THE ULTIMATE GUIDE TO PROBABILITY AND STATISTICS

COMPILED BY HAYAH AHMED

TABLE OF CONTENTS

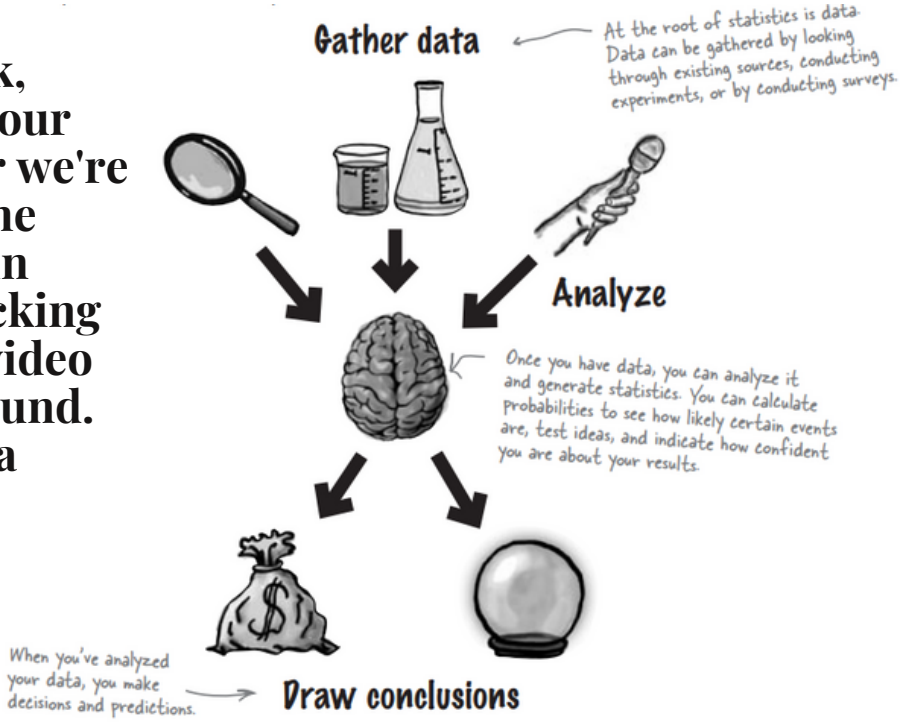
- Introduction to Statistics
- Pie Chart | Bar Chart | Histogram | Line Chart
- Measures of Central Tendency
- Measures of Dispersion
- Box Plot & Whisker Plot
- Probability Basics
- Discrete Probability Distributions
- Permutation & Combination
- Random Variables
- Probability Distributions (e.g., Binomial, Poisson, Normal)
- Expected Value and Variance
- Normal Distributions
- Statistical Sampling
- Estimating populations & samples
- Constructing Confidence Intervals
- Hypothesis testing
- Correlation and Regression Analysis
- Scatter Plot
- Extras

RESOURCES

- https://byjus.com/maths/types-of-statistics/
- all-of-statistics.pdf
- head-first-statistic
- https://www.investopedia.com/terms/p/probabilitydistribution.asp
- https://www.riosalado.edu/web/oer/WRKDEV100-20011_INTER_0000_v1/lessons/Mod05_MeanMedianMode.shtml

Introduction to Statistics

Everywhere we look, statistics permeate our daily lives. Whether we're scrolling through the internet, engaging in sports, or even checking the top scores of a video game, statistics abound. But what exactly is a statistic?



In essence, **statistics** are numbers that condense raw facts and figures into meaningful summaries. They serve to highlight key ideas that may not be immediately discernible from the raw data alone. When we talk about data, we're referring to facts or figures from which conclusions can be drawn. For instance, instead of sifting through countless football scores, we rely on statistics to swiftly provide us with the league position of our favorite team. Statistics offer a quick and efficient means of accessing the information we seek.

- As an example, Two interpretations of the same set of data: profits made by a company in the latter half of last year illustrates how statistics can lead to different conclusions based on interpretation.

The Study of Statistics:

- Empowers individuals to make objective decisions and accurate predictions.
- Enables effective communication of messages.
- Statistics conveniently summarize key truths about data.
- Statistics can be used to tell the truth or to deceive.

Importance of Visualization

Month	Jul	Aug	Sep	Oct	Nov	Dec
Profit (millions)	2.0	2.1	2.2	2.1	2.3	2.4

What we need is some way of visualizing them. If you need to visualize information, there's no better way than using a chart or graph. They can be a quick way of summarizing raw information and can help you get an impression of what's going on at a glance. But you need to be careful because even the simplest chart can be used to subtly mislead and misdirect you.

Types of Statistics:

Descriptive Statistics

Summarizes and describes features of a dataset. Includes measures such as mean, median, mode, range, variance, standard deviation. In this type of statistics, the data is summarised through the given observations Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

Inferential Statistics

Draws conclusions or inferences about a population based on sample data. Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.This type of statistics is used to interpret the meaning of Descriptive statistics.

Inferential statistics can be categorized into several main categories:

- Estimation
- Hypothesis Testing
- Regression Analysis
- Analysis of Variance (ANOVA)
- Non-parametric Methods
- Bayesian Inference

Estimation involves using sample data to approximate population parameters, providing either point estimates or confidence intervals. Hypothesis testing evaluates the significance of observed results through null and alternative hypotheses, test statistics, and p-values. Regression analysis explores relationships between variables, with linear regression for continuous outcomes and logistic regression for categorical outcomes. ANOVA compares means across groups, while non-parametric methods handle non-normal data or categorical variables. Bayesian inference integrates prior beliefs with data to derive posterior probabilities. These methods collectively facilitate drawing conclusions and making predictions about populations based on sample data.

Categories vs. numbers

When you're working with charts, one of the key things you need to figure out is what sort of data you're dealing with. Once you've figured that out, you'll find it easier to make key decisions about what chart you need to best represent your data.

Numerical or quantitative data

Numerical data, on the other hand, deals with numbers. It's data where the values have meaning as numbers, and that involves measurements or counts.

Categorical or qualitative data

The data is split into categories that describe qualities or characteristics. For this reason, it's also called qualitative data. An example of qualitative data is game genre; each genre forms a separate category.

Pie Chart | Bar Chart | Histogram | Line Chart

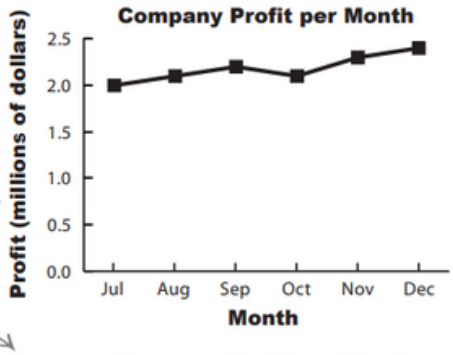
VISUALIZING YOUR DATA: Which chart or graph is right for you?

Transforming data into an effective visualization (any kind of chart or graph) or dashboard is the first step towards making your data make an impact.

Both charts are based on the same underlying data, but they each send a different message.

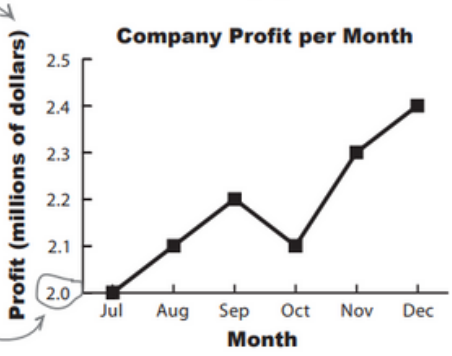
The first chart shows that the profit is relatively steady. It achieves this by having the vertical axis start at 0, and then plotting the profit for each month against this.

Look, the vertical axes are different on each chart.



The second chart gives a different impression by making the vertical axis start at a different place and adjusting the scale accordingly. At a glance, the profits appear to be rising dramatically each month. It's only when you look closer that you see what's really going on.

The axis for this chart starts at 2.0, not 0. No wonder the profit looks so awesome.



Bar Charts

Bar charts are among the most commonly used data visualizations. They offer a clear way to highlight differences between categories, identify trends and outliers, and visualize historical highs and lows.

On a bar chart, each bar represents a particular category, and the length of the bar indicates the value. The longer the bar, the greater the value. All the bars have the same width, which makes it easier to compare them.

Bar charts can allow for more accuracy

Just like pie charts, bar charts allow you to compare relative sizes, but the advantage of using a bar chart is that they allow for a greater degree of precision. They're ideal in situations where categories are roughly the same size, as you can tell with far greater precision which category has the highest frequency. It makes it easier for you to see small differences.

TYPES OF BAR CHARTS

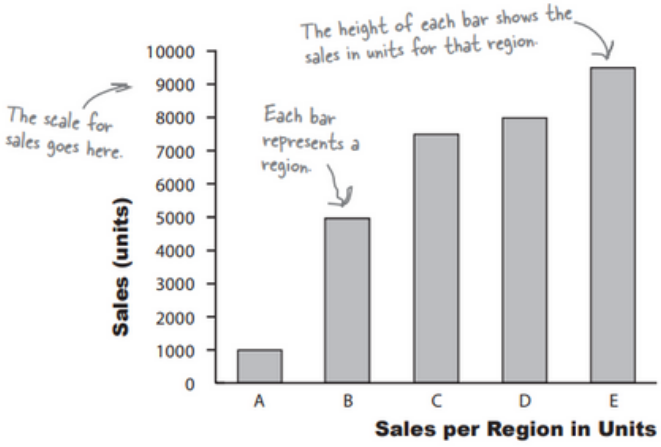
Vertical bar chart

Vertical bar charts show categories on the horizontal axis, and either frequency or percentage on the vertical axis.

The height of each bar shows the sales in units for that region.

The scale for sales goes here.

Each bar represents a region.

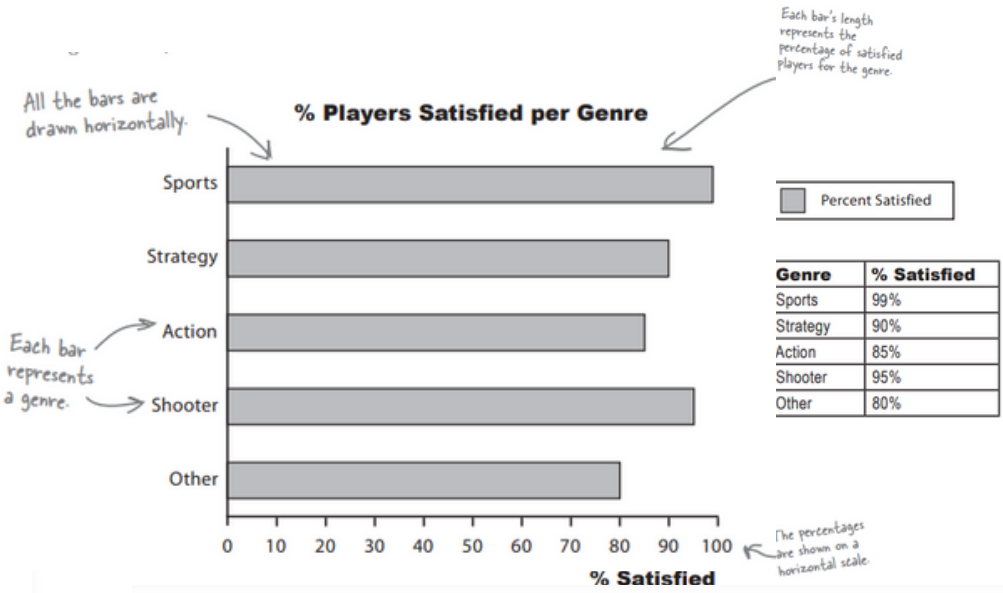


Sales (units)

Region	Sales (units)
A	1,000
B	5,000
C	7,500
D	8,000
E	9,500

Horizontal bar chart

Horizontal bar charts are just like vertical bar charts except that the axes are flipped round. With horizontal bar charts, you show the categories on the vertical axis and the frequency or percentage on the horizontal axis.

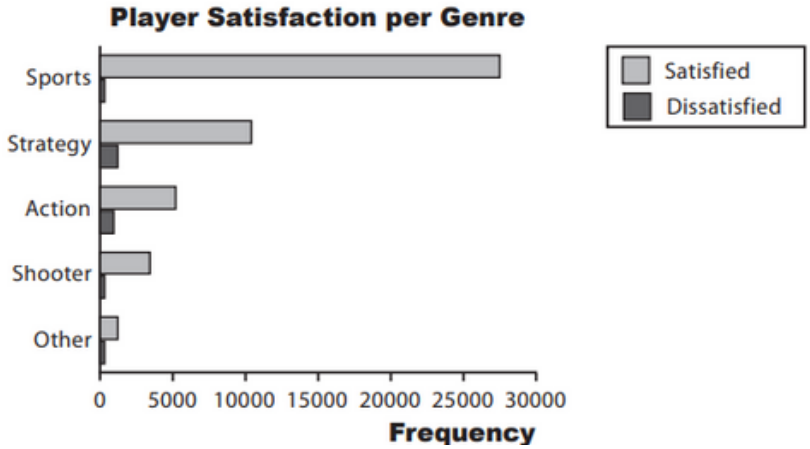


Genre	% Satisfied
Sports	99%
Strategy	90%
Action	85%
Shooter	95%
Other	80%

Grouped Bar Chart

In a grouped bar chart, multiple bars are grouped together side by side, with each group representing a distinct category or subcategory.

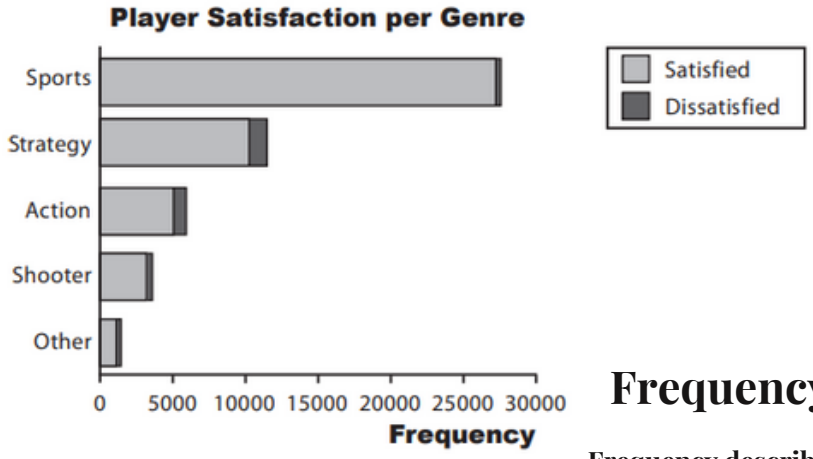
The height of each bar represents the value of a specific category or subcategory, and the different groups allow for easy comparison between categories.



Stacked Bar Chart

In a stacked bar chart, each bar is divided into segments, with each segment representing a different part or category.

This sort of chart is useful if you want to **compare frequencies**, but it's difficult to see proportions and percentages.



Pie Chart

Pie charts work by splitting your data into distinct groups or categories. The chart consists of a circle split into wedge-shaped slices, and each slice represents a group. The size of each slice is proportional to how many are in each group compared with the others. The larger the slice, the greater the relative popularity of that group. The number in a particular group is called the frequency.

Pie charts divide your entire data set into distinct groups. This means that if you add together the frequency of each slice, you should get 100%.

So when are pie charts useful?

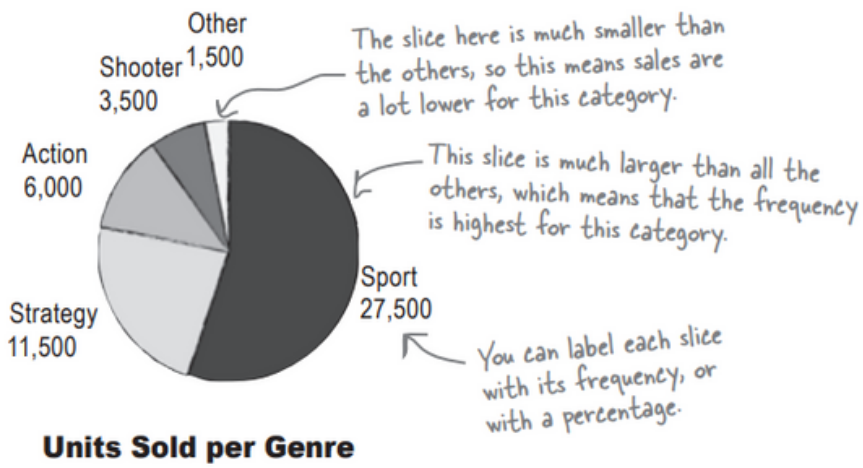
We've seen that the size of each slice represents the relative frequency of each group of data you're showingg. Because of this, pie charts can be useful if you want to compare basic proportions. It's usually easy to tell at a glance which groups have a high frequency compared with the others. Pie charts are less useful if all the slices have similar sizes, as it's difficult to pick up on subtle differences between the slice sizes.

Frequency

Frequency describes how many items there are in a particular group or interval. It's like a count of how many there are.

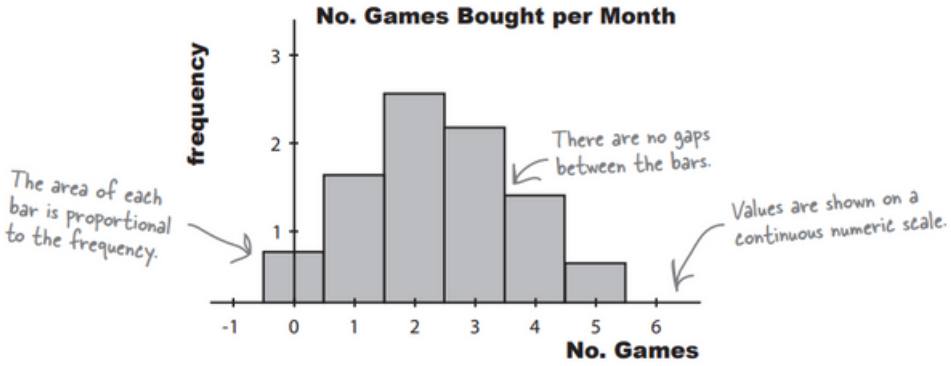
Pie charts show proportions

Genre	Units sold
Sports	27,500
Strategy	11,500
Action	6,000
Shooter	3,500
Other	1,500



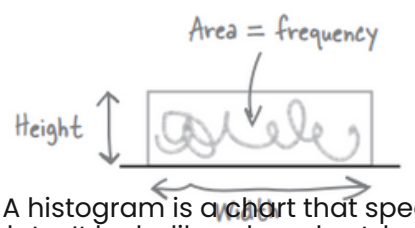
Histogram

Histograms are like bar charts but with two key differences. The first is that the area of each bar is proportional to the frequency, and the second is that there are no gaps between the bars on the chart



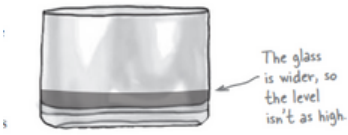
- The **frequency** is a statistical way of saying how many items there are in a category.
- **Pie charts** are good for showing basic proportions.
- **Bar charts** give you more flexibility and precision.
- Horizontal bar charts are used for **categorical data**, particularly where the category names are long.
- Vertical bar charts are used for **numerical data**, or categorical data if the category names are short.
- Frequency density relates to how concentrated the frequencies are for grouped data. It's calculated using

Frequency density = Frequency/ Group width

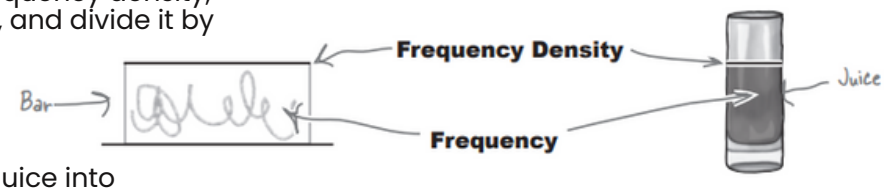


- A histogram is a chart that specializes in grouped data. It looks like a bar chart, but the height of each bar equates to frequency density rather than frequency.
- You can show **multiple sets of data on a bar chart**, and you have a choice of how to do this. You can compare frequencies by showing related bars sideby-side on a split-category bar chart. You can show proportions and total frequencies by stacking the bars on top of each other on a segmented bar chart.
- **Bar chart scales** can show either percentages or frequencies.
- Each chart comes in a number of different varieties.
- When drawing histograms, the width of each bar is proportional to the width of its group.The bars are shown on a continuous numeric scale.
- In a histogram, the frequency of a group is given by the area of its bar.
- A histogram has no gaps between its bars.

Frequency Density



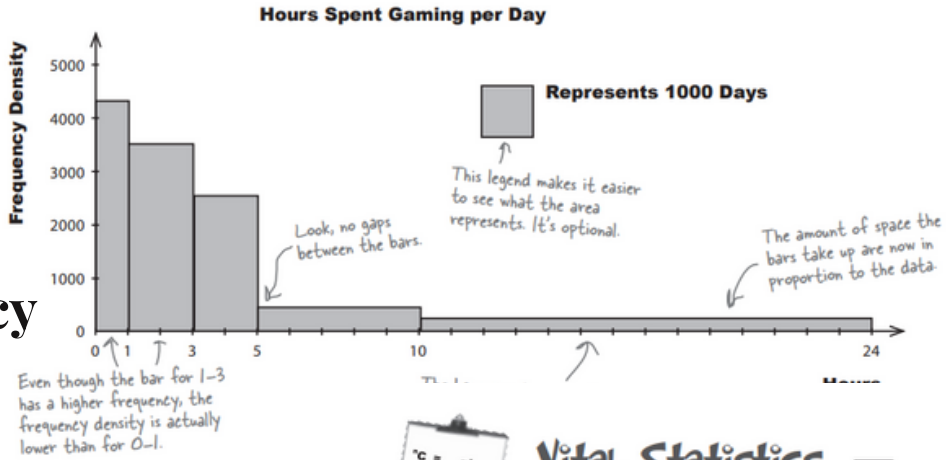
Frequency density is a way of indicating how concentrated values are in a particular interval. It gives you a way of comparing different intervals that may be different widths. It makes the frequency proportional to the area of a bar, rather than height. To find the frequency density, take the frequency of an interval, and divide it by the width.



Juice = Frequency

Imagine that instead of pouring juice into glasses, you're "pouring" frequency into the bars on your chart. Just as you know the width of the glass, you know what width your bars are. And just like the space the juice occupies in the glass (width x height) tells you the quantity of juice in the glass, the area of the bar on the graph is equivalent to its frequency.
The frequency density is then equal to the height of the bar. Keeping with our analogy, it's equivalent to the level your juice comes to in each glass. Just as a wider glass means the juice comes to a lower level, a wider bar means a lower frequency density.

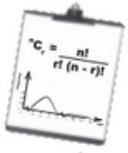
Hours	Frequency
0–1	4,300
1–3	6,900
3–5	4,900
5–10	2,000
10–24	2,100



Cumulative frequency

Imagine if you would want to see at a glance how many people play for less than a certain number of hours. Like, instead of seeing how many people play for between 3 and 5 hours, could we have a graph that shows how many people play for up to 5 hours? Thats where cumulative frequency comes in handy

Cumulative frequencies can never decrease.



Vital Statistics

Cumulative Frequency

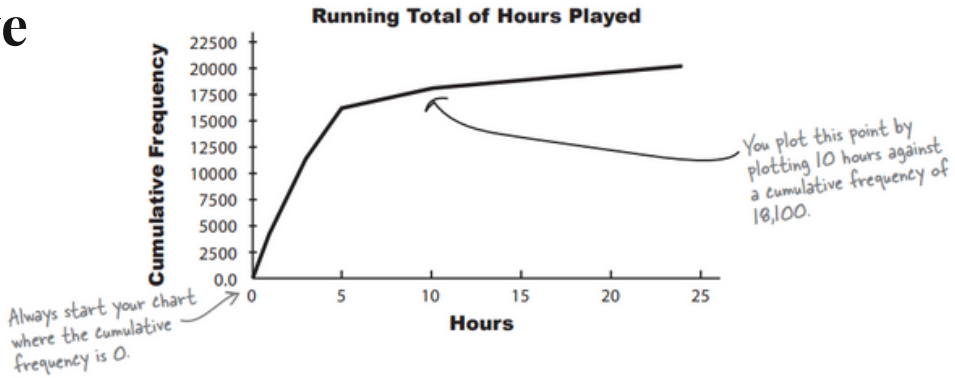
The total frequency up to certain value. It's basically a running total of the frequencies.

Working

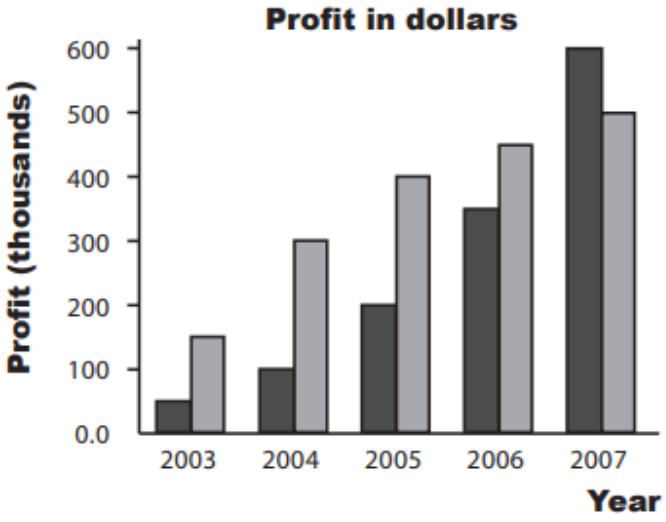
Hours	Frequency	Upper limit	Cumulative frequency
0	0	0	0
0–1	4,300	1	4,300
1–3	6,900	3	4,300+6,900 = 11,200
3–5	4,900	5	4,300+6,900+4,900 = 16,100
5–10	2,000	10	4,300+6,900+4,900+2,000 = 18,100
10–24	2,100	24	4,300+6,900+4,900+2,000+2,100 = 20,200

We've added in 0, as you can't play games for LESS than 0 hours a week.

Cumulative frequency Graphs

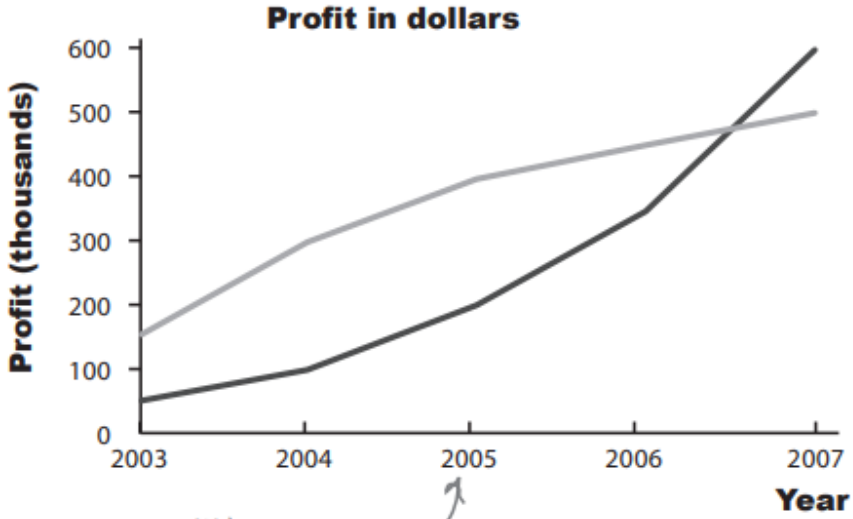


LINE CHART VS BAR CHAT



The bar chart does a good job of comparing the profit on a year-by-year basis, and it's great if you want to compare profits in an individual year. As an example, we can see that up to 2007, the competitor made a bigger profit, but in 2007 Manic Mango did.

A weaknesses of this chart is that if the CEO suddenly decided to add a third competitor, it might make the chart a bit harder to take in at a single glance.



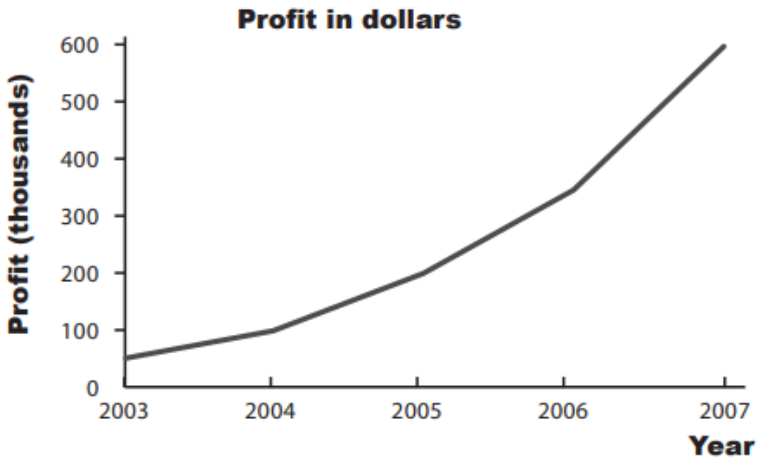
We'd choose the line chart, as the overall trend is clearer than on the bar chart. But don't worry if you chose the other; the chart you use depends on which key facts you want to emphasize.

The line chart is better at showing a trend, the year-on-year profits for each company. The trend line for each company is well-defined, which means we easily see the pattern profits: Manic Mango profits are climbing well, where its competition is beginning to slacken off. It would also be easy to add another company without swamping the chart.

A weakness is that you can also compare year-by-year profit, but perhaps the bar chart is clearer.

LINE CHART

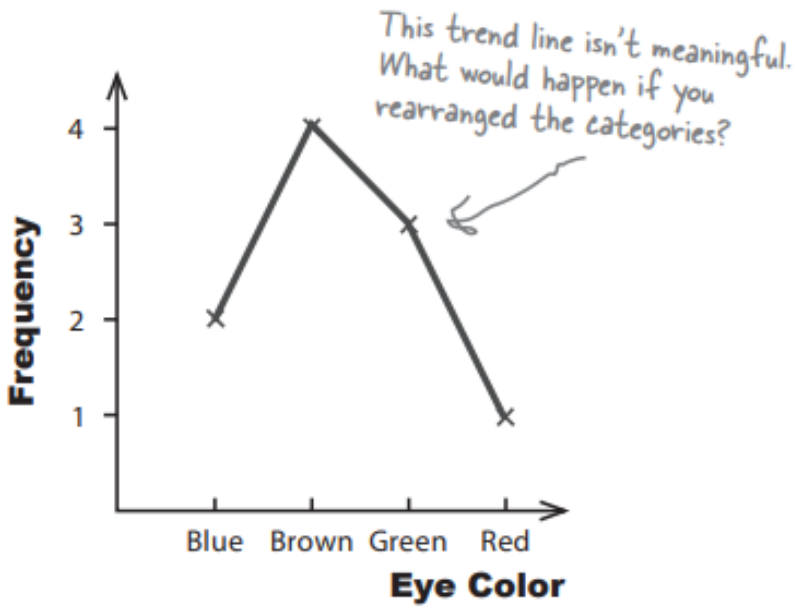
Line charts are good at showing trends in your data. For each set of data, you plot your points and then join them together with lines. You can easily show multiple sets of data on the same chart without it getting too cluttered. Just make sure it's clear which line is which.



Line charts are often used to show time measurements. Time always goes on the horizontal axis, and frequency on the vertical. You can read off the frequency for any period of time by choosing the time value on the horizontal axis, and reading off the corresponding frequency for that point on the line.

- You can show more than one set of data on a line chart. Use one line for each set of data, and make sure it's clear which line is which.
- Don't use line charts to show categorical data unless you're showing trends for each category, for example over time. If you do this, draw one line per category.

Don't use line charts to show categorical data



Measure of Central Tendancy

MEAN

"on average, it takes 20 minutes to drive to work," it means that in general, for most days or trips, the driving time is around 20 minutes. This doesn't mean every single trip will take exactly 20 minutes; some trips might take longer, and some might be shorter. But when we consider all the trips together and calculate the average time, it comes out to be 20 minutes.

So, **mean** is a way of expressing what's typical or expected based on the collective data, even though individual instances may vary.

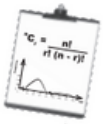
$$\mu = \frac{\Sigma fx}{\Sigma f}$$

Multiply each number by its frequency, then add the results together.

Sum of the frequencies

Handling frequencies

When you calculate the mean of a set of numbers, you'll often find that some of the numbers are repeated.It's really important to make sure that you include the ***frequency*** of each number when you're working out the mean.



Vital Statistics

Outlier

An extreme high or low value that stands out from the rest of the data

CASE STUDY

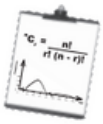
A man in his late fifties who wants an exercise class composed of other middle-aged folks but ended up in the Kung Fu class with lots of young 'uns and a few ancient masters

Power Workout Classmate Ages

Age	19	20	21
Frequency	1	3	1

Kung Fu Classmate Ages

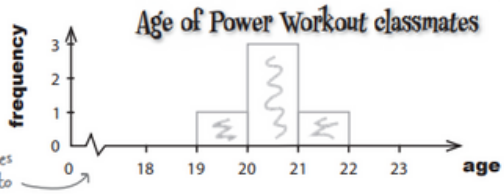
Age	19	20	21	145	147
Frequency	3	6	3	1	1



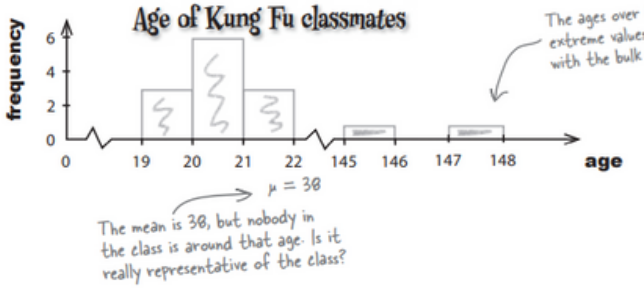
Vital Statistics

Skewed Data

When outliers "pull" the data to the left or right



We left out the ages between 0 and 18 to save space.



The ages over here are outliers. They're extreme values that don't really fit in with the bulk of the data.

The mean is 38, but nobody in the class is around that age. Is it really representative of the class?

Our data has outliers

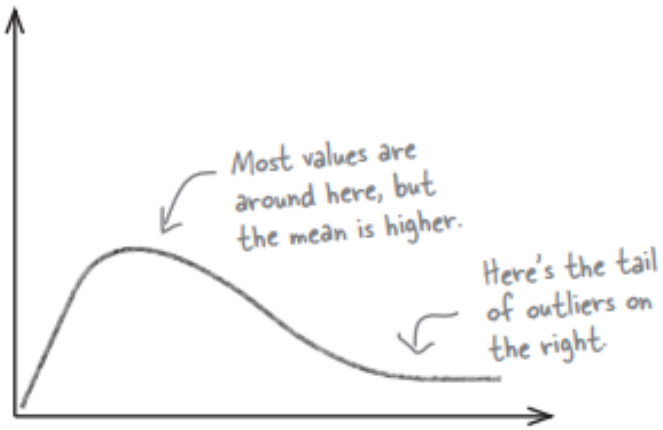
The shape of the chart for the Kung Fu class isn't as straightforward. Most of the ages are around 20, but there are two masters whose ages are much greater than this. Extreme values such as these are called outliers.

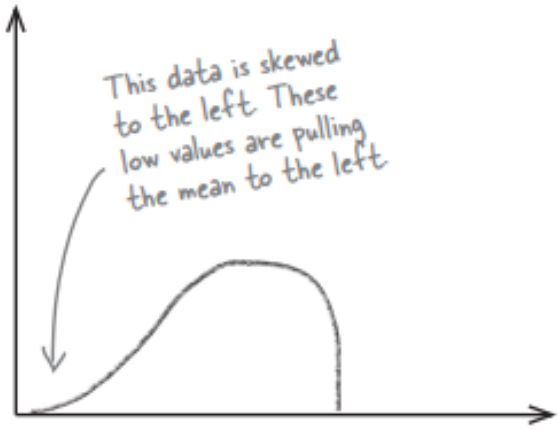
The Kung Fu class data is skewed to the right because if you line the data up in ascending order, the outliers are on the right.We can't just ignore the ancient masters, though; they're still part of the class. Unfortunately, the presence of people who are way above the "typical" age of the class distorts the mean, pulling it upwards

SKEWED DATA

Skewed to the right

Data that is skewed to the right has a "tail" of high outliers that trail off to the right. If you look at a right-skewed chart, you can see this tail. The high outliers in the Kung Fu class data distort the mean, pulling it higher—that is, to the right.



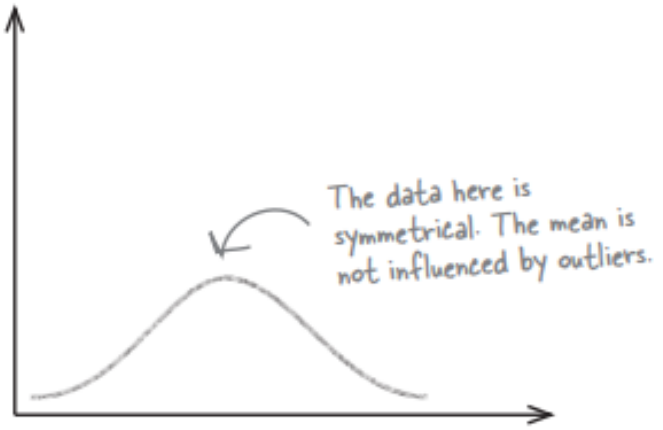


Skewed to the left

Here's a chart showing data that is skewed to the left. Can you see the tail of outliers on the left? This time the outliers are low, and they pull the mean over to the left. In this situation, the mean is lower than the majority of values.

Symmetric data

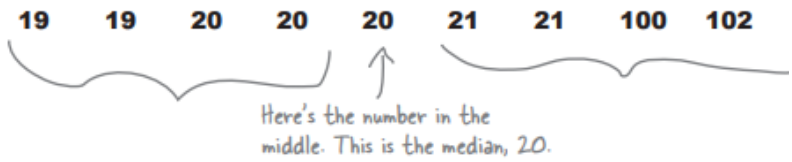
In an ideal world, you'd expect data to be symmetric. If the data is symmetric, the mean is in the middle. There are no outliers pulling the mean in either direction, and the data has about the same shape on either side of the center.



MEDIAN

If the mean becomes misleading because of skewed data and outliers, then we need some other way of saying what a typical value is. We can do this by, quite literally, taking the middle value. This is a different sort of average, and it's called the median.

The median is always in the middle. It's the middle value.



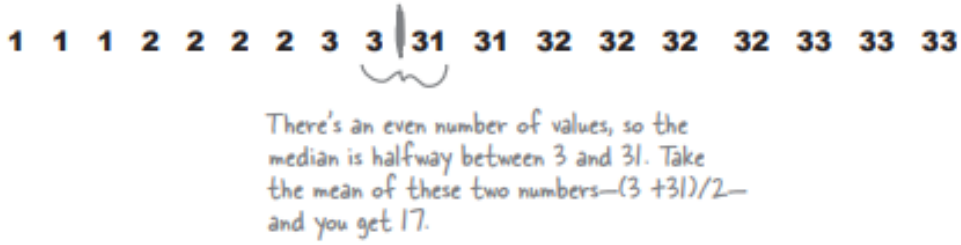
If you have an even set of numbers, just take the mean of the two middle numbers (add them together, and divide by 2), and that's your median. In this case, the median is 20.5.

the whole point of the mean is that it gives a typical value. It's the average. The big danger is that the mean will give a value that doesn't exist in the data set. Take the Kung Fu class as an example. If you were to go into the class and pick a person at random, the chances are that person would be around 20 years old because most people in the class are that sort of age. Just going with the mean doesn't give you that impression. Finding the median can give you a more accurate perspective on the data.

How to find the median in three steps:

1. Line your numbers up in order, from smallest to largest.
2. If you have an odd number of values, the median is the one in the middle. If you have *n* numbers, the middle number is at position $(n + 1) / 2$.
3. If you have an even number of values, get the median by adding the two middle ones together and dividing by 2. You can find the midpoint by calculating $(n + 1) / 2$. The two middle numbers are on either side of this point.

LIMITATIONS OF MEAN AND MEDIAN



The mean and median for the class are both 17, even though there are no 17-year-olds in the class!

But what if there had been an odd number of people in the class. Both the mean and median would still have been misleading. Take a look:

1 1 1 2 2 2 2 2 3 3 31 31 32 32 32 32 33 33 33

If we add another 2-year-old to the class, the median becomes 3. But what about the adults?

If another two-year-old were to join the class, like we see above, the median would still be 3. This reflects the age of the children, but doesn't take the adults into account.

1 1 1 2 2 2 2 2 3 31 31 31 32 32 32 32 33 33 33

If we add another 31-year-old to the class, the median instead becomes 31. This time, we ignore the kids!

If another 33-year-old were added to the class instead, the median would be 31. But that fails to reflect all the kids in the class.

Whichever value we choose for the average age, it seems misleading.

Mode

The mode has to be in the data set. It's the only average that works with categorical data.

What should we do for data like this?

Mode

In addition to the mean and median, there's a third type of average called the mode. The mode of a set of data is the most popular value, the value with the highest frequency. Unlike the mean and median, the mode absolutely has to be a value in the data set, and it's the most frequent value.

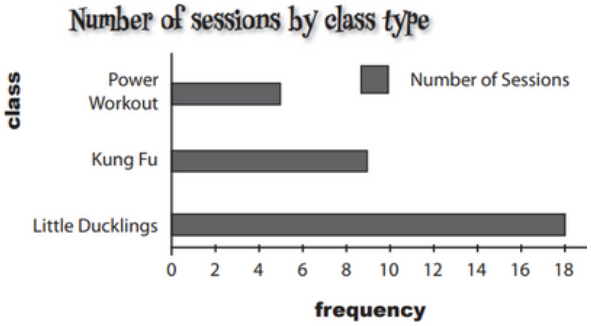
Sometimes data can have more than one mode. If there is more than one value with the highest frequency, then each one of these values is a mode. If the data looks as though it's representing more than one trend or set of data, then we can give a mode for each set. If a set of data has two modes, then we call the data bimodal.

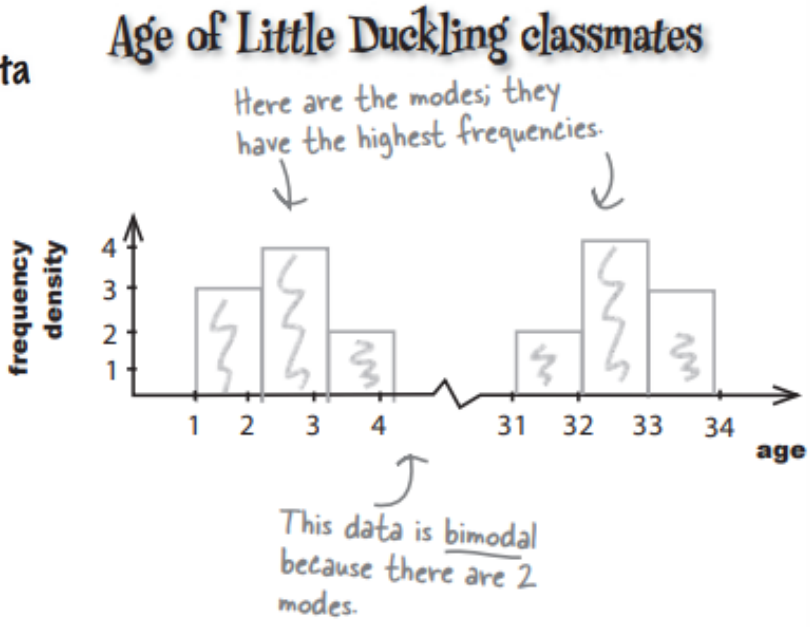
Age	1	2	3	31	32	33
Frequency	3	4	2	2	4	3

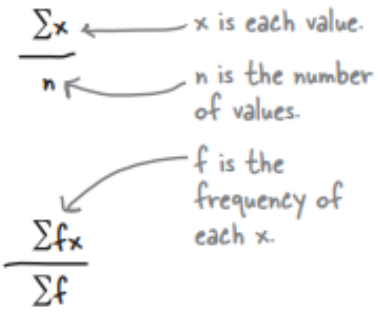
These two values are the most popular, so they are both modes.

It even works with categorical data

When you're dealing with categorical data, the mode is the most frequently occurring category.The category or group with the highest frequency is called the modal class.



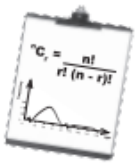


Average	How to calculate	When to use it
Mean (μ)	Use either  	When the data is fairly symmetric and shows just the one trend.
Median	Line up all the values in ascending order. If there are an odd number of values, the median is the one in the middle. If there are an even number of values, add the two middle ones together, and divide by two.	When the data is skewed because of outliers.
Mode	Choose the value(s) with the highest frequency. If the data is showing two clusters of data, report a mode for each group.	When you're working with categorical data. When the data shows two or more clusters of data. The only type of average you can calculate for categorical data is the mode.

Charts	When to Use	Real Life Example Justification
Bar Chart	To compare different categories or groups	Comparing sales performance of different products in a retail store
Line Chart	To show trends or changes over time	Tracking stock prices over a year
Histogram	To display frequency distributions	Showing distribution of ages in a population
Pie Chart	To represent parts of a whole	Illustrating market share of different smartphone brands in a region

Measure of Dispersion

Averages do a great job of giving you a typical value in your data set, but they don't tell you the full story. OK, so you know where the center of your data is, but often the mean, median, and mode alone aren't enough information to go on when you're summarizing a data set. In this chapter, we'll show you how to take your data skills to the next level as we begin to analyze *ranges and variation*.



Vital Statistics

Range

The range is a way of measuring how spread out a set of values are. It's given by

Upper bound – Lower bound

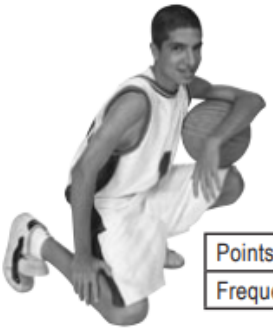
where the upper bound is the highest value, and the lower bound the lowest.

We need to compare player scores

Here are the scores of the three players:

Points scored per game	7	8	9	10	11	12	13
Frequency	1	1	2	2	2	1	1

Here, frequency tells us the number of games where the player got each score. This player scored 9 points in 2 games, and 12 points in 1 game.



Points scored per game	7	9	10	11	13
Frequency	1	2	4	2	1

Points scored per game	3	6	7	10	11	13	30
Frequency	2	1	2	3	1	1	1

Each player has a mean, median, and mode score of 10 points, but if you look at their scores, you'll see they've all achieved it in different ways. There's a difference in how consistently the players have performed, which the average can't measure. What we need is a way of differentiating between the three sets of scores so that we can pick the most suitable player for the team. We need some way of comparing the sets of data in addition to the average

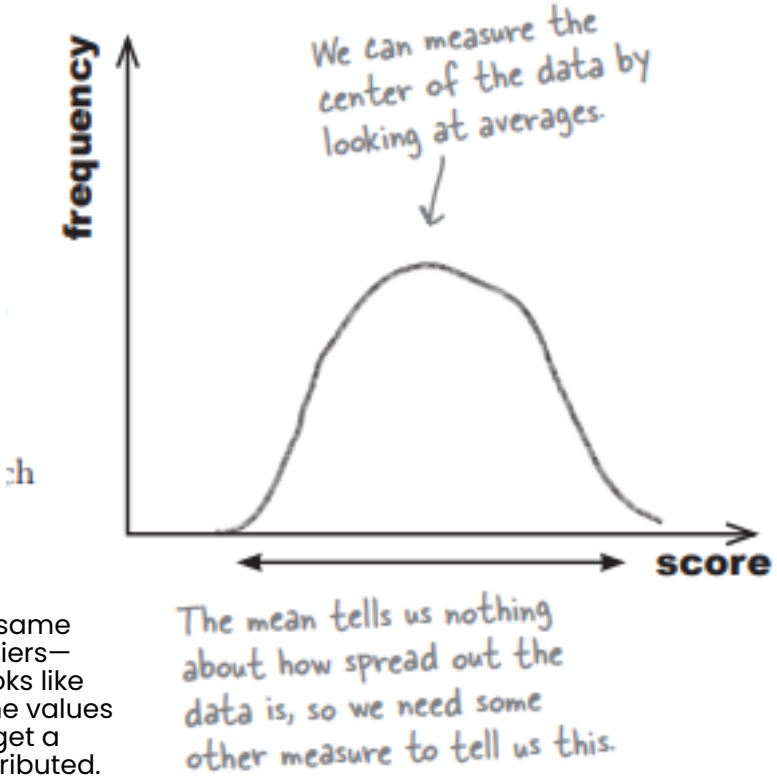
Use the range to differentiate between data sets

Averages give us a way of determining where the center of a set of data is, but they don't tell us how the data varies.

We can differentiate between each set of data by looking at the way in which the scores spread out from the average. Each player's scores are distributed differently, and if we can measure how the scores are dispersed, the coach will be able to make a more informed decision.

The range tells us over how many numbers the data extends, a bit like measuring its width.The range is a simple and easy way of measuring how spread out values are, and it gives us another way of comparing sets of data.

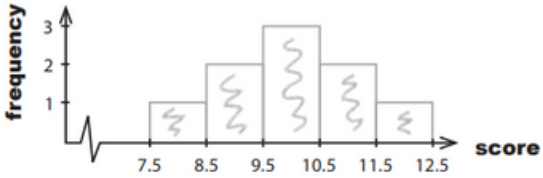
The range only describes the width of the data, not how it's dispersed between the bounds.



Both sets of data above have the same range, but the second set has outliers—extreme high and low values. It looks like the range can measure how far the values are spread out, but it's difficult to get a real picture of how the data is distributed.

The problem with outliers

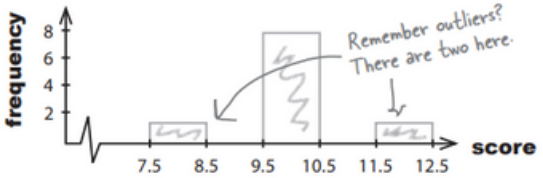
Score	8	9	10	11	12
Frequency	1	2	3	2	1



$\mu = 10$
Lower bound = 8
Upper bound = 12
Range = 12 − 8
= 4

Look, these results are the same even though the data's different.

Score	8	9	10	11	12
Frequency	1	0	8	0	1

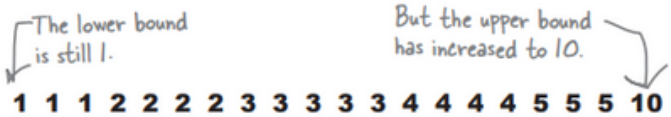


$\mu = 10$
Lower bound = 8
Upper bound = 12
Range = 12 − 8
= 4

The range is a simple way of saying what the spread of a set of data is, but it's often not the best way of measuring how the data is distributed within that range. If your data has outliers, using the range to describe how your values are dispersed can be very misleading because of its sensitivity to outliers.

But what happens if we introduce an outlier, like the number 10?

Imagine you have a set of numbers as follows:



Our lower bound is the same, but the upper bound has gone up to 10, giving us a new range of 9. The range has increased by 5 just because we added one extra number, an outlier.

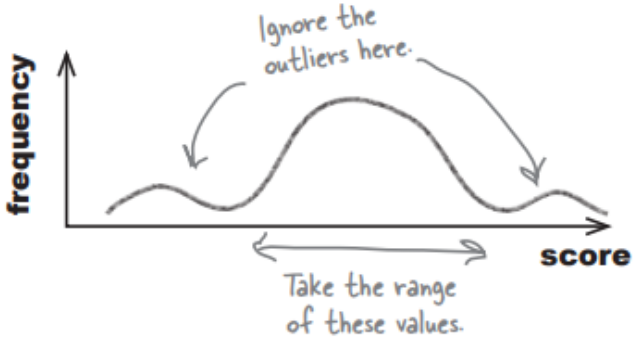
The range is a great quick-and-dirty way to get an idea of how values are distributed, but it's a bit limited.

The range tells you how far apart your highest and lowest values are, but that's about it. It only provides a very basic idea of how the values are distributed.

We need to get away from outliers

SOLUTION 1

One way out of this problem is to look at a kind of mini range, one that ignores the outliers. Instead of measuring the range of the whole set of data, we can find the range of part of it, the part that doesn't contain outliers

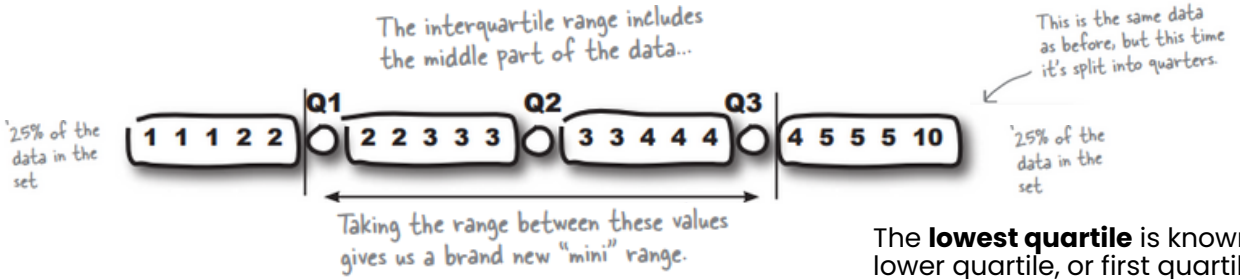


One of the problems with ignoring outliers on an ad hoc basis is that it's difficult to compare sets of data. How do we know that all sets of data are omitting outliers in exactly the same way?

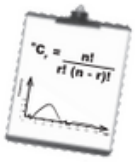
SOLUTION 2

Quartiles come to the rescue

We can construct a range in this way by first lining up the values in ascending order, and then splitting the data into four equally sized chunks, with each chunk containing one quarter of the data.



The values that split the data into equal chunks are known as **quartiles**, as they split the data into quarters. Finding quartiles is a bit like finding the median. Instead of finding the value that splits the data in half, we're finding the values that split the data into quarters.



Vital Statistics

Quartiles

Quartiles are values that split your data into quarters. The lowest quartile is called the lower quartile, and the highest quartile is called the upper quartile.

The middle quartile is the median.

The **lowest quartile** is known as the lower quartile, or first quartile (Q1), and the **highest quartile** is known as the upper quartile, or third quartile (Q3). The quartile in the middle (Q2) is the median, as it splits the data in half. The range of the values in **these two quartiles is called the interquartile range (IQR)**

Interquartile range =

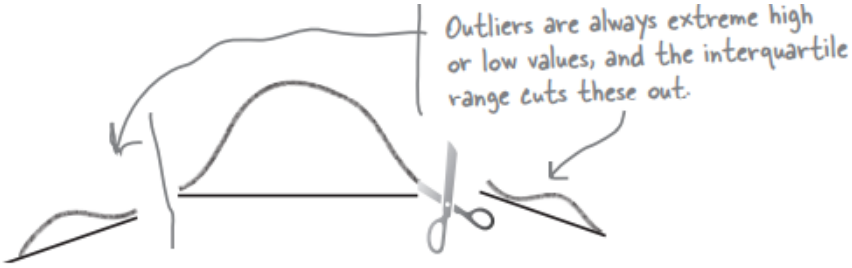
Upper quartile – Lower quartile

The interquartile range gives us a standard, repeatable way of measuring how values are dispersed. It's another way in which we can compare different sets of data.

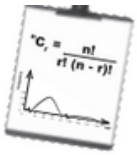
The interquartile range excludes outliers

The upper and lower quartiles are positioned so that the lower quartile has 25% of the data below it, and the upper quartile has 25% of the data above it. This means that the interquartile range only uses the central 50% of the data, so outliers are disregarded. As we've said before, outliers are extreme high or low values in the data, so by only considering values around the center of the data, we automatically exclude any outliers.

As the interquartile range only uses the central 50% of the data, outliers are excluded irrespective of whether they are extremely high or extremely low. They can't be in the middle. This means that any outliers in the data are effectively cut out.



Excluding the outliers with the interquartile range means that we now have a way of comparing different sets of data without our results being distorted by outliers.

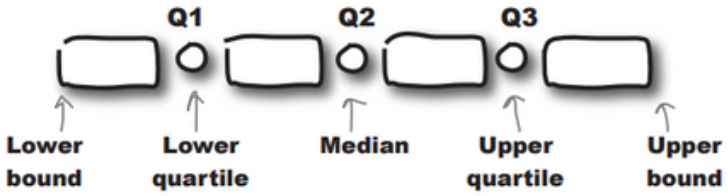


Vital Statistics

Interquartile Range

A "mini range" that's less sensitive to outliers. You find it by calculating

Upper quartile – Lower quartile



Finding the position of the lower quartile

1. First, start off by calculating n ÷ 4.
2. If this gives you an integer, then the lower quartile is positioned halfway between this position and the next one. Take the average of the numbers at these two positions to get your lower quartile.
3. If n ÷ 4 is *not* an integer, then round it up. This gives you the position of the lower quartile.

As an example, if you have 6 numbers, start off by calculating 6 ÷ 4, which gives you 1.5. Rounding this number up gives you 2, which means that the lower quartile is at position 2.

Finding the position of the upper quartile

1. Start off by calculating 3n ÷ 4.
2. If it's an integer, then the upper quartile is positioned halfway between this position and the next. Add the two numbers at these positions together and divide by 2.
3. If 3n ÷ 4 is *not* an integer, then round it up. This new number gives you the position of the upper quartile.

- The upper and lower bounds of the data are the highest and lowest values in the data set.
- The range is a simple way of measuring how values are dispersed. It's given by:

range = upper bound - lower bound

- The range is very sensitive to outliers.
- The interquartile range is less sensitive to outliers than the range.
- Quartiles are values that split your data into quarters. The highest quartile is called the upper quartile, and the lowest quartile is called the lower quartile. The middle quartile is the median.
- The **interquartile range** is the range of the central 50% of the data. It's given by calculating

upper quartile - lower quartile

If you break up a set of data into percentages, the values that split the data are called **percentiles**. In the case above, our data is split into tenths, so the values are called **deciles**. We can use percentiles to construct a new range called the i**nterpercentile range**.

Percentiles

Percentiles are values that split your data into percentages in the same way that quartiles split data into quarters. Each percentile is referred to by the percentage with which it splits the data, so the 10th percentile is the value that is 10% of the way through the data. In general, the xth percentile is the value that is k% of the way through the data. It's usually denoted by Pk .

Quartiles are actually a type of percentile. The lower quartile is P25, and the upper quartile is P75. The median is P50.

Percentile uses

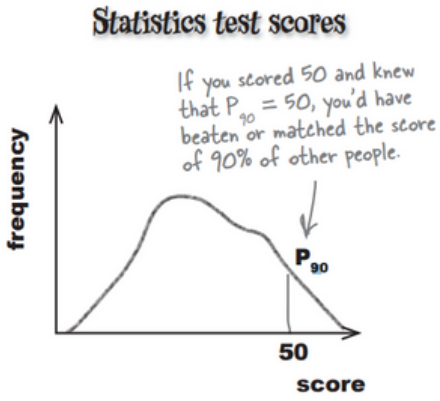
Even though the interpercentile range isn't that commonly used, the percentiles themselves are useful for benchmarking and determining rank or position. They enable you to determine how high a particular value is relative to all the others. As an example, suppose you heard you scored 50 on your statistics test. With just that number by itself, you'd have no idea how well you'd done relative to anyone else. But if you were told that the 90th percentile for the exam was 50, you'd know that you scored the same as or better than 90% of the other people.

Finding percentiles

You can find percentiles in a similar way to how you find quartiles.

1. First of all, line all your values up in ascending order.
2. To find the position of the kth percentile out of n numbers, start off by calculating $k\left(\frac{n}{100}\right)$.
3. If this gives you an integer, then your percentile is halfway between the value at position $k\left(\frac{n}{100}\right)$ and the next number along. Take the average of the numbers at these two positions to give you your percentile.
4. If $k\left(\frac{n}{100}\right)$ is *not* an integer, then round it up. This then gives you the position of the percentile.

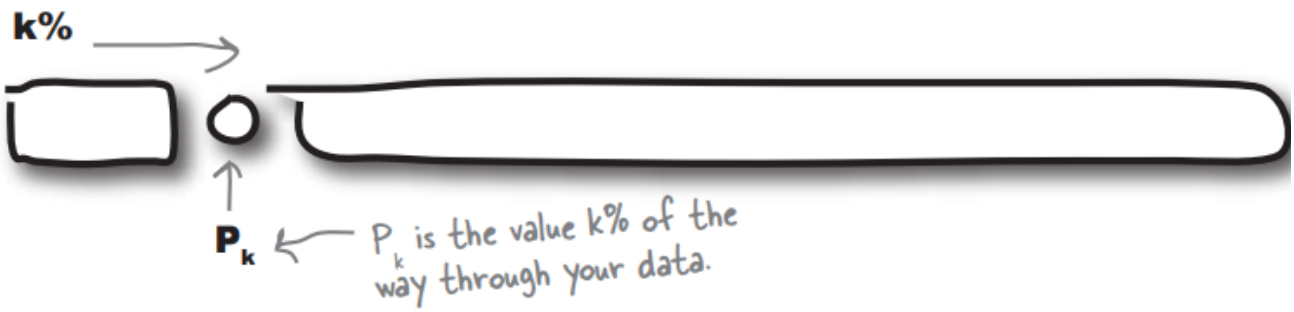
As an example, if you have 125 numbers and want to find the 10th percentile, start off by calculating $10 \times 125 \div 100$. This gives you a value of 12.5. Rounding this number up gives you 13, which means that the 10th percentile is the number at position 13.



Percentile

The kth percentile is the value that's k% of the way through your data. It's denoted by

P_k



How can we more accurately measure variability?

VARIANCE

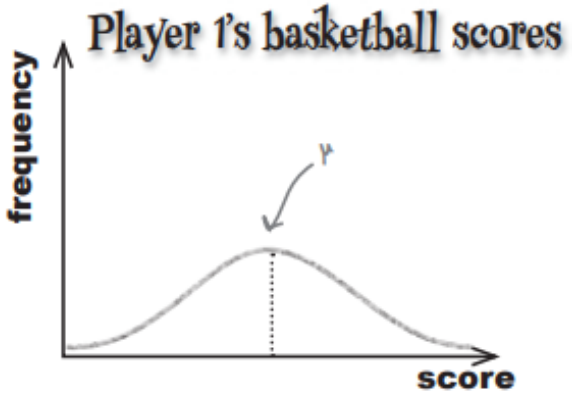
One way of understanding variance is to look at how far away each value is from the mean. The smaller the result, the closer values are to the mean.

Variance

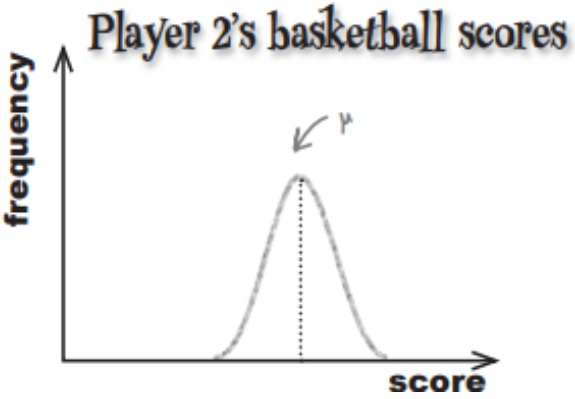
The variance is a way of measuring spread, and it's the average of the distance of values from the mean squared.

The variance is the average of the distances from the mean squared.

$$\text{Variance} = \frac{\Sigma(x - \mu)^2}{n}$$



The values here are spread out quite a long way from the mean. If the coach picks this player for the team, he's unlikely to be able to predict how the player will perform on game day. The player may achieve a very high score if he's having a good day. On a bad day, however, he may not score highly at all, and that means he'll potentially lose the game for the team.



The values for this second set of data are much closer to the mean and vary less. If the coach picks this player, he'll have a good idea of how well the player is likely to perform in each game.

Standard deviation

The problem with the variance is that it can be quite difficult to think about spread in terms of distances squared. There's an easy way to correct this. All we need to do is take the square root of the variance. We call this the *standard deviation.*

The smaller the standard deviation, the closer values are to the mean. The smallest value the standard deviation can take is 0.

$$\sigma = \sqrt{\text{variance}}$$

$$\sigma^2 = \text{variance}$$

BE the coach Solution

Here are the scores for the three players. The mean for each of them is 10. Your job is to play like you're the coach, and work out the standard deviation for each player. Which player is the most reliable one for your team?

Player 1

Score	7	9	10	11	13
Frequency	1	2	4	2	1

$$\text{Variance} = \frac{7^2 + 2(9^2) + 4(10^2) + 2(11^2) + 13^2}{10} - 100$$

$$= \frac{49 + 162 + 400 + 242 + 169}{10} - 100$$

$$= 2.2$$

$$\text{Standard Deviation} = \sqrt{2.2} = 1.48$$

Player 2

Score	7	8	9	10	11	12	13
Frequency	1	1	2	2	2	1	1

$$\text{Variance} = \frac{7^2 + 8^2 + 2(9^2) + 2(10^2) + 2(11^2) + 12^2 + 13^2}{10} - 100$$

$$= \frac{49 + 64 + 162 + 200 + 242 + 144 + 169}{10} - 100$$

$$= 3$$

$$\text{Standard Deviation} = \sqrt{3} = 1.73$$

Player 3

Score	3	6	7	10	11	13	30
Frequency	2	1	2	3	1	1	1

$$\text{Variance} = \frac{2(3^2) + 6^2 + 2(7^2) + 3(10^2) + 11^2 + 13^2 + 30^2}{11} - 100$$

$$= \frac{18 + 36 + 98 + 300 + 121 + 169 + 900}{11} - 100$$

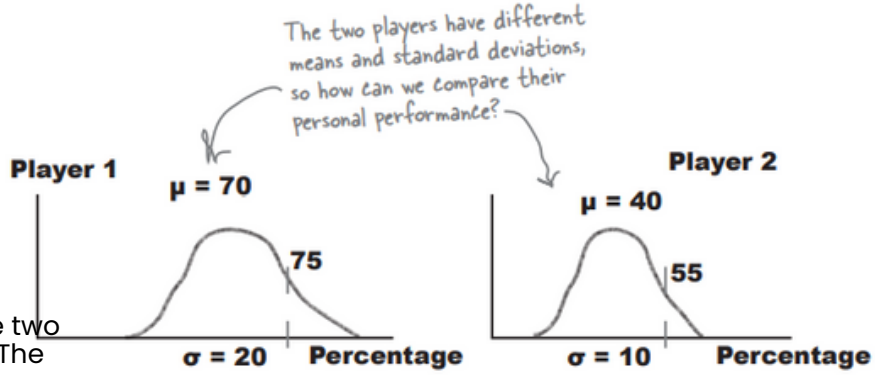
$$= 49.27$$

$$\text{Standard Deviation} = \sqrt{49.27} = 7.02$$

Player 1 and Player 2 both have small standard deviations, so the values are clustered around the mean. But Player 3 has a standard deviation of 7.02, meaning scores are typically 7.02 points away from the mean. So Player 1 is the most reliable, and Player 3 is the least.

Interpreting standard scores

Imagine a situation in which you have two basketball players of different ability. The first player gets the ball into the net an average of 70% of the time, and he has a standard deviation of 20%. The second player has a mean of 40% and a standard deviation of 10%. In a particular practice session, Player 1 gets the ball into the net 75% of the time, and Player 2 makes a basket 55% of the time. Which player does best against their personal track record?

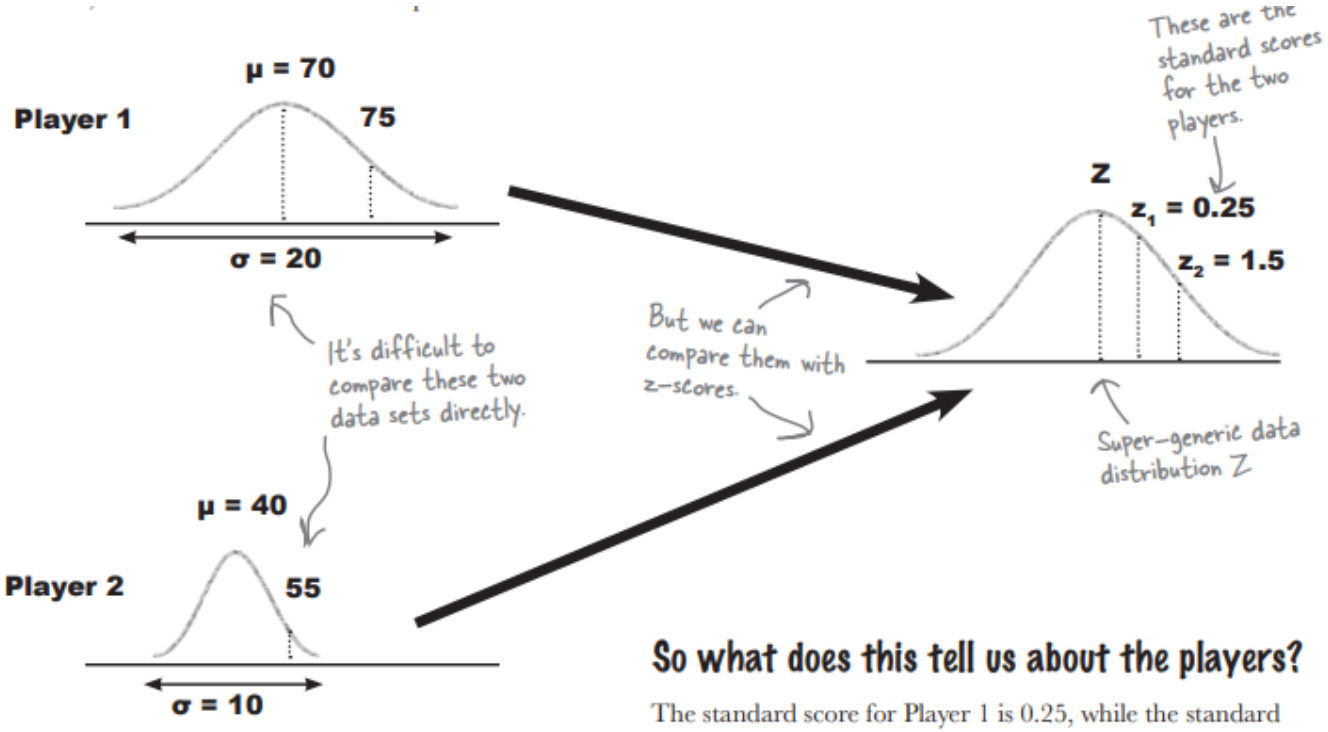


$$z = \frac{x - \mu}{\sigma}$$

These are the mean and standard deviation of the set of data containing the value x.

we can achieve this with the standard score, or z-score.

Standard scores give you a way of comparing values across different sets of data where the mean and standard deviation differ



So what does this tell us about the players?

The standard score for Player 1 is 0.25, while the standard score for Player 2 is 1.5. In other words, when we standardize the scores, the score for Player 2 is higher.

This means that even though Player 1 is generally a better shooter and put balls into the net at a higher rate than Player 2, Player 2 performed better relative to his own track record. Player 2 performed better…for him.

Imagine we have a bunch of numbers, like scores in a test, and we want to understand how each score compares to the average (mean) score and how spread out the scores are from that average.

Standard Score (or z-score):

IIt's like a special scale that we usto compare different numbers.
We transform our original numbers into this special scale where the average score is 0 and the spread of scores is 1.

Positive and Negative Z-scores:

- IIf a score is higher than the average, it gets a positive z-score.
- If a score is lower than the average, it gets a negative z-score.
- If a score is exactly at the average, its z-score is 0.

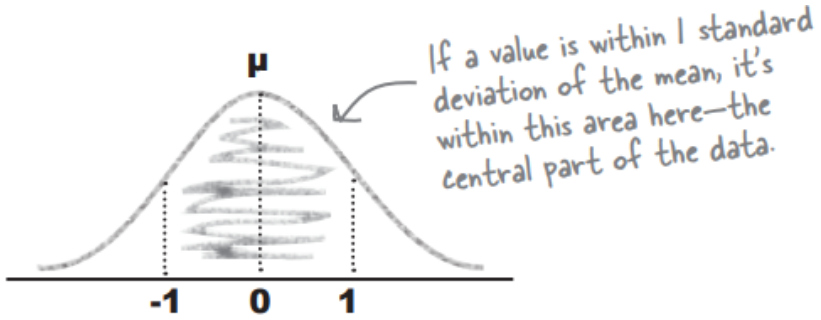
Distance from the Mean:

IThe size of the z-score tells us how far away a number is from the average. For example, if a z-score is -2, it means the number is two times the spread below the average. If it's +3, it means the number is three times the spread above the average.

Standard Deviations from the Mean:

We've seen that using z-scores transforms your data set into a generic distribution with a mean of 0 and a standard deviation of 1. If a value is within 1 standard deviation of the mean, this tells us that the standard score of the value is between -1 and 1. Similarly, if a value is within 2 standard deviations of the mean, the standard score of the value would be somewhere between -2 and 2.

Standard score = number of standard deviations from the mean.



- The **variance** and **standard deviation** measure how values are dispersed by looking at how far values are from the mean.
- The variance is calculated using

$$\frac{\sum (x-\mu)^2}{n}$$

- An alternate form is

$$\frac{\sum x^2}{n} - \mu^2$$

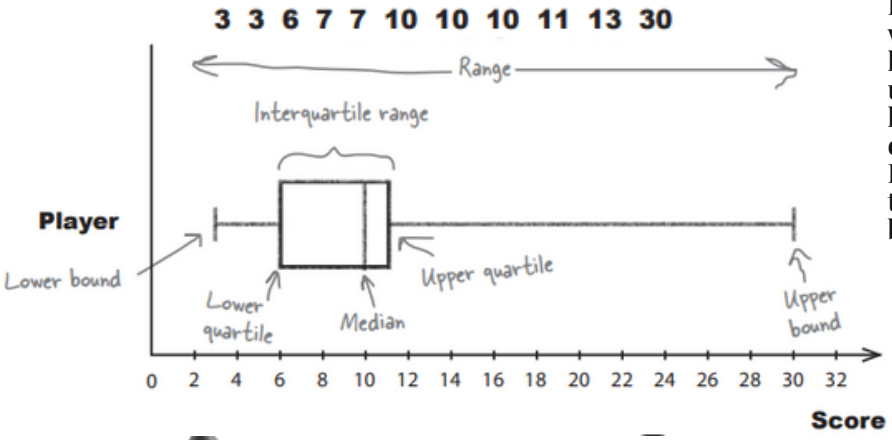
- The standard deviation is equal to the square root of the variance, and the variance is the standard deviation squared.
- **Standard scores**, or **z-scores**, are a way of comparing values across different sets of data where the means and standard deviations are different. To find the standard score of a value x, use:

$$z = \frac{x-\mu}{\sigma}$$

Box Plot & Whisker Plot

A box and whisker diagram shows the range, interquartile range, and median of a set of data. More that one set of data can be represented on the same chart, which means it's a great way of comparing data sets.

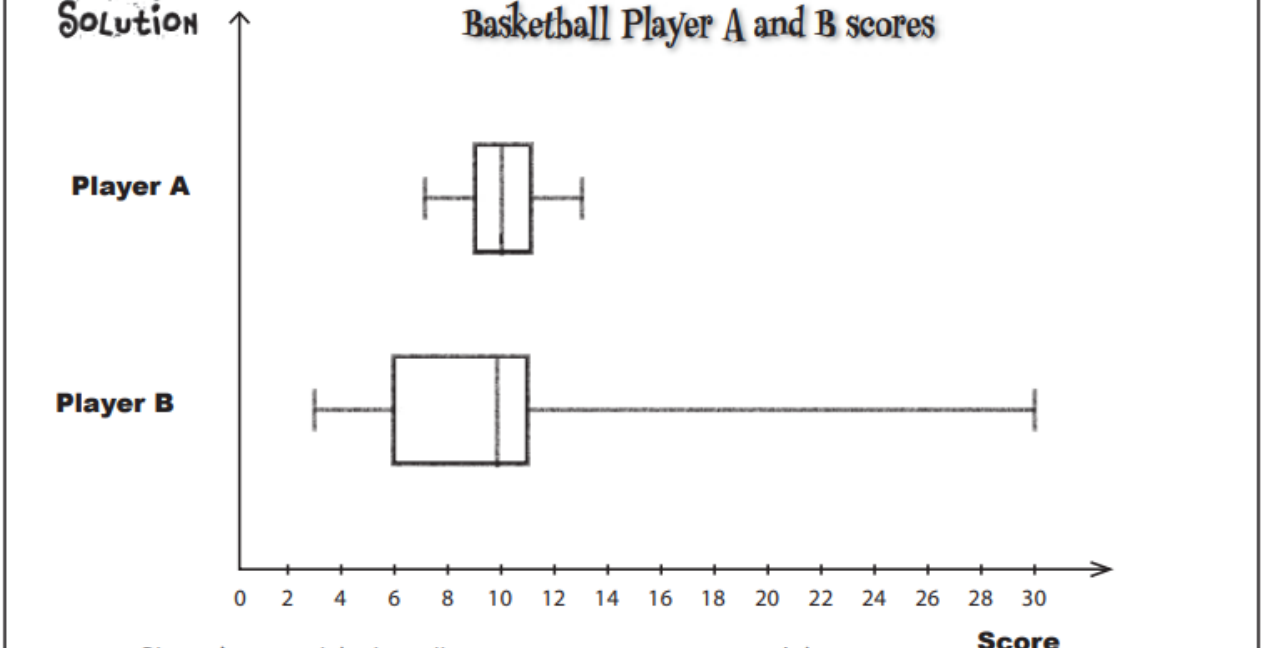
To create a box and whisker diagram, first you draw a box against a scale with the left and right sides of the box representing the lower and upper quartiles, respectively. Then, draw a line inside the box to mark the value of the median. This box shows you the extent of the interquartile range. After that, you draw "whiskers" to either side of your box to show the lower and upper bounds and the extent of the range



If your data has outliers, the range will be wider. On a box and whisker diagram, the length of the whiskers increases in line with the upper and lower bounds. You can get an idea of how data is skewed by looking at the whiskers on the box and whisker diagram.
If the box and whisker diagram is symmetric, this means that the underlying data is likely to be fairly symmetric, too.

Exercise Solution

Here are box and whisker diagrams for each basketball player. Compare the ranges of their scores. If you had to choose between having player A or player B on the team, which would you pick? Why?



Player A has a relatively small range, and his median score is a bit higher than Player B's.

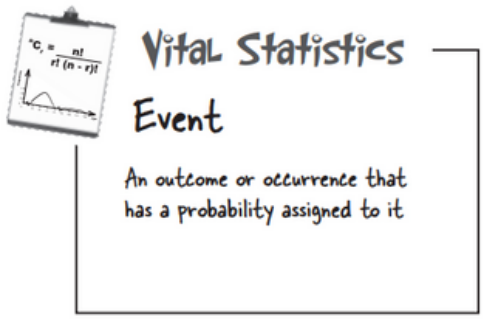
Player B has a very large range. Sometimes this player scores a lot higher than Player A, but sometimes a lot lower.

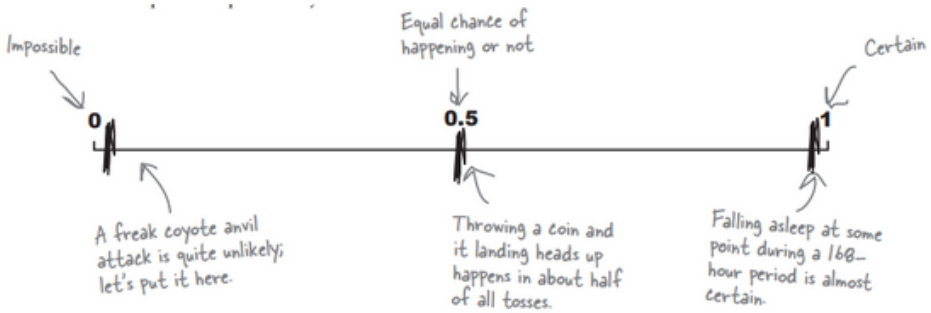
Player A plays more consistently and usually scores higher than Player B (compare the medians and interquartile range), so we'd pick Player A.

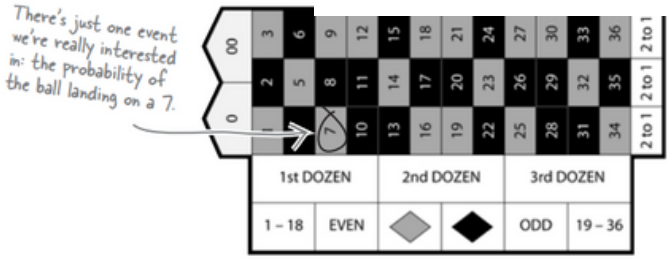
Probability Basics

Probability

Probability is a way of measuring the chance of something happening. You can use it to indicate how likely an occurrence is (the probability that you'll go to sleep some time this week), or how unlikely (the probability that a coyote will try to hit you with an anvil while you're walking through the desert). An **event** is any outcome where you can say how likely it is to occur. Probability is measured on a scale of 0 to 1. If an event is impossible, it has a probability of 0. If it's an absolute certainty, then the probability is 1.







These are all possible outcomes, as the ball could land in any of these pockets.

To find the probability of winning, we take the number of ways of winning the bet and divide by the number of possible outcomes like this

We can write this in a more general way, too. For the probability of any event A:

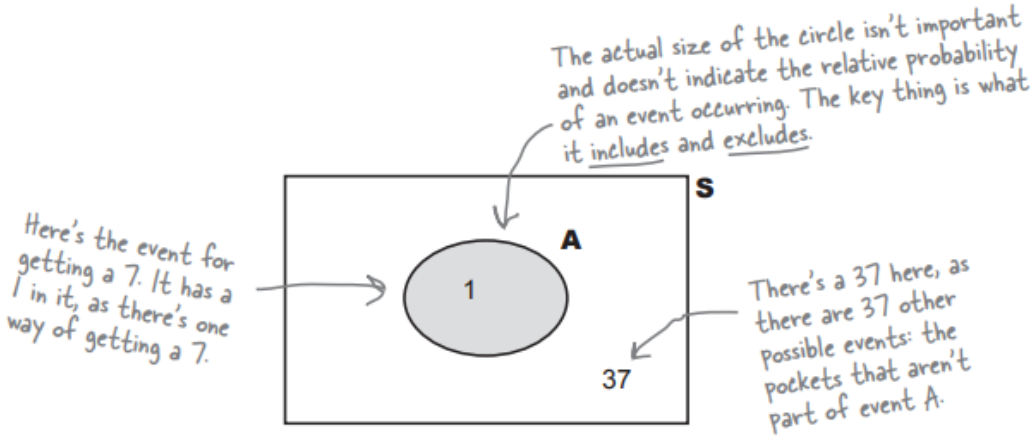
Probability of event A occurring → $P(A) = \frac{n(A)}{n(S)}$ ← Number of ways of getting an event A; The number of possible outcomes

Probability = $\frac{\text{number of ways of winning}}{\text{number of possible outcomes}}$ ← There's one way of getting a 7, and there are 38 pockets.

S is known as the **possibility space**, or **sample space**. It's a shorthand way of referring to all of the possible outcomes. Possible events are all subsets of S.

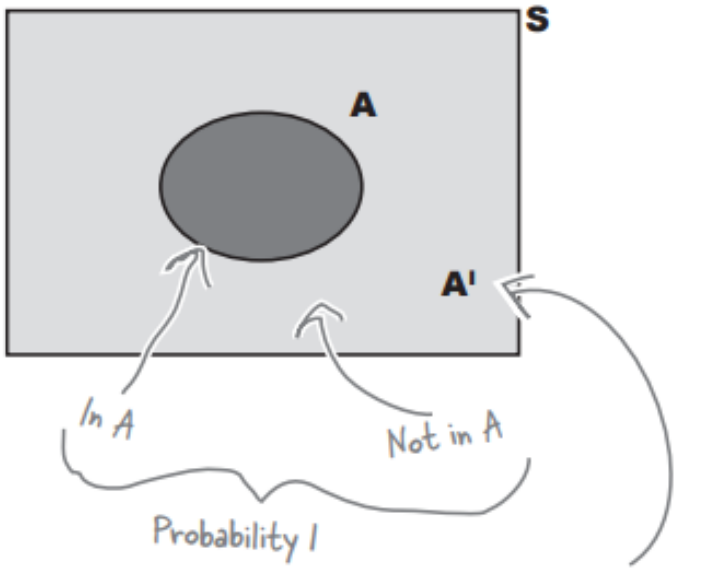
You can visualize probabilities with a Venn diagram

. Instead of numbers, you have the option of using the actual probabilities of each event in the diagram. It all depends on what kind of information you need to help you solve the problem.



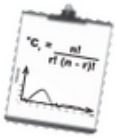
Complementary events

A` covers every possibility that's not in event A, so between them, A and A` must cover every eventuality. If something's in A, it can't be in A` , and if something's not in A, it must be in Al . This means that if you add P(A) and P(A`) together, you get 1. In other words, there's a 100% chance that something will be in either A or A` . This gives us P(A) + P(A`) = 1



In this diagram, A' is used instead of 37 to indicate all the possible events that aren't in A

P(A') = 1 - P(A)



Vital Statistics

Probability

To find the probability of an event A, use

$$P(A) = \frac{n(A)}{n(S)}$$

Vital Statistics

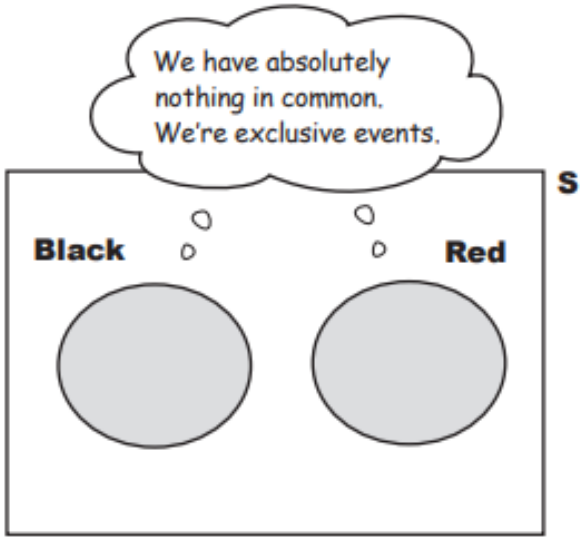
A'

A' is the complementary event of A. It's the probability that event A does not occur.

$$P(A') = 1 - P(A)$$

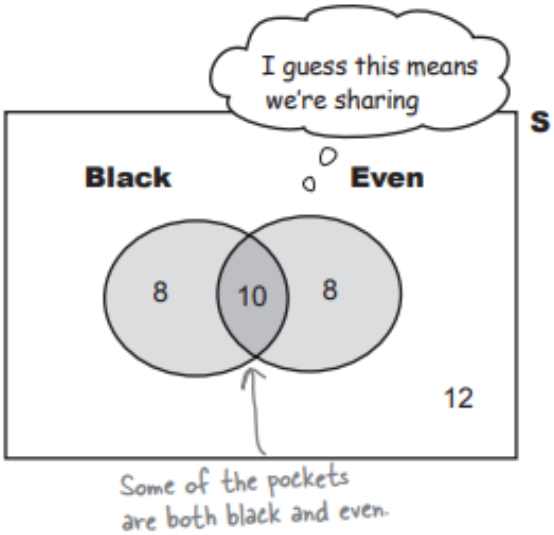
Exclusive events and intersecting events

When we were working out the probability of the ball landing in a black or red pocket, we were dealing with two separate events, the ball landing in a black pocket and the ball landing in a red pocket. These two events are ***mutually exclusive*** because it's impossible for the ball to land in a pocket that's both black and red.



If two events are mutually exclusive, only one of the two can occur.

What about the black and even events? This time the events aren't mutually exclusive. It's possible that the ball could land in a pocket that's both black and even. The two events intersect.



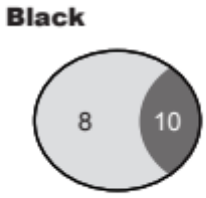
If two events intersect, it's possible they can occur simultaneously.

simultaneously.

Problems at the intersection

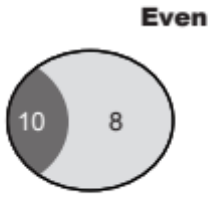
Calculating the probability of getting a black or even went wrong because we included black and even pockets twice. Here's what happened.

First of all, we found the probability of getting a black pocket and the probability of getting an even number.

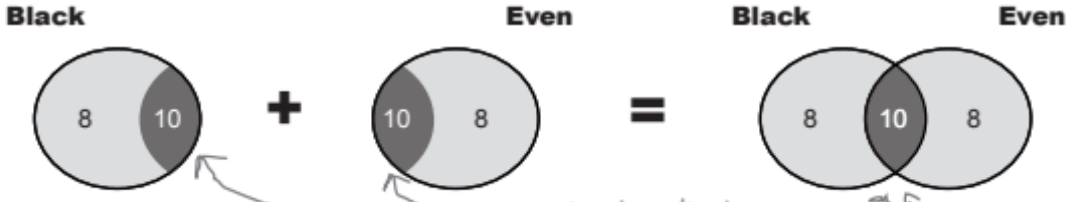


$$P(\text{Black}) = \frac{18}{38} = 0.474$$

$$P(\text{Even}) = \frac{18}{38} = 0.474$$



When we added the two probabilities together, we counted the probability of getting a black and even pocket twice.



$$P(\text{Black} \cap \text{Even}) = \frac{10}{38} = 0.263$$

To get the correct answer, we need to subtract the probability of getting both black and even. This gives us

P(Black or Even) = P(Black) + P(Even) - P(Black *and* Even)

We only need one of these, so let's subtract one of them.

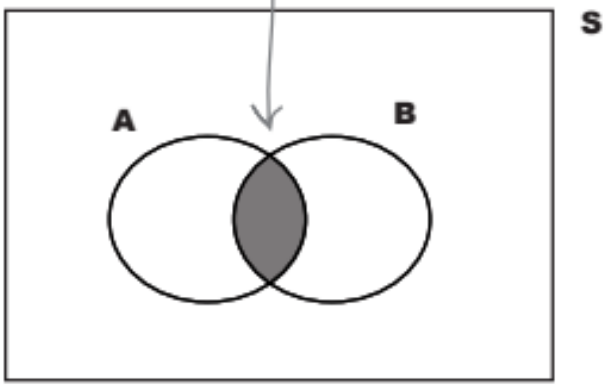
We can now substitute in the values we just calculated to find P(Black or Even):

P(Black or Even) = 18/38 + 18/38 - 10/38 = 26/38 = 0.684

Some more notation

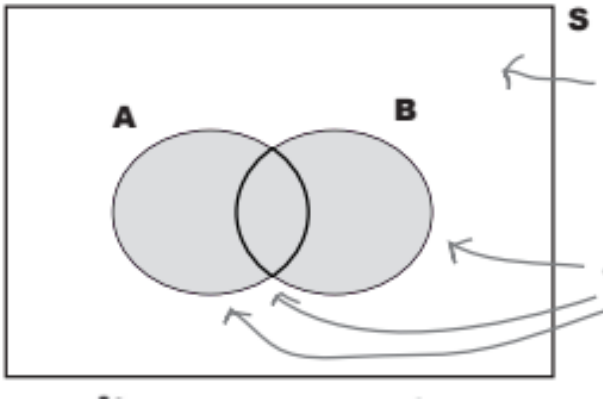
There's a more general way of writing this using some more math shorthand.

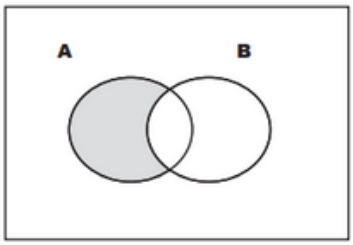
First of all, we can use the notation A ∩ B to refer to the intersection between A and B. You can think of this symbol as meaning "and." It takes the common elements of events.



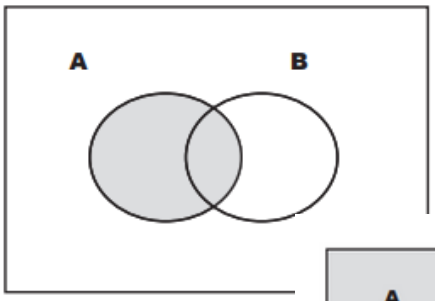
A ∪ B, on the other hand, means the union of A and B. It includes all of the elements in A and also those in B. You can think of it as meaning "or."

If A ∪ B = 1, then A and B are said to be ***exhaustive***. Between them, they make up the whole of S. They exhaust all possibilities.

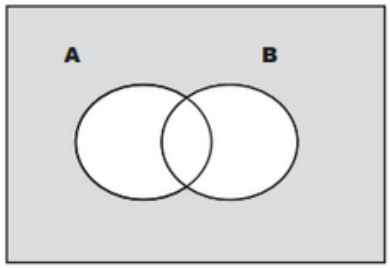




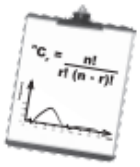
P(A ∪ B) - P(B)



P(A ∩ B) + P(A ∩ B')



P(A' ∩ B')



Vital Statistics

A or B

To find the probability of getting event A or B, use

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

∪ means OR

∩ means AND

Watch it!

There's a difference between exclusive and exhaustive.

If events A and B are exclusive, then

$P(A \cap B) = 0$

If events A and B are exhaustive, then

$P(A \cup B) = 1$

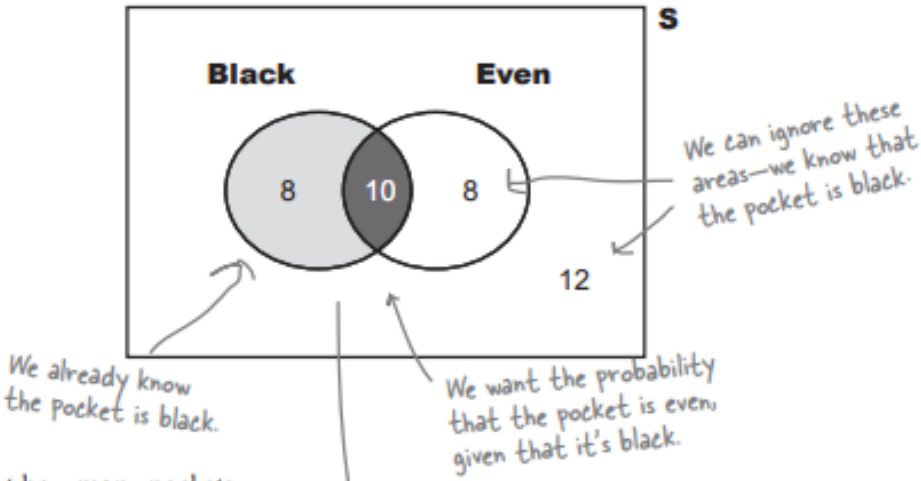
Conditions apply

The croupier says the ball has landed in a black pocket.
What's the probability that the pocket is also even?

But we've already done this; it's just the probability of getting black and even.

This is a slightly different problem

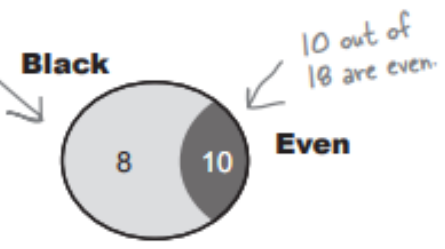
We don't want to find the probability of getting a pocket that is both black and even, out of all possible pockets. Instead, we want to find the probability that the pocket is even, given that we already know it's black.



In other words, we want to find out how many pockets are even out of all the black ones. Out of the 18 black pockets, 10 of them are even, so

$$P(\text{Even given Black}) = \frac{10}{18}$$

$$= 0.556 \text{ (to 3 decimal places)}$$



It turns out that even with some inside information, our odds are actually lower than before. The probability of even given black is actually less than the probability of black or even.

However, a probability of 0.556 is still better than 50% odds, so this is still a pretty good bet. Let's go for it.

Find conditional probabilities

So how can we generalize this sort of problem? First of all, we need some more notation to represent ***conditional probabilities***, which measure the probability of one event occurring relative to another occurring.

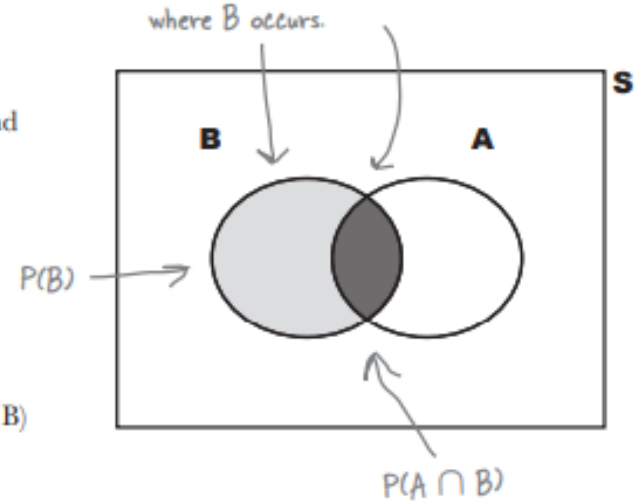
If we want to express the probability of one event happening given another one has already happened, we use the "|" symbol to mean "given." Instead of saying "the probability of event A occurring given event B," we can shorten it to say

$P(A \mid B)$ ← The probability of A given that we know B has happened.

So now we need a general way of calculating $P(A \mid B)$. What we're interested in is the number of outcomes where both A and B occur, divided by all the B outcomes. Looking at the Venn diagram, we get:

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

Because we're trying to find the probability of A given B, we're only interested in the set of events where B occurs.



We can rewrite this equation to give us a way of finding $P(A \cap B)$

$$P(A \cap B) = P(A \mid B) \times P(B)$$

It doesn't end there. Another way of writing $P(A \cap B)$ is $P(B \cap A)$. This means that we can rewrite the formula as

$$P(B \cap A) = P(B \mid A) \times P(A)$$

In other words, just flip around the A and the B.

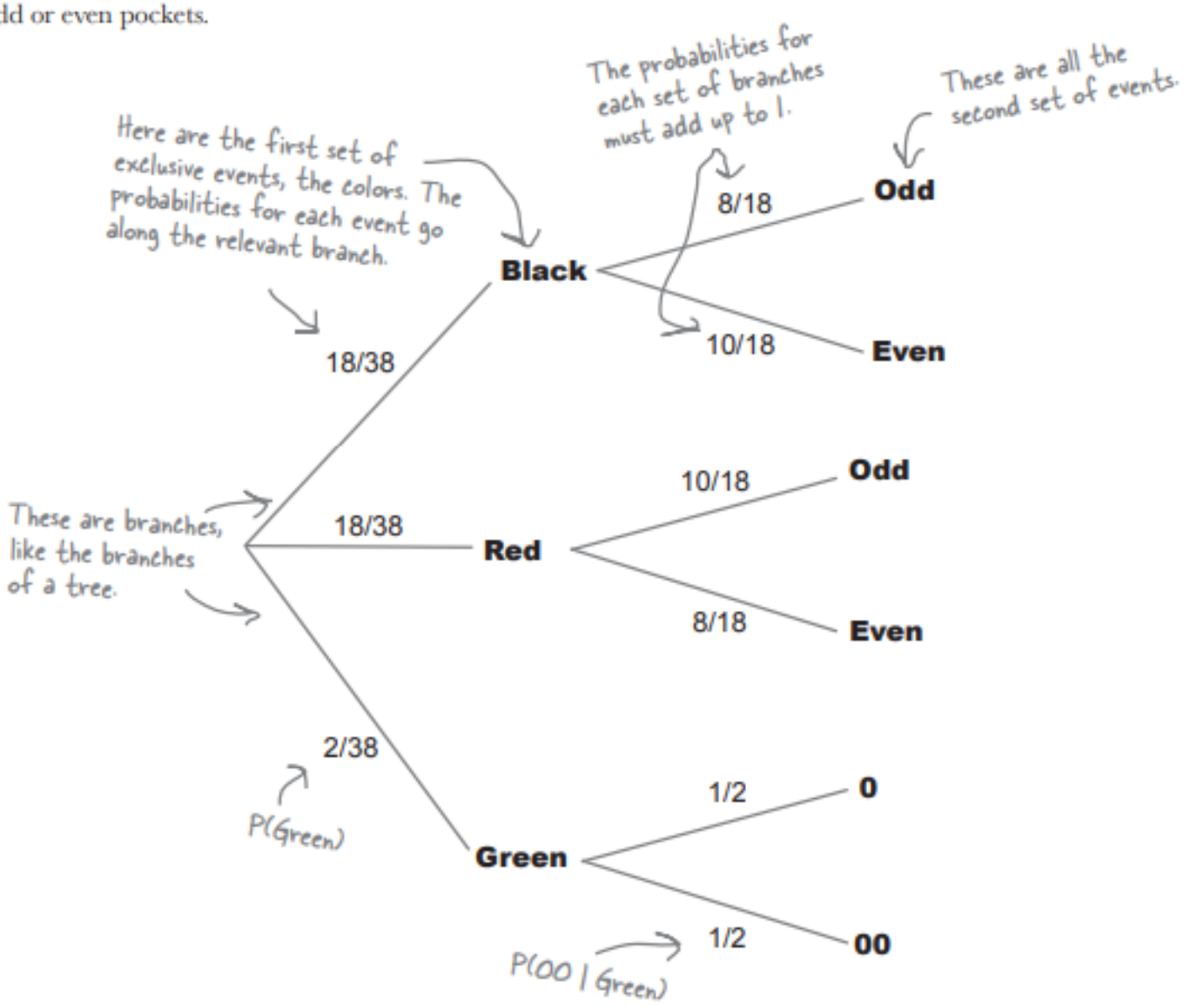
It looks like it can be difficult to show conditional probability on a Venn diagram. I wonder if there's some other way.

Venn diagrams aren't always the best way of visualizing conditional probability.

Don't worry, there's another sort of diagram you can use—a probability tree.

You can visualize conditional probabilities with a probability tree

It's not always easy to visualize conditional probabilities with Venn diagrams, but there's another sort of diagram that really comes in handy in this situation—the ***probability tree***. Here's a probability tree for our problem with the roulette wheel, showing the probabilities for getting different colored and odd or even pockets.



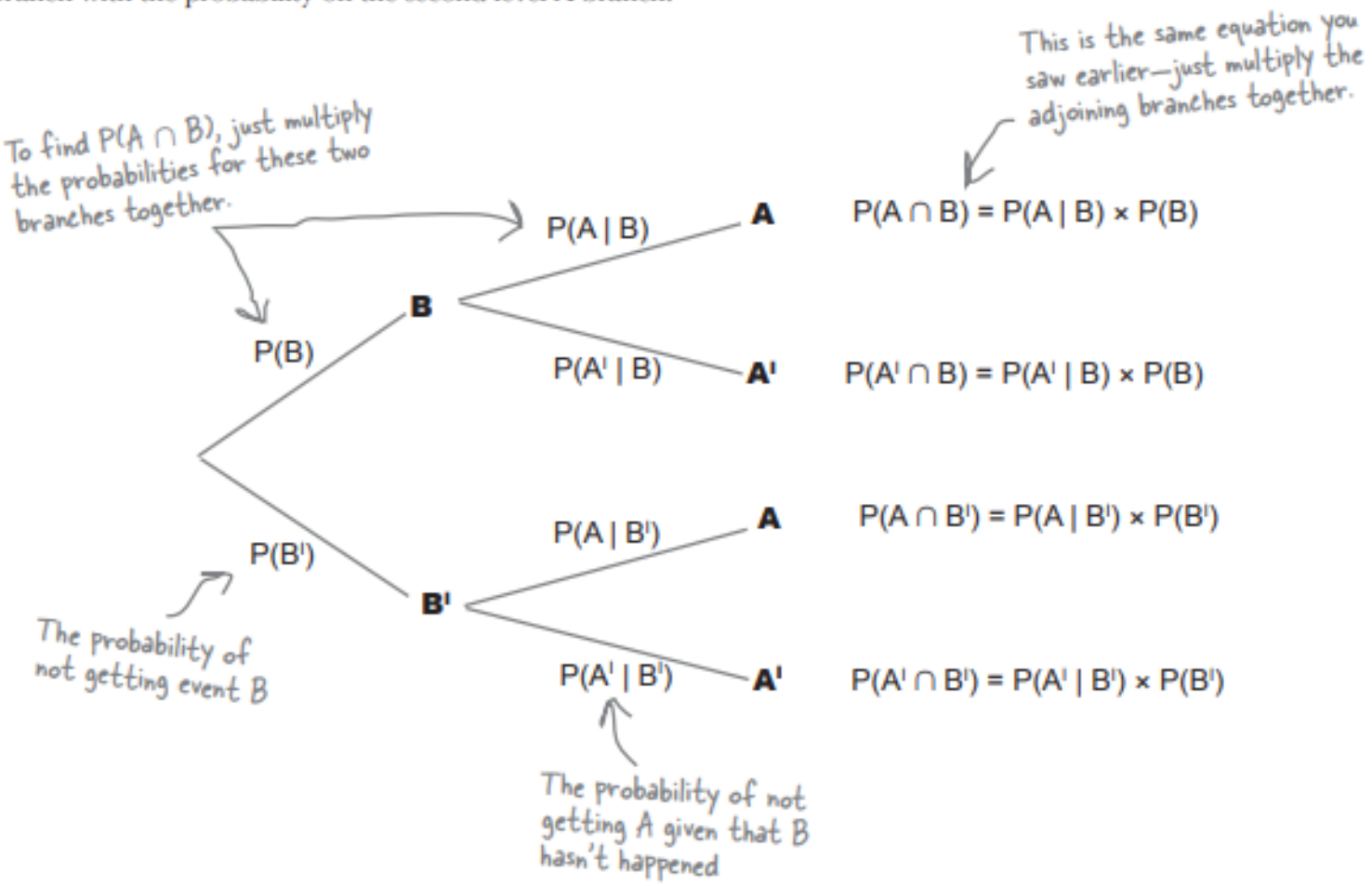
The first set of branches shows the probability of each outcome, so the probability of getting a black is 18/38, or 0.474. The second set of branches shows the probability of outcomes **given the outcome of the branch it is linked to**. The probability of getting an odd pocket given we know it's black is 8/18, or 0.444.

Trees also help you calculate conditional probabilities

Probability trees don't just help you visualize probabilities; they can help you to calculate them, too.

Let's take a general look at how you can do this. Here's another probability tree, this time with a different number of branches. It shows two levels of events: A and A^I and B and B^I. A^I refers to every possibility not covered by A, and B^I refers to every possibility not covered by B.

You can find probabilities involving intersections by multiplying the probabilities of linked branches together. As an example, suppose you want to find $P(A \cap B)$. You can find this by multiplying $P(B)$ and $P(A \mid B)$ together. In other words, you multiply the probability on the first level B branch with the probability on the second level A branch.



Using probability trees gives you the same results you saw earlier, and it's up to you whether you use them or not. Probability trees can be time-consuming to draw, but they offer you a way of visualizing conditional probabilities.

Vital Statistics

Law of Total Probability

If you have two events A and B, then

$$P(B) = P(B \cap A) + P(B \cap A')$$

$$= P(A)\,P(B \mid A) + P(A')\,P(B \mid A')$$

The Law of Total Probability is the denominator of Bayes' Theorem.

Vital Statistics

Bayes' Theorem

If you have n mutually exclusive and exhaustive events, A_1 through to A_n, and B is another event, then

$$P(A \mid B) = \frac{P(A)\,P(B \mid A)}{P(A)\,P(B \mid A) + P(A')\,P(B \mid A')}$$

If events affect each other, they are dependent

The probability of getting black followed by black is a slightly different problem from the probability of getting an even pocket given we already know it's black. Take a look at the equation for this probability:

P(Even | Black) = 10/18 = 0.556

For P(Even | Black), the probability of getting an even pocket is affected by the event of getting a black. We already know that the ball has landed in a black pocket, so we use this knowledge to work out the probability. We look at how many of the pockets are even out of all the black pockets.

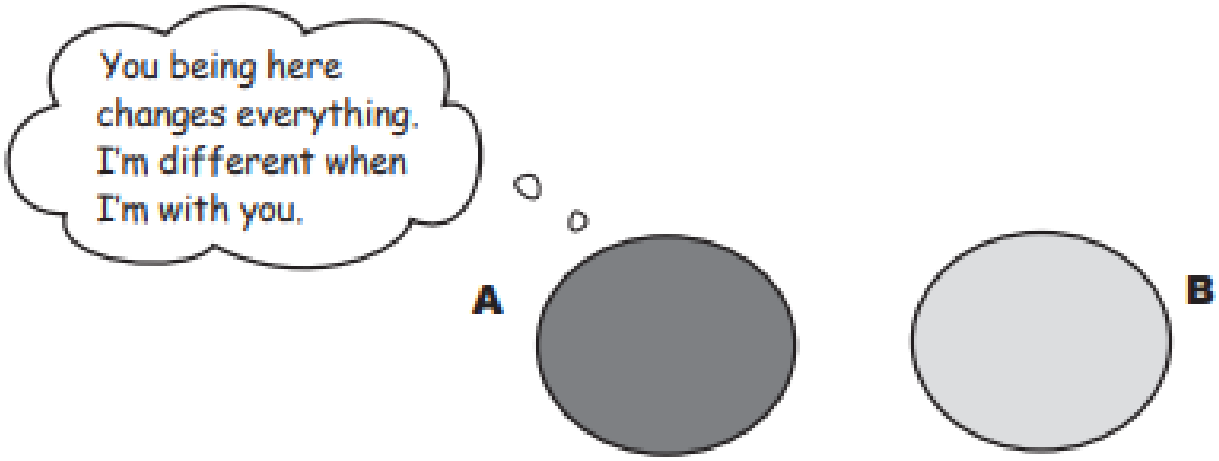
If we didn't know that the ball had landed on a black pocket, the probability would be different. To work out P(Even), we look at how many pockets are even out of all the pockets

P(Even) = 18/38 = 0.474

These two probabilities are different

P(Even | Black) gives a different result from P(Even). In other words, the knowledge we have that the pocket is black changes the probability. These two events are said to be ***dependent***.

In general terms, events A and B are said to be dependent if P(A | B) is different from P(A). It's a way of saying that the probabilities of A and B are affected by each other.



More on calculating probability for independent events

It's easier to work out other probabilities for independent events too, for example $P(A \cap B)$.

We already know that

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

If A and B are independent, $P(A \mid B)$ is the same as $P(A)$. This means that

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

or

$$\mathbf{P(A \cap B) = P(A) \times P(B)}$$

for independent events. In other words, if two events are independent, then you can work out the probability of getting both events A and B by multiplying their individual probabilities together.

Watch it!

If A and B are mutually exclusive, they can't be independent, and if A and B are independent, they can't be mutually exclusive.

If A and B are mutually exclusive, then if event A occurs, event B cannot. This means that the outcome of A affects the outcome of B, and so they're dependent.

Similarly if A and B are independent, they can't be mutually exclusive.

Vital Statistics

Independence

If two events A and B are independent, then

$$P(A \mid B) = P(A)$$

If this holds for any two events, then the events must be independent. Also

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

The second coin throw isn't affected by the first.

	Dependent	Independent
Throwing a coin and getting heads twice in a row.	☐	☑
Removing socks from a drawer until you find a matching pair.	☑	☐
Choosing chocolates at random from a box and picking dark chocolates twice in a row.	☑	☐
Choosing a card from a deck of cards, and then choosing another one.	☑	☐
Choosing a card from a deck of cards, putting the card back in the deck, and then choosing another one.	☐	☑
The event of getting rain given it's a Thursday.	☐	☑

When you remove one sock, there are fewer socks to choose from the next time, and this affects the probability.

It's no more or less likely to rain just because it's Thursday, so these two events are independent.

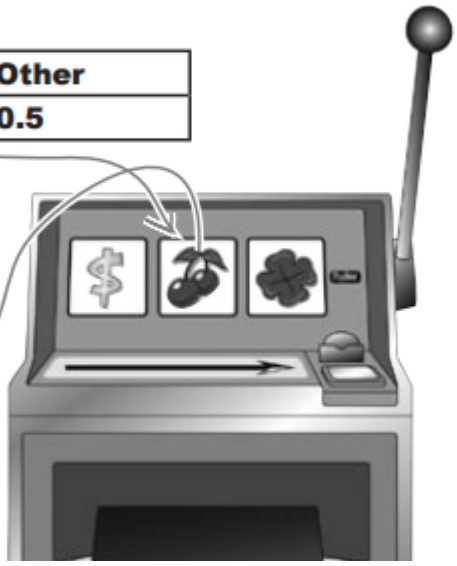
Discrete Probability Distributions

Scenario

This sounds like something we can calculate. Here are the probabilities of a particular image appearing in a particular window:

$	Cherry	Lemon	Other
0.1	0.2	0.2	0.5

The three windows are independent of each other, which means that the image that appears in one of the windows has no effect on the images that appear in any of the others.



The probability of a cherry appearing in this window is 0.2.

probability of $ $ $

$P(\$, \$, \$) = P(\$) \times P(\$) \times P(\$)$

$= 0.1 \times 0.1 \times 0.1$

$= 0.001$

The probability of a dollar sign appearing in a window is 0.1

probability of $ $ cherry (any order)

There are three ways of getting this:

$P(\$, \$, \text{cherry}) + P(\$, \text{cherry}, \$) + P(\text{cherry}, \$, \$)$

$= (0.1^2 \times 0.2) + (0.1^2 \times 0.2) + (0.1^2 \times 0.2)$

$= 0.006$

probability of lemon lemon lemon

$P(\text{lemon, lemon, lemon}) = P(\text{lemon}) \times P(\text{lemon}) \times P(\text{lemon})$

$= 0.2 \times 0.2 \times 0.2$

$= 0.008$

A lemon appearing in a window is independent of ones appearing in the other two windows, so you multiply the three probabilities together.

probability of cherry cherry cherry

$P(\text{cherry, cherry, cherry}) = P(\text{cherry}) \times P(\text{cherry}) \times P(\text{cherry})$

$= 0.2 \times 0.2 \times 0.2$

$= 0.008$

probability of winning nothing

This means we get none of the winning combinations.

Rather than work out all the possible ways in which you could lose, you can say P(losing) = 1 – P(winning).

$P(\text{losing}) = 1 - P(\$, \$, \$) - P(\$, \$, \text{cherry (any order)}) - P(\text{cherry, cherry, cherry}) - P(\text{lemon, lemon, lemon})$

$= 1 - 0.001 - 0.006 - 0.008 - 0.008$

$= 0.977$

These are the four probability values we calculated above.

We can compose a probability distribution for the slot machine

Here are the probabilities of the different winning combinations on the slot machine.

This is just a summary of the probabilities we just worked out.

Combination	None	Lemons	Cherries	Dollars/cherry	Dollars
Probability	0.977	0.008	0.008	0.006	0.001

Probability Distributions

This table is a probability distribution of your winnings/losses on the slot machine. It allows you to see the chance of each outcome and helps you make informed decisions.

For example, the table shows a very high probability (0.977) of losing $1 (not getting any winning combinations). This can help you understand the risk involved in playing the slot machine.

Combination	None	Lemons	Cherries	$s/cherry	Dollars
Gain	-$1	$4	$9	$14	$19
Probability	0.977	0.008	0.008	0.006	0.001

We lose $1 if we don't hit a winning combination.

These are the same probabilities, just written in terms of how much we'll gain.

Our gain for hitting each winning combination: the payoff minus the $1 we paid to play.

The table gives us the probability distribution of the winnings, a set of the probabilities for every possible gain or loss for our slot machine.

When you derived the probabilities of the slot machine, you calculated the probability of making each gain or loss. In other words, you calculated the probability distribution of a **random variable**, which is a variable that can takes on a set of values, where each value is associated with a specific probability.

Random Variable

Random variable is a variable whose value depends on the outcome of a random event. In simpler terms, it's a numerical representation of the uncertain results of an experiment or event. Here's a breakdown to understand it better:

- **Variable**: Imagine a variable as a container that can hold different values. For example, the variable "number of heads on a coin toss" can hold the values 0 (for tails) or 1 (for heads).
- **Random Event**: A random event is something that has an uncertain outcome. Flipping a coin, rolling a dice, or picking a card from a deck are all random events.

Discrete Random Variable

A discrete random variable is a specific type of random variable that can only take on a countable number of distinct values. It's like a variable that flips through a limited set of channels, unlike a continuous random variable that can tune to any frequency on a dial.

Once you've calculated a probability distribution, you can use this information to determine the expected outcome.

Here's our slot machine probability distribution written using this notation:

The value of each combination's winnings is represented by x.

Here x is 19.

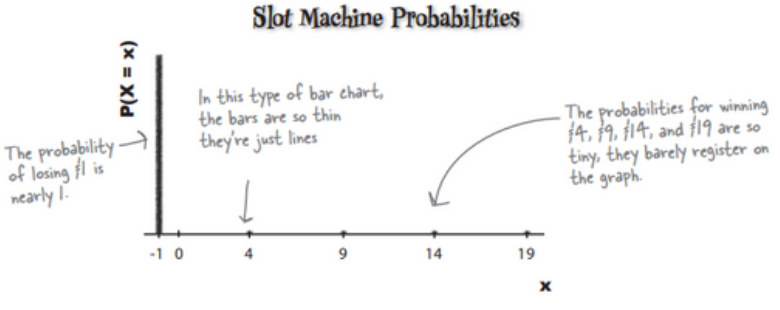
Combination	None	Lemons	Cherries	$s/cherry	Dollars
x	-1	4	9	14	19
P(X = x)	0.977	0.008	0.008	0.006	0.001

X is the variable.

The probability that the variable X is 9—in other words, that the value of the winnings is $9.

The variable is ***discrete***. This means that it can only take exact values.

As well as giving a table of the probability distribution, we can also show the distribution on a chart to help us visualize it. Here is a bar chart showing the slot machine probabilities.



Expectation

You have a probability distribution for the amount you could gain on the slot machines, but now you need to know how much you can expect to win or lose long-term. You can do this by calculating how much you can typically expect to win or lose in each game. In other words, you can find the expectation. The expectation of a variable X is a bit like the mean, but for probability distributions. You even calculate it in a similar way. To find the expectation, you multiply each value x by the probability of getting that value, and then sum the results.

I'm the expectation. Treat me like I'm mean.

Multiply each value by its probability.

$E(X)$ is the expectation of X

$$E(X) = \sum xP(X = x)$$

Once you've done multiplying, add the whole lot up together.

x	-1	4	9	14	19
P(X = x)	0.977	0.008	0.008	0.006	0.001

$$\begin{aligned} E(X) &= (-1 \times 0.977) + (4 \times 0.008) + (9 \times 0.008) + (14 \times 0.006) + (19 \times 0.001) \\ &= -0.977 + 0.032 + 0.072 + 0.084 + 0.019 \\ &= -0.77 \end{aligned}$$

This is the amount in £'s you can expect to gain on each pull of the lever—and it's negative!

The expectation tells you how much on average you can expect to win or lose with each game but it doesn't tell you anything about how the values are spread out.

Variances and probability distributions

. For our slot machine, this will tell us more about the variation of our potential winnings.

μ is the alternative way of writing $E(X)$.

This is the variance. A shorthand way of referring to the variance of X is $Var(X)$.

$$Var(X) = E(X - \mu)^2$$

We need to find the expectation of $(X - \mu)^2$—but how?

So how do we calculate $E(X - \mu)^2$?

Finding $E(X - \mu)^2$ is actually quite similar to finding $E(X)$.

When we calculate $E(X)$, we take each value in the probability distribution, multiply it by its probability, and then add the results together. In other words, we use the calculation

$$E(X) = \Sigma xP(X = x)$$

When we calculate the variance of X, we calculating $(x - \mu)^2$ for every value x, multiply it by the probability of getting that value x, and then add the results together.

Go through each value x and work out what $(x - \mu)^2$ is. Then multiply it by the probability of getting x...

$$E(X - \mu)^2 = \sum (x - \mu)^2 P(X = x)$$

...and then add these results together.

In other words, instead of multiplying x by its probability, you multiply $(x - \mu)^2$ by the probability of getting that value of x.

Standard Deviation and probability distributions

As well as having a variance, probability distributions have a standard deviation.

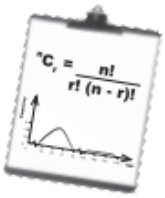
It serves a similar function to the standard deviation of a set of values. It's a way of measuring how far away from the center you can expect your values to be.

As before, the standard deviation is calculated by taking the square root of the variance like this:

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

We can use the same symbol for standard deviation as before.

This means that the standard deviation of the slot machine winnings is $\sqrt{2.6971}$, or 1.642. This means that on average, our winnings per game will be 1.642 away from the expectation of -0.77.

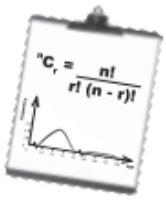


Vital Statistics

Expectation

Use the following formula to find the expectation of a variable X:

$$E(X) = \Sigma x P(X = x)$$



Vital Statistics

Variance

Use the following formula to calculate the variance

$$\text{Var}(X) = E(X - \mu)^2$$

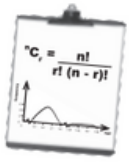
lINEAR TRANSFORMATION

Imagine you're on a game show and you have to guess the amount of money in hidden boxes. You calculate the average amount you expect to find in each box (let's call this average the "expectation").

Suddenly, the host tells you there's a twist! The prizes are being changed. They'll now be $10 less than double the original amount. This is a linear transform, a fancy way of saying the new prizes are based on a straight-line rule.

Here's how it works:

Let X represent the original amount of money in a box.
The new amount, Y, is calculated by multiplying the original amount (X) by 2 and then subtracting $10. So, Y = 2X - 10.



Vital Statistics

Linear Transforms

If you have a variable X and numbers a and b, then:

$$E(aX + b) = aE(X) + b$$

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

- **Probability distributions** describe the probability of all possible outcomes of a given variable.
- The **expectation** is the expected average long-term outcome. It's represented as either E(X) or μ, and is calculated using $E(X) = \Sigma x P(X=x)$.
- The **expectation of a function of X** is given by $E(f(X)) = \Sigma f(x) P(X=x)$
- The **variance of a probability distribution** is given by $\text{Var}(X) = E(X - \mu)^2$
- The **standard deviation of a probability distribution** is given by $\sigma = \sqrt{\text{Var}(X)}$
- **Linear transforms** are when a variable X is transformed into aX + b, where a and b are constants. The expectation and variance are given by:
  $E(aX + b) = aE(X) + b$
  $\text{Var}(aX + b) = a^2\text{Var}(X)$

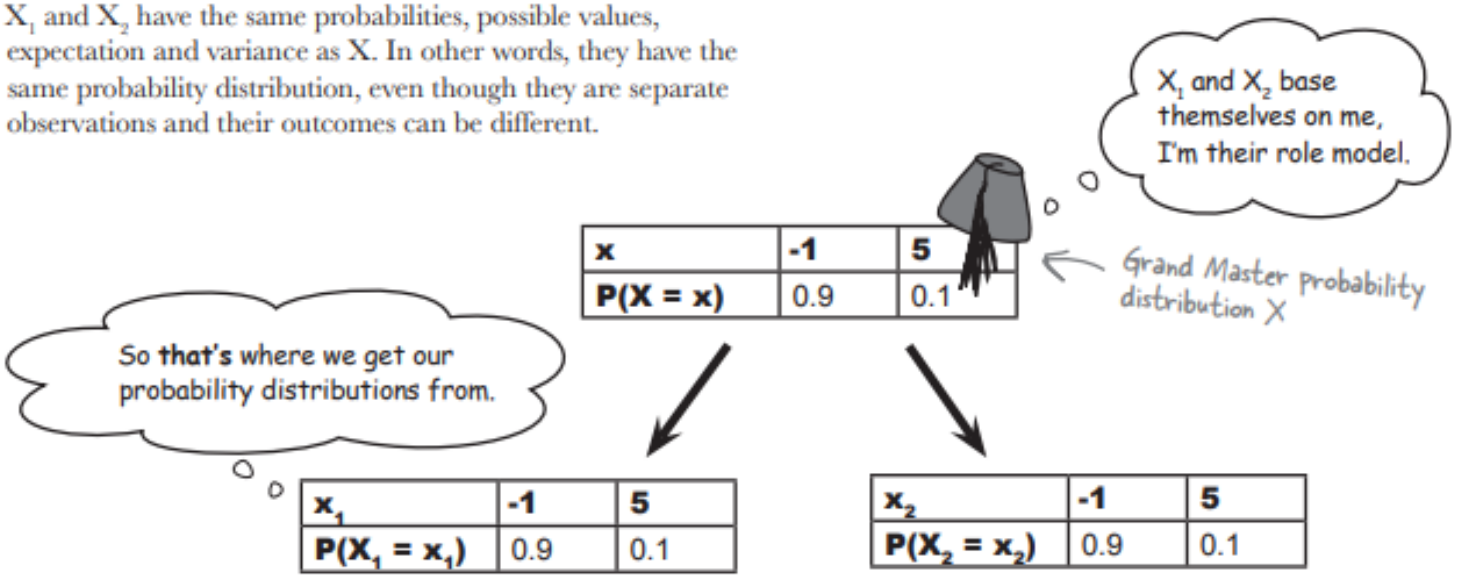
Every pull of the lever is an independent observation

When we play multiple games on the slot machine, each game is called an event, and the outcome of each game is called an ***observation***. Each observation has the same expectation and variance, but their outcomes can be different. You may not gain the same amount in each game.

We need some way of differentiating between the different games or observations. If the probability distribution of the slot machine gains is represented by X, we call the first observation X_1 and the second observation X_2.

Each game is called an event. The outcome of each game is called an observation.

X_1 and X_2 have the same probabilities, possible values, expectation and variance as X. In other words, they have the same probability distribution, even though they are separate observations and their outcomes can be different.



x	-1	5
P(X = x)	0.9	0.1

x_1	-1	5
$P(X_1 = x_1)$	0.9	0.1

x_2	-1	5
$P(X_2 = x_2)$	0.9	0.1

When we want to find the expectation and variance of two games on the slot machine, what we really want to find is the expectation and variance of $X_1 + X_2$. Let's take a look at some shortcuts.

Vital Statistics

Independent Observations

Use the following formula to calculate the variance

$E(X_1 + X_2 + \ldots + X_n) = nE(X)$

$Var(X_1 + X_2 + \ldots + X_n) = nVar(X)$

Let's find the expectation and variance of $X_1 + X_2$.

Expectation

First of all, let's deal with $E(X_1 + X_2)$.

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$
$$= E(X) + E(X)$$
$$= 2E(X)$$

$E(X_1)$ and $E(X_2)$ are both equal to $E(X)$ as X_1 and X_2 follow the same probability distribution as X

In other words, if we have the expectation of two observations, we multiply E(X) by 2. This means that if we were to play two games on a slot machine where E(X) = -0.77, the expectation would be -0.77×2, or -1.54.

We can extend this to deal with multiple observations. If we want to find the expectation of *n* observations, we can use

If there are n observations, we just multiply E(X) by n.

$$E(X_1 + X_2 + \ldots X_n) = nE(X)$$

Watch it!

$X_1 + X_2$ is not the same as 2X.

$X_1 + X_2$ means you are considering two observations of X. 2X means you have one observation, but the possible values have doubled.

Variance

So what about $\text{Var}(X_1 + X_2)$? Here's the calculation.

$$\begin{aligned}\text{Var}(X_1 + X_2) &= \text{Var}(X_1) + \text{Var}(X_2) \\ &= \text{Var}(X) + \text{Var}(X) \\ &= 2\text{Var}(X)\end{aligned}$$

$\text{Var}(X_1)$ and $\text{Var}(X_2)$ are the same as $\text{Var}(X)$ as X_1 and X_2 follow the same probability distribution as X.

This means that if we were to play two games on a slot machine where $\text{Var}(X) = 2.6971$, the variance would be 2.6971×2, or 5.3942.

We can extend this for any number of independent observations. If we have n independent observations of X

Multiply Var(X) by n, the number of observations.

$$\textbf{Var}(X_1 + X_2 + \ldots X_n) = n\textbf{Var}(X)$$

In other words, to find the expectation and variance of multiple observations, just multiply E(X) and Var(X) by the number of observations.

BULLET POINTS

- Probability distributions describe the probability of all possible outcomes of a given random variable.
- The expectation of a random variable X is the expected long-term average. It's represented as either E(X) or μ. It's calculated using

  $E(X) = \Sigma xP(X=x)$
- The variance of a random variable X is given by

  $\text{Var}(X) = E(X - \mu)^2$
- The standard deviation σ is the square root of the variance.
- Linear transforms are when a random variable X is transformed into aX + b, where a and b are numbers. The expectation and variance are given by

  $E(aX + b) = aE(X) + b$
  $\text{Var}(aX + b) = a^2\text{Var}(X)$

Statistic	Shortcut or formula
$E(aX + b)$	$aE(X) + b$
$\text{Var}(aX + b)$	$a^2\text{Var}(X)$
$E(X)$	$\sum xP(X = x)$
$E(f(X))$	$\sum f(x)P(X = x)$
$\text{Var}(aX - bY)$	$a^2\text{Var}(X) + b^2\text{Var}(Y)$
$\text{Var}(X)$	$E(X - \mu)^2 = E(X^2) - \mu^2$
$E(aX - bY)$	$aE(X) - bE(Y)$
$E(X_1 + X_2 + X_3)$	$3E(X)$
$\text{Var}(X_1 + X_2 + X_3)$	$3\text{Var}(X)$
$E(X^2)$	$\sum x^2P(X = x)$
$\text{Var}(aX - b)$	$a^2\text{Var}(X)$

LINEAR TRANSFORM OR INDEPENDENT OBSERVATION? SOLUTION

Below are a series of scenarios. Assuming you know the distribution of each X, and your task is to say whether you can solve each problem using linear transforms or independent observations.

	Linear transform	Independent observation
The amount of coffee in an extra large cup of coffee; X is the amount of coffee in a normal-sized cup.	☑	☐
Drinking an extra cup of coffee per day; X is the amount of coffee in a cup.	☐	☑
Finding the net gain from buying 10 lottery tickets; X is the net gain of buying 1 lottery ticket.	☐	☑
Finding the net gain from a lottery ticket after the price of tickets goes up; X is the net gain of buying 1 lottery ticket.	☑	☐
Buying an extra hen to lay eggs for breakfast; X is the number of eggs laid per week by a certain breed of hen.	☐	☑

The winnings from each lottery ticket are independent of the others.

Changing the price of a ticket changes the expected winnings, but not the probability of winning, so this can be solved with linear transforms.

BULLET POINTS

- **Independent observations of X** are different instances of X. Each observation has the same probability distribution, but the outcomes can be different.
- If $X_1, X_2, ..., X_n$ are independent observations of X then:

$$E(X_1 + X_2 + ... + X_n) = nE(X)$$
$$Var(X_1 + X_2 + ... X_n) = nVar(X)$$

- If X and Y are independent random variables, then:

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

- The expectation and variance of linear transforms of X and Y are given by

$$E(aX + bY) = aE(X) + bE(Y)$$
$$E(aX - bY) = aE(X) - bE(Y)$$
$$Var(aX + bY) = a2Var(X) + b2Var(Y)$$
$$Var(aX - bY) = a2Var(X) + b2Var(Y)$$

You can also add and subtract linear transformations

It doesn't stop there. As well as adding and subtracting random variables, we can also add and subtract their linear transforms.

Imagine what would happen if Fat Dan changed the cost and prizes on both machines, or even just one of them. The last thing we'd want to do is work out the entire probability distribution in order to find the new expectations and variances.

Fortunately, we can take another shortcut.

Suppose the gains on the X and Y slot machines are changed so that the gains for X become aX, and the gains for Y become bY. a and b can be any number.

$$X \rightarrow aX$$

$$Y \rightarrow bY$$

a and b can be any number.

To find the expectation and variance for combinations of aX and bY, we can use the following shortcuts.

Adding aX and bY

If we want to find the expectation and variance of aX + bY, we use

$$E(aX + bY) = aE(X) + bE(Y)$$

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y)$$

It's a linear transform, so we square the numbers here.

We square the numbers because it's a linear transform, just like before.

Subtracting aX and bY

If we subtract the random variables and calculate E(aX - bY) and Var(aX - bY), we use

$$E(aX - bY) = aE(X) - bE(Y)$$

$$Var(aX - bY) = a^2Var(X) + b^2Var(Y)$$

Remember to add the variances.

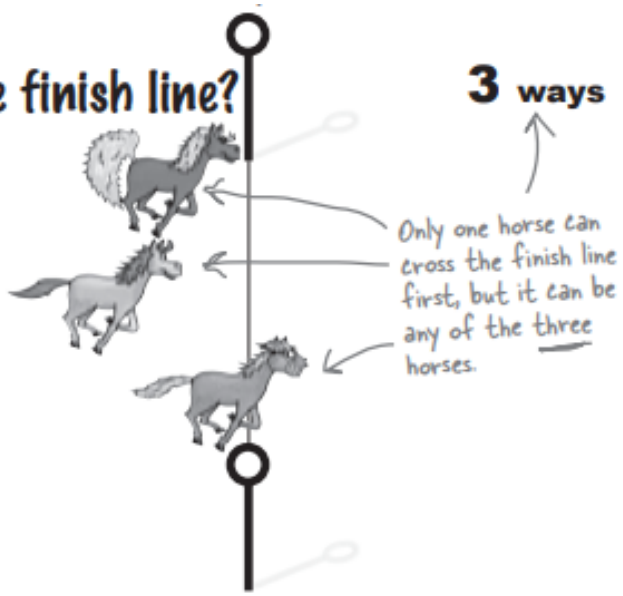
Just as before, we add the variances, even though we're subtracting the random variables.

Permutations & Combinations

How many ways can they cross the finish line?

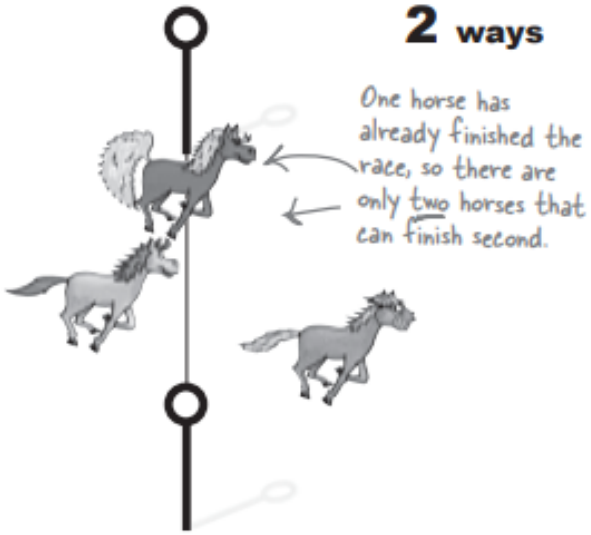
Let's start by looking at the first position of the race.

One of the horses has to win the race, and this can be any one of the three horses taking part. This means that there are three ways of filling the number one position.

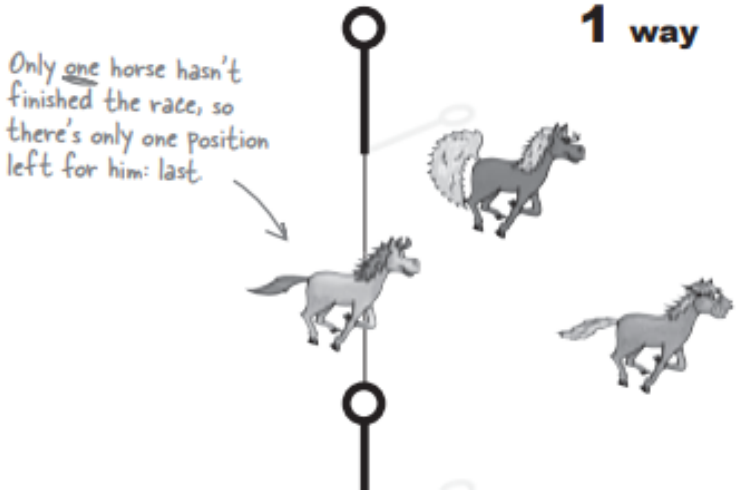


So what about the second position in the race?

If one of the horses has finished the race, this means there are two horses left. Either of these can come second in the race. This means that there are two ways of filling the number two position, no matter which horse came first.



Once two horses have finished the race, there's only one position left for the final horse—third place.



So how does this help us calculate all the possible finishing orders?

Calculate the number of arrangements

We just saw that there were 3 ways of filling the first position, and for each of these, there are 2 ways of filling the second position. And no matter how those first two slots are filled, there's only one way of filling the last position. In other words, the number of ways in which we can fill all three positions is:

2 ways of filling the 2nd position

3 ways of filling the 1st position → **3 × 2 × 1 = 6** ← 6 ways of filling all 3 positions

1 way of filling the 3rd position

This means that we can tell there are 6 different ways of ordering the three horses, without us having to figure out each of the arrangements.

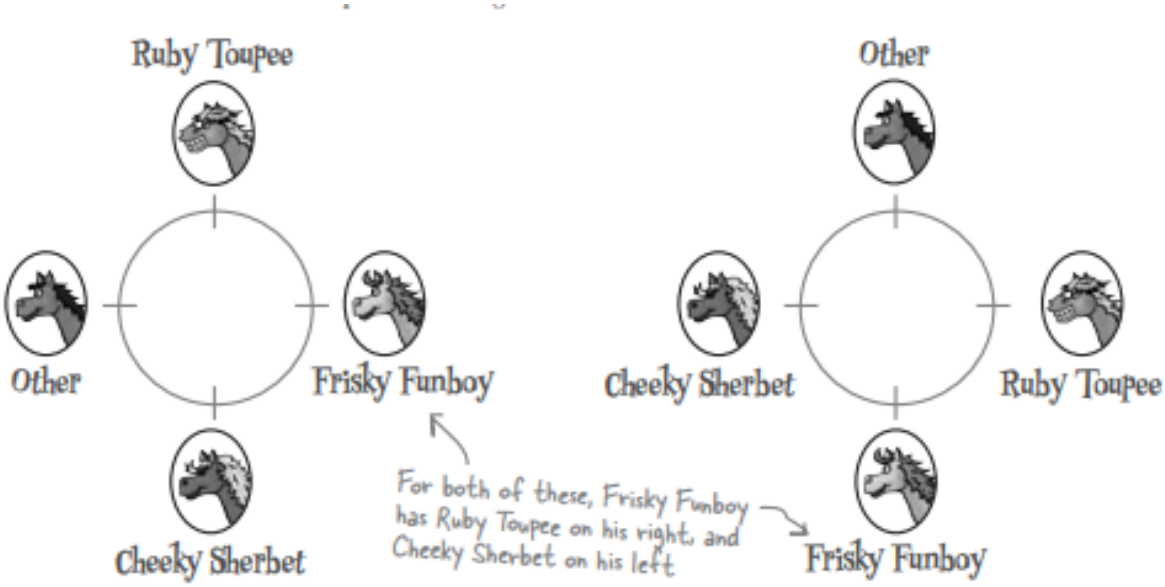
This type of calculation is called the factorial of a number

$$n! = n \times (n - 1) \times (n - 2) \times \ldots \times 3 \times 2 \times 1$$

If, for example, you want to find the number of arrangements of 4 separate objects, all you have to do is calculate 4!, giving you 4 × 3 × 2 × 1 = 24 separate arrangements.

Going round in circles

There's one exception to this rule, and that's if you're arranging objects in a circle.



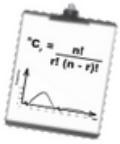
At first glance, these two arrangements look different, but they're actually the same. The horses are in exactly the same positions relative to each other, the only difference is that in the second arrangement, the horses have walked a short distance round the circle. This means that some of the ways in which you can order the horses are actually the same.

So how do we solve this sort of problem?

The key here is to fix the position of one of the horses, say Frisky Funboy. With Frisky Funboy standing in a fixed position, you can count the number of ways in which the remaining 3 horses can be ordered, and this will give you the right result without any duplicates.

In general, if you have n objects you need to arrange in a circle, the number of possible arrangements is given by

(n - 1)! ← The number of ways of arranging n objects in a circle



Vital Statistics

Formulas for arrangements

If you want to find the number of possible arrangements of n objects, use n! where

$n! = n \times (n-1) \times \ldots \times 3 \times 2 \times 1$

In other words, multiply together all the numbers from n down to 1.

If you are arranging n objects in a circle, then there are $(n - 1)!$ possible arrangements.

Generalize a formula for arranging duplicates

Imagine you need to count the number of ways in which n objects can be arranged. Then imagine that k of the objects are alike.

To find the number of arrangements, start off by calculating the number of arrangements for the n objects as if they were all unique. Then divide by the number of ways in which the k objects (the ones that are alike) can be arranged. This gives you:

There are n objects in total. → $\frac{n!}{k!}$ ← If you have n objects where k are alike the number of arrangements is given by n!/k!

k of the objects are alike. →

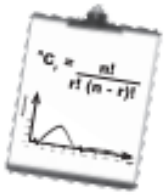
We can take this further.

Imagine you want to arrange n objects, where k of one type are alike, and j of another type are alike, too. You can find the number of possible arrangements by calculating:

There are n objects in total. → $\frac{n!}{j!k!}$ ← The number of ways of arranging n objects where j of one type are alike, and so are k of another type.

j of one type of object are alike, and so are k of another type. →

In general, when calculating arrangements that include duplicate objects, divide the total number of arrangements (n!) by the number of arrangements of each set of alike objects (j!, k!, and so on).



Vital Statistics

Arranging by type

If you want to arrange n objects where j of one type are alike, k of another type are alike, so are m of another type and so on, the number of arrangements is given by

$$\frac{n!}{j!k!m!\ldots}$$

Examining permutations

So how can we rewrite the calculation in terms of factorials?

The number of arrangements is $20 \times 19 \times 18$. Let's rewrite it and see where it gets us.

$$20 \times 19 \times 18 = \frac{20 \times 19 \times 18 \times (17 \times 16 \times \ldots \times 3 \times 2 \times 1)}{(17 \times 16 \times \ldots \times 3 \times 2 \times 1)}$$

If we multiply it by 17!/17!, this will still give us the same answer.

$$= \frac{20!}{17!}$$

This is the same expression written in terms of factorials.

This is the same expression that we had before, but this time written in terms of factorials.

The number of arrangements of 3 objects taken from 20 is called the number of ***permutations***. As you've seen, this is calculated using

$$\frac{20!}{(20 - 3)!}$$

$$= \frac{2,432,902,008,176,640,000}{355,687,428,096,000}$$

This is the same answer we got earlier

$$= 6,840$$

In general, the number of permutations of r objects taken from n is the number of possible way in which each set of r objects can be ordered. It's generally written ${}^{n}P_{r}$ where

This is the total number of objects

This is the number of positions we want to fill

$$^{n}P_{r} = \frac{n!}{(n - r)!}$$

So if you want to know how many ways there are of ordering r objects taken from a pool of n, permutations are the key.

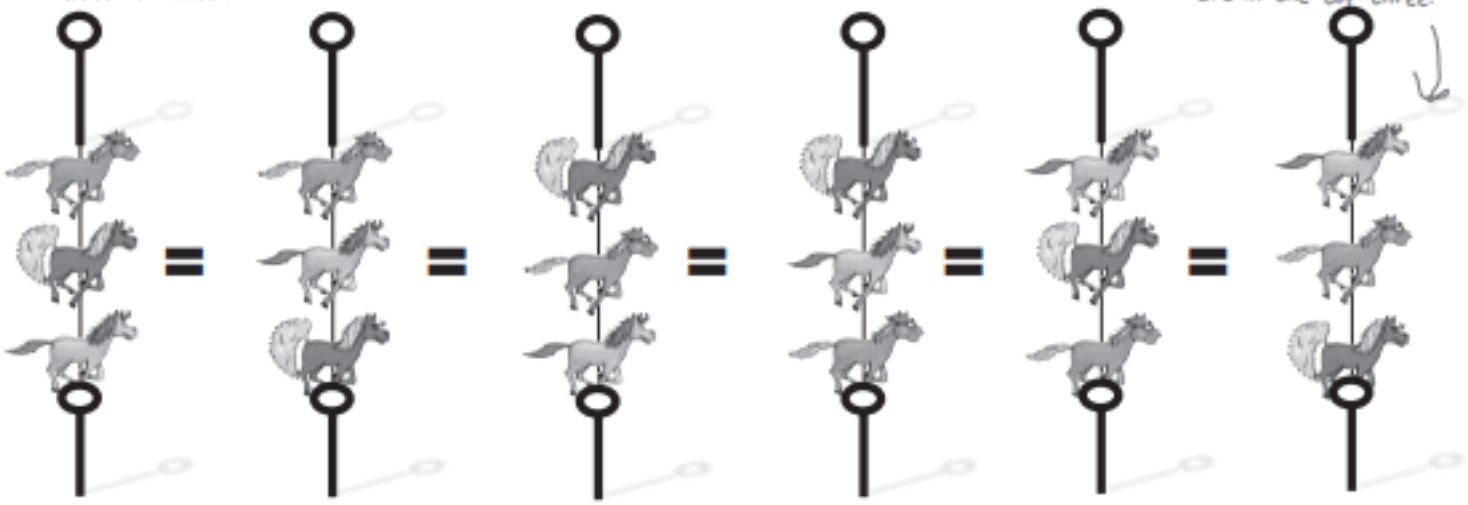
Permutations give the total number of ways you can order a certain number of objects (r), drawn from a larger pool of objects (n).

What if horse order doesn't matter

So far we've found the number of permutations of ordering three horses taken from a group of twenty. This means that we know how many exact arrangements we can make.

This time around, we don't want to know how many different permutations there are. We want to know the number of ***combinations*** of the top three horses instead. We still want to know how many ways there are of filling the top three positions, but this time the exact arrangement doesn't matter.

We don't need to know the precise order in which the horses finish the race, it's enough to know which horses are in the top three.



So how can we solve this sort of problem?

At the moment, the number of permutations includes the number of ways of arranging the 3 horses that are in the top three. There are 3! ways of arranging each set of 3 horses, so let's divide the number of permutations by 3!. This will give us the number of ways in which the top three positions can be filled but *without* the exact order mattering.

The result is

$$\frac{20!}{3!17!} = \frac{6{,}840}{3!}$$

$$= 1{,}140$$

This means that there are 6,840 permutations for filling the first three places in the race, but if you're not concerned about the order, there are 1,140 combinations.

Examining combinations

Earlier on we found a general way of calculating permutations. Well, there's a way of doing this for combinations too.

In general, the number of combinations is the number of ways of choosing r objects from n, without needing to know the exact order of the objects. The number of combinations is written ${}^{n}C_{r}$, where

This is the total number of objects.

This is the number of positions we want to fill.

$$^{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

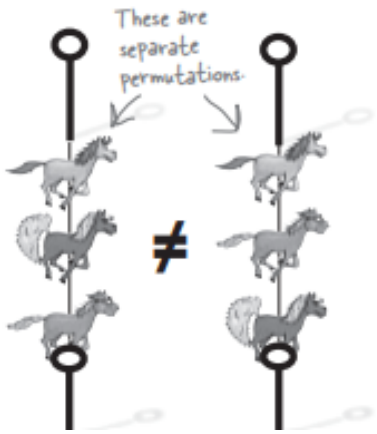
This bit is calculated in the same way as a permutation.

You divide by an extra r! if it's a combination.

Permutations

A ***permutation*** is the number of ways in which you can choose objects from a pool, and where the order in which you choose them counts. It's a lot more specific than a combination as you want to count the number of ways in which you fill each position.

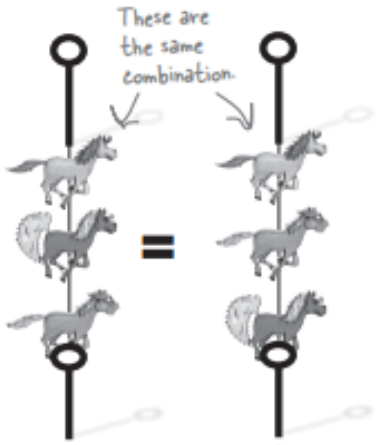
Permutation: order matters.



Combinations

A ***combination*** is the number of ways in which you can choose objects from a pool, without caring about the exact order in which you choose them. It's a lot more general than a permutation as you don't need to know how each position has been filled. It's enough to know which objects have been chosen.

Combination: order doesn't matter.



Vital Statistics

Permutations

If you choose r objects from a pool of n, the number of permutations is given by

$${}^{n}P_{r} = \frac{n!}{(n-r)!}$$

Combinations

If you choose r objects from a pool of n, the number of combinations is given by

$${}^{n}C_{r} = \frac{n!}{r!(n-r)!}$$

Geometric, Binomial & Poisson Distributions

Geometric Distributions

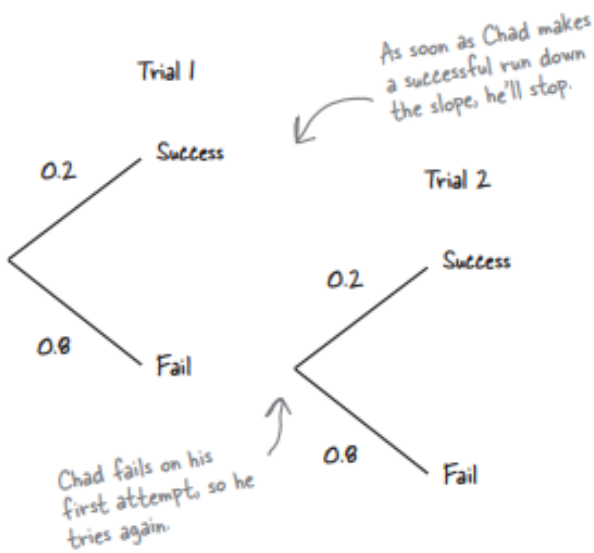
Calculating probability distributions takes time. So far we've looked at how to calculate and use probability distributions, but wouldn't it be nice to have something easier to work with, or just quicker to calculate? In this chapter, we'll show you some special probability distributions that follow very definite patterns. Once you know these patterns, you'll be able to use them to calculate probabilities, expectations, and variances in record time

But what if you needed to look at the probability of him needing fewer than 10 attempts (for insurance reasons), or even 20 or 100

Even though it's neverending, there's still a way of figuring out this type of probability distribution. This is actually a special kind of probability distribution, with special properties that makes it easy to calculate probabilities, along with the expectation and variance.

successful run, he's going to stop..

Here's a probability tree for the first two trials, as these are all that's needed to work out the probabilities.



If we say X is the number of trials needed to get down the slopes, then

$P(X = 1) = P(\text{Success in trial 1})$
$= 0.2$

$P(X = 2) = P(\text{Success in trial 2} \cap \text{Failure in trial 1})$
$= 0.2 \times 0.8$
$= 0.16$

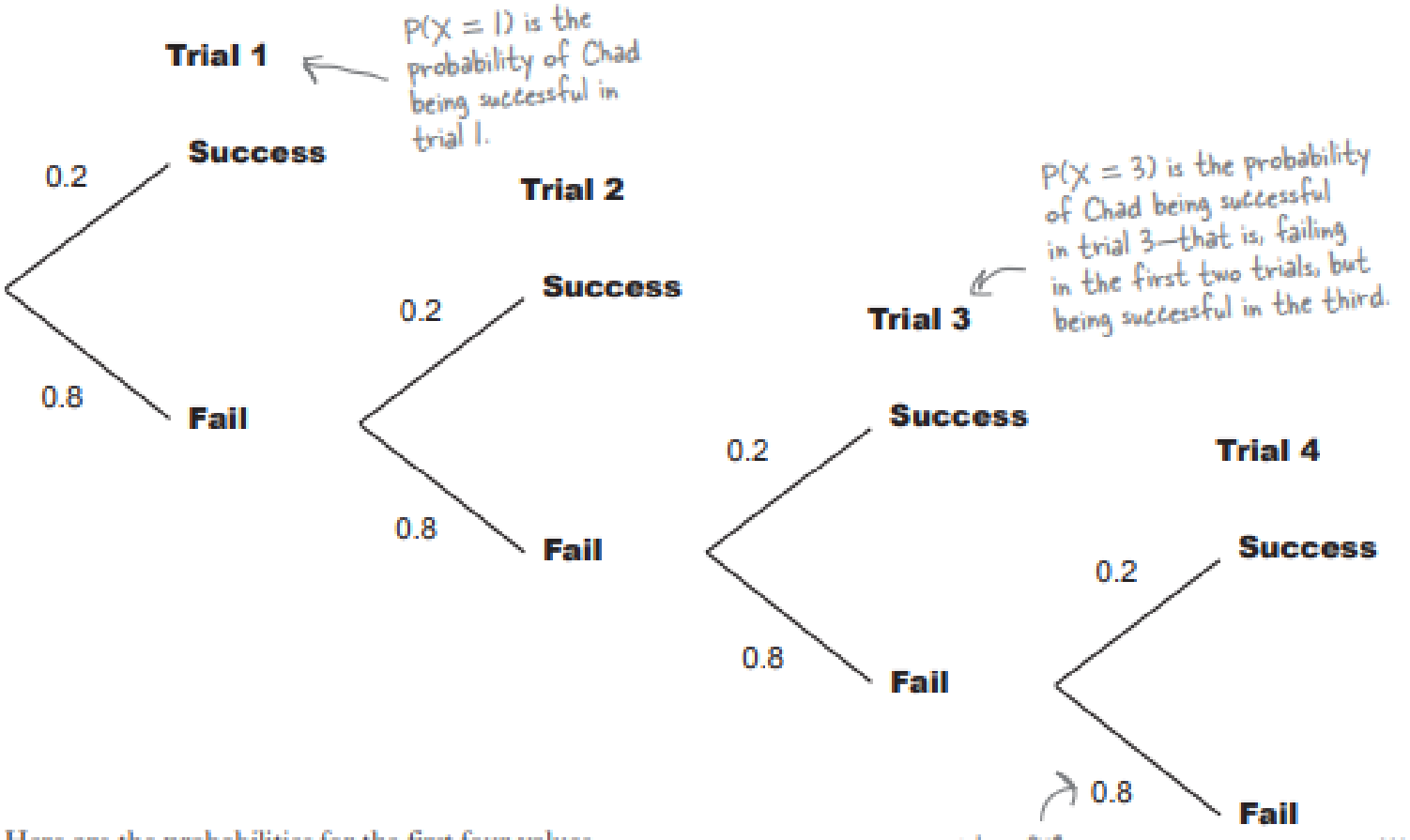
$P(X \leq 2) = P(X = 1) + P(X = 2)$ ← We can add these probabilities because they're independent.
$= 0.2 + 0.16$
$= 0.36$

It's time to exercise your probability skills. The probability of Chad making a successful run down the slopes is 0.2 for any given trial (assume trials are independent). What's the probability he'll need two trials? What's the probability he'll make a successful run down the slope in one **or** two trials? Remember, when he's had his first successful run, he's going to stop..

There's a pattern to this probability distribution

Let's define the variable X to be the number of trials needed for Chad to make a successful run down the slope. Chad only needs to make one successful run, and then he'll stop.

Let's start off by examining the first four trials so that we can calculate probabilities for the first four values of X. By doing this, we can see if there's some sort of pattern that will help us to easily work out the probabilities of other values.



Here are the probabilities for the first four values of X.

These probabilities are calculated using the probability tree.

x	P(X = x)
1	0.2
2	$0.8 \times 0.2 = 0.16$
3	$0.8 \times 0.8 \times 0.2 = 0.128$
4	$0.8 \times 0.8 \times 0.8 \times 0.2 = 0.1024$

Notice each probability is composed by multiplying different powers of 0.8 and 0.2 together.

The probability distribution can be represented algebraically

As you can see, the probabilities of Chad's snowboarding trials follow a particular pattern. Each probability consists of multiples of 0.8 and 0.2. You can quickly work out the probabilities for any value *r* by using:

$$P(X = r) = 0.8^{r-1} \times 0.2$$

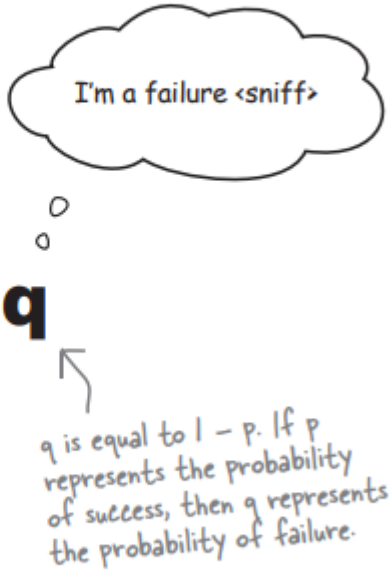
In other words, if you want to find P(X = 100), you don't have to draw an enormous probability tree to work out the probability, or think your way through exactly what happens in every trial. Instead, you can use:

$$P(X = 100) = 0.8^{99} \times 0.2$$

We can generalize this even further. If the probability of success in a trial is represented by ***p*** and the probability of failure is ***1 - p***, which we'll call ***q***, we can work out any probability of this nature by using:

$$\mathbf{P(X = r) = q^{r-1} p}$$

(r – 1) failures and 1 success. In our case, p = 0.2 and q = 0.8.



This formula is called the ***geometric distribution***.

We said that Chad's snowboarding exploits are an example of the ***geometric distribution***. The geometric distribution covers situations where:

1. You run a series of independent trials.
2. There can be either a success or failure for each trial, and the probability of success is the same for each trial.
3. The main thing you're interested in is how many trials are needed in order to get the first successful outcome.

So if you have a situation that matches this set of criteria, you can use the geometric distribution to help you take a few shortcuts. The important thing to be aware of is that we use the word "success" to mean that the event we're interested in happens. If we're looking for an event that has negative connotations, in statistical terms it's still counted as a success.

Let's use the variable X to represent **the number of trials needed to get the first successful outcome**—in other words, the number of trials needed for the event we're interested in to happen.

To find the probability of X taking a particular value *r*, you can get a quick result by using:

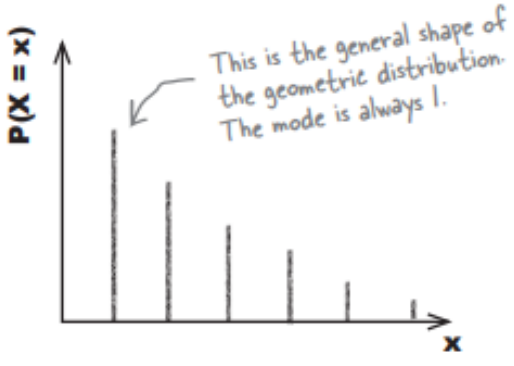
$$\mathbf{P(X = r) = p\, q^{r-1}}$$

where *p* is the probability of success, and q = 1 – p, the probability of failure. In other words, to get a success on the *r*th attempt, there must first have been (r – 1) failures.

The geometric distribution has a distinctive shape.

P(X = r) is at its highest when r = 1, and it gets lower and lower as r increases. Notice that the probability of getting a success is highest for the first trial. This means that **the mode of any geometric distribution is always 1**, as this is the value with the highest probability.

This may sound counterintuitive, but it's most likely that only one attempt will be needed for a successful outcome.



A quick guide to the geometric distribution

Here's a quick summary of everything you could possibly need to know about the Geometric distribution

When do I use it?

Use the Geometric distribution if you're running independent trials, each one can have a success or failure, and you're interested in how many trials are needed to get the first successful outcome

How do I calculate probabilities?

Use the following handy formulae. p is the probability of success in a trial, q = 1 - p, and X is the number of trials needed in order to get the first successful outcome. We say X ~ Geo(p).

$$P(X = r) = p\,q^{r-1}$$

The probability of the first success being in the r'th trial

$$P(X > r) = q^r$$

The probability you'll need more than r trials to get your first success

$$P(X \leq r) = 1 - q^r$$

The probability you'll need r trials or less to get your first success

What about the expectation and variance?

Just use the following

$$E(X) = 1/p \qquad Var(X) = q/p^2$$

Binomial Distribution

There are three different ways of answering exactly one question correctly out of three questions. Another way of looking at this is that there are 3 different combinations.

Just to remind you, a combination ${}^nC_r$ is the number of ways of choosing r objects from n, without needing to know the exact order. This is exactly the situation we have here. We need to choose r correct questions from 3.

This means that the probability of getting r questions correct out of 3 is given by

$$P(X = r) = {}^3C_r \times 0.25^r \times 0.75^{3-r}$$

So, by this formula, the probability of getting 1 question correct is:

$$\begin{aligned} P(X = r) &= {}^3C_1 \times 0.25 \times 0.75^{3-1} \\ &= 3!/(3-1)! \times 0.25 \times 0.5625 \\ &= 6/2 \times 0.0625 \times 0.75 \\ &= 0.422 \end{aligned}$$

Let's generalize the probability further

So far you've seen that the probability of getting r questions correct out of 3 is given by

$$P(X = r) = {}^{3}C_{r} \times 0.25^{r} \times 0.75^{3-r}$$

where the probability of answering a question correctly is 0.25, and the probability of answering incorrectly is 0.75.

The next round of Who Wants To Win A Swivel Chair has 5 questions instead of 3. Rather than rework this probability for 5 questions, let's rework it for n questions instead. That way we'll be able to use the same formula for every round of Who Wants To Win A Swivel Chair.

So what's the formula for the probability of getting r questions right out of n? It's actually

Just replace the 3 with n.

$$P(X = r) = {}^{n}C_{r} \times 0.25^{r} \times 0.75^{n-r}$$

What if the probability of getting a question right changes? I wonder if we can generalize this further.

Yes, we can generalize this further.

Imagine the probability of getting a question right is given by p, and the probability of getting a question wrong is given by 1 – p, or q. The probability of getting r questions right out of n is given by

$$P(X = r) = {}^{n}C_{r} \times p^{r} \times q^{n-r}$$

This sort of problem is called the ***binomial distribution***. Let's take a closer look.

Binomial Distribution Up Close

Guessing the answers to the questions on Who Wants To Win A Swivel Chair is an example of the ***binomial distribution***. The binomial distribution covers situations where

1. You're running a series of independent trials.
2. There can be either a success or failure for each trial, and the probability of success is the same for each trial.
3. There are a finite number of trials.

These two are like the Geometric distribution.

This is different

Just like the geometric distribution, you're running a series of independent trials, and each one can result in success or failure. The difference is that this time you're interested in the number of successes.

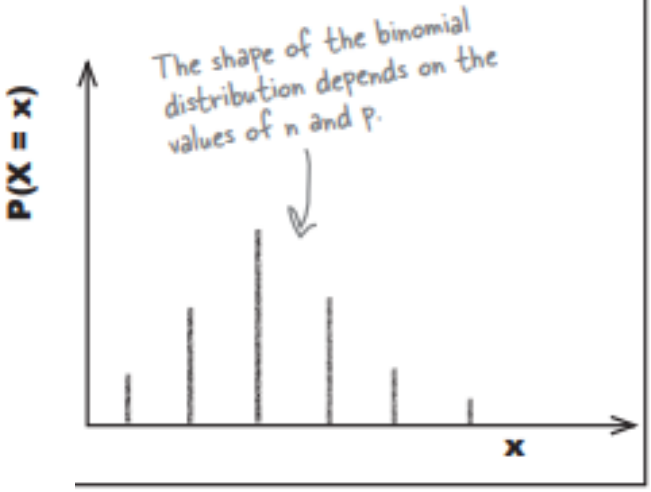
Let's use the variable X to represent **the number of successful outcomes out of n trials**. To find the probability there are r successes, use:

$$P(X = r) = {}^{n}C_{r}\, p^{r}\, q^{n-r} \quad \text{where} \quad {}^{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

p is the probability of a successful outcome in each trial, and n is the number of trials. We can write this as

$$X \sim B(n, p)$$

The exact shape of the binomial distribution varies according to the values of n and p. The closer to 0.5 p is, the more symmetrical the shape becomes. In general it is skewed to the right when p is below 0.5, and skewed to the left when p is greater than 0.5.



Binomial expectation and variance

Let's summarize what we just did. First of all, we took at one trial, where the probability of success is *p*, and where the distribution is binomial. Using this, we found the expectation and variance of a single trial.

We then considered n independent trials, and used shortcuts to find the expectation and variance of *n* trials. We found that if X ~ B(n, p),

$$E(X) = np$$

$$Var(X) = npq$$

These formulae work for any binomial distribution.

This is useful to know as it gives us a quick way of finding the expectation and variance of any probability distribution, without us having to work out lots of individual probabilities.

Your quick guide to the binomial distribution

Here's a quick summary of everything you could possibly need to know about the binomial distribution

When do I use it?

Use the binomial distribution if you're running a fixed number of independent trials, each one can have a success or failure, and you're interested in the number of successes or failures

How do I calculate probabilities?

Use

$$P(X = r) = {}^{n}C_{r}\, p^{r}\, q^{n-r}$$

$${}^{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

where p is the probability of success in a trial, q = 1 - p, n is the number of trials, and X is the number of successes in the n trials.

What about the expectation and variance?

$$E(X) = np \qquad Var(X) = npq$$

Poisson Distribtuion

The popcorn machine malfunctions frequently, with an average of 3.4 breakdowns per week. This is a big concern for the cinema manager because there's a big promotion coming up next week, and the manager wants everything to run smoothly.
A breakdown during the promotion could upset customers and hurt future business.We can't predict the exact number of breakdowns next week. It could be zero, a few, or even more than usual. Regular probability methods (like counting trials) don't work well here because the malfunctions are random events.

The Poisson distribution considers the average malfunction rate (3.4 per week) and allows us to calculate the probability of specific outcomes, like the probability of zero breakdowns next week.

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

Poisson Distribution Up Close

The Poisson distribution covers situations where:

1. Individual events occur at random and independently in a given interval. This can be an interval of time or space—for example, during a week, or per mile.

2. You know the mean number of occurrences in the interval or the rate of occurrences, and it's finite. The mean number of occurrences is normally represented by the Greek letter λ (lambda).

Let's use the variable X to represent **the number of occurrences in the given interval**, for instance the number of breakdowns in a week. If X follows a Poisson distribution with a mean of λ occurrences per interval or rate, we write this as:

$$X \sim Po(\lambda)$$

We're not going to derive it here, but to find the probability that there are r occurrences in a specific

$$P(X = r) = \frac{e^{-\lambda} \lambda^r}{r!}$$

ppearances put you off. straightforward to practice.

The formula for the proba
is some number. It's a stan
even though the formula r
straightforward to use in practice.

mathematical
int. It always
s for 2.718, so you
ust substitute in
this number for e in the Poisson formula. Many scientific calculators have an e^x key that will calculate powers of e for you.

As an example, if $X \sim Po(2)$

$$P(X = 3) = \frac{e^{-2} \times 2^3}{3!}$$

Use the formula and substitute in $r = 3$ and $\lambda = 2$.

$$= \frac{e^{-2} \times 8}{6}$$

$$= e^{-2} \times 1.333$$

$$= 0.180$$

So if X follows a Poisson distribution, what's its expectation and variance? It's easier than you might think...

Expectation and variance for the Poisson distribution

Finding the expectation and variance for the Poisson distribution is a lot easier than finding it for other distributions.

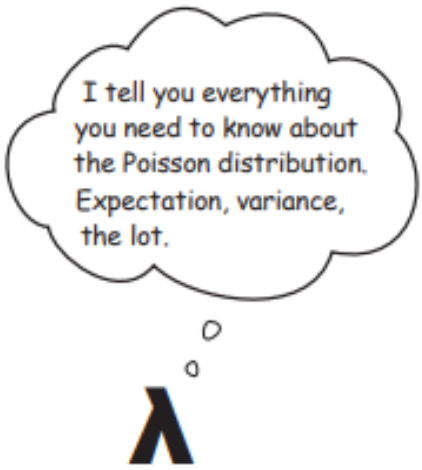
If $X \sim Po(\lambda)$, E(X) is the number of occurrences we can expect to have in a given intervals, so for the popcorn machine, it's the number of breakdowns we can expect to have in a typical week. In other words, E(X) is the mean number of occurrences in the given interval.

Now, if $X \sim Po(\lambda)$, then the mean number of occurrences is given by λ. In other words, E(X) is equal to λ, the parameter that defines our Poisson distribution.

To make things even simpler, the variance of the Poisson distribution is also given by λ, so if $X \sim Po(\lambda)$,

$$\mathbf{E(X) = \lambda} \qquad \mathbf{Var(X) = \lambda}$$

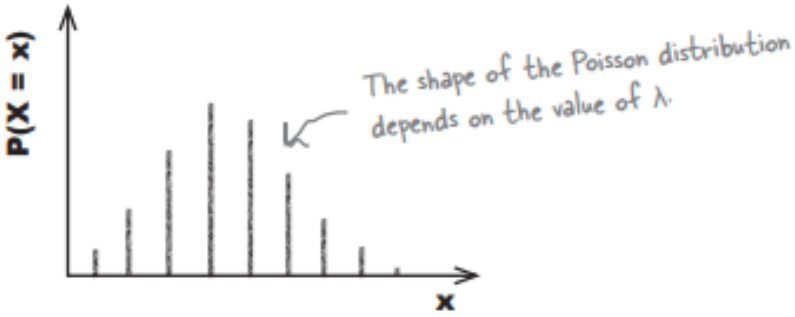
In other words, if you're given a Poisson distribution $Po(\lambda)$, you don't have to calculate anything at all to find the expectation and variance. It's the parameter of the Poisson distribution itself.



What does the Poisson distribution look like?

The shape of the Poisson distribution varies depending on the value of λ. If λ is small, then the distribution is skewed to the right, but it becomes more symmetrical as λ gets larger.

If λ is an integer, then there are two modes, λ and $\lambda - 1$. If λ is not an integer, then the mode is λ.



Your quick guide to the Poisson distribution

Here's a quick summary of everything you could possibly need to know about the Poisson distribution

When do I use it?

Use the Poisson distribution if you have independent events such as malfunctions occurring in a given interval, and you know λ, the mean number of occurrences in a given interval. You're interested in the number of occurrences in one particular interval.

How do I calculate probabilities, and the expectation and variance?

Use

$$\mathbf{P(X = r) = \frac{e^{-\lambda} \lambda^r}{r!}} \qquad \mathbf{E(X) = \lambda} \qquad \mathbf{Var(X) = \lambda}$$

How do I combine independent random variables?

If $X \sim Po(\lambda_x)$ and $Y \sim Po(\lambda_y)$, then

$$\mathbf{X + Y \sim Po(\lambda_x + \lambda_y)}$$

What connection does it have to the binomial distribution?

If $X \sim B(n, p)$, where n is large and p is small, then X can be approximated using

$$\mathbf{X \sim Po(np)}$$

Combine Poisson variables

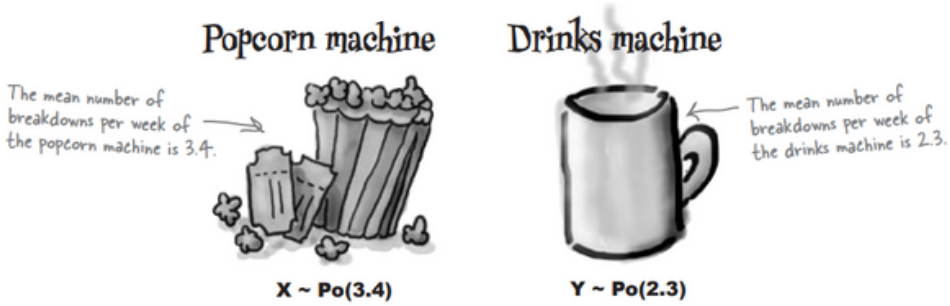
You saw in previous chapters that if X and Y are independent random variables, then

$$P(X + Y) = P(X) + P(Y)$$
$$E(X + Y) = E(X) + E(Y)$$

This means that if $X \sim Po(\lambda_x)$ and $Y \sim Po(\lambda_y)$,

$$\mathbf{X + Y \sim Po(\lambda_x + \lambda_y)}$$

This means that if X and Y both follow Poisson distributions, then so does X + Y. In other words, we can use our knowledge of the way both X and Y are distributed to find probabilities for X + Y.



BULLET POINTS

- The **geometric distribution** applies when you run a series of independent trials, there can be either a success or failure for each trial, the probability of success is the same for each trial, and the main thing you're interested in is how many trials are needed in order to get your first success.
- If the conditions are met for the geometric distribution, X is the number of trials needed to get the first successful outcome, and p is the probability of success in a trial, then

  $X \sim Geo(p)$

- The following probabilities apply if $X \sim Geo(p)$:

  $P(X = r) = pq^{r-1}$
  $P(X > r) = q^r$
  $P(X \leq r) = 1 - q^r$

- If $X \sim Geo(p)$ then

  $E(X) = 1/p$
  $Var(X) = q/p^2$

- The **binomial distribution** applies when you run a series of finite independent trials, there can be either a success or failure for each trial, the probability of success is the same for each trial, and the main thing you're interested in is the number of successes in the n independent trials.
- If the conditions are met for the binomial distribution, X is the number of successful outcomes out of n trials, and p is the probability of success in a trial, then

  $X \sim B(n, p)$

- If $X \sim B(n, p)$, you can calculate probabilities using

  $P(X = r) = {}^nC_r\, p^r\, q^{n-r}$

  where

  $${}^nC_r = \frac{n!}{r!\,(n-r)!}$$

- If $X \sim B(n, p)$, then

  $E(X) = np$
  $Var(X) = npq$

- The **Poisson distribution** applies when individual events occur at random and independently in a given interval, you know the mean number of occurrences in the interval or the rate of occurrences and this is finite, and you want to know the number of occurrences in a given interval.
- If the conditions are met for the Poisson distribution, X is the number of occurrences in a particular interval, and λ is the rate of occurrences, then

  $X \sim Po(\lambda)$

- If $X \sim Po(\lambda)$ then

  $$P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!}$$
  $E(X) = \lambda$
  $Var(X) = \lambda$

- If $X \sim Po(\lambda_x)$, $Y \sim Po(\lambda_y)$ and X and Y are independent,

  $X + Y \sim Po(\lambda_x + \lambda_y)$

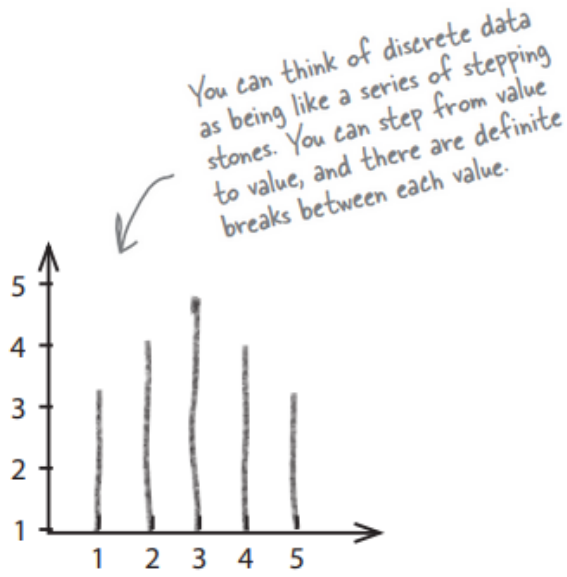
- If $X \sim B(n, p)$ where n is large and p is small, you can approximate it with $X \sim Po(np)$.

Normal Distribution

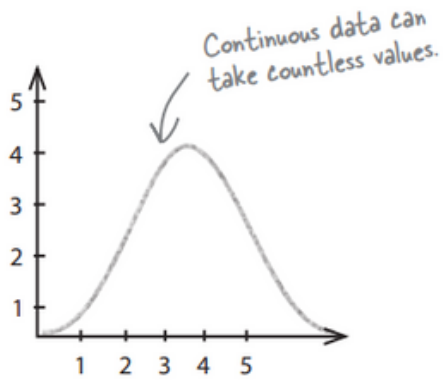
Discrete probability distributions can't handle every situation. So far we've looked at probability distributions where we've been able to specify exact values, but this isn't the case for every set of data. Some types of data just don't fit the probability distributions we've encountered so far. In this chapter, we'll take a look at how continuous probability distributions work

Discrete data takes exact values...

By this we mean the data is composed of distinct numeric values, and we're been able to calculate the probability of each of these values. As an example, when we looked at the probability distribution for the winnings on a slot machine, the possible amounts we could win on each game were very precise



suppose you were asked to accurately measure pieces of string that are between 10 inches and 11 inches long. You could have measurements of 10 inches, 10.1 inches, 10.01 inches, and so on, as the length could be anything within that range. Numeric data like this is called continuous. It's frequently data that is measured in some way rather than counted, and a lot depends on the degree of precision you need to measure to.



The problem is that a lot of real-world problems involve continuous data, and discrete probability distributions just don't work with this sort of data. To find probabilities for continuous data, you need to know about continuous data and **continuous probability distributions.**

Normal Probability Distritbution

We need to find the probability that Julie will have to wait for more than 5 minutes for her date to turn up. The trouble is, the amount of time Julie has to wait is continuous data, which means the probability distributions we've learned thus far don't apply.

When we were dealing with discrete data, we were able to produce a specific probability distribution. We could do this by either showing the probability of each value in a table, or by specifying whether it followed a defined probability distribution, such as the binomial or Poisson distribution. By doing this, we were able to specify the probability of each possible value. As an example, when we found the probability distribution for the winnings per game for one of Fat Dan's slot machines, we knew all of the possible values for the winnings and could calculate the probability of each one..

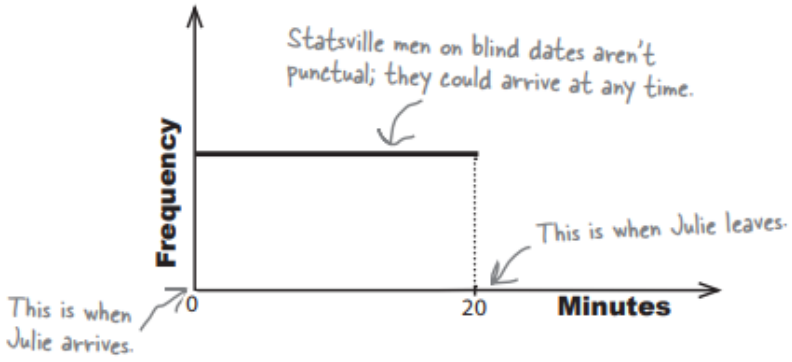
With discrete data, we could give the probability of each value.

x	-1	4	9	14	19
P(X = x)	0.977	0.008	0.008	0.006	0.001

For continuous data, it's a different matter. We can no longer give the probability of each value because it's impossible to say what each of these precise values is. As an example, Julie's date might turn up after 4 minutes, 4 minutes 10 seconds, or 4 minutes 10.5 seconds. Counting the number of possible options would be impossible. Instead, we need to focus on a particular level of accuracy and the probability of getting a **range** of values.

I get it. For discrete probability distributions, we look at the probability of getting a particular **value**; for continuous probability distributions, we look at the probability of getting a particular **range**.

Here's a sketch of the frequency showing the amount of time Julie spends waiting for her date to arrive:

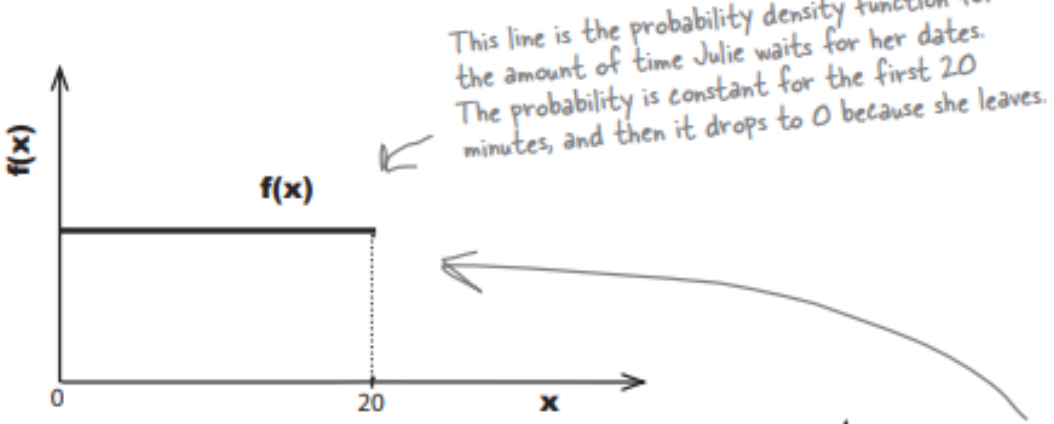


Probability density functions can be used for continuous data

We can describe the probability distribution of a continuous random variable using a ***probability density function***.

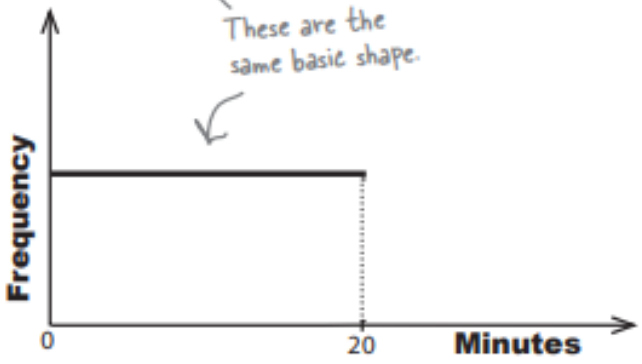
A probability density function f(x) is a function that you can use to find the probabilities of a continuous variable across a range of values. It tells us what the shape of the probability distribution is.

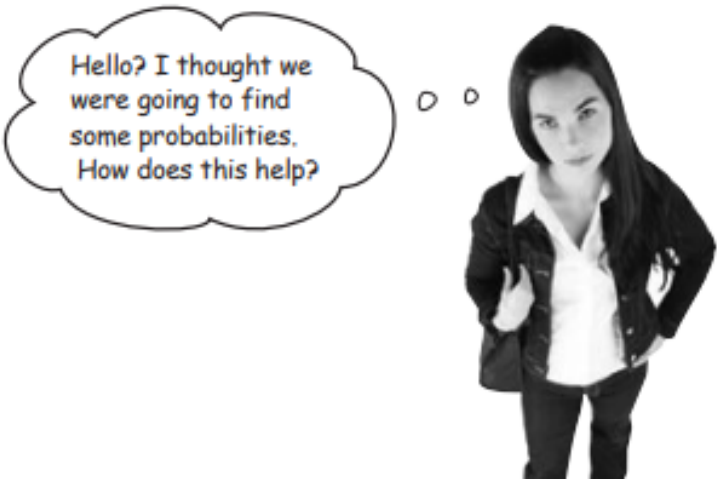
Here's a sketch of the probability density function for the amount of time Julie spends waiting for her date to turn up:



Can you see how it matches the shape of the frequency? This isn't just a coincidence.

Probability is all about how likely things are to happen, and the frequency tells you how often values occur. The higher the relative frequency, the higher the probability of that value occurring. As the frequency for the amount of time Julie has to wait is constant across the 20 minute period, this means that the probability density function is constant too.

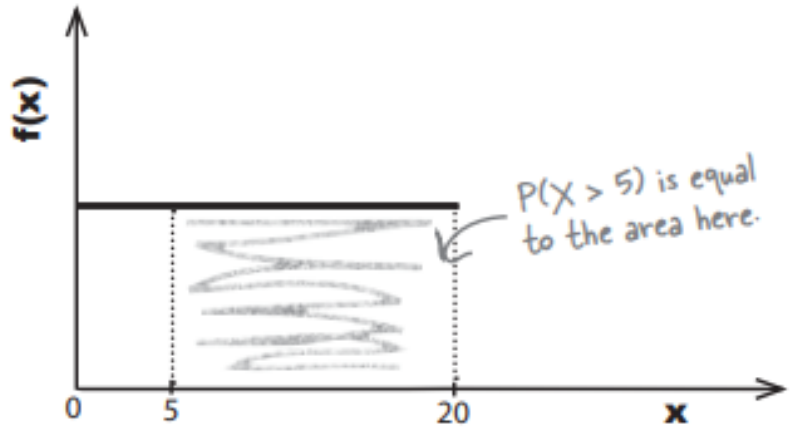




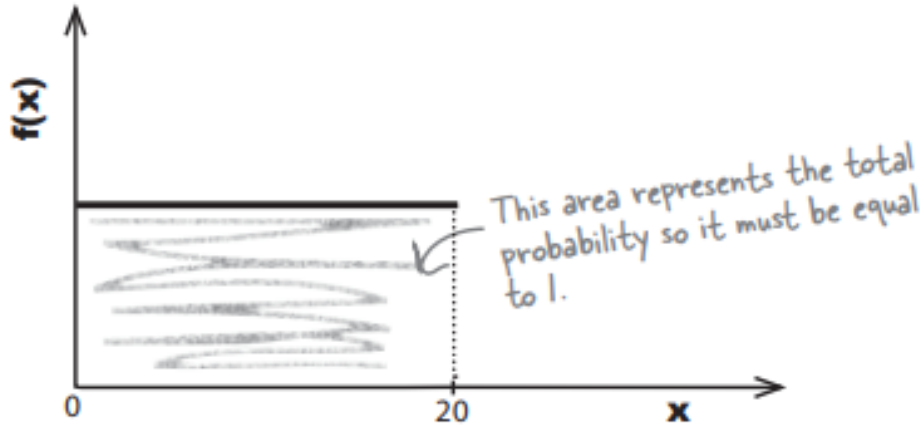
Probability = area

For continuous random variables, probabilities are given by area. To find the probability of getting a particular range of values, we start off by sketching the probability density function. The probability of getting a particular range of values is given by the area under the line between those values.

As an example, we want to find the probability that Julie has to wait for between 5 and 20 minutes for her date to turn up. We can find this probability by sketching the probability density function, and then working out the area under it where x is between 5 and 20.



The total area under the line must be equal to 1, as the total area represents the total probability. This is because for any probability distribution, the total probability must be equal to 1, and, therefore, the area must be too.

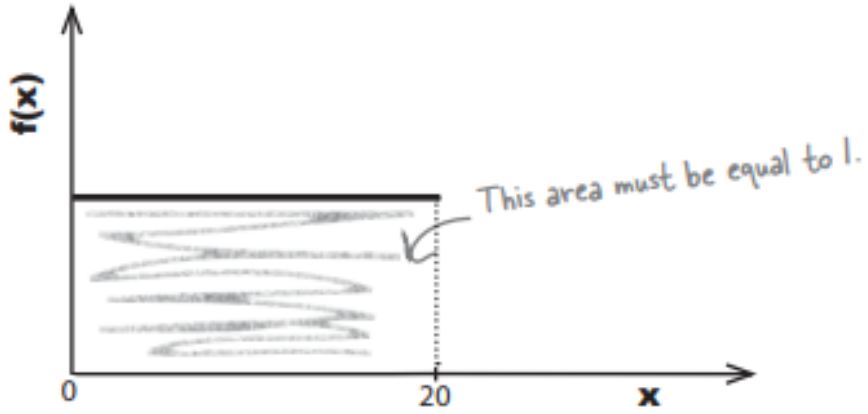


Let's use this to help us find the probability that Julie will need to wait for over 5 minutes for her date to arrive.

To calculate probability, start by finding f(x)...

Before we can find probabilities for Julie, we need to find f(x), the probability density function.

So far, we know that f(x) is a constant value, and we know that the total area under it must be equal to 1. If you look at the sketch of f(x), the area under it forms a rectangle where the width of the base is 20. If we can find the height of the rectangle, we'll have the value of f(x).



We find the area of a rectangle by multiplying its width and height together. This means that

$$1 = 20 \times \text{height}$$

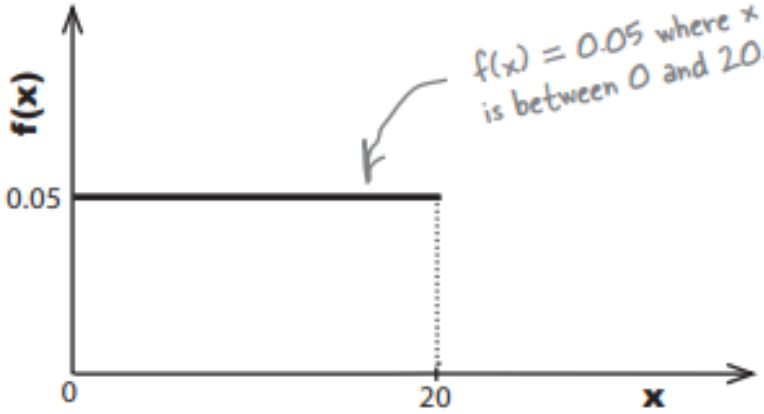
$$\text{height} = 1/20$$

$$= 0.05$$

This means that f(x) must be equal to 0.05, as that ensures the total area under it will be 1. In other words,

$$f(x) = 0.05 \qquad \text{where x between 0 and 20}$$

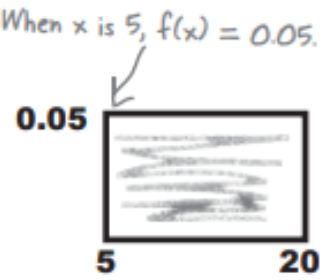
Here's a sketch:



Now that we've found the probability density function, we can find P(X > 5).

...then find probability by finding the area

The area under the probability density line between 5 and 20 is a rectangle. This means that calculating the area of this rectangle will give us the probability P(X > 5).



$$P(X > 5) = (20 - 5) \times 0.05$$
$$= 0.75$$

Area of rectangle = base x height.

So the probability that Julie will have to wait for more than 5 minutes is 0.75.

Do I **have** to use area to find probability? Can't I just pick all the exact values in that range and add their probabilities together? That's what we did for discrete probabilities.

That doesn't work for continuous probabilities.

For continuous probabilities, we *have* to find the probability by calculating the area under the probability density line.

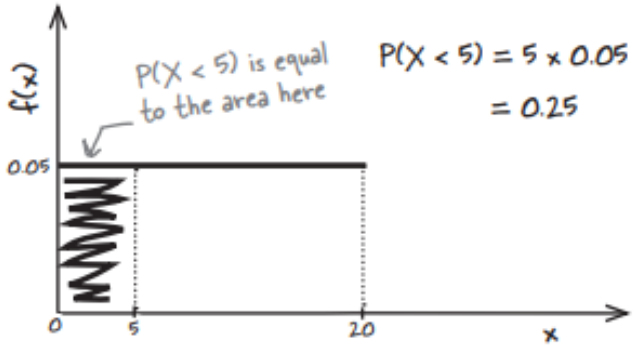
We can't add together the probability of getting each value within the range as there are an infinite number of values. It would take forever.

The only way we can find the probability for continuous probability distributions is to work out the area underneath the curve formed by the probability density function.

When dealing with continuous data, you calculate probabilities for a range of values.

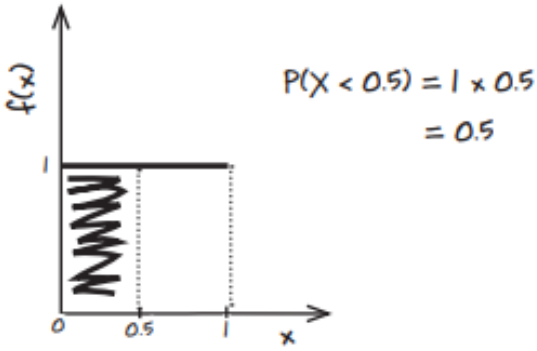
1. f(x) = 0.05 where 0 < x < 20

Find P(X < 5)



2. f(x) = 1 where 0 < x < 1

Find P(X < 0.5)

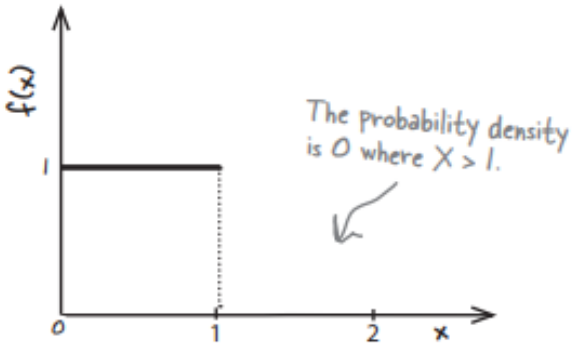


3. f(x) = 1 where 0 < x < 1

Find P(X > 2)

The upper limit of x for this probability density function is 1, which means that it's 0 above this.

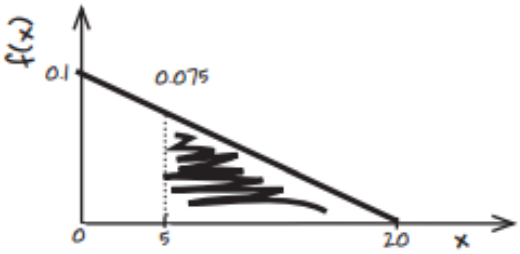
P(X > 2) = 0



4. f(x) = 0.1 – 0.005x where 0 < x < 20

Find P(X > 5)

When x = 5, f(x) = 0.075. This means we have to find the area of a right-angled triangle with height 0.075 and width 15.



P(X > 5) = (0.075 × 15)/2

= 1.125/2

= 0.5625

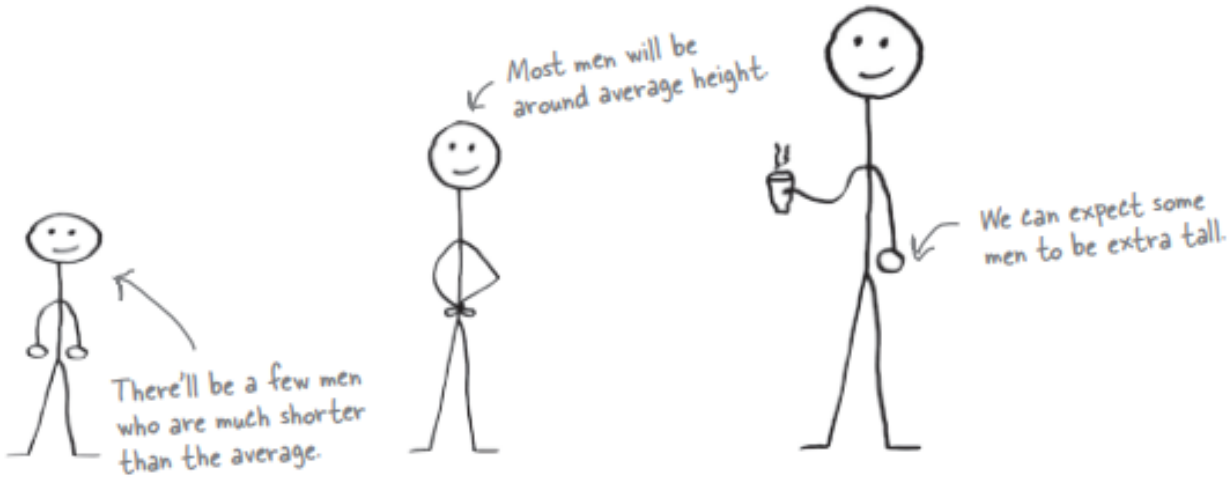
The area of a triangle is 1/2 the base multiplied by the height.

BULLET POINTS

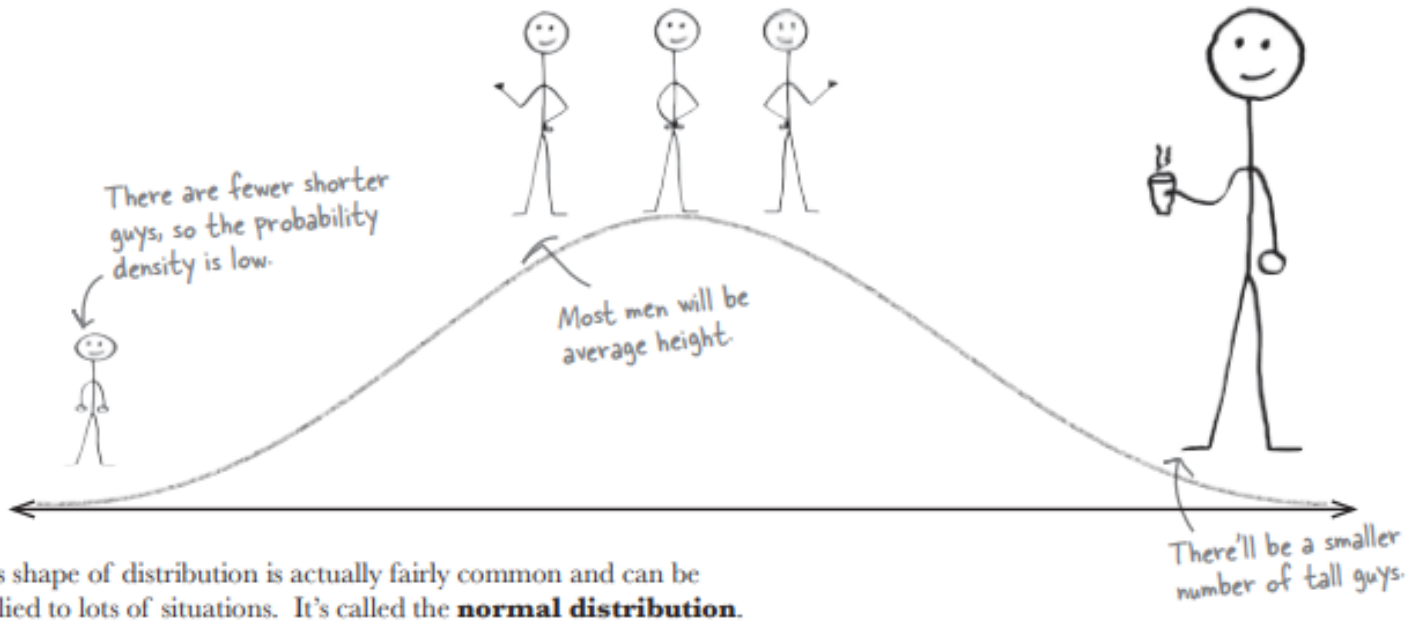
- **Discrete data** is composed of distinct numeric values.
- **Continuous data** covers a range, where any value within that range is possible. It's frequently data that is measured in some way, rather than counted.
- Continuous probability distributions can be described with a probability density function.
- You find the probability for a range of values by calculating the area under the probability density function between those values. So to find P(a < X < b), you need to calculate the area under the probability density function between a and b.
- The total area under the probability density function must equal 1.

Male modelling

So far we've looked at very simple continuous distributions, but it's unlikely these will model the heights of the men Julie might be dating. It's likely we'll have several men who are quite a bit shorter than average, a few really tall ones, and a lot of men somewhere in between. We can expect most of the men to be average height.



Given this pattern, the probability density of the height of the men is likely to look something like this.



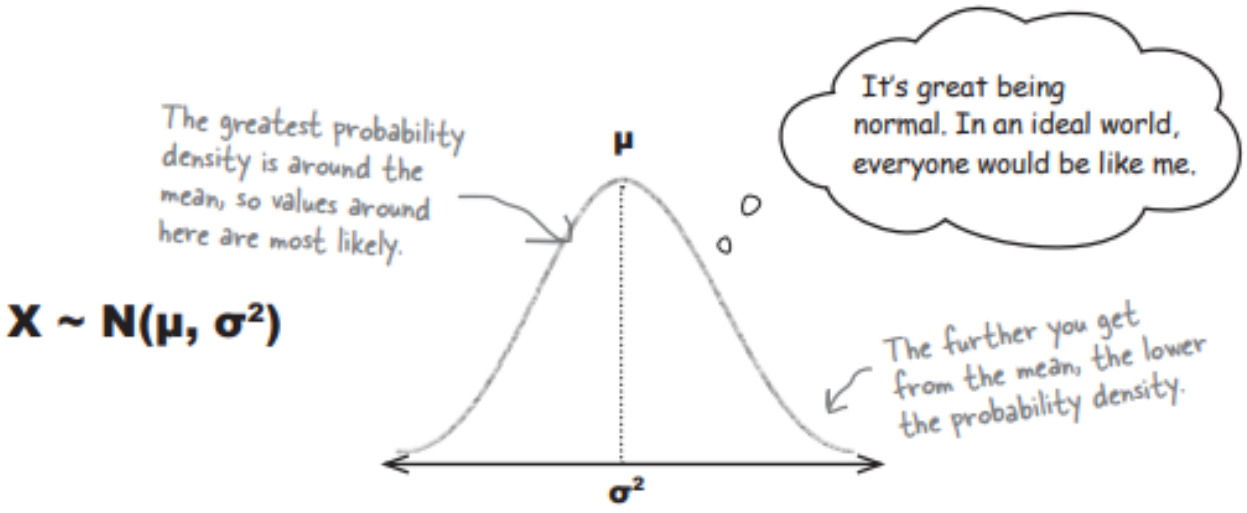
This shape of distribution is actually fairly common and can be applied to lots of situations. It's called the **normal distribution**.

The normal distribution is an "ideal" model for continuous data

The normal distribution is called normal because it's seen as an ideal. It's what you'd "normally" expect to see in real life for a lot of continuous data such as measurements.

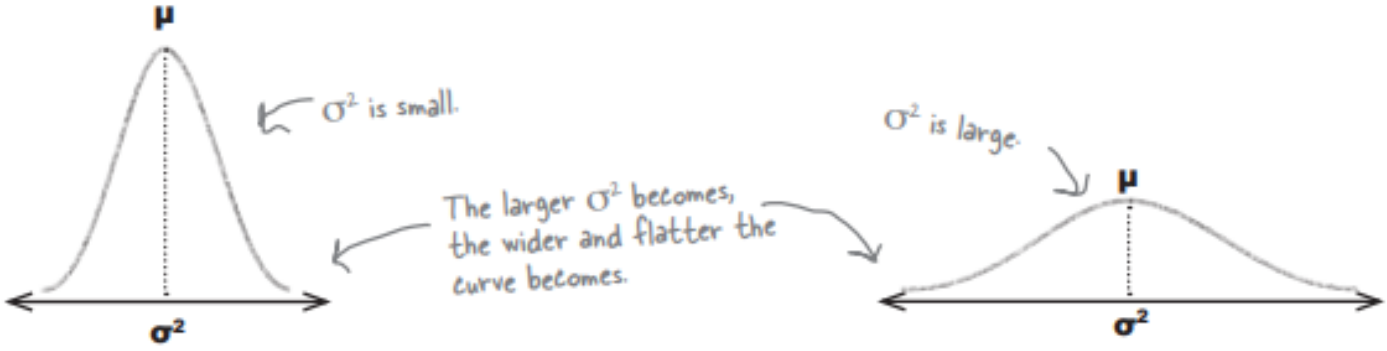
The normal distribution is in the shape of a bell curve. The curve is symmetrical, with the highest probability density in the center of the curve. The probability density decreases the further away you get from the mean. Both the mean and median are at the center and have the highest probability density.

The normal distribution is defined by two parameters, μ and σ^2. μ tells you where the center of the curve is, and σ gives you the spread. If a continuous random variable X follows a normal distribution with mean μ and standard deviation σ, this is generally written $X \sim N(\mu, \sigma^2)$.



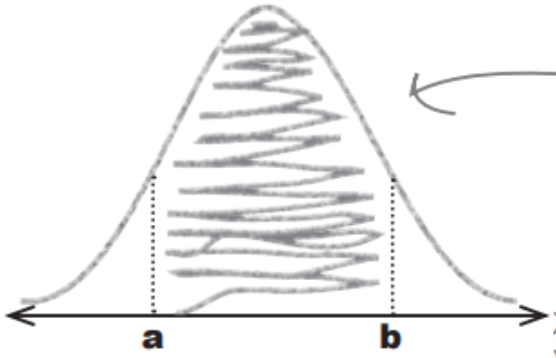
So what effect do μ and σ really have on the shape of the normal distribution?

We said that μ tells you where the center of the curve is, and σ^2 indicates the spread of values. In practice, this means that as σ^2 gets larger, the flatter and wider the normal curve becomes.



So how do we find normal probabilities?

As with any other continuous probability distribution, you find probabilities by calculating the area under the curve of the distribution. The curve gives the probability density, and the probability is given by the area between particular ranges. If, for instance, you wanted to find the probability that a variable X lies between a and b, you'd need to find the area under the curve between points a and b.



The shaded area gives the probability between that X is between a and b.

X is the variable we're trying to find probabilities for.

Standard Score

$$Z = \frac{X - \mu}{\sigma}$$

The mean of X

The standard deviation of X

You can find normal probability tables in the appendix at the back of the book.

Here's the column for .06, the second decimal place for z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170

Here's the row for $z = -1.5x$, where x is some number.

This is where -1.5 and .06 meet. It gives the value of $P(Z < z)$.

So, looking up the value of –1.56 in the probability table gives us a probability of 0.0594. In other words, $P(Z < -1.56) = 0.0594$. This means that

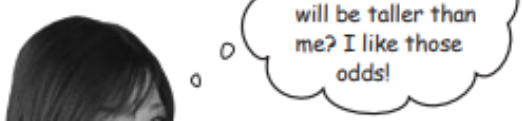
$$P(Z > -1.56) = 1 - P(Z < -1.56)$$

The total probability is 1, so the total area under the curve is 1.

$$= 1 - 0.0594$$

$$= 0.9406$$

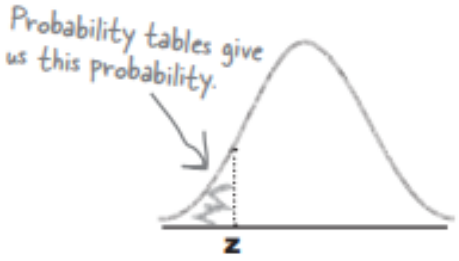
In other words, the probability that Julie's date is taller than her is 0.9406.



Probability Tables Up Close

Probability tables allow you to look up the probability $P(Z < z)$ where z is some value. The problem is you don't always want to find this sort of probability; sometimes you want to find the probability that a continuous random variable is greater than z, or between two values. How can you use probability tables to find the probability you need?

The big trick is to find a way of using the probability tables to get to what you want, usually by finding a whole area and then subtracting what you don't need.



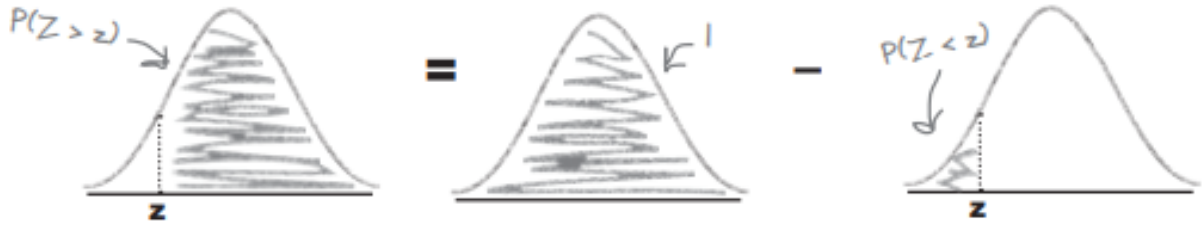
Finding P(Z > z)

We can find probabilities of the form $P(Z > z)$ using

$$P(Z > z) = 1 - P(Z < z)$$

We've already used this to find the probability that Julie is taller than her date.

In other words, take the area where $Z < z$ away from the total probability.



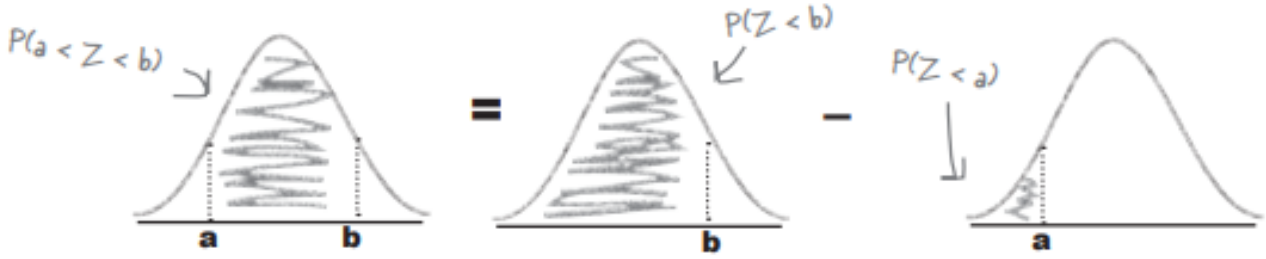
Finding P(a < Z < b)

Finding this sort of probability is slightly more complicated to calculate, but it's still possible. You can calculate this sort of probability using

$$P(a < Z < b) = P(Z < b) - P(Z < a)$$

You could use this to find the probability that the height of Julie's date is within a particular range.

In other words, calculate $P(Z < b)$, and take away the area for $P(Z < a)$.



BULLET POINTS

- The normal distribution forms the shape of a symmetrical bell curve. It's defined using $N(\mu, \sigma^2)$.
- To find normal probabilities, start by identifying the probability range you need. Then find the standard score for the limit of this range using

$$Z = \frac{X - \mu}{\sigma} \quad \text{where } Z \sim N(0, 1).$$

- You find normal probabilities by looking up your standard score in probability tables. Probability tables give you the probability of getting this value or lower.

Statistical Sampling

Statistics deal with data, but where does it come from? Some of the time, data's easy to collect, such as the ages of people attending a health club or the sales figures for a games company. But what about the times when data isn't so easy to collect? Sometimes the number of things we want to collect data about are so huge that it's difficult to know where to start. In this chapter, we'll take a look at how you can effectively gather data in the real world, in a way that's efficient, accurate, and can also save you time and money to boot.

Sample

A statistical sample is a selection of items taken from a population. You choose your sample so that it's fairly representative of the population as a whole; it's a representative subset of the population. For Mighty Gumball, a sample of gumballs means just a small selection of gumballs rather than every single one of them.

Sample survey

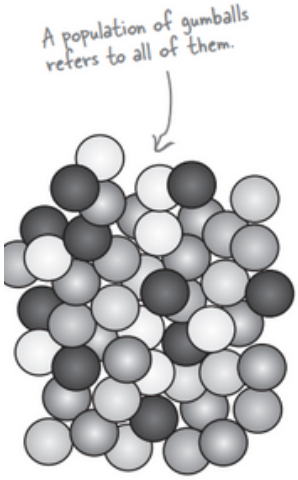
A study or survey involving just a sample of the population is called a sample survey. A lot of the time, conducting a survey is more practical than a census. It's usually less time-consuming and expensive, as you don't have to deal with the entire population. And because you don't use the whole population, taking a sample survey of the gumballs means that there'll be plenty left over when you're done.

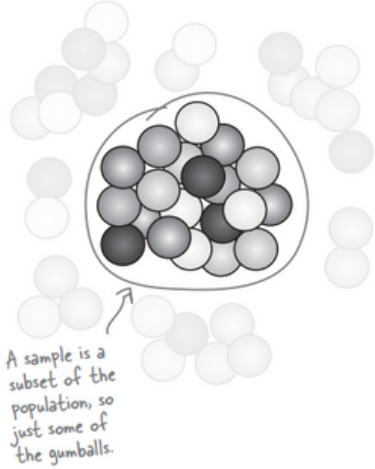
Populations

A statistical population refers to the entire group of things that you're trying to measure, study, or analyze. It can refer to anything from humans to scores to gumballs. The key thing is that a population refers to all of them

Census

A census is a study or survey involving the entire population, so in the case of Mighty Gumball, they're conducting a census of their gumball population by tasting every single one of them. A census can provide you with accurate information about your population, but it's not always practical. When populations are large or infinite, it's just not possible to include every member.



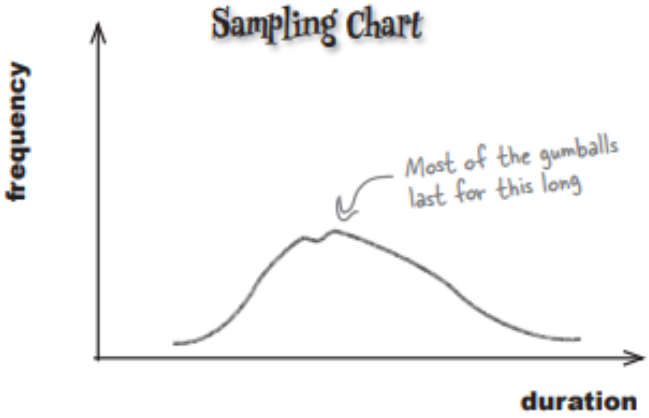


How sampling works

The key to creating a good sample is to choose one that is as close a match to your population as possible. If your sample is representative, this means it has similar characteristics to the population

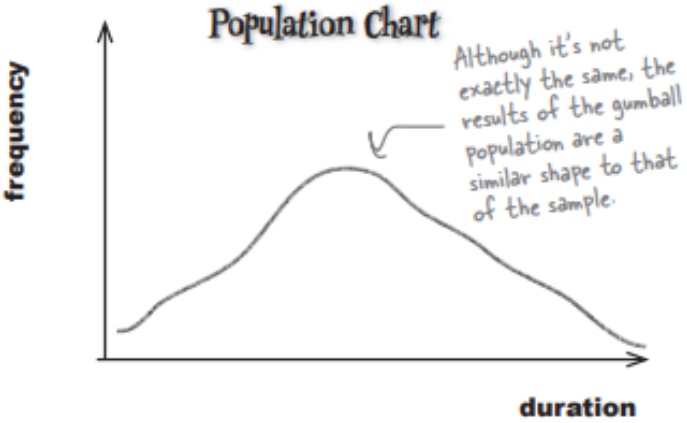
Even though you've only tried a small sample of gumballs, you still have an impression of the shape of the distribution, and the more gumballs you try, the clearer the shape is. As an example, you can get a rough impression of where the center of the population distribution is by looking at the shape of the sample distribution.

Let's compare this with the actual population:



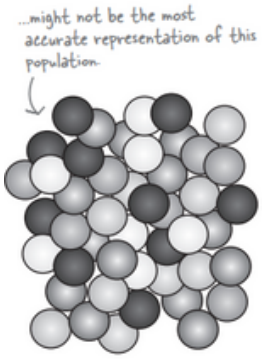
Here's the chart for the population. Can you see how closely the sample and population distributions agree?

If you compare the two charts, the overall shape is very similar, even though one is for *all* of the gumballs and the other is for just *some* of them. They share key characteristics such as where the center of the data is, and this means you can use the sample data to make predictions about the population.

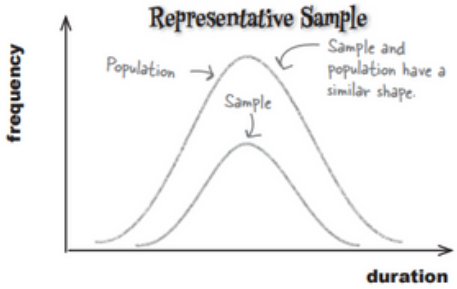


When sampling goes wrong

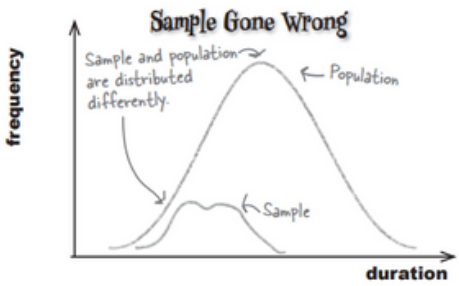
Using the wrong sample could lead you to draw wrong conclusions about population parameters, such as the mean or standard deviation. You might be left with a completely different view of your data, and this could lead you to make the wrong decisions. The trouble is, you might not know this at the time. You might think your population is one thing when in fact it's not. We need to make sure we have some mechanism for making sure our samples are a reliable representation of the population.



We want this:

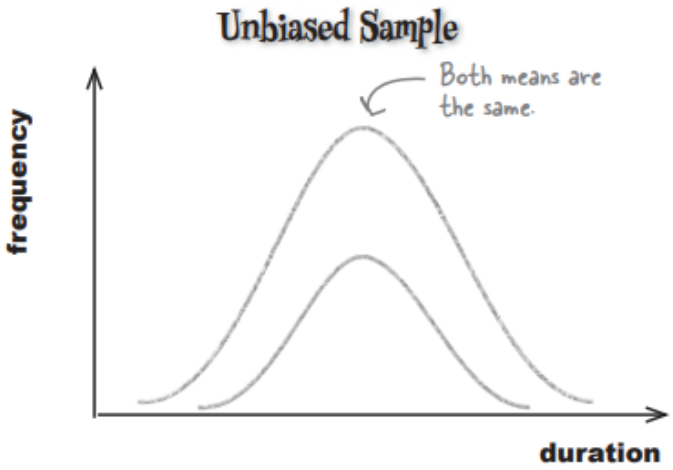


Instead of this:



Sometimes samples can be biased

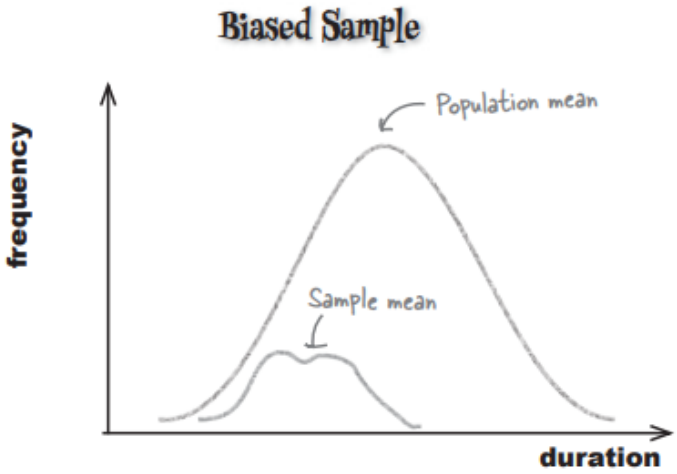
Bias is a sort of favoritism that you can unwittingly (or maybe knowingly) introduce into your sample, meaning that your sample is no longer randomly selected from your population.If a sample is unbiased, then it's representative of the population. It's a fair reflection of what the population is like.



Unbiased samples

An ***unbiased sample*** is representative of the target population. This means that it has similar characteristics to the population, and we can use these to make inferences about the population itself.

The shape of the distribution of an unbiased sample is similar to the shape of the population it comes from. If we know the shape of the sample distribution, we can use it to predict that of the population to a reasonable level of confidence.



Biased samples

A ***biased sample*** is not representative of the target population. We can't use it to make inferences about the population because the sample and population have different characteristics. If we try to predict the shape of the population distribution from that of the sample, we'd end up with the wrong result.

Sources of bias

w does bias creep into samples? Through any of the following and more:

A sampling frame where items have been left off, such that not everything in the target population is included. If it's not in your sampling frame, it won't be in your sample.

An incorrect sampling unit. Instead of individual gumballs, maybe the sampling unit should have been boxes of gumballs instead.

Individual sampling units you chose for your sample weren't included in your actual sample. As an example, you might send out a questionnaire that not everybody responds to.

Poorly designed questions in a questionnaire. Design your questions so that they're neutral and everyone can answer them. An example of a biased question is "Mighty Gumball candy is tastier than any other brand, do you agree?" It would be better to ask the person being surveyed for the name of their favorite brand of confectionary.

Samples that aren't random. As an example, if you're conducting a survey on the street, you may avoid questioning anyone that looks too busy to stop, or too aggressive. This means that you exclude aggressive or busy-looking people from your survey.

How to choose your sample

We've looked at how to design your sample and explored types of bias that need to be avoided. Now we need to select our actual sample from the sample frame. But how should we go about this?

Simple random sampling

One option is to choose the sample at random. Imagine you have a population of N sampling units, and you need to pick a sample of n sampling units. ***Simple random sampling*** is where you choose a sample of n using some random process, and all possible samples of size n are equally likely to be selected.

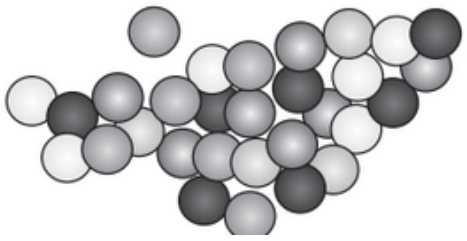
With simple random sampling, you have two options. You can either sample **with replacement** or **without replacement**.

Sampling with replacement

Sampling with replacement means that when you've selected each unit and recorded relevant information about it, you put it back into the population. By doing this, there's a chance that a sampling unit might be chosen more than once. You'd be sampling with replacement if you decided to question people on the street at random without checking if you had already questioned them before. If you stop a person for questioning and then let them go once you've finished asking them questions, you are in effect releasing them back into the population. It means that you may question them more than once.

Sampling without replacement

Sampling without replacement means that the sampling unit isn't replaced back into the population. An example of this is the gumball taste test; you wouldn't want to put gumballs that have been tasted back into the population.



You wouldn't want to replace gumballs once you've tasted them, so this would be simple random samping without replacement.

How to choose a simple random sample

There are two main ways of using simple random sampling: by drawing lots or using random numbers.

Drawing lots

Drawing lots is just like pulling names out of a hat. You write the name or number of each member of the sampling frame on a piece of paper or ball, and then place them all into a container. You then draw out n names or numbers at random so that you have enough for your sample.

Random number generators

If you have a large sampling frame, drawing lots might not be practical, so an alternative is to use a random number generator, or random number tables. For this, you give each member of the sampling frame a number, generate a set of n random numbers, and then pick the members of the set whose assigned numbers correspond to the random numbers that were generated.

It's important to make sure that each number has an equal chance of occuring so that there's no bias.

There are other types of sampling

Even simple random sampling has its problems.

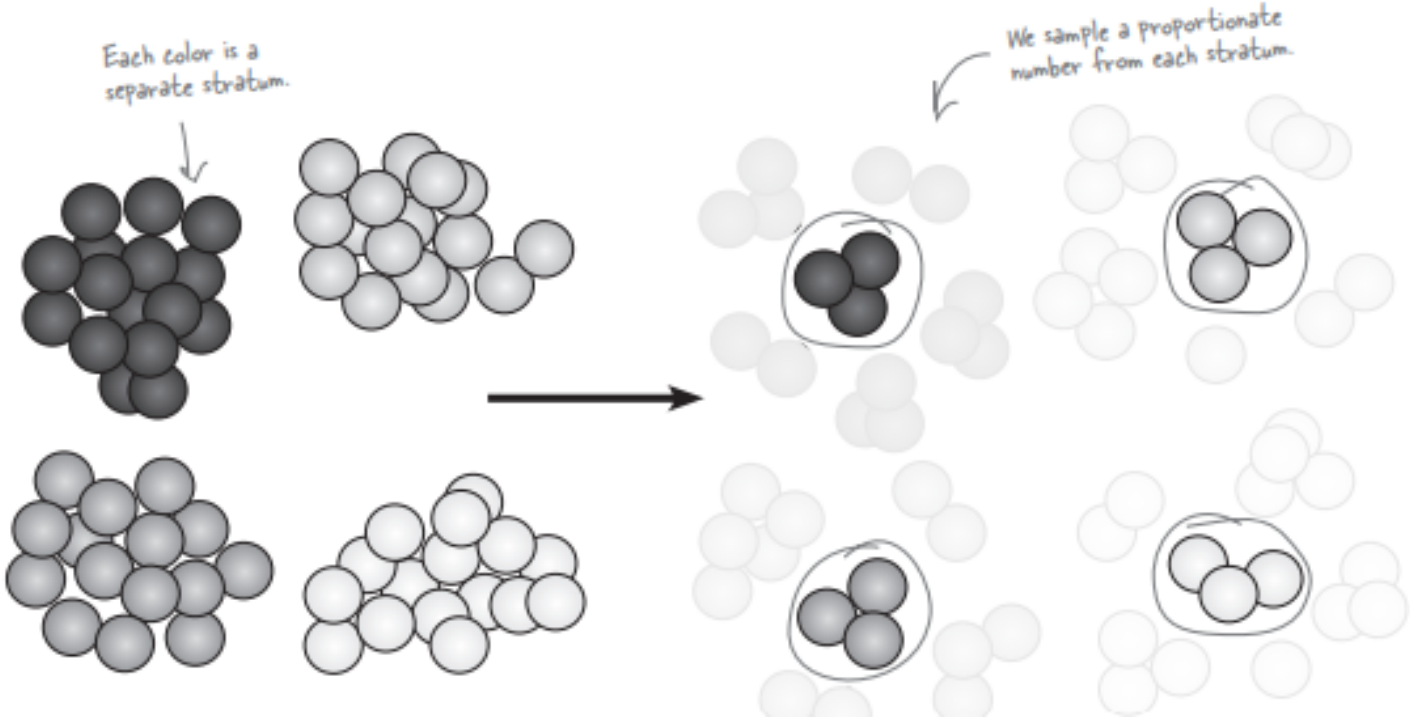
With simple random sampling, there's still a chance that your sample will not represent the target population. For example, you might end up randomly drawing only yellow gumballs for your sample, and the other colors would be left out.

So how can we avoid this?

We can use stratified sampling...

An alternative to simple random sampling is ***stratified sampling***. With this type of sampling, the population is split into similar groups that share similar characteristics. These charateristics or groups are called ***strata***, and each individual group is called a ***stratum***. As an example, we could split up the gumballs into the different colors, yellow, green, red, and pink, so that each color forms a different stratum.

Once you've done this, you can perform simple random sampling on each stratum to ensure that each group is represented in your overall sample. To do this, look at the proportions of each stratum within the overall population and take a proportionate number from each. As an example, if 50% of the gumballs that Mighty Gumball produce are red, half of your sample should consist of red gumballs.



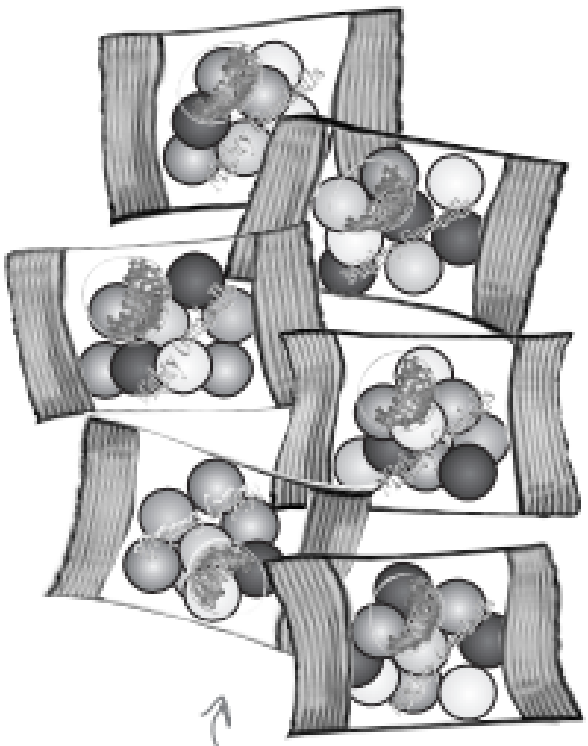
...or we can use cluster sampling...

Cluster sampling is useful if the population has a number of similar groups or clusters. As an example, gumballs might be sold in packets, with each packet containing a similar number of gumballs with similar colors. Each packet would form a ***cluster***.

With cluster sampling, instead of taking a simple random sample of units, you **draw a simple random sample of clusters**, and then survey everything within each of these clusters. As an example, you could take a simple random sample of *packets* of gumballs, and then taste all the gumballs in these packets.

Cluster sampling works because each cluster is similar to the others, and an added advantage is that you don't need a sampling frame of the whole population in order to achieve it. As an example, if you were surveying trees and used particular forests as your cluster, you would only need to know about each tree within only the forests you'd selected.

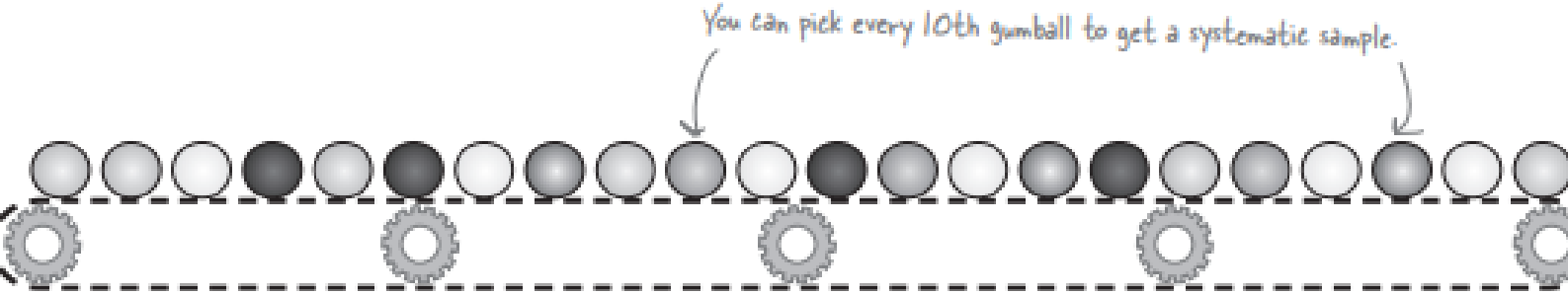
The problem with cluster sampling is that it might not be entirely random. As an example, it's likely that all of the gumballs in a packet will have been produced by the same factory. If there are differences between the factories, you may not pick these up.



...or even systematic sampling

With systematic sampling, you list the population in some sort of order, and then survey every kth item, where k is some number. As an example, you could choose to sample every 10th gumball.

Systematic sampling is relatively quick and easy, but there's one key disadvantage. If there's some sort of cyclic pattern in the population, your sample will be biased. As an example, if gumballs are produced such that every 10th gumball is red, you will end up only sampling red gumballs, and this could lead to you drawing misleading conclusions about your population.

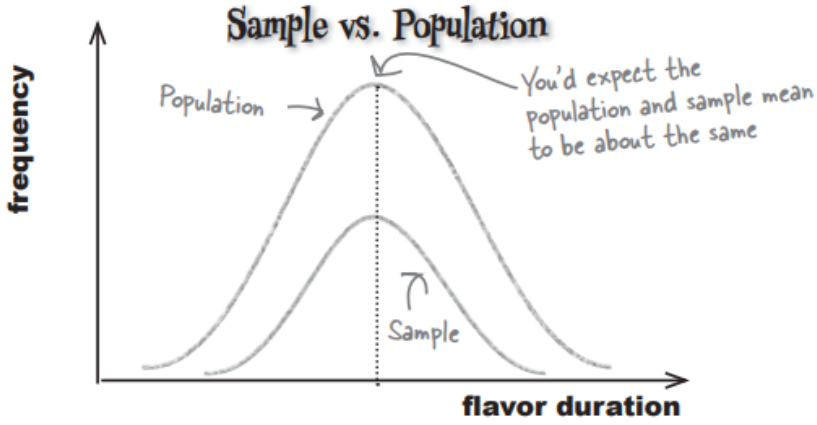


BULLET POINTS

- A **population** is the entire collection of things you are studying.
- A **sample** is a relatively small selection taken from the population that you can use to draw conclusions about the population itself.
- To take a sample, start off by defining your target population, the population you want to study. Then decide on your sampling units, the sorts of things you need to sample. Once you've done that, draw up a sampling frame, a list of all the sampling units in your target population.
- A sample is **biased** if it isn't representative of your target population.
- **Simple random sampling** is where you choose sampling units at random to form your sample. This can be with or without replacement. You can perform simple random sampling by drawing lots or using random number generators.
- **Stratified sampling** is where you divide the population into groups of similar units or strata. Each stratum is as different from the others as possible. Once you've done this, you perform simple random sampling within each stratum.
- **Cluster sampling** is where you divide the population into clusters where each cluster is as similar to the others as possible. You use simple random sampling to choose a selection of clusters. You then sample every unit in these clusters.
- **Systematic sampling** is where you choose a number, k, and sample every kth unit.

Estimating populations and samples

So how can we use the results of the sample taste test to tell us the mean amount of time gumball flavor lasts for in the general gumball population?
The answer is actually pretty intuitive. We assume that the mean flavor duration of the gumballs in the sample matches that of the population. In other words, we find the mean of the sample and use it as the mean for the population too.



Estimating the population mean

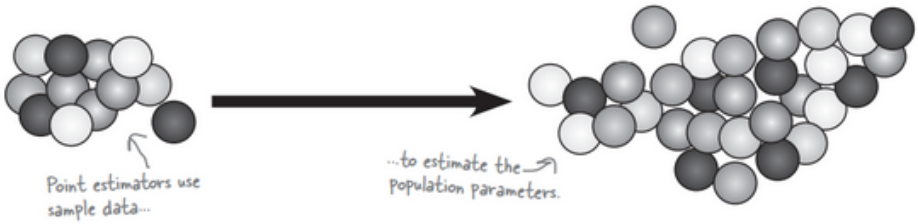
the mean of the sample is the best estimate we can make for the mean of the population. It's the most likely value for the population mean that we can come up with based on the information that we have. The mean of the sample is called a **point estimator** for the population mean. In other words, it's a calculation based on the sample data that provides a good estimate for the mean of the population.

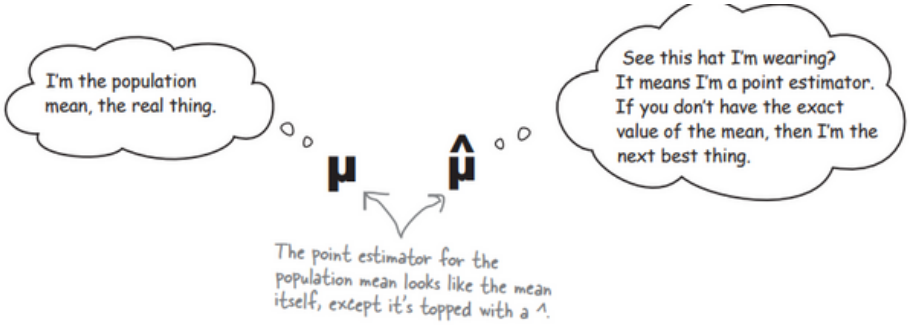
Point estimators can approximate population parameters

This time around we don't know the exact value of the population parameters. Instead of calculating them using the population, we estimate them using the sample data instead

A point estimator of a population parameter is some function or calculation that can be used to estimate the value of the population parameter. As an example, the point estimator of the population mean is the mean of the sample, as we can use the sample mean to estimate the population mean.

We differentiate between an actual population parameter and its point estimator using the ^ symbol. As an example, we use the symbol μ to represent the population mean, and to represent its estimator. So to show that you're dealing with the point estimator of a particular population parameter, take the symbol of the population parameter, and top it with a ^.





$\bar{x}$ is the sample equivalent of μ, and you calculate it in the same way you would the population mean. You add together all the data in your sample, and then divide by however many items there are. In other words, if your sample size is n,

$\bar{x}$ is the mean of the sample.

$$\bar{x} = \frac{\Sigma x}{n}$$

Add together the numbers in the sample, and divide by how many there are.

We can use this to write a shorthand expression for the point estimator for the population. Since we can estimate the population mean using the mean of the sample, th

The variance of the data in the sample may not be the best way of estimating the population variance.

You already know that the variance of a set of data measures the way in which values are dispersed from the mean. When you choose a sample, you have a smaller number of values than with the population, and since you have fewer values, there's a good chance they're more clustered around the mean than they would be in the population. More extreme values are less likely to be in your sample, as there are generally fewer of them.

imple.

We estimate the mea of the population...

BULLET POIN

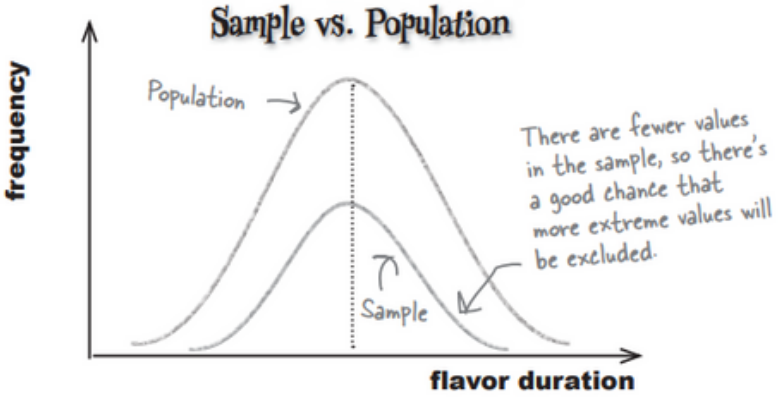
- A **point estimator** is an population parameter, de
- The ^ symbol is added t when you're talking abou example, the point estim
- The mean of a sample is mean of the sample, use

ean is found by

nate for the true se the mean of

$$\bar{x} = \frac{\Sigma x}{n}$$

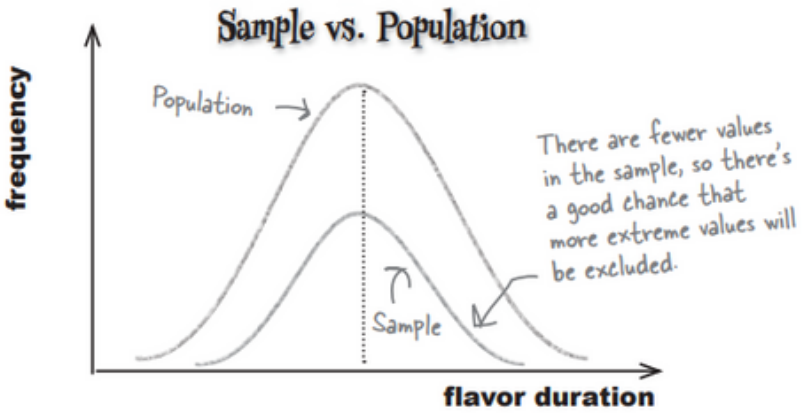
where *x* represents the values in the sample, and *n* is the sample size.



Estimating the population variance

Variance of a set of data measures the way in which values are dispersed from the mean. When you choose a sample, you have a smaller number of values than with the population, and since you have fewer values, there's a good chance they're more clustered around the mean than they would be in the population. More extreme values are less likely to be in your sample, as there are generally fewer of them.

The variance of the data in the sample may not be the best way of estimating the population variance.



We need a different point estimator than sample variance

The sample variance tends to be slightly less than the variance of the population, and the degree to which this holds depends on the number of values in the sample. If the number in the sample is small, there's likely to be a bigger difference between the sample and population variances than if the size of the sample is large

So what is the estimator?

Rather than take the variance of all the data in the sample to estimate the population variance, there's something else we can use instead. If the size of the sample is n, we can estimate the population variance using

This formula is a closer match to the value of the population variance.The population variance tends to be higher than the variance of the data in the sample. This means that this formula is a slightly better point estimator for the population variance.

Point estimator for the population variance, based on your sample.

$$\hat{\sigma}^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

$\bar{x}$: Sample mean

$n - 1$, not n, where n is the size of the sample. This time it's an estimate..

Population variance

Population variance $\sigma^2 = \frac{\Sigma(x - \mu)^2}{n}$ (μ: Population mean; n: Size of the population)

If you want to find the exact variance of a population and you have data for the whole population, use

Point estimator for the population variance $\hat{\sigma}^2 = s^2$

where

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

s^2 gives the formula based on the sample data

Divide by n – 1 for a sample because most of the time, you use your sample data to estimate the variance of the population. Dividing by n – 1 gives you a slightly more accurate result than dividing by n. This is because the variance of values in the sample is likely to be slightly lower than the population variance.

Predicting population proportion

If we use X to represent the number of successes in the population, then X follows a binomial distribution with parameters n and p. n is the number of people in the population, and p is the proportion of successes.

In the same way that our best estimate of the population mean is the mean of the sample, our best guess for the proportion of successes in the population has to be the proportion of successes in the sample

In other words, we can use the proportion of successes in the sample as a point estimator for the proportion of successes in the population

Point estimator for the proportion of successes in the population → $\hat{p} = p_s$ ← Proportion of successes in the sample

where

$$p_s = \frac{\text{number of successes}}{\text{number in sample}}$$

Probability and proportion are related

$$p = \text{probability} = \text{proportion}$$

Imagine you have a population for which you want to find the proportion of successes. To calculate this proportion, you take the number of successes, and divide by the size of the population. Now suppose you want to calculate the probability of choosing a success from the population at random. To derive this probability, you take the number of successes in the population, and divide by the size of the population. In other words, you derive the probability of getting a success in exactly the same way as you derive the proportion of successes.

BULLET POINTS

- The point estimator for the population variance is given by

$$\hat{\sigma}^2 = s^2$$

where s^2 is given by

$$\frac{\Sigma(x - \bar{x})^2}{n - 1}$$

- The population proportion is represented using p. It's the proportion of successes within the population.

- The point estimator for p is given by p_s, where p_s is the proportion of successes in the sample.

$$\hat{p} = p_s$$

- You calculate p_s by dividing the number of successes in the sample by the size of the sample.

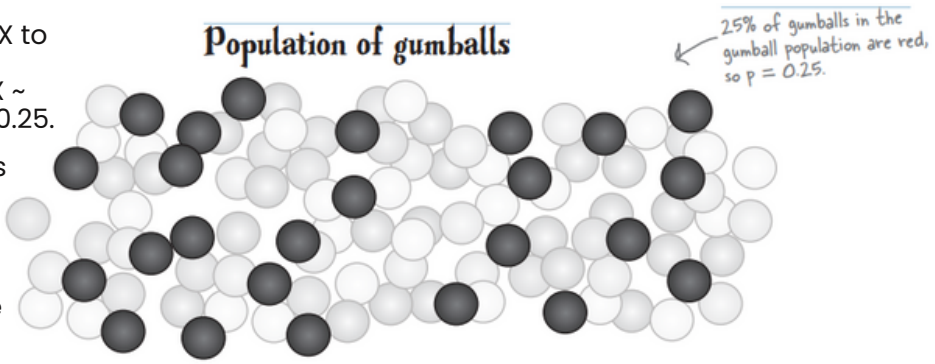
$$p_s = \frac{\text{number of successes}}{\text{number in sample}}$$

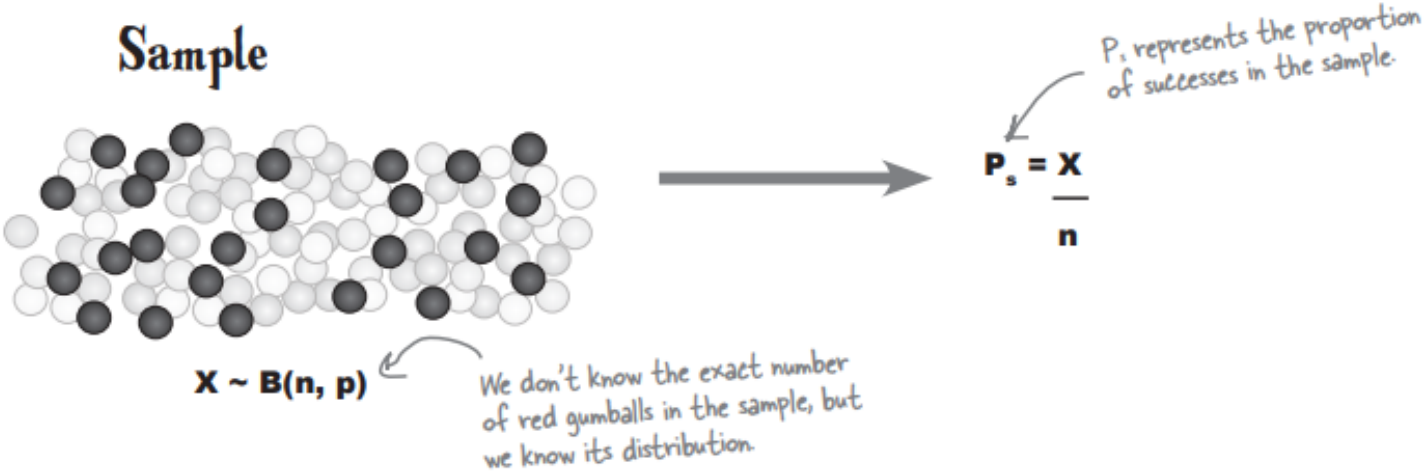
Sampling Distribution of Proportions:

- Imagine you have a big box of gumballs, and you want to know the proportion of red gumballs in it.
- Each time you take a smaller sample (let's say 100 gumballs), you might get a slightly different proportion of red gumballs.
- The sampling distribution of proportions tells us how these proportions vary across different samples of the same size.

If we use the random variable X to represent the number of red gumballs in the sample, then X ~ B(n, p), where n = 100 and p = 0.25.

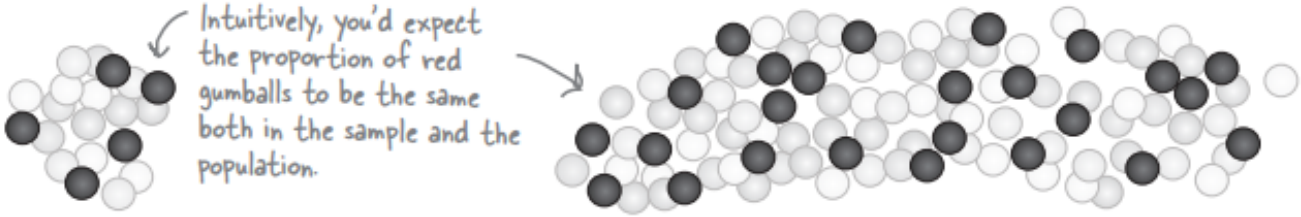
The proportion of red gumballs in the sample depends on X, the number of red gumballs in the sample. This means that the proportion itself is a random variable. We can write this as Ps , where Ps = X/n





Expectation of Sample Proportion:

- We expect the proportion of red gumballs in a sample to match the proportion in the whole box.
- So, if 25% of all gumballs are red (p = 0.25), we'd expect about 25% of the gumballs in a sample to be red.



We want to find $E(P_s)$, where $P_s = X/n$. In other words, we want to find the expected value of the sample proportion, where the sample proportion is equal to the number of red gumballs divided by the total number of gumballs in the sample. This gives us

$$E(P_s) = E\left(\frac{X}{n}\right)$$

$$= \frac{E(X)}{n}$$

Now X is the number of red gumballs in the sample. If we count the number of red gumballs as the number of successes, then $X \sim B(n, p)$.

You've already seen that for a binomial distribution, $E(X) = np$. This means that

$$E(P_s) = \frac{E(X)}{n}$$

$$= \frac{np}{n} \leftarrow E(X) = np$$

$$= p$$

This result ties in with what we intuitively expect. We can expect the proportion of successes in the sample to match the proportion of successes in the population.

Variance of Sample Proportion:

- The variance tells us how much the sample proportions vary from the population proportion.
- If we have more red gumballs (successes) or more non-red gumballs (failures) in our sample, it affects the variance.

$$\text{Var}(P_s) = \text{Var}\left(\frac{X}{n}\right)$$

$$= \frac{\text{Var}(X)}{n^2}$$

This comes from $\text{Var}(aX) = a^2\text{Var}(X)$. In this case, $a = 1/n$.

As we've said before, X is the number of number of red gumballs in the sample. If we count the number of red gumballs as the number of successes, then $X \sim B(n, p)$. This means that $Var(X) = npq$, as this is the variance for the binomial distribution. This gives us

$$Var(P_s) = \frac{Var(X)}{n^2}$$

$$= \frac{npq}{n^2} \quad \leftarrow Var(X) = npq$$

$$= \frac{pq}{n}$$

Taking the square root of the variance gives us the standard deviation of P_s, and this tells us how far away from p the sample proportion is likely to be. It's sometimes called the ***standard error of proportion***, as it tells you what the error for the proportion is likely to be in the sample.

$$\textbf{Standard error of proportion} = \sqrt{\frac{pq}{n}}$$

Distribution of Sample Proportions:

P_s follows a normal distribution

When n is large, the distribution of P_s becomes approximately normal. By large, we mean greater than 30. The larger *n* gets, the closer the distribution of P_s gets to the normal distribution.

We've already found the expectation and variance of P_s, so this means that if n is large,

$$P_s \sim N\left(p, \frac{pq}{n}\right)$$

As P_s follows a normal distribution for $n > 30$, this means that we can use the normal distribution to solve our gumball problem. We can use the normal distribution to calculate the probability that the proportion of red gumballs in a jumbo box of gumballs will be at least 40%.

There's just one thing to remember: the sampling distribution needs a continuity correction.

Watch it!

Sometimes statisticians disagree about how large n needs to be.

If you're taking a statistics exam, make sure that you check how your exam board defines this.

This means we can use the properties of the normal distribution to make predictions about sample proportions.

Ps —continuity correction required

Since the number of successes in each sample is discrete (you can't have a fraction of a gumball), we need to adjust for this when using the normal distribution.
This adjustment is called a continuity correction and ensures better accuracy when approximating probabilities using the normal distribution.
It's typically ±(1/2n), where n is the sample size.

In other words, if you use the normal distribution to approximate probabilities for Ps , make sure you apply a continuity correction of ±1/2n. The exact continuity correction depends on the value of n.

If n is very large, the continuity correction can be left out.

As *n* gets larger, the continuity correction becomes very small, and this means that it makes very little difference to the overall probability. Some textbooks omit the continuity correction altogether.

BULLET POINTS

- The **sampling distribution of proportions** is what you get if you consider all possible samples of size *n* taken from the same population and form a distribution out of their proportions. We use P_s to represent the sample proportion random variable.
- The expectation and variance of P_s are defined as

$$E(P_s) = p$$

$$Var(P_s) = pq/n$$

where *p* is the population proportion.

- The **standard error of proportion** is the standard deviation of this distribution. It's given by

$$\sqrt{Var(P_s)}$$

- If $n > 30$, then P_s follows a normal distribution, so

$$P_s \sim N(p, pq/n)$$

for large *n*. When working with this, you need to apply a continuity correction of

25% of the gumball population are red. What's the probability that in a box of 100 gumballs, at least 40% will be red? We'll guide you through the steps.

1. If P_s is the proportion of red gumballs in the box, how is P_s distributed?

Let's use p to represent the probability that a gumball is red. In other words, $p = 0.25$.

Let's use P_s to represent the proportion of gumballs in the box that are red.

$P_s \sim N(p, pq/n)$, where $p = 0.25$, $q = 0.75$, and $n = 100$. As pq/n is equal to $0.25 \times 0.75 / 100 = 0.001875$, this gives us

$P_s \sim N(0.25, 0.001875)$

2. What's the value of $P(P_s \geq 0.4)$? Hint: Remember that you need to apply a continuity correction.

$$P(P_s \geq 0.4) = P(P_s > 0.4 - 1/(2 \times 100))$$
$$= P(P_s > 0.395)$$

As $P_s \sim N(0.25, 0.001875)$, we need to find the standard score of 0.395 so that we can look up the result in probability tables. This gives us

$$z = \frac{0.395 - 0.25}{\sqrt{0.001875}}$$
$$= 3.35$$

$$P(Z > z) = 1 - P(Z < 3.35)$$
$$= 1 - 0.9996$$
$$= 0.0004$$

In other words, the probability that in a box of 100 gumballs, at least 40% will be red, is 0.0004.

Sampling Distribution of Proportions Up Close

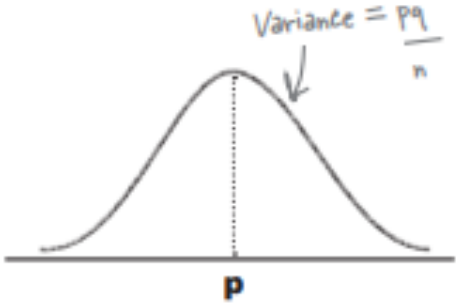
The sampling distribution of proportions is the distribution formed by taking the proportions of all possible samples of size n. The proportion of successes in a sample is represented by P_s, and it is is distributed as

$$E(P_s) = p$$

$$Var(P_s) = \frac{pq}{n}$$

When n is large, say bigger than 30, the distribution of P_s becomes approximately normal, so

$$P_s \sim N\left(p, \frac{pq}{n}\right)$$



Knowing the probability distribution of P_s is useful because it means that given a particular population, we can calculate probabilities for the proportion of successes in the sample. We can approximate this with the normal distribution, and the larger the size of the sample, the more accurate the approximation.

The sampling distribution continuity correction

When you use the normal distribution in this way, it's important to apply a ***continuity correction***. This is because the number of successes in the sample is discrete, and it's used in the calculation of proportion.

If X represents the number of successes in the sample, then $P_s = X/n$. The continuity correction for X is $\pm(1/2)$, so this means the continuity correction is given by

$$\text{Continuity correction} = \frac{\pm 1}{2n}$$

In other words, if you use the normal distribution to approximate probabilities for the sampling proportion, make sure you apply a continuity correction of $\pm 1/2n$.

- The **sampling distribution of means** is what you get if you consider all possible samples of size n taken from the same population and form a distribution out of their means. We use $\overline{X}$ to represent the sample mean random variable.
- The **expectation and variance of $\overline{X}$** are defined as

$$E(\overline{X}) = \mu$$

$$Var(\overline{X}) = \sigma^2/n$$

- The **standard error of the mean** is the standard deviation of this distribution. It's given by

$$\sqrt{Var(\overline{X})}$$

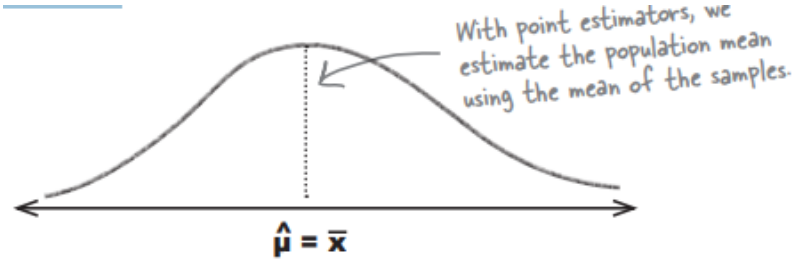
- If $X \sim N(\mu, \sigma^2)$, then $\overline{X} \sim N(\mu, \sigma^2/n)$.
- The **central limit theorem** says that if n is large and X doesn't follow a normal distribution, then

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

Constructing Confidence Intervals

Sometimes samples don't give quite the right result. You've seen how you can use point estimators to estimate the precise value of the population mean, variance, or proportion, but the trouble is, how can you be certain that your estimate is completely accurate? After all, your assumptions about the population rely on just one sample, and what if your sample's off? In this chapter, you'll see another way of estimating population statistics, one that allows for uncertainty

Using the point estimator, we've been able to give a very precise estimate for the mean duration of the flavor. Here's a sketch showing the distribution of flavor duration in the sample of gumballs.



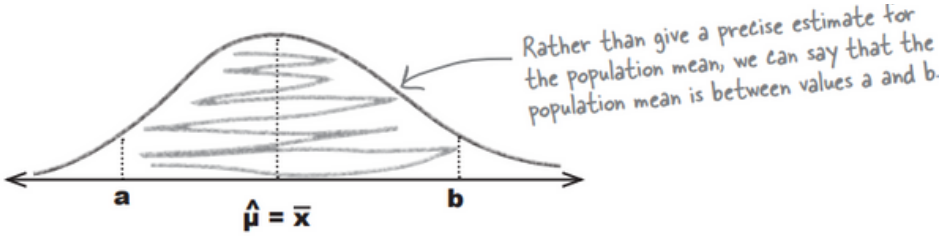
The interval limits are chosen so that there's a specified probability of the population mean being between a and b. As an example, you may want to choose a and b so that there's a 95% chance of the interval containing the population mean. In other words, you choose a and b so that

Point estimators are valuable, but they may give slight errors.

Because we're not dealing with the entire population, all we're doing is giving a best estimate. If the sample we use is unbiased, then the estimate is likely to be close to the true value of the population. The question is, how close is close enough? Rather than give a precise value as an estimate for the population mean, there's another approach we can take instead. We can specify some interval as an estimation of flavor duration rather than a very precise length of time. As an example, we could say that we expect gumball flavor to last for between 55 and 65 minutes. This still gives the impression that flavor lasts for approximately one hour, but it allows for some margin of error. The question is, how do we come up with the interval? It all depends how confident you want to be in the results...

The problem with deriving point estimators is that we rely on the results of a single sample to give us a very precise estimate. We've looked at ways of making the sample as representative as possible by making sure the sample is unbiased, but we don't know with absolute certainty that it's 100% representative, purely because we're dealing with a sample.

Rather than specify an exact value, we can specify two values we expect flavor duration to lie between. We place our point estimator for the mean in the center of the interval and set the interval limits to this value plus or minus some margin of error.



We represent this interval as (a, b). As the exact value of a and b depends on the degree of confidence you want to have that the interval contains the population mean, (a, b) is called a **confidence interval**

Four steps for finding confidence intervals

Here are the broad steps involved in finding confidence intervals. Don't worry if you don't get what each step is about just yet, we'll go through them in more detail soon.

You encountered sampling distributions in the last chapter.

1. **Choose your population statistic** ← This is the population statistic you want to construct a confidence interval for.
2. **Find its sampling distribution**
3. **Decide on the level of confidence** ← The probability that your interval contains the statistic
4. **Find the confidence limits** ← To find the confidence limits, we need to know the level of confidence and the sampling distribution.

1. Choose your Population Statistic:

Identify the statistic you want to construct a confidence interval for. This could be the population mean, proportion, variance, etc. In this case, we wanted to find a confidence interval for the mean duration of gumball flavor, so we aimed to construct a confidence interval for the population mean, denoted as μ.

2. Find its Sampling Distribution:

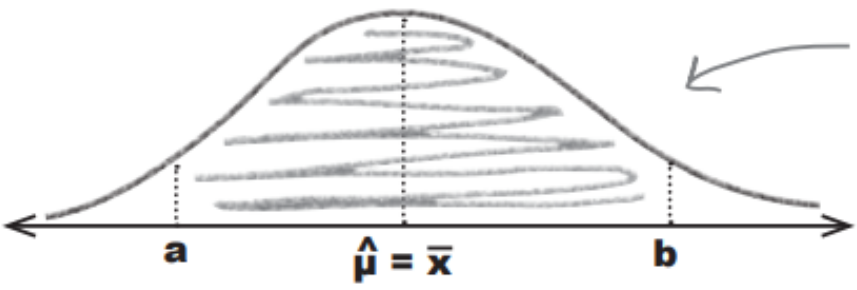
Determine the sampling distribution associated with the chosen population statistic. This involves understanding the expectation and variance of the sampling distribution. In our case, for the population mean, we found that the sampling distribution of means follows a normal distribution with an expectation of μ and a variance of σ^2 / n, where σ^2 is the population variance and n is the sample size.

$$E(\overline{X}) = \mu \qquad Var(\overline{X}) = \frac{s^2}{n}$$

← This is the point estimator for the variance. We don't know what the true value of the population variance is, so we use the sample variance to estimate it instead..

3. Decide on the Level of Confidence:

Determine the level of confidence you want for your interval. Common choices include 90%, 95%, or 99%. This represents the probability that the interval contains the population parameter. In this case, a 95% confidence level was chosen, indicating a 95% probability that the interval contains the true population mean.The level of confidence lets you say how sure you want to be that the confidence interval contains your population statistic. As an example, suppose we want a confidence level of 95% for the population mean. This means that the probability of the population mean being inside the confidence interval is 0.95.

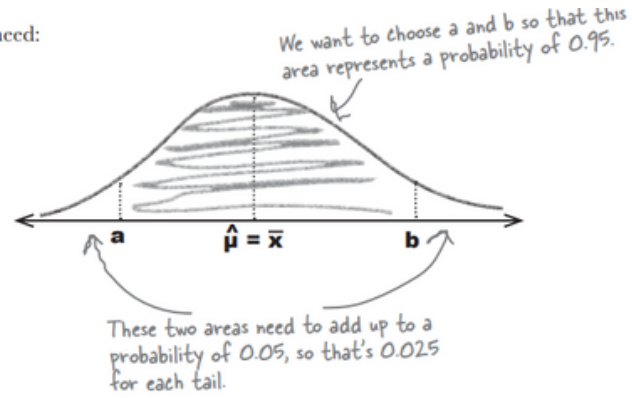


← The confidence level is the probability of the population mean being inside the confidence interval. For a confidence lev of 95%, the probability here is 0.95.

4. Find the Confidence Limits:

Calculate the confidence limits, which define the range within which the population parameter is expected to lie with the chosen level of confidence. This involves finding values a and b such that the probability that the parameter falls between them is equal to the chosen confidence level. In our case, we used a normal distribution to find the confidence limits, using the standard normal probabilities to determine the values of a and b.

Here's a sketch of what we need:



How to select an appropriate confidence level

So who decides what the level of confidence should be? What's the right level of confidence?

The answer to this really depends on your situation and how confident you need to be that your interval contains the population statistic. A 95% confidence level is common, but sometimes you might want a different one, such as 90% or 99%. As an example, the Mighty Gumball CEO might want to have a higher degree of confidence that the population mean falls inside the confidence interval, as he intends to use it in his television advertisements.

The key thing to remember is that the higher the confidence level is, the wider the interval becomes, and the more chance there is of the confidence interval containing the population statistic.

As X follows a normal distribution, we can use the normal distribution to find the confidence interval.

Let's summarize the steps

Let's look back at the steps we went through in order to construct the confidence interval.

The first thing we did was **choose the population statistic** that we needed to construct a confidence interval for. We needed to find a confidence interval for the mean duration of gumball flavor, and this meant that we needed to construct a confidence interval for μ.

Once we'd figured out which population we needed to construct a confidence interval for, we had to **find its sampling distribution**. We found the expectation and variance of the sampling distribution of means, substituting in values for every statistic except for μ. We then figured out that we could use a normal distribution for $\overline{X}$.

After that, we decided on the **level of confidence** we needed for the confidence interval. We decided to use a confidence level of 95%.

Finally, we had to **find the confidence limits** for the confidence interval. We used the level of confidence and sampling distribution to come up with a suitable confidence interval.

Handy shortcuts for confidence intervals

Here are some of the shortcuts you can take when you calculate confidence intervals. All you need to do is look at the population statistic you want to find, look at the distribution of the population and the conditions, and then slot in the population statistic or its estimator. The value c depends on the level of confidence

Population statistic	Population distribution	Conditions	Confidence interval
μ	Normal	You know what σ^2 is n is large or small $\bar{x}$ is the sample mean	$\left(\bar{x} - c\frac{\sigma}{\sqrt{n}}, \bar{x} + c\frac{\sigma}{\sqrt{n}}\right)$
μ	Non-normal	You know what σ^2 is n is large (at least 30) $\bar{x}$ is the sample mean	$\left(\bar{x} - c\frac{\sigma}{\sqrt{n}}, \bar{x} + c\frac{\sigma}{\sqrt{n}}\right)$
μ	Normal or non-normal	You don't know what σ^2 is n is large (at least 30) $\bar{x}$ is the sample mean s^2 is the sample variance	$\left(\bar{x} - c\frac{s}{\sqrt{n}}, \bar{x} + c\frac{s}{\sqrt{n}}\right)$
p	Binomial	n is large p_s is the sample proportion q_s is $1 - p_s$	$\left(p_s - c\sqrt{\frac{p_s q_s}{n}}, p_s + c\sqrt{\frac{p_s q_s}{n}}\right)$

The value of c depends on the level of confidence you need. These values work whenever you use the normal distribution as the basis of your test statistic.

Level of confidence	Value of c
90%	1.64
95%	1.96
99%	2.58

What's the interval in general?

In general, the confidence interval is given by

statistic ± (margin of error)

The margin of error is given by the value of c multiplied by the standard deviation of the test statistic.

margin of error = c × (standard deviation of statistic)

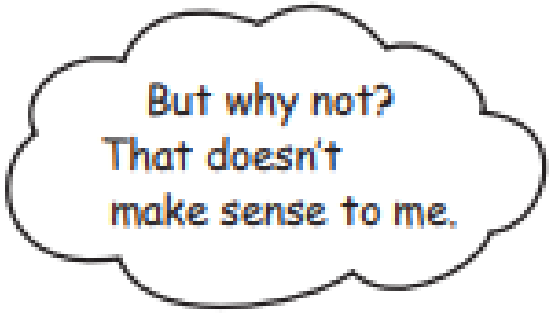
The normal distribution isn't a good approximation for every situation.

All of the sampling distributions we've seen so far either follow a normal distribution or can be approximated by it. The trouble is that we can't use the normal distribution for every single confidence interval. Unfortunately, this situation is one of them.

So why can't we use the normal distribution here?

When sample sizes are large, the normal distribution is ideal for finding confidence intervals. It gives accurate results, irrespective of how the population itself is distributed.

Here we have a different situation. Even though X itself is distributed normally, $\overline{X}$ isn't.



There are two key reasons.

The first is that we don't know what the true variance is of the population, so this means we have to estimate σ^2 using the sample data. We can easily do this using point estimators, but there's a problem: the size of the sample is so small that there are likely to be significant errors in our estimate, much larger errors than if we used a larger sample of gumballs. The potential errors we're dealing with mean that the normal distribution won't give us accurate enough probabilities for $\overline{X}$, which means it won't give us an accurate confidence interval.

So what sort of distribution does $\overline{X}$ follow? It actually follows a ***t-distribution***. Let's find out more.

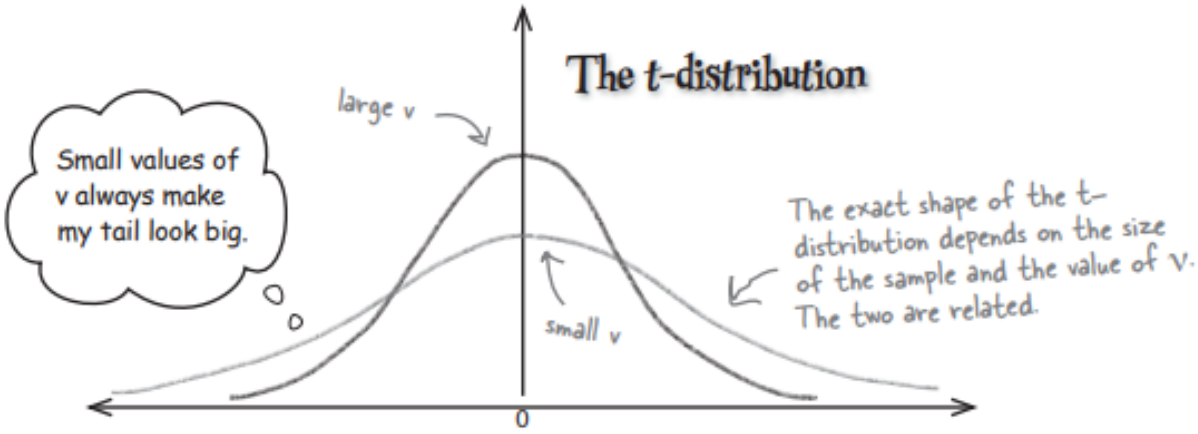
$\overline{X}$ follows the t-distribution when the sample is small

The t-distribution is a probability distribution that specializes in exactly the sort of situation we have here. It's the distribution that $\overline{X}$ follows where the population is normal, σ^2 is unknown, and you only have a small sample at your disposal.

The t-distribution looks like a smooth, symmetrical curve, and it's exact shape depends on the size of the sample. When the sample size is large, it looks like the normal distribution, but when the sample size is small, the curve is flatter and has slightly fatter tails. It takes one parameter, ν, where ν is equal to $n - 1$. n is the size of the sample, and ν is called the ***number of degrees of freedom***.

We'll look at degrees of freedom in more depth in Chapter 14.

Let's take a look at this. Here's a sketch of the t-distribution for different values of ν. Can you see how the value of ν affects the shape of the distribution?



A shorthand way of saying that T follows the t-distribution with ν degrees of freedom is

T is the test statistic. You'll see how to calculate it on the next page

$$T \sim t(\nu)$$

t(ν) means we're using the t-distribution with ν degrees of freedom. $\nu = n - 1$.

The t-distribution works in a similar way to the normal distribution. We start off by converting the limit of the probability area into a standard score, and then we use probability tables to get the result we want.

Let's start with the standard score.

Find the standard score for the t-distribution

We calculate the standard score for the t-distribution in the same way we did for the normal distribution. As with the the normal distribution, we standardize by subtracting the expectation of the sampling distribution and then dividing by its standard deviation. The only difference is that we represent the result with T instead of Z, as we're going to use it with the t-distribution.

We need to find the distribution of $\overline{X}$, so this means we need to use the expectation and standard deviation of $\overline{X}$. The expectation of $\overline{X}$ is μ, and the standard deviation is σ/n. As we need to estimate the value of σ with s, this means that the standard score for the t-distribution is given by

This is the same formula as for Z—subtract the mean and divide by the standard deviation.

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

This is the population mean we're finding a confidence interval for.

This is the standard deviation of $\overline{X}$.

All we need to do is substitute in the values for $\overline{X}$, $\hat{\sigma}$, and n.

Step 4: Find the confidence limits

You find confidence limits with the t-distribution in a similar way to how you find them with the normal distribution. Your confidence interval is given by

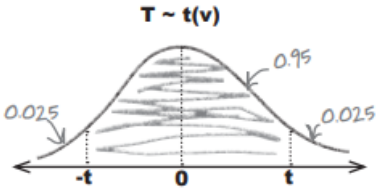
$$\left(\overline{x} - t\,\frac{s}{\sqrt{n}}, \overline{x} + t\,\frac{s}{\sqrt{n}}\right)$$

This is the same as we had before, just replace c with t.

where

$$P(-t \le T \le t) = 0.95$$

This is 0.95, as we want to find the 95% confidence interval.



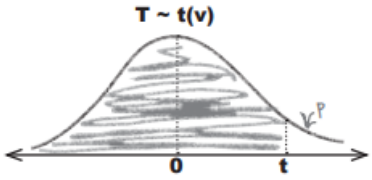
We can find the value of t using t-distribution probability tables.

Using t-distribution probability tables

t-distribution probability tables give you the value of t where $P(T > t) = p$. In our case, $p = 0.025$.

To find t, use the first column to look up ν, and the top row to look up p. The place where they intersect gives the value of t. As an example, if we look up $\nu = 7$ and $p = 0.05$, we get $t = 1.895$.

Once you've found the value of t, you can use it to find your confidence interval.



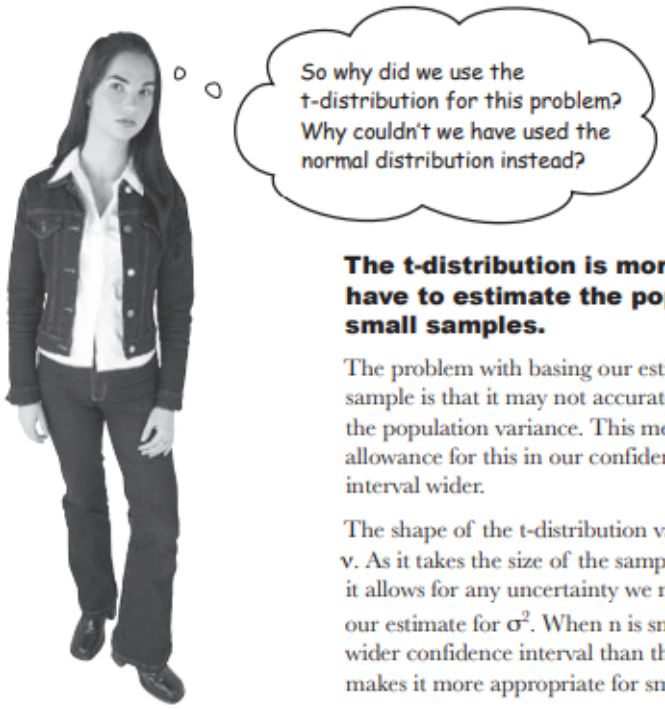
p = 0.05

	Tail probability p											
ν	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587

ν = 7

This is where 7 and .05 meet.

The t-distribution vs. the normal distribution



The t-distribution is more accurate when we have to estimate the population variance for small samples.

The problem with basing our estimate of σ^2 on just a small sample is that it may not accurately reflect the true value of the population variance. This means we need to make some allowance for this in our confidence interval by making the interval wider.

The shape of the t-distribution varies in line with the value of ν. As it takes the size of the sample into account, this means that it allows for any uncertainty we may feel about the accuracy of our estimate for σ^2. When n is small, the t-distribution gives a wider confidence interval than the normal distribution, which makes it more appropriate for small-sized samples.

Handy shortcuts for confidence intervals - the t-distribution

Here's a quick reminder of when you need to use the t-distribution, and what the confidence interval is for μ. Just substitute in your values.

Population statistic	Population distribution	Conditions	Confidence interval
μ	Normal or non-normal	You don't know what σ^2 is n is small (less than 30) $\bar{x}$ is the sample mean s^2 is the sample variance	$\left(\bar{x} - t(\nu)\frac{s}{\sqrt{n}}, \bar{x} + t(\nu)\frac{s}{\sqrt{n}}\right)$

To find $t(\nu)$, you need to look it up in t-distribution probability tables. To do this, use $\nu = n - 1$ and your level of confidence to find the critical region.

Hypothesis Testing

Not everything you're told is absolutely certain. The trouble is, how do you know when what you're being told isn't right? Hypothesis tests give you a way of using samples to test whether or not statistical claims are likely to be true. They give you a way of weighing the evidence and testing whether extreme results can be explained by mere coincidence, or whether there are darker forces at work. Come with us on a ride through this chapter, and we'll show you how you can use hypothesis tests to confirm or allay your deepest suspicions.

Case Study

Statsville's leading drug company has introduced a new remedy for snoring, claiming a 90% success rate within two weeks. However, a doctor at Statsville Surgery conducted her own trial with 15 patients and found only 11 were cured. This raises doubts about the drug's efficacy. The discrepancy could be due to chance, biased sampling, or flawed testing by the drug company.

Examine the claim

Take the claim of the drug company.

Examine the evidence

See how much evidence we need to reject the drug company's claim, and check this against the evidence we have. We do this by looking at how rare the doctors results would be if the drug company is correct.

Make a decision

Depending on the evidence, accept or reject the claims of the drug company.

In general, this process is called **hypothesis testing**, as you take a hypothesis or claim and then test it against the evidence.

This is the claim that we're putting on trial

1. **Decide on the hypothesis you're going to test**
2. **Choose your test statistic** — We need to pick the statistic that best tests the claim.

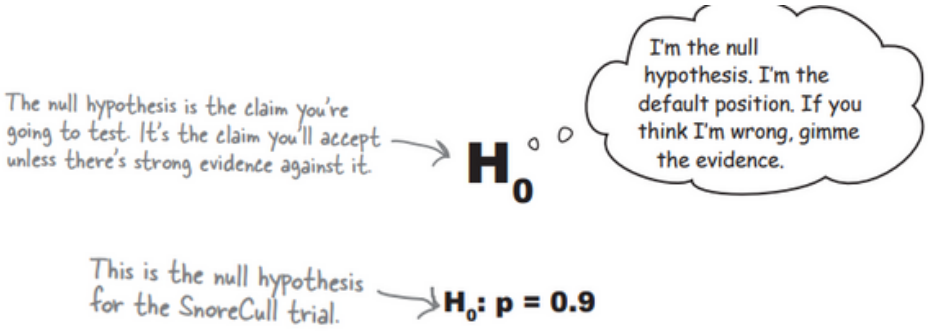
We need a certain level of certainty.

3. **Determine the critical region for your decision**
4. **Find the p-value of the test statistic** — We need to see how rare our results are, assuming the claims are true.
5. **See whether the sample result is within the critical region** — We then see if it's within our bounds of certainty.
6. **Make your decision**

Step 1: Decide on the hypothesis

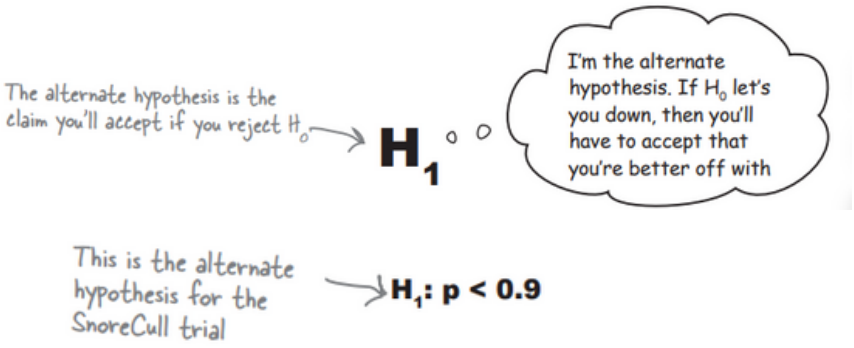
The claim that we're testing is called the **null hypothesis**. It's represented by H0 , and it's the claim that we'll accept unless there is strong evidence against it.

The null hypothesis for SnoreCull is the claim of the drug company: that it cures 90% of patients. This is the claim that we're going to go along with, unless we find strong evidence against it. We need to test whether at least 90% of patients are cured by the drug, so this means that the null hypothesis is that p = 90%.



The counterclaim to the null hypothesis is called the **alternate hypothesis**. It's represented by H1 , and it's the claim that we'll accept if there's strong enough evidence to reject H0 .

The alternate hypothesis for SnoreCull is the claim you'll accept if the drug company's claim turns out to be false. If there's sufficiently strong evidence against the drug company, then it's likely that the doctor is right. The doctor believes that SnoreCull cures less than 90% of people, so this means that the alternate hypothesis is that p < 90%



When hypothesis testing, you assume the null hypothesis is true. If there's sufficient evidence against it, you reject it and accept the alternate hypothesis.

Step 2: Choose your test statistic

The **test statistic** is the statistic that you use to test your hypothesis. It's the statistic that's most relevant to the test

In our hypothesis test, we want to test whether SnoreCull cures 90% of people or more. To test this, we can look at the probability distribution according to the drug company, and see whether the number of successes in the sample is significant. If we use X to represent the number of people cured in the sample, this means that we can use X as our test statistic. There are 15 people in the sample, and the probability of success according to the drug company is 0.9. As X follows a binomial distribution, this means that the test statistic is actually:

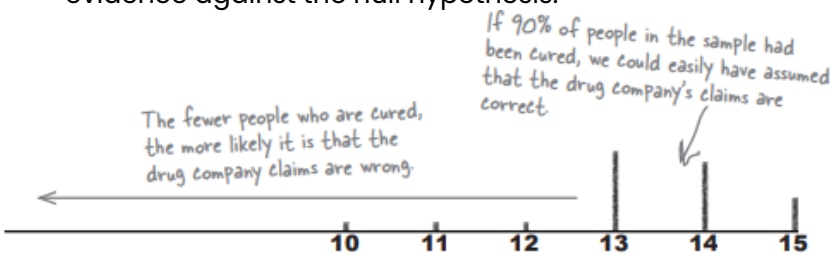
$X \sim B(15, 0.9)$ ← This is the test statistic for our hypothesis test.

We choose the test statistic according to H_0, the null hypothesis.

We need to test whether there is sufficient evidence against the null hypothesis, and we do this by first assuming that H0 is true. We then look for evidence that contradicts H0 . For the SnoreCull hypothesis test, we assume that the probability of success is 0.9 unless there is strong evidence against this being true.We do this by finding a critical region.

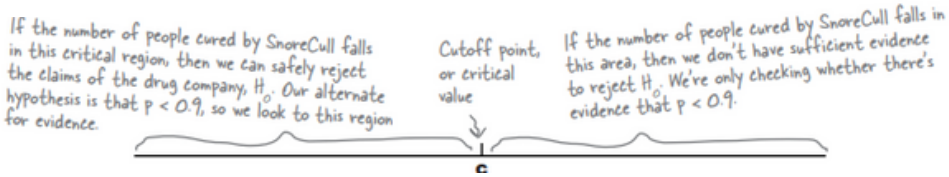
Step 3: Determine the critical region

The **critical region** of a hypothesis test is the set of values that present the most extreme evidence against the null hypothesis.



If the test statistic falls within the critical region, then the researcher can reject the null hypothesis. This means that there is statistically significant evidence to suggest that the null hypothesis is false.

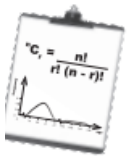
If the number of snorers cured falls outside the critical region, then we'll accept that there isn't sufficient evidence to reject the null hypothesis, and we'll accept the claims of the drug company. We'll call the cut off point for the critical region c, the critical value



To find the critical region, first decide on the significance level

The **significance level** of a test is a measure of how unlikely you want the results of the sample to be before you reject the null hypothesis Ho . Just like the confidence level for a confidence interval, the significance level is given as a percentage.

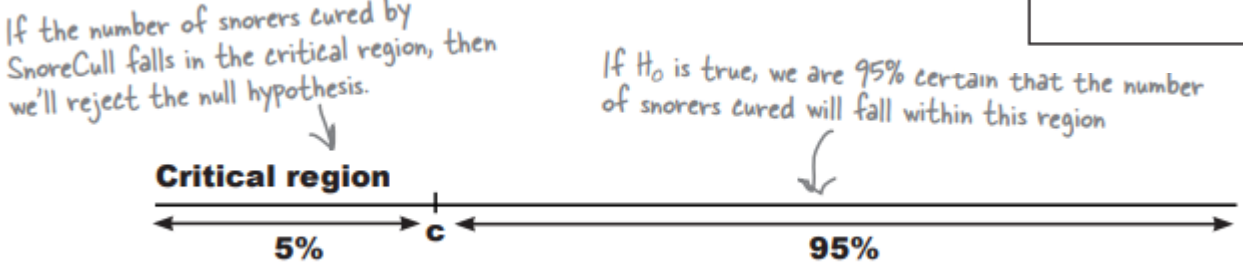
As an example, suppose we want to test the claims of the drug company at a 5% level of significance. This means that we choose the critical region so that the probability of fewer than c snorers being cured is less than 0.05. It's the lowest 5% of the probability distribution.



Vital Statistics

Significance level

The significance level is represented by α. It's a way of saying how unlikely you want your results to be before you'll reject H_0.



So what significance level should we use?

Let's use a significance level of 5% in our hypothesis test. This means that if the number of snorers cured in the sample is in the lowest 5% of the probability distribution, then we will reject the claims of the drug company. If the number of snorers cured lies in the top 95% of the probability distribution, then we'll decide there isn't enough evidence to reject the null hypothesis, and accept the claims of the drug company.

If we use X to represent the number of snorers cured, then we define the critical region as being values such that

$$P(X < c) < \alpha$$

where

$$\alpha = 5\%$$

Critical Regions Up Close

When you're constructing a critical region for your test, another thing you need to be aware of is whether you're conducting a **one-tailed** or **two-tailed** test. Let's look at the difference between the two, and what impact this has on the critical region?

One-tailed tests

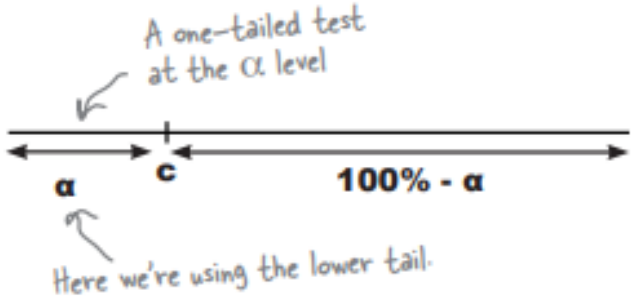
A ***one-tailed test*** is where the critical region falls at one end of the possible set of values in your test. You choose the level of the test—represented by α—and then make sure that the critical region reflects this as a corresponding probability.

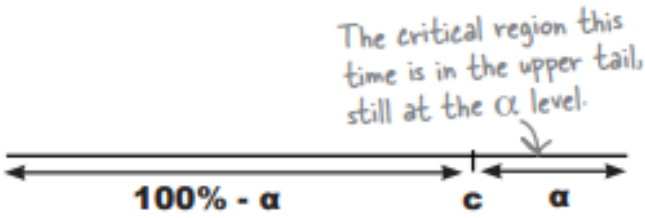
The tail can be at either end of the set of possible values, and the end you use depends on your alternate hypothesis H_1.

If your alternate hypothesis includes a < sign, then use the **lower tail**, where the critical region is at the lower end of the data.

If your alternate hypothesis includes a > sign, then use the **upper tail**, where the critical region is at the upper end of the data.

We're using a one-tailed test for the SnoreCull hypothesis test with the critical region in the lower tail, as our alternate hypothesis is that $p < 0.9$.



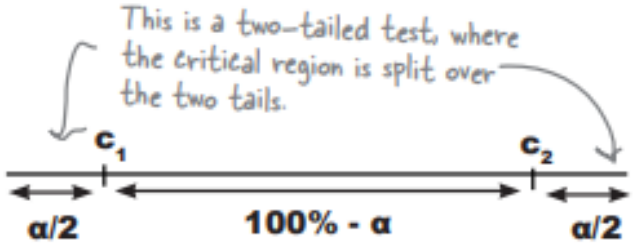


Two-tailed tests

A ***two-tailed test*** is where the critical region is split over both ends of the set of values. You choose the level of the test α, and then make sure that the overall critical region reflects this as a corresponding probability by splitting it into two. Both ends contain $\alpha/2$, so that the total is α.

You can tell if you need to use a two-tailed test by looking at the alternate hypothesis H_1. If H_1 contains a $\neq$ sign, then you need to use a two-tailed test as you are looking for some change in the parameter, rather than an increase or decrease.

We would have used a two-tailed test for our SnoreCull if our alternate hypothesis had been $p \neq 0.9$. We would have had to check whether significantly more or significantly fewer than 90% of patients had been cured

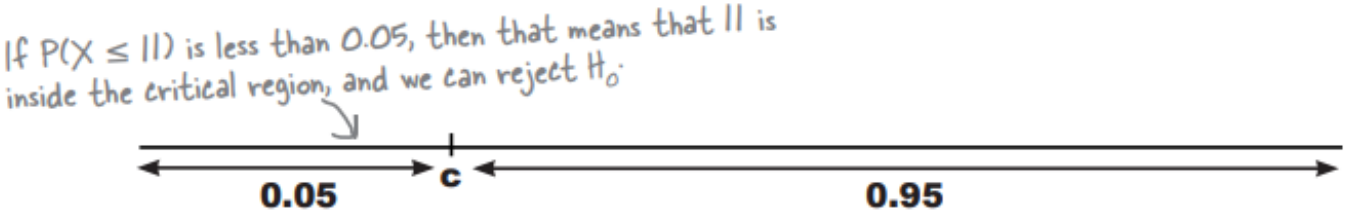


Step 4: Find the p-value

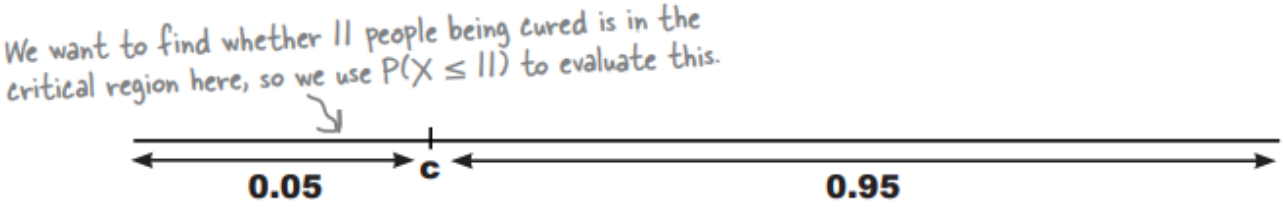
A p-value is the probability of getting a value up to and including the one in your sample in the direction of your critical region. It's a way of taking your sample and working out whether the result falls within the critical region for your hypothesis test. In other words, we use the p-value to say whether or not we can reject the null hypothesis.

How do we find the p-value?

For the SnoreCull test, 11 people were cured, and our critical region is the lower tail of the distribution. This means that our p-value is P(X 11), where X is the distribution for the number of people cured in the sample. As the significance level of our test is 5%, this means that if P(X 11) is less than 0.05, then the value 11 falls within the critical region, and we can reject the null hypothesis.

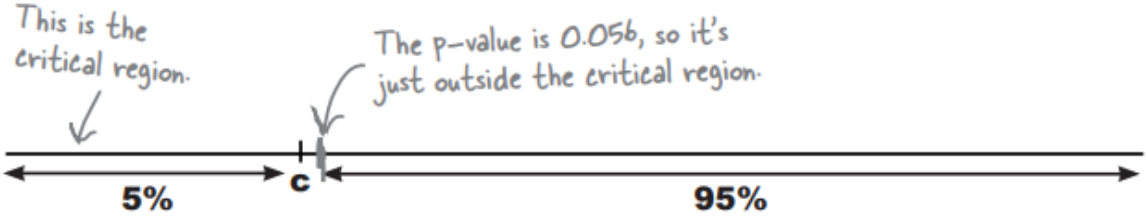


A p-value is the probability of getting the results in the sample, or something more extreme, in the direction of the critical region.



Step 5: Is the sample result in the critical region?

Now that we've found the p-value, we can use it to see whether the result from our sample falls within the critical region. If it does, then we'll have sufficient evidence to reject the claims of the drug company. Our critical region is the lower tail of the probability distribution, and we're using a significance level of 5%. This means that we can reject the null hypothesis if our p-value is less that 0.05. As our p-value is 0.0555, this means that the number of people cured by SnoreCull in the sample doesn't fall within the critical region



Step 6: Make your decision

The p-value of the hypothesis test falls just outside the critical region of the test. This means that there isn't sufficient evidence to reject the null hypothesis. In other words:

BULLET POINTS

- In a hypothesis test, you take a claim and test it against statistical evidence.
- The claim that you're testing is called the null hypothesis test. It's represented as H_0, and it's the claim that's accepted unless there's strong statistical evidence against it.
- The alternate hypothesis is the claim we'll accept if there's strong enough evidence against H_0. It's represented by H_1.
- The test statistic is the statistic you use to test your hypothesis. It's the statistic that's most relevant to the test. You choose the test statistic by assuming that H_0 is true.
- The significance level is represented by α. It's a way of saying how unlikely you want your results to be before you'll reject H_0.
- The critical region is the set of values that presents the most extreme evidence against the null hypothesis test. You choose your critical region by considering the significance level and how many tails you need to use.
- A one-tailed test is when your critical region lies in either the upper or the lower tail of the data. A two-tailed test is when it's split over both ends. You choose your tail by looking at your alternate hypothesis.
- A p-value is the probability of getting the result of your sample, or a result more extreme in the direction of your critical region.
- If the p-value lies in the critical region, you have sufficient reason to reject your null hypothesis. If your p-value lies outside your critical region, you have insufficient evidence.

What if the sample size is larger?
use normal distribution.

Distribution	Description	Typical Cases
Binomial Distribution	Describes the number of successes in a fixed number of independent Bernoulli trials.	Small sample size, discrete outcomes, success/failure trials.
Normal Distribution	Describes continuous data and is widely used due to the Central Limit Theorem for large sample sizes.	Large sample size, approximately normal data.
t-Distribution	Similar to the normal distribution but accounts for smaller sample sizes and unknown population variance.	Small sample size, when population variance is unknown.
Chi-Squared Distribution	Used to test hypotheses about the variance of a normally distributed population.	Testing variance or goodness-of-fit tests.
F-Distribution	Used in analysis of variance (ANOVA) and regression analysis to compare variances between groups.	Comparing variances between multiple groups.

Correlation & Regression

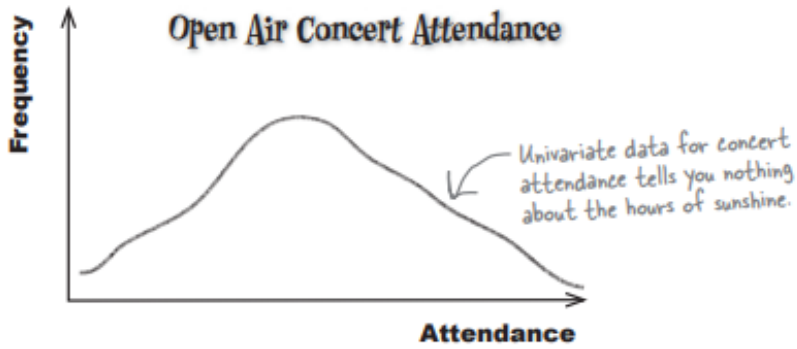
Have you ever wondered how two things are connected? So far we've looked at statistics that tell you about just one variable—like men's height, points scored by basketball players, or how long gumball flavor lasts—but there are other statistics that tell you about the connection between variables. Seeing how things are connected can give you a lot of information about the real world, information that you can use to your advantage.

Exploring types of data

Up until now, the sort of data we've been dealing with has been univariate.

Univariate data concerns the frequency or probability of a single variable. As an example, univariate data could describe the winnings at a casino or the weights of brides in Statsville. In each case, just one thing is being described.

What univariate data can't do is show you connections between sets of data. For example, if you had univariate data describing the attendance figures at an open air concert, it wouldn't tell you anything about the predicted hours of sunshine on that day. It would just give you figures for concert attendance.



So what if we do need to know what the connection is between variables? While univariate data can't give us this information, there's another type of data that can—**bivariate data**.

All about bivariate data

Bivariate data gives you the value of *two* variables for each observation, not just one. As an example, it can give you both the predicted hours of sunshine and the concert attendance for a single event or observation, like this.

Bivariate data gives you the value of two variables for each observation.

Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100's)	22	33	30	42	38	49	42	55

If one of the variables has been controlled in some way or is used to explain the other, it is called the ***independent*** or ***explanatory*** variable. The other variable is called the ***dependent*** or ***response*** variable. In our example, we want to use sunshine to predict attendance, so sunshine is the independent variable, and attendance is the dependent.

Visualizing bivariate data

Just as with univariate data, you can draw charts for bivariate data to help you see patterns. Instead of plotting a value against its frequency or probability, you plot one variable on the x-axis and the other variable against it on the y-axis. This helps you to visualize the connection between the two variables.

This sort of chart is called a ***scatter diagram*** or ***scatter plot***, and drawing one of these is a lot like drawing any other sort of chart.

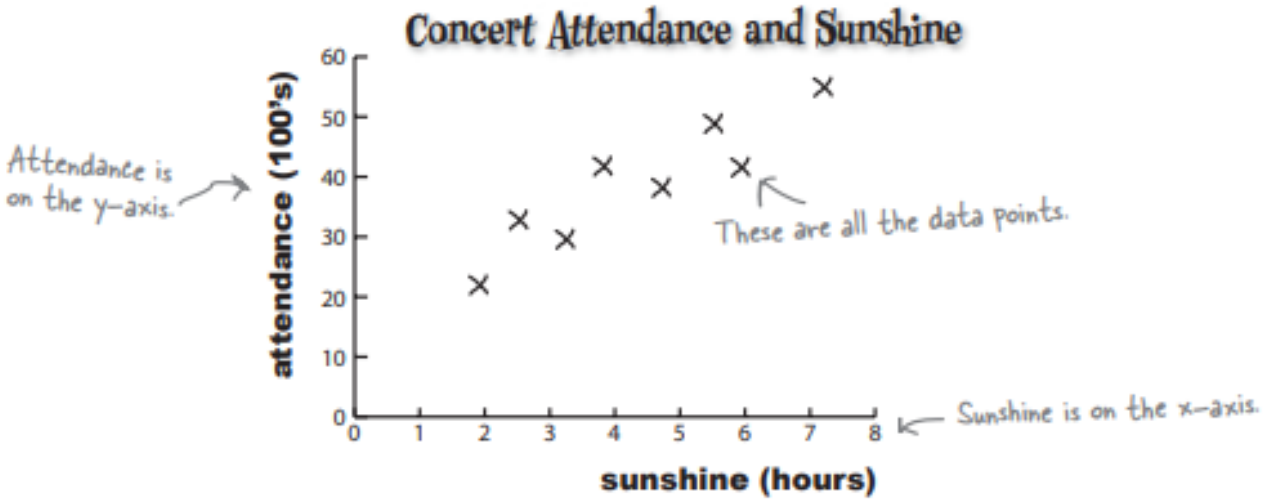
Start off by drawing two axes, one vertical and one horizontal. Use the x-axis for one variable and the y-axis for the other. The independent variable normally goes along the x-axis, leaving the dependent variable to go on the y-axis. Once you've drawn your axes, you then take the values for each observation and plot them on the scatter plot.

Here's a scatter plot showing the number of hours of sunshine and concert attendance figures for particular events or observations. As the predicted number of hours sunshine is the independent variable, we've plotted it on the x-axis. The concert attendance is the dependent variable, so that's on the y-axis.

Hours sunshine goes on the x-axis, attendance on the y-axis.

x (sunshine)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
y (attendance)	22	33	30	42	38	49	42	55

Here's the data.



Can you see how the scatter diagram helps you visualize patterns in the data? Can you see how this might help us to define the connection between open air concert attendance and predicted number of hours sunshine for the day?

Scatter diagrams show you patterns

As you can see, scatter diagrams are useful because they show the actual pattern of the data. They enable you to more clearly visualize what connection there is between two variables, if indeed there's any connection at all.

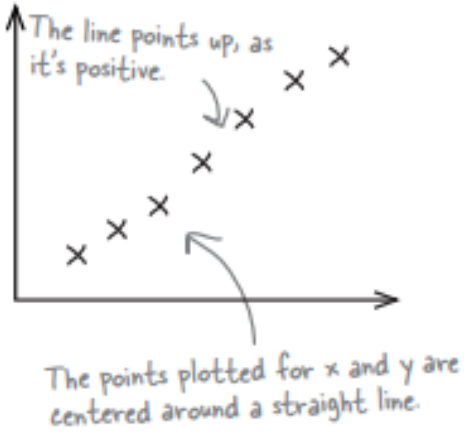
The scatter diagram for the concert data shows a distinct pattern—the data points are clustered along a straight line. We call this a ***correlation***.

Linear Correlations Up Close

Scatter diagrams show the ***correlation*** between pairs of values.

Correlations are mathematical relationships between variables. You can identify correlations on a scatter diagram by the distinct patterns they form. The correlation is said to be ***linear*** if the scatter diagram shows the points lying in an approximately straight line.

Let's take a look at a few common types of correlation between two variables:

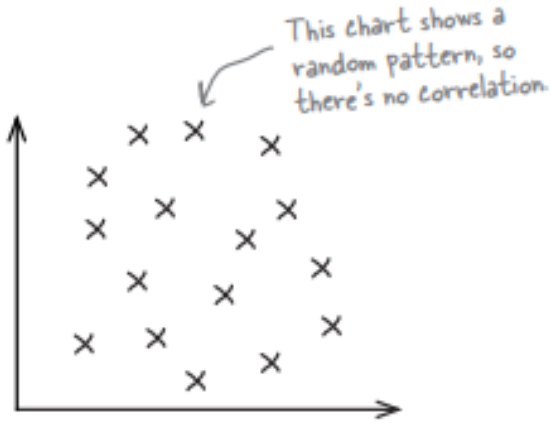


Positive linear correlation

Positive linear correlation is when low values on the x-axis correspond to low values on the y-axis, and higher values of x correspond to higher values of y. In other words, y tends to increase as x increases.

Negative linear correlation

Negative linear correlation is when low values on the x-axis correspond to high values on the y-axis, and higher values of x correspond to lower values of y. In other words, y tends to decrease as x increases.



No correlation

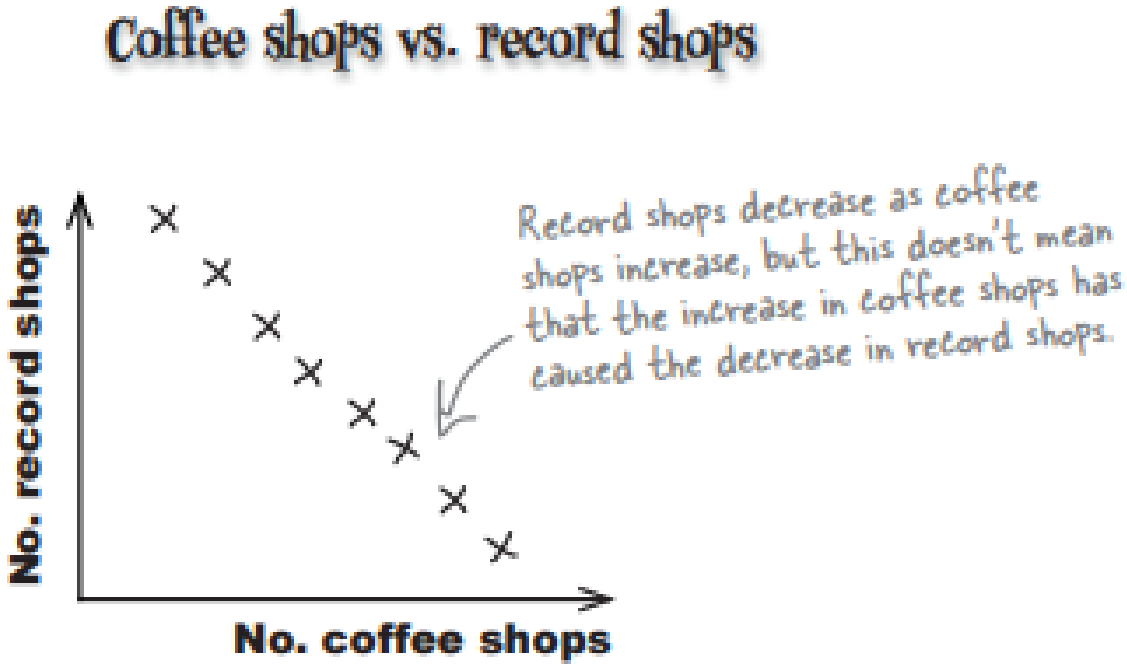
If the values of x and y form a random pattern, then we say there's no correlation.

Correlation vs Causation

A correlation between two variables doesn't necessarily mean that one caused the other or that they're actually related in real life.

A correlation between two variables means that there's some sort of **mathematical relationship** between the two. This means that when we plot the values on a chart, we can see a pattern and make predictions about what the missing values might be. What we don't know is whether there's an **actual relationship** between the two variables, and we certainly don't know whether one caused the other, or if there's some other factor at work.

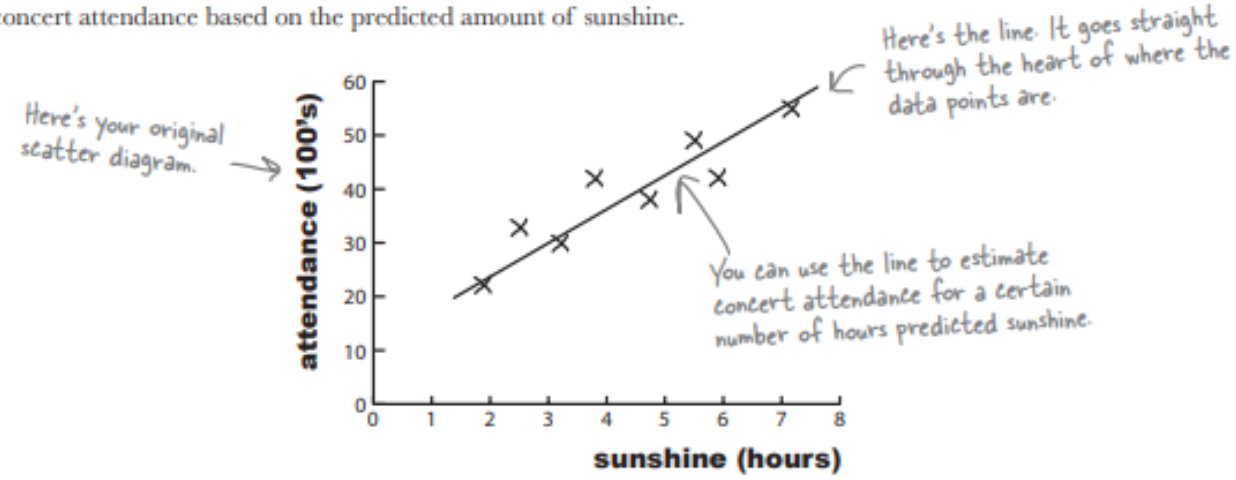
As an example, suppose you gather data and find that over time, the number of coffee shops in a particular town increases, while the number of record shops decreases. While this may be true, we can't say that there is a real-life relationship between the number of coffee shops and the number of record shops. In other words, we can't say that the increase in coffee shops caused the decline in the record shops. What we *can* say is that as the number of coffee shops increases, the number of record shops decreases.



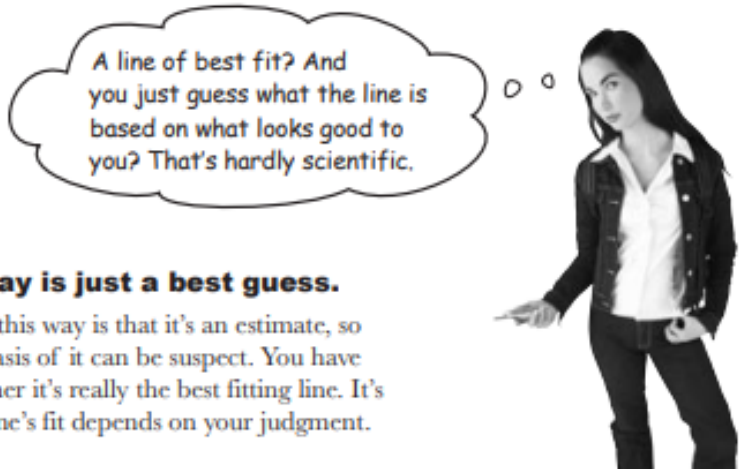
Predict values with a line of best fit

So far you've seen how scatter diagrams can help you see whether there's a correlation between values, by showing you if there's some sort of pattern. But how can you use this to predict concert attendance, based on the predicted amount of sunshine? How would you use your existing scatter diagram to predict the concert attendance if you know how many hours of sunshine are expected for the day?

One way of doing this is to draw a straight line through the points on the scatter diagram, making it fit the points as closely as possible. You won't be able to get the straight line to go through every point, but if there's a linear correlation, you should be able to make sure every point is reasonably close to the line you draw. Doing this means that you can read off an estimate for the concert attendance based on the predicted amount of sunshine.



The line that best fits the data points is called the ***line of best fit***.

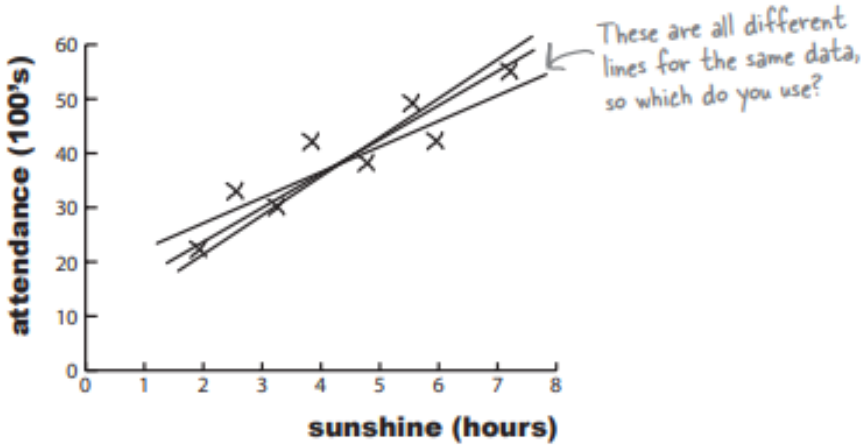


Drawing the line in this way is just a best guess.

The trouble with drawing a line in this way is that it's an estimate, so any predictions you make on the basis of it can be suspect. You have no precise way of measuring whether it's really the best fitting line. It's subjective, and the quality of the line's fit depends on your judgment.

Your best guess is still a guess

Imagine if you asked three different people to draw what each of them think is the line of best fit for the open air concert data. It's quite likely that each person would come up with a slightly different line of best fit, like this:



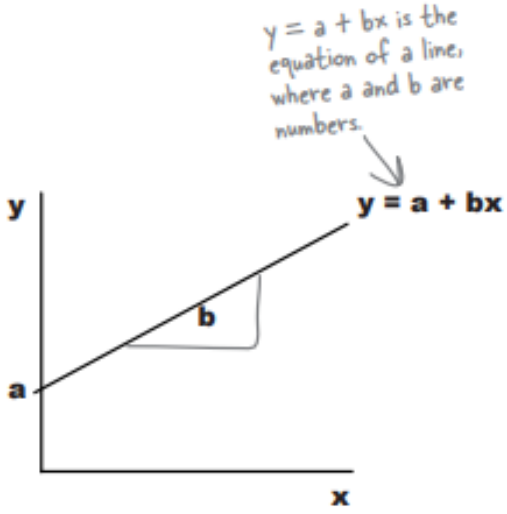
All three lines could conceivably be a line of best fit for the data, but what we can't tell is which one's really best.

What we really need is some alternative to drawing the line of best fit by eye. Instead of guessing what the line should be, it will be more reliable if we had a mathematical or statistical way of using the data we have available to find the line that fits best.

We need to find the equation of the line

The equation for a straight line takes the form $y = a + bx$, where a is the point where the line crosses the y-axis, and b is the slope of the line. This means that we can write the line of best fit in the form $y = a + bx$.

In our case, we're using x to represent the predicted number of hours of sunshine, and y to represent the corresponding open air concert figures. If we can use the concert attendance data to somehow find the most suitable values of a and b, we'll have a reliable way to find the equation of the line, and a more reliable way of predicting concert attendance based on predicted hour of sunshine.

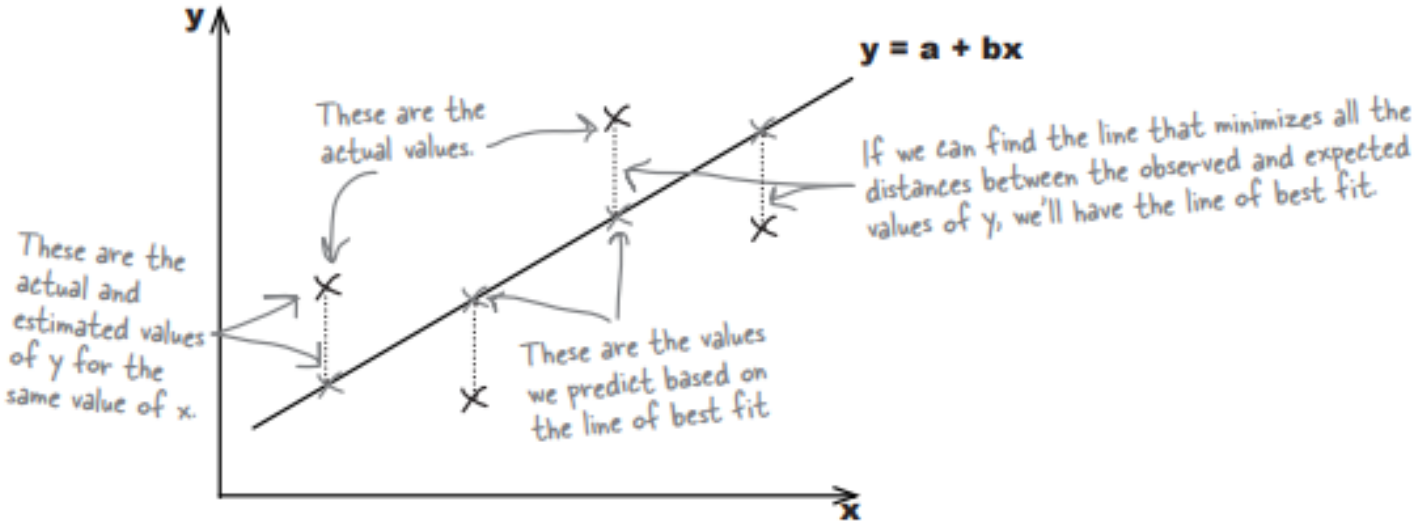


We need to minimize the errors

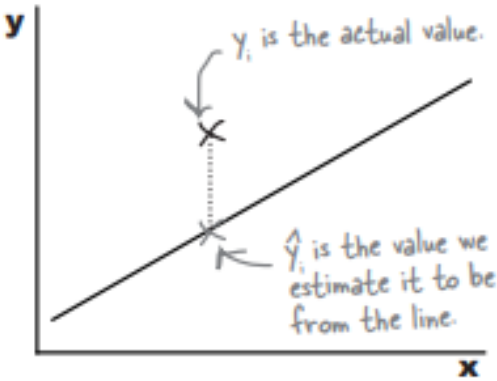
Let's take a look at what we need from the line of best fit, $y = a + bx$.

The best fitting line is the one that most accurately predicts the true values of all the points. This means that for each known value of x, we need each of the y variables in the data set to be as close as possible to what we'd estimate them to be using the line of best fit. In other words, given a certain number of hours sunshine, we want our estimates for open air concert attendance to be as close as possible to the actual values.

The line of best fit is the line $y = a + bx$ that minimizes the distances between the actual observations of y and what we estimate those values of y to be for each corresponding value of x.



Let's represent each of the y values in our data set using y_i, and its estimate using the line of best fit as $\hat{y}_i$. This is the same notation that we used for point estimators in previous chapters, as the ^ symbol indicates estimates.



We want to minimize the total distance between each actual value of y and our estimate of it based on the line of best fit. In other words, we need to minimize the total differences between y_i and $\hat{y}_i$. We could try doing this by minimizing

$$\Sigma(y_i - \hat{y}_i)$$

but the problem with this is that all of the distances will actually cancel each other out. We need to take a slightly different approach, and it's one that we've seen before.

Introducing the sum of squared errors

Can you remember when we first derived the variance? We wanted to look at the total distance between sets of values and the mean, but the total distances cancelled each other out. To get around this, we added together all the distances squared instead to ensure that all values were positive.

We have a similar situation here. Instead of looking at the total distance between the actual and expected points, we need to add together the distances *squared*. That way, we make sure that all the values are positive.

The total sum of the distances squared is called the **sum of squared errors**, or **SSE**. It's given by:

The sum of squared errors → $$SSE = \Sigma(y - \hat{y})^2$$ ← The difference between the real values of y, and what we predict from the line of best fit

In other words, we take each value of y, subtract the predicted value of y from the line of best fit, square it, and then add all the results together.

Least Squares Regression Up Close

The mathematical method we've been using to find the line of best fit is called **least squares regression**.

Least squares regression is a mathematical way of fitting a line of best fit to a set of bivariate data. It's a way of fitting a line y = a + bx to a set of values so that the sum of squared errors is minimized—in other words, so that the distance between the actual values and their estimates are minimized. The sum of squared errors is given by

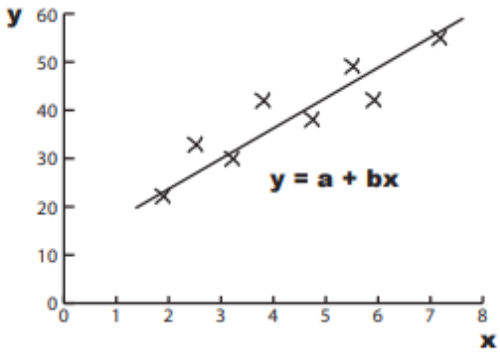
$$SSE = \Sigma(y - \hat{y})^2$$

To perform least squares regression on a set of data, you need to find the values of a and b that best fit the data points to the line y = a + bx and minimizes the SSE. You can do this using:

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

and

$$a = \bar{y} - b\bar{x}$$



Once you've found the line of best fit, y = a + bx, you can use it to predict the value of y, given a value b. To do this, just substitute your x value into the equation y = a + bx.

The line y = a + bx is called the ***regression line***.

Watch it!

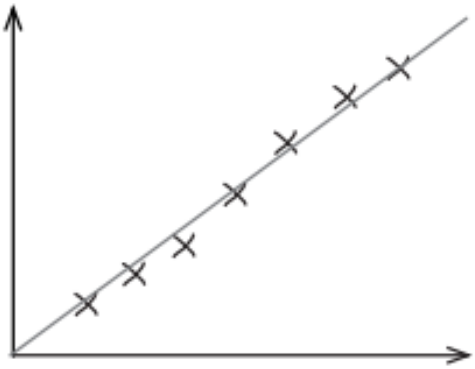
When you're predicting values of y for a particular value of x, be wary of predicting values that fall outside the area you have data points for.

Linear regression is just an estimate based on the information you have, and it shows the relationship between the data points you know about. This doesn't mean that it applies well beyond the limits of the data

Let's look at some correlations

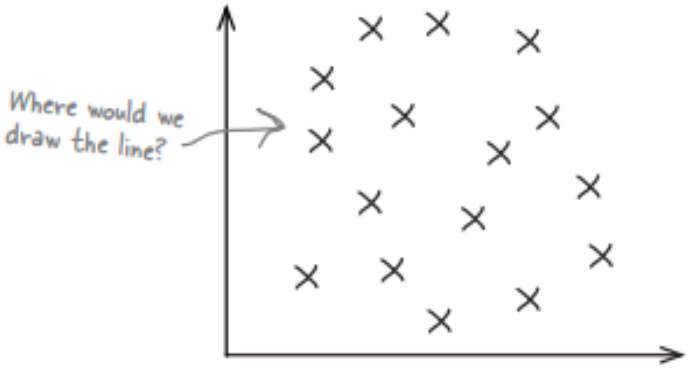
The line of best fit of a set of data is the best line we can come up with to model the mathematical relationship between two variables.

Even though it's the line that fits the data best, it's unlikely that the line will fit precisely through every single point. Let's look at some different sets of data to see how closely the line fits the data.



Accurate linear correlation

For this set of data, the linear correlation is an accurate fit of the data. The regression line isn't 100% perfect, but it's very close. It's likely that any predictions made on the basis of it will be accurate.



No linear correlation

For this set of data, there is no linear correlation. It's possible to calculate a regression line using least squares regression, but any predictions made are unlikely to be accurate.

Can you see what the problem is?

Both sets of data have a regression line, but the actual fit of the data varies quite a lot. For the first set of data, the correlation is very tight, but for the second, the points are scattered too widely for the regression line to be useful.

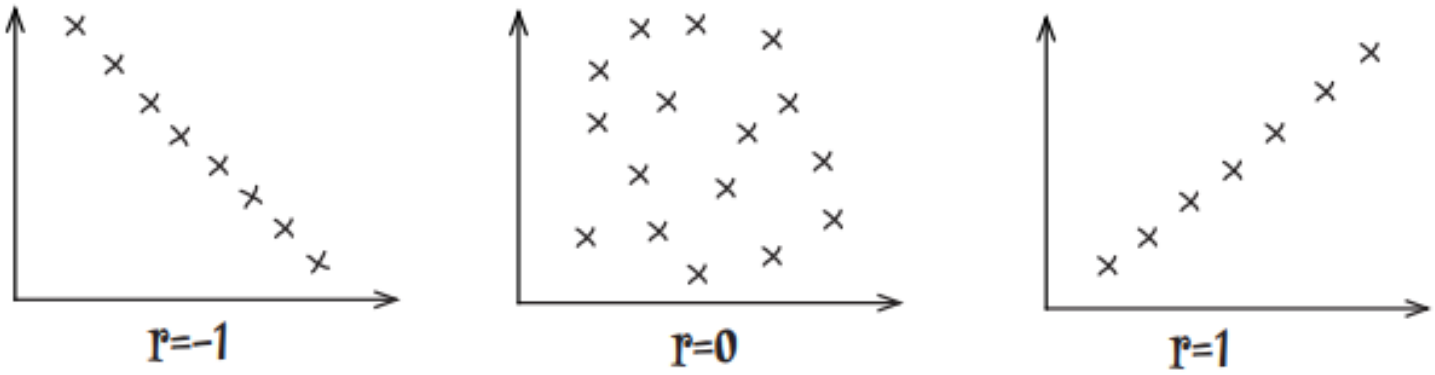
Least squares estimates can be used to predict values, which means they would be helpful if there was some way of indicating how tightly the data points fit the line, and how accurate we can expect any predictions to be as a result.

There's a way of calculating the fit of the line, called the **correlation coefficient**.

The correlation coefficient measures how well the line fits the data

The ***correlation coefficient*** is a number between -1 and 1 that describes the scatter of data points away from the line of best fit. It's a way of gauging how well the regression line fits the data. It's normally represented by the letter r.

If r is -1, the data is a **perfect negative linear correlation**, with all of the data point in a straight line. If r is 1, the data is a **perfect positive linear correlation**. If r is 0, then there is **no correlation**.



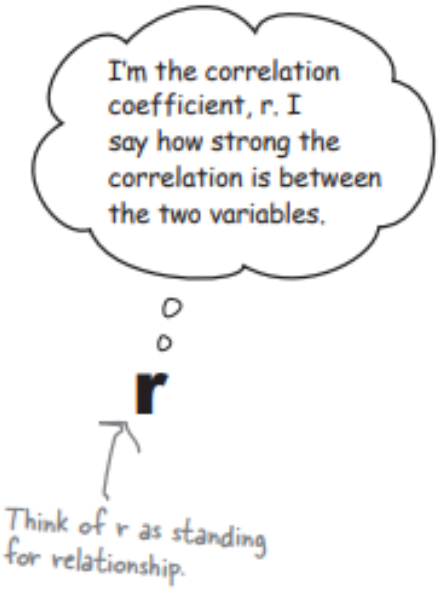
Usually r is somewhere between these values, as -1, 0, and 1 are all extreme.

If r is negative, then there's a **negative linear correlation** between the two variables. The closer r gets to -1, the stronger the correlation, and the closer the points are to the line.

If r is positive, then there's a **positive linear correlation** between the variables. The closer r gets to 1, the stronger the correlation.

In general, as r gets closer to 0, the **linear correlation gets weaker**. This means that the regression line won't be able to predict y values as accurately as when r is close to 1 or -1. The pattern might be random, or the relationship between the variables might not be linear.

If we can calculate r for the concert data, we'll have an idea of how accurately we can predict concert attendance based on the predicted hours of sunshine. So how do we calculate r? Turn the page and we'll show you how.



There's a formula for calculating the correlation coefficient, r

So how do we calculate the correlation coefficient, r?

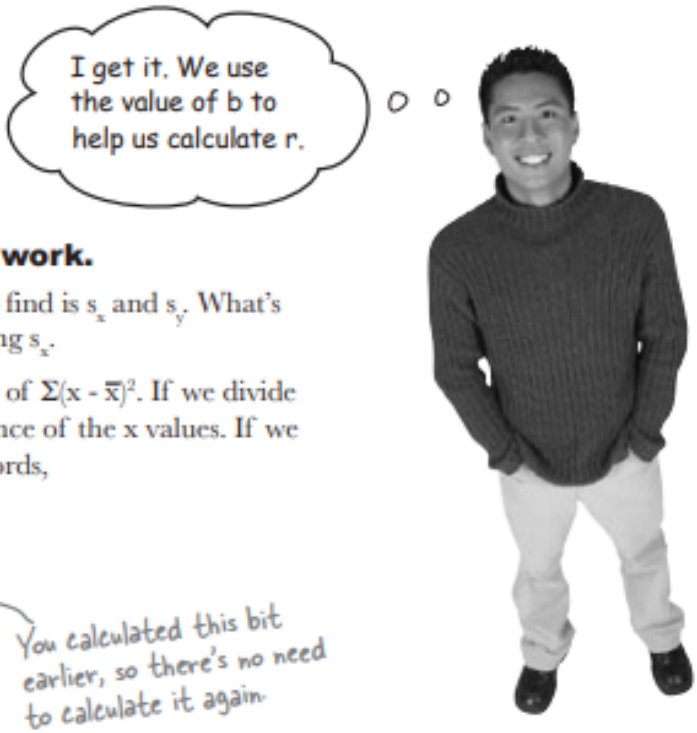
We're not going to show you the proof for this, but the correlation coefficient r is given by

$$r = b \frac{s_x}{s_y}$$

b is the slope of the line of best fit that you've already found.

s_x is the standard deviation of the x values in the sample. s_y is the standard deviation of the y values.

where s_x is the standard deviation of the x values in the sample, and s_y is the standard deviation of the y values.



We've already done most of the hard work.

Since we've already calculated b, all we have left to find is s_x and s_y. What's more, we're already most of the way towards finding s_x.

When we calculated b, we needed to find the value of $\Sigma(x - \bar{x})^2$. If we divide this by n - 1, this actually gives us the sample variance of the x values. If we then take the square root, we'll have s_x. In other words,

This is the standard deviation of the x values in the sample, it's the same formula you've seen before

$$s_x = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

You calculated this bit earlier, so there's no need to calculate it again.

The only remaining piece of the equation we have to find is s_y, the standard deviation of the y values in the sample. We calculate this in a similar way to finding s_x.

$$s_y = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n - 1}}$$

This is the standard deviation of the x values in the sample, and you've done these sorts of calculations before.

Let's try finding what r is for the concert attendance data.

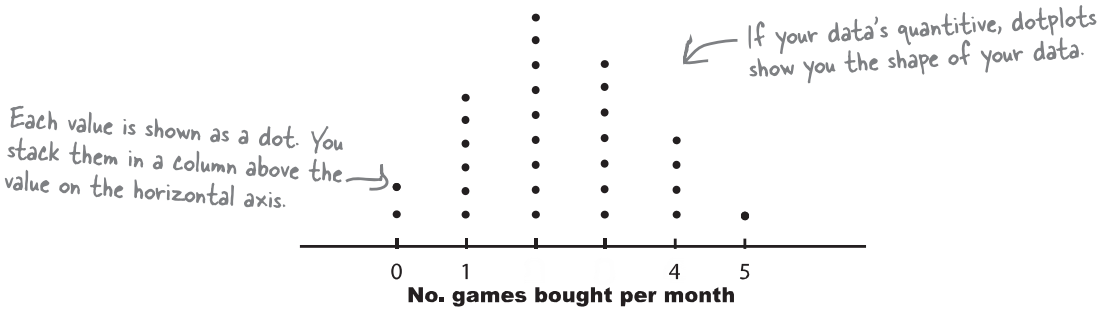
The Top Ten Things (we didn't cover)

#1. Other ways of presenting data

We showed you a number of charts in the first chapter, but here are a couple more that might come in useful.

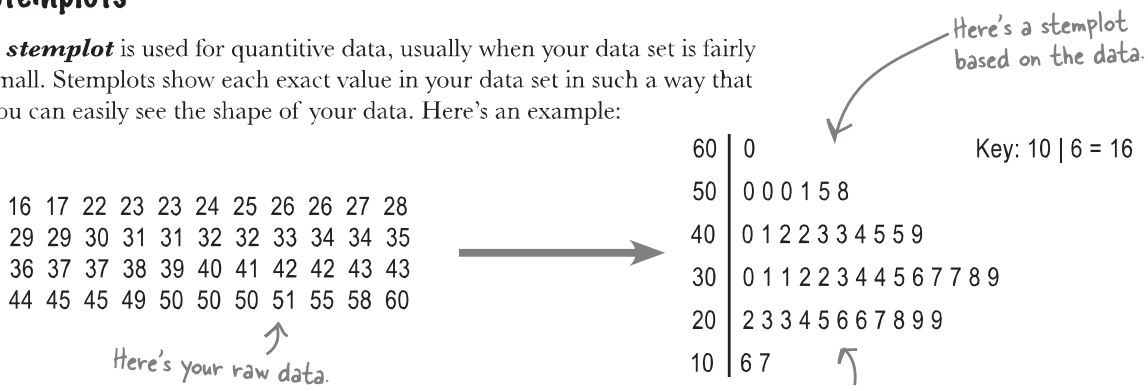
Dotplots

A ***dotplot*** shows your data on a chart by representing each value as a dot. You put each dot in a stacked column above the corresponding value on the horizontal axis like this:



Stemplots

A ***stemplot*** is used for quantitive data, usually when your data set is fairly small. Stemplots show each exact value in your data set in such a way that you can easily see the shape of your data. Here's an example:



The entries on the left are called ***stems***, and the entries on the right are called ***leaves***. In this stemplot, the stem shows tens, and the leaves show units. To find each value in the raw data, you take each leaf and add it to its stem. As an example, take the line

10 | 6 7

This represents two numbers, 16 and 17. You get 16 by adding the leaf 6 to its stem 10. Similarly, you get value 17 by adding the leaf 7 to the stem 10.

There's usually a key to help you interpret the stemplot correctly. In this case, the key is 10 | 6 = 16.

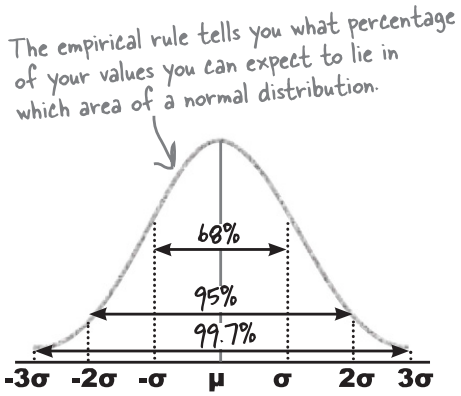
#2. Distribution anatomy

There are two rules that tell you where most of your data values lie in a probability distribution.

The empirical rule for normal distributions

The ***empirical rule*** applies to any set of data that follows a normal distribution. It states that almost all of the values lie within three standard deviations of the mean. In particular,

- About 68% of your values lie within 1 standard deviation of the mean.
- About 95% of your values lie within 2 standard deviations of the mean.
- About 99.7% of your values lie within 3 standard deviations of the mean.



Just knowing the number of standard deviations from the mean can give you a rough idea about the probability.

Chebyshev's rule for any distribution

A similar rule applies to any set of data called ***Chebyshev's rule***, or ***Chebyshev's inequality***. It states that for any distribution

- At least 75% of your values lie within 2 standard deviations of the mean.
- At least 89% of your values lie within 3 standard deviations of the mean.
- At least 94% of your values lie within 4 standard deviations of the mean.

Chebychev's rule isn't as precise as the empirical rule, as it only gives you the minimum percentages, but it still gives you a rough idea of where values fall in the probability distribution. The advantage of Chebyshev's rule is that it **applies to any distribution**, while the empirical rule just applies to the normal distribution.

#3. Experiments

Experiments are used to test cause and effect relationships between variables. As an example, an experiment could test the effect of different doses of SnoreCull on snorers.

In an experiment, ***indpendent variables*** or factors are manipulated so that we can see the effect on ***dependent variables***. As an example, we might want to examine the effect that different doses of SnoreCull have on the number of hours spent snoring in a night. The doses of SnoreCull would be the independent variable, and the number of hours spent snoring would be the dependent variable.

The subjects that you use for your experiment are called ***experimental units***—in this case, snorers.

So what makes for a good experiment?

There are three basic principles you need to bear in mind when you design an experiment: ***controls***, ***randomization***, and ***replication***. Just as with sampling, a key aim is to minimize bias.

- **You need to control the effects of external influences or natural variability.**
  When you conduct an experiment, you need to minimize effects that are not part of the experiment. To do this, the first thing is to have a ***control group***, a neutral group that receives no treatments, or only neutral treatments. You can assess the effectiveness of the treatment by comparing the results of your treated groups with the results of your control group.

  A ***placebo*** is a neutral treatment, one that has no effect on the dependent variable. Sometimes the subjects of your experiment can respond differently to having a neutral treatment as opposed to having no treatment at all, so giving a placebo to a group is a way of controlling this effect. If the group taking a placebo doesn't know that it's a placebo, then this is called ***blinding***, and it's called ***double blinding*** if even those administering the treatments don't know.

- **You need to assign subjects to treatments at random.**
  You'll see more about this on the next page.

- **You need to replicate treatments.**
  Each treatment should be given to many subjects. You need to use many snorers per treatment to gauge the effects, not just one snorer.

Another factor to be aware of is confounding. ***Confounding*** occurs when the controls in an experiment don't eliminate other possible causes for the effect on the dependent variable. As an example, imagine if you gave doses to SnoreCull to men, but placebos to women. If you compared the results of the two groups, you wouldn't be able to tell whether the effect on the men was because of the drug, or because one gender naturally snores more than another.

Designing your experiment

We said earlier that you need to randomly assign subjects to experiments. But what's the best way of doing this?

Completely randomized design

One option is to use a ***completely randomized design***. For this, you literally assign treatments to subjects at random. If we were to conduct an experiment testing the effect of doses of SnoreCull on snorers, we would randomly assign snorers to particular treatment groups. As an example, we could give half of the snorers a placebo and the other half a single dose of SnoreCull.

Completely randomized design is similar to simple random sampling. Instead of choosing a sample at random, you assign treatments at random.

Placebo	SnoreCull
500	500

If there were 1,000 subjects, we could give half a placebo and the other half a dose of SnoreCull

Randomized block design

Another option is to use ***randomized block design***. For this, you divide the subjects into similar groups, or blocks. As an example, you could split the snorers into males and females. Within each block, you assign treatments at random, so for each gender, you could give half the snorers a dose of SnoreCull and give the other half a placebo. The aim of this is to minimize confounding, as it reduces the effect of gender.

Randomized block design is similar to stratified random sampling. Instead of splitting your population into strata, you split your subjects into blocks.

	Placebo	SnoreCull
Male	250	250
Female	250	250

If there were 500 men and 500 women, we could give half of each gender a placebo and the other half a dose of SnoreCull

Matched pairs design

Matched pairs design is a special case of randomized block design. You can use it when there are only two treatment conditions and subjects can be grouped into like pairs. As an example, the SnoreCull experiment could have two treatment conditions, to give a placebo or to give a single dose, and snorers could be grouped into similar pairs according to gender and age. You then give one of each pair a placebo, and the other a dose of SnoreCull. If one pair consisted of two men aged 30, for instance, you would give one of the men a placebo and the other man a dose of SnoreCull.

	Placebo	SnoreCull
Male 30	1	1
Male 30	1	1
Female 30	1	1
Female 30	1	1
...	...	...

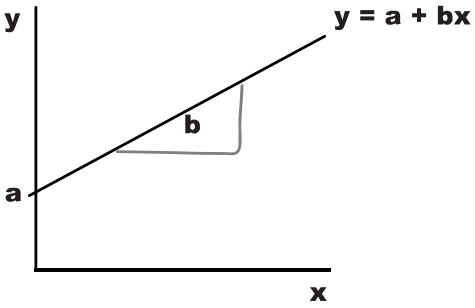
You could also form matched pairs using gender and age to negate confounding due to these variables.

#4. Least square regression alternate notation

In Chapter 15 you saw how a least squares regression line takes the form y = a + bx, where

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

This is the formula for the slope of the line.



There's another form of writing this that a lot of people find easier to remember, and that's to rewrite it in terms of variances. If we use the notation

$$s_x^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

Sample variance of the x values

$$s_y^2 = \frac{\Sigma(y - \bar{y})^2}{n - 1}$$

Sample variance of the y values

$$s_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1}$$

then you can rewrite the formula for the slope of a line as

$$b = \frac{s_{xy}}{s_x^2}$$

This is the same calculation written in a different way.

You can do something similar with the correlation coefficient. Instead of writing

$$r = b\frac{s_x}{s_y}$$

you can write the equation for the correlation coefficient as

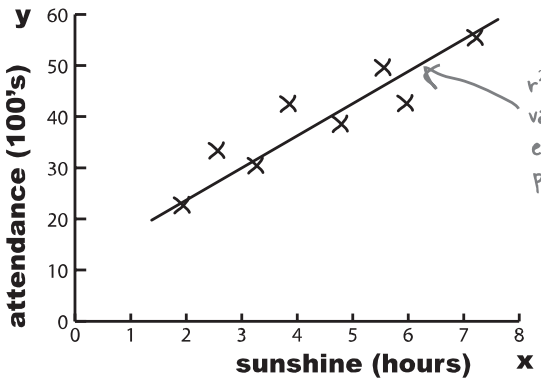
This is the formula for the correlation coefficient.

$$r = \frac{s_{xy}}{s_x s_y}$$

s_{xy} is called the ***covariance***. Just as the variance of x describes how the x values vary, and the variance of y describes how the y values vary, the covariance of x and y is a measure of how x and y vary together.

#5. The coefficient of determination

The ***coefficient of determination*** is given by $\mathbf{r^2}$ or $\mathbf{R^2}$. It's the percentage of variation in the y variable that's explainable by the x variable. As an example, you can use it to say what percentage of the variation in open-air concert attendance is explainable by the number of hours of predicted sunshine.



r^2 tells you what percentage of the variation in concert attendance is explainable by the number of hours predicted sunshine.

If $r^2 = 0$, then this means that you can't predict the y value from the x value.

If $r^2 = 1$, then you can predict the y value from the x value without any errors.

Usually r^2 is between these two extremes. The closer the value of r^2 is to 1, the more predictable the value of y is from x, and the closer to r^2 it is, the less predictable the value of y is.

Calculating r^2

There are two ways of calculating r^2. The first way is to just square the correlation coefficient r.

This is just the correlation coefficient squared.

$$r^2 = \left(\frac{s_{xy}}{s_x s_y}\right)^2$$

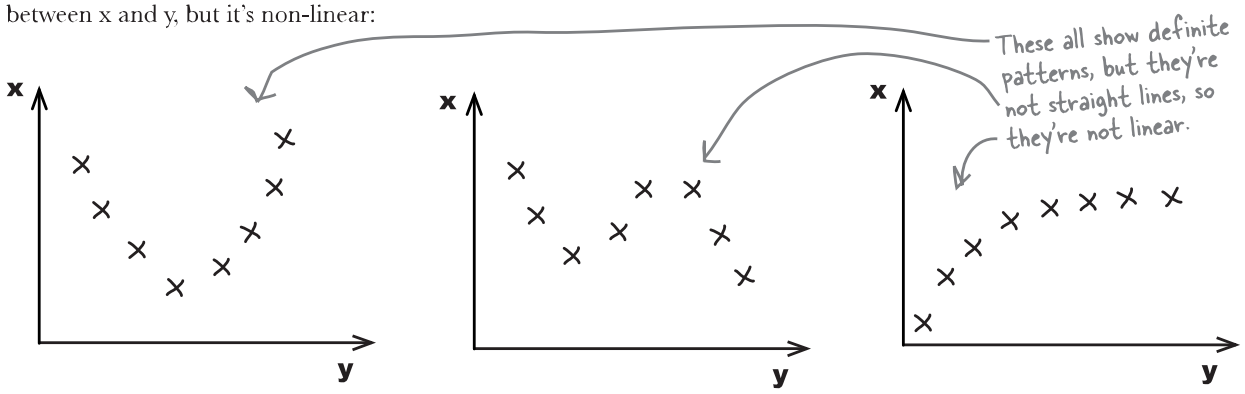
Another way of calculating it is to add together the squared distances of the y values to their estimates, and then divide by the result of adding together squared distances of the y values to $\bar{y}$.

$$r^2 = \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

This gives you the same value as above; it's just a different way of calculating it.

#6. Non-linear relationships

If two variables are related, their relationship isn't necessarily linear. Here are some examples of scatter plots where there's a clear mathematical relationship between x and y, but it's non-linear:



Linear regression assumes that the relationship between two variables can be described by a straight line, so performing least squares regression on raw data like this won't give you a good estimate for the equation of the line.

There is a way around this, however. You can sometimes transform x and y in such a way that the transformation is close to being linear. You can then perform linear regression on the transformation to find the values of a and b. The big trick is to try and transform your non-linear equation of the line so that it takes the form

$$y' = a + bx'$$

where y' and x' are functions of x.

As an example, you might find that your line of best fit takes the form

$$y = 1/(a + bx)$$

This can be rewritten as

$$1/y = a + bx$$

It's now in the form y' = a + bx, so we can use linear regression.

so that y' = 1/y. In other words, you can perform least squares regression using the line y' = a + bx, where y' = 1/y. Once you've transformed your y values, you can use least squares regression to find the values of a and b, then substitute these back into your original equation.

This is just a quick overview, so you know what's possible.

If your line of best fit isn't linear, you can sometimes transform it to a linear form.

#7. The confidence interval for the slope of a regression line

You've seen how you can find confidence intervals for μ and σ^2. Well, you can also find one for the slope of the regression line $y = a + bx$.

The confidence interval for b takes the form

$$\hat{b} \pm \text{(margin of error)}$$

But what's the margin of error?

The margin of error for b

The margin of error is given by

$$\textbf{margin of error} = t(\nu) \times \textbf{(standard deviation of b)}$$

where $\nu = n - 2$, and n is the number of observations in your sample. To find the value of $t(\nu)$, use t-distribution probability tables to look up ν and your confidence level.

The standard deviation of the sampling distribution of b is given by

This is the standard deviation of the sampling distribution of b →

$$s_b = \frac{\sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}}}{\sqrt{\Sigma(x - \bar{x})^2}}$$

To calculate this, add together the differences squared between each actual y observation and what you estimate it to be from the regression line. Then divide by $n - 2$, and take the square root. Once you've done this, divide the whole lot by the square root of the total differences squared between each x observation and $\bar{x}$.

This gives us a confidence interval of

$$(\hat{b} - t(v)\, s_b,\ \hat{b} + t(v)\, s_b)$$

You use the t-distribution with n – 2 degrees of freedom.

$$v = n - 2$$

Knowing the standard deviation of b has other uses too. As an example, you can also use it in hypothesis tests to test whether the slope of a regression line takes a particular value.

If you're taking a statistics exam where you have to use s_b, the formula will be given to you.

This means that you don't have to memorize it; you just need to know how to apply it.

#8. Sampling distributions - the difference between two means

Sometimes it's useful to know what the sampling distribution is like for the difference between the means of two normally distributed populations. You may want to use this to construct a confidence interval or conduct a hypothesis test. As an example, you may want to conduct a hypothesis test based on the means of two normally distributed populations being equal.

If $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ where X and Y are independent, then the expectation and variance of the distribution $\overline{X} - \overline{Y}$ are given by

This is because $E(\overline{X} - \overline{Y}) = E(\overline{X}) - E(\overline{Y})$

$$E(\overline{X} - \overline{Y}) = \mu_x - \mu_y$$

$$Var(\overline{X} - \overline{Y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$$

Similarly, $Var(\overline{X} - \overline{Y}) = Var(\overline{X}) + Var(\overline{Y})$

If the population variances σ_x^2 and σ_y^2 are known, then $\overline{X} - \overline{Y}$ is distributed normally. In other words

$$\overline{X} - \overline{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

You can use this to find a confidence interval for $\overline{X} - \overline{Y}$. Confidence intervals take the form (statistic) ± (margin of error), so in this case, the confidence interval is given by

$$\overline{x} - \overline{y} \pm c\sqrt{Var(\overline{X} - \overline{Y})}$$

This is your confidence interval for $\overline{X} - \overline{Y}$.

The value of c depends on the level of confidence you need for your confidence interval:

Level of confidence	Value of c
90%	1.64
95%	1.96
99%	2.58

Your level of confidence gives you your value of c

If σ_x^2 and σ_y^2 are unknown, then you will need to approximate them with s_x^2 and s_y^2. If the samples sizes are large, then you can still use the normal distribution. If the sample sizes are small, then you will need to use the t-distribution instead.

#9. Sampling distributions - the difference between two proportions

There's also a sampling distribution for the difference between the proportions of two binomial populations. You can use this to construct a confidence interval or conduct a hypothesis test. As an example, you may want to conduct a hypothesis test based on the proportions of two populations being equal.

If $X \sim B(n_x, p_x)$ and $Y \sim B(n_y, p_y)$ where X and Y are independent, then the expectation and variance of the distribution $P_x - P_y$ are given by

As before, $E(P_x - P_y) = E(P_x) - E(P_y)$

$$E(P_x - P_y) = p_x - p_y$$

$$Var(P_x - P_y) = \frac{p_x q_x}{n_x} + \frac{p_y q_y}{n_y}$$

$Var(P_x - P_y) = Var(P_x) + Var(P_y)$

If np and nq are both greater than 5 for each population, then $P_x - P_y$ can be approximated with a normal distribution. In other words

$$P_x - P_y \sim N\left(p_x - p_y, \frac{p_x q_x}{n_x} + \frac{p_y q_y}{n_y}\right)$$

You can use this to find a confidence interval for $P_x - P_y$. Confidence intervals take the form (statistic) ± (margin of error), so in this case the confidence interval is given by

$$p_x - p_y \pm c\sqrt{Var(P_x - P_y)}$$

This is your confidence interval for $P_x - P_y$

The value of c depends on the level of confidence you need for your confidence interval. They're the same values of c as on the opposite page.

If you're taking a statistics exam where you have to use the sampling distribution between two means or two proportions, the variance of the sampling distribution will be given to you.

This means that you don't have to memorize them; you just need to know how to apply them.

#10. E(X) and Var(X) for continuous probability distributions

When we found the expectation and variance of ***discrete*** probability distributions, we used the equations

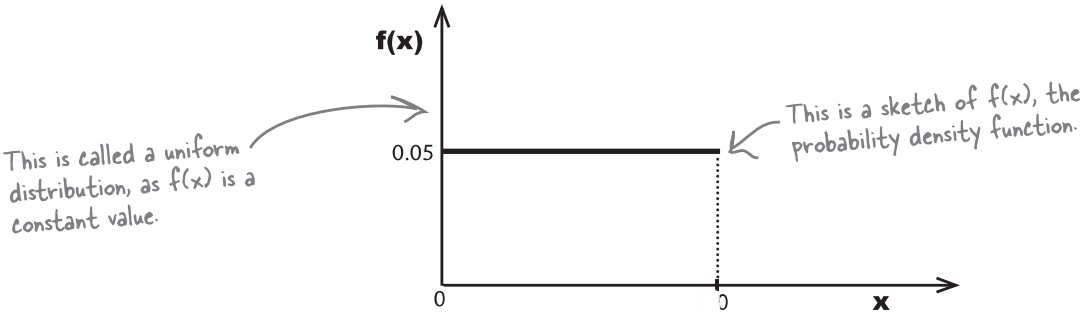
$$E(X) = \Sigma xP(X = x)$$

$$Var(X) = \Sigma x^2P(X = x) - E^2(X)$$

When your probability distribution is ***continuous***, you find the expectation and variance using area.

As an example, suppose you have a continuous probability distribution where the probability density function is given by

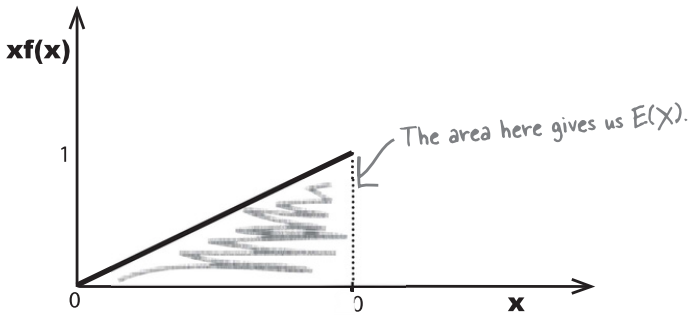
$$f(x) = 0.05 \qquad 0 \leq x \leq 20$$



Finding E(X)

To find the expectation, we'd need to find the area under the curve xf(x) for the range of the probability distribution. Here we need to find the area under the line 0.05x where x is between 0 and 20

Multiply x by the original function to get xf(x).

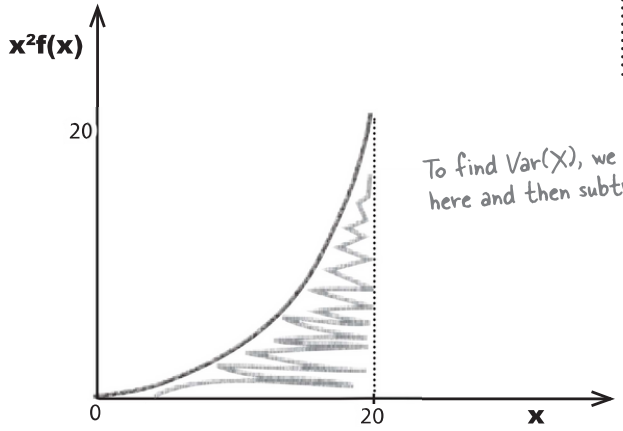


You don't often need to find the expectation and variance of a continuous random variable.

A lot of the time you'll be working with distributions like the normal, and in this, case the expectation and variance are given to you.

Finding Var(X)

To find the variance, you need to find the area under the curve $x^2f(x)$ and subtract $E^2(X)$. In other words, we need to find the area under the curve $0.05x^2$ between 0 and 20 and subtract the square of E(X).



To find Var(X), we find the area here and then subtract $E^2(X)$.

In general, you can find the expectation and variance of a continuous random variable using

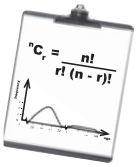
$$E(X) = \int xf(x)dx$$

$$Var(X) = \int x^2f(x)dx - E^2(X)$$

Finding the expectation and variance of a continuous random variable often involves using calculus.

[Note from Marketing: Can we put in a plug for Head First Calculus—coming soon]

over the entire range of x.



Vital Statistics

Uniform Distribution

If X follows a uniform distribution then

$f(x) = 1/(b - a)$ where $a \le x \le b$

$E(X) = (a + b)/2$

$Var(X) = (b - a)^2/12$