IS606 - Homework 4

Daniel Dittenhafer

October 11, 2015

4.4 Heights of adults (p204)

Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height an gender, for 507 physically active individuals.

a) What is the point estimate for the average height of active individuals? What about the median? The mean would be the point estimate for the average height, which is 171.1. The median is 170.3.

b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR? Given the distribution is mostly normal, we can use the sample standard deviation as the point estimate for the population standard deviation (p173), SD=9.4.

The IQR would be derived from the sample IQR of Q3 - Q1 = 177.8 - 163.8 = 14.

c) Is a person who is 1m 80cm (180cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning. After codifying the sample mean and standard deviation, we compute the Z score for the 180cm tall scenario.

```
meanHeight <- 171.1
sdHeight <- 9.4

x <- 180
zTall <- (x - meanHeight) / sdHeight

pTall <- pnorm(zTall)
pTall
```

```
## [1] 0.8281318
```

Being 180cm tall puts one at 0.9468085 standard deviations above the mean with 17.19 % of people taller. As such, I would *not* consider being 180cm tall particularly unusual, though it is taller than 82.81 % of the sample.

```
x <- 155
zShort <- (x - meanHeight) / sdHeight
pShort <- pnorm(zShort)
pShort
```

```
## [1] 0.0433778
```

Being 155cm tall puts one at -1.712766 standard deviations below the mean with 95.66 % of people taller. As such, I would consider being 155cm tall unusual, with just 4.34 % of the sample being shorter

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 identical mean and standard deviation in the new sample unless by some coincidence the second sample cases where identical to the first. With that said, any new sample mean and standard deviation would be normally distributed around the population mean and standard deviation.

e) We quantify the variability of the point estimates through the Standard Error (SE) which is the standard deviation of the sampling distribution. Using `R` we compute the SE for the sample below:

```
n <- 507
seHeight <- sdHeight / sqrt(n)
seHeight
```

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

4.14 Thanksgiving spending, Part I (p208)

a) We are 95% confident that the average spending of these 436 American adults is between $80.31 and $89.11. More accurately, I would say we are 95% confident that the average spending of the population is between $80.31 and $89.11. It seems we should be 100% confident the average spending of the specific individuals in the survey is between the upper and lower bound. The confidence level of the interval is meant to measure the likelyhood that the population parameter falls within the range.

b) This confidence interval is not valid since the distribution of spending in the sample is right skewed. Although the sample's distribution is right skewed, this does not affect the *sampling* distribution's shape normal the confidence interval built from it when the sample size is sufficiently large. In our current case, n=436, which is well above 100 so we rely on the Central Limit Theorem for a normal sampling distribution.

c) 95% of random samples have a sample mean between $80.31 and $89.11 While this statement might be true in some particular scenario, it is not known to be true as a result of the 95% confidence internal of this data set. Rather, we expect 95% of confidence intervals at this level from random samples would contain the population's mean.

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

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. Correct. This is due to the fact that the number of standard deviations surronding the mean that encompass 90% (1.645) of the normal distribution's center is less than the number associated with 95% (1.96)

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. First we can compute the Standard Error by reversing the Margin of Error computation:

```
n <- 436
SE <- 4.4 /  1.96
SE
```

```
## [1] 2.244898
```

The Standard Error comes from the Standard Deviation divided by the square root of the number of observations.

```
stdev <- SE * sqrt(n)
stdev
```

```
## [1] 46.87485
```

Does tripling the sample size achieve a margin of error of 1.4666667 assuming standard deviation stays the same?

```
newN <- n * 3
newSE <- stdev / sqrt(newN)
newMoE <- newSE * 1.96
newMoE
```

```
## [1] 2.540341
```

The new margin of error is 2.5403412, which is not a third of the current margin of error.

In order to acheive a margin of error of 1.4666667 given a standard deviation of 46.8748456 what sample size would we need (again assuming the same standard deviation)?

```
desiredMoE <- 4.4 / 3
reqN <- ((1.96 * stdev) / desiredMoE)^2
reqN
```

```
## [1] 3924
```

```
checkMoE <- 1.96 * (stdev / sqrt(reqN))
checkMoE
```

```
## [1] 1.466667
```

Therefore, inorder to achieve a margin off error of 1.4666667, a sample of 3924 would be needed assuming current standard deviation stays the same.

g) The margin of error is 4.4. Correct. 4.4 is the result of 1.96 * SE.

4.24 Gifted children, Part I (p211)

stat	value
n	36
min	21
mean	30.69
sd	4.31
max	39

a) Are conditions for inference satisfied? Sample size is small, at 36, but above the minimum of 30. The distribution is a very rough normal shape. Maybe we can accept this as meeting the conditions for inference.

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 first count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10. Setting up the hypothesis test as follows:

$H_0 : \mu_g = 32$ (The gifted children's average months is equal to 32 (the average for children in general).)

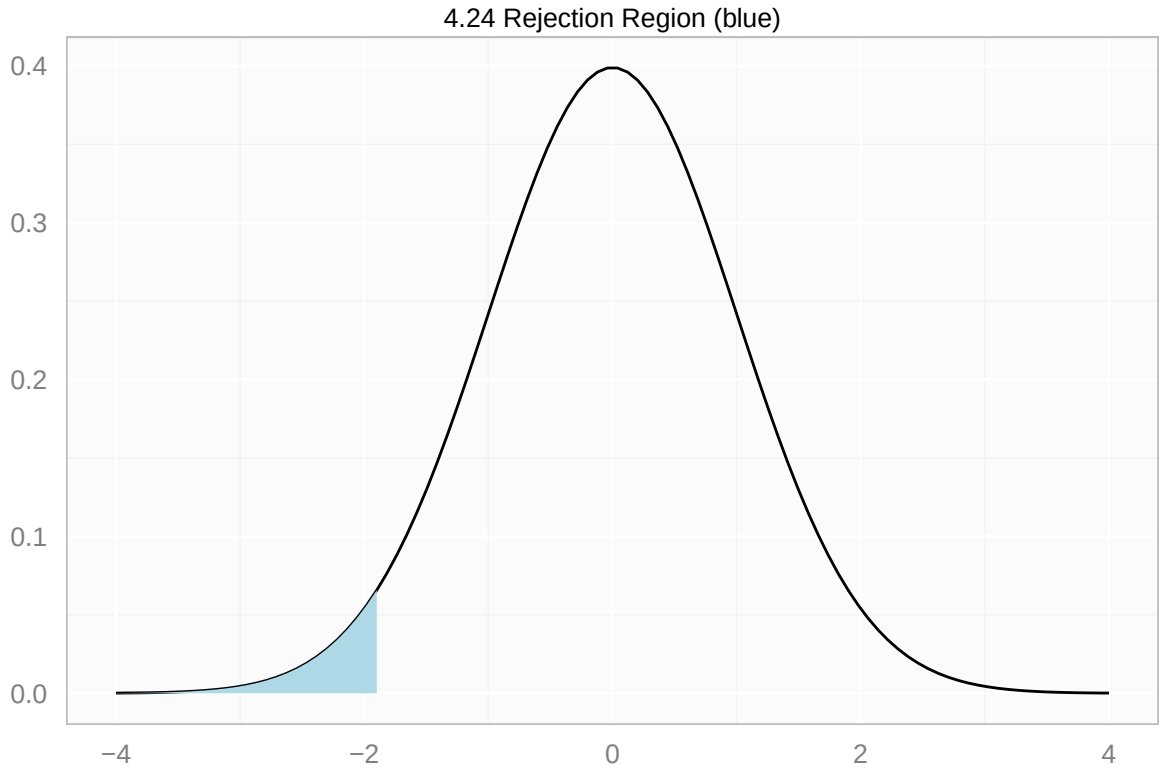
$H_A : \mu_g < 32$ (The gifted children's average months is less than 32.)

$\alpha = 0.10$

```
a <- 0.10
xbar <- 30.69
sdX <- 4.31
n <- 36
SEx <- sdX / sqrt(n)
zXbar <- (xbar - 32) / SEx
zXbar
```

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

This is a one-sided hypothesis test as shown below:



```r
pval <- pnorm(zXbar)
pval
```

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

c) Interpret the p-value in context of the hypothesis test and the data. The p-value of 0.0341 is much lower than the significance level $\alpha = 0.1$. This suggests the gifted months mean of 30.69 is not even close to the 32 month average. Therefore, I conclude to reject the null hypothesis in favor of the alternative. In other words, it is implausible that we would see a mean from our sample as low as we did if there wasn't a significant different between gifted and non-gifted children.

d) Calculate the 90% confidence interval for the average age at which gifted children first count to 10 successfully. The following R code computes the 90% confidence interval.

```r
# Determine z score of 0.10
theZ <- abs(qnorm(a))
theZ
```

```
## [1] 1.281552
```

```r
# Compute the confidence interval
lower <- xbar - (theZ * SEx)
upper <- xbar + (theZ * SEx)
ci <- c(lower, upper)
ci
```

```
## [1] 29.76942 31.61058
```

The 90% confidence interval of the gifted children is 29.7694188 - 31.6105812.

e) Do your results from the hyposthesis test and the confidence interval agree? Explain. The results agree because the range of the confidence interval does not overlap the average of 32 for non-gifted children. If the CI range had overlapped, this would indicate that 32 might be the population mean for the gifted children and would have caused us to fail to reject the null hypothesis.

4.26 Gifted Children, Part II (p212)

a) Perform a hypothesis test to evaluate if these data provide a convincing evidence that the average IQ of mothers of gifted children is different that the average IQ for the population at large, which is 100. Use a significance level of 0.10. Setting up the hypothesis test as follows:

$H_0 : \mu_g = 100$ (The gifted children's mother's IQ is equal to the average IQ.)

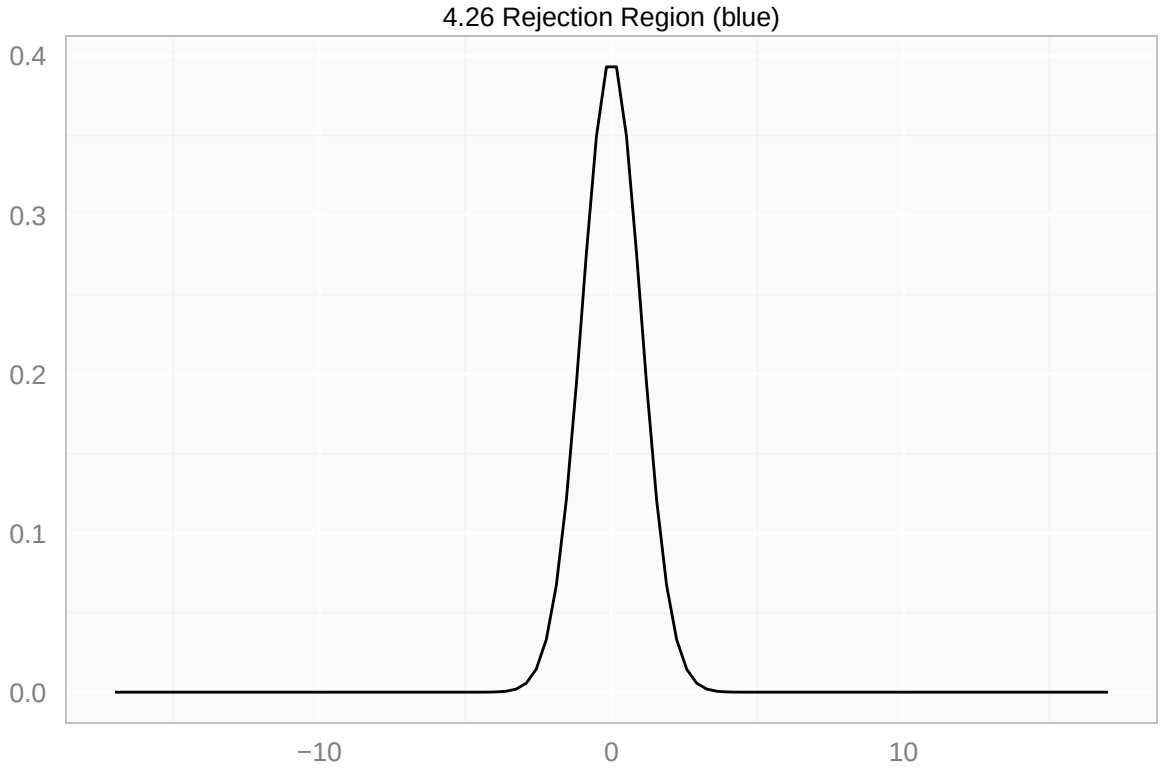
$H_A : \mu_g > 100$ (The gifted children's mother's IQ is greater than the average IQ)

$\alpha = 0.10$

```r
n <- 36
xbar <- 118.2
sdX <- 6.5
SEx <- sdX / sqrt(n)
zXbar <- (xbar - 100) / SEx
zXbar
```

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

This is a one-sided, upper tail hypothesis, though the rejection region is so small that it is not visible in the visualization.



Computing the p-value. Since this is an upper tail test, we subtract from 1.

```
pval <- 1 - pnorm(zXbar)
pval
```

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

The p-value of 0 is much lower than the significance level $\alpha = 0.1$. This suggests the IQs of mothers of gifted children mean of 118.2 is not even close to the 32 month average. Therefore, I conclude to reject the null hypothesis in favor of the alternative. In other words, it is implausible that we would see a mean from our sample as high as we did if there wasn't a significant different between gifted and non-gifted mother's IQ.

b) Calculate the 90% confidence interval for the average IQ of mothers of gifted children. The following R code computes the 90% confidence interval.

```
# Determine z score of 0.10
theZ <- abs(qnorm(a))
theZ
```

```
## [1] 1.281552
```

```r
# Compute the confidence interval
lower <- xbar - (theZ * SEx)
upper <- xbar + (theZ * SEx)
ci <- c(lower, upper)
ci
```

```
## [1] 116.8117 119.5883
```

The 90% confidence interval of the IQ of mother's of gifted children is 116.8116525 - 119.5883475.

c) Do your results from the hypothesis test and the confidence interval agree? Explain. Yes, the results agree. The confidence interval for Mother's IQ for gifted children is well above the 100 average for mother's of non-gifted children.

4.34 CLT (p214)

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 values from repeated samples from a population. The shape is approximately normal, with a center at the population mean. The shape more closely approximates the normal distribution as more samples are taken and included. This also will move the center closer to the population mean. Likewise, the spread of the sampling distribution will narrow around the population mean as more samples are included.

4.40 CFLBs (p216)

A manufacturer of compact fourescent 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? Basically we want to find the p-value of 10,500 hours.

```r
# First compute the z-score for 10,500 hours
z <- (10500 - 9000) / 1000
z
```

```
## [1] 1.5
```

```r
# Then determine the area under the normal curve at said z score.
# Since we want the area of the upper tail, we'll substract from 1.
p <- 1 - pnorm(z)
p
```

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

The probability that a randomly chosen light bulb last more than 10,500 hours is 0.0668.

b) Describe the distribution of the mean lifespan of 15 light bulbs. Assuming random sampling of 15 independent light bulbs, the distribution of the mean lifespan would be centered near population mean (claimed to be 9000 hours) and having a nearly normal shape.

c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours? Assuming the question is driving at the probability that all of some 15 randomly selected lights each have a lifespan of more than 10,500 hours, since we know the probability of one randomly chosen light bulb having a lifespan of 10,500 hours is 0.0668, the probability for 15 independent lights bulbs would be $P(15 \text{ bulbs with lifespan } = 10,500) = 0.0668^{15} = 2.3569435 \times 10^{-18} \approx 0$

Another way to look at this is how likely would a given sample of 15 light bulbs have a mean lifespan of 10,500 hours?

$H_0 : \bar{x} = 10,500$

$H_A : \bar{x} \neq 10,500$

```
mu <- 9000
s <- 1000
# Compute standard error of the mean
se <- s / sqrt(15)
se
```

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

```
# What is the z-score within the sampling distribution of a mean of 10,500
z10k5 <- (10500 - 9000) / se
z10k5
```

```
## [1] 5.809475
```

```
# Lookup the p value for this Z score.
pv <- pnorm(z10k5, mean=mu, sd=s)
pv
```

```
## [1] 1.189897e-19
```

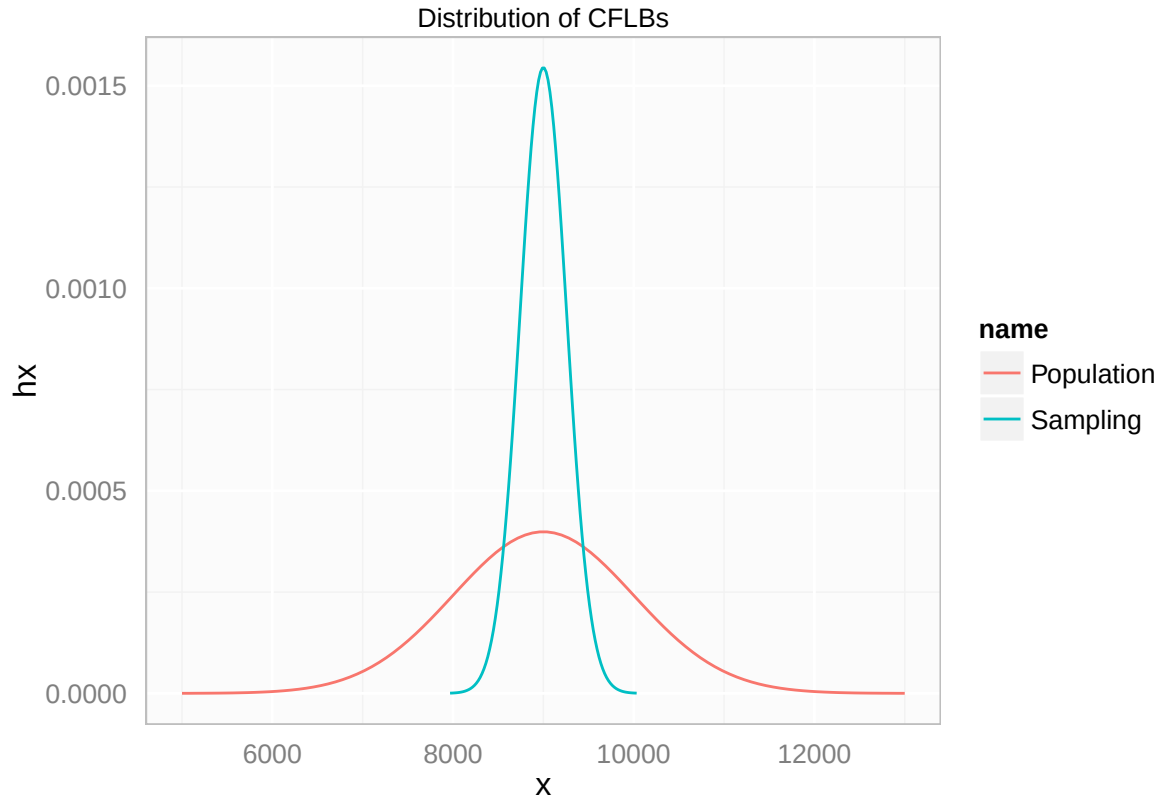
The area under normal curve for a Z score of 5.809475 is so small it is effectively zero.

d) Sketch the two distributions (population and sampling) on the same scale. Visualization is shown below the code segment that produces it:

```
x <- seq(mu - (4 * s), mu + (4 * s), length=100)
hx <- dnorm(x, mean=mu, sd=s)
df <- data.frame(name="Population", x, hx)

smpl <- seq(mu - (4 * se), mu + (4 * se), length=100)
hxSmpl <- dnorm(smpl, mean=mu, sd=se)
df <- rbind(df, data.frame(name="Sampling", x=smpl, hx=hxSmpl))

g1 <- ggplot() +
    geom_line(data=df, aes(x=x, y=hx, color=name)) +
    myTheme +
    labs(title="Distribution of CFLBs")
g1
```



e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution? I don't think my estimate for (a) would be very useful. Based on information from the OpenIntro text, I understand that there are techniques to deal with skewed distributions. Either way, the sampling distribution would still tend toward a normal shape and our existing tools could be used to estimate (c).

4.48 Same observation, different sample size (p218)

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.

First let's examine what happens to the standard error value:

```
s <- 1
se1 <- s / sqrt(50)
se2 <- s / sqrt(500)
c(se1, se2)
```

```
## [1] 0.14142136 0.04472136
```

Standard error gets smaller as sample size increases, in our example from 0.1414214 to 0.0447214. What happens next? The standard error would likely be used in the hypothesis test to determine the z score of the alternate hypothesis.

```r
xbar <- 100.3
zXbar1 <- (xbar - 100) / se1
zXbar2 <- (xbar - 100) / se2
c(zXbar1, zXbar2)
```

```
## [1] 2.121320 6.708204
```

Z score gets more extreme as sample size increases... Ok, so what does this do to the p-value?

```r
pval1 <- 1 - pnorm(zXbar1)
pval2 <- 1 - pnorm(zXbar2)
c(pval1, pval2)
```

```
## [1] 1.694743e-02 9.851675e-12
```

p-value gets smaller as sample size increases, which would give stronger evidence of a difference if any.

Conclusion: p-value of 0.08 would decrease to something signficantly smaller as shown through the example.