Chapter 5 - Foundations for Inference

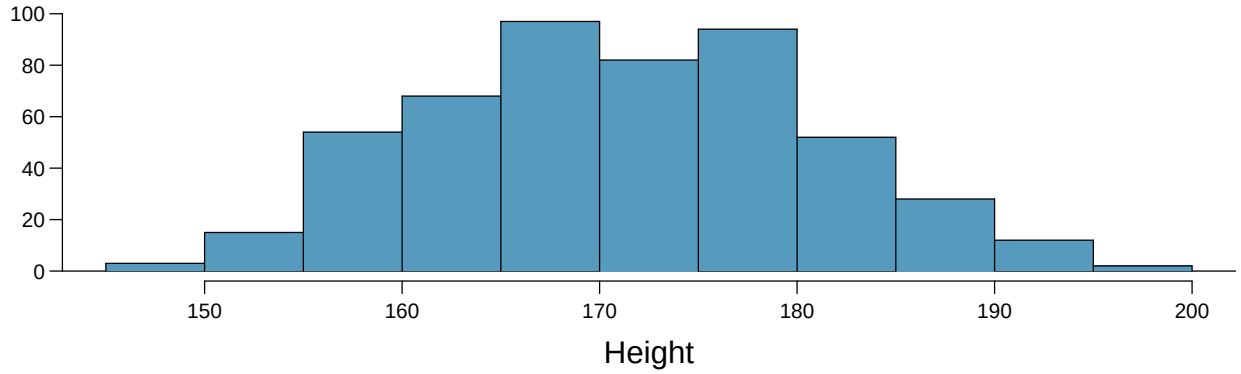
Donald Butler

10/09/2021

```
library(openintro)
library(tidyverse)
library(DATA606)
```

```
##
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 4th Edition. You can read this by typing
## vignette('os4') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.
```

Heights of adults. (7.7, p. 260) Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram below shows the sample distribution of heights in centimeters.



(a) What is the point estimate for the average height of active individuals? What about the median?

```
mean(bdims$hgt)
```

```
## [1] 171.1438
```

```
median(bdims$hgt)
```

```
## [1] 170.3
```

The point estimate for average height is 171 cm and the estimate for median height is 170.

(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

```
sd(bdims$hgt)
```

```
## [1] 9.407205
```

```
IQR(bdims$hgt)
```

```
## [1] 14
```

The point estimate for standard deviation is 9.4 and the interquartile range is 14.

(c) Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

```
(180-mean(bdims$hgt))/sd(bdims$hgt)
```

```
## [1] 0.9414287
```

```
(155-mean(bdims$hgt))/sd(bdims$hgt)
```

```
## [1] -1.716109
```

A person that is 180cm tall would not be unusually tall because his height is less than one standard deviation from the mean. A person that is 155 cm tall is also not unusually short because their height is less than two standard deviations from the mean.

(d) The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

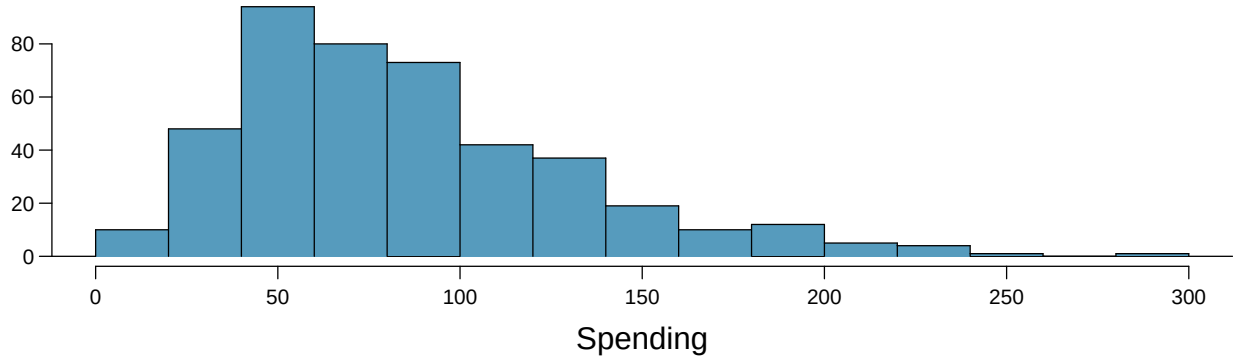
The sample statistics, including standard deviation, would not be the same in a new sample. Each sample would vary from the true population statistic by the sampling error.

(e) The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate (Hint: recall that $SD_x = \frac{\sigma}{\sqrt{n}}$)? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

```
sd(bdims$hgt)/sqrt(nrow(bdims))
```

```
## [1] 0.4177887
```

Thanksgiving spending, Part I. The 2009 holiday retail season, which kicked off on November 27, 2009 (the day after Thanksgiving), had been marked by somewhat lower self-reported consumer spending than was seen during the comparable period in 2008. To get an estimate of consumer spending, 436 randomly sampled American adults were surveyed. Daily consumer spending for the six-day period after Thanksgiving, spanning the Black Friday weekend and Cyber Monday, averaged $84.71. A 95% confidence interval based on this sample is ($80.31, $89.11). Determine whether the following statements are true or false, and explain your reasoning.



(a) We are 95% confident that the average spending of these 436 American adults is between $80.31 and $89.11.

Technically true because we know the average of these 436 American adults is $84.71.

(b) This confidence interval is not valid since the distribution of spending in the sample is right skewed.

False, the sample observations were randomly choosen and the sample size is greater than 30 with no particularly extreme outliers.

(c) 95% of random samples have a sample mean between $80.31 and $89.11.

False, other random samples could have a sample mean outside of this range, yet still have their confidence interval contain the population mean.

(d) We are 95% confident that the average spending of all American adults is between $80.31 and $89.11.

True, assuming the sample is representative of the entire population, we can conclude with 95% certainty that the average spending is within the confidence interval.

(e) A 90% confidence interval would be narrower than the 95% confidence interval since we don't need to be as sure about our estimate.

True, a lower confidence level will narrow the confidence interval.

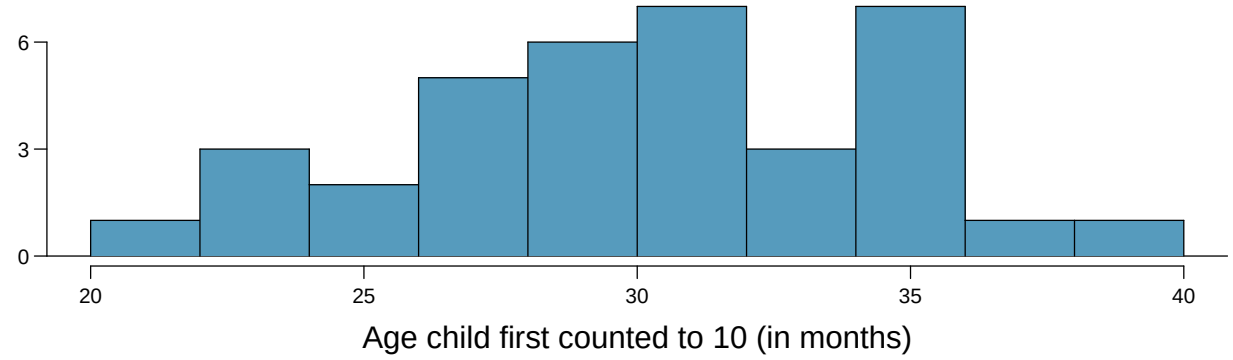
(f) In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger.

False, the margin of error is proportional to $\frac{1}{\sqrt{n}}$, so we would need a sample size 9 times larger.

(g) The margin of error is 4.4.

True, the margin of error is half the confidence interval width.

Gifted children, Part I. Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. The following histogram shows the distribution of the ages (in months) at which these children first counted to 10 successfully. Also provided are some sample statistics.



n	36
min	21
mean	30.69
sd	4.31
max	39

(a) Are conditions for inference satisfied?

Yes, the sample was randomly selected and contains more than 30 observations.

(b) Suppose you read online that children first count to 10 successfully when they are 32 months old, on average. Perform a hypothesis test to evaluate if these data provide convincing evidence that the average age at which gifted children first count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.

$H_0 : \mu = 32$
$H_A : \mu < 32$

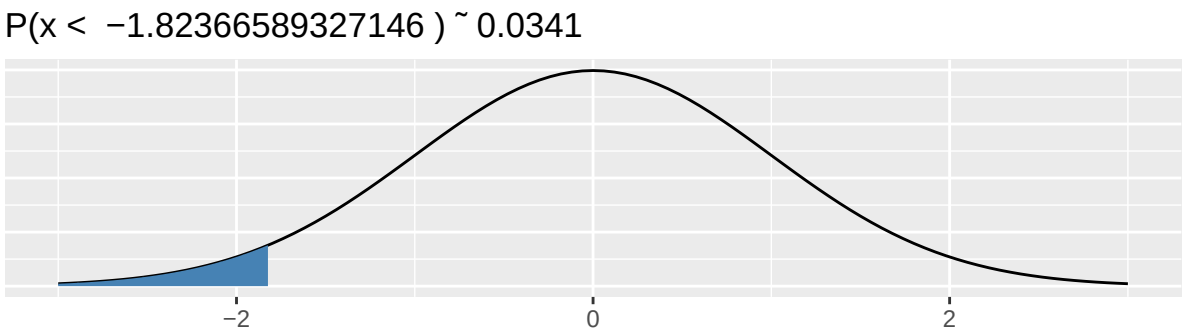
```
(SE = 4.31 / sqrt(36))
```

```
## [1] 0.7183333
```

```
(Z = (30.69 - 32) / SE)
```

```
## [1] -1.823666
```

```
DATA606::normal_plot(cv = Z, tails = 'less')
```



Since the p-value is less than $\alpha = .10$, we reject the null hypothesis accept that gifted children first count to 10 before 32 months.

(c) Interpret the p-value in context of the hypothesis test and the data.

If the null hypothesis were true, we would see a sample mean of 30.69 only 3.4% of samples. Since the p-value is less than the significance level of .10, we reject the null hypothesis.

(d) Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.

```
30.69 - 1.645 * 4.31 / sqrt(36)
```

```
## [1] 29.50834
```

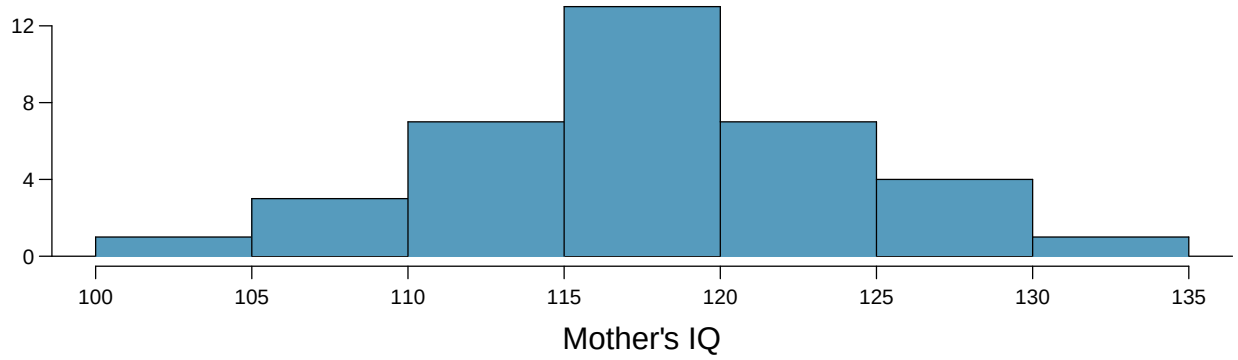
```
30.69 + 1.645 * 4.31 / sqrt(36)
```

```
## [1] 31.87166
```

(e) Do your results from the hypothesis test and the confidence interval agree? Explain.

Yes, the 90% confidence interval suggests the average age a gifted child first counts to 10 is between 29.5 and 31.9 months, which is less than average age of the general population.

Gifted children, Part II. Exercise above describes a study on gifted children. In this study, along with variables on the children, the researchers also collected data on the mother's and father's IQ of the 36 randomly sampled gifted children. The histogram below shows the distribution of mother's IQ. Also provided are some sample statistics.



n	36
min	101
mean	118.2
sd	6.5
max	131

(a) Perform a hypothesis test to evaluate if these data provide convincing evidence that the average IQ of mothers of gifted children is different than the average IQ for the population at large, which is 100. Use a significance level of 0.10.

$H_0 : \mu = 100$
$H_A : \mu > 100$

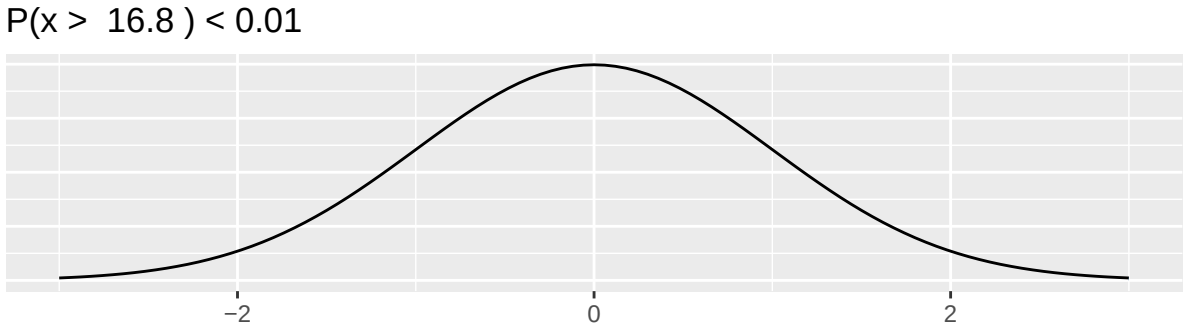
```
(SE = 6.5 / sqrt(36))
```

```
## [1] 1.083333
```

```
(Z = (118.2 - 100) / SE)
```

```
## [1] 16.8
```

```
DATA606::normal_plot(cv = Z, tails = 'greater')
```



Since the p-value is less than $\alpha = .10$, we reject the null hypothesis accept that the IQ of mothers with gifted children is greater than 100.

(b) Calculate a 90% confidence interval for the average IQ of mothers of gifted children.

```
118.2 - 1.645 * 6.5 / sqrt(36)
```

```
## [1] 116.4179
```

```
118.2 + 1.645 * 6.5 / sqrt(36)
```

```
## [1] 119.9821
```

(c) Do your results from the hypothesis test and the confidence interval agree? Explain.

Yes, the 90% confidence interval is greater than 100.

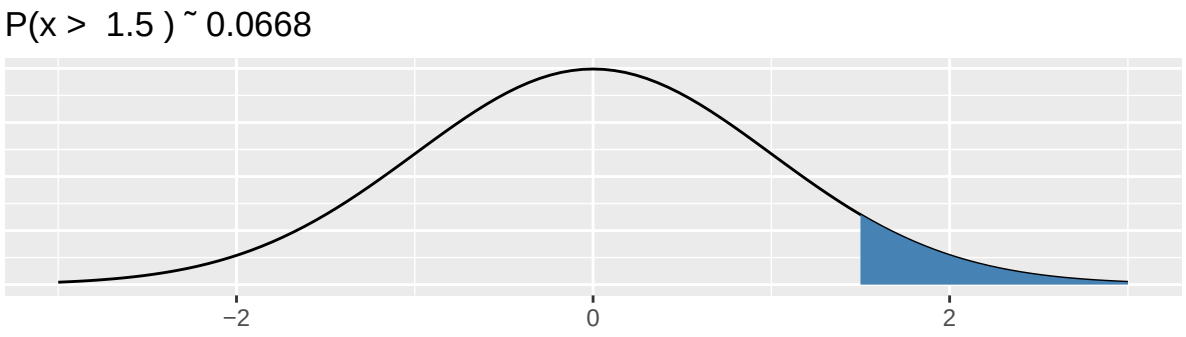
CLT. Define the term "sampling distribution" of the mean, and describe how the shape, center, and spread of the sampling distribution of the mean change as sample size increases.

Sampling distribution of the mean is the process of taking multiple samples and calculating the mean for each sample. If enough samples are selected, the distribution of means is normal. Increasing the sample size of each will narrow the spread of the sampling distribution.

CFLBs. A manufacturer of compact fluorescent light bulbs advertises that the distribution of the lifespans of these light bulbs is nearly normal with a mean of 9,000 hours and a standard deviation of 1,000 hours.

(a) What is the probability that a randomly chosen light bulb lasts more than 10,500 hours?

```
DATA606::normal_plot(cv = (10500 - 9000) / 1000, tails = 'greater')
```



The probability that a randomly chosen light bulb lasts more than 10,500 hours is 6.7%.

(b) Describe the distribution of the mean lifespan of 15 light bulbs.

By selecting less than 30 samples, the distribution may not be normal even though the population that the sample is drawn from is.

(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?

```
1 - pnorm(10500,9000,1000/sqrt(15))
```

```
## [1] 3.133452e-09
```

(d) Sketch the two distributions (population and sampling) on the same scale.

(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution?

No, the distributions must be nearly normal to allow calculation with the normal probability density function.

Same observation, different sample size. Suppose you conduct a hypothesis test based on a sample where the sample size is n = 50, and arrive at a p-value of 0.08. You then refer back to your notes and discover that you made a careless mistake, the sample size should have been n = 500. Will your p-value increase, decrease, or stay the same? Explain.

The Z value is proportional to the $\sqrt{n}$, so a 10 fold increase in sample size will increase the Z value by $\sqrt{10}$, and reduce the p-value.