Raghu_606_HW_chapter4

Raghu Ramnath

3/6/2017

Chapter 4.

```
library(ggplot2)
require(ggplot2)
```

4.4 Heights of adults.

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?

ANS: point estimate or Mean:- 171.1, Median:-170.3

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

ANS: SD is 9.4 , IQR is Q3-Q1= 177.8-163.8 =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.

ANS:

```
se <- 9.4 / sqrt(507)
se

## [1] 0.4174687

lower <- 155 - 1.96 * se
upper <- 180 + 1.96 * se
c(lower, upper)

## [1] 154.1818 180.8182
```

The 95% CI is given by (154.1818,180.8182). So 180cm is not unusually tall, and 155 is not 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.

ANS:No the mean and the standard deviation would be different due to randomness in the sampling.

(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 $SD_{\bar{x}} = pn$)? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

ANS: We use standard error to quantify the variability. Standard error of mean is 0.418 .

4.14 Thanksgiving spending, Part I.

The 2009 holiday retail season, which kicked o↵ 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.

ANS: False. The sample mean is always in the confidence interval. The 95% confidence interval covers the population mean (a parameter) with 95% probability and not all 436 american adults.

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

ANS:False. The confidence interval may still be valid when the sample distribution is slightly skewed.

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

ANS: False. Samples of different size may have different confidence intervals. mean value of 95% of the random samples of size 436 lie within the confidence interval.

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

ANS: True. As per the definition of the confidence interval, the confidence interval covers the parameter value (which in this case is the average spending of an average American adult) with probability 95%.

(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.

ANS:False

(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.

ANS:

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

## [1] 2.244898

sd <- SE * sqrt(n)
sd

## [1] 46.87485

new_n <- n * 3
new_SE <- sd / sqrt(new_n)
new_MoE <- new_SE * 1.96
new_MoE

## [1] 2.540341

d_MoE <- 4.4 / 3
reqN <- ((1.96 * sd) / d_MoE)^2
reqN

## [1] 3924

c_MoE <- 1.96 * (sd / sqrt(reqN))
c_MoE

## [1] 1.466667
```

(g) The margin of error is 4.4.

ANS: True. The margin of error is given by □∗×□□, equivalently, we have (89.11-80.31)/2=4.4.

4.24 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.

(a) Are conditions for inference satisfied?

ANS: sample size 36 is small but above the minimum 30. distribution is normal shape. it satisfies the condition 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 fist count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.

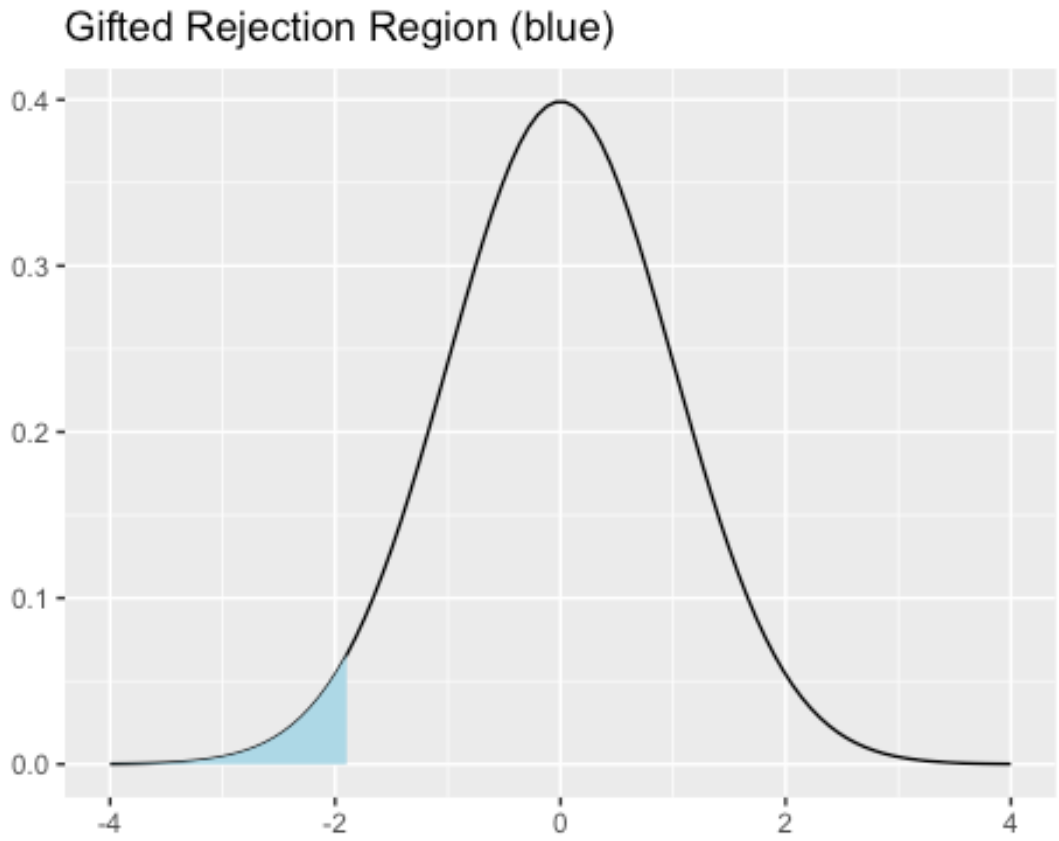
ANS:

Setting up the hypothesis test as follows:

(H_0: mean = 32) (The gifted children's average months is equal to 32) (H_A: mean < 32) (The gifted children's average months is less than 32.) (alpha = 0.10)

```
a <- 0.10
m <- 30.69
sd <- 4.31
n <- 36
SE <- sd / sqrt(n)
SE

## [1] 0.7183333

z <- (m - 32) / SE
z

## [1] -1.823666

pval <- pnorm(z)
pval

## [1] 0.0341013

x <- seq(-4, 4, length=100)
hx <- dnorm(x)
df <- data.frame(x, hx)
dfRegion <- df[df$x < z, ]

g1 <- ggplot() +
  geom_line(aes(x=x, y=hx)) +
  geom_ribbon(data=dfRegion, aes(ymin=0, ymax=hx, x=x),
fill="lightblue") +
 labs(title="Gifted Rejection Region (blue)", y="", x="")
```

```
#  myTheme
g1
```



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

ANS: Pvalue of .0341 is lower than significance level 0.10. the gifted months mean of 30.69 is not close to the 32 month average. Therefore, I conclude to reject the null hypothesis in favor of the alternative.

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

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

## [1] 1.281552

# Compute the confidence interval
lower <- m - (Z * SE)
upper <- m + (Z * SE)
ci <- c(lower, upper)
ci

## [1] 29.76942 31.61058
```

```
lower - upper
```

```
## [1] -1.841162
```

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

ANS: The results agree because the range of the confidence interval does not overlap the average of 32 for non-gifted children. If the Confidence Interval range had overlapped, this would indicate that 32 might be the population mean for the gifted children and a reson to fail to reject the null hypothesis.

4.26 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 hypothesistes 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: mean = 100) (The gifted children's mother's IQ is equal to 100) (H_A: mean > 100) (The gifted children's mother's IQ is greater than 100.) (alpha = 0.10)

```
a <- 0.10
m <- 118.2
sd <- 6.5
n <- 36
SE <- sd / sqrt(n)
SE
```

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

```
z <- (m - 100) / SE
z
```

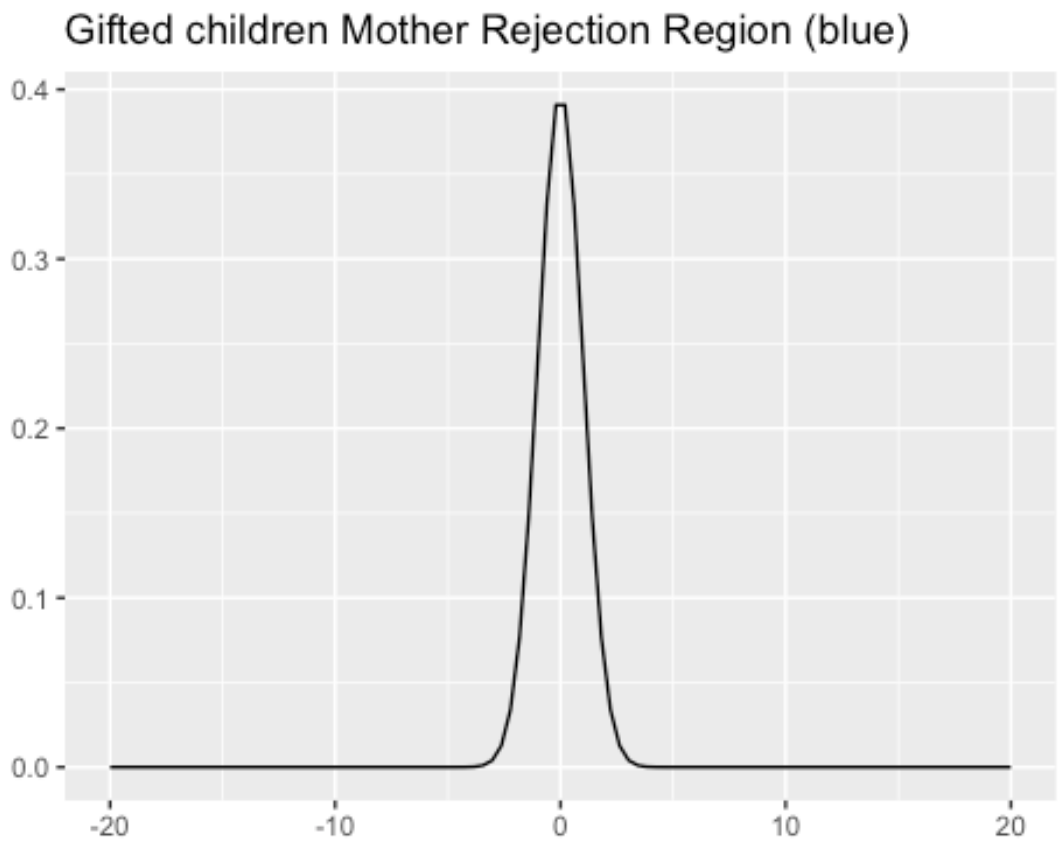
```
## [1] 16.8
```

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

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

```
x <- seq(-20, 20, length=100)
hx <- dnorm(x)
df <- data.frame(x, hx)
dfRegion <- df[df$x > z, ]

g1 <- ggplot() +
```

```
  geom_line(aes(x=x, y=hx)) +
  geom_ribbon(data=dfRegion,aes(ymin=0, ymax=hx, x=x),
fill="lightblue") +
 labs(title="Gifted children Mother Rejection Region (blue)", y="",
x="")
#  myTheme
g1
```



We conclude by rejecting H0 in favor of the alternative. We see no significant overlap. A p value of 0 implies no possibility

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

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

## [1] 1.281552

# Compute the confidence interval
lower <- m - (Z * SE)
upper <- m + (Z * SE)
```

```
ci <- c(lower, upper)
ci

## [1] 116.8117 119.5883

lower - upper

## [1] -2.776695
```

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

ANS: Yes, the results agree. The confidence interval for Mother's IQ for gifted children is above 100.

4.34 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.

ANS: The sampling distribution of the mean is the distribution of sample mean from repeated samples of a population. The shape is approximately normal, with a center at the population mean. The shape 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. 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? ANS:

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

## [1] 1.5

# 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
```

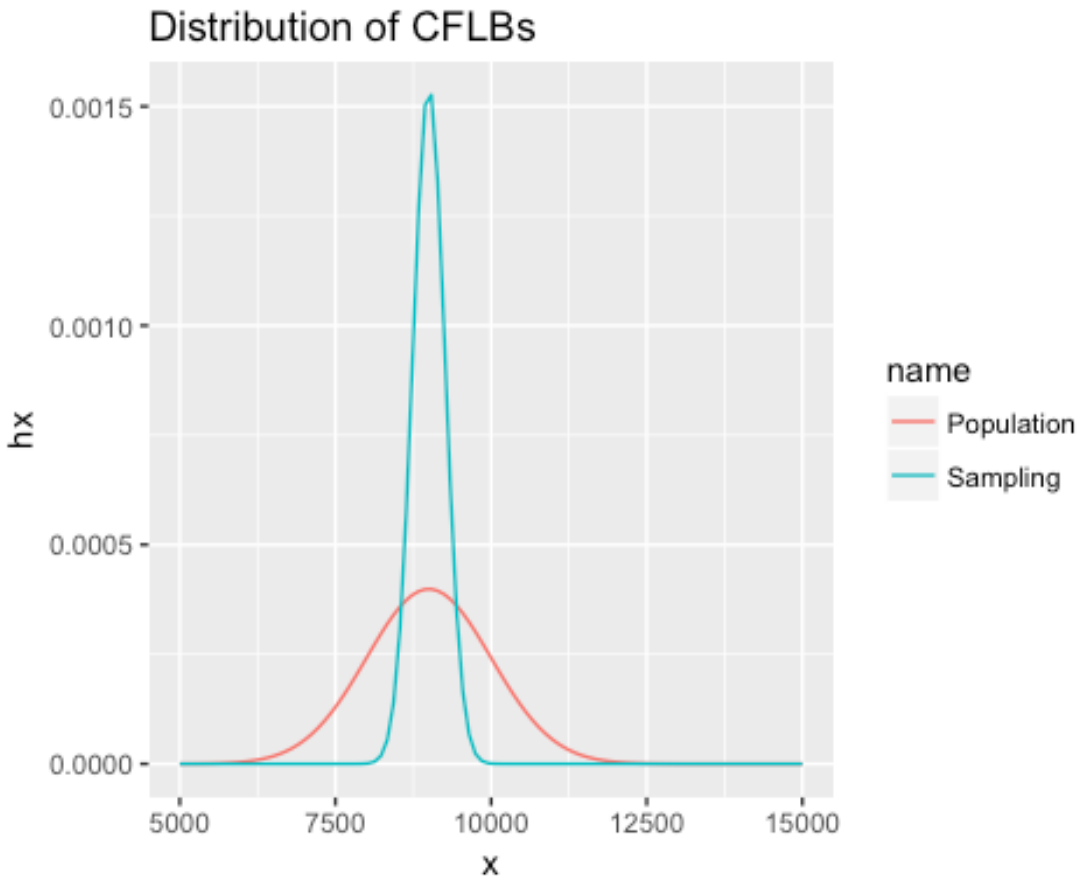
(b) Describe the distribution of the mean lifespan of 15 light bulbs. ANS: It should be close to 9000 hrs which is the mean.

(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours? ANS: probability is 0, given the very small p value.

```
me <- 9000
sd <- 1000
# standard error of the mean
se <- sd / sqrt(15)
se

## [1] 258.1989

# z-score within the sampling distribution of a mean of 10,500
z1 <- (10500 - me) / se
z1

## [1] 5.809475

# p value for this Z score.
pv <- pnorm(z1, mean=me, sd=sd)
pv

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

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

```
x <- seq(5000, 15000, length=100)
hx <- dnorm(x, mean=me, sd=sd)
df <- data.frame(name="Population", x, hx)

smpl <- seq(5000, 15000, length=100)
hxSmpl <- dnorm(smpl, mean=me, 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?

ANS: No. Since the shape and tools required to estimate is not known.

4.48 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.

ANS:The p value should decrease as the sample size increases. The standard error decreases, the z-score increases, and therefore the p value will decrease.