# The distribution of the means

## Introduction to the concept of the “standard error of the mean”

Imagine we wish to know the answer to a question such as:
“*What is the average height of 20 year old Italian men?*” (this does actually relate to other data later on...!)
We can attempt to answer this question by taking a sample of ten 20 year old Italian men and averaging their height[1].

| Height (cm) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 176.7 | 178.0 | 181.4 | 178.9 | 176.5 | 175.2 | 179.0 | 169.2 | 175.2 | 175.7 |

Find the mean (average) of this data set: $\bar{x} =$ _____

We can obtain an estimate of the spread of our data by finding the standard deviation:

| $x_i$ (cm) | 176.7 | 178.0 | 181.4 | 178.9 | 176.5 | 175.2 | 179.0 | 169.2 | 175.2 | 175.7 |
|---|---|---|---|---|---|---|---|---|---|---|
| $d = x_i - \bar{x}$ | | | | | | | | | | |
| $d^2 = (x_i - \bar{x})$ | | | | | | | | | | |

Calculate the standard deviation: $\sigma = \left(\frac{\Sigma(x_i - \bar{x})}{n}\right)^2$ : _____

Perhaps we then decide we want a better estimate of the average height, so take a sample of 100 men. For this larger sample we find a mean of 177.3cm and a standard deviation of 4.8cm.

We would expect this larger sample should give us a more accurate measure of the mean of the population of 20 year old men - but how do we determine the uncertainty in our measurement? The standard deviation of the population is not the right measure - this stays ***constant*** as we take more measurements and so more closely sample the true underlying distribution of heights.

Instead, we need to know something about how the standard deviation of the distribution of the ***means*** of a set of data change as we increase the number of measurements we average over.

[1] This data is based on the height of european men born in 1980 from https://ourworldindata.org/human-height

Tammy Humphrey (2018)

An example

If we take 10 men and find the mean of their height, and then take *another* 10 men and take the mean of their height, and another 10 men and find the mean of their height and so on, we could use this data to calculate *the standard deviation of the average height of 10 men.* We use a new variable, $\sigma_{\bar{x}}$ to represent the *standard deviation of the mean*.

| Average Height of 10 men (cm) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 175.72 | 172.0` | 173.7 | 179.2 | 177.7 | 176.5 | 176.9 | 178.1 | 177.6 | 177.0 |

Standard deviation of the average height of 10 men: $\sigma_{\bar{x}} = 1.8cm$

What if we now do the same thing, but we average the height of 100 men and repeat this ten times, as shown in the data below. The standard deviation of this data set is much lower than the standard deviation in the average height of ten men.

| Average Height of 100 men (cm) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 177.1 | 177.2 | 176.8 | 176.1 | 175.8 | 177.3 | 176.1 | 176.4 | 176.7 | 176.9 |

Standard deviation of the average height of 100 men: $\sigma_{\bar{x}} = 0.41cm$

A normal distribution has the property that:

- 68% of values lie within one standard deviation of the mean
- 95% of values lie within two standard deviations of the mean
- 99.7% of values lie within three standard deviations of the mean.

This means that:

If you measure the height of 10 men 100 times, then 95 of these measurements of average height will lie within within $2\sigma_{\bar{x}} = 3.6cm$ of the “true” average height of all 20 year old men.

If you measure the height of 100 men 100 times, then 95 of these measurements of average height will lie within within $2\sigma_{\bar{x}} = 0.82cm$ of the “true” average height of all 20 year old men.

In other words, there is less than a 5% chance ($p<0.05$) of your measurement of average height being more than 3.6cm away from the “true” height if you sample 10 men. Similarly there is less than a 5% chance ($p<0.05$) of your measurement of average height being more than 0.82cm away from the “true” height if you sample 100 men.

**We can take $2\sigma_{\bar{x}}$ as the uncertainty in your estimate of the mean of a set of data.**

It can be shown that the standard deviation of the mean is related to the standard deviation of the population the data is drawn from as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where $n$ is the number of measurements that have been averaged. **It is important to note that we don’t have to actually take many different measurements of the mean to know $\sigma_{\bar{x}}$!**

Tammy Humphrey (2018)

# Spreadsheet activity to demonstrate the distribution of the means

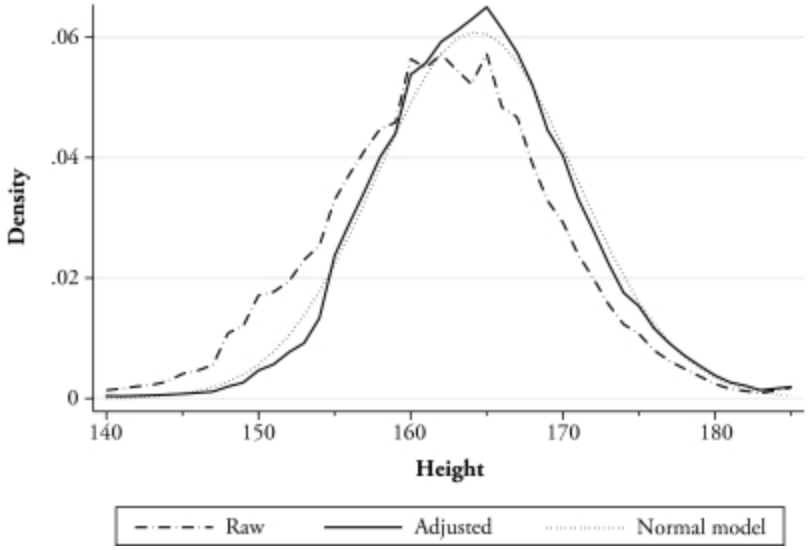
The purpose of this exercise is to

- Introduce the normal distribution (and as extension material, the central limit theorem)
- Use a spreadsheet to investigate the the normal distribution
- Investigate the distribution of the means of a group of numbers - the “standard error”

The following data is from

A'Hearn, B., Peracchi, F., & Vecchi, G. (2009). Height and the normal distribution: evidence from Italian military data. *Demography*, *46*(1), 1-25.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2831262/

It shows the distribution of the heights of 20 year old men entering Italian military service in 1920. The heights are approximately normally distributed, with a mean of about 164 cm, and a standard deviation of about 4 cm.
The article is investigating differences in the heights of men born at different times and in different areas of Italy. This involves comparing the means of different datasets.




In this lab we will generate similar (but synthetic) height data so that we know the true underlying mean and standard deviation, and investigate the properties of the mean values that we extract from our artificially generated data.

Tammy Humphrey (2018)

# Instructions

1. Open a spreadsheet in Google sheets.

2. Generate a normally distributed random number (to represent the measurement of the height of one person).
   To do this: click on the cell A1 and paste the command *=norminv(rand(),164,4)*. This generates a single random number drawn from a Normal distribution with mean 164 and standard deviation 4.

3. Copy the formula into the range A1:D1000 (i.e. create a block of 4000 random numbers, in four columns and 1000 rows).
   To do this: First copy the formula in cell A1 using CTRL V, then holding down the shift key then clicking on the 'A', 'B', 'C' and 'D' at the top of the first four columns. This selects all cells in the first four columns (here 1000 cells as this is the default number in the spreadsheet). Press CTRL V to paste the formula into all these cells.

4. Highlight the range A1:A1000 (by clicking on the top of the column) and choose insert->chart. This should put in a histogram. If it does not, select the histogram option for the type of chart. Under the customise -> histogram tab, you can set the size of each histogram bin. I suggest you use 1.

5. Find the mean of the numbers in row 1, columns A to D for all 1000 rows, so that you have 1000 numbers in column E, each of which is the average of 4 different heights.
   To do this: Select cell E1 then paste =AVERAGE(A1:D1) which calculates the average of the numbers in cells A1 to D1. Click on cell E1, and press CTRL V to copy this formula. Select all the cells in column E then paste this formula into all these cells.

6. Add the range E1:E1000 to the histogram plot
   To do this: Double click on the histogram. Select DATA and click "Add series". Type E1:E1000.

7. Add a cell that finds the standard deviation of the values in A1:A1000
   To do this: Click F1 (for example) and paste in =STDEV(A1:A1000)

8. Add a cell that finds the standard deviation of the values in E1:E1000
   To do this: Click G1 and paste in =STDEV(E1:E1000)

Tammy Humphrey (2018)

## Questions

What is the relation between the standard deviation of the numbers generated by the *=norminv(rand(),164,4)* command (i.e. columns A to D), and the standard deviation of the numbers generated by averaging 4 of those numbers (Column E) ? Is this what you would expect?

Why do we often take the average of several measurements ?

Can we be sure that the average height of Italian men has increased in the last 100 years? (to be continued in Module 3!)

## Extension - The Central Limit Theorem

Not all data is normally distributed, yet the normal distribution is assumed in a lot of statistics. Why is this ?
Instead of starting with normally distributed data, open a new sheet (by clicking the '+' symbol at the bottom left) and generate uniformly distributed random numbers on the interval 0 to 50 with 50*rand(). Plot these values as a histogram with a "bin size" of 2.
Make 9 more columns in a similar way (so you have 10 columns of data altogether)
In the next column, find the average of the 10 columns of data (so you have a whole column of data which is averages of 10 values). Plot this data on the same histogram (with a bin size of 2)

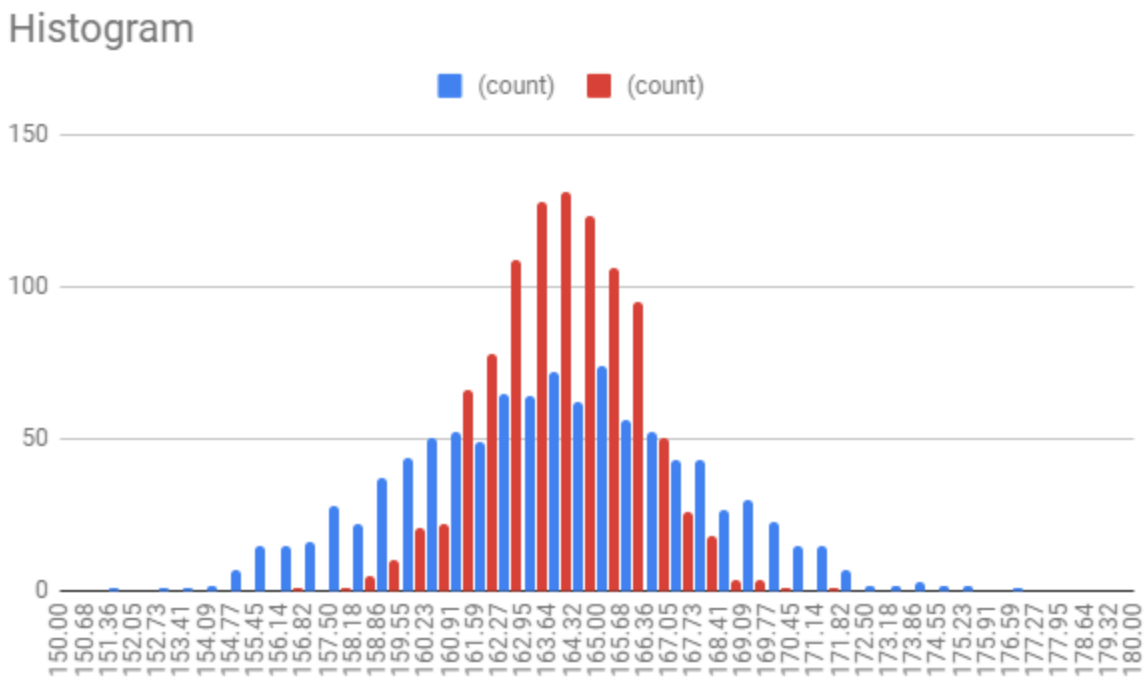
What does the distribution of the averages look like ?
What if you average more values ?

Tammy Humphrey (2018)

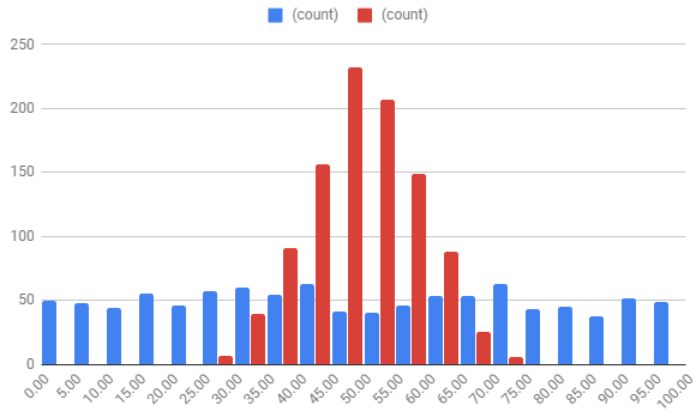
# Answers

First page calculations:
Average of 10 heights: 176.6cm StDev: 4.2cm

Histogram:
Heights of Italian men in 1920 (blue). The standard deviation of this data is 4cm
Mean height of 4 Italian men in 1920 (red). The standard deviation of the mean is $\sigma_{\bar{x}} = 2cm$, which is as predicted by our formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4cm}{\sqrt{4}} = 2cm$




Central limit theorem histogram:




Tammy Humphrey (2018)