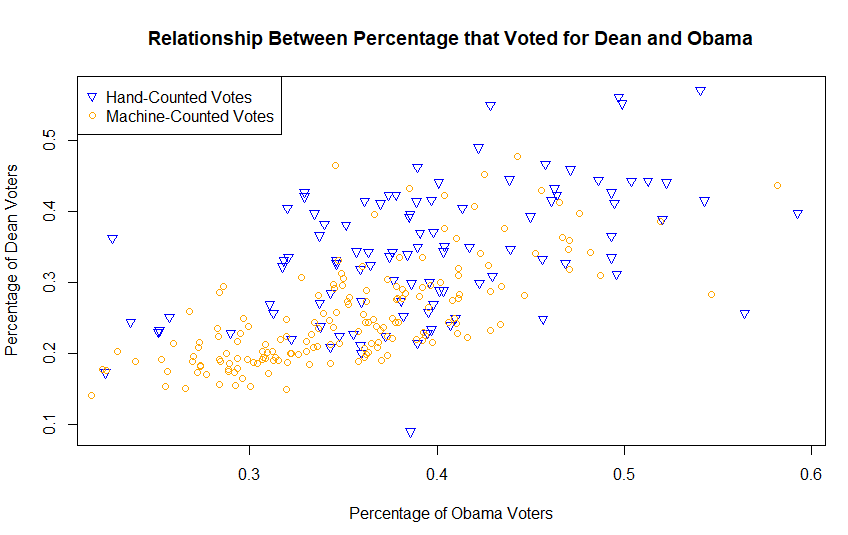
Luke Ehrenstrom

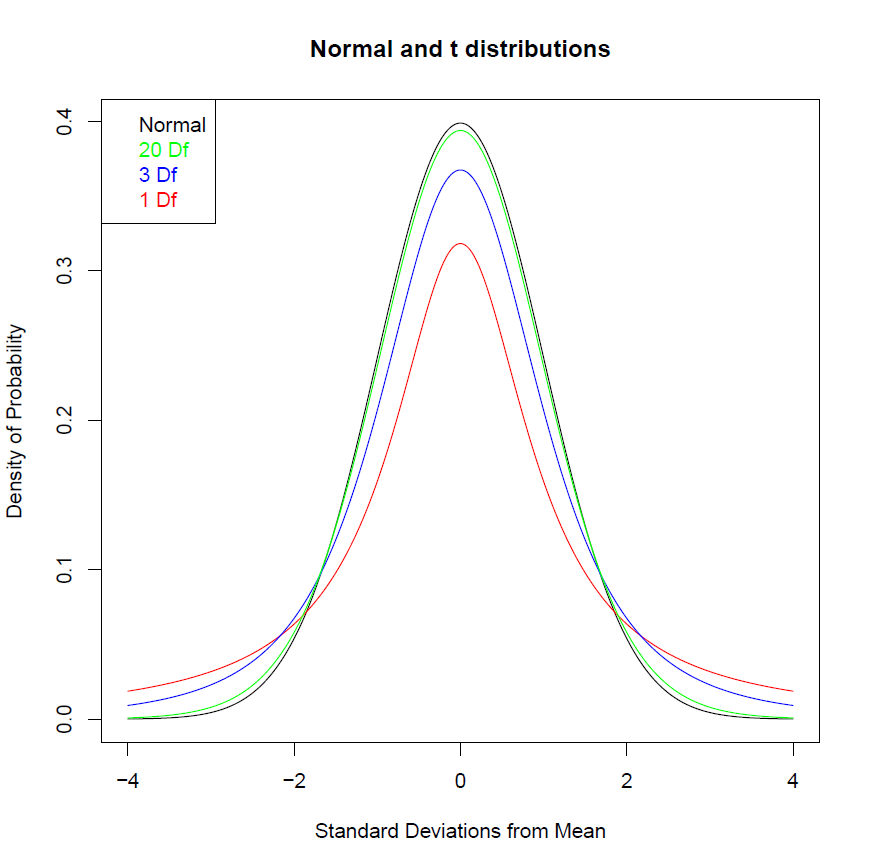
Question 1

1. install.packages("faraway")
2. library("faraway")
3. data("newhamp")
4. help("newhamp")
5. data(newhamp)
6. ObamaPerHand = newhamp[newhamp$pObama & newhamp$votesys=="H",]
7. ObamaPerMach = newhamp[newhamp$pObama & newhamp$votesys=="D",]
8. DeanPerHand = newhamp[newhamp$Dean & newhamp$votesys=="H",]
9. DeanPerMach = newhamp[newhamp$Dean & newhamp$votesys=="D",]
11. plot(ObamaPerHand$pObama, DeanPerHand$Dean, pch = 6, xlab = "Percentage of Obama Voters", ylab = "Percentage of Dean Voters",
12. main = "Relationship Between Percentage that Voted for Dean and Obama", col="blue")
13. points(ObamaPerMach$pObama, DeanPerMach$Dean, pch = 1, col="orange")
14. legend("topleft", legend = c("Hand-Counted Votes", "Machine-Counted Votes"), pch = c(6, 1), col = c("blue", "orange"))



Research on best plot for comparison between two variables: <https://mste.illinois.edu/courses/ci330ms/youtsey/scatterinfo.html>

Question 2

1. As the degrees of freedom increase, the t-distribution begins to approximate a normal distribution.
2. 
3. x = seq(-4, 4, length=1000)
4. norm = dnorm(x)
5. setwd("C:/Users/lehre/Desktop/QPM")
6. getwd()
7. pdf("PS3Question2Graph.pdf")
8. plot(x, norm, type = "l", main = "Normal and t distributions", ylab = "Density of Probability", xlab = "Standard Deviations from Mean")
9. t1 = dt(x, 1)
10. t3 = dt(x, 3)
11. t20 = dt(x, 20)
12. lines(x, t1, type = "l", lwd = 1, col="red")
13. lines(x, t3, type = "l", lwd = 1, col="blue")
14. lines(x, t20, type = "l", lwd = 1, col="green")
15. legend("topleft", legend = c("Normal", "20 Df", "3 Df", "1 Df"), text.col = c("black", "green", "blue", "red"))
16. dev.off()
17. The plot indicates that increasing degrees of freedom lead to a closer approximation of a normal distribution. As the sample size of a survey increases, one can (in accordance with the Central Limit Theorem) utilize z-tests that correspond with a normal distribution in estimations of population parameters.

Additional resources used for creating plots, adding additional lines, a legend: https://stackoverflow.com/questions/2564258/plot-two-graphs-in-same-plot-in-r

https://stackoverflow.com/questions/10543443/how-to-draw-a-standard-normal-distribution-in-r/10543555

<https://www.r-graph-gallery.com/119-add-a-legend-to-a-plot/>

Question 3

1. *H*0: μ = 50

*H*a : μ ≠ 50

α = .05

1. Standard Error: 0.4714

p = .460

z-score = -0.739

1. install.packages("Zelig")
2. library("Zelig")
3. data("voteincome")
4. ?voteincome
5. data(voteincome)
6. View(voteincome)
7. Voters = voteincome[voteincome$vote==1,]
8. S = sd(Voters$age)
9. xbar = mean(Voters$age)
10. n = length(Voters$age)
11. StandardError = S/sqrt(n)
12. mu = 50
13. zscoretest = (xbar-mu)/StandardError
14. p = pnorm(zscoretest)\*2
15. With a p-value exceeding our level of significance 0.05, we fail to reject the null hypothesis that the mean age of voters is 50 years old, providing some evidence that the mean age of voters is 50 years old.
16. zscore = 1.96
17. xbar = mean(Voters$age)
18. leftCI = xbar - zscore\*StandardError
19. rightCI = xbar + zscore\*StandardError
20. CI = c(leftCI, rightCI)
21. Both answers provide an estimate for the value of the population parameter μ. For the hypothesis test, we are unable to reject the null hypothesis, providing some evidence that the mean age of voters is 50 years old. We can be 95% confident that the parameter mean is included in the interval (48.377, 50.145) drawn from our sample, an interval that includes 50 years old.

Question 4

1. The low sample size (in accordance with the Central Limit Theorem, the sample size is less than 30) and the lack of a population standard deviation means that a t-test should be used as we lack the requirements to utilize a z-test because the sample size is not sufficiently large enough to estimate the population standard deviation.
2. We must assume that the number of books purchased per year is approximately interval, randomly sampled, and that the number of books purchased is normally distributed.
3. t-score: -1.667

p-value: .05816

We fail to reject the null hypothesis of *H*0: μ = 10 and thus there is some evidence to suggest that the mean number of books that patrons purchase is equal to 10.

1. ybar = 9.5
2. s = 1.2
3. n = 16
4. df = n - 1
5. mu = 10
6. StandardError = s/sqrt(n)
7. tscoretest = (ybar - mu)/StandardError
8. p = pt(tscoretest, df)
9. With access to a population standard deviation, you are able to use a z-test to perform hypothesis testing.
10. We must assume that the number of books purchased per year is approximately interval, randomly sampled, and that the number of books purchased is normally distributed.
11. Standard Error: 0.3

z-score: -1.667

p = .04779

Because the p-value does not exceed our level of significance of 0.05, we have sufficient cause to reject the null hypothesis of *H*0: μ = 10 and there is some evidence to suggest that the mean number of books that patrons purchase is less than 10.

1. ybar = 9.5
2. pops = 1.2
3. n = 16
4. mu = 10
5. StandardError = pops/sqrt(n)
6. zscoretest4 = (ybar-mu)/StandardError
7. p = pnorm(zscoretest4)
8. In c), we failed to reject the null hypothesis, while in f) we successfully rejected the null hypothesis. This is because in c) we utilized a t-test with low degrees of freedom that has thicker tails than the normal distribution, with the probability density spread further along the t-distribution than the standard normal distribution used in the z-test for f).

Source for reminder on assumptions required for each test: <http://www.psychology.emory.edu/clinical/bliwise/Tutorials/TOM/meanstests/assump.htm>

Question 5

1. The population distribution would resemble a binomial distribution because the choice in the survey is binary: either voting for a candidate or against a candidate. The sampling distribution of the survey would be normally distributed.
2. pi-hat = 0.4885
3. pihat = 341/698
4. Standard Error: 0.0189
5. SEPiHat = sqrt((pihat\*(1-pihat))/698)
6. Confidence interval:
7. zscore = 1.96
8. CILeft = pihat - zscore\*SEPiHat
9. CIRight = pihat + zscore\*SEPiHat
10. CI = c(CILeft, CIRight)

Question 6

1. The causal claim being made by the authors is that face-to-face voter mobilization leads to an increase in voter turnout.
2. The treatment variable is whether the group of registered voters experienced face-to-face contact or not.
3. The outcome variable is whether the registered voter voted or not.
4. After comparing the voter turnout rates between registered voters who received face-to-face contact and those who did not, increased rates of voter turnout among those received face-to-face contact were found to be statistically significant and thus providing some evidence that canvassing effectively increases voter turnout.

Question 7

1. *H*0: μ ₁ - μ ₂ = 0

*H*a : μ ₁ - μ ₂ ≠ 0

α = .05

1. n1 = 1117
2. n2 = 870
3. mu1 = 2.99
4. mu2 = 2.86
5. samplemudist = mu1 - mu2
6. stddev1 = 2.34
7. stddev2 = 2.22
8. SEMean1 = .070
9. SEMean2 = .075
10. stderror = sqrt(((stddev1^2)/n1)+((stddev2^2)/n2))
11. p = pnorm(samplemudist, 0, sd = stderror, lower.tail = FALSE)\*2
12. z = samplemudist/stderror

p = .206

z = 1.265

Because the p-value of .206 exceeds the level of significance 0.05, we fail to reject the null hypothesis and there is some evidence to suggest that there is no difference between the hours of TV watched between men and women.

1. Yes, the confidence interval would include zero. There was insufficient evidence from the sample to reject the null hypothesis that there is no difference between men and women’s TV watching habits. We can be 95% confident that the interval created from the sample statistics includes the population mean. Thus, considering we failed to reject the null hypothesis that the difference between men and women’s TV watching habits is zero, the confidence interval would include zero as a possible value for the difference between the two means.
2. In using a z-test to estimate the population mean, we must assume normality to find a p-value. In reality, there would likely be large numbers on the extremes of either watching no television or a lot of television. However, the sampling distribution of the mean hours of TV watched would be approximately normal, and that is what is used to predict the mean hours of TV watched. While the sampling distribution would far more closely resemble a normal distribution than the actual population distribution, they would likely have similar means, and that is what allows us to use a sampling distribution to predict the parameter mean.

Question 8

1. *H*0: μ ₁ - μ ₂ = 0

*H*a : μ ₁ - μ ₂ ≠ 0

α = .05

1. n1 = 11
2. n2 = 16
3. df = n1 + n2 - 2
4. mu1 = 2.99
5. mu2 = 2.86
6. samplemudist = mu1 - mu2
7. stddev1 = 2.34
8. stddev2 = 2.22
9. SEMean1 = .070
10. SEMean2 = .075
11. stddevhat = sqrt((((n1 - 1)\*(stddev1^2))+((n2 - 1)\*(stddev2^2)))/((n1+n2-2)))
12. SEttest = stddevhat\*sqrt((1/n1)+(1/n2))
13. t = (samplemudist)/SEttest
14. p2 = pt(t, df=df, lower.tail=FALSE)\*2

p = 0.885

t = .1462

Because the p-value exceeds the level of significance 0.05, we fail to reject the null hypothesis that there is no difference between the mean number of hours of TV watched by men and women, providing some evidence that there is no difference between the numbers of hours of TV watched by men and women.

Collaborated with: Sinclair Bowman, Michael Rimmey