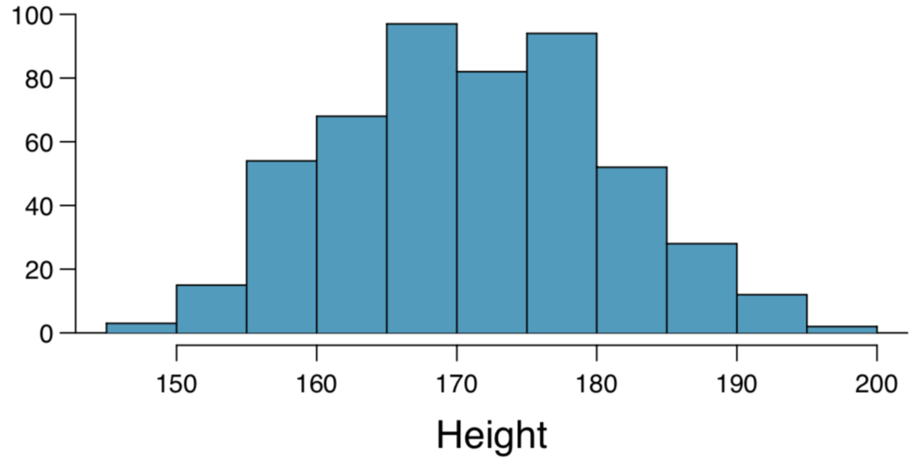
week_5_assignment

Salma Elshahawy

10/9/2019

7.8 Heights of adults. Researchers studying anthropometry collected body girth measurements and skele- tal 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.



Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
SD	9.4
Q3	177.8
Max	198.1

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

the average height is 171.1cm, the median is 170.3cm

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

the standard deviation for the height is 9.4cm, the IQR is Q3-Q1 = 177.8 - 163.8 = 14.1cm

```
IQR = 177.9 - 163.8
IQR
```

```
## [1] 14.1
```

(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.

```
mean <- 171.8
sd <- 9.4
z_score180 <- (180 - mean)/sd
z_score180
```

```
## [1] 0.8723404
```

```r
z_score150 <- (155 - mean)/sd
z_score150
```

```
## [1] -1.787234
```

NO! 180cm is within one standard deviation from the mean, not considered unusually tall; And 155 is within 2 standard deviations from the mean, still not considered unusually short.

(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.

I would not expect the mean and standard deviation of another sample to be exact because of the sample is random.

(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? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

```r
n <- 507
sd/sqrt(n)
```

```
## [1] 0.4174687
```

Variability of the estimate can be measured by standard error = 0.417

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 95% confident that the average spending of ALL American adults is between \$80.31 and \$89.11.

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

False. The sample is sufficiently large and not strongly right skewed.

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

False. Sample statistics vary from sample to sample.

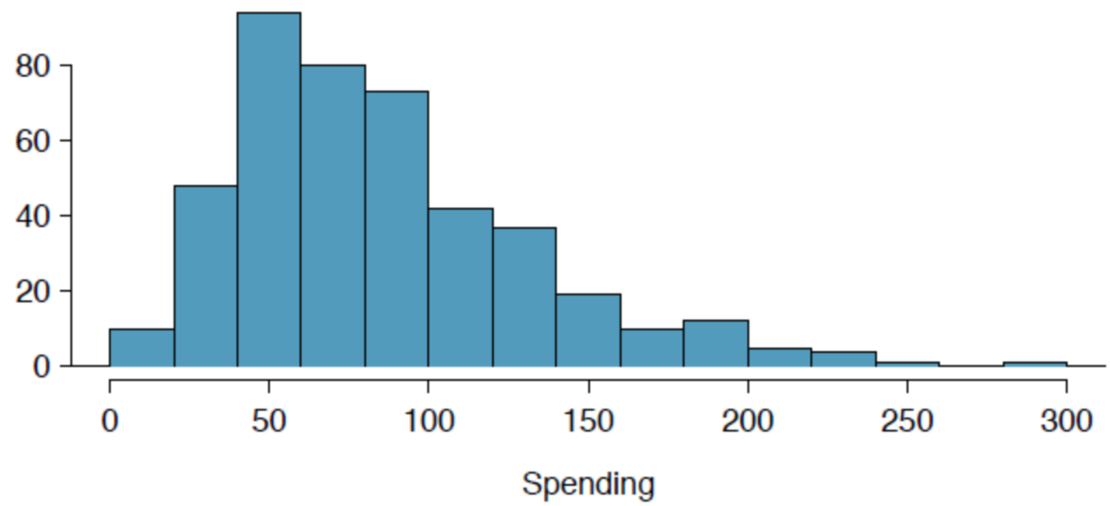


Figure 1: figure2

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

True. That's the definition of a 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. If we want to be more certain that we capture the population parameter, we would use a wider interval. So, the 90% interval is narrower than the 95% 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. To obtain 1/3 of the margin of error, we need 9 times as large as the sample size because we are taking the square root of n in the denominator.

g. The margin of error is 4.4.

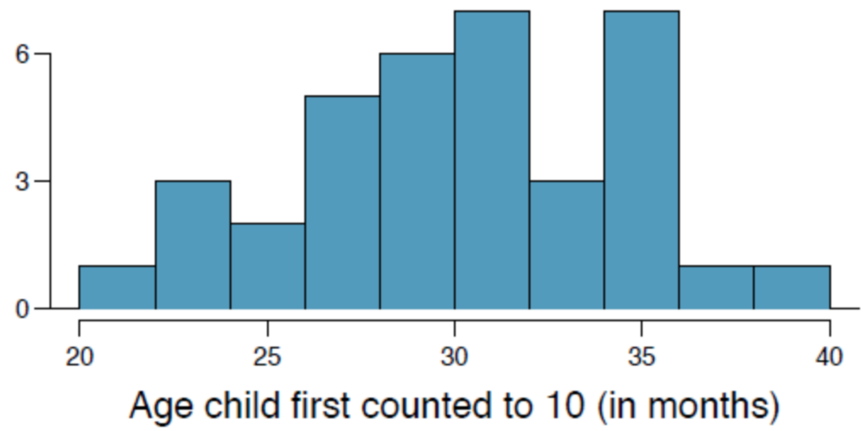
```
mean2 <- 84.71
err_lower <- mean2 - 80.31 #margin of error from lower bound
err_upper <- 89.11 - mean2 #margin of error from upper bound
err_upper
```

```
## [1] 4.4
```

```
err_lower
```

```
## [1] 4.4
```

True.



n	36
min	21
mean	30.69
sd	4.31
max	39

Figure 2: figure3

$$H_0: \mu_{\bar{x}} = 32 \quad H_1: \mu_{\bar{x}} < 32 \quad \alpha = 0.10$$

Figure 3: equation_1

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.

a. Are conditions for inference satisfied?

- Independence: **Yes. Random sampling is used, and n $<$ 10% of the population (when sampling without replacement)**
- Sample size/skew: **Yes. The more skewed the population distribution, the larger sample size we need for the CLT to appply. In this case we have a moderately skewed distribution, and n $>$ 30 satisfied the condition.**

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.

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

```
mean3 <- 30.69 #sample mean
se <- 4.31/sqrt(36) #standard error

Z <- (mean3-32)/se #test statistic
p_value <- pnorm(Z) #p-value
p_value
```

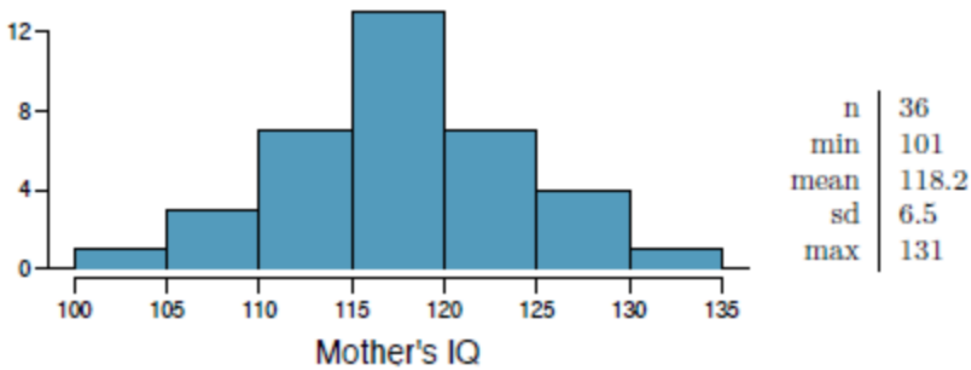


Figure 4: figure4

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

P-value = 0.034 is the probability of observing a sample mean less than 30.69 for a sample size of 36, assuming that the null hypothesis is true.

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

```
z <- 1.645
mean3 - z*se
```

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

```
mean3 + z*se
```

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

The confidence interval is (29.5, 31.87)

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

P-value is small, thus we reject the null hypothesis. The confidence interval doesn't capture and is below 32, which agrees with the result from the hypothesis test.

Gifted children, Part II. Exercise 4.24 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.

```r
se <- 6.5/sqrt(36)
se
```

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

```r
z <- (118.2-100)/se
z
```

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

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

```r
#118.2 +/- se
#upper limit
118.2 + 1.083
```

```
## [1] 119.283
```

```r
#lower limit
118.2 - 1.083
```

```
## [1] 117.117
```

confidence interval is (117.117, 119.283)

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

Agree. They both suggest that the IQ of mothers of gifted children is different from 100. In fact, larger 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 distribution of average sample means. As sample size increases, we would expect samples to yield more consistent sample means, hence the variability among the sample means would be lower and the spread would be narrower.

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?

```r
1 - pnorm(10500, mean = 9000, sd = 1000)
```

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

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

```r
se <- 1000/sqrt(15)
se
```

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

The distribution of the mean lifespan of 15 light bulbs may be approximated by a normal model given they are independent

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

```r
Z <-(10500-9000)/ se

1 - pnorm(Z) #prob. of obtaining more than 10500
```

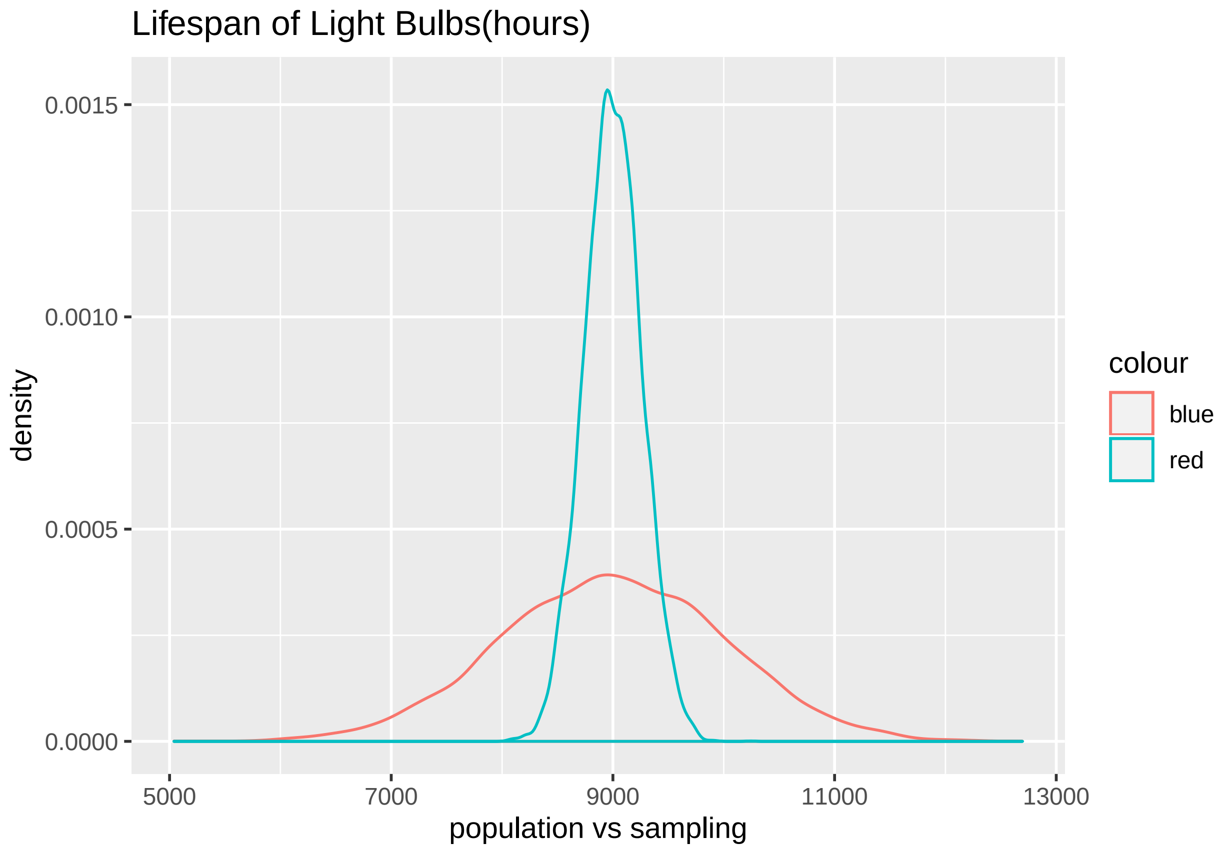
```
## [1] 3.133452e-09
```

The probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours equals to zero.

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

```r
population <- rnorm(10500, mean = 9000, sd = 1000)
sampling <- rnorm(10500, mean = 9000, sd = se)

library(ggplot2)
ggplot() +
  geom_density(aes(population, col="blue")) +
    geom_density(aes(sampling, col="red")) +
      labs(title = "Lifespan of Light Bulbs(hours)", x = "population vs sampling")
```



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

Because this analysis based on a normal distribution analysis, we cannot get an estimates if it is skewed.

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.

As n increases se decreases, then Z will increase and p-value will decrease.