or tails) with exactly 3 heads. That is, we have to choose exactly 3 out of 10 flips to be heads. This is the same question as in the previous example.

$$\binom{10}{3} = \frac{10!}{3!\,7!} = \frac{10\cdot 9\cdot 8}{3\cdot 2\cdot 1} = 120.$$

(ii) Each flip has 2 possible outcomes (heads or tails). So the rule of product says there are $2^{10} = 1024$ sequences of 10 flips. Since the coin is fair each sequence is equally probable. So the probability of 3 heads is

$$\frac{120}{1024} = 0.117\,.$$

Probability: Terminology and Examples
Class 2, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the definitions of sample space, event and probability function.
2. Be able to organize a scenario with randomness into an experiment and sample space.
3. Be able to make basic computations using a probability function.

2 Terminology

2.1 Probability cast list

- Experiment: a repeatable procedure with well-defined possible outcomes.
- Sample space: the set of all possible outcomes. We usually denote the sample space by Ω, sometimes by S.
- Event: a subset of the sample space.
- Probability function: a function giving the probability for each outcome.

Later in the course we will learn about

- Probability density: a continuous distribution of probabilities.
- Random variable: a random numerical outcome.

2.2 Simple examples

Example 1. Toss a fair coin.
Experiment: toss the coin, report if it lands heads or tails.
Sample space: $\Omega = \{H, T\}$.
Probability function: $P(H) = 0.5, \quad P(T) = 0.5$.

Example 2. Toss a fair coin 3 times.
Experiment: toss the coin 3 times, list the results.
Sample space: $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.
Probability function: Each outcome is equally likely with probability 1/8.

For small sample spaces we can put the set of outcomes and probabilities into a probability table.

Outcomes	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
Probability	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Example 3. Measure the mass of a proton
Experiment: follow some defined procedure to measure the mass and report the result.
Sample space: $\Omega = [0, \infty)$, i.e. in principle we can get any positive value.
Probability function: since there is a continuum of possible outcomes there is no probability function. Instead we need to use a *probability density*, which we will learn about later in the course.

Example 4. Taxis (An infinite discrete sample space)
Experiment: count the number of taxis that pass 77 Mass. Ave during an 18.05 class.
Sample space: $\Omega = \{0, 1, 2, 3, 4, \ldots\}$.
This is often modeled with the following probability function known as the Poisson distribution. (Do not worry about mastering the Poisson distribution just yet):

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!},$$

where λ is the average number of taxis. We can put this in a table:

Outcome	0	1	2	3	...	k	...
Probability	$e^{-\lambda}$	$e^{-\lambda}\lambda$	$e^{-\lambda}\lambda^2/2$	$e^{-\lambda}\lambda^3/3!$	...	$e^{-\lambda}\lambda^k/k!$	...

Question: Accepting that this is a valid probability function, what is $\displaystyle\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}$?

Solution: This is the total probability of all possible outcomes, so the sum equals 1. (Note, this also follows from the Taylor series $e^{\lambda} = \displaystyle\sum_{n=0}^{\infty}\frac{\lambda^n}{n!}$.)

In a given setup there can be more than one reasonable choice of sample space. Here is a simple example.

Example 5. Two dice (Choice of sample space)
Suppose you roll one die. Then the sample space and probability function are

Outcome	1	2	3	4	5	6
Probability:	1/6	1/6	1/6	1/6	1/6	1/6

Now suppose you roll two dice. What should be the sample space? Here are two options.
1. Record the pair of numbers showing on the dice (first die, second die).
2. Record the sum of the numbers on the dice. In this case there are 11 outcomes $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. These outcomes are not all equally likely.

As above, we can put this information in tables. For the first case, the sample space is the product of the sample spaces for each die

$$\{(1,1),\ (2,1),\ (3,1), \ldots (6,6)\}$$

Each of the 36 outcomes is equally likely. (Why 36 outcomes?) For the probability function we will make a two dimensional table with the rows corresponding to the number on the first die, the columns the number on the second die and the entries the probability.

Die 1 \ Die 2	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

Two dice in a two dimensional table

In the second case we can present outcomes and probabilities in our usual table.

outcome	2	3	4	5	6	7	8	9	10	11	12
probability	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

The sum of two dice

Think: What is the relationship between the two probability tables above?

We will see that the best choice of sample space depends on the context. For now, simply note that given the outcome as a pair of numbers it is easy to find the sum.

Note. Listing the experiment, sample space and probability function is a good way to start working systematically with probability. It can help you avoid some of the common pitfalls in the subject.

Events.

An event is a collection of outcomes, i.e. an event is a subset of the sample space Ω. This sounds odd, but it actually corresponds to the common meaning of the word.

Example 6. Using the setup in Example **??** we would describe the event that you get exactly two heads in words by $E =$ 'exactly 2 heads'. Written as a subset this becomes

$$E = \{HHT, HTH, THH\}.$$

You should get comfortable moving between describing events in words and as subsets of the sample space.

The probability of an event E is computed by adding up the probabilities of all of the outcomes in E. In this example each outcome has probability $1/8$, so we have $P(E) = 3/8$.

2.3 Definition of a discrete sample space

Definition. A discrete sample space is one that is listable, it can be either finite or infinite.

Examples. {H, T}, {1, 2, 3}, {1, 2, 3, 4, ...}, {2, 3, 5, 7, 11, 13, 17, ...} are all discrete sets. The first two are finite and the last two are infinite.

Example. The interval $0 \le x \le 1$ is *not* discrete, rather it is *continuous*. We will deal with continuous sample spaces in a few days.

2.4 The probability function

So far we've been using a casual definition of the probability function. Let's give a more precise one.

Careful definition of the probability function.
For a discrete sample space S a probability function P assigns to each outcome ω a number $P(\omega)$ called the probability of ω. P must satisfy two rules:

- Rule 1. $0 \le P(\omega) \le 1$ (probabilities are between 0 and 1).
- Rule 2. The sum of the probabilities of all possible outcomes is 1 (something must occur)

In symbols, Rule 2 says: if $S = \{\omega_1, \omega_2, \dots, \omega_n\}$ then $P(\omega_1) + P(\omega_2) + \dots + P(\omega_n) = 1$. Or, using summation notation: $\sum_{j=1}^{n} P(\omega_j) = 1$.

The probability of an event E is the sum of the probabilities of all the outcomes in E. That is,

$$P(E) = \sum_{\omega \in E} P(\omega).$$

Think: Check Rules 1 and 2 on Examples 1 and 2 above.

Example 7. Flip until heads (A classic example)
Suppose we have a coin with probability p of heads and we have the following scenario.

Experiment: Toss the coin until the first heads. Report the number of tosses.
Sample space: $\Omega = \{1, 2, 3, \dots\}$.
Probability function: $P(n) = (1-p)^{n-1}p$.

Challenge 1: show the sum of all the probabilities equals 1 (hint: geometric series).

Challenge 2: justify the formula for $P(n)$ (we will do this soon).

Stopping problems. The previous toy example is an uncluttered version of a general class of problems called stopping rule problems. A stopping rule is a rule that tells you when to end a certain process. In the toy example above the process was flipping a coin and we stopped after the first heads. A more practical example is a rule for ending a series of medical treatments. Such a rule could depend on how well the treatments are working, how the patient is tolerating them and the probability that the treatments would continue to be effective. One could ask about the probability of stopping within a certain number of treatments or the average number of treatments you should expect before stopping.

3 Some rules of probability

For events A, L and R contained in a sample space Ω.

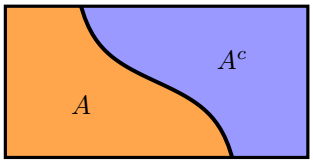
Rule 1. $P(A^c) = 1 - P(A)$.

Rule 2. If L and R are disjoint then $P(L \cup R) = P(L) + P(R)$.

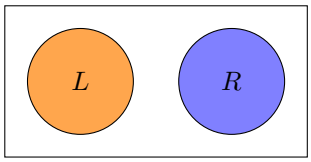
Rule 3. If L and R are not disjoint, we have the **inclusion-exclusion principle**:

$$P(L \cup R) = P(L) + P(R) - P(L \cap R)$$

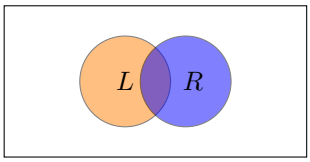
We visualize these rules using Venn diagrams.



$\Omega = A \cup A^c$, no overlap



$L \cup R$, no overlap



$L \cup R$, overlap $= L \cap R$

We can also justify them logically.

Rule 1: A and A^c split Ω into two non-overlapping regions. Since the total probability $P(\Omega) = 1$ this rule says that the probabiity of A and the probability of 'not A' are complementary, i.e. sum to 1.

Rule 2: L and R split $L \cup R$ into two non-overlapping regions. So the probability of $L \cup R$ is is split between $P(L)$ and $P(R)$

Rule 3: In the sum $P(L) + P(R)$ the overlap $P(L \cap R)$ gets counted twice. So $P(L) + P(R) - P(L \cap R)$ counts everything in the union exactly once.

Think: Rule 2 is a special case of Rule 3.

For the following examples suppose we have an experiment that produces a random integer between 1 and 20. The probabilities are not necessarily uniform, i.e., not necessarily the same for each outcome.

Example 8. If the probability of an even number is 0.6 what is the probability of an odd number?

Solution: Since being odd is complementary to being even, the probability of being odd is 1-0.6 = 0.4.

Let's redo this example a bit more formally, so you see how it's done. First, so we can refer to it, let's name the random integer X. Let's also name the event 'X is even' as A. Then the event 'X is odd' is A^c. We are given that $P(A) = 0.6$. Therefore $P(A^c) = 1 - 0.6 = \boxed{0.4}$.

Example 9. Consider the 2 events, A: 'X is a multiple of 2'; B: 'X is odd and less than 10'. Suppose $P(A) = 0.6$ and $P(B) = 0.25$.
(i) What is $A \cap B$?
(ii) What is the probability of $A \cup B$?

Solution: (i) Since all numbers in A are even and all numbers in B are odd, these events are disjoint. That is, $\boxed{A \cap B = \emptyset.}$

(ii) Since A and B are disjoint $\boxed{P(A \cup B) = P(A) + P(B) = 0.85.}$

Example 10. Let A, B and C be the events X is a multiple of 2, 3 and 6 respectively. If $P(A) = 0.6$, $P(B) = 0.3$ and $P(C) = 0.2$ what is $P(A \text{ or } B)$?

Solution: Note two things. First we used the word 'or' which means union: 'A or B' = $A \cup B$. Second, an integer is divisible by 6 if and only if it is divisible by both 2 and 3.

This translates into $C = A \cap B$. So the inclusion-exclusion principle says

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + 0.3 - 0.2 = \boxed{0.7}.$$

Conditional Probability, Independence and Bayes' Theorem
Class 3, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the definitions of conditional probability and independence of events.
2. Be able to compute conditional probability directly from the definition.
3. Be able to use the multiplication rule to compute the total probability of an event.
4. Be able to check if two events are independent.
5. Be able to use Bayes' formula to 'invert' conditional probabilities.
6. Be able to organize the computation of conditional probabilities using trees and tables.
7. Understand the base rate fallacy thoroughly.

2 Conditional Probability

Conditional probability answers the question 'how does the probability of an event change if we have extra information'. We'll illustrate with an example.

Example 1. Toss a fair coin 3 times.
(a) What is the probability of 3 heads?

Solution: Sample space $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.
All outcomes are equally probable, so $P(3 \text{ heads}) = 1/8$.

(b) Suppose we are told that the first toss was heads. Given this information how should we compute the probability of 3 heads?

Solution: We have a new (reduced) sample space: $\Omega' = \{HHH, HHT, HTH, HTT\}$.
All outcomes are equally probable, so

$$P(3 \text{ heads given that the first toss is heads}) = 1/4.$$

This is called conditional probability, since it takes into account additional conditions. To develop the notation, we rephrase (b) in terms of *events*.

Rephrased (b) Let A be the event 'all three tosses are heads' $= \{HHH\}$.
Let B be the event 'the first toss is heads' $= \{HHH, HHT, HTH, HTT\}$.

The conditional probability of A knowing that B occurred is written

$$P(A|B)$$

This is read as

'the conditional probability of A given B'

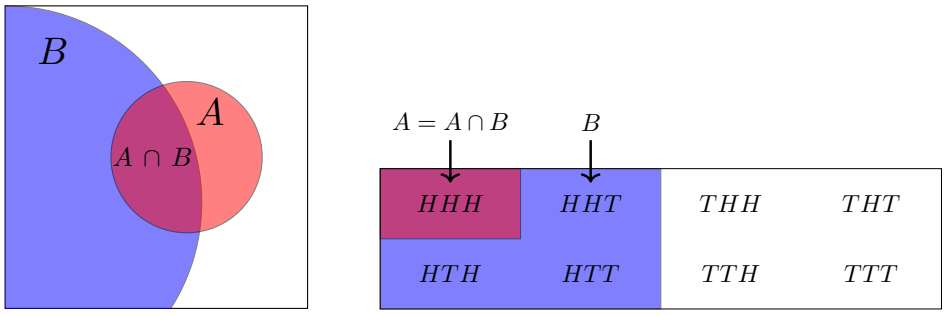
or

'the probability of A conditioned on B'

or simply

'the probability of A given B'.

We can visualize conditional probability as follows. Think of $P(A)$ as the proportion of the area of the *whole* sample space taken up by A. For $P(A|B)$ we restrict our attention to B. That is, $P(A|B)$ is the proportion of area of B taken up by A, i.e. $P(A \cap B)/P(B)$.



Conditional probability: Abstract visualization and coin example

Note, $A \subset B$ in the right-hand figure, so there are only two colors shown.

The formal definition of conditional probability catches the gist of the above example and visualization.

Formal definition of conditional probability
Let A and B be events. We define the conditional probability of A given B as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ provided } P(B) \neq 0. \tag{1}$$

Let's redo the coin tossing example using the definition in Equation (1). Recall $A =$ '3 heads' and $B =$ 'first toss is heads'. We have $P(A) = 1/8$ and $P(B) = 1/2$. Since $A \cap B = A$, we also have $P(A \cap B) = 1/8$. Now according to (1),

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/8}{1/2} = 1/4,$$

which agrees with our answer in Example 1 b.

We start with a simple example where we can find all the probabilities directly by counting.

Example 2. Draw two cards from a deck. Define the events: $S_1 =$ 'first card is a spade' and $S_2 =$ 'second card is a spade'. What is the $P(S_2|S_1)$?

Solution: We'll use formula (1) to compute the conditional probability. We have to compute $P(S_1)$, $P(S_2)$ and $P(S_1 \cap S_2)$: We know that $P(S_1) = 1/4$ because there are 52 equally probable ways to draw the first card and 13 of them are spades. The same logic says that there are 52 equally probable ways the second card can be drawn, so $P(S_2) = 1/4$.

Aside: The probability $P(S_2) = 1/4$ may seem surprising since the value of first card certainly affects the probabilities for the second card. However, if we look at *all* possible two card sequences we will see that every card in the deck has equal probability of being

the second card. Since 13 of the 52 cards are spades we get $P(S_2) = 13/52 = 1/4$. Another way to say this is: if we are not given value of the first card then we have to consider all possibilities for the second card.

Continuing, we compute $P(S_1 \cap S_2)$ by counting:

Number of ways to draw a spade followed by a second spade: $13 \cdot 12$.
Number of ways to draw any card followed by any other card: $52 \cdot 51$.

Thus,

$$P(S_1 \cap S_2) = \frac{13 \cdot 12}{52 \cdot 51} = 3/51.$$

Now, using (1) we get

$$P(S_2|S_1) = \frac{P(S_2 \cap S_1)}{P(S_1)} = \frac{3/51}{1/4} = 12/51.$$

This case is simple enough that we can check our answer by computing the conditional probability directly: if the first card is a spade then of the 51 cards remaining, 12 are spades. So, the probability the second card is also a spade is

$$P(S_2|S_1) = 12/51.$$

Warning: In more complicated problems it will be be much harder to compute conditional probability by counting. Usually we have to use Equation (1).

Think: For S_1 and S_2 in the previous example, what is $P(S_2|S_1^c)$?

3 Multiplication Rule

The following formula is called the multiplication rule.

$$P(A \cap B) = P(A|B) \cdot P(B). \tag{2}$$

This is simply a rewriting of the definition in Equation (1) of conditional probability. We will see that our use of the multiplication rule is very similar to our use of the rule of product in counting. In fact, the multiplication rule is just a souped up version of the rule of product.

We start by verifying the multiplication rule for the previous example.

Example 3. Draw two cards from a deck. Define the events: S_1 = 'first card is a spade' and S_2 = 'second card is a spade'. Verify the multiplication rule.

Solution: From the previous example, we know $P(S_2|S_1) = 12/51$, $P(S_1 \cap S_2) = 3/51$, $P(S_1) = 1/4$. From this it is easy to check that

$$P(S_2|S_1) \cdot P(S_1) = \frac{12}{51} \cdot \frac{1}{4} = \frac{3}{51} = P(S_1 \cap S_2).$$

4 Law of Total Probability

The law of total probability will allow us to use the multiplication rule to find probabilities in more interesting examples. It involves a lot of notation, but the idea is fairly simple. We

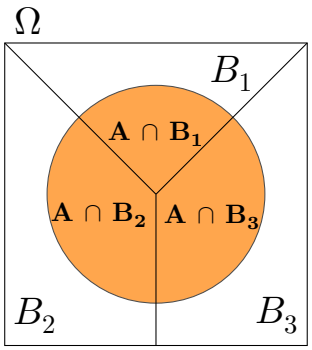
state the law when the sample space is divided into 3 pieces. It is a simple matter to extend the rule when there are more than 3 pieces.

Law of Total Probability

Suppose the sample space Ω is divided into 3 disjoint events B_1, B_2, B_3 (see the figure below). Then for any event A:

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \\ P(A) &= P(A|B_1)\,P(B_1) + P(A|B_2)\,P(B_2) + P(A|B_3)\,P(B_3) \end{aligned} \tag{3}$$

The top equation says 'if A is divided into 3 pieces then $P(A)$ is the sum of the probabilities of the pieces'. The bottom equation (3) is called the law of total probability. It is just a rewriting of the top equation using the multiplication rule.



The sample space Ω and the event A are each divided into 3 disjoint pieces.

The law holds if we divide Ω into any number of events, so long as they are *disjoint* and *cover* all of Ω. Such a division is often called a *partition* of Ω.

Our first example will be one where we already know the answer and can verify the law.

Example 4. An urn contains 5 red balls and 2 green balls. Two balls are drawn one after the other. What is the probability that the second ball is red?

Solution: The sample space is $\Omega = \{\text{rr, rg, gr, gg}\}$.

Let R_1 be the event 'the first ball is red', G_1 = 'first ball is green', R_2 = 'second ball is red', G_2 = 'second ball is green'. We are asked to find $P(R_2)$.

Let's compute this same value using the law of total probability (3). First, we'll find the conditional probabilities. This is a simple counting exercise.

$$P(R_2|R_1) = 4/6, \qquad P(R_2|G_1) = 5/6.$$

Since R_1 and G_1 partition Ω the law of total probability says

$$\begin{aligned} P(R_2) &= P(R_2|R_1)P(R_1) + P(R_2|G_1)P(G_1) \\ &= \frac{4}{6}\cdot\frac{5}{7} + \frac{5}{6}\cdot\frac{2}{7} \\ &= \frac{30}{42} = \frac{5}{7}. \end{aligned} \tag{4}$$

Of course, this example is simple enough that we could have computed $P(R_2)$ directly the same way we found $P(S_2)$ directly in the card example. But, we will see that in more complicated examples the law of total probability is truly necessary.

Probability urns

The example above used probability urns. Their use goes back to the beginning of the subject and we would be remiss not to introduce them. This toy model is very useful. We quote from Wikipedia: https://en.wikipedia.org/wiki/Urn_problem

> In probability and statistics, an urn problem is an idealized mental exercise in which some objects of real interest (such as atoms, people, cars, etc.) are represented as colored balls in an urn or other container. One pretends to draw (remove) one or more balls from the urn; the goal is to determine the probability of drawing one color or another, or some other properties. A key parameter is whether each ball is returned to the urn after each draw.

It doesn't take much to make an example where (3) is really the best way to compute the probability. Here is a game with slightly more complicated rules.

Example 5. An urn contains 5 red balls and 2 green balls. A ball is drawn. If it's green a red ball is added to the urn and if it's red a green ball is added to the urn. (The original ball is not returned to the urn.) Then a second ball is drawn. What is the probability the second ball is red?

Solution: The law of total probability says that $P(R_2)$ can be computed using the expression in Equation (4). Only the values for the probabilities will change. We have

$$P(R_2|R_1) = 4/7, \qquad P(R_2|G_1) = 6/7.$$

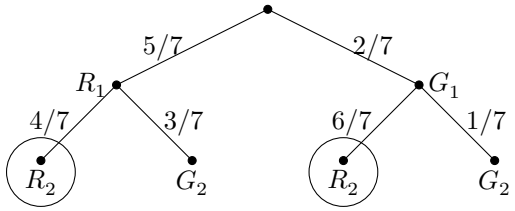
Therefore,

$$P(R_2) = P(R_2|R_1)P(R_1) + P(R_2|G_1)P(G_1) = \frac{4}{7}\cdot\frac{5}{7} + \frac{6}{7}\cdot\frac{2}{7} = \frac{32}{49}.$$

5 Using Trees to Organize the Computation

Trees are a great way to organize computations with conditional probability and the law of total probability. The figures and examples will make clear what we mean by a tree. As with the rule of product, the key is to organize the underlying process into a sequence of actions.

We start by redoing Example 5. The sequence of actions are: first draw ball 1 (and add the appropriate ball to the urn) and then draw ball 2.



You interpret this tree as follows. Each dot is called a node. The tree is organized by levels. The top node (root node) is at level 0. The next layer down is level 1 and so on. Each level shows the outcomes at one stage of the game. Level 1 shows the possible outcomes of the first draw. Level 2 shows the possible outcomes of the second draw starting from each node in level 1.

Probabilities are written along the branches. The probability of R_1 (red on the first draw) is 5/7. It is written along the branch from the root node to the one labeled R_1. At the next level we put in conditional probabilities. The probability along the branch from R_1 to R_2 is $P(R_2|R_1) = 4/7$. It represents the probability of going to node R_2 given that you are already at R_1.

The muliplication rule says that the probability of getting to any node is just the product of the probabilities along the path to get there. For example, the node labeled R_2 at the far left really represents the event $R_1 \cap R_2$ because it comes from the R_1 node. The multiplication rule now says

$$P(R_1 \cap R_2) = P(R_1) \cdot P(R_2|R_1) = \frac{5}{7} \cdot \frac{4}{7},$$

which is exactly multiplying along the path to the node.

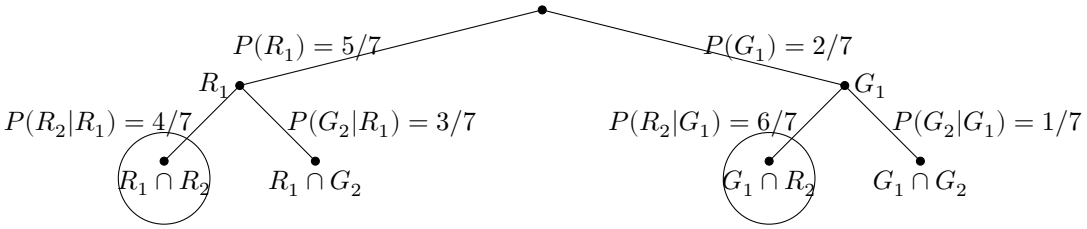
The law of total probability is just the statement that $P(R_2)$ is the sum of the probabilities of all paths leading to R_2 (the two circled nodes in the figure). In this case,

$$P(R_2) = \frac{5}{7} \cdot \frac{4}{7} + \frac{2}{7} \cdot \frac{6}{7} = \frac{32}{49},$$

exactly as in the previous example.

5.1 Shorthand vs. precise trees

The tree given above involves some shorthand. For example, the node marked R_2 at the far left really represents the event $R_1 \cap R_2$, since it ends the path from the root through R_1 to R_2. Here is the same tree with everything labeled precisely. As you can see this tree is more cumbersome to make and use. We usually use the shorthand version of trees. You should make sure you know how to interpret them precisely.



6 Independence

Two events are independent if knowledge that one occurred does not change the probability that the other occurred. Informally, events are independent if they do not influence one another.

Example 6. Toss a coin twice. We expect the outcomes of the two tosses to be independent of one another. In real experiments this always has to be checked. If my coin lands in honey and I don't bother to clean it, then the second toss might be affected by the outcome of the first toss.

More seriously, the independence of experiments can by undermined by the failure to clean or recalibrate equipment between experiments or to isolate supposedly independent observers from each other or a common influence. We've all experienced hearing the same 'fact' from different people. Hearing it from different sources tends to lend it credence until we learn that they all heard it from a common source. That is, our sources were not independent.

Translating the verbal description of independence into symbols gives

$$A \text{ is independent of } B \quad \text{if} \quad P(A|B) = P(A). \tag{5}$$

That is, knowing that B occurred does not change the probability that A occurred. In terms of events as subsets, knowing that the realized outcome is in B does not change the probability that it is in A.

If A and B are independent in the above sense, then the multiplication rule gives

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

This justifies the following technical definition of independence.

Formal definition of independence: Two events A and B are independent if

$$P(A \cap B) \;=\; P(A) \cdot P(B) \tag{6}$$

This is a nice symmetric definition which makes clear that A is independent of B if and only if B is independent of A. Unlike the equation with conditional probabilities, this definition makes sense even when $P(B) = 0$. In terms of conditional probabilities, we have:

1. If $P(B) \neq 0$ then A and B are independent if and only if $P(A|B) = P(A)$.
2. If $P(A) \neq 0$ then A and B are independent if and only if $P(B|A) = P(B)$.

Independent events commonly arise as different trials in an experiment, as in the following example.

Example 7. Toss a fair coin twice. Let H_1 = 'heads on first toss' and let H_2 = 'heads on second toss'. Are H_1 and H_2 independent?

Solution: Since $H_1 \cap H_2$ is the event 'both tosses are heads' we have

$$P(H_1 \cap H_2) = 1/4 = P(H_1)P(H_2).$$

Therefore the events are independent.

We can ask about the independence of any two events, as in the following two examples.

Example 8. Toss a fair coin 3 times. Let H_1 = 'heads on first toss' and A = 'two heads total'. Are H_1 and A independent?

Solution: We know that $P(A) = 3/8$. Since this is not 0 we can check if the formula in Equation 5 holds. Now, H_1 = {HHH, HHT, HTH, HTT} contains exactly two outcomes (HHT, HTH) from A, so we have $P(A|H_1) = 2/4$. Since $P(A|H_1) \neq P(A)$ these events are not independent.

Example 9. Draw one card from a standard deck of playing cards. Let's examine the independence of 3 events 'the card is an ace', 'the card is a heart' and 'the card is red'.

Define the events as A = 'ace', H = 'hearts', R = 'red'.

(a) We know that $P(A) = 4/52$ (4 out of 52 cards are aces), $P(A|H) = 1/13$ (1 out of 13 hearts are aces). Since $P(A) = P(A|H)$ we have that A is independent of H.

(b) $P(A|R) = 2/26 = 1/13 = P(A)$. So A is independent of R. That is, whether the card is an ace is independent of whether it is red.

(c) Finally, what about H and R? Since $P(H) = 1/4$ and $P(H|R) = 1/2$, H and R are not independent. We could also see this the other way around: $P(R) = 1/2$ and $P(R|H) = 1$, so H and R are not independent. That is, the suit of a card is not independent of the color of the card's suit.

6.1 Paradoxes of Independence

An event A with probability 0 is independent of itself, since in this case both sides of equation (6) are 0. This appears paradoxical because knowledge that A occurred certainly gives information about whether A occurred. We resolve the paradox by noting that since $P(A) = 0$ the statement 'A occurred' is vacuous.

Think: For what other value(s) of $P(A)$ is A independent of itself?

7 Bayes' Theorem

Bayes' theorem is a pillar of both probability and statistics and it is central to the rest of this course. For two events A and B Bayes' theorem (also called Bayes' rule and Bayes' formula) says

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}. \tag{7}$$

Comments: 1. Bayes' rule tells us how to 'invert' conditional probabilities, i.e. to find $P(B|A)$ from $P(A|B)$.
2. In practice, $P(A)$ is often computed using the law of total probability.

Proof of Bayes' rule
The key point is that $A \cap B$ is symmetric in A and B. So the multiplication rule says

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

Now divide through by $P(A)$ to get Bayes' rule.

A common mistake is to confuse the meanings of $P(A|B)$ and $P(B|A)$. They can be very different. This is illustrated in the next example.

Example 10. Toss a coin 5 times. Let H_1 = 'first toss is heads' and let H_A = 'all 5 tosses are heads'. Then $P(H_1|H_A) = 1$ but $P(H_A|H_1) = 1/16$.

For practice, let's use Bayes' theorem to compute $P(H_1|H_A)$ using $P(H_A|H_1)$.The terms

are $P(H_A|H_1) = 1/16$, $P(H_1) = 1/2$, $P(H_A) = 1/32$. So,

$$P(H_1|H_A) = \frac{P(H_A|H_1)P(H_1)}{P(H_A)} = \frac{(1/16)\cdot(1/2)}{1/32} = 1,$$

which agrees with our previous calculation.

7.1 The Base Rate Fallacy

The base rate fallacy is one of many examples showing that it's easy to confuse the meaning of $P(B|A)$ and $P(A|B)$ when a situation is described in words. This is one of the key examples from probability and it will inform much of our practice and interpretation of statistics. You should strive to understand it thoroughly.

Example 11. The Base Rate Fallacy
Consider a routine screening test for a disease. Suppose the frequency of the disease in the population (base rate) is 0.5%. The test is fairly accurate with a 5% false positive rate and a 10% false negative rate.

You take the test and it comes back positive. What is the probability that you have the disease?

Solution: We will do the computation three times: using trees, tables and symbols. We'll use the following notation for the relevant events:
D^+ = 'you have the disease'
D^- = 'you do not have the disease
T^+ = 'you tested positive'
T^- = 'you tested negative'.

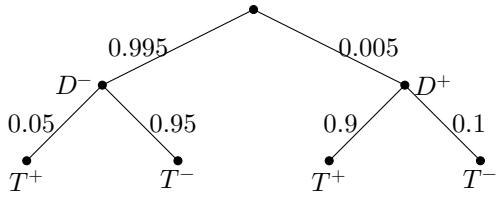
We are given $P(D+) = 0.005$ and therefore $P(D-) = 0.995$. The false positive and false negative rates are (by definition) conditional probabilities.

$$P(\text{false positive}) = P(T^+|D^-) = 0.05 \quad \text{and} \quad P(\text{false negative}) = P(T^-|D^+) = 0.1.$$

The complementary probabilities are known as the true negative and true positive rates:

$$P(T^-|D^-) = 1 - P(T^+|D^-) = 0.95 \quad \text{and} \quad P(T^+|D^+) = 1 - P(T^-|D^+) = 0.9.$$

Trees: All of these probabilities can be displayed quite nicely in a tree.



The question asks for the probability that you have the disease given that you tested positive, i.e. what is the value of $P(D^+|T^+)$. We aren't given this value, but we do know $P(T^+|D^+)$, so we can use Bayes' theorem.

$$P(D^+|T^+) = \frac{P(T^+|D^+)\cdot P(D^+)}{P(T^+)}.$$

The two probabilities in the numerator are given. We compute the denominator $P(T^+)$ using the law of total probability. Using the tree, we just have to sum the probabilities for each of the nodes marked T^+

$$P(T^+) = 0.995 \times 0.05 + 0.005 \times 0.9 = 0.05425$$

Thus,

$$P(D^+|T^+) = \frac{0.9 \times 0.005}{0.05425} = 0.082949 \approx 8.3\%.$$

Remarks: This is called the base rate fallacy because the base rate of the disease in the population is so low that the vast majority of the people taking the test are healthy, and even with an accurate test most of the positives will be healthy people. Ask your doctor for his/her guess at the odds.

To summarize the base rate fallacy with specific numbers

95% of all tests are accurate does not imply 95% of positive tests are accurate

We will refer back to this example frequently. It and similar examples are at the heart of many statistical misunderstandings.

Other ways to work Example 11

Tables: Another trick that is useful for computing probabilities is to make a table. Let's redo the previous example using a table built with 10000 total people divided according to the probabilites in this example.

We construct the table as follows. Pick a number, say 10000 people, and place it as the grand total in the lower right. Using $P(D^+) = 0.005$ we compute that 50 out of the 10000 people are sick (D^+). Likewise 9950 people are healthy (D^-). At this point the table looks like:

	$D+$	$D-$	total
T^+			
T^-			
total	50	9950	10000

Using $P(T^+|D^+) = 0.9$ we can compute that the number of sick people who tested positive as 90% of 50 or 45. The other entries are similar. At this point the table looks like the table below on the left. Finally we sum the T^+ and T^- rows to get the completed table on the right.

	D^+	D^-	total
T^+	45	498	
T^-	5	9452	
total	50	9950	10000

	D^+	D^-	total
T^+	45	498	543
T^-	5	9452	9457
total	50	9950	10000

Using the complete table we can compute

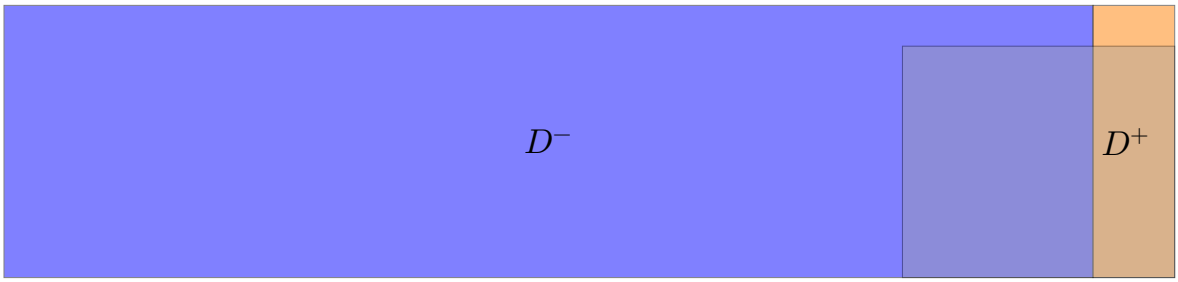
$$P(D^+|T^+) = \frac{|D^+ \cap T^+|}{|T^+|} = \frac{45}{543} = 8.3\%.$$

Symbols: For completeness, we show how the solution looks when written out directly in

symbols.

$$\begin{aligned} P(D^+|T^+) &= \frac{P(T^+|D^+)\cdot P(D^+)}{P(T^+)} \\ &= \frac{P(T^+|D^+)\cdot P(D^+)}{P(T^+|D^+)\cdot P(D^+) + P(T^+|D^-)\cdot P(D^-)} \\ &= \frac{0.9\times 0.005}{0.9\times 0.005 + 0.05\times 0.995} \\ &= 8.3\% \end{aligned}$$

Visualization: The figure below illustrates the base rate fallacy. The large blue rectangle represents all the healthy people. The much smaller orange rectangle represents the sick people. The shaded rectangle represents the people who test positive. The shaded area covers most of the orange area and only a small part of the blue area. Even so, the most of the shaded area is over the blue. That is, most of the positive tests are of healthy people.



7.2 Bayes' rule in 18.05

As we said at the start of this section, Bayes' rule is a pillar of probability and statistics. We have seen that Bayes' rule allows us to 'invert' conditional probabilities. When we study statistics we will see that the art of statistical inference involves deciding how to proceed when one (or more) of the terms on the right side of Bayes' rule is unknown.

Discrete Random Variables
Class 4, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the definition of a discrete random variable.
2. Know the Bernoulli, binomial, and geometric distributions and examples of what they model.
3. Be able to describe the probability mass function and cumulative distribution function using tables and formulas.
4. Be able to construct new random variables from old ones.

2 Random Variables

This topic is largely about introducing some useful terminology, building on the notions of sample space and probability function. The key words are

1. Random variable
2. Probability mass function (pmf)
3. Cumulative distribution function (cdf)

2.1 Recap

A discrete sample space Ω is a finite or listable set of outcomes $\{\omega_1, \omega_2 \ldots\}$. The probability of an outcome ω is denoted $P(\omega)$.

An event E is a subset of Ω. The probability of an event E is $P(E) = \sum_{\omega \in E} P(\omega)$.

2.2 Random variables as payoff functions

Example 1. A game with 2 dice.

Roll a die twice and record the outcomes as (i, j), where i is the result of the first roll and j the result of the second. We can take the sample space to be

$$\Omega = \{(1,1), (1,2), (1,3), \ldots, (6,6)\} = \{(i,j) \,|\, i, j = 1, \ldots 6\}.$$

The probability function is $P(i, j) = 1/36$.

In this game, you win \$500 if the sum is 7 and lose \$100 otherwise. We give this payoff function the name X and describe it formally by

$$X(i,j) = \begin{cases} 500 & \text{if } i+j=7 \\ -100 & \text{if } i+j \neq 7. \end{cases}$$

Example 2. We can change the game by using a different payoff function. For example

$$Y(i,j) = ij - 10.$$

In this example if you roll $(6,2)$ then you win \$2. If you roll $(2,3)$ then you win -\$4 (i.e., lose \$4).

Question: Which game is the better bet?

Solution: We will come back to this once we learn about expectation.

These payoff functions are examples of random variables. A random variable assigns a number to each outcome in a sample space. More formally:

Definition: Let Ω be a sample space. A discrete random variable is a function

$$X : \Omega \to \mathbf{R}$$

that takes a discrete set of values. (Recall that $\mathbf{R}$ stands for the real numbers.)

Why is X called a random variable? It's 'random' because its value depends on a random outcome of an experiment. And we treat X like we would a usual variable: we can add it to other random variables, square it, and so on.

2.3 Events and random variables

For any value a we write $X = a$ to mean the event consisting of all outcomes ω with $X(\omega) = a$.

Example 3. In Example 1 we rolled two dice and X was the random variable

$$X(i,j) = \begin{cases} 500 & \text{if } i+j=7 \\ -100 & \text{if } i+j \neq 7. \end{cases}$$

The event $X = 500$ is the set {(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)}, i.e. the set of all outcomes that sum to 7. So $P(X = 500) = 1/6$.

We allow a to be any value, even values that X never takes. In Example 1, we could look at the event $X = 1000$. Since X never equals 1000 this is just the empty event (or empty set)

$$\text{`}X = 1000\text{'} \;=\; \{\} \;= \emptyset \qquad P(X = 1000) = 0.$$

2.4 Probability mass function and cumulative distribution function

It gets tiring and hard to read and write $P(X = a)$ for the probability that $X = a$. When we know we're talking about X we will simply write $p(a)$. If we want to make X explicit we will write $p_X(a)$. We spell this out in a definition.

Definition: The probability mass function (pmf) of a discrete random variable is the function $p(a) = P(X = a)$.

Note:
1. We always have $0 \le p(a) \le 1$.
2. We allow a to be any number. If a is a value that X never takes, then $p(a) = 0$.

Example 4. Let Ω be our earlier sample space for rolling 2 dice. Define the random variable M to be the maximum value of the two dice, i.e.

$$M(i, j) = \max(i, j).$$

For example, the roll (3,5) has maximum 5, i.e. $M(3, 5) = 5$.

We can describe a random variable by listing its possible values and the probabilities associated to these values. For the above example we have:

value a:	1	2	3	4	5	6
pmf $p(a)$:	1/36	3/36	5/36	7/36	9/36	11/36

For example, $p(2) = 3/36$.

Question: What is $p(8)$? **Solution:** $p(8) = 0$.

Think: What is the pmf for $Z(i, j) = i + j$? Does it look familiar?

2.5 Events and inequalities

Inequalities with random variables describe events. For example $X \le a$ is the set of all outcomes ω such that $X(w) \le a$.

Example 5. If our sample space is the set of all pairs of (i, j) coming from rolling two dice and $Z(i, j) = i + j$ is the sum of the dice then

$$Z \le 4 = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1)\}$$

2.6 The cumulative distribution function (cdf)

Definition: The cumulative distribution function (cdf) of a random variable X is the function F given by $F(a) = P(X \le a)$. We will often shorten this to distribution function.

Note well that the definition of $F(a)$ uses the symbol less than **or equal to**. This will be important for getting your calculations exactly right.

Example. Continuing with the example M, we have

value a:	1	2	3	4	5	6
pmf $p(a)$:	1/36	3/36	5/36	7/36	9/36	11/36
cdf $F(a)$:	1/36	4/36	9/36	16/36	25/36	36/36

$F(a)$ is called the cumulative distribution function because $F(a)$ gives the total probability that accumulates by adding up the probabilities $p(b)$ as b runs from $-\infty$ to a. For example, in the table above, the entry 16/36 in column 4 for the cdf is the sum of the values of the pmf from column 1 to column 4. In notation:

As events: '$M \le 4$' $= \{1, 2, 3, 4\}$; $\quad F(4) = P(M \le 4) = 1/36+3/36+5/36+7/36 = 16/36$.

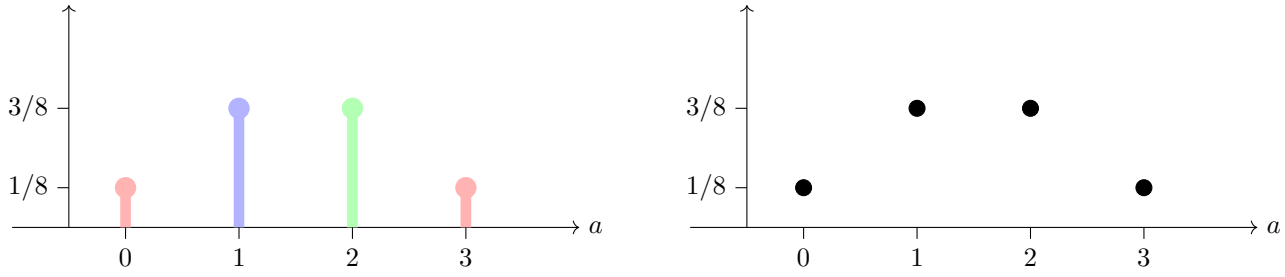
Just like the probability mass function, $F(a)$ is defined for all values a. In the above example, $F(8) = 1$, $F(-2) = 0$, $F(2.5) = 4/36$, and $F(\pi) = 9/36$.

2.7 Graphs of $p(a)$ and $F(a)$

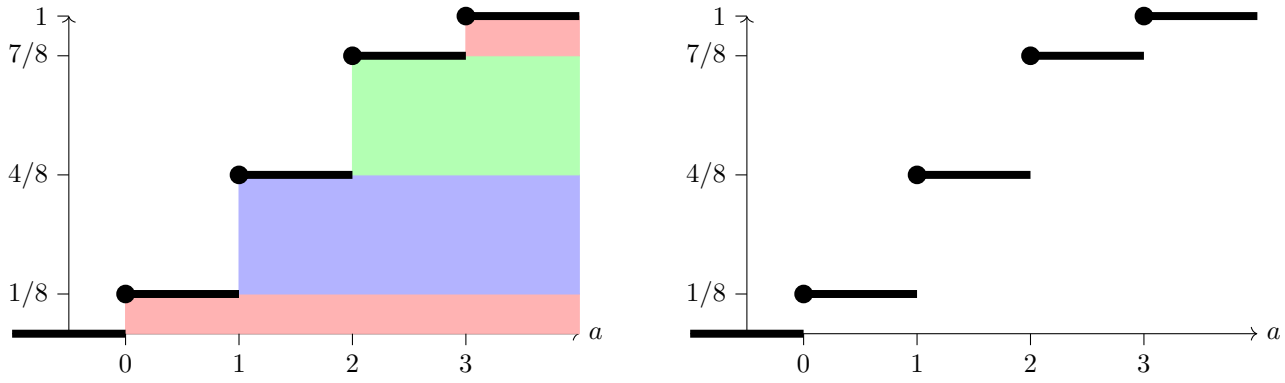
We can visualize the pmf and cdf with graphs. For example, let X be the number of heads in 3 tosses of a fair coin:

value a:	0	1	2	3
pmf $p(a)$:	1/8	3/8	3/8	1/8
cdf $F(a)$:	1/8	4/8	7/8	1

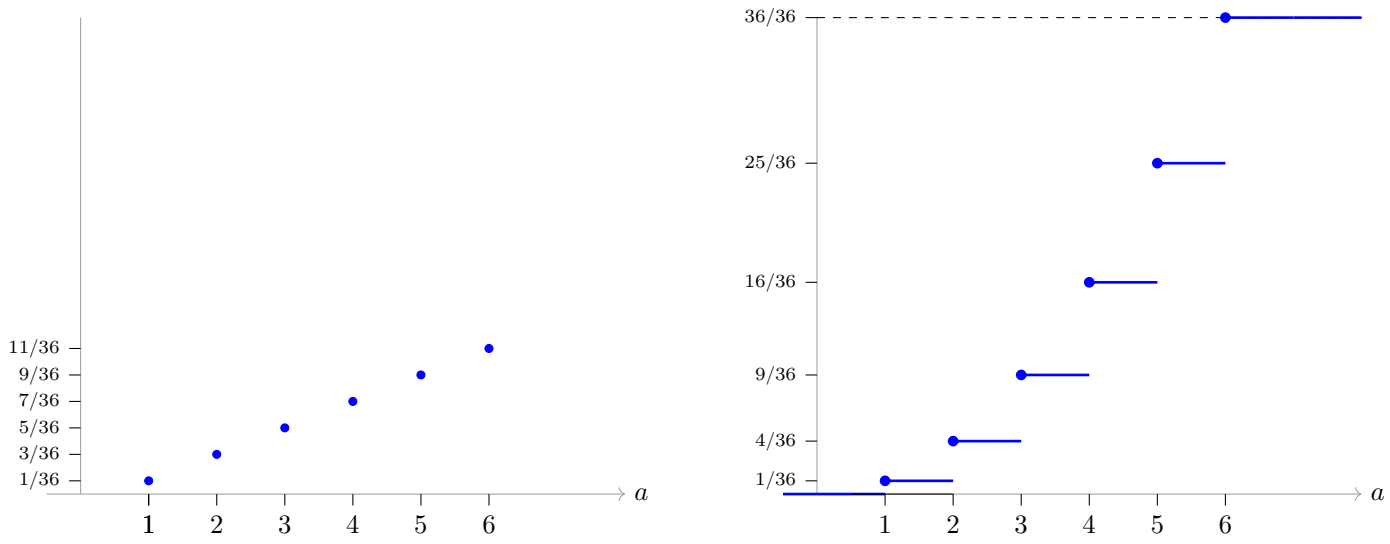
The colored graphs show how the cumulative distribution function is built by accumulating probability as a increases. The black and white graphs are the more standard presentations.



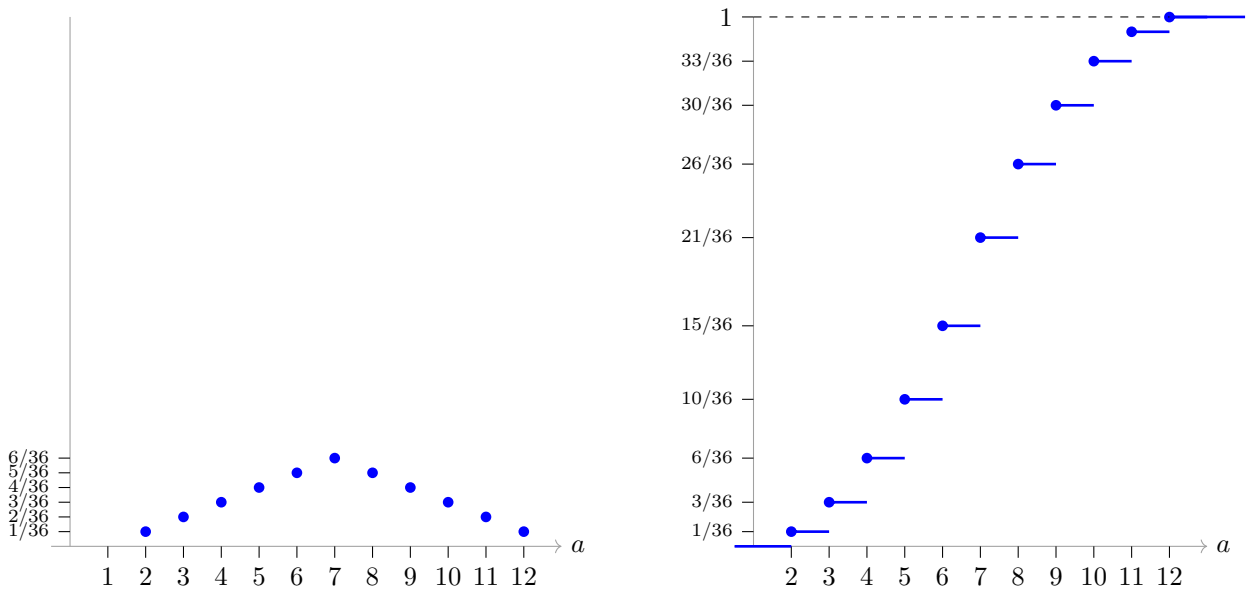
Probability mass function for X



Cumulative distribution function for X



pmf and cdf for the maximum of two dice (Example 4)



pmf and cdf for the sum of two dice

Histograms: Later we will see another way to visualize the pmf using histograms. These require some care to do right, so we will wait until we need them.

2.8 Properties of the cdf F

The cdf F of a random variable satisfies several properties:

1. F is non-decreasing. That is, its graph never goes down, or symbolically if $a \le b$ then $F(a) \le F(b)$.

2. $0 \le F(a) \le 1$.

3. $\lim_{a\to\infty} F(a) = 1, \quad \lim_{a\to-\infty} F(a) = 0$.

In words, (1) says the cumulative probability $F(a)$ increases or remains constant as a increases, but never decreases; (2) says the accumulated probability is always between 0 and 1; (3) says that as a gets very large, it becomes more and more certain that $X \le a$ and as a gets very negative it becomes more and more certain that $X > a$.

Think: Why does a cdf satisfy each of these properties?

3 Specific Distributions

3.1 Bernoulli Distributions

Model: The Bernoulli distribution models one trial in an experiment that can result in either success or failure This is the most important distribution and is also the simplest. A random variable X has a Bernoulli distribution with parameter p if:

1. X takes the values 0 and 1.

2. $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

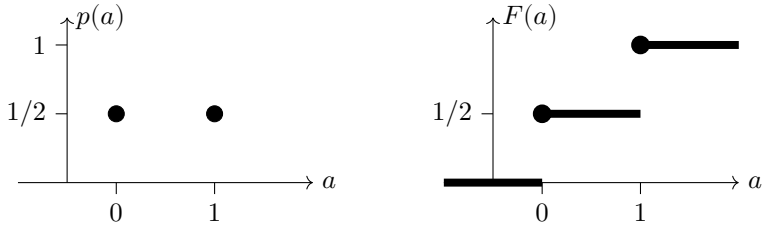
We will write $X \sim$ Bernoulli(p) or Ber(p), which is read "X follows a Bernoulli distribution with parameter p" or "X is drawn from a Bernoulli distribution with parameter p".

A simple model for the Bernoulli distribution is to flip a coin with probability p of heads, with $X = 1$ on heads and $X = 0$ on tails. The general terminology is to say X is 1 on success and 0 on failure, with success and failure defined by the context.

Many decisions can be modeled as a binary choice, such as votes for or against a proposal. If p is the proportion of the voting population that favors the proposal, than the vote of a random individual is modeled by a Bernoulli(p).

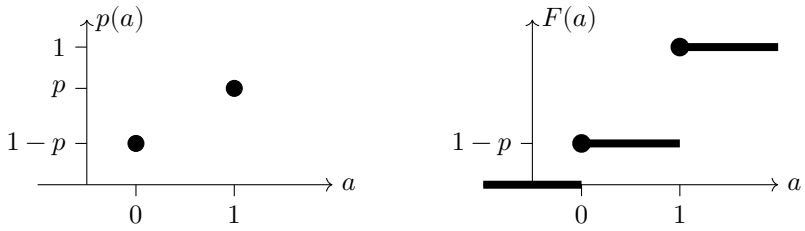
Here are the table and graphs of the pmf and cdf for the Bernoulli(1/2) distribution and below that for the general Bernoulli(p) distribution.

value a:	0	1
pmf $p(a)$:	1/2	1/2
cdf $F(a)$:	1/2	1



Table, pmf and cmf for the Bernoulli(1/2) distribution

values a:	0	1
pmf $p(a)$:	1-p	p
cdf $F(a)$:	1-p	1



Table, pmf and cmf for the Bernoulli(p) distribution

3.2 Binomial Distributions

The binomial distribution Binomial(n,p), or Bin(n,p), models the number of successes in n independent Bernoulli(p) trials.

There is a hierarchy here. A single Bernoulli trial is, say, one toss of a coin. A single binomial trial consists of n Bernoulli trials. For coin flips the sample space for a Bernoulli trial is $\{H, T\}$. The sample space for a binomial trial is all sequences of heads and tails of length n. Likewise a Bernoulli random variable takes values 0 and 1 and a binomial random variables takes values 0, 1, 2, …, n.

Example 6. Binomial(1,p) is the same as Bernoulli(p).

Example 7. The number of heads in n flips of a coin with probability p of heads follows a Binomial(n, p) distribution.

We describe $X \sim \text{Binomial}(n, p)$ by giving its values and probabilities. For notation we will use k to mean an arbitrary number between 0 and n.

We remind you that 'n choose k = $\binom{n}{k}$ = ${}_nC_k$ is the number of ways to choose k things out of a collection of n things and it has the formula

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}. \tag{1}$$

(It is also called a binomial coefficient.) Here is a table for the pmf of a Binomial(n, k) random variable. We will explain how the binomial coefficients enter the pmf for the binomial distribution after a simple example.

values a:	0	1	2	⋯	k	⋯	n
pmf $p(a)$:	$(1-p)^n$	$\binom{n}{1}p^1(1-p)^{n-1}$	$\binom{n}{2}p^2(1-p)^{n-2}$	⋯	$\binom{n}{k}p^k(1-p)^{n-k}$	⋯	p^n

Example 8. What is the probability of 3 or more heads in 5 tosses of a fair coin?

Solution: The binomial coefficients associated with $n = 5$ are

$$\binom{5}{0} = 1, \quad \binom{5}{1} = \frac{5!}{1!\,4!} = \frac{5\cdot4\cdot3\cdot2\cdot1}{4\cdot3\cdot2\cdot1} = 5, \quad \binom{5}{2} = \frac{5!}{2!\,3!} = \frac{5\cdot4\cdot3\cdot2\cdot1}{2\cdot1\cdot3\cdot2\cdot1} = \frac{5\cdot4}{2} = 10,$$

and similarly

$$\binom{5}{3} = 10, \quad \binom{5}{4} = 5, \quad \binom{5}{5} = 1.$$

Using these values we get the following table for $X \sim$ Binomial(5,p).

values a:	0	1	2	3	4	5
pmf $p(a)$:	$(1-p)^5$	$5p(1-p)^4$	$10p^2(1-p)^3$	$10p^3(1-p)^2$	$5p^4(1-p)$	p^5

We were told $p = 1/2$ so

$$P(X \geq 3) = 10\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 + 5\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^5 = \frac{16}{32} = \frac{1}{2}.$$

Think: Why is the value of 1/2 not surprising?

3.3 Explanation of the binomial probabilities

For concreteness, let $n = 5$ and $k = 2$ (the argument for arbitrary n and k is identical.) So $X \sim$ binomial$(5, p)$ and we want to compute $p(2)$. The long way to compute $p(2)$ is to list all the ways to get exactly 2 heads in 5 coin flips and add up their probabilities. The list has 10 entries:

HHTTT, HTHTT, HTTHT, HTTTH, THHTT, THTHT, THTTH, TTHHT, TTHTH, TTTHH

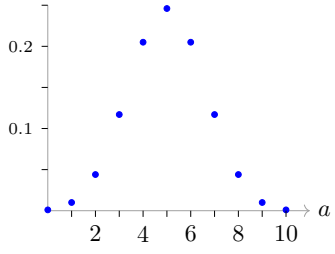
Each entry has the same probability of occurring, namely

$$p^2(1-p)^3.$$

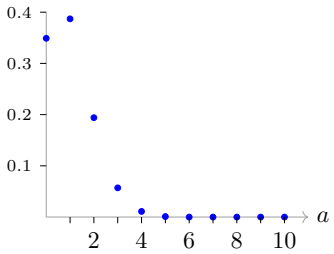
This is because each of the two heads has probability p and each of the 3 tails has probability $1-p$. Because the individual tosses are independent we can multiply probabilities. Therefore, the total probability of exactly 2 heads is the sum of 10 identical probabilities, i.e. $p(2) = 10p^2(1-p)^3$, as shown in the table.

This guides us to the shorter way to do the computation. We have to count the number of sequences with exactly 2 heads. To do this we need to choose 2 of the tosses to be heads and the remaining 3 to be tails. The number of such sequences is the number of ways to choose 2 out of 5 things, that is $\binom{5}{2}$. Since each such sequence has the same probability, $p^2(1-p)^3$, we get the probability of exactly 2 heads $p(2) = \binom{5}{2}p^2(1-p)^3$.

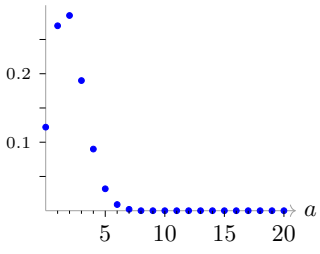
Here are some binomial probability mass functions:



Binomial(10, 0.5)



Binomial(10, 0.1)



Binomial(20, 0.1)

3.4 Geometric Distributions

A geometric distribution models the number of tails before the first head in a sequence of coin flips (Bernoulli trials).

Example 9. (a) Flip a coin repeatedly. Let X be the number of tails before the first heads. So, X can equal 0, i.e. the first flip is heads, 1, 2, …. In principle it takes any nonnegative integer value.

(b) Give a flip of tails the value 0, and heads the value 1. In this case, X is the number of 0's before the first 1.

(c) Give a flip of tails the value 1, and heads the value 0. In this case, X is the number of 1's before the first 0.

(d) Call a flip of tails a success and heads a failure. So, X is the number of successes before the first failure.

(e) Call a flip of tails a failure and heads a success. So, X is the number of failures before the first success.

You can see this models many different scenarios of this type. The most neutral language is the number of tails before the first head.

Formal definition. The random variable X follows a geometric distribution with parameter p if

- X takes the values 0, 1, 2, 3, …
- its pmf is given by $p(k) = P(X = k) = (1-p)^k p$.

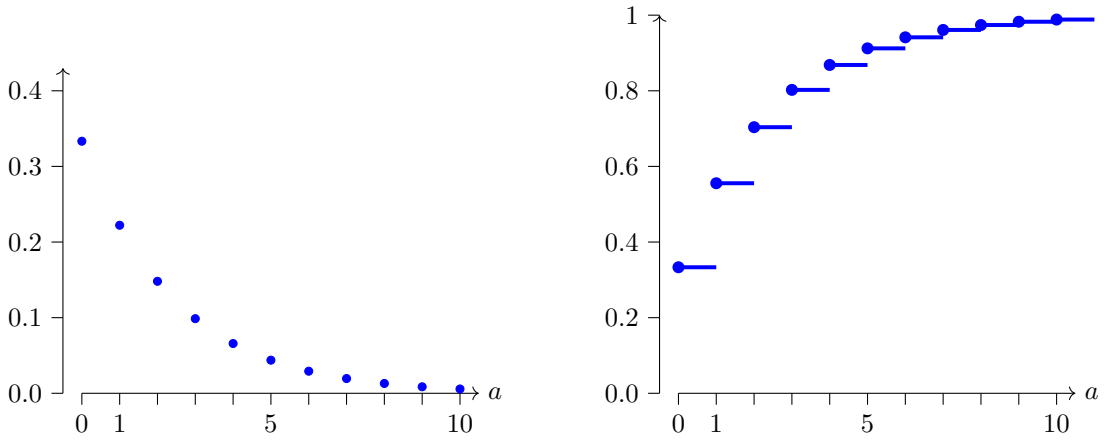
We denote this by $X \sim$ geometric(p) or geo(p). In table form we have:

value	a:	0	1	2	3	…	k	…
pmf	$p(a)$:	p	$(1-p)p$	$(1-p)^2p$	$(1-p)^3p$	…	$(1-p)^kp$	…

Table: $X \sim$ geometric(p): $X =$ the number of 0s before the first 1.

We will show how this table was computed in an example below.

The geometric distribution is an example of a discrete distribution that takes an infinite number of possible values. Things can get confusing when we work with successes and failure since we might want to model the number of successes before the first failure or we might want the number of failures before the first success. To keep straight things straight you can translate to the neutral language of the number of tails before the first heads.



pmf and cdf for the geometric(1/3) distribution

Example 10. Computing geometric probabilities. Suppose that the inhabitants of an island plan their families by having babies until the first girl is born. Assume the probability of having a girl with each pregnancy is 0.5 independent of other pregnancies, that all babies survive and there are no multiple births. What is the probability that a family has k boys?

Solution: In neutral language we can think of boys as tails and girls as heads. Then the number of boys in a family is the number of tails before the first heads.

Let's practice using standard notation to present this. So, let X be the number of boys in a (randomly-chosen) family. So, X is a geometric random variable. We are asked to find $p(k) = P(X = k)$. A family has k boys if the sequence of children in the family from oldest to youngest is

$$BBB \dots BG$$

with the first k children being boys. The probability of this sequence is just the product of the probability for each child, i.e. $(1/2)^k \cdot (1/2) = (1/2)^{k+1}$. (Note: The assumptions of equal probability and independence are simplifications of reality.)

Think: What is the ratio of boys to girls on the island?

More geometric confusion. Another common definition for the geometric distribution is the number of tosses until the first heads. In this case X can take the values 1, i.e. the first flip is heads, 2, 3, ... This is just our geometric random variable plus 1. The methods of computing with it are just like the ones we used above.

3.5 Uniform Distribution

The uniform distribution models any situation where all the outcomes are equally likely.

$$X \sim \text{uniform}(N).$$

X takes values $1, 2, 3, \dots, N$, each with probability $1/N$. We have already seen this distribution many times when modeling to fair coins ($N = 2$), dice ($N = 6$), birthdays ($N = 365$), and poker hands ($N = \binom{52}{5}$).

3.6 Discrete Distributions Applet

The applet at https://mathlets.org/mathlets/probability-distributions/ gives a dynamic view of some discrete distributions. The graphs will change smoothly as you move the various sliders. Try playing with the different distributions and parameters.

This applet is carefully color-coded. Two things with the same color represent the same or closely related notions. By understanding the color-coding and other details of the applet, you will acquire a stronger intuition for the distributions shown.

3.7 Other Distributions

There are a million other named distributions arising is various contexts. We don't expect you to memorize them (we certainly have not!), but you should be comfortable using a resource like Wikipedia to look up a pmf. For example, take a look at the info box at the top right of https://en.wikipedia.org/wiki/Hypergeometric_distribution. The info box lists many (surely unfamiliar) properties in addition to the pmf.

4 Arithmetic with Random Variables

We can do arithmetic with random variables. For example, we can add subtract, multiply or square them.

There is a simple, but extremely important idea for counting. It says that if we have a sequence of numbers that are either 0 or 1 then the sum of the sequence is the number of 1s.

Example 11. Consider the sequence with five 1s

$$1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0.$$

It is easy to see that the sum of this sequence is 5 the number of 1s.

We illustrate this idea by counting the number of heads in n tosses of a coin.

Example 12. Toss a fair coin n times. Let X_j be 1 if the jth toss is heads and 0 if it's tails. So, X_j is a Bernoulli(1/2) random variable. Let X be the total number of heads in the n tosses. Assuming the tosses are independent we know $X \sim \text{binomial}(n, 1/2)$. We can also write

$$X = X_1 + X_2 + X_3 + ... + X_n.$$

Again, this is because the terms in the sum on the right are all either 0 or 1. So, the sum is exactly the number of X_j that are 1, i.e. the number of heads.

The important thing to see in the example above is that we've written the more complicated binomial random variable X as the sum of extremely simple random variables X_j. This will allow us to manipulate X algebraically.

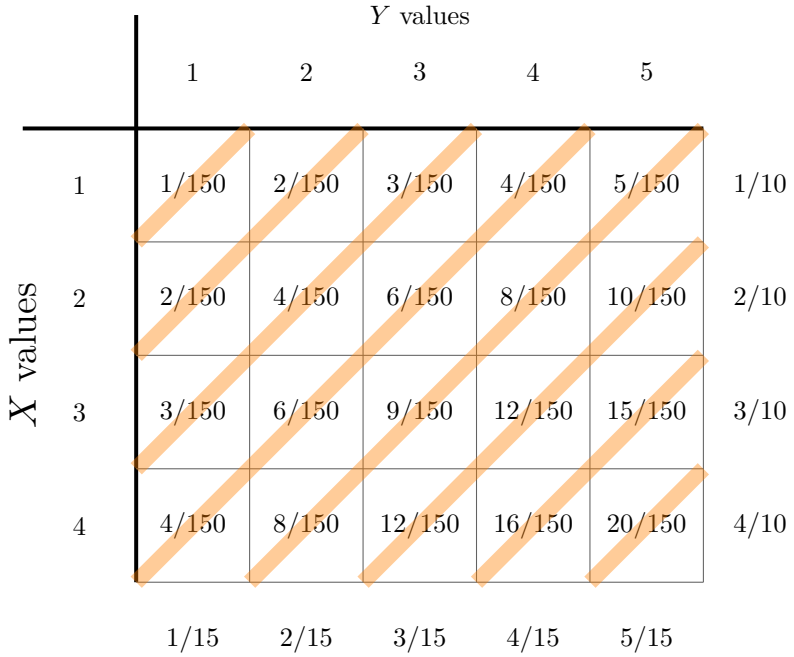
Think: Suppose X and Y are independent and $X \sim \text{binomial}(n, 1/2)$ and $Y \sim \text{binomial}(m, 1/2)$. What kind of distribution does $X + Y$ follow? (**Answer:** $\text{binomial}(n + m, 1/2)$. Why?)

Example 13. Suppose X and Y are independent random variables with the following tables.

Values of X	x:	1	2	3	4	
pmf	$p_X(x)$:	1/10	2/10	3/10	4/10	
Values of Y	y:	1	2	3	4	5
pmf	$p_Y(y)$:	1/15	2/15	3/15	4/15	5/15

Check that the total probability for each random variable is 1. Make a table for the random variable $X + Y$.

Solution: The first thing to do is make a two-dimensional table for the product sample space consisting of pairs (x, y), where x is a possible value of X and y one of Y. To help do the computation, the probabilities for the X values are put in the far right column and those for Y are in the bottom row. Because X and Y are independent the probability for (x, y) pair is just the product of the individual probabilities.



The diagonal stripes show sets of squares where $X + Y$ is the same. All we have to do to compute the probability table for $X + Y$ is sum the probabilities for each stripe.

$X + Y$ values:	2	3	4	5	6	7	8	9
pmf:	1/150	4/150	10/150	20/150	30/150	34/150	31/150	20/150

When the tables are too big to write down we'll need to use purely algebraic techniques to compute the probabilities of a sum. We will learn how to do this in due course.

Discrete Random Variables: Expected Value
Class 4, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know how to compute expected value (mean) of a discrete random variable.
2. Know the expected value of Bernoulli, binomial and geometric random variables.

2 Expected Value

In the R reading questions for this lecture, you simulated the average value of rolling a die many times. You should have gotten a value close to the exact answer of 3.5. To motivate the formal definition of the average, or expected value, we first consider some examples.

Example 1. Suppose we have a six-sided die marked with five 5 3's and one 6. (This was the red one from our non-transitive dice.) What would you expect the average of 6000 rolls to be?

Solution: If we knew the value of each roll, we could compute the average by summing the 6000 values and dividing by 6000. Without knowing the values, we can compute the expected average as follows.

Since there are five 3's and one six we expect roughly 5/6 of the rolls will give 3 and 1/6 will give 6. Assuming this to be exactly true, we have the following table of values and counts:

value:	3	6
expected counts:	5000	1000

The average of these 6000 values is then

$$\frac{5000 \cdot 3 + 1000 \cdot 6}{6000} = \frac{5}{6} \cdot 3 + \frac{1}{6} \cdot 6 = 3.5$$

We consider this the expected average in the sense that we 'expect' each of the possible values to occur with the given frequencies.

Example 2. We roll two standard 6-sided dice. You win \$1000 if the sum is 2 and lose \$100 otherwise. How much do you expect to win on average per trial?

Solution: The probability of a 2 is 1/36. If you play N times, you can 'expect' $\frac{1}{36} \cdot N$ of the trials to give a 2 and $\frac{35}{36} \cdot N$ of the trials to give something else. Thus your total expected winnings are

$$1000 \cdot \frac{N}{36} - 100 \cdot \frac{35N}{36}.$$

To get the expected average per trial we divide the total by N:

$$\text{expected average} \;=\; 1000 \cdot \frac{1}{36} - 100 \cdot \frac{35}{36} = -69.44.$$

Think: Would you be willing to play this game one time? Multiple times?

Notice that in both examples the sum for the expected average consists of terms which are a value of the random variable times its probabilitiy. This leads to the following definition.

Definition: Suppose X is a discrete random variable that takes values x_1, x_2, ..., x_n with probabilities $p(x_1)$, $p(x_2)$, ..., $p(x_n)$. The expected value of X is denoted $E[X]$ and defined by

$$E[X] = \sum_{j=1}^{n} p(x_j)\, x_j = p(x_1)x_1 + p(x_2)x_2 + \ldots + p(x_n)x_n.$$

Notes:

1. The expected value is also called the mean or average of X and often denoted by μ ("mu").
2. As seen in the above examples, the expected value need not be a possible value of the random variable. Rather it is a weighted average of the possible values.
3. Expected value is a summary statistic, providing a measure of the location or central tendency of a random variable.
4. If all the values are equally probable then the expected value is just the usual average of the values.

Example 3. Find $E[X]$ for the random variable X with table:

values of X:	1	3	5
pmf:	1/6	1/6	2/3

Solution: $E[X] = \dfrac{1}{6}\cdot 1 + \dfrac{1}{6}\cdot 3 + \dfrac{2}{3}\cdot 5 = \dfrac{24}{6} = 4$

Example 4. Let X be a Bernoulli(p) random variable. Find $E[X]$.

Solution: X takes values 1 and 0 with probabilities p and $1-p$, so

$$E[X] = p\cdot 1 + (1-p)\cdot 0 = p.$$

Important: This is an important example. Be sure to remember that the expected value of a Bernoulli(p) random variable is p.

Think: What is the expected value of the sum of two dice?

2.1 Mean and center of mass

You may have wondered why we use the name 'probability mass function'. Here's one reason: if we place an object of mass $p(x_j)$ at position x_j for each j, then $E[X]$ is the position of the center of mass. Let's recall the latter notion via an example.

Example 5. Suppose we have two masses along the x-axis, mass $m_1 = 500$ at position $x_1 = 3$ and mass $m_2 = 100$ at position $x_2 = 6$. Where is the center of mass?

Solution: Intuitively we know that the center of mass is closer to the larger mass.

From physics we know the center of mass is

$$\overline{x} = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2} = \frac{500 \cdot 3 + 100 \cdot 6}{600} = 3.5.$$

We call this formula a 'weighted' average of the x_1 and x_2. Here x_1 is weighted more heavily because it has more mass.

Now look at the definition of expected value $E[X]$. It is a weighted average of the values of X with the weights being probabilities $p(x_i)$ rather than masses! We might say that "The expected value is the point at which the distribution would balance". Note the similarity between the physics example and Example 1.

2.2 Algebraic properties of $E[X]$

When we add, scale or shift random variables the expected values do the same. The shorthand mathematical way of saying this is that $E[X]$ is linear.

1. If X and Y are random variables on a sample space Ω then

$$E[X + Y] = E[X] + E[Y]$$

2. If a and b are constants then

$$E[aX + b] = aE[X] + b.$$

We will think of $aX + b$ as scaling X by a and shifting it by b.

Before proving these properties, let's see them in action with a few examples.

Example 6. Roll two dice and let X be the sum. Find $E[X]$.

Solution: Let X_1 be the value on the first die and let X_2 be the value on the second die. Since $X = X_1 + X_2$ we have $E[X] = E[X_1] + E[X_2]$. Earlier we computed that $E[X_1] = E[X_2] = 3.5$, therefore $E[X] = 7$.

Example 7. Let $X \sim \text{binomial}(n, p)$. Find $E[X]$.

Solution: Recall that X models the number of successes in n Bernoulli(p) random variables, which we'll call $X_1, \ldots X_n$. The key fact, which we highlighted in the previous reading for this class, is that

$$X = X_1 + X_2 + \ldots + X_n = \sum_{j=1}^{n} X_j.$$

Now we can use the Algebraic Property (1) to make the calculation simple.

$$X = \sum_{j=1}^{n} X_j \;\Rightarrow\; E[X] = \sum_j E[X_j] = \sum_j p = \boxed{np}.$$

We could have computed $E[X]$ directly as

$$E[X] = \sum_{k=0}^{n} kp(k) = \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k}.$$

It is possible to show that the sum of this series is indeed np. We think you'll agree that the method using Property (1) is much easier.

Example 8. (For infinite random variables the mean does not always exist.) Suppose X has an infinite number of values according to the following table.

values x:	2	2^2	2^3	...	2^k	...
pmf $p(x)$:	$1/2$	$1/2^2$	$1/2^3$	...	$1/2^k$	...

Try to compute the mean.

Solution: The mean is

$$E[X] = \sum_{k=1}^{\infty} 2^k \frac{1}{2^k} = \sum_{k=1}^{\infty} 1 = \infty.$$

The mean does not exist! This can happen with infinite series.

Example 9. Mean of a geometric distribution

Let $X \sim \text{geo}(p)$. Recall this means X takes values $k = 0, 1, 2, \ldots$with probabilities $p(k) = (1-p)^k p$. (X models the number of tails before the first heads in a sequence of Bernoulli trials.) The mean is given by

$$E[X] = \frac{1-p}{p}.$$

To see this requires a clever trick. Mathematicians love this sort of thing and we hope you are able to follow the logic and enjoy it. In this class we will not ask you to come up with something like this on an exam.

Here's the trick.: to compute $E[X]$ we have to sum the infinite series

$$E[X] = \sum_{k=0}^{\infty} k(1-p)^k p.$$

Now, we know the sum of the geometric series: $\displaystyle\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$.

Differentiate both sides: $\displaystyle\sum_{k=0}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}$.

Multiply by x: $\displaystyle\sum_{k=0}^{\infty} kx^k = \frac{x}{(1-x)^2}$.

Replace x by $1-p$: $\displaystyle\sum_{k=0}^{\infty} k(1-p)^k = \frac{1-p}{p^2}$.

Multiply by p: $\displaystyle\sum_{k=0}^{\infty} k(1-p)^k p = \frac{1-p}{p}$.

This last expression is the mean.

$$E[X] = \frac{1-p}{p}.$$

Example 10. Flip a fair coin until you get heads for the first time. What is the expected number of times you flipped tails?

Solution: The number of tails before the first head is modeled by $X \sim \text{geo}(1/2)$. From the previous example $E[X] = \dfrac{1/2}{1/2} = 1$. This is a surprisingly small number.

Example 11. Michael Jordan, perhaps the greatest basketball player ever, made 80% of his free throws. In a game what is the expected number he would make before his first miss.

Solution: Here is an example where we want the number of successes before the first failure. Using the neutral language of heads and tails: success is tails (probability $1-p$) and failure is heads (probability $= p$). Therefore $p = 0.2$ and the number of tails (made free throws) before the first heads (missed free throw) is modeled by a $X \sim \text{geo}(0.2)$. We saw in Example 9 that this is

$$E[X] = \frac{1-p}{p} = \frac{0.8}{0.2} = 4.$$

2.3 Expected values of functions of a random variable

(The change of variables formula.)

If X is a discrete random variable taking values x_1, x_2, …and h is a function then $h(X)$ is a new random variable. Its expected value is

$$E[h(X)] = \sum_j h(x_j)p(x_j).$$

We illustrate this with several examples.

Example 12. Let X be the value of a roll of one die and let $Y = X^2$. Find $E[Y]$.

Solution: Since there are a small number of values we can make a table.

X	1	2	3	4	5	6
Y	1	4	9	16	25	36
prob	1/6	1/6	1/6	1/6	1/6	1/6

Notice the probability for each Y value is the same as that of the corresponding X value. So,

$$E[Y] = E[X^2] = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \ldots + 6^2 \cdot \frac{1}{6} = 15.167.$$

Example 13. Roll two dice and let X be the sum. Suppose the payoff function is given by $Y = X^2 - 6X + 1$. Is this a good bet?

Solution: We have $E[Y] = \displaystyle\sum_{j=2}^{12} (j^2 - 6j + 1)p(j)$, where $p(j) = P(X = j)$.

We show the table, but really we'll use R to do the calculation.

X	2	3	4	5	6	7	8	9	10	11	12
Y	-7	-8	-7	-4	1	8	17	28	41	56	73
prob	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Here's the R code I used to compute $E[Y] = 13.833$.

```
x = 2:12
y = x^2 - 6*x + 1
p = c(1 2 3 4 5 6 5 4 3 2 1)/36
ave = sum(p*y)
```

It gave $E[Y] = 13.833$.

To answer the question above: since the expected payoff is positive it looks like a bet worth taking.

Quiz: If $Y = h(X)$ does $E[Y] = h(E[X])$? **Solution: NO!!!** This is not true in general!

Think: Is it true in the previous example?

Quiz: If $Y = 3X + 77$ does $E[Y] = 3E[X] + 77$?

Solution: Yes. By property (2), scaling and shifting does behave like this.

2.4 Proofs of the algebraic properties of $E[X]$

We finish by proving the algebraic properties of $E[X]$.

1. For random variables X and Y on a sample space Ω: $E[X + Y] = E[X] + E[Y]>$

2. For constants a, b and random variable X: $E[aX + b] = aE[X] + b$.

The proof of Property (1) is simple, but there is some subtlety in even understanding what it means to add two random variables. Recall that the value of random variable is a number determined by the outcome of an experiment. To add X and Y means to add the values of X and Y for the same outcome. In table form this looks like:

outcome ω:	ω_1	ω_2	ω_3	...	ω_n
value of X:	x_1	x_2	x_3	...	x_n
value of Y:	y_1	y_2	y_3	...	y_n
value of $X + Y$:	$x_1 + y_1$	$x_2 + y_2$	$x_3 + y_3$	...	$x_n + y_n$
prob. $P(\omega)$:	$P(\omega_1)$	$P(\omega_2)$	$P(\omega_3)$	...	$P(\omega_n)$

The proof of (1) follows immediately:

$$E[X + Y] = \sum (x_i + y_i)P(\omega_i) = \sum x_i P(\omega_i) + \sum y_i P(\omega_i) = E[X] + E[Y].$$

The proof of Property (2) only takes one line.

$$E[aX + b] = \sum p(x_i)(ax_i + b) = a \sum p(x_i)x_i + b \sum p(x_i) = aE[X] + b.$$

The b term in the last expression follows because $\sum p(x_i) = 1$.

Variance of Discrete Random Variables
Class 5, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to compute the variance and standard deviation of a random variable.
2. Understand that standard deviation is a measure of scale or spread.
3. Be able to compute variance using the properties of scaling and linearity.

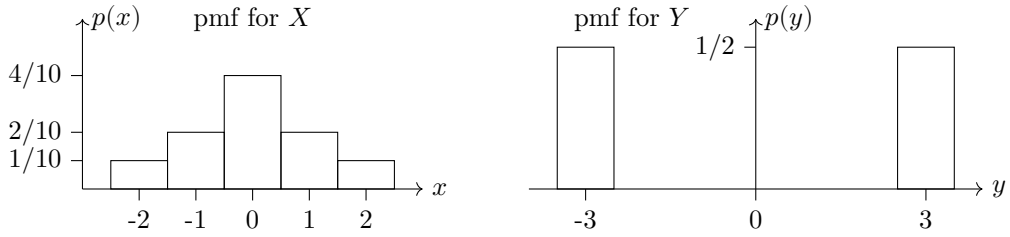
2 Spread

The expected value (mean) of a random variable is a measure of location or central tendency. If you had to summarize a random variable with a single number, the mean would be a good choice. Still, the mean leaves out a good deal of information. For example, the random variables X and Y below both have mean 0, but their probability mass is spread out about the mean quite differently.

values X	-2	-1	0	1	2
pmf $p(x)$	1/10	2/10	4/10	2/10	1/10

values Y	-3	3
pmf $p(y)$	1/2	1/2

It's probably a little easier to see the different spreads in plots of the probability mass functions. We use bars instead of dots to give a better sense of the mass.



pmf's for two different distributions both with mean 0

In the next section, we will learn how to quantify this spread.

3 Variance and standard deviation

Taking the mean as the center of a random variable's probability distribution, the variance is a measure of how much the probability mass is spread out around this center. We'll start with the formal definition of variance and then unpack its meaning.

Definition: If X is a random variable with mean $E[X] = \mu$, then the variance of X is defined by

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

The standard deviation σ of X is defined by

$$\sigma = \sqrt{\text{Var}(X)}.$$

If the relevant random variable is clear from context, then the variance and standard deviation are often denoted by σ^2 and σ ('sigma'), just as the mean is μ ('mu').

What does this mean? First, let's rewrite the definition explicitly as a sum. If X takes values $x_1, x_2, \dots, x_n$ with probability mass function $p(x_i)$ then

$$\text{Var}(X) = E[(X-\mu)^2] = \sum_{i=1}^{n} p(x_i)(x_i-\mu)^2.$$

In words, the formula for $\text{Var}(X)$ says to take a weighted average of the squared distance to the mean. By squaring, we make sure we are averaging only non-negative values, so that the spread to the right of the mean won't cancel that to the left. By using expectation, we are weighting high probability values more than low probability values. (See Example 2 below.)

Note on units:
1. σ has the same units as X.
2. $\text{Var}(X)$ has the same units as the square of X. So if X is in meters, then $\text{Var}(X)$ is in meters squared.
Because σ and X have the same units, the standard deviation is a natural measure of spread.

Let's work some examples to make the notion of variance clear.

Example 1. Compute the mean, variance and standard deviation of the random variable X with the following table of values and probabilities.

value x	1	3	5
pmf $p(x)$	1/4	1/4	1/2

Solution: First we compute $E[X] = 7/2$. Then we extend the table to include $(X-7/2)^2$.

value x	1	3	5
$p(x)$	1/4	1/4	1/2
$(x-7/2)^2$	25/4	1/4	9/4

Now the computation of the variance is similar to that of expectation:

$$\text{Var}(X) = \frac{25}{4}\cdot\frac{1}{4} + \frac{1}{4}\cdot\frac{1}{4} + \frac{9}{4}\cdot\frac{1}{2} = \frac{11}{4}.$$

Taking the square root we have the standard deviation $\sigma = \sqrt{11/4}$.

Example 2. For each random variable X, Y, Z, and W plot the pmf and compute the mean and variance.

(i)

value x	1	2	3	4	5
pmf $p(x)$	1/5	1/5	1/5	1/5	1/5

(ii)

value y	1	2	3	4	5
pmf $p(y)$	1/10	2/10	4/10	2/10	1/10

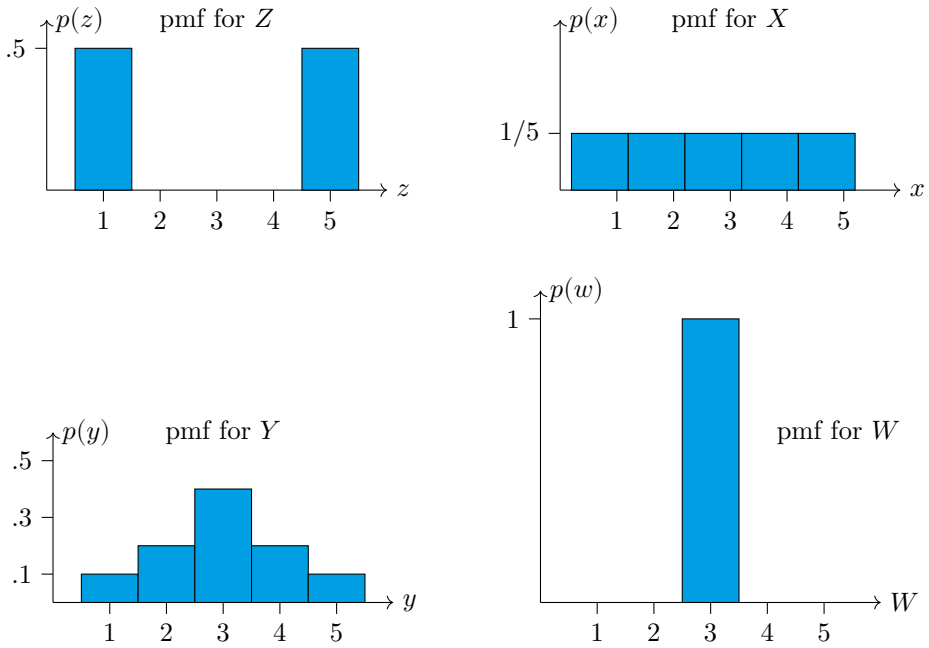
(iii)

value z	1	2	3	4	5
pmf $p(z)$	5/10	0	0	0	5/10

(iv)

value w	1	2	3	4	5
pmf $p(w)$	0	0	1	0	0

Solution: Each random variable has the same mean 3, but the probability is spread out differently. In the plots below, we order the pmf's from largest to smallest variance: Z, X, Y, W.



Next we'll verify our visual intuition by computing the variance of each of the variables. All of them have mean $\mu = 3$. Since the variance is defined as an expected value, we can compute it using the tables.

(i)

value x	1	2	3	4	5
pmf $p(x)$	1/5	1/5	1/5	1/5	1/5
$(X-\mu)^2$	4	1	0	1	4

$\text{Var}(X) = E[(X-\mu)^2] = \frac{4}{5} + \frac{1}{5} + \frac{0}{5} + \frac{1}{5} + \frac{4}{5} = \boxed{2}$.

(ii)

value y	1	2	3	4	5
$p(y)$	1/10	2/10	4/10	2/10	1/10
$(Y-\mu)^2$	4	1	0	1	4

$\text{Var}(Y) = E[(Y-\mu)^2] = \frac{4}{10} + \frac{2}{10} + \frac{0}{10} + \frac{2}{10} + \frac{4}{10} = \boxed{1.2}$.

(iii)

value z	1	2	3	4	5
pmf $p(z)$	5/10	0	0	0	5/10
$(Z-\mu)^2$	4	1	0	1	4

$\text{Var}(Z) = E[(Z-\mu)^2] = \frac{20}{10} + \frac{20}{10} = \boxed{4}$.

(iv)

value w	1	2	3	4	5
pmf $p(w)$	0	0	1	0	0
$(W-\mu)^2$	4	1	0	1	4

$\mathrm{Var}(W) = \boxed{0}$. Note that W doesn't vary, so it has variance 0!

3.1 The variance of a Bernoulli(p) random variable.

Bernoulli random variables are fundamental, so we should know their variance.

If $X \sim$ Bernoulli(p) then

$$\mathrm{Var}(X) = p(1-p).$$

Proof: We know that $E[X] = p$. We compute $\mathrm{Var}(X)$ using a table.

values X	0	1
pmf $p(x)$	$1-p$	p
$(X-\mu)^2$	$(0-p)^2$	$(1-p)^2$

$\mathrm{Var}(X) = (1-p)p^2 + p(1-p)^2 = (1-p)p(1-p+p) = \boxed{(1-p)p.}$

As with all things Bernoulli, you should remember this formula.

Think: For what value of p does Bernoulli(p) have the highest variance? Try to answer this by plotting the PMF for various p.

3.2 A word about independence

So far we have been using the notion of independent random variable without ever carefully defining it. For example, a binomial distribution is the sum of independent Bernoulli trials. This may (should?) have bothered you. Of course, we have an intuitive sense of what independence means for experimental trials. We also have the probabilistic sense that random variables X and Y are independent if knowing the value of X gives you no information about the value of Y.

In a few classes we will work with continuous random variables and joint probability functions. After that we will be ready for a full definition of independence. For now we can use the following definition, which is exactly what you expect and is valid for discrete random variables.

Definition: The discrete random variables X and Y are independent if

$$P(X = a,\, Y = b) \;=\; P(X = a)P(Y = b)$$

for any values a, b. That is, the probabilities multiply.

3.3 Properties of variance

The three most useful properties for computing variance are:

1. If X and Y are independent then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

2. For constants a and b, $\text{Var}(aX+b) = a^2\text{Var}(X)$.

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

For Property 1, note carefully the requirement that X and Y are independent. We will return to the proof of Property 1 in a later class.

Property 3 gives a formula for $\text{Var}(X)$ that is often easier to use in hand calculations. The computer is happy to use the definition! We'll prove Properties 2 and 3 after some examples.

Example 3. Suppose X and Y are independent and $\text{Var}(X) = 3$ and $\text{Var}(Y) = 5$. Find: (i) $\text{Var}(X+Y)$, (ii) $\text{Var}(3X+4)$, (iii) $\text{Var}(X+X)$, (iv) $\text{Var}(X+3Y)$.

Solution: To compute these variances we make use of Properties 1 and 2.
(i) Since X and Y are independent, $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) = 8$.

(ii) Using Property 2, $\text{Var}(3X+4) = 9 \cdot \text{Var}(X) = 27$.

(iii) Don't be fooled! Property 1 fails since X is certainly not independent of itself. We can use Property 2: $\text{Var}(X+X) = \text{Var}(2X) = 4 \cdot \text{Var}(X) = 12$. (Note: if we mistakenly used Property 1, we would the wrong answer of 6.)

(iv) We use both Properties 1 and 2.

$$\text{Var}(X+3Y) = \text{Var}(X) + \text{Var}(3Y) = 3 + 9 \cdot 5 = 48.$$

Example 4. Use Property 3 to compute the variance of $X \sim \text{Bernoulli}(p)$.

Solution: From the table

X	0	1
$p(x)$	$1-p$	p
X^2	0	1

we have $E[X^2] = p$. So Property 3 gives

$$\text{Var}(X) = E[X^2] - E[X]^2 = p - p^2 = p(1-p).$$

This agrees with our earlier calculation.

Example 5. Redo Example 1 using Property 3.

Solution: From the table

X	1	3	5
$p(x)$	1/4	1/4	1/2
X^2	1	9	25

we have $E[X] = 7/2$ and

$$E[X^2] = 1^2 \cdot \frac{1}{4} + 3^2 \cdot \frac{1}{4} + 5^2 \cdot \frac{1}{2} = \frac{60}{4} = 15.$$

So $\text{Var}(X) = 15 - (7/2)^2 = 11/4$ –as before in Example 1.

3.4 Variance of binomial(n,p)

Suppose $X \sim \text{binomial}(n,p)$. Since X is the sum of *independent* $\text{Bernoulli}(p)$ variables and each Bernoulli variable has variance $p(1-p)$ we have

$$X \sim \text{binomial}(n,p) \;\Rightarrow\; \text{Var}(X) = np(1-p).$$

3.5 Proof of properties 2 and 3

Proof of Property 2: This follows from the properties of $E[X]$ and some algebra.

Let $\mu = E[X]$. Then $E[aX + b] = a\mu + b$ and

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

Proof of Property 3: We use the properties of $E[X]$ and a bit of algebra. Remember that μ is a constant and that $E[X] = \mu$.

$$\begin{aligned} E[(X-\mu)^2] &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2 \\ &= E[X^2] - E[X]^2. \qquad \text{QED} \end{aligned}$$

4 Tables of Distributions and Properties

Distribution	range X	pmf $p(x)$	mean $E[X]$	variance $\text{Var}(X)$
Bernoulli(p)	0, 1	$p(0) = 1-p, \quad p(1) = p$	p	$p(1-p)$
Binomial(n, p)	0, 1,..., n	$p(k) = \binom{n}{k} p^k(1-p)^{n-k}$	np	$np(1-p)$
Uniform(n)	1, 2, ..., n	$p(k) = \dfrac{1}{n}$	$\dfrac{n+1}{2}$	$\dfrac{n^2-1}{12}$
Geometric(p)	0, 1, 2,...	$p(k) = p(1-p)^k$	$\dfrac{1-p}{p}$	$\dfrac{1-p}{p^2}$

Let X be a discrete random variable with range x_1, x_2, ...and pmf $p(x_j)$.

Expected Value:		**Variance**:
Synonyms:	mean, average	
Notation:	$E[X]$, μ	$\text{Var}(X)$, σ^2
Definition:	$E[X] = \sum_j p(x_j)x_j$	$E[(X-\mu)^2] = \sum_j p(x_j)(x_j-\mu)^2$
Scale and shift:	$E[aX+b] = aE[X] + b$	$\text{Var}(aX+b) = a^2\text{Var}(X)$
Linearity:	(for any X, Y) $E[X+Y] = E[X] + E[Y]$	(for X, Y independent) $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$
Functions of X:	$E[h(X)] = \sum p(x_j)\, h(x_j)$	
Alternative formula:		$\text{Var}(X) = E[X^2] - E[X]^2 = E[X^2] - \mu^2$

Continuous Random Variables
Class 5, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the definition of a continuous random variable.
2. Know the definition of the probability density function (pdf) and cumulative distribution function (cdf).
3. Be able to explain why we use probability density for continuous random variables.

2 Introduction

We now turn to continuous random variables. All random variables assign a number to each outcome in a sample space. Whereas discrete random variables take on a discrete set of possible values, continuous random variables have a continuous set of values.

Computationally, to go from discrete to continuous we simply replace sums by integrals. It will help you to keep in mind that (informally) an integral is just a continuous sum.

Example 1. Since time is continuous, the amount of time Jon is early (or late) for class is a continuous random variable. Let's go over this example in some detail.

Suppose you measure how early Jon arrives to class each day (in units of minutes). That is, the outcome of one trial in our experiment is a time in minutes. We'll assume there are random fluctuations in the exact time he shows up. Since in principle Jon could arrive, say, 3.43 minutes early, or 2.7 minutes late (corresponding to the outcome -2.7), or at any other time, the sample space consists of all real numbers. So the random variable which gives the outcome itself has a continuous range of possible values.

It is too cumbersome to keep writing 'the random variable', so in future examples we might write: Let T = "time in minutes that Jon is early for class on any given day."

3 Calculus Warmup

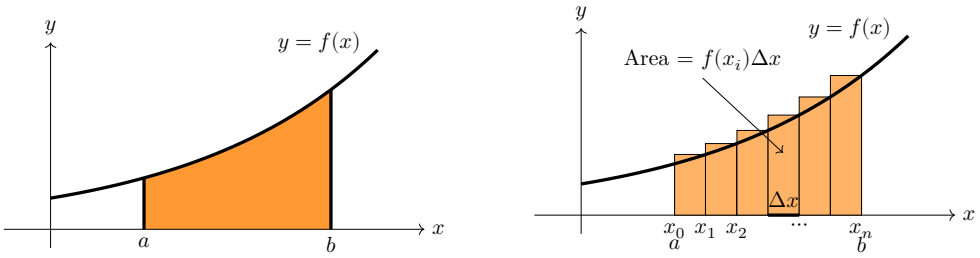
While we will assume you can compute the most familiar forms of derivatives and integrals by hand, we do not expect you to be calculus whizzes. For tricky expressions, we'll let the computer do most of the calculating. Conceptually, you should be comfortable with two views of a definite integral.

1. $\displaystyle\int_a^b f(x)\,dx$ = area under the curve $y = f(x)$.
2. $\displaystyle\int_a^b f(x)\,dx$ = 'sum of $f(x)\,dx$'.

The connection between the two is:

$$\text{area} \approx \text{sum of rectangle areas} \quad = \quad f(x_1)\Delta x + f(x_2)\Delta x + \ldots + f(x_n)\Delta x = \sum_1^n f(x_i)\Delta x.$$

As the width Δx of the intervals gets smaller the approximation becomes better.



Area is approximately the sum of rectangles

Note: In calculus you learned to compute integrals by finding antiderivatives. This is important for calculations, but don't confuse this method for the reason we use integrals. Our interest in integrals comes primarily from its interpretation as a 'sum' and to a much lesser extent its interpretation as area.

4 Continuous Random Variables and Probability Density Functions

A continuous random variable takes a range of values, which may be finite or infinite in extent. Here are a few examples of ranges: $[0, 1]$, $[0, \infty)$, $(-\infty, \infty)$, $[a, b]$.

Definition: A random variable X is continuous if there is a function $f(x)$ such that for any $c \leq d$ we have

$$P(c \leq X \leq d) \; = \; \int_c^d f(x)\, dx. \tag{1}$$

The function $f(x)$ is called the probability density function (**pdf**).

The pdf always satisfies the following properties:

1. $f(x) \geq 0$ (f is nonnegative).
2. $\displaystyle\int_{-\infty}^{\infty} f(x)\, dx = 1$ (This is equivalent to: $P(-\infty < X < \infty) = 1$).

The probability density function $f(x)$ of a continuous random variable is the analogue of the probability mass function $p(x)$ of a discrete random variable. Here are two important differences:

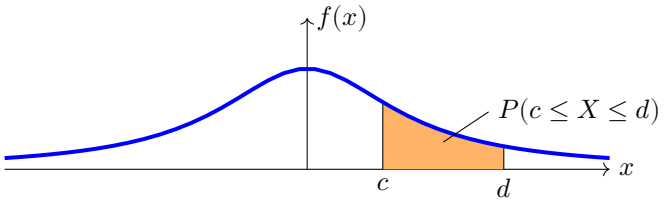
1. Unlike $p(x)$, the pdf $f(x)$ is *not* a probability. You have to integrate it to get probability. (See section 4.2 below.)
2. Since $f(x)$ is not a probability, there is no restriction that $f(x)$ be less than or equal to 1.

Note: In Property 2, we integrated over $(-\infty, \infty)$ since we did not know the range of values taken by X. Formally, this makes sense because we just define $f(x)$ to be 0 outside of the range of X. In practice, we would integrate between bounds given by the range of X.

4.1 Graphical View of Probability

If you graph the probability density function of a continuous random variable X then

$$\boxed{P(c \le X \le d) = \text{area under the graph between } c \text{ and } d.}$$



Think: What is the total area under the pdf $f(x)$?

4.2 The terms 'probability mass' and 'probability density'

Why do we use the terms mass and density to describe the pmf and pdf? What is the difference between the two? The simple answer is that these terms are completely analogous to the mass and density you saw in physics and calculus. We'll review this first for the probability mass function and then discuss the probability density function.

Mass as a sum:

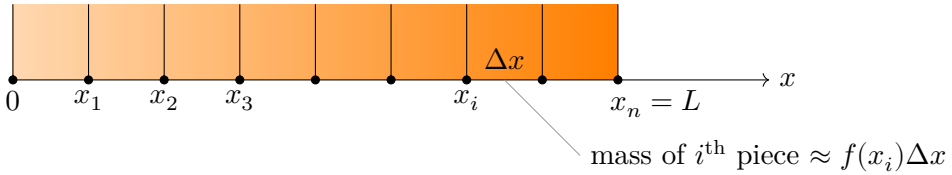
If masses m_1, m_2, m_3, and m_4 are set in a row at positions x_1, x_2, x_3, and x_4, then the total mass is $m_1 + m_2 + m_3 + m_4$.

We can define a 'mass function' $p(x)$ with $p(x_j) = m_j$ for $j = 1, 2, 3, 4$, and $p(x) = 0$ otherwise. In this notation the total mass is $p(x_1) + p(x_2) + p(x_3) + p(x_4)$.

The probability mass function behaves in exactly the same way, except it has the dimension of probability instead of mass.

Mass as an integral of density:

Suppose you have a rod of length L meters with varying density $f(x)$ kg/m. (Note the units are mass/length.)



If the density varies continuously, we must find the total mass of the rod by integration:

$$\textbf{total mass} = \int_0^L f(x)\,dx.$$

This formula comes from dividing the rod into small pieces and 'summing' up the mass of each piece. That is:

$$\text{total mass} \approx \sum_{i=1}^{n} f(x_i)\,\Delta x$$

In the limit as Δx goes to zero the sum becomes the integral.

The probability density function behaves exactly the same way, except it has units of probability/(unit x) instead of kg/m. Indeed, equation (1) is exactly analogous to the above integral for total mass.

While we're on a physics kick, note that for both discrete and continuous random variables, the expected value is simply the center of mass or balance point.

Example 2. Suppose X has pdf $f(x) = 3$ on $[0, 1/3]$ (this means $f(x) = 0$ outside of $[0, 1/3]$). Graph the pdf and compute $P(0.1 \le X \le 0.2)$ and $P(0.1 \le X \le 1)$.

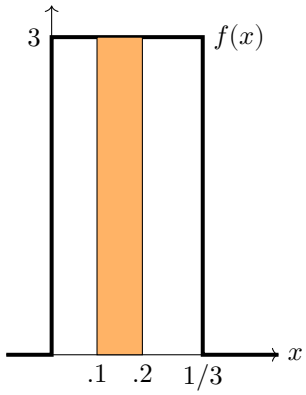
Solution: $P(0.1 \le X \le 0.2)$ is shown below at left. We can compute the integral:

$$P(0.1 \le X \le 0.2) = \int_{0.1}^{0.2} f(x)\,dx = \int_{0.1}^{0.2} 3\,dx = 0.3.$$

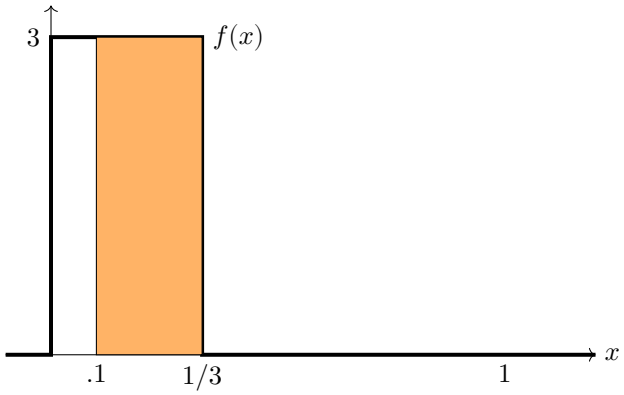
Or we can find the area geometrically:

$$\text{area of rectangle} = 3 \cdot 0.1 = 0.3.$$

$P(0.1 \le X \le 1)$ is shown below at right. Since there is only area under $f(x)$ up to $1/3$, we have $P(0.1 \le X \le 1) = 3 \cdot (1/3 - 0.1) = 0.7$.



$P(0.1 \le X \le 0.2)$



$P(0.1 \le X \le 1)$

Think: In the previous example $f(x)$ takes values greater than 1. Why does this not violate the rule that probabilities are always between 0 and 1?

Note on notation. We can define a random variable by giving its range and probability density function. For example we might say, let X be a random variable with range [0,1]

and pdf $f(x) = x/2$. Implicitly, this means that X has no probability density outside of the given range. If we wanted to be absolutely rigorous, we would say explicitly that $f(x) = 0$ outside of [0,1], but in practice this won't be necessary.

Example 3. Let X be a random variable with range [0,1] and pdf $f(x) = Cx^2$. What is the value of C?

Solution: Since the total probability must be 1, we have

$$\int_0^1 f(x)\,dx = 1 \quad \Leftrightarrow \quad \int_0^1 Cx^2\,dx = 1.$$

By evaluating the integral, the equation at right becomes

$$C/3 = 1 \;\Rightarrow\; \boxed{C = 3}.$$

Note: We say the constant C above is needed to normalize the density so that the total probability is 1.

Example 4. Let X be the random variable in the Example 3. Find $P(X \le 1/2)$.

Solution: $P(X \le 1/2) = \displaystyle\int_0^{1/2} 3x^2\,dx = x^3\Big|_0^{1/2} = \boxed{\frac{1}{8}}.$

Think: For this X (or any continuous random variable):

- What is $P(a \le X \le a)$?
- What is $P(X = 0)$?
- Does $P(X = a) = 0$ mean that X can never equal a?

In words the above questions get at the fact that the probability that a random person's height is exactly 5'9" (to infinite precision, i.e. no rounding!) is 0. Yet it is still possible that someone's height is exactly 5'9". So the answers to the thinking questions are 0, 0, and No.

4.3 Cumulative Distribution Function

The cumulative distribution function (**cdf**) of a continuous random variable X is defined in exactly the same way as the cdf of a discrete random variable.

$$F(b) \;=\; P(X \le b).$$

Note well that the definition is about probability. When using the cdf you should first think of it as a probability. Then when you go to calculate it you can use

$$F(b) \;=\; P(X \le b) \;=\; \int_{-\infty}^{b} f(x)\,dx, \quad \text{where } f(x) \text{ is the pdf of } X.$$

Notes:

1. For discrete random variables, we defined the cumulative distribution function but did

not have much occasion to use it. The cdf plays a far more prominent role for continuous random variables.

2. As before, we started the integral at $-\infty$ because we did not know the precise range of X. Formally, this still makes sense since $f(x) = 0$ outside the range of X. In practice, we'll know the range and start the integral at the start of the range.

3. In practice we often say 'X has distribution $F(x)$' rather than 'X has cumulative distribution function $F(x)$.'

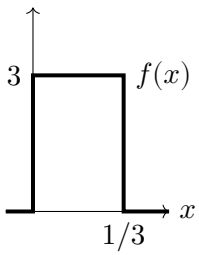
Example 5. Find the cumulative distribution function for the density in Example 2.

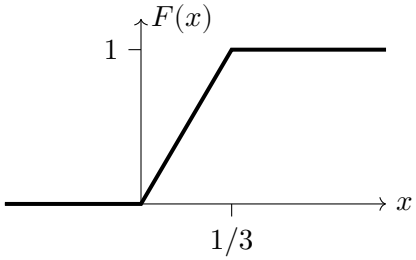
Solution: For a in [0,1/3] we have $F(a) = \int_0^a f(x)\,dx = \int_0^a 3\,dx = 3a$.

Since $f(x)$ is 0 outside of [0,1/3] we know $F(a) = P(X \le a) = 0$ for $a < 0$ and $F(a) = 1$ for $a > 1/3$. Putting this all together we have

$$F(a) = \begin{cases} 0 & \text{if } a < 0 \\ 3a & \text{if } 0 \le a \le 1/3 \\ 1 & \text{if } 1/3 < a. \end{cases}$$

Here are the graphs of $f(x)$ and $F(x)$.





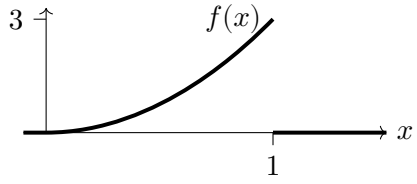
Note the different scales on the vertical axes. Remember that the vertical axis for the pdf represents probability density and that of the cdf represents probability.

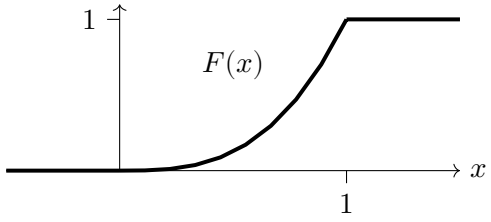
Example 6. Find the cdf for the pdf in Example 3, $f(x) = 3x^2$ on $[0, 1]$. Suppose X is a random variable with this distribution. Find $P(X < 1/2)$.

Solution: $f(x) = 3x^2$ on [0,1] $\Rightarrow$ $F(a) = \int_0^a 3x^2\,dx = a^3$ on [0,1]. Therefore,

$$F(a) = \begin{cases} 0 & \text{if } a < 0 \\ a^3 & \text{if } 0 \le a \le 1 \\ 1 & \text{if } 1 < a \end{cases}$$

Thus, $P(X < 1/2) = F(1/2) = 1/8$. Here are the graphs of $f(x)$ and $F(x)$:





4.4 Properties of cumulative distribution functions

Here is a summarry of the most important properties of cumulative distribution functions (cdf)

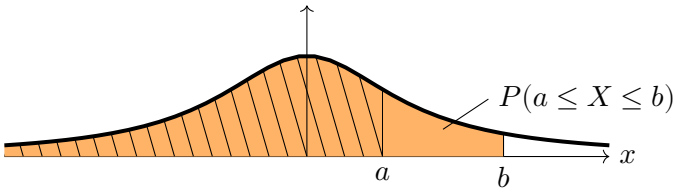
1. (Definition) $F(x) = P(X \le x)$
2. $0 \le F(x) \le 1$
3. $F(x)$ is non-decreasing, i.e. if $a \le b$ then $F(a) \le F(b)$.
4. $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$
5. $P(a \le X \le b) = F(b) - F(a)$
6. $F'(x) = f(x)$.

Properties 2, 3, 4 are identical to those for discrete distributions. The graphs in the previous examples illustrate them.

Property 5 can be seen algebraically:

$$\int_{-\infty}^{b} f(x)\,dx = \int_{-\infty}^{a} f(x)\,dx + \int_{a}^{b} f(x)\,dx$$
$$\Leftrightarrow \int_{a}^{b} f(x)\,dx = \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx$$
$$\Leftrightarrow P(a \le X \le b) = F(b) - F(a).$$

Property 5 can also be seen geometrically. The orange region below represents $F(b)$ and the striped region represents $F(a)$. Their difference is $P(a \le X \le b)$.



Property 6 is the fundamental theorem of calculus.

4.5 Probability density as a dartboard

We find it helpful to think of sampling values from a continuous random variable as throwing darts at a funny dartboard. Consider the region underneath the graph of a pdf as a dartboard. Divide the board into small equal size squares and suppose that when you throw a dart you are equally likely to land in any of the squares. The probability the dart lands in a given region is the fraction of the total area under the curve taken up by the region. Since the total area equals 1, this fraction is just the area of the region. If X represents the x-coordinate of the dart, then the probability that the dart lands with x-coordinate between a and b is just

$$P(a \le X \le b) \quad = \quad \text{area under } f(x) \text{ between } a \text{ and } b \quad = \quad \int_{a}^{b} f(x)\,dx.$$

Gallery of Continuous Random Variables
Class 5, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to give examples of what uniform, exponential and normal distributions are used to model.
2. Be able to give the range and pdf's of uniform, exponential and normal distributions.

2 Introduction

Here we introduce a few fundamental continuous distributions. These will play important roles in the statistics part of the class. For each distribution, we give the range, the pdf, the cdf, and a short description of situations that it models. These distributions all depend on parameters, which we specify.

As you look through each distribution do not try to memorize all the details; you can always look those up. Rather, focus on the shape of each distribution and what it models.

Although it comes towards the end, we call your attention to the normal distribution. It is easily the most important distribution defined here.

2.1 Parametrized distributions

When we studied discrete random variables we learned, for example, about the Bernoulli(p) distribution. The probability p used to define the distribution is called a parameter and Bernoulli(p) is called a parametrized distribution. For example, tosses of fair coin follow a Bernoulli distribution where the parameter $p = 0.5$. When we study statistics one of the key questions will be to estimate the parameters of a distribution. For example, if I have a coin that may or may not be fair then I know it follows a Bernoulli(p) distribution, but I don't know the value of the parameter p. I might run experiments and use the data to estimate the value of p.

As another example, the binomial distribution Binomial(n, p) depends on two parameters n and p.

In the following sections we will look at specific parametrized continuous distributions. The applet https://mathlets.org/mathlets/probability-distributions/ allows you to visualize the pdf and cdf of these distributions and to dynamically change the parameters.

3 Uniform distribution

1. Parameters: a, b.
2. Range: $[a, b]$.

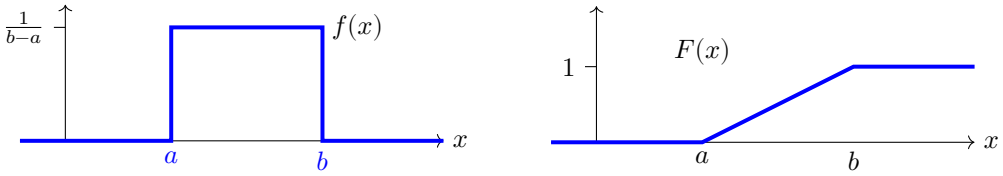
3. Notation: uniform(a, b) or U(a, b).

4. Probability density function (pdf): $f(x) = \dfrac{1}{b-a}$ for $a \le x \le b$.

5. Cumulative distribution function (cdf): $F(x) = (x-a)/(b-a)$ for $a \le x \le b$.

6. Models: Situtations where all outcomes in the range have equal probability (more precisely all outcomes have the same probability density).

Graphs:



pdf and cdf for uniform(a,b) distribution.

Example 1. 1. Suppose we have a tape measure with markings at each millimeter. If we measure (to the nearest marking) the length of items that are roughly a meter long, the rounding error will be uniformly distributed between -0.5 and 0.5 millimeters.

2. Many board games use spinning arrows (spinners) to introduce randomness. When spun, the arrow stops at an angle that is uniformly distributed between 0 and 2π radians.

3. In most pseudo-random number generators, the basic generator simulates a uniform distribution and all other distributions are constructed by transforming the basic generator.

4 Exponential distribution

1. Parameter: λ.

2. Range: $[0, \infty)$.

3. Notation: exponential(λ) or exp(λ).

4. Probability density function (pdf): $f(x) = \lambda e^{-\lambda x}$ for $0 \le x$.

5. Cumulative distribution function (cdf): (This is an easy integral.)

$$F(x) = 1 - e^{-\lambda x} \text{ for } x \ge 0$$

6. Right tail distribution: $P(X > x) = 1 - F(x) = e^{-\lambda x}$. (Note: this is defined as $P(X > x)$, i.e. that X is to the right of x on the number line.)

7. Models: The waiting time for a continuous process to change state.

Example 2. If I step out to 77 Mass Ave after class and wait for the next taxi, my waiting time in minutes is exponentially distributed. We will see that in this case λ is given by 1/(average number of taxis that pass per minute).

Example 3. The exponential distribution models the waiting time until an unstable isotope undergoes nuclear decay. In this case, the value of λ is related to the half-life of the isotope.

Memorylessness: There are other distributions that also model waiting times, but the exponential distribution has the additional property that it is memoryless. Here's what this means in the context of Example 2: suppose that the probability that a taxi arrives within the first five minutes is p. If I wait five minutes and, in this case, no taxi arrives, then the probability that a taxi arrives within the next five minutes is still p. That is, my previous wait of 5 minutes has no impact on the length of my future wait!

By contrast, suppose I were to instead go to Kendall Square subway station and wait for the next inbound train. Since the trains are coordinated to follow a schedule (e.g., roughly 12 minutes between trains), if I wait five minutes without seeing a train then there is a far greater probability that a train will arrive in the next five minutes. In particular, waiting time for the subway is not memoryless, and a better model would be the uniform distribution on the range [0,12].

The memorylessness of the exponential distribution is analogous to the memorylessness of the (discrete) geometric distribution, where having flipped 5 tails in a row gives no information about the next 5 flips. Indeed, the exponential distribution is precisely the continuous counterpart of the geometric distribution, which models the waiting time for a discrete process to change state. More formally, memoryless means that the probability of waiting t more minutes is independent of the amount of time already waited. In symbols,

$$P(X > s + t \,|\, X > s) = P(X > t).$$

Proof of memorylessness: We know that

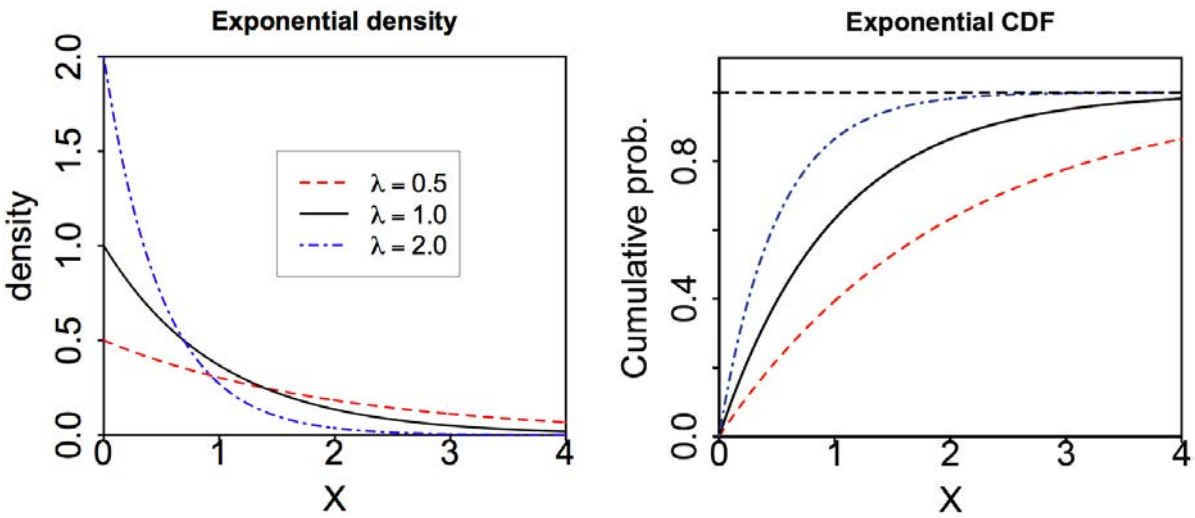
$$(X > s + t) \cap (X > s) = (X > s + t),$$

since the event 'waited at least s minutes' contains the event 'waited at least $s + t$ minutes'. Therefore the formula for conditional probability gives

$$P(X > s + t \,|\, X > s) = \frac{P(X > s + t)}{P(X > s)} = \frac{\mathrm{e}^{-\lambda(s+t)}}{\mathrm{e}^{-\lambda s}} = \mathrm{e}^{-\lambda t} = P(X > t).$$

The probability $P(X > s + t) = \mathrm{e}^{-\lambda(s+t)}$ is the formula for the right tail probability given above.

Graphs:



5 Normal distribution

In 1809, Carl Friedrich Gauss published a monograph introducing several notions that have become fundamental to statistics: the normal distribution, maximum likelihood estimation, and the method of least squares (we will cover all three in this course). For this reason, the normal distribution is also called the Gaussian distribution, and it is by far the most important continuous distribution.

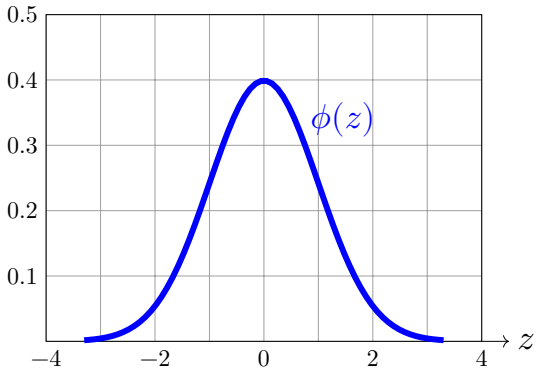
1. Parameters: μ, σ.
2. Range: $(-\infty, \infty)$.
3. Notation: Normal(μ, σ^2) or $N(\mu, \sigma^2)$.
4. Probability density function (pdf): $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$.
5. Cumulative distribution function (cdf): $F(x)$ has no formula, so use tables or software such as `pnorm` in R to compute $F(x)$.
6. Models: Measurement error, intelligence/ability, height, averages of lots of data.

The standard normal distribution $N(0, 1)$ has mean 0 and variance 1. We reserve Z for a standard normal random variable, $\phi(z) = \dfrac{1}{\sqrt{2\pi}} e^{-z^2/2}$ for the standard normal density, and $\Phi(z)$ for the standard normal distribution.

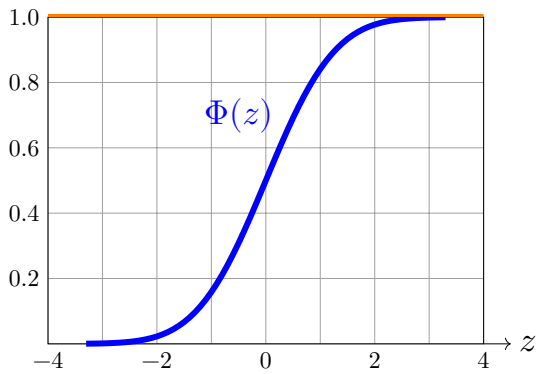
Note: we will define mean and variance for continuous random variables next time. They have the same interpretations as in the discrete case. As you might guess, the normal distribution $N(\mu, \sigma^2)$ has mean μ, variance σ^2, and standard deviation σ.

Here are some graphs of normal distributions. Note that they are shaped like a bell curve. Note also that as σ increases they become more spread out.

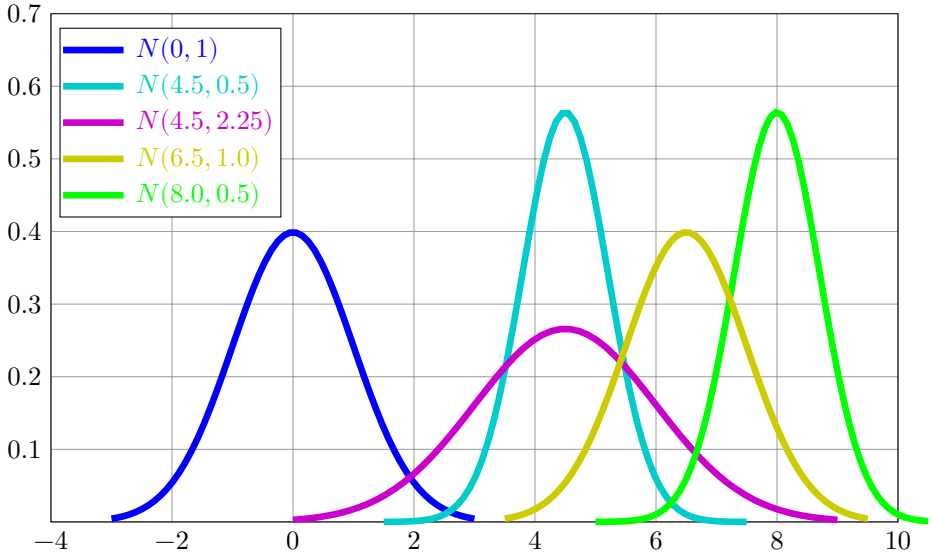
The bell curve: First we show the standard normal probability density and cumulative distribution functions. Below that is a selection of normal densities. Notice that the graph is centered on the mean and the bigger the variance the more spread out the curve.



Standard normal pdf



Standard normal cdf



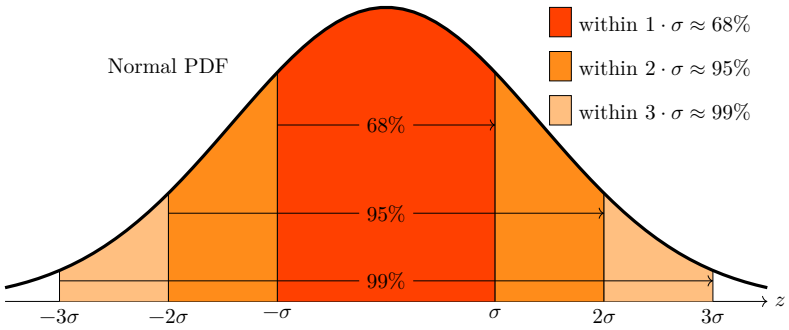
Notation note. In the figure above we use our notation $N(\mu, \sigma^2)$. So, for example, $N(8, 0.5)$ has variance 0.5 and standard deviation $\sigma = \sqrt{0.5} \approx 0.7071$.

5.1 Normal probabilities

To make approximations it is useful to remember the following rule of thumb for three approximate probabilities from the standard normal distribution:

$$P(-1 \le Z \le 1) \approx 0.68, \qquad P(-2 \le Z \le 2) \approx 0.95, \qquad P(-3 \le Z \le 3) \approx 0.99.$$

The figure below shows these probabilities as areas under the graph of the standard normal pdf $\phi(z)$.

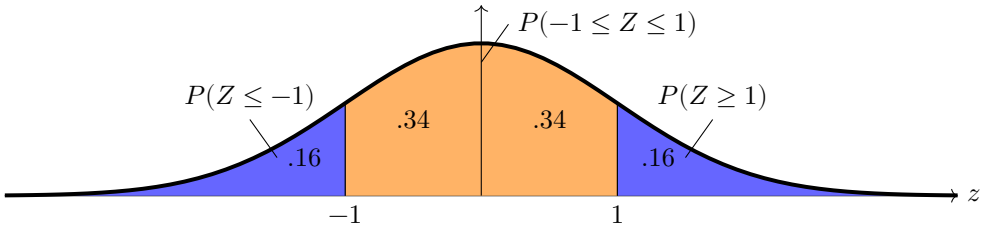


Symmetry calculations

We can use the symmetry of the standard normal distribution about $z = 0$ to make some calculations.

Example 4. The rule of thumb says $P(-1 \le Z \le 1) \approx 0.68$. Use this to estimate $\Phi(1)$.

Solution: $\Phi(1) = P(Z \le 1)$. In the figure, the two tails (in blue) have combined area $1 - 0.68 = 0.32$. By symmetry the left tail has area 0.16 (half of 0.32), so $P(Z \le 1) \approx 0.68 + 0.16 = 0.84$.



5.2 Using R to compute the standard normal cdf

```
# Use the R function pnorm(x,μ,σ) to compute F(x) for N(μ,σ^2)
pnorm(1,0,1)
[1] 0.8413447

pnorm(0,0,1)
[1] 0.5

pnorm(1,0,2)
[1] 0.6914625

pnorm(1,0,1) - pnorm(-1,0,1)
[1] 0.6826895

pnorm(5,0,5) - pnorm(-5,0,5)
[1] 0.6826895

# Of course z can be a vector of values
pnorm(c(-3,-2,-1,0,1,2,3),0,1)
[1] 0.001349898 0.022750132 0.158655254 0.500000000 0.841344746 0.977249868 0.998650102
```

Note: The R function pnorm(x, μ, σ) uses σ whereas our notation for the normal distribution $\mathrm{N}(\mu, \sigma^2)$ uses σ^2.

Here's a table of values with fewer decimal points of accuracy

z:	-2	-1	0	0.3	0.5	1	2	3
$\Phi(z)$:	0.0228	0.1587	0.5000	0.6179	0.6915	0.8413	0.9772	0.9987

Example 5. Use R to compute $P(-1.5 \le Z \le 2)$.

Solution: This is $\Phi(2) - \Phi(-1.5) =$ `pnorm(2,0,1) - pnorm(-1.5,0,1) = 0.91044`

6 Pareto and other distributions

In 18.05, we only have time to work with a few of the many wonderful distributions that are used in probability and statistics. We hope that after this course you will feel comfortable learning about new distributions and their properties when you need them. Wikipedia is often a great starting point.

The Pareto distribution is one common, beautiful distribution that we will not have time to cover in depth.

1. Parameters: $m > 0$ and $\alpha > 0$.
2. Range: $[m, \infty)$.
3. Notation: Pareto(m, α).
4. Density: $f(x) = \dfrac{\alpha\, m^\alpha}{x^{\alpha+1}}$.
5. Distribution: (easy integral)

$$F(x) = 1 - \frac{m^\alpha}{x^\alpha}, \text{ for } x \ge m$$

6. Tail distribution: $P(X > x) = m^\alpha / x^\alpha$, for $x \ge m$.
7. Models: The Pareto distribution models a power law, where the probability that an event occurs varies as a power of some attribute of the event. Many phenomena follow a power law, such as the size of meteors, income levels across a population, and population levels across cities. See Wikipedia for loads of examples:

   https://en.wikipedia.org/wiki/Pareto_distribution#Applications

Manipulating Continuous Random Variables
Class 5, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to find the pdf and cdf of a random variable defined in terms of a random variable with known pdf and cdf.

2 Transformations of Random Variables

We frequently transform a known random variable into a new one by applying a formula. For example we might look at $Y = aX + b$ or $Y = X^2$. In this section we will see how to find the probability density and cumulative distribution of Y from those of X.

For discrete random variables it was often possible do this by looking at probability tables. For continuous random variables we will need to use systematic algebraic techniques. We will see that transforming the pdf is just the change of variables ('u-substitution') from calculus. To transform the cdf directly we will rely on its definition as a probability.

Let's remind ourselves of the basics:

- The cdf of X is $F_X(x) = P(X \leq x)$.
- The pdf of X is related to F_X by $f_X(x) = F_X'(x)$.

2.1 Transforming the cdf

Example 1. Suppose X has range $[0, 2]$ and cdf $F_X(x) = x^2/4$. What is the range, pdf and cdf of $Y = X^2$?

Solution: The range is easy: $[0, 4]$.

To find the cdf we work systematically from the definition. For this example we will break it down into tiny steps, so you can see the thought process in detail.

Step 1. Use definition:

$$F_Y(y) = P(Y \leq y).$$

Step 2. Replace Y by its formula in X:

$$P(Y \leq y) = P(X^2 \leq y).$$

Step 3. Algebraically manipulate this to isolate the X:

$$P(X^2 \leq y) = P(X \leq \sqrt{y})$$

Step 4. Notice that this is exactly the definition of F_X:

$$P(X \leq \sqrt{y}) = F_X(\sqrt{y})$$

Step 5. Use the known formula for F_X:

$$F_X(\sqrt{y}) = (\sqrt{y})^2/4 = y/4.$$

Following the chain from step 1 to step 5 we have the cdf:

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(X \le \sqrt{y}) = F_X(\sqrt{y}) = y/4.$$

Finally, to find the pdf we can just differentiate the cdf:

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{1}{4}.$$

2.2 Transforming the pdf directly

An alternative way to find the pdf directly is by change of variables. We present this for completeness and for anyone who prefers it as a method. Our observation is that most people find the cdf easier to transform.

In calculus you learned the 'u'-substitution. We'll do a calculus example to remind you how this goes and then apply it to the pdf.

Example 2. Calculus example. Convert the integral $\int (x^2+1)^7\,dx$ into an integral in $u = x^2 + 1$.

Solution: We have to convert each part of the integral from x to u:

$$(x^2+1)^7 = u^7$$

$$du = 2x\,dx, \quad \text{therefore} \quad dx = \frac{du}{2x} = \frac{du}{2\sqrt{u-1}}$$

Now we replacing each piece in the integral we get

$$\int (x^2+1)^7\,dx = \int u^7 \frac{du}{2\sqrt{u-1}}.$$

Example 3. Find the pdf of Y in Example 1 directly using the method of 'u'-substitution. (In this case, 'u' will actually be 'y'.)

Solution: The trick is to remember that probability is given by an integral $\int f_X(x)dx$. We are given the change of variable $y = x^2$, so we change the integral from one in x to one in y.

$$y = x^2 \;\Rightarrow\; dy = 2x\,dx, \quad \text{therefore} \quad dx = \frac{dy}{2\sqrt{y}}.$$

We are given $F_X(x) = x^2/4$, so we can compute $f_X(x) = F_X'(x) = x/2$. Changing this to y we have

$$f_X(x) = \sqrt{y}/2.$$

Putting the two pieces together we have the transformation

$$f_X(x)\,dx = \frac{\sqrt{y}}{2}\frac{dy}{2\sqrt{y}} = \frac{1}{4}\,dy$$

Since this is a probability, the factor in front of dy is the probability density. That is, $f_Y(y) = 1/4$, exactly as in Example 1.

Here are a few more examples. We do them a little more quickly than the above examples.

Example 4. Let $X \sim \exp(\lambda)$, so $f_X(x) = \lambda e^{-\lambda x}$ on $[0, \infty]$. What is the probability density of $Y = X^2$?

Solution: We will do this using the change of variables for the pdf.

$$y = x^2 \;\Rightarrow\; dy = 2x\,dx, \qquad \text{therefore} \qquad dx = \tfrac{dy}{2\sqrt{y}}$$
$$f_X(x) = \lambda e^{-\lambda x} = \lambda e^{-\lambda\sqrt{y}}.$$

Combining these we get,

$$f_X(x)\,dx = \lambda e^{-\lambda\sqrt{y}}\,\frac{dy}{2\sqrt{y}} = f_Y(y)\,dy.$$

So we conclude that $\boxed{f_Y(y) = \dfrac{\lambda}{2\sqrt{y}}\,e^{-\lambda\sqrt{y}}}$.

Example 5. Redo the previous example using the cdf.

Solution: The cdf for the exponential random variable X is $F_X(x) = 1 - e^{-\lambda x}$. Therefore, for $Y = X^2$ we have

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(X \le \sqrt{y}) = F_X(\sqrt{y}) = 1 - e^{-\lambda\sqrt{y}}.$$

We have found $F_Y(y)$. If we wanted $f_Y(y)$ we could take the derivative. We would get the same answer as in the previous example.

Example 6. Assume $X \sim N(5, 3^2)$ then $Z = \dfrac{X-5}{3}$ is standard normal, i.e., $Z \sim N(0, 1)$.

Solution: Again using the change of variables and the formula for $f_X(x)$ we have

$$z = \frac{x-5}{3} \;\Rightarrow\; dz = \frac{dx}{3}, \qquad \text{therefore} \qquad dx = 3\,dz$$

For this example we will transform $f_X(x)\,dx$ in one line instead of two.

$$f_X(x)\,dx = \frac{1}{3\sqrt{2\pi}}e^{-(x-5)^2/(2\cdot 3^2)}\,dx = \frac{1}{3\sqrt{2\pi}}e^{-z^2/2}\,3\,dz = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}\,dz = f_Z(z)\,dz$$

Therefore $f_Z(z) = \dfrac{1}{\sqrt{2\pi}}\,e^{-z^2/2}$. Since this is exactly the density for $N(0, 1)$ we have shown that Z is standard normal.

This example shows an important general property of normal random variables which we state as a theorem.

Theorem. Standardization of normal random variables.
Assume $X \sim N(\mu, \sigma^2)$. Show that $Z = \dfrac{X-\mu}{\sigma}$ is standard normal, i.e., $Z \sim N(0, 1)$.

Proof. This is exactly the same computation as the previous example with μ replacing 5 and σ replacing 3. We show the computation without comment.

$$z = \frac{x-\mu}{\sigma} \;\Rightarrow\; dz = \frac{dx}{\sigma} \;\Rightarrow\; dx = \sigma\, dz$$

$$f_X(x)\, dx = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\cdot\sigma^2)}\, dx = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2}\, \sigma\, dz = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = f_Z(z)\, dz$$

Therefore $f_Z(z) = \dfrac{1}{\sqrt{2\pi}}\, e^{-z^2/2}$. This shows Z is standard normal.

We call the change from X to Z in this theorem standardization because it converts X from an arbitrary normal random variable to a standard normal variable.

Expectation, Variance and Standard Deviation for Continuous Random Variables
Class 6, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to compute and interpret expectation, variance, and standard deviation for continuous random variables.
2. Be able to compute and interpret quantiles for discrete and continuous random variables.

2 Introduction

So far we have looked at expected value, standard deviation, and variance for discrete random variables. These summary statistics have the same meaning for continuous random variables:

- The expected value $\mu = E[X]$ is a measure of location or central tendency.
- The standard deviation σ is a measure of the spread or scale.
- The variance $\sigma^2 = \mathrm{Var}(X)$ is the square of the standard deviation.

To move from discrete to continuous, we will simply replace the sums in the formulas by integrals. We will do this carefully and go through many examples in the following sections. In the last section, we will introduce another type of summary statistic, quantiles. You may already be familiar with the 0.5 quantile of a distribution, otherwise known as the median or 50^{th} percentile.

3 Expected value of a continuous random variable

Definition: Let X be a continuous random variable with range $[a, b]$ and probability density function $f(x)$. The expected value of X is defined by

$$E[X] = \int_a^b x f(x)\, dx.$$

Let's see how this compares with the formula for a discrete random variable:

$$E[X] = \sum_{i=1}^{n} x_i p(x_i).$$

The discrete formula says to take a weighted sum of the values x_i of X, where the weights are the probabilities $p(x_i)$. Recall that $f(x)$ is a probability **density**. Its units are

prob/(unit of X). So $f(x)\,dx$ represents the probability that X is in an infinitesimal range of width dx around x. Thus we can interpret the formula for $E[X]$ as a weighted integral of the values x of X, where the weights are the probabilities $f(x)\,dx$.

As before, the expected value is also called the mean or average.

3.1 Examples

Let's go through several example computations. Where the solution requires an integration technique, we push the computation of the integral to the appendix.

Example 1. Let $X \sim \text{uniform}(0,1)$. Find $E[X]$.

Solution: X has range $[0,1]$ and density $f(x) = 1$. Therefore,

$$E[X] = \int_0^1 x\,dx = \frac{x^2}{2}\Big|_0^1 = \boxed{\frac{1}{2}}.$$

Not surprisingly the mean is at the midpoint of the range.

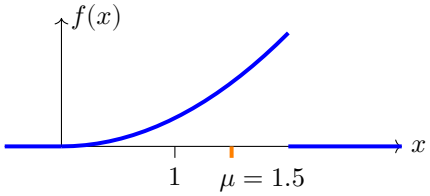
Example 2. Let X have range $[0,2]$ and density $\frac{3}{8}x^2$. Find $E[X]$.

Solution:

$$E[X] = \int_0^2 x f(x)\,dx = \int_0^2 \frac{3}{8}x^3\,dx = \frac{3x^4}{32}\Big|_0^2 = \boxed{\frac{3}{2}}.$$

Does it make sense that this X has mean is in the right half of its range?

Solution: Yes. Since the probability density increases as x increases over the range, the average value of x should be in the right half of the range.

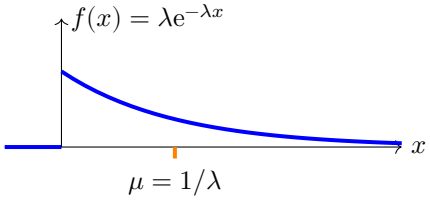


μ is "pulled" to the right of the midpoint 1 because there is more mass to the right.

Example 3. Let $X \sim \exp(\lambda)$. Find $E[X]$.

Solution: The range of X is $[0,\infty)$ and its pdf is $f(x) = \lambda e^{-\lambda x}$. So (details in appendix)

$$E[X] = \int_0^\infty x\lambda e^{-\lambda x}\,dx = -xe^{-\lambda x} - \frac{e^{-\lambda x}}{\lambda}\Big|_0^\infty = \boxed{\frac{1}{\lambda}}.$$

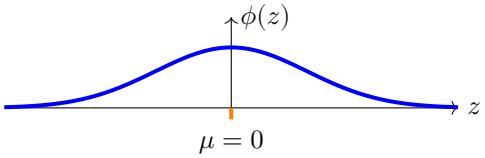


Mean of an exponential random variable

Example 4. Let $Z \sim \mathrm{N}(0,1)$. Find $E[Z]$.

Solution: The range of Z is $(-\infty, \infty)$ and its pdf is $\phi(z) = \dfrac{1}{\sqrt{2\pi}} e^{-z^2/2}$. So (details in appendix)

$$E[Z] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z e^{-z^2/2}\, dz = -\frac{1}{\sqrt{2\pi}} e^{-z^2/2}\Big|_{-\infty}^{\infty} = \boxed{0}.$$



The standard normal distribution is symmetric and has mean 0.

3.2 Properties of $E[X]$

The properties of $E[X]$ for continuous random variables are the same as for discrete ones:

1. If X and Y are random variables on a sample space Ω then

$$E[X+Y] = E[X] + E[Y]. \qquad \text{(linearity I)}$$

2. If a and b are constants then

$$E[aX+b] = aE[X] + b. \qquad \text{(linearity II)}$$

Example 5. In this example we verify that for $X \sim \mathrm{N}(\mu, \sigma)$ we have $E[X] = \mu$.

Solution: Example (4) showed that for standard normal Z, $E[Z] = 0$. We could mimic the calculation there to show that $E[X] = \mu$. Instead we will use the linearity properties of $E[X]$. In the class 5 notes on manipulating random variables we showed that if $X \sim \mathrm{N}(\mu, \sigma^2)$ is a normal random variable we can **standardize** it:

$$Z = \frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1).$$

Inverting this formula we have $X = \sigma\, Z + \mu$. The linearity of expected value now gives

$$E[X] = E[\sigma\, Z + \mu] = \sigma\, E[Z] + \mu = \mu$$

3.3 Expectation of Functions of X

This works exactly the same as the discrete case. if $h(x)$ is a function then $Y = h(X)$ is a random variable and

$$E[Y] = E[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)\, dx.$$

Example 6. Let $X \sim \exp(\lambda)$. Find $E[X^2]$.

Solution: Using integration by parts we have

$$E[X^2] = \int_0^{\infty} x^2 \lambda e^{-\lambda x}\, dx = \left[-x^2 e^{-\lambda x} - \frac{2x}{\lambda} e^{-\lambda x} - \frac{2}{\lambda^2} e^{-\lambda x}\right]_0^{\infty} = \boxed{\frac{2}{\lambda^2}}.$$

4 Variance

Now that we've defined expectation for continuous random variables, the definition of variance is identical to that of discrete random variables.

Definition: Let X be a continuous random variable with mean μ. The variance of X is

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

4.1 Properties of Variance

These are exactly the same as in the discrete case.

1. If X and Y are independent then $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.
2. For constants a and b, $\text{Var}(aX+b) = a^2\text{Var}(X)$.
3. **Theorem**: $\text{Var}(X) = E[X^2] - E[X]^2 = E[X^2] - \mu^2$.

For Property 1, note carefully the requirement that X and Y are independent.

Property 3 gives a formula for $\text{Var}(X)$ that is often easier to use in hand calculations. The proofs of properties 2 and 3 are essentially identical to those in the discrete case. We will not give them here.

Example 7. Let $X \sim \text{uniform}(0,1)$. Find $\text{Var}(X)$ and σ_X.

Solution: In Example 1 we found $\mu = 1/2$. Next we compute

$$\text{Var}(X) = E[(X-\mu)^2] = \int_0^1 (x-1/2)^2\,dx = \boxed{\frac{1}{12}}.$$

Example 8. Let $X \sim \exp(\lambda)$. Find $\text{Var}(X)$ and σ_X.

Solution: In Examples 3 and 6 we computed

$$E[X] = \int_0^\infty x\lambda e^{-\lambda x}\,dx = \frac{1}{\lambda} \quad \text{and} \quad E[X^2] = \int_0^\infty x^2\lambda e^{-\lambda x}\,dx = \frac{2}{\lambda^2}.$$

So by Property 3,

$$\text{Var}(X) = E[X^2] - E[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \quad \text{and} \quad \sigma_X = \frac{1}{\lambda}.$$

We could have skipped Property 3 and computed this directly from $\text{Var}(X) = \int_0^\infty (x-1/\lambda)^2\lambda e^{-\lambda x}\,dx$.

Example 9. Let $Z \sim \text{N}(0,1)$. Show $\text{Var}(Z) = 1$.

Note: The notation for normal variables is $X \sim \text{N}(\mu, \sigma^2)$. This is certainly suggestive, but as mathematicians we need to prove that $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$. Above we showed $E[X] = \mu$. This example shows that $\text{Var}(Z) = 1$, just as the notation suggests. In the next example we'll show $\text{Var}(X) = \sigma^2$.

Solution: Since $E[Z] = 0$, we have

$$\text{Var}(Z) = E[Z^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2}\,dz.$$

(using integration by parts with $u = z$, $v' = ze^{-z^2/2} \Rightarrow u' = 1$, $v = -e^{-z^2/2}$)

$$= \frac{1}{\sqrt{2\pi}}\left(-ze^{-z^2/2}\Big|_{-\infty}^{\infty}\right) + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-z^2/2}\,dz.$$

The first term equals 0 because the exponential goes to zero much faster than z grows at both $\pm\infty$. The second term equals 1 because it is exactly the total probability integral of the pdf $\varphi(z)$ for $N(0,1)$. So $\mathrm{Var}(X) = 1$.

Example 10. Let $X \sim N(\mu, \sigma^2)$. Show $\mathrm{Var}(X) = \sigma^2$.

Solution: This is an exercise in change of variables. Letting $z = (x-\mu)/\sigma$, we have

$$\mathrm{Var}(X) = E[(X-\mu)^2] = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty}(x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2}\,dx$$
$$= \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^2 e^{-z^2/2}\,dz = \sigma^2.$$

The integral in the last line is the same one we computed for $\mathrm{Var}(Z)$.

5 Quantiles

Definition: The median of X is the value x for which $P(X \le x) = 0.5$, i.e. the value of x such that $P(X \le x) = P(X \ge x)$. In other words, X has equal probability of being above or below the median, and each probability is therefore $1/2$. In terms of the cdf $F(x) = P(X \le x)$, we can equivalently define the median as the value x satisfying $F(x) = 0.5$.

Think: What is the median of Z?

Solution: By symmetry, the median is 0.

Example 11. Find the median of $X \sim \exp(\lambda)$.

Solution: The cdf of X is $F(x) = 1 - e^{-\lambda x}$. So the median is the value of x for which $F(x) = 1 - e^{-\lambda x} = 0.5$.. Solving for x we find: $\boxed{x = (\ln 2)/\lambda}$.

Think: In this case the median does not equal the mean of $\mu = 1/\lambda$. Based on the graph of the pdf of X can you argue why the median is to the left of the mean.

Definition: The p^{th} quantile of X is the value q_p such that $P(X \le q_p) = p$.

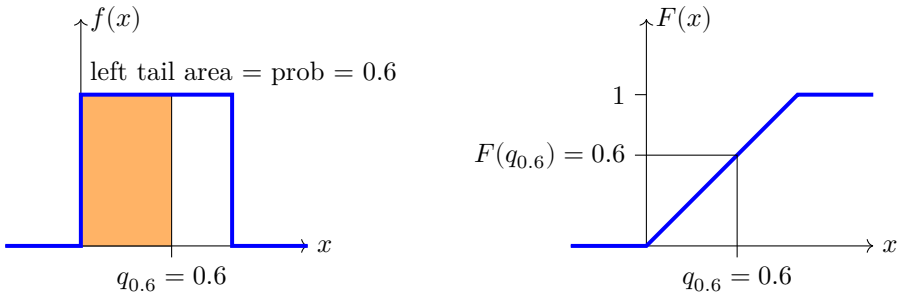
Notes. 1. In this notation the median is $q_{0.5}$.

2. We will usually write this in terms of the cdf: $F(q_p) = p$.

With respect to the pdf $f(x)$, the quantile q_p is the value such that there is an area of p to the left of q_p and an area of $1-p$ to the right of q_p. In the examples below, note how we can represent the quantile graphically using either area of the pdf or height of the cdf.

Example 12. Find the 0.6 quantile for $X \sim U(0,1)$.

Solution: The cdf for X is $F(x) = x$ on the range [0,1]. So $q_{0.6} = 0.6$.

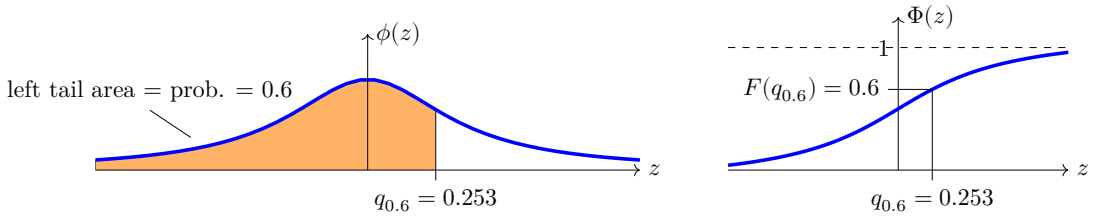


$q_{0.6}$: left tail area $= 0.6 \Leftrightarrow F(q_{0.6}) = 0.6$

Example 13. Find the 0.6 quantile of the standard normal distribution.

Solution: We don't have a formula for the cdf, so we use the R 'quantile function' `qnorm`.

$$q_{0.6} = \texttt{qnorm(0.6, 0, 1)} = 0.25335$$



$q_{0.6}$: left tail area $= 0.6 \Leftrightarrow F(q_{0.6}) = 0.6$

Quantiles give a useful measure of **location** for a random variable. We will use them more in coming lectures.

5.1 Percentiles, deciles, quartiles

For convenience, quantiles are often described in terms of percentiles, deciles or quartiles. The 60^{th} percentile is the same as the 0.6 quantile. For example you are in the 60^{th} percentile for height if you are taller than 60 percent of the population, i.e. the **probability** that you are taller than a randomly chosen person is 60 percent.

Likewise, deciles represent steps of 1/10. The third decile is the 0.3 quantile. Quartiles are in steps of 1/4. The third quartile is the 0.75 quantile and the 75^{th} percentile.

6 Appendix: Integral Computation Details

From Example 3 Let $X \sim \exp(\lambda)$. Find $E[X]$.

The range of X is $[0, \infty)$ and its pdf is $f(x) = \lambda e^{-\lambda x}$. Therefore

$$E[X] = \int_0^\infty x f(x)\, dx = \int_0^\infty \lambda x e^{-\lambda x}\, dx$$

(using integration by parts with $u = x$, $v' = \lambda e^{-\lambda x} \Rightarrow u' = 1$, $v = -e^{-\lambda x}$)

$$= -xe^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x}\,dx$$
$$= 0 - \frac{e^{-\lambda x}}{\lambda}\Big|_0^\infty = \frac{1}{\lambda}.$$

We used the fact that $xe^{-\lambda x}$ and $e^{-\lambda x}$ go to 0 as $x \to \infty$.

From Example 4 Let $Z \sim N(0,1)$. Find $E[Z]$.

The range of Z is $(-\infty, \infty)$ and its pdf is $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$. By symmetry the mean must be 0. The only mathematically tricky part is to show that the integral converges, i.e. that the mean exists at all (some random variable do not have means, but we will not encounter this very often.) For completeness we include the argument, though this is not something we will ask you to do. We first compute the integral from 0 to ∞:

$$\int_0^\infty z\phi(z)\,dz = \frac{1}{\sqrt{2\pi}} \int_0^\infty ze^{-z^2/2}\,dz.$$

The u-substitution $u = z^2/2$ gives $du = z\,dz$. So the integral becomes

$$\frac{1}{\sqrt{2\pi}} \int_0^\infty ze^{-z^2/2}\,dz. = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-u}\,du = -e^{-u}\Big|_0^\infty = 1$$

Similarly, $\displaystyle\int_{-\infty}^0 z\phi(z)\,dz = -1$. Adding the two pieces together gives $E[Z] = 0$.

From Example 6 Let $X \sim \exp(\lambda)$. Find $E[X^2]$.

$$E[X^2] = \int_0^\infty x^2 f(x)\,dx = \int_0^\infty \lambda x^2 e^{-\lambda x}\,dx$$

(using integration by parts with $u = x^2$, $v' = \lambda e^{-\lambda x} \Rightarrow u' = 2x$, $v = -e^{-\lambda x}$)

$$= -x^2 e^{-\lambda x}\Big|_0^\infty + \int_0^\infty 2xe^{-\lambda x}\,dx$$

(the first term is 0, for the second term use integration by parts: $u = 2x$, $v' = e^{-\lambda x} \Rightarrow u' = 2$, $v = -\frac{e^{-\lambda x}}{\lambda}$)

$$= -2x\frac{e^{-\lambda x}}{\lambda}\Big|_0^\infty + \int_0^\infty \frac{e^{-\lambda x}}{\lambda}\,dx$$
$$= 0 - 2\frac{e^{-\lambda x}}{\lambda^2}\Big|_0^\infty = \frac{2}{\lambda^2}.$$

Central Limit Theorem and the Law of Large Numbers
Class 6, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Understand the statement of the law of large numbers.
2. Understand the statement of the central limit theorem.
3. Be able to use the central limit theorem to approximate probabilities of averages and sums of independent identically-distributed random variables.

2 Introduction

We all understand intuitively that the average of many measurements of the same unknown quantity tends to give a better estimate than a single measurement. Intuitively, this is because the random error of each measurement cancels out in the average. In these notes we will make this intuition precise in two ways: the law of large numbers (LoLN) and the central limit theorem (CLT).

Briefly, both the law of large numbers and central limit theorem are about many independent samples from same distribution. The LoLN tells us two things:

1. The average of many independent samples is (with high probability) close to the mean of the underlying distribution.
2. The density histogram of many independent samples is (with high probability) close to the graph of the density of the underlying distribution.

To be absolutely correct mathematically we need to make these statements more precise, but as stated they are a good way to think about the law of large numbers.

The central limit theorem says that the sum or average of many independent copies of a random variable is approximately a normal random variable. The CLT goes on to give precise values for the mean and standard deviation of the normal variable.

These are both remarkable facts. Perhaps just as remarkable is the fact that often in practice n does not have to be all that large.

2.1 There is more to experimentation than mathematics

The mathematics of the LoLN says that the average of a lot of independent samples from a random variable will almost certainly approach the mean of the variable. The mathematics cannot tell us if the tool or experiment is producing data worth averaging. For example, if the measuring device is defective or poorly calibrated then the average of many measurements will be a highly accurate estimate of the wrong thing! This is an example of

systematic error or sampling bias, as opposed to the random error controlled by the law of large numbers.

3 The law of large numbers

Suppose X_1, X_2, ..., X_n are independent random variables with the same underlying distribution. In this case, we say that the X_i are independent and identically-distributed, or i.i.d. In particular, the X_i all have the same mean μ and standard deviation σ.

Let $\overline{X}_n$ be the average of $X_1, \dots, X_n$:

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Note that $\overline{X}_n$ is itself a random variable. The law of large numbers and central limit theorem tell us about the value and distribution of $\overline{X}_n$, respectively.

LoLN: As n grows, the probability that $\overline{X}_n$ is close to μ goes to 1.

CLT: As n grows, the distribution of $\overline{X}_n$ converges to the normal distribution $N(\mu, \sigma^2/n)$.

Before giving a more formal statement of the LoLN, let's unpack its meaning through a concrete example (we'll return to the CLT later on).

Example 1. Averages of Bernoulli random variables
Suppose each X_i is an independent flip of a fair coin, so $X_i \sim$ Bernoulli(0.5) and $\mu = 0.5$. Then $\overline{X}_n$ is the proportion of heads in n flips, and we expect that this proportion is close to 0.5 for large n. Randomness being what it is, this is not guaranteed; for example we could get 1000 heads in 1000 flips, though the probability of this occurring is very small.

So our intuition translates to: with high probability the sample average $\overline{X}_n$ is close to the mean 0.5 for large n. We'll demonstrate by doing some calculations in R. You can find the code used for 'class 6 prep' in the usual place on our website.

To start we'll look at the probability of being within 0.1 of the mean. We can express this probability as

$$P(|\overline{X}_n - 0.5| < 0.1) \quad \text{or equivalently} \quad P(0.4 \le \overline{X}_n \le 0.6)$$

The law of large numbers says that this probability goes to 1 as the number of flips n gets large. Our R code produces the following values for $P(0.4 \le \overline{X}_n \le 0.6)$.

$n = 10$:	`pbinom(6, 10, 0.5) - pbinom(3, 10, 0.5)`	$= 0.65625$
$n = 50$:	`pbinom(30, 50, 0.5) - pbinom(19, 50, 0.5)`	$= 0.8810795$
$n = 100$:	`pbinom(60, 100, 0.5) - pbinom(39, 100, 0.5)`	$= 0.9647998$
$n = 500$:	`pbinom(300, 500, 0.5) - pbinom(199, 500, 0.5)`	$= 0.9999941$
$n = 1000$:	`pbinom(600, 1000, 0.5) - pbinom(399, 1000, 0.5)`	$= 1$

As predicted by the LoLN the probability goes to 1 as n grows.

We redo the computations to see the probability of being within 0.01 of the mean. Our R code produces the following values for $P(0.49 \le \overline{X}_n \le 0.51)$.

$n = 10$:	`pbinom(5, 10, 0.5) - pbinom(4, 10, 0.5)`	$= 0.2460937$
$n = 100$:	`pbinom(51, 100, 0.5) - pbinom(48, 100, 0.5)`	$= 0.2356466$
$n = 1000$:	`pbinom(510, 1000, 0.5) - pbinom(489, 1000, 0.5)`	$= 0.49334$
$n = 10000$:	`pbinom(5100, 10000, 0.5) - pbinom(4899, 10000, 0.5)`	$= 0.9555742$

Again we see the probability of being close to the mean going to 1 as n grows. Since 0.01 is smaller than 0.1 it takes larger values of n to raise the probability to near 1.

This convergence of the probability to 1 is the LoLN in action! Whenever you're confused, it will help you to keep this example in mind. So we see that the LoLN says that with high probability the average of a large number of independent trials from the same distribution will be very close to the underlying mean of the distribution. Now we're ready for the formal statement.

3.1 Formal statement of the law of large numbers

Theorem (Law of Large Numbers): Suppose X_1, X_2, ..., X_n, ... are i.i.d. random variables with mean μ. For each n, let $\overline{X}_n$ be the average of the first n variables. Then for any $a > 0$, we have

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| < a) \;=\; 1.$$

This says precisely that as n increases the probability of being within a of the mean goes to 1. Think of a as a small tolerance of error from the true mean μ.

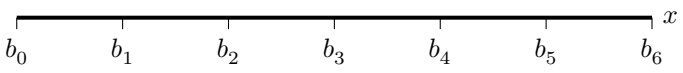
Looking back at Example 1, we see that for tosses of a fair coin: If we choose the number of tosses $n = 500$, then with probability $p = 0.99999$, the experimental frequency of heads $\overline{X}_n$ will be within $a = 0.1$ of 0.5. In words, this tells us that, on average, only 1 in 100,000 experiments will produce an experimental frequency outside this range. If we decrease the tolerance a and/or increase the probability p, then n will need to be larger.

4 Histograms

We can summarize multiple samples $x_1, \dots, x_n$ of a random variable in a histogram. Here we want to carefully construct histograms so that they resemble the area under the pdf. We will then see how the LoLN applies to histograms.

The step-by-step instructions for constructing a density or frequency histogram are as follows.

1. Pick an interval of the real line and divide it into m intervals, with endpoints b_0, b_1, ..., b_m. Usually these are equally sized, so let's assume this to start.



Six equally-sized bins

Each of the intervals is called a bin. For example, in the figure above the first bin is $[b_0, b_1]$ and the last bin is $[b_5, b_6]$. Each bin has a bin width, e.g. $b_1 - b_0$ is the first bin width. Usually the bins all have the same width, called the bin width of the histogram.

2. Place each x_i into the bin that contains its value. If x_i lies on the boundary of two bins, we'll put it in the left bin (this is the R default, though it can be changed).

3. To draw a frequency histogram: put a vertical bar above each bin. The height of the bar should equal the number of x_i in the bin.

4. To draw a density histogram: put a vertical bar above each bin. The area of the bar should equal the fraction of all data points that lie in the bin.

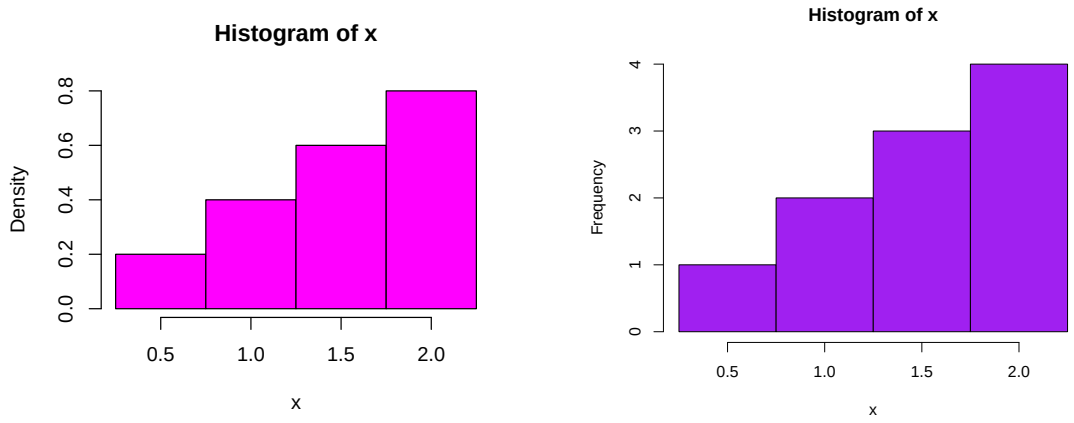
Notes:
1. When all the bins have the same width, the frequency histogram bars have area proportional to the count. So the density histogram results from simply by dividing the height of each bar by the total area of the frequency histogram. Ignoring the vertical scale, the two histograms look identical.

2. Caution: if the bin widths differ, the frequency and density histograms may look very different. There is an example of this below. Don't let anyone fool you by manipulating bin widths to produce a histogram that suits their mischievous purposes!

In 18.05, we'll stick with equally-sized bins. In general, we prefer the density histogram since its vertical scale is the same as that of the pdf.

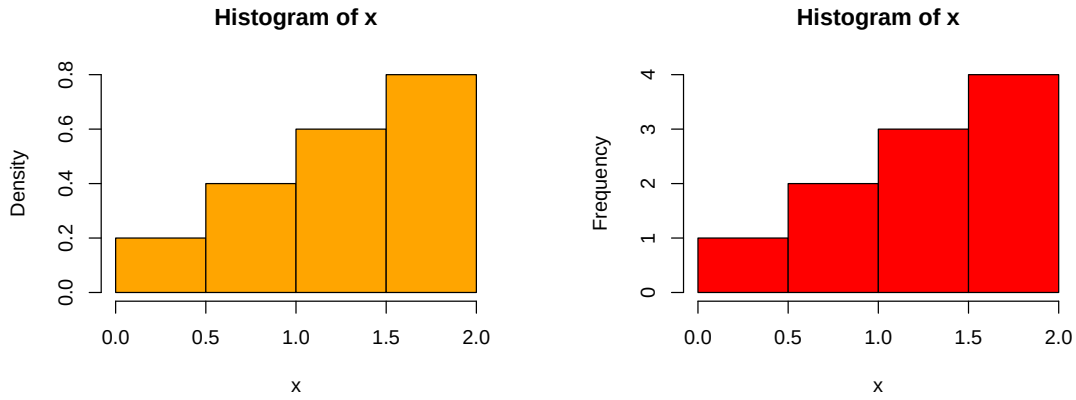
Examples. Here are some examples of histograms, all with the data [0.5,1,1,1.5,1.5,1.5,2,2,2,2]. The R code that drew them is in the file 'class6-prep-b.r'. You can find it in the usual place on our website.

1. Here the density and frequency plots look the same but have different vertical scales.



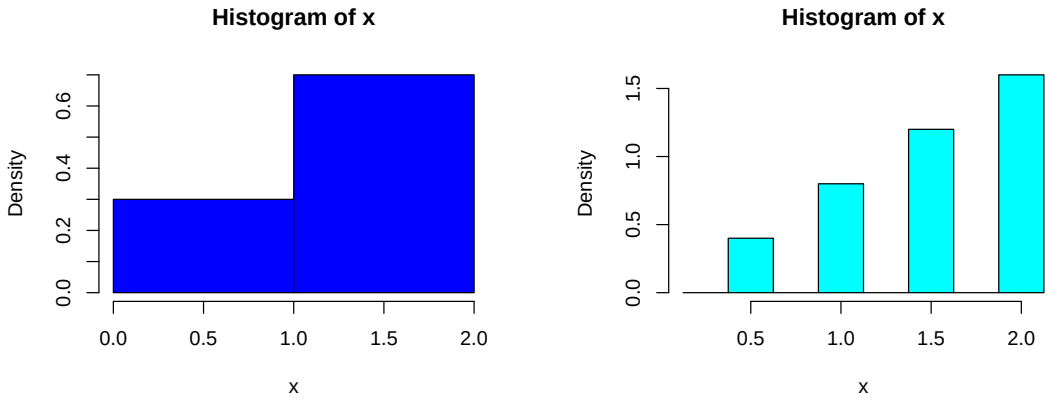
Bins centered at 0.5, 1, 1.5, 2, i.e. width 0.5, bounds at 0.25, 0.75, 1.25, 1.75, 2.25.

2. Note the values are all on the bin boundaries and are put into the left-hand bin. That is, the bins are right-closed, e.g the first bin is for values in the right-closed interval $(0, 0.5]$.



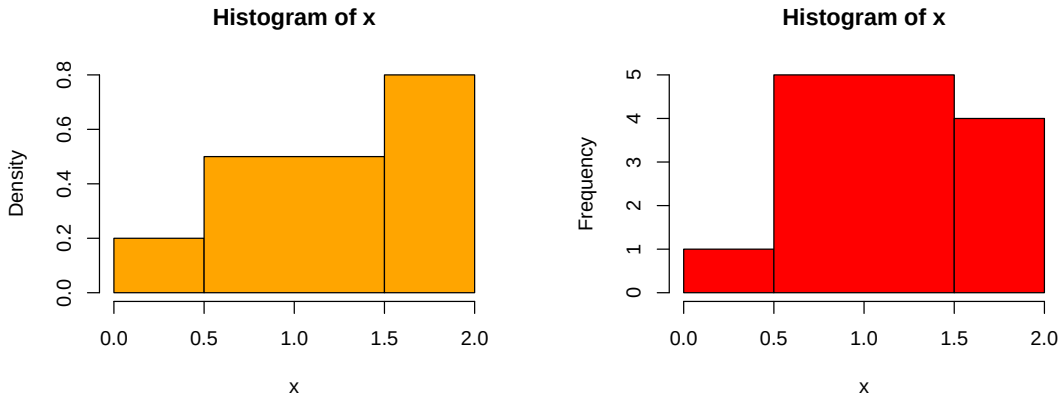
Bin bounds at 0, 0.5, 1, 1.5, 2.

3. Here we show density histograms based on different bin widths. Note that the scale keeps the total area equal to 1. The gaps are bins with zero counts.



Left: wide bins; Right: narrow bins.

4. Here we use unqual bin widths, so the density and frequency histograms look different



Don't be fooled! These are based on the same data.

The density histogram is the better choice with unequal bin widths. In fact, R will complain

if you try to make a frequency histogram with unequal bin widths. Compare the frequency histogram with unequal bin widths with all the other histograms we drew for this data. It clearly looks different. What happened is that by combining the data in bins $(0.5, 1]$ and $(1, 1.5]$ into one bin $(0.5, 1.5)$ we effectively made the height of both smaller bins greater.

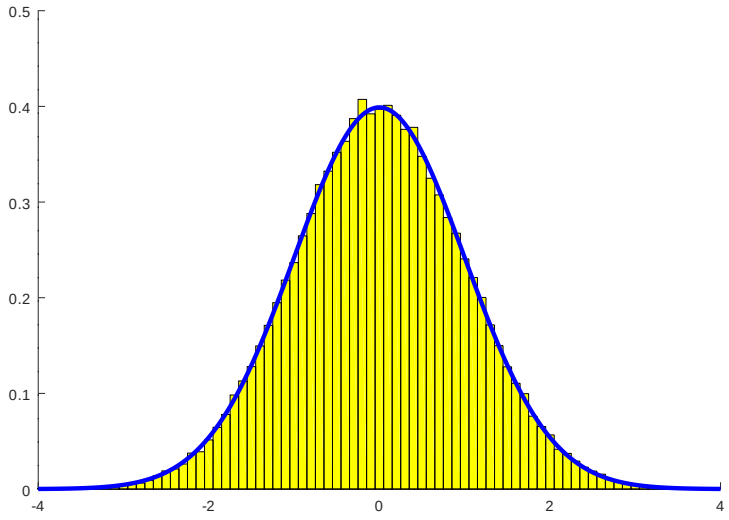
The reason the density histogram is nice is discussed in the next section.

4.1 The law of large numbers and histograms

The law of large number has an important consequence for density histograms.

LoLN for histograms: With high probability the density histogram of a large number of samples from a distribution is a good approximation of the graph of the underlying pdf $f(x)$ over the range of the histogram.

Let's illustrate this by generating a density histogram with bin width 0.1 from 100000 draws from a standard normal distribution. As you can see, the density histogram very closely tracks the graph of the standard normal pdf $\phi(z)$.



Density histogram of 10000 draws from a standard normal distribution, with $\phi(z)$ in blue.

5 The Central Limit Theorem

We now prepare for the statement of the CLT.

5.1 Standardization

Given a random variable X with mean μ and standard deviation σ, we define its standardization of X as the new random variable

$$Z = \frac{X - \mu}{\sigma}.$$

Note that Z has mean 0 and standard deviation 1. Note also that if X has a normal distribution, then the standardization of X is the standard normal distribution Z with mean 0 and variance 1. This explains the term 'standardization' and the notation of Z above.

5.2 Statement of the Central Limit Theorem

Suppose X_1, X_2, ..., X_n, ...are i.i.d. random variables each having mean μ and standard deviation σ. For each n, let S_n denote the sum and let $\overline{X}_n$ be the average of $X_1, \dots, X_n$.

$$S_n = X_1 + X_2 + \dots + X_n = \sum_{i=1}^{n} X_i$$
$$\overline{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{S_n}{n}.$$

The properties of mean and variance show

$$\begin{array}{llllll} E[S_n] & = n\mu, & \mathrm{Var}(S_n) & = n\sigma^2, & \sigma_{S_n} & = \sqrt{n}\,\sigma \\ E[\overline{X}_n] & = \mu, & \mathrm{Var}(\overline{X}_n) & = \dfrac{\sigma^2}{n}, & \sigma_{\overline{X}_n} & = \dfrac{\sigma}{\sqrt{n}}. \end{array}$$

Since they are multiples of each other, S_n and $\overline{X}_n$ have the same standardization

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

Central Limit Theorem: For large n,

$$\overline{X}_n \approx \mathrm{N}(\mu, \sigma^2/n), \qquad S_n \approx \mathrm{N}(n\mu, n\sigma^2), \qquad Z_n \approx \mathrm{N}(0,1).$$

Notes: 1. In words: $\overline{X}_n$ is approximately a normal distribution with the same mean as X but a smaller variance.
2. S_n is approximately normal.
3. Standardized $\overline{X}_n$ and S_n are approximately standard normal.

The central limit theorem allows us to approximate a sum or average of i.i.d random variables by a normal random variable. This is extremely useful because it is usually easy to do computations with the normal distribution.

A precise statement of the CLT is that the cdf's of Z_n converge to $\Phi(z)$:

$$\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z).$$

The proof of the Central Limit Theorem is more technical than we want to get in 18.05. It is accessible to anyone with a decent calculus background.

5.3 Standard Normal Probabilities

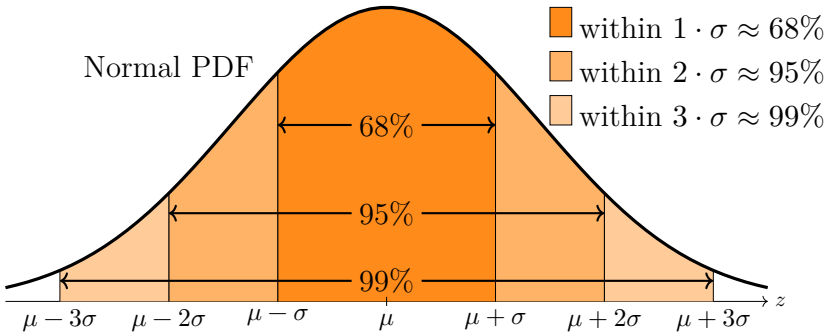
To apply the CLT, we will want to have some normal probabilities at our fingertips. The following probabilities appeared in Class 5. Let $Z \sim N(0,1)$, a standard normal random variable. Then with rounding we have:

1. $P(|Z| < 1) \approx 0.68$
2. $P(|Z| < 2) \approx 0.95$; more precisely $P(|Z| < 1.96) \approx 0.95$.
3. $P(|Z| < 3) \approx 0.997$

These numbers are easily computed in R using `pnorm`. However, they are well worth remembering as rules of thumb. You should think of them as:

1. The probability that a normal random variable is within 1 standard deviation of its mean is 0.68.
2. The probability that a normal random variable is within 2 standard deviations of its mean is 0.95.
3. The probability that a normal random variable is within 3 standard deviations of its mean is 0.997.

This is shown graphically in the following figure.



Claim: From these numbers we can derive:

1. $P(Z < 1) \approx 0.84$
2. $P(Z < 2) \approx 0.977$
3. $P(Z < 3) \approx 0.999$

Proof: We know $P(|Z| < 1) = 0.68$. The remaining probability of 0.32 is in the two regions $Z > 1$ and $Z < -1$. These regions are referred to as the right-hand tail and the left-hand tail respectively. By symmetry each tail has area 0.16. Thus,

$$P(Z < 1) = P(|Z| < 1) + P(\text{left-hand tail}) = 0.84$$

The other two cases are handled similarly.

5.4 Applications of the CLT

Example 2. Flip a fair coin 100 times. Estimate the probability of more than 55 heads.

Solution: Let X_j be the result of the j^{th} flip, so $X_j = 1$ for heads and $X_j = 0$ for tails. The total number of heads is

$$S = X_1 + X_2 + ... + X_{100}.$$

We know $E[X_j] = 0.5$ and $\text{Var}(X_j) = 1/4$. Since $n = 100$, we have

$$E[S] = 50, \quad \text{Var}(S) = 25 \quad \text{and} \quad \sigma_S = 5.$$

The central limit theorem says that the standardization of S is approximately $\text{N}(0, 1)$. The question asks for $P(S > 55)$. Standardizing and using the CLT we get

$$P(S > 55) = P\left(\frac{S - 50}{5} > \frac{55 - 50}{5}\right) \approx P(Z > 1) = 0.16.$$

Here Z is a standard normal random variable and $P(Z > 1) = 1 - P(Z < 1) \approx 0.16$.

Example 3. Estimate the probability of more than 220 heads in 400 flips.

Solution: This is nearly identical to the previous example. Now $\mu_S = 200$ and $\sigma_S = 10$ and we want $P(S > 220)$. Standardizing and using the CLT we get:

$$P(S > 220) = P\left(\frac{S - \mu_S}{\sigma_S} > \frac{220 - 200}{10}\right) \approx P(Z > 2) = 0.025.$$

Again, $Z \sim \text{N}(0, 1)$ and the rules of thumb show $P(Z > 2) = 0.025$.

Note: Even though $55/100 = 220/400$, the probability of more than 55 heads in 100 flips is larger than the probability of more than 220 heads in 400 flips. This is due to the LoLN and the larger value of n in the latter case.

Example 4. Estimate the probability of between 40 and 60 heads in 100 flips.

Solution: As in the first example, $E[S] = 50$, $\text{Var}(S) = 25$ and $\sigma_S = 5$. So

$$P(40 \le S \le 60) = P\left(\frac{40 - 50}{5} \le \frac{S - 50}{5} \le \frac{60 - 50}{5}\right) \approx P\left(-2 \le Z \le 2\right)$$

We can compute the right-hand side using our rule of thumb. For a more accurate answer we use R:

$$\texttt{pnorm(2) - pnorm(-2)} = 0.954\,...$$

Recall that in Section 3 we used the binomial distribution to compute an answer of 0.965…. So our approximate answer using CLT is off by about 1%.

Think: Would you expect the CLT method to give a better or worse approximation of $P(200 < S < 300)$ with $n = 500$?

We encourage you to check your answer using R.

Example 5. Polling. When taking a political poll the results are often reported as a number with a margin of error. For example 52% ± 3% favor candidate A. The rule of thumb is that if you poll n people then the margin of error is $\pm 1/\sqrt{n}$. We will now see exactly what this means and that it is an application of the central limit theorem.

Suppose there are 2 candidates A and B. Suppose further that the fraction of the population who prefer A is p_0. That is, if you ask a random person who they prefer then the probability they'll answer A is p_o

To run the poll a pollster selects n people at random and asks 'Do you support candidate A or candidate B?' Thus we can view the poll as a sequence of n independent Bernoulli(p_0) trials, X_1, X_2, ..., X_n, where X_i is 1 if the i$^{\text{th}}$ person prefers A and 0 if they prefer B. The fraction of people polled that prefer A is just the average $\overline{X}$.

We know that each $X_i \sim$ Bernoulli(p_0) so,

$$E[X_i] = p_0 \quad \text{and } \sigma_{X_i} = \sqrt{p_0(1-p_0)}.$$

Therefore, the central limit theorem tells us that

$$\overline{X} \approx \mathrm{N}(p_0, \sigma^2/n), \qquad \text{where } \sigma = \sqrt{p_0(1-p_0)}.$$

In a normal distribution 95% of the probability is within 2 standard deviations of the mean. This means that in 95% of polls of n people the sample mean $\overline{X}$ will be within $2\sigma/\sqrt{n}$ of the true mean p_0. The final step is to note that for any value of p_0 we have $\sigma \le 1/2$. (It is an easy calculus exercise to see that 1/4 is the maximum value of $\sigma^2 = p_0(1-p_0)$.) This means that we can conservatively say that in 95% of polls of n people the sample mean $\overline{X}$ is within $1/\sqrt{n}$ of the true mean. The frequentist statistician then takes the interval $\overline{X} \pm 1/\sqrt{n}$ and calls it the 95% confidence interval for p_0.

A word of caution: it is tempting and common, **but wrong**, to think that there is a 95% probability the true fraction p_0 is in a particular confidence interval. This is subtle, but the error is the same one as thinking you have a disease if a test with a 95% true positive rate comes back positive. We will go into this in much more detail when we learn about confidence intervals.

5.5 Why use the CLT

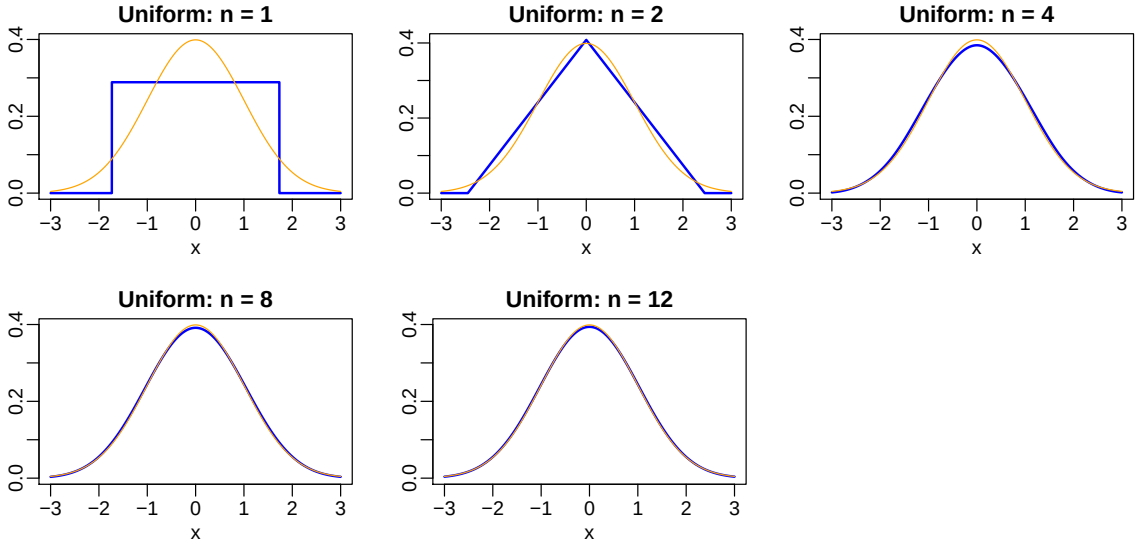
Since the probabilities in the above examples can be computed exactly using the binomial distribution, you may be wondering what is the point of finding an approximate answer using the CLT. In fact, we were only able to compute these probabilities exactly because the X_i were Bernoulli and so the sum S was binomial. In general, the distribution of the X_i may not be familiar, or may not even be known, so you will not be able to compute the probabilities for S exactly. It can also happen that the exact computation is possible in theory but too computationally intensive in practice, even for a computer. The power of the CLT is that it applies whenever X_i has a mean and a variance. Though the CLT applies to many distributions, we will see in the next section that some distributions require larger n for the approximation to be a good one.

5.6 How big does n have to be to apply the CLT?

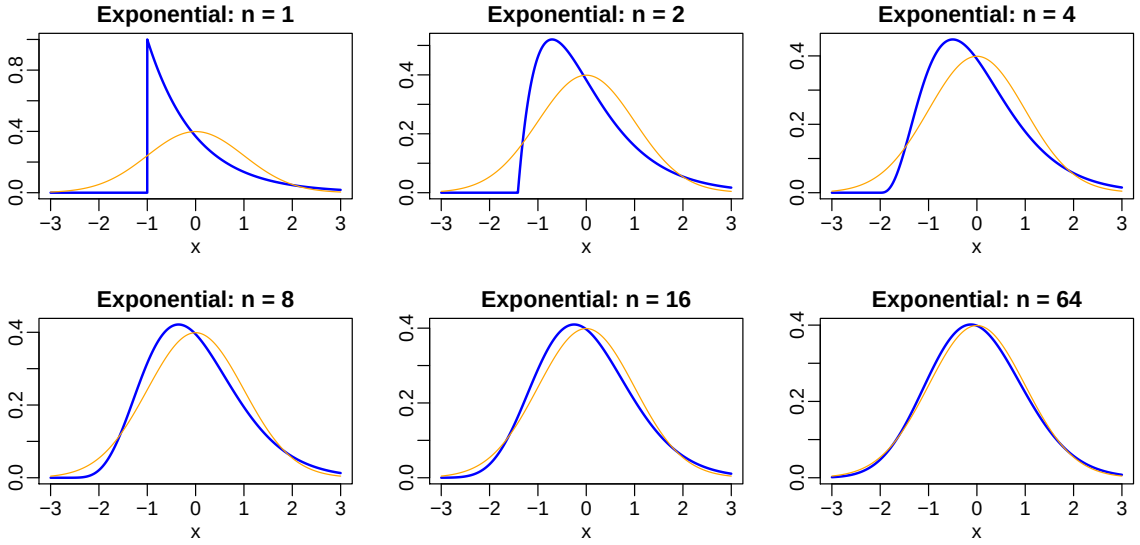
Short answer: often, not that big.

The following sequences of pictures show the convergence of averages to a normal distribution.

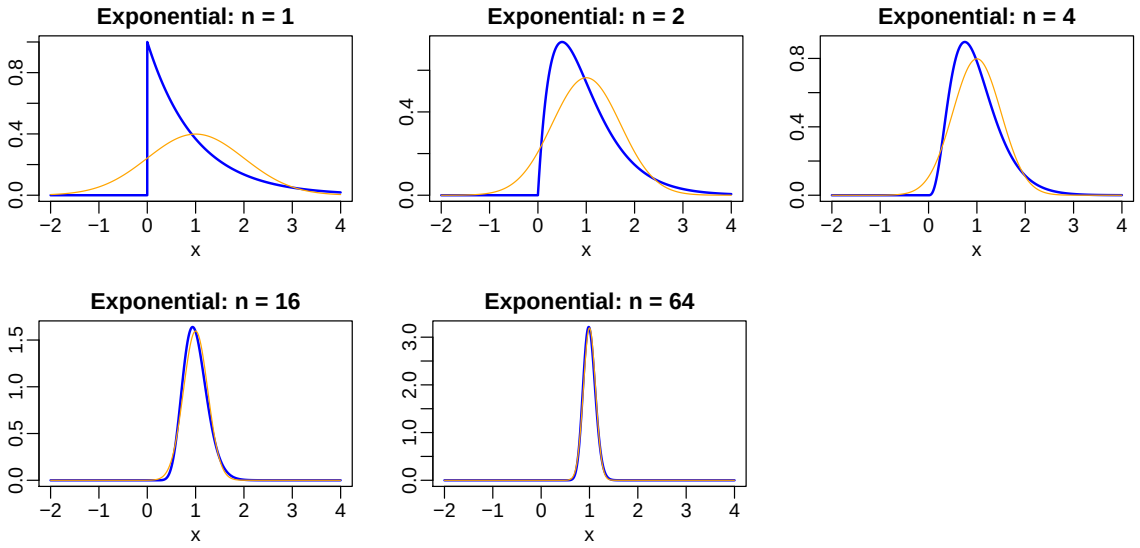
First we show the standardized average of n i.i.d. **uniform** random variables with $n = 1, 2, 4, 8, 12$. The pdf of the average is in blue and the standard normal pdf is in red. By the time $n = 12$ the fit between the standardized average and the true normal looks very good.



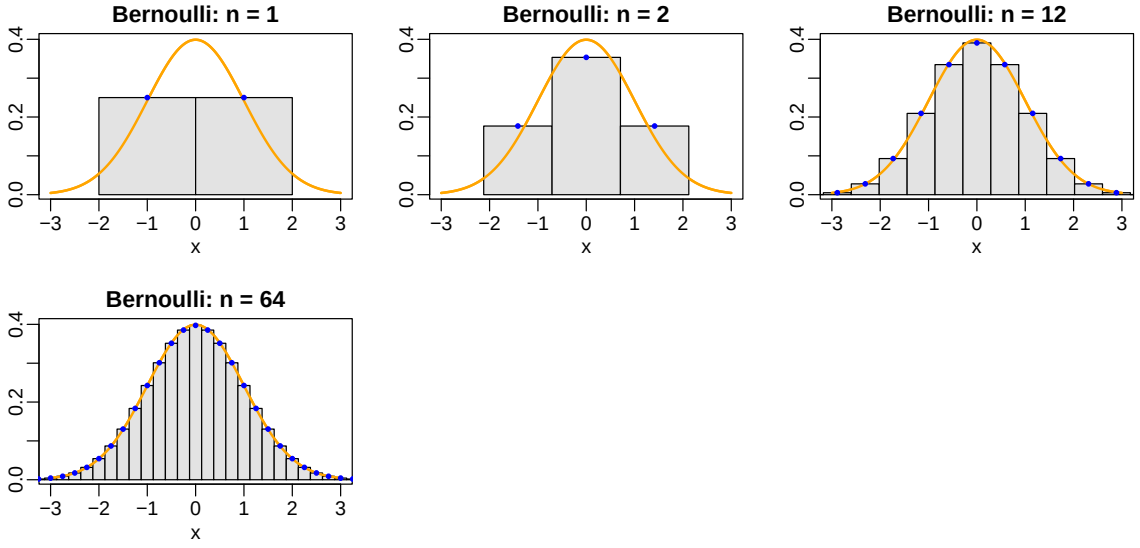
Next we show the standardized average of n i.i.d. **exponential** random variables with $n = 1, 2, 4, 8, 16, 64$. Notice that this asymmetric density takes more terms to converge to the normal density.



Next we show the (non-standardized) average of n exponential random variables with $n = 1, 2, 4, 16, 64$. Notice how this standard deviation shrinks as n grows, resulting in a spikier (more peaked) density.

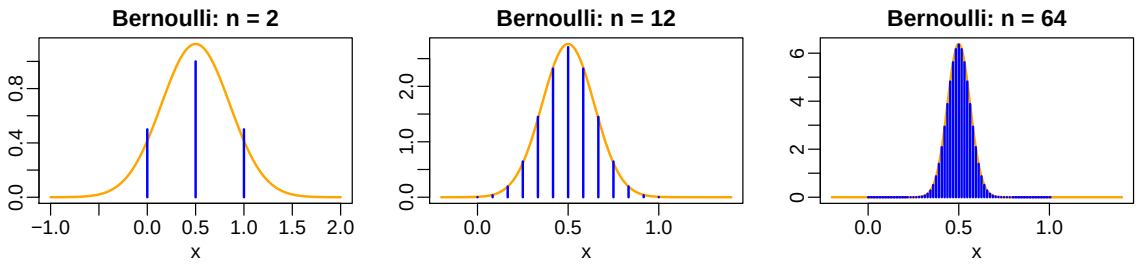


The central limit theorem works for discrete variables also. Here is the standardized average of n i.i.d. Bernoulli(0.5) random variables with $n = 1, 2, 12, 64$. Notice that as n grows, the average can take more values, which allows the discrete distribution to 'fill in' the normal density.



Note. In order to put the binomial (sum of Bernoulli) and normal distribution on the same axes, we had to convert the binomial probability mass function to a density. We did this by making it a bar graph with bars centered on each value and with bar width equal to the distance between values. Then the height of each bar is chosen so that the area equals the probability of the corresponding value.

Finally we show the (non-standardized) average of n Bernoulli(0.5) random variables, with $n = 4, 12, 64$. Notice how the standard deviation gets smaller resulting in a spikier (more peaked) density. (In these figures, rather than plotting colored bars, we made the bars white and only plotted a blue line at the center of each bar.


Bernoulli: n = 2
0.0
0.4
0.8
−1.0
0.0
0.5
1.0
1.5
2.0
x
Bernoulli: n = 12
0.0
1.0
2.0
0.0
0.5
1.0
x
Bernoulli: n = 64
0
2
4
6
0.0
0.5
1.0
x

Appendix
Class 6, 18.05
Jeremy Orloff and Jonathan Bloom

1 Introduction

In this appendix we give more formal mathematical material that is not strictly a part of 18.05. This will not be on homework or tests. We give this material to emphasize that in doing mathematics we should be careful to specify our hypotheses completely and give clear deductive arguments to prove our claims. We hope you find it interesting and illuminating.

2 With high probability the density histogram resembles the graph of the probability density function:

We stated that one consequence of the law of large numbers is that as the number of samples increases the density histogram of the samples has an increasing probability of matching the graph of the underlying pdf or pmf. This is a good rule of thumb, but it is rather imprecise. It is possible to make more precise statements. It will take some care to make a sensible and precise statement, which will not be quite so sweeping.

Suppose we have an experiment that produces data according to the random variable X and suppose we generate n independent samples from X. Call them

$$x_1, x_2, \dots, x_n.$$

By a bin we mean a range of values, i.e. $(b_1, b_2]$. The data point x_k is in this bin if $b_1 < x_k \leq b_2$. (For the left-most bin, we would use an interval closed on both sides.) To make a density histogram of the data we divide the range of X into m bins and calculate the fraction of the data in each bin.

Now, let p_k be the probability a random data point is in the kth bin. This is this probability for an indicator (Bernoulli) random variable $B_{k,j}$ which is 1 if the jth data point is in the bin and and 0 otherwise.

Statement 1. Let $\bar{p}_k$ be the fraction of the data in bin k. As the number n of data points gets large the probability that $\bar{p}_k$ is close to p_k approaches 1. Said differently, given any small number, call it a the probability $P(|\bar{p}_k - p_k| < a)$ depends on n, and as n goes to infinity this probability goes to 1.

Proof. Let $\bar{B}_k$ be the average of $B_{k,j}$. Since $E[B_{k,j}] = p_k$, the law of large number says exactly that

$$P(|\bar{B}_k - p_k| < a) \quad \text{approaches 1 as } n \text{ goes to infinity.}$$

But, since the $B_{k,j}$ are indicator variables, their average is exactly $\bar{p}_k$, the fraction of the data in bin k. Replacing $\bar{B}_k$ by $\bar{p}_k$ in the above equation gives

$$P(|\bar{p}_k - p_k| < a) \quad \text{approaches 1 as } n \text{ goes to infinity.}$$

This is exactly what Statement 1 claimed.

Statement 2. The same statement holds for a finite number of bins simultaneously. That is, for bins 1 to m we have

$P(\,(|\bar{B}_1 - p_1| < a),\,(|\bar{B}_2 - p_2| < a),\,\ldots,\,(|\bar{B}_m - p_m| < a)\,)$ approaches 1 as n goes to infinity.

Proof. First we note the following probability rule, which is a consequence of the inclusion exclusion principle: If two events A and B have $P(A) = 1 - \alpha_1$ and $P(B) = 1 - \alpha_2$ then $P(A \cap B) \geq 1 - (\alpha_1 + \alpha_2)$.

Now, Statement 1 says that for any α we can find n large enough that $P(|\bar{B}_k - p_k| < a) > 1 - \alpha/m$ for each bin separately. By the probability rule, the probability of the intersection of all these events is at least $1 - \alpha$. Since we can let α be as small as we want by letting n go to infinity, in the limit we get probability 1 as claimed.

Statement 3. If $f(x)$ is a continuous probability density with range $[a, b]$ then by taking enough data and having a small enough bin width we can insure that with high probability the density histogram is as close as we want to the graph of $f(x)$.

Proof. We will only sketch the argument. Assume the bin around x has width is Δx. If Δx is small enough then the probability a data point is in the bin is approximately $f(x)\Delta x$. Statement 2 guarantees that if n is large enough then with high probability the fraction of data in the bin is also approximately $f(x)\Delta x$. Since this is the area of the bin we see that its height will be approximately $f(x)$. That is, with high probability the height of the histogram over any point x is close to $f(x)$. This is what Statement 3 claimed.

Note. If the range is infinite or the density goes to infinity at some point we need to be more careful. There are statements we could make for these cases.

3 The Chebyshev inequality

One proof of the LoLN follows from the following key inequality.

The Chebyshev inequality. Suppose Y is a random variable with mean μ and variance σ^2. Then for any positive value a, we have

$$P(|Y - \mu| \geq a) \leq \frac{\mathrm{Var}(Y)}{a^2}.$$

In words, the Chebyshev inequality says that the probability that Y differs from the mean by more than a is bounded by $\mathrm{Var}(Y)/a^2$. Morally, the smaller the variance of Y, the smaller the probability that Y is far from its mean.

Proof of the LoLN: Since $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X)/n$, the variance of the average $\bar{X}_n$ goes to zero as n goes to infinity. So the Chebyshev inequality for $Y = \bar{X}_n$ and fixed a implies that as n grows, the probability that $\bar{X}_n$ is farther than a from μ goes to 0. Hence the probability that $\bar{X}_n$ is within a of μ goes to 1, which is the LoLN.

Proof of the Chebyshev inequality: The proof is essentially the same for discrete and continuous Y. We'll assume Y is continuous and also that $\mu = 0$, since replacing Y by $Y - \mu$

does not change the variance. So

$$\begin{aligned} P(|Y| \geq a) &= \int_{-\infty}^{-a} f(y)\,dy + \int_{a}^{\infty} f(y)\,dy \leq \int_{-\infty}^{-a} \frac{y^2}{a^2} f(y)\,dy + \int_{a}^{\infty} \frac{y^2}{a^2} f(y)\,dy \\ &\leq \int_{-\infty}^{\infty} \frac{y^2}{a^2} f(y)\,dy = \frac{\mathrm{Var}(Y)}{a^2}. \end{aligned}$$

The first inequality uses that $y^2/a^2 \geq 1$ on the intervals of integration. The second inequality follows because including the range $[-a, a]$ only makes the integral larger, since the integrand is positive.

4 The need for variance

We didn't lie to you, but we did gloss over one technical fact. Throughout we assumed that the underlying distributions had a variance. For example, the proof of the law of large numbers made use of the variance by way of the Chebyshev inequality. But there are distributions which do not have a mean and variance because the sums or integrals for these do not converge to a finite number. For such distributions the law of large numbers may not be true.

In 18.05 we won't have to worry about this, but if you go deeper into statistics this may become important. For those who are interested: a standard example you can look up or play with in R is the Cauchy distribution.

Joint Distributions, Independence
Class 7, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Understand what is meant by a joint pmf, pdf and cdf of two random variables.
2. Be able to compute probabilities and marginals from a joint pmf or pdf.
3. Be able to test whether two random variables are independent.

2 Introduction

In science and in real life, we are often interested in two (or more) random variables at the same time. For example, we might measure the height and weight of giraffes, or the IQ and birthweight of children, or the frequency of exercise and the rate of heart disease in adults, or the level of air pollution and rate of respiratory illness in cities, or the number of Facebook friends and the age of Facebook members.

Think: What relationship would you expect in each of the five examples above? Why?

In such situations the random variables have a joint distribution that allows us to compute probabilities of events involving both variables and understand the relationship between the variables. This is simplest when the variables are independent. When they are not, we use covariance and correlation as measures of the nature of the dependence between them.

3 Joint Distribution

3.1 Discrete case

Suppose X and Y are two discrete random variables and that X takes values $\{x_1, x_2, \dots, x_n\}$ and Y takes values $\{y_1, y_2, \dots, y_m\}$. The ordered pair (X, Y) take values in the product $\{(x_1, y_1), (x_1, y_2), \dots (x_n, y_m)\}$. The joint probability mass function (joint pmf) of X and Y is the function $p(x_i, y_j)$ giving the probability of the joint outcome $X = x_i$, $Y = y_j$.

We organize this in a joint probability table as shown:

$X\backslash Y$	y_1	y_2	...	y_j	...	y_m
x_1	$p(x_1,y_1)$	$p(x_1,y_2)$	$\cdots$	$p(x_1,y_j)$	$\cdots$	$p(x_1,y_m)$
x_2	$p(x_2,y_1)$	$p(x_2,y_2)$	$\cdots$	$p(x_2,y_j)$	$\cdots$	$p(x_2,y_m)$
...	...	...	...	...	...	...
...	...	...	...	...	...	...
x_i	$p(x_i,y_1)$	$p(x_i,y_2)$	$\cdots$	$p(x_i,y_j)$	$\cdots$	$p(x_i,y_m)$
...	...	...	...	...	...	
x_n	$p(x_n,y_1)$	$p(x_n,y_2)$	$\cdots$	$p(x_n,y_j)$	$\cdots$	$p(x_n,y_m)$

Example 1. Roll two dice. Let X be the value on the first die and let Y be the value on the second die. Then both X and Y take values 1 to 6 and the joint pmf is $p(i,j) = 1/36$ for all i and j between 1 and 6. Here is the joint probability table:

$X\backslash Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

Example 2. Roll two dice. Let X be the value on the first die and let T be the total on both dice. Here is the joint probability table:

$X\backslash T$	2	3	4	5	6	7	8	9	10	11	12
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36

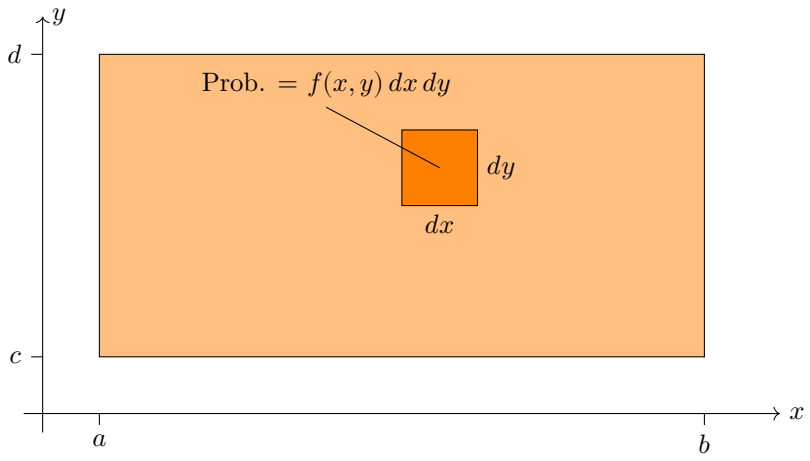
A joint probability mass function must satisfy two properties:
1. $0 \le p(x_i, y_j) \le 1$
2. The total probability is 1. We can express this as a double sum:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j) = 1$$

3.2 Continuous case

The continuous case is essentially the same as the discrete case: we just replace discrete sets of values by continuous intervals, the joint probability mass function by a joint probability density function, and the sums by integrals.

If X takes values in $[a, b]$ and Y takes values in $[c, d]$ then the pair (X, Y) takes values in the product $[a, b] \times [c, d]$. The joint probability density function (joint pdf) of X and Y is a function $f(x, y)$ giving the probability density at (x, y). That is, the probability that (X, Y) is in a small rectangle of width dx and height dy around (x, y) is $f(x, y)\, dx\, dy$.



A joint probability density function must satisfy two properties:
1. $0 \le f(x, y)$
2. The total probability is 1. We now express this as a double integral:

$$\int_c^d \int_a^b f(x, y)\, dx\, dy = 1$$

Note: as with the pdf of a single random variable, the joint pdf $f(x, y)$ can take values greater than 1; it is a probability density, **not** a probability.

In 18.05 we won't expect you to be experts at double integration. Here's what we will expect.

- You should understand double integrals conceptually as double sums.
- You should be able to compute double integrals over rectangles.
- For a non-rectangular region, when $f(x, y) = c$ is constant, you should know that the double integral is the same as the $c \times$ (the area of the region).

3.3 Events

Random variables are useful for describing events. Recall that an event is a set of outcomes and that random variables assign numbers to outcomes. For example, the event '$X > 1$' is the set of all outcomes for which X is greater than 1. These concepts readily extend to pairs of random variables and joint outcomes.

Example 3. In Example 1, describe the event $B =$ '$Y - X \geq 2$' and find its probability.

Solution: We can describe B as a set of (X, Y) pairs:

$$B = \{(1,3), (1,4), (1,5), (1,6), (2,4), (2,5), (2,6), (3,5), (3,6), (4,6)\}.$$

We can also describe it visually

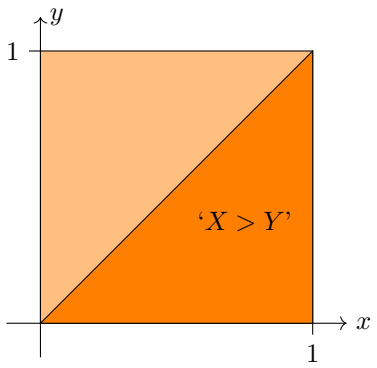
$X\backslash Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

The event B consists of the outcomes in the shaded squares.

The probability of B is the sum of the probabilities in the orange shaded squares, so $P(B) = 10/36$.

Example 4. Suppose X and Y both take values in [0,1] with uniform density $f(x, y) = 1$. Visualize the event '$X > Y$' and find its probability.

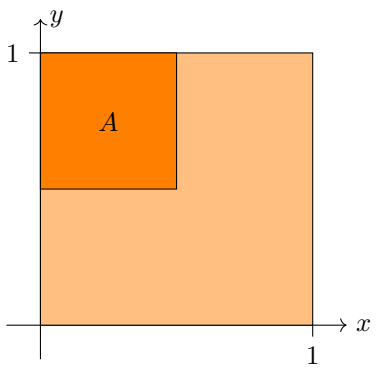
Solution: Jointly X and Y take values in the unit square. The event '$X > Y$' corresponds to the shaded lower-right triangle below. Since the density is constant, the probability is just the fraction of the total area taken up by the event. In this case, it is clearly 0.5.



The event '$X > Y$' in the unit square.

Example 5. Suppose X and Y both take values in [0,1] with density $f(x, y) = 4xy$. Show $f(x, y)$ is a valid joint pdf, visualize the event $A =$ '$X < 0.5$ and $Y > 0.5$' and find its probability.

Solution: Jointly X and Y take values in the unit square.



The event A in the unit square.

To show $f(x, y)$ is a valid joint pdf we must check that it is positive (which it clearly is) and that the total probability is 1.

$$\text{Total probability} = \int_0^1 \int_0^1 4xy\,dx\,dy = \int_0^1 \left[2x^2y\right]_0^1\,dy = \int_0^1 2y\,dy = 1. \quad \text{QED}$$

The event A is just the upper-left-hand quadrant. Because the density is not constant we must compute an integral to find the probability.

$$P(A) = \int_0^{0.5} \int_{0.5}^1 4xy\,dy\,dx = \int_0^{0.5} \left[2xy^2\right]_{0.5}^1\,dx = \int_0^{0.5} \frac{3x}{2}\,dx = \boxed{\frac{3}{16}}.$$

3.4 Joint cumulative distribution function

Suppose X and Y are jointly-distributed random variables. We will use the notation '$X \le x$, $Y \le y$' to mean the event '$X \le x$ and $Y \le y$'. The joint cumulative distribution function (joint cdf) is defined as

$$F(x, y) = P(X \le x,\, Y \le y)$$

Continuous case: If X and Y are continuous random variables with joint density $f(x, y)$ over the range $[a, b] \times [c, d]$ then the joint cdf is given by the double integral

$$F(x, y) = \int_c^y \int_a^x f(u, v)\,du\,dv.$$

To recover the joint pdf, we differentiate the joint cdf. Because there are two variables we need to use partial derivatives:

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

Discrete case: If X and Y are discrete random variables with joint pmf $p(x_i, y_j)$ then the joint cdf is give by the double sum

$$F(x, y) = \sum_{x_i \le x} \sum_{y_j \le y} p(x_i, y_j).$$

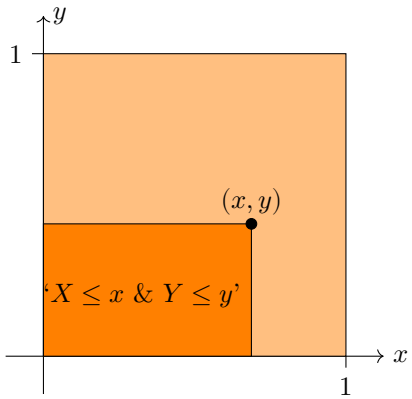
3.5 Properties of the joint cdf

The joint cdf $F(x, y)$ of X and Y must satisfy several properties:

1. $F(x, y)$ is non-decreasing: i.e. if x or y increase then $F(x, y)$ must stay constant or increase.
2. $F(x, y) = 0$ at the lower-left of the joint range.
   If the lower left is $(-\infty, -\infty)$ then this means $\lim_{(x,y)\to(-\infty,-\infty)} F(x, y) = 0$.
3. $F(x, y) = 1$ at the upper-right of the joint range.
   If the upper-right is (∞, ∞) then this means $\lim_{(x,y)\to(\infty,\infty)} F(x, y) = 1$.

Example 6. Find the joint cdf for the random variables in Example 5.

Solution: The event '$X \le x$ and $Y \le y$' is a rectangle in the unit square.



To find the cdf $F(x, y)$ we compute a double integral:

$$F(x, y) = \int_0^y \int_0^x 4uv \, du \, dv = \boxed{x^2y^2}.$$

Example 7. In Example 1, compute $F(3.5, 4)$.

Solution: We redraw the joint probability table. Notice how similar the picture is to the one in the previous example.

$F(3.5, 4)$ is the probability of the event '$X \le 3.5$ and $Y \le 4$'. We can visualize this event as the shaded rectangles in the table:

$X\backslash Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

The event '$X \le 3.5$ and $Y \le 4$'.

Adding up the probability in the shaded squares we get $F(3.5, 4) = 12/36 = 1/3$.

Note. One unfortunate difference between the continuous and discrete visualizations is that for continuous variables the value increases as we go up in the vertical direction while the opposite is true for the discrete case. We have experimented with changing the discrete tables to match the continuous graphs, but it causes too much confusion. We will just have to live with the difference!

3.6 Marginal distributions

When X and Y are jointly-distributed random variables, we may want to consider only one of them, say X. In that case we need to find the pmf (or pdf or cdf) of X without Y. This is called a marginal pmf of the joint pmf (or pdf or cdf). The next example illustrates the way to compute this and the reason for the term 'marginal'.

3.7 Marginal pmf

Example 8. In Example 2 we rolled two dice and let X be the value on the first die and T be the total on both dice. Compute the marginal pmf for X and for T.

Solution: In the table each row represents a single value of X. So the event '$X = 3$' is the third row of the table. To find $P(X = 3)$ we simply have to sum up the probabilities in this row. We put the sum in the right-hand margin of the table. Likewise $P(T = 5)$ is just the sum of the column with $T = 5$. We put the sum in the bottom margin of the table.

$X\backslash T$	2	3	4	5	6	7	8	9	10	11	12	$p(x_i)$
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0	1/6
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	1/6
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	1/6
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	1/6
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	1/6
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(t_j)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	1

Computing the marginal probabilities $P(X = 3) = 1/6$ and $P(T = 5) = 4/36$.

Note: Of course in this case we already knew the pmf of X and of T. It is good to see that our computation here is in agreement!

As motivated by this example, marginal pmfs are obtained from the joint pmf by summing:

$$p_X(x_i) = \sum_j p(x_i, y_j), \qquad p_Y(y_j) = \sum_i p(x_i, y_j)$$

The term marginal refers to the fact that the values are written in the margins of the table.

3.8 Marginal pdf

For a continous joint density $f(x,y)$ with range $[a,b]\times[c,d]$, the marginal pdfs are:

$$f_X(x) = \int_c^d f(x,y)\,dy, \qquad f_Y(y) = \int_a^b f(x,y)\,dx.$$

Compare these with the marginal pmfs above; as usual the sums are replaced by integrals. We say that to obtain the marginal for X, we integrate out Y from the joint pdf and vice versa.

Example 9. Suppose (X,Y) takes values on the square $[0,1]\times[1,2]$ with joint pdf $f(x,y) = \frac{8}{3}x^3y$. Find the marginal pdfs $f_X(x)$ and $f_Y(y)$.

Solution: To find $f_X(x)$ we integrate out y and to find $f_Y(y)$ we integrate out x.

$$f_X(x) = \int_1^2 \frac{8}{3}x^3y\,dy = \left[\frac{4}{3}x^3y^2\right]_1^2 = \boxed{4x^3}$$

$$f_Y(y) = \int_0^1 \frac{8}{3}x^3y\,dx = \left[\frac{2}{3}x^4y^1\right]_0^1 = \boxed{\frac{2}{3}y}.$$

Example 10. Suppose (X,Y) takes values on the unit square $[0,1]\times[0,1]$ with joint pdf $f(x,y) = \frac{3}{2}(x^2+y^2)$. Find the marginal pdf $f_X(x)$ and use it to find $P(X<0.5)$.

Solution:

$$f_X(x) = \int_0^1 \frac{3}{2}(x^2+y^2)\,dy = \left[\frac{3}{2}x^2y + \frac{y^3}{2}\right]_0^1 = \boxed{\frac{3}{2}x^2 + \frac{1}{2}}.$$

$$P(X<0.5) = \int_0^{0.5} f_X(x)\,dx = \int_0^{0.5} \frac{3}{2}x^2 + \frac{1}{2}\,dx = \left[\frac{1}{2}x^3 + \frac{1}{2}x\right]_0^{0.5} = \boxed{\frac{5}{16}}.$$

3.9 Marginal cdf

Finding the marginal cdf from the joint cdf is easy. If X and Y jointly take values on $[a,b]\times[c,d]$ then

$$F_X(x) = F(x,d), \qquad F_Y(y) = F(b,y).$$

If d is ∞ then this becomes a limit $F_X(x) = \lim_{y\to\infty} F(x,y)$. Likewise for $F_Y(y)$.

Example 11. The joint cdf in the last example was $F(x,y) = \frac{1}{2}(x^3y + xy^3)$ on $[0,1]\times[0,1]$. Find the marginal cdfs and use $F_X(x)$ to compute $P(X<0.5)$.

Solution: We have $F_X(x) = F(x,1) = \frac{1}{2}(x^3+x)$ and $F_Y(y) = F(1,y) = \frac{1}{2}(y+y^3)$. So $P(X<0.5) = F_X(0.5) = \frac{1}{2}(0.5^3+0.5) = \frac{5}{16}$: exactly the same as before.

3.10 3D visualization

We visualized $P(a<X<b)$ as the area under the pdf f(x) over the interval $[a,b]$. Since the range of values of (X,Y) is already a two dimensional region in the plane, the graph of

$f(x, y)$ is a surface over that region. We can then visualize probability as volume under the surface.

Think: Summoning your inner artist, sketch the graph of the joint pdf $f(x, y) = 4xy$ and visualize the probability $P(A)$ as a volume for Example 5.

4 Independence

We are now ready to give a careful mathematical definition of independence. Of course, it will simply capture the notion of independence we have been using up to now. But, it is nice to finally have a solid definition that can support complicated probabilistic and statistical investigations.

Recall that events A and B are independent if

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

Random variables X and Y define events like '$X \leq 2$' and '$Y > 5$'. So, X and Y are independent if any event defined by X is independent of any event defined by Y. The formal definition that guarantees this is the following.

Definition: Jointly-distributed random variables X and Y are independent if their joint cdf is the product of the marginal cdfs:

$$F(X, Y) = F_X(x)F_Y(y).$$

For discrete variables this is equivalent to the joint pmf being the product of the marginal pmfs.:

$$p(x_i, y_j) = p_X(x_i)p_Y(y_j).$$

For continous variables this is equivalent to the joint pdf being the product of the marginal pdfs.:

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

Once you have the joint distribution, checking for independence is usually straightforward although it can be tedious.

Example 12. For discrete variables independence means the probability in a cell must be the product of the marginal probabilities of its row and column. In the first table below this is true: every marginal probability is 1/6 and every cell contains 1/36, i.e. the product of the marginals. Therefore X and Y are independent.

In the second table below most of the cell probabilities are not the product of the marginal probabilities. For example, none of marginal probabilities are 0, so none of the cells with 0 probability can be the product of the marginals.

$X\backslash Y$	1	2	3	4	5	6	$p(x_i)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(y_j)$	1/6	1/6	1/6	1/6	1/6	1/6	1

$X\backslash T$	2	3	4	5	6	7	8	9	10	11	12	$p(x_i)$
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0	1/6
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	1/6
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	1/6
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	1/6
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	1/6
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(y_j)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	1

Example 13. For continuous variables independence means you can factor the joint pdf or cdf as the product of a function of x and a function of y.

(i) Suppose X has range $[0, 1/2]$, Y has range $[0, 1]$ and $f(x, y) = 96x^2y^3$ then X and Y are independent. The marginal densities are $f_X(x) = 24x^2$ and $f_Y(y) = 4y^3$.

(ii) If $f(x, y) = 1.5(x^2+y^2)$ over the unit square then X and Y are not independent because there is no way to factor $f(x, y)$ into a product $f_X(x)f_Y(y)$.

(iii) If $F(x, y) = \frac{1}{2}(x^3y + xy^3)$ over the unit square then X and Y are not independent because the cdf does not factor into a product $F_X(x)F_Y(y)$.

Covariance and Correlation
Class 7, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Understand the meaning of covariance and correlation.
2. Be able to compute the covariance and correlation of two random variables.

2 Covariance

Covariance is a measure of how much two random variables vary together. For example, height and weight of giraffes have positive covariance because when one is big the other tends also to be big.

Definition: Suppose X and Y are random variables with means μ_X and μ_Y. The covariance of X and Y is defined as

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

2.1 Properties of covariance

1. $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$ for constants a, b, c, d.
2. $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$.
3. $\text{Cov}(X, X) = \text{Var}(X)$
4. $\text{Cov}(X, Y) = E[XY] - \mu_X\mu_Y$.
5. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$ for any X and Y.
6. If X and Y are independent then $\text{Cov}(X, Y) = 0$.
   Warning: The converse is false: zero covariance does not always imply independence.

Notes. 1. Property 4 is like the similar property for variance. Indeed, if $X = Y$ it is exactly that property: $\text{Var}(X) = E[X^2] - \mu_X^2$.

By Property 5, the formula in Property 6 reduces to our earlier formula $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ when X and Y are independent.

We give the proofs below. However, understanding and using these properties is more important than memorizing their proofs.

2.2 Sums and integrals for computing covariance

Since covariance is defined as an expected value we compute it in the usual way as a sum or integral.

Discrete case: If X and Y have joint pmf $p(x_i, y_j)$ then

$$\text{Cov}(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)(x_i-\mu_X)(y_j-\mu_Y) = \left(\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)x_iy_j\right) - \mu_X\mu_Y.$$

Continuous case: If X and Y have joint pdf $f(x,y)$ over range $[a,b]\times[c,d]$ then

$$\text{Cov}(X,Y) = \int_c^d\int_a^b (x-\mu_x)(y-\mu_y)f(x,y)\,dx\,dy = \left(\int_c^d\int_a^b xyf(x,y)\,dx\,dy\right) - \mu_x\mu_y.$$

2.3 Examples

Example 1. Flip a fair coin 3 times. Let X be the number of heads in the first 2 flips and let Y be the number of heads on the last 2 flips (so there is overlap on the middle flip). Compute $\text{Cov}(X,Y)$.

Solution: We'll do this twice, first using the joint probability table and the definition of covariance, and then using the properties of covariance.

With 3 tosses there are only 8 outcomes {HHH, HHT,...}, so we can create the joint probability table directly.

$X\backslash Y$	0	1	2	$p(x_i)$
0	1/8	1/8	0	1/4
1	1/8	2/8	1/8	1/2
2	0	1/8	1/8	1/4
$p(y_j)$	1/4	1/2	1/4	1

From the marginals we compute $E[X] = 1 = E[Y]$. Now we use use the definition:

$$\text{Cov}(X,Y) = E[(X-\mu_x)(Y-\mu_Y)] = \sum_{i,j} p(x_i,y_j)(x_i-1)(y_j-1)$$

We write out the sum leaving out all the terms that are 0, i.e. all the terms where $x_i = 1$ or $y_i = 1$ or the probability is 0.

$$\text{Cov}(X,Y) = \frac{1}{8}(0-1)(0-1) + \frac{1}{8}(2-1)(2-1) = \frac{1}{4}.$$

We could also have used property 4 to do the computation: From the full table we compute

$$E[XY] = 1\cdot\frac{2}{8} + 2\frac{1}{8} + 2\frac{1}{8} + 4\frac{1}{8} = \frac{5}{4}.$$

So $\text{Cov}(XY) = E[XY] - \mu_X\mu_Y = \dfrac{5}{4} - 1 = \dfrac{1}{4}$.

Next we redo the computation of $\text{Cov}(X, Y)$ using the properties of covariance. As usual, let X_i be the result of the i^{th} flip, so $X_i \sim \text{Bernoulli}(0.5)$. We have

$$X = X_1 + X_2 \quad \text{and} \quad Y = X_2 + X_3.$$

We know $E[X_i] = 1/2$ and $\text{Var}(X_i) = 1/4$. Therefore using Property 2 of covariance, we have

$$\text{Cov}(X, Y) = \text{Cov}(X_1+X_2, X_2+X_3) = \text{Cov}(X_1, X_2)+\text{Cov}(X_1, X_3)+\text{Cov}(X_2, X_2)+\text{Cov}(X_2, X_3).$$

Since the different tosses are independent we know

$$\text{Cov}(X_1, X_2) = \text{Cov}(X_1, X_3) = \text{Cov}(X_2, X_3) = 0.$$

Looking at the expression for $\text{Cov}(X, Y)$ there is only one non-zero term

$$\text{Cov}(X, Y) = \text{Cov}(X_2, X_2) = \text{Var}(X_2) = \boxed{\frac{1}{4}}.$$

Example 2. (Zero covariance does not imply independence.) Let X be a random variable that takes values $-2, -1, 0, 1, 2$; each with probability $1/5$. Let $Y = X^2$. Show that $\text{Cov}(X, Y) = 0$ but X and Y are not independent.

Solution: We make a joint probability table:

$Y \backslash X$	-2	-1	0	1	2	$p(y_j)$
0	0	0	1/5	0	0	1/5
1	0	1/5	0	1/5	0	2/5
4	1/5	0	0	0	1/5	2/5
$p(x_i)$	1/5	1/5	1/5	1/5	1/5	1

Using the marginals we compute means $E[X] = 0$ and $E[Y] = 2$.

Next we show that X and Y are not independent. To do this all we have to do is find one place where the product rule fails, i.e. where $p(x_i, y_j) \neq p(x_i)p(x_j)$:

$$P(X = -2,\, Y = 0) = 0 \quad \text{but} \quad P(X = -2) \cdot P(Y = 0) = 1/25.$$

Since these are not equal X and Y are not independent. Finally we compute covariance using Property 4:

$$\text{Cov}(X, Y) = \frac{1}{5}(-8 - 1 + 1 + 8) - \mu_X\mu_y = 0.$$

Discussion: This example shows that $\text{Cov}(X, Y) = 0$ does not imply that X and Y are independent. In fact, X and X^2 are as dependent as random variables can be: if you know the value of X then you know the value of X^2 with 100% certainty.

The key point is that $\text{Cov}(X, Y)$ measures the linear relationship between X and Y. In the above example X and X^2 have a quadratic relationship that is completely missed by $\text{Cov}(X, Y)$.

Continuous covariance works the same way, except our computations are done with integrals instead of sums. Here is an example.

Example 3. Continuous covariance. Suppose X and Y are jointly distributed random variables, with range on the unit square $[0, 1] \times [0, 1]$ and joint pdf $f(x, y) = 2x^3 + 2y^3$.

(i) Verify the $f(x, y)$ is a valid probability density.

(ii) Compute μ_X and μ_Y.

(iii) Compute the covariance of $\text{Cov}(X, Y)$

Solution: Part of the point of this example is to show how to set up and compute the integrals using a joint density function. Since the pdf here is a polynomial, these computations are relatively easy.

(i) A valid pdf has two properties: it is nonnegative and the total integral over the entire joint range is 1.

Nonnegativity is clear: $f(x, y) \geq 0$. The integral is not hard to compute

$$\int_0^1 \int_0^1 f(x, y)\, dx\, dy = \int_0^1 \int_0^1 2x^3 + 2y^3\, dx\, dy$$

$$\text{Inner integral: } \int_0^1 2x^3 + 2y^3\, dx = \frac{x^4}{2} + 2xy^3 \Big|_0^1 = \frac{1}{2} + 2y^3.$$

$$\text{Outer integral: } \int_0^1 \frac{1}{2} + 2y^3\, dy = \frac{y}{2} + \frac{y^4}{2} \Big|_0^1 = 1.$$

So, the integral over the entire joint range is 1. Thus, $f(x, y) = x + y$ is a valid probability density.

(ii) We need to compute integrals to find the means. We will write down the integrals, but not show the details of their computation. (Also, by symmetry, we know the two means are the same.)

$$\mu_X = \int_0^1 \int_0^1 x f(x, y)\, dx\, dy = \int_0^1 \int_0^1 2x^4 + 2xy^3)\, dx\, dy = \frac{13}{20}$$

$$\mu_Y = \int_0^1 \int_0^1 y f(x, y)\, dx\, dy = \int_0^1 \int_0^1 2yx^3 + 2y^4\, dx\, dy = \frac{13}{20}$$

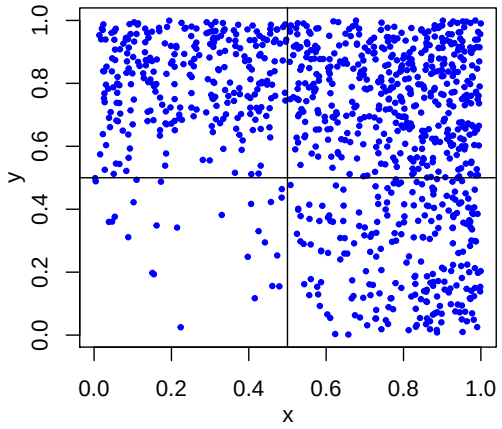
(iii) We know $\text{Cov}(X, Y) = E\left[(X - \mu_x)(Y - \mu_Y)\right]$. This is an integral. Again, we will write down the integral, but not show details of its computation,

$$\text{Cov}(X, Y) = E\left[(X - \mu_x)(Y - \mu_Y)\right] = \int_0^1 \int_0^1 (x - 13/20)(y - 13/20) f(x, y)\, dx\, dy$$

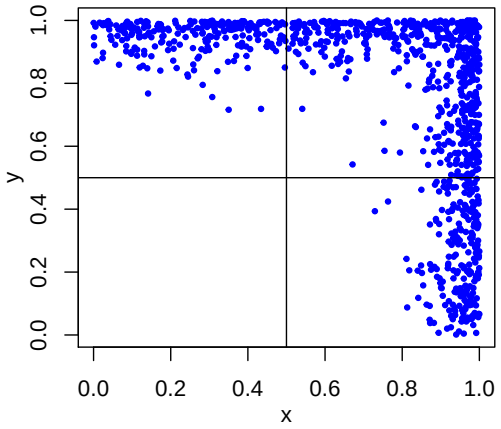
$$= \int_0^1 \int_0^1 (x - 7/12)(y - 7/12)(2x^3 + 2y^3)\, dx\, dy = -\frac{9}{400}$$

(In fact, we wrote down the integral in the most straightforward way, but secretly we did the computation by computing $E[XY] - E[X]E[Y]$.)

Here's a plot of the pseudo-random samples generated from this distribution. Because the R code could do it easily, we also include a plot with a more extreme density function.



Samples from $f(x, y) = 2x^3 + 2y^3$.



Samples from $f(x, y) = 10x^{19} + 10y^{19}$.

3 Correlation

The units of covariance $\mathrm{Cov}(X, Y)$ are 'units of X times units of Y'. This makes it hard to compare covariances: if we change scales then the covariance changes as well. Correlation is a way to remove the scale from the covariance.

Definition: The correlation coefficient between X and Y is defined by

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

3.1 Properties of correlation

1. ρ is the covariance of the standardizations of X and Y.
2. ρ **is dimensionless** (it's a ratio!).
3. $-1 \le \rho \le 1$. Furthermore,
   $\rho = +1$ if and only if $Y = aX + b$ with $a > 0$,
   $\rho = -1$ if and only if $Y = aX + b$ with $a < 0$.

Property 3 shows that ρ measures the linear relationship between variables. If the correlation is positive then when X is large, Y will tend to large as well. If the correlation is negative then when X is large, Y will tend to be small.

Example 2 above shows that correlation can completely miss higher order relationships.

Example 4. We continue Example 1. To compute the correlation we divide the covariance by the standard deviations. In Example 1 we found $\mathrm{Cov}(X, Y) = 1/4$ and $\mathrm{Var}(X) =$

$2\text{Var}(X_j) = 1/2$. So, $\sigma_X = 1/\sqrt{2}$. Likewise $\sigma_Y = 1/\sqrt{2}$. Thus

$$\text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{1/4}{1/2} = \frac{1}{2}.$$

We see a positive correlation, which means that larger X tend to go with larger Y and smaller X with smaller Y. In Example 1 this happens because toss 2 is included in both X and Y, so it contributes to the size of both.

Example 5. Look back at Example 3. See if you can compute the following.

$$\text{Var}(X) = 31/400, \text{ so } \sigma_X = \sqrt{31/400} \approx 0.28$$
$$\text{Var}(Y) = \text{Var}(X), \text{ so } \sigma_Y \approx 0.28$$
$$\text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} \approx -0.29.$$

3.2 Bivariate normal distributions

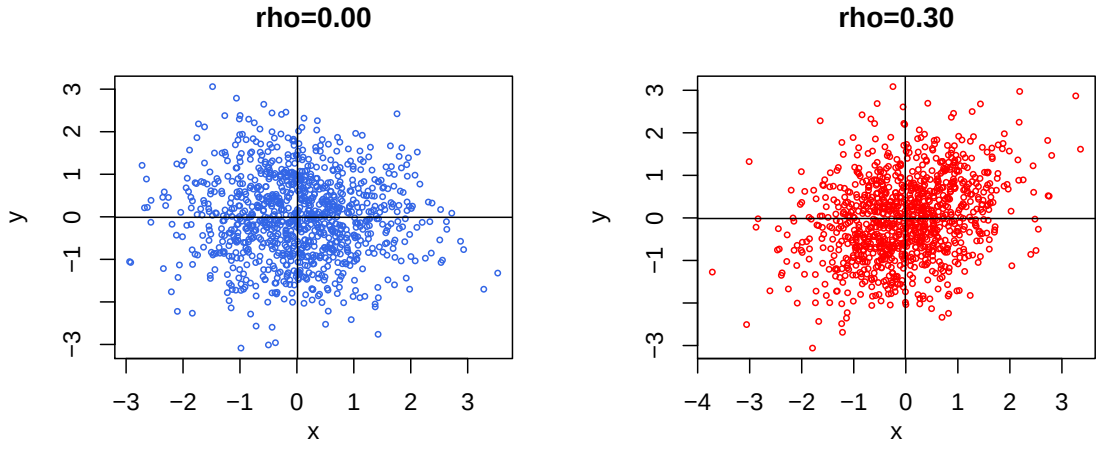
The bivariate normal distribution has density

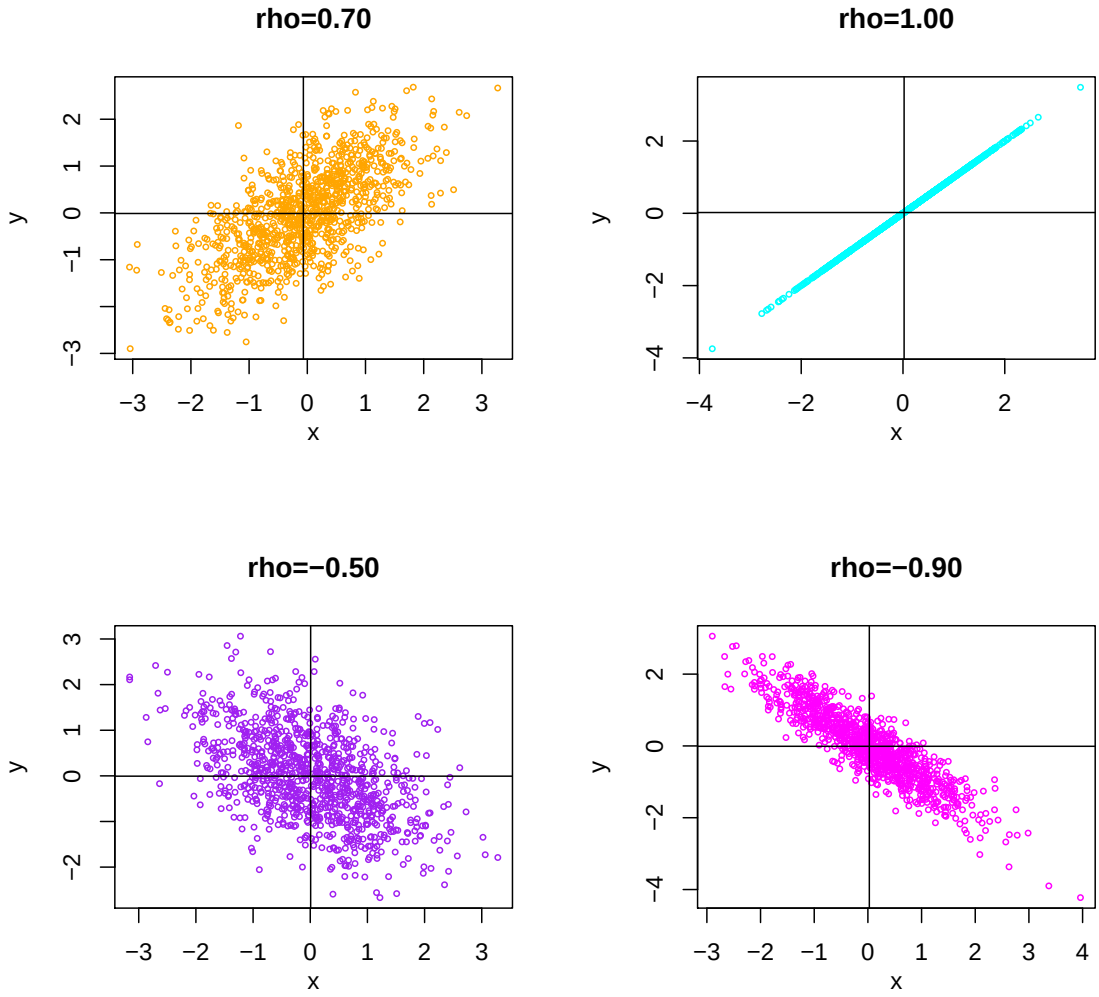
$$f(x,y) = \frac{e^{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

For this distribution, the marginal distributions for X and Y are normal and the correlation between X and Y is ρ.

In the figures below we used R to simulate the distribution for various values of ρ. Individually X and Y are standard normal, i.e. $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$. The figures show scatter plots of the results.

These plots and the next set show an important feature of correlation. We divide the data into quadrants by drawing a horizontal and a verticle line at the means of the y data and x data respectively. A positive correlation corresponds to the data tending to lie in the 1st and 3rd quadrants. A negative correlation corresponds to data tending to lie in the 2nd and 4th quadrants. You can see the data gathering about a line as ρ becomes closer to ± 1.

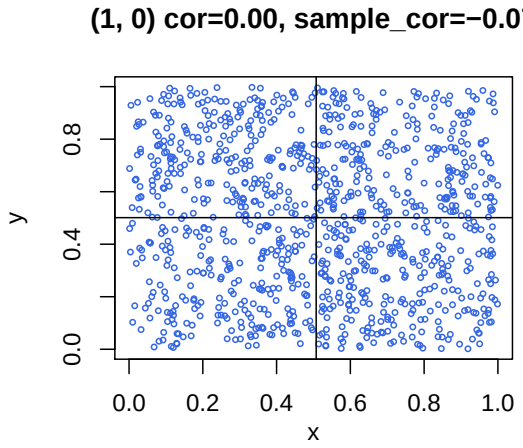


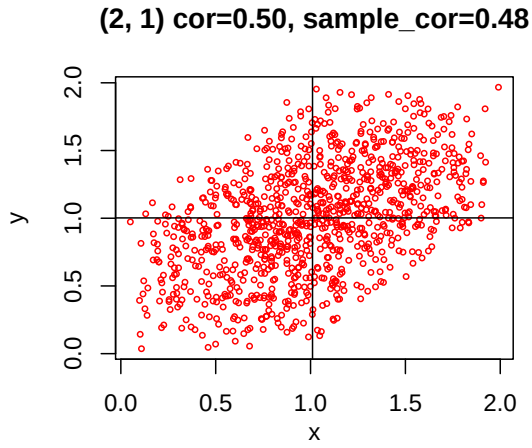


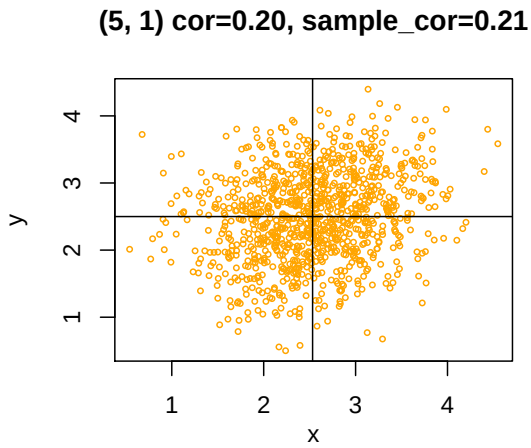
3.3 Overlapping uniform distributions

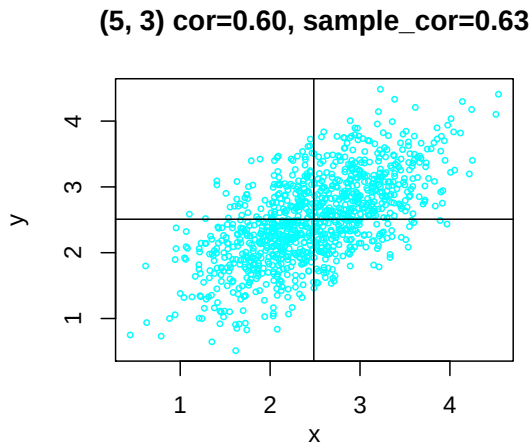
We ran simulations in R of the following scenario. X_1, X_2, ..., X_{20} are i.i.d and follow a $U(0,1)$ distribution. X and Y are both sums of the same number of X_i. We call the number of X_i common to both X and Y the overlap. The notation in the figures below indicates the number of X_i being summed and the number which overlap. For example, 5,3 indicates that X and Y were each the sum of 5 of the X_i and that 3 of the X_i were common to both sums. (The data was generated using `rand(1,1000);`)

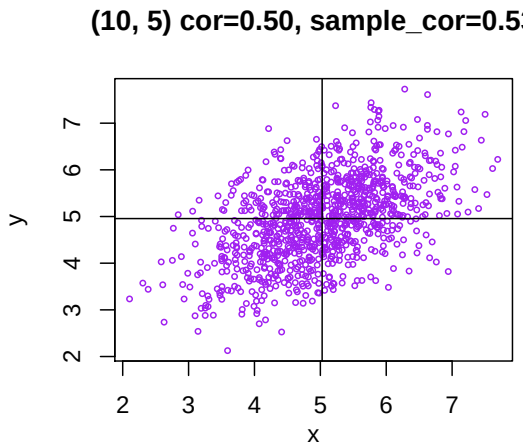
Using the linearity of covariance it is easy to compute the theoretical correlation. For each plot we give both the theoretical correlation and the correlation of the data from the simulated sample.

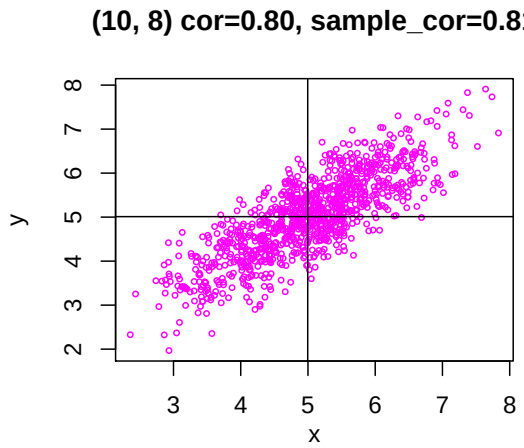

(1, 0) cor=0.00, sample_cor=−0.07
y
x
0.0
0.2
0.4
0.6
0.8
1.0


(2, 1) cor=0.50, sample_cor=0.48
y
x
0.0
0.5
1.0
1.5
2.0


(5, 1) cor=0.20, sample_cor=0.21
y
x
1
2
3
4


(5, 3) cor=0.60, sample_cor=0.63
y
x
1
2
3
4


(10, 5) cor=0.50, sample_cor=0.53
y
x
2
3
4
5
6
7


(10, 8) cor=0.80, sample_cor=0.81
y
x
2
3
4
5
6
7
8

4 Proof of the properties of covariance and correlation

4.1 Proofs of the properties of covariance

1 and 2 follow from similar properties for expected value.

3. This is the definition of variance:

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

4. Recall that $E[X - \mu_x] = 0$. So

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\ &= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X \mu_Y \\ &= E[XY] - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y \\ &= E[XY] - \mu_X \mu_Y.\end{aligned}$$

5. Using properties 3 and 2 we get

$$\text{Var}(X+Y) = \text{Cov}(X+Y, X+Y) = \text{Cov}(X, X)+2\text{Cov}(X, Y)+\text{Cov}(Y, Y) = \text{Var}(X)+\text{Var}(Y)+2\text{Cov}(X, Y).$$

6. If X and Y are independent then $f(x, y) = f_X(x) f_Y(y)$. Therefore

$$\begin{aligned}\text{Cov}(X, Y) &= \int\int (x - \mu_X)(y - \mu_Y) f_X(x) f_Y(y)\, dx\, dy \\ &= \int (x - \mu_X) f_X(x)\, dx \int (y - \mu_Y) f_Y(y)\, dy \\ &= E[X - \mu_X] E[Y - \mu_Y] \\ &= 0.\end{aligned}$$

4.2 Proof of Property 3 of correlation

(*This is for the mathematically interested.*)

$$0 \le \text{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = \text{Var}\left(\frac{X}{\sigma_X}\right) + \text{Var}\left(\frac{Y}{\sigma_Y}\right) - 2\text{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) = 2 - 2\rho$$

This implies $\rho \le 1$

Likewise $0 \le \text{Var}\left(\dfrac{X}{\sigma_X} + \dfrac{Y}{\sigma_Y}\right)$, so $-1 \le \rho$.

If $\rho = 1$ then $0 = \text{Var}\left(\dfrac{X}{\sigma_X} - \dfrac{Y}{\sigma_Y}\right) \Rightarrow \dfrac{X}{\sigma_X} - \dfrac{Y}{\sigma_Y} = c.$ ∎

18.05 Introduction to Probability and Statistics
Spring 2022