Post 2: Z-testing

Sierra Park

11/25/2017

```
#load library
library(ggplot2)
```

Introduction

Dataset: Student Admissions at UC Berkeley

I will explore hypothesis testing using the UC Berkeley Student Admissions built-in dataset in R. This is an aggregate data on applicants to UC Berekely graduate school for the six largest departments in 1973 classified by admission and sex. I will perform a z-test to see if the admission rates for males and females are the same across the six departments.

What is Hypothesis Testing?

The main purpose of hypothesis testing is to test if the data from an experiment or observation supports a hypothesis.

This process starts with forming a null hypothesis that assumes that the chances of all events are equal and that all differences are merely due to chance. On the other hand, an alternative hypothesis states that the differences are real–there is an underlying phenomenom behind the differences we see in the results.

After forming the hypotheses, we state the alpha level in which you will use as a threshold to determine whether we should accept or fail to accept the null hypothesis. We generally choose alpha to be 5%.

The metric that we use to determine whether the null hypothesis is true is the p-value, so-called the observed significance level. P-value determines how “likely” or “unlikely” we observe a more extreme test statistic in the direction of the alternative hypothesis than the one observed. If the p-value is small (let’s say, less than or equal to alpha), then it is “unlikely”; if the p-value is large (le’ts say, more than alpha), then it is “likely.” In other words, p-value is the chance, under the null hypothesis, that the test statistic is equal to the value that was observed in the data or is even further in the direction of the alternative. If a p-value is small, it means that the tail beyond it is small, so the observed statistic is away from what the null predicts, implying that the dat asupport the alternative hypothesis better than the null hypothesis.

If the p-value is less than 5%, then the result is statistically significant; if the p-value is even smaller–less than 1%–then the result is highly statistically significant, and we reject the null in both cases since the null hypothesis can no longer be true.

For this post, we will use two-sided hypotheesis, just to denote that there is a difference between the admission rate of males and females.

Forming Hypothesis

In our case, null hypothesis is:

$$H = \text{There is no difference in the admission rate for males and females}$$

and alternative hypothesis is:

$$H_{\alpha} = \text{There is a difference in the admission rate for males and females; the difference is due to chance.}$$

Observe the Data

Let’s take a look at the UC Berkeley Admissions data.

```
UCBAdmissions
```

```
## , , Dept = A
##
##           Gender
## Admit      Male Female
##   Admitted  512     89
##   Rejected  313     19
##
## , , Dept = B
##
##           Gender
## Admit      Male Female
##   Admitted  353     17
##   Rejected  207      8
##
## , , Dept = C
##
##           Gender
## Admit      Male Female
##   Admitted  120    202
##   Rejected  205    391
##
## , , Dept = D
##
##           Gender
## Admit      Male Female
##   Admitted  138    131
##   Rejected  279    244
##
## , , Dept = E
##
##           Gender
## Admit      Male Female
##   Admitted   53     94
##   Rejected  138    299
##
## , , Dept = F
##
##           Gender
## Admit      Male Female
##   Admitted   22     24
##   Rejected  351    317
```

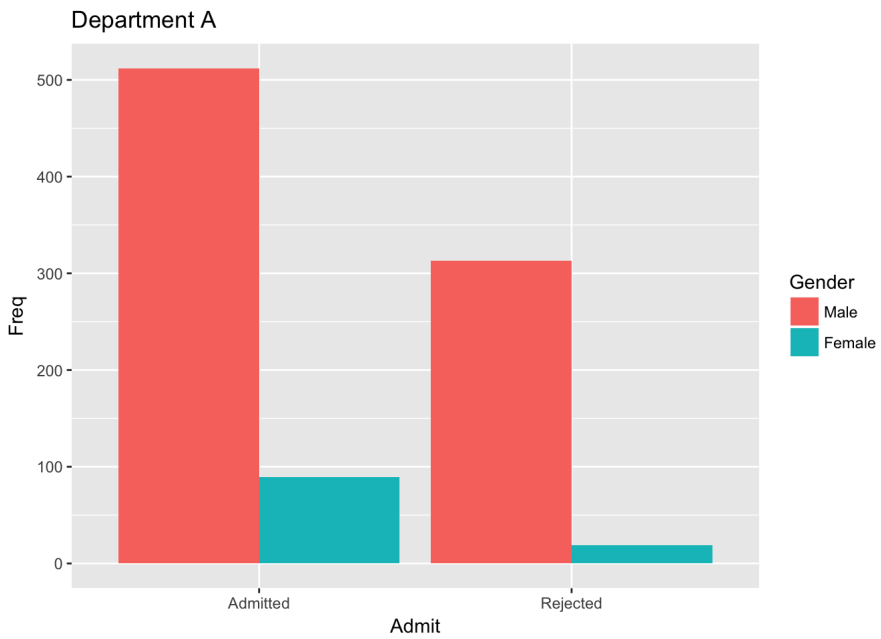
We see that there are 6 departments and each of them have the number of males and females admitted and rejected students. Hence, we will calculate the percentages of admission rate for males and females in eaach department using a for loop.

However, before doing so, we will graph the statistics to get a sense of what they look like.

Analysis of the Data

For Department A, there were more admitted students than rejected students, and more admitted and rejected males than females. This seems plausible since there were a lot more male students who have applied than female students.

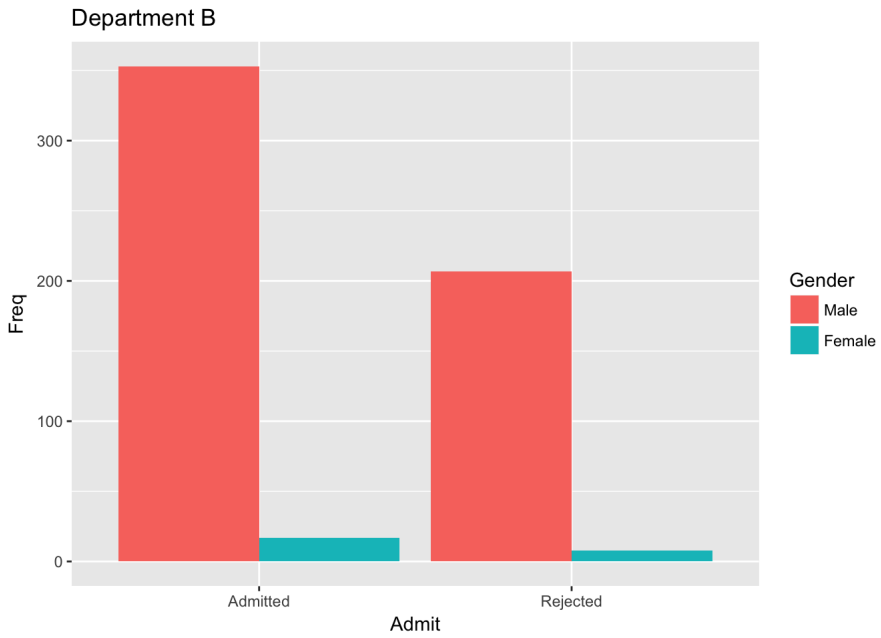
```
ggplot(data.frame(UCBAdmissions[,,LETTERS[1]]), aes(Admit , Freq, fill = Gender)) + geom_bar(stat="identity", posi
tion = "dodge") + ggtitle("Department A")
```



It seems as if the same pattern as Department A occurs to Department B as well. For Department B, there were also more admitted students than rejected students, and more admitted and rejected males than females. However, the total number of applicants in Department B is less

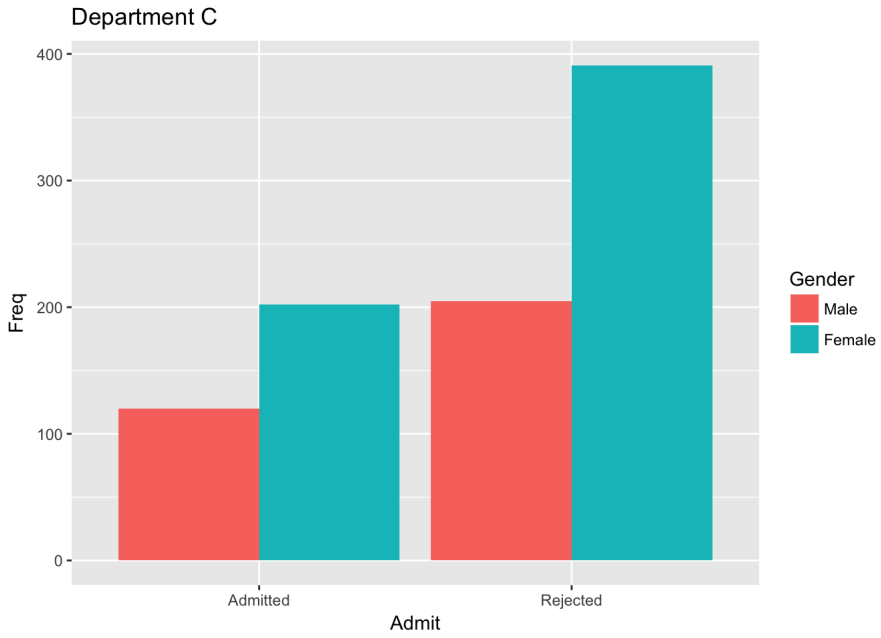
than Department A.

```
ggplot(data.frame(UCBAdmissions[,,LETTERS[2]]), aes(Admit, Freq, fill = Gender)) + geom_bar(stat="identity", posit
ion = "dodge") + ggtitle("Department B")
```



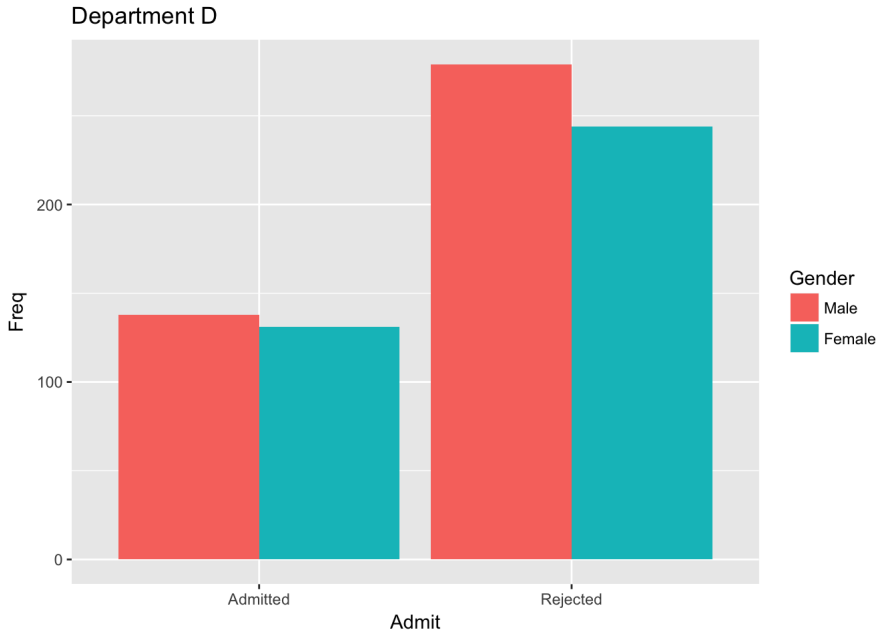
The trend changes, however, for Department C. For Department C, there were more rejected students than admitted students, and more admitted and rejected females than males.

```
ggplot(data.frame(UCBAdmissions[,,LETTERS[3]]), aes(Admit, Freq, fill = Gender)) + geom_bar(stat="identity", posit
ion = "dodge") + ggtitle("Department C")
```



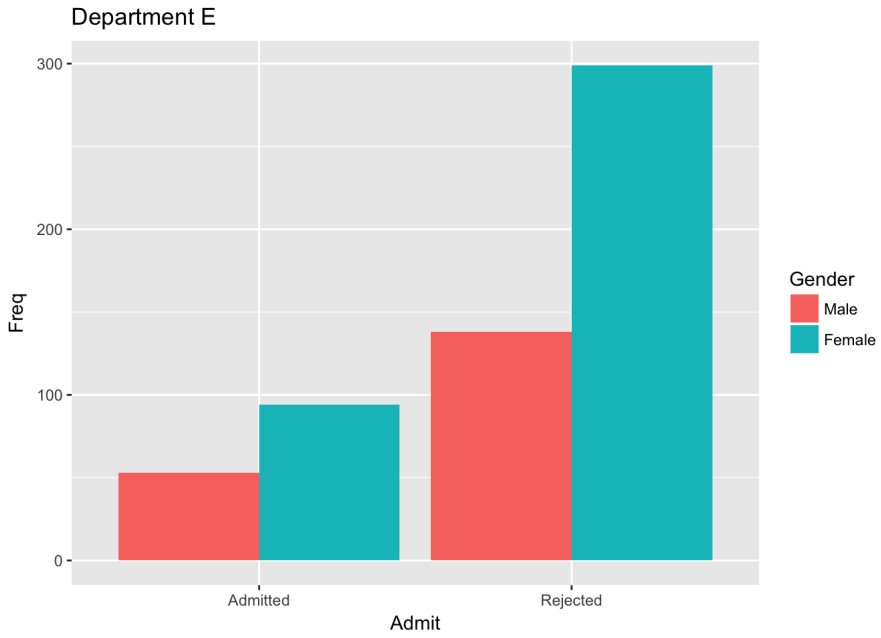
For Department D, there were more rejected students than admitted students, and more admitted and rejected males than females.

```
ggplot(data.frame(UCBAdmissions[,,LETTERS[4]]), aes(Admit, Freq, fill = Gender)) + geom_bar(stat="identity", posit
ion = "dodge") + ggtitle("Department D")
```



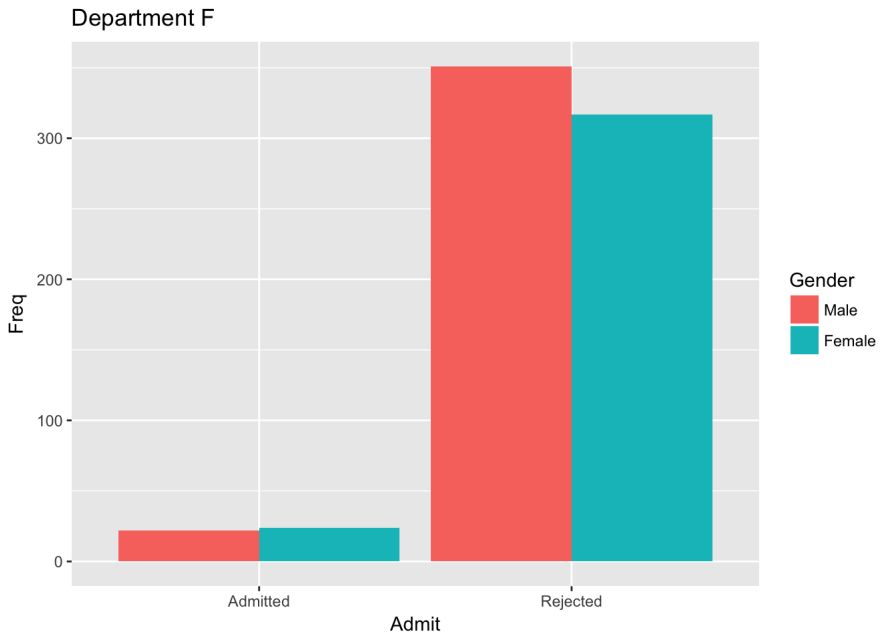
For Department E, there were more rejected students than admitted students in general, and more admitted and rejected females than males. There is the biggest gap between

```
ggplot(data.frame(UCBAdmissions[,,LETTERS[5]]), aes(Admit, Freq, fill = Gender)) + geom_bar(stat="identity", posit
ion = "dodge") + ggtitle("Department E")
```



For Department F, there were more rejected students than admitted students. There are more admitted females than admitted males, but more rejected males than rejected females. The gap between admitted vs. rejected in both genders is the biggest in this department.

```
ggplot(data.frame(UCBAdmissions[,,LETTERS[6]]), aes(Admit, Freq, fill = Gender)) + geom_bar(stat="identity", posit
ion = "dodge") + ggtitle("Department F")
```



We have examined the data for each department, and we can conclude that there is a recurring pattern among some departments. Specifically, we can divide them into two categories: one where there were more admitted than rejected students (Department A and B), and the other with more rejected than admitted students (Department C, D, E, and F). We note that Department A had the highest number of applicants.

Performing the Hypothesis Test

Now, let's perform the z-test for percent admitted males and females. I have created a for loop to calculate the admitted percentages for males and females of each department.

```
percentadmitmale <-  rep(0, 6) #Create an empty vector for percent admitted males
percentadmitfemale <-  rep(0,6) #Create an empty vector percent admitted females
numbermale <- rep(0,6)
numberfemale <- rep(0,6)
for (i in 1:6){ #Calculate the admitted percentage for males and females of each department
  dep <- UCBAdmissions[,,LETTERS[i]]
  numbermale[i] <- sum(dep[1,1], dep[2,1])
  numberfemale[i] <- sum(dep[2,2], dep[1,2])
  percentadmitmale[i] <- dep[1,1]/(dep[1,1] + dep[2,1])
  percentadmitfemale[i] <- dep[2,1]/(dep[2,2] + dep[1,2])
}
```

We have two vectors, *percentadmitmale* and *percentadmitfemale*, that show the percent admission for males and females, respectively.

```
percentadmitfemale
```

```
## [1] 2.8981481 8.2800000 0.3456998 0.7440000 0.3511450 1.0293255
```

```
percentadmitmale
```

```
## [1] 0.62060606 0.63035714 0.36923077 0.33093525 0.27748691 0.05898123
```

Examining the two vectors, it seems as if all Departments besides B had some difference between the percent admitted females and males.

NOTE: Hypotheses are stated above.

Let our observed statistic be the difference between the admitted percentages of females and males.

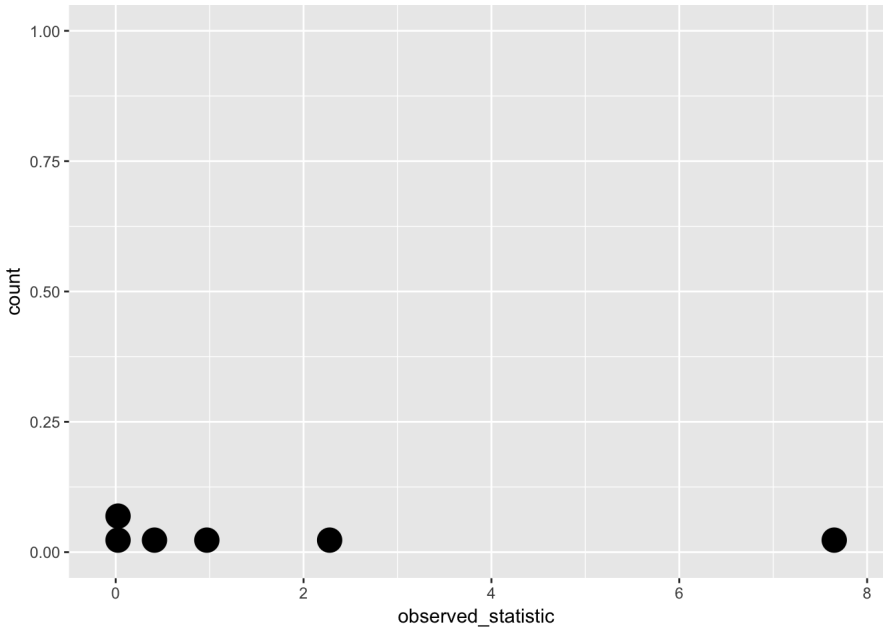
```
observed_statistic = (percentadmitfemale - percentadmitmale)
observed_statistic
```

```
## [1]  2.27754209  7.64964286 -0.02353094  0.41306475  0.07365813  0.97034428
```

In order to see the distribution/spread of the observed statistic, I have used a dot plot to represent the data.

```
ggplot(data.frame(observed_statistic)) + geom_dotplot(aes(observed_statistic))
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



All of the statistic is above 0, which means that the percent admitted males is greater than that of females.

Now We will perform z-test for difference in proportions.

There are a few things that we need to calculate:

- Pooled sample proportion: Since the null hypothesis states that admission rate of females is the same as the admission rate of males, we use a pooled sample proportion (p) to compute the standard error of the sampling distribution.

```
p <- (percentadmitfemale *numberfemale + percentadmitmale * numbermale) / (numberfemale + numbermale)
p
```

```
## [1] 0.8842444 0.9572650 0.3540305 0.5265152 0.3270548 0.5224090
```

- Standard error: Compute the standard error (SE) of the sampling distribution difference between two proportions:

```
SE <- sqrt(p*(1-p)*((1/numbermale) + (1/numberfemale)))
SE
```

```
## [1] 0.03273853 0.04134491 0.03300492 0.03553345 0.04138029 0.03742406
```

- Test statistic: The test statistic is a z-score defined by:

$$z = (\text{percentadmitfemale} - \text{percentadmitmale})/SE$$

Hence, in our case, we define z to be:

```
z <- (percentadmitfemale - percentadmitmale)/SE
z
```

```
## [1]  69.5676340 185.0201772  -0.7129525  11.6246724   1.7800291  25.9283545
```

Now that we have a z-score that we would like, we can use that to compute the p-value. This is done in the next section.

Translate Z-score to P-value

Since having a z-score less than -1.96 or greater than 1.96 means more than two standard deviation away, the corresponding p-value would be less than 0.05 on either side of the bell curve. Hence, looking at our z vector, Department A, C, and D have absolute value of z-value greater than 1.96.

Therefore, we conclude that the admission rate between the males and females were insignificant in all departments except Department A, C, and D.

Conclusion

In this post, I ued a built-in dataset of UC Berkeley Admissions to analyze the admission rate between males and femals across 6 different departments. Hypothesizing that the admission rate is equal for both genders, I conducted a z-test, and found out that the p-value for three departments were significant: A, C, and D. This means that in those three departments, there was a noticeable difference in the admission rates between the two genders. This conclusion is supported by observing the gender-grouped barplots and noting the difference of the proportions between the genders.

Sources:

1. http://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/
2. https://onlinecourses.science.psu.edu/statprogram/node/138

3. https://www.inferentialthinking.com/chapters/10/2/terminology-of-testing.html
4. http://stattrek.com/hypothesis-test/difference-in-proportions.aspx
5. http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm
6. http://ggplot2.tidyverse.org/reference/labs.html
7. http://ggplot2.tidyverse.org/reference/geom_bar.html
8. https://stackoverflow.com/questions/17721126/simplest-way-to-do-grouped-barplot