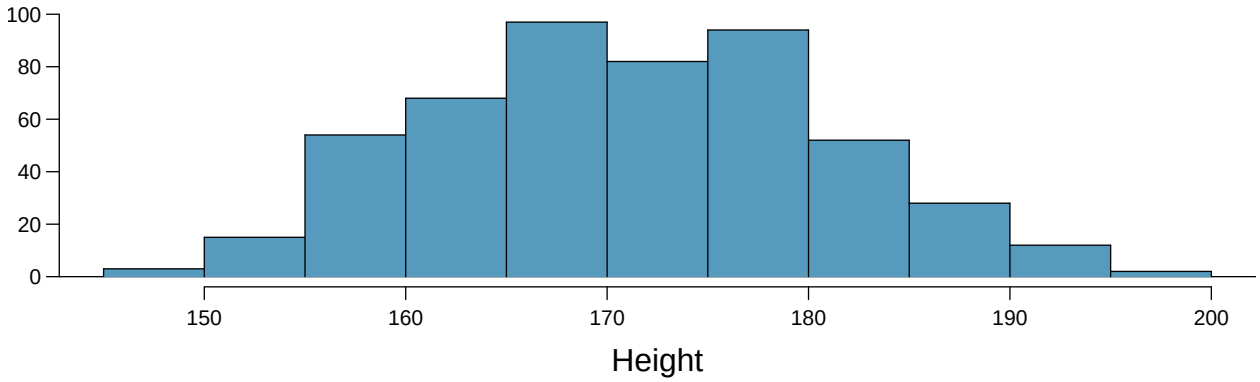
Chapter 5 - Foundations for Inference

Jose A. Mawyin

10/13/2019

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.1 <- mean(bdims$hgt)
median.1 <- median(bdims$hgt)
cat("The point estimate for the average height of active individuals is:", mean.1)
```

```
## The point estimate for the average height of active individuals is: 171.1438
```

```
cat("\nThe point estimate for the median height of active individuals is:", median.1)
```

```
##
## The point estimate for the median height of active individuals is: 170.3
```

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

```
sd.1 <- sd(bdims$hgt)
IQR.1 <- IQR(bdims$hgt)
cat("The point estimate for the standard deviation of the heights of active individuals is:" ,sd.1)
```

```
## The point estimate for the standard deviation of the heights of active individuals is: 9.407205
```

```
cat("\nThe point estimate for the IQR of active individuals is:" , IQR.1)
```

```
##
## The point estimate for the IQR of active individuals 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.

```
Height.1a <- 180
Height.1b <- 155
(abs(Height.1a-mean.1)/sd.1) %>% round(2)
```

```
## [1] 0.94
```

```
(abs(Height.1b-mean.1)/sd.1) %>% round(2)
```

```
## [1] 1.72
```

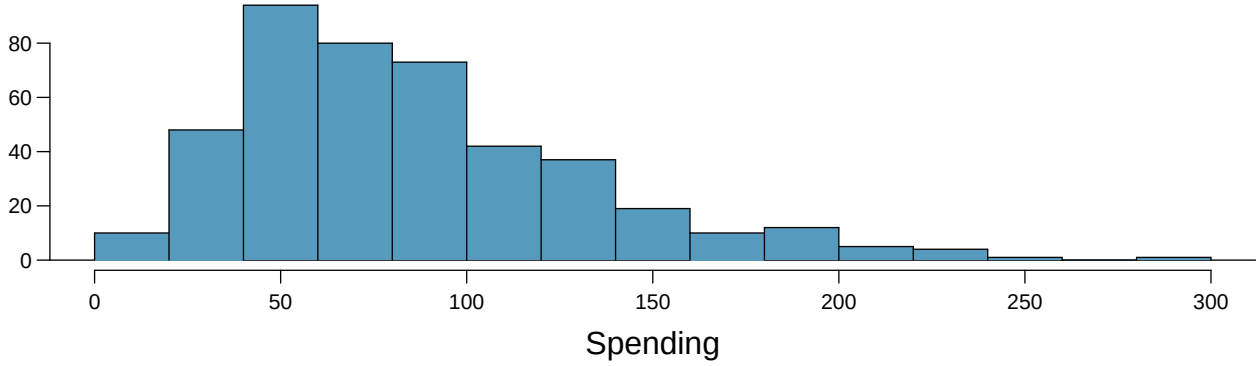
```
cat("The person who is 1m 80cm (180 cm) tall is just within a standard deviation of the sampled populati
```

```
## The person who is 1m 80cm (180 cm) tall is just within a standard deviation of the sampled populatio
## The second person, 155cm tall is more than 1.75 standard deviations from the mean of the sampled popu
```

(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 mean and standard deviation of the new sample can (most likely would not be exactley the same) be different than the previous sample. Nevertheles, I expect the new sample mean to be close to the previous mean and definetely within a standard deviation of the previous sample.*

(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.

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. 95% of random samples will have a mean 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. You can still calculate a confidence interval with non-normal skewed data.

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

True. This is the correct interpretation of 95% confidence interval.

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

False. See defintion above.

(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 90% confidence interval will have a narrower lower and higher boundary intervals.

(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.

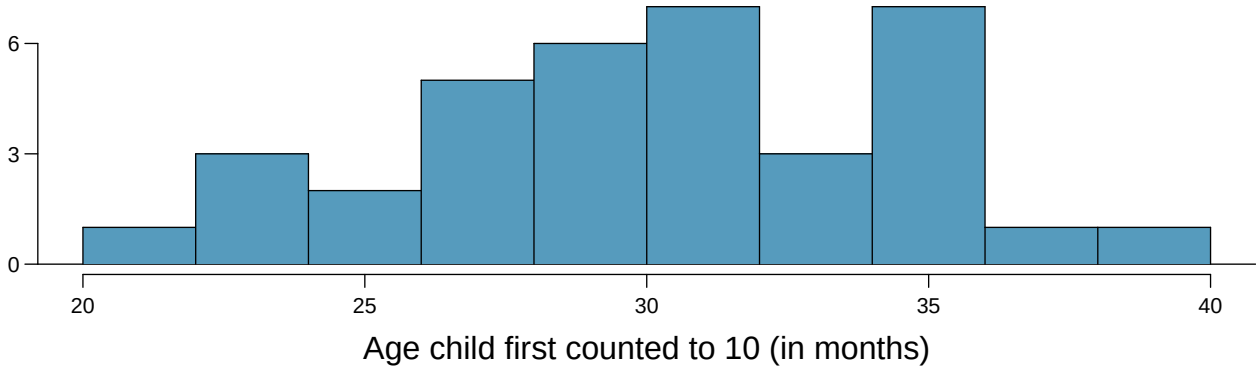
False The margin of error is proportional to the sample size as:

$$MoE \quad \propto \quad \frac{1}{\sqrt{n}}$$

Therefore, the sample size will need to be 9 times largers to decrease the margin of error by 1/3.

(g) The margin of error is 4.4.

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 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?

No tall conditions are fully satisfied. The conditions for inference is that the sample must be random, normal and large. In this case, the sample is described as random. The sample population of 36 is enough for analysis. However, the histogram shown above does not look fully normal.

(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.

Calculate Standard Error and e Z-score:

$$SE = \frac{\sigma}{\sqrt{n}} \qquad z - score = \frac{\mu - value_{null}}{SE}$$

```
n.3 <- 36
mean.3 <- 30.69
SD.3 <- 4.31
null.3 <- 32
sign.lvl.3b <- 0.1
SE.3 <- SD.3/(sqrt(n.3))
Z.S.3 <- (mean.3 - null.3)/SE.3
P.value.3b <- pnorm(Z.S.3, 0, 1)
Test.3b <- P.value.3b > sign.lvl.3b
cat("Given a null hypothesis value of", null.3," and a significance level of", sign.lvl.3b," we reject
```

```
## Given a null hypothesis value of 32  and a significance level of 0.1  we reject the null hypothesis a
```

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

The p value indicates that the children in the gifted school read earlier than normal children.

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

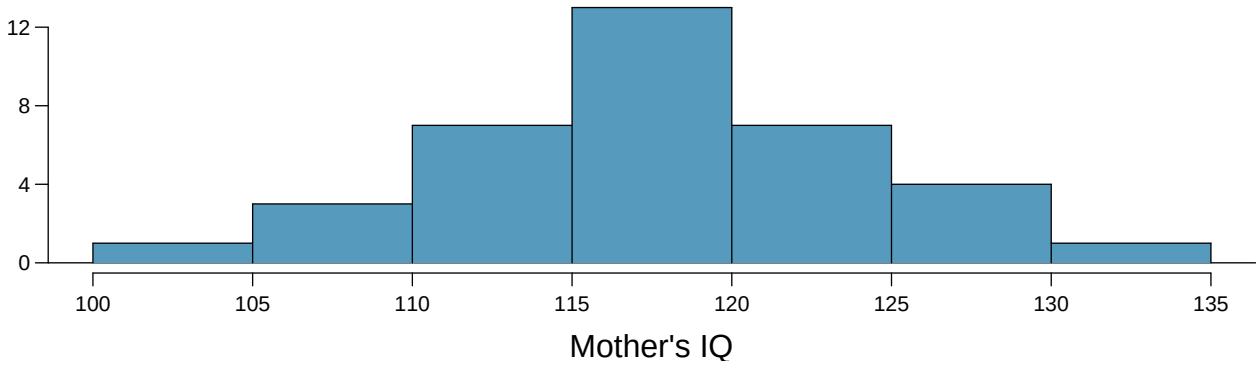
```
LowerTail.3d <- (mean.3 - 1.645 * SD.3) %>% round(2)
UpperTail.3d <- (mean.3 + 1.645 * SD.3) %>% round(2)
cat("The range", LowerTail.3d, "to", UpperTail.3d, "months is the 90% confidence interval for the averag
```

```
## The range 23.6 to 37.78 months is the 90% confidence interval for the average age at which gifted chi
```

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

Yes, results agreed with the discounted null hypotheses of 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) Performa 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.

```
n.4 <- 36
mean.4 <- 100
SD.4 <- 6.5
null.4 <- 118.2
sign.lvl.4a <- 0.1
SE.4 <- SD.4/(sqrt(n.4))
Z.S.4 <- (mean.4 - null.4)/SE.4
P.value.4a <- pnorm(Z.S.4, 0, 1)
Test.4a <- P.value.4a > sign.lvl.4a
cat("Testing that the average IQ of mothers of gifted children is different than the average IQ for the
```

```
## Testing that the average IQ of mothers of gifted children is different than the average IQ for the p
```

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

```
LowerTail.4b <- (mean.4 - 1.29 * SD.4) %>% round(2)
UpperTail.4b <- (mean.4 + 1.29 * SD.4) %>% round(2)
cat("The range", LowerTail.4b, "to", UpperTail.4b, "IQ points is the 90% confidence interval for the av
```

```
## The range 91.61 to 108.39 IQ points is the 90% confidence interval for the average IQ of mothers of
```

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

The results match the rejected null hypothesis values of 118.2 IQ points.

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.

Given a population with a finite mean μ and a finite non-zero variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of $\frac{\sigma^2}{N}$ as N, the 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.

```
mean.cflb <- 9000
sd.cflb <- 1000
```

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

```
CFLB.a <- (1-pnorm(10500, mean = mean.cflb, sd = sd.cflb, lower.tail = TRUE, log.p = FALSE)) %>% round(1
cat("What is the probability that a randomly chosen light bulb lasts more than 10,500 hours is:", CFLB.a
```

```
## What is the probability that a randomly chosen light bulb lasts more than 10,500 hours is: 0.07
```

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

The distribution of the main population is normal but we could not be assure that the distribution of the 15 light bulbs would be normal as weel because of the small sampling population.

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

```
x <- 10500
n <- 15
SE.CFLB.c <- sd.cflb/sqrt(n)
prob15 <- 1-pnorm(x, mean.cflb, SE.CFLB.c)
Prob.CFLBs.c <- round(prob15, 4)
cat("The probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours
```

```
## The probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours is
```

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

Both distributions will have similar means. However, the spread of the small sample population will be much larger.

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

The sample size is soo small that the Central Limit Theorem is not applicable in this case. Not possible to estimate the probability.

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 the sample size increases, p-values tend to become smaller. As the sample size increases, our uncertainty about where the population mean could be (the proportion of heads in our example) decreases.