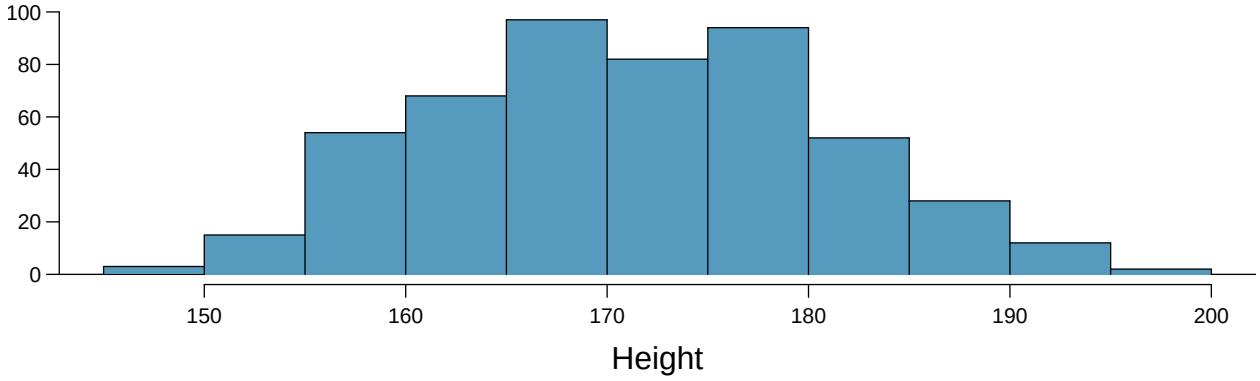
DATA606_HW5

RJM

2019-10-18

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.




```
head(bdims)
```

```
##   bia.di bii.di bit.di che.de che.di elb.di wri.di kne.di ank.di sho.gi
## 1   42.9   26.0   31.5   17.7   28.0   13.1   10.4   18.8   14.1  106.2
## 2   43.7   28.5   33.5   16.9   30.8   14.0   11.8   20.6   15.1  110.5
## 3   40.1   28.2   33.3   20.9   31.7   13.9   10.9   19.7   14.1  115.1
## 4   44.3   29.9   34.0   18.4   28.2   13.9   11.2   20.9   15.0  104.5
## 5   42.5   29.9   34.0   21.5   29.4   15.2   11.6   20.7   14.9  107.5
## 6   43.3   27.0   31.5   19.6   31.3   14.0   11.5   18.8   13.9  119.8
##   che.gi wai.gi nav.gi hip.gi thi.gi bic.gi for.gi kne.gi cal.gi ank.gi
## 1   89.5   71.5   74.5   93.5   51.5   32.5   26.0   34.5   36.5   23.5
## 2   97.0   79.0   86.5   94.8   51.5   34.4   28.0   36.5   37.5   24.5
## 3   97.5   83.2   82.9   95.0   57.3   33.4   28.8   37.0   37.3   21.9
## 4   97.0   77.8   78.8   94.0   53.0   31.0   26.2   37.0   34.8   23.0
## 5   97.5   80.0   82.5   98.5   55.4   32.0   28.4   37.7   38.6   24.4
## 6   99.9   82.5   80.1   95.3   57.5   33.0   28.0   36.6   36.1   23.5
##   wri.gi age  wgt   hgt sex
## 1   16.5  21 65.6 174.0   1
## 2   17.0  23 71.8 175.3   1
## 3   16.9  28 80.7 193.5   1
## 4   16.6  23 72.6 186.5   1
## 5   18.0  22 78.8 187.2   1
## 6   16.9  21 74.8 181.5   1
```

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

```
x <- bdims$hgt
mean(x)
```

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

```
# The average height is 171.144 cm.

median(x)
```

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

```
# The median height is 170.3 cm.
```

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

```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   147.2   163.8   170.3   171.1   177.8   198.1
```

```
sd <- sd(x)
# The standard deviation of the heights is 9.407.

IQR(x)
```

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

```
# The IQR of the heights is 14 (177.8 - 163.8).
```

(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.

```
# Looking at the summary, 180 cm is within the top 25% and 155 cm is within the low 25%. So, these # he
```

(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.

```
# It is highly unlikely that the mean and standard deviation would be same, but it should be      # sim
```

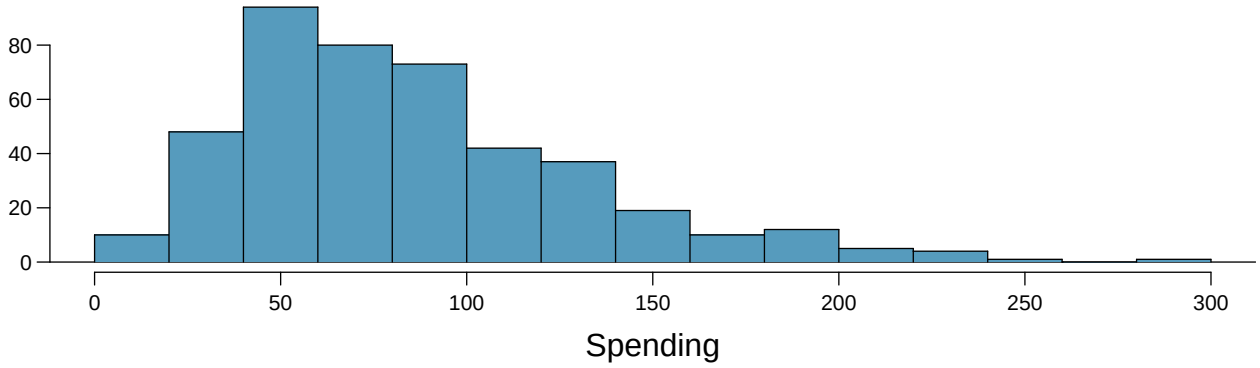
(e) The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate (Hint: recall that $SD_x = \frac{\sigma}{\sqrt{n}}$)? Compute this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

```
n <- length(bdims$hgt)

SD_x <- sd/sqrt(n)

SD_x
```

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

```
# The standard error of the mean is 0.418.
```

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 - The CI assumption is for the whole population not just the sample.
```

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

```
# False - The right skew is very slight and the majority of the data points seem to fall within   # the
```

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

```
# False - This assumption is not true as the samples can have different ranges. There is no       # cor
```

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

```
# True - We can make this assumption for the whole population based on the CI of the results.
```

(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 - The ranges gets narrower as the CI becomes smaller.
```

(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 sample size needs to be 9 times in this case.
```

(g) The margin of error is 4.4.

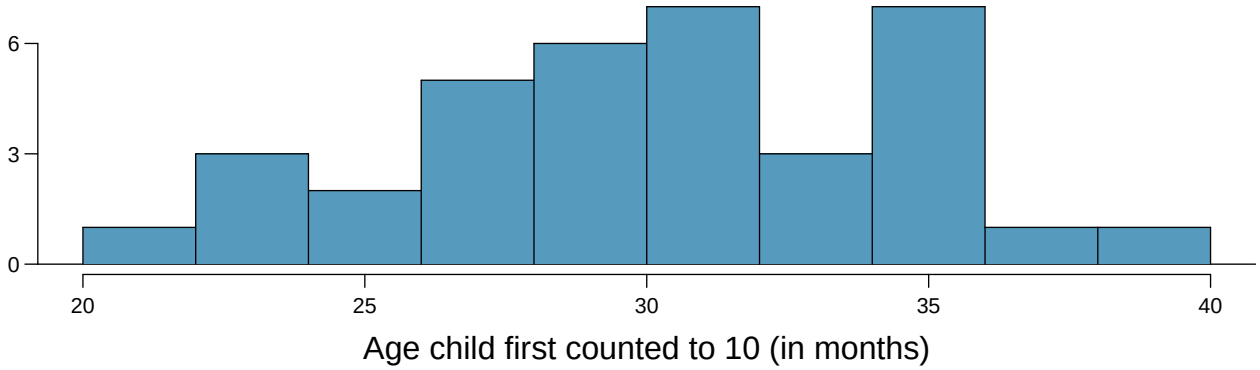
```r
summary(tgSpending)
```

```
##     spending
##  Min.   :  5.719
##  1st Qu.: 49.177
##  Median : 75.792
##  Mean   : 84.707
##  3rd Qu.:112.255
##  Max.   :282.803
```

```r
sd <- sd(tgSpending$spending)
n <- length(tgSpending$spending)
# ME = upper bound of CI - mean
# ME = 89.11 - 84.7 = 4.4

SD_x <- sd / sqrt(n)

SD_x
```

```
## [1] 2.247468
```

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.



n	36
min	21
mean	30.69
sd	4.31
max	39

(a) Are conditions for inference satisfied?

```
# The sample size is 36 (> 30), is randomly picked, and the results are not skewed, so the        # con
```

(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.

```
# H_0 = First count at 32 months
# H_A = First count at <32 months
# alpha = 0.10
sd <- 4.31
n <- 36
mean <- 30.69
month_age <- 32
SD_y <-sd / sqrt(n)

SD_y
```

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

```
z_score <- (mean - month_age) / SD_y
z_score
```

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

```r
p_value <- (1 - pnorm(abs(z_score)))
p_value
```

```
## [1] 0.0341013
```

```r
# Since p-value (0.034) is less than the aplpha (0.10), we have to reject the null hypothesis.
```

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

```r
# As per the above, the p-value is 0.0341.
```

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

```r
# For 90% CI:
z_value <- 1.645
right_CI <- mean + z_value*SD_y
left_CI <- mean - z_value*SD_y
right_CI
```

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

```r
left_CI
```

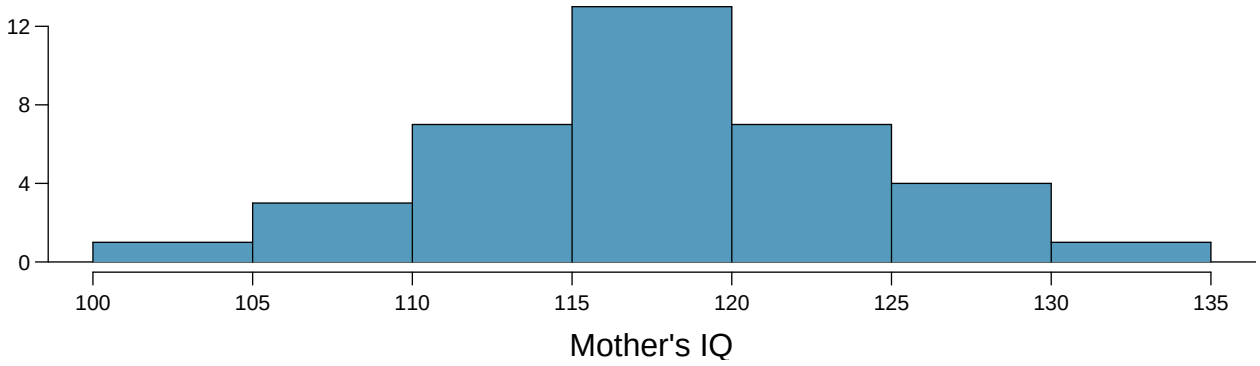
```
## [1] 29.50834
```

```r
# The CI is between 29.508 and 31.872.
```

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

The results from the hypothesis test and CI agree as they both point towards the average age of children's first count to 10 being under 32 months.

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 = IQ is 100.
# H_A = IQ is not 100.
# alpha = 0.10
sd_m <- 6.5
n <- 36
mean_m <- 118.2
mother_IQ <- 100
SD_b <- sd_m/ sqrt(n)

SD_b
```

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

```
z_score_m <- (mean_m - mother_IQ) / SD_b
z_score_m
```

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

```
p_value_m <- (1 - pnorm(abs(z_score_m)))
p_value_m
```

```
## [1] 0
```

```r
# Since p-value (0.0) is less than the aplpha (0.10), we have to reject the null hypothesis.
```

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

```r
# For 90% CI:
z_value <- 1.645
right_CI_m <- mean_m + z_value*SD_b
left_CI_m <- mean_m - z_value*SD_b
right_CI_m
```

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

```r
left_CI_m
```

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

```r
# The CI is between 29.508 and 31.872.
```

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

Yes, both results seem to agree as we have rejected the null hypothesis of IQ equal to 100 and the we can say with 90% confidence that the IQ for mothers averages between 116.418 and 119.982.

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 distribution of mean of each sample from multiple samples. As the sample size increases the shape, center, and spread of the sampling distribution tend to appear more normal and tend to form a bell curve.

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?

```
prob_bulb1 <- 1-pnorm(q=10500, mean=9000, sd=1000)
prob_bulb1
```

```
## [1] 0.0668072
```

The probability is 6.7%.

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

```
sd_bulb <- 1000
mean_bulb <- 9000
n_bulb <- 15

bulb_sd <- sd_bulb/sqrt(n_bulb)
bulb_sd
```

```
## [1] 258.1989
```

The distribution of the mean lifespan of 15 light bulbs is 258.199.

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

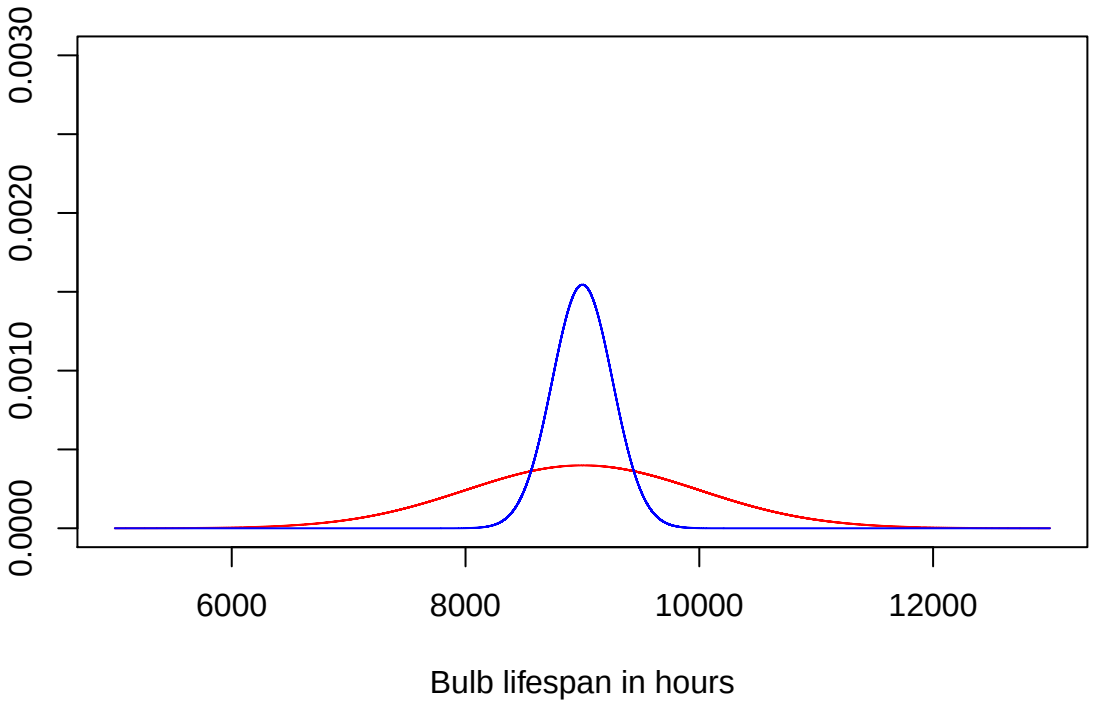
```
p_bulb <- 1 - pnorm(q=10500, mean=9000, sd=258.20)
p_bulb
```

```
## [1] 3.13392e-09
```

The probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours is very low (approaching 0).

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

```
d <- seq(5000,13000,0.01)
plot(d, dnorm(d,mean_bulb, sd_bulb), type="l", ylim = c(0,0.003), ylab = "", xlab = "Bulb lifespan in h
lines(d, dnorm(d,mean_bulb, bulb_sd), col="blue")
```



The population's distribution is red and the sample's is in blue.

(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution? # The assumption is that the distribution is normal in both parts. A skewed distribution would have a different mean and median, and we would need the IQR as a better measure than the standard deviation. For (c) it might be possible to use the mean due to the small sample size, but we do not have enough information in both cases to find the IQR.

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. # An increase in n will help decrease the SE, which in turn increases the z-value. With an increase in z-value, the p-value will be lowered.