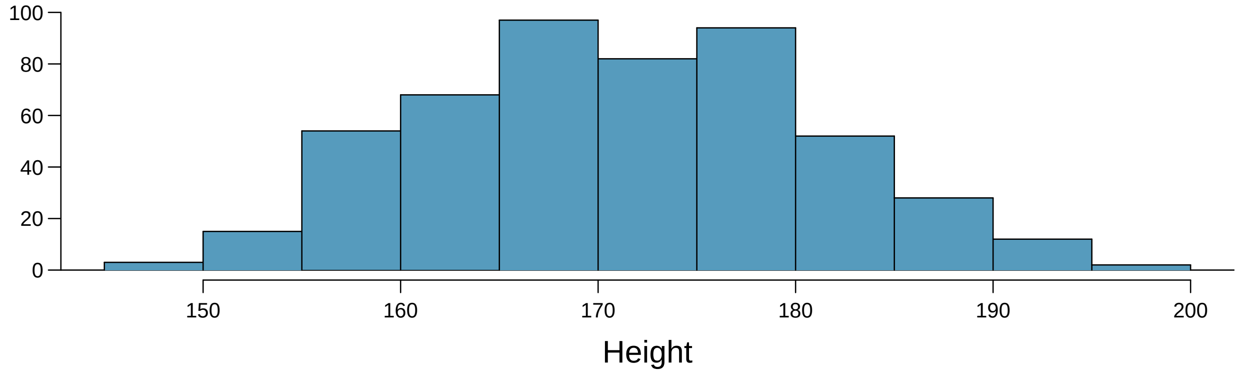
Chapter 5 - Foundations for Inference

Md Forhad Akbar

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?
Answer:

```
summary(bdims$hgt)
```

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

The point estimate for the average height is 171.1 and median is 170.3.
(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?
Answer:

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

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

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

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

The point estimate for the standard deviation is 9.4. The IQR is 3rd Qu - 1st Qu which is 177.8 - 163.8 = 14 We can also write IQR command as above to get IQR 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.
Answer:

```
z_180 <- (180-171.1)/9.4
z_180
```

```
## [1] 0.9468085
```

```
z_150 <- (155-171.1)/9.4
z_150
```

```
## [1] -1.712766
```

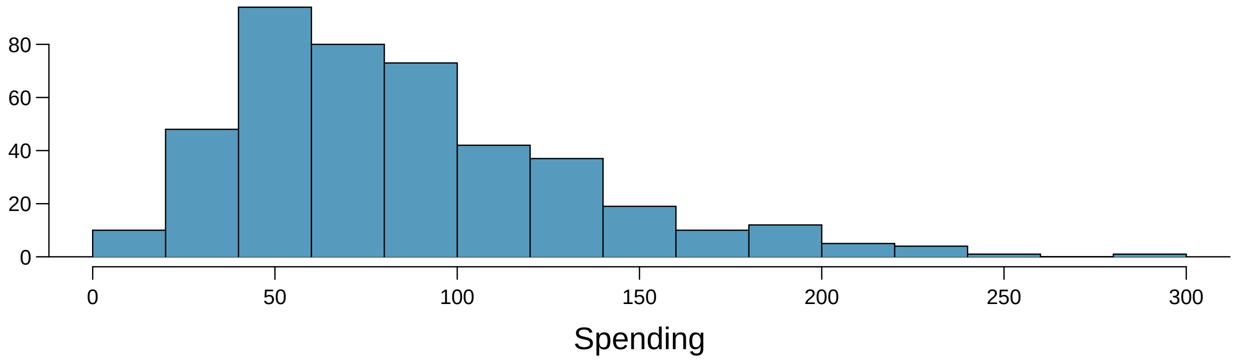
180cm height is 0.95 standard deviation above from the mean. 155cm height is 1.71 standard deviation below from the mean. They are both not unusual heights.
(d) The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.
Answer: If we take another random sample of physically active individuals i would expect the mean and the standard deviation of this new sample to be different from above as the new sampling will be done taking observation randomly. For a randomly picked sample it is usually going to provide different means and standard deviations because of the different observations however it is theoretically possible to get the same values but it is very unlikely.
(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.
Answer: We typically use the term standard error rather than standard deviation to quantify the variability, and the notation SEˆp is used for the standard error associated with the sample proportion.

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

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

The standard error of the original sample is 0.42

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.
Answer: This is true because the 95% confidence interval for this sample is between those two values.

(b) This confidence interval is not valid since the distribution of spending in the sample is right skewed.
Answer: False: The sampled observations are independent, they are randomly sampled. Sample is slighlty right skewed however it is not extremly skewed and the sample size is greater than 30.

(c) 95% of random samples have a sample mean between $80.31 and $89.11.
Answer: False: 95% confident means that 95% of the intervals contains the true population mean. This does not mean that 95% of the random samples have sample means between 80.31 and 89.11

(d) We are 95% confident that the average spending of all American adults is between $80.31 and $89.11.
Answer: False, this only true for our sample size. Statistically speaking this sample cannot represent the population.

(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.
Answer: True, In order to increase the confidence level, we create wider intervals. So instead of 95% confidence level, we are norrowing down the interval for 90% confidence level. Interval would be narrower with less possible values in terms of spending in 90% confidence interval.

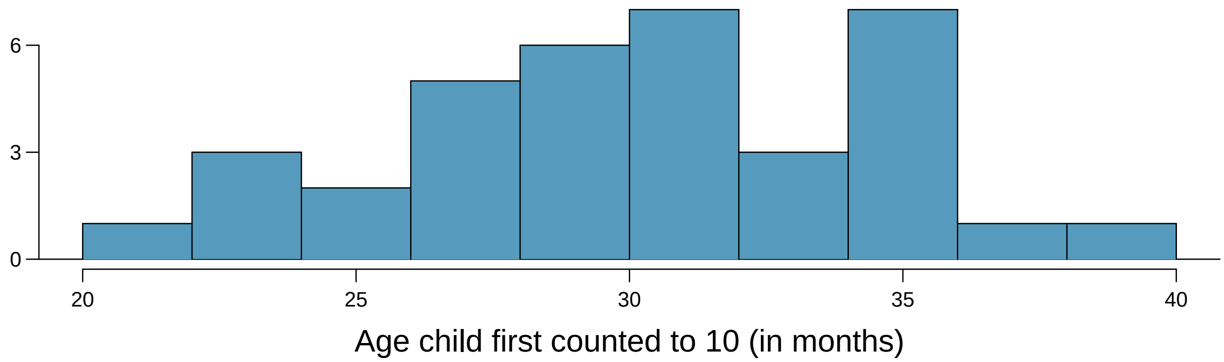
(f) In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger.
Answer: False, We will need 9 times larger sample following $SD_x = \frac{\sigma}{\sqrt{n}}$) where n is the sample size.

(g) The margin of error is 4.4.
Answer: Ture

```
1.96*sd(tgSpending$spending)/sqrt(436)
```

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

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?
Answer: Yes. The observations are independent and randomly sampled. The sample size is less than 10% of the population (considering the population of a large city). The distribution is not extremly skewed.

(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.
Answer:
Ho the null hypothesis: The average age of gifted children to successfully count to 10 is 32 months.
HA the alternative hypothesis: The average age of gifted children to successfully count to 10 is less than 32 months.
We can quantify to see if the hyphothesis by calculating p-value. If the p value is lower than significance level (0.10) we reject the Ho

```
se <- 4.31/sqrt(36)
z <- (32-(30.69))/se
se
```

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

```
pnorm(q=30.69, mean = 32, sd= se, lower.tail = TRUE)
```

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

We can reject the null hypothesis as P value is less than 0.10.

(c) Interpret the p-value in context of the hypothesis test and the data.
Answer: The probability of our null hypothesis to be true is very unlikely, it has a 3.4% chance.

Therefore, gifted children are more likely to count to 10 before 32 months.

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

```
se <- 4.31 / sqrt(36)
lower <- 30.69 - 1.65 * se
lower
```

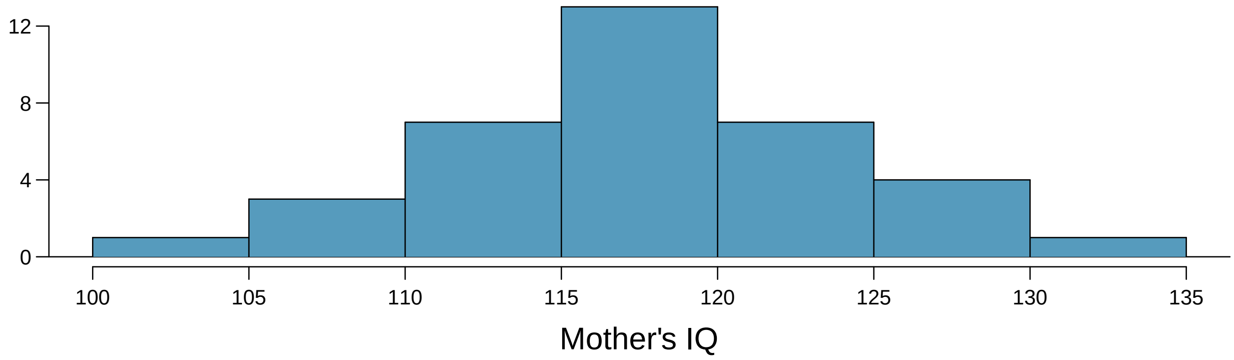
```
## [1] 29.50475
```

```
upper <- 30.69 + 1.65 * se
upper
```

```
## [1] 31.87525
```

90% confidence interval for the average age at which gifted children first count to 10 successfully is 29.50 to 31.88
(e) Do your results from the hypothesis test and the confidence interval agree? Explain.
Answer: Yes. Both (29.51) and (31.88) are below 32.

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 the se 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.
Answer:

```
summary(gifted$motheriq)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   101.0   113.8   118.0   118.2   122.2   131.0
```

```
sd(gifted$motheriq)
```

```
## [1] 6.504943
```

```
miq <- gifted$motheriq
t.test(miq, mu = 100)
```

```
## 
##  One Sample t-test
## 
## data:  miq
## t = 16.756, df = 35, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 100
## 95 percent confidence interval:
##  115.9657 120.3676
## sample estimates:
## mean of x 
##  118.1667
```

```
# calculate p value
pnorm(q=118.2, mean=100, sd=6.5, lower.tail = TRUE)
```

```
## [1] 0.9974449
```

The p value is 0.99 with significance level of 0.10, Therefore, we can reject the Ho hypthosis.
(b) Calculate a 90% confidence interval for the average IQ of mothers of gifted children.
Answer:

```
se <- 6.5 / sqrt(36)
lower <- 118.2 - 1.65 * se
lower
```

```
## [1] 116.4125
```

```
upper <- 118.2 + 1.65 * se
upper
```

```
## [1] 119.9875
```

90% confidence interval for the average IQ of mothers of gifted children is 116.41 to 119.99

(c) Do your results from the hypothesis test and the confidence interval agree? Explain.
Answer: 90% confidence interval for the average IQ of mothers of gifted children is 116.41 to 119.99. 100 is far from the confidence interval. Therefore, we can reject Ho hypothesis.

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.
Answer: Sampling distribution of the mean can be defined as probability distribution of means for ALL possible random samples OF A GIVEN SIZE from some population. For example, Take a sample of 10 random students from a population of 100 of SAT math scores. You might get a mean of 502 for that sample. Then, you do it again with a new sample of 10 students. You might get a mean of 480 this time. Then, you do it again. And again. And again...... and get the following means for each of those three new samples of 10 people: 550, 517, 472 and so on.

sampling distribution shape will become more symmetric , center would become about the same and spread would become smaller as sample size increases.

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?
Answer:

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

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

The probability that a randomly chosen light bulb lasts more than 10,500 hours is 0.0668072
(b) Describe the distribution of the mean lifespan of 15 light bulbs.
Answer: The lifespan of the lightbulbs are nearly normally distributed. The sample size 15. mean is 9000.

```
se_bulbs <- 1000/sqrt(15)
se_bulbs
```

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

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

```
z <- (10500 - 9000) / se_bulbs
pnorm(-abs(z))
```

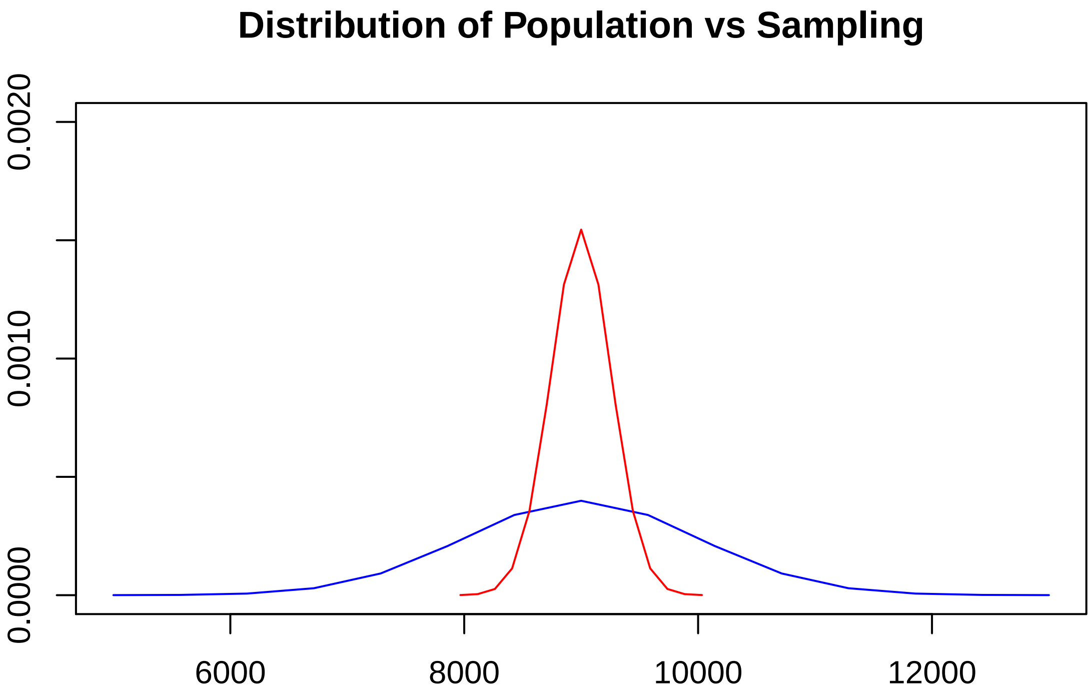
```
## [1] 3.133452e-09
```

We can see that the probability of the sample mean of 15 light bulbs being more than 10,500 is almost zero.
(d) Sketch the two distributions (population and sampling) on the same scale.
Answer:

```
sd <- 1000
mean <- 9000
se <- 1000/sqrt(15)

normsample <- seq(mean - (4 * sd), mean + (4 * sd), length=15)
randomsample<- seq(mean - (4 * se), mean + (4 * se), length=15)
hnorm <- dnorm(normsample,mean,sd)
hrandom<- dnorm(randomsample,mean,se)

plot(normsample, hnorm, type="l",col="blue",
xlab="",ylab="",main="Distribution of Population vs Sampling", ylim=c(0,0.002))
lines(randomsample, hrandom, col="red")
```



(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution?
Answer: It is likely not possible to get the estimates if the distribution is not somewhat normal, (or presumed normal). The calculation relies on conditions being met, one is that the distribution is not strongly skewed; if it is, then the sample should be sufficiently large.

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.
Answer: If we increase sample size then our the z-score will increase. Increasing z-score will decrease the probability of event occurring. A decreased probability correlates to a lower p-value.
This can be explained using an example. Lets assume (SD=10), for sample size (n) = 50 z-score(50) = 1.41421356237309, then z-score will be used to calculate p-value, and p-value effects hypothesis testing. z-score(500) = 0.447213595499958, in this case z-score decresed, this will impact p-value, and hypothesis testing As the sample size(n) increases, p-value decreases.

```r
pnorm(1.41421356237309)
```

```
## [1] 0.9213504
```

```r
pnorm(0.447213595499958)
```

```
## [1] 0.6726396
```