Chapter 1: Two Sample t-Test

September 2, 2016

Previously in Stat 217...

We tested a hypothesis that people perceived as unattractive may experience some types of discrimination. The permutation test was used to compare the sentence length that male jurors gave to female defendants that were already divided into “Average” and “Unattractive” categories. We found some evidence that the discrimination exists, but the scope of inference is limited. The result only applies to the participated subjects and we cannot say there is an effect of attractiveness to sentence lengths.

Randomization methods, like permutation test, are what called **nonparametric** that often make fewer assumptions. We used the computer to shuffle, sample, and simulate observations.

Today we will focus on a **parametric** method, which makes assumptions about the distribution of the test statistic. We use theoretical distributions (like a Normal or t distribution) to compare a standardized test statistic to and get p-value. We’ll use the **two independent sample t-test** method.

This is the code to help you start with necessary data set.

```
# Load packages
require(heplots)
require(mosaic)
require(beanplot)

# Load the data set
data(MockJury)
# Remove Beautiful group
MockJury2 <- MockJury[MockJury$Attr != "Beautiful", ]
MockJury2$Attr <- factor(MockJury2$Attr)
```

Data Analysis by Two Independent Sample t-Test Method

1. Hypotheses

We use the same null and alternative hypothesis as last time

$H_0 : \mu_A = \mu_U$ vs. $H_A : \mu_A \neq \mu_U$

State the null and alternative hypotheses in words.

2. Check The Assumptions

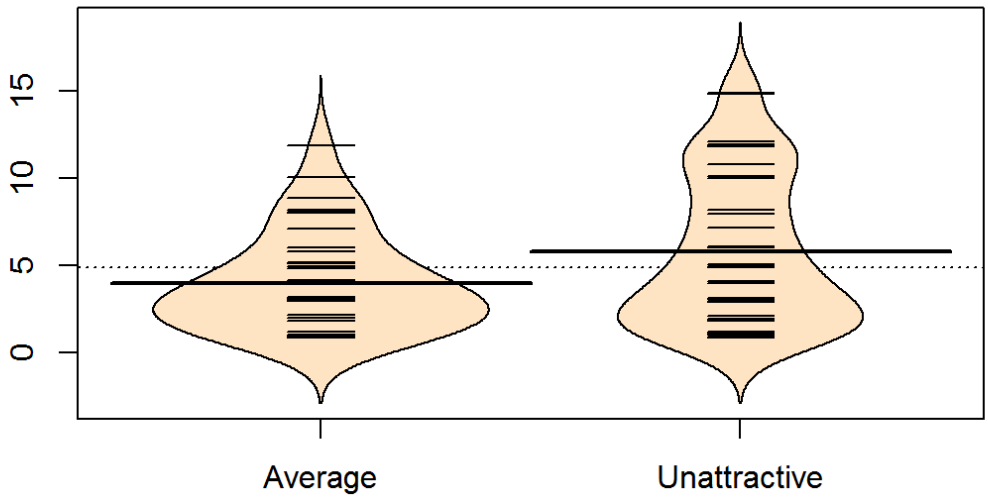
As the name of test **two independent sample t-test** method suggests, we need to check the data set for the following assumptions before we can proceed with the chosen method.

1. **Independent observations**. An observation in the data set is unrelated to all other observations.

Explain why our data should meet this assumption.

2. **Equal variances**. Use beanplots to compare the variability between groups. Be careful if group's sizes are greatly different.

```
beanplot(Years ~ Attr, data = MockJury2, log = "", col = "bisque", method =
"jitter")
```



What do you think about this assumption? Explain.

3.

a. (For t-test) **Normal distributions** in each group. Looking for skewness and outliers in the beanplots. Those are problems.
b. (For permutation test) **Similar distributions between the groups**. This assumption is **relaxed** if permutation method is used.

In the beanplots above, is there any problem with assumption (3a)? With assumption (3b)?

3. t-Statistic

Let's for now ignore the assumptions requirements for two independent t-test above (1, 2, and 3a) and proceed with the next step in the non-parametric method.

In this class we're going to use the following formula for t:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ where } s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

> Note: You may recall using different formula in Stat 216. It turns out there are actually a couple different versions of the test statistic t. The one above may not be correct if two samples are dependent, or don't have equal variances.

The formula looks quite scary, but we can leave it to R to calculate t-statistic for us. The function `t.test`, then accessing to its `statistic` attribute, will give us the t-statistic of the test for difference in mean sentence lengths between "Average" and "Unattractive" female defendants.

```
t.test(Years ~ Attr, data = MockJury2, var.equal = TRUE)$statistic
```

```
##         t
## -2.17023
```

> Note: t-statistic is a negative number because it uses opposite subtraction order (A - U) from `diffmean` (which uses U - A). Don't worry, it won't affect the result.

Core difference between parametric and non-parametric methods: If all the assumptions are met, under the null hypothesis, the t-statistic will follow the t-distribution with df = $n_1 + n_2 - 2$. So to find the evidence against the null, instead of comparing the observed result with the simulated null distribution in non-parametric method, we will compare the t-statistic with the t_{df} distribution.

> Note: t-distribution always associates with a degree of freedom. There is no specific t-distribution without degree of freedom.

In this analysis, use the following output to find the degree of freedom for this t-distribution.

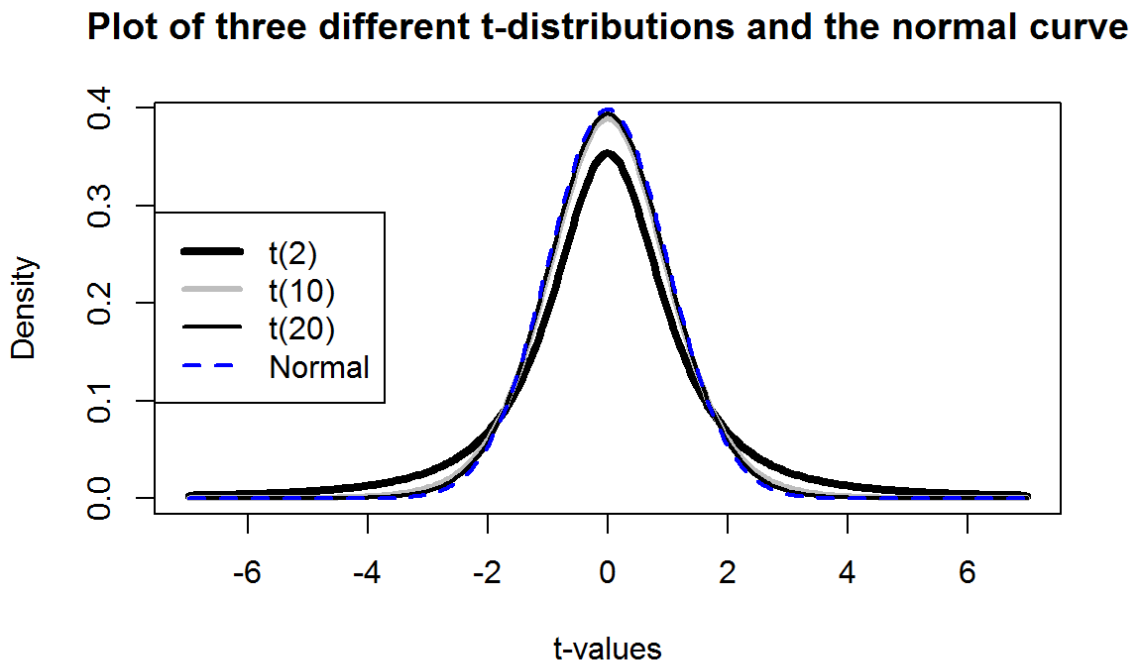
```
favstats(Years ~ Attr, data = MockJury2)
```

```
##           Attr min Q1 median Q3 max     mean       sd  n missing
## 1      Average   1  2      3  5  12 3.973684 2.823519 38       0
## 2 Unattractive   1  2      5 10  15 5.810811 4.364235 37       0
```

1. What is n_1 ?
2. What is n_2 ?
3. What is the degree of freedom?

Submit Clear

Review t-distributions



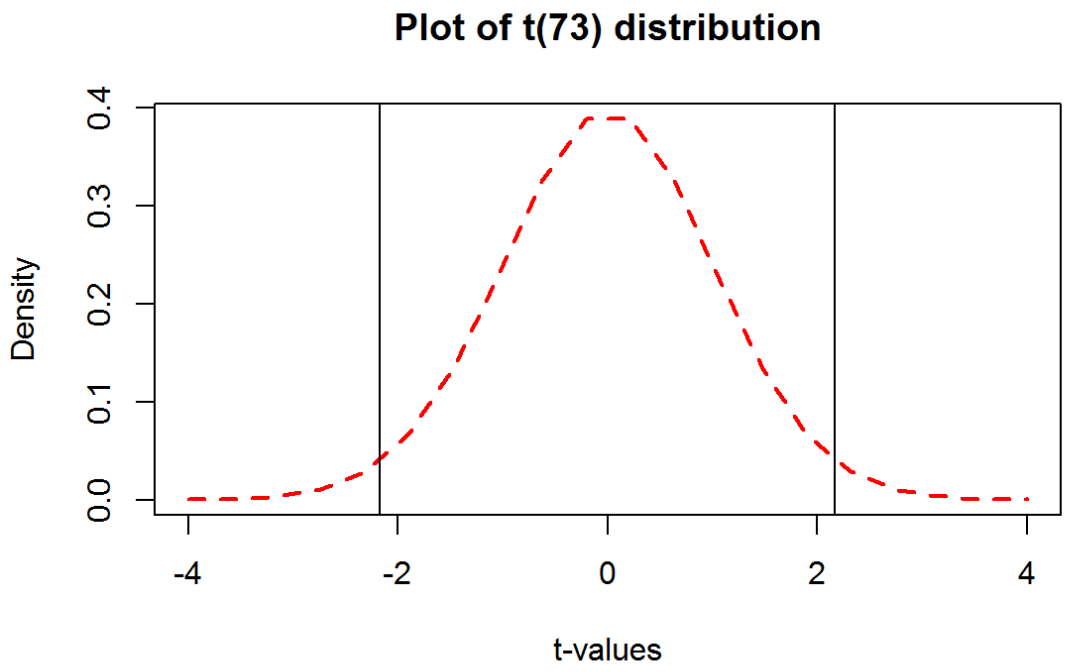
What is the relation between t distribution and normal distribution?

- ○ **t distribution looks more and more like normal distribution when the df gets smaller.**
- ○ **t distribution looks more and more like normal distribution when the df gets larger.**

Submit Clear

4. Find p-value

R has multiple ways to find the p-value for two independent sample t-test. One way of doing it is to follow the definition of p-value. As stated above, p-value can be calculated by comparing the t-statistic (-2.1702) to t_{73} .



In the plot, the dashed line is t_{73} distribution, the vertical lines are t-statistic (-2.1702) and its opposite value (2.1702). Why do we also put the positive t-statistic in the plot?

- ◯ **Because we're doing one-tail test.**
- ◯ **Because we're doing two-tail test.**

Submit Clear

`pt()` function in R will return the p-value.

```
pt(-2.1702, df = 73, lower.tail = T)   # p-value if it's left-tail test
```

```
## [1] 0.01662286
```

```
2 * pt(-2.1702, df = 73, lower.tail = T)   # p-value if it's two-tail test
```

```
## [1] 0.03324571
```

What is the correct p-value for this analysis?

- ◯ **0.0167**
- ◯ **0.033**

Submit Clear

There's another way to find the p-value for this test, which I called **lazy method**. All you need to do is call `t.test()` function, give it the data set, and make sure `var.equal = TRUE` to match the second assumption. It will find out the degree of freedom, t-statistic, p-value, and another piece of information we'll see more in the next class as well.

```
t.test(Years ~ Attr, data = MockJury2, var.equal = TRUE)  # By default, it's two-tailed test
```

```
##
##  Two Sample t-test
##
## data:  Years by Attr
## t = -2.1702, df = 73, p-value = 0.03324
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.5242237 -0.1500295
## sample estimates:
##      mean in group Average mean in group Unattractive
##                   3.973684                   5.810811
```

The test that the above function performs is two-tailed test by default, which is what we want for the MockJury analysis. If a one-tailed test is needed, you need to specify `alternative=` argument like below.

```
t.test(Years ~ Attr, data = MockJury2, var.equal = TRUE, alternative = "greater")  # Right-tailed test
t.test(Years ~ Attr, data = MockJury2, var.equal = TRUE, alternative = "less")  # Left-tailed test
```

5. Make Decision

What is your decision for the MockJury 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

6. Write Conclusion

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

Choosing Between Two Methods

From what you've found when you check the assumptions, what would be the better method to use for this analysis?

- ◯ **Non-parametric permutation test.**
- ◯ **Parametric two independent sample t-test.**

Submit Clear