S520 Problem Set 9 Solutions

Arturo Valdivia

Due on 3/28/2022

Q1

1.i.

c. $H_0 : \mu \leq 0.02$ vs $H_1 : \mu > 0.02$

d. We construct the vector `sample1` with 40 diseased chicken out of $n = 1000$ and include it as a variable into data frame `df1`:

```
sample1 = c(rep("diseased",40), rep("normal",1000-40))
df1 = data.frame(sample1)
```

Now we can run the appropriate code. While not needed, I use a random seed for replication purposes:

```
set.seed(100)
null_sim <- df1 |>
  specify(response = sample1, success = "diseased") |>
  hypothesize(null = "point", p = .02) |>
  generate(reps = 10000, type = "draw") |>
  calculate(stat = "prop")

null_sim |>
  get_p_value(obs_stat = 0.04 , direction = "right")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of `reps` chosen in the `generate()` step. See
## `?get_p_value()` for more information.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

The p-value is equal to zero, because out of 10000 bootstrap sample (under H_0) none of them produced a proportion that was as extreme as or more extreme than 0.04. We reject the null hypothesis and conclude that more than 0.02 of the chickens are diseased (and they should be killed).

This result may seem surprising, but it just happened that 0.04 was too large of a proportion to appear by chance. Let's see the 5-number summary of the proportion that do appear by chance on a simulated data:

```
summary(null_sim)
```

```
##    replicate         stat
##  1       :   1   Min.   :0.00600
##  2       :   1   1st Qu.:0.01700
##  3       :   1   Median :0.02000
##  4       :   1   Mean   :0.02004
##  5       :   1   3rd Qu.:0.02300
##  6       :   1   Max.   :0.03900
##  (Other):9994
```

The largest (most extreme) proportion was 0.039, just below the observed proportion in the sample.

1.ii.

We have:

```
mu = 0.02
n = 100
xbar = 4/100
sigma = sqrt(mu*(1-mu))
z.v = (xbar - mu)/(sigma/sqrt(n))
1 - pnorm(z.v)
```

```
## [1] 0.07656373
```

Using $\alpha = 0.025$, we fail to reject the null hypothesis.

1.iii.

```
sample2 = c(rep("diseased",4), rep("normal",100-40))
df2 = data.frame(sample2)
set.seed(100)
null_sim2 <- df2 |>
  specify(response = sample2, success = "diseased") |>
  hypothesize(null = "point", p = .02) |>
  generate(reps = 10000, type = "draw") |>
  calculate(stat = "prop")

null_sim2 |>
  get_p_value(obs_stat = 0.04 , direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.138
```

The p-value now is 0.1384 (simulation results may vary slightly) and we fail to reject the null hypothesis. Observe that the p-value is much larger than in the theory-based approach, likely because the sample was perhaps not large enough for the distribution of the proportion to be approximately normal.

Q2.

We want to find evidence that AD perform better in the morning than in the afternoon when describing the picture. The experimental unit is an AD patient and two measurements are taken per patient. We can use those measurements to define a single value obtained per patient; i.e., let X_i be defined as the number of information units for Picture A minus the number of information units for Picture B for the ith patient, where $i = 1, \ldots, 60$.. In this context, $\bar{X}_{60}$ is the sample mean of 60 patients, and μ is the mean or expected value for X_i and also for $\bar{X}_{60}$. The hypotheses are:

$H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$

since the scientist wonders if asking in the morning is equivalent to asking in the afternoon. This is a two-sided (or two-tailed) test and we need to find the area on both tails, using a significance level $\alpha = 0.025$ as directed in the problem set instructions.

We now calculate the t test statistic:

$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{-0.1833-0}{5.18633/\sqrt{60}} = -0.274$

Here are the results in R:

```
t.v = (-0.1833 - 0)/(5.18633/sqrt(60))
p_value <- 2 * pt(t.v, 60-1)
p_value
```

```
## [1] 0.7852214
```

Since the p-value is greater than the significance level of 0.025, we fail to reject the null hypothesis. The data do not provide enough evidence to that there is a difference in the quality of discourse between describing Picture A in the morning and describing Picture B in the afternoon.

In this problem, we do not have a sample of data from which we can obtain bootstrap samples. The sample mean and standard deviation are not enough to be able to generate these samples. Therefore, the simulation-based approach cannot be used.

Question 3.

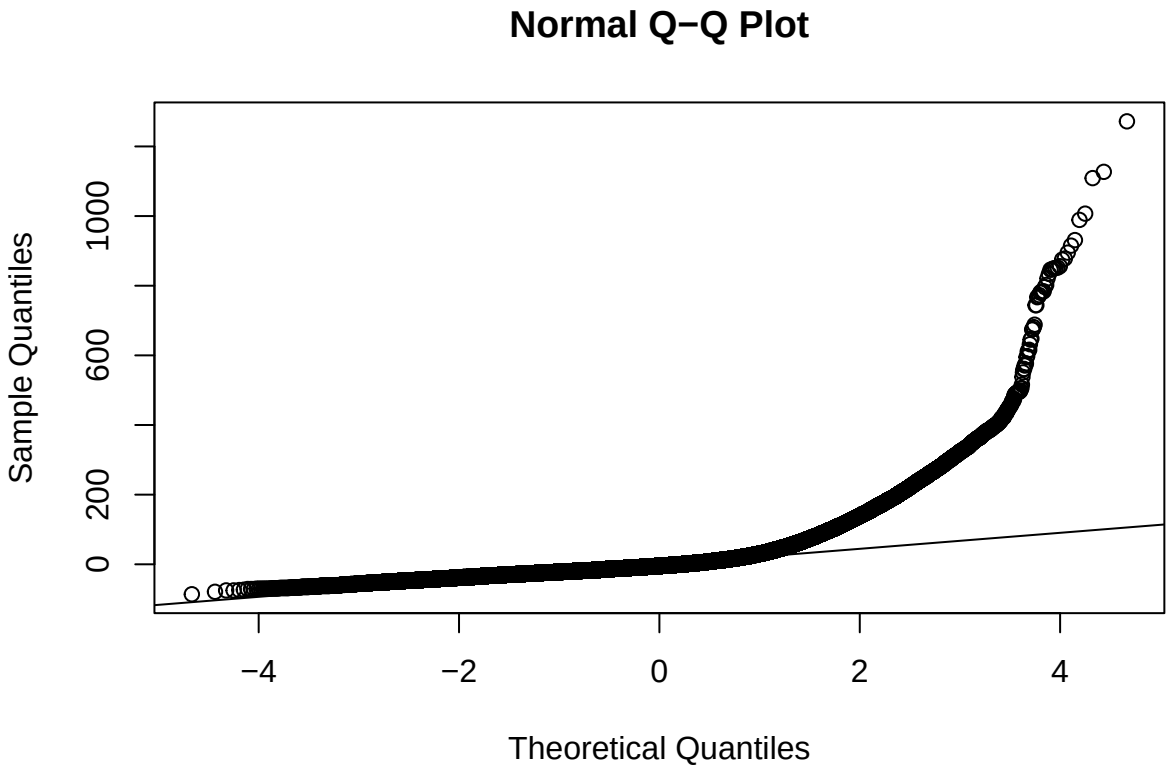
3a

Here is the code:

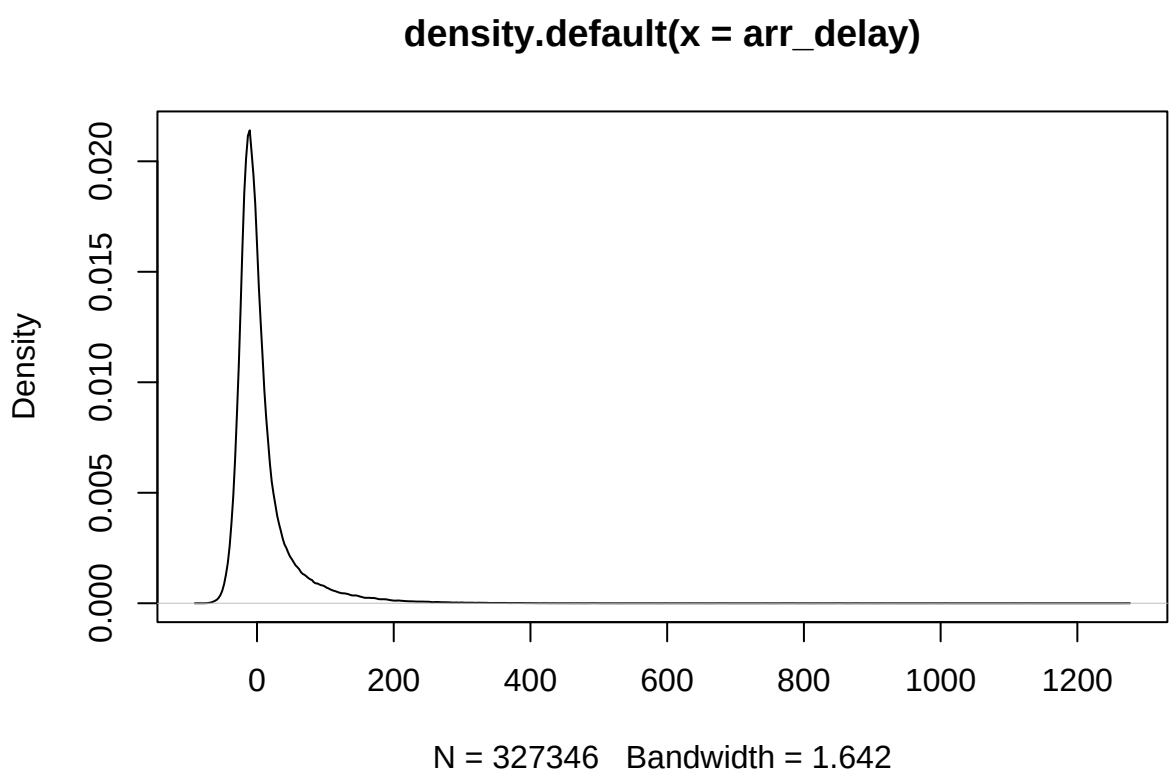
```
library(nycflights13)
arr_delay=flights$arr_delay
arr_delay=na.omit(arr_delay)
```

3b

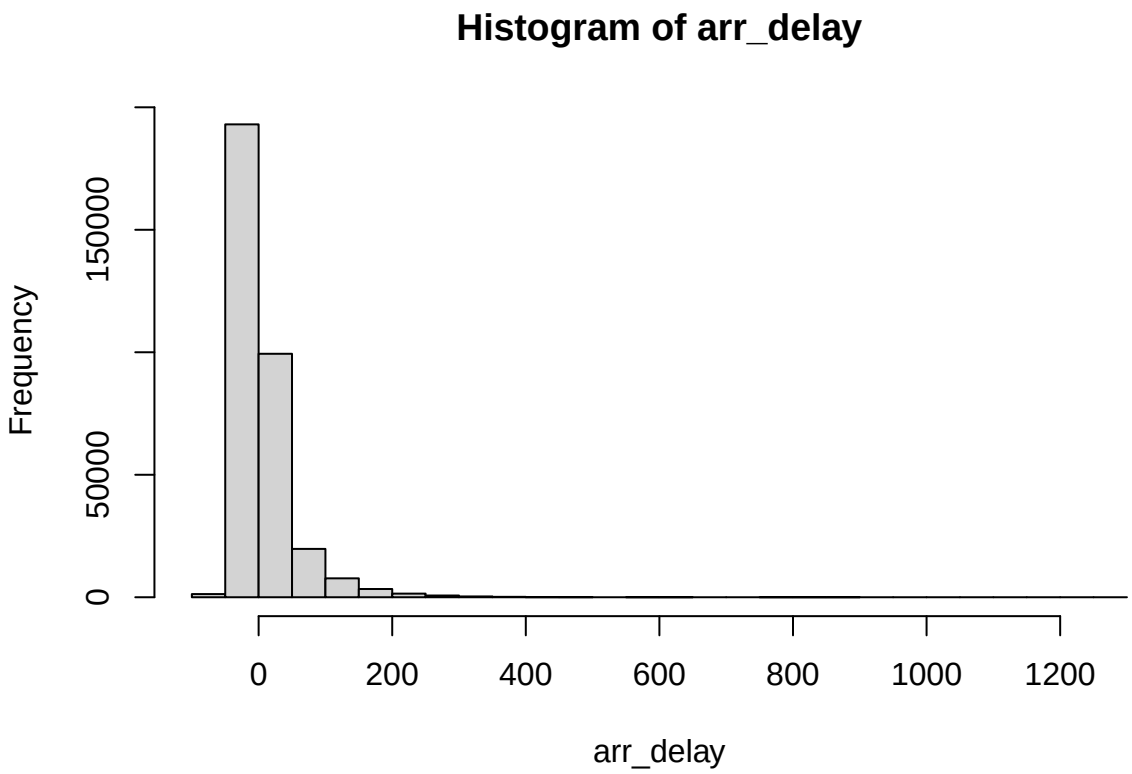
```
qqnorm(arr_delay)
qqline(arr_delay)
```




```
plot(density(arr_delay))
```




```
hist(arr_delay)
```



Looking at the plots, the data is clearly right-skewed and does not seem to be drawn from a normal distribution.

3c

```
set.seed(520)
r_sample=sample(arr_delay,100,replace=T)
```

The hypotheses are

$H_0 : \mu \geq 10$ vs $H1 : \mu < 10$$

Here is the test in R:

```
mu0 = 10
xbar=mean(r_sample)
n = 100
s=sd(r_sample)
t.v = (xbar - mu0)/(s/sqrt(n))
pt(t.v, n - 1)
```

```
## [1] 0.007522487
```

Using $\alpha = 0.025$, we reject the null hypothesis. We have enough evidence to conclude that the average arrival delay is less than 10 minutes.

3d

We want to check if the proportion of flights without arrival delays is greater than 50%. So, we are only concerned about whether or not arrival delays happened. This can be modeled as a Bernoulli trial for each flight where success is a flight without arrival delay, so $X_i \sim Bernoulli(p)$ and $E\bar{X}_{100} = EX_i = p = \mu$. The hypotheses can be written as:

$H_0 : p \leq 0.5$ vs $H_1 : p > 0.5$

We can solve this problem, as customary, in R:

```
p0 = 0.5
n = length(r_sample)
phat=mean(r_sample<=0)
sigma = sqrt(p0*(1-p0))
z = (phat - p0)/(sigma/sqrt(n))
1-pnorm(z)
```

```
## [1] 0.03593032
```

Using $\alpha = 0.025$, we fail to reject the null hypothesis. We do not have enough evidence to conclude that more than the flights have no arrival delay.

4

4a

Here is the code in R:

```
df4 <- data.frame(r_sample)

# Infer code to find the sample mean
x_bar <- df4 %>%
  specify(response = r_sample)  %>%
  calculate(stat = "mean")
x_bar
```

```
## Response: r_sample (numeric)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  2.58
```

```
# Generate bootstrap samples under the null distribution

null_sim <- df4 %>%
  specify(response = r_sample) %>%
  hypothesize(null = "point", mu = 10) %>%
  generate(reps = 30000, type = "bootstrap") %>%
  calculate(stat = "mean")
```

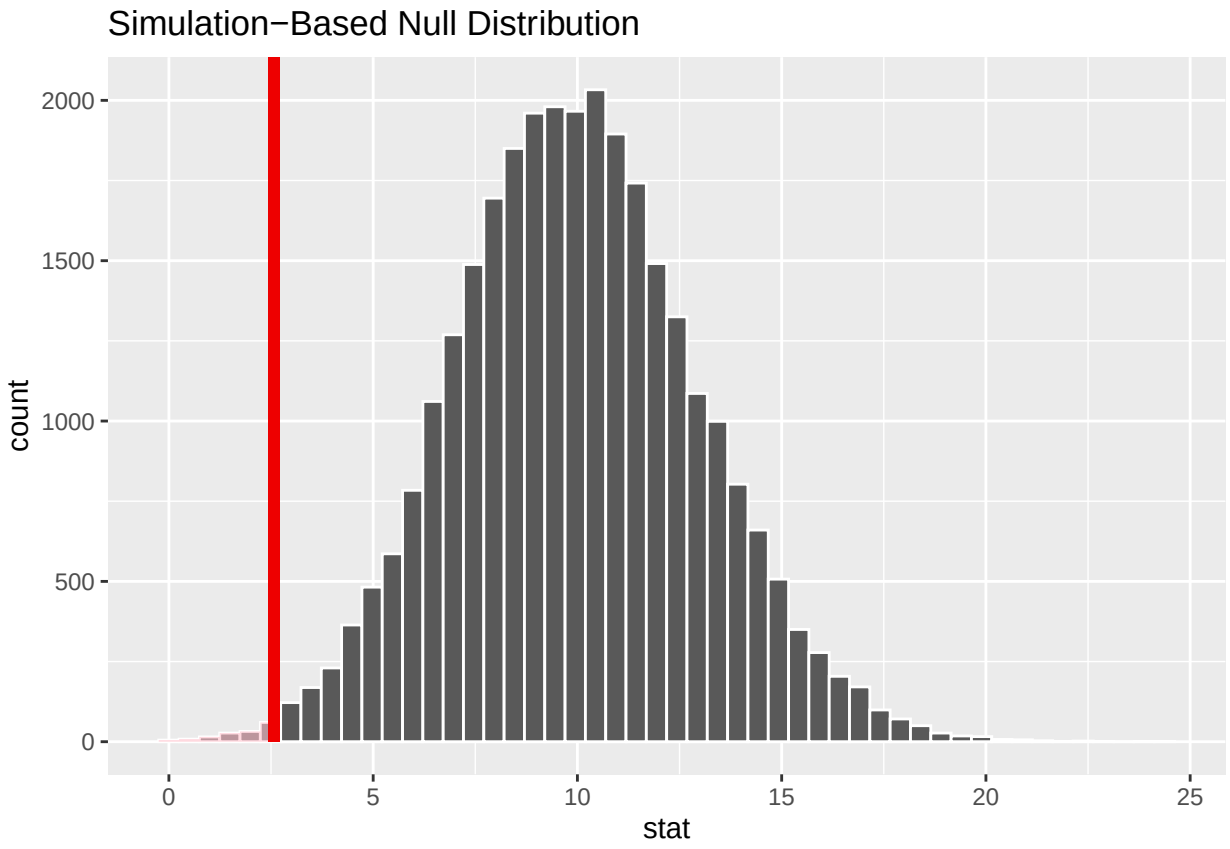
```
# Obtaining p-value based on the bootstrap samples

null_sim %>%
  get_p_value(obs_stat = x_bar, direction = "left")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1 0.00423
```

The p-value is small enough. As in 3c and using $\alpha = 0.025$, we reject the null hypothesis. We have enough evidence to conclude that the average arrival delay is less than 10 minutes. We can also construct the plot to visualize this test:

```
visualize(null_sim, method = "simulation", bins = 50) +
  shade_p_value(obs_stat = x_bar, direction = "left")
```



4b

We startHere is the code in R:

```
on_time <- as.factor(r_sample<=0)
df4b <- data.frame(on_time)
```

```r
null_sim_p <- df4b |>
  specify(response = on_time, success = "TRUE") |>
  hypothesize(null = "point", p = .5) |>
  generate(reps = 10000, type = "draw") |>
  calculate(stat = "prop")

#' Infer code to find the sample proportion

phat <- df4b |>
  observe(response = on_time, success = "TRUE", stat = "prop")

#' Obtaining p-value based on the simulated samples

null_sim_p |>
  get_p_value(obs_stat = phat , direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0465
```

As our p value is larger than 0.025, therefore we fail to reject the null hypothesis, as it was done in 3d; we don't have evidence that mroe than 50% of flights have no arrival delay.