ST 551 Homework 2

Luis Garcia Lamas

Fall 2022

Instructions

This assignment is due by 11:59 PM, October 17th on Canvas via Gradescope. **You should submit your assignment as a typed PDF which you can compile using the provide .Rmd (R Markdown) template.** Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope. You should also use complete, grammatically correct sentences for your solutions.

Question 1 (27 points)

Suppose a random sample of $n = 50$ Oregon Douglas Fir seed cones has a sample mean of $\bar{X} = 34$ seeds per cone. The population variance is $\sigma^2 = 5$ (seeds per cone)^2.

a. (3 points) Compute the z-statistic for testing the null hypothesis $H_0 : \mu = 35$.

```r
#Setting values
n <- 50 #sample size
x_bar <- 34 #sample mean
null_mean <- 35 #null hypothesis mean
var <- 5 #population variance

z <- (x_bar - null_mean)/sqrt(var/n) #Computing z statistic

print(z) #showing value
```

```
## [1] -3.162278
```

b. (3 points) Perform a level $\alpha = 0.1$ test of $H_0 : \mu = 35$ vs. the one-sided lower alternative $H_A : \mu < 35$. Perform this test "by hand" rather than using R's built-in testing functions. Note: Computing values in R is fine, just refrain from using the `t.test()` function.

Since we have the population variance, we look at the normal distribution.

```r
alpha <- .1
z_alpha <- qnorm(alpha)
#qnorm gets us the z statistic for probability to the left of the statistic

z < z_alpha #testing if the z test statistic is lower than the alpha critical value
```

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

Due to the z statistic being less than the z critical value for a one-sided lower alternate test we would reject the null hypothesis that the population mean is 35.

c. (3 points) What is the one-sided lower p-value for the statistic you computed in part (a)?

```r
pnorm(z)
```

```
## [1] 0.0007827011
```

d. (3 points) What is the two-sided p-value for the statistic you computed in part (a)?

```r
2*pnorm(z)
```

```
## [1] 0.001565402
```

e. (3 points) Construct a 95% confidence interval for the population mean number of seeds per cone.

```
z <- qnorm(.975)
SE <- sqrt(var/n)

up_int <- x_bar + z*SE
low_int <- x_bar - z*SE

CI <- c(low_int,up_int)
print(CI)
```

```
## [1] 33.3802 34.6198
```

f. (3 points) Based on your confidence interval from part (e), and without conducting a formal hypothesis test, would a level $\alpha = 0.05$ two-sided hypothesis test reject or fail to reject the null hypothesis that the population mean is 34.5 seeds per cone? Explain your reasoning.

We would fail to reject the null hypothesis as the population mean 34.5 seeds per cone is within the interval of (33.38, 34.62). The reasoning is that the 95% confidence interval corresponds to $\alpha = 0.05$ 2-sided. Anything within the confidence interval is something we can expect with $\alpha = 0.05$, so 95% of the time that we sample from a population we would expect our confidence interval to contain the true population mean.

g. (3 points) Without computing anything, would a 99% confidence interval be narrower or wider than the interval you constructed in part (e)? Explain your reasoning.

It would be wider. We would be more confident that we contain the true population mean because it would be catching a larger selection of values that would not reject the null. Looking directly at the equation, the z statistic would increase causing the confidence interval to be wider.

h. (3 points) Based on your answers to part (e) and (g), and without computing anything, would a level $\alpha = 0.01$ two sided hypothesis test reject or fail to reject the null hypothesis that the population mean is 34.5 seeds per cone? Explain your reasoning.

It would likely fail to reject the null hypothesis because the population mean = 34.5 will still be within the confidence interval given that my previous confidence interval also contained it, and the 99% confidence interval will be even larger. This implies that the value will still be contained within the 99% confidence interval.

i. (3 points) Now suppose that we perform a z-test but incorrectly assume that the population variance is **3** (seeds per cone)^2 instead of the true value of **5** (seeds per cone)^2. If the null hypothesis is true, what proportion of the time will we *incorrectly* reject the null hypothesis if we are using a level $\alpha = 0.1$ critical value? Is this a problem for conducting statistical analyses? Why or why not?

```r
n <- 50 #sample size
x_bar <- 34 #sample mean
null_mean <- 35 #null hypothesis mean
var <- 3 #population variance
true_var <- 5
alpha <- .1

z <- (x_bar - null_mean)/sqrt(var/n) #Computing z statistic

z_alpha_up <- qnorm(1-alpha/2)
z_alpha_low <- qnorm(alpha/2)

z_crit_up1<- z_alpha_up*SE+null_mean
z_crit_low1<- z_alpha_low*SE+null_mean

z_crit_up2<- z_alpha_up*SE+x_bar
z_crit_low2 <- z_alpha_low*SE+x_bar

sum1 <- pnorm(z_crit_up1, mean = null_mean, sd = sqrt(5))-pnorm(z_crit_up2, mean = null_mean, sd = sqrt
sum2 <- pnorm(z_crit_low1, mean = null_mean, sd = sqrt(5))-pnorm(z_crit_low2, mean = null_mean, sd = sq
sum <- sum1 + sum2
print(sum)
```

```
## [1] 0.3366532
```

I believe here what we want to do is find the nominal points which we would reject the null hypothesis for both mean and population mean. We would then test this against the sample mean of 34 and n = 50 and find the nominal points for that normal distribution with assumed sample mean and variance of 3. We would then take the probability difference of the test statistic of the sample mean and population mean with variance 3 respectively.

Question 2 (12 points)

In your own words, give a precise, one-sentence description for each of the following concepts:

a. (2 points) Confidence interval:

An interval of values associated with an alpha value that contains the true population mean (1-alpha)*100% of the time when a population is sampled multiple times.

b. (2 points) p-value:

The probability of a given critical value in a specified probability distribution.

c. (2 points) Null hypothesis:

The assumed true population mean, proportion, or etc in a given population distribution.

d. (2 points) Test statistic:

A statistic that is determined based on sampled data.

e. (2 points) Rejection region:

A set of values such that a test statistic within it would reject the null hypothesis.

f. (2 points) Hypothesis test consistency:

If an alternative hypothesis is true, the null hypothesis is rejected with near 100% probability as we increase our sample size very large.

Question 3 (12 points)

You are interested in testing the hypothesis that the average body temperature for people who live in Corvallis is 98.6°F. You gather a random sample of $n = 50$ Corvallis residents, measure each person's body temperature, and then perform a two-sided level $\alpha = 0.05$ z-test to test the null hypothesis $H_0 : \mu = 98.6°F$ vs. $H_A : \mu \neq 98.6°F$. Assume that you know the population variance of body temperatures in Corvallis is $\sigma^2 = 2°F^2$ (i.e. the variance is 2 and the units are squared degrees Fahrenheit).

For this question, please perform all calculations "by hand" (i.e. don't use a built-in R function. Computing the numeric answers in R is fine).

a. (3 points) If the true population mean body temperature in Corvallis is $99°F$, what is the power of this test?

```
#Setting values
null_mean <- 98.6
true_mean <- 99
n <- 50
var <- 2
alpha = .05

SE <- sqrt(var/n) #standard error
z_low <- qnorm(alpha/2) #standard normal z stat for alpha = .05 2 tailed
z_up <- qnorm(1-alpha/2)

low <- z_low*SE+null_mean #lower critical value
up <- z_up*SE+null_mean #upper crtical value

z_low <- (low-true_mean)/SE #getting z stat with respect of given true mean
z_up <- (up-true_mean)/SE

power = pnorm(z_low)+ (1-pnorm(z_up)) #calculating 2 sided power

print(power)
```

```
## [1] 0.5160053
```

b. (3 points) If instead of $n = 50$ people in your sample, you obtained $n = 100$ people in your sample, what would the power of the test be?

```r
null_mean <- 98.6
true_mean <- 99
n <- 100 #changing sample size
var <- 2
alpha = .05

SE <- sqrt(var/n)
z_low <- qnorm(alpha/2)
z_up <- qnorm(1-alpha/2)

low <- z_low*SE+null_mean
up <- z_up*SE+null_mean

z_low <- (low-true_mean)/SE
z_up <- (up-true_mean)/SE

power = pnorm(z_low)+ (1-pnorm(z_up))

print(power)
```

```
## [1] 0.8074304
```

c. (3 points) If instead of performing a level $\alpha = 0.05$ z-test with your original $n = 50$ people, you performed a level $\alpha = 0.1$ z-test with that same sample size, what would the power of the test be?

```r
#Setting values
null_mean <- 98.6
true_mean <- 99
n <- 50
var <- 2
alpha = .1

SE <- sqrt(var/n) #standard error
z_low <- qnorm(alpha/2) #standard normal z stat for alpha = .05 2 tailed
z_up <- qnorm(1-alpha/2)

low <- z_low*SE+null_mean #lower critical value
up <- z_up*SE+null_mean #upper crtical value

z_low <- (low-true_mean)/SE #getting z stat with respect of given true mean
z_up <- (up-true_mean)/SE

power = pnorm(z_low)+ (1-pnorm(z_up)) #calculating 2 sided power

print(power)
```

```
## [1] 0.6388938
```

d. (3 points) Can you generalize these results? How does power change as you increase sample size (Describing whether it increases or decreases is fine)? How does power change as you increase the significance level?

Power increases as sample size increases, and decreases as sample size decreases. Power also increases as alpha increases, and decreases as alpha decreases.

Question 4 (17 points)

In this problem, you'll explore the performance (i.e. the *exactness*) of the t-test for data from different distributions. The code below can be modified to perform the necessary simulations. This code simulates `nsim` (10,000) data sets, each a sample of size `n` (10) from the Uniform distribution [Unif(0, 1)]. The sample mean and sample standard deviation for each data set is computed and then used to calculate the t-statistic for testing the null hypothesis that the population mean is equal to the true mean.

Use each scenario below to generate 10,000 simulated data sets (Note: Update the line of code `data <- runif(n, 0, 1)` with the appropriate R function):

Scenario	Distribution	Sample Size (`n`)	True Mean (`true_mean`)	R function
1	Uniform(0,1)	10	0.5	`runif(n, 0, 1)`
2	Uniform(0,1)	20	0.5	`runif(n, 0, 1)`
3	Uniform(0,1)	50	0.5	`runif(n, 0, 1)`
4	Chi-squared(1)	10	1.0	`rchisq(n, 1)`
5	Chi-squared(1)	20	1.0	`rchisq(n, 1)`
6	Chi-squared(1)	50	1.0	`rchisq(n, 1)`
7	Beta(0.5, 0.5)	10	0.5	`rbeta(n, 0.5, 0.5)`
8	Beta(0.5, 0.5)	20	0.5	`rbeta(n, 0.5, 0.5)`
9	Beta(0.5, 0.5)	50	0.5	`rbeta(n, 0.5, 0.5)`
10	Exponential(1)	10	1.0	`rexp(n, 1)`
11	Exponential(1)	20	1.0	`rexp(n, 1)`
12	Exponential(1)	50	1.0	`rexp(n, 1)`

a. (1 point for each blank space) For each scenario listed above, record the proportion of the resulting p-values which are less than $\alpha = 0.05$. Aside: The value $\alpha = 0.05$ is often referred to as the *nominal (significance) level* of the test; the actual probability that a p-value is less than 0.05 is the *achieved (significance) level.*

Scenario	Distribution	Sample Size (n)	Proportion of p-values less than $\alpha = 0.05$
1	Uniform(0,1)	10	.055
2	Uniform(0,1)	20	.0512
3	Uniform(0,1)	50	.0485
4	Chi-squared(1)	10	.1354
5	Chi-squared(1)	20	.11
6	Chi-squared(1)	50	.0774
7	Beta(0.5, 0.5)	10	.054
8	Beta(0.5, 0.5)	20	.0496
9	Beta(0.5, 0.5)	50	.0493
10	Exponential(1)	10	.1039
11	Exponential(1)	20	.0772
12	Exponential(1)	50	.0647

b. (5 points) Write a short (3-5) sentence summary of how the t-test performs in these different settings: Is it "close enough" to exact that you would be comfortable using it even when the underlying distribution is as far from Normal as these examples?

It seems that for all distributions tested here they got fairly close to .05 once the sample size increased to 50. However, for certain distributions such as the exponential and chi squared, they did much worse in low sample sizes. I would feel comfortable using it if we had a large enough sample size of around 50, for certain distributions such as Uniform and Beta I would feel more comfortable. It is neat though that the t test can perform well even with far different distributions than the Normal.