TOOL

Quantifying Uncertainty With Permutation

To check for association between a numerical variable, Y, and a categorical variable, X, we often compare the distributions of Y across two different levels of X. If they are different, we conclude that Y has some association with X.

To assess the uncertainty associated with our result, we want to get a sense of how our conclusion would have changed if we had worked with a different random sample from the same population. You can use a resampling method called permutation to do this without collecting multiple samples. Permutation is a technique that randomly shuffles all of the observations of Y in your data without regard to which group of X to which they belong.

When you use permutation on your data, you begin with the assumption that the distributions of Y are indeed the same for each of the two levels of X. There are two possible outcomes:

1. Permuting the values of Y across the two groups of X leads to a difference in the sample means of Y that is similar to the sample difference that we observed in our data.
2. Permuting the values of Y across the two groups of X leads to a difference in the sample means of Y that is smaller than the sample difference that we observed in our data.

If you see the first outcome, permuting the data didn't meaningfully change the difference between groups, and suggests the Y is not different between the groups. If the second outcome occurs, permuting the data lessens the difference between the groups, and suggests that Y is different between the two groups.

Use this tool as a guide to using permutation to understand the uncertainty present when you compare two numerical variables.

Steps to Test a Hypothesis With Permutation in R

1. Write down your null and alternative hypotheses as well as your observed sample statistic.
2. Calculate the number of observations, `n`, in a data set, `dat`, that contains observations from the both groups of X, using the command `nrow(dat)`.
3. Use the command `perm.id = sample(1:n, size = n, replace = FALSE)` to randomly permute the observations.

4. Use the commands `dat.perm = dat` and `dat.perm$Y = dat$Y[perm.id]` to create a new data frame `dat.perm` where the values of the numerical variable Y are randomly permuted, but the values of the categorical variable X remain the same.
5. Use a for loop to draw many different permuted samples. Calculate the sample statistic of each sample and store it in the vector `sample_statistic`.
6. Create a histogram of sample statistics and calculate the p-value by calculating the proportion of observations in the histogram which are higher in magnitude than the observed sample statistic (difference of samples).
7. If the p-value is less than 0.05, reject the null hypothesis and conclude that the population distributions of Y are different across the two groups.

Using R With This Tool

The portions of this tool with a grey background are code text you can use to do the examples included in this tool. You can also modify them to use with your own data. In these examples:

- Commands are the lines of code that don't begin with a pound sign (#). Type these lines into R to carry out the command.
- Commented text begins with one pound sign and explains the lines of code.
- The example code output begins with two pound signs.

Data Set Information

The **Prestige** data set contains information about different job types in Canada in 1971, and is part of the **carData** package in R. In this data set, **education** refers to the average number of years of education the employees that hold that job have had, and **prestige** refers to the Pineo-Porter prestige value which measures the pride an employee has in their job.

You can load and view the data set with the following code:

```
install.packages("carData")                  # Install the carData package
library(carData)                             # Load the carData package
data(Prestige)                               # Load the Prestige data set
Prestige = Prestige[!is.na(Prestige$type),]  # Remove rows that do not
                              # contain information on the profession type
head(Prestige)                # View the first 6 rows of the data set

##                    education income women prestige census type
## gov.administrators     13.11  12351 11.16     68.8   1113 prof
## general.managers       12.26  25879  4.02     69.1   1130 prof
## accountants            12.77   9271 15.70     63.4   1171 prof
## purchasing.officers    11.42   8865  9.11     56.8   1175 prof
## chemists               14.62   8403 11.68     73.5   2111 prof
## physicists             15.64  11030  5.13     77.6   2113 prof
```

Example: Testing a Hypothesis With Permutation

The **Prestige** data set contains information on occupations and has grouped them into several groups including blue and white-collar workers. The boxplots and summary statistics of those two groups suggest that they are different. To understand the level of uncertainty associated with these summaries, you can compare the median income of workers in those two groups with permutation.

Step 1: Write down your null and alternative hypotheses and calculate the observed sample statistic.

Null hypothesis (H_0): Prestige score distributions of blue collar and white collar jobs are the same.

Alternative hypothesis (H_A): Prestige score distributions of blue collar and white collar jobs are different.

Sample statistic: Mean difference in prestige scores.

```
# Load the data set:

Prestige.wb = Prestige[Prestige$type != "prof",]
prestige.wc = Prestige.wb$prestige[Prestige.wb$type == "wc"]
prestige.bc = Prestige.wb$prestige[Prestige.wb$type == "bc"]

# Calculate the observed sample statistic:
obs_stat = mean(prestige.wc) - mean(prestige.bc)
obs_stat
## [1] 6.716206
```

Step 2: Determine the number of samples in this subset of the `Prestige` data set:

```
n = nrow(Prestige.wb)
```

Step 3: Use the sample function to randomly permute a vector of length `n`, which is the length of the `Prestige` data set.

```
perm.id = sample(1:n, size = n, replace = FALSE)
```

Step 4: Use the permuted vector to create a vector that contains the new, permuted `Prestige` data:

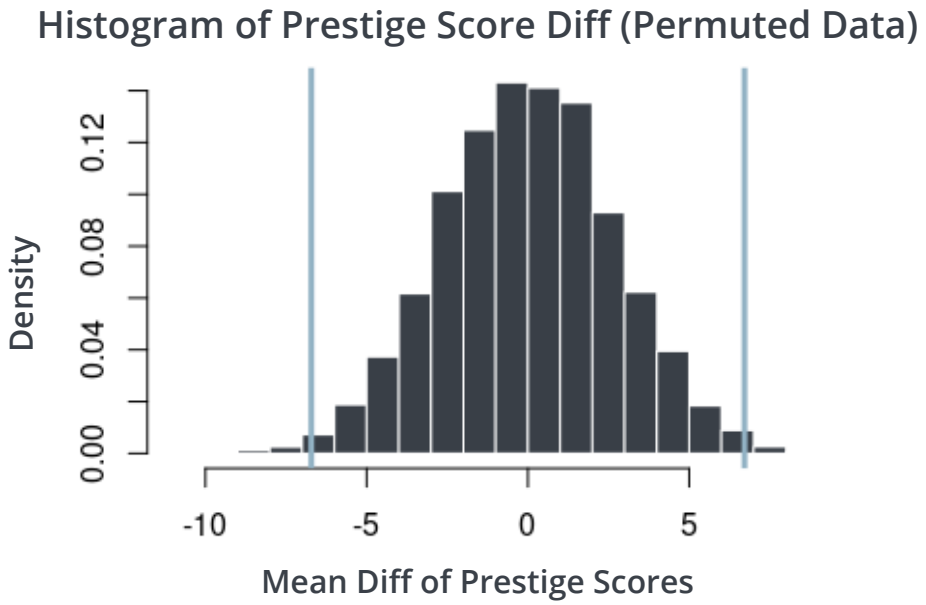
```
Prestige.wb$income = Prestige.wb$income[perm.id]
```

Step 5: Use a for loop to draw many permuted samples from the `Prestige` data set and store the sample statistics in the vector `store_mean_diff`:

```
set.seed(1) # Set seed for reproducibility
P = 10000 # Number of permuted samples
store_mean_diff = rep(0, P) # Vector in which to store the sample statistic

for (n in 1:P){
  Prestige.wb.perm = Prestige.wb
  Prestige.wb.perm$prestige = sample(Prestige.wb$prestige, replace = FALSE)
  prestige.perm.wc = Prestige.wb.perm$prestige[Prestige.wb.perm$type ==
    "wc"]
  prestige.perm.bc = Prestige.wb.perm$prestige[Prestige.wb.perm$type ==
    "bc"]

  store_mean_diff[n] = mean(prestige.perm.wc) - mean(prestige.perm.bc)
}
```

Step 6: Create a histogram of sample statistics and calculate the p-value of your observed difference in prestige scores between blue and white collar workers, given the distribution of the mean difference in the prestige scores between the two groups from the distribution of permuted data sets.

```
hist(store_mean_diff, breaks = 20, freq = FALSE, col = ecBlack, border =
     "white",
     xlab = "Mean Diff of Prestige Scores",
     main = "Histogram of Prestige Score Diff (Permuted Data)")
abline(v = obs_stat, col = skyBlue, lwd = 3)
mean(abs(store_mean_diff) >= abs(obs_stat)) # p-value

## [1] 0.0106

abline(v = -obs_stat, col = skyBlue, lwd = 3)
```



Step 7: Make a decision.

If the null hypothesis were true, the chances of seeing a 6.7 unit or larger difference in average prestige scores is 1%. This means the p-value is 0.01, and that we should reject our null hypothesis. Thus, we can conclude that the prestige-score distributions of blue and white collar jobs in the population are different.