Module 5

Sampling: Randomly Representative

James Baglin

Last updated: 13 July, 2020

Overview

##Summary

Often a population is immeasurable. Therefore, statistical investigations must often rely on the use of a sample from the population. This module will introduce sampling from populations.

Learning Objectives

The learning objectives associated with this module are:

- Describe the purpose of sampling for estimation and inference.
- Define and distinguish between different sampling methods.
- Define a sampling distribution for a statistic.
- Define the expected value and variance of a sampling distribution.
- Use technology to simulate and explore the characteristics of sampling distributions.
- Define and apply the Central Limit Theorem (CLT).

Module Video

This a nice video that explains the challenge of sampling in ecology research. The video discusses populations, samples and random samples.

What's up, buttercup? Population sampling techniques | A ...

Populations and Samples

In this module we will dive deeper into the world of inferential statistics first introduced back in Module 1. Recall, statistical inference refers to methods for estimating population parameters using sample data. A **population** is the larger group that a researcher wants to generalise their results to. For example, a researcher may need to know the average battery life for a new model of mobile phone, or estimate the average transfer speeds for a new computer hard disk drive. It would be too expensive or infeasible to test every unit manufactured. Therefore, the researcher must use another method.

A researcher's confidence in their ability to infer what's happening in the population comes down to the quality of the sample and quality of the data collected. In this section we will deal with samples. A **sample** is a smaller subset of a population. If the sample is chosen appropriately, it can provide a fairly accurate account of the population. Why do we use samples? Cost, time and practical constraints often make measuring the population impossible. Think back to the mobile phone and hard disk drive example in the previous paragraph. When an entire population is measured, it is called a **census**. The Australian Bureau of Statistics

(ABS) conducts the Australian Census every five years at an estimated cost of $440 million (Based on 2011 Census). As you can understand, this amount of time and money is well beyond the means of most statistical investigations.

There are good and bad ways of gathering samples. Probability-based methods maximise the chances of gathering a randomly representative sample. Common probability-based methods include simple random sampling, cluster sampling and stratified sampling. Non-probability based methods make no effort to ensure the sample is randomly representative. The best example of these types of methods are convenience sampling, purposive sampling, quota sampling and snowballing. Let’s take a closer look at the probability-based methods.

Sampling Methods

Simple Random Sampling (SRS)

In SRS, every unit in a population has an equal chance of being selected. For example, every new model mobile phone manufactured has an equal chance of being selected to undertake a battery test. This is the most simple and effective probability-based sampling method. However, it can be tricky to implement. For example, if we were looking at the population, how do we get a list of every single Australian so you can ensure everyone has an equal chance of being selected? A phone book is a good start, but what about people without landlines? This course will focus mainly on simple random sampling. Watch the following video by Steve Mays for a nice overview of SRS.

Stratified Sampling

Stratified sampling divides the population into subpopulations, called strata (e.g. gender, age bands, ethnicity), and then takes a SRS from each strata proportional to the strata’s representation in the population. For example, the Australian population is approximately 49% male and 51% female. A stratified sample for gender would divide the population into males and females and then proceed to take SRSs of males and females so the resulting sample is approximately 49:51 male:female. Stratified sampling can be more complex as there is no limit to the number of strata and levels within the strata. For example, a researcher may wish to stratify the population by gender, age bands and ethnicity. This would result in a sample that is more likely to be representative, but would require substantially more time and effort.

Cluster Sampling

Cluster sampling first divides the population into naturally occurring and homogeneous clusters, e.g. postcodes, towns, manufacturing batches, factories, etc. The investigator then randomly selects a defined number of clusters. For example, referring back to the hard disk drive example, the company may have manufactured 100 batches of hard disk drives on different days. They may decide to randomly select 10 batches which they define as the clusters. Using these randomly sampled clusters, the investigator would then proceed with the use of SRS within each cluster to select their sample. For example, the investigator might decide to randomly select 10 hard disk drives from each of the 10 batches making a total sample size of 100. Cluster sampling can be more economical and less time-consuming than SRS. This is because the researcher is required to perform SRS only within a limited number of clusters and not the entire population.

Convenience Sampling

Convenience sampling methods, or non-probabilistic sampling, make no effort to randomly sample from the population. Therefore, the degree to which a convenience sample is randomly representative to the population is always unknown. Convenience samples have a high probability of being biased. A **biased sample** is a sample that cannot be considered representative of the target population. It is possible for a convenience sample to be representative, but the probability is always unknown. Substantial caution must be placed on inferences drawn from the use of convenience samples. Regardless, convenience samples are probably the most common samples used in research due to their low cost and relative ease. Very few researchers have the time and money to use probabilistic methods. That's not to say you shouldn't try, but if you're forced to use a convenience sample, you should always note its limitations.

It's important to note that probability-based sampling methods do not guarantee a representative sample either. That's why we say the sample is **randomly representative**. There is still uncertainty. This is particularly true for small samples. We can take another look at the info-graphic provided by Wild, Pfannkuch and Horton (2011) that looks at populations, samples and sample size. Note that Wild et al. are referring to probability-based sampling methods.

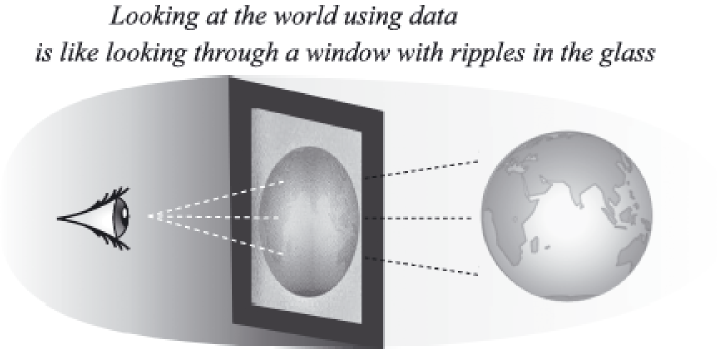

Looking at the world using data
is like looking through a window with ripples in the glass

Fig. 2. 'What I see is not quite the way it really is'

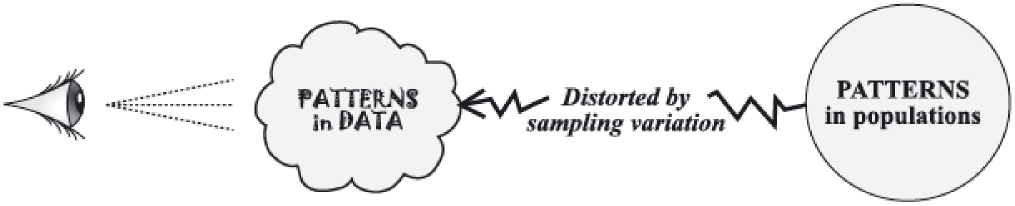


Fig. 3. Distortions due to sampling

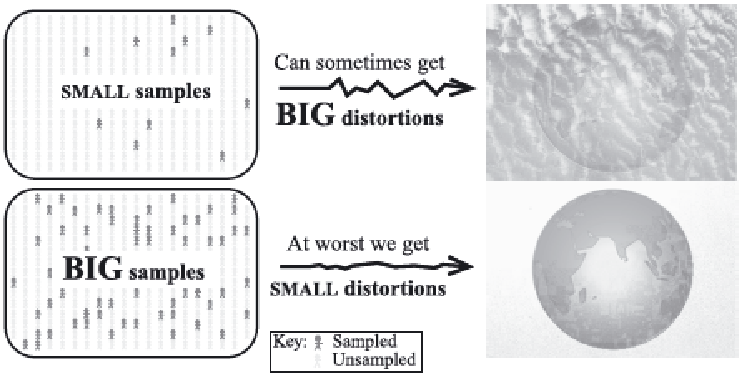


Fig. 4. Distortions related to sample size

Figures 2 - 4 have been taken from Wild, Pfannkuch and Horton (2011).

So, from this we can conclude the following:

The larger a random sample, the more likely it is to represent the population.

This is an important lesson. Sample size does matter and should always be considered an important consideration when planning an investigation. In the next section will be explore this concept further when we consider sampling distributions.

Sampling Distributions

Take a random sample from the population, of say size $n = 100$, measure a quantitative variable on each unit and calculate the sample mean. Write down the sample mean in a data recording sheet. Now place the sample back into the population and take another random sample of the same size. Measure your variable, calculate the sample mean, and record its value in the same recording sheet. The sample mean won't be the same, because it's a slightly different sample. Remember, this is called sampling variability or error. Now, repeat this process many, many times, say one thousand. Now, take all the one thousand sample means you recorded and plot them using a histogram. This histogram of the sample means is an example of a **sampling distribution**.

A **sampling distribution** is a hypothetical distribution of a sample statistic, such as a mean, median or proportion, constructed through the repeated sampling from a population. A sampling distribution describes what would happen if we were to repeat a study many times over and for each replicated study we recorded a summary statistic. Sampling distributions are influenced by two major factors. The first factor is the underlying distribution of the random variable. For example, if your random variable is distributed normally, binomially, or exponentially, this will have an effect on the characteristics of the sampling distribution. The second major factor is the sample size n. The sample size of the hypothetical repeated studies has some interesting effects on the sampling distribution, as we will discover shortly.

The Cal Poly Sampling Distribution Shiny app (https://calpolystat1.shinyapps.io/sampling_distribution/) will allow you to commence exploring sampling distributions.

Activity 1

1. Set the following inputs:
    - **Population Distribution**: Normal
    - **Population mean**: 0
    - **Population standard deviation**: 1
    - **Sample size**: 10
    - **Statistic**: Mean
    - **Number of samples**: 1
2. Click **Draw samples**. This draws one random sample ($n = 10$) from the population. The sample values are displayed in the first histogram. The sample mean is calculated and plotted on the sampling distribution of the mean plot.
3. **Number of samples**: 1000
4. Click **Draw samples**. The app will quickly draw 1000 random samples and plot their means. Note the difference between a sample distribution versus a sampling distribution of the mean.

Activity 2

1. Now let's change the underlying population distribution and increase the sample size. Set the following inputs:
    - **Population Distribution**: Left-skewed
    - **Population mean**: 0
    - **Population standard deviation**: 1
    - **Sample size**: 100
    - **Statistic**: Mean
    - **Number of samples**: 1000
2. Click **Draw samples**. Note that while the population distribution is heavily left-skewed, the sampling distribution of the means is not. You will learn more about this when we look at the *central limit theorem*.

YouTube Data

We will use the `YouTube.csv` (data/Youtube.csv) data to explore the concepts of a sampling distribution. The data were originally sourced from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Online+Video+Characteristics+and+Transcoding+Time+Dataset). The dataset contains the basic video characteristics of over 24,000 YouTube clips. The variables are as follows:

- **id**: Youtube vide id
- **duration**: duration of video
- **bitrate**: bitrate(total in Kbits)
- **bitrate.video**: bitrate(video bitrate in Kbits)
- **height**: height of video in pixles
- **width**: width of video in pixles
- **framerate**: actual video frame rate
- **frame.rate.est.**: estimated video frame rate
- **codec**: coding standard used for the video
- **category**: YouTube video category

Here is a random sample of the full dataset:

Show 10 ▼ **entries** **Search:**

	id	duration	bitrate	bitrate.video.	height	width	f
1	KUWH-dsjQPA	82	628	500	480	360	
2	vT-f-S2zVnM	39	581	454	640	480	
3	QhEd-ifJg60	39	1500	1421	720	1280	
4	dEMF-bnF7dA	42	2547	2400	720	1280	
5	ubtH-GrxEmI	288	4102	3945	1920	1080	
6	uO9o-NooIJk	196	486	411	480	360	
7	iVaw-wy8G58	31	236	126	176	144	
8	90SdmjuCAqw	249	2163	2046	1280	720	
9	GyIT-ZmRySA	27	315	255	320	240	
10	fVj9-aQBCAw	200	634	501	480	320	

Showing 1 to 10 of 100 entries

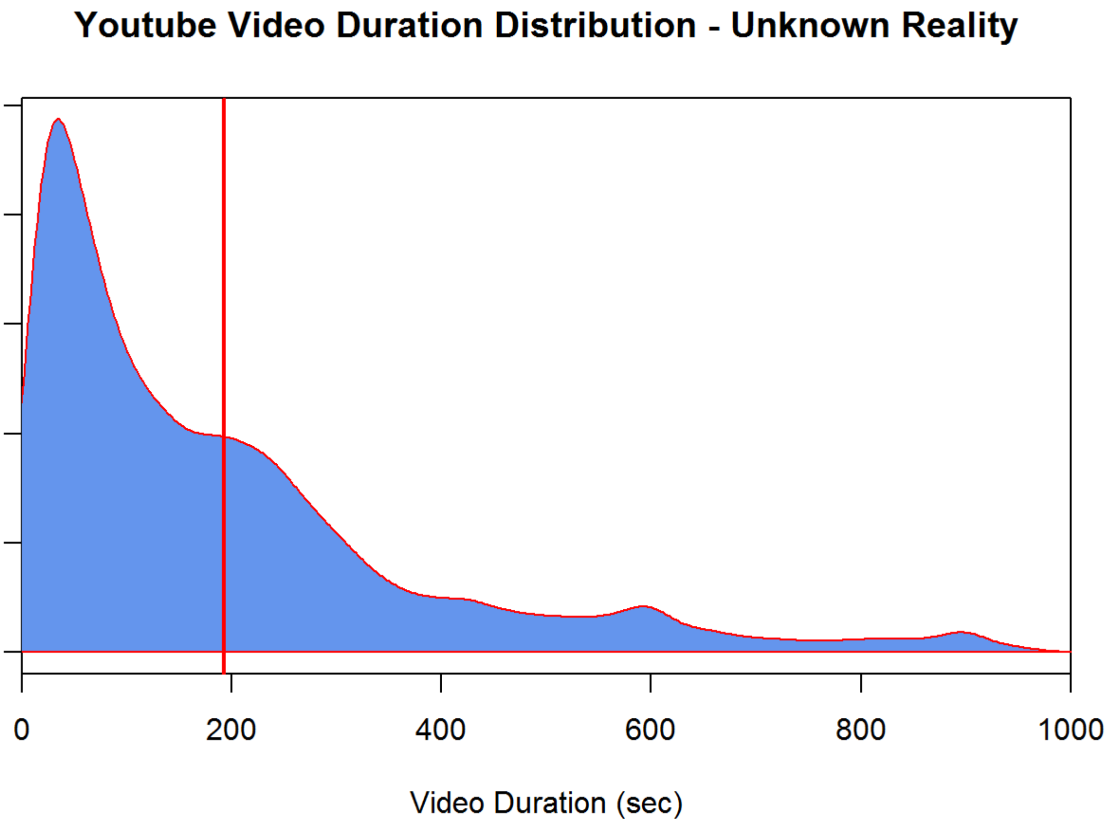
Previous 1 2 3 4 5 … 10 Next

Population Distribution

The `Youtube.csv` data will be treated as the unknown population. As the dataset contains over 24,000 video characteristics, this isn't too difficult to imagine, even though the total population size of YouTube is much, much, much higher (I am yet to find a credible estimate! If you find one email me). For the sake of the example, we will imagine this to be the unknown reality that we are trying to estimate. We will look at estimating the average YouTube video duration, measured in seconds (sec). The data has been cleaned to enable a better a visualisation of the population distribution. Extreme outliers (Duration > Q_3 + [$IQR * 3$]) have been removed to help lessen the extent of the extreme values in the right tail of the distribution. I used the following R code and saved the filtered data object as `YouTube_clean`.

```
YouTube_clean <- YouTube %>% filter(duration < (281 + ((281-52)*3)))
```

This step only removed around 3% of the original data. The YouTube video duration distribution is visualised in the following density plot. A density plot is similar to a histogram, but uses a smoothing algorithm to remove the need for bins. The mean is depicted using a red line. The distribution is skewed to the right. Visually:



Here are the population's parameters:

```
YouTube_clean$duration %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1      50     131     193     261     966
```

The population mean, which we denote as μ, rounds to 193 secs (or 3 minutes 21 secs). However, variability is high with a population standard deviation, which we denote as σ, of 193 secs.

Simulations in R

We can use R to simulate repeated random sampling from the population of YouTube video durations. Let's run a simulation that generates 10,000 random samples of size $n = 10$. The simulator will save the sample means in order to create a sampling distribution. Here's a simple simulator:

```
set.seed(123456) #Set random seed number to allow replication

n <- 10 #set sample size

sims <- 10000 #Set number of random samples to be drawn

x_bar <- data.frame(x=as.numeric()) #Create a data.frame to store sample means
  from simulation

for (i in 1:sims) { #create a loop to perform the simulations
  samp <- YouTube_clean %>% sample_n(n) #Generate a random sample
  x_bar[i,1] <- samp$duration %>% mean() #Store the sample mean in the data.fr
  ame
}

x_bar$x %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    34.5   148.6   187.0   192.7   231.8   476.4
```

The `set.seed()` function forces R’s built-in random number generator to start at a particular seeding value. This ensures that others, like yourself, can re-run this code, and get the same results. If we used a different seed, the results would be a little different. Try for yourself if you wish.

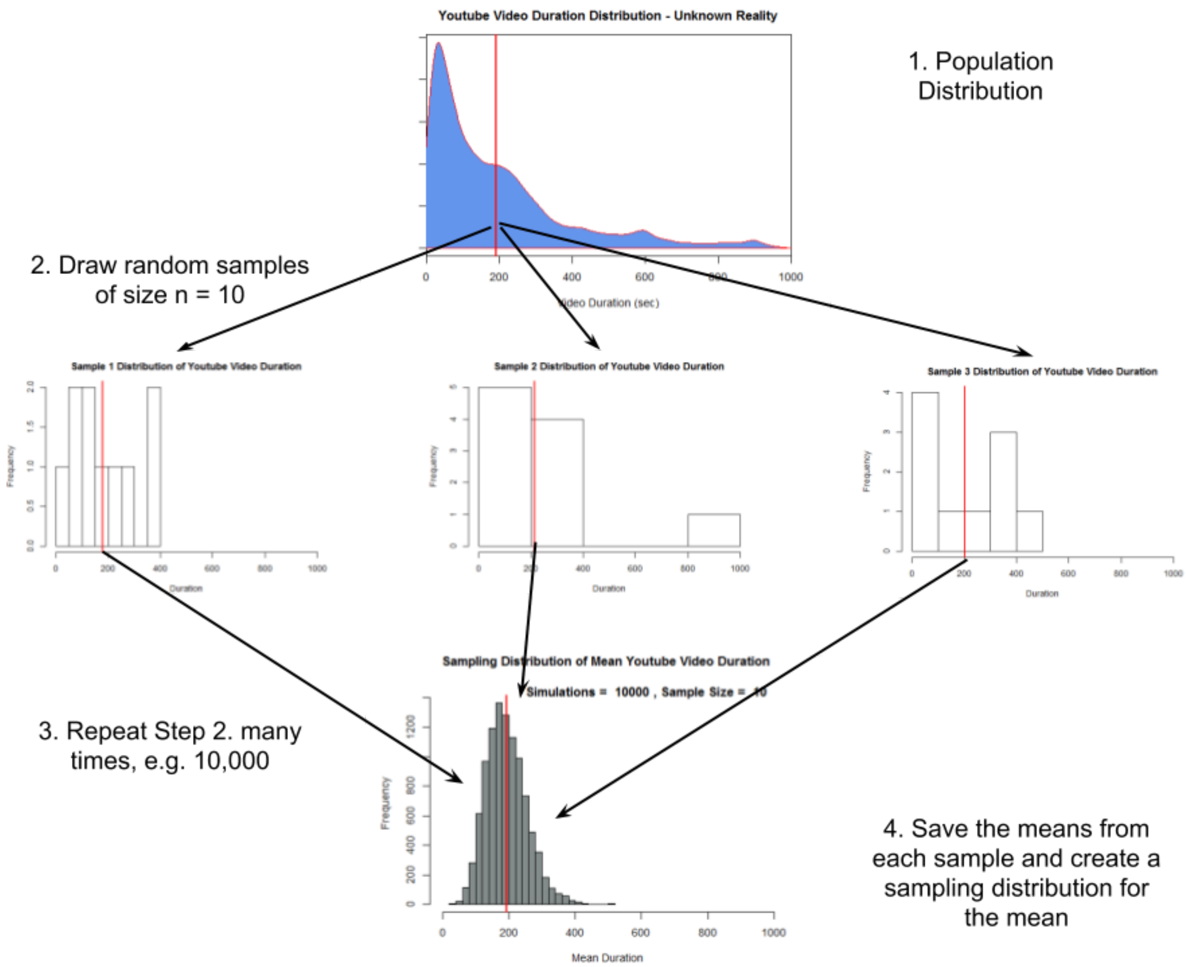
Next, we set the sample size `n <- 10`, and the number of random samples to draw from the population distribution, `sims <- 10000`. Generally, we should do this at least 10,000 times, but keep in mind that requires computational time.

Next, we want to same the results of the simulation to an object, which we will define as a `data.frame` named `x_bar`. The loop function, `for (i in 1:n){}` tells R to loop through as function starting at `1` and ending at `n`. We use the sims object to tell R who many times to loop. If we change the sims object, we can quickly change the number of simulations, e.g. `sims <- 1000` or `sims <-100000`.

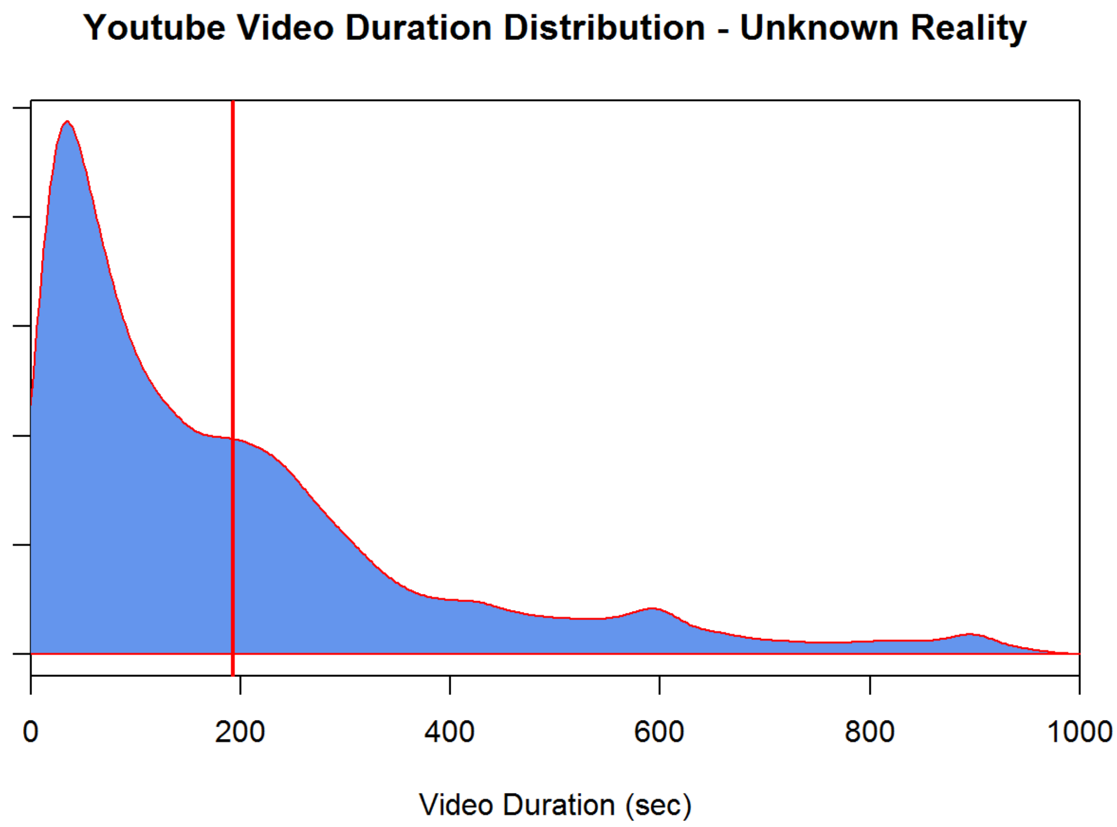
Inside the loop, we use the `sample_n` function to take a random sample of size `n` from the `YouTube_clean` data frame. Then we take the mean duration of the sample and assign it to the i^{th} row in the data frame. i is set as the loop number.

Once the loop finishes, the simulation is complete, and we can use the `x_bar` object to analyse the simulated sampling distribution of the mean.

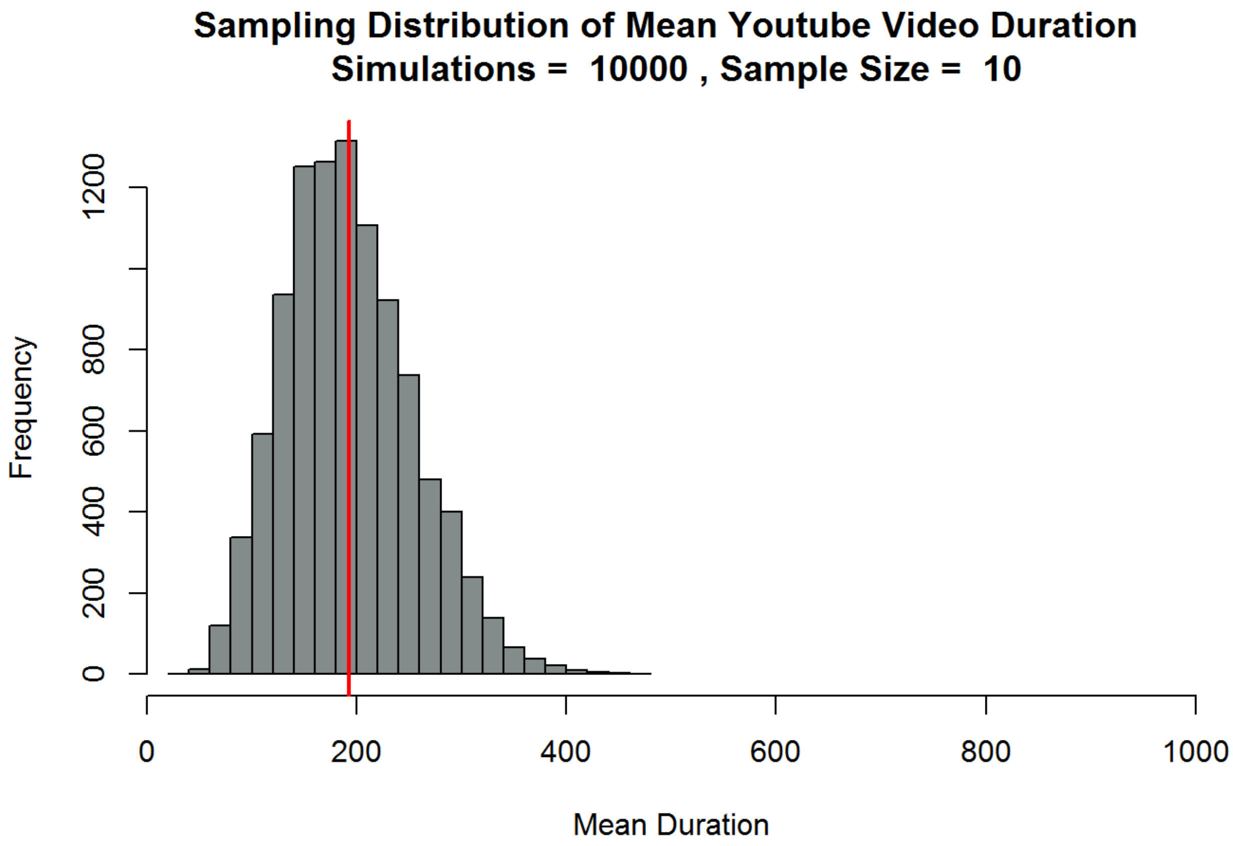
Let’s visualise these steps:


Youtube Video Duration Distribution - Unknown Reality
Video Duration (sec)
0
200
400
600
800
1000
1. Population Distribution
2. Draw random samples of size n = 10
Sample 1 Distribution of Youtube Video Duration
Sample 2 Distribution of Youtube Video Duration
Sample 3 Distribution of Youtube Video Duration
Frequency
Duration
3. Repeat Step 2. many times, e.g. 10,000
Sampling Distribution of Mean Youtube Video Duration
Simulations = 10000 , Sample Size = 10
Frequency
0
200
400
600
800
1200
Mean Duration
4. Save the means from each sample and create a sampling distribution for the mean

Let’s take a closer look at the sampling distribution visualised using a histogram. Also included for comparison is the population distribution. Pay attention to the differences. Notice how the variability of the sampling distribution is much smaller and the mean of the sampling distribution is approximately the same as the population mean?



Youtube Video Duration Distribution - Unknown Reality



Sampling Distribution of Mean Youtube Video Duration
Simulations = 10000 , Sample Size = 10

Expected Value and Variance

These observations introduce two important concepts related to the sampling distributions. For sampling distributions of the mean, the expected value, $E(\bar{x})$, variance, $Var(\bar{x})$ and standard deviation, $\sigma_{\bar{x}}$, are as follows:

$$E(\bar{x}) = \mu_{\bar{x}} = \mu$$

$$Var(\bar{x}) = \sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

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

where $\bar{x}$ refers to a sample mean, and n = sample size. Let's demonstrate that this is true. The population parameters are as follows:

$$\mu = 193$$

$$\sigma^2 = 193^2$$

$$\sigma = 193$$

Note that, conincidently, $\mu = \sigma = 193$.

According to the formula above, the mean, variance and standard deviation of a sampling distribution of the mean using a sample size $n = 10$, are...

$$E(\bar{x}) = \mu_{\bar{x}} = 193$$

$$Var(\bar{x}) = \sigma^2_{\bar{x}} = \frac{\sigma^2}{n} = \frac{193^2}{10} = 3724.9$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{193}{\sqrt{10}} = 61.03$$

Now, keeping in mind that we used a simulation, and we expect there to be some random error (especially for the variance), recall the descriptive statistics of the sampling distribution simulated in R...

```
x_bar$x %>% mean()
```

```
## [1] 192.7484
```

```
x_bar$x %>% var()
```

```
## [1] 3723.879
```

```
x_bar$x %>% sd()
```

```
## [1] 61.02359
```

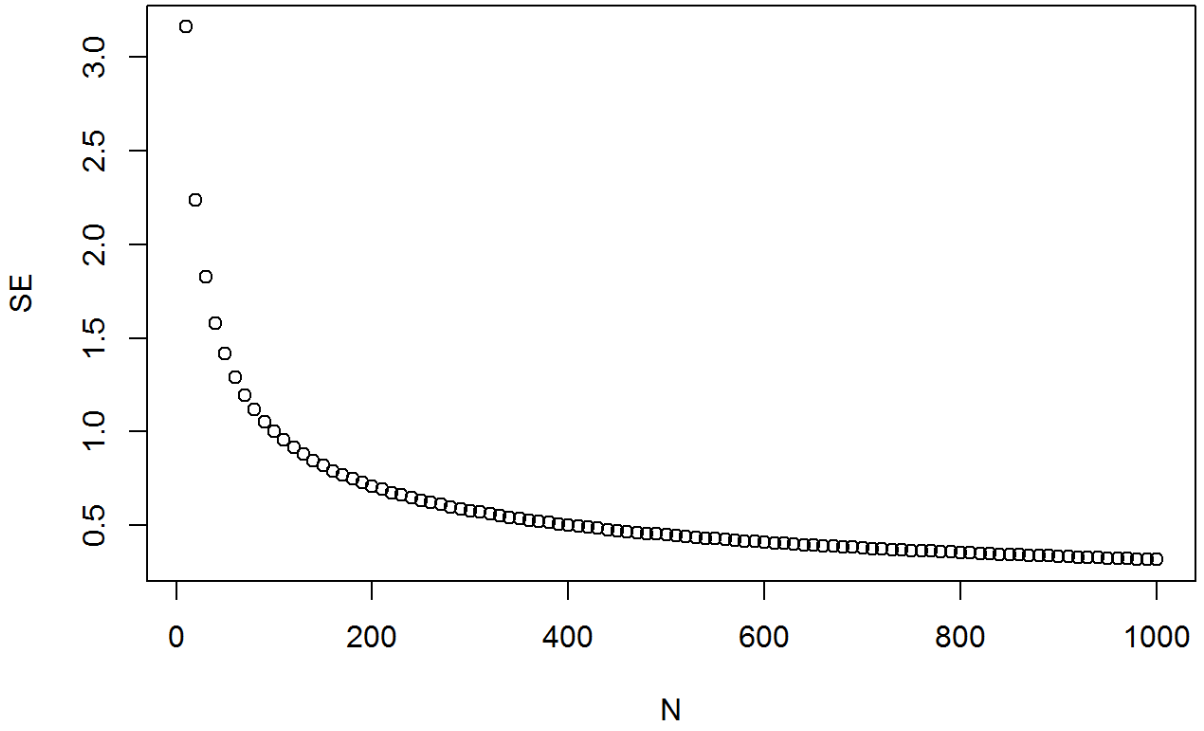
These simulation estimates are very close! We get closer if we increase the simulation size, but this becomes impractical due to the extra computational time.

Standard Error

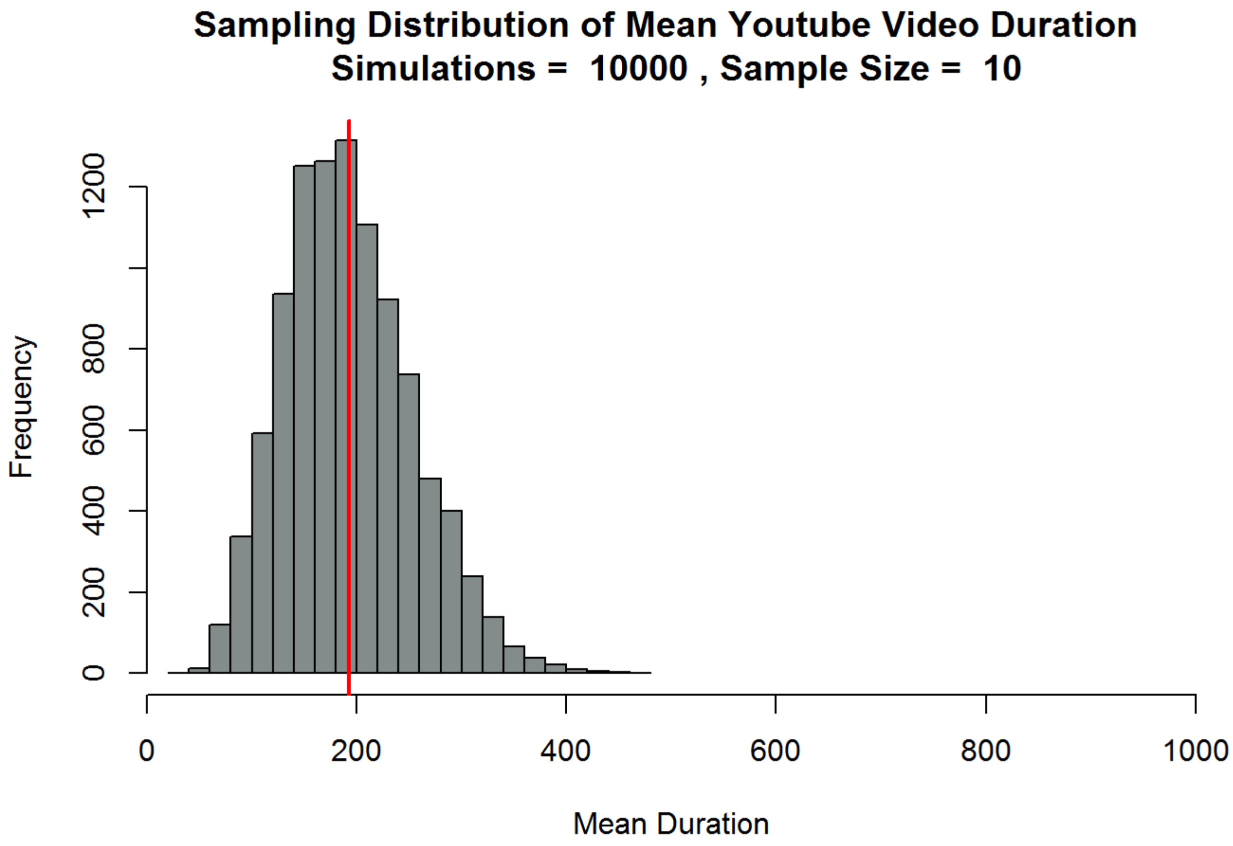
The standard deviation for a sampling distribution is known as the **standard error (SE)**. So, we could write the standard error for the mean as:

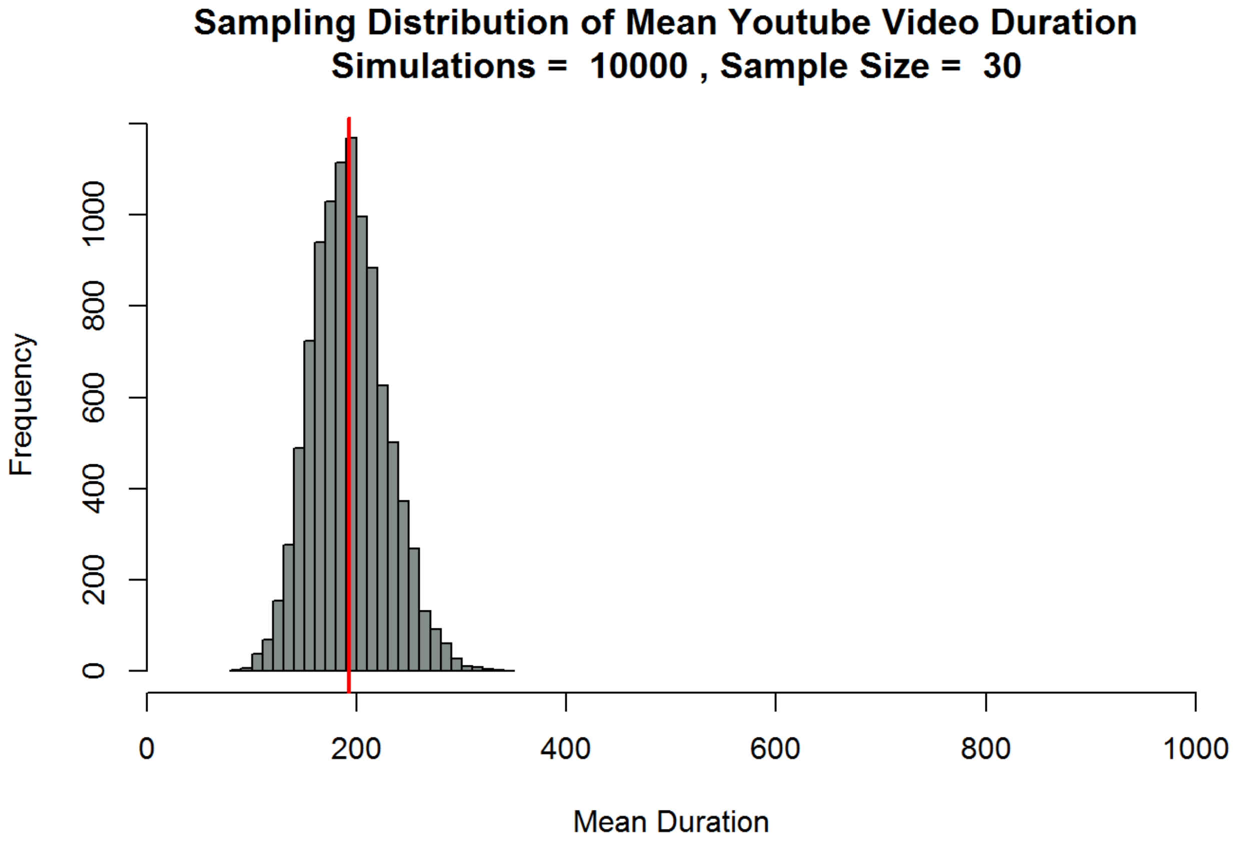
$$SE = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

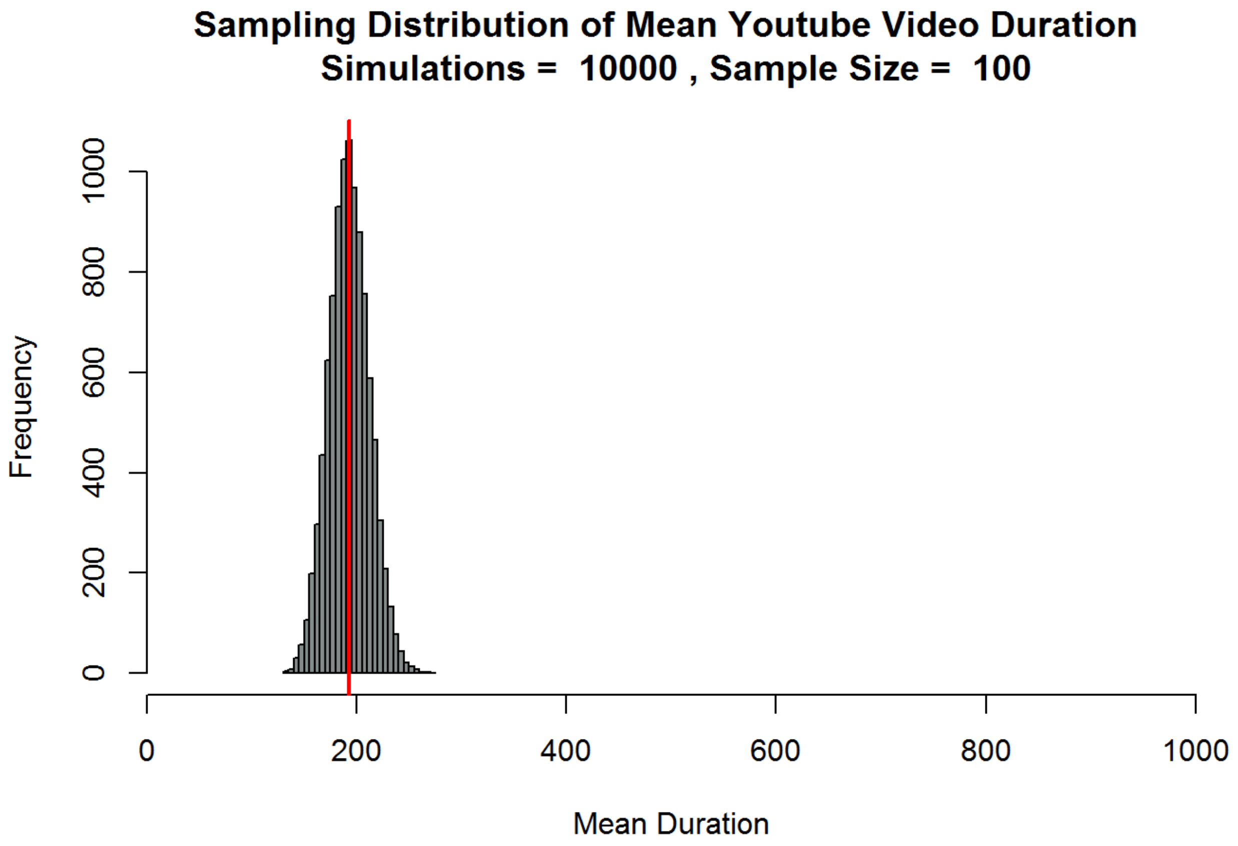
The size of the sample and the standard error share an inverse relationship. As sample size increases, the SE for the mean decreases. Consider the following plot visualising this relations assuming $\sigma = 10$.



Why? Larger random samples provide more reliable estimates of population parameters, therefore, less error. Let’s demonstrate this further by running the simulation outlined above for three different sample sizes, n = 10, 30, and 100. The results for the simulations are summarised in the following three histograms:


Sampling Distribution of Mean Youtube Video Duration
Simulations = 10000 , Sample Size = 10
Frequency
0
200
400
600
800
1200
0
200
400
600
800
1000
Mean Duration


Sampling Distribution of Mean Youtube Video Duration
Simulations = 10000 , Sample Size = 30
Frequency
0
200
400
600
800
1000
0
200
400
600
800
1000
Mean Duration



As you can see, as the sample size increases, the standard error decreases. You might also detect that the shape changes from being slightly skewed to symmetric. This brings us to the next important concept.

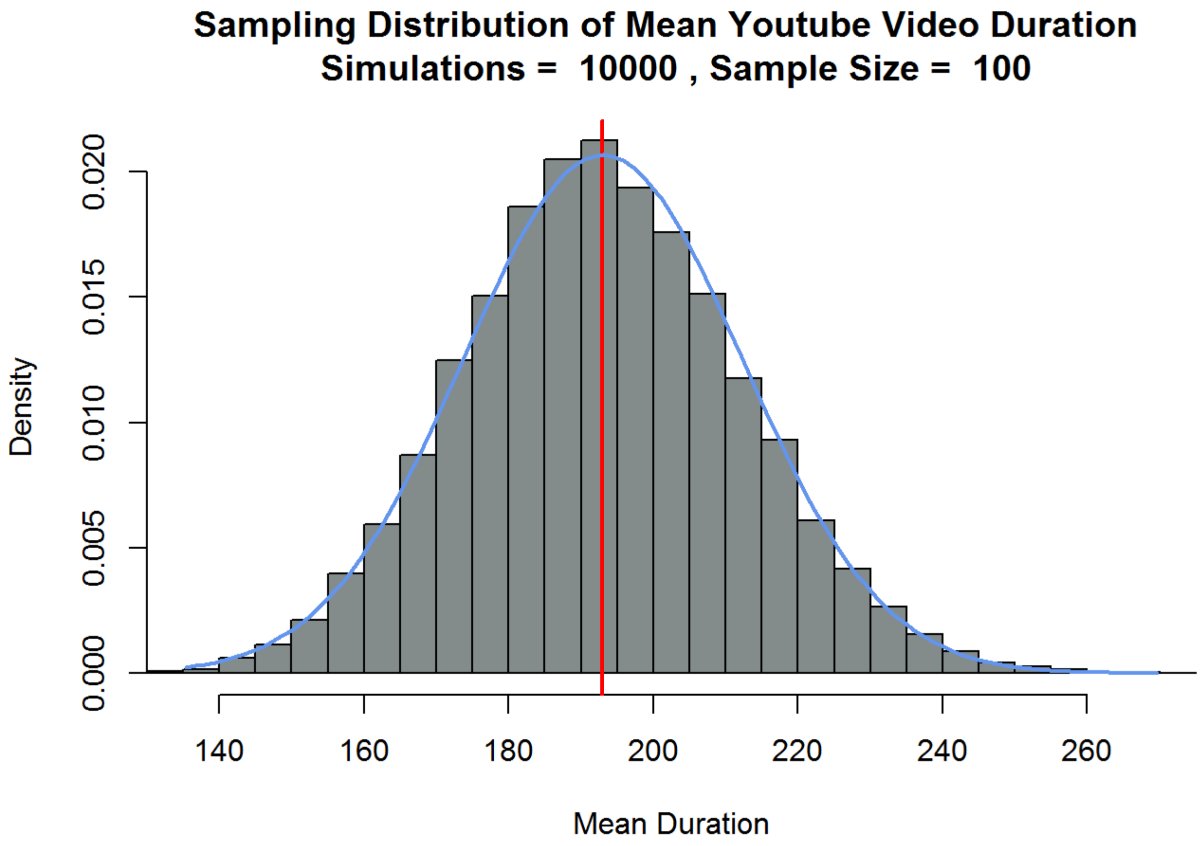
Central Limit Theorem

There are a few useful rules we need to know about sampling distributions of the mean. **If the underlying population distribution of a variable is normally distributed, the resulting sampling distribution of the mean will be normally distributed**. This rule is referred to the **Central Limit Theorem** and can be written as:

$$\text{If } x \sim N(\mu, \sigma) \text{ then } \bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

This makes sense. However, what if the population distribution isn't normally distributed as was the case with YouTube video durations? We got a hint in the previous figure. Let's take a closer look.

We use our simulator to create a sampling distribution of the mean duration using 10,000 samples of size 100. We plot the distribution and overlay a hypothetical normal distribution (blue line) with $\mu_{\bar{x}} = 193$ and $\sigma_{\bar{x}} = 19.3$:



The blue (theoretical) line is a near perfect fit to the sampling distribution. This is another important property of the Central Limit Theorem. **When the sample size we use is large, typically defined as $n > 30$, the sampling distribution of the mean is approximately normal, regardless of the variable's underlying population distribution**.

What are sampling distributions used for?

We will start using sampling distributions in greater depth to help us answer interesting questions about the results of statistical investigation later in Module 7 (MATH1324_Module_07.html). For now, we go through a few examples to start getting a sense of how sampling distributions are used in statistics. We will use the CLT to quickly calculate probabilities of observing different sample means for various sample sizes, assuming the population mean and standard deviation of YouTube video duration is 193 secs.

1. What is the probability of randomly selecting a sample of size $n = 100$ that has a sample mean duration of less than 150 secs?

Because we have a large sample size, we can invoke the CLT, which means the sampling distribution of the mean can be approximated as:

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) = N(193, 19.3)$$

Now we can use R's normal distribution functions to determine $Pr(\bar{x} < 150)$. In R we use the formula:

```
pnorm(q = 150, mean = 193, sd = 19.3)
```

```
## [1] 0.01294095
```

The probability, $Pr(\bar{x} < 150)$), was found to equal .013. Therefore, there is a 1.3% chance that an investigator will randomly select a sample with a mean below 150 secs. In other words, this would be unusual.

2. What is the probability of randomly selecting a sample of size n = 100 that has a mean duration greater than four minutes?

We need to find $Pr(\bar{x} > 240)$. In R, we use the formula:

```
pnorm(q = 240, mean = 193, sd = 19.3, lower.tail = FALSE)
```

```
## [1] 0.007441098
```

The answer is found to be $Pr(\bar{x} > 240) = 0.007$.

3. What is the probability of randomly selecting a sample of size n = 200 that has a mean duration greater than five minutes?

We need to change the standard error as the example calls for a larger sample. We can do this directly in R using:

```
pnorm(q = 300, mean = 193, sd = 193/sqrt(200), lower.tail = FALSE)
```

```
## [1] 2.244519e-15
```

Note how we calculated SE directly in the formula using the `sqrt()` function. We find $Pr(\bar{x} > 300) = 2.244519e - 15$. What does that mean? When you see e-15, that means to move the decimal place to the right 15 places from 2.244519, so, $Pr(\bar{x} > 300) = 0.0000000000000002244519$. In other words, the probability is really, really small.

Why has the probability substantially dropped? As a larger sample size was used, the sampling distribution has a smaller standard error. Therefore, observing a sample mean duration of 300 secs would be very unlikely when the sample size was $n = 200$ vs. $n = 100$. Larger random samples provide more reliable estimates of population parameters.

References

Wild, C., M. Pfannkuch, M. Regan, and N. J. Horton. 2011. "Towards more accessible conceptions of statistical inference." *Journal of the Royal Statistical Society* 174: 247–95. https://doi.org/10.1111/j.1467-985X.2010.00678.x (https://doi.org/10.1111/j.1467-985X.2010.00678.x).