Hypothesis Testing Using Randomization

Neba Nfonsang
University of Denver

Data

In [1]:

```
# view the first 10 rows of the data
head(Orange, 10)
```

Tree	age	circumference
1	118	30
1	484	58
1	664	87
1	1004	115
1	1231	120
1	1372	142
1	1582	145
2	118	33
2	484	69
2	664	111

Two-Sample Hypothesis Testing (Formulation)

- A hypothesis test is a procedure for making statistical inference or drawing conclusions about a population using sample data.

Research question

- We want to test whether there is a significant difference between the circumferences of trees from group 1 and trees from group 2.
- In other words, do trees from group 1 and group 2 have the same circumference averagely?

- Statistically, this question could be framed as: how likely do the two samples come from the same population? This is what we want to investigate with our hypothesis testing.
- Note: sample 1 and sample 2 are the observed samples and consist of circumferences of trees from group 1 and group 2 respectively. A group can be viewed as a population.

State the test statistics

- A test statistic is a sample characteristic used to make inference about the population parameter.
- The test statistic in this case is the **difference between the means** of circumferences of trees from group 1 and group 2. So, the observed test statistic would be the difference between the means of the two observed samples.

State the hypotheses

- The hypotheses are stated in terms of the populations not in terms of the samples. We rather use sample data to test the null hypothesis about the population.
- Null Hypothesis: There is no statistically significant difference between the circumferences of the trees from group 1 and group 2. In other words:
- There is no statistically significant difference between the means of the circumferences of trees from group 1 and group 2.
- mu1 = mu2 , or the mean of the circumferences of trees from group 1 = mean of the circumferences of trees from group 2.
- mu1 - mu2 = 0
- Alternative Hypothesis: There is a statistically significant difference between means of the circumferences of trees from group 1 and group 2.

How would the null hypothesis be tested?

- **Using randomization:** randomization is a resampling technique which is useful especially when testing for differences in means or proportions.
- With randomization we assume that the null hypothesis is true. To implement this in code, we could combine the data into a single larger "population" since the null hypothesis is that the samples came from the same population. Note that the null hypothesis is always tested not the alternative.
- Then, we randomly draw the samples from the population and calculate the mean differences and record the mean differences. Repeat the random sampling several times, say 10,000 times and record the mean difference for each iteration.
- The recorded mean differences constitute a randomization distribution (which is a kind of sampling distribution, and represents our null distribution or hypothesis).
- Then, we use the randomization distribution to find how likely the observed statistic or values more extreme than the observed statistic will occur under the null hypothesis (this is the p-value).
- In other words, the p-value is calculated as the proportion of simulated statistic (mean differences) equal to or more extreme than the observed statistic under the null hypothesis.

Conclusion and interpretation of results

- After computing the p-value, we compare the p-value to 0.05:

- If the p-value is less than 0.05 or 5%, we conclude that it is unlikely that the two samples came from the same population, hence we reject the null hypothesis. That is, there is a statistically significant difference between the means of the circumferences of the trees from group 1 and group 2.
- If the p-value is greater than 0.05, we conclude that it is more likely that the observed samples came from the same populations. That is, there is no statistically significant difference between the means of the tree circumferences.

Implement the Hypothesis Testing

In [2]:

```
# check the different categories of trees in the data
Orange$Tree
```

```
1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 3 4 4 4 4
4 4 4 5 5 5 5 5 5 5
```

In [3]:

```
# or you could check the unique categories of trees in the data
unique(Orange$Tree)
```

```
1 2 3 4 5
```

In [4]:

```
# sample 1: extract the circumferences of trees
# belonging to group 1
tree1.cir <-  Orange$circumference[Orange$Tree==1]
tree1.cir
```

```
30 58 87 115 120 142 145
```

In [5]:

```
# sample 2: extract the circumferences of trees
# belonging to group 2
tree2.cir <-  Orange$circumference[Orange$Tree==2]
tree2.cir
```

```
33   69   111   156   172   203   203
```

In [6]:

```
# length of sample 1
n1 <- length(tree1.cir)
n1
```

```
7
```

In [7]:

```
# length of sample 2
n2 <- length(tree2.cir)
n2
```

```
7
```

In [8]:

```
# combine the two samples to form the "population"
# under the null hypothesis
pop <- c(tree1.cir, tree2.cir)
pop
```

```
30   58   87   115   120   142   145   33   69   111   156   172   203   203
```

In [9]:

```
# length of observations in the population
n <- length(pop)
n
```

```
14
```

In [10]:

```
# sum of observations in the population
pop.tot = sum(pop)
pop.tot
```

```
1644
```

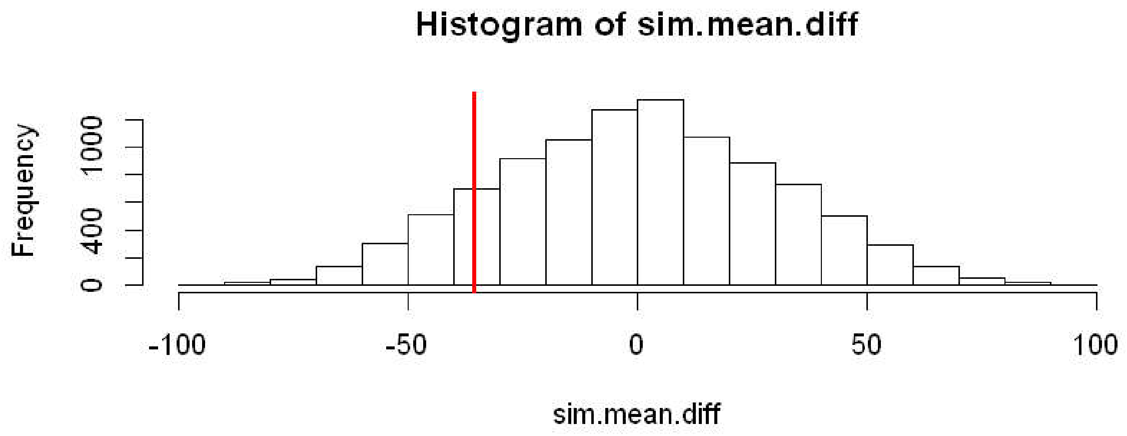
In [11]:

```
# observed mean difference
obs.mean.diff <- mean(tree1.cir) - mean(tree2.cir)
obs.mean.diff
```

```
-35.7142857142857
```

In [12]:

```
# conduct the randomization
set.seed(1234)

iterations = 10000


sim.mean.diff <- rep(NA, iterations)
for (i in 1:iterations){
    # sample from the population with equal propability
    # without replacement
    samp1 <- sample(pop, size = n1, replace = F)
    mean1 <- mean(samp1)
    mean2 <- (pop.tot - sum(samp1))/n2
    sim.mean.diff[i] <- mean1 - mean2

}
options(repr.plot.height=3)
hist(sim.mean.diff)
abline(v = obs.mean.diff, col="red", lw=3)
```



P - Value

- Compute the proportion of simulated mean differences equal to or more extreme than the observed value.
- We will use the two-sided p-value since our null hypothesis was non-directional.
- The two-sided p-value is calcauated as follows:

In [13]:

```
mean(abs(sim.mean.diff) >= abs(obs.mean.diff))
```

0.2535

- This proportion is 25.35%, which is too high, higher than 5%. So, it is more likely that the two samples came from the same population. That is, there is no statistically significant difference between the means of the circumferences of the trees.

Verify your Results with a Traditional T-Test

- We can use a traditional t-test to verify our results to see if our simulation worked well.

In [14]:

```
t.test(tree1.cir, tree2.cir)
```

```
	Welch Two Sample t-test

data:  tree1.cir and tree2.cir
t = -1.193, df = 10.327, p-value = 0.2596
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -102.13059   30.70202
sample estimates:
mean of x mean of y 
 99.57143 135.28571 
```

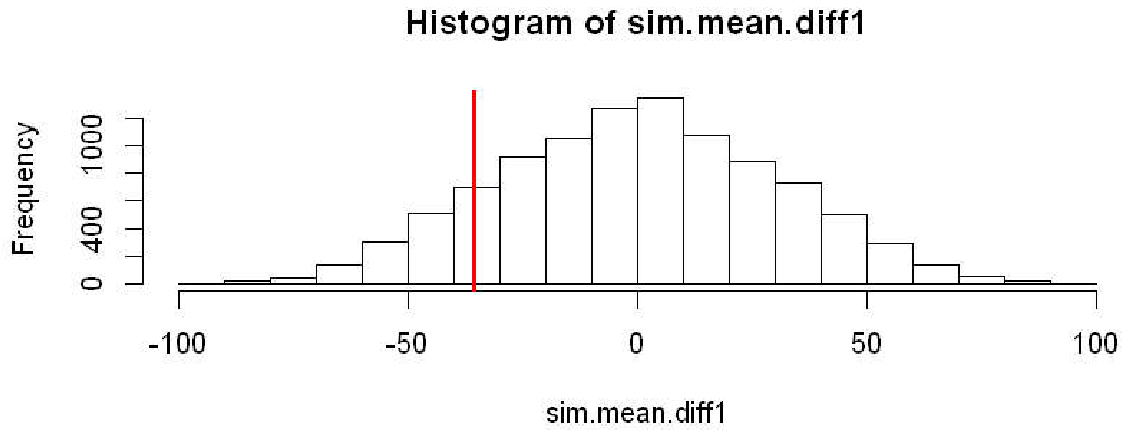
- The p-value from a traditional t-test is very similar to the p-value obtained from the simulation, and is greater than 0.05 (or 5%) so we conclude that there is no statistically significant difference between the means of the circumferences of trees. That is, the two samples came from the same population.

An Alternative Code for Randomization

In [15]:

```
set.seed(1234)
# conduct the randomization
iterations = 10000


sim.mean.diff1 <- rep(NA, iterations)
for (i in 1:iterations){

    # sample from the population with equal propability
    # without replacement, and we are sampling the index
    pop.ind <- 1:n
    ind1 <- sample(pop.ind, size=n1, replace = F)
    samp1 <- pop[ind1]
    samp2 <- pop[-(ind1)]
    sim.mean.diff1[i] <- mean(samp1) - mean(samp2)

}
options(repr.plot.height=3)
hist(sim.mean.diff1)
abline(v = obs.mean.diff, col="red", lw=3)

# p-value
mean(abs(sim.mean.diff1) >= abs(obs.mean.diff))
```

```
0.2535
```


Histogram of sim.mean.diff1
Frequency
sim.mean.diff1