

Lab 2: Intro to Hypothesis Testing - Permutation Tests

Objective:

Understand how to use R to perform permutation tests.

Definitions

- hypothesis testing, statistical significance
- null vs. alternative hypothesis
- test statistic, observed test statistic, P -value
- permutation test: permutation resample, permutation distribution

Introduction:

Recall the question posed in class on Tuesday: "If you could stop time and live forever in good health at a particular age, at what age would you like to live?"

Suppose we are interested in testing the claim that the average ideal age for women is *greater* than men. A random sample of 3 women and 3 men were asked this question resulting in the following responses:

	Women	Men
	49 42 38	29 38 50

The average response for the women is 43, and the average age response for the men is 39.

The difference in mean response between the women and men is $43 - 38 = 4$ years.

In the observed sample, the average response for women is greater than men, but this result could be due to random chance alone, rather than an actual difference between men and women. If there is no real difference, then the split of the 6 observations into the two groups is essentially random. We could have just as easily observed:

	Women	Men
	29 42 38	49 38 50

Now the difference in mean response between the women and men is -9.3 years.

So how do we determine if the result we actually observe provides evidence of a claim? We use probability and determine how easily pure random chance would produce a given result. This is the core idea of **statistical significance** or classical **hypothesis testing**, to calculate the probability that pure random chance would give an effect as extreme as that observed in the data, in the absence of any real effect. If that probability (referred to as the **P -value**) is small enough, we conclude that the data provide convincing evidence of a real effect.

Activities:

Getting Started: Navigate to your class folder structure. Within your "Labs" folder make a subfolder called "Lab2". Next, download the lab notebook for this lab from Blackboard and save it in your "Lab2" folder. You will again be working with the `FlightDelays` dataset on this lab. You should either re-download the data file into your "Lab2" folder from Blackboard, or just copy the file from your "Lab1" folder into the "Lab2" folder.

Within RStudio, navigate to your "Lab2" folder via the file browser in the lower right pane and then click "More > Set as working directory". Get set to write your observations and R commands in an R Markdown file by opening the "lab2_notebook.Rmd" file in RStudio. Remember to add your name as the author in line 3 of the document. For this lab, enter all of your commands into code chunks in the lab notebook. You can still experiment with code in an R script, if you want. To set up an R Script in RStudio, in the upper left corner click "File > New File > R script". A new tab should open up in the upper left pane of RStudio.

Ideal Age Example: If we let μ_W denote the true mean response for women to the question posed above, and we let μ_M denote the true mean response for men, then we set up in class on Tuesday the following **null** and **alternative hypotheses** to test the claim that the average ideal age for women is greater than men.

$$\begin{aligned} H_0 : \mu_W - \mu_M &= 0 \quad (\text{there is no difference in the mean response between women and men}) \\ \text{vs.} \\ H_A : \mu_W - \mu_M &> 0 \quad (\text{the mean response for women is greater than the mean response for men}) \end{aligned}$$

We also stated that the **test statistic** used to determine the result of the above test is given by the difference in the respective sample means. So, if we let \bar{X}_W denote the sample mean of the responses from the women, and \bar{X}_M denotes the sample mean for the men, the test statistic is $T = \bar{X}_W - \bar{X}_M$. Given the above data, the **observed test statistic** in this case is $t = \bar{x}_W - \bar{x}_M = 43 - 39 = 4$.

To determine whether or not the observed difference of 4 indicates a real difference, we will compute the associated P -value by working out all possible splits of the 6 observed responses into two groups and calculating how many produce a test statistic as large or larger than what was actually observed. Formally, the P -value is given by

$$P\text{-value} = P(T \geq 4).$$

Pause for Reflection #1:

How many different splits of the 6 numbers {29, 38, 38, 42, 49, 50} into two groups of 3 (ignoring ordering within each group) are possible? (*Hint:* How did we count the number of unordered things in probability?)

With your neighbor, write out all the possible splits in your lab notebook and calculate the corresponding test statistic for each of the possible splits. Then find the P -value by calculating the proportion of splits that resulted in a difference in mean response between women and men as large or larger than what was actually observed. Code has been started in the lab notebook for you to do these calculations with R.

Based on the P -value you find, do you think the true mean response from women is greater than men? Discuss this with your neighbor.

Statistically Significant: A result is considered **statistically significant** if it would rarely occur by chance. This begs the question, "how rare does the result need to be?" The answer: It depends on the context! But, for example, a P -value of 0.0002 would indicate that assuming the null hypothesis is true, the observed outcome would occur just 2 out of 10000 times by chance alone, which in most circumstances seems pretty rare and you would conclude that the evidence supports the alternative hypothesis.

Flight Delays Example: Recall the `FlightDelays` dataset from Lab 1, which contains information on 4029 departures of United Airlines and American Airlines from LaGuardia Airport during May and June 2009. In this lab, you will focus on the variable `Delay`, which gives the minutes that a flight was delayed (note that negative values indicate early departures). So, first load the dataset and then create a new object for easy reference to the `Delay` variable:

```
FlightDelays = read.csv("FlightDelays.csv", header = TRUE, sep = ",")
delay = FlightDelays$Delay
```

Recall from Lab 1, that a higher proportion of United flights were delayed more than 30 minutes. So we are going to test the claim that the mean delay for United flights is more than the mean delay for American flights. We can compute the average delay for the two airlines using the `tapply()` function in R. The `tapply()` function allows you to compute numeric summaries of

quantitative variables based on levels of a categorical variable. For instance, the following finds the sample mean flight delay length by airline in the `FlightDelays` dataset:

```
tapply(delay, FlightDelays$Carrier, mean)
```

```
##           AA           UA  
## 10.09738    15.98308
```

The mean delay for the sample of United flights is $\bar{x}_U = 15.98$ and the mean delay for American flights is $\bar{x}_A = 10.10$. The sample means are clearly different, but the difference ($15.98 - 10.10 = 5.88$ min) could have arisen by chance. Can the difference *easily* be explained by chance alone? If not, we will conclude that there are genuine differences in the mean delay times for the two airlines.

Hypotheses: In order to perform a hypothesis test of the stated claim, let μ_U denote the true mean delay time for United flights, and let μ_A denote the true mean delay for American flights. We will use $T = \bar{X}_U - \bar{X}_A$ as the test statistic, with an observed value of $t = 5.88$ min.

Pause for Reflection #2:

In your lab notebook, write the null and alternative hypotheses to test the claim that United flights have a longer mean delay than American flights. Note that you can use LaTeX in R Markdown files, which will help you typeset the notation used in stating the hypotheses. The syntax has been provided in the lab notebook.

Permutation Resampling: Suppose there really is no difference in the mean delay between the two airlines. Then the 4029 observed delay times come from a single population, the way they were divided into two groups (by labeling some as American flights and others as United) is essentially random, and any other division is equally likely. We could proceed, as in the ideal age example, calculating the difference in means for *every* possible way to split the data into two samples. However, that would entail looking at the number of ways to choose 1123 objects (the number of United flights in the dataset) from a total of 4029 objects (the total number of observations). This number is *astronomical*! Instead, we use sampling.

We create a **permutation resample** by randomly drawing $m = 1123$ observations *without* replacement from the pooled data to be one sample (the United flights), leaving the remaining $n = 2906$ observations to be the second sample (the American flights). We then calculate the test statistic for the new samples. Repeating this process many times (1000 or more), we can then calculate the P -value by finding the proportion of times the resulting test statistic equals or exceeds the original observed test statistic.

The distribution of the test statistic across all permutation resamples is the **permutation distribution**. This may be exact (i.e., calculated exhaustively as in the ideal age example) or approximate (i.e., implemented by sampling, as you will do next for the flight delays).

Two-sample Permutation Test: The following code walks you through performing a permutation test of the claim that the mean delay for United flights is longer than the mean delay for American flights.

Type each command below into the code chunk provided under the "Two-sample Permutation Test" heading in the lab notebook.

First, create an object to store the value of the observed test statistic:

```
observed = 15.98308 - 10.09738
```

To draw a permutation resample, you will draw a random sample of size 1123 from the numbers 1 through 4029 (there are 4029 observations total). The times corresponding to these positions in the `delay` vector you created earlier will be values for the United flights and the remaining ones for the American flights. The difference in means for this permutation will be stored in an object called `result`. This will be repeated many times.

```
N = 10^5 - 1          # number of times to repeat this process
result = numeric(N)   # space to save the random differences in each permutat.
for (i in 1:N)
{ # sample of size 1123, from 1 to 4029, without replacement
  index = sample(4029, size = 1123, replace = FALSE)
  result[i] = mean(delay[index]) - mean(delay[-index])
}
```

To analyze the results, first create a histogram of the (approximate) permutation distribution and add a vertical line at the observed test statistic.

```
hist(result, xlab = "xbarU - xbarA", main = "Permutation Distribution for delays")
abline(v = observed, col = "blue")      # add line at observed mean difference
```

Finally, compute the P -value by finding how many times a permutation resample produced a test statistic as large or larger than the observed value.

```
(sum(result >= observed) + 1)/(N + 1)    # P-value
```

The code snippet `result >= observed` results in a vector of TRUE's and FALSE's depending on whether or not the mean difference computed for a resample is greater than the observed mean difference. `sum(result >= observed)` then counts the number of TRUE's.

Pause for Reflection #3:

Consider the histogram and P -value that the above code produced. Is the result statistically significant? In other words, do you think there is a real difference in the mean delay times between United and American flights? Type a response in your lab notebook.

Choice of Test Statistic: In the examples above, we used the difference in means. We could have equally well used a variety of other test statistics, e.g., a difference in medians. It turns out, that if two statistics are monotonically related, i.e., one is always larger than the other, then the choice of one or the other as test statistic will result in exactly the same P -value. Let's explore this.

Repeat the permutation test of flight delays using (i) the difference in means, (ii) the mean of the United delay times, (iii) the sum of United delay times, and (iv) the difference in medians. You want to compute these statistics for the same permutation resamples, so find them all in the same `for` loop. The following code has already been added to the lab notebook:

```
result1 = numeric(N)      # space to save the differences in means
result2 = numeric(N)      # space to save the United means
result3 = numeric(N)      # space to save the sums of United delays
result4 = numeric(N)      # space to save the differences in medians
for (i in 1:N)
{ # sample of size 1123, from 1 to 4029, without replacement
  index = sample(4029, size = 1123, replace = FALSE)
  result1[i] = mean(delay[index]) - mean(delay[-index])
  result2[i] = mean(delay[index])
  result3[i] = sum(delay[index])
}
```

```
result4[i] = median(delay[index]) - median(delay[-index])
}
```

Pause for Reflection #4:

Compute and compare the P -values obtained for the four different test statistics used in your lab notebook. What do you observe?

Note: You will need to compute the corresponding observed test statistic for the three new test statistics. To do so, make use of the `tapply()` function.

Adding One: When computing the P -value for the permutation test, we add one to both the numerator and denominator. This corresponds to including the original data as an extra resample.

Pause for Reflection #5:

Discuss with your neighbor why you should add one, and include the original data, when computing the P -value. Record your thoughts in your lab notebook.

One- and Two-sided Tests: In the flight delays example, we had an initial hunch that United flights had a longer mean delay than American, so we performed a one-sided permutation test for a claim of "increase". However, we could have also tested the claim as a statement of "decrease", i.e., that American flights have a shorter mean delay than United. This would still be a one-sided test.

Instead of performing a one-sided test altogether though, we could have also performed a two-sided test, which would simply be a test of no difference, not claiming that one airline's mean delay time is longer or shorter than the other. When performing a two-sided permutation test, we calculate both one-sided P -values, multiply the smaller by 2, and if necessary round down to 1.0.

Two-sided P -values are the default in statistical practice: you should perform a two-sided test unless there is a clear reason to pick a one-sided alternative hypothesis. It is not fair to look at the data before deciding to use a one-sided hypothesis.

Pause for Reflection #6:

Consider the one-sided permutation test for a claim of "decrease" in the flight delays example. Write down in your lab notebook what the alternative hypothesis would be in this case and describe how you would calculate the P -value. Also, comment on why we multiply by 2 when calculating a two-sided P -value.

Lab 2: Intro to Hypothesis Testing - Permutation Tests is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.