How meaningful is your p-value?

Yuan Wang, Jinlin He

April 3, 2017

## Objective

In statistical hypothesis testing, many researchers use **p-value** as a measure of how strong/significant the results are such that the null hypothesis is true, 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](http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124) 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, such that students can understand how p-value changes with the sample even with exactly the same population.

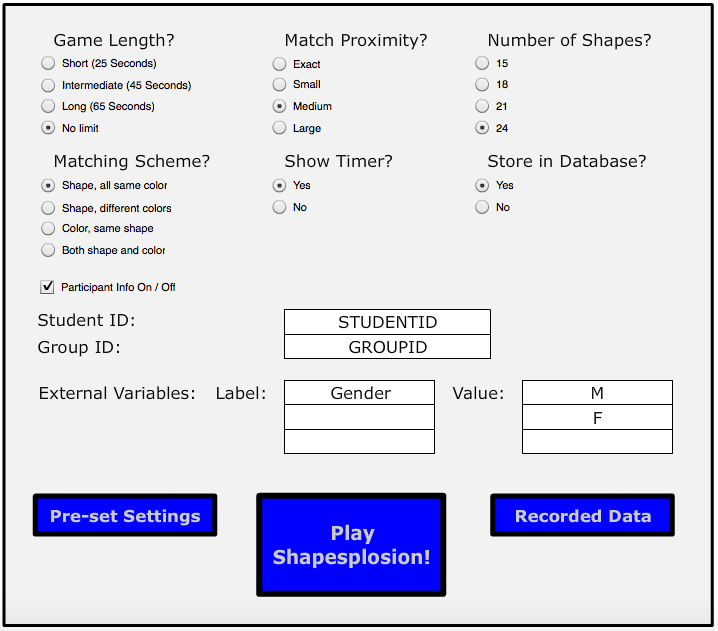
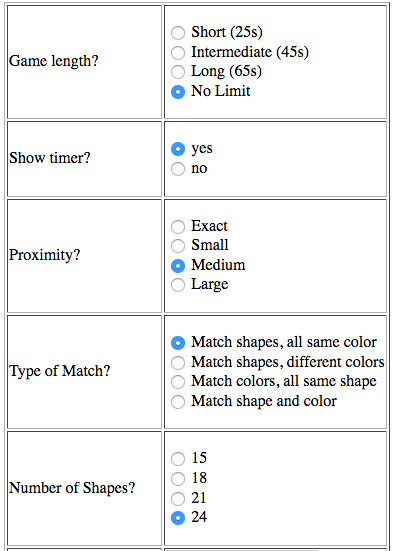
## Outline

* Introduction to data source
* **Warning**: a preliminary study
* **Hypothesis test**: the first study
* **Hypothesis test**: the second study
* **Hypothesis tests**: the third study
* Discussion
* Further exploration

## Data

Shapesplosion is an on-line game in which a person is expected 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. The following link allows you to play the game (<http://web.grinnell.edu/individuals/kuipers/stat2labs/Perfection.html>).

Here are two screenshot of the game where the users can choose their preferred settings for the game. The image on the left is the start page which all users would encounter when playing the game, and the image on the left shows more specifics with regards to each variable.

In this lab, we will review data from multiple student groups that focused on a specific research question:

*“Does gender affect the time used to play Shapesplosion game?”*

## A preliminary study

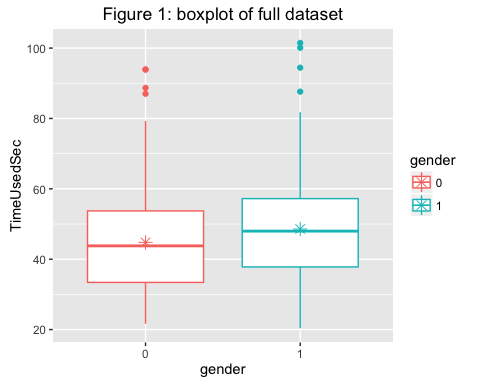
Before we start, the following code will import your data as well as the R libraries for our analysis.

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 useful/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(legend.position="right") + labs(title="Figure 1: boxplot of full dataset") + stat\_summary(fun.y = mean, geom = "point", pch = 8, cex = 3)



#### 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.

## The second study

Let’s star our official analysis on a new study, using the group\_ID “MATH22015” with a sample size of 37, which is a subset of the full dataset.

MATH22015<-group\_data[group\_data$groupID=="MATH22015",]  
dim(MATH22015)[1]

## [1] 37

MAT\_female <- MATH22015[MATH22015$gender == 0,]$TimeUsedSec  
MAT\_male <- MATH22015[MATH22015$gender == 1,]$TimeUsedSec  
MAT\_t\_test <- t.test(MAT\_male,MAT\_female,alternative="two.sided")  
MAT\_t\_test

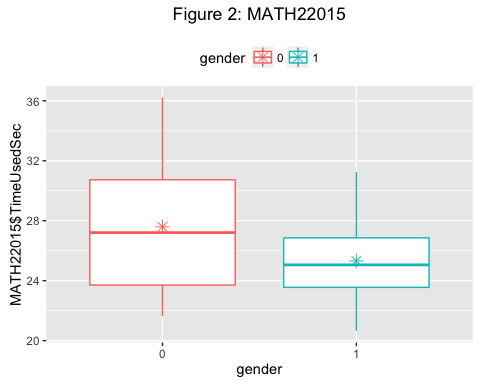
##   
## Welch Two Sample t-test  
##   
## data: c(23.705, 29.542, 31.245, 20.657, 29.102, 30.141, 26.102, 25.11, ... and c(36.229, 21.646, 32.117, 27.249, 30.276, 33.171, 31.19, 27.632, ...  
## t = -1.8214, df = 32.9, p-value = 0.07764  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -4.8235423 0.2668288  
## sample estimates:  
## mean of x mean of y   
## 25.31622 27.59458

**QUestion 2**: *What can you conclude from the study?*

**Answer**: *This study resulted in a p-value of 0.0776. Using an alpha level of 0.1, this study would have evidence to reject the null hypothesis and concluded that gender indeed makes a difference in play time of the game.*

The graph also suggests that female players on average take less time to play the game than male players, contrary to our previous findings. Render the graph with the code below.

MATH22015$gender= as.factor(MATH22015$gender)  
  
ggplot(data=MATH22015, aes(x=gender, y=MATH22015$TimeUsedSec)) + geom\_boxplot() + theme(legend.position="top") + labs(title="Figure 2: MATH22015") + aes(colour = gender)+   
 stat\_summary(fun.y = mean, geom = "point", pch = 8, cex = 3)



## The second study

Take another subdataset with groupID = hjf190f14 with sample size 27, let us again conduct a hypothesis test on this group.

mth22602 <-group\_data[group\_data$groupID=="mth22602",]  
dim(mth22602)[1]

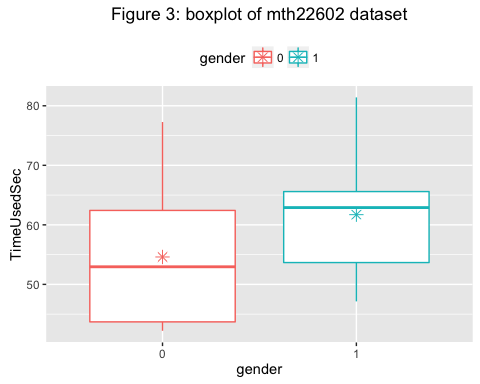
## [1] 27

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

##   
## Welch Two Sample t-test  
##   
## data: c(53.765, 65.589, 74.76, 72.605, 58.194, 47.776, 50.06, 47.142, ... and c(56.77, 49.308, 64.307, 42.205, 43.649, 56.603, 77.27, 69.601, ...  
## t = 1.5173, df = 16.24, p-value = 0.1484  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -2.809845 17.019504  
## sample estimates:  
## mean of x mean of y   
## 61.70853 54.60370

With a p-value of 0.15, we surely fail to reject the null hypothesis when alpha = 0.1. Now let us take a look at the graph, from which we observe that male mean seems to be much higher than female mean, but still the hypothesis test failed.

mth22602$gender<- as.factor(mth22602$gender)  
mth22602\_mean <- data.frame(gender = c(1,0), value = c(mean(mth22602[mth22602$gender ==1,]$TimeUsedSec), mean(mth22602[mth22602$gender ==0,]$TimeUsedSec)))  
ggplot(data=mth22602, aes(x=gender, y=TimeUsedSec)) + geom\_boxplot() + theme(legend.position="top") + labs(title="Figure 3: boxplot of mth22602 dataset") + aes(colour = gender) + stat\_summary(fun.y = mean, geom = "point", pch = 8, cex = 3)



**Question 3**: *Write two to three 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?*

**Answer**: *Statistics vary whenever we perform a study. Though population and methods remain identical*

## The third study:

#### Comparing multiple hypothesis tests

group\_data$gender <- as.factor(group\_data$gender)

In Part 1 of this activity, we 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 multiple years. The following code conducts t-test and creates boxplots for several of these groups.

### Add margin to add titles  
par(mar=c(2,2,2,2))  
par(mfrow = c(4,5))  
groupName <- c()  
pvalues <- c()  
  
tb <- as.data.frame(table(group\_data$groupID))  
##Select groups that are under size 50  
tb2 <- tb[tb$Freq >= 5 & tb$Freq <= 50,]  
##Create a vector of groupID's whose size is between 25 and 50  
selected\_groupID <- as.character(tb2$Var1)  
for (i in 1:length(selected\_groupID)) {  
 female <- group\_data[group\_data$groupID == selected\_groupID[i] & group\_data$gender == 0,]$TimeUsedSec  
 male <- group\_data[group\_data$groupID == selected\_groupID[i] & group\_data$gender == 1,]$TimeUsedSec  
 if (length(female) > 1 & length(male) > 1) {  
 groupName <- cbind(groupName, selected\_groupID[i])  
 p <- round(t.test(female, male)$p.value, digits = 3)  
 pvalues <- cbind(pvalues, p)  
 data1 <- group\_data[group\_data$groupID == selected\_groupID[i],]  
 if (p < 0.1) {  
 if (mean(female) < mean(male)) {  
 boxplot(TimeUsedSec ~ gender,data=data1, col="green", main=paste("n=", dim(data1)[1], ", pval=", p), xlab="Gender (1=Male, 0=Female)", ylab="Time Used Seconds")  
 } else {  
 boxplot(TimeUsedSec ~ gender,data=data1, col="red", main=paste("n=", dim(data1)[1], ", pval=", p), xlab="Gender (1=Male, 0=Female)", ylab="Time Used Seconds")  
 }  
 } else {  
 boxplot(TimeUsedSec ~ gender,data=data1, main=paste("n=", dim(data1)[1], ", pval=", p), xlab="Gender (1=Male, 0=Female)", ylab="Time Used Seconds")  
 }  
 }  
}  
  
mtext("Figure 4", outer=TRUE, cex=1, line=-1)

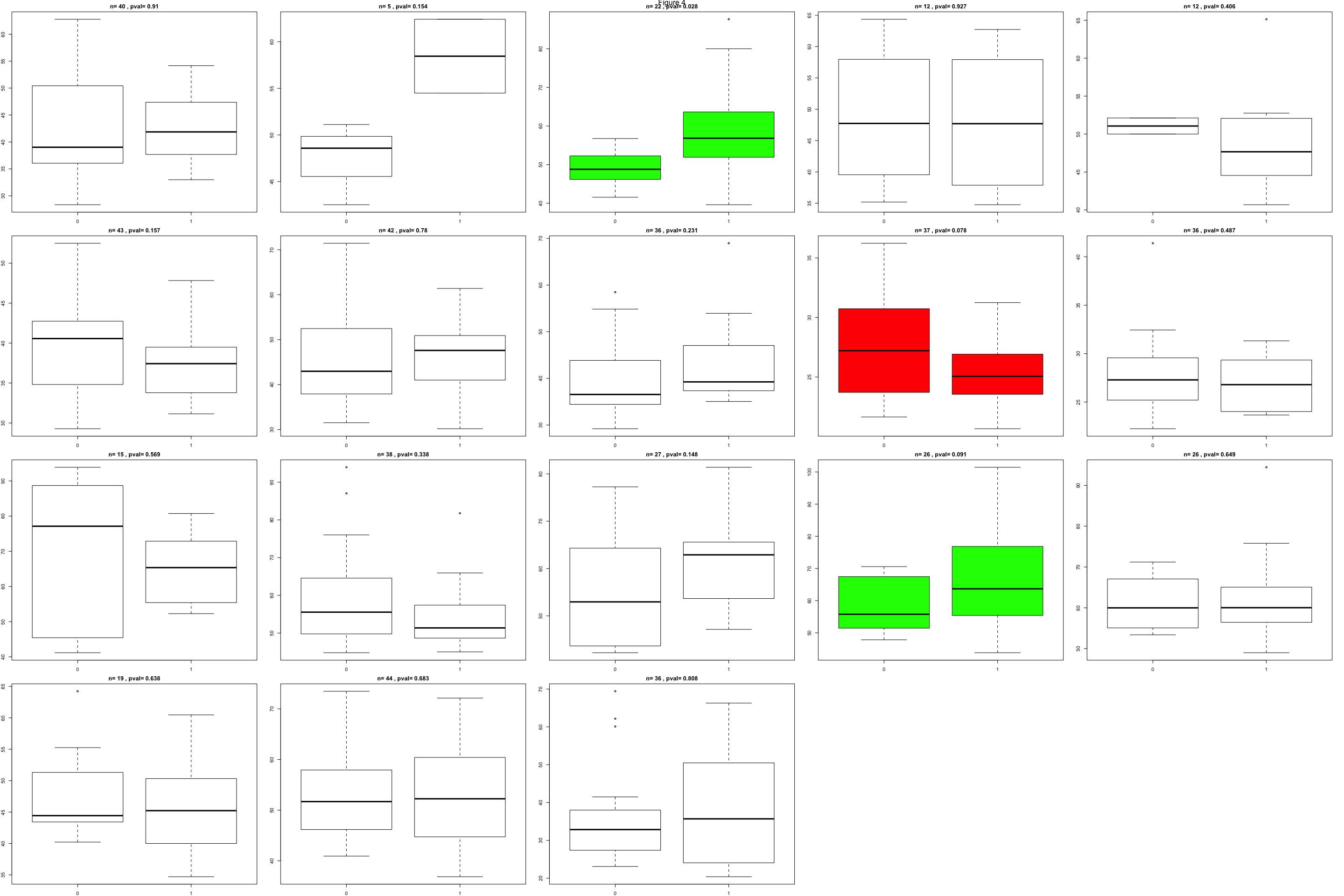


Figure 4 is a graphical representation of all 18 groups from the group dataset. The colored boxplots have p-value < 0.1.

**Question 4**: *What is the range of p-values observed in these studies?*

**Question 5**: *How many groups had higher mean times for females? How many groups had higher mean times for males?*

**Question 6**: *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?*

**Question 7**: *Why do the p-values differ?*

## Discussion: What exactly is p-value?

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 imply information as meaningful as we thought.

## How credible is p-value?

#### The first Shiny applicaitons

It is noticeable that p-value is not reliable with repetition either. In our previous studies we observed that p-values differ when we sellect different groups of students and conduct hypothesis tests on each group. Is it a "group-specific thing"? Follow the LINK and play around with this application. This application takes the entire data set as the population and randomly sample from the population with sample size of your choice. Take many samples and observe the variation of p-values. Now answer this, **how credible is p-value?**

#### The second Shiny application

Is this phenomenon due to our 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](https://www.youtube.com/watch?v=5OL1RqHrZQ8) to get a better understanding of the **incredibility** of p-values.