PSet #4 - R Section

Kylyn Smith

2024-02-29

Part 1

Q1, Q3, & Q4

See previous pages of this PDF.

Q2

Suppose the prices of stocks in a particular market follow a normal distribution with a mean price of $500 and a standard deviation of $40.

(a) Determine the proportion of stocks that have prices higher than $600.

```
p_under600 <- pnorm(600, mean=500, sd=40)
paste("The proportion of stocks with price higher than $600 is", round(1-
p_under600, 6))
```

```
## [1] "The proportion of stocks with price higher than $600 is 0.00621"
```

(b) Find the stock price that corresponds to the 85th percentile.

```
z85 <- qnorm(0.85)
x85 <- z85*40 + 500
paste("The stock price corresponding to the 85th percentile is $", round(x85,
3))
```

```
## [1] "The stock price corresponding to the 85th percentile is $ 541.457"
```

(c) For a research study, a random sample of 47 stocks is drawn from the market. What is the distribution family, the expected value, and the variance of the standardized sample mean stock price, denoted as Z? State any assumptions you made to obtain your answer.

The expected value of Z-bar is $0 due to the properties of standardization of a large sample (since the sample mean is a random variable which would fall around the true mean randomly, the standardized sample mean yields zero). This uses the Central Limit Theorem and the assumption that 47 is a large sample. The variance of a standardized variable is equal to 1 simply because of how standardization calculation is derived. Z-bar is a random variable drawn from the population which has a normal distribution, and its own mean and variance are 0 and 1 respectively, so Z-bar follows the underlying distribution and is thus a standard normal distribution. We could have also stated that Z-bar is normally distributed because of the Central Limit Theorem and the fact that it is a large sample.

(d) Given your answer in (c), find the probability that the sample mean stock prices lies between $480 and $520.

```
z480 <- (480 - 500) / (40 / sqrt(47))
z520 <- (520 - 500) / (40 / sqrt(47))
z480
```

```
## [1] -3.427827
```

```
z520
```

```
## [1] 3.427827
```

The |Z-score| of being within $20 of the sample mean (representing the range from $480-$520) is $\pm$ 3.43. Using the Z-table provided, the probability of the mean being above $520 is 0.9997, the probability of the mean being below $480 is 0.0003, so the probability of the sample mean stock prices falling within this range is 0.9997-0.0003 = **0.9994**.

Part 2

Q5

First, we will replicate the left panel of the bar chart (only the left three bars). We will modify their chart by adding confidence intervals to each average so we can visually judge the amount of sampling variability. Begin by restricting the sample to jobs in the healthcare sector which do not require a certificate, using the variables health and nodeg_req. You should be left with a subsample of 948 resumes.

```
setwd("C:/Users/Smith Family/Documents/Images for Kylyns
Schoolwork/Yale/econ117/econ117_pset4_folder")

dyagk <- read_dta("DYAGK_audit_data_processed.dta")

dyagk_noreq <- subset(dyagk, (dyagk$health == 1) & (dyagk$nodeg_req == 1))
```

(a) Calculate average callback rates for resumes with no certificate, a for-profit certificate, and a public certificate. (Note: These three categories are mutually exclusive and exhaustive in this subsample.)

```
pubs <- subset(dyagk_noreq, dyagk_noreq$pub==1)
fps <- subset(dyagk_noreq, dyagk_noreq$fp==1)
nocerts <- subset(dyagk_noreq, (dyagk_noreq$pub==0) & (dyagk_noreq$fp==0))
paste("The average callback rate for resumes with a public certificate is",
round(mean(pubs$any_call), 4))
```

```
## [1] "The average callback rate for resumes with a public certificate is
0.0886"
```

```
paste("The average callback rate for resumes with a for-profit certificate
is", round(mean(fps$any_call), 4))
```

```
## [1] "The average callback rate for resumes with a for-profit certificate
is 0.0423"
```

```
paste("The average callback rate for resumes with no certificate is",
round(mean(nocerts$any_call), 4))
```

```
## [1] "The average callback rate for resumes with no certificate is 0.0591"
```

(b) Write down the formula for calculating 95% confidence interval for the population mean.

```
qnorm(0.025)
```

```
## [1] -1.959964
```

$$P(Z < -z_{\alpha/2}) = P(Z < -1.9560) = 0.025$$

$$C = [X_N - z_{\alpha/2}(s/\sqrt{N}), X_N + z_{\alpha/2}(s/\sqrt{N})] = [X_N - 1.9560(s/\sqrt{N}), X_N + 1.9560(s/\sqrt{N})]$$

(c) Without using the t.test() command, write code to calculate a 95% confidence interval for the average callback rate for those without a certificate. (Hint: Find the sample mean, sample standard deviation, and number of observations. Then apply the formula you wrote down in part (b).)

```
mean_nocert <- mean(nocerts$any_call)
sd_nocert <- sd(nocerts$any_call)
N_nocert <- nrow(nocerts)

low_nocert <- mean_nocert - 1.9560*(sd_nocert/sqrt(N_nocert))
high_nocert <- mean_nocert + 1.9560*(sd_nocert/sqrt(N_nocert))

paste("The 95% confidence interval for the average callback rate for resumes
with no certificate is", round(low_nocert, 5), "to", round(high_nocert, 5))
```

```
## [1] "The 95% confidence interval for the average callback rate for resumes
with no certificate is 0.04623 to 0.0719"
```

(d) Calculate 95% confidence intervals for the average callback rates of the remaining two groups, for-profit and public certificate holders. Here you may use the t.test() command if you wish.

```
mean_fp <- mean(fps$any_call)
sd_fp <- sd(fps$any_call)
N_fp <- nrow(fps)

low_fp <- mean_fp - 1.9560*(sd_fp/sqrt(N_fp))
high_fp <- mean_fp + 1.9560*(sd_fp/sqrt(N_fp))

paste("The 95% confidence interval for the average callback rate for resumes
with a for-profit certificate is", round(low_fp, 5), "to", round(high_fp, 5))
```

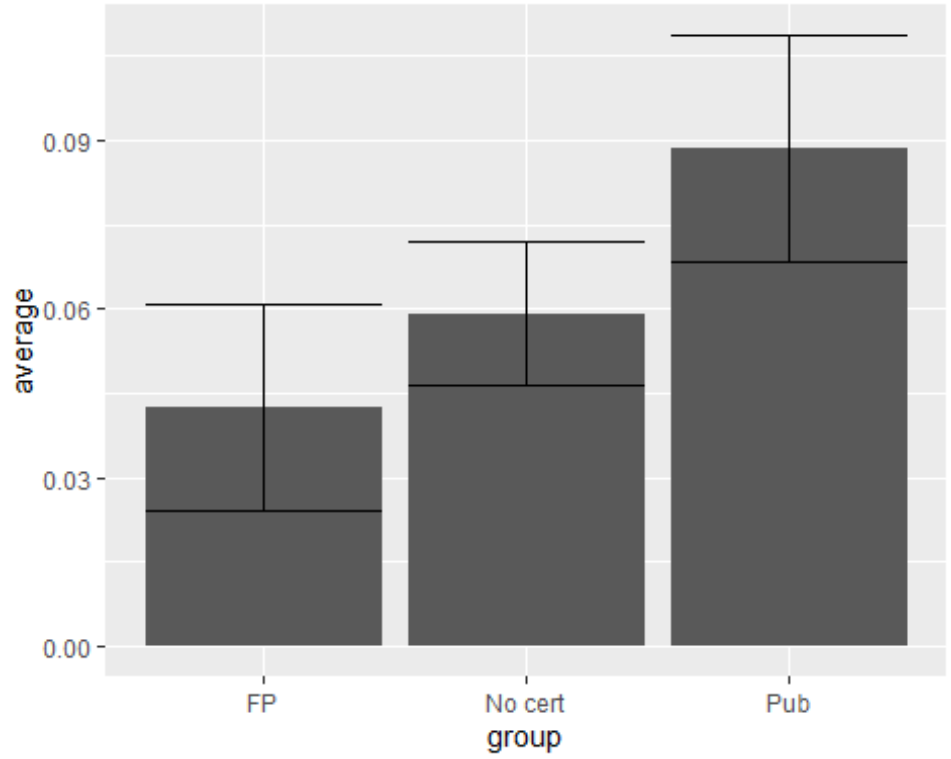
```
## [1] "The 95% confidence interval for the average callback rate for resumes
with a for-profit certificate is 0.02392 to 0.06063"
```

```r
mean_pub <- mean(pubs$any_call)
sd_pub <- sd(pubs$any_call)
N_pub <- nrow(pubs)

low_pub <- mean_pub - 1.9560*(sd_pub/sqrt(N_pub))
high_pub <- mean_pub + 1.9560*(sd_pub/sqrt(N_pub))

paste("The 95% confidence interval for the average callback rate for resumes
with a public certificate is", round(low_pub, 5), "to", round(high_pub, 5))
```

```
## [1] "The 95% confidence interval for the average callback rate for resumes
with a public certificate is 0.06852 to 0.10869"
```

(e) Plot the average callback rates in a bar chart like Figure 2, but add error bars to visualize the 95% confidence intervals you just calculated.

```r
means_and_CIs <- data.frame(group = c("No cert", "FP", "Pub"),
average = c(mean_nocert, mean_fp, mean_pub),
lowerCI = c(low_nocert, low_fp, low_pub),
upperCI = c(high_nocert, high_fp, high_pub))

ggplot(means_and_CIs) +
geom_bar(aes(x = group, y = average), stat = "identity") +
geom_errorbar(aes(x = group, ymin = lowerCI, ymax = upperCI))
```



Q6

An economist from Harvard sees your bar chart and is surprised that for-profit certificate holders seem to perform worse than those without a certificate. He suggests a statistical test of this claim. Since applicants without a certificate get callbacks 5.9% of the time, he reasons, we can test if for-profit graduates have a lower callback rate using the null hypothesis μfp = 0.059 and one-sided alternative μfp < 0.059.

(a) Carry out a test of this hypothesis with α = 0.05, without using the t.test() command. What do you find?

```
z_fp <- (mean_fp - 0.059) / (sd_fp / sqrt(N_fp))
pnorm(z_fp)

## [1] 0.0373888
```

The probability that the sample mean of a group (of the same size as this sample) of for-profit certificate holders receiving callbacks is less than 0.059, if the actual mean is 0.059, is 0.03739.

With α = 0.05, the p-value of this hypothesis test (0.03739) falls below the significance level, so I reject the null that μ_{fp}=0.059.

(b) Propose a modified one-sided hypothesis test which recognizes our uncertainty about the population callback rate for the no-certificate group. (Hint: Your hypotheses should involve two population means, not just one.) Carry out this modified test, again with α = 0.05 and without using the t.test() command. Do your results differ? Why?

Null Hypothesis: μ_{fp} - μ_{nocert} = 0

Alternative Hypothesis: μ_{fp} - μ_{nocert} < 0

```
z_new <- ((mean_fp - mean_nocert) - 0) /
sqrt((sd_fp^2/N_fp)+(sd_nocert^2/N_nocert))
print(z_new)

## [1] -1.465763

pnorm(z_new)

## [1] 0.0713564
```

With an α of 0.05, I fail to reject the null hypothesis of this two-sample, one-sided hypothesis test, since the p-value (probability of getting a difference in sample means between for-profit and no-certification resumes less than zero in a sample of this size given that the difference in population means = 0) is 0.0713564 > 0.05. Clearly, the results of these two tests differ. This is due to the fact that we are now accounting for the uncertainty in the no-certification mean, which makes our test/comparison of these individual samples less significant on reflecting on the population as a whole.

(c) Comment on the validity of each test. Which gives a better answer to the question of whether the for-profit callback rate is less than the no-certificate callback rate?

The first test (one-sided one-sample) makes sense when trying to compare simply to an objective standard or threshold, WITHOUT recognizing the idea that this threshold we chose to represent definitely the mean of the no-certification population could in fact not be the exact mean of the no-certification population, since it was derived from the sample. The second test (one-sided two-sample) takes into account this uncertainty and instead tries to compare the results of both sample statistics with each other. The latter test gives a more valid answer (which is that the for-profit/no-certificate callback rate disparity is indeterminate) because we cannot take the 0.059 mean for the no-certificate sample to represent exactly the mean of the respective population.

Q7

Now consider the right panel of the bar chart, which plots callback rates for jobs requiring a certificate. The authors summarize these results by saying, "we find no significant difference in callback rates by type of postsecondary institution for health jobs (such as practical nursing and pharmacy technician) that require. . . a certificate."

(a) Formally frame a null and alternative hypothesis designed to test their claim.

Null Hypothesis: $\mu_{fp_{cert}} - \mu_{pub_{cert}} = 0$

Alternative Hypothesis: $\mu_{fp_{cert}} - \mu_{pub_{cert}} \neq 0$

(b) Is this a one-sided or two-sided alternative hypothesis?

This is a two-sided alternative hypothesis since we are only looking for difference, not in a particular direction.

(c) Now subset to this alternative subsample of healthcare jobs requiring a certificate. You should find a sample of 1,396 such resumes. Calculate the mean callback rates for for-profit and public certificates. Before running any tests, would you say there is a "significant difference" between these sample callback rates?

```
needcert <- dyagk_req <- subset(dyagk, (dyagk$health == 1) & (dyagk$nodeg_req
== 0))
pubscert <- subset(dyagk_req, dyagk_req$pub==1)
fpscert <- subset(dyagk_req, dyagk_req$fp==1)
paste("The average callback rate for resumes with a public certificate is",
round(mean(pubscert$any_call), 4))
```

```
## [1] "The average callback rate for resumes with a public certificate is
0.0574"
```

```
paste("The average callback rate for resumes with a for-profit certificate
is", round(mean(fpscert$any_call), 4))
```

```
## [1] "The average callback rate for resumes with a for-profit certificate
is 0.0488"
```

There does not seem to be a significant difference between the average callback rates (only about 1%).

(d) Implement the hypothesis test you proposed in part (a), with α = 0.05 (using the t.test() command). What do you conclude?

```
mean_pubcert <- mean(pubscert$any_call)
sd_pubcert <- sd(pubscert$any_call)
N_pubcert <- nrow(pubscert)

mean_fpcert <- mean(fpscert$any_call)
sd_fpcert <- sd(fpscert$any_call)
N_fpcert <- nrow(fpscert)

z_new_cert <- ((mean_pubcert - mean_fpcert) - 0) /
sqrt((sd_fpcert^2/N_fpcert)+(sd_pubcert^2/N_pubcert))

2*(1-pnorm(z_new_cert))

## [1] 0.1851505

t.test(x=pubscert$any_call, y=fpscert$any_call,
       alternative = "two.sided",
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)

##
##  Welch Two Sample t-test
##
## data:  pubscert$any_call and fpscert$any_call
## t = 1.3251, df = 1394, p-value = 0.1854
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.004129804  0.021321781
## sample estimates:
##  mean of x  mean of y
## 0.05737062 0.04877463
```

According to either method (t.test() or calculating directly), the p-value result of about 0.185 exceeds the significance threshold of 0.05, so I fail to reject the null hypothesis that the average callback rates between the two groups are different.

(e) Based on the R output, what is the p-value? In at most two sentences, provide an interpretation of this number.

The p-value is about 0.185. This means that, if the actual difference in population means between callback rates for for-profit and public resumes were to be zero, there would be approximately an 18.5% chance of achieving a sample mean difference this far or farther from the population mean difference (0) with these sample sizes. This is not rare enough for us to conclude that the population means are surely different.

(f) Overall, healthcare employers favor public-college degrees over for-profit degrees, but the gap is much smaller (maybe even zero) for healthcare jobs with certificate requirements. What might explain this? Note that, unlike jobs without a certificate requirement, these degree-requiring healthcare jobs usually require a standardized occupational license too. (Hint: You might consider the role of college choice as a signal of ability, but there is more than one good answer to this question.)

One possible explanation for the lower reliance on college type as a determination factor of ability for job postings which require a certification as opposed to the higher reliance on college type for non-certificate postings could be the standardization/reliability of the certification test itself in proving ability. Healthcare workers who expressly need to have a competency in a certain arena can explicitly prove they have such a competency by completing a third-party certification. Jobs which do not require a certification can look instead to other indicators of aptitude, such as college admittance. By requiring a certificate, employers can objectively know that the applicants who have a certificate are qualified for the job, regardless of where they were educated, since the certificate is a base standard.