Lab 4: Sampling variability

STAT218

In this lab you'll explore sampling variation. There are two learning goals:

1. Demonstrate through simulation the variability of summary statistics, particularly the mean, across random samples.
2. Explore the effect of sample size on sampling variability.

We'll use 3,179 responses to the 2009-2010 NHANES and consider specifically the total cholesterol variable. We will treat these observations as a population, and simulate the effect of sampling on summary statistics by generating small to moderate subcollections of observations drawn at random without replacement.

The dataset has been stored as a separate file `nhanes.RData`. Load the dataset, extract the variable of interest, and preview the first handful of observations using the commands below.

```
# load nhanes dataset
load('data/nhanes.RData')

# extract total cholesterol column
cholesterol <- nhanes$TotChol

# view summary
str(cholesterol)
```

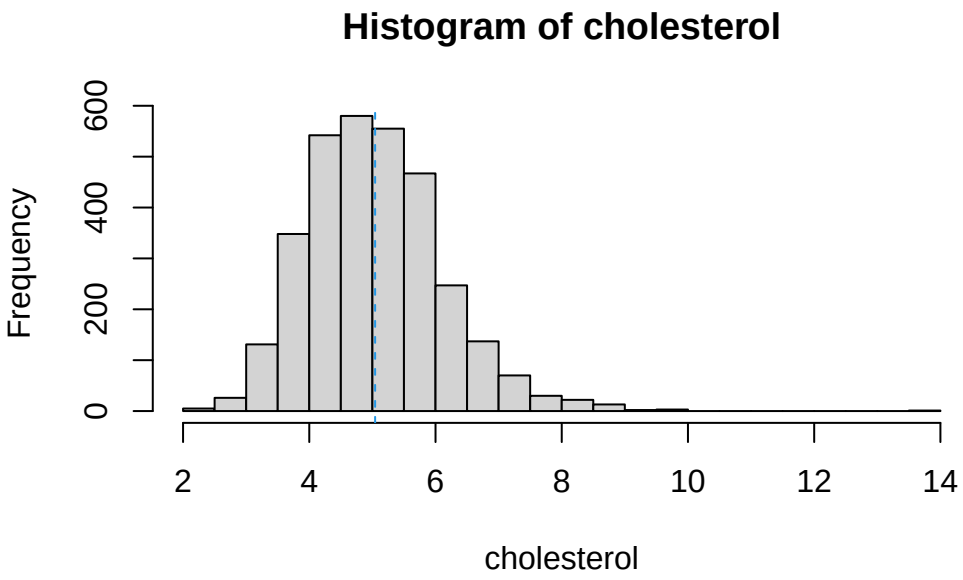
```
num [1:3179] 3.49 3.49 3.49 6.7 5.82 5.82 5.82 4.99 4.24 6.41 ...
```

Since we are pretending that these 3,179 respondents form a population, we can examine the population distribution of the cholesterol variable and also calculate population statistics.

```
# numeric summary
summary(cholesterol)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.330   4.240   4.970   5.043   5.690  13.650
```

```r
# compute population mean and sd
pop_mean <- mean(cholesterol)
pop_sd <- sd(cholesterol)

# construct histogram
hist(cholesterol, breaks = 30)
abline(v = pop_mean, col = 4, lty = 2)
```



Exploring sampling variability

The commands below will draw a sample, compute the mean and standard deviation, and compare them with the corresponding population values.

- does the histogram look similar to the population distribution?
- do the sample statistics closely align with the population values?

```r
# draw a sample -- try running this line a few times
sample(cholesterol, size = 50, replace = F)

# now store a sample
samp <- sample(cholesterol, size = 50)
```

```r
# calculate mean and sd
samp_mean <- mean(samp)
samp_sd <- sd(samp)
samp_mean
samp_sd

# estimation error
samp_mean - pop_mean
samp_sd - pop_sd

# make a histogram
hist(samp, breaks = 10)

# add lines at sample mean and population mean
abline(v = samp_mean, col = 2) # red line
abline(v = pop_mean, col = 4, lty = 2) # blue line
```

Because you are drawing a random sample, results will differ each time you run the above commands. Likewise, results will differ from your groupmates.

Your turn

Run the lines above and compare results with your group. Manually enter your means and standard deviations in a vector. Repeat and add entries until you have at least 6 sample means and 6 standard deviations. Discuss:

1. Does it appear that the sample statistics tend to be close to the population values?
2. Do the sample statistics vary much across samples? How might you measure this using your simulated means and standard deviations?

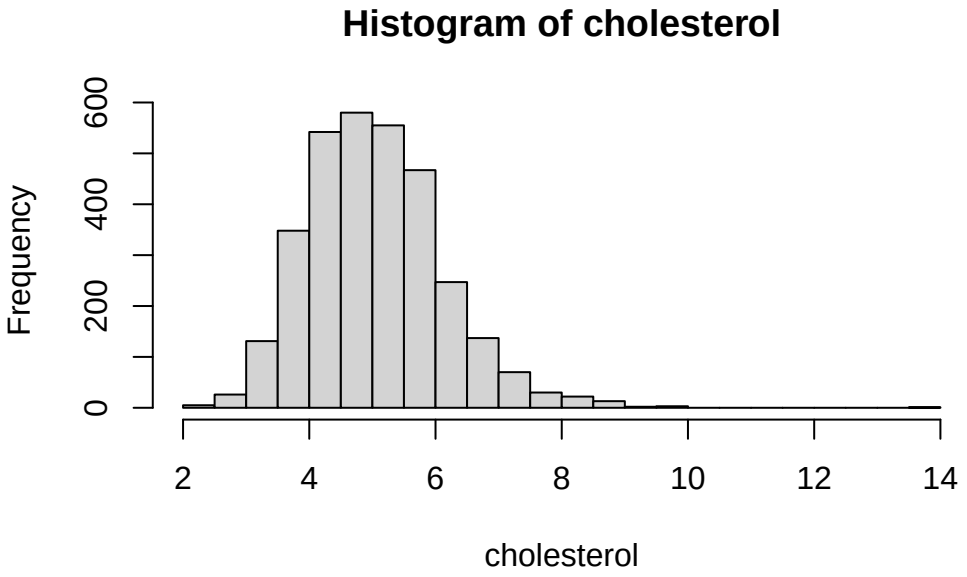
```r
# make vectors of 6 or more means and standard deviations
samp_means <- ...
samp_sds <- ...

# calculate errors
samp_means - pop_mean
samp_sds - pop_sd

# how would you measure variability?
```

Effect of sample size

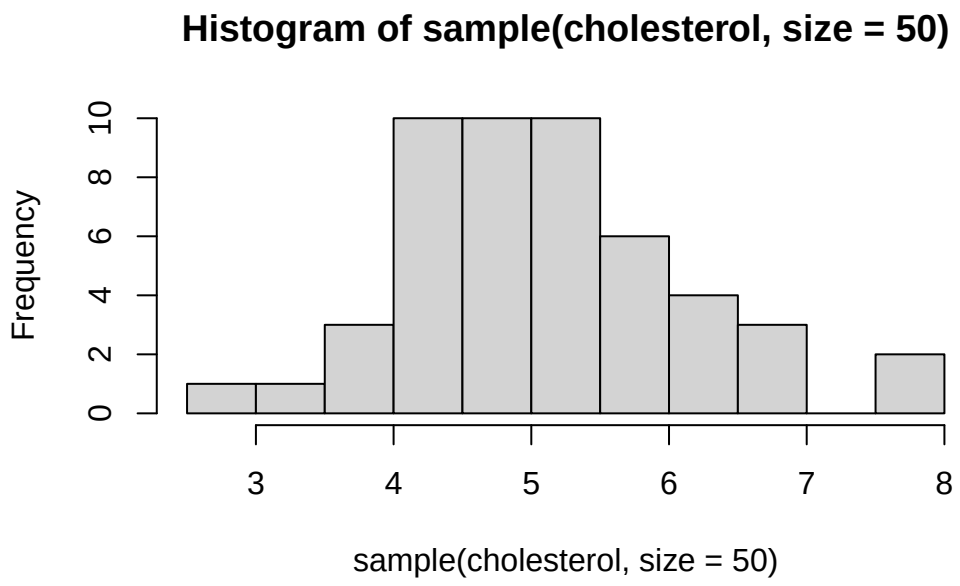
Now let's explore how sample size affects sampling variation of summary statistics.

First let's examine how the frequency distribution of a single sample changes when we increase the sample size. Notice that the histogram of the larger sample more closely matches the population distribution.

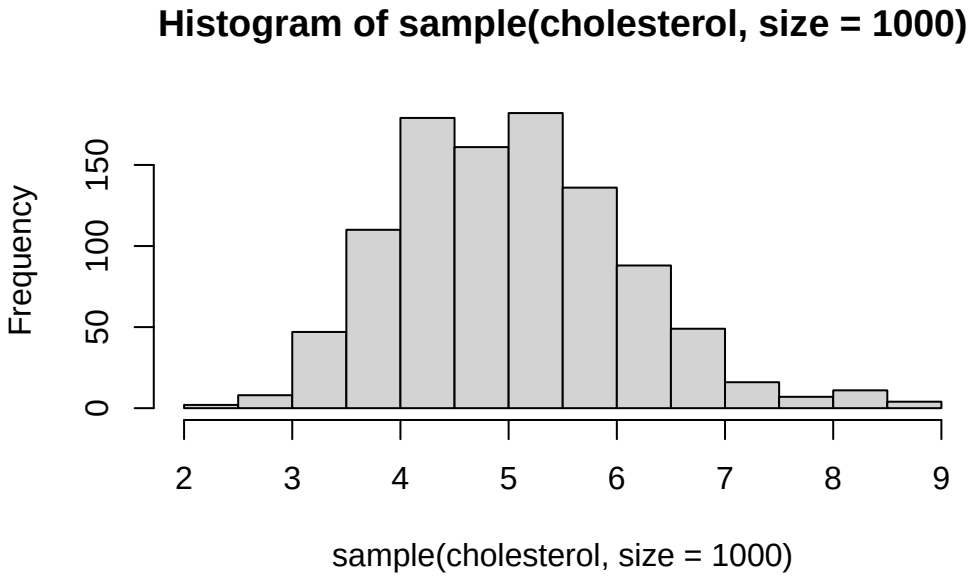
```
# population distribution
hist(cholesterol, breaks = 30)
```




```
# small sample size (try running a few times)
hist(sample(cholesterol, size = 50), breaks = 15)
```




```r
# large sample size (try running a few times)
hist(sample(cholesterol, size = 1000), breaks = 20)
```



We should therefore expect that summary statistics more closely approximate population values for larger samples. Our measure of sampling variability reflects this expectation. Recall that the theoretical standard deviation of the sample mean is:

$$SD(\bar{x}) = \frac{\sigma}{\sqrt{n}} \quad \left(\frac{\text{population SD}}{\sqrt{\text{sample size}}}\right)$$

This will diminish as n increases, indicating less sampling variation for larger samples. Let's explore that empirically.

The commands below simulate `nsim` samples of size `n` and calculate the mean of each sample. Don't worry about understanding the code that creates `samp_means`; your job is to use the results to measure sampling variability. First run this without any changes.

```
# number of samples to simulate
nsim <- 1000

# this generates nsim sample means from samples of size 10
samp_means_10 <- sapply(1:nsim,
                        function(i){
                          mean(sample(cholesterol, size = 10))
                        })

# repeat, but for samples of size 100
samp_means_100 <- sapply(1:nsim,
                         function(i){
                           mean(sample(cholesterol, size = 100))
                         })


# inspect
str(samp_means_10)
```

```
num [1:1000] 5.31 5.23 5.12 5.3 5.45 ...
```

```
str(samp_means_100)
```

```
num [1:1000] 5.14 5.09 5.01 5.07 5.12 ...
```

Effectively, you've just simulated a sampling distribution for the mean cholesterol of a sample of size 10 by generating means from lots of random samples. Notice already the additional variability for the smaller sample size — the means seem to deviate more often by a larger amount from the population value.

> **Your turn**
>
> Now try using your simulated means to measure the sampling variability of the mean; compare this with the theoretical standard deviation.

Do this for both the $n = 10$ and $n = 100$ cases. What changes?

```
# calculate standard deviation of simulated means
sim_sd_10 <- ...

# calculate theoretical standard deviation
theory_sd_10 <- ...

# calculate average error
avg_error_10 <- ...

# calculate standard deviation of simulated means
sim_sd_100 <- ...

# calculate theoretical standard deviation
theory_sd_100 <- ...

# calculate average error
avg_error_100 <- ...
```

Take note of the fact that in practice, it would not be possible to simulate sampling distributions in this way, because you'd lack data for a complete population. In practice, you'll only have *one* sample, and will need to use this to estimate the sampling variability based on theory.

The simulations above don't provide an actionable method for estimating sampling variability, but are rather an exercise to aid in understanding exactly what theoretical estimates of sampling variability are designed to measure.