How meaningful is your p-value?

Jinlin He, Yuan Wang

February 13, 2017

//////////////search for packages to easy format .rmd

Objective

In order to better understand the meaning of a p-value, we are going to compare multiple studies that all conducted similar tests.

Data

Need to add more explanation of the game and a screenshot of the game webpage

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 is a screenshot of the start page where the users can choose their preferred settings for the game.

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?"

Part One

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

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

```
## Consucting a two-sided t-test using the entire dataset
female <- group_data[group_data$gender == 0,]$TimeUsedSec
male <- group_data[group_data$gender == 1,]$TimeUsedSec
t_test<- t.test(male,female,alternative="two.sided")
t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  male and female
## t = 2.9171, df = 505.699, p-value = 0.00369
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.239033 6.350631
## sample estimates:
## mean of x mean of y
##  48.57590  44.78107
```

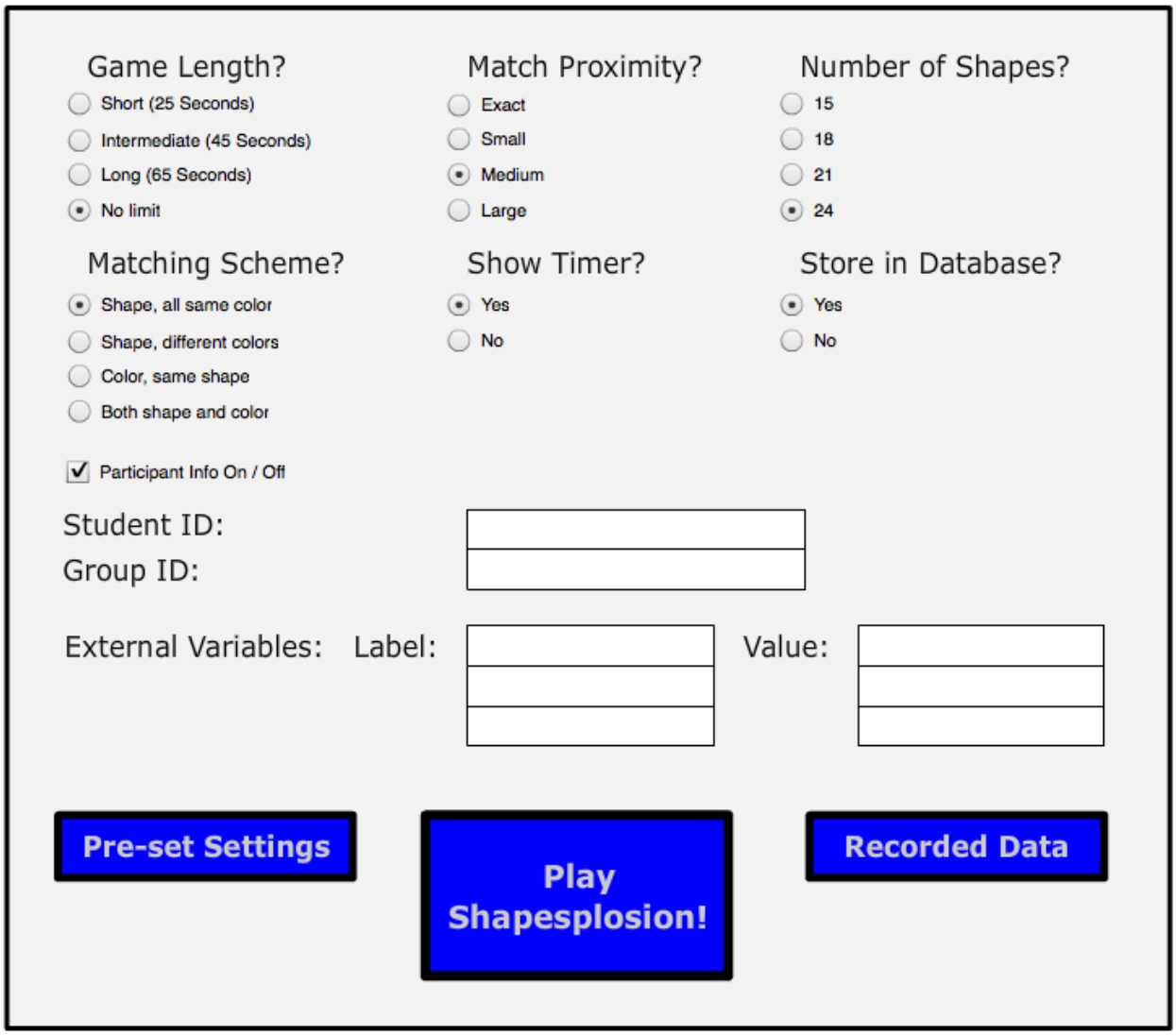
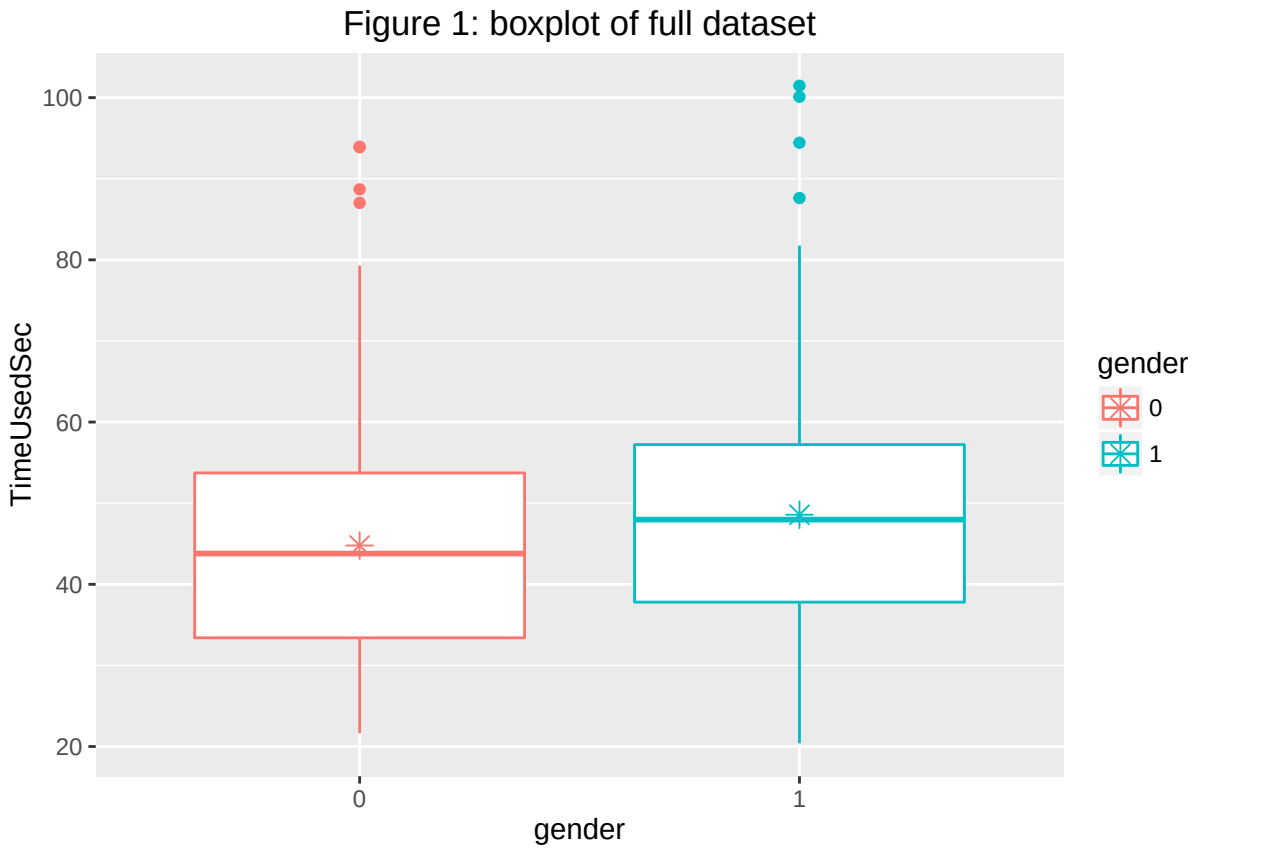

Game Length?
Short (25 Seconds)
Intermediate (45 Seconds)
Long (65 Seconds)
No limit
Match Proximity?
Exact
Small
Medium
Large
Number of Shapes?
15
18
21
24
Matching Scheme?
Shape, all same color
Shape, different colors
Color, same shape
Both shape and color
Show Timer?
Yes
No
Store in Database?
Yes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

Figure 1:

Question 1: Use the code below to create a dotplot and boxplot of the data. Does it seem reasonable to use a two sample t-test for this data?

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




```
## Set figure size
```

Question 2: 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 _____ students in an introductory statistics class.

As the result above suggested, the p value for two-sided t-test performed above on the overall group is 0.00369. It suggests that on an alpha level of 0.05, the probability of obtaining a mean difference as extreme as 2.92(test statistics) or -2.92 is less than 0.05. Therefore, just based on the result from this sample, we may conclude that gender does have an effect on the play time of the game.

Conducting two hypothesis tests on a second study

Let's repeat our analysis on a new study, using the group_ID "MATH22015", which is a subset of the full dataset.

//////////////////##Description of the study time stamp/sample size

```
MATH22015<-group_data[group_data$groupID=="MATH22015",]
MAT_female <- MATH22015[MATH22015$gender == 0,]$TimeUsedSec
MAT_male <- MATH22015[MATH22015$gender == 1,]$TimeUsedSec
MAT_t_test <- t.test(MAT_male,MAT_female,alternative="two.sided")
```

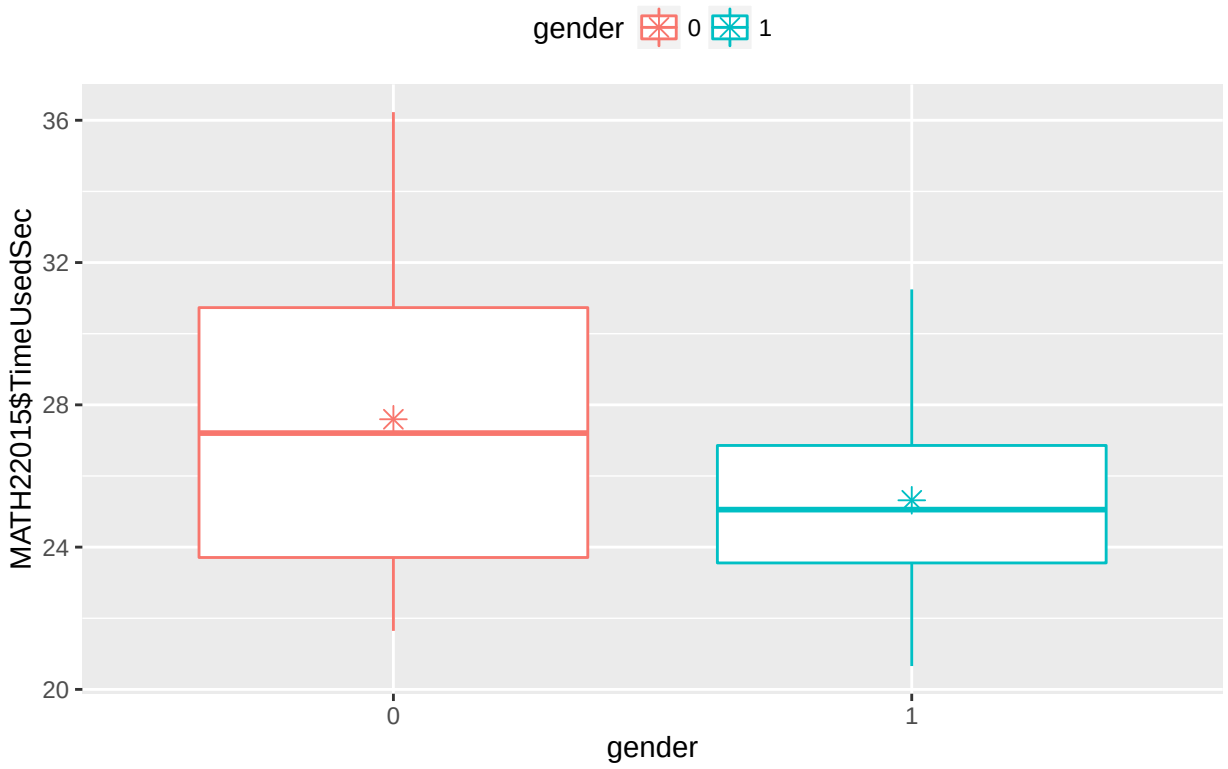
This study resulted in a p-value of 0.0776. Using an alpha level of 0.1, this study would have evidence to reject the null hypothesis and concluded that gender indeed makes a difference in play time of the game.

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

```
MATH22015$gender= as.factor(MATH22015$gender)

ggplot(data=MATH22015, aes(x=gender, y=MATH22015$TimeUsedSec)) + geom_boxplot()  + theme(legend.positio
  stat_summary(fun.y = mean, geom = "point", pch = 8, cex = 3)
```

Figure 2: boxplot of MATH22015 dataset



Now conducting hypothesis test on another with groupID = hjf190f14, let us conduct a second hypothesis test on this group.

```
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
```

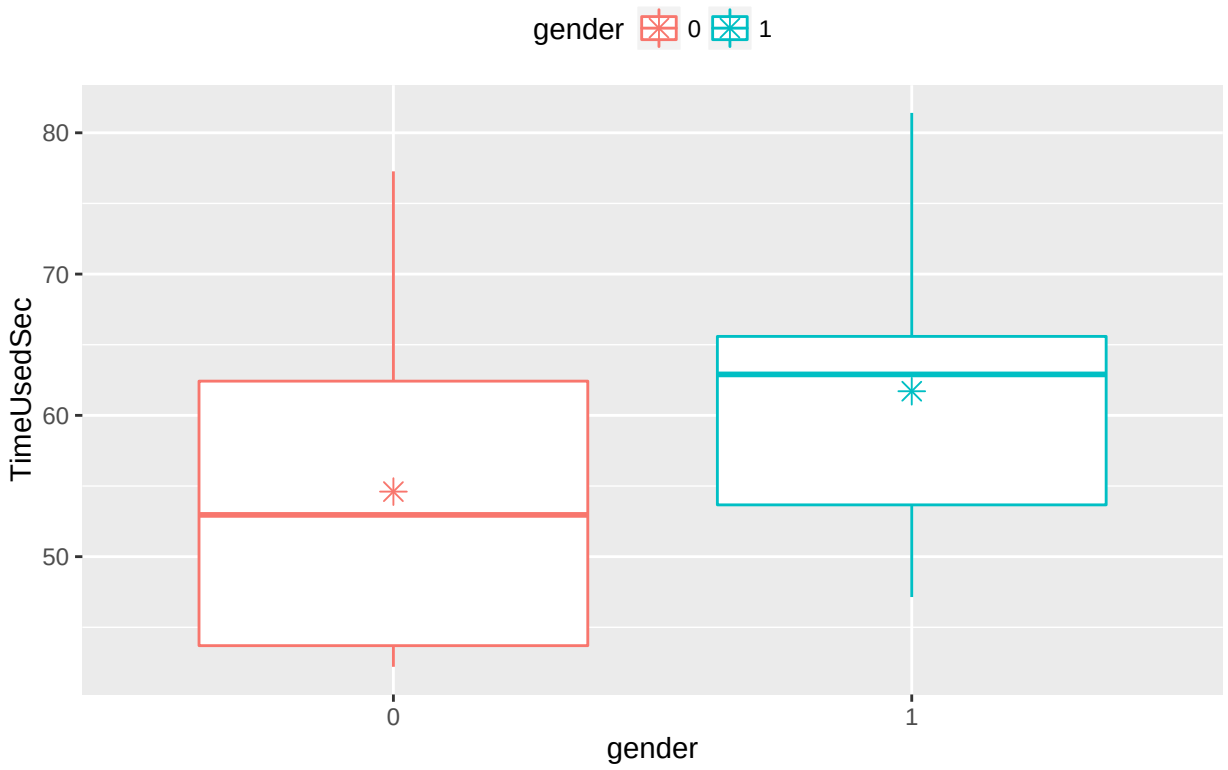
```
##
##  Welch Two Sample t-test
##
## data:  mth_male and mth_female
## t = 1.5173, df = 16.24, p-value = 0.1484
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -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)
ggplot(data=mth22602, aes(x=gender, y=TimeUsedSec)) + geom_boxplot() + theme(legend.position="top") + la
```

Figure 3: boxplot of mth22602 dataset



Question 3: 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?

//////////////////////interval of possible p-value range even when the H1 are

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

Part Two

Comparing multiple hypothesis tests

//////////////adding two graphs

```
group_data$gender <- as.factor(group_data$gender)
```

In Part 1 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 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="
      } else {
        boxplot(TimeUsedSec ~ gender,data=data1, col="red", main=paste("n=", dim(data1)[1], ", pval=", p
      }
    } else {
      boxplot(TimeUsedSec ~ gender,data=data1, main=paste("n=", dim(data1)[1], ", pval=", p), xlab="Gend
    }
  }
}

mtext("Figure 4", outer=TRUE,  cex=1, line=-1)
```

Figure 4

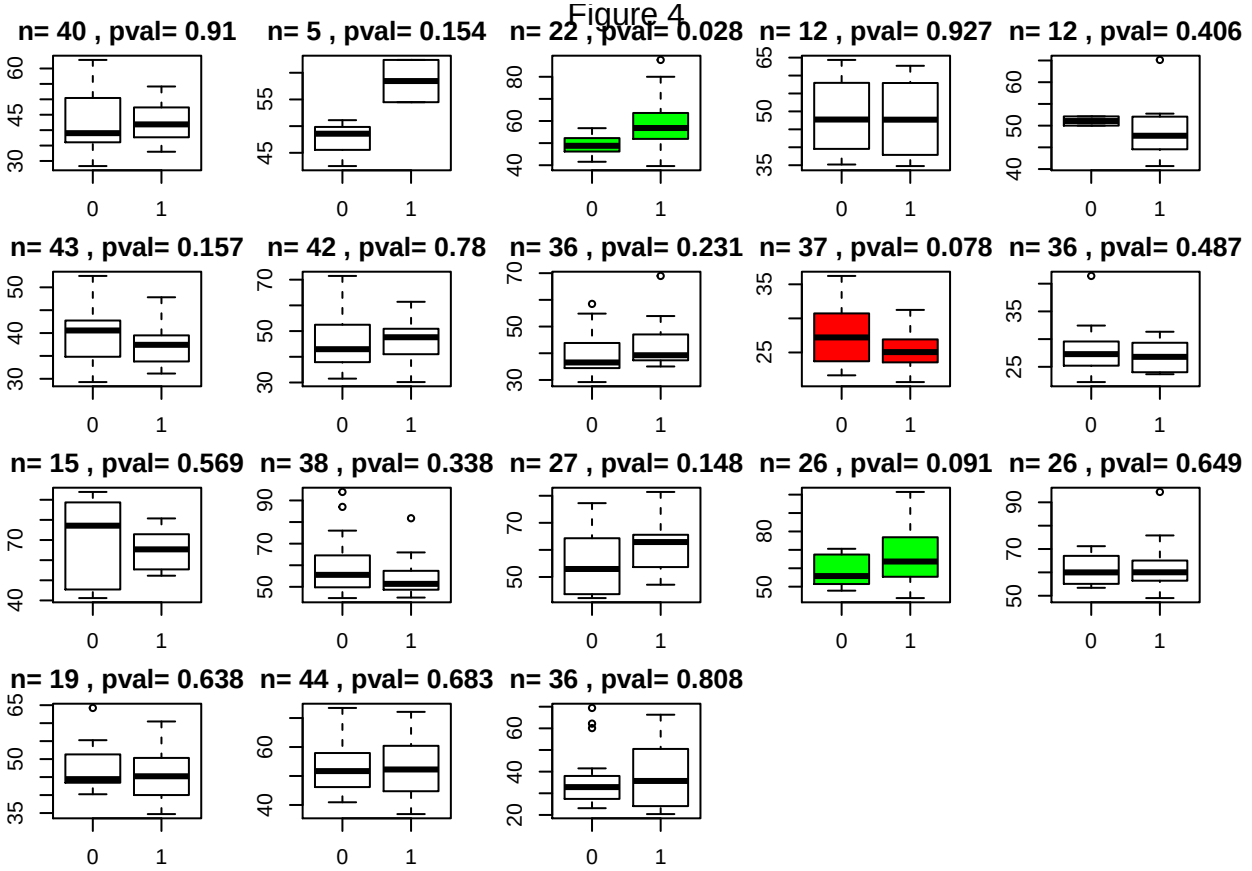


Figure 4 is a graphical representation of all 18 groups from the group dataset. The colored boxplots have p-value < 0.1.

Question 1: What is the range of p-values observed in these studies?

Question 2: How many groups had higher mean times for females? How many groups had higher mean times for males?

Question 3: 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 4: Why do the p-values differ?

Discussion: How credible 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.

What exactly is p-value?