Chapter 1: R and Permutation Test

August 31, 2016

Introduction

Research suggests that people perceived as unattractive may experience some types of discrimination. We're going to look at the data collected from a study in 1989 that investigated the relationship between perceived physical attractiveness and sentence length that male jurors gave to female defendants. Here an excerpt of the data.

	Attr	Crime	Years	Serious	independent	sincere
1	Beautiful	Burglary	10	8	9	8
2	Beautiful	Burglary	3	8	9	3
3	Beautiful	Burglary	5	5	6	3
4	Beautiful	Burglary	1	3	9	8
...	...	...	...	...	...	...
111	Average	Swindle	7	4	9	1
112	Average	Swindle	6	3	5	2
113	Average	Swindle	12	9	9	1
114	Average	Swindle	8	8	1	5

What is the type of `Attr` variable?

- ◯ **Quantitative**
- ◯ **Categorical**

Submit | Show Hint | Show Answer | Clear

What is the type of `Years` variable?

- ◯ **Quantitative**
- ◯ **Categorical**

Submit | Show Hint | Show Answer | Clear

Data Analysis

1. Load neccessary packages

Today we will use following packages. If R creates error on any package, it means you haven't installed it, then install it and re-run that line (by pressing `Ctrl+Enter` or `Command+Enter`)

```
require(heplots)  # which has MockJury data
require(beanplot)  # which has beanplot()
require(mosaic)  # which has favstat(), diffmean(), shuffle()
```

2. Data Summary

To answer the research question, we need to focus on the two variables, `Attr` and `Years` . Let's start by examining a summary of these variables. You can obtain this summary by using the `favstats` function or by using `summary` function.

```
> Note: `favstat` only works on quantitative variables, while `summary` can
work on both quantitative and categorical.
```

```
summary(MockJury$Attr)
```

```
##    Beautiful      Average Unattractive
##           39           38           37
```

```
favstats(MockJury$Year)
```

```
##  min Q1 median Q3 max     mean       sd   n missing
##    1  2      3  7  15 4.692982 3.633977 114       0
```

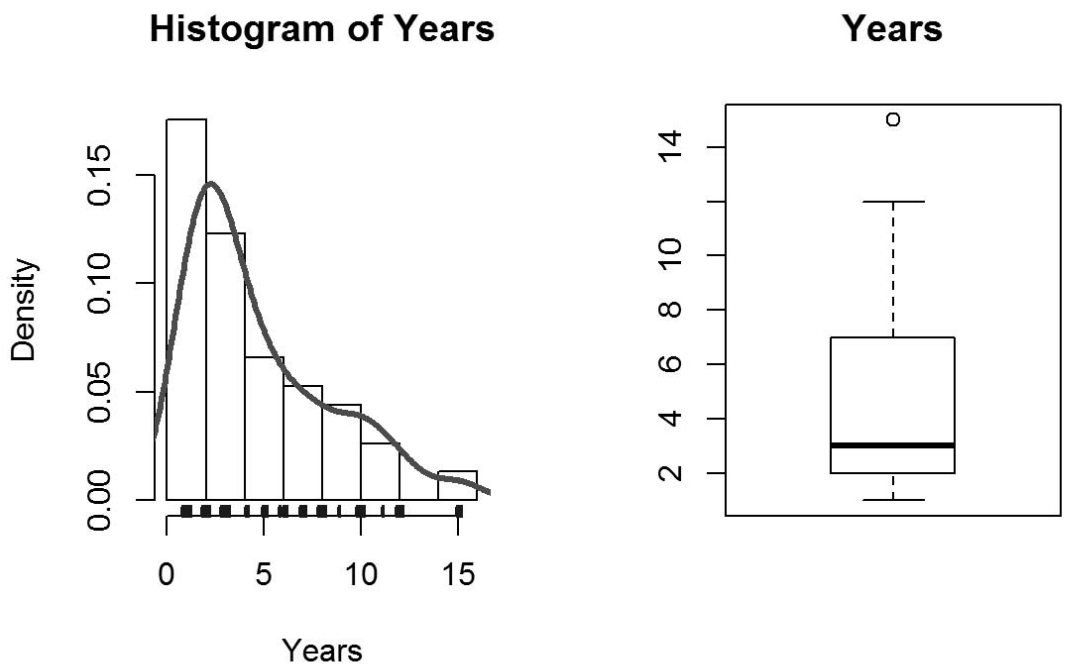
What information about the data can you glean from this summary?

3. Plots

We can also make some plots, particularly of our response variable (years, in our case). This helps us get an idea of the shape and distribution of our response. On the left is a histogram where taller bars indicate higher frequencies of sentence lengths and on the right is a boxplot. Notice that a density curve is overlaying the histogram and a rug plot is below it. The density curve is scaled so that the total area under the curve is 1. A rug plot is a plot in which a tick mark is made for each observation.

```
# Histogram with Density curve, and Rug
hist(MockJury$Years, freq = F, xlab = "Years", main = "Histogram of Years")
lines(density(MockJury$Years), lwd = 3, col = "red")
rug(jitter(MockJury$Years), col = "Blue", lwd = 2)

# Boxplot
boxplot(MockJury$Years, main = "Years")
```

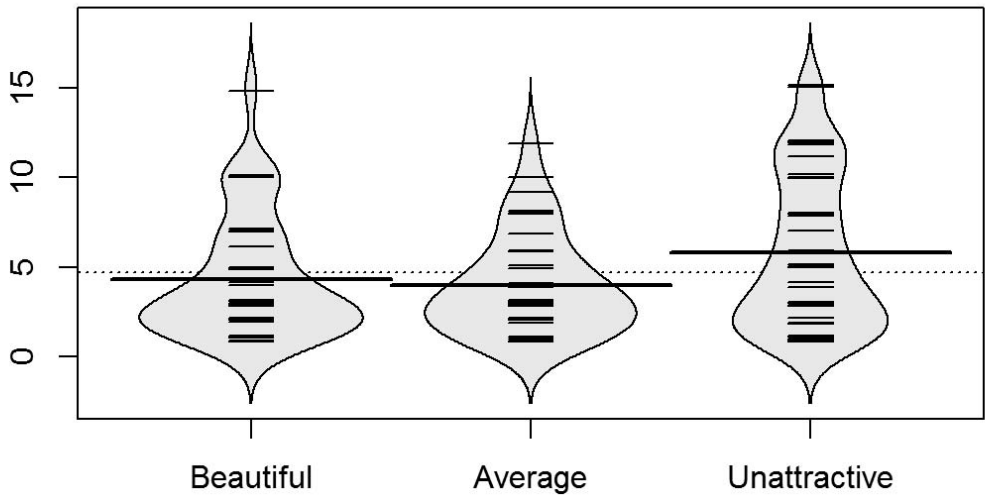


What can you say about the distribution of the response variable from these plots?

We originally wanted to examine the relationship between sentence length and attractiveness. So it's better to have a side-by-side plot, where each group of interest is shown using the same y-axis scaling. We can make a plot called a **beanplot** to look at the distribution of responses across levels of attractiveness.

> Note: `Years ~ Attr` is how you tell R you want to see how the response (`Years`) differs based on groups (`Attr`). We don't use `$` sign here because we explicitly told R where these variables come from in `data = MockJury`.

```
beanplot(Years ~ Attr, data = MockJury, log = "", col = "bisque", method =
"jitter")
```



Beanplot has the curve in density plot mirrored. The jitter is included. The bold line is the mean, **not the median** as in box plot.

What does the width of the plot relate to? Which group has the largest range of sentence lengths? Based on these plots, is it plausible to suspect that unattractive women are given longer sentences?

4. Compare Two Groups

We will start with comparing the *Average* and *Unattractive* groups. The following code will remove *Beautiful* group from the `MockJury` data set and assign it to a separate data set called `MockJury2` (we still want to keep full data in MockJury).

```
MockJury2 <- MockJury[MockJury$Attr != "Beautiful", ]  # Remove Beautiful
summary(MockJury2$Attr)  # Check
```

```
##    Beautiful      Average Unattractive
##            0           38           37
```

```
MockJury2$Attr <- factor(MockJury2$Attr)  # Beautiful is still there, not go
od! Re-adjusts the factor levels
summary(MockJury2$Attr)  # Check again
```

```
##      Average Unattractive
##           38           37
```

Let's look at the mean sentence lengths for each level of attractiveness.

```
mean(Years ~ Attr, data = MockJury2)
```

```
##      Average Unattractive
##     3.973684     5.810811
```

If we use the `diffmean` function, it will compute the difference in the group means for us.

```
diffmean(Years ~ Attr, data = MockJury2)
```

```
## diffmean
## 1.837127
```

We see that the mean sentence length for unattractive women is almost two years longer than for women of average attractiveness, while the sentence length for the beautiful women lies in the middle. There seems to be a big difference between the average and the unattractive groups, but we need statistics to decided whether this difference could be due to chance.

5. Hypotheses

Think back to Stat 216. What is the parameter of interest in words and notation when we want to test for the difference between average (v) and unattractive (u) women?

- ○ $\mu_u - \mu_v$ **. The difference in true mean sentence lengths given to female defendants between two groups.**
- ○ $\bar{x}_u - \bar{x}_v$ **. The difference in the mean of sentence lengths given to female defendants between two groups.**
- ○ $p_u - p_v$ **. The difference in true proportion of sentence lengths given to female defendants between two groups.**
- ○ $\hat{p}_u - \hat{p}_v$ **. The difference in true proportion of sentence lengths given to female defendants between two groups.**

Submit Clear

What are the null hypothesis and alternative hypothesis?

- ○ $H_0 : \mu_u = \mu_v$ **;** $H_A : \mu_u < \mu_v$
- ○ $H_0 : \mu_u = \mu_v$; $H_A : \mu_u \neq \mu_v$
- ○ $H_0 : \mu_u = \mu_v$ **;** $H_A : \mu_u > \mu_v$
- ○ **none of these are correct**

Submit Clear

These hypotheses establish the null and alternative model:

- Null model: $y_{ij} = \mu + \varepsilon_{ij}$
- Alternative model: $y_{ij} = \mu_j + \varepsilon_{ij}$

where y_{ij} is the i^{th} observation from the j^{th} group, ε_{ij} is error, and assumed to have normal distribution with mean 0 and some variance.

For example, the following table is the sentence lengths given to female defendants separated into two columns corresponding to Average and Unattractive group.

	Years	**Attr**
1	1	Unattractive
2	4	Unattractive
3	3	Unattractive
4	2	Unattractive
5	8	Unattractive
...	...	...
37	12	Unattractive
1	5	Average
2	5	Average
3	4	Average
4	3	Average
4	6	Average
...	...	...
38	8	Average

Ex. Suppose we are considering the alternative model for the 4th observation (i = 4) from the second group ($j = 2$)

1. What is y_{42} ?
2. What is the null model?
3. What is the alternative model?

Submit Clear

6. Permutation Test

In Stat 216 we used computer applets to make null distributions when we wanted to test our null hypothesis using randomization methods. In this class we're going to use R. But the idea is the same. Let's review the idea.

How can we shuffle the dots in two groups?

- ☐ **When the alternative hypothesis is assumed to be true, two groups are the same.**
- ☐ **The null hypothesis is that there is no difference between groups, so the observations can move around.**
- ☐ **Groups have no effect on the response, so changing group shouldn't affect the result if the null is true.**
- ☐ **none of these are correct**

Submit Clear

Why do we start by assuming the null is true?

- ◯ **Because we know it should be true, and we're trying to prove it.**
- ◯ **To build a null distribution and see how unusual the evidence is on that distribution.**
- ◯ **We don't have to assume anything.**

Submit Clear

```
perm1 <- with(MockJury2, data.frame(Years, Attr, Perm.Attr = shuffle(Attr)))
head(perm1)
```

```
##   Years         Attr    Perm.Attr
## 1     1 Unattractive Unattractive
## 2     4 Unattractive      Average
## 3     3 Unattractive Unattractive
## 4     2 Unattractive      Average
## 5     8 Unattractive      Average
## 6     8 Unattractive Unattractive
```

Defendants 4, 2, and 8 are three examples of being switched groups. Because the parameter of interest is the difference in means, let's find the new difference in means in this shuffle.

```
mean(Years ~ Perm.Attr, data = perm1)
```

```
##      Average Unattractive
##     4.736842     5.027027
```

```
diffmean(Years ~ Perm.Attr, data = perm1)
```

```
##  diffmean
## 0.2901849
```

One permutation is good for getting an idea of what's going on, but if we want to create a distribution, we need lots and lots of permutations (like we did in 216). To do that, we're going to build a **for loop**.

```
# Save the observed difference in means in Tobs
Tobs <- diffmean(Years ~ Attr, data = MockJury2)
# Number of permutations you want
B <- 1000
# Create 1000 empty slots in Tstar to save the permutation statistics in
Tstar <- matrix(NA, nrow = B)
for (b in 1:B) {
    Tstar[b] <- diffmean(Years ~ shuffle(Attr), data = MockJury2)
}
```

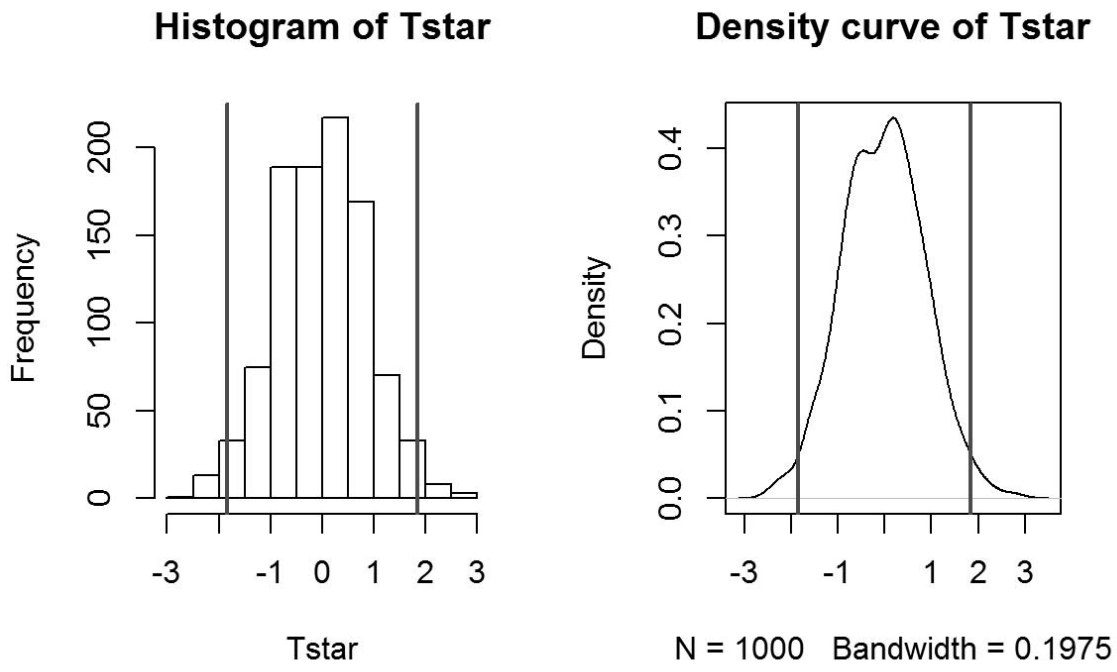
Now, let's explain how the for loop works line by line. Note how I explain the inner function first, outer function later, the right hand side of the assignment first, the left hand side of the assignment later.

1. For each b that takes the value of 1, then 2, then 3, ...1000.
2. Shuffle the name of the groups in Attr.
3. Find the difference in mean of Years between groups in the shuffled Attr. Years and Attr come from data set called MockJury2.
4. Put that difference in means in the slot b of Tstar.

Let's plot the distribution of those 1000 permutation statistics using histogram and density plot, and compare our observed difference in means with that distribution.

```
# Histogram
hist(Tstar)
abline(v = Tobs, lwd = 2, col = "red")  # adding a line for the observed sta
tistic
abline(v = -Tobs, lwd = 2, col = "red")  # adding a line for the negative ob
served statistic

# Density plot
plot(density(Tstar), main = "Density curve of Tstar")
abline(v = Tobs, lwd = 2, col = "red")
abline(v = -Tobs, lwd = 2, col = "red")
```



What can you see from the above plots? Where is the distribution centered? Why does it make sense?

The last piece of information we need for our hypothesis test is a p-value, which I informally think as **"the measure of surprise"**. Formally, it's the **probability of getting a value that is as extreme or more extreme than the observed result, assuming that the null hypothesis is true**. We can get this from R using the following code.

```
pdata(Tstar, Tobs, lower.tail = F)  # from Tobs to the right
```

```
## diffmean
##    0.017
```

```
pdata(Tstar, -Tobs, lower.tail = T)  # from negative Tobs to the left
```

```
## diffmean
##    0.018
```

```
pdata(abs(Tstar), abs(Tobs), lower.tail = F)  # both tails
```

```
## diffmean
##    0.035
```

What is the correct p-value for this analysis?

- ◯ **0.017**
- ◯ **0.018**
- ◯ **0.035**

Submit Clear

7. Conclusion and Scope of Inference

What is your decision for the 217 data? Use a significance level of 0.05?

- ◯ **Reject the null**
- ◯ **Accept the null**
- ◯ **Fail to reject the null**
- ◯ **none of the above**

Submit Clear

What do you conclude? Write a conclusion in full and coherent sentence to answer the research question. Include the scope of inference.