## Introduction

In this lesson, we'll briefly introduce basics of statistical inference, the process of drawing conclusions "about a population using noisy statistical data where uncertainty must be accounted for". In other words, statistical inference lets scientists formulate conclusions from data and quantify the uncertainty arising from using incomplete data.

Which of the following is NOT an example of statistical inference?

1: Recording the results of a statistics exam

2: Testing the efficacy of a new drug

3: Constructing a medical image from fMRI data

4: Polling before an election to predict its outcome

So statistical inference involves formulating conclusions using data AND quantifying the uncertainty associated with those conclusions. The uncertainty could arise from incomplete or bad data.

Which of the following would NOT be a source of bad data?

1: A randomly selected sample of population

2: A poorly designed study

3: Small sample size

4: Selection bias

So with statistical inference we use data to draw general conclusions about a population. Which of the following would a scientist using statistical inference techniques consider a problem?

1: Our data sample is representative of the population

2: Our study has no bias and is well-designed

3: Contaminated data

Which of the following is NOT an example of statistical inference in action?

1: Counting sheep

2: Estimating the proportion of people who will vote for a candidate

3: Determining a causative mechanism underlying a disease

4: Testing the effectiveness of a medical treatment

We want to emphasize a couple of important points here. First, a statistic (singular) is a number computed from a sample of data. We use statistics to infer information about a population. Second, a random variable is an outcome from an experiment. Deterministic processes, such as computing means or variances, applied to random variables, produce additional random variables which have their own distributions. It's important to keep straight which distributions you're talking about.

Finally, there are two broad flavors of inference. The first is frequency, which uses "long run proportion of times an event occurs in independent, identically distributed repetitions." The second is Bayesian in which the probability estimate for a hypothesis is updated as additional evidence is acquired. Both flavors require an understanding of probability so that's what the next lessons will cover.

## Probability – Part 1

In this module we discuss probability, the foundation of statistical analysis. Probability assigns a number between 0 and 1 to events to give a sense of the "chance" of the event. Probability has become our default model for apparently random phenomena. Our eventual goal is to use **probability models,** our formal mechanism for connecting our data to a population. However, before we get to probability models, we need to understand the basics of probability calculus.

In this lesson, we'll review basic ideas of probability, the study of quantifying the likelihood of particular events occurring. Note the similarity between the words probability and probably. Every time you use the latter word you're implying that an event is more likely than not to occur.

The first step in understanding probability is to see if you understand what outcomes of an experiment are possible. For instance, if you were rolling a single, fair die once, how many outcomes are possible?

1: Too many

2: 4

3: 6

4: 1

The probability of a particular outcome of an experiment is the ratio of the number of ways that outcome can occur to all possible outcomes of the experiment. Since there are 6 possible outcomes to the experiment of rolling a die, and we assume the die is fair, each outcome is equally likely.

So what is the probability of rolling a 2?

1: 1/6

2: 2/6

3: 0

4: 1/3

What is the probability of rolling an even number?

1: 1

2: 1/2

3: 1/3

4: 0

Since the probability of a particular outcome or event E is the ratio of ways that E could occur to the number of all possible outcomes or events, the probability of E, denoted P(E), is always between 0 and 1. Impossible events have a probability of 0 (since they can't occur) and events that are certain to occur have a probability of 1.

If you're doing an experiment with n possible outcomes, say e1, e2, ..., en, then the sum of the probabilities of all the outcomes is 1. If all the outcomes are equally likely, as in the case of a fair die, then the probability of each is 1/n.

If A and B are two independent events then the probability of them both occurring is the product of the probabilities. P(A&B) = P(A) \* P(B)

Suppose you rolled the fair die twice in succession. What is the probability of rolling two 4's?

1: 2/6

2: 0

3: 1/36

4: 1/6

Suppose you rolled the fair die twice. What is the probability of rolling the same number two times in a row?

1: 2/6

2: 0

3: 1/6

4: 1/36

Since we don't care what the outcome of the first roll is, its probability is 1. The second roll of the dice has to match the outcome of the first, so that has a probability of 1/6. The probability of both events occurring is 1 \* 1/6.

Now consider the experiment of rolling 2 dice, one red and one green. Assume the dice are fair and not loaded. How many distinct outcomes are possible?

1: 1

2: 12

3: 11

4: 36

If an event E can occur in more than one way and these ways are disjoint (mutually exclusive) then P(E) is the sum of the probabilities of each of the ways in which it can occur.

Rolling these two dice, what's the probability of rolling a 10?

1: 2/36

2: 3/36

3: 0

4: 1/6

Since the highest possible dice roll is a '6', the only combinations which give '10' are 4+6 and 5+5. The first occurs in two ways - red dice gets '4' and green gets '6' OR red gets '6' and green gets '4'.

What sum is the most likely when rolling these two dice?

1: 1

2: 7

3: 2

4: 12

5: 9

The choice of '1' is impossible rolling two dice, and '2' and '12' each occur in only one possible way (snake-eyes and double 6's, respectively), so '7' and '9' are the only reasonable choices. To roll '7' you can use any number from '1' to '6', while '9' can only use outcomes of '3' and above.

The probability of at least one of two events, A and B, occurring is the sum of their individual probabilities minus the probability of their intersection. P(A U B) = P(A) + P(B) - P(A&B).

It's easy to see why this is. Calculating P(A) and P(B) counts outcomes that are in both A and B twice, so they're overcounted. The probability of the intersection of the two events, denoted as A&B, must be subtracted from the sum.

Back to rolling two dice. Which expression represents the probability of rolling an even number or a number greater than 8?

1: (18+10-2)/36

2: (18+4-2)/36

3: (18+10-4)/36

4: (18+10)/36

It follows that if A and B are disjoint or mutually exclusive, i.e. only one of them can occur, then P(A U B) = P(A) + P(B) .

Which of the following expressions represents the probability of rolling a number greater than 10?

1: (2+1)/36

2: (1+1)/36

3: (3+1-1)/36

4: (3+1)/36

Use the answer to the previous question and the fact that the sum of all outcomes must sum to 1 to determine the probability of rolling a number less than or equal to 10.

> 33/36

[1] 0.9166667

Now we'll apply the concepts of probability to playing cards.

A deck of cards is a set of 52 cards, 4 suits of 13 cards each. There are two red suits, diamonds and hearts, and two black suits, spades and clubs. Each of the 13 cards in a suit has a value – an ace which is sometimes thought of as 1, a number from 2 to 10, and 3 face cards, king, queen, and jack. We've created a deck in R for you. Type 'deck' to see it now.

> deck

spades hearts diamonds clubs

[1,] "A:spades" "A:hearts" "A:diamonds" "A:clubs"

[2,] "2:spades" "2:hearts" "2:diamonds" "2:clubs"

[3,] "3:spades" "3:hearts" "3:diamonds" "3:clubs"

[4,] "4:spades" "4:hearts" "4:diamonds" "4:clubs"

[5,] "5:spades" "5:hearts" "5:diamonds" "5:clubs"

[6,] "6:spades" "6:hearts" "6:diamonds" "6:clubs"

[7,] "7:spades" "7:hearts" "7:diamonds" "7:clubs"

[8,] "8:spades" "8:hearts" "8:diamonds" "8:clubs"

[9,] "9:spades" "9:hearts" "9:diamonds" "9:clubs"

[10,] "10:spades" "10:hearts" "10:diamonds" "10:clubs"

[11,] "J:spades" "J:hearts" "J:diamonds" "J:clubs"

[12,] "Q:spades" "Q:hearts" "Q:diamonds" "Q:clubs"

[13,] "K:spades" "K:hearts" "K:diamonds" "K:clubs"

When drawing a single card, how many outcomes are possible?

> 52

[1] 52

What is the probability of drawing a jack?

> 4/52

[1] 0.07692308

If you're dealt a hand of 5 cards, what is the probability of getting all 5 of the same value?

> 0

[1] 0

What is the probability of drawing a face card?

> 12/52

[1] 0.2307692

Suppose you draw a face card and don't replace it in the deck. What is the probability that when you draw a second card it also will be a face card?

1: 11/52

2: 0

3: 11/51

4: 12/51

Suppose you draw a face card and don't replace it in the deck. What is the probability that when you draw a second card it also will be a face card of the same suit?

> 1\*2/51

[1] 0.03921569

## Probability – Part 2

Recall that a random variable is a numerical outcome of an experiment. It can be discrete (take on only a countable number of possibilities), or continuous (take on any value on the real line or subset of it).

If you had a ruler of infinite precision, would measuring the height of adults around the world be continuous or discrete?

1: continuous

2: discrete

Is the drawing of a hand of cards continuous or discrete?

1: continuous

2: discrete

Continuous random variables are usually associated with measurements of time, distance, or some biological process since they can take on any value, often within some specified range. Limitations of precision in taking the measurements may imply that the values are discrete; we in fact consider them continuous.

A probability mass function (PMF) gives the probability that a discrete random variable is exactly equal to some value.

For instance, suppose we have a coin which may or may not be fair. Let x=0 represent a 'heads' outcome and x=1 represent a 'tails' outcome of a coin toss. If p is the probability of 'heads' which of the following represents the PMF of the coin toss? The variable x is either 0 (heads) or 1 (tails).

1:

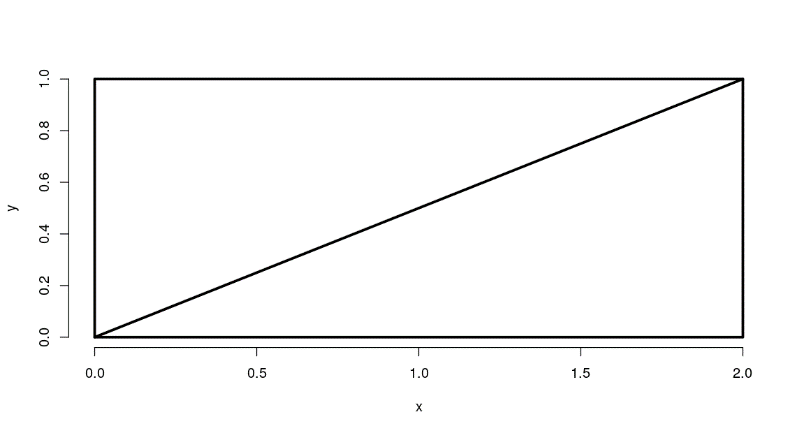
2:

The correct expression has an exponent of 1 for p when x is 0.

A probability density function is associated with a continuous random variable. To quote from Wikipedia, it "is a function that describes the relative likelihood for this random variable to take on a given value. The probability of the random variable falling within a particular range of values is given by ... the area under the density function but above the horizontal axis and between the lowest and greatest values of the range."

We'll repeat two requirements of a probability density function. It must be nonnegative everywhere, and the area under it must equal one."

Consider this figure - a rectangle with height 1 and width 2 with a diagonal line drawn from the lower left corner (0,0) to the upper right (2,1). The area of the entire rectangle is 2 and elementary geometry tells us that the diagonal divides the rectangle into 2 equal areas.

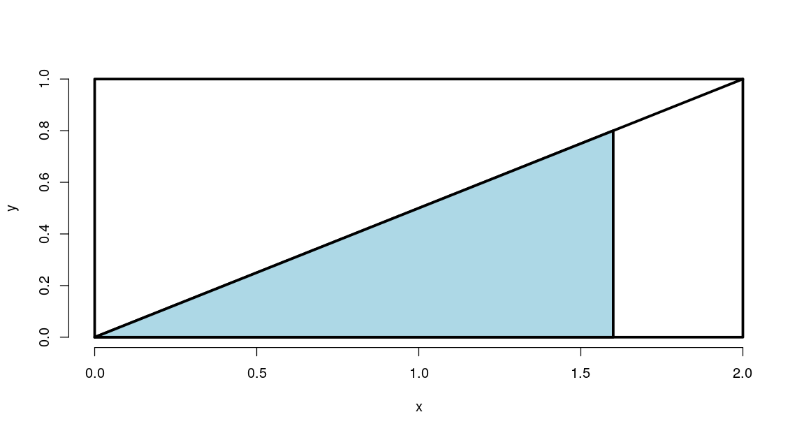


Could the diagonal line represent a probability density function for a random variable with a range of values between 0 and 2? Assume the lower side of the rectangle is the x axis.

1: No

2: Yes

Now consider the shaded portion of the triangle - a smaller triangle with a base of length 1.6 and height determined by the diagonal.



We'll answer the question, "What proportion of the big triangle is shaded?"

This proportion represents the probability that throwing a piece of cat kibble at the bigger triangle (below the diagonal) hits the blue portion.

We have to compute the area of the blue triangle. (Since the area of the big triangle is 1, the area of the blue triangle is the proportion of the big triangle that is shaded.) We know the base, but what is its height?

1: .25

2: I can't tell

3: .8

4: .5

What is the area of the blue triangle?

> 1.6\*0.8/2

[1] 0.64

So, what is the probability that the kibble we throw at the bigger triangle will hit the blue portion?

> 0.64

[1] 0.64

This artificial example was meant to illustrate a simple probability density function (PDF). Most PDFs have underlying formulae more complicated than lines. We'll see more of these in future lessons.

The cumulative distribution function (CDF) of a random variable X, either discrete or continuous, is the function F(x) equal to the probability that X is less than or equal to x. In the example above, the area of the blue triangle represents the probability that the random variable was less than or equal to the value 1.6.

In the triangle example from above, which of the following expressions represents F(x), the CDF?

1: x^2

2: x\*x/4

3: x\*x/2

4: x\*2x/2

If you're familiar with calculus you might recognize that when you're computing areas under curves you're actually integrating functions.

When the random variable is continuous, as in the example, the PDF is the derivative of the CDF. So integrating the PDF (the line represented by the diagonal) yields the CDF. When you evaluate the CDF at the limits of integration the result is an area.

To see this in the example, we've defined the function mypdf for you. This is the equation of the line represented by the diagonal of the rectangle. As the PDF, it is the derivative of F(x), the CDF. Look at mypdf now.

> mypdf

function(x){x/2}

<environment: 0x5646acf510c0>

Now use the R function integrate to integrate mypdf with the parameters lower equal to 0 and upper equal to 1.6. See if you get the same area (probability) you got before.

> integrate(mypdf, 0, 1.6)

0.64 with absolute error < 7.1e-15

The survivor function S(x) of a random variable X is defined as the function of x equal to the probability that the random variable X is greater than the value x. This is the complement of the CDF F(x), in our example, the portion of the lower triangle that is not shaded.

In our example, which of the following expressions represents the survival function?

1: 1-x^2

2: 1-x\*x/4

3: 1-x\*x/2

4: 1-x\*2x/2

The quantile v of a CDF is the point x\_v at which the CDF has the value v. More precisely, F(x\_v)=v. A percentile is a quantile in which v is expressed as a percentage.

What percentile is the median?

1: 25-th

2: I can't tell

3: 50-th

4: 95-th

What is the 50th percentile of the CDF F(x)=(x^2)/4 from the example above?

> 2^0.5

[1] 1.414214

What does this mean with respect to the kibble we're tossing at the triangle?

1: Half of it falls to the left of 1.41

2: All of it falls to the left of 1.41

3: All of it falls on the vertical line at 1.41

4: All of it falls to the right of 1.41

We'll close by repeating some important points.

A probability model connects data to a population using assumptions.

Be careful to distinguish between population medians and sample medians.

A sample median is an estimator of a population median (the estimand).

## Conditional Probability

Conditional probability is a very intuitive idea, "What is the probability given partial information about what has occurred?". The probability of getting hit by lightning is small. However, it's much larger for people playing outside in open fields [during a lightning storm](https://xkcd.com/795/)!

If you were given a fair die and asked what the probability of rolling a 3 is, what would you reply?

1: 1

2: 1/4

3: 1/2

4: 1/3

5: 1/6

Suppose the person who gave you the dice rolled it behind your back and told you the roll was odd. Now what is the probability that the roll was a 3?

1: 1/4

2: 1/3

3: 1

4: 1/6

5: 1/2

The probability of this second event is conditional on this new information, so the probability of rolling a 3 is now one third.

We represent the conditional probability of an event A given that B has occurred with the notation P(A|B). More specifically, we define the conditional probability of event A, given that B has occurred with the following.

P(A|B) = P(A & B)/ P(B) . P(A|B) is the probability that BOTH A and B occur divided by the probability that B occurs.

Back to our dice example. Which of the following expressions represents P(A&B), where A is the event of rolling a 3 and B is the event of the roll being odd?

1: 1

2: 1/3

3: 1/4

4: 1/2

5: 1/6

Continuing with the same dice example. Which of the following expressions represents P(A&B)/P(B), where A is the event of rolling a 3 and B is the event of the roll being odd?

1: 1/6

2: (1/6) / (1/2)

3: (1/3) / (1/2)

4: (1/2) / (1/6)

From the definition of P(A|B), we can write P(A&B) = P(A|B) \* P(B), right? Let's use this to express P(B|A).

P(B|A) = P(B&A) / P(A) = P(A|B) \* P(B) / P(A). This is a simple form of Bayes' Rule which relates the two conditional probabilities.

Suppose we don't know P(A) itself, but only know its conditional probabilities, that is, the probability that it occurs if B occurs and the probability that it occurs if B doesn't occur. These are P(A|B) and P(A|~B), respectively. We use ~B to represent 'not B' or 'B complement'.

We can then express P(A) = P(A|B) \* P(B) + P(A|~B) \* P(~B) and substitute this is into the denominator of Bayes' Formula.

P(B|A) = P(A|B) \* P(B) / ( P(A|B) \* P(B) + P(A|~B) \* P(~B) )

Bayes' Rule has applicability to medical diagnostic tests. Suppose we know the accuracy rates of the test for both the positive case (positive result when the patient has HIV) and negative (negative test result when the patient doesn't have HIV). These are referred to as test sensitivity and specificity, respectively.

Let 'D' be the event that the patient has HIV, and let '+' indicate a positive test result and '-' a negative. What information do we know? Recall that we know the accuracy rates of the HIV test.

1: P(+|D) and P(-|~D)

2: P(+|~D) and P(-|~D)

3: P(+|~D) and P(-|D)

4: P(+|D) and P(-|D)

Suppose a person gets a positive test result and comes from a population with a HIV prevalence rate of .001. We'd like to know the probability that he really has HIV. Which of the following represents this?

1: P(D|-)

2: P(+|D)

3: P(~D|+)

4: P(D|+)

By Bayes' Formula, P(D|+) = P(+|D) \* P(D) / ( P(+|D) \* P(D) + P(+|~D) \* P(~D) )

We can use the prevalence of HIV in the patient's population as the value for P(D). Note that since  
P(~D) = 1-P(D) and P(+|~D) = 1-P(-|~D) we can calculate P(D|+). In other words, we know values for all the terms on the right side of the equation. Let's do it!

Disease prevalence is .001. Test sensitivity (+ result with disease) is 99.7% and specificity (- result without disease) is 98.5%.

First compute the numerator, P(+|D) \* P(D). (This is also part of the denominator.)

> 0.997\*0.001

[1] 0.000997

Now solve for the remainder of the denominator, P(+|~D)\*P(~D).

> (1-0.985)\*(1-0.001)

[1] 0.014985

Now put the pieces together to compute the probability that the patient has the disease given his positive test result, P(D|+). Plug your last two answers into the formula P(+|D) \* P(D) / ( P(+|D) \* P(D) + P(+|~D) \* P(~D) ) to compute P(D|+).

> 0.000997/(0.000997+0.014985)

[1] 0.06238268

So the patient has a 6% chance of having HIV given this positive test result. The expression P(D|+) is called the positive predictive value. Similarly, P(~D|-), is called the negative predictive value, the probability that a patient does not have the disease given a negative test result.

The diagnostic likelihood ratio of a positive test, DLR\_+, is the ratio of the two + conditional probabilities, one given the presence of disease and the other given the absence. Specifically, DLR\_+ = P(+|D) / P(+|~D). Similarly, the DLR\_- is defined as a ratio. Which of the following do you think represents the DLR\_-?

1: P(+|~D) / P(-|D)

2: I haven't a clue.

3: P(-|D) / P(+|~D)

4: P(-|D) / P(-|~D)

Recall that P(+|D) and P(-|~D), (test sensitivity and specificity respectively) are accuracy rates of a diagnostic test for the two possible results. They should be close to 1 because no one would take an inaccurate test, right? Since DLR\_+ = P(+|D) / P(+|~D) we recognize the numerator as test sensitivity and the denominator as the complement of test specificity.

Since the numerator is close to 1 and the denominator is close to 0 do you expect DLR\_+ to be large or small?

1: I haven't a clue.

2: Large

3: Small

Now recall that DLR\_- = P(-|D) / P(-|~D). Here the numerator is the complement of sensitivity and the denominator is specificity. From the arithmetic and what you know about accuracy tests, do you expect DLR\_- to be large or small?

1: I haven't a clue.

2: Small

3: Large

Now a little more about likelihood ratios. Recall Bayes Formula. P(D|+) = P(+|D) \* P(D) / ( P(+|D) \* P(D) + P(+|~D) \* P(~D) ) and notice that if we replace all occurrences of 'D' with '~D', the denominator doesn't change. This means that if we formed a ratio of P(D|+) to P(~D|+) we'd get a much simpler expression (since the complicated denominators would cancel each other out). Like this....

P(D|+) / P(~D|+) = P(+|D) \* P(D) / (P(+|~D) \* P(~D)) = P(+|D)/P(+|~D) \* P(D)/P(~D).

The left side of the equation represents the post-test odds of disease given a positive test result. The equation says that the post-test odds of disease equals the pre-test odds of disease (that is, P(D)/P(~D) ) times

1: I haven't a clue.

2: the DLR\_-

3: the DLR\_+

In other words, a DLR\_+ value equal to N indicates that the hypothesis of disease is N times more supported by the data than the hypothesis of no disease.

Taking the formula above and replacing the '+' signs with '-' yields a formula with the DLR\_-. Specifically, P(D|-) / P(~D|-) = P(-|D) / P(-|~D) \* P(D)/P(~D). As with the positive case, this relates the odds of disease post-test, P(D|-) / P(~D|-), to those of disease pre-test, P(D)/P(~D).

The equation P(D|-) / P(~D|-) = P(-|D) / P(-|~D) \* P(D)/P(~D) says what about the post-test odds of disease relative to the pre-test odds of disease given negative test results?

1: I haven't a clue.

2: post-test odds are greater than pre-test odds

3: post-test odds are less than pre-test odds

Let's cover some basics now.

Two events, A and B, are independent if they have no effect on each other. Formally, P(A&B) = P(A)\*P(B). It's easy to see that if A and B are independent, then P(A|B)=P(A). The definition is similar for random variables X and Y.

We've seen examples of independence in our previous probability lessons. Let's review a little. What's the probability of rolling a '6' twice in a row using a fair die?

1: 1/36

2: 1/2

3: 2/6

4: 1/6

You're given a fair die and asked to roll it twice. What's the probability that the second roll of the die matches the first?

1: 2/6

2: 1/6

3: 1/2

4: 1/36

If the chance of developing a disease with a genetic or environmental component is p, is the chance of both you and your sibling developing that disease p\*p?

1: No

2: Yes

The events aren't independent since genetic or environmental factors likely will affect the outcome.

We'll conclude with iid. Random variables are said to be iid if they are independent and identically distributed. By independent we mean "statistically unrelated from one another". Identically distributed means that "all have been drawn from the same population distribution".

Random variables which are iid are the default model for random samples and many of the important theories of statistics assume that variables are iid. We'll usually assume our samples are random and variables are iid.

## Expectations

The empirical average is a very intuitive idea; it's the middle of our data in a sense. But, what is it estimating? We can formally define the middle of a population distribution. This is the expected value. Expected values are very useful for characterizing populations and usually represent the first thing that we're interested in estimating.

The expected value of a random variable X, E(X), is a measure of its central tendency. For a discrete random variable X with PMF p(x), E(X) is defined as a sum, over all possible values x, of the quantity x\*p(x). E(X) represents the center of mass of a collection of locations and weights, {x, p(x)}.

Another term for expected value is mean. Recall your high school definition of arithmetic mean (or average) as the sum of a bunch of numbers divided by the number of numbers you added together. This is consistent with the formal definition of E(X) if all the numbers are equally weighted.

Consider the random variable X representing a roll of a fair dice. By 'fair' we mean all the sides are equally likely to appear. What is the expected value of X?

> 3.5

[1] 3.5

We've defined a function for you, expect\_dice, which takes a PMF as an input. For our purposes, the PMF is a 6-long array of fractions. The i-th entry in the array represents the probability of i being the outcome of a dice roll. Look at the function expect\_dice now.

> expect\_dice

function(pmf){ mu <- 0; for (i in 1:6) mu <- mu + i\*pmf[i]; mu}

<environment: 0x5646b5cf7ec0>

We've also defined PMFs for three dice, dice\_fair, dice\_high and dice\_low. The last two are loaded, that is, not fair. Look at dice\_high now.

> dice\_high

[1] 0.04761905 0.09523810 0.14285714 0.19047619 0.23809524 0.28571429

Using the function expect\_dice with dice\_high as its argument, calculate the expected value of a roll of dice\_high.

> expect\_dice(dice\_high)

[1] 4.333333

See how the expected value of dice\_high is higher than that of the fair dice. Now calculate the expected value of a roll of dice\_low.

> expect\_dice(dice\_low)

[1] 2.666667

You can see the effect of loading the dice on the expectations of the rolls. For high-loaded dice the expected value of a roll (on average) is 4.33 and for low-loaded dice 2.67. We've stored these off for you in two variables, edh and edl. We'll need them later.

One of the nice properties of the expected value operation is that it's linear. This means that, if c is a constant, then E(cX) = c\*E(X). Also, if X and Y are two random variables then E(X+Y)=E(X)+E(Y). It follows that E(aX+bY)=aE(X)+bE(Y).

Suppose you were rolling our two loaded dice, dice\_high and dice\_low. You can use this linearity property of expectation to compute the expected value of their average. Let X\_hi and X\_lo represent the respective outcomes of the dice roll. The expected value of the average is E((X\_hi + X\_lo)/2) or .5 \*( E(X\_hi)+E(X\_lo) ). Compute this now. Remember we stored the expected values in edh and edl.

> 0.5\*(edh+edl)

[1] 3.5

Did you expect that?

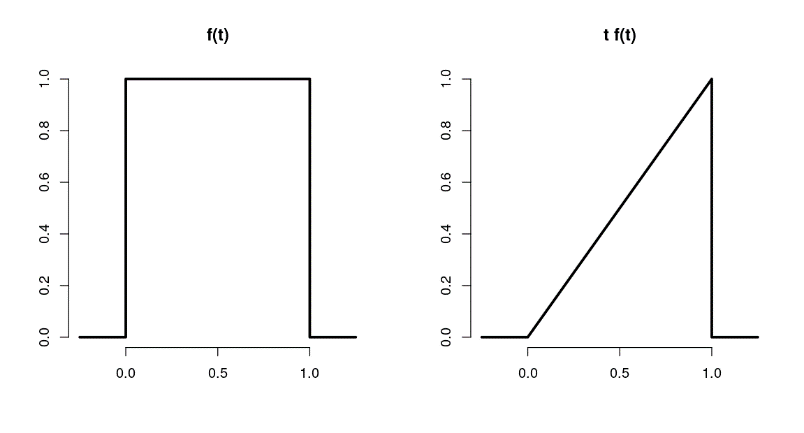
1: Yes

2: No

For a continuous random variable X, the expected value is defined analogously as it was for the discrete case. Instead of summing over discrete values, however, the expectation integrates over a continuous function.

It follows that for continuous random variables, E(X) is the area under the function t\*f(t), where f(t) is the PDF (probability density function) of X. This definition borrows from the definition of center of mass of a continuous body.

The figure below shows the constant (1) PDF on the left and the graph of t\*f(t) on the right.



Knowing that the expected value is the area under the triangle, t\*f(t), what is the expected value of the random variable with this PDF?

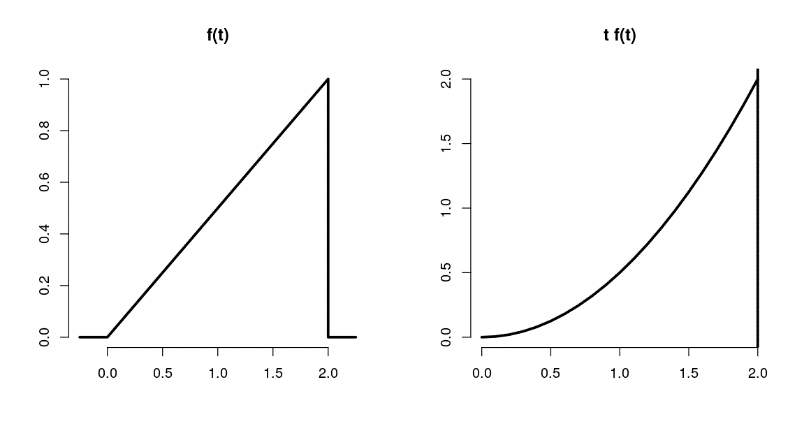
1: .5

2: 2.0

3: .25

4: 1.0

For the purposes of illustration, here's another figure using a PDF from our previous probability lesson. It shows the triangular PDF f(t) on the left and the parabolic t\*f(t) on the right. The area under the parabola between 0 and 2 represents the expected value of the random variable with this PDF.



To find the expected value of this random variable you need to integrate the function t\*f(t). Here f(t)=t/2, the diagonal line. (You might recall this from the last probability lesson.) The function you're integrating over is therefore t^2/2. We've defined a function myfunc for you representing this. You can use the R function 'integrate' with parameters myfunc, 0 (the lower bound), and 2 (the upper bound) to find the expected value. Do this now.

> integrate(myfunc, 0, 2)

1.333333 with absolute error < 1.5e-14

As all the examples have shown, expected values of distributions are useful in characterizing them. The mean characterizes the central tendency of the distribution. However, often populations are too big to measure, so we have to sample them and then we have to use sample means. That's okay because sample expected values estimate the population versions. We'll show this first with a very simple toy and then with some simple equations.

We've defined a small population of 5 numbers for you, spop. Look at it now.

> spop

[1] 1 4 7 10 13

The R function mean will give us the mean of spop. Do this now.

> mean(spop)

[1] 7

Suppose spop were much bigger and we couldn't measure its mean directly and instead had to sample it with samples of size 2. There are 10 such samples, right? We've stored this for you in a 10 x 2 matrix, allsam. Look at it now.

> allsam

[,1] [,2]

[1,] 1 4

[2,] 1 7

[3,] 1 10

[4,] 1 13

[5,] 4 7

[6,] 4 10

[7,] 4 13

[8,] 7 10

[9,] 7 13

[10,] 10 13

Each of these 10 samples will have a mean, right? We can use the R function apply to calculate the mean of each row of the matrix allsam. We simply call apply with the arguments allsam, 1, and mean. The second argument, 1, tells 'apply' to apply the third argument 'mean' to the rows of the matrix. Try this now.

> apply(allsam, 1, mean)

[1] 2.5 4.0 5.5 7.0 5.5 7.0 8.5 8.5 10.0 11.5

You can see from the resulting vector that the sample means vary a lot, from 2.5 to 11.5, right? Not unexpectedly, the sample mean depends on the sample. However...

... if we take the expected value of these sample means we'll see something amazing. We've stored the sample means in the array smeans for you. Use the R function mean on the array smeans now.

> mean(smeans)

[1] 7

Look familiar? The result is the same as the mean of the original population spop. This is not because the example was specially cooked. It would work on any population. The expected value or mean of the sample mean is the population mean. What this means is that the sample mean is an unbiased estimator of the population mean.

Formally, an estimator e of some parameter v is unbiased if its expected value equals v, i.e., E(e)=v. We can show that the expected value of a sample mean equals the population mean with some simple algebra.

Let X\_1, X\_2, ... X\_n be a collection of n samples from a population with mean mu. The mean of these is (X\_1 + X\_2 + ... + X\_n)/n.

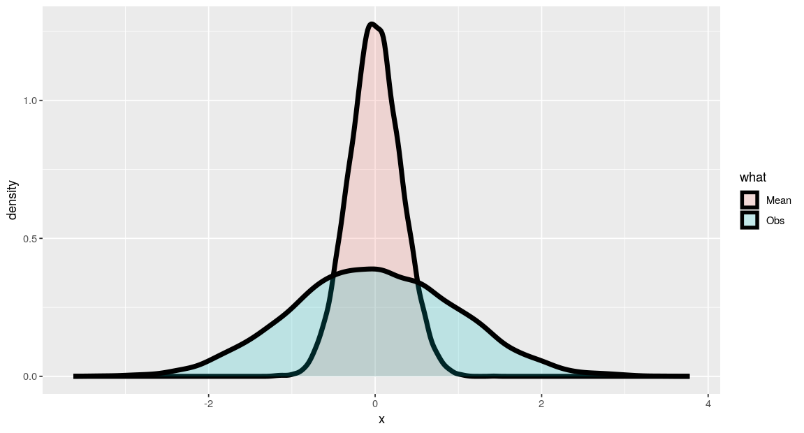
What's the expected value of the mean? Recall that E(aX)=aE(X), so E( (X\_1+..+X\_n)/n ) =

1/n \* (E(X\_1) + E(X\_2) + ... + E(X\_n)) = (1/n)\*n\*mu = mu. Each E(X\_i) equals mu since X\_i is drawn from the population with mean mu.

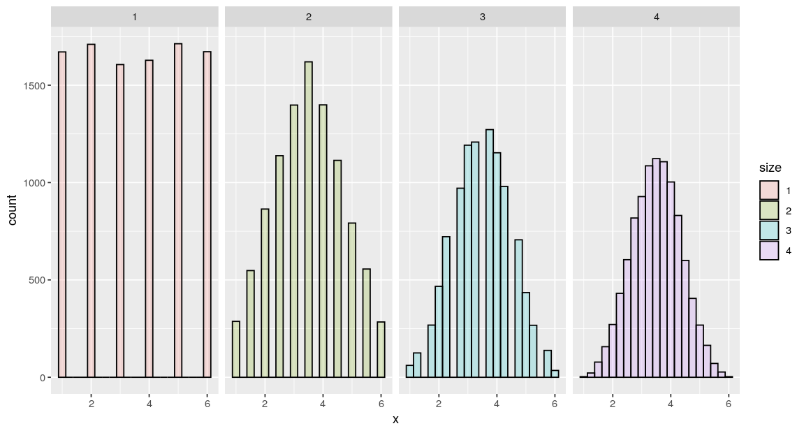
We expect, on average, a random X\_i will equal mu.

Now that was theory. We can also show this empirically with more simulations.

Here's another figure. It shows how a sample mean and the mean of averages spike together. The two shaded distributions come from the same data. The blue portion represents the density function of randomly generated standard normal data, 100000 samples. The pink portion represents the density function of 10000 averages, each of 10 random normals. (The original data was stored in a 10000 x 10 array and the average of each row was taken to generate the pink data.)



And yet another figure. Rolling a single die 10000 times yields the first figure. Each of the 6 possible outcomes appears with about the same frequency. The second figure is the histogram of outcomes of the average of rolling two dice. Similarly, the third figure is the histogram of averages of rolling three dice, and the fourth four dice. As we showed previously, the centre or mean of the original distribution is 3.5 and that's exactly where all the panels are centred.



Let's recap. Expected values are properties of distributions. The average, or mean, of random variables is itself a random variable and its associated distribution itself has an expected value. The centre of this distribution is the same as that of the original distribution.

Now let's review!

Expected values are properties of what?

1: distributions

2: demanding parents

3: fulcrums

4: variances

A population mean is a centre of mass of what?

1: a sample

2: a family

3: a population

4: a distribution

A sample mean is a centre of mass of what?

1: a population

2: observed data

3: a family

4: a distribution

True or False? A population mean estimates a sample mean.

1: True

2: False

True or False? A sample mean is unbiased.

1: False

2: True

True or False? The more data that goes into the sample mean, the more concentrated its density / mass function is around the population mean.

1: False

2: True