ST101 Unit 1: Visualizing Relationships in Data

Contents

Welcome

Welcome to Statistics 101. Let’s start with a teaser. This is a challenging teaser, and it may provoke you.

I believe you should be unhappy. Not because our class is bad, because I will prove in a moment that you are unpopular. The reason we’re doing this is to show how deep statistics is, and how we can easily fool ourselves.

For simplicity, let's say that there are two types of people, type A and type B. Type A are popular. They have 80 friends. Type B are less popular. They only have 20 friends.

Now, you may now say that I don't know which type you are. We will calculate the expected, or average number of friends. To do this, we assume that half of the people are of type A and the other half are of type B.

Average Friends Quiz

What is the average number of friends you have, if you have a 50% chance of being a type A and a 50% chance of being a type B?

Expected Friend Type Quiz

Now, each of your friends on Facebook or Google+ will have either 80 friends or 20 friends. If you pick one friend at random, what is the chance that you have picked a Type A friend? What is the chance that they are type B? (The two numbers should sum to 1).

Unpopular Quiz

Let’s get back to the real question. How many friends should you expect the friend that you picked to have?

Course Overview

Most of the material we will cover in this class is very basic. It is the first class you would have in college if you're not a statistics major. We’ll teach you how to visualize data, how to summarize it, how to run tests, and even how to find trends. But there are also a few challenging nuggets in there.

The challenges are optional, and they’re clearly marked as optional. You will prove some theorems along the way, and – most importantly – you’ll get the chance to program the things you've learned (using the Python programming language).

Again, programming is optional. You don’t need to have a programming background to do this course. But we would suggest you give it a try. Many people learn the material much better by programming than any other way.

Looking at Data

The basis of statistics is that the world is full of data, and we have to make decisions. Statistics comes to our rescue. It takes data and turns it into information that we can use to make decisions. Whatever field you are in, the chances are that it is driven by data. Statistics is important to know and to understand. It's universal, useful, and Sebastian promises it is fun too!

Valuing Houses

One of the standard problems that people study in statistics has to do with purchasing decisions. Suppose you want to buy a house. There are houses of various sizes, but you really like one particular house. This house has a specific asking price, let's say $92,000.00. You want to know whether this is too much, or perhaps too little?

In statistics, the way to find out is by looking at data. Let's assume there's a database of previous house sales in the same neighbourhood. For simplicity, we'll assume we know two things, the size of the home and the sale price.

Size (ft^2)	Cost ($)
1400	112,000
2400	192,000
1800	144,000
1900	152,000
1300	104,000
1100	88,000

Valuing Houses Quiz

The house you wish to purchase is 1300 square feet in size. How much should you expect to pay?

Valuing Houses Quiz 2

How much should you expect to pay for a house with 1800 square feet?

Valuing Houses Quiz 3

What if the house you want to buy is 2100 square feet?

Valuing Houses Quiz 4

What about 1500 square feet?

Valuing Houses Quiz 5

What is the cost of a home per square foot?

Scatter-Plots

Most Important Part Quiz

What do think is the most important thing that a statistics person does?

- Look at data
- Program computers
- Run statistics
- Eat pizza

Linear Relationship 1 Quiz

Here is another data set:

Size (ft^2)	Cost ($)
1400	98,000
2400	168,000
1800	126,000
1900	133,000
1400	91,000
1100	77,000

Is there a fixed cost per square foot for this data set?

Linear Relationship 2 Quiz

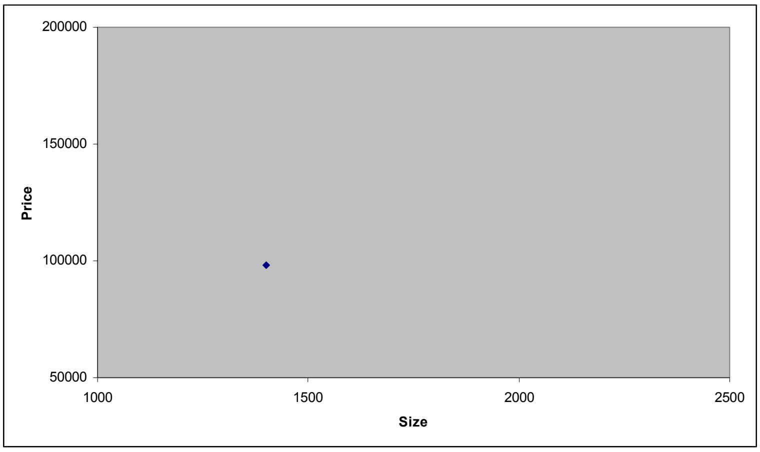
What if I change the second 1400 square foot house to 1300 square feet:

Size (ft^2)	Cost ($)
1400	98,000
2400	168,000
1800	126,000
1900	133,000
1400	91,000
1100	77,000

Is there a fixed cost per square foot for this data set now?

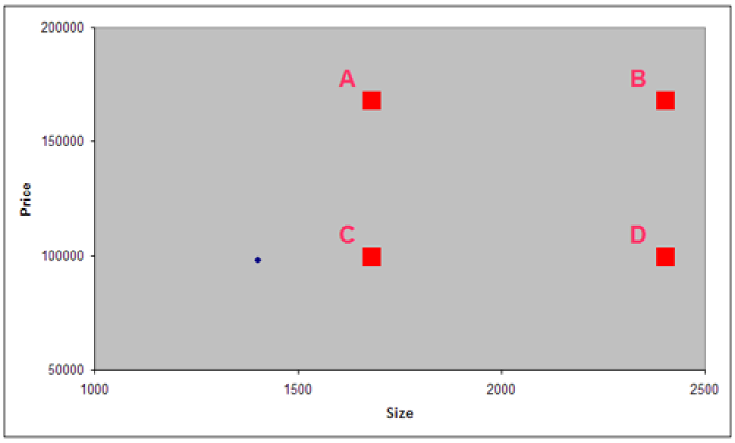
As we have seen, data can carry a lot of information. There is a trick called a scatter-plot that you can use to visualise the data. Take a pencil and a piece of paper and arrange the data in a graph where the x axis is house size, and the y axis is the price.

In a scatter plot, each data item becomes a dot on the graph. The first house would appear on the graph as follows:



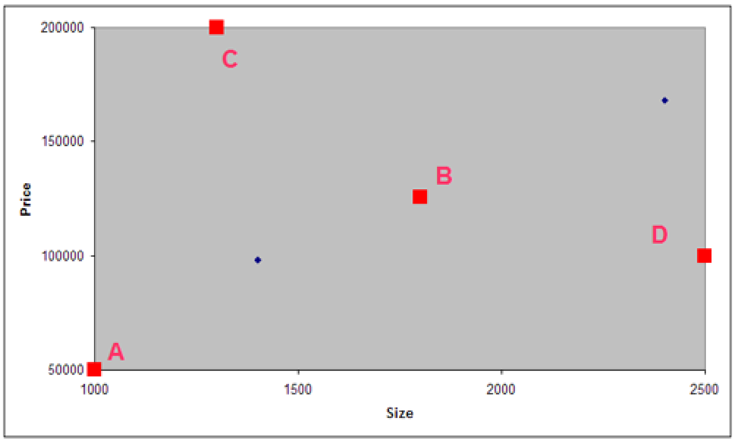
Scatter Plot Quiz

Where would the second house be plotted?

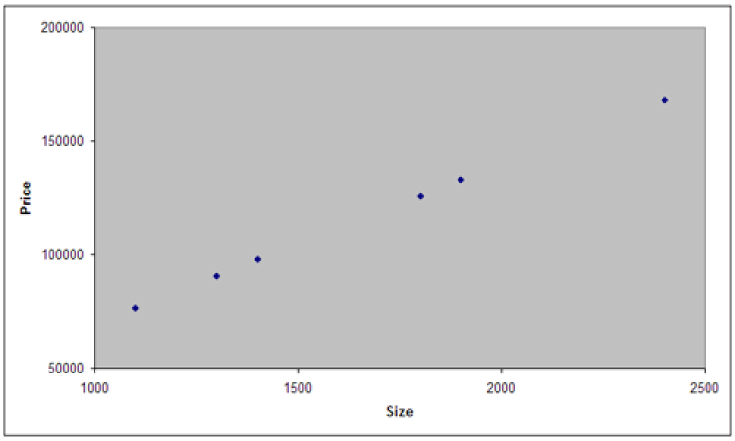


Picking Points Quiz

What about the third house?



Now, we chose a 2-dimensional list to make for a convenient 2-dimensional scatter-plot. These are the most popular scatter plots because surfaces like paper are 2-D. When we add the remaining data points, we get:



This is a nice scatter-plot that allows us to draw a straight line through all the points. When this happens, and there's a relationship between the data that is governed by a straight line we call the data **linear**. Linearity is fairly rare in statistics. More often you will find deviations, because the size of a house is not the only factor that determines its cost (or perhaps also because most of us are bad negotiators!). When a data set is linear, it is really easy to predict the prices of houses in between, just be looking at the data. We're doing what a statistician ought to do.

Make Your Own Quiz

Plot the following data as a scatterplot. Is the relationship between price and size linear?

Size (ft^2)	Cost ($)
1700	51,000
2100	63,000
1900	57,000
1300	39,000
1600	48,000
2200	66,000

Fixed Price Quiz

Do you think there's a fixed price per square foot for this data?

Price Per Square Foot Quiz

Now do you think there’s a fixed price per square foot for this data?

Size (ft^2)	Cost ($)
1700	53,000
2100	65,000
1900	59,000
1300	41,000
1600	50,000
2200	68,000

Make Your Own Quiz 2

Plot the following data. Is the data linear? Can you fit a line through the scatter plot data?

Size (ft^2)	Cost ($)
1700	53,000
2100	65,000
1900	59,000
1300	41,000
1600	50,000
2200	68,000

Find The Constant Quiz

That is a surprising result. Even though there’s no fixed cost per square foot, the relationship between the data is linear. Actually, in this case the price per square foot is linear, plus or minus a constant dollar amount. Can you work out the amounts?

Price = $________ per square foot x size + $________

Is it Linear? Quiz

Plot the modified data below. Is the relationship between the data linear?

Size (ft^2)	Cost ($)
1700	53,000
2100	44,000
1900	59,000
1300	82,000
1600	50,000
2200	68,000

Congratulations

So, now we know a lot about scatter plots. They tend to be 2-dimensional, and a simple eye-ball of the data can tell us a lot the relationship of one variable to another.

Scatter-plots aren’t great when there is what is called "noise" in the data. This happens when the data deviates from expectation in some random, noisy way. Next, we’ll look at another simple plotting technique called bar charts that address the issue of noisy data by grouping data points into a single cumulative bar.

Bar Charts

In this section we’re going to look at bar charts. These are a common statistical data visualization tool.

Checking Linearity Quiz

Let's look at our housing data again. This time the data is ordered by increasing house size. Is this data linear?

Size (ft^2)	Cost ($)
1300	88,000
1400	72,000
1600	94,000
1900	86,000
2100	112,000
2300	98,000

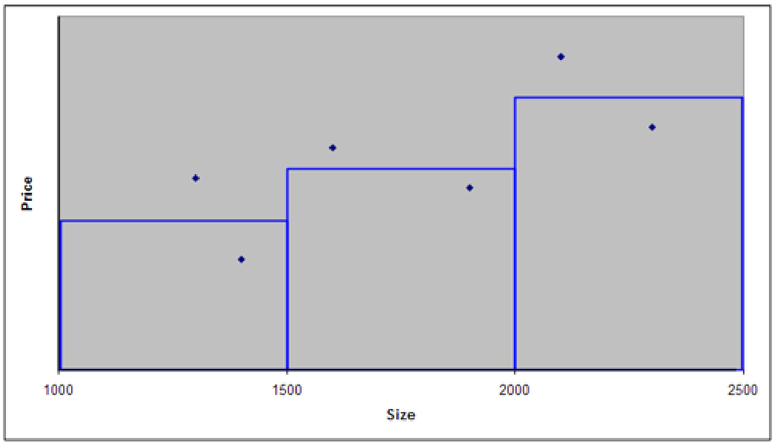
Interpolation Quiz

If we now ask how much you should pay for a 2200 square foot house, using the interpolation method we learned earlier, what figure would you get? Do you trust that number?

There is good reason not to trust this value. The cost of a 2300 square-foot house is less than the 2100 square-foot house. These deviations from the linear relationship are called **noise**. This is the term that statisticians use. Maybe one house has a great view, while another is an old house. Perhaps a third is on the coast, or maybe one needs a new kitchen. There are a whole range of possible factors that effect the cost over and above the size of the property.

If these factors aren’t included in the data, a statistician will call it **random noise**. Bar charts are one way to alleviate the problem.

In a bar chart, we take the raw data and pool it together into ‘bands’. For example, in our house data, we may group all the data for house sizes between 1000 and 1500 square-feet into one bar. Then group the data for house sizes between 1500 and 2000 square-feet into another bar, and so on:



Grouping Data Quiz

What should the height of the first bar (from 1000 to 1500 square-feet) be in dollars?

Grouping Data Quiz 2

What about the dollar values of the heights of the second and third bars?

Bar Charts

When we look at the bar chart, we will see that it is a much finer representation of the data. Pooling multiple data points together to form a single bar, can give a much clearer picture of the dependence of cost on size. While the bar chart doesn’t show the linear relationship in the same way as the scatter-plot (actually, in this case the relationship is non-linear), it really gives a clear sense that, as house size increases, the cost increases. Something that may not have been obvious from just looking at the individual data points.

The bar chart lets us pool groups of data together into single bars and so understand global trends. Now, these global trends might not be that important if you only have six data points, but imagine that you have 60,000 data points. In this case, small variations in individual data points may not tell us much, but the bar chart can really help us to understand the data.

One of the jobs of the statistician is to use cumulative tools, such as bar graphs, to gain an understanding of the underlying data.

Histograms

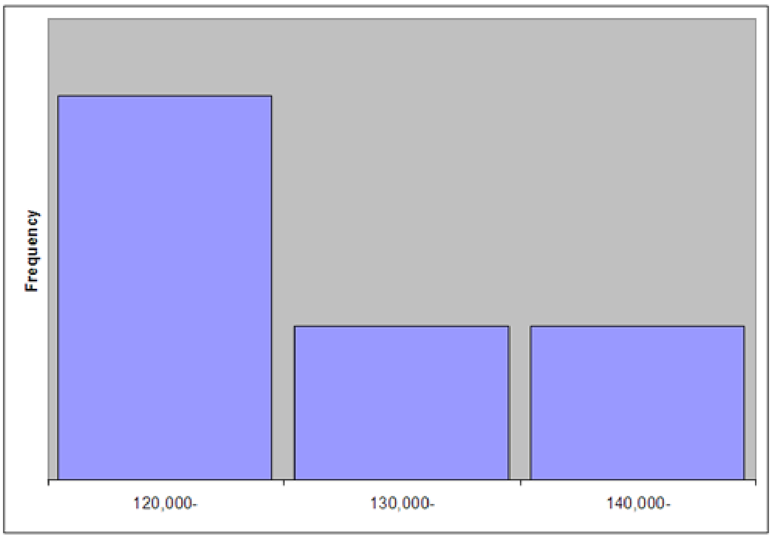
Now, we are going to introduce histograms as a special case of the bar chart.

The key difference is that the bar charts that we have discussed so far have dealt with 2-dimensional data. Histograms only consider 1-dimensional data. Let's consider an example.

Let's suppose that we asked a group of software engineers how much they earn, and got the following responses:

$132,754
$137,192
$122,177
$147,121
$143,000
$126,010
$129,200
$124,312
$128,132

For the histogram, we are going to create a bar chart that is only concerned with frequency. This is basically a count that groups the salaries into a series of buckets, say from $120,000 to $130,000, from $130,000 to $140,000, and over $140,000.

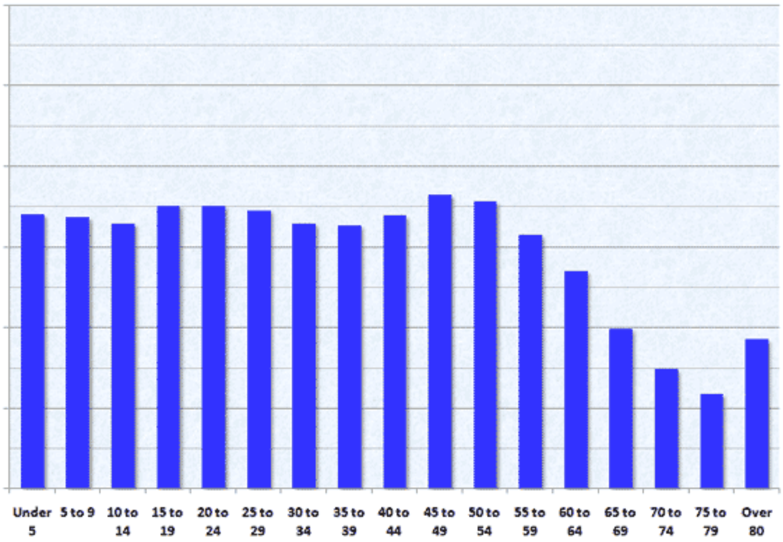


Histogram Quiz

What is the frequency count for the salaries that fall into the three brackets?

Age Distribution Quiz

A well-known histogram can be obtained by looking at age distributions. For the USA, the distribution looks something like this:



Let’s create a rather simplified histogram, looking at people between 0 and 40 using the following data set:

21, 17, 9, 27, 35, 4, 12, 12, 32, 14, 38, 9, 19, 22, 21, 14, 3, 8, 31, 15, 33, 29

Group the data into the ranges:

0 – 10, 11 – 20, 21 – 30, and 31 – 40.

What are the heights of the bars for the four ranges?

Summary

In this unit we learned about bar charts and histograms. Both use vertical bars, and both aggregate data. The big difference is that bar charts are defined over 2-D data, one dimension applied to the x axis and the other to the y axis. Histograms only apply to 1-D data, and the y axis becomes the count of that data.

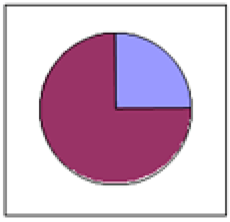
Pie Charts

Most of us have seen pie charts before. In statistics, we use pie charts to visualise data. Specifically, relative data, and we will see what that means in a moment.

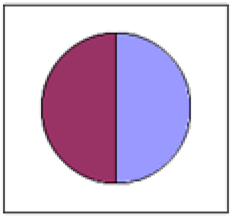
Voting Quiz 1

Let's say that there is an election and there are just two parties. Both parties are getting the same number of votes, i.e. 50%. Which of these pie charts reflect the outcome of the election?

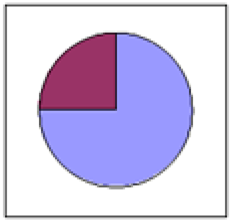
A



B



C



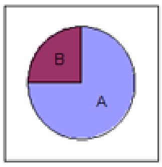
Voting Quiz 2

Now, we said that pie charts are good for relative data. Suppose Party A got 724,000 votes and Party B got 181,000 votes. What percentage of the vote did Part A get?

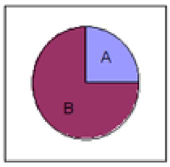
Voting Quiz 3

Now, given this, which of these charts most closely resembles the election result?

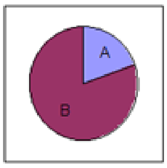
A



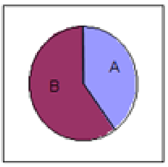
B



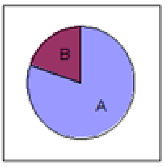
C



D



E



Relative Data Quiz

Now, given that we know that the distribution of the votes was 80% to 20%, if we know that 23,000 people voted for party B, how many people voted for party A?

So, a remarkable property of pie charts is that they are invariant to the actual numbers of votes. What it actually depicts is the ***relative*** numbers of votes. In this case, it show that Party A got many more votes than Party B. It shows this graphically, so you can see this without having to study the actual number of votes cast.

Pick the Breaks Quiz

You are studying a Udacity course. You find the following age distribution among the students on the class:

Age	# Students
13 – 19	12,000
20 – 32	96,000
33 - 999	36,000

Construct the pie chart for this data.

Build a chart Quiz

Let’s build another pie chart. Let’s assume this time that our election ran with four parties. This was the result:

Party A	175,000
Party B	50,000
Party C	25,000
Party D	50,000

Construct the pie chart for this data.

Inferring Counts Quiz

In another election, the distribution of votes between the four parties remained unchanged, but the total number of votes was 240,000. How many votes were cast for each of the parties?

Party A	140,000
Party B	40,000
Party C	20,000
Party D	40,000

Now, the chart tells us nothing about the absolute number of votes cast, but it does tell us a lot about the distribution of the votes. We can easily see that A is the dominant party with more than 50% of the votes cast.

Summary

So we just learned about pie charts. We learned that they are great for relative data, and they're wonderful for comparing which slice of the pie is biggest. We will look at relative data again later with a case study about gender discrimination in college admissions, using a study originally performed at UC Berkeley in California.

Programming Charts (Optional)

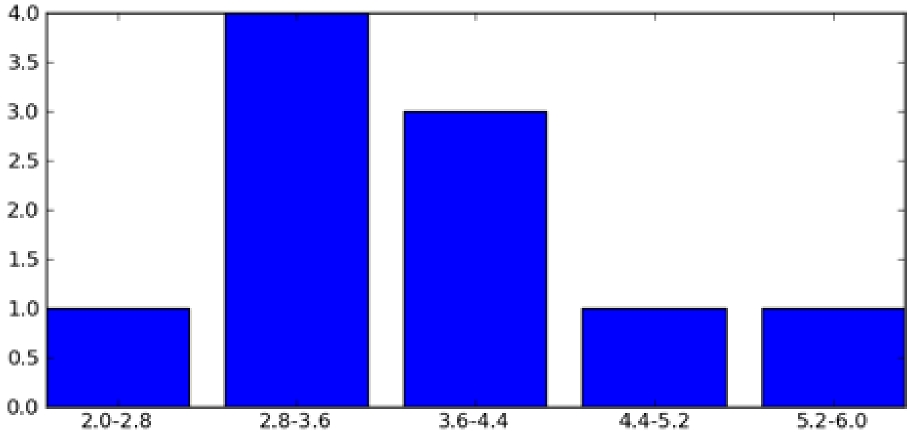
In this section, you get the chance to experience how to visualise data first hand. We will provide the data, and you will plot it and then answer some simple questions about that data.

Here are three lines of code, 3 instructions that the computer has carried out for me:

```
from plotting import *
data = [3, 4, 2, 4, 3, 5, 3, 6, 4, 3]
histplot (data)
```

If you are not a programmer, you can ignore the first instruction. It tells the computer that we want to plot things and it will always be there. The second and third lines are the important ones.

The second line defines a data set. It is a list of 10 elements. The third line tells the computer to make a histogram plot of my data. We then get the result:



As you can see, in this case the range 2.8 to 3.6 was most frequent.

So, there are three things we need to tell the computer. We tell it we want to plot things. We define the data, and give it a name. We tell it the type of plot we want (in this case a histogram).

Plot Height Quiz

Here is the data about the height of a group of people:

```
Height=[65.78, 71.52, 69.4, 68.22, 67.79, 68.7, 69.8, 70.01, 67.9, 66.78,
 66.49, 67.62, 68.3, 67.12, 68.28, 71.09, 66.46, 68.65, 71.23, 67.13, 67.83,
68.88, 63.48, 68.42, 67.63, 67.21, 70.84, 67.49, 66.53, 65.44, 69.52, 65.81,
67.82, 70.6, 71.8, 69.21, 66.8, 67.66, 67.81, 64.05, 68.57, 65.18, 69.66, 67.97,
65.98, 68.67, 66.88, 67.7, 69.82, 69.09]
```

Plot the data as a histogram. What is the most frequent height?

Plot Weight Quiz

Let’s try that again with the weights of the same group of people.

```
Weight=[112.99, 136.49, 153.03, 142.34, 144.3, 123.3, 141.49, 136.46,
112.37, 120.67, 127.45, 114.14, 125.61, 122.46, 116.09, 140.0, 129.5, 142.97,
137.9, 124.04, 141.28, 143.54, 97.9, 129.5, 141.85, 129.72, 142.42, 131.55,
108.33, 113.89, 103.3, 120.75, 125.79, 136.22, 140.1, 128.75, 141.8, 121.23,
131.35, 106.71, 124.36, 124.86, 139.67, 137.37, 106.45, 128.76, 145.68, 116.82,
143.62, 134.93]
```

What is the most frequent weight?

Scatter Plot Quiz

Let’s look at the combined data for height and weight. Create a scatter plot of the height of the individuals against their weight. The command to create a scatter plot is:

```
scatterplot(Height, Weight)
```

Is the data:

- exactly linear
- approximately linear
- height and weight are completely unrelated

Barchart Quiz

Replace the scatter-plot with a bar-chart using the same two arguments as before.

Now the chart show clearly that as the height increases, so does the weight, but the differences between the heights of the bars suggests that the relationship is not exactly linear. In reality, of course we know that the relationship between height and weight in a population isn't linear, but for the sake of this exercise, it is the best of the three options we provided for you.

Wages Quiz

Write a line of code to print a scatterplot of Age on the horizontal axis against Wage on the vertical axis. What is the youngest age at which a person earns $267,000?

Wage Bar-Chart Quiz

Make a bar chart using the same data. Is the relationship between age and wage:

- exactly linear
- approximately linear
- there is no relationship between age and wage.

Most Common Age Quiz

Create a graph to answer the question, what is the most common age?

Conclusion

In this section we created the Python code to generate our own bar charts, histograms, and scatter plots. We can use this to study the data, and perhaps learn something about it.

Admissions Case Study

Statistics is not just a superficial field. It can be really deep. The problem that we will examine here derives from an actual study by the University of California Berkeley. They wanted to know whether their admissions procedure was gender biased. Sebastian looked at various admission statistics to understand whether their admission policies had a preference for a certain gender.

The numbers that we will be using are not the same as those from UC Berkeley. This is a simplified version of that problem, but the paradox that it illustrates is the same. It is called "Simpson's Paradox".

Admissions Quiz 1

Among male students, 900 applied for Major A and 450 were admitted. What is the acceptance rate as a percentage?

Admissions Quiz 2

In a second major, Major B, 100 male students applied and 10 were accepted. What is the acceptance rate as a percentage?

Admissions Quiz 3

The same statistic was run for female students. Females applied predominantly for Major B. There were 900 applications for major B, of whom 180 were admitted. Just 100 female students applied for Major A, of whom 80 were admitted.

What is the acceptance rate for Major A as a percentage for the female student population?

Admissions Quiz 4

What is the acceptance rate for Major B as a percentage for the female student population?

Gender bias quiz

So, just looking at these numbers for the two different majors, do we believe – in terms of the acceptance rate – that there is a gender bias? Is it in favour of male or female students?

Superficially, it appears that female students are favoured because for both majors, they have a better admission rate than the corresponding rate for male students. But what happens if we look at the admission statistics independently of the major?

Aggregation Quiz 1

A total of 1000 male students applied and 460 were admitted. What is the acceptance rate for male students across both majors?

Aggregation Quiz 2

Now do the same for female students. A total of 1000 students applied and 260 were admitted.

Gender Bias Revisited Quiz

So, across both majors, do we believe – in terms of the acceptance rate – that there is a gender bias? Is it in favour of male or female students?

Perhaps surprisingly, given our earlier findings, when we look at both majors together, we find that males have a much higher admissions rate than females. This is not made up. The actual numbers we are using may be made up, but this effect was actually observed University of California at Berkley many years ago.

Looking at majors individually, we find that in each major individually the acceptance rate for females trumps that of males, and yet when we look at the overall statistics we find the opposite. We haven't added anything. We just regrouped the data.

This example shows just how ambiguous statistics can be. In choosing how to graph your data, you can have a major impact on what people believe. A famous saying states “I never believe statistics that I didn't doctor myself”.

The key lesson here is that statistics can be deep and are often manipulated. You should always be sceptical of statistics, whether they are your own results or other peoples, and you really need to understand how raw data is turned into decisions or conclusions.

Answers

Average Friends Quiz

Now, I don’t know which type you are, but there’s a 50% chance that you’re Type A, in which case you’ll have 80 friends, and a 50% chance that you’re Type B, and you’ll have 20 friends. I can calculate your expected number of friends as:

(80 x 0.5) + (20 x 0.5) = 40 + 10 = 50 friends.

Expected Friend Type Quiz

Because Type A people are so much more popular, your chances of linking to a type A is 0.8. That chance that you picked a Type B friend is therefore 0.2.

Unpopular Quiz

Let’s get back to the real question. How many friends should you expect the friend that you picked to have?

There is an 80% chance that you picked a Type A friend, who will have 80 friends. Similarly, there’s a 20% chance that you picked a Type B friend who has 20 friends. This gives an expected number of friends:

(80 x 0.8) + (20 x 0.2) = 64 + 4 = 68 friends

You would only expect to have 50 friends, so this suggests that you are unpopular!

Valuing Houses Quiz

$104,000

Valuing Houses Quiz 2

$144,000

Valuing Houses Quiz 3

$168,000

Valuing Houses Quiz 4

$120,000

Valuing Houses Quiz 5

$80

Most Important Part Quiz

- **Look at data**
- Program computers
- Run statistics
- Eat pizza

Linear Relationship 1 Quiz

No. Two houses of the same size sold for different prices

Linear Relationship 2 Quiz

Yes. The cost per square feet is 70 dollars.

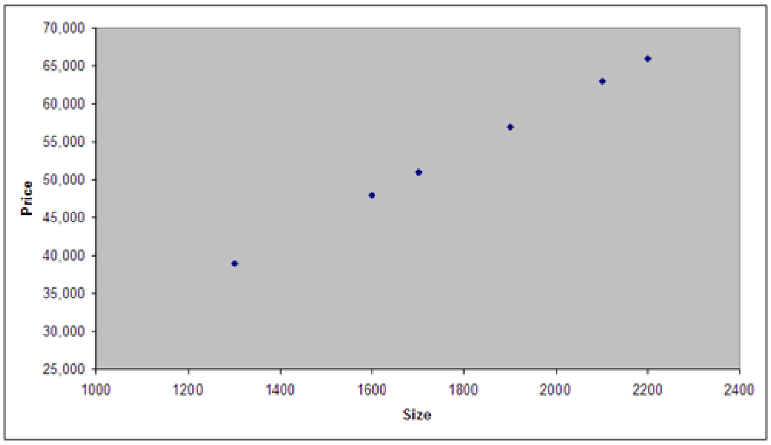
Scatter Plot Quiz

B

Picking Points Quiz

B

Make Your Own Quiz



Fixed Price Quiz

No

Price Per Square Foot Quiz

No. They're almost the same but not quite.

Make Your Own Quiz 2

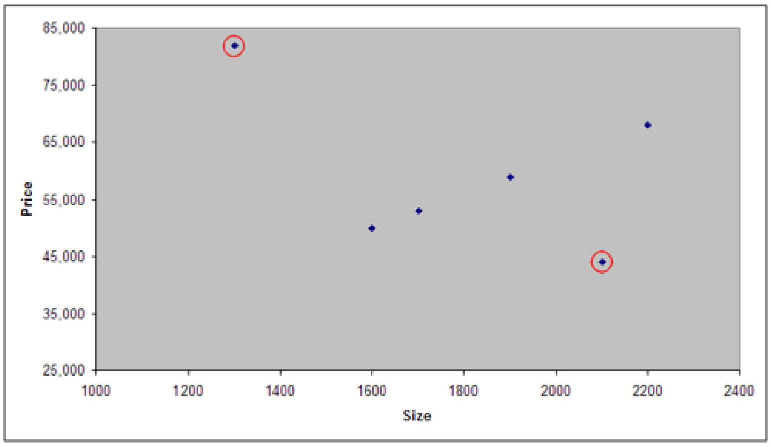
Yes. The data is linear.

Find The Constant Quiz

Price = \$30 per square foot x size + \$2000

Is it Linear? Quiz

No. The two highlighted values are '**outliers**'. We'll talk more about outliers later. There is no way to fit a linear function through all these data points.



Checking Linearity Quiz

No

Interpolation Quiz

\$105,000

No, the value cannot be trusted.

Grouping Data Quiz

$80,000. The average (mean) of $88,000 and $72,000.

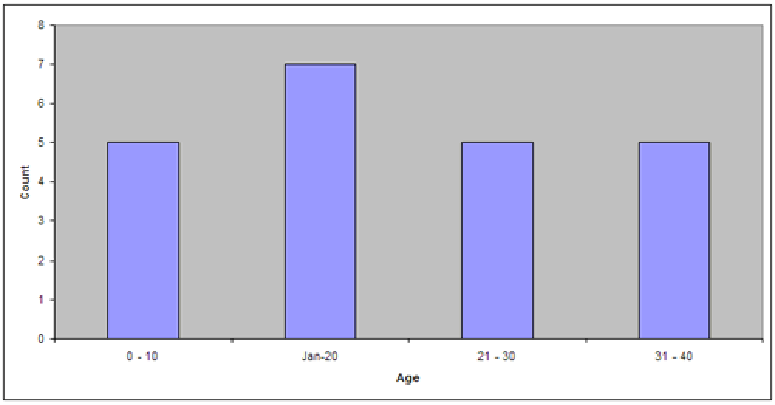
Grouping Data Quiz 2

$90,000 for the second bar and $105,000 for the third.

Histogram Quiz

$120,000 to $130,000	**5**
$130,000 to $140,000	**2**
$140,000 to $150,000	**2**

Age Distribution Quiz



Voting Quiz 1

B

Voting Quiz 2

80%

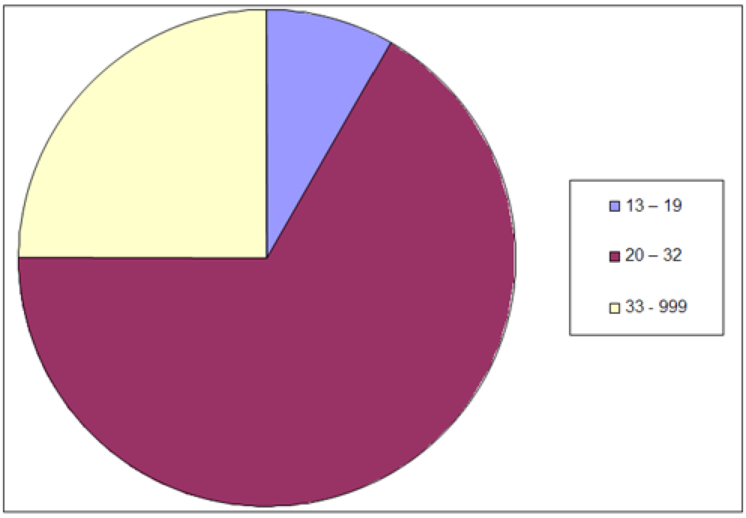
Voting Quiz 3

E

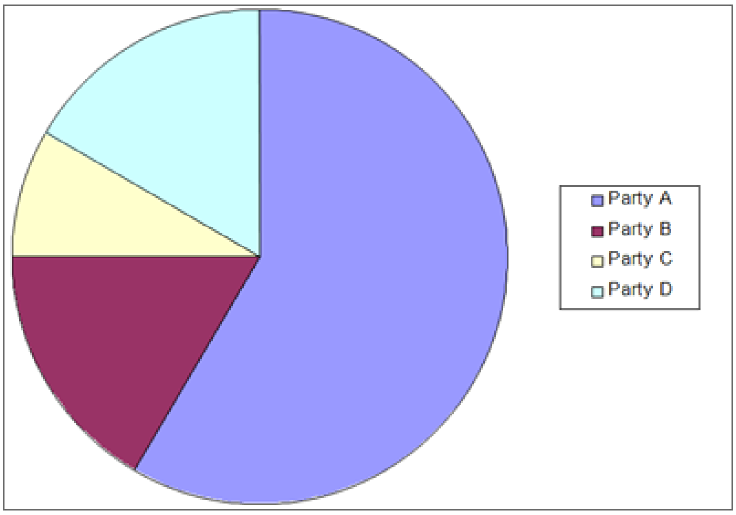
Relative Data Quiz

92000

Pick the Breaks Quiz



Build a chart Quiz



Inferring Counts Quiz

Party A	140,000
Party B	40,000
Party C	20,000
Party D	40,000

Plot Height Quiz

histplot (Height)

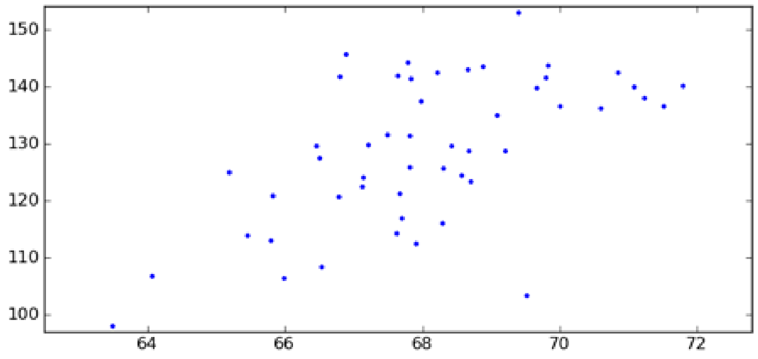
66 – 68 inches

Plot Weight Quiz

histplot (Weight)

119 – 130 lbs

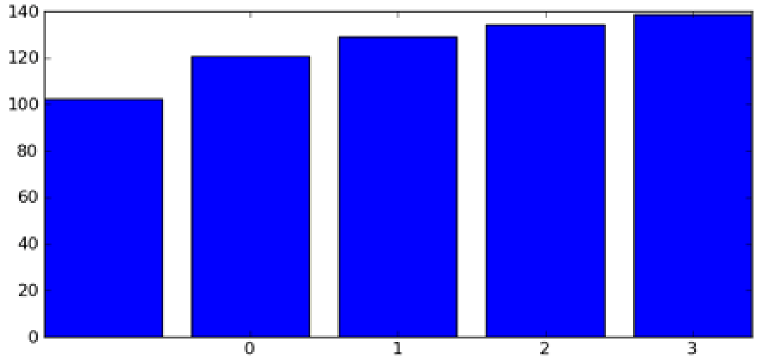
Scatter Plot Quiz



Approximately linear.

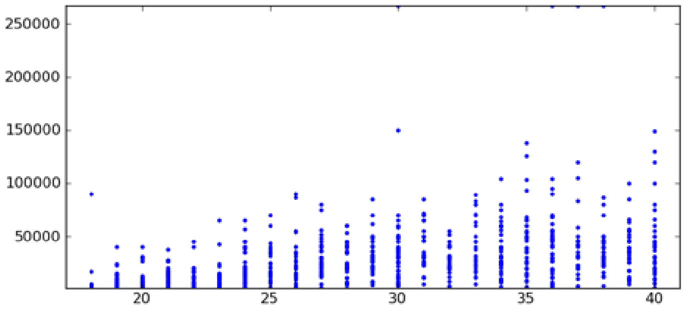
Barchart Quiz

barchart (Height, Weight)



Wages Quiz

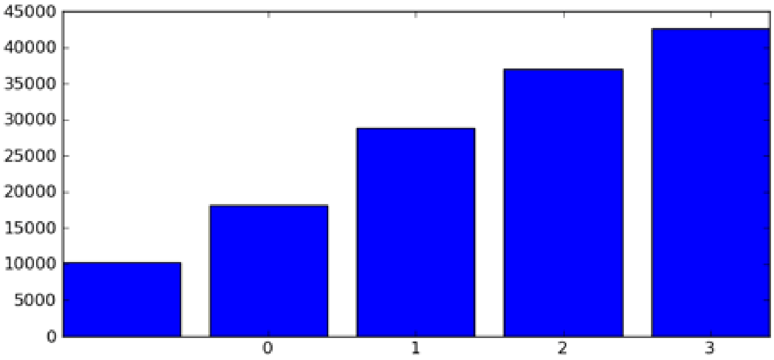
scatterplot (Age, Wage)



30.

Wage Bar-Chart Quiz

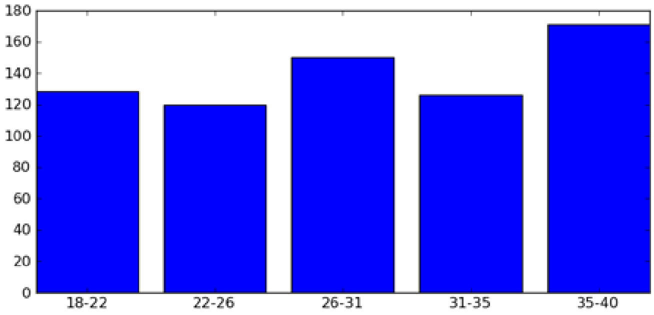
barchart (Age, Wage)



Approximately linear.

Most Common Age Quiz

histplot (Age)



35 - 40

Admissions Quiz 1

50%

Admissions Quiz 2

10%

Admissions Quiz 3

80%

Admissions Quiz 4

20%

Gender bias quiz

Yes, in favour of female students.

Aggregation Quiz 1

46%

Aggregation Quiz 2

26%

Gender Bias Revisited Quiz

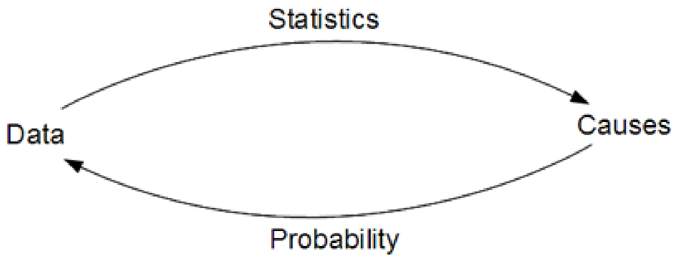
Yes, in favour of male students.

ST101 Unit 2: Probabilities in Data

Contents

Probability

In this unit we will be talking about probability. In a sense, probability is just the opposite of statistics. Put differently, in statistics we are given data and try to infer possible causes, whereas in probability we are given a description of the causes and we try to predict the data.



The reason that we are studying probability rather than statistics is that it will give us a language to describe the relationship between data and the underlying causes.

Flipping coins

Flipping a coin creates data. Each flip of the coin will result in either a 'head' or a 'tail' result. A fair coin is one that has a 50% chance of coming up heads and a 50% chance of coming up tails. Probability is a method for describing the anticipated outcomes of these coin flips.

Fair Coin

The probability of a coin coming up heads is written using the notation:

P(Heads) =

In a fair coin, the chance of a coin flip coming up heads is 50%. In probability, this is given as a probability of 0.5:

P(Heads) = 0.5

A probability of 1 means that the outcome will always happen. A probability of 0 means that it will never happen. In a fair coin, the probability that a flip will come up tails is:

P(Tails) = 0.5

The sum of the probabilities of all possible outcomes is always 1. So:

P(Heads) + P(Tails) = 1

Loaded Coin

A loaded coin is one that comes up with one outcome much more frequently than the other.

Loaded Coin Quiz

Suppose the probability of heads is 0.75 for a particular coin. What is the probability than the coin-flip will come up tails?

Complementary Outcomes

If we know the probability of an outcome, A, then the probability of the opposite outcome, ¬A ("not A"), is given by:

$$P(\neg A) = 1 - P(A)$$

This is a very basic law of probability.

Two Flips

So what happens when we flip the same, unbiased, coin twice? What is the probability of getting two heads in a row, assuming that P(H) = 0.5?

We can derive the answer to this type of problem using a **truth table**. A truth table enumerates every possible outcome of the experiment. In this case:

Flip-1	Flip-2	Probability
H	H	0.25
H	T	0.25
T	H	0.25
T	T	0.25

In the case of two coin flips, there are four possible outcomes, and because heads and tails are equally likely, each of the four outcomes is equally likely. Since the total probability must equal 1, the probability of each outcome is 0.25.

Another way to consider this is that the probability that we will see a head, followed by another head is the product of the probabilities of the two events:

$$P(H, H) \ = \ P(H) \text{ x } P(H) \ = \ 0.5 \text{ x } 0.5 \ = \ 0.25$$

So what happens if the coin is loaded?

Well, if the probability of getting a head, P(H), is 0.6, then the probability that we will see a tail, P(T), is going to be 0.4, and the truth table will be:

Flip-1	Flip-2	Probability
H	H	0.6 x 0.6 = 0.36
H	T	0.6 x 0.4 = 0.24
T	H	0.4 x 0.6 = 0.24
T	T	0.4 x 0.4 = 0.16

Notice that the total probability is still 1:

$$0.36 + 0.24 + 0.24 + 0.16 = 1$$

The truth table lists all possible outcomes, so the sum of the probabilities will always be 1.

Two Flips Quiz

Suppose the probability of heads, P(H) = 1. What is the probability of seeing two heads on successive flips, P(H, H)?

One Head

The truth table can get more interesting when we ask different questions. Suppose we flip the coin twice, but what we care about is that *exactly* one of the two flips reveals a head. For a fair coin, where P(H) = 0.5, the probability is:

$$P(\text{Exactly one H}) = 0.5$$

We can see from the truth table that there are exactly two possible outcomes with exactly one head:

Flip-1	Flip-2	Probability
H	H	0.25
H	T	0.25
T	H	0.25
T	T	0.25

The probability of these outcomes is 0.25 + 0.25 = 0.5.

One of Three Quiz

Suppose we take a fair coin where P(H) = 0.5, and we flip is three times. What is the probability that exactly one of those flips will be a head?

One of Three Quiz 2

What about if the coin is loaded with P(H) = 0.6. What is the probability that exactly one flip out of three will be a head?

Even Roll Quiz

Say you have a fair 6-sided die. The probability of each number appearing on any given throw of the die is 1/6.

What is the probability that a throw will be even?

Doubles Quiz

Suppose we throw a fair die twice. What is the probability that we throw the same number on each throw (i.e. a "double")?

Summary

In this section we learned that if we know the probability of an event, P(A) the probability of the opposite event is just 1 – P(A).

We also learned about composite events where the probability is given by:

$$P(A) \text{ x } P(A) \text{ x } \ldots \text{ x } P(A)$$

Now technically, these conditional events imply **independence**. This just means that the outcome of the second coin flip does not depend on the outcome of the first. In the next section we will look at dependence.

Conditional Probability

In real life, things depend on each other. For example, people can be born smart or dumb. For simplicity, let's assume that whether they're born smart or dumb is just nature's equivalent of the flip of a coin.

Now, whether they become a Stanford professor is not entirely independent of their intelligence. In general, becoming a Stanford professor is not very likely. The probability may only be 0.0001, but it also depends on their intelligence. If they are born smart, the probability may be higher.

In the previous section, subsequent events like coin tosses were independent of what had happened before. We are now going to look at some more interesting cases where the outcome of the first event does have an impact on the probability of the outcome of the second.

Cancer Example

Let's suppose that there is a patient who may be suffering from cancer. Let's say that the probability of a person getting this cancer is 0.1:

$$P(\text{Cancer}) = 0.1$$

$$P(\neg\text{Cancer}) = 0.9$$

Now, we don't know whether the person actually has cancer, but there is a blood test that we can give. The outcome of the test may be positive, or it may be negative, but like any good test, it tells us something about the thing we really care about – in this case whether or not the person has cancer.

Let's say that the probability of a positive test when a person has cancer is 0.9:

$$P(\text{Positive} \mid \text{Cancer}) = 0.9$$

and, $P(\text{Negative} \mid \text{Cancer}) = 0.1$

The sum of the possible test outcomes will always e equal to 1.

This is called the **sensitivity** of the test. Now, this notation says that the result of the test depends on whether or not the person has cancer. This is known as a **conditional probability**.

In order to fully specify the test, we also need to specify the probability of a positive test in the case of a person who doesn't have cancer. In this case, we will say that this is 0.2:

$$P(\text{Positive} \mid \neg\,\text{Cancer}) = 0.2$$

$$P(\text{Negative} \mid \neg\,\text{Cancer}) = 0.8$$

This is the **specificity** of the test. We now have all the information we need to derive the truth table:

Cancer	Test	P( )
Y	Positive	0.1 x 0.9 = 0.09
Y	Negative	0.1 x 0.1 = 0.01
N	Positive	0.9 x 0.2 = 0.18
N	Negative	0.9 x 0.8 = 0.72
		Σ = 1.0

We can now use the truth table to find the probability that we will see a positive tets result

$$P(\text{Positive}) = 0.09 + 0.18 = 0.27$$

Total Probability

Let’s put this into mathematical notation. We were given the probability of having cancer, $P(C)$, from which we were able to derive the probability of not having cancer:

$$P(\neg C) = 1 - P(C)$$

We also had the two conditional probabilities, $P(+ | C)$ and $P(+ | \neg C)$, and from these we were able to derive the probabilities of a negative test:

$$P(- | C) = 1 - P(+ | C)$$

and $P(- | \neg C) = 1 - P(+ | \neg C)$

Then, the probability of a positive test result was:

$$P(+) = P(C).P(+ | C) + P(\neg C).P(+ | \neg C)$$

This is known as **total probability**. Let’s consider another example.

Two Coins

Image that we have a bag containing two coins. We know that coin1 is fair, and coin2 is loaded, so that:

$$P_1(H) = 0.5 \text{ and } P_1(T) = 0.5$$

$$P_2(H) = 0.9 \text{ and } P_2(T) = 0.1$$

We now pick a coin from the bag. Each coin has an equal probability of being picked from the bag. We flip the coin once.

Two Coins Quiz 1

What is the probability that the coin comes up heads?

Two Coins Quiz 2

Let's say that we now flip the coin twice. What is the probability that we will see a head first followed by a tail?

Two Coins Quiz 3

Now the bag contains two new coins. both are loaded:

$$P(H \mid 1) = 1$$

$$P(H \mid 2) = 0.6$$

The probability of picking coin 1 is still 0.5:

$$P(1) = 0.5$$

What is the probability of flipping the coin twice and seeing two tails?

Bayes Rule

In this section, we introduce what may be the Holy Grail of probabilistic inference. It's called Bayes' Rule. The rule is based on work by Reverent Thomas Bayes who used the principle to infer the existence of God. In doing so, he created a new family of methods that have vastly influenced artificial intelligence and statistics.

Let's think about the cancer example from the previous section. Say that there is a specific cancer that occurs in 1% of the population. There is a test for this cancer that has a 90% chance of a positive result if the person has cancer. The specificity of the test is 90%, i.e. there is a 90% chance of a negative test result if the person doesn't have cancer:

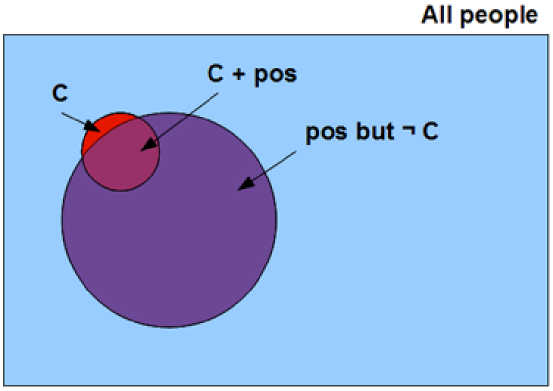
$$P(C) = 0.01$$

$$P(+ \mid C) = 0.9$$

$$P(- \mid \neg C) = 0.9$$

So, here is the question. What is the probability that a person has cancer, given that they have had a positive test?

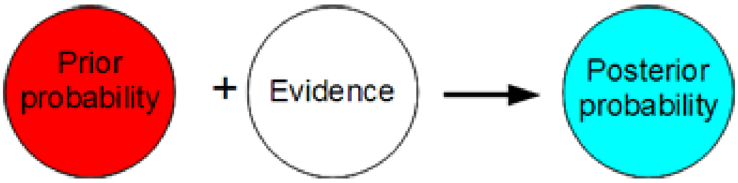
Let's show the figures on a diagram:



Only 1% of the people have this cancer. 99% are cancer-free. The test for this cancer catches 90% of those who have the cancer, which is 90% of the cancer circle. But the test can also give a positive result even when the person doesn't have cancer. In our case a false-positive can occur in 10% of cases – *which is 10% of the total population*. The remaining area represents the case of people who don't have the cancer and get a negative result from the test.

In fact, the area **C + pos** in the diagram above is actually about 8.3% of the total area representing a positive test result. So a positive test has only raised the probability that the person has cancer by a factor of about 8.

So this is the basis of Bayes' Rule. We start with some prior probability before we run the test, and then we get some evidence from the test itself, which leads us to what is known as a posterior probability:



In our example, we have the prior probability, P(C), and we obtain the posterior probabilities as follow. First we calculate what are known as the **joint probabilities**:

$$P(C \mid pos) = P(C) \,.\, P(Pos \mid C)$$

$$P(\neg C \mid pos) = P(\neg C) \,.\, P(Pos \mid \neg C)$$

Given the values in our example we get:

$$P(C \mid pos) = 0.01 \times 0.9 = 0.009$$

$$P(\neg C \mid pos) = 0.99 \times 0.1 = 0.099$$

These values are *non-normalised* – they do not sum to 1. In terms of our diagram above, they are the absolute areas of the regions representing a positive test.

We obtain the posterior probabilities by normalising the joint probabilities. To do this, we divide each of the joint probabilities by the probability of a positive test result:

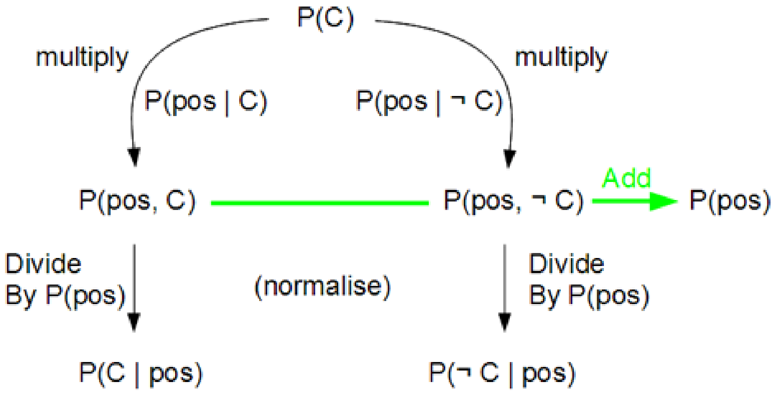
$$P(\text{pos}) = P(C \mid \text{pos}) + P(\neg C \mid \text{pos})$$

So the posterior probabilities are:

$$P(C \mid pos) = \frac{P(C).P(pos \mid C)}{P(pos)}$$

$$P(\neg C \mid pos) = \frac{P(\neg C).P(pos \mid \neg C)}{P(pos)}$$

We can represent the process of calculating Bayes' Rule in a diagram:



Let's say we get a positive test. We have a prior probability, and a test with a given sensitivity and specificity:

Prior: P(C)

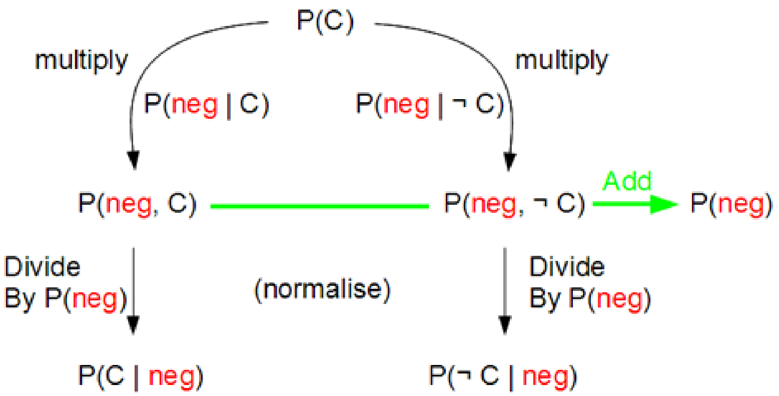
Sensitivity: P(pos | C)

Specificity: P(pos | ¬ C)

We multiply the prior by the sensitivity and by the specificity. This gives us a number that combines the cancer hypothesis with the test result for each of the cases – cancer or non-cancer. We add these numbers (normally, they do not add up to 1), to get the total probability of a positive test.

Now all we need to do to obtain the posterior probabilities is to normalise the two numbers by dividing by the total probability, P(pos).

This is our algorithm for Bayes Rule, and we can produce an almost exactly similar diagram for a negative test result as shown below:



Let’s work through an example. We start with our prior probability, sensitivity and specificity:

$$P(C) = 0.001$$

$$P(pos \mid C) = 0.9$$

$$P(neg \mid \neg C) = 0.9$$

Cancer Probabilities Quiz

Calculate the probabilities of

- $P(\neg C)$
- $P(neg \mid C)$
- $P(pos \mid \neg C)$

Probability Given Test Quiz

Assume that the test comes back negative. Calculate

- $P(C \mid neg)$ #the combined probability of having cancer given the negative test result
- $P(\neg C \mid neg)$ #the combined probability of being cancer-free given the negative test result

Normaliser Quiz

Calculate the normaliser, P(neg)

Normalising Probability Quiz

What is the posterior probability of cancer, given that we had a negative test result?

What is remarkable about the result is what the posterior probabilities actually mean. Before the test, we had a 1% chance of having cancer. After a negative test result this has gone down by about a factor of 9.

Conversely, before the test there was a 99% chance that we were cancer-free. That number has now gone up to 99.89%, greatly increasing our confidence that we are cancer free.

Let's consider another example. In this case the prior probability, sensitivity and specificity are:

$$P(C) = 0.1$$

$$P(pos \mid C) = 0.9$$

$$P(neg \mid \neg C) = 0.5$$

So the sensitivity is high, but the specificity is much lower.

Disease Test Quiz 1

What are the values of the probabilities:

- $P(\neg C)$
- $P(neg \mid C)$
- $P(pos \mid \neg C)$

Disease Test Quiz 2

What are the values of the probabilities:

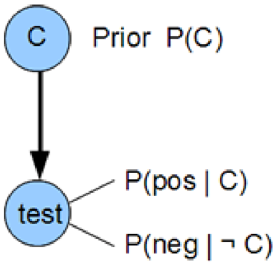
- $P(C, neg)$
- $P(\neg C, neg)$
- $P(neg)$

Disease Test Quiz 3

What are the values of the probabilities:

- P(C | neg)
- P(¬ C | neg)

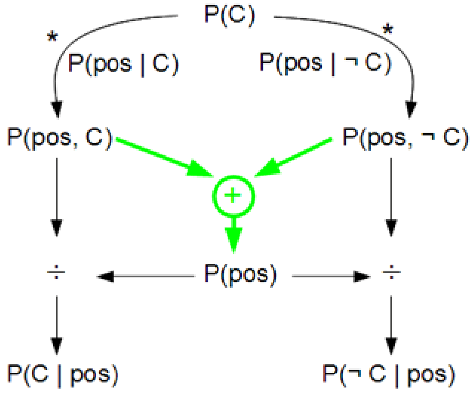
Bayes Rules Summary



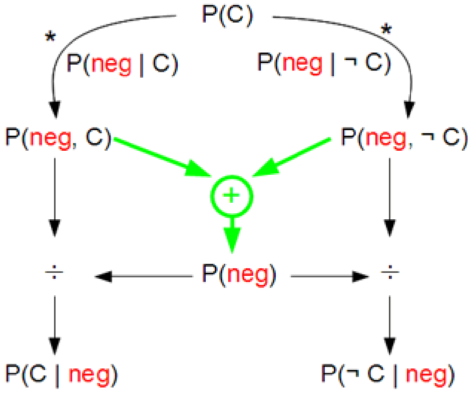
In Bayes Rule, we have a hidden variable that we care about, but can't measure directly. Instead we have a test. We have a prior probability of how often the variable is true. The test characterised by how often it gives a positive result when the variable is true (sensitivity), and how often it gives a negative result when the variable is false (specificity).

Bayes rule then applies the algorithm we saw earlier to calculate the posterior probabilities for the variable given a test outcome:

Positive test:



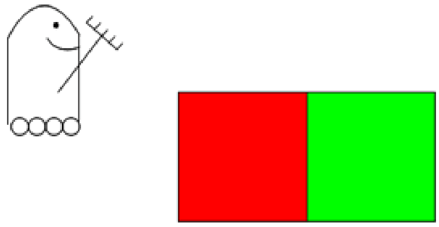
Negative Test:



Robot Sensing

Let’s practice using Bayes’ Rule with a different example.

Consider a robot living in a world that has exactly two places. There is a red place, R, and a green place, G:



Initially, the robot has no idea of it’s location, so the prior probabilities are:

$$P(R) = P(G) = 0.5$$

The robot has sensors that allow it to ‘see’ its environment, but these sensors are somewhat unreliable:

$$P(\text{see R} \mid \text{in R}) = 0.8$$

$$p(\text{see G} \mid \text{in G}) = 0.8$$

Robot Sensing Quiz 1

Suppose the robot sees red. What is the posterior probability that the robot is in the red cell? What is the posterior probability that it is in the green cell?

Robot Sensing Quiz 2

Suppose the prior probabilities are now:

$$P(R) = 0$$

$$P(G) = 1$$

Once again the robot sees red. What is the posterior probability that the robot is in the red cell? What is the posterior probability that it is in the green cell?

Robot Sensing Quiz 3

Given the probabilities:

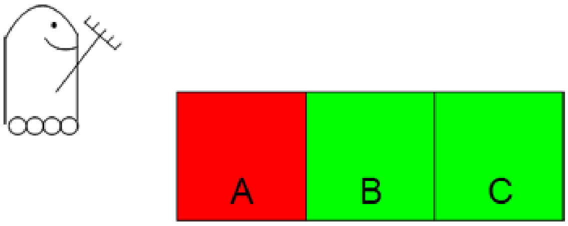
$$P(R) = P(G) = 0.5$$

$$P(\text{see } R \mid \text{in } R) = 0.8$$

$$p(\text{see } G \mid \text{in } G) = 0.5$$

Now the robot sees red. Calculate the posterior probability that the robot is in the red cell, and the posterior probability that it is in the green cell.

Let’s make things a little more complicated. Suppose that there are now three places in the robot’s world, one red and two green. For simplicity, we will label these A, B and C:



So the hidden variable now has three states. We will assume that each place has the same prior probability:

$$P(A) = P(B) = P(C) = 1/3$$

The robot sees red, and we know that:

$$P(R \mid \mathbf{A}) = 0.9$$

$$P(G \mid \mathbf{B}) = 0.9$$

$$P(G \mid \mathbf{C}) = 0.9$$

We can solve for the posterior probabilities exactly as before.

$$P(A, R) = P(A) \text{ x } P(R \mid A) = 1/3 \text{ x } 0.9 = 0.3$$

$$P(B, R) = P(B) \text{ x } P(R \mid B) = 1/3 \text{ x } 0.1 = 0.0333$$

$$P(C, R) = P(C) \text{ x } P(R \mid C) = 1/3 \text{ x } 0.1 = 0.0333$$

So, the normaliser is

$$P(R) = 0.3 + 0.0333 + 0.0333 = 0.3667$$

Which gives the posterior probabilities:

P(A | R) = 0.82

P(B | R) = 0.09

P(C | R) = 0.09

Generalising

The last example showed that there may be more than two states of the hidden variable that we are interested in. There may be 3, 4, 5 or any other number. We can solve these cases using exactly the same methods, but we have to keep track of more values.

In fact, there can be more than just two outcomes of the test. For example, the robot may see red, green or blue. This means that our measurement probabilities will be more elaborate, but the actual method for calculating the posterior probabilities will remain the same.

We can now deal with very large problems, that have many possible hidden causes, by applying Bayes' Rule to determine the posterior probabilities.

Sebastian at Home Quiz

Sebastian has a problem. He travels a lot. It has got so bad that he sometimes wakes up in bed not knowing what country he is in. He is only at home 40% of the time

P(away) = 0.6

P(home) = 0.4

In the summer, Sebastian lives in California. Typically, it doesn't rain in California in the summer. Whereas, there is a much higher chance of rain in many of the countries that Sebastian travels to:

P(rain | home) = 0.01

P(rain | away) = 0.3

Let's say that he wakes up and hears it raining. What is the probability that he is at home in California?

Programming Bayes' Rule (optional)

This section is optional.

To print a number in Python, we just use the print statement thus:

```
print 0.3
```

will cause 0.3 to be printed by the Python interpreter.

A **function** or procedure in Python takes some inputs and produces outputs. A function lets us use the same code to operate on different data by passing that data as the input to the function. We define a Python function as follows:

```
def <Name>(<Parameters>):
   <Block>
```

For example, the following simple function, **f**, just returns whatever parameter **p** it is given:

```
def f(p):
   return p

print f(0.3)
```

The print statement just prints the output of the function, given the input 0.3. This has exactly the same effect as before, but in this case, the print statement isn't printing 0.3 directly, but rather it is printing the output of the function .

Let's say that we are going to provide the probability of an event (P = 0.3) to our function, and we want the function to return the probability of the inverse event. To do this we just modify the function as follows:

```
def f(p):
   return 1 - p

print f(0.3)
```

This will print the result 0.7. If we change the input value, the interpreter will print a different output.

Two Flips Quiz

Suppose we have a coin with P(H) = p. Write a function that returns the probability of seeing heads twice, i.e. P(H, H)

Three Flips Quiz

Let’s say that we now flip the coin three times. Write a function that returns the probability of seeing heads exactly once.

Flip Two Coins Quiz

We now have two coins. Coin 1 has a probability P(H) = p1, and coin 2 has a probability P(H) = p2. Our function will now need to take two arguments as inputs:

```
def f(p1, p2):
```

Write a function that returns the probability of that both coins come up heads.

Flip One of Two Quiz

We have two coins. Coin 1 has a probability P(H) = p1, and coin 2 has a probability P(H) = p2. We pick one coin from a bag. The probability that we pick coin 1 is P_0, and the probability that we pick coin 2 is $1 - P_0$:

$$P(C_1) = P_0$$

$$P(C_2) = 1 - P_0$$

What is the probability that we get heads when we flip the coin?

Write a function with three input arguments, that calculates the probability that we get heads when we flip the coin.

Cancer Example Quiz

Let’s return to our cancer example. We have our prior probability, P_0, the probability of a positive test, given cancer, P_1, and there’s the probability of a negative test for not-cancer which we’ll call P_2.

P© = P_0

$$P(pos \mid C) = P_1$$

$$P(neg \mid \neg C) = P_2$$

What is the formula to calculate probability of a positive test P(pos)?

Calculate Total Quiz

Write a function to calculate the probability of a positive test.

Program Bayes' Rule Quiz

What is the formula to calculate the posterior probability of having cancer following a positive test (using the variable defined above)?

Write a function to calculate the posterior probability of having cancer following a positive test.

Program Bayes' Rule Quiz 2

Write a function to calculate the posterior probability of having cancer following a *negative* test, P(C | neg).

Correlation vs. Causation

In the last unit, we described Simpson's Paradox and showed how it was surprisingly easy to draw false conclusions from data. In this section we will try to give you an insight into a common mistake that is made when interpreting statistical data as a result of confusing **correlation** with **causation**. Newspaper articles frequently confuse correlation with causation.

We will show an example where data is correlated, and show why it is tempting to confuse correlation with causation.

Suppose that you are sick. In fact, you are so sick that you fear that you may die. Fortunately, you're not sick enough so that you can't apply the lessons of this class in order to make a rational decision about whether to go to the hospital.

You consult the data, and find that at your local hospital 40 people were hospitalised, and 4 of these people died. You also find that the majority of the population of your town, 8000 people, didn't visit the hospital, and 20 of these people died at home.

So, 10% of those who were admitted to hospital died, and 0.25% of those who were at home died. This means that the chances of dying in hospital are 40 times greater than the chances of dying at home. There is therefore a ***correlation*** between whether or not you die, and whether or not you are in hospital.

But this doesn't mean that hospitals cause sick people to die.

The statement:

> “The chances of dying in hospital are 40 times greater than dying at home”

shows that there is a ***correlation*** between whether or not you die, and whether or not you are in hospital. Whereas, the statement:

> “Being in a hospital increases your probability of dying by a factor of 40”

is a causal statement. It says that being in hospital causes you to die, not just that being in hospital coincides with the fact that you die. People frequently get this wrong. They observe a correlation, but they suggest that the correlation is causal.

To understand why this could be wrong, let’s look a little deeper into our example.

Considering Health

Let’s say that of the people in the hospital, 36 were sick, and 4 of these died. Four of the people in the hospital were actually healthy, and they all survived.

Of the people who were at home, 40 were actually sick, and 20 of these people died. The remaining 7960 were healthy, but 20 of these people also died (perhaps due to accidents etc.).

These statistics are consistent with our earlier statistics. We have just added another variable - whether a person is sick or healthy.

The percentages of people who died are tabulated below:

	In Hospital	Died	
Sick	36	4	11.11%
Healthy	4	0	0%

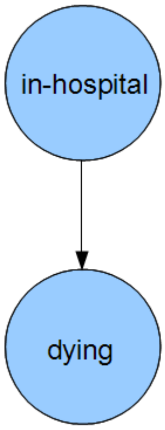
	At Home	Died	
Sick	40	20	50%
Healthy	7960	20	0.2513%

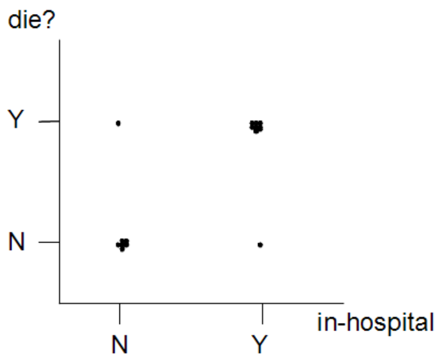
Now, if you are sick, your chances of dying at home are 50% compared with about 11% in the hospital, so you should really make your way to the hospital.

Correlation

So why does the hospital example lead us to draw such a wrong conclusion?

We looked at two variables, being in hospital and the chance of dying, and we rightfully observed that these two things are correlated. If we had a scatter-plot with two categories – whether a person was in hospital, & whether or not that person died - we would see increased occurrence of data points as shown below:





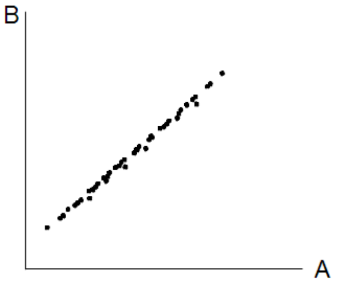
This shows that the data correlates.

So what is correlation? Well, in any plot, data is correlated if knowledge about on variable tells us something about the other.

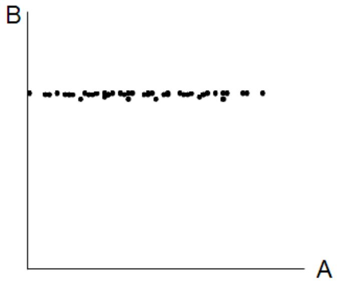
Correlation Quiz

Are the following data pots correlated?

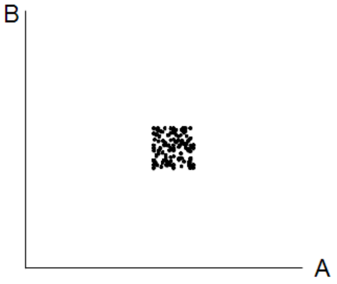
A



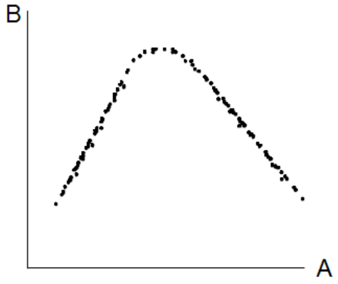
B



C



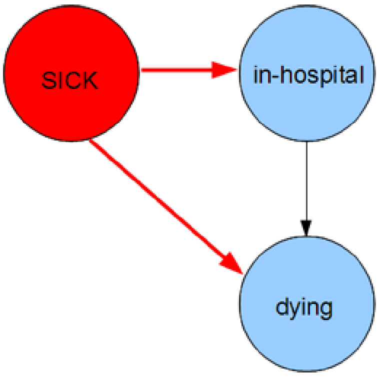
D



Causation Structure

In the example above, there is clearly a correlation between whether a person is in hospital, and whether or not they die. But we initially left out an important variable: whether or not a person was sick.

In fact, it was the sickness that caused people to die. If we add arcs of causation to our diagram, we find that sickness causes death, and that sickness also causes people to go into hospital:



In fact, in our example, once a person knew that they were sick, being in a hospital *negatively correlated* with them dying. That is, being in a hospital made it less likely that they would die, given that they were sick.

In statistics, we call this a confounding variable. It can be very tempting to just omit this from your data, but if you do, you might find correlations that have absolutely nothing to do with causation.

Fire Correlation

Suppose we study a number of fires. We recorded the number of fire-fighters and the surface area (size) of the fire.

# Fire-fighters	# Size fire
10	100
40	400
200	2000
70	700

Clearly, the number of fire-fighters is correlated with the size of the fire. But fire-fighters don't cause the fires! Getting rid of all the fire-fighters will not get rid of all the fires. This is actually a case of **reverse causation**. The size of the fire determines the number of fire-fighters that will be sent to deal with it.

However, it is impossible to know this just from the data. We only know that this is a case of reverse causation because we already know something about fire and fire-fighters.

Assignment

Check out old news articles in newspapers, or online, and find some that takes data which shows a correlation, and from the data suggests causation, or tells you what to do based on that data.

You will find that the news is full of such abuses of statistics.

Answers

Loaded Coin Quiz

P(Tails) = 1 – P(Heads) = 0.25

Two Flips Quiz

P(H, H) = 1

One of Three Quiz

0.375

Flip-1	Flip-2	Flip-3	Probability
H	H	H	0.125
H	H	T	0.125
H	T	H	0.125
H	T	T	**0.125**
T	H	H	0.125
T	**H**	T	**0.125**
T	T	**H**	**0.125**
T	T	T	0.125

One of Three Quiz 2

P(H) = 0.6
P(T) = 0.4

Flip-1	Flip-2	Flip-3	Probability
H	H	H	0.216
H	H	T	0.144
H	T	H	0.144
H	T	T	**0.096**
T	H	H	0.144
T	**H**	T	**0.096**
T	T	**H**	**0.096**
T	T	T	0.064

P(exactly 3 Heads) = 0.288

Even Roll Quiz

Three possible outcomes are even (2, 4, 6), so the probability that a throw is even is: 3 x 1/6 = 0.5

Doubles Quiz

The truth table has 36 possible outcomes. Each outcome in the truth table will have a probability of 1/36. Six of these outcomes are “doubles”, so the probability of a double is:

6 x 1/36 = 1/6 =0.16667

Two Coins Quiz 1

Pick	Flip	P( )
1	H	0.25
1	T	0.25
2	H	0.45
2	T	0.05

So, P(H) = 0.25 + 0.45 = 0.7

Two Coins Quiz 2

Pick	Flip1	Flip2	P()
1	H	H	0.5 x 0.5 x 0.5 = 0.125
1	H	T	0.5 x 0.5 x 0.5 = 0.125
1	T	H	0.5 x 0.5 x 0.5 = 0.125
1	T	T	0.5 x 0.5 x 0.5 = 0.125
2	H	H	0.5 x 0.9 x 0.9 = 0.405
2	H	T	0.5 x 0.9 x 0.1 = 0.045
2	T	H	0.5 x 0.1 x 0.9 = 0.045
2	T	T	0.5 x 0.1 x 0.1 = 0.005

P(H, T) = 0.125 + 0.045 = 0.17

Two Coins Quiz 3

Pick	Flip1	Flip2	P()
1	H	H	0
1	H	T	0
1	T	H	0
1	T	T	0
2	H	H	0.5 x 0.6 x 0.6 = 0.18
2	H	T	0.5 x 0.6 x 0.4 = 0.12
2	T	H	0.5 x 0.4 x 0.6 = 0.12
2	T	T	0.5 x 0.4 x 0.4 = 0.08

P(T, T) = 0.08

Cancer Probabilities Quiz

- P(¬ C) = 0.99
- P(neg | C) = 0.1
- P(pos | ¬ C) = 0.1

Probability Given Test Quiz

Assume that the test comes back negative. Calculate

- P(C | neg) = 0.01 x 0.1 = 0.001
- P(¬ C | neg) = 0.99 x 0.9 = 0.891

Normaliser Quiz

p(neg) = 0.892

Normalising Probability Quiz

P(C | neg) = 0.0011
P(¬ C | neg) = 0.9989

Disease Test Quiz 1

- P(¬ C) = 0.9
- P(neg | C) = 0.1
- P(pos | ¬ C) = 0.5

Disease Test Quiz 2

- P(C, neg) = 0.1 x 0.1 = 0.01
- P(¬ C, neg) = 0.9 x 0.5 = 0.45
- P(neg) = 0.46

Disease Test Quiz 3

- P(C | neg) = 0.0217
- P(¬ C | neg) = 0.9783

Robot Sensing Quiz 1

P(at R | see R) = 0.8
P(at G | see R) = 0.2

Robot Sensing Quiz 2

P(at R | see R) = 0
P(at G | see R) = 1

Robot Sensing Quiz 3

P(at R | see R) = 0.615
P(at G | see R) = 0.385

Sebastian at Home Quiz

P(rain) = P(home) x P(rain | home) + P(away) x P(rain | away)
P(rain) = 0.4 x 0.01 + 0.6 x 0.3 = 0.184

P(home | rain) = P(home) x P(rain | home) / P(rain)
P(home | rain) = 0.4 x 0.01 / 0.184) = **0.0217**

Two Flips Quiz

```
def f(p):
   return (p * p)
```

Three Flips Quiz

```
def f(p):
   return 3 * p * (1-p) * (1-p)
```

Flip Two Coins Quiz

```
def f(p1,p2):
   return p1 * p2
```

Flip One of Two Quiz

$P(H) = (P_0 \text{ x } P1) + ((1 - P_0) \text{ x } P2)$

```
def f(p0,p1,p2):
   return (p0 * p1) + ((1 - p0) * p2)
```

Cancer Example Quiz

$P(pos) = (P_1 \text{ x } P_0) + ((1 - P_2) \text{ x } (1 - P_0))$

Calculate Total Quiz

```
def f(p0,p1,p2):
   return (p0 * p1) + ((1 - p0) * (1 - p2))
```

Program Bayes' Rule Quiz

$P(C \mid pos) = P_0 \text{ x } P_1 / ((P_0 \text{ x } P_1) + ((1 - P_0) \text{ x } (1 - P_2))$

```
def f(p0,p1,p2):
   return ((p0 * p1)/((p0 * p1) + ((1 - p0) * (1 - p2))))
```

Program Bayes' Rule Quiz 2

```
def f(p0,p1,p2):
   return (p0 * (1 - p1))/((p0 * (1 - p1)) + ((1 - p0) * p2))
```

Correlation Quiz

A. Yes
B. No
C. No
D. Yes

ST101 Unit 3: Estimation

Contents

Estimation

This section is all about estimators. We will introduce the Maximum Likelihood Estimator and the Laplacian Estimator. We will see how to use these techniques to derive probabilities from observed data, such as coin flips.

The subtitle for this section might be “To Fake or Not to Fake?”. The answer may surprise you!

Probability of Heads

Suppose we flip a coin six times and get the outcome H-T-T-H-T-H. Based solely on this data, we would say that the probability of heads, P(H), for this coin is the number of heads we observed divided by the number of times we flipped the coin:

$$P(H) = \frac{3}{6} = 0.5$$

Now suppose we take a different coin and flip it five times. We observe H-H-T-H-H. based on this data we would assume that P(H) is:

$$P(H) = \frac{4}{5} = 0.8$$

We call these the empirical frequencies. They are just the ratio of the number of outcomes in which the event occurs to the total number of trials. If we took a coin and flipped it seven times and we saw T-T-T-T-T-T-T we would assume that the probability of heads was:

$$P(H) = \frac{0}{7} = 0$$

Identify the Estimator Quiz

Which of the following formulae captures the method we used to calculate the empirical frequencies above?

A	B
$\sum_i X_i$	$\prod_i X_i$
C	D
$\frac{1}{N}\sum_i X_i$	$\frac{1}{N}\prod_i X_i$

We call this the Maximum Likelihood Estimator, and its value will always be between 0 and 1, which is a valid probability. It is a really good way to estimate the underlying probability that may have produced a given data set.

What happens when we have more than two outcomes.

Let’s say that we have a six-sided die. This therefore has six possible outcomes. We throw the die ten times, and get the following outcomes:

1-6-6-3-2-6-5-4-6-2

Now, N = 10 so:

$$P(1) = \frac{1}{10} \times 1 = 0.1$$

$$P(2) = \frac{1}{10} \times 2 = 0.2$$

$$P(3) = \frac{1}{10} \times 1 = 0.1$$

$$P(4) = \frac{1}{10} \times 1 = 0.1$$

$$P(5) = \frac{1}{10} \times 1 = 0.1$$

$$P(1) = \frac{1}{10} \times 4 = 0.4$$

The sum of the probabilities is 1, which is what we would expect since these are all the possible outcomes.

Likelihood

The estimation problem that we are trying to overcome is, given some data, what is the probability, P, that would give rise to the observed data?

In earlier units, we have seen how to estimate the data that we would observe given the probability of some event occurring. We can record a ‘1’ if the event happens, and a ‘0’ if it does not.

Likelihood Quiz

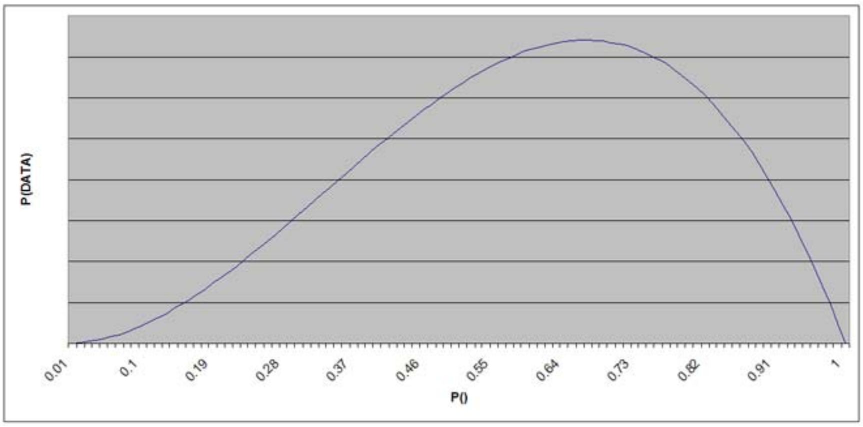
Suppose we see the following data:

1-0-1

For each value of P() what is the value of P(DATA)?

1. P() = ½
2. P() = ⅓
3. P() = ⅔
4. P() = 1

If we plot the probabilities on a graph, with P() on the x-axis and P(DATA) on the y-axis, we get:



The point at P() = 0.75 maximises the likelihood of the data. This is called the Maximum Likelihood Estimator, MLE.

Weakness

Suppose we flip a coin exactly once. It comes up heads. Using the Maximum Likelihood Estimator we would end up with a probability of heads:

$$P(H) = \frac{1}{1} = 1$$

Similarly, if we had seen tails, MLE would have given us:

$$P(H) = \frac{0}{1} = 0$$

Weakness Quiz

Does this mean that, from a single coin flip, the maximum likelihood estimator will always assume a loaded or biased coin?

Weakness Quiz 2

Suppose we make 111 coin flips. Will the maximum likelihood estimator always assume that the coin is loaded or biased?

Faking It

There is a solution. We can fake it!

More precisely, we can add fake data to the original data to create a larger pool of data points. In the case of a coin-flip, there are two possible outcomes – heads or tails – so we would add two fake data points to our data, one heads and one tails.

If we had just one result, we would see the following:

1	MLE $= P(H) = \frac{1}{1} = 1$
1-0-1	MLE $= P(H) = \frac{2}{3} = 0.667$

Fake Data Quiz

Calculate the MLE, with and without an extra fake data-point, for the following sequences of results:

- 1
- 0-0-1
- 1-0-0-1
- 1-1

Fake Data Summary

There are a couple of things that we should note here.

In general, adding the fake data points pulls everything towards 0.5, ‘smoothing’ the estimate. However, the first and last examples in the previous quiz showed that, the more data we get, the more we are willing to move away from 0.5. In the limiting case, as we see infinitely many ‘heads’, P(H) will indeed approach to 1.

In general, adding fake data gives better estimates in practice. The reason for this is that it is really quite reckless to suppose that a coin will always come up heads after just a single coin-flip. It is better to say that we have some evidence that heads may be more likely, but that we’re not yet convinced. The “not quite convinced” is the same as having a prior. These priors are called Dirichlet Priors.

More importantly, the technique of adding fake data to improve the estimator is called a Laplacian Estimator. When we have plenty of data, the Laplacian Estimator gives about the same result as the Maximum Likelihood Estimator, but when data is scarce, the Laplacian Estimator usually works much, much better.

Dice Example

Suppose we roll a die and get the results: 1-2-3-2.

What is the Maximum Likelihood Estimate and the Laplacian Estimate for each of the following?

- P(1)
- P(2)
- P(3)

Summary

In this section we introduced the Maximum Likelihood Estimator, and derived its mathematical formula:

$$MLE = \frac{1}{N}\sum X_i$$

This is a really simple formula, called the **empirical count**.

We also discussed the Laplacian Estimator that added k fake data points, one for each possible outcome, giving us the slightly more complicated formula:

$$LE = \frac{1}{N+k}(1+\sum X_i)$$

In both cases, N is the number of experiments, and k is the number of outcomes.

We the identified cases where the Laplacian Estimator gives much better results than the Maximum Likelihood Estimator. Specifically, cases where there isn't much data.

Averages

In this section, we will teach you about the three Ms in statistics: The 'mean', the 'median', and the 'mode'. These terms are really important to know. They are useful for looking at data, and are often confused. Let's begin with the mean.

Here is a list of house prices:

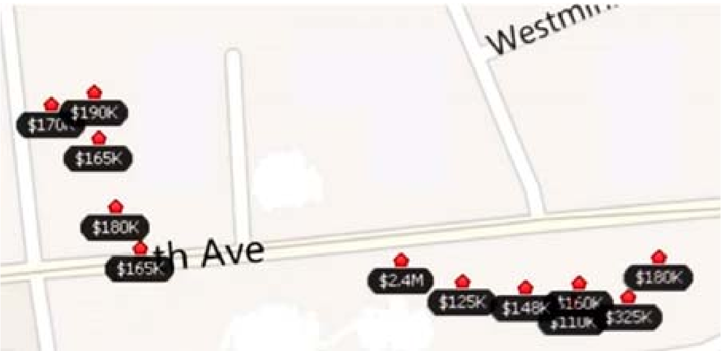
House prices
$190 k
$170 k
$165 k
$180 k
$165 k

The **mean** calculates the average of these prices using the formula we saw in the last section:

$$mean = \frac{1}{N}\sum X_i$$

In this case, the sum of the house prices is $870k, and there are 5 prices, so N = 5 and the mean is $174k. But why is the mean useful?

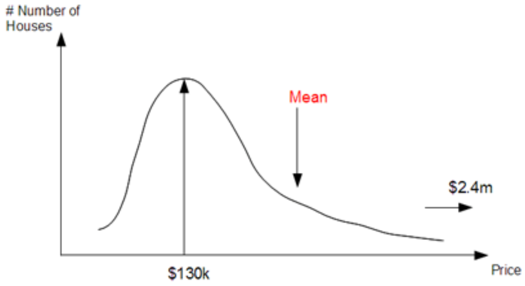
These prices are actually taken from a small area in Pittsburgh, Pennsylvania:



The prices in our area range from $165k to $190k. Our mean value of $174k really characterises this neighbourhood.

In a second neighbourhood on our map, most house prices are a little cheaper - $110k, $125k, $148k, and $160k. However, there are two outliers - $325k and $2.4 million. These outliers distort the mean house price for the second neighbourhood which is $492k.

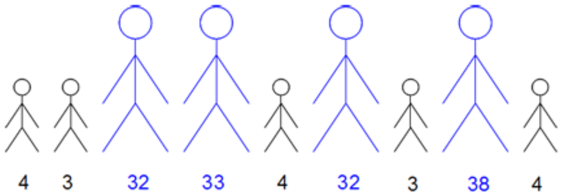
Clearly we ought to be suspicious of this number. Most of the homes are in the $100k range, but the mean doesn't reflect this. It gives the impression that the average house price in the neighbourhood is $492k, but that isn't a good description of reality. If we plotted the prices on a graph, we would get something like:



We have a peak around $130k, but there is a really long tail that goes all the way out to $2.4 million. The effect of this is to drag the mean away from the peak at $130k towards the outliers. It has a really strong impact on the mean.

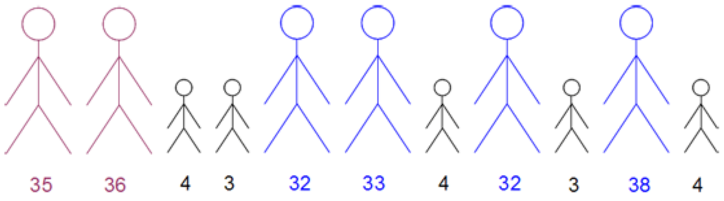
This is where the second M, the '**median**', comes into its own. The median is a different statistic that attempts to find the 'typical' value in a list, by identifying the one in the middle. To find the median, we just sort our list and pick the value in the middle. In this case, the median price is $160k, and if we look back at our map, this is a really good characterisation of house prices in this neighbourhood, and certainly a much better characterisation that that given by the mean value of $492k.

There is another limitation of the mean that may not be overcome by using the median. To illustrate this, we will consider a child's birthday party. This party has a number of children and some of their parents. We will say there are five children and four adults with ages as shown:



The mean age of the group is 17, although clearly this isn't very informative. There aren't any teenagers at this party. This is another case where the mean is giving us misleading statistics.

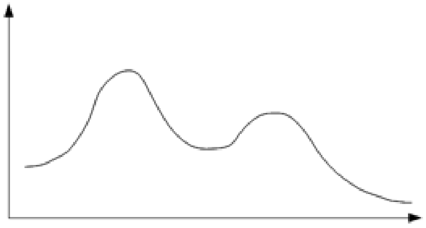
The median age for this group is 4. Is this meaningful? Well, not really. Suppose another two parents aged 35 and 36 turn up:



The median is now 32! A small change to the people at the party shifted the median from 4 to 32.

The '**mode**' is the value that occurs most often in a list. We already met modes when we looked at bar charts and talked about finding the most frequent bar. In this case, because the ages are discrete it is easy to see that the most frequent age at this party is 4, and that it was unchanged by the arrival of the two new parents.

So, the mode is a very useful statistic in cases where the data is **multi-modal**. Multi-modal distributions have curves with more than one peak:



This particular example is actually a **bi-modal** curve since it has exactly two peaks. The mode is the value of the highest peak.

Obviously, although the mean will always be determined precisely, sometimes there may be ties for the median or the mode. If we have an even number of elements, there will be two elements at the centre, and either could be chosen as the median. Equally, two elements could appear in the list the same number of times, and either one could be the mode. In these cases, we just assume that ties are broken at random. We can pick either number.

Three Averages Quiz

Calculate the mean, median and mode for the data:

5, 9, 100, 9, 97, 6, 9, 98, 9

Three Averages Quiz 2

Calculate the mean, median and mode for the data:

3, 9, 3, 8, 2, 9, 1, 9, 2, 4

Variance

This section is all about variance, and its close cousin standard deviation.

Consider a college student who has five close friends and five close family members. Their ages are as follows:

Friends:	17, 19, 18, 17, 19
Family:	7, 38, 4, 23, 18

The mean age of both groups is 18, but the friends are clustered very close to 18 which the family are much more widely distributed. In neither case does the mean, the median or the mode capture the spread of the data.

To capture the spread of the data we need to calculate the **variance**. To do this, we first normalise the data by subtracting the mean from each value:

Friends:	17	19	18	17	19
(normalised)	-1	1	0	-1	1

Family:	7	38	4	23	18
(normalised)	-11	20	-14	5	0

We should note that the mean of both normalised sequences is 0. We can see that the normalised values for the ages of the friends is much closer to zero than those for the family members, which shows that the spread of the family members' ages is much larger.

Now we calculate the squares of the normalised values:

Friends:	17	19	18	17	19
(normalised)	-1	1	0	-1	1
	1	1	0	1	1

Family:	7	38	4	23	18
(normalised)	-11	20	-14	5	0
	121	400	196	25	0

We now calculate the variance by adding the squared values and dividing by the number of values:

$$\text{variance} = \frac{1}{N}\sum(X_i - \mu)^2 \qquad (\mu \text{ is the mean})$$

Variance Quiz

Calculate the variance for both the friends and the family group.

Measuring Spread

The variance is a measure of how far the data is spread. It is really small if all the data is clustered close to the mean, and can be really large if data points occur a long way from the mean.

The variance, by its very nature computes in the quadratic. It is the average quadratic deviation from the mean:

$$\text{variance} = \frac{1}{N}\sum(X_i - \mu)^2$$

We can take the square root of the variance to obtain the standard deviation (σ) which is not a quadratic:

$$\text{standard deviation} = \sigma = \sqrt{\text{var}\,iance}$$

For our group of friends, the standard deviation is $\sqrt{0.8}$ = 0.894, and for the family members it is $\sqrt{148.4}$ = 12.182. The standard deviation provides a measure of the amount by which we might expect the age of an average member of the group to deviate from the mean.

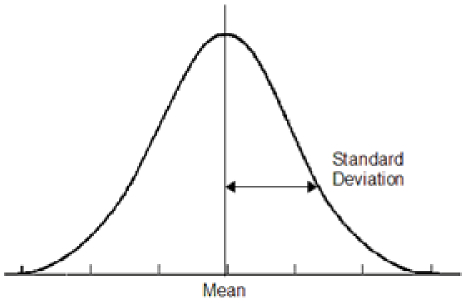
Standard Deviation Quiz

Calculate the mean, the variance, and the standard deviation for the following data sequences:

- 3, 4, 5, 6, 7
- 8, 9, 10, 11, 12
- 15, 20, 25, 30, 35
- 3, 3, 3, 3, 3
- 4

Formulae

Plotting the data onto a graph may give something like this:



The formulae that we used to calculate the values are:

$$\mu = \frac{1}{N}\sum_i X_i$$

$$\sigma^2 = \frac{1}{N}\sum (X_i - \mu)^2$$

$$\sigma = \sqrt{\sigma^2}$$

A problem with these formulae is that they require two passes through the data.

First we have to go through the data to compute the mean. We do this by summing all the data and dividing by the total number of data items. For this we need to maintain two things:

1. the total number of data items, N (which we increment each time we see a new item)
2. the sum of all X$_i$, $\sum_i X_i$

Once we have done this, we have a value for the mean, μ, and now we have to go through and compute the variation for which we need to maintain $\sum (X_i - \mu)^2$.

It turns out that we can use a trick whereby, instead of maintaining $\sum (X_i - \mu)^2$, we can instead maintain $\sum X_i^2$, and so calculate the variation and standard deviation in a single pass.

We have $\sigma^2 = \frac{1}{N}\sum (X_i - \mu)^2$

So, $\sigma^2 = \frac{1}{N}\sum (X_i - \mu)(X_i - \mu) = \frac{1}{N}\sum \left[X_i^2 - 2X_i\mu + \mu^2\right]$

$$\sigma^2 = \frac{1}{N}\sum X_i^2 - \frac{2\mu}{N}\sum X_i + \mu^2$$

Now, $\mu = \frac{1}{N}\sum_i X_i$

so, $\sigma^2 = \frac{1}{N}\sum X_i^2 - 2\mu^2 + \mu^2 \quad = \quad \frac{1}{N}\sum X_i^2 - \mu^2$

$$\sigma^2 = \frac{1}{N}\sum X_i^2 - \frac{1}{N^2}\left(\sum X_i\right)^2$$

Using this formula, the counters or statistics $\sum X_i$, $\sum X_i^2$, and N are all we need to calculate the variance in a single pass through the data.

Alternative Formula Quiz 1

Calculate $\sum X_i$, $\sum X_i^2$, and N for the data sequence:

3, 4, 5, 6, 7

Alternative Formula Quiz 2

If we plug these results into the formulae above, what values do we get for μ and σ^2?

We have covered a lot in this section. We introduced **variance**, which is the spread of the data squared, and then when on to explain **standard deviation**, which is the same, but without the square. We also now have a way to compute the values in a single pass through the data using only these running counters:

$\sum X_i$, $\sum X_i^2$, N

Raise Quiz

Suppose that we are considering giving all our employees a raise. We want to think about what the effect of the raise would be on the mean and standard deviation of the distribution of salaries within the company.

We are considering two types of raises:

1. A fixed amount of $1000
2. A relative raise of 20%

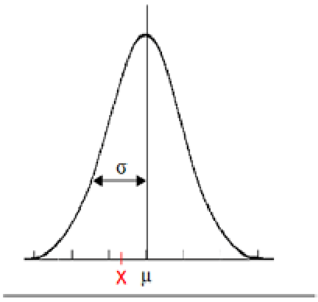
These will change the mean and standard deviation to new values which we will call μ' and σ'. The change will be either multiplicative or additive.

What are the multiplicative or additive factors in each case.

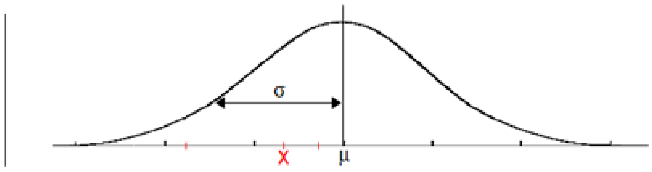
Standard Scores

We now want to introduce the concept of a standard score.

The basic idea is, that for any Gaussian, no matter what the mean and covariance are, we can state how far in or out a point, x, is. Let’s think about an example. Suppose we have the point x on this Gaussian curve:



We can locate the corresponding point on a second Gaussian relative to the mean and the standard deviation, even if that curve is actually much wider with a different mean and standard deviation:



Mathematically, the standard score, Z, is defined as:

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

Standard Score Quiz

We have the data set:

3, 4, 5, 6, 7

The mean for this data set is 5, and the standard deviation is $\sqrt{2}$.

What is the standard score for x = 2, relative to the Gaussian that fits our data set?

Standard Score Quiz

What happens to Z if we change the new data point to x = 5?

Programming Estimators (Optional)

This section is completely optional, but it will provide the opportunity to program the techniques that we have been looking at throughout this unit.

Programming Mean Quiz

Create a Python function to return the mean for a list of data.

Programming Median Quiz

Write a Python function to return the median value of a list of data. Assume all the lists have an odd number of elements.

Programming Mode Quiz

Write a Python function to return the mode for a data set. Assume that if there are multiple modes you may return any one of them.

Programming Variance Quiz

Write a Python function to calculate the variance for a set of floating-point values. You have already written the code to calculate the mean, so you should use it.

Programming Standard Deviation Quiz

For the last programming quiz, we want you to calculate a Python function to calculate the standard deviation for a set of floating-point values.

Answers

Identify the Estimator Quiz

C

Likelihood Quiz

1. P(DATA) = ½ x ½ x ½ = 0.125
2. P(DATA) = ⅓ x ⅔ x ⅓ = 0.074
3. P(DATA) = ⅔ x ⅓ x ⅔ = 0.148
4. P(DATA) = 1 x 0 x 1 = 0

Weakness Quiz

Yes.

Weakness Quiz 2

Yes.

A fair coin always has P(H) = 0.5

For this to be true over N flips, the number of heads must be exactly N/2. This is just not possible for an odd number of observations like 111, so the coin will always look slightly biased.

Fake Data Quiz

Data	**MLE**	**MLE with extra data point**
1	P(H) = 1	P(H) = 0.667
0-0-1	P(H) = 0.333	P(H) = 0.4
1-0-0-1	P(H) = 0.5	P(H) = 0.5
1-1	P(H) = 1	P(H) = 0.75

Dice Example

For the MLE, the data sequence is: 1-2-3-2

For the Laplacian Estimator the data sequence becomes: 1-2-3-2-1-2-3-4-5-6

	MLE	Laplace
P(1)	$\frac{1}{4} = 0.25$	$\frac{2}{10} = 0.2$
P(2)	$\frac{2}{4} = 0.5$	$\frac{3}{10} = 0.3$
P(3)	$\frac{1}{4} = 0.25$	$\frac{2}{10} = 0.2$

Three Averages Quiz

Mean = 38
Median = 9
Mode = 9

Three Averages Quiz 2

Mean = 5
Median = 3 or 4
Mode = 9

Variance Quiz

Friends: variance = 4/5 = 0.8
Family: variance = 742/5 = 148.4

Standard Deviation Quiz

Data	Mean	Variance	Std Deviation
3, 4, 5, 6, 7	5	2	1.414
8, 9, 10, 11, 12	10	2	1.414
15, 20, 25, 30, 35	25	50	7.071
3, 3, 3, 3, 3	3	0	0
4	4	0	0

Alternative Formula Quiz 1

N = 5

$\sum X_i = 25$

$\sum X_i^2 = 135$

Alternative Formula Quiz 2

μ = 5

σ2 = 2

Raise Quiz

Fixed raise of \$1000	Relative raise 20%
μ' = μ + \$1000	μ' = 1.2 x μ
$\sigma^2 = \frac{1}{N}\sum\left(X_i + 1000 - (\mu + 1000)\right)^2$ σ' = σ	$\sigma^2 = \frac{1}{N}\sum(1.2 \times X_i - 1.2 \times \mu)^2$ σ' = 1.2 x σ

Standard Score Quiz

Z = -2.121

Standard Score Quiz 2

Z = 0

Programming Mean Quiz

```
def mean(data):
    return sum(data)/len(data)
```

Programming Median Quiz

```
def median(data):
    sdata = sorted(data)
    return sdata[int(len(data)/2)]
```

Programming Mode Quiz

```
def mode(data):
    mode = 0
    for item in data:
        if data.count(item) > data.count(mode):
            mode = item
    return mode
```

Programming Variance Quiz

```
def mean(data):
    return sum(data)/len(data)

def square(a):
    return a * a

def variance(data):
    ndata = []
    mu = mean(data)
    for i in data:
        ndata.append(square(i - mu))
    return sum(ndata)/len(ndata)
```

Programming Standard Deviation Quiz

```
def mean(data):
    return sum(data)/len(data)

def variance(data):
    mu=mean(data)
    return mean([(x-mu)**2 for x in data])

def stddev(data):
    return sqrt(variance(data))
```

ST101 Unit 4: Outliers and Normal Distribution

Contents:

Outliers

Ignoring Data

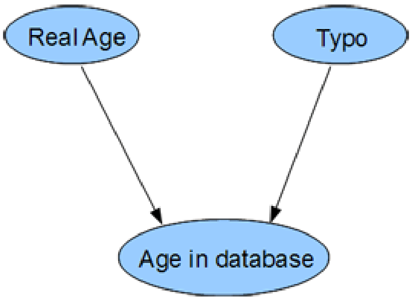
We all know that many politicians and cult leaders happily ignore data when it suits them to do so. But did you know that statisticians also ignore data?

Consider the example of a sports club with the following members:

Name	**Age**
Joseph	22
Maria	21
Susan	24
Marc	20
Tom	211
Jack	23

If we wanted to compute the mean age of this group, should we ignore any of the data?

And in this case the answer is yes. Tom's age is obviously a mistake. It was probably the result of a typo when entering the data onto a computer. The ages in our database can all be explained as either the real ages of members or as typos:



So how do we identify, and so ignore, these outlying data points?

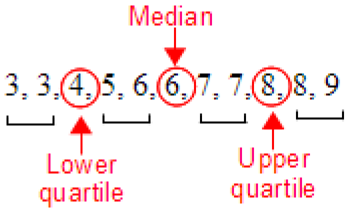
Quartiles

The easiest way to ignore outliers is by using quartiles, or percentiles.

Suppose we have the following dataset of 11 items sorted into numerical order:

3, 3, 4, 5, 6, 6, 7, 7, 8, 8, 9

As the name suggests, quartiles partition the data into four subsets, with a single digit gap between each subset:



The element in the middle of the sorted group is the median, which we met in the last unit. The other two elements are called the lower quartile and the upper quartile. The range between these elements is called the **inter-quartile range**. It is this range that we would use to calculate things like the mean. Data falling outside this range is then ignored.

This is a simple, but often very effective technique for removing outliers. It will remove extreme values that are often attributable to things other than what you are trying to understand.

Now this works well for a dataset with 11 items, and also for a dataset of 15 items, 19 items, 21 items, and so on. Any number of elements that satisfies

$$4N + 3$$

will give a nice symmetrical division into quartiles, since we need four quartiles of N elements, plus the three separating elements. If the dataset contains a slightly different number of values we will end up breaking the symmetry slightly, but this is usually not significant as most data sets are relatively large.

Compute Quartiles Quiz

Here is our age distribution with the frequencies for each age:

Age	19	20	21	22	23	24	25
Frequency	2	1	1	3	2	1	1

What are the lower quartile, the median, and the upper quartile?

What is the “trimmed” mean after removing the outliers (i.e. the mean of the inter-quartile range)?

Compute Quartiles Quiz 2

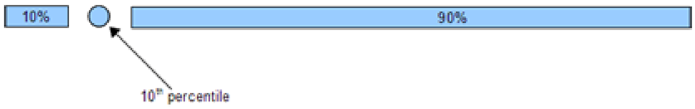
Let’s consider a second example:

33, -99, 17, 13, 1489

What is the mean of this group of numbers? What is the mean after outliers are removed?

Percentiles

Percentiles are similar to quartiles. The k-th percentile is the value that occurs k% of the way through the data. So, for the 10th percentile, we would have:



For our original age data:

Name	Age
Joseph	22
Maria	21
Susan	24
Marc	20
Tom	211
Jack	23

removing the upper 20th-percentile will remove Tom’s age from our dataset.

Binomial Distribution

We have looked at ways to calculate the probabilities of events like coin flips, but what happens when we have a large number of events?

For example, how do we calculate the probability that a coin having P(H) = 0.8 will come up heads nine times out of a total of 12 flips?

Suppose we flip two coins. We know that there are two possible outcomes in which we will see the same number of heads and tails (i.e. exactly one head and one tail):

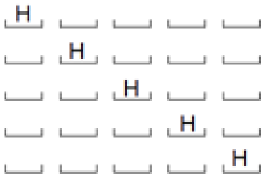
Heads	Heads
Heads	Tails
Tails	Heads
Tails	Tails

If we flip four coins, there are now $2^4 = 16$ possible outcomes, and we will see exactly the same number of heads and tails in six of these outcomes:

Heads	Heads	Heads	Heads
Heads	Heads	Heads	Tails
Heads	Heads	Tails	Heads
Heads	Heads	Tails	Tails
Heads	Tails	Heads	Heads
Heads	Tails	Heads	Tails
Heads	Tails	Tails	Heads
Heads	Tails	Tails	Tails
Tails	Heads	Heads	Heads
Tails	Heads	Heads	Tails
Tails	Heads	Tails	Heads
Tails	Heads	Tails	Tails
Tails	Tails	Heads	Heads
Tails	Tails	Heads	Tails
Tails	Tails	Tails	Heads
Tails	Tails	Tails	Tails

Clearly this can only work for an even number of coin flips. We cannot have the same number of heads and tails if we flip an odd number of coins.

If we were to flip five coins, there are just five possible outcomes where we would see exactly one head (or exactly one tail):



How many outcomes will give us exactly two heads?

Well, the first head could come up in any of the positions shown above, so there are five possibilities for the first head. This leaves four possible positions for the second head, giving us 4 x 5 = 20 outcomes. But this over-counts by exactly a factor of two. The reason is that we have counted the two cases below as different outcomes, depending on which coin came up heads first:

In reality of course, these should be counted as just one outcome, and the number of outcomes where we see exactly two heads is:

$$\frac{5\times 4}{2}=10$$

For three heads, we can work out the number of outcomes in two ways. Firstly, seeing three heads from five coin flips is exactly the same as seeing two tails, so we can use the same calculation we used to establish the number of outcomes having exactly two heads above.

The second way is to recognise that there are five places where the first head can appear, four for the second, and three for the third. But this is over-counting again. In each arrangement of three heads, there are three possible positions where we could have placed the first head, two for the second, and just one for the third. i.e. there are six arrangements, and our original calculation counted each of these as a different outcome, so the actual number of outcomes where we will see exactly three heads out of our five coin flips is:

$$\frac{5\times 4\times 3}{3\times 2\times 1}=10$$

So, following this logic, if we have ten coins, and we want to know the number of outcomes where we will see exactly four heads, we can just use:

$$\frac{10\times 9\times 8\times 7}{4\times 3\times 2\times 1}=\frac{5040}{24}=210$$

Combinatorics Quiz

We flip ten coins. How many possible outcomes have exactly five heads showing?

Formulae

There is a special notation that will help us express these examples in more general terms. This is called factorials. The factorial of any number, n is written as n!, and evaluates as:

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

So, $10! \; = \; 10\times 9\times 8\times 7\times 6\times 5\times 4\times 3\times 2\times 1$

Now, we can actually write:

$$10\times 9\times 8\times 7\times 6=\frac{10!}{5!}$$

So, to calculate how many outcomes show k heads out of n coin flips, our formula becomes:

$$\frac{n!}{k!\cdot(n-k)!}$$

Arrangements Quiz

We flip 125 coins. How many outcomes end with exactly 3 heads showing?

Binomial

What happens when we introduce probabilities? Let's say that we have a fair coin (i.e. P(H)=0.5), and we flip the coin five times. What is the probability that we will get exactly one head?

Well, we know that there are exactly five ways that we can observe exactly one head in five flips:

$$\frac{5!}{1!\cdot(5-1)!}=\frac{5!}{1!\cdot 4!}=5$$

We also know that there are $2^5 = 32$ possible outcomes (this is the size of the truth table).

So the probability of seeing exactly one head when we flip a fair coin five times is:

P(#HEADS = 1) = 5/32 = 0.15625

If we wanted to know the probability that we would see three heads out of five flips of a fair coin, we can calculate it using:

$$\text{P(\#HEADS = 3)} = \frac{\frac{5!}{3!(5-3)!}}{2^5}=\frac{10}{32}=0.3125$$

What if the coin is loaded? Does that make a difference?

Let's say we have a loaded coin where P(H) = 08. We want to calculate the probability that we will see heads exactly once when we flip the coin three times.

We can answer this using a truth table:

Heads	Heads	Heads
Heads	Heads	Tails
Heads	Tails	Heads
Heads	Tails	Tails
Tails	Heads	Heads
Tails	Heads	Tails
Tails	Tails	Heads
Tails	Tails	Tails

Because the coin is loaded, not all outcomes in the truth table are equally likely. For the highlighted rows, where we have exactly one head, and two tails, the probability for that outcome will be:

$$P(H)\times\left(1-P(H)\right)\times\left(1-P(H)\right)=0.8\times0.2\times0.2=0.032$$

There are three outcomes, so the total probability of seeing exactly one head from three flips is:

$$3\times0.032=0.096$$

Binomial Quiz

Let's say that we flip our loaded coin with P(H) = 0.8 five times. What is the probability that we will see exactly four heads?

Binomial Quiz 2

Let's say that we flip our loaded coin with P(H) = 0.8 five times. What is the probability that we will see exactly three heads?

Binomial Quiz 3

Let's say that we flip our loaded coin with P(H) = 0.8 twelve times. What is the probability that we will see exactly nine heads?

Conclusion

So, for any coin having P(H) = p, that we flip n times, the probability of seeing k heads is given by:

$$\text{P(\#HEADS = k)} = \frac{n!}{(n-k)!\cdot k!}\cdot p^{k}\cdot(1-p)^{n-k}$$

This formula gives the probability of what is known as the binomial distribution. This is the cumulative outcome of many identical coin flips.

As you can see, we can take very large experiments, with very large numbers of coin flips, and compute the probability that heads will appear a specified number of times using the relatively simple formula above.

Central Limit Theorem Programming (Optional)

In this section, we will take some steps towards one of the deepest insights in all of statistics. It is called the Central Limit Theorem. The way that we will get to this insight is through a series of programming exercises.

Now, all of the programming on this course is optional, and it is fine to skip this section, but this is probably the most interesting way to understand the central limit theorem and statistics involving large numbers.

Programming Flips

Write a function, flip(N), that simulates flipping a coin 1000 times.

Having done this, compute the mean and the standard deviation of your results.

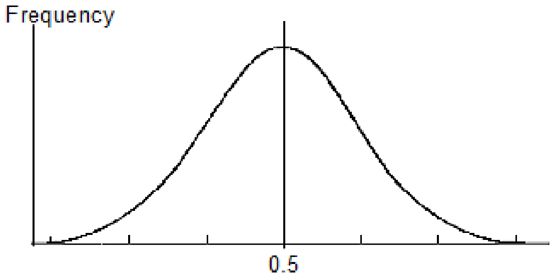
Hint:

The function random.random() gives a random output that sits between 0 and 1.

Sets of Flips

Write a function sample(N), that stores the means of N iterations of the function flip(N) in a list. Print the resulting list as a histogram using 30 bins.

The interesting thing here is that in this binomial distribution the frequency of outcomes seems to be centred around the expected outcome (in this case 0.5), and falls off according to this characteristic curve.



This curve is often known as a **bell curve**.

The significance of these bell curves, and their relationship to the central limit theorem, will be discussed in the next section.

The Normal Distribution

This section describes one of the most transformative things in modern statistics. We will start with the binomial distribution which we met earlier, and then move on to the central limit theorem. This effectively means that we are taking the number of coin flips to infinity. From that, we will arrive at the normal distribution, which is the basis for so much in statistics.

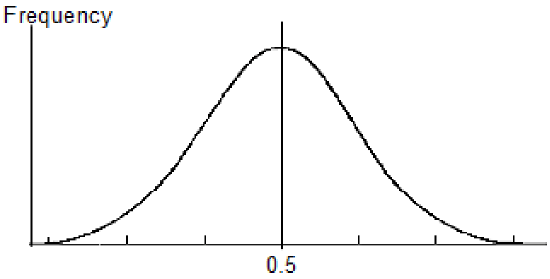
The reason why this is important is that experiments often generate thousands of data points (e.g. the thousands of patients in a drug trial), rather than the one or two coin flips we considered earlier. In order to analyse this amount of data, it is often more practical to start from the normal distribution as an approximation to the binomial distribution.

We start with our well-established formula for binomial distributions:

$$\frac{n!}{(n-k)!\cdot k!}\cdot p^k \cdot (1-p)^{n-k}$$

- n is the number of coin flips
- k is the number of times it comes up heads
- p is P(HEADS)

This expression is maximised when k = n/2. The other interesting thing is that the frequency of outcomes falls off as we move away from the maximum value according to this characteristic curve:



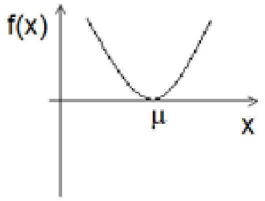
This is the bell curve we discussed earlier, and the question arises: can we identify a better formula to describe this curve? It turns out that there is a better formula, and that it can be applied to almost any distribution that is sampled many times.

We can define a normal distribution with a specific mean, μ, and variance, σ^2.

Now, for any outcome x, we can write the function:

$$f(x) = (x-\mu)^2$$

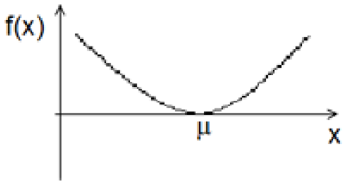
This function produces a characteristic quadratic curve with a minima where x = μ:



Now we divide our function f(x) by σ^2:

$$f(x) = \frac{(x-\mu)^2}{\sigma^2}$$

This scales down the output, f(x), by a factor of σ^2 and has the effect of 'widening' or flattening the quadratic curve:

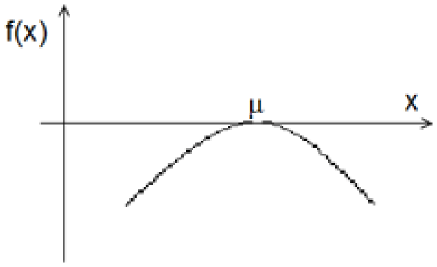


Note that large variances will result in very wide quadratic curves and small variances will result in relatively narrow or 'sharp' quadratic curves.

Next, we multiply our function by -½ to get:

$$f(x) = -\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}$$

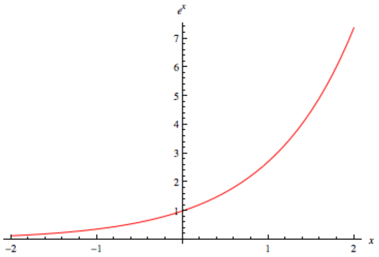
This flattens and inverts our quadratic curve so we get the following curve that has a maximum of zero when x = μ, but is otherwise strictly negative:



Finally, we make our function the exponent of e:

$$f(x) = e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

Now the exponential function $f(x) = e^x$ has a well known characteristic curve:



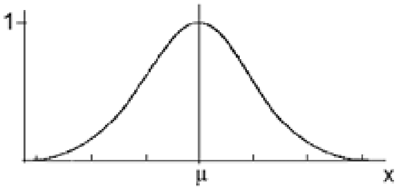
The exponential function is monotonic, that is, the larger the value of the exponent, the larger the value of the function.

The maximum value of our exponent is zero, and this occurs when x = μ. So our function f(x) is maximised when x = μ.

When x = μ, $(x-\mu) = 0$ and $f(x) = e^0 = 1$

The value of the function will be minimised when x = ± ∞. At these values of x we obtain $f(x) = e^{-\infty} = 0$

So we now have a function f(x) that has the value 1 when x = μ, and decays to 0 when x approaches ± ∞. This function also looks like the bell curve (although that is not entirely obvious!):

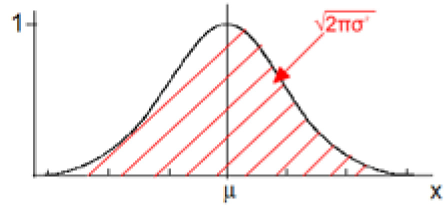


So this relatively simple formula describes the limit of computing the mean over any set of experiments, such as making infinitely many coin flips. No matter what kinds of experiment we do, when we drive n to very large numbers, we will obtain a bell curve like this.

The only flaw with this curve as it stands is that the area under the curve doesn't always add up to 1. It turns out that we do need the area under the curve to add up to 1 (just as we wanted the coin flip and its compliment to add up to one).

The area under the curve is given by:

$$\sqrt{2\pi \cdot \sigma^2}$$



We can normalise our function to ensure that the area under the curve always equals 1 by multiplying our function by the inverse of this to give the true normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot e^{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}}$$

This is the normal distribution for any value x, indexed by the parameters μ and σ^2. We can write this using mathematical notation as follows:

$$N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot e^{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}}$$

Formula Summary

So, we have our normal distribution, which can also be written as:

$$N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right\}$$

Now, if you are new to this, then this expression will look really cryptic. In time, if you continue working with statistics, this formula will become second nature to you. For now, it is important that you just understand how the formula is constructed, with the quadratic penalty term for deviations from the expected mean of the expression.

We can extract values from the normal distribution just as we did with flipping coins before. The way to do this is to recognise that any value of x has a probability given by the equation above. Therefore, a value x, that has twice the value as some other value x' is twice as likely to be drawn.

P(x) = 2.P(x')

Now, the normal distribution has an entire continuous range of outcomes. Obviously, this renders each individual outcome to be probability 0, but in essence, we can think of the height of the curve at point x as being proportional to the probability of that value being drawn.

Central Limit Theorem

There are a number of types of experiment we can carry out. We might characterise them as:

1. single coin flip
2. many coin flips - mean
3. infinitely many coin flips - mean

These examples have parallels in other fields. A medical doctor with a single patient can treat them much like the single coin-flip. If they have 10 patients, then it is more like the binomial distribution in case 2. Alternatively, if they have many thousands of patients - as they do on many drug-trials, for example - then they will be treated as a normal distribution.

In each case, there is an appropriate formula to determine the probability distribution:

1. p
2. $\frac{n!}{(n-k)!\cdot k!}\cdot p^k\cdot(1-p)^{n-k}$
3. $\frac{1}{\sqrt{2\pi\cdot\sigma^2}}\cdot e^{-\frac{1}{2}\cdot\frac{(x-p)^2}{\sigma^2}}$ with $\sigma^2=\frac{p(1-p)}{n}$

To clarify:

For 1 coin flip: This is the formula for figuring out the probability of the flip being heads?

For many coin flips: This is the formula for figuring out the probability of a given number of flips (k) being heads?

For infinitely many flips: What is the formula for figuring out the probability of a given proportion of flips being heads?

For formula 2:

$$\sigma^2=\frac{p(1-p)}{n}$$

What is interesting is that it turns out that it is the central limit theorem that governs the transition from a single coin-flip to many coin-flips, and right through to infinitely many coin-flips.

The exponential function in the third expression captures the distribution of a possible mean if it transitions from a discrete space of many, but finite, outcomes to a space that is continuous and which has infinitely many outcomes.

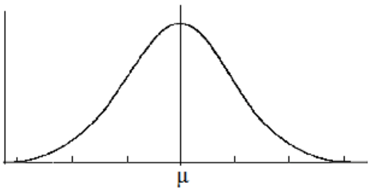
It turns out that the transition to a normal distribution works not just for coin-flips, but also for many other distributions that fall outside the scope of this class.

Manipulating Normals

Now that you understand something of the theory of normal distributions, we're going to return to something intuitive, that is how to manipulate normal distributions.

As we said in the last section, you can draw values from the normal distribution in just the same way as you would by flipping coins. So perhaps we should try to understand what happens when we manipulate normal distributions?

Suppose all the salaries at Udacity are normally distributed. That is to say, if we were to look at the salaries, there will be a well-defined mean and the salary distribution would approximate to the normal distribution curve, with fewer and fewer people having salaries as we get further and further from the mean:



Let's say that the mean is \$60,000 per year and the standard deviation is 10,000:

$\mu = \$60{,}000$
$\sigma = 10{,}000$

Now, we already know from previous units that if we give everybody a raise of a fixed amount, say \$10,000, the new mean and standard deviation will be:

$\mu' = \$70{,}000$
$\sigma' = 10{,}000$

In the context of the normal distribution, if we ignore the normalisation constant, we know that all the salaries are drawn from a distribution that looks like this:

$$\exp\left\{-\frac{1}{2}\cdot\frac{(x-60000)^2}{10000^2}\right\}$$

Now, the new salary, x' = x + 10,000

Or, alternatively, x = x' - 10,000

We can substitute this into our exponent to give:

$$\exp\left\{-\frac{1}{2}\cdot\frac{(x'-10000-60000)^2}{10000^2}\right\} = \exp\left\{-\frac{1}{2}\cdot\frac{(x'-70000)^2}{10000^2}\right\}$$

So the mean has increased to \$70,000, but the standard deviation remained unchanged.

OK, so let's say that Udacity is doing really well and they decide to double everybody's salary. Once again, we met this situation in earlier units, but let's now look at it in the context of the normal distribution.

Say the mean salary is now $70,000, with a standard deviation of 10,000, so:

$\mu = \$70{,}000$
$\sigma = 10{,}000$
$x' = 2x$

Substituting into the formula gives:

$$\exp\left\{-\frac{1}{2}\cdot\frac{\left(\frac{1}{2}x'-70000\right)^2}{10000^2}\right\} = \exp\left\{-\frac{1}{2}\cdot\frac{\left(\frac{1}{2}x'-\frac{1}{2}140000\right)^2}{10000^2}\right\}$$

$$= \exp\left\{-\frac{1}{2}\cdot\frac{1}{4}\frac{(x'-140000)^2}{10000^2}\right\} = \exp\left\{-\frac{1}{2}\cdot\frac{(x'-140000)^2}{(2\times 10000)^2}\right\}$$

$$= \exp\left\{-\frac{1}{2}\cdot\frac{(x'-140000)^2}{(20000)^2}\right\}$$

So we can see that:

$\mu = \$140{,}000$
$\sigma = 20{,}000$

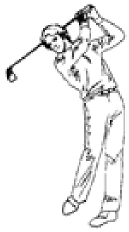
Throwing Quiz

Suppose that we are out on a playing field learning to throw a ball. On average we are able to throw it 30m, but because there is an element of randomness in the ball we have a standard deviation of 5m.

After training we have improved our performance by 10% and we learned to step forward more before we throw the ball, giving a further improvement of 2m on the distance we are able to throw the ball.

For a Gaussian outcome like this, how does this improvement change the mean, μ', and standard deviation σ'?

Golfer Quiz 1



Suppose that a golfer hits a golf ball down the fairway. The average distance is 100m with a variance of 30m^2. The second shot down the fairway also has an average distance of 100m and a variance of 30m^2.

For the combined strokes, what would you expect the mean distance and variance to be?

Golfer Quiz 2

Consider the same situation, but with the problem expressed in terms of standard deviation rather than variance:

μ = 100m
σ = 10m

We know that the combined μ will be 200m, but what about the new standard deviation?

So, when we add Gaussian variables, the means and variances add up, but the standard deviations **do not** add up.

Constants Quiz

Let's say that we draw the value A from a normal distribution:

$N(\mu, \sigma^2)$ μ is the mean and σ^2 is the variance.

Calculate the mean and variance of **aA + b** where a and b are constants.

Adding Normals Quiz

Let's say we have a normal distribution with μ and σ^2 from which we draw A, and we do the same with B from a distribution having the same μ and σ^2.

$A \sim N(\mu, \sigma^2)$ $B \sim N(\mu, \sigma^2)$

We add A + B. What are the new μ and σ^2?

Subtracting Normals Quiz

What happens to μ and σ^2 if we subtract A - B?

Summary

We have covered a lot in this section. We have seen that if X is normal, having the parameters μ and σ^2 then,

$$\mathbf{aX + b}$$

will always have the parameters:

$$\mathbf{a\mu + b} \text{ and } \mathbf{a^2\sigma^2}$$

We also say that if we add X and Y (both normals), the result has a mean that is the sum of the means of X and Y, and has a variance that is the sum of the variances of X and Y.

You have now practiced some basic math on normals, and should be beginning to get a feel for how they change as we manipulate them.

Statistical Mythbusters

It is a known fact that most drivers believe that they drive better than the average driver. Is it possible that they could be right?

Most people also believe that they have a higher IQ than the average person. Could they also be right about this?

It is even said that most people believe that they can run faster than the average person. They can't possibly be right, can they?

In fact, people could also be correct in all of these cases. Let's consider a very simple example to explain why.

Here is a hand.

It has five fingers. No real surprise there.

In most cases people have five fingers on their right hand. If you were to survey a group of 20 people, and count the number of fingers on their hand, you might get a data set that looks something like this:

5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4, 5, 5, 5, 5, 5, 5, 5

Now, you will find people who have fewer fingers, say 3 or 4 fingers, but that is quite rare (and even the occasional person with an *extra* finger, though that is even more rare!).

If we take the mean of our data sample in this case, we find that the average number of fingers for the group is 4.95.

So 95% of people in our sample group have more than the average number of fingers! 19 out of 20 hands have more than 4.95 fingers. In reality, with a larger sample size, the percentage would probably be very much larger.

So the three statements that we began with are all entirely possible from a statistical point of view, even though none of them may be based on any actual scientific evidence.

Answers

Compute Quartiles Quiz

Lower quartile = 20
Median = 22
Upper quartile = 23

Trimmed mean = 21.857

Compute Quartiles Quiz 2

Mean = 290.6
Trimmed mean = 21

Combinatorics Quiz

$$\frac{10\times 9\times 8\times 7\times 6}{5\times 4\times 3\times 2\times 1}=\frac{30240}{120}=252$$

Arrangements Quiz

$$\frac{125!}{3!\cdot(125-3)!}=\frac{125\times 124\times 123}{3\times 2}=317750$$

Binomial Quiz

The number of outcomes with exactly four heads is given by:

$$\frac{5!}{4!\cdot(5-4)!}=\frac{5!}{4!}=5$$

So, P(#HEADS = 4) = $5\times 0.8\times 0.8\times 0.8\times 0.2=0.4096$

Binomial Quiz 2

The number of outcomes with exactly four heads is given by:

$$\frac{5!}{3! \cdot (5-3)!} = \frac{5!}{3! \cdot 2!} = 10$$

So, P(#HEADS = 3) = $10 \times (0.8)^3 \times (0.2)^2 = 0.2048$

Binomial Quiz 3

The number of outcomes with exactly nine heads is given by:

$$\frac{12!}{9! \cdot (12-9)!} = \frac{12 \times 11 \times 10}{3!} = 220$$

So, P(#HEADS = 3) = $220 \times (0.8)^9 \times (0.2)^3 = 0.236$

Programming Flips

```
import random
from math import sqrt

def mean(data):
   return float(sum(data))/len(data)

def variance(data):
   mu=mean(data)
   return sum([(x-mu)**2 for x in data])/len(data)

def stddev(data):
   return sqrt(variance(data))


def flip(N):
   fred = []
   i = 0
   while i < N:
      if random.random() > 0.5:
         head = 1
      else:
         head = 0
      fred.append(head)
      i += 1
   return fred
```

Sets of Flips

```
import random
from math import sqrt
from plotting import *

def mean(data):
   return float(sum(data))/len(data)

def variance(data):
   mu=mean(data)
   return sum([(x-mu)**2 for x in data])/len(data)

def stddev(data):
   return sqrt(variance(data))


def flip(N):
   return [random.random()>0.5 for x in range(N)]

def sample(N):
   return [mean(flip(N)) for x in range(N)]


N=1000
outcomes=sample(N)
histplot(outcomes,nbins=30)

print mean(outcomes)
print stddev(outcomes)
```

Throwing Quiz

$\mu' = (1.1 \text{ x } \mu) + 2\text{m} = 35\text{m}$

$\sigma' = 1.1 \text{ x } \sigma = 5.5\text{m}$

Golfer Quiz 1

$\mu' = 200\text{m}$

$\sigma^{2\prime} = 60\text{m}^2$

Golfer Quiz 2

$\sigma^2 = 100\text{m}^2$

$2\sigma^2 = 200\text{m}^2$

so $\sigma' = 14.14\text{m}$

Constants Quiz

We have seen that multiplying values by a constant, a, results in a mean multiplied by that constant, $a\mu$. Adding a constant, b, to the values results in a mean $\mu+b$. So:

$$\mu = a\mu + b$$

We also saw that when we multiplied values by a constant, a, the standard deviation, σ, also increased to $a\sigma$. So:

$$\sigma^2 = a^2\sigma^2$$

Adding Normals Quiz

$\mu = 2\mu$
$\sigma^2 = 2\sigma^2$

Subtracting Normals Quiz

$\mu = 0$
$\sigma^2 = 2\sigma^2$

ST101 Unit 5: Inference

Contents:

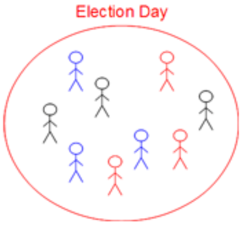
Confidence Intervals

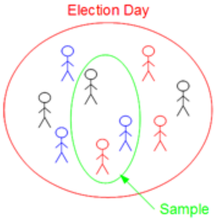
In this section we will talk about confidence intervals, and later in this unit we will look at testing hypotheses. These are fundamental concepts that every applied statistician uses almost every day.

To introduce the concept of the confidence interval, let's look at an example.

Imagine that there is an election. As a statistician, you will often want to try and understand what the outcome of the election will be before the actual vote happens:

Of course, you could just go out and ask every voter, but the effort involved would be almost the same as running the entire election in the first place! Although this may be possible for a small group of people, say 5 or 10 perhaps, but for an electorate numbering in the tens, or even hundreds of millions it is just not feasible.





What statisticians actually do is to choose a random sample and hope that the responses from this sample are representative of the whole group.

The sample will be a (hopefully) randomly drawn sample of people from the pool of voters who we then assume to be representative of the pool at large.

Using the sample, the statisticians will come up with an estimate of how they expect people to vote. These estimates are often reported along the lines of:

> "*60% will vote for party A, and 40 % for party B*"

In practice however, the statisticians actually report back their estimate together with a margin of error, for example:

Party A:	60% ±3%
Party B:	40% ±3%

This margin of error means that what is being returned isn't just a single number like 60%, but rather a range or interval. We call this the **confidence interval**. In the case of Party A in the example above, the lower bound of the confidence interval is 57%, and the upper bound is 63%.

What this means is that, given our current sample, we are fairly confident that, come election day, Party A will achieve an outcome within the range 57-63%.

Confidence intervals are a fundamental concept in statistics. They can be applied to coin flips and many of the other examples we saw earlier.

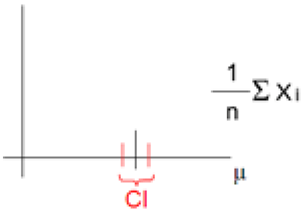
A candidate standing in an election may have a true chance, p, that any given voter will vote for him. Now, if $p > 0.5$ in a two-candidate run-off election, then he will win the election in most cases (although not always).

As statisticians, however, we cannot assess the true chance. What we do is we form a sample group:

$$X_1 \ldots X_n$$

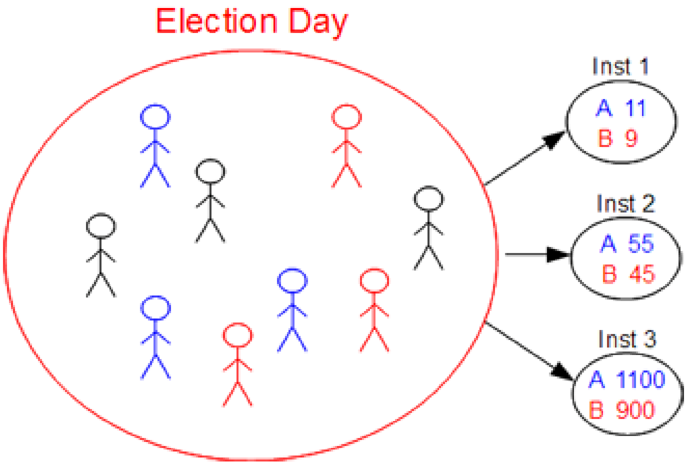
In coin-flipping, we flip n coins; in an election we ask n people chosen at random. The results from this sample group will give us an estimated mean, μ, and estimated variance, σ^2 (which then gives us the standard deviation).

What is new is the confidence interval. The confidence interval, CI, is not the same as the variance. What we are saying when we quote the confidence interval is that, based on the outcome with the sample group, we believe that the parameter μ (usually based on the MLE) falls within the range given by the CI:



Very often, the CI is defined as the range where there is a 95% chance that the outcome will occur. So how do we compute the confidence interval?

Let's stick with our election example, and for simplicity we will say that there are only two parties. Let's also say that there are three institutions sampling the voters to try to predict the election result. The first institution samples 20 voters, the second institution samples 100 voters and the third institution samples 1000 voters:



In each case, using the maximum likelihood estimator, the probability that a voter will vote for Party A, P(A), is 0.55. Clearly, we would have more confidence in the prediction from institution 3 because they used a much larger sample size.

Consider an extreme case where a company only sampled one voter. If that voter said that they intended to vote for Party A, then the company will be forced to conclude that P(A) = 1. Would we trust this prediction? Of course not! In general, the more data that is sampled (assuming that the sampling is fair and independent), the more trust we will have in the result. More trust means a smaller confidence interval.

So, if we increase the sample size, N, the confidence interval will get smaller. By contrast, the standard deviation will be unchanged if we increase the sample size. The standard deviation distribution is not dependent on the sample.

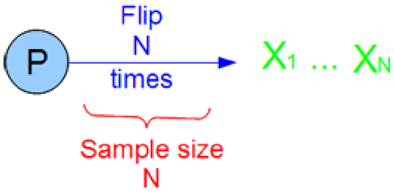
NOTE: Standard Deviation in this case means the standard deviation of the population from which the elements are sampled.

This is calculated as: $\sum_{i=0}^{n} \frac{(x_i - \mu)^2}{n}$

In fact, it turns out that $CI \propto \frac{\sigma}{\sqrt{N}}$

Let's go back probability, and to our simple coin-flip example.

We'll assume that the probability of the coin coming up heads in P, and that we flip the coin N times. This gives us a sample set of size N:



In the past, we have calculated the empirical mean, μ, and the variance, σ^2. Now we are going to calculate the confidence interval, CI.

We know that if the probability of heads P(H) = 0.5 then the expected mean will be:

μ = 0.5

We also know that the variance is defined as the quadratic difference of the sum of the differences of the outcomes from the mean:

$$\sigma^2 = \frac{1}{N} \sum (X_i - \mu)^2$$

In this case, there are two possible outcomes, heads, 1, or tails, 0. Both outcomes have a probability of 0.5.

If the outcome is 1, the squared difference from the mean is:

$$(1 - 0.5)^2 = 0.25$$

Similarly, if the outcome is 0, the squared difference from the mean is:

$$(0 - 0.5)^2 = 0.25$$

So the variance is:

$$\frac{(1-0.5)^2 + (0-0.5)^2}{2} = 0.25$$

Now, let's say we have three sample groups of size 1, 2, and 10. We know that the mean, μ, is 0.5, and that the variance of each individual coin-flip, σ^2, is 0.25. We saw that when we add Gaussian variables, the means and variances add up, so we can calculate the mean and variance of the sums of all the outcomes as follows:

	MEAN(ΣX_i)	VAR(ΣX_i)
N = 1	0.5	0.25
N = 2	1	0.5
N = 10	5	2.5

We can also calculate the variance of the means, $VAR\left(\frac{1}{N}\sum X_i\right)$.

Since variance is a quadratic expression,

$$VAR(aX) = a^2 VAR(X)$$

and so we have:

	MEAN(ΣX_i)	VAR(ΣX_i)	$VAR\left(\frac{1}{N}\sum X_i\right)$
N = 1	0.5	0.25	$\frac{1}{N}VAR\left(\sum X_i\right) = \frac{1}{1} \times 0.25 = 0.25$
N = 2	1	0.5	$\frac{1}{N}VAR\left(\sum X_i\right) = \frac{1}{2} \times 0.5 = 0.125$
N = 10	5	2.5	$\frac{1}{N}VAR\left(\sum X_i\right) = \frac{1}{10} \times 2.5 = 0.025$

This tells us something quite profound. The variance of the sum, VAR(ΣX_i), increases as the sample size, N, increases. However, the variance of the mean, or *spread of the mean*, actually decreases as the sample size increases.

As we already know, the standard deviation of the means will be just the square root of the variances of the means, thus:

	MEAN(ΣX$_i$)	VAR(ΣX$_i$)	$VAR\left(\frac{1}{N}\sum X_i\right)$	$SD\left(\frac{1}{N}\sum X_i\right)$
N = 1	0.5	0.25	0.25	0.5
N = 2	1	0.5	0.125	0.354
N = 10	5	2.5	0.025	0.05

Now we are able to calculate the confidence interval. It turns out that if we multiply the value we just calculated for the standard deviation of the mean by 1.96 we get the confidence interval for that sample size:

	MEAN(ΣX$_i$)	VAR(ΣX$_i$)	$VAR\left(\frac{1}{N}\sum X_i\right)$	$SD\left(\frac{1}{N}\sum X_i\right)$	C.I.
N = 1	0.5	0.25	0.25	0.5	0.98
N = 2	1	0.5	0.125	0.354	0.686
N = 10	5	2.5	0.025	0.158	0.3136

Now a note of caution. This trick of multiplying by 1.96 to calculate the confidence interval isn't mathematically correct for very small sample sizes. It normally assumes that we have at least 30 samples.

Confidence at 100 Quiz

Calculate the values for a sample size N = 100.

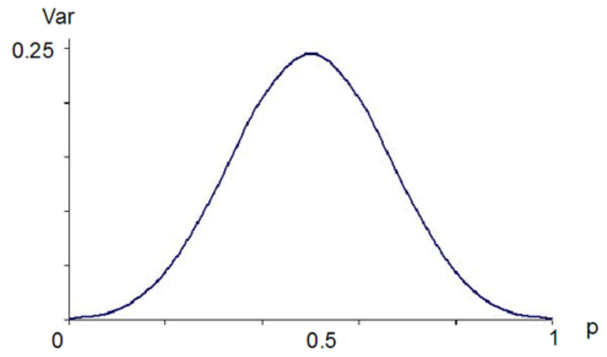
Variance Formulae

Now, we have studied the special case of a fair coin where p = 0.5, let's look at the more general case for an arbitrary value of p.

It turns out that the real challenge isn't calculating the confidence interval, but actually calculating the variance, σ^2. Let's consider the extreme cases where p = 0 or p = 1.

When p = 0, the coin always comes up tails, and the variance is therefore 0. Similarly, when p = 1, the coin always comes up heads, and the variance is also 0.

Now, we already know that when p = 0.5 the variance $\sigma^2 = 0.25$ and if we were to plot variance against p, we would expect to see something like this:



In fact, the formulae for calculating the mean and variance for a coin where P(H) = p are:

$$\mu = p$$

$$\sigma^2 = p(1 - p)$$

Now, given that p is the mean of X, we can derive the formula for the variance as follows.

There are two possible outcomes for the coin-flip. If the coin comes up heads, the variance is the product of p times the quadratic difference of 1 (for heads) minus the mean, p:

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

If the coin comes up tails, we get a similar result with 0 for tails:

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

Therefore:

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

Multiplying this out gives:

$$\begin{aligned}\text{Var}(X) &= p(1 - 2p + p^2) + p^2 - p^3 \\ &= p - 2p^2 + p^3 + p^2 - p^3 \\ &= p - p^2 \\ &= p(1 - p)\end{aligned}$$

CI Quiz

You have a loaded coin where p = 0.1

Calculate the variance Var(p) and the confidence intervals for sample sizes:

- N = 1
- N = 10
- N = 100

Use the following formula to calculate CI: $1.96\sqrt{\frac{(1-p)}{N}}$

CI Quiz 2

Suppose you flip a coin and observe the following sample set:

0, 1, 1, 1

Calculate the mean, the variance, and the confidence interval. You should calculate the actual variance from the data sequence, not the variance given by the formula.

Note: You should use the formula:

$$1.96\sqrt{\frac{\sigma^2}{N}}$$

to calculate the confidence interval, even though the sample size, N, is only 4.

In practice, you would only use this formula for samples where N ≥30

CI Quiz 3

Now calculate the mean, the variance, and the confidence interval for the sample set:

0, 0, 0, 1, 1, 1, 1, 1, 1, 1

CI Quiz 4

What happens to the mean, the variance, and the confidence interval if we flip the coin 1000 times and get the following result:

- 400 x tails
- 600 x heads

Normal Quantiles

In the last section we introduced the “magic number”, 1.96, as the multiplier for calculating the confidence interval. We defined the confidence interval to be:

$$\text{Mean} \ \pm \ 1.96\frac{\sigma}{\sqrt{N}}$$

where N is the number of samples. But where did the “magic” number 1.96 come from?

You will recall that when we looked at the central limit theorem we found that, when we have a large sample size, N, the mean outcome for N independently drawn values, X_i

$$\text{Mean} = \frac{1}{N}\sum X_i$$

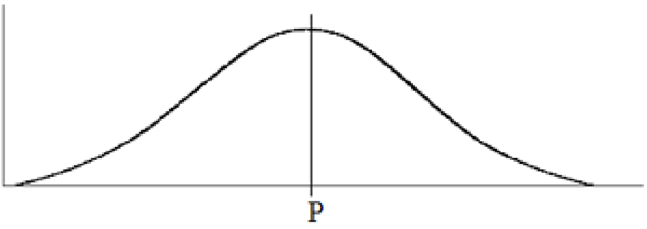
becomes more and more normal. It runs out that the “magic” number, 1.96, is directly related to this finding from the central limit theorem.

Take our coin-flip example. The coin has a true probability, P, but we cannot measure this. What we can do is to take our sample of N flips and calculate the empirical mean using the formula:

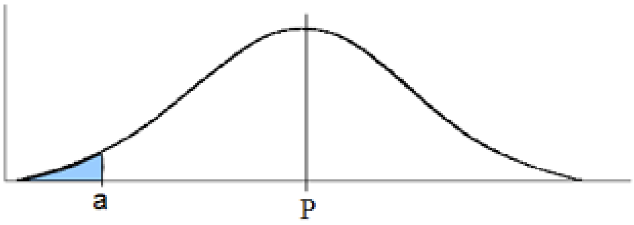
$$\mu = \frac{1}{N}\sum X_i$$

Now, we have already seen that it is quite likely that $p \neq \mu$ since the best that we can hope for from a finite number of coin-flips is just an estimate of the true probability, P.

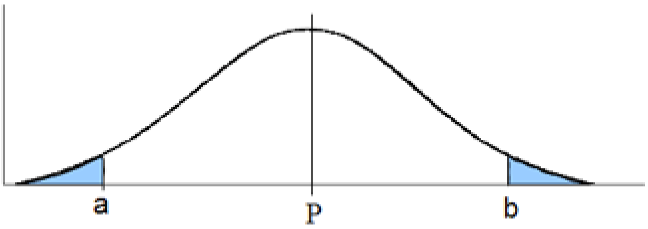
We know that for large sample sizes ($N \geq 30$) the distribution of μ that you might observe is Gaussian:



Now, the probability of observing any particular value of μ is the height of the curve at that point. so, for any value a, the chance of observing a value of μ smaller than a, P(μ < a), is given by the area under the curve as shown:



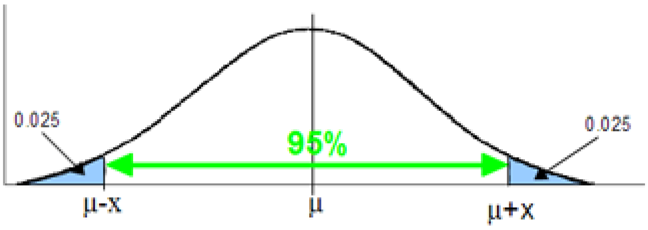
Also, by symmetry, the chance of observing a value of μ > b, where b is the same distance from p as a, is given by the area under the curve to the right of b:



If we compute the value of x such that when

$a = \mu - x$ and $b = \mu + x$

we have exactly 2.5% of the area under the curve enclosed on either side, then we can be 95% certain that μ will be between these limits:



This **95% confidence interval** occurs when x = 1.96. Making x smaller reduces the size of the confidence interval, while increasing x increases the size of the confidence interval:

90% confidence interval	x = 1.64
99% confidence interval	x = 2.58

The values for x are called **quantiles**.

T-Tables

If we have less than 30 samples, we are likely to experience problems if we try to use quantiles to calculate our confidence intervals. In this situation we can use a tool called a T-table. These are listed in most statistical textbooks.

Here is a selection of values from a T-table:

Significance Level = α

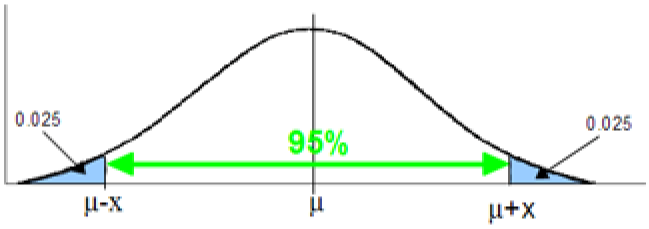
Degrees Of Freedom	**0.005(1 tail)**	**0.01(1 tail)**	**0.025(1 tail)**	**0.05(1 tail)**	**0.10(1 tail)**	**0.25(1 tail)**
	0.01(2 tails)	**0.02(2 tails)**	**0.05(2 tails)**	**0.1(2 tails)**	**0.20(2 tails)**	**0.50(2 tails)**
1	63.657	31.821	12.706	6.314	3.078	1.000
2	9.925	6.965	4.303	2.920	1.886	0.816
3	5.841	4.541	3.182	2.353	1.638	0.765
4	4.604	3.747	2.776	2.132	1.533	0.741
5	4.032	3.365	2.571	2.015	1.476	0.727
6	3.707	3.143	2.447	1.943	1.440	0.718
7	3.500	2.998	2.365	1.895	1.415	0.711
8	3.355	2.896	2.306	1.860	1.397	0.706
9	3.250	2.821	2.262	1.833	1.383	0.703
10	3.169	2.764	2.228	1.812	1.372	0.700
11	3.106	2.718	2.201	1.796	1.363	0.697
12	3.054	2.681	2.179	1.782	1.356	0.696
13	3.012	2.650	2.160	1.771	1.350	0.694
14	2.977	2.625	2.145	1.761	1.345	0.692
15	2.947	2.602	2.131	1.753	1.341	0.691
16	2.921	2.584	2.120	1.746	1.337	0.690
17	2.898	2.567	2.110	1.740	1.333	0.689
18	2.878	2.552	2.101	1.734	1.330	0.688
19	2.861	2.540	2.093	1.729	1.328	0.688
20	2.845	2.528	2.086	1.725	1.325	0.687

This may need some explanation!

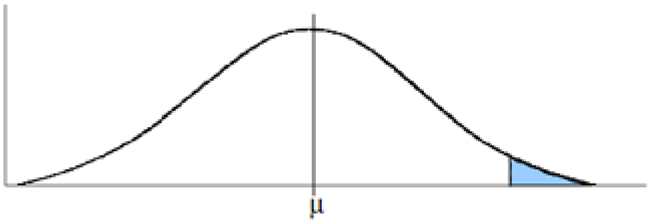
For our purposes, if we have N samples, the **Degrees of Freedom** is (N – 1). So, if our sample set contained 17 samples we would look to the row where Degrees of Freedom is 16.

The numbers in the header row are the (1 – CI), so if you want the 95% confidence interval you look to the column headed (1 – 0.95) = 0.05.

You will have noticed that the header row is divided in two with values for “1-tail” and also for “2-tails”. So far, we have been considering the “2-tail” case where we cut-off to both left and right of our confidence interval:



There are occasions (we will meet these later), where we only want to cut off one side:



These often occur in the context of testing hypotheses, but for the time being, we can limit ourselves to the “2-tail” row.

The t family of distributions arises in a case where the underlying population is approximately normally distributed (e.g. heights) and both the mean and variance must be estimated. The additional uncertainty in the variance results in thicker tails than the normal distribution. For larger numbers of degrees of freedom the t-distribution becomes increasingly similar to the normal distribution.

You can find more about the use of the t-distribution vs the normal at Khan Academy and Wikipedia.

So, as an example, if we have 8 samples in our sample set, and we want the 90% confidence interval we would look in the row where Degrees of Freedom = 7, and the column headed “0.1 (2-tails)” and locate the value 1.895.

Reading Tables Quiz

Suppose we want to achieve a 95% confidence interval, and we want a factor that is:

≥ 2.415

What is the minimum number of data points we need to collect to achieve this?

Reading Tables Quiz 2

Let's say that we flip a coin 5 times and get the sequence:

1, 1, 0, 0, 0

Calculate the following values:

- σ^2
- $\sqrt{\frac{\sigma^2}{N}}$
- CI

The confidence interval should be given in the form :

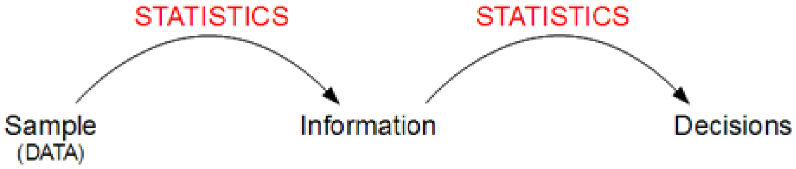
$$\text{MEAN} \pm x\sqrt{\frac{\sigma^2}{N}}$$

where x is the value from the T-table.

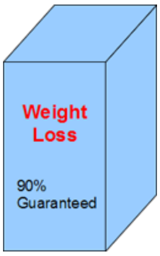
Note that in practice the t-distribution is not used to estimate binomial probabilities. Since the underlying distribution is not normal, on small samples exact confidence intervals are computed as described in the next section.

Hypothesis Test

Hypothesis testing is really all about decision making. So far on this course, we have talked about data a lot, and we have used statistics to extract information from that data. Now, we are going to use more statistics to make decisions based on that information:



These decisions will be binary, that is the outcome will be either 'yes' or 'no'.



Let's start with an example. Imagine that you have been given a carton of weight-loss pills. On the box it says that if you take these pills a substantial weight loss is guaranteed in 90% of cases.

It is your job to try to establish whether or not the pill manufacturer's claim is accurate. You contact people who took the pills and asked the question "Did you lose weight?".

You asked 15 people, and all of them answered. Eleven of these people answered 'Yes', while 4 people answered 'No'. Now this is only a 73.3% success rate, which is far lower than the manufacturer's claim on the carton. However, as we now know, this could just be a sample error. We will use hypothesis testing to see whether we should accept the manufacturer's claim on the basis of our data, or whether we should contact the manufacturer to complain that their claim is incorrect.

In this case we start with the hypothesis that the manufacturer's claim is correct. This starting hypothesis is often called the **Null-hypothesis**, H_0. We also have a counter hypothesis - one that invalidates the Null-hypothesis. This is called the **Counter-hypothesis**, H_1.

In this case, we can write:

$$H_0: p = 0.9$$

$$H_1: p < 0.9$$

Now we have these two hypotheses, and we are going to test them. We accept the Null-hypothesis until it is proved to be wrong, and the Counter-hypothesis is correct (i.e. we have sufficient evidence to show that there is a very high likelihood that the Null-hypothesis is wrong).

Critical Region

In the case of our weight-loss pills, the result from our sample group was:

YES: 11

NO: 4

and our hypotheses are:

H_0: $p = 0.9$

H_1: $p < 0.9$

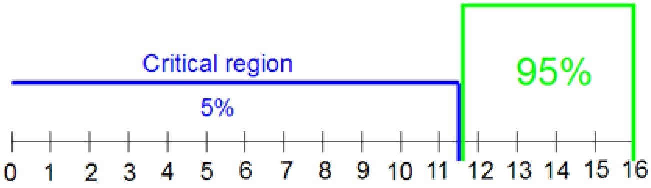
Now this sample set is obviously binomial ('binomial' just means binary results with just two possible outcomes, like coin-flips), and with this sample set there are exactly 16 possible outcomes, from nobody saying they lost weight, to all 15 members of the sample group saying that they had lost weight. We can calculate the probabilities for the binomial distribution using:

$$\frac{N!}{k!(N-k)!} p^k (1-p)^{(N-k)}$$

we get the result:

0	0	0	0	0	0	0	0	0.0003	0.002	0.01	0.04	0.13	0.27	0.34	0.21
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

What we want to do is to identify the group of outcomes on the left of this distribution that total no more than 5% of the total probability (for a 95% confidence level). Everything in this group is called the **critical region**. Everything to the right of this group is the **acceptance region**:



Outcomes in the critical region invalidate the Null-hypothesis. Outcomes in the acceptance region validate it.

In our case, summing all the values from 0 to 10 gives a result less than 0.05 (5%), but adding the value for 11 causes the result to exceed this threshold. The critical region in this case is therefore 0 – 10.

CRITICAL REGION											ACCEPTANCE REGION				
0	0	0	0	0	0	0	0	0.0003	0.002	0.01	0.04	0.13	0.27	0.34	0.21
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Loaded Coin Example

Let’s work through another example. Last weekend, Sebastian bought a loaded coin from the Magic Shop. He was told that for his new coin

P(HEADS) = 0.3

Sebastian suspects that the shop sold him a fair coin (or something close to a fair coin), so this is a 1-sided hypothesis.

Obviously, in this case the Null-hypothesis will be:

H_0: p = 0.3

and the Counter-hypothesis will be:

H_1: p > 0.3

Sebastian flipped the coin eleven times and got the sample set:

T, H, H, T, T, T, H, T, T, H, H

The observed value for p is therefore:

$$p = \frac{5}{1} = 0.45$$

Should Sebastian return the coin, given that the observed probability for this sample set is much higher than the advertised probability of 0.3?

Once again, we can calculate the binomial probability distribution using:

$$\frac{N!}{k!(N-k)!} p^k (1-p)^{(N-k)}$$

and we get:

0.02	0.09	0.12	0.25	0.22	0.13	0.06	0.02	0	0	0	0
0	1	2	3	4	5	6	7	8	9	10	11

Now, we can see that the most-likely outcome for 11 coin flips is 3 Heads, and the critical region will be somewhere to the right of 3 in this table.

In fact, for a 95% confidence level, the critical region is the columns 7 to 11 in the table, since including column 6 would push the total probability over 0.05.

So the observed five heads out of eleven flips falls well within the safe region. Sebastian would only need to return the coin if he had seen seven or more heads.

Fair Coin Example

So now we go to a bank to get what we hope is a fair coin. Our Null-hypothesis is thus:

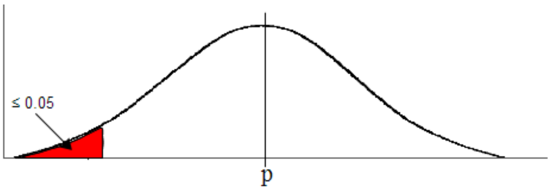
H_0: p = 0.5

and the Counter-hypothesis will be:

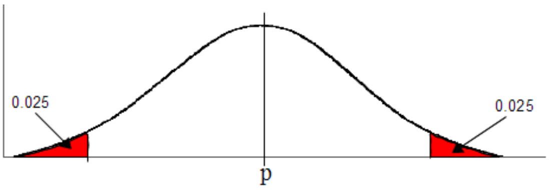
H_1: $p \neq 0.5$

Note that this time we have a 2-sided hypothesis. The probability may be smaller than 0.5 or higher than 0.5.

The way to look at this conceptually is that all we have done so far is to assume that H0 is correct, computed some sort of distribution, and then cut out a critical region such that the area under the curve did not exceed 0.05, or 5% (assuming we wanted 95% confidence):



In the 2-sided test, we cut out a smaller region on the left of the curve, but also one on the right, such that the area under the curve on each side does not exceed 0.025 or 2.5% for 95% confidence:



The total area under the curve (both left and right) still does not exceed 5%.

This is called a **2-tailed test**, and you'll have noticed that it looks an awful lot like the confidence interval!

So now we flip the coin. In 14 flips we got the following results:

T, T, T, H, H, T, H, T, T, T, T, T, T, T

So let's do the analysis.

In this table, we've listed the probabilities calculated for the binomial distribution:

0	0	0.005	0.022	0.06	0.12	0.18	0.21	0.18	0.12	0.06	0.022	0.005	0	0
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

The critical region in this case is going to be those columns on the left where the sum of the probabilities is ≤ 0.025, and the columns on the right where the sum of the probabilities is ≤ 0.025. i.e. columns 0 – 2, and columns 12 – 14 as shown:

Critical region			Acceptance Region									Critical region		
0	0	0.005	0.022	0.06	0.12	0.18	0.21	0.18	0.12	0.06	0.022	0.005	0	0
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

So it appears that our 3 heads falls within the acceptance region for our coin.

Cancer Treatment

A treatment for cancer is advertised. The manufacturer claims that it works in 80% of all cases. You suspect that the drug may not work as well as advertised.

Suppose that ten people are treated with the drug. The most likely outcome, if the manufacturer's claims are true is that eight of them with be cured.

Cancer Treatment Quiz

Using the 95% confidence level, what is the largest number of healthy people in the critical region that would cause you to reject the manufacturer's claims for this drug?

Cancer Treatment Quiz 2

In the experiment, 5 out of the 10 subjects in the sample set were found to be healthy, and five still had cancer.

Do we reject the manufacturer's claim that the treatment will work in 80% of all cases?

Summary

Congratulations! You now understand the basics of hypothesis testing. You should understand what is meant by the terms **critical region** and **null-hypothesis**, and you have seen how to apply them in both a 1-sided and 2-sided test.

Essentially, if the actual outcome observed from our sample set falls into the critical region then we will become suspicious and reject the null hypothesis. Whereas, if the observed outcome falls into the 95% acceptance region then we would accept the hypothesis as valid. That is the essence of hypothesis testing.

Hypothesis Test 2

In this section, we will combine what we learned in the last section with what we learned about confidence intervals.

Height Test Example

According to Wikipedia, in the 2007-2008 season, the average NBA basketball player was 6' 6.98" (200 cm) tall. Let's question this. Is this acceptable?

We visit an NBA training game and measure a group of players. We observe the following heights:

199cm, 200cm, 201cm, 202cm, 203cm, 204cm, 205cm, 206cm

On the basis of these measurements, and with a 95% confidence level, should we reject the claim in Wikipedia?

We have already seen how to calculate confidence intervals using the, now familiar, formula:

$$\frac{1}{N}\sum X_i \quad \pm a\sqrt{\frac{\sigma^2}{N}}$$

Where a is the factor from the T-table. In this case, a = 2.365, and if we plug our observed data into the formula we get:

202.5 ± 1.916

This means that the 95% confidence interval stops at 200.58cm. The figure of 200cm from Wikipedia therefore falls into the critical region, and we should reject the 200cm hypothesis based on our sample of eight people.

Now this would not be correct in practice. In fact, 6' 6.98" = 200.61cm, which falls within the confidence interval. The figure was rounded to 200cm for the purposes of this example. Also, we might find a very different sample if we were to visit a different town. The sampling wasn't completely independent.

However, the principle illustrated here is valid. Hypothesis testing can be really simple if we use our confidence interval and check whether the observed outcome lies within or outside that confidence interval.

We now know how to compute the confidence interval for any sample. If our null-hypothesis H0 falls inside the confidence interval then we accept that hypothesis. If, on the other hand, the null-hypothesis falls outside the confidence interval, then we reject the null-hypothesis and we accept the alternate-hypothesis.

In summary, given a sample of data, X1 … XN, we the calculate the mean and variance using:

$$\mu = \frac{1}{N}\sum X_i$$

$$\sigma^2 = \frac{1}{N}\sum (X_i - \mu)^2$$

We then get the T-value, at some desired error probability, p, from the tables:

T(N-1, p) = a

Remembering to select the correct T-value according to whether we are considering a 1-sided or a 2-sided test. Then the ± term in the confidence interval is simply:

$$a\sqrt{\frac{\sigma^2}{N}}$$

The lower bound of the confidence interval will therefore be $\mu - a\sqrt{\frac{\sigma^2}{N}}$ and the upper bound will be $\mu + a\sqrt{\frac{\sigma^2}{N}}$

NOTE: You should be aware that in other statistics courses, and for smaller samples, the variance is computed using N-1 as the divisor instead of n. So:

$$\sigma^2 = \frac{1}{N-1}\sum (X_i - \mu)^2$$

Club Age

A dance club operator advertises that the average age of its clients is 26. You visit the club and encounter 30 people with the following age distribution:

Number of people	Age
4	21
6	24
7	26
11	29
2	40

Do you trust the club operator's claim, based on this data sample?

Calculating the mean and the variance is straightforward, and we get:

$$\mu = 28.97$$

$$\sigma^2 = 19.57$$

Club Age Quiz

Noticing that N = 30, what value of a will we use for a 2-tailed, 95% confidence level?

Club Age Quiz 2

What is the ± term in the confidence interval?

So there is no real reason to doubt the club operator's claim based upon the sample that we have drawn.

Once again, in reality we would need to be aware that on any given night there might be a reason why particularly young people come to the club, or that on another night older people might come, perhaps due to the style of music being played. As a statistician, we would be wary of judgements made based on a sample taken on a single night. To ensure that the sample is truly independent you would need to attend on random nights, and pick a random person each night.

It's a tough life being a statistician!

Programming Tests and Intervals (Optional)

This is the optional programming section.

What we want here is really simple.

We want code that takes a sample and a hypothesis and returns either 'Yes' or 'No' depending on whether the result falls within the confidence interval.

For simplicity, we will assume 95% confidence and a 2-sided test.

Confidence Intervals Quiz

Write a function conf() that computes the ± term in the confidence interval. Use the functions mean() and var() that we wrote earlier as needed.

Hypothesis Test Quiz

Write a function test(), that takes as input a sample, l, and a hypothesis, h, and returns True if we believe the hypothesis in a 2-sided test, and otherwise returns False.

Answers

Confidence at 100 Quiz

	MEAN(ΣX_i)	VAR(ΣX_i)	$VAR\left(\frac{1}{N}\sum X_i\right)$	$SD\left(\frac{1}{N}\sum X_i\right)$	C.I.
N = 100	50	25	0.0025	0.05	0.098

CI Quiz

Var(p) = 0.09

- N = 1 CI = 0.588
- N = 10 CI = 0.186
- N = 100 CI = 0.0588

CI Quiz 2

$\mu = 0.75$

$$\sigma^2 = \frac{(0-0.75)^2 + (1-0.75)^2 + (1-0.75)^2 + (1-0.75)^2}{4} = 0.185$$

CI = 0.424

CI Quiz 3

$\mu = 0.7$

$$\sigma^2 = \frac{3\times(0-0.7)^2 + 7\times(1-0.7)^2}{10} = 0.21$$

CI = 0.284

CI Quiz 4

$\mu = 0.6$

$$\sigma^2 = \frac{400\times(0-0.6)^2 + 600\times(1-0.6)^2}{1000} = 0.24$$

CI = 0.0304

Reading Tables Quiz

15

Reading Tables Quiz 2

- σ2 = 0.24
- $\sqrt{\frac{\sigma^2}{N}} = 0.219$
- CI = 0.4 ± 0.467

Cancer Treatment Quiz

This is clearly a 1-sided test (we're not concerned if the test works better than advertised).

H_0 p = 0.8
H_1 p < 0.8

The binomial distribution looks like this, with the critical region shown:

0	0	0	0	0.005	0.026	0.088	0.2	0.3	0.27	0.1
0	1	2	3	4	5	6	7	8	9	10

So the largest number of healthy people in the critical region 1s 5.

Cancer Treatment Quiz 2

Yes.

Club Age Quiz

1.96

> **Note:** While this is not required for this course, you may want to know that 1.96 is based on the assumption that the sampling distribution is close enough to normal.
>
> The actual value is from a t-distribution with 29 degrees of freedom and is closer to 2.045.

Club Age Quiz 2

$$1.96\sqrt{\frac{19.57}{30}} = 1.58$$

Confidence Intervals Quiz

```
from math import sqrt

def mean(l):
   return float(sum(l))/len(l)

def var(l):
   m = mean(l)
   return sum([(x-m)**2 for x in l])/len(l)

def factor(l):
   return 1.96


def conf(l):
   return factor(l) * sqrt(var(l)/len(l))
```

Hypothesis Test Quiz

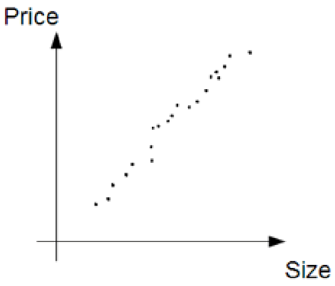
```
from math import sqrt

def mean(l):
   return float(sum(l))/len(l)

def var(l):
   m = mean(l)
   return sum([(x-m)**2 for x in l])/len(l)

def factor(l):
   return 1.96


def conf(l):
   return factor(l) * sqrt(var(l) / len(l))


def test(l, h):
   m = mean(l)
   c = conf(l)
   return abs(h - m) < c
```

ST101 Unit 6: Regression

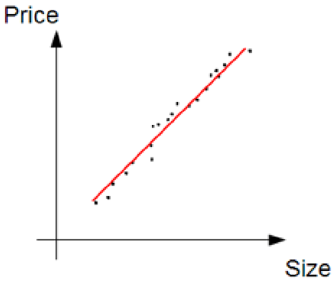
Contents:

Regression

You should remember from the very beginning of this course that we can have data sets with more than one dimension. For example, the size of a house relative to its price:



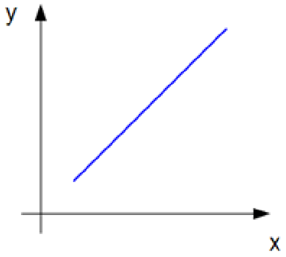
In the first unit we saw different ways of presenting the data, like scatter-plots or bar charts, but we didn't look at what might be thought of as the 'Holy Grail' of statistics, fitting a line to the data points:



By the end of this unit you will be able to fit a line to data points like these, and you will even be able to state what the residual error is in that fit. This will allow you not only to understand the data, but also to make predictions about points that you have never seen before.

Lines

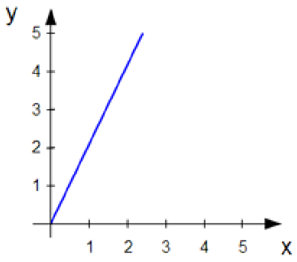
How do we specify a line? Well, suppose we have a horizontal axis, x, and a vertical axis, y. A straight line is commonly described by a functional relationship between x and y of the form:



$$y = bx + a$$

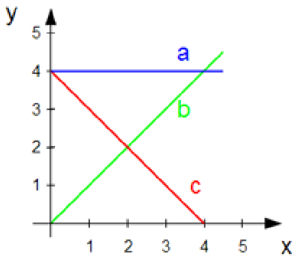
Consider the line described by the function y = 2x.

If x = 0 then y = 0. This means that the line must go through the origin. If x = 2 then y = 4. We now have two points and we can draw our line:



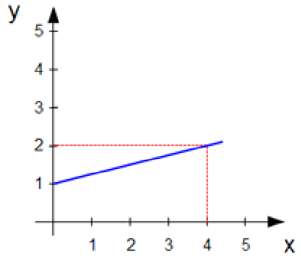
Pick the Line Quiz

What about the line, y = -x +4 i.e. where b = –1 and a = 4. Which of these lines best matches this function?



Find Coefficients Quiz

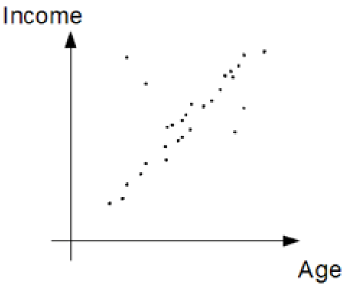
What are the coefficients a and b for the line shown below:

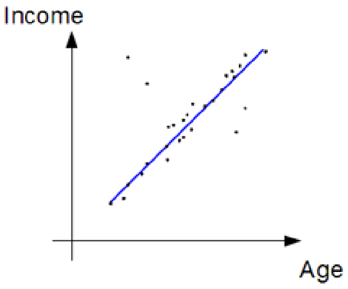


$y = bx + a$

Linear Regression

If we have 2-dimensional data, for example the age of a person and that person's income, linear regression is a technique for trying to fit a line that best describes that data:

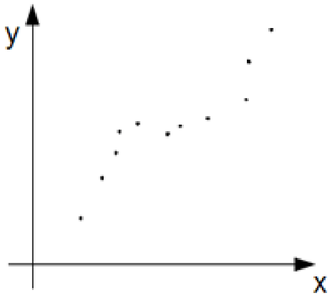




In linear regression we are given data (having more than 1-dimension), and we attempt to find the best line to fit the data. To put this differently, we are trying to identify the parameters a and b for the function:

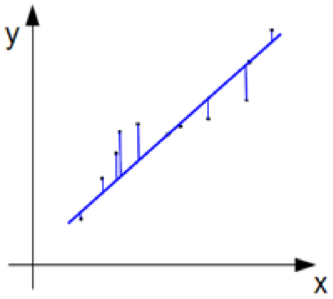
$$y = bx + a$$

We are trying to find the line that is the best-fit to our data, and the word 'best' is interesting in this context. Obviously it is impossible to draw a straight line that passes through every point in this data set:



This is what is known as **non-linear data** - the data points go up and down. Data often looks like this, even when the relationship between x and y is linear. This is because there is usually what is known as **noise** in the data. Noise is a random element in the data that we cannot explain.

In trying to find the 'best-fit' to the data, we are trying to find a line that minimises the differences between the data points and the line in the y-direction:



The reason for this is that we are assuming that our data results from some unknown linear function plus noise:

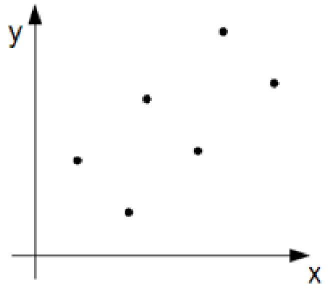
$$y = bx + a + noise$$

If the noise is assumed to be Gaussian, then minimising the quadratic deviation between the data points and the line provides the correct mathematical answer. In practice, what we are doing is adding, over all data points, the difference between our function and the y-value of the data point, squared:

$$\sum_{x_i} (bx_i + a - y_i)^2$$

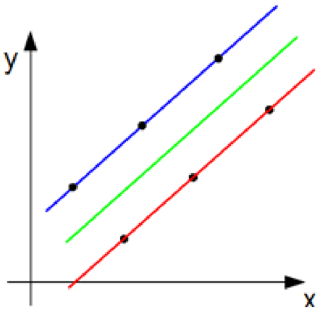
This is the distance that we are minimising.

Consider the following six data points:

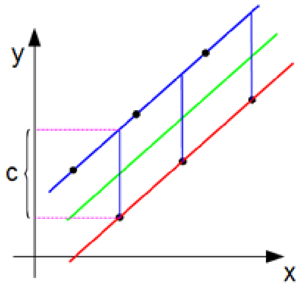


What is the best fit line for this data?

Well, consider the three possible lines shown:



In the case of the blue line, there is no loss for the three points that it passes through, but a fairly substantial loss for the other three points:



If the distance along the y-axis between the blue line and the data points is c, then the error, e, for the blue line is:

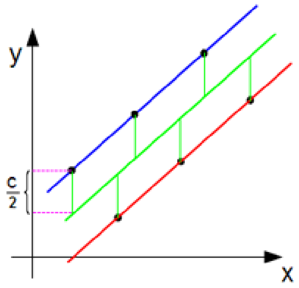
$$e = 3 \times c^2$$

since there are three points, and we are using the quadratic distance.

Similarly for the red line, it suffers no loss for the three points that it passes through, but a fairly substantial loss for the other three points, and the error, e , for the red line is also:

$$e = 3 \times c^2$$

In the case of the green line there are errors for all six points, but the error is only half as big in each case:



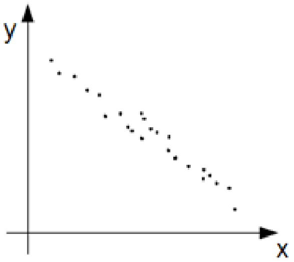
Now, the error is:

$$e = 6\times\left(\frac{c}{2}\right)^2 = \frac{6}{4}c^2 = \frac{3}{2}c^2$$

This means that the total quadratic error for the green line is half as big as it is for either the blue or the red lines. This is because, when we use the quadratic, larger errors count much, much more than smaller errors. Clearly, in this case, the green line would be the best fit for these data points.

Negative or Positive Quiz

Will the parameters a and b be negative or positive for the data shown below?



$$y = bx + a$$

Regression Formula

The function $y = bx + a$ is the Holy Grail of linear regression, and much of statistics is concerned with how to use the data to determine the value of b, and the value of a. If we can do this with the data, then we have solved the problem of fitting the best line.

If the data comes in pairs:

x_1	x_2	x_3	…	x_N
y_1	y_2	y_3	…	y_N

Then we can calculate b using the formula:

$$b = \frac{\sum_i \left[(x_i - \overline{x})(y_i - \overline{y})\right]}{\sum_i (x_i - \overline{x})^2}$$

Note that $\overline{x} = mean(X_i)$ and $\overline{y} = mean(Y_i)$.

Previously, we've used μ for the mean, but now that we have more than one variable we are going to use the bar-notation.

This formula shouldn't be entirely unfamiliar. When we calculated the variance, we used: $(x_i - \overline{x})^2$, whereas now we have 2-dimensional data we are using $(x_i - \overline{x})(y_i - \overline{y})$.

Now that we know b, and we know that:

$$y = bx + a$$

It turns out that this function is also true for the average values, $\overline{x}$ and $\overline{y}$. We can calculate the value for a using:

$$a = \overline{y} - b\overline{x}$$

Let's try it out in an example.

Suppose we have the following data sample:

x	y
6	7
2	3
1	2
-1	0

Now, with this data set we notice that the y-value is always exactly 1 larger than the x-value, so:

$$y = x + 1$$

This means that b = 1 and a = 1.

Let's see what we get when we use the formula. The first thing we do is to calculate the means:

$$\overline{x} = \frac{6+2+1-1}{4} = \frac{8}{4} = 2$$

$$\overline{y} = \frac{7+3+2+0}{4} = \frac{12}{4} = 3$$

Now we can evaluate b using:

$$b = \frac{(6-2)(7-3) + (2-2)(3-3) + (1-2)(2-3) + (-1-2)(0-3)}{(6-2)^2 + (2-2)^2 + (1-2)^2 + (-1-2)^2}$$

$$b = \frac{16+0+1+9}{16+0+1+9} = 1$$

Now we can calculate a using:

$$a = \overline{y} - b\overline{x} = 3 - (1 \times 2) = 1$$

Which agrees with our initial observation.

Regression Quiz

Calculate the values for a and b for the following data set:

x	y
4	7
3	9
7	1
2	11

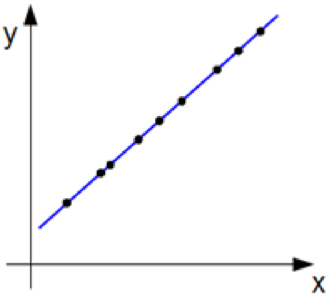
Correlation

This section is all about correlation. Correlation is a measure of how closely two variables are related. We can calculate something called the correlation coefficient which gives us a measure of how closely two variables are related.

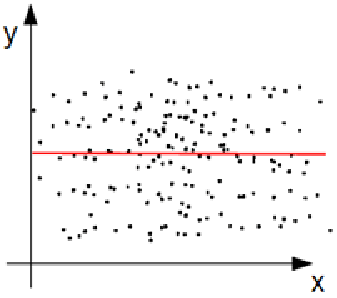
The correlation coefficient, r, will have a value between -1 and 1, so that:

$$-1 < r < 1$$

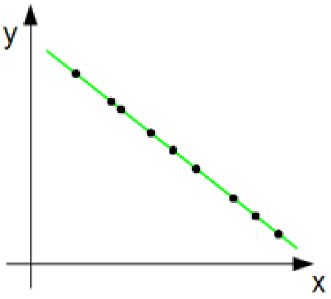
If the data is completely linear, without noise, as shown below, then we can fit a line exactly to the data and the correlation will be 1:



If the data is completely unrelated then the correlation will be 0:



The correlation coefficient can also be -1 in the case where the data is still perfectly aligned to the line, but where there is a negative relationship between the two variables as shown:



Correlation From Regression Quiz

Suppose that we ran linear regression on our data, and we found that:

b = 4

a = -3

Where these values describe this linear relationship between x and y:

y = 4x – 3

Which of the following statements about the correlation coefficient, r, is true?

- r is positive
- r is negative
- r = 0
- We can't tell

Correlation Formula

So, to summarise what we know so far about the correlation coefficient:

- It has a value between -1 and 1
- It tells us how "related" two variable are
- Both +1 and -1 describe perfectly linear data

So how do we compute a value for r?

Well, one way to compute the correlation coefficient is very similar to the way that we calculated the value of b when we looked at linear regression:

$$r = \frac{\sum_i \left[\left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right) \right]}{\sqrt{\sum_i \left(x_i - \bar{x} \right)^2 \cdot \sum_i \left(y_i - \bar{y} \right)^2}}$$

Where: $\bar{x} = \frac{1}{N} \sum x_i$ and $\bar{y} = \frac{1}{N} \sum y_i$

Now this is probably the most complex formula that you have encountered so far in this class, but you will see it is related to a lot of the stuff we have seen before, like variance and so on, and that using it is not that difficult in practice.

If we look at the denominator:

$$\sqrt{\sum_i \left(x_i - \overline{x}\right)^2 \cdot \sum_i \left(y_i - \overline{y}\right)^2}$$

We notice that the term within the square root is the product of the non-normalised variance of x and the non-normalised variance of y.

The numerator is a kind of “mixed-variance”:

$$\sum_i \left[\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)\right]$$

This is sometimes called the **covariance**, because it calculates the variance over two co-occurring variables. However, you will have noticed that, once again, there is a missing normaliser. What has happened is that the normalisers on the numerator and denominator cancel each other out.

Computing Correlation

Let’s try the formula out on some data.

x:	3	4	5
y:	7	8	9

It is easy to see that the mean of x is:

$$\overline{x} = \frac{1}{3}(3+4+5) = \frac{12}{3} = 4$$

and the mean of y is:

$$\overline{y} = \frac{1}{3}(7+8+9) = \frac{24}{3} = 8$$

Armed with these mean values we can calculate the following:

$x - \overline{x}$	-1	0	1
$y - \overline{y}$	-1	0	1

We can plug these values into our equation, and we get:

$$r = \frac{(-1\times -1)+(0\times 0)+(1\times 1)}{\sqrt{\left[(-1)^2+(0)^2+(1)^2\right]\cdot\left[(-1)^2+(0)^2+(1)^2\right]}} = \frac{2}{\sqrt{2\times 2}} = 1$$

Let’s look at a different dataset,:

x:	3	4	5
y:	2	5	8

Again, we calculate the means:

$$\bar{x} = \frac{1}{3}(3+4+5) = \frac{12}{3} = 4$$

$$y = \frac{1}{3}(2+5+8) = \frac{15}{3} = 5$$

Guess r Quiz

Before we do the calculation, let’s see what your intuition says about the value for r in this case. Which of the following do you think is correct?

- r = 1
- r = 3
- r = 2
- r = 0

Now let’s tabulate the differences from the means:

$x-\bar{x}$	-1	0	1
$y-\bar{y}$	-3	0	3

Again, we can plug the values into our formula to get:

$$r = \frac{(-1\times -3)+(0\times 0)+(1\times 3)}{\sqrt{\left[(-1)^2+(0)^2+(1)^2\right]\cdot\left[(-3)^2+(0)^2+(3)^2\right]}} = \frac{6}{\sqrt{2\times 18}} = 1$$

Which is as we expected.

Reverse Order

Let's see what happens if we change the order of the y-values as shown:

x:	3	4	5
y:	8	5	2

Since only the order has been changed and the actual values are unchanged, the mean values will be the same as before:

$$\bar{x} = \frac{1}{3}(3+4+5) = \frac{12}{3} = 4$$

$$y = \frac{1}{3}(8+5+2) = \frac{15}{3} = 5$$

Now, since the y-values decrease as the x-values increase we would expect the correlation coefficient to be negative. Let's see what happens when we calculate the value of r.

Tabulating the differences from the means gives:

$x-\bar{x}$	-1	0	1
$y-\bar{y}$	3	0	-3

And once again, we just plug the values into our formula:

$$r = \frac{(-1\times 3)+(0\times 0)+(1\times -3)}{\sqrt{\left[(-1)^2+(0)^2+(1)^2\right]\cdot\left[(3)^2+(0)^2+(-3)^2\right]}} = \frac{-6}{\sqrt{2\times 18}} = -1$$

So the data is perfectly *negatively* correlated.

Uncorrelated Data

Let’s try something a little bit tricky. Let’s change the y-values to:

x:	3	4	5
y:	8	5	8

What seems to be happening is that the y-values start to decrease as the x-values increase, but then they start to increase again. While there may be a relationship between x and y, it doesn’t appear to be a linear one, and we would expect the correlation to be zero.

Let’s do the calculation. The means are now:

$$\bar{x} = \frac{1}{3}(3+4+5) = \frac{12}{3} = 4$$

$$y = \frac{1}{3}(8+5+8) = \frac{21}{3} = 7$$

Now when we tabulate the differences from the means we get:

$x-\bar{x}$	-1	0	1
$y-\bar{y}$	1	-2	1

And if we plug the values into our formula we get the correlation coefficient:

$$r = \frac{(-1\times1)+(0\times-2)+(1\times1)}{\sqrt{\left[(-1)^2+(0)^2+(1)^2\right]\cdot\left[(1)^2+(-2)^2+(1)^2\right]}} = \frac{0}{\sqrt{2\times6}} = 0$$

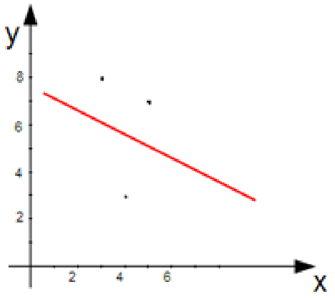
Of course, we didn’t actually need to calculate the denominator. Once we knew the numerator was zero we also knew that r = 0.

Final Example

Let’s look at a final example. This is our data:

x:	3	4	5
y:	8	3	7

Clearly, this doesn’t look very correlated, although it looks *more* correlated than the data in the last example since the final y-value hasn’t increased by as much. We might therefore expect that the correlation coefficient will be less than 1, and if we plot the data on a graph, we see that it should also be negative:



As usual, the first thing we do is to calculate the mean values for x and y:

$$\bar{x} = \frac{1}{3}(3+4+5) = \frac{12}{3} = 4$$

$$y = \frac{1}{3}(8+3+7) = \frac{18}{3} = 6$$

Now when we tabulate the differences from the means we get:

$x-\bar{x}$	-1	0	1
$y-\bar{y}$	2	-3	1

And if we plug the values into our formula we get the correlation coefficient:

$$r = \frac{(-1\times 2)+(0\times -3)+(1\times 1)}{\sqrt{\left[(-1)^2+(0)^2+(1)^2\right]\cdot\left[(2)^2+(-3)^2+(1)^2\right]}} = \frac{-1}{\sqrt{2\times 14}} = -0.189$$

So there is a negative correlation, as we suspected, but it is weak. This data really isn’t well described by a linear function.

Summary

You now understand the basics of correlation coefficients. As we said earlier, the correlation coefficient:

- has a value between -1 and 1
- tells us how “related” two variable are

We also now know that:

- If $r > 0$ there is a positive relationship between x and y.
- If $r < 0$ there is a negative relationship between x and y.
- If $r = 0$ there is no relationship between x and y.

Furthermore, we have seen that $|r| \to 1$ as the relationship between x and y becomes increasingly linear. Both $r = +1$ and $r = -1$ describe perfectly linear data without any noise or deviation from the line.

Correlation is a very powerful tool. For any data set with multiple variables, such as salary versus age, you can now tell how closely the variables relate to each other using the relatively simple formula:

$$r = \frac{\sum_i \left[\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]}{\sqrt{\sum_i \left(x_i - \bar{x}\right)^2 \cdot \sum_i \left(y_i - \bar{y}\right)^2}}$$

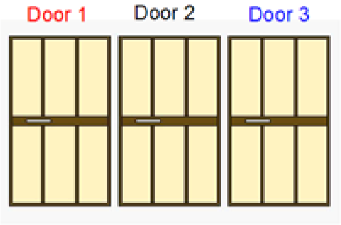
Monty Hall Problem (Optional)

This section should be fun. It is, however, entirely optional.

Some of you may be familiar with Monty Hall. He ran a US game show called **Let's Make a Deal** from 1963. The show ran for many years.

At the heart of the game show was a puzzle. This puzzle puzzles statisticians to the present day. In this unit, you get the chance to solve the Monty Hall program, and you also get the chance to program it, and verify that the assertion – which in not entirely obvious – is actually correct.

In the game there are three doors. Behind one door is a car. Monty knows which door, but he won't tell you where the car is.



The game is then played as follows:

You get to choose a door. Say, for the sake of argument that you pick door number 2. If the car is behind door 2, then you win the car, otherwise you win nothing. So far, so good.

Now comes the interesting bit. Obviously, at least one of the two remaining doors doesn't have a car behind it. Now, Monty knows which door has the car, and he reveals one of the doors that you didn't choose that doesn't have the car.

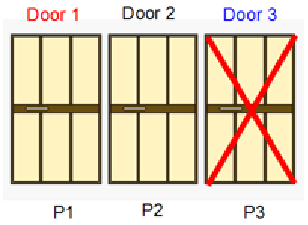
Monty then asks you if you want to switch from the door you have chosen to the other closed door. Do you want to change your choice in the hope that you will increase your chance of winning the car?

What makes this problem interesting is that when Monty opened the door, he really didn't tell you anything you didn't already know. You already knew in advance that one of the two doors didn't contain the car. The fact that Monty has just opened a door should give you zero information about which of the remaining doors the car is behind.

If you chose door 2, and Monty then opened door 3, all you now know is that door 3 doesn't contain the car. Why should door 1 now be more likely than door 2?

Door Chance Quiz

Given that you chose door 2, and Monty then revealed that door 3 didn't contain the car, what are the probabilities, P1, P2 and P3 that the car is behind each door?



Of course it is easy to see that P3 = 0. We already knew that the car wasn't behind door 3., But don't worry if you weren't able to figure out the other two probabilities, the answer to this problem is entirely non-intuitive.

We can perhaps best understand what is happening by building a truth table.

There are three possible true locations for the car, and we know that each of these possible locations has a probability of 1/3.

When we choose a door, Monty will open one of the other doors. We can construct the truth table as shown:

	True location	Monty Hall	Probability	Normalised probability
P(1) = 1/3	1	1	0	
	1	2	0	
	1	3	1/3	
P(2) = 1/3	2	1	1/6	
	2	2	0	
	2	3	1/6	
P(3) = 1/3	3	1	1/3	
	3	2	0	
	3	3	0	

Let's say, for the sake of argument that we have chosen door number 2. Now, we know that on the first row of the truth table, if the car is behind door number 1 then the probability that Monty will open door number 1 is zero. On the second line, the probability that Monty will open door 2 is also zero because we have chosen door number 2.

So the only option remaining is for Monty to open door 3. This has a probability of 1, but the posterior probability must take account that the total probability for P(1) = 3, so the probability becomes $1 \times 1/3 = 1/3$.

Similarly for the case where the car is behind door number 3, the probability that Monty will open door number 3 is zero, as is the probability that he will open the door that we have chosen (door number 2). The only remaining choice is door number 1.

If the car is behind door number 2, and we have chosen door number 2 then there is a fifty chance for door number 1 or door number 3. The posterior probability for both doors is therefore $1/2 \times 1/3 = 1/6$.

Now we have our truth table based on the fact that we chose door number 2.

Let’s say that Monty opens door number 3. Now every other case, in which he might have picked door number 1 or door number 2 has zero probability. The only events that remain are highlighted here:

	True location	Monty Hall	Probability	Normalised probability
P(1) = 1/3	~~1~~	~~1~~	~~0~~	
	~~1~~	~~2~~	~~0~~	
	1	**3**	**1/3**	**2/3**
P(2) = 1/3	~~2~~	~~1~~	~~1/6~~	
	~~2~~	~~2~~	~~0~~	
	2	**3**	**1/6**	**1/3**
P(3) = 1/3	~~3~~	~~1~~	~~1/3~~	
	~~3~~	~~2~~	~~0~~	
	3	**3**	**0**	**0**

Now we just need to normalise these values so that the total probability adds up to 1. Currently the sum of the probabilities is $1/3 + 1/6 = 1/2$. If we divide each individual probability by this sum we get the normalised probabilities as shown. These are the true posterior probabilities given that we chose door number 2, and Monty opened door number 3.

Clearly we should switch our choice every time, as this will double our chances of winning the car!

Simulation

Write a function simulate() that runs 1000 iterations of a simulation of the Monty Hall problem and so empirically verify the probabilities that we have just calculated.

The function should count how many times you win the car in a variable, K, and then return K divided by the number of iterations, N.

Python includes the built-in function randint(). This generates random integers in the range specified by the function arguments, so:

```
randint(1,3)
```

with return a random integer in the range 1 to 3 (which is exactly what you will need in order to pick a random door).

You will also have to simulate the actions of Monty Hall. Sometimes these actions will be deterministic, at other times they will be stochastic (random). Recall that we saw both types of action when we constructed the truth-table.

Once the "Monty-simulator" has picked a door, flip your choice to the remaining door. If that matches the true location then you should increment K, otherwise not.

When you run this 1000 times, the output from your function should be approximately equal to 2/3.

The assignment will require some real knowledge of Python and is therefore considered to be a challenging problem. Good luck.

Answers

Pick the Line Quiz

c

Find Coefficients Quiz

a = 1
b = 0.25

Negative or Positive Quiz

a is positive. As x increases, y decreases, so b is negative. This is an example of ***negative correlation***.

Regression Quiz

$\bar{x} = 4$

$\bar{y} = 7$

$b = \frac{28}{14} = 2$

a = 15

Correlation From Regression Quiz

- **r is positive**
- r is negative
- r = 0
- We can’t tell

Guess r Quiz

r = 1

Door Chance Quiz

P1 = 0.667
P2 = 0.333
P3 = 0

Simulation

```
from random import randint

N = 1000

def simulate(N):
    K = 0
    for i in range(N):
        TrueLoc = randint(1,3)
        guess = randint(1,3)
        if TrueLoc == guess:
            monty = randint(1,3)
            while monty == TrueLoc:
                monty = randint(1,3)
        else:
            monty = 6 - TrueLoc - guess
            switch = 6 - guess - monty
            if switch == TrueLoc:
                K = K + 1
    return float(K) / float(N)

print simulate(N)
```