Descriptive statistics and the sampling distribution

In this lecture we will discuss basic descriptive statistics, viewing them not just as ways of summarising variables but also as simple models of the world. Then, we will "go a little meta" by talking about how these statistics are themselves variables with their own distributions and descriptive stats.

AUTHOR
Milan Valášek

AFFILIATION
University of Sussex

PUBLISHED
Sept. 2, 2021

Contents

Introduction

In Lecture 4 we started talking about how quantitative methodology deals with measurable things (variables) and aims to explain and predict the world around us by modelling relationships between these things. The models of the relationships are mathematical/statistical in nature and they are based on numeric descriptions of variables. In other words such a model is essentially an equation with the predicted (dependent/outcome) variable on one side and predictor (independent) variables on the other.

When we're learning statistics, we're learning how to create these models and how to find the right equation that captures the relationships between the variables that we are modelling as accurately as possible. To be able to start formulating statistical models, we first need to understand how we can use maths to describe basic properties of things. This is the task of the first part of this lecture.

Describing things with maths

We already know that variables are any characteristics whose values differ (vary) in the population. They differ in their range and distribution as well as *from population to population*. Take, for instance, temperature in degrees Celsius (centigrade). Air temperature on Earth ranges from about −90°C to about 60°C with a fairly *symmetrical distribution* where, most of the time, the temperatures are relatively mild. However, temperature produced by humans has a much larger range. In laboratory conditions, scientists have produced temperatures as low as about −273°C and as high as ~5.5 trillion °C! Obviously, most of the temperatures produced by humans are neither this hot nor this cold with the bulk of them in the tens, hundreds, maybe thousands of degrees, making the distribution *positively skewed* (with a long right tail). Similarly, as discussed in Lecture 6, height in humans is *normally* distributed, while the distribution of wealth is, once again, heavily *skewed* to the right. To reiterate, not only two different variables but even *the same variable* (*e.g.*, temperature) in different populations can have wildly different ranges and distributions.

It's really important to understand that the term population, in the context of quantitative research and statistics, **does not only refer to people**! Yes, we can use the term to refer to all people, kittens, or plants but it's a much broader concept. We can also apply it to intangible, abstract, even hypothetical things and events, such as songs, games of chess, or possible outcomes of elections.

Central tendency and spread

There are numerous ways of describing variables and their distributions mathematically. The most basic way in which we can do it is in terms of **central tendency** and **spread**. By central tendency, we mean the "average" value of a variable, *i.e.*, where the *"most typical"*, or **central** value is located along the possible range of values. Spread, on the other hand, refers to how much *variability* there is in the individual values of the variable in the sample or population, *i.e.*, how much the values are **spread** along the range of values

There are various measures of both central tendency and spread, each with its advantages and disadvantages. What's important to understand is that, no matter the particular measure, all of them are mathematical *abstractions*. They provide useful information about the variables at hand but they can't really tell us the whole story. This may sound like a drawback but it isn't really. If we want to model relationships between variables in generalisible ways, we have to engage in some degree of abstraction. In fact we're often interested in finding the simplest models that still do a relatively good job at describing and predicting the world.

Measures of central tendency

Measures of central tendency tell us about the *"most typical"* value of a variable. What "most typical" or *average* means, however, is up to debate. For example, look at Figure 1 below. It shows average annual salary in US dollars on a sample of 78 countries. Each bar of this plot represents the number of countries in a given salary bracket ($0-$10k, $10-$20k, ...).

Plot R code

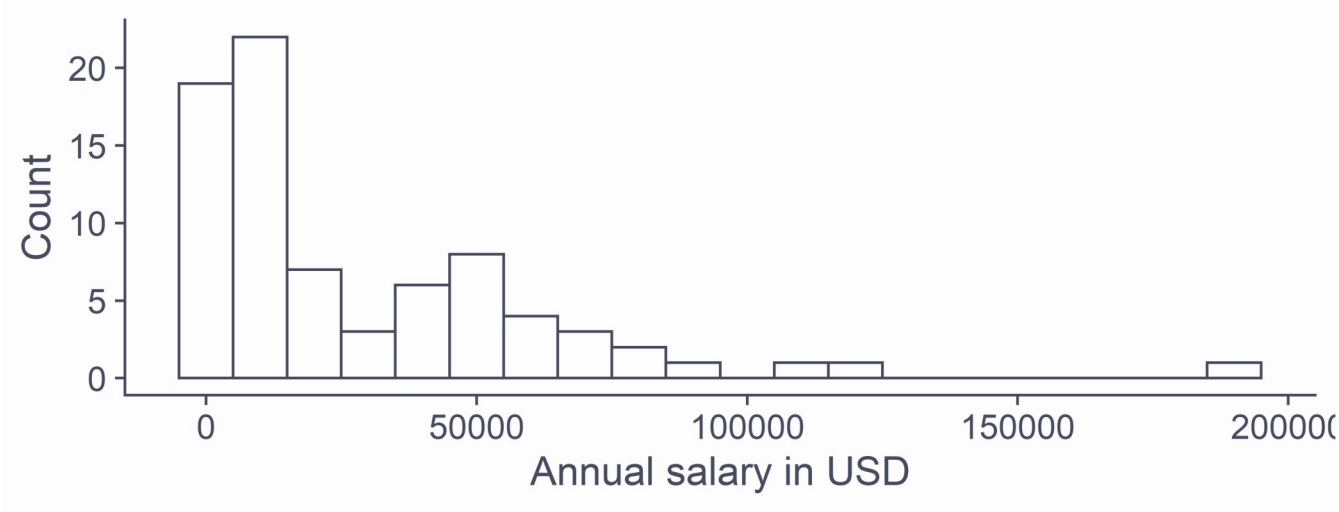


Figure 1: Average national annual salary [source: https://www.worlddata.info/average-income.php]

The distribution is heavily right-skewed, with a lot of countries, where people only get paid on average less than $30,000 per year and with a handful of countries where people make on average more than 100,000 US dollars. Given this distribution, what should be considered *most typical* salary bracket? The one with the highest number of countries in it? If so, that would mean that the most typical salary on this planet is between $10k and $20k, or about £7,300 - £14,600 per year. Or should we maybe pick the value where half the countries have a lower average salary and half have a higher one? Choosing this option leads to an estimate of the most typical salary of $12,855 (about £9,342) based on our sample data.

As you can see, the question of what should be considered average or most typical is a legitimate one. Different measures of central tendency provide different answers to this question Let's talk about the three main kinds of average: the mode, the median, and the arithmetic mean.

Mode

The mode is a term that refers to the most frequent value in the distribution. It is exactly the kind of "average" we discussed above, when we said that the most typical salary on this planet is between $10k and $20k a year. The easiest way to spot the mode is to just plot your variable on a *histogram* the way we did in Figure 1 and look for the tallest bar.

A distribution of a variable can have one or more modes. If it only has one, it's referred to as a unimodal distribution. A bimodal distribution is one with two modes and a distribution with three or more modes is usually called multimodal. Take a look at Figure 2. In panel A is a sample of 7,000 observations of a normally distributed variable. Because the normal distribution is bell-shaped and symmetrical, it only has one mode. The most likely values are in the middle and the tails include values that happen less often. The plot in panel B shows a histogram of a sample of 8,000 observations of a bimodal variable. Finally, in panel C, we have a histogram of 20,000 observations sampled from a *uniform* distribution, where each value is equally likely to be observed. This is a special kind of multimodal distribution where each value is the mode!

Plot R code

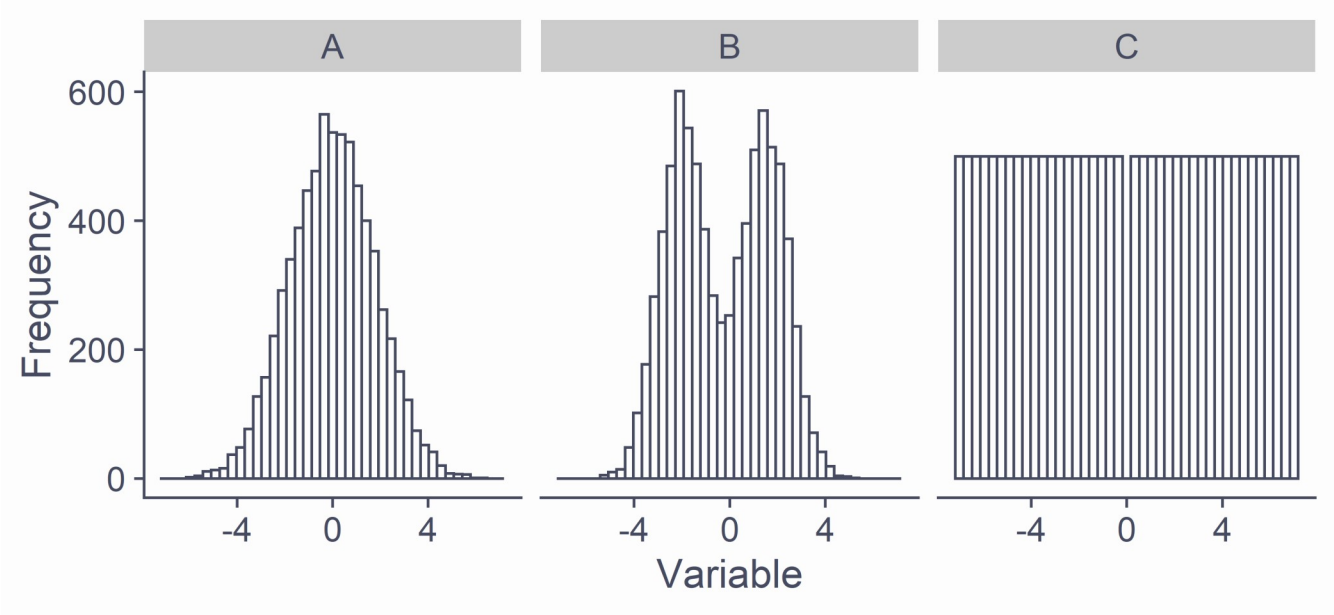


Figure 2: Examples of a (A) unimodal, (B) bimodal, and (C) multimodal distribution.

The mode is the only measure of central tendency we can apply to data measured at the **nominal/categorical** level.

When it comes to truly **continuous** variables, such as height, they are likely all multimodal. After all no two people are *exactly* equally tall and so every person's height is the mode of the distribution. For that reason, the mode is not often used with continuous variables measured at the **interval** or **ratio** levels.

Median

The **median** is the second kind of average we talked about above; the point at which half of the distribution lies below it and half lies above it. To find the median, we first need to sort our data from smallest to largest and then find the mid-point. For instance, let's say we roll a 6-sided die 5 times and get:

To calculate the median, let's do the two steps:

1. Sort data from smallest to largest: 1, 1, 3, 4, 6
2. Find the mid-point: We have five observations so the third one in the sorted sequence is the mid-point.

So, out of our 5 rolls of the die, the median is 3 (and the mode is 1). If we added one more roll and it came out, let's say ⏸, the sorted sequence would be: 1, 1, 3, 4, 4, 6.

If the number of observations is even, the median is half-way between the two mid-points. In this case, the median is half-way between 3 and 4, at 3.5.

Figure 3 below shows the annual salary in USD per each of the countries in the data set, sorted from lowest to highest. Notice, that this time, we're not grouping countries in salary brackets and looking at how many there are in each one as was the case in Figure 1. Here, each bar represents a country.

Because we have an even numbers of countries ino our dataset (78), there are two mid-points. These are highlighted in orange in the plot below. To get the median annual national salary, we need to find the value half-way between the average salary in Romania and Venezuela, which in this data set turns out to be $12,855.

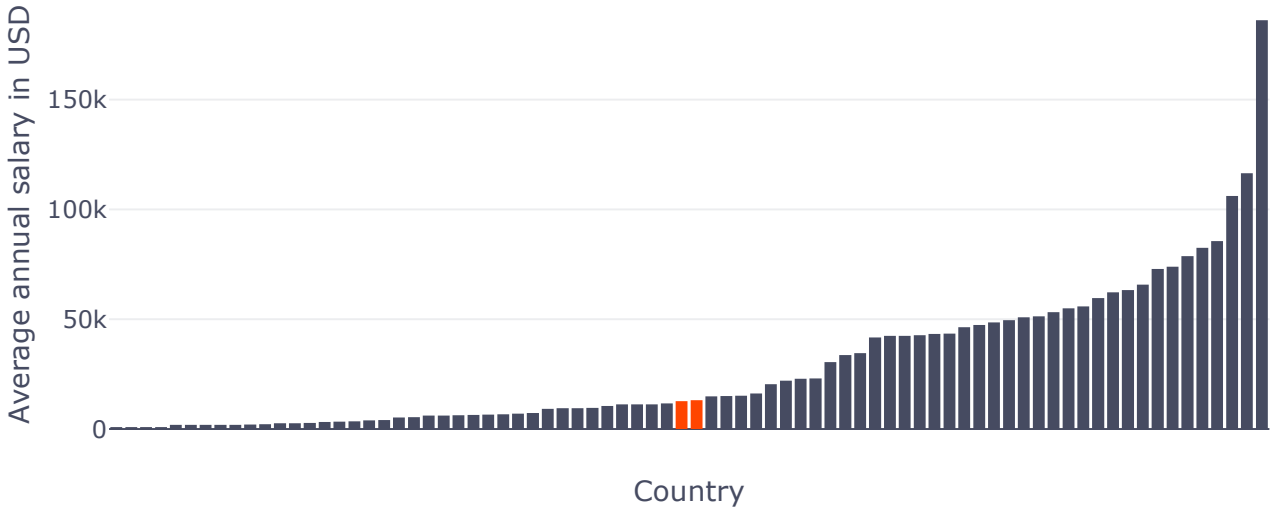


Figure 3: Average national annual salary per country sorted from lowest to highest [source: https://www.worlddata.info/average-income.php]

Hover over bars to see the name of the country and the value of average annual salary.

If plots are not showing, just refresh the page.

R know-how

Click to expand

In order to be able to calculate a meaningful median, the variable in question must be measured on **at least the ordinal level**!

(Arithmetic) Mean

The arithmetic **mean** [1] is what most people mean by *average*.

There is a difference between the **population mean** and the **sample mean**. We can never know the mean of the population because it is impossible to measure all instances of a variable. For that reason, the best thing we can do is calculate the mean of our sample and make *inferences* about the likely value of the population mean, as discussed in Lecture 6.

To distinguish between these two, we use the Greek letter μ /mew/ for population mean and $\bar{x}$ /x-bar/ for sample mean.

In fact, we usually use Greek letters for **population parameters** (characteristics) and Latin letters for **sample statistics**.

It's calculated by adding up all the values and dividing the sum by the number of values you added up.

The formula that describes this operation for a *sample* of *N* observations of a variable **x** is:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Let's break it down:

As we already know $\bar{x}$ is the sample mean. The symbol that vaguely looks like E is the capital Greek letter *sigma* and denotes **sum**:

$\sum_{i=1}^{N} x_i$ simply means *"add all the x_i together starting from $i = 1$ all the way to $i = N$, where N is the number of observations of the variable x"*:

$$\sum_{i=1}^{N} x_i = x_1 + x_2 + x_3 + \cdots + x_N$$

To illustrate this summation operation, let's imagine we have a set of observation of a variable x, let's say, 4, 5, 4, 10, 0. There are five observation of x, and so $N = 5$. So how to we add all of these values? Well, we simply take the 1st value (4), add to it the 2nd value (5), then the 3rd value (4), then the 4th value (10), and finally the last, 5th value (0). In symbolic terms that's: $x_1 + x_2 + x_3 + x_4 + x_5$ or $4 + 5 + 4 + 10 + 0$.

If we had 10 values to add, we would go on until we added x_{10}. And so for any N observations, we need to add $x_1 + x_2 + x_3 + \ldots + x_N$. Instead of writing things out laboriously like this, we can just replace the counter subscripts $1, 2, 3 \ldots N$ with an index variable i and say that we'll start at $i = 1$, add x_i and then do the same for $i = 2$, $i = 3$, and so on until we get to $i = N$. Once the do, we will have added up all the values in x.

That's exactly what the expression $\sum_{i=1}^{N} x_i$ means: set i to one and start adding up x_i for each value of i all the way to $i = N$.

Once we've added all the observations, all that's left to do is divide the sum by N, the number of observations.

R know-how

In order to be able to calculate a meaningful mean, the variable in question must be measured on **at least the interval level**!

Mean Vs Median

Both of these measures have their advantages and disadvantages. The mean has a proper *algebraic* formula which means that we can do all sorts of maths (and stats) with it, such as use calculus. There is no such formula for the median and so the range of mathematical tools we can apply to it is more limited. For that reason, most, although not all, of the statistical methods we will be learning are based on the mean. Don't worry though, we won't do the calculus ourselves!

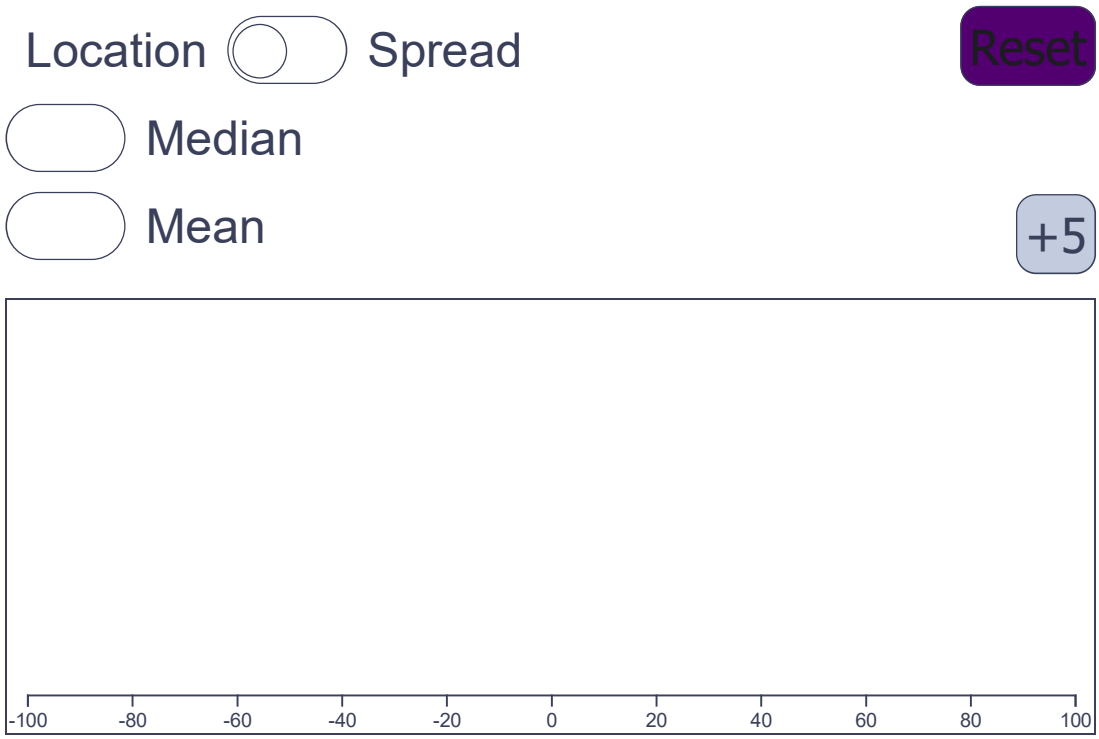
Compared to other measures of central tendency, *means taken from different samples of the same population are **relatively similar to each other***. If, on the other hand, we calculated medians of different samples from the same population, there would be more variability in the values we'd obtain.

However, it's not all just rainbows and butterflies with the mean. Unfortunately, the mean is relatively **sensitive to extreme values**. This means that, no matter the size of the sample, adding a single sufficiently small or large value can shift the mean to an arbitrary value! This makes it not a great measure of the centre for variables with very skewed distributions.

For example, the mean of the numbers 5, 3, 1, 7, 10, and 4 is 5 but if we add 1,000,000 to the numbers, the mean will shoot up to the value of 1,428,576. Conversely, if we add −1,000,000 instead, we'll get a mean of -1,428,567.

The median is not this sensitive: if we calculate the medians in these three cases, we get, 3.5, 4, and 3, respectively.

Use the interactive visualisation below to explore these characteristics of the mean and the median. Add a few points, turn on the mean and median switches and then drag one of the points left and right to see what happens with the calculated values.



Click inside the plotting region to add single points. The +5 button adds five points sampled at random from the normal distribution.

You can drag the points around and select multiple points by clicking and dragging from an empty space. Selected points can be moved around as a group.

Clicking on an existing point removes it from the plot. Clicking on a point from a selected group removes the entire group.

You can toggle between measures of central tendency (location) and spread using the switch.

Check your understanding #1

Mini quizzes like this one are designed for you to test yourself on the material you just read about. Some of the questions are fairly straight-forward, others require a little bit of thinking. Nevertheless, you should be able to answer all of them correctly so, if you get a question wrong, make sure you go back to the relevant section.

QUESTION 1

What is the only measure of central tendency that can have multiple values in a single sample?

Mode
Range
Arithmetic mean
Median

Here's a bunch of numbers:

34, 5, 42, 3.14, −18.5, 0.4, 5

QUESTION 2

What is the mode of these numbers?

Submit

QUESTION 3

What is the median of the same bunch of numbers?

Submit

QUESTION 4

What is their mean?

Submit

QUESTION 5

Which of these measures will change most if we add −577 to our numbers and recalculate them?

Median
Impossible to tell
Mode

In a normal distribution, which of these measures will have the largest value?

Arithmetic mean

Mode

Median

They are all the same

QUESTION 7

In a positively skewed unimodal distribution, which of these measures will have the largest value?

They are all the same

Mode

Median

Arithmetic mean

QUESTION 8

Which of these measures corresponds to the 50^{th} percentile irrespective of the shape of the distribution in question?

Arithmetic mean

Impossible to tell

Median

Mode

Measures of spread

The mode, median, and mean tell us about the central point of a variable's distribution but they don't tell us how spread the data are around this point, *e.g.*, how much variability there is in the variable.

Look at Figure 4 below: It's perfectly possible for two distributions to be centred around the same point but have very different amounts of variability. Both of the distributions in the plot have a mean=0 but, as you can see, the values of one are spread out much more widely than the values of the other one.

Plot R code

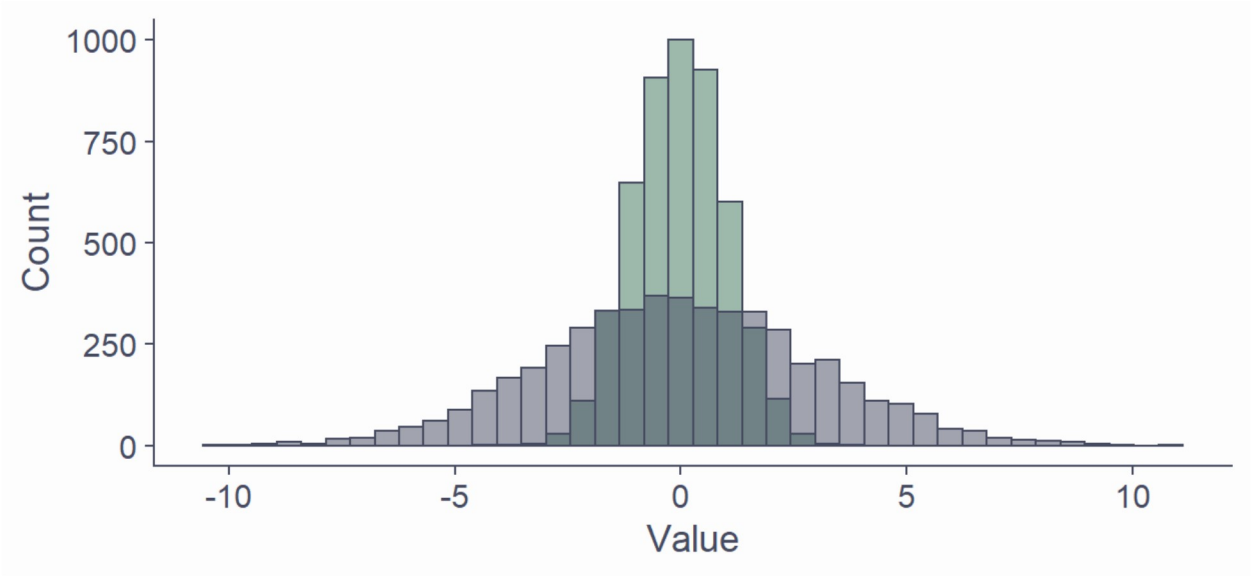


Figure 4: Histogram of two distributions with equal means but different spread. *N*=5,000 in each case.

This is why, in addition to measures of central tendency, we also need measures that tell us about the spread, or *dispersion*, of a variable. Once again, there are several measures of spread available and we'll talk about five of them:

- Range
- Inter-quartile range
- Deviation
- Variance
- Standard deviation

You can visualise all of them (except for variance) using the interactive visualisation above. Just flip the switch from "Location" to "Spread" to reveal more options.

Range

The range of a variable is simply the distance between its smallest and largest values. For instance, if we gather a sample of 100 participants and the youngest one is 17 years old, while the oldest one is 67, the range of the age variable in this sample is 67−17=50 years.

This measure is not without its merits but has a couple of issues. Firstly, it is *extremely* sensitive to outliers (unusually small or large values). Perhaps 99 of the participants in the example above a between 17 and 21 years old and it only takes a single 67-year-old to inflate the range. Secondly, because of the way it's calculated, the range doesn't tell us anything about the shape of the distribution. It can only tell us how far the minimum and maximum values are from one another.

Despite that, it's useful to look at the ranges of all the variables in your data just to see if they make sense.

R know-how

Inter-quartile range

The inter-quartile range or IQR, is the distance between the 1st and the 3rd quartile of a distribution. You may be familiar with the term *percentile*. Percentiles are just points along the range of a variable that divide the data into 100 groups, each with equal number of observations. For example, the first percentile is a value of a variable, at which 1% of the observations lie below it and 99% lie above it. So, if you are told that your exam result is in the 93rd percentile, that means, that you scored better than 93% of the people who sat the exam. Quartiles are like this but instead of 100 groups, they split the observations into just 4 groups. The first quartile is the same as the 25th percentile. The second quartile is the 50th percentile, AKA the median, and so on...

In other words, IQR is the range of the inner (bulk) 50% of the variable. One benefit of the IQR, compared to the range is that it's not at all sensitive to extreme values. On the flip-side, this is a bit of a mixed blessing because the IQR completely disregards the outer half of the data and provides no information about how spread this half is. What's good about it is that, even if two variables have the same range, the one that's more spread out will have a larger IQR that one where the vast majority of the points are tightly packed in the middle. Have a look for yourself! Using the applet above, add a few points to the plot and turn on IQR. Then, leaving the largest and smallest value as they are, play around with the spread of the other points and observe how the IQR changes.

R know-how

Deviation

Both the range and the IQR work by looking at the distance between only two observations in the entire variable. That's fine for some situations but both of these are very rough measures of variability in the data. To get a more fine-grain indicator of spread, we could calculate the distance of every value in the variable from its centre. For reasons that are outside of the scope of this lecture (but are explained in the Extra box below), it's reasonable to calculate the distance of every point from the mean. This is known as deviation:

$$x_i - \bar{x}$$

where x_i is every single data point.

Because we are calculating this for every data point, there are as many deviations as there are values in our variable. To get a *single* measure, we could add up the deviations. However, there are a couple of problems with this approach.

One is that if we just add the numbers up, the sum of deviations gets bigger with bigger samples. That's not good because even big samples can have small amount of variation, while smaller samples can vary a lot. We want our measure of spread to be able to capture this.

For the second, admittedly bigger, problem, go back to the applet and put two points on the plot. Then, turn on "Deviation" and "Values".

Honestly, do it...

As you can see, they add up to zero. That is not a coincidence: Add another point to see that it's still true. After all, mean is *by definition* the mid-point of the data so the combined distance of smaller values from it is the same as the combined distance of those larger than the mean.

We get around the problem of deviations adding up to 0 by taking the square of the deviations before adding them up, because the square of a negative number is a positive number and so we'll only be adding up positive numbers. Go back to the visualisation and switch on "Deviation" and "Squares" and play around with a few points to get an intuition for what squaring deviations does.

The sum of squared deviations is called the ***S**um of **S**quares* and its formula is:

$$\text{SS} = \sum_{i=1}^{N} (x_i - \bar{x})^2$$

Again, this simply means *"take the square of the deviation of every value of your variable from the mean and add those squared values up."*

EXTRA OK, but why squares and not absolute values?

And as for the problem arising from the fact that the sum of deviations gets bigger as we're adding more points, we can get around that by dividing the sum of squares by the number of values, *N*. This *average squared deviation* from the mean is called **variance**.

Just like with the mean, we need to distinguish between *population variance* and *sample variance*. Population variance is a fixed value – it is what it is – but, unfortunately, we can never know its actual value because we can't observe the entire population. All we can do is collect data from a sample and calculate sample variance.

Population variance, σ^2 /sigma squared/, is exactly what we talked about above: the average squared deviation from the mean. The population mean, μ, is subtracted from each value in the variable and the result is squared. Then, all of these squared deviations are added up to produce the sum of squares. Finally, the sum of squares is divided by N, the number of values in the population.

The formula for population variance then is (notice the μ for the population mean):

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Notice that the formula is quite similar to the formula for the mean. That's because variance is basically the mean squared deviation of the variable.

Sample variance, s^2, is calculated in a very similar way, just with two differences. Firstly, we use the sample mean, $\bar{x}$, instead of the unknown population mean. Secondly, we divide the sum of squares by $N - 1$, instead of N. This is shown in the formula for sample variance:

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N - 1}$$

The reason for why we use $N - 1$ instead of N is quite technical but without it, the sample standard deviation would be *biased*: it would produce values that are too small more often than they are too large. For the technical details, see Bessel's correction.

R know-how

Standard deviation

Variance is a good measure of dispersion and it is widely used. In fact, most of the statistical tests we will be using are based around variance. However, there is one minor inconvenience about this measure when it comes to interpretability: it's measured in *squared units*! For example, if salary is measured in US dollars, s^2_{salary} is expressed in USD^2, whatever those may be.

Fortunately, the solution to this problem is easy: we can simply take the square root of variance. This is called the standard deviation. Just like with variance, there is population standard deviation, σ, and sample standard deviation, s or SD.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

You can think of *SD* as a measure of the differences of a set of scores from their mean. If variance is the mean *squared* deviation in the variable, standard deviation is the **mean deviation**.

R know-how

A nice property of the standard deviation is that it scales proportionally to units of measurement. From the R output above you can see that the standard deviation of average national annual salary in our sample is 33094.66. Now, salary is measured in US dollars. If, however, we measured salary in 1000s of USD, $s_{\text{salary1000}}$ would be **proportional** to s_{salary}: it would be (33.09, as opposed to 33094.66.

Check your understanding #2

QUESTION 9

For the same numbers as before (34, 5, 42, 3.14, −18.5, 0.4, 5) calculate thir range.

Submit

QUESTION 10

Without checking, is this value going to be smaller or larger than their interquartile range?

Impossible to tell without calculating
The same
Larger
Smaller

QUESTION 11

What is the interquartile range of these numbers?

Submit

QUESTION 12

Which yields a larger value, the formula for population variance (σ^2) or for the sample variance (s^2) if applied to the same numbers?

Sample variance
Population variance
They are the same
Depends on the numbers

QUESTION 13

Treating our numbers as a **random sample** from a population, what is their standard deviation?

Give answer to 2 decimal places

Submit

QUESTION 14

University students

From sample to population

Now that we are familiar with the basic ways of describing variables using measures of central tendency and spread, let's take a step back and look at the bigger picture. As stated in the introduction, we do quantitative research because we want to make claims about the world in general. We do this by taking a sample of the world and exploring it using statistics.

Perhaps the most important thing about quantitative research and statistics is that *we don't actually care all that much about samples*, ***we care about populations***. However, we cannot measure the entire population so we have to make do with samples. And so we end up making claims about the world based on what we know from the sample

The problem with samples, as discussed in Lecture 6, is that **we cannot be sure that our sample accurately represents the population**. Because of that, there's always **uncertainty** associated with any empirical claims we make based on a sample of data.

Statistics is the science of uncertainty

Let's illustrate this point on the example of a game of Scrabble. In this vocabulary-based word game, you place tiles with letters on a board to form words. Every tile has a point value and you are awarded points based on the tiles you use in the words you create. A full set of Scrabble tiles contains 100 tiles with a mean tile value of 1.87 points and a *SD* of 1.83.

Figure 5 shows a histogram of tiles by their point value in a full set.

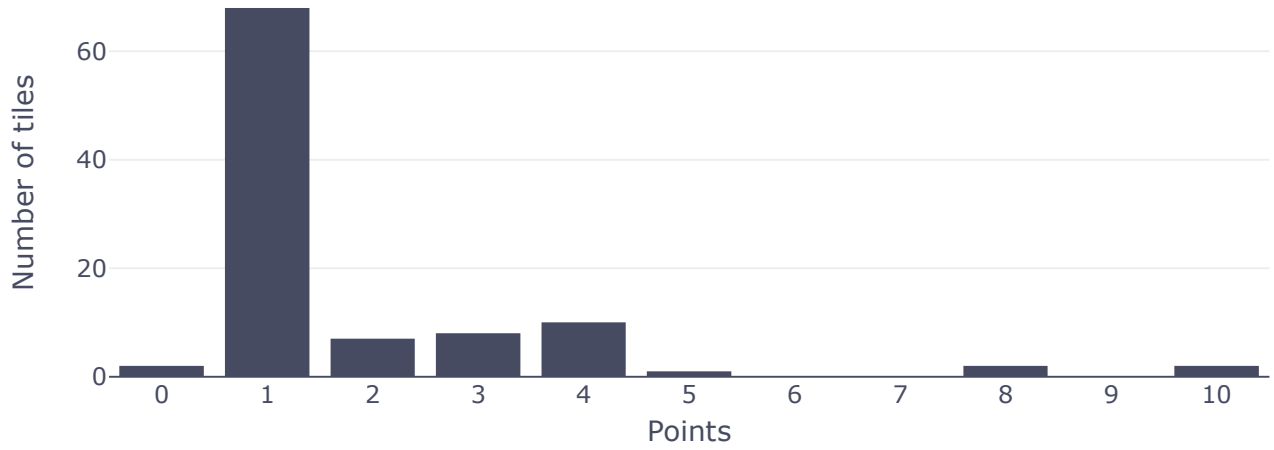


Figure 5: Distribution of Scrabble tiles by point value.

Hover over bars to see which tiles are worth the given number of points.

For the purpose of this example, let's consider the set of 100 tiles our **population**. At the beginning of a game, all tiles are placed in a green bag and each player picks seven tiles *at random*. Each player's pick constitutes one **sample**. Now if you've ever played Scrabble, you'll know that sometimes it just so happens that you pick only vowels whcih are all worth a single point and sometimes you get only the high-value Zs, Qs, Ks, or Ws. This happens due to *statistical fluctuation* related to **random sampling**. However, most often, you pick a mix of low-score and high-score tiles that is roughly similar to the distribution in 5.

Try it out for yourself using the interactive visualisation below! First, press the Draw button to pick seven tiles *at random* from the bag. The point value is shown on each tile except for the blank tiles worth zero points. When you do, you'll see the mean of the points of the sampled tiles appear below the tiles. Although we don't know what the value of the mean of the seven tiles will be, chances are it will *not* be the same as the population mean (1.87). Remember, this is due to randomness in sampling!

Do it and come back...

Now, press Draw again to get another selection of seven tiles and check the mean. See? it's different again!
And again.
And again..
And again...

OK, here comes the cool bit. Flip the switch below the tiles to display a histogram of the means of the samples of seven tiles you've drawn so far. The purple vertical line in the background shows the population mean, $\mu = 1.87$. The dashed blue line, on the other hand shows the mean of this distribution: the mean of the means of the 7-tile samples...
Meta, right?

And now the finale.

Press the ⊙ button to start sampling and resampling automatically. Each time, the new sample mean will get added to the distribution in the plot below. If you want, you can use the slider to speed things up.

Have a go and when you're ready, read on.

Interactive visualisation of the sampling distribution of the mean

Controls:

1. Draw randomly picks seven tiles from the bag.
2. ▶ / ⏸ triggers and pauses automatic resampling.
3. When you push ▶, a speed slider appears. You can use it to control the speed of the animation.
4. Switch under the tiles reveals/hides a histogram of means from samples.
5. Reset, well... resets things.

Purple vertical line shows the population mean of the scrabble tiles, $\mu = 1.87$. The light-purple ribbon around the line shows ±1 standard error of the mean, $SEM = 0.69$.

Dashed blue line shows the mean of the distribution and gets updated every time a sample is drawn. Notice how it eventually overlaps the purple line.

OK, I hope you've got a nice tall distribution going there. You can pause the animation or let it run. It will pause itself when the tallest bar hits 10,000.

There are a few **very important** things to notice about this distribution:

- Its shape has relatively quickly stabilised
- It looks much **more normal** (symmetrical and bell-shaped) that the population distribution (in Figure 5)
- Its mean (the dashed blue line) is now **identical to the population mean** (purple line)

An idealised version of this distribution is called the **sampling distribution of the mean**.

Sampling distribution

The sampling distribution is the distribution of a statistic (*e.g.*, the mean) based on **all possible samples of a given size taken from the same population**

Sampling distribution is NOT *the distribution of the sample!*

In our Scrabble example, it's the distribution of the means of **all possible 7-tile draws**.

The sampling distribution of a statistic has some really interesting and useful properties:

The first one is that its mean of the sampling distribution is equal to the population value of the calculated statistic

- The mean of the sampling distribution of **the mean** *is equal to the population mean*
- The mean of the sampling distribution of **variance** *is equal to population variance*

The second one is that, in the case of the sampling distribution **of the mean** (but not necessarily other parameters), its shape gets *more and more normal* (bell-shapes and symmetrical) as the size of the sample increases. In the applet above, we only sampled seven tiles. That's definitely not a large sample. But already, you can discern an emerging bell-like shape despite the population distribution, shown in Figure 5, being very much not normal. Figure 6 shows how the shape of the sampling distribution of the mean *approximates* the normal distribution as sample size (*N*) goes from 3 to 100. Notice how the mean, indicted by the vertical orange line, stays at pretty much exactly the value of the population mean $\mu = 1.87$, as per the paragraph above.

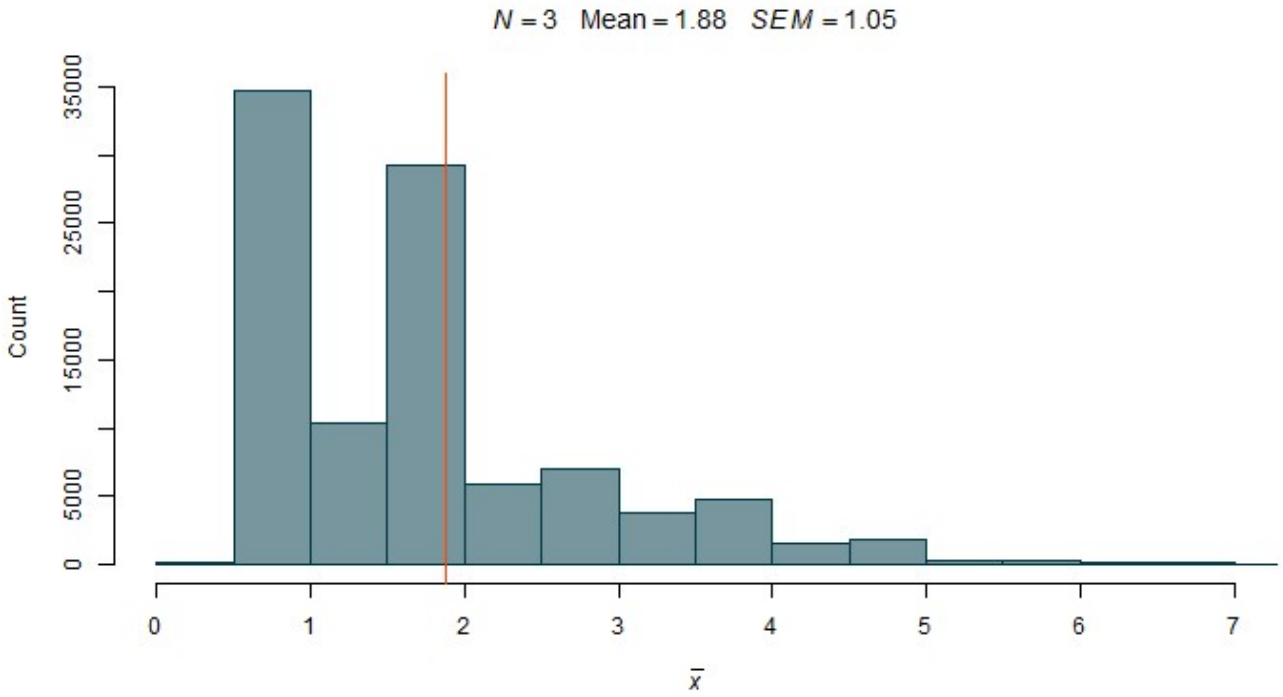


Figure 6: the sampling distribtuion gets more and more normal as sample size gets larger.

Standard error

Finally, the standard deviation of the sampling distribution, called the standard error (*SE*), gets smaller as *N* gets bigger. You can see that in Figure 6: the spread of the distributions gets smaller as the sample size that each of the sampling distributions is derived from gets bigger. In fact, there is a formula for the standard error that tells us how much spread there is in the sampling distribution.

To reiterate, the standard error, *SE*, is the standard deviation of the sampling distribution.

For the standard error of the mean (*SEM*), the formula is:

$$SEM = \frac{\sigma}{\sqrt{N}}$$

Because of the relationship between variance (σ^2) and standard deviation (σ), this is the same as:

$$SEM = \sqrt{\frac{\sigma^2}{N}}$$

This formula tells you two important things. Firstly, *SE* of the mean **gets bigger the more variability there is in the population**. That why there's the population standard deviation (σ) in the *numerator* of the formula. This makes sense if you think about it. If, let's say, mean annual salary doesn't change much from country to country, every sample you pick from the population is going to be very similar. As a result, the sampling distribution of the mean of these samples will not be very spread out and so the *SE* will be relatively small. Conversely, if the mean national annual salary varies a lot between countries, as is the case in this world, then we can expect to draw samples whose means will differ (vary) a lot. In that case, the *SE* will be larger.

Secondly, as mentioned above, *SE* of the mean gets smaller as sample size, N, gets bigger. That is why there's the N in the *denominator* of the formula. You saw this in Figure 6 where the spread of the sampling distribution of the mean gets smaller and smaller as we change the sample size that the given sampling distribution is based on.

In the salary example, based on our sample of 78 countries, the *SEM* is equal to $3,747.23. That means that if we took *all the possible samples* of 78 countries from the population of all countries on Earth [2], the majority of the sample means – in fact about 68% of them – would be within +1 *SEM* from the **true population mean**. We can be sure of that, even though we don't know what the true population mean is!

> The standard error tells us how much variability there is in the sampling distribution of a given statistic. In other words, it tells us how much we can expect the statistic we are calculating to differ from sample to sample. Often, we use *SE* of the mean but it can be calculated for any other statistic.

The standard error of the mean for samples of seven Scrabble tiles is about 0.69, which tells us that most 7-tile samples will be within the range from $\mu - 0.69$ to $\mu + 0.69$. In this rather rare example, we actually know the mean population score because we know the values of each of the 100 Scrabble tiles: $\mu = 1.87$. So we know that most 7-tile samples will have means ranging from 1.18 to 2.56. If you look at the sampling distribution in the Scrabble applet, you'll see that this is true.

This is why *SE* is considered a **measure of uncertainty** about our estimates.

> The concepts of the sampling distribution and standard error will be of crucial importance later, when we are talking about testing hypotheses and statistical modelling

Check your understanding #3

QUESTION 15

Which of these samples will yield the least reliable estimate of the population mean?

- Small sample with little variability
- Large sample with a lot of variability
- Large sample with a little variability
- Small sample with a lot of variability

QUESTION 16

Which is larger, variability in the sample or variability of an estimate based on that sample?

- Variability in the sample
- Depends on the sample
- Variability of an estimate
- They are the same

QUESTION 17

What is the relationship between the mean and its standard error?

- As one gets bigger, the other increases too
- There is no relationship
- Depends on the variance in the sample
- As one gets bigger, the other one gets smaller

Recap

In this lecture, we learned how to describe distributions ("shapes of variables") using maths. We talked about measures of *central tendency* and measures of *spread* of variables.

Central tendency refers to the mid-point of a variable. We discussed three types of "average":

- Mode: the most frequent value
- Median: the mid-point if we sort our values from smallest to largest
- Mean: the centre of the distribution, such that the combined distances (deviations) of all the points add up to zero.

Spread refers to the amount variability in the variable. We talked about several measures of spread:

- Range: the distance between smallest and largest value in the variable
- IQR: the distance between the 1^{st} and 3^{rd} quartile
- Variance: the average *squared* deviation in the variable
- Standard deviation: the average deviation in the variable

Each measure has its properties, benefits, and drawback and is useful in different situations. You need to be aware of these.

After talking about the descriptive measures, we considered the bigger picture.

We talked about how when doing quantitative research, we don't really care about samples, we care about populations. However, we have to rely on on samples because we don't have access to populations. Different samples have different properties (*e.g.*, means) even though they are sampled from the same population.

The sampling distribution is the distribution of a given statistic from all possible samples of the same size drawn from the same population.

The standard deviation of the sampling distribution is the standard error. It quantifies the variability of a statistic from sample to sample and therefore the uncertainty about the relationship between the sample statistic (*e.g.*, $\bar{x}$) and the population parameter (*e.g.*, μ). Two important properties of the standard error are:

- The larger the sample, the smaller the *SE*
- More variable populations lead to larger *SE*s

Footnotes

1. There are other kinds of mean, such as harmonic or geometric but, generally speaking, when someone just say “mean”, they are referring to the arithmetic mean [↩]

2. There are about 200 countries but it’s a complicated geopolitical mess [↩]