An Application of Hypothesis Testing

ACTIVITY 36

The previous activity focused on confidence intervals and hypothesis testing. This activity is an application of the lessons from the previous activities as well as an introduction to the R scripting of the ANOVA test used for testing the means of a continuous variable across multiple categories.

Purpose

The purpose of this activity is to give you the opportunity to apply multivariate hypothesis testing in R

Learning Objective

Recognize statistical functions in R and diagnose the difference of multivariate data

Required Resources

- R, R Studio

Time Allocated

70 minutes out-of-class

Tasks

A. Testing Means

For this activity, we need the incident records located in the incidents schema within the class database. Go ahead and load the whole data set into R. *duration* is a column within the incidents website in the form HH:MM:SS, but is not recognized by R as a difference in time. This generates errors when attempting to make anything coherent out of the duration values. The simple solution is to regenerate the values by generating a loop that calculates the value duration using the incident start time and the incident last updated time.

```
for (i in 1:nrow(a)){a$dur[i]<-(as.numeric(a$lastupdatetime[i])
                                -as.numeric(a$starttime[i]))}
```

Or, an alternative way of generating a new column from existing data is

```
c$dur<-(as.numeric(c$lastupdatetime[1:nrow(c)]) -as.numeric(c$starttim
e[1:nrow(c)]))
```

Another alternative is to use the *lubridate* package `hms()function` to define the time value as seconds.

Now that the duration values are in seconds and sensible, we can look over the *incidents* data and determine which columns are discrete and can be used to develop subsets of the data. An appropriate choice is the number of lanes affected in the incidents which has heading *numlanesaffected*.

a. What are the different values within this column?

Go ahead and subset the data frame by the different values within *c$numlanesaffected* and generate the summary statistics accompanied by a boxplot. Normally full distribution diagnostics would be in order, but for the purposes of this activity we will negate this to focus on hypothesis testing.

b. Are there any noticeable differences in the means of the durations for the respective affected lanes?

Performing a two-sample hypothesis test requires the function `t.test()` and can be used to test if there are differences between the means of two samples within the same population. Since there are

more than two categories within the *numlanesaffected* column multiple tests are required to test if the means are the same or not the same.

c. Generate the two-sample tests for all sensible combinations, and discuss your findings by accepting or rejecting H_0. Use $\alpha = 0.05$.

Although R makes processing multiple hypothesis tests relatively easy and fast with large numbers of "categories", other methods exist that directly relate to hypothesis testing of multiple categories. In the case of the number of lanes affected by an incident, a statistical test of the means referred to as ANOVA, or Analysis of Variance exists which can test three or more means simultaneously to see if they are all equal or not all equal. The important distinction in the null and alternative hypothesis is that H_0: $\mu1=\mu2=\mu3$ and H_1: The means are not all equal. Thus, the ANOVA test only tests if there exists a difference between the means and does not decipher the difference between the means. The ANOVA test is run using the function `aov()` and using an ~ to separate the categorical variable from the continuous variable. But first we need to ensure that the categorical values are being read as such, which means we are required to force the structure from integer to a simple factor using the `as.factor()` function and the generation of a new column.

```
c$lanes<-as.factor(c$numlanesaffected)
```

Now that the lanes column has been established the `aov()` function can be called. Or,

```
duranova<-aov(dur~lanes, data=c)
summary(duranova)
```

Which returns the following table ($\alpha = 0.05$),

```
              Df     Sum Sq   Mean Sq F value    Pr(>F)
lanes          3 7.8564e+08 261881218  6.4838 0.0002589 ***
Residuals    520 2.1003e+10  40390039
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-value, 6.4838, is defined as a 6.4838 times greater variation between the groups than within the groups. Using an F-distribution table the upper critical value is between 2.6049 (df = ∞) and 2.6802 (df = 120) which our F-value and the corresponding Pr(>F) or adjusted p-value states is significant and thus we reject H_0 finding significant evidence that the means differ. However, which means differ still remains an issue and can be found using post-hoc tests when there is significant evidence to reject H_0.

A common post-hoc test for the ANOVA is the `TukeyHSD()`, and is used to expose means which are significantly different. The call is simple and is as follows,

```
TukeyHSD(duranova)
```

Returning the following table,

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = dur ~ lanes, data = c)
```

```
$lanes
         diff         lwr       upr     p adj
1-0  2124.282    607.5456  3641.019 0.0019031
2-0  3606.428    745.8330  6467.023 0.0067358
3-0  1264.833  -8241.2457 10770.912 0.9861203
2-1  1482.146  -1453.8869  4418.178 0.5626325
3-1  -859.449 -10388.5004  8669.602 0.9955727
3-2 -2341.595 -12174.4647  7491.276 0.9276890
```

This table describes the categorical variables 0-3 by providing an adjusted p-value describing the whether or not a relationship exists, and again using $\alpha = 0.05$ significant difference is found between 0,1 and 0,2.

Please annotate your script with the appropriate answers and comment on any differences or similiarities between question 3 and the ANOVA and Tukey HSD tests. As always, make sure the script executes.

B. Complete an ANOVA analysis for the duration of incidents by *incidenttypeid*, *detectiontypeid*, *impacttypeid*, and *incidentlevel*. Interpret your results and prepare prepare a write up complete with plots and descriptions of your analysis. Be sure to comment on your interpretation of the plots and the ANOVA output. Try to add some explanation as to why one category may have longer or shorter duration.

Deliverable

This is a discovery activity. That means that it puts together some things that you have learned to date. You don't need to expound too much on why it worked, but upload your R code (cleaned up and commented upon) to the class dropbox.

Assessment

This activity will be graded on the following criteria:

Activity 36 Grading Rubric

	Excellent (10)	Good (8)	Poor (6)	NONE
Script	Organized, complete, accurate and executes.	Missing minor parts, but executes and is otherwise organized and accurate.	Missing significant portions of the activity, unorganized, inaccurate, but executes.	Code does not execute
Annotation	Annotations are complete and describe what the code is accomplishing.	Some annotations are incomplete or do not describe what the code is accomplishing.	No annotations were provided.	Code does not execute
Commentary	Demonstrated active engagement with exploring the various options of the functions within the activity.	Demonstrated some minor changes to the instructors code.	No changes and thus no commentary that demonstrates active engagement with exploring the options within the activity.	Code does not execute

Student Notes