Lab 5: Hypothesis Testing
Due at 11:59PM on December, 9th as a PDF via websubmit.

Vidya Akavoor

December 9, 2016

Collaborators: Sreeja Keesara, Lauren DiSalvo
Sources:

1. Piazza `https://piazza.com/class/is95ssbu4rq3t5?cid=434#`
2. Z-Test `http://www.investopedia.com/terms/z/z-test.asp?lgl=no-infinite`
3. One-sided z-values `https://people.ucsc.edu/~dgbonett/docs/psyc204/204Tables.pdf`
4. How to plot a normal density curve in R `http://gastonsanchez.com/how-to/2015/05/15/normal-curve/`
5. Index of /R-manual/R-devel/library/stats/html `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/`

Late days for this assignment: 0
Total late days this semester: 3 days left

z-test for the population mean (σ is known)

3.1 Learning Example:

Problem 1: Null Hypothesis H_0: Average Math SAT score $\mu \leq 500$
Alternative Hypothesis H$_1$:$\mu > 500$

Problem 2: We can use the z-test on our analysis because since the distribution is a normal distribution, it does not matter what the sample size is and we can use the test on the four random samples that we have.

Problem 3: $\overline{\mathrm{x}} = 550$

Problem 4: $\sigma = 100$

Problem 5: sample size = 4

Problem 6:

$$z = \frac{550 - 500}{\frac{100}{\sqrt{4}}} = 1$$

This means that our sample mean is one standard deviation away from the actual mean.

Problem 7: p-value = .1587

Problem 8: We cannot reject the null hypothesis because .1587 is greater than .05.

Problem 9: The z-value that comes closest to giving us a p-value less that .05 is 1.65 so:

$$1.65 = \frac{550 - 500}{\frac{100}{\sqrt{n}}}$$

n = 10.89 so we need 11 people as the sample size to reject the null hypothesis with a 95% significance level.

Problem 10:

```
significance <- function(n, sig, mu, x){
  z = (x-mu)/(sig/(sqrt(n)))
  p = 1- pnorm(z)
  paste("z = ", z, " p = ", p)
}
```

output: ”z = 1 p = 0.158655253931457”

Problem 11:

```
significance <- function(n, sig, mu, x){
  z = (x-mu)/(sig/(sqrt(n)))
  p = 1- pnorm(z)
  significant = ifelse(p<=0.05, "yes", "no")
  paste("n = ", n, " z= ", z, " p = ", p, " significant = ", significant)
}
```

```
significance(4:14, 100,500, 550)
”n = 4 z = 1 p = 0.158655253931457 significant = no”
”n = 5 z = 1.11803398874989 p = 0.131776238641486 significant = no”
”n = 6 z = 1.22474487139159 p = 0.110335680959923 significant = no”
”n = 7 z = 1.3228756555323 p = 0.0929383661829379 significant = no”
”n = 8 z = 1.4142135623731 p = 0.0786496035251425 significant = no”
”n = 9 z = 1.5 p = 0.0668072012688581 significant = no”
”n = 10 z = 1.58113883008419 p = 0.0569231490033291 significant = no”
”n = 11 z = 1.6583123951777 p = 0.0486272142195017 significant = yes”
”n = 12 z = 1.73205080756888 p = 0.0416322583317752 significant = yes”
```

”n = 13 z = 1.80277563773199 p = 0.035711728864926 significant = yes”

”n = 14 z = 1.87082869338697 p = 0.0306844145697011 significant = yes”

The minimum sample size for which we can reject the null hypothesis is 11 as the p-value at 10 is over .05 and the p-value at 11 is under .05. This also follows with our theoretical value from before.

3.2 Problem:

Problem 12: Null Hypothesis, H_0: The combined city and highway miles per gallon (mpg) of two-seater automobiles, $\mu \leq 20$.
Alternative Hypothesis, H_1 :$\mu > 20$

Problem 13: Since the sample size is 71 (which is greater than 30) it does not matter that we do not know the distribution and the conditions for the z-test have been met.

Problem 14:

$$z = \frac{20.38028169 - 20}{\frac{4.7}{\sqrt{71}}} = 0.68$$

The p-value based on z and the normal table is 0.2483.

Problem 15:

```
significance <- function(n, sig, mu, x, alt){
  z = (x-mu)/(sig/(sqrt(n)))
  if (alt == "two.sided"){
    p = (1 - pnorm(z))*2
  }
  else{
    p = 1- pnorm(z)
  }
  paste("x = ", x, " n = ", n, " z = ", z, " p = ", p)
}
```

Problem 16: output:
> significance(71, 4.7, 20, 20.38028169, ”greater”)
”x = 20.38028169 n = 71 z = 0.681768186369495 p = 0.247692771885673”

The z and p-values are very close to the theoretical ones that we calculated.

Problem 17: Based on the p-value, we do not reject the null hypothesis because .24 is greater than .05.

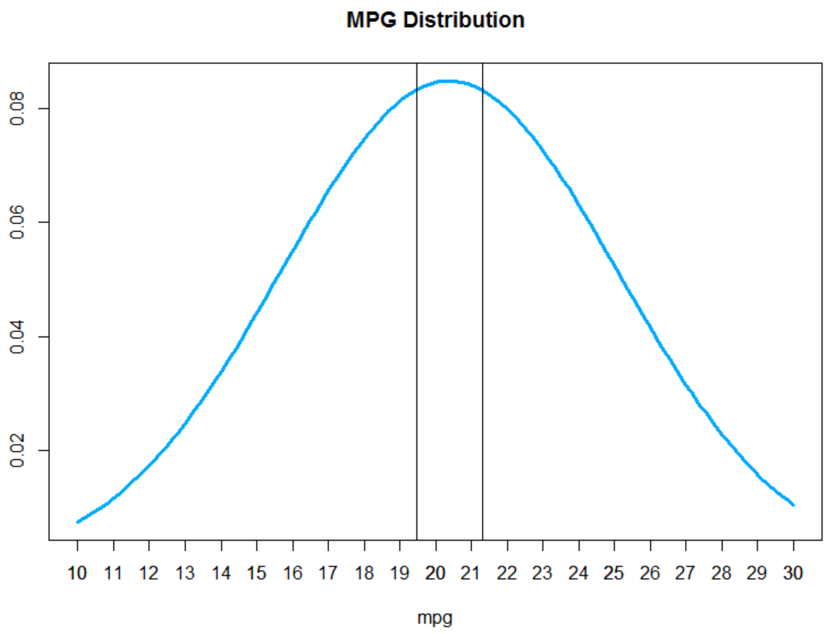
Problem 18:

$$UpperBound : 20.38028169 + 1.645 * (\frac{4.7}{\sqrt{71}}) = 21.297841929$$

$$LowerBound : 20.38028169 - 1.645 * (\frac{4.7}{\sqrt{71}}) = 19.462721451$$

Problem 19:

```
x <- seq(10, 30, length.out = 100)
y <- dnorm(x, mean = 20.38028169, sd = 4.7)
plot(x, y, type = ’l’, col = ’#00ABFF’, lwd = 3, axes = TRUE, xlab = ’mpg’, ylab = ’’,
    main = ’MPG Distribution’)
axis(side = 1, at = seq(10,30, by = 1))
abline(v=21.297841929)
abline(v=19.462721451)
```



Problem 20: The z-value that comes closest to giving us a p-value less that .05 is 1.65 so:

$$1.65 = \frac{20.38028169 - 20}{\frac{4.7}{\sqrt{n}}}$$

$n = 415.86$ so we need 416 cars as the sample size to reject the null hypothesis with a 95% significance level.

3.3 Relating Hypothesis Tests and Confidence Intervals

Problem 21:

$$UpperBound : 550 + 1.645 * (\frac{100}{\sqrt{4}}) = 632.25$$

$$LowerBound : 550 - 1.645 * (\frac{100}{\sqrt{4}}) = 467.75$$

Problem 22: The population mean ($\mu_0 = 500$) falls inside the confidence intervals.

Problem 23:

$$UpperBound : 550 + 1.645 * (\frac{100}{\sqrt{11}}) = 599.60$$

$$LowerBound : 550 - 1.645 * (\frac{100}{\sqrt{11}}) = 500.40$$

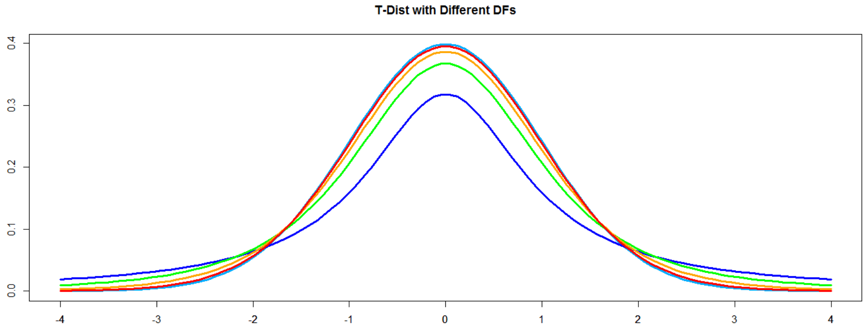
Problem 24: The population mean ($\mu_0 = 500$) falls outside the confidence intervals.

Problem 25:

```
x <- seq(-4, 4, length.out = 100)
y <- dnorm(x, mean = 0, sd = 1)
plot(x, y, type = ’l’, col = ’#00ABFF’, lwd = 3, axes = TRUE, xlab = ’’, ylab = ’’, main
    = ’T-Dist with Different DFs’)
axis(side = 1, at = seq(-4, 4, by = 1))

t1 <- dt(x,df=1)
t3 <- dt(x,df=3)
t8 <- dt(x,df=8)
t30 <- dt(x,df=30)


lines(x,t1,type="l",col="blue",yaxt=’n’,ann=FALSE, lwd = 3)
par(new=TRUE)
lines(x,t3,type="l",col="green",yaxt=’n’,ann=FALSE, lwd = 3)
par(new=TRUE)
lines(x,t8,type="l",col="orange",yaxt=’n’,ann=FALSE, lwd = 3)
par(new=TRUE)
lines(x,t30,type="l",col="red",yaxt=’n’,ann=FALSE, lwd = 3)
```



Key:
Dark Blue: df=1
Green: df=3
Orange: df=8

Red: df=30
Light Blue: normal

Problem 26: As we increase the degrees of freedom, the t-distribution looks more like the normal distribution.

Problem 27: The t-test has the same conditions as the z-test (that the samples need to be random and either follow a normal distribution or be more than 30 in number) and since we can use the z-test on the SAT analysis, we can also use the t-test.

Problem 28: We have 3 degrees of freedom because df = sample size $-1 = 4 - 1 = 3$

Problem 29:

$$t = \frac{550 - 500}{\frac{100}{\sqrt{4}}} = 1$$

This means that our sample mean is one sample standard deviation away from the actual mean.

Problem 30: 1 - pt(1, df = 3)
p-value = 0.1955011

Problem 31: The p-value from the t-test was larger than the p-value from the z-test. This is not surprising because the t-distribution is more spread out, so there is a higher probability of being closer to the tails.

Problem 32: We cannot reject the null hypothesis because the p-value from the t-test (.1955011) is greater than .05 (the 5% significance level).

Problem 33: qt(0.05, 3) = 2.353363
confidence intervals:

$$UpperBound : 550 + 2.353363 * (\frac{100}{\sqrt{4}}) = 667.67$$

$$LowerBound : 550 - 2.353363 * (\frac{100}{\sqrt{4}}) = 432.33$$

The confidence interval is wider than the one computed using the z-test. This is because the t-distribution is more spread out, so it takes a wider interval to capture the same probability.