Introduction to Statistical Inference (QTM 100 Lab)

Lecture 10: Analysis of Variance (ANOVA) and Two Sample t-Test

---

Justin Eloriaga — Emory University

Fall 2024

Gameplan

Preliminaries

Analysis of Variance

Pairwise Comparisons

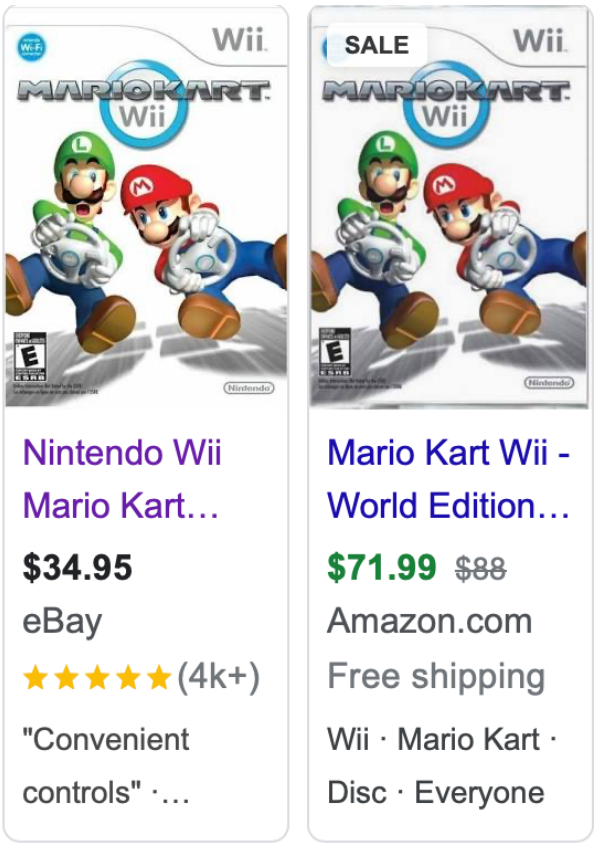
Two-Sample t-tests

Preliminaries

Overview

- We further explore analysis on numerical variables
- We now actually discuss some things more commonly used in empirical research
    - Two sample t-test
    - ANOVA

Mario Kart



- Remember Mario Kart? I'm sure at least 3/4 of the class played this before.
- We will explore a dataset that includes all auctions on ebay for a full week in October 2009.

Variable	Description
`ID`	Auction ID assigned by Ebay.
`duration`	Auction length, in days.
`n_bids`	Number of bids.
`cond`	Game condition, either new or used.
`start_pr`	Starting price of the auction.
`ship_pr`	Shipping price.
`total_pr`	Total price, which equals the auction price plus the shipping price.
`ship_sp`	Shipping speed or method.
`seller_rate`	The seller's rating on Ebay (number of positive ratings minus the number of negative ratings).
`stock_photo`	Whether or not the auction feature photo was a "stock" photo.
`wheels`	Number of Wii wheels included in the auction.
`title`	The title of the auctions.

Shipping Speed and the Price?

- **Research Question**: Do Mario Kart games that were shipped using more expensive methods cost more?
- Use the `hist()` and `boxplot()` functions to have an idea of the distributions of the total price (i.e. `total_pr`)
- It is unusual that a video game sells for an unusually high price!!! $100 f*** dollars for Mario Kart! Crazy!

Inspecting the Crazy Price

Let's begin by viewing the cases where the price was over $100

```
mariokart[mariokart$total_pr>100,]
```

We have two suspects!

```
            id duration n_bids cond start_pr ship_pr total_pr ship_sp seller_rate stock_photo wheels
20 110439174663        7     22 used     1.00   25.51   326.51  parcel         115          no      2
65 130335427560        3     27 used     6.95    4.00   118.50  parcel          41          no      0
                                                 title
20    Nintedo Wii Console Bundle Guitar Hero 5 Mario Kart
65 10 Nintendo Wii Games - MarioKart Wii, SpiderMan 3, etc
```

Let's take this out of our dataset and "clean" our data by not using these two datapoints any more. We can do this using the `subset()` function

```
# Create a new dataset
mkClean <- subset(mariokart, mariokart$total_pr<100)
```

Check using the `hist()` command if the outlying observations are gone. Then, we use use `mkClean` for all subsequent codes.

Deeper Dive on the Price

Look at the average cost for each shipping speed

```
tapply(mkClean$total_pr, mkClean$ship_sp, mean)
```

What about the number of observations for shipping speed?

```
table(mkClean$ship_sp)
```

```
firstClass      media      other     parcel   priority   standard    ups3Day  upsGround
  42.53182   51.02857   46.99667   47.49429   43.97783   46.56091   47.00000   52.80290

firstClass      media      other     parcel   priority   standard    ups3Day  upsGround
        22         14          3         14         23         33          1         31
```

Analysis of Variance

Creating `newship`

Let's create a new variable called `newship`. This variable can take four values

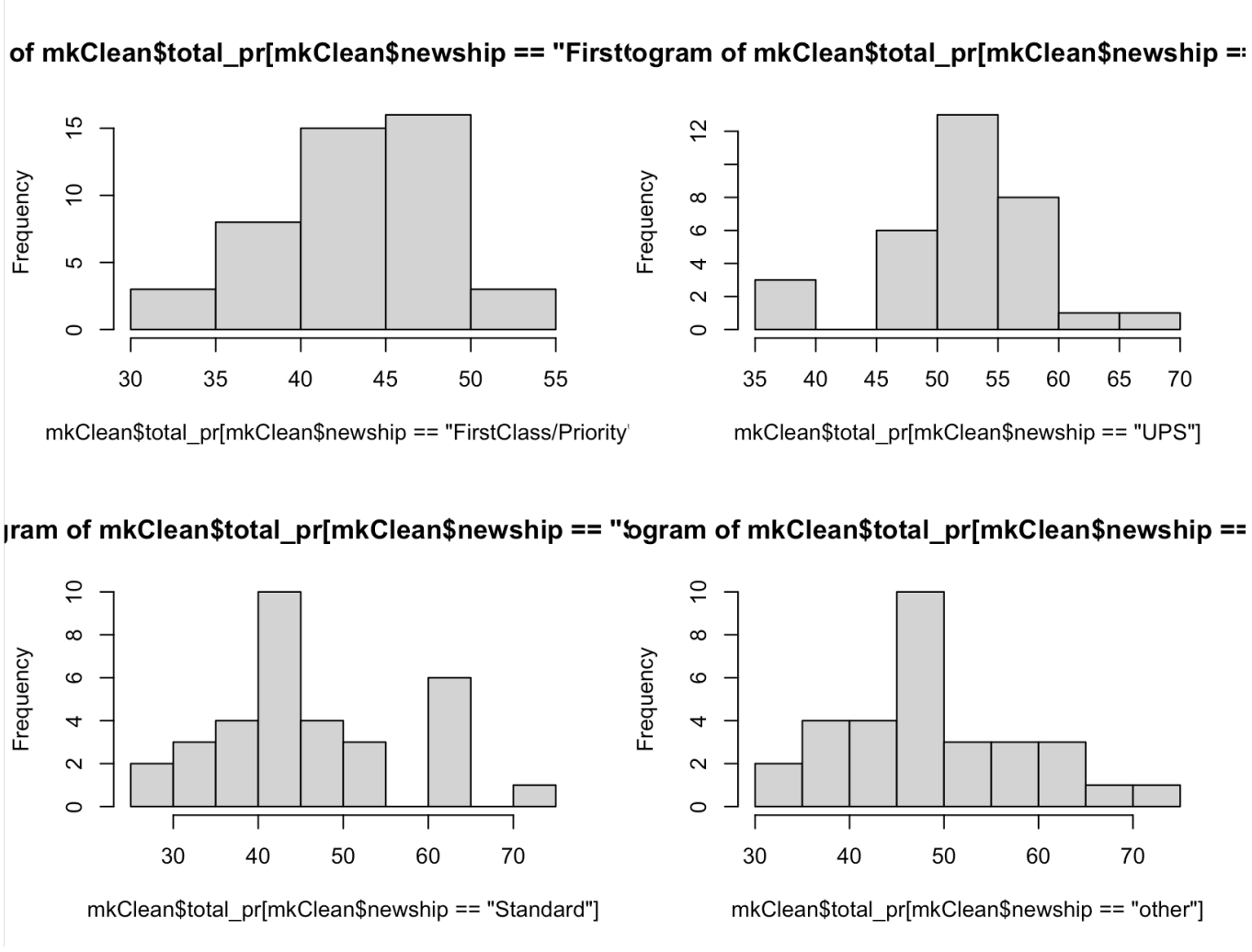
- FirstClass/Priority (i.e. `firstClass` and `priority`)
- UPS (i.e. `ups3Day` and `upsGround`)
- Standard (i.e. `standard`)
- other (i.e. `media`, `parcel`, and `other`)

Total Price and Shipping Type

Let's use `boxplot()` to visualize the relationship between `total_pr` and `newship`

Digging Deeper

Let's try and do individual histograms for each shipping type.



Using the `aov` command

Performing an ANOVA entails the use of the `aov` command. Performing an ANOVA allows us to see if the observed differences of price across different shipping methods is statistically significant.

```
anova.ship <- aov(mkClean$total_pr ~
mkClean$newship)
```

One can then use the `summary()` command to see the results. With $p < 0.001$, we reject H_0 and argue that $\exists$ at least one mean that differs from all the others.

Using the `aov` command

Performing an ANOVA entails the use of the `aov` command. Performing an ANOVA allows us to see if the observed differences of price across different shipping methods is statistically significant.

```
anova.ship <- aov(mkClean$total_pr ~
mkClean$newship)
```

One can then use the `summary()` command to see the results. With $p < 0.001$, we reject H_0 and argue that $\exists$ at least one mean that differs from all the others.

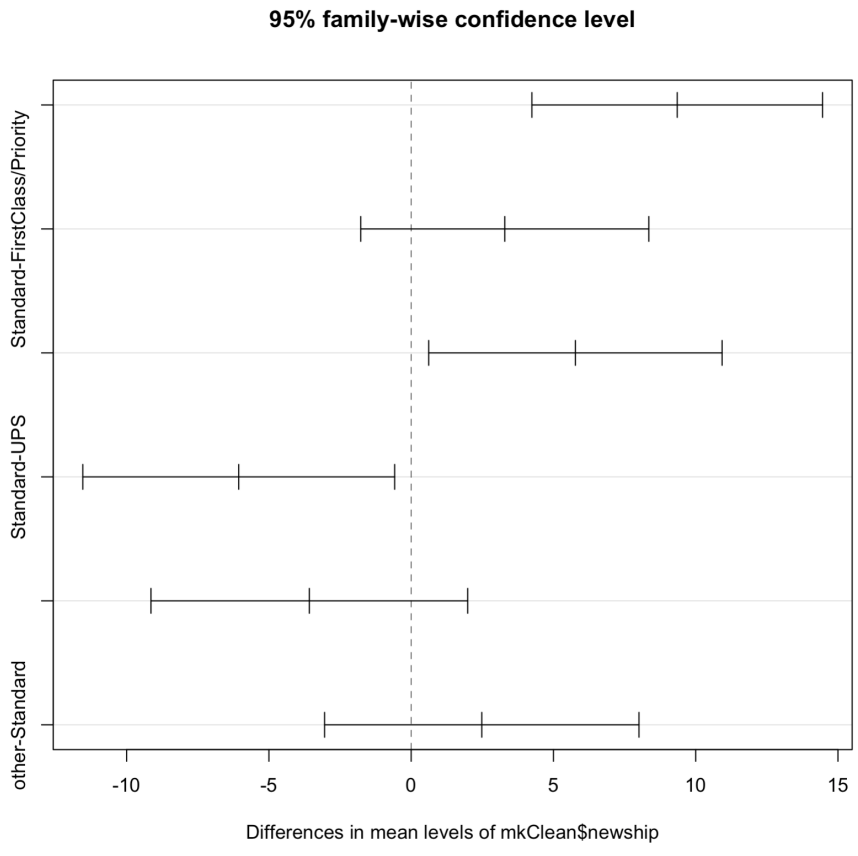
But..... which one differs?????

```
                Df Sum Sq Mean Sq F value   Pr(>F)
mkClean$newship   3   1746   582.1   8.071 5.45e-05 ***
Residuals       137   9882    72.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pairwise Comparisons

Tukey Test

To get better granularity on which specific shipping types are really statistically significant, we use a *Tukey test*.



- 3 significant differences in means
- Total price for items shipped with UPS is higher than items shipped with First-Class/Priority.
- Total price for items shipped with other methods is higher than items shipped with FirstClass/Priority.
- Items shipped with Standard shipping have a lower total price than items shipped with UPS.

Two-Sample t-tests

Two-sample t test

Suppose you just had *two* things to compare, UPS and Standard

- **Question**: Are the average prices the same across these two shipping types?
- You could technically use ANOVA, but its too complicated for such a simple question.
- Better to just use a two-sample t test!

First obtain the prices sold under each category

```
UPS <- mkClean$total_pr[mkClean$newship == 'UPS']
Standard <- mkClean$total_pr[mkClean$newship ==
'Standard']
```

Then, run the two-sample t test

```
t.test(UPS, Standard, var.equal=TRUE)
```

Results

```
        Two Sample t-test

data:  UPS and Standard
t = 2.4992, df = 63, p-value = 0.01507
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.214589 10.906718
sample estimates:
mean of x mean of y
 52.62156  46.56091
```

- There seems to be some significant difference in the price between UPS and Standard.