Lab 10: ANOVA and 2 Sample T-test

Overview

In this lab, we continue our analysis of numerical variables. We have already looked at inferences for single and related samples. How about two or more samples? We will learn how to perform one-way ANOVA and two-sample t-test to make inferences for numerical variables with multiple groups that need to be compared.

The Data: Mario Kart

This data set includes all auctions on Ebay for a full week in October 2009. Auctions were included in the data set if they satisfied a number of conditions. (1) They were included in a search for "wii mario kart" on ebay.com, (2) items were in the Video Games > Games > Nintendo Wii section of Ebay, (3) the listing was an auction and not exclusively a "Buy it Now" listing (sellers sometimes offer an optional higher price for a buyer to end bidding and win the auction immediately, which is an optional Buy it Now auction), (4) the item listed was the actual game, (5) the item was being sold from the US, (6) the item had at least one bidder, (7) there were no other items included in the auction with the exception of racing wheels, either generic or brand-name being acceptable, and (8) the auction did not end with a Buy It Now option. All prices are in US dollars.

This data set, mariokart, contains 12 variables:

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.

Accessing the data

Import the mariokart.csv data set into RStudio and get to know the data by identifying the data types, and examining the data values.

```
# Set working directory
setwd("~/Your Working Directory/")
```

```r
# Read in data
mariokart<-read.csv("mariokart.csv",header=T)

# Examine structure of mariokart
str(mariokart)

# Examine mariokart variables
summary(mariokart)
```

Getting started

For this analysis, we are interested in the relationship between shipping speed and price. Do mario kart games that were shipped using more expensive methods also cost more?

Begin by viewing the distribution of the variable of interest, `total_pr`.

```r
# Histogram of total price
hist(mariokart$total_pr)

# Boxplot of total price
boxplot(mariokart$total_pr)
```

It is clear that a couple of packages sold for an unusually high price. Inspect these observations closer by examining all packages that had a total selling price over $100.

```r
# View cases where total price was over 100 (prints all variables)
mariokart[mariokart$total_pr>100,]
```

It is clear that these two packages came with multiple games and so are not consistent with the Mario Kart price alone. Create a new data set that excludes these two packages, and review the distribution of `total_pr`.

```r
# Create subset
mkClean<-subset(mariokart,mariokart$total_pr<100)

# View histogram
hist(mkClean$total_pr)
```

Use the data set mkClean for all subsequent lab work.

Now, let us view the average cost for each shipping speed using our `tapply` function.

```r
# Look at the average cost for each shipping speed
tapply(mkClean$total_pr, mkClean$ship_sp, mean)

# Look at the number of observations for each shipping speed
table(mkClean$ship_sp)
```

When we inspect the data, we can see several categories with very few observations. In order to ease our analysis, we will combine them into four groups: First Class/Priority, UPS, Standard, and other. You should recognize the following code from our earlier lab manual on recoding factor variables.

```r
#Create new variable with less shipping categories
mkClean$newship <- factor(NA, levels= c("FirstClass/Priority", "UPS", "Standard", "other"))

#Assign each sale to its new category
mkClean$newship[mkClean$ship_sp=="firstClass" |
                  mkClean$ship_sp=="priority"] <- "FirstClass/Priority"

mkClean$newship[mkClean$ship_sp=="ups3Day" | mkClean$ship_sp=="upsGround"] <- "UPS"

mkClean$newship[mkClean$ship_sp=="media" |
                  mkClean$ship_sp=="parcel" |
                  mkClean$ship_sp=="other"] <- "other"

mkClean$newship[mkClean$ship_sp=="standard"] <- "Standard"

#Verify recoding
table(mkClean$newship,mkClean$ship_sp)
```

We already know that our outcome variable of interest, total price, is normally distributed at the population level. Now we need to verify whether it is normal for each group of interest.

```r
#Start by visualizing the relationship
boxplot(mkClean$total_pr ~ mkClean$newship,
        main = 'Total Price by Shipping Carrier',
        xlab = 'Shipping Carrier',
        ylab = 'Total Price ($)')

#Inspect the average total price for each new shipping group
tapply(mkClean$total_pr, mkClean$newship, mean)

#Histogram for first class and priority
hist(mkClean$total_pr[mkClean$newship=="FirstClass/Priority"])

#Histogram for UPS
hist(mkClean$total_pr[mkClean$newship=="UPS"])

#Histogram for Standard
hist(mkClean$total_pr[mkClean$newship=="Standard"])

#Histogram for other
hist(mkClean$total_pr[mkClean$newship=="other"])
```

We can see that it is not perfectly normal for all groups, but the distribution is acceptable overall. The average prices seem somewhat similar across groups, with `UPS` having the highest average price (52.62) and `FirstClass/Priority` having the lowest average price (43.27).

We use the `aov` function to model a numerical response (`total_pr`) by the categorical grouping (`newship`) variable to see if our observed differences are statistically significant. Note that we are actually storing the results of our `aov` function in a new R object (`anova.ship`). It means that we can reaccess the results later and use them for other R functions. The `summary` function here pulls the important numbers out of our model so that we can view them.

```r
# Conduct ANOVA
anova.ship <- aov(mkClean$total_pr~mkClean$newship)

# View results
summary(anova.ship)
```

With `p < 0.001`, we reject the null hypothesis and conclude that at least one mean differs from the others. It indicates that shipping type is associated with `price.` Without further analyses, we cannot tell which shipping speeds have significantly different prices.

Pairwise comparisons

When the null hypothesis is rejected in ANOVA, it is of interest to determine precisely which pairs of means differ. It results in multiple pairwise tests. When conducting multiple tests, special statistical techniques are used to control the overall Type I error rate. There are many methods available, but we only present the Tukey method for simplicity.

```r
#Perform Tukey Test
TukeyHSD(anova.ship)

#Visually inspect results
plot(TukeyHSD(anova.ship))
```

The results provide an estimate for the difference in means, the lower and upper bounds for the confidence interval of the difference of means, and a p-value corresponding to the null hypothesis that the two means are equal. Of the six pairwise comparisons, three show significant differences in the means. In the plot, pairwise comparisons that rejected the null hypothesis can be seen as confidence intervals that do not intersect with 0 (the dotted line in the middle), representing Ho: mean1 - mean2 = 0. Note that the "difference" is calculated by subtracting the mean of the second item listed from the first. So if the number is positive, this means that Item One > Item Two. If it is negative, it means that Item One < Item Two.

Our significant results were:

1. The total price for items shipped with UPS is significantly higher than items shipped with First-Class/Priority.
2. The total price for items shipped with other methods is significantly higher than items shipped with FirstClass/Priority.
3. Items shipped with Standard shipping have a significantly lower average total price than items shipped with UPS.

Two sample t-tests

Let's imagine that we were given only two of the shipping speeds to compare, UPS and Standard, and we were asked to find if their average prices were the same. In the technical sense, we could perform a one-way ANOVA test to determine whether the average prices for the two shipping speeds were equal or not. However, we could easily see from our earlier ANOVA analysis that the ANOVA test outputs information that might be extraneous to answering our simple question. A much more straightforward test that we could perform is the two-sample t-test. Let's take a look at it.

```r
# Obtain total prices for UPS shipping
UPS <- mkClean$total_pr[mkClean$newship == 'UPS']

# Obtain total prices for standard shipping
Standard <- mkClean$total_pr[mkClean$newship == 'Standard']

# Perform two sample t-test with equal variances
t.test(UPS, Standard, var.equal=TRUE)

## Note to set unequal variances, we would set var.equal to FALSE
```

The test statistic is t = 2.50 with 63 degrees of freedom. At the $\alpha = 0.05$ level of significance, we reject the null hypothesis that the average prices for UPS and Standard shipping are equal (p = 0.02); we conclude that average prices are significantly different. We are 95% confident that the true average price difference is in the interval of 1.21 to 10.91. This interval does not contain zero, so it is not plausible that the average prices are the same.

We observe that the p-values from the pairwise comparison of Standard-UPS from the ANOVA test and two-sample t-test are approximately the same. Therefore, when comparing two groups, you can use either the `aov()` or `t.test`, but with more than two groups, you should use the `aov()`.