Problem Set 3 – Andrew Celli

April 2019

**PROBLEM 1**

# PROBLEM 1 --------------------------------------------------

install.packages("faraway")

library("faraway")

data("newhamp")

NewHampH = newhamp[newhamp$votesys == "H",]

NewHampD = newhamp[newhamp$votesys == "D",]

plot(NewHampH$pObama, NewHampH$Dean, col = "blue", main = "Proportion of votes in hand-counted counties",ylab = "DEAN",xlab = "OBAMA")

points(NewHampD$pObama, NewHampD$Dean, col="red", main = "Proportion of votes in machine-counted counties",ylab = "DEAN", xlab = "OBAMA")

legend("topright", inset=0, c("Hand-counted - Blue", "Machine-counted - Red"), cex=.65)

[RESULTING IMAGE BELOW]

****

**PROBLEM 2**

1. As degrees of freedom (N – 1) increases, the distribution approximates normal, as the tails get narrower.
2. (partially adapted from statmethods.net)
3. The plot shows that the tails of the red line (df = 20) nears zero faster than the other two t distributions, but not quitre faster than the normal. Since df is related to sample size, the greater the sample size, the narrower the tails.

# PROBLEM 2 --------------------------------------------------

set.seed(20)

par(mfrow=c(1,1))

x = seq(-4,4,.1)

dnormx = dnorm(x)

dev.off

dt1 = dt(x, 1)

dt3 = dt(x, 3)

dt20 = dt(x, 20)

plot(x, col = "white", xlab="x value", cex = 1, ylim = c(0,.38), xlim = c(-4,4))

lines(x, dnormx, col = "black")

lines(x, dt1, col = "blue")

lines(x, dt3, col = "green")

lines(x, dt20, col = "red")

title(main = "t Distribution with varying Degrees of Freedom (df)", xlab = "Occurence value; N = 100")

legend("topright", inset=0, c("normal - black", "df of 1 - blue","df of 3 - green","df of 20 - red"), cex=.95)

A close up of a device

Description automatically generated

**PROBLEM 3**

* 1. Null hypothesis is AverageAge = 50; alternative is AverageAge != 50 (Age < 50; Age > 50) (2-tailed)
  2. SE = 0.4511027; Mean = 49.26133; Z-Score = 1.637469; P Value 0.102
  3. Can’t reject the null.
  4. (48.37717, 50.14549)
  5. D shows that the value 50 is within the 95% confidence interval. Because of that, the null hypothesis can’t be rejected, as long as there is a .05 confidence level.

# PROBLEM 3 --------------------------------------------------

dir()

setwd("Desktop") #had trouble getting it in normally, so found file on comp.

voteincome = read.csv("voteincome.csv")

age = as.numeric(unlist(voteincome["age"])) #learned from StackOverflow

MeanAge = mean(age)

MeanAge

sd(age)

StandardErrorAge = sd(age)/(sqrt(1500)) # Equals 0.4511027

ZScoreAge = (50-MeanAge)/StandardErrorAge # Equals 1.637469

pnorm(ZScoreAge, lower.tail = FALSE)\*2

NinetyFivePCTInterval = c(MeanAge - 1.96\*StandardErrorAge, MeanAge + 1.96\*StandardErrorAge) #Equals (48.37717 , 50.14549)

**PROBLEM 4**

1. T-score, due to the lack of SD parameter. Additionally, the sample size is <30.
2. Normal distribution of quantitative data and patrons were sampled randomly
3. Cannot reject the null, as Pval = .05815953 (T Score = -1. 6666666666666666666666666666666666667) (*rounded lol*)
4. You *could* use a Z-score, but sample size is still small so you *should* stick with T since the sample size is the same.
5. Normal distribution, large sample size and patrons were sampled randomly
6. Can reject null. Pval is now .04779032
7. PValue with the ZTest was lower and allows us to reject the null, but it was probably innapropriate to use a Zscore considering N.

# PROBLEM 4 --------------------------------------------------

SELib = 1.2/sqrt(16)

TscoreLib = (9.5-10)/SELib #EQUALS -1.666667

pt(-1.666667, df = 15, lower.tail = TRUE, log.p = FALSE)

pnorm(-1.666667, lower.tail = TRUE)#0.04779032

**PROBLEM 5**

1. Population distribution is binomial. The estimate proportion distribution is normal, applicable to a Z-test.
2. 0.4885387 is the estimated proportion who voted for Trump
3. 0.01892031
4. (0.4514549, 0.5256225)

# PROBLEM 5 --------------------------------------------------

estimateProportion = 341/(341+357) #EQUALS 0.4885387

SEofProp = sqrt((estimateProportion\*(1-estimateProportion))/(341+357)) #EQUALS 0.0189203

c(estimateProportion-1.96\*SEofProp, estimateProportion+1.96\*SEofProp) #EQUALS (0.4514549, 0.5256225)

**PROBLEM 6**

1. The authors claim that specific types of social pressure cause different degrees of voters turnout.
2. The treatment was the type of social pressure content (if any) participants received in the mail. I.e. what aspect of social pressure was highlighted in the mailer.
3. If they voted/turnout.
4. The sample size was extremely large and evenly distributed among treatment types, randomization was sufficient, and they used a control and multiple treatment types. With all this in mind, the results had sufficiently low P-values across all models. Additionally, reverse causality is practically impossible because voting behavior didn’t affect who received the “neighbors” mailer.

**PROBLEM 7**

1. Null Hypo: WomenHours = MenHours (OR estimate = 0); work in rcode below
   1. Conclusion: Can’t Reject Null.
2. The 95% Confidence in this case would include 0.
3. It is slightly skewed to the right, because it is impossible to watch less than 0 hours of TV, yet a perfectly normal distribution of the data does extend into the negative numbers. With that said, there is a large sample size, so assuming random selection, the estimates are probably close to the parameter and results are fine.

# PROBLEM 7 --------------------------------------------------

TVSe = sqrt(0.070^2+.075^2) #EQUALS 0.1025914

TvZ = (2.99-2.86)/TVSe #EQUALS 1.2671

pnorm(-TvZ)\*2 #0.2050972

pnorm(0, mean = 2.99, sd = 2.34)#EQUALS .10, so more than the .025 (PART B)

**PROBLEM 8**

* 1. Null hypo is MaleAverage == FemaleAverage (OR estimate = 0)
  2. Y1 = 2.99; Y2 = 2.86 || S1 = 2.34; S2 = 2.22 || N1 = 11; N2 = 16
  3. Calculations using formulas in Section 29 of Coursebook; results weird though…
     1. SE = .88
     2. T = .14
     3. P = .8273556
  4. Conclusion: Can’t Reject Null.

# PROBLEM 8 --------------------------------------------------with mk

SDSmallSample = sqrt(

(((11-1)\*2.34^2)+((16-1)\*2.22^2))

/

(11+16-2)

)

SESmallSample = SDSmallSample \* sqrt(((1/11)+(1/16)))

TSmall = (2.99-2.86)/SESmallSample #EQUALS .14

pt(-TSmall, df = 25)\*2 #0.8848619