DATA 606 Homework 5

Ilya Kats

April 1, 2017

5.6 Working bakwards, Part II.

90% confidence interval is (65, 77), n = 25, df = 24

The sample mean is midpoint in the confidence interval, $\bar{x} = \frac{65+77}{2} = 71$.

Margin of error is $ME = 77 - 71 = 6$.

```r
t24 <- qt(0.95, df=24)
t24
```

```
## [1] 1.710882
```

$ME = t^*_{24} SE$, so $SE = \frac{ME}{t^*_{24}} = \frac{s}{\sqrt{n}}$

$s = \frac{ME\sqrt{n}}{t^*_{24}} = \frac{6*\sqrt{25}}{1.71088} = 17.53481$

ANSWER:

$\bar{x} = 71, ME = 6, s = 17.53481$

5.14 SAT scores.

$\sigma = 250, ME = 25$

a. Confidence interval is 90%, so $z^* = 1.645$ $ME = z^* \frac{\sigma}{\sqrt{n}}$, so $n = (\frac{z^*\sigma}{ME})^2$ $n = \frac{1.645^2 * 250^2}{25^2} = 1.645^2 * 100 = 270.6025$ Raina needs to collect **271 observations**.

b. Since Luke wants a narrower confidence interval, he needs to collect a larger sample to have higher confidence that the sample is representative of the population.

c. Using the formula from (a) with $z^* = 2.58$, $n = 2.58^2 * 100 = 665.64$. Luke needs to collect **666 observations**.

5.20 High School and Beyond, Part I.

$n = 200$

a. There is **no clear difference** in the average reading and writing scores. Box plots show some difference, but the mean scores are fairly close to each other. Additionally, it looks like the histogram of the difference may be centered at 0.

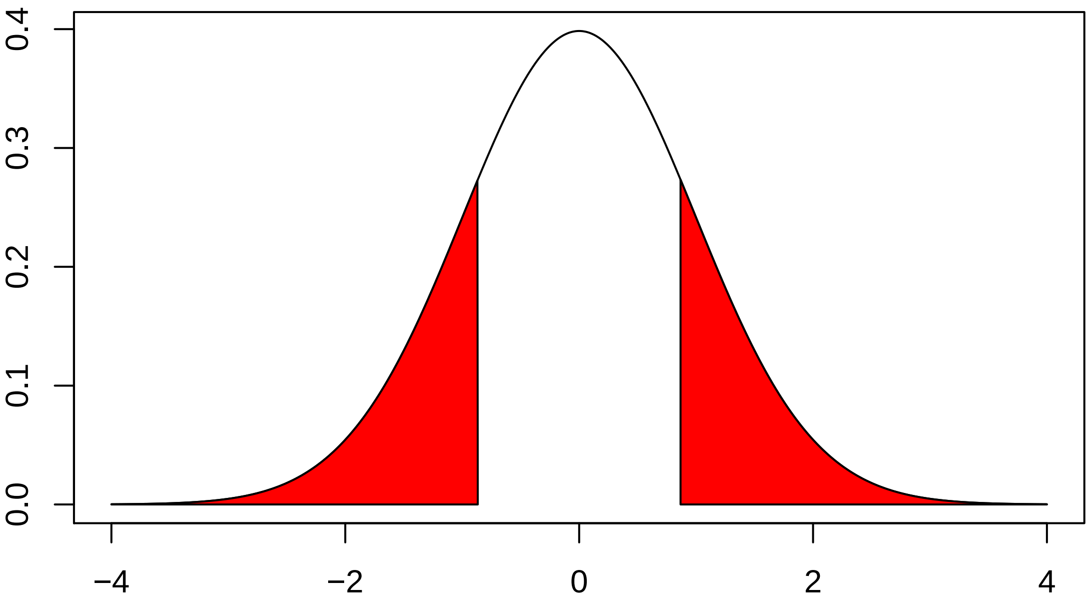
b. Reading and writing scores are **paired** rather than independent.

c. H_0 : $\mu_{read-write} = 0$ (There is no difference in the average scores in reading and writing.) H_A : $\mu_{read-write} \neq 0$ (There is a difference in average scores.)

d. The observations are based on a simple random sample and we can assume that 200 students is less than 10% of the student population, so **independence is reasonable**. The sample is **larger than 30** observations. The distribution is symmetrical and **no skew** is evident.

e. $\bar{x}_{read-write} = -0.545$, $s = 8.887$ $SE = \frac{s}{\sqrt{n}} = \frac{8.887}{\sqrt{200}} = 0.6284$ $T_{199} = \frac{-0.545-0}{0.6284} = 0.867282$

```
plot(seq(-4, 4, 0.01), dt(seq(-4, 4, 0.01), df=199), type="l", xlab = "", ylab = "")
polygon(c(-4, seq(-4, -0.867, 0.01), -0.867),
        c(0, dt(seq(-4, -0.867, 0.01), df = 199), 0),
        col = "red")
polygon(c(0.867, seq(0.867, 4, 0.01), 4),
        c(0, dt(seq(0.867, 4, 0.01), df = 199), 0),
        col = "red")
```



```
# Calculate p-value
pt(-0.867282, df=199) * 2
```

```
## [1] 0.3868321
```

Assuming 95% confidence interval, $\alpha = 0.05$, since $p - value = 0.3868 > 0.05$, we fail to reject H_0. This sample **does not provide convincing evidence** of a difference between reading and writing scores.

f. **Type II error** is possible, since we might have incorrectly failed to reject the null hypothesis.

g. We have failed to reject the null hypothesis which included the value of 0, so we would expect a confidence interval to include 0.

5.32 Fuel efficiency of manual and automatic cars, Part I.

$\bar{x}_{automatic} = 16.12, \bar{x}_{manual} = 19.85$

$s_{automatic} = 3.58, s_{manual} = 4.51$

$n_{automatic} = n_{manual} = n = 26$

$H_0 : \mu_{automatic} - \mu_{manual} = 0$ $H_A : \mu_{automatic} - \mu_{manual} \neq 0$

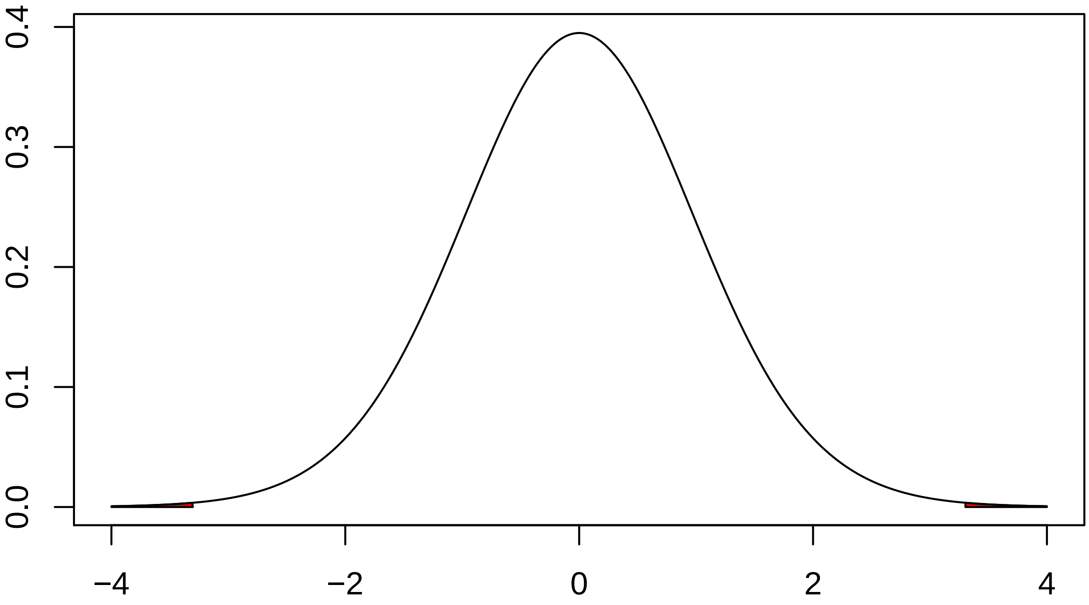
$\bar{x}_{automatic} - \bar{x}_{manual} = 16.12 - 19.85 = -3.73$

$SE = \sqrt{\frac{s^2_{automatic}}{n} + \frac{s^2_{manual}}{n}} = \sqrt{\frac{3.58^2}{26} + \frac{4.51^2}{26}} = 1.12927$

$df = 25$

$T = \frac{-3.73 - 0}{1.12927} \approx -3.303$

```
plot(seq(-4, 4, 0.01), dt(seq(-4, 4, 0.01), df=25), type="l", xlab = "", ylab = "")
polygon(c(-4, seq(-4, -3.303, 0.01), -3.303),
        c(0, dt(seq(-4, -3.303, 0.01), df = 25), 0),
        col = "red")
polygon(c(3.303, seq(3.303, 4, 0.01), 4),
        c(0, dt(seq(3.303, 4, 0.01), df = 25), 0),
        col = "red")
```



```
# Calculate p-value
pt(-3.303, df=25) * 2
```

```
## [1] 0.002883755
```

Assume confidence interval of 95%, $\alpha = 0.05$. Since $p - value = 0.003 < 0.05$, we reject H_0. The difference in average fuel efficiency of cars with automatic and manual transmissions is not due to chance.

5.48 Works hours and education.

a. H_0: Average number of hours worked is the same for all groups. H_A: Average number of hours worked varies for some (or all) groups.

b. We have to assume that the respondents were selected for the survey randomly. 1,172 observations clearly comprise less than 10% of population, so we can **assume independence within each group and across all groups**. Box plots for all groups appear symmetrical. There are outliers for some groups, but considering a relatively large sample, we can assume that the **distribution is normal**. Box plots and mean and SD value are somewhat similar between groups, so **variability is about equal**.

c. Assume confidence interval of 95%, $\alpha = 0.05$. Since $p - value = 0.0682 > \alpha$, we fail to reject H_0.

```r
# Store given values
k <- 5
n <- 1172
MSG <- 501.54
SSE <- 267382
p <- 0.0682

# Find Df
dfG <- k-1
dfE <- n-k
dfT <- dfG + dfE
df <- c(dfG, dfE, dfT)

# Find Sum Sq
SSG <- dfG * MSG
SST <- SSG + SSE
SS <- c(SSG, SSE, SST)

# Find Mean Sq
MSE <- SSE / dfE
MS <- c(MSG, MSE, NA)

# Find F-value
Fv <- MSG / MSE

# Combine all values and display
annovatb <- data.frame(df, SS, MS, c(Fv, NA, NA), c(p, NA, NA))
colnames(annovatb) <- c("Df", "Sum Sq", "Mean Sq", "F Value", "Pr(>F)")
rownames(annovatb) <- c("degree", "Residuals", "Total")

annovatb
```

```
##             Df    Sum Sq  Mean Sq  F Value Pr(>F)
## degree       4   2006.16 501.5400 2.188992 0.0682
## Residuals 1167 267382.00 229.1191       NA     NA
## Total     1171 269388.16       NA       NA     NA
```