Foundations for Inference

Code ▾

DS606-HW5

George Cruz

2020-10-20

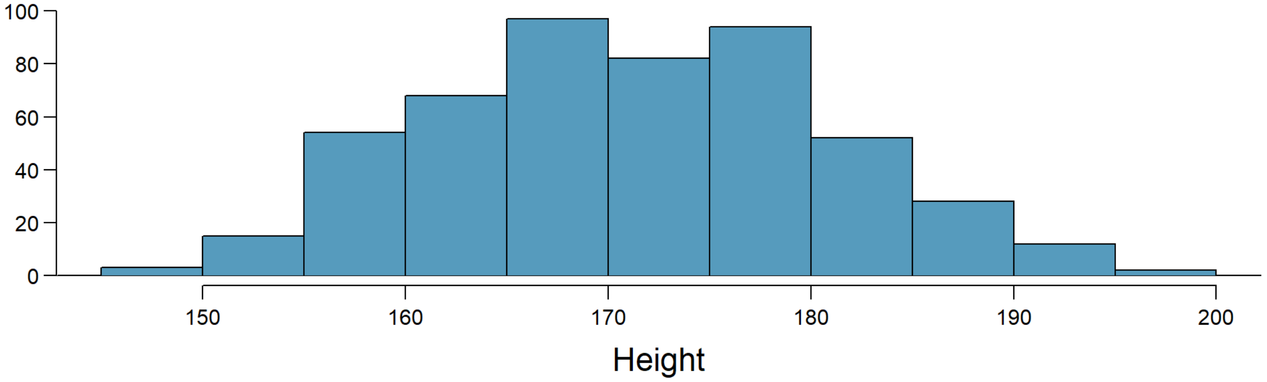
Foundations for Inference

Hide

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

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?

Hide

```
c(avg=mean(bdims$hgt), med=median(bdims$hgt))
```

```
##      avg      med
## 171.1438 170.3000
```

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

Hide

```
c(sd=sd(bdims$hgt), iqr=IQR(bdims$hgt))
```

```
##        sd       iqr
##  9.407205 14.000000
```

(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. The first person is above average but within 1 standard deviation, so we could call that person tall. The second person is about 1.7 sd away from the mean, we could call that person short. Because none of these values fall within 2 standard deviations, we can not call them unusual.

Hide

```
c(too_tall=(180-171.14)/9.41, too_short=(155-171.14)/9.41)
```

```
##   too_tall  too_short
##  0.9415515 -1.7151966
```

(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.

No. I would expect some variations between the two samples.

(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.

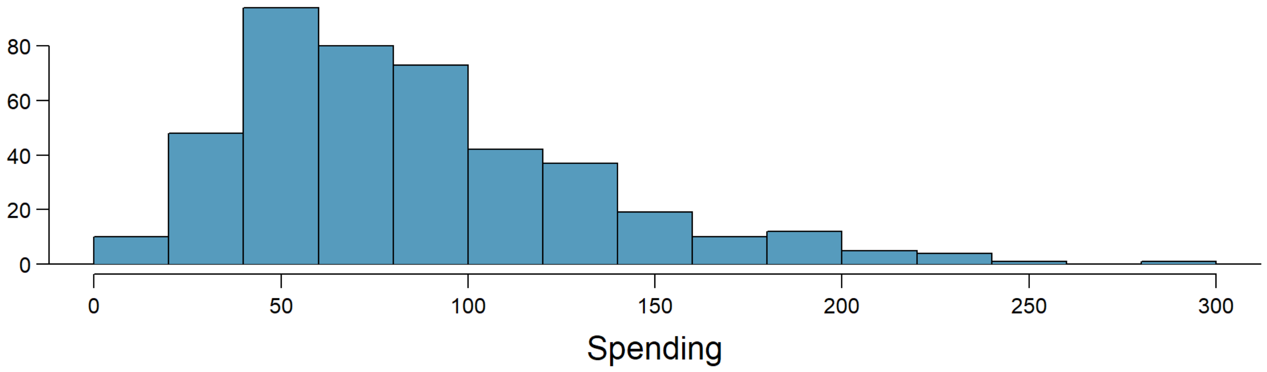
Hide

```
sqrt(var(bdims$hgt)/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. False. We are 100% confident this is the case.

Hide

```
mean(thanksgiving_spend$spending)
```

```
## [1] 84.70677
```

(b) This confidence interval is not valid since the distribution of spending in the sample is right skewed. False. This confidence interval is valid regardless of the shape of the distribution due to the size of the sample. Sample size > 30.

(c) 95% of random samples have a sample mean between $80.31 and $89.11. Mostly true. 95% of random samples should have a sample within this range.

(d) We are 95% confident that the average spending of all American adults is between $80.31 and $89.11. True. Because the survey was random and with a large number of observations.

(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.

(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. We would need to increase the sample size by 9.

(g) The margin of error is 4.4.

Hide

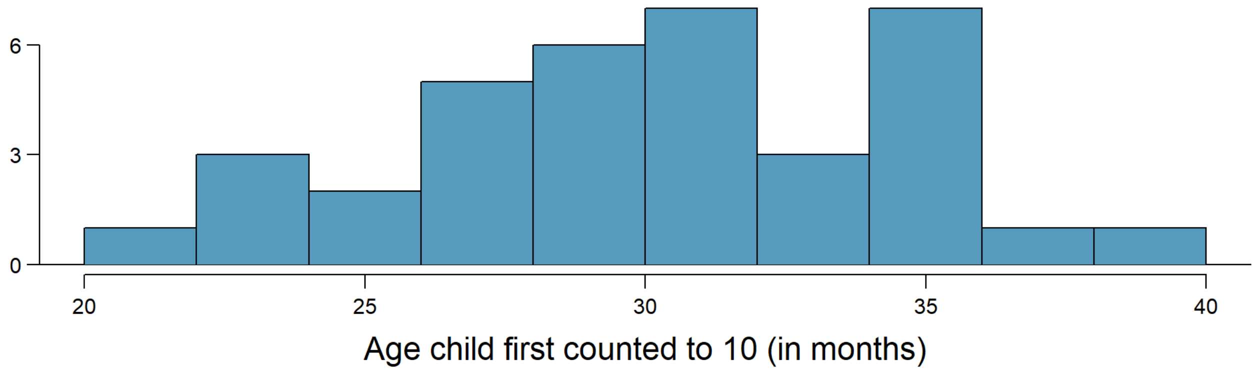
```
z = 1.96
se = sqrt(var(thanksgiving_spend$spending)/nrow(thanksgiving_spend))

se * z
```

```
## [1] 4.405038
```

True

Gifted children, Part I.

Researchers investigating characteristics of gifted children col- lected 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 dis- tribution of the ages (in months) at which these children first counted to 10 successfully. Also provided are some sample statistics.



(a) Are conditions for inference satisfied? Yes. We have an almost normal distribution, a random sample of over 30 independent 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 fist count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.

Hide

```
summary(gifted$count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   21.00   28.00   31.00   30.69   34.25   39.00
```

H0: Gifted children first count to 10 at the same avg age as the population H1: Gifted children first count to 10 at a different avg age than the pop.

Hide

```
gifted_sd <- sd(gifted$count)
gifted_sd
```

```
## [1] 4.314887
```

Hide

```
se <- gifted_sd/sqrt(36)
Z <- (30.69 - 32)/(se)
p <- pnorm(Z) * 2
p
```

```
## [1] 0.06851567
```

(c) Interpret the p-value in context of the hypothesis test and the data. Because the significant level is less than 0.10 we reject the null hypothesis.

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

Hide

```
cv <- 1.64

ci_low <- round(30.69 - cv * se,2)
ci_high <- round(30.69 + cv * se,2)

c(ci_low, ci_high)
```

```
## [1] 29.51 31.87
```

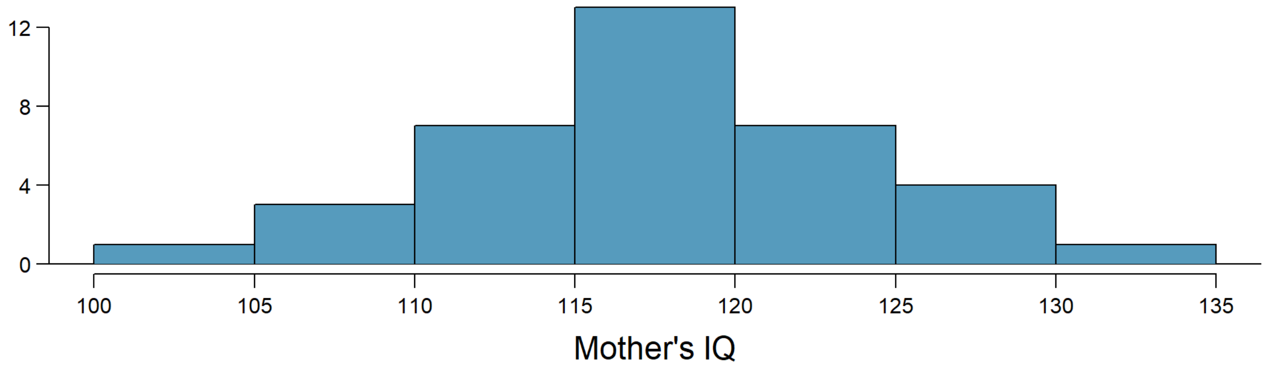
The 90% confidence interval for the avg counting age is: (29.51, 31.87)

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

They do. Because we rejected the hypothesis, we know the avg age for gifted children is different than the avg age for normal children to start counting. Based on this confidence interval, the avg age will be between 29.5 and 31.8 which would be lower than the avg for the normal 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.



(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. H0: The Avg IQ of mothers of gifted children is different as the population H1: The Avg IQ of mothers of gifted children is the same as the populations

Hide

```
cv <- 1.64
se <- sd(gifted$motheriq)/sqrt(36)
Z <- (100 - 118.16)/(se)
p <- pnorm(Z) * 2
format(round(p,3), nsmall=3)
```

```
## [1] "0.000"
```

The p value is so small that it approaches 0. We accept the null hypothesis.

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

Hide

```
avg <- mean(gifted$motheriq)
avg
```

```
## [1] 118.1667
```

Hide

```
ci_low <- round(118.16 - cv * se,2)
ci_high <- round(118.16 + cv * se,2)

c( ci_low, ci_high)
```

```
## [1] 116.38 119.94
```

The 90% confidence interval for mother IQ's of gifted children is (116.38, 119.94)

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

Yes. Because we accepted the null hypothesis we know that the avg IQ for mothers of gifted children is different than the normal population. This confidence interval points to that avg being between 116 and 120, which is higher 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.

The sampling distribution of the mean is the mean of the population from which the scores were sampled. Therefore, if a population has a mean μ, then the mean of the sampling distribution of the mean is also μ.

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?

Hide

```
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

Hide

```
v <- 1 - pnorm(10500, 9000, 1000)

percent(v,accuracy = 0.01)
```

```
## [1] "6.68%"
```

(b) Describe the distribution of the mean lifespan of 15 light bulbs. Due to the fact that this is a nearly normal distribution, we can assume the distribution of 15 bulbs will look normal as well.

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

Hide

```
v <- 1 - pnorm(10500, 9000, 1000/sqrt(15))
percent(v,accuracy = 0.000000001)
```

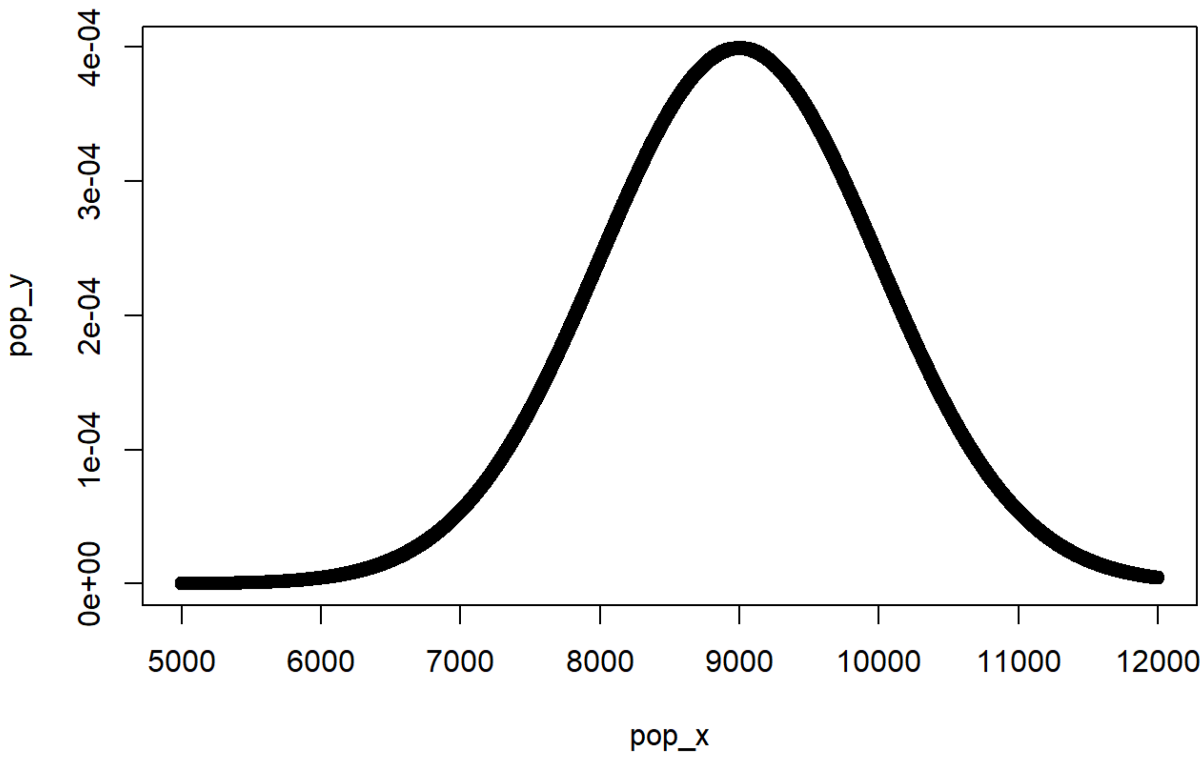
```
## [1] "0.000000313%"
```

Really, really small.

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

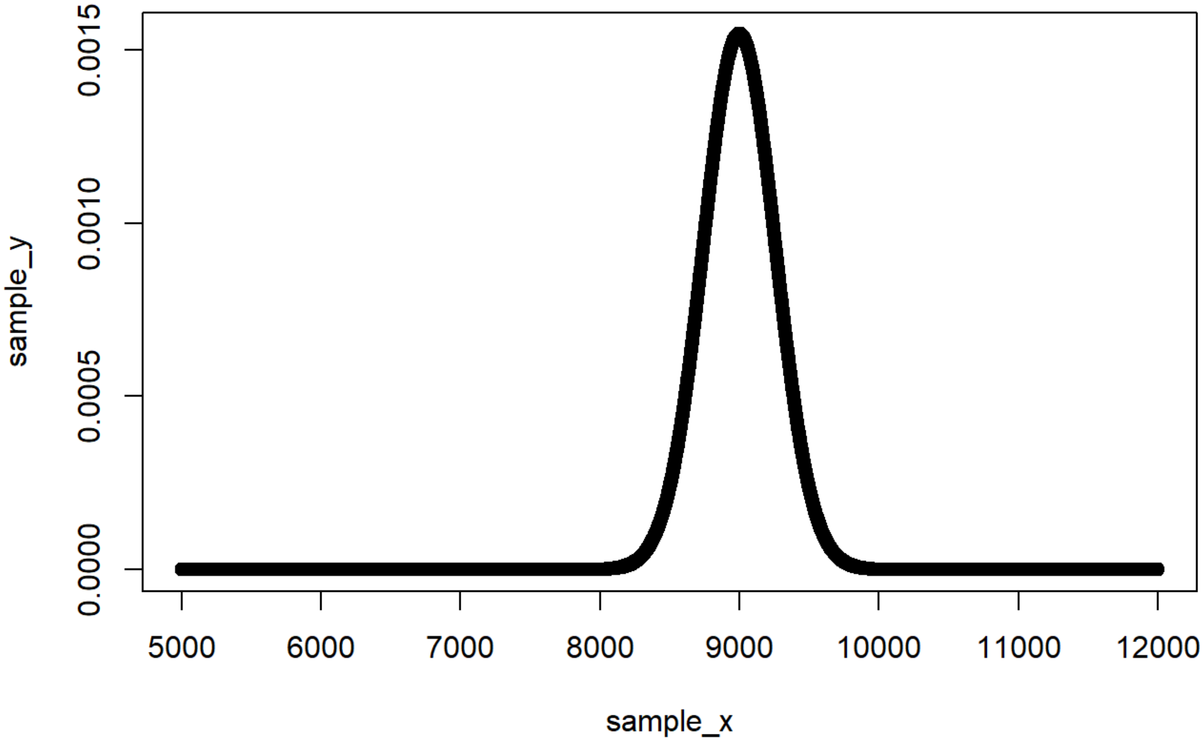
Hide

```
pop_x <- 5000:12000
pop_y <- dnorm(pop_x,mean=9000,sd=1000)

sample_x <- 5000:12000
sample_y <- dnorm(sample_x,mean=9000,sd=1000/sqrt(15))

plot(pop_x,pop_y)
```



Hide

```
plot(sample_x, sample_y)
```



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

Probably not.

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 p value should decrease as the denominator increased in size. …