AE 16: Intro to CLT

Dav King

3/3/22

```
library(tidyverse)
library(tidymodels)
```

Learning goals

- Understand and apply simulation-based methods to test a claim about independence between two groups
- Understand and apply simulation-based methods to calculate confidence interval to estimate difference in proportions
- Review simulation-based methods

Intro to CLT (if time)

Variability of sample statistics

- Each sample from the population yields a slightly different sample statistic (sample mean, sample proportion, etc.)
- The variability of these sample statistics is measured by the **standard error**
- Previously we quantified this value via simulation
- Today we'll discuss some of the theory underlying **sampling distributions**, particularly as they relate to *sample means*.

Recall

Statistical inference is the act of generalizing from a sample in order to make conclusions regarding a population. As part of this process, we quantify the degree of certainty we have.

We are interested in population parameters, which we do not observe. Instead, we must calculate statistics from our sample in order to learn about them.

Sampling distribution of the mean

Suppose we're interested in the resting heart rate of students at Duke, and are able to do the following:

1. Take a random sample of size n from this population, and calculate the mean resting heart rate in this sample, $\bar{X}_1$
2. Put the sample back, take a second random sample of size n, and calculate the mean resting heart rate from this new sample, $\bar{X}_2$
3. Put the sample back, take a third random sample of size n, and calculate the mean resting heart rate from this sample, too... and so on.

After repeating this many times, we have a dataset that has the sample averages from the population: $\bar{X}_1$, $\bar{X}_2$, $\cdots$, $\bar{X}_K$ (assuming we took K total samples).

Sampling distribution of the mean

Question: Can we say anything about the distribution of these sample means? *(Keep in mind, we don't know what the underlying distribution of mean resting heart rate looks like in Duke students!)*

As it turns out, we can...

The Central Limit Theorem

For a population with a well-defined mean μ and standard deviation σ, these three properties hold for the distribution of sample average $\bar{X}$,assuming certain conditions hold:

1. The mean of the sampling distribution is identical to the population mean μ,
2. The standard deviation of the distribution of the sample averages is $\sigma/\sqrt{n}$, or the **standard error** (SE) of the mean, and
3. For n large enough (in the limit, as $n \to \infty$), the shape of the sampling distribution of means is approximately *normal* (Gaussian).

What is the normal (Gaussian) distribution?

The normal distribution is unimodal and symmetric and is described by its *density function*:

If a random variable X follows the normal distribution, then

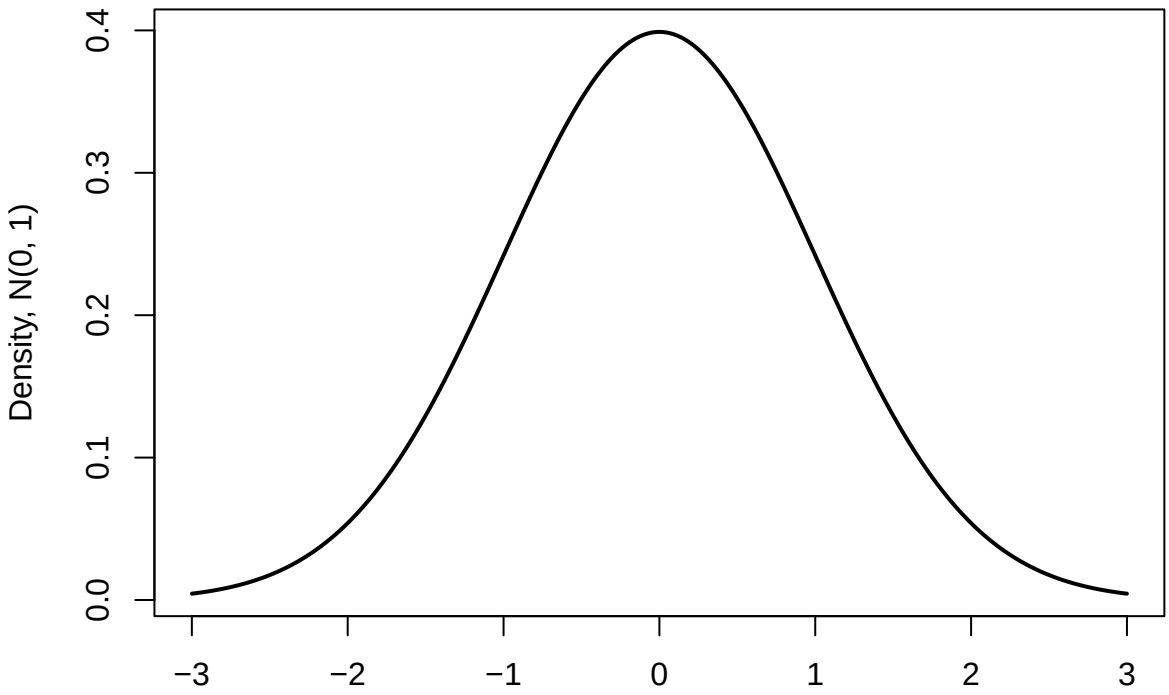
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$$

where μ is the mean and σ^2 is the variance.

We often write $N(\mu, \sigma^2)$ to describe this distribution.

The normal distribution (graphically)

We will talk about probability densities and using them to define probabilities later, but for now, just know that the normal distribution is the familiar "bell curve":



But we didn't know anything about the underlying distribution!

The central limit theorem tells us that sample averages are normally distributed, if we have enough data. This is true even if our original variables are not normally distributed.

Check out this interactive demonstration!

Conditions

What are the conditions we need for the CLT to hold?

- **Independence:** The sampled observations must be independent. This is difficult to check, but the following are useful guidelines:
  - the sample must be random
  - if sampling without replacement, sample size must be less than 10% of the population size
- **Sample size / distribution:**
  - if data are numerical, usually n $\geq$ 30 is considered a large enough sample, but if the underlying population distribution is extremely skewed, more might be needed

- if we know for sure that the underlying data are normal, then the distribution of sample averages will also be exactly normal, regardless of the sample size
- if data are categorical, at least 10 successes and 10 failures.

Practice using CLT & Normal distribution

Suppose the bone density for 65-year-old women is normally distributed with mean $809 mg/cm^3$ and standard deviation of $140 mg/cm^3$.

Let x be the bone density of 65-year-old women. We can write this distribution of x in mathematical notation as

$$x \sim N(809, 140)$$

Exercise 1

What bone densities correspond to Q_1 (25th percentile), Q_2 (50th percentile), and Q_3 (the 75th percentile) of this distribution? Use the `qnorm()` function to calculate these values.

```
qnorm(p = 0.25, mean = 809, sd = 140) #Q1
```

```
## [1] 714.5714
```

```
qnorm(p = 0.5, mean = 809, sd = 140) #Q2
```

```
## [1] 809
```

```
qnorm(p = 0.75, mean = 809, sd = 140) #Q3
```

```
## [1] 903.4286
```

Exercise 2

The densities of three woods are below:

- Plywood: 540 mg/cubic centimeter
- Pine: 600 mg/cubic centimeter
- Mahogany: 710 mg/cubic centimeter
- What is the probability that a randomly selected 65-year-old woman has bones less dense than Pine?
- Would you be surprised if a randomly selected 65-year-old woman had bone density less than Mahogany? What if she had bone density less than Plywood? Use the respective probabilities to support your response.

```
pnorm(q = 600, mean = 809, sd = 140) #Pine
```

```
## [1] 0.06773729
```

```r
pnorm(q = 540, mean = 809, sd = 140) #Plywood
```

```
## [1] 0.02733885
```

```r
pnorm(q = 710, mean = 809, sd = 140) #Mahogany
```

```
## [1] 0.2397389
```

It would be quite unsurprising if she had a bone density less than Mahogany - that's less than one standard deviation below the mean. It would be far more surprising if she had bones less dense than plywood - for which there is only a scant chance.

Exercise 3

Suppose you want to analyze the mean bone density for a group of 10 randomly selected 65-year-old women.

- Are the conditions met use the Central Limit Theorem to define the distribution of $\bar{x}$, the mean density of 10 randomly selected 65-year-old women?

No - Independence?

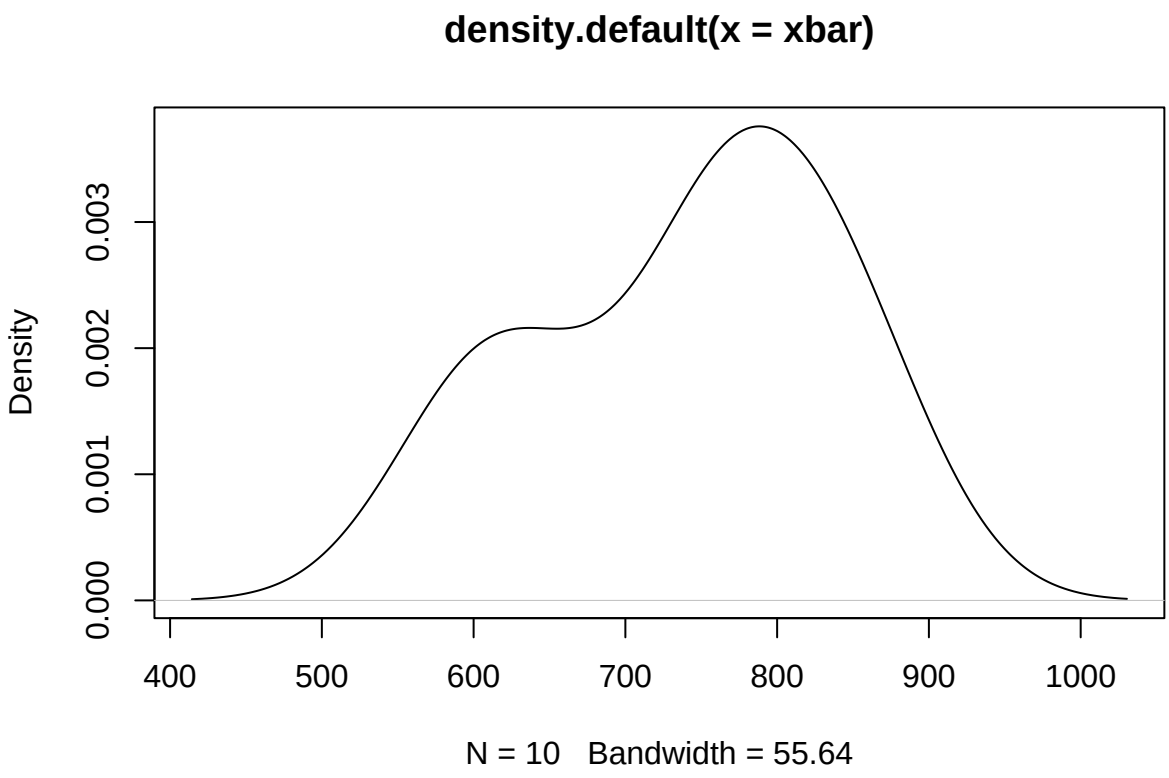
Yes - this is less than 10% of the total population of 65-year-old women. - Sample size/distribution?

No - we do not have $n \geq 30$ and we do not know for certain that the underlying distribution is normal.

- What is the shape, center, and spread of the distribution of $\bar{x}$, the mean bone density for a group of 10 randomly selected 65-year-old women?

We have no way to say for certain what any of these would be - it would vary depending on what sample we drew. However, if I draw a random sample:

```r
set.seed(10)
xbar <- rnorm(10, mean = 809, sd = 140)
dx <- density(xbar)
plot(dx)
```




```
mean(xbar)
```

```
## [1] 740.308
```

```
sd(xbar)
```

```
## [1] 97.98509
```

We can see that, for this sample, it is somewhere between unimodal and bimodal, has a mean of 740.308 and a standard deviation of 97.985.

- Write the distribution of $\bar{x}$ using mathematical notation.

$$\bar{x} \sim N(740.308, 97.985)$$

Exercise 4

- What is the probability that the mean bone density for the group of 10 randomly-selected 65-year-old women is less dense than Pine?

```
pnorm(q = 600, mean = mean(rnorm(10, mean = 809, sd = 140)),
      sd = sd(rnorm(10, mean = 809, sd = 140)))
```

```
## [1] 0.005342778
```

- Would you be surprised if a group of 10 randomly-selected 65-year old women had a mean bone density less than Mahogany? What if the group had a mean bone density less than Plywood? Use the respective probabilities to support your response.

```
pnorm(q = 710, mean = mean(rnorm(10, mean = 809, sd = 140)),
      sd = sd(rnorm(10, mean = 809, sd = 140))) #Mahogany
```

```
## [1] 0.4414268
```

```
pnorm(q = 540, mean = mean(rnorm(10, mean = 809, sd = 140)),
      sd = sd(rnorm(10, mean = 809, sd = 140))) #Plywood
```

```
## [1] 0.00768235
```

I would be surprised to see a group of 10 randomly-selected 65 year old women with a mean bone density less than Mahogany (for which there is not a great chance), but much more surprised to see it be less than Pine (a much lower density, more than an additional standard deviation away from the population mean). I would write about specific probabiilities, but without setting the seed (which we were not asked to do), there is no way to do so.

Exercise 5

Please explain how your answers differ in Exercises 2 and 4.

The probabilities in exercise 4 are (generally) less than they are in exercise 2. This has a very simple explanation. While it's already unlikely to find someone whose bone density is far away from the mean, this is the entire point of random sampling - these values exist within the normal distribution, and there is always a chance of receiving them. However, as seen by these probabilities, they are unlikely, Thus, it is much more unlikely to find ten people whose bone density can all average out to these extreme values - since that would require many more instances of choosing events with low probabilities.