# How meaningful is your p-value?

*Jinlin He, Yuan Wang*

*April 18, 2017*

**Objective**

In statistical hypothesis testing, many researchers use **p-value** as a measure of how strong or significant the results to prove or disprove the null hypothesis, i.e. the level of significance of the evidence in showing if there is a mean difference. However, without the understanding of the meaning of p-value and adequate prudence, people could falsely use p-value in scientific research that leads to misleading conclusions. John P. A. Ioannidis (2005) frankly expressed her concern in her paper Why Most Published Research Findings Are False that "...simulations show that for most study designs and settings, it is more likely for a research claim to be false than true...it can be proven that most claimed research findings are false."

This lab aims to walk through a series of studies that all conduct similar tests but yielding various results, and show students that how p-value changes with the sample from the same population and when the tests are conducted correctly with identical procesdures.

---

**Outline**

- Introduction to data source
- **Warning**: a preliminary study
- Hypothesis test on sample One
- Hypothesis test on sample Two
- Comparing multiple hypothesis tests across groups
- Exploration
- Discussion
- Further Readings

---

**1.Data**

In this tutorial, we are using data generated by an online game (Shapesplosion). The game requires players to place specifically shaped pegs into the appropriate holes within a short time period. For several years students have used the Shapesplosion game to design an experiment and collect data.
*Link to the game: (http://web.grinnell.edu/individuals/kuipers/stat2labs/Perfection.html).*

Here is a screenshot of the start page for the players can choose their preferred game settings. The game also allows researchers to create additional variables of their design. As shown below, "Gender" can be added as a factor of interest in this way.

In this lab, we focus on data collected from experiments that have added "Gender" as a factor of interest. Thus, these data can be used to address a specific research question:

- **"Does gender affect the time used to play Shapesplosion game?"**

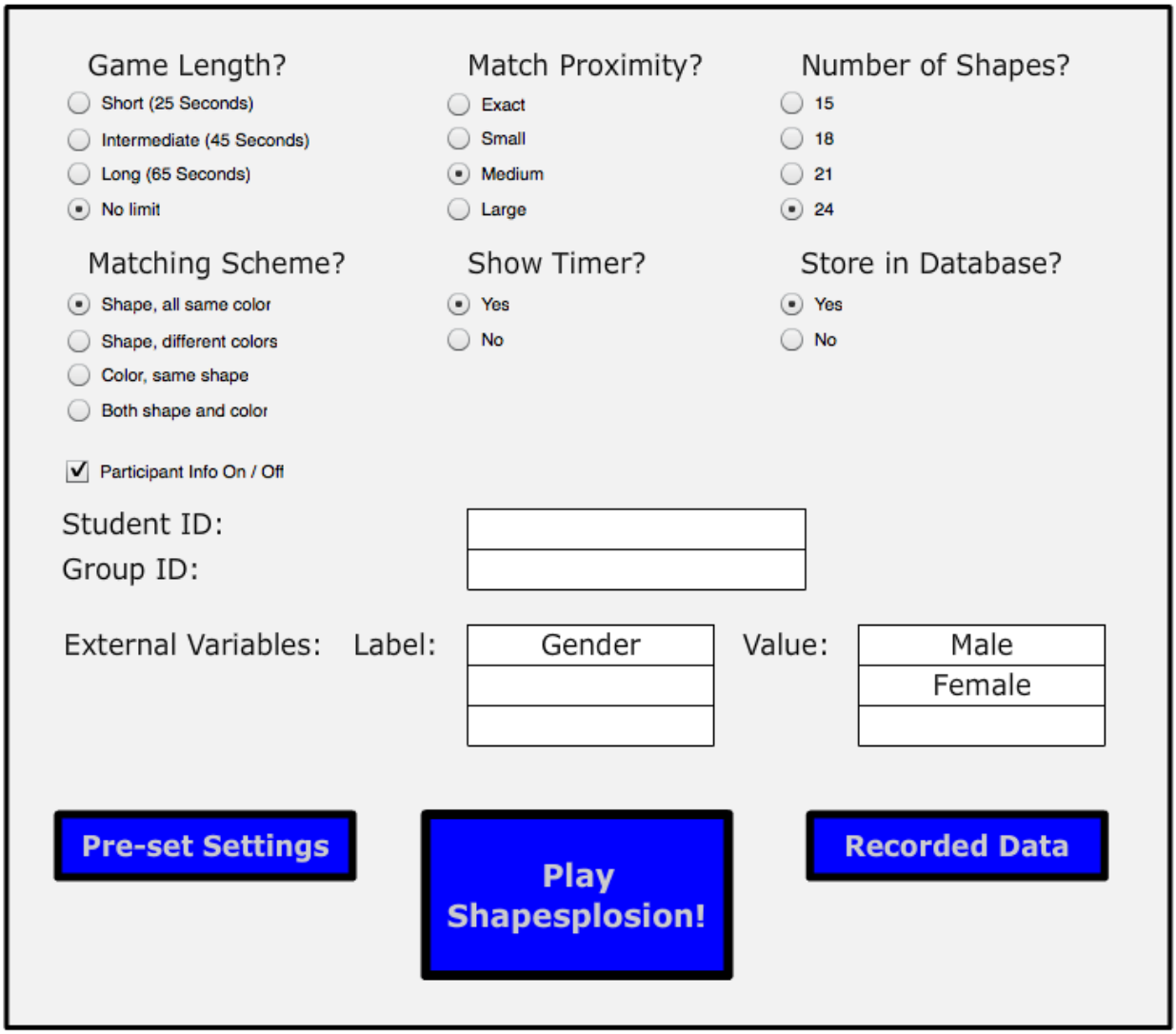

Game Length?
Short (25 Seconds)
Intermediate (45 Seconds)
Long (65 Seconds)
No limit
Match Proximity?
Exact
Small
Medium
Large
Number of Shapes?
15
18
21
24
Matching Scheme?
Shape, all same color
Shape, different colors
Color, same shape
Both shape and color
Show Timer?
Yes
No
Store in Database?
Yes
No
Participant Info On / Off
Student ID:
Group ID:
External Variables:
Label:
Gender
Value:
Male
Female
Pre-set Settings
Play Shapesplosion!
Recorded Data

Figure 1:

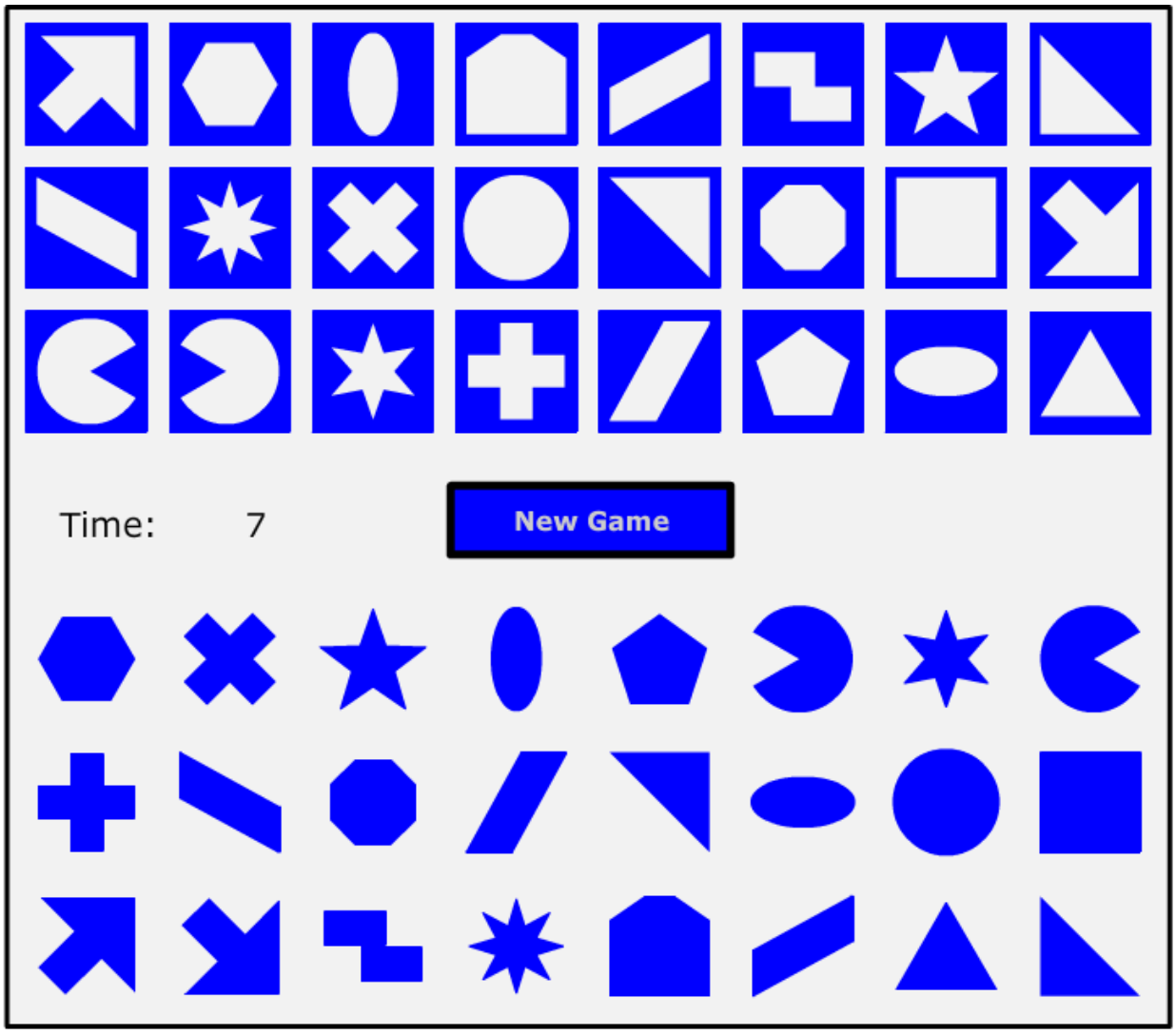



Figure 2:

## 2. A preliminary study

Before we start, the following code will import your data as well as the R libraries for our analysis.

```
library(mosaic)
library(ggplot2)
```

```
## Reading in the data
group_data <- read.csv("cleaned_gender.csv")
```

---

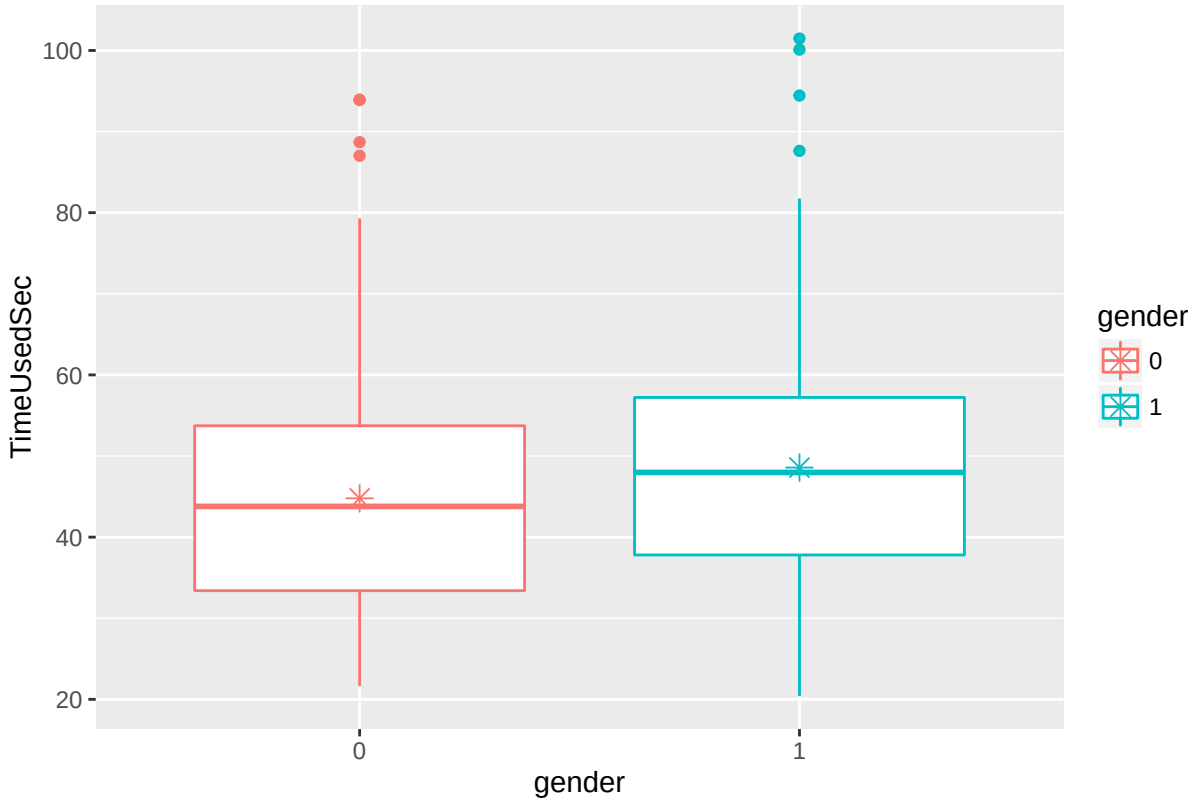
**Question:**

1) Use the code below to create a dotplot and boxplot. These plots use the entire data set. How useful is this plot? In other words, is it reasonable to conduct a hypothesis test on gender based on the entire data set?

---

```
group_data$gender= as.factor(group_data$gender)
ggplot(data=group_data, aes(x=gender, y=TimeUsedSec)) + geom_boxplot()  + aes(colour=gender) + theme(leg
```

Figure 1: boxplot of full dataset




**Warning: DO NOT ever conduct a hypothesis test on the entire data set without understanding the data! We normally require similar sample units as a criterion for conducting a meaningful hypothesis test; however, with the entire data set, this criterion is usually not met.**

---

## 3. Hypothesis Test on Sample One

### Import data

We will start our analysis by running a two-sample t-test on the data of a group named "MATH22015". The timestamps of the data show that this group conducted all their experiments on August 29th. The group has collected 37 responses, with 19 responses from female participants and 18 responses from male participants. The following code will import your data and pull out the group data corresponding to the group "MATH22015".

```
group_data <- read.csv("cleaned_gender.csv")
group_data$gender <- as.factor(group_data$gender)
MATH22015<-group_data[group_data$groupID=="MATH22015",]
MAT_female <- MATH22015[MATH22015$gender == 0,]$TimeUsedSec
MAT_male <- MATH22015[MATH22015$gender == 1,]$TimeUsedSec
```

### Check for Assumptions

Before we conduct t-test on the sample data, we need to question what are the assumptions of conducting a two-sample t-test?
For example, recall that one of the assumptions would be the equal variance for both groups. Use the code below to create a dotplot and boxplot of the data.

---

**Question:**

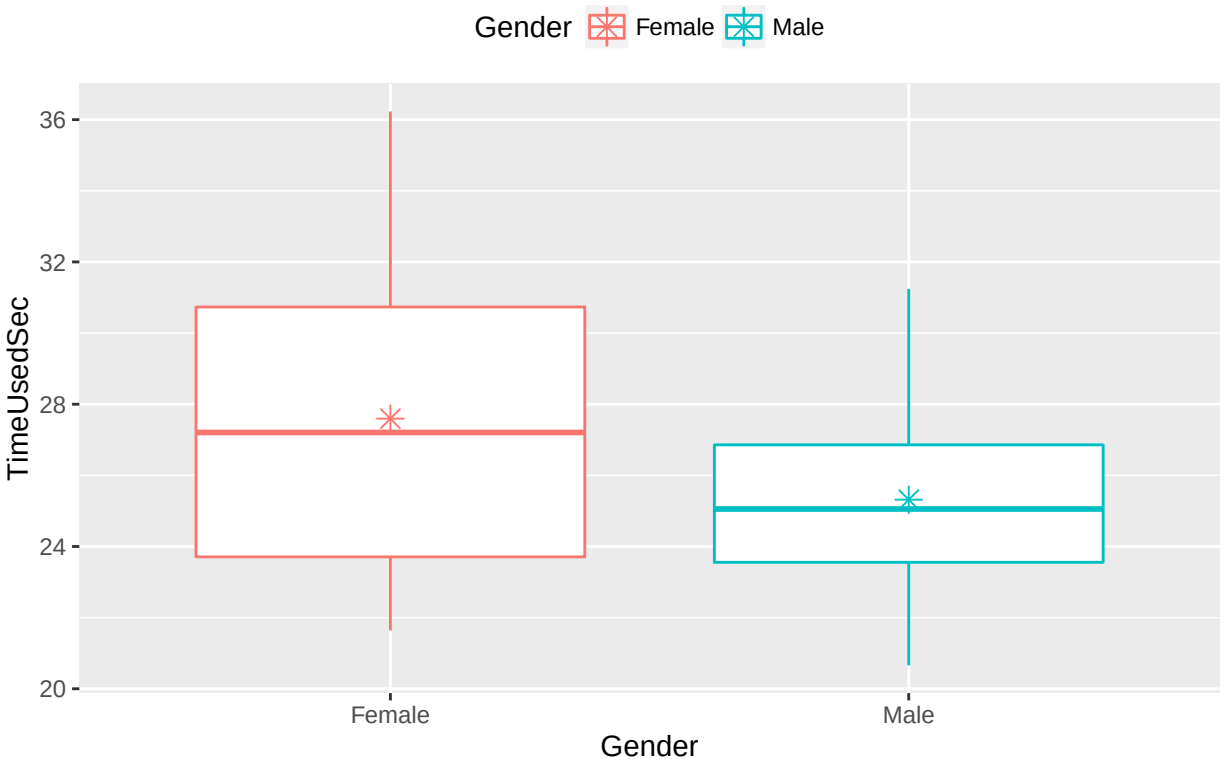
2) Does it seem reasonable to use a two-sample t-test for this data?

---

```
library(mosaic)
library(ggplot2)
```

```
## Boxplot
ggplot(data=MATH22015, aes(x=gender, y=MATH22015$TimeUsedSec)) + geom_boxplot()  + theme(legend.positio
labs(title="Fig.1: boxplot of MATH22015 sample", x = "Gender", y ="TimeUsedSec") + aes(colour = gender)
```

Fig.1: boxplot of MATH22015 sample




*Star represents respective sample mean*

**Conduct the two-sample t-test**

After checking for assumptions, we proceed to conduct the two-sample t-test for the sample data of group "MATH22015":

```
## Conduct t-test
test1 <- t.test(MAT_male,MAT_female,alternative="two.sided")
test1$p.value
```

```
## [1] 0.0776448
```

---

**Questions:**

3) With a reference to our research question ("Does gender affect the time used to play Shapesplosion game?"), what is the null hypothesis of the t-test?

4) Write two to three sentences clearly stating conclusions can you draw from this study. Please assume that the data was collected properly from a class of 37 students in an introductory statistics class.

---

Recall that the p-value means the probability of observing a mean difference in play time between the female group and male group as extreme as 2.278 (observed mean difference) is 0.078. On an alpha level of 0.1, this study provides sufficient evidence for us to reject the null hypothesis that gender does not affect the play time of Shapesplosion. In other words, the result of the hypothesis test leads us to conclude that female players indeed have different play time of Shapesplosionas compared to their male conterparts.

**Remark**

- Though the hypothesis test is two-sided, and we are unable to conclude whether female players take shorter or longer time than their male counterparts, the boxplot in Fig.1 seems to suggest that female players on average take more time to play the game than male players.

---

### 4. Hypothesis Test on Sample Two

Now we take the sample data from another group with the groupID of "mth22602". Timestamps show that this group conducted their experiments in October 25th, 2013. The group has 27 observations in total with 10 observations from the female participants and 17 observations from the male participants.

```
mth22602 <-group_data[group_data$groupID=="mth22602",]
mth_female <- mth22602[mth22602$gender == 0,]$TimeUsedSec
mth_male <- mth22602[mth22602$gender == 1,]$TimeUsedSec
mth_t_test <- t.test(mth_male, mth_female, alternative="two.sided")
mth_t_test$p.value
```

```
## [1] 0.1484163
```

After similar process of conducting a two-sample t-test on the data, we obtained a p-value of 0.1484.

**Questions:**

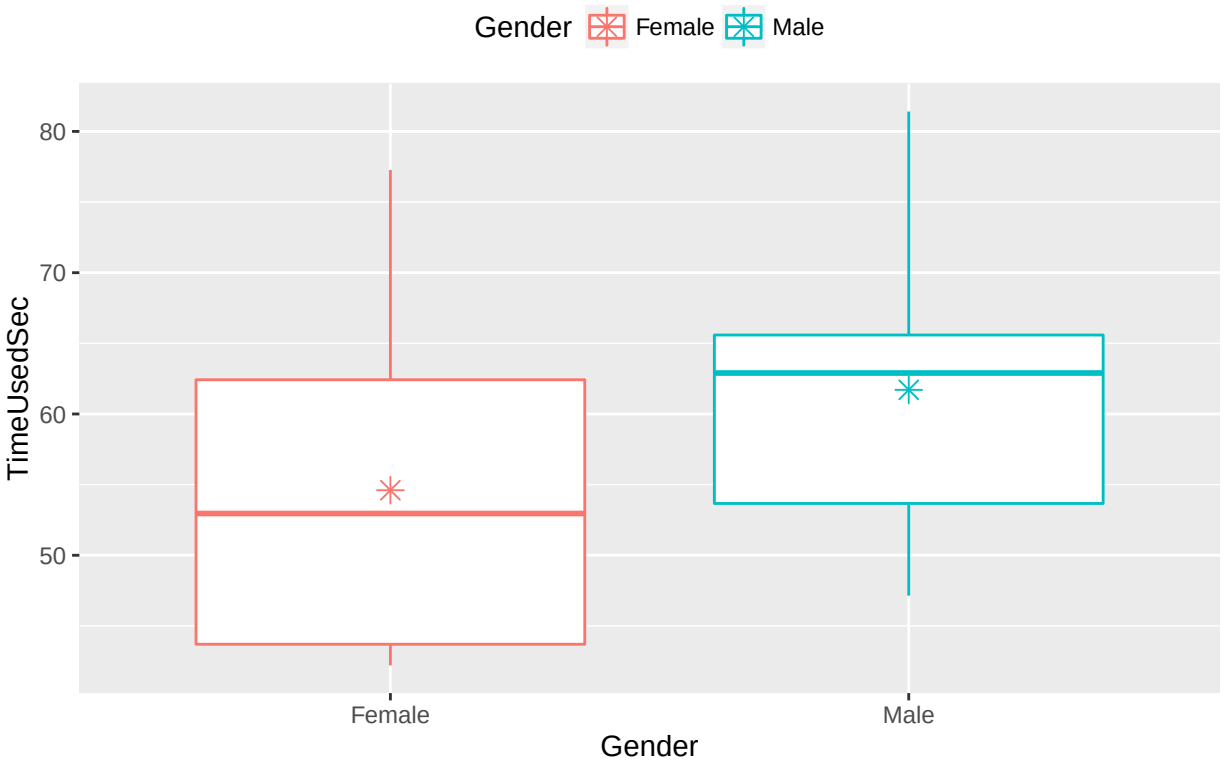
5) Based on this p-value, what conclusion can we draw from this data based on an alpha level of 0.1?

6) Now let us take a look at the graph generated below, what does the boxplot in Fig.2 suggest differently from the boxplot in Fig.1?

```
mth22602_mean <- data.frame(gender = c(1,0), value = c(mean(mth22602[mth22602$gender ==1,]$TimeUsedSec)

ggplot(data=mth22602, aes(x=gender, y=TimeUsedSec)) + geom_boxplot() + theme(legend.position="top") +
  scale_colour_discrete(name="Gender",labels=c("Female", "Male"))+labs(title="Fig.2: boxplot of mth2260
  scale_x_discrete(labels=c("Female", "Male"))
```

Fig.2: boxplot of mth22602 sample




*Star represents respective sample mean*

---

**Question:**

7) Write three to four sentences clearly explaining how two studies asking the same research questions with similar methodologies would get different results? Does this show evidence that one of the groups made an error somewhere in their data collection or analysis?

---

### 5. Comparing multiple hypothesis tests across groups

In Part One of this activity, you compared two different studies that evaluated the effect of gender on completion time of the shapesplosion game. Several additional studies on gender were conducted by multiple groups over the years. The following code conducts t-test and creates boxplots for each of the groups.

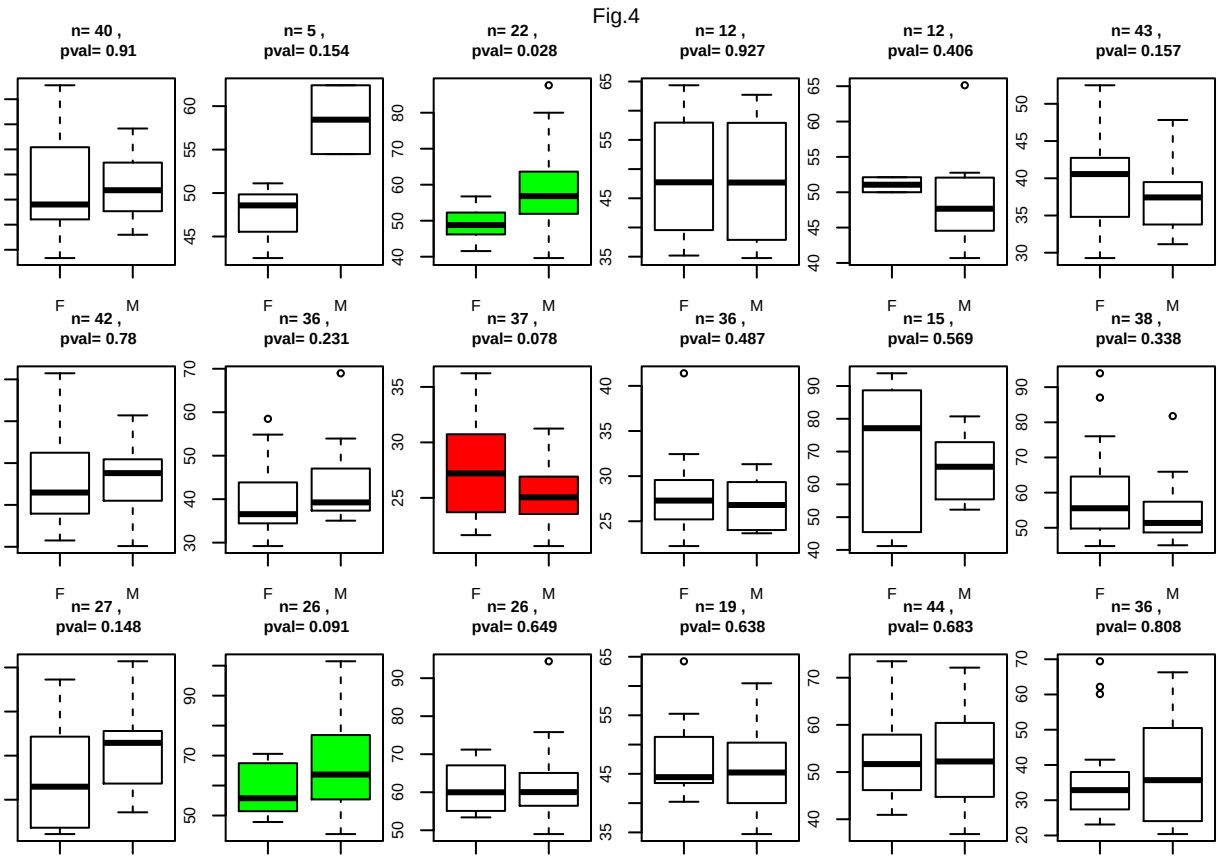



Fig.4 is a graphical representation of all 18 groups from the group dataset. The colored boxplots have p-value $< 0.1$. In addition, the green plots indicate results that female participants have lower mean play times than male participants, while red plot shows that male participants have lower mean play time than female participants.

---

**Questions**

8) What is the range of p-values observed in these studies?

9) Amongst all groups, how many of them had higher mean times for females? How many groups had higher mean times for males?

10) Amongst groups that did not produce significant p-values (plots that are uncolored), which graph visibly appears to show the biggest difference between genders? What reasons could explain why this group did not observe a significant p-value?

11) Why do the p-values differ?

---

## 6. Exploration

### The first Shiny applicaitons

So far we have observed that p-values differ when we sellect different groups of students and conduct hypothesis tests on each group. Does the variability of P-value happen to occur for the groups chosen in this tutorial? Follow the LINK and examine the P-value from all other group samples from the dataset. This application

loops through every group sample and conduct hypothesis test of gender on each group sample. Take as many group samples and observe the variation of p-values. Now answer this, **how credible is p-value?**

**The second Shiny application**

Is this phenomenon occur specifically for our game data set? Play around with this APP . Suppose we can collect a population with a theoretical distribution and randomly sample from the population with parameters of your choice: theoretical distribution, sample size, sample mean, and sample standard deviation. Then conduct multiple tests. **How credible is p-value** in this case?

**"The dance of p-value"**

Now take a look at this YouTube video to get a better understanding of the **incredibility** of p-values.

---

## 7. Discussion

**Now we come back to our initial quesiton: how credible is p-value?**

**Firstly, it is important to remember the definition of p-vlaue. In this context, it is the probability of obtaining a mean difference in play time between male and female players as extreme as we observed in our repective samples, on the premise that the null hypothesis (male and female players spend equal time on the game) is true. Therefore, if the null hypothesis is false in the first place, that is if the population mean of female playtime and male playtime are indeed different, p-value does not make any meaningful suggestions to our research question.**

**Secondly, statistics vary whenever we perform a study. Though population and methods remain identical. To quote the finding from an article(Cumming, 2008), "It has been suggested that if an initial experiment results in two-tailed p = .05, there is an 80% chance the one-tailed p value from a replication will fall in the interval (.00008, .44) [. . . ] Remarkably, the interval – termed a p interval – is this wide however large the sample size", we can thus rationalize the obeservation in the exercise above that a repeat of the same experiment resulted in a substantially different P value.**

---

## 8. Further Reading:

An overview of variability of p-value: **The fickle P value generates irreproducible results** (http://www.nature.com/nmeth/journal/v12/n3/pdf/nmeth.3288.pdf)

Alternative ways to measure variability of p-value: **Assessment of P-value variability in the current replicability crisis** (https://arxiv.org/pdf/1609.01664.pdf)

**References:**


1. Cumming G. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspectives on Psychological Science. 2008;3(4):286–300