Hyothesis Test

Zhan Li

```
library(tidyverse)
```

Question 1

(a)

```
url_1 <- "http://ritsokiguess.site/STAC32/pop.csv"
data_1 <- read_csv(url_1)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   v = col_double()
## )
```

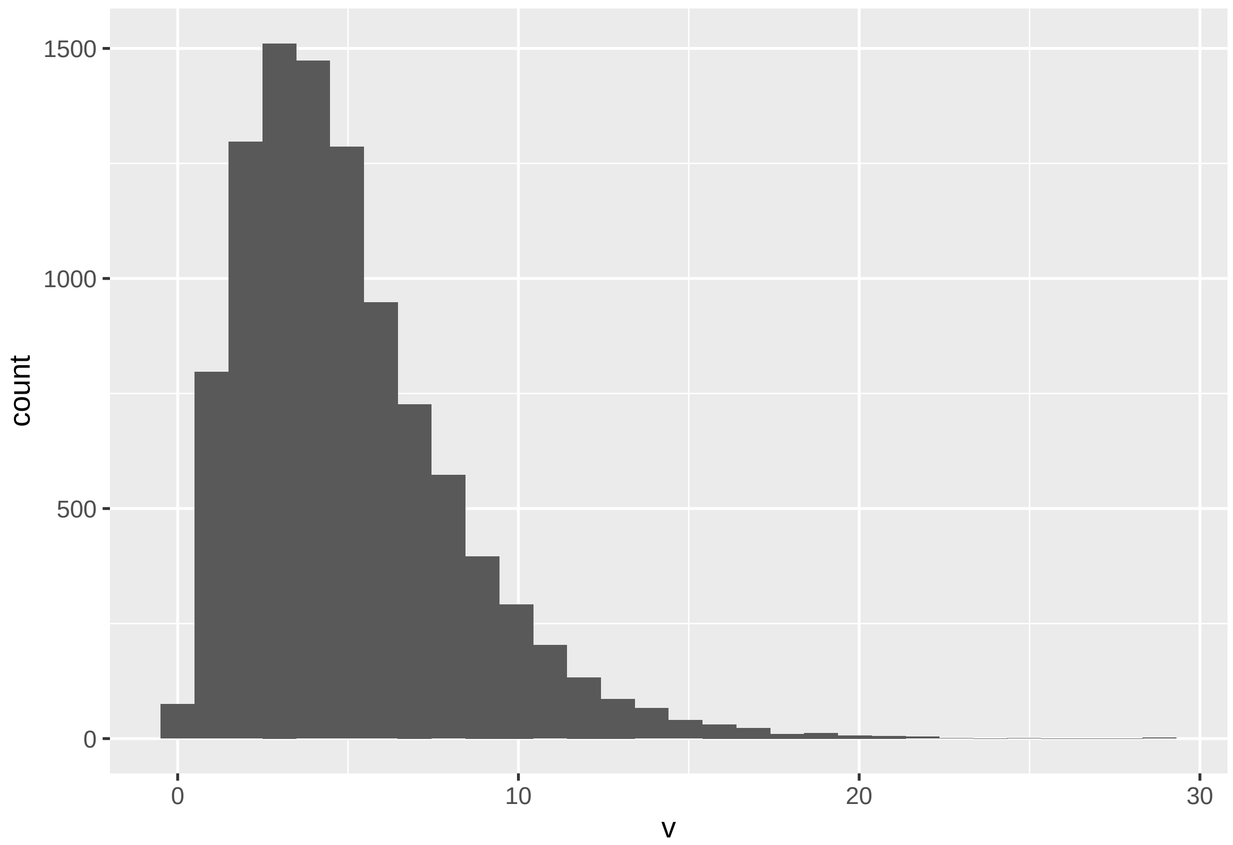
```
data_1
```

```
## # A tibble: 10,000 x 1
##         v
##     <dbl>
##  1  9.97
##  2  2.18
##  3  6.20
##  4  2.11
##  5  6.30
##  6  1.54
##  7  5.77
##  8  2.94
##  9 16.8
## 10  1.95
## # ... with 9,990 more rows
```

Since our link specifies it's a csv file, therefore we use the function "read_csv()".

(b)

```
ggplot(data_1, aes(x = v)) + geom_histogram(bins=30)
```



Since we are most likely interested in the shape of the distribution and whether it's normal or not, I chose to draw a histogram with v being the clear choice of the x variable. Since the data size is rather large, I decided to use a larger than usual bin size=30. I noticed that the distribution is clearly right-skewed, because the mode of the histogram is concentrated around 3-4.

(c)

```
set.seed(6859)
rerun(1000, sample(data_1$v, size=10, replace=FALSE )) %>%
  map( ~ t.test(., mu=4, alternative="greater")) %>%
  map_dbl("p.value") %>%
  enframe(value="pvals")%>%
  count(pvals <= 0.05)
```

```
## # A tibble: 2 x 2
##   `pvals <= 0.05`     n
## * <lgl>           <int>
## 1 FALSE             843
## 2 TRUE              157
```

Since we are just taking samples of size 10, replace should be FALSE. Alternative should be greater since our alternative hypothesis is $\mu > 4$. Notice from the simulation, the power to reject the null hypothesis is $\frac{157}{1000} = 0.157 > 0.05$. Therefore, it's about 15.7% that we reject $\mu = 4$.

(d)

```
set.seed(6859)
rerun(1000, sample(data_1$v, size=50, replace=FALSE )) %>%
  map( ~ t.test(., mu=4, alternative="greater")) %>%
  map_dbl("p.value") %>%
  enframe(value="pvals")%>%
  count(pvals <= 0.05)
```

```
## # A tibble: 2 x 2
##   `pvals <= 0.05`     n
## * <lgl>           <int>
## 1 FALSE             276
## 2 TRUE              724
```

Note with 50 sample size, the power to reject the null hypothesis becomes $\frac{724}{1000} = 0.724$, which is larger than the previous 0.157. This makes sense because as our sample size gets larger, we expect $p \leq 0.05$ to increase.

(e)

```
set.seed(6859)
rerun(1000, sample(data_1$v, size=50, replace=FALSE )) %>%
  map( ~ t.test(., mu=5, alternative="greater")) %>%
  map_dbl("p.value") %>%
  enframe(value="pvals")%>%
  count(pvals <= 0.05)
```

```
## # A tibble: 2 x 2
##   `pvals <= 0.05`     n
## * <lgl>           <int>
## 1 FALSE             978
## 2 TRUE               22
```

```
mean(data_1$v)
```

```
## [1] 5
```

When $\mu = 5$, the power to reject becomes $\frac{22}{1000} = 0.022$. However, after we calculated the true mean which is 5, we can see the decision is to reject true null hypothesis, so this becomes a type 1 error. So the 0.022 isn't really power, it's less than what we expected because it's a type 1 error.

Question 2

(a)

```
url_2 <- "http://ritsokiguess.site/STAC33/protein.txt"
data_2 <- read_table(url_2)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   protein = col_double()
## )
```

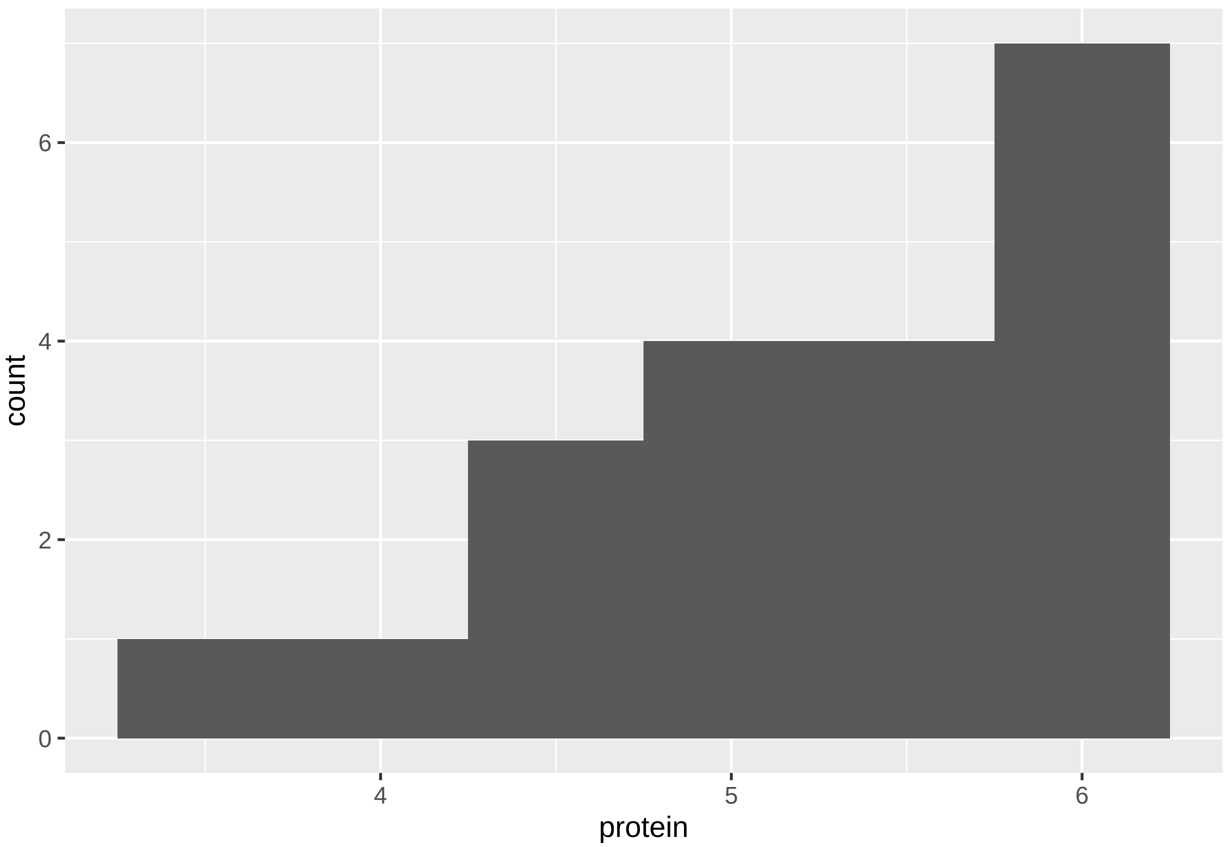
```
data_2
```

```
## # A tibble: 20 x 1
##    protein
##      <dbl>
##  1     5.1
##  2     4.2
##  3     6.1
##  4     5.1
##  5     5.7
##  6     5.5
##  7     4.9
##  8     6.1
##  9     6.1
## 10     5.8
## 11     5.2
## 12     4.3
## 13     4.7
## 14     3.6
## 15     4.4
## 16     5.5
## 17     5.6
## 18     5.8
## 19     6.1
## 20     6.1
```

Notice there are 20 rows of data, which matches our sample size of 20.

(b)

```
ggplot(data_2, aes(x = protein)) + geom_histogram(bins = 6)
```



Once again, we used histogram because we'd like to know whether the data follows normal distribution. Protein is our clear choice of the x variable. I chose bins = 6 because the sample size is 20 which is relatively small.

(c)

As we can see from the histogram, the distribution is very left skewed, not normal, plus the sample size is relatively small, so we can't really apply CLT and use the t-test to assess the mean. Median would be a better form of assessment for this scenario, therefore we use the sign test.

(d)

```
library(smmr)
sign_test(data_2, protein, 6)
```

```
## $above_below
## below above
##    15     5
##
## $p_values
##   alternative    p_value
## 1       lower 0.02069473
## 2       upper 0.99409103
## 3   two-sided 0.04138947
```

As we can see from the data.frame, I looked at the two-sided p-value since we are looking protein that's equal to 6 ounces. So the two sided p value is approximately 0.04, which is less than 0.05. So we should reject the null hypothesis in favor of our alternative hypothesis. The null hypothesis is each package contains 6 ounces of protein, the alternative hypothesis is each package does not contain 6 ounces of protein. Notice the p-value when the median is lower than 6 is 0.02, which is much less than the p-value(0.99) when the median is greater than 6, so between lower than 6 and greater than 6, I favor towards the hypothesis that the protein ounce is less than 6.

(e)

Sign test uses median instead of mean From the above R console, we can see the median below 6 vs above 6 is 15:5. We know if the median is 6, then below vs above should be 10:10. However, 15:5 implies the true median should be below 6, which means the p value is smaller than 0.05 so that we can reject the null hypothesis.

(f)

```r
ci_median(data_2, protein, conf.level = 0.90)
```

```
## [1] 4.905273 5.793750
```

As we can see from the 90 percent confidence interval, 6 is not within the interval, so we should reject the null hypothesis.