DATA 557 - Exercise 2

Will Wright

1/16/2019

Question 1

A study is done to determine if enhanced seatbelt enforcement has an effect on the proportion of drivers wearing seatbelts. Prior to the intervention (enhanced enforcement) the proportion of drivers wearing their seatbelt was 0.8. The researcher wishes to test the null hypothesis that the proportion of drivers wearing their seatbelt after the intervention is equal to 0.8 (i.e., unchanged from before). The alternative hypothesis is that the proportion of drivers wearing their seatbelt is not equal to 0.8 (either < 0.8 or > 0.8). After the intervention, a random sample of 100 drivers was selected and the number of drivers wearing their seatbelt was found to be 85.

1.1. Conduct a test of the null hypothesis with type I error probability 0.05. Should the null hypothesis be rejected? Explain the steps you used, including showing the Z statistic, the rejection rule, and state the conclusion of the test.

```
x <- 85
n <- 100
p0 <- 0.8

z <- (x-n*p0)/sqrt(n*p0*(1-p0))
z
```

```
## [1] 1.25
```

```
abs(z)>qnorm(1-0.05/2)
```

```
## [1] FALSE
```

```
# or with a p-val and alpha of 0.05
a <- 0.05
p <- 2*(1-pnorm(z))
p
```

```
## [1] 0.2112995
```

```
ifelse(p<a, print("Reject"), print("Do not reject"))
```

```
## [1] "Do not reject"
```

```
## [1] "Do not reject"
```

Because p>0.05, we fail to reject the null hypothesis.

1.2. Calculate the exact type I error probability for your rejection rule using the binomial distribution. Compare with the value obtained using the normal approximation.

```
dbinom(x,n,p0)
```

```
## [1] 0.04806179
```

```
#hmmmmmmmmmmmmmmmmmmmm
x-qnorm(1-0.05/2)*sqrt(n*p0*(1-p0))
```

```
## [1] 77.16014
```

```
x+qnorm(1-0.05/2)*sqrt(n*p0*(1-p0))
```

```
## [1] 92.83986
```

```
sum(dbinom(c(0:floor(n*p0-qnorm(0.975)*sqrt(n*p0*191-p0)),ceiling(n*p0+qnorm(0.975)*sqrt(n*p0*191-p0)):n
```

```
## [1] 2.037036e-10
```

1.3. Calculate the power of your hypothesis test to detect an alternative proportion of 0.9. Use the binomial probabilities for the calculation. Is the power adequate to detect this alternative hypothesis?

1.4. Calculate a 95% confidence interval for the population proportion of drivers wearing seatbelts after the intervention. Compare the confidence interval with the result of the hypothesis test. Do they give the same conclusions?

Question 2

A researcher is interested in measurements of a pollutant in water samples. In particular, there is a question about whether the value changes if the sample is tested when it is older compared with being tested right after it is collected. The researcher does not know whether aging could increase or decrease the pollutant concentration. To test the hypothesis 15 samples of water were taken from a lake. Each sample was divided into 2 aliquots, one to be analysed right away and the other to be analysed 1 month later. The difference between pollutant concentrations was recorded for each of the samples. The values obtained for the differences (fresh sample - aged sample), arranged from smallest to largest, were as follows: -5, -2, -1, -1, 0, 0, 2, 3, 4, 4, 5, 5, 6, 6, 11.

2.1. Define the null hypothesis and alternative hypothesis in words and also using a mathematical equation in terms of the parameter of interest.

H0 is that there is no difference in pollutants (that the means are the same); Xbar = 0 H1 is that there is a difference in pollutants; Xbar $\neq$ 0

2.2. Perform a test of the null hypothesis with type I error probability 0.05. Assume the value for the variance of the differences is equal to 15. Define the rejection rule, state whether or not you would reject the null hypothesis, provide the p-value for the test, and state your conclusion from the experiment in words.

```
sample_diff <- c(-5, -2, -1, -1, 0, 0, 2, 3, 4, 4, 5, 5, 6, 6, 11)
n <- length(sample_diff)
p <- 0.05
var <- 15
se <- sqrt(var)/sqrt(n)
se
```

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

```
xbar <- mean(sample_diff)
```

```r
rejectionRegion <- c(0-se*qnorm(0.975),0+se*qnorm(0.975))
xbar
```

```
## [1] 2.466667
```

```r
cat("Because the sample mean is greater than 1.96, we reject the null hypothesis that there is no diffe
```

```
## Because the sample mean is greater than 1.96, we reject the null hypothesis that there is no differe
```

```r
z <- abs(xbar)/se
z
```

```
## [1] 2.466667
```

```r
p <- 2*(1-pnorm(z))
cat(paste0("The p-value is ",round(p,4)))
```

```
## The p-value is 0.0136
```

2.3. Calculate a 95% confidence interval for the mean difference in concentration between fresh and aged samples. Compare with the results of the hypothesis test. Do the confidence interval and hypothesis test give the same conclusions?

```r
confInterval <- c(xbar-se*qnorm(0.975),xbar+se*qnorm(0.975))
confInterval
```

```
## [1] 0.5067027 4.4266307
```

2.4. Design and conduct a simulation study to assess the validity of the hypothesis testing procedure. Choose a plausible distribution for the differences in pollutant measurements, with the requirement that the variance of your distribution is equal to 15 (there is no unique correct answer for this!). Choose a reasonable number of samples to simulate. Estimate the type I error probability under your simulation model. What do you conclude about the validity of the hypothesis test for this experiment?

Question 3 (continuation of question 2)

3.1. Perform a test of the null hypothesis and provide a 95% confidence interval for the mean difference without the assumption that the variance is known to be equal to 15. What is the name of the test? Compare results with the results from question 2. Explain how and why they differ.

```r
# H0 <-
```

3.2. Design and conduct a simulation study to assess the validity of the test used in 3.1 for this experiment. Choose a distribution for the population that is a plausible distribution for the differences in pollutant measurements (for this question, it is not necessary to require that the variance is equal to 15). Estimate the type I error of the test. Is the test valid? Explain any differences. Compare to the results for your simulation study in question 2.4.

3.3. Propose an alternative distribution for the population that might give different results for the type I error. Estimate the type I error of the test under this distribution and compare with results of the previous simulation. Explain any differences.

3.4. Suppose that it was determined that the last data value (11) was an error due to failure of the measuring equipment. Re-run the test and confidence interval with this value excluded. How did the results change?