Testing Lab 2: p values

Data Science Team

Today’s Activity

Today’s activity will go more in-depth into p-values. Start a new RNotebook and code on your own as we go. Then, with a partner, complete the activities below. (Don’t forget to upload to Canvas.)

Activity 1

Using code from Testing Lab 1, simulate 1000 p-values and plot their histogram. Do they appear to be uniformly distributed? Why or why not?

Activity 2

For normally-distributed outcomes with unknown variance (typical in practice), a t-test is conducted.

(a) Download the movies dataset.

```r
movies <- read.csv("http://s3.amazonaws.com/dcwoods2717/movies.csv")
```

(b) Plot the density of rating – does it appear to be roughly normally distributed?

```r
t.test(movies$rating[movies$genre == "Comedy"], movies$rating[movies$genre != "Comedy"])
```

(c) Perform a `t.test()` of the hypothesis that comedies have the same average rating as other kinds of movies. Do you reject the null?

```r
t.test(movies$rating[movies$genre == "Comedy"], movies$rating[movies$genre != "Comedy"])$p.value
```

(d) Notice you can extract just the p-value like so:

```r
t.test(movies$rating[movies$genre == "Comedy"], movies$rating[movies$genre != "Comedy"])$p.value
```

(e) Run a simulation to see how the uncertainty level is affected by sample size. Specifically, consider the sample sizes $N = 25, 50, 75, 100, 125, 150, 175, 200$. Run the simulation 1000 times and record the results. Create a boxplot of the p-values (x will be the sample size, y will be the p-value).

There are many ways to do this. Here is one recipe:

(i) Start by making a store matrix with two columns. How many rows does the storage matrix need? Also, give the matrix two column names, N and p.

(ii) Create vector of sample sizes.

(iii) Write a double loop. The outer loop will run from 1 to 1000 simulations. The inner loop will run from 1 to 8, the length of the sample size vector.

(iv) Outside of the loop, create a counter variable called r for row and set it equal to 1. Inside of the loop, increment it when the sampling is complete:

```r
r <- r + 1
```

(v) Sample without replacement in the inner loop and conduct a p-value. The size will depend on j, the counter for the inner loop. The row of the store matrix you are placing N and the p-value into will be controlled by r, not i. (This will typically take 3-5 lines inside of the loop to conduct the test and store both the sample size and the p-value in the appropriate row, column).