Module 1: Hypothesis Tests

Girten

## Introduction to Hypothesis Tests

Hypotheses are written in terms of population parameters. What parameters have you used in previous hypothesis tests?

Hypothesis have been written in terms of means (), proportions (p), variances, IQR, medians, etc…..

**Definition:** The *null hypothesis*, denoted is a statement that corresponds to no real effect. The *alternative hypothesis*, denoted is a statement that there is some type of effect present in the population. Typically the alternative hypothesis takes one of three forms: less than, greater than, or not equal to.

**Definition:** A *test statistic* is a numerical function of the data whose value is used to determine the result of the test. The function itself will be denoted by , where X denotes the data. The test statistic is a random variable - its value will change based upon the sample collected. Once the test statistic has been calculated, it is referred to as the *observed test statistic* and is written in lower case .

**Definition:** The *p-value* is the probability that chance alone would produce a test statistic as extreme as the observed test statistics **if the null hypothesis were true**. The direction of extremity depends upon the alternative hypothesis tested.

### Hypothesis Test Example

Suppose scientists invent a new drug that supposedly will inhibit a mouse’s ability to run through a maze. Three mice are randomly chosen to receive the drug, and another 3 mice serve as a control. The time each mouse takes to go through a maze is measured in seconds. The results are:

|  |  |
| --- | --- |
| Drug | Control |
| 30 | 18 |
| 25 | 21 |
| 20 | 22 |
| = 25 | = 20.3333333 |

Define the parameters that can be used to compare typical maze times. Then, state the hypotheses that will be tested.

= mean maze time for mice receiving the experimental drug

= mean maze time for mice receiving the control drug

If we were to test this claim, the *classical* approach to a hypothesis test would use the t-distribution:

We can use the R function **t.test** to conduct the hypothesis test. The required inputs in the function are the data values in the two samples and the direction of the alternative.

t.test(x=drug, y=control, alternative = "greater")

##   
## Welch Two Sample t-test  
##   
## data: drug and control  
## t = 1.4924, df = 2.6731, p-value = 0.1215  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -3.073938 Inf  
## sample estimates:  
## mean of x mean of y   
## 25.00000 20.33333

#inputs are the two vectors containing the sample data followed by the direction of the alternative hypothesis

the p-value reported is 0.125, meaning that 12% of samples would result in data this extreme or more if there is no difference in the mean maze completion time for the two treatments (i.e. the null is true). Because this is not an unusual probability, the null hypothesis would not be rejected. There is not significant evidence to suggest a difference in mean maze completion time for the two treatments.

Questions: \* Is the test statistic intuitive? \* How are we assuming the null hypothesis is true?

There are some potential problems with using a t-test to test the claim. Assumptions have to be satisfied in order to use the t-distribution for inference:

* The populations from which the data is sampled are normally distributed.
* Large sample sizes can compensate for non-normality (the largeness depends on the degree of non-normality).
* The variability in the populations is approximately equal. If this assumption is not satisfied, there is an approximation for the degrees of freedom (Welch approximation in R; Satterthwaite in other programs).

In our example, the sample size was so small that it’s impossible to determine if the samples suggest normality. Ideally, we would like an alternative to the classical hypothesis test.  
\*\*\*\*\*

## Introduction to Permutation Tests

When a hypothesis test is conducted, the test statistic is calculated based upon the assumption that the null hypothesis is true. In the numerator of the test statistic, we compared the difference in the sample means to the difference in the population means. However, the null said the two population means were equal; that means we were comparing the difference in the sample means to a null value of 0.

Suppose we wanted to use the following test statistic to test the claim: . Unlike the previous test statistic, this test statistic does not have a known distribution. Without a distribution, there is no way to calculate a p-value. We need a way to determine if our observed test statistic is extreme.

**Definition:** The *null distribution* is the distribution of a test statistic if the null hypothesis is assumed true.

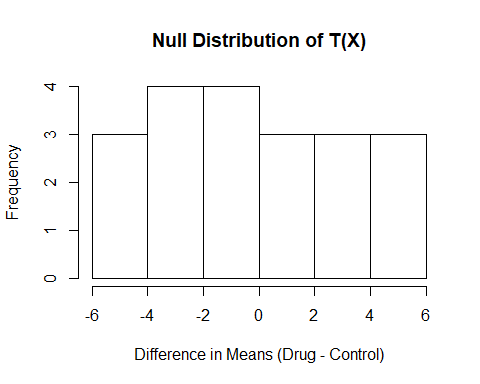
If the null hypothesis is true and the drug does not really impact maze completion, then the split of the 6 observed times into two groups is completely random. How many ways can the 6 times be split into two groups of 3?

The Excel file posted on Canvas contains all the permutations of the 6 maze times. This can be read into R, and we can find the sample mean for the times assigned to the drug and those assigned to the placebo.

maze\_permutations = read.csv(file = "C:/Users/sgirt/Documents/STA 394/maze+times.csv", header = TRUE, row.names = NULL)  
  
maze\_permutations

## Drug1 Drug2 Drug3 Control1 Control2 Control3  
## 1 18 20 21 22 25 30  
## 2 18 20 22 21 25 30  
## 3 18 20 25 21 22 30  
## 4 18 20 30 21 22 25  
## 5 18 22 21 20 25 30  
## 6 18 25 21 22 20 30  
## 7 18 30 21 22 25 20  
## 8 22 20 21 18 25 30  
## 9 25 20 21 22 18 30  
## 10 30 20 21 22 25 18  
## 11 18 22 25 20 21 30  
## 12 18 22 30 20 25 21  
## 13 18 25 30 22 20 21  
## 14 22 25 21 18 20 30  
## 15 22 30 21 18 25 20  
## 16 25 30 21 22 18 20  
## 17 22 20 25 18 21 30  
## 18 25 20 30 22 18 21  
## 19 22 20 30 18 25 21  
## 20 22 25 30 18 20 21

drug\_means = numeric(20) #creates a numeric vector of length 20  
control\_means = numeric(20)  
  
for (i in 1:20){  
 drug\_means[i] = mean(as.matrix(maze\_permutations[i, 1:3]))  
 control\_means[i] = mean(as.matrix(maze\_permutations[i, 4:6]))  
}  
  
simulated\_ts = drug\_means - control\_means  
  
# Create a histogram of the null distribution  
hist(simulated\_ts, main="Null Distribution of T(X)", xlab="Difference in Means (Drug - Control)")



mean(simulated\_ts)

## [1] 0

Now that the null distribution has been calculated, we can use it to find the p-value of the test. There are 20 values calulated in our null distribution; we need to find the proportion that are as extreme or more than the test statistic calculated from our observed data.

obs\_ts = mean(drug) - mean(control)  
  
#p-value  
sum(simulated\_ts >= obs\_ts) / length(simulated\_ts)

## [1] 0.15

### General Logic of Permutation Tests

1. Choose a test statistic that measures the effect of interest. This doesn’t have to be traditional test statistic; it can be something that is more intuitive like what was chosen in the mice example.
2. Using simulation, construct the distribution of this statistic under the assumption that the null hypothesis is true. The null distribution is a reference distribution much like we use the normal distribution, t-dist, f-dist, etc. In the previous example, we found the *exact* null distribution of listing out all possible combinations of the maze times. In general, this is not feasible and we rely on the approximate null distribution created via simulation.
3. Calculate the p-value of the test using the null distribution.

**Example:** Tennis elbow is thought to be aggravated by the impact experienced when hitting the ball. An investigator measured the force (Newtons) on the hand just after impact on a on-handed backhand drive for 7 advanced players and 9 intermediate players. Does the data provide evidence that the typical force after impact is greater for advanced tennis players?

Define the parameters usesd to compare typical force measurements for the two typese of player. Then, state the hypothesis to be tested:

= mean force after impact for advanced tennis players

= mean force after impact for intermediate players

To test this claim, the test statistic can be used.

If we were to construct the null distribution for T(X) for this data, there would be = 1.14410^{4} combinations. While we could exhaustively list these, it would take a considerable amount of time.

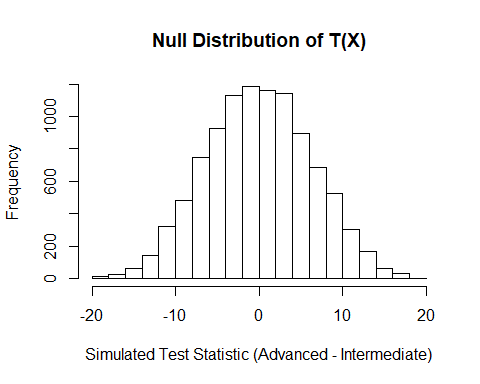
Instead, we can simulate the approximate null distribution of T(X):

* Combine the 16 force measurements. This assumes the null hypothesis is true - i.e. player skill has no impact on mean force after impact.
* Randomly sample 7 observations that are going to be assigned to the advanced players. The remaining 9 observations will be assigned to intermediate players.
* Calculate the mean force for the values assigned to the two groups of players and find their difference (advanced - intermediate).
* Repeat this process 9999 times.

What do you expect to see when a histogram of the null distribution is constructed?

We would expect the null distribution to pile up (center) around a value of 0 because under the assumption of the null we expect a value of 0 in the difference.

# combines the 16 observations into one vector  
force = c(advanced, intermediate)  
  
# create a storage vector for our simulated differences  
simulated\_ts = numeric(9999)  
  
for (i in 1:9999){  
 # create a vector of 7 position values that correspond to forces assignmed to the advanced group  
 index = sample(16, 7, replace = F)  
 simulated\_ts[i] = mean(force[index]) - mean(force[-index])  
}  
  
hist(simulated\_ts, main = "Null Distribution of T(X)", xlab = "Simulated Test Statistic (Advanced - Intermediate)")



mean(simulated\_ts)

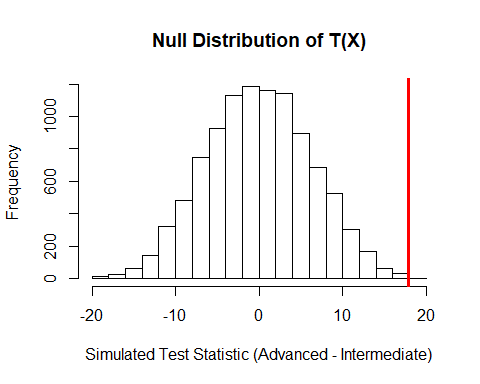
## [1] -0.001532741

sd(simulated\_ts)

## [1] 6.266036

In order to calculate the p-value in our hypothesis test, we need to find our observed test statistic from the original sample:

obs\_ts = mean(advanced) - mean(intermediate)  
  
hist(simulated\_ts, main = "Null Distribution of T(X)", xlab = "Simulated Test Statistic (Advanced - Intermediate)")  
  
# Add a vertical line to the histogram representing our observed test statistic  
abline(v = obs\_ts, col="red", lwd = 3)



Calculate the p-value:

sum(simulated\_ts >= obs\_ts + 1) / (9999 + 1)

## [1] 1e-04

The p-value reported was very small. The data collected is unlikely to occur if the null hypothesis is true. There is significanct evidence to conclude the mean force on impact for advanced tennis players is greater than that for intermediate tennis players.

### Implementation Details

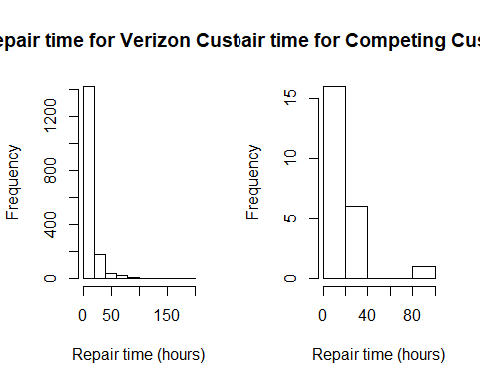
* Calculation of the p-value: When conducting the permutation test, we used 9999 resamples, but the p-value was calculated as p-value = . Why add 1 to the numerator & denominator? The +1 corresponds to including the original observed data as one of the resamples. It allows us to avoid obtaining a p-value of 0 from the permutation test.
* What about two-sided p-values? The authors recommend finding the p-values for both one-sided tests (less than, greater than) and doubling the smaller probability. But, this assumes the null distribution is roughly symmetric and that isn’t guaranteed. So, I recommend one of the following:
  + Calculate a one-sided p-value on each side of the null distribution (by reversing the order of subtraction) & sum the values.
  + If the test statistic is based on differences, construct the null distribution using the absolute value of the difference. The p-value will be the proportion of values greater than or equal to our observed difference.

**Example:** Verizon is the primary local telephone company (ILEC) for a large area of the easetern US. It is responsible for providing repair service for the customers of other companies known as competing carriers (CLEC). A sample of 1664 Verizon (ILEC) repair times is selected, and a sample of 23 competitor (CLEC) repair times is selected. Does the data provide convincing evidence the typical repair times differ for the two types of customers?

The dataset can be loaded from the resampledata library available on the [website](https://sites.google.com/site/ChiharaHesterberg) for the textbook:

Let’s investigate the distributions of repair times to determine the best measure for typical repair time. We can create side-by-side histograms of the repair times:

par(mfrow = c(1, 2))  
# par() sets graphical parameters  
#mfrow = c(rows, columns) tells how many rows &columns you want in the graphics window. Our line of code will create one row with two columns (or a side-by-side plot)  
  
hist(verizon, main = "Repair time for Verizon Customers", xlab = "Repair time (hours)")  
  
hist(comp, main = "Repair time for Competing Customers", xlab = "Repair time (hours)")



Both distributions are strongly right-skewed, which will inflate the value of the mean. In order to compare typical repair times, let’s compare the median repair times.

= median repair time for Verizon customers

= median repair time for competing customers

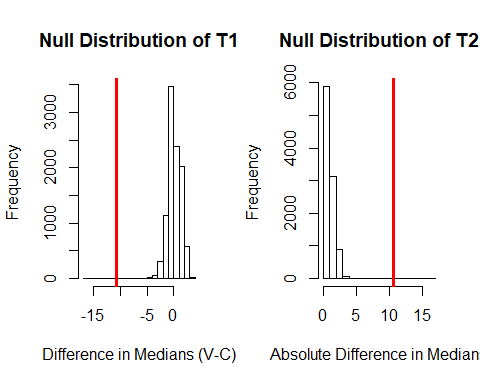
To test the two-sided claim, we can use either of the following test statistics:

* $T\_1 = \_V - \_C

To simulate the null distribution, we’ll calculate the test statistics in the same FOR loop to allow for the p-values to be compared.

$\abs{\tilde{M}\_V}$

repairs = c(verizon, comp)  
  
simulated\_ts1 = numeric(9999)  
simulated\_ts2 = numeric(9999)  
  
for ( i in 1:9999){  
 index = sample(1687, 1664, replace=F)  
 # This creates a sample of 1664 positions out of the 1687 possible to represent Verizon repair times.   
 simulated\_ts1[i] = median(repairs[index]) - median(repairs[-index])  
 simulated\_ts2[i] = abs(median(repairs[index]) - median(repairs[-index]))  
}  
  
obs\_ts1 = median(verizon) - median(comp)  
obs\_ts2 = abs(median(verizon) - median(comp))  
  
par(mfrow = c(1, 2))  
  
hist(simulated\_ts1, main = "Null Distribution of T1", xlab = "Difference in Medians (V-C)")  
abline(v = obs\_ts1, col = "red", lwd = 3)   
abline(v = -1\*obs\_ts1, col = "red", lwd = 3)  
  
hist(simulated\_ts2, main = "Null Distribution of T2", xlab = "Absolute Difference in Medians")  
abline(v = obs\_ts2, col = "red", lwd = 3)



#p-value1  
(sum(simulated\_ts1 <= obs\_ts1) + sum(simulated\_ts1 >= -1\*obs\_ts1) + 1) / (9999 + 1)

## [1] 0.0011

#p-value2  
(sum(simulated\_ts2 >= obs\_ts2) + 1) / 10000

## [1] 0.0011

Because the p-value is so small, there is evidence of a significant difference in the median repair times. The observed medians allow us to conclude the median repair time for Verizon customers is significantly less than that of their competitors.

* Observation on the two-sided p-value
  + If we used the method recommended by the textbook, we would have found the two-sided p-values and doubled the smallest. The greater than p-value would have been the smallest at 0.0001. Doubling that would give a p-value of 0.0002 for the test. The discrepancy between this p-value and those we calculated comes from the fact that the null distribution for the test statistic wasn’t symmetric.