Experiment Planning and Design

Lecture 3: Simple Statistical tools

Claus Aranha
caranha@cs.tsukuba.ac.jp

Department of Computer Science

April 28, 2015

Class Outline

In this class, we will learn some basic statistical tools and methods. These tools are useful to analyse the results of experiments.

- Statistical Inference
- Hypothesis Testing
- “R” Tutorial

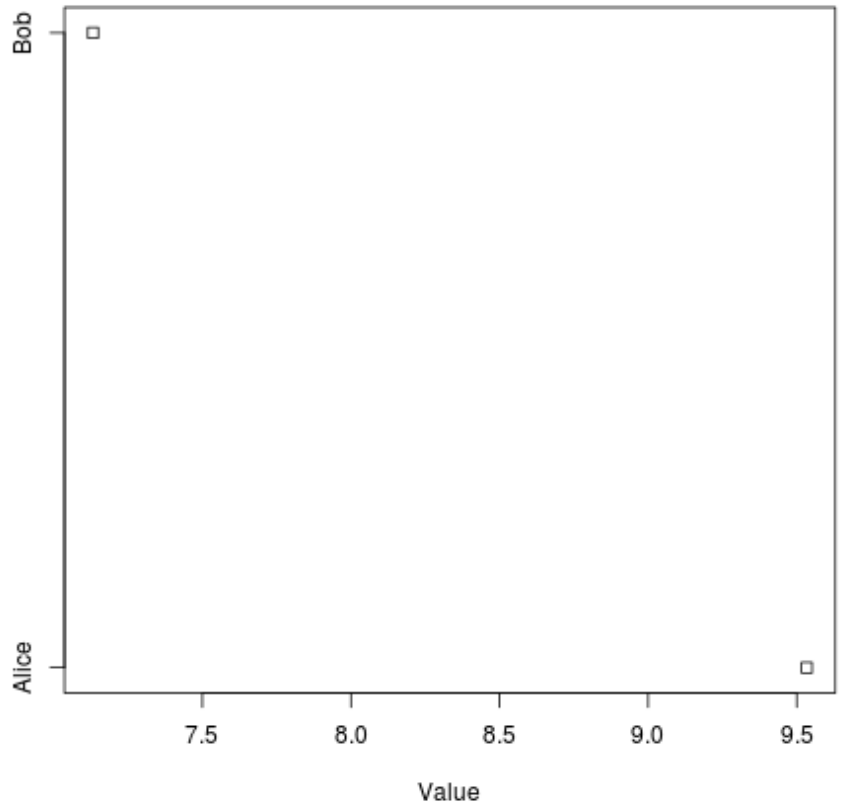
The Story of Alice and Bob

Alice and Bob are two scientists, trying to create a flying robot. Each has a different technology that they claim is the best one.

The Story of Alice and Bob (2)

Alice and Bob test their robots. They record the maximum heigh of the robot. After the test, Alice claims that her technology is the best.

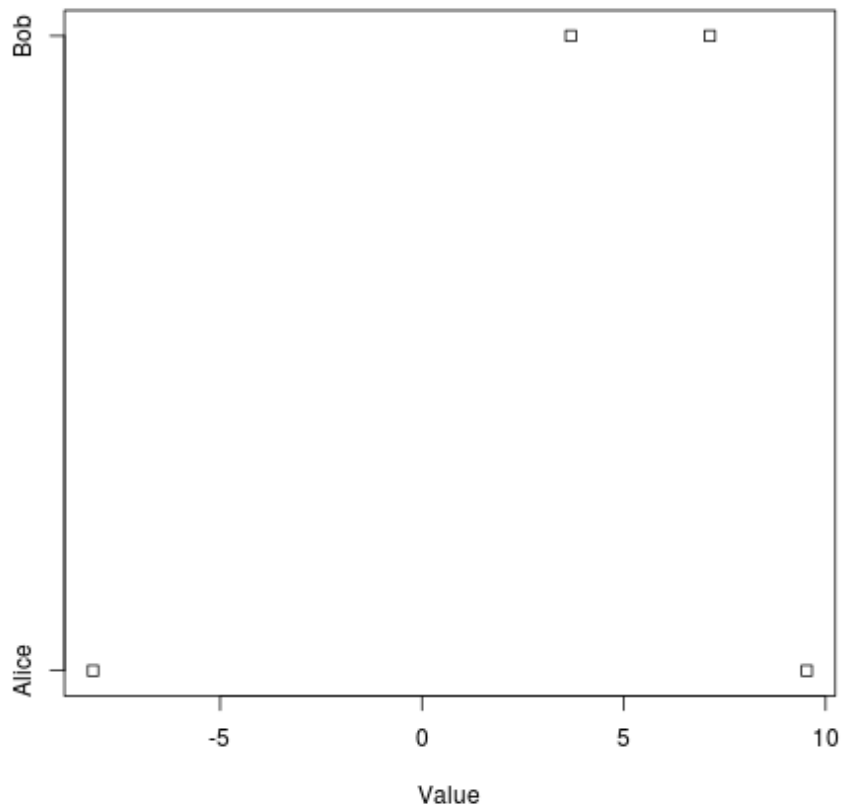
Alice	Bob
9.533777	7.132107



The Story of Alice and Bob (3)

Bob claims that the wind got in the way of his robot, and asks for a rematch. This time, Alice's robot Crashes! Bob is now claiming that his technology is safer and better.

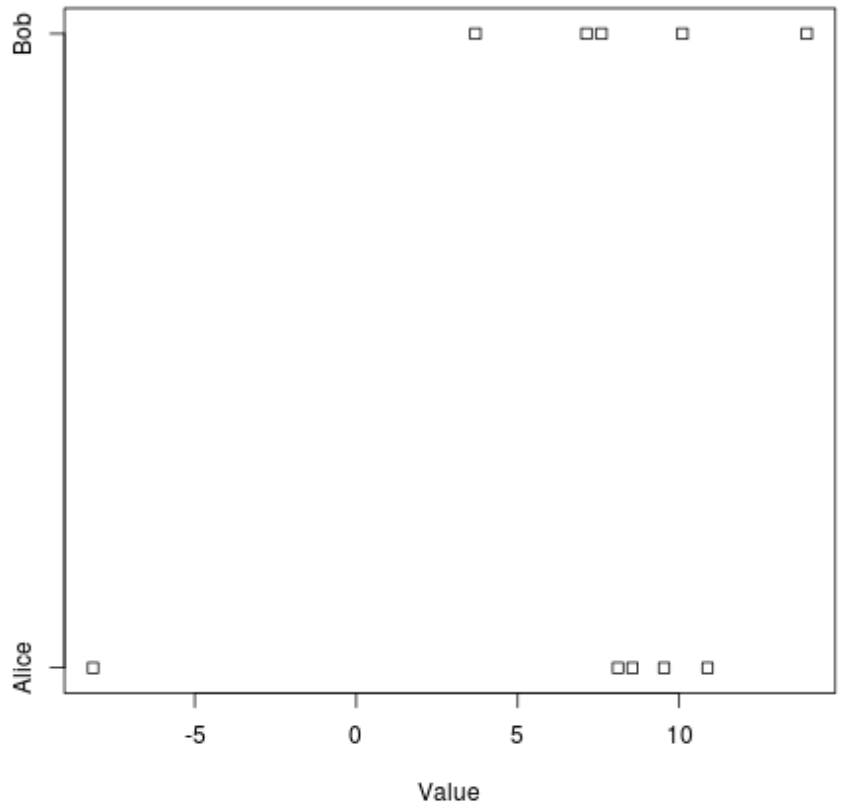
Alice	Bob
9.533777	7.132107
-8.149627	3.689753



The Story of Alice and Bob (4)

Alice and Bob do a few more experiments. After five repetitions, Bob claims that, on average, his robot will fly more than Alice's. So it proves that his technology is better.

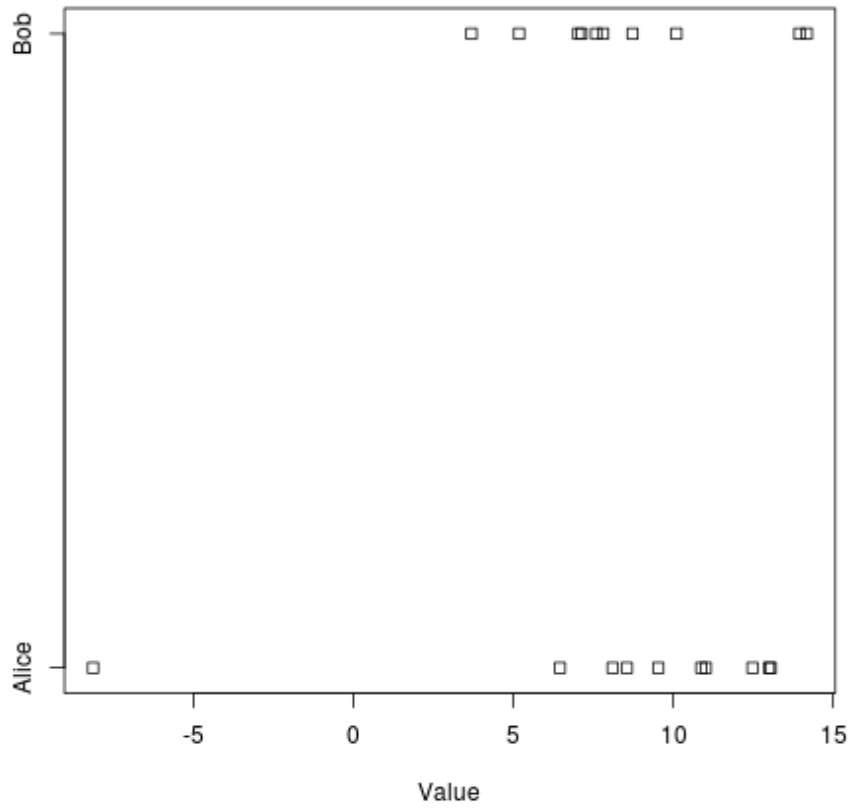
Alice	Bob
9.533777	7.132107
-8.149627	3.689753
8.55298	13.951185
8.105736	10.098746
10.879449	7.59397
Mean	Mean
5.784463	8.493152



The Story of Alice and Bob (5)

If we make a few more experiments, the difference between the means becomes really small. Is this the correct value? How do we know when to stop making experiments?

Alice	Bob
9.533777	7.132107
-8.149627	3.689753
8.55298	13.951185
8.105736	10.098746
10.879449	7.59397
12.482068	5.179808
12.992992	7.030863
11.022459	14.178277
6.454991	8.728171
13.055410	7.806157
Mean	Mean
8.493023	8.538904

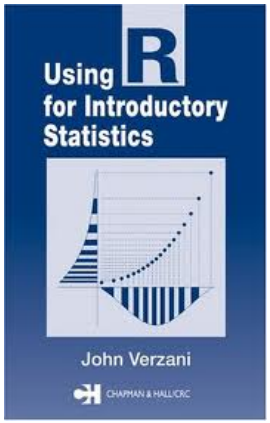


The Story of Alice and Bob – Conclusion

In reality, the values for the experiments of Alice and Bob came from the same distribution. This means that their technologies are equal, and any differences were only luck.

- Using statistical techniques, we can answer the questions:
  - Are these two groups of values REALLY different?
  - If they are different, HOW MUCH are they different?
- We will apply those techniques using the tool “R”

The statistical package “R”



- Open-Source (GNU) language for statistical computing;
- `http://www.r-project.org`;
- Developed since 1990
- “Using R for Introductory Statistics” by John Verzani

Getting Started

First, let's install the library from the "Using R" book.
(This can be used to install other libraries as well)

Install the package from the Internet

```
> install.packages("UsingR")
```

Load Library after installation

```
> library(UsingR)
```

Basic R Commands

R works as a commandline program. You enter commands and get their results. For example, we can use R as a simple calculator, by trying out the following commands, and pressing <ENTER>

```
> 2 + 2
```

```
> 4^8
```

```
> 4**8
```

Basic R commands (2)

Functions are very important for R. You can specify parameters for the functions, or use default values.

```
> factorial(10)
> sin(pi)

> log(27)
> log(27,3)
> log(base=3,27)

> squareroot(2)
> sqrt(-2)
```

In the last example, we got a nice error message! Error messages can be Errors or Warnings

Basic R commands (3)

Assigning Variables

```
> x = 3
> x <- 3
> x -> 3
```

Names can be letters, numbers and a period. Case is important!

```
> x = 3
> X = 4
> x + 1
```

Arrays in R

In statistical analysis, we usually work with many values: tables or arrays. For example, data sets or experiment results.

```
Creating an array in R:
> a = c(1,2,3,4)
> a

> b = c("alpha", "beta", "gama")
> b

Be careful! Each array can only have on type!
> c = c("one", 2, "three")
```

Functions and Arrays

Many functions in R accept arrays as data

```
> x = c{1,2,3,4,5,6}
> factorial(x)
> sqrt(x)
> log(x)
> x+x
> x*3
```

There are also many array specific functions

```
> sum(x)
> mean(x)
> len(x)

Also max(),min(),range(),sort(),cumsum(),diff() ...
```

POP QUIZ!

Calculate the Standard Error of this data:

10,11,12,13,12,11,10,11,12,13

Answer:

POP QUIZ!

Calculate the Standard Error of this data:

10,11,12,13,12,11,10,11,12,13

Answer:

```
c=(10,11,12,13,12,11,10,11,12,13)
error = sqrt(sum((mean(x) - x)**2)/length(x))
```

Accessing Arrays

We can use numbers or arrays to access particular numbers

```
> x = c(2,4,6,8,10,12)
> x[2]
> x[c(1,2,3)]
```

Some cool ways to access arrays

```
> 1:3
> x[1:3]
> x[(1:3)*2]
> x[3:1]

> x > 7
> x[x>7]
```

Listing and Saving your work

List all variables that exist in the environment

```
> ls()
or
> objects()
```

Saving the work done so far

```
> dump(ls(),file="variables.R")
> dump(c(x,y,result),file="xy.txt",append=TRUE)
```

Reading a file with R commands

```
> source("variables.R")
> source("xy.txt")
```

Getting Help

```
> help()                    # main help window
> help(``topic'')           # help for ``topic''
> help(mean)          # help for an specific function

> ??''topic''               # keyword search for ``topic''
> help.search("mean")

> help.start()              # Open help in browser

> example(mean)             # Executes an example for the functi
```

Questions before we move on?
Play a bit with what we have seen so far!

Data Analysis

Before you begin experimenting, take a hard, long look at the data.

First step of the experiment workflow

- By examining the data, we can learn many things about our problem, before we even do any experiments;
  - (Difficulty of the problem, dimensionality, desired values, hard cases, easy cases, problematic factors, etc)
- Data examination can be textual, or graphical;
- Very important for Optimization, Signal analysis, Field studies. (Important in other fields as well)

When to visualize data

- Before the experiment: Investigate dataset properties;
- After the experiment: Investigate results;

Data in R: Frames

Multivariate data in R is stored in data frames. Think of these frames as matrices, where each row is a different data point (experiment, sample), and each columns is a different parameter or factor.

```
> x = c{"Alice","Alice","Alice","Bob","Bob","Bob")
> y = c(8,-8,7,7,3,6)
> t = data.frame(x,y)
# Warning! X and Y must be of the same length!

> names(t) = c("Experimenter","Test Result")
# What does this do?

> z = c(20,30,20,10,40,20)
> t = data.frame(x,y,z)
```

Loading Frames from disk

- You can read a data frame into R by using the read.table() command.

```
> t = read.table(fnet.txt)
# Warning! If loading a big table, don't forget to
# Assign a variable :-)
```

- There are many parameters in this command, such as separator character, and whether to use headers or not.
- The default parameters expect a data file like this:

```
10 15 30
30 10 10
40 5  30
70 1  10
```

Accessing Data

- We can access an item in a Data frame using the notation: frame[row,column];

```
> AB = read.table("AliceBob.table",header = TRUE)
> AB[1,2]
> AB[,2]
> AB[1,]
> AB[1:10,2]
```

- We can also attach a frame, and access the columns by name;

```
> attach(AB)
> names(AB)
> Value
> AB[Researcher=="Alice",]
> AB[Value > 8,]

# Be careful! Attached variables don't change the main table!
> Value[3] = 5
> AB[3,] # Did not change!
```

Textual Analysis of the Data

Things to ask yourself

- What are the ranges of the factors?
- What are the means and the medians?
- What is the “shape” of the data?

This is important for both the problem data set, and for the analysis of the results!

Getting Data from the Data

```
## Loading the data and examining the columns
> Test = read.table("fnet.data", header = TRUE)
> names(Test)
> attach(Test)

## Finding the limits
> range(M)
> range(Depth)
> mean(M)
> median(M)
```

- When is the mean different from the median?
- “Half of all people are below the mean” - true or false?

More Data commands!

```
# Loading a smaller dataset: "kid.weights"
> library(UsingR)
> name(kid.weights)

# stem tells us the "shape" of the data
> stem(age); stem(height); stem(weight)

# quantiles of the data
> quantile(height)
> quantile(M)
> IQR(M) # Distance between 2nd and 3rd quartiles

# List of all "Factor Levels"
> table(Mo)
> summary(kid.weights)
```

One image is worth more than 1000 data points

We can use R to plot many different data plots!

- Central Tendency
- Modality
- Skew
- Distribution
- Outliers

Before we begin

R has a large array of “output devices”: to the screen, png files, pdf files, etc. The default is outputting to the screen, but that is not good for old computers.

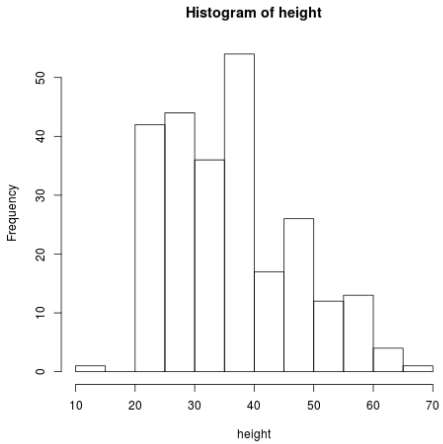
```
> png() # change the output device to png
> dev.cur() # list the current devices

# To get more information
> ?Devices # list many, many devices
```

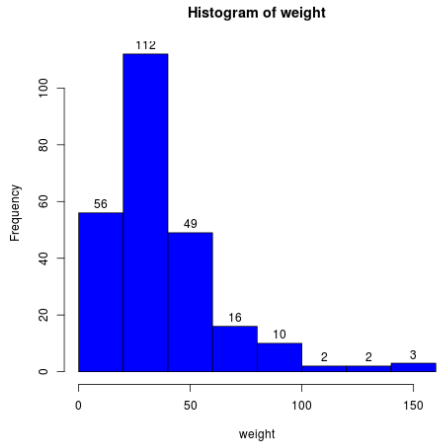
Data Visualization

Histograms

- Can show the probability distribution of a variable with great detail;
- Best for samples with large number of observations;
- Size of the “bins” can be adjusted to increase/decresar resolution;

```
> hist(heights)
> hist(weights, labels=TRUE, col="blue")
```



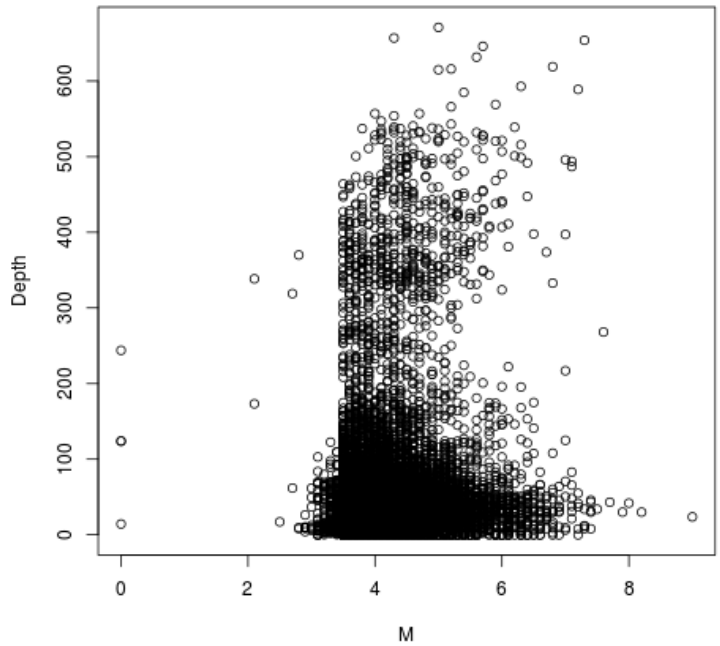


DataVisualization (2)

Scatterplots

For data with two variables, a scatterplot is a good place to begin the data analysis.

```
> png()
> quake = read.table("fnet.txt",
                     header=TRUE)
> attach(quake)
> plot(M,Depth)
> dev.off()
```



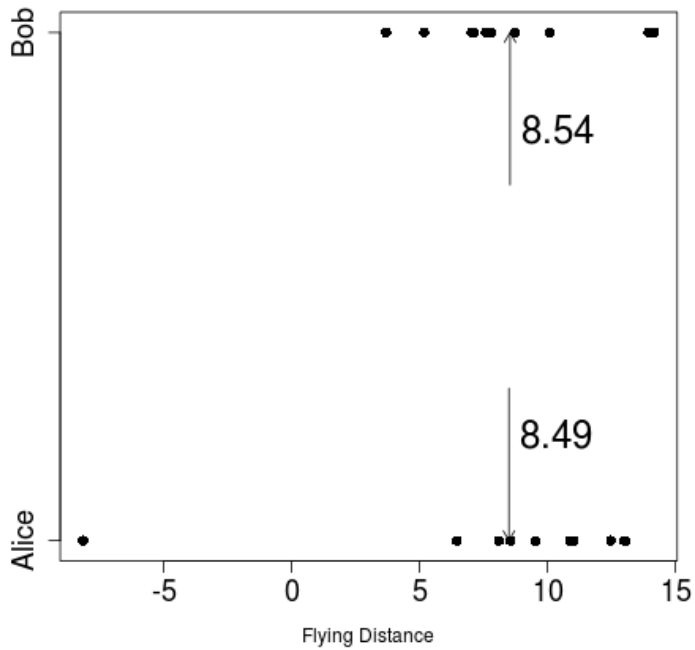
Data Visualization (3)

Point Diagram

If you want to compare the same value over different classes, a stripchart will be useful.

Let's try to see Alice's and Bob's story again!

```
## Many small tricks in this one -- check all the help files!
> AB = read.table("AliceBob.table",header=T)
> attach(AB)
> AB.means = tapply(Value,Researcher,mean)
> stripchart(Value~Researcher, xlab=expression("Flying Distance"),
             pch=16, cex=1.2,cex.axis=1.5)
> arrows(AB.means,c(1.3,1.7), AB.means,c(1,2),length=.1)
> text(AB.means,c(1.2,1.8),round(AB.means,2),pos=4,cex=2)
```



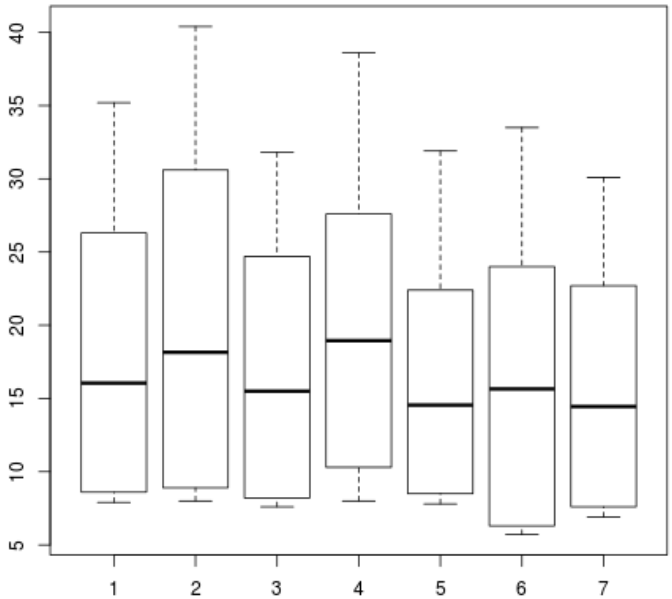
Data Visualization (4)

Box Plots

Very important for understanding multi-variate data:

- Can be used for various sample sizes;
- Gives an idea of the distribution of the data;
- Can be used for multiple factors at the same time;

```
> library(UsingR)
> png()
> attach(ewr)
> boxplot(AA,CO,DL,HP,NW,TW,US)
> dev.off()
```

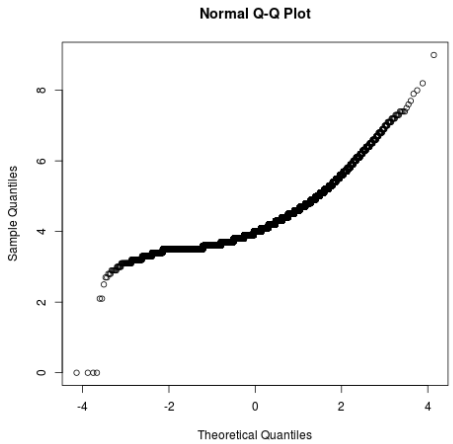


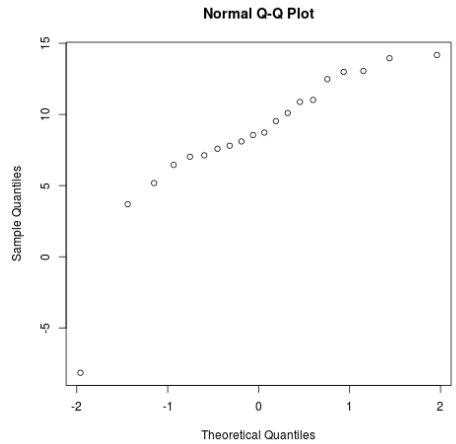
Data Visualization (5)

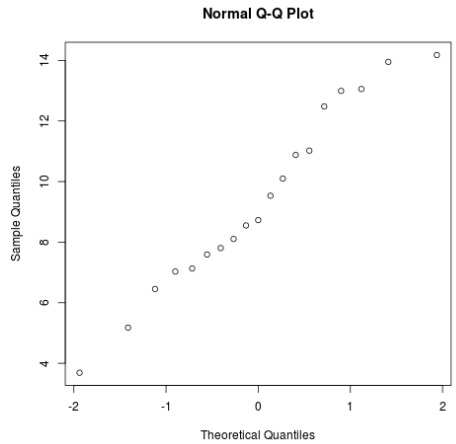
QQ Plot

- Compares the quantiles of two samples
- Allow you to compare the shapes of two distributions
- For example, you can compare the shape of a normal distribution with the shape of your data.

```
> qqnorm(M)
> qqnorm(Value)
> qqnorm(Value[Value > 0])
```







This is NOT all

There are many, many different ways of presenting and graphically analyzing data. This just scratches the surface! We might see a few more during this course, but you should also look for examples online and read manuals and FAQs.

Now let us talk about Scientific Hypothesis

Plato and his Cave

Population

- “Population” regards all the possible values for a variable;
- Usually it is impossible to directly observe;

Sample

- “Sample” are the values for the variable observed in an experiment;
- It is a subset, drawn from the population;

Plato and his Cave

Population

- "Population" regards all the possible values for a variable;
- Usually it is impossible to directly observe;

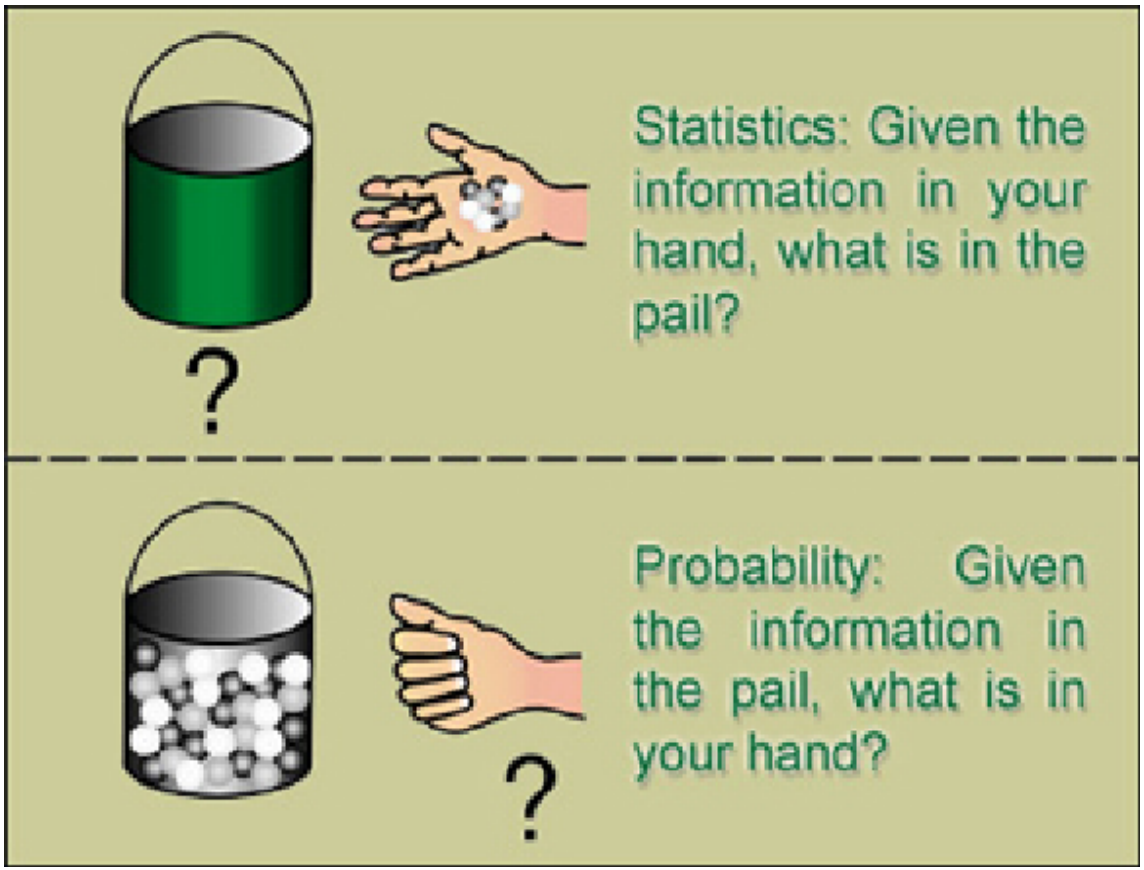
Sample

- "Sample" are the values for the variable observed in an experiment;
- It is a subset, drawn from the population;

In an statistical analysis, we want to discover information about the population, by extrapolating from the sample we got from the experiment.

Sampling and Sampling Distributions (1)

Statistical Inference: Making conclusions about a population from a sample taken from this population.



Sampling and Sampling Distributions (2)

Some more concepts

- Degrees of Freedom: Number of Independent factors, or sources of variation in a given statistic (parameters)
- Distributions: Description of the sample (or population)
  - Normal;
  - Power Law;
  - χ^2
  - t
  - F
  - etc.

Creating a Sample (and analyzing it)

- The “sample()” command;
- Mean, Variance, Standard Deviation; – now with different meanings
- Z-scores;
- Correlation (and spearman rank)

Distributions in R

The d, p, q, r family

- d: probability distribution function
- p: cumulative distribution function
- q: quantiles
- r: random samples

Some sample distributions

unif, binom, norm, exp, lnorm, etc...

Central Limit Theorem

Let $y_1, ..., y_n$ a series of random variables with mean μ and variance σ^2 finites, and let $x = \sum y_i$. Then:

$$Z_n = \frac{x - n\mu}{\sqrt{n\sigma^2}} \tag{1}$$

can be approximated to a normal(0,1) distribution.

What does this mean?

When the error in an experiment is additive and resulting from independent sources, it is reasonable to model the combined error from the experiment as a normal distribution.

This is useful to "normalize" errors from multiple experiments in CS.

Statistical Hypothesis (1)

What is an hypothesis?

Proposed explanation to an observed phenomenom.

Statements about the parameters of the Population
(No hypothesis about the sample, we KNOW the sample)

- Scientific Hypothesis:
  - Testifiable;
  - Falsifiable;
- The Hypothesis-Deduction model:
  - Create a falsifiable hypothesis;
  - Refute or Confirm the hypothesis by the data;
  - Compare rival hypothesis;
  - Predictive ability of the hypothesis;

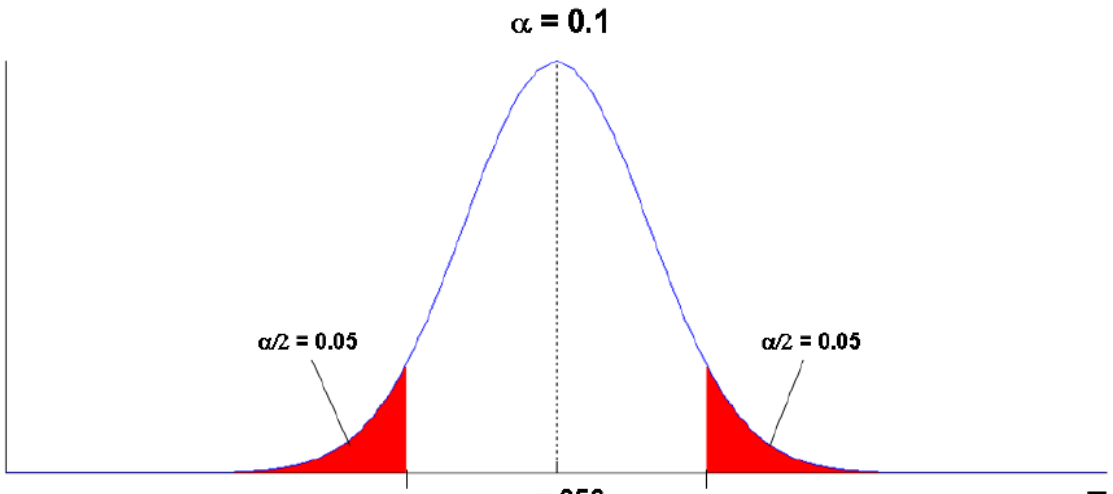
Statistical Hypothesis (2)

Null Hypothesis H_0

- Point value for the parameter being studied ($\mu = \mu_0$);
- "Nothing is different" hypothesis;

Alternative Hypothesis H_a

- Hypothesis for an observed result;
- Defines an interval of interest for this result:
  - one-sided: ($\mu_a < \mu_0$ or $\mu_a > \mu_0$)
  - two-sided: ($\mu_a \neq \mu_0$)



Statistical Hypothesis (3)

Defining the Null Hypothesis

- Previous knowledge about the process (to find out if there was a change in the parameter);
- Values obtained from theory or models (validation of the model);
- Requirements of a project (conformity tests);

Hypothesis Test

- Obtain the sampled data;
- Calculate the test statistics;
- Make a decision based on the value found;

Statistical Hypothesis (example)

- We define a null hypothesis as “the cans have an average of 350ml of beer in it.”;

$$H_0 : \mu = 350ml$$

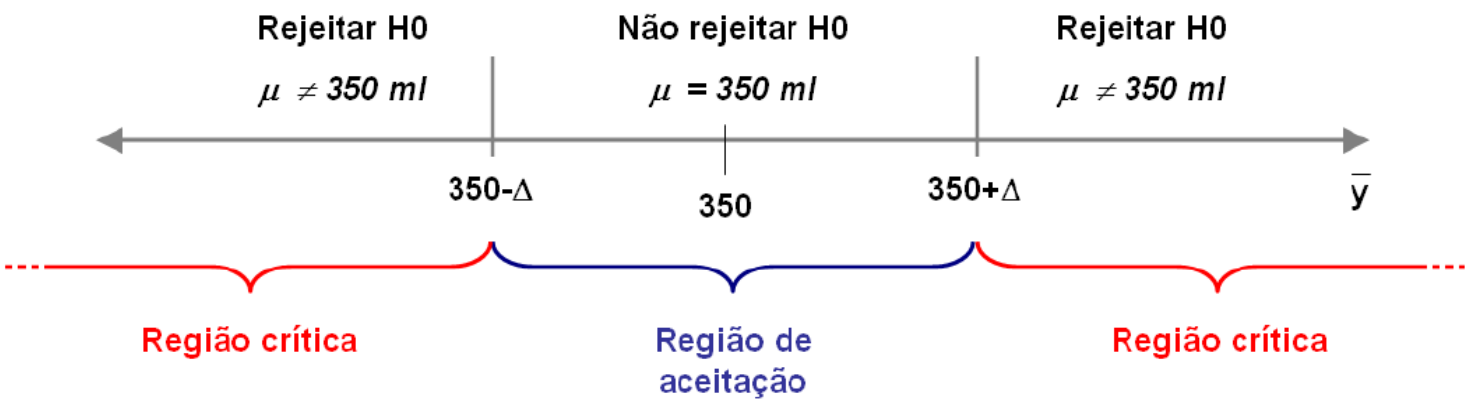
$$H_a : \mu \neq 350ml$$

- We take a sample of 20 cans, and measure the contents of each;
- The sample average $\bar{y}$ is an estimator for the population’s mean μ;
  - if $\bar{y} = 350$, the null hypothesis is not refuted;
  - if $\bar{y} \neq 350$, the null hypothesis is refuted - the reason must be studied.
  - this is the statistic we want to test;

Statistical Hypothesis (example - 2)

The alternate hypothesis

- $\bar{y}$ can assume a range of values;
- Thus, we define the H_a as a range of values different from H_0;



Errors in Hypothesis Tests

Decision	H_0 is True	H_0 is False
Fail to Reject H_0	no error	type II error
Reject H_0	type I error	no error

- Type I error (False positive): rejecting the null hypothesis when it is true;
- Type II error (False negative): not rejecting the null hypothesis when it is false;

Errors in Hypothesis Tests

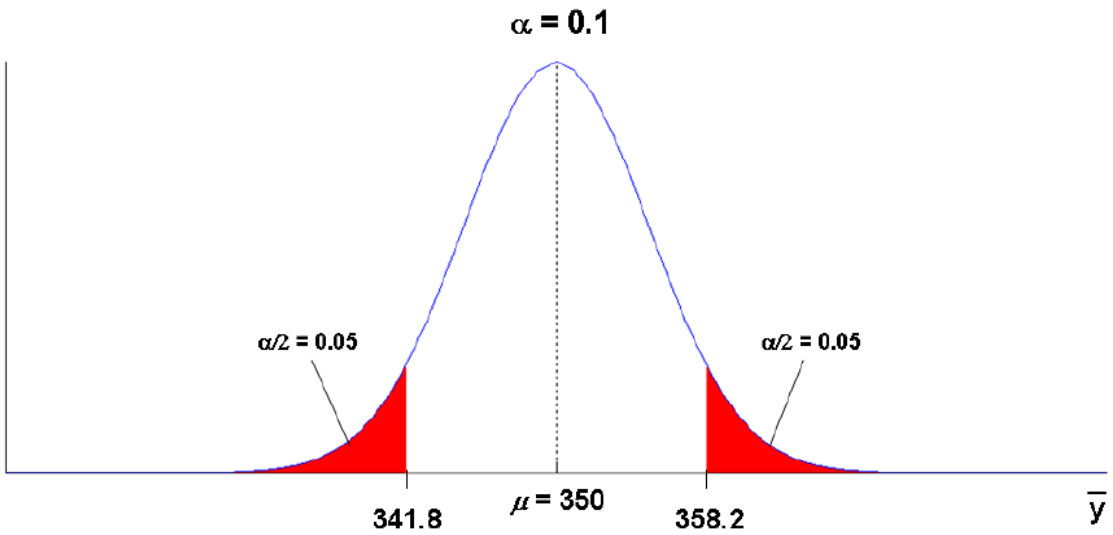
Decision	H_0 is True	H_0 is False
Fail to Reject H_0	no error	type II error
Reject H_0	type I error	no error

- Type I error (False positive): rejecting the null hypothesis when it is true; We don't want this!
- Type II error (False negative): not rejecting the null hypothesis when it is false;

Significance Level

Significance levels are the probabilities for one of the errors to happen:

- α: Probability of a type I error;
- β: Probability of a type II error;

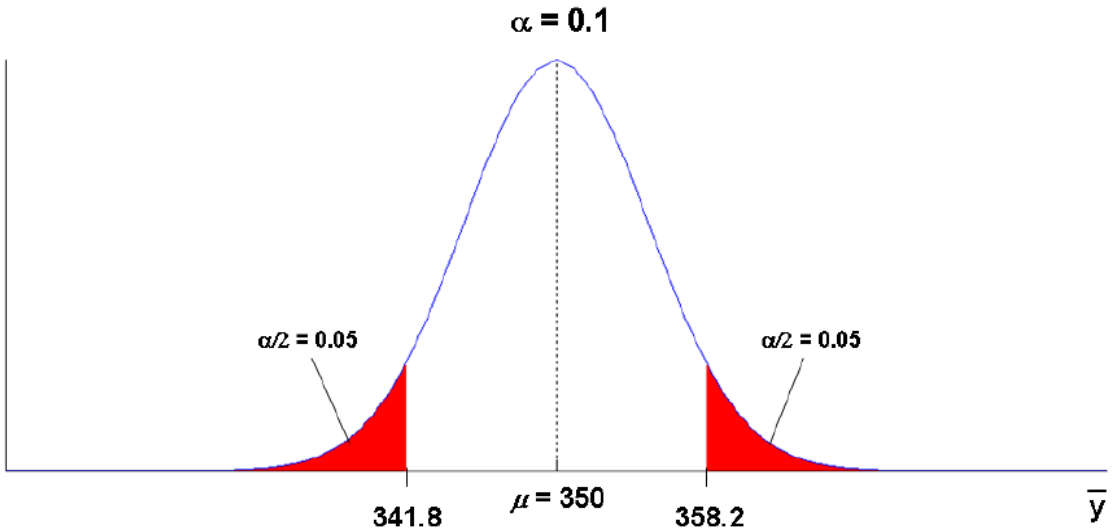


The probability distribution above are possible values for the population mean.

Significance Level

Significance levels are the probabilities for one of the errors to happen:

- α: Probability of a type I error;
- β: Probability of a type II error;

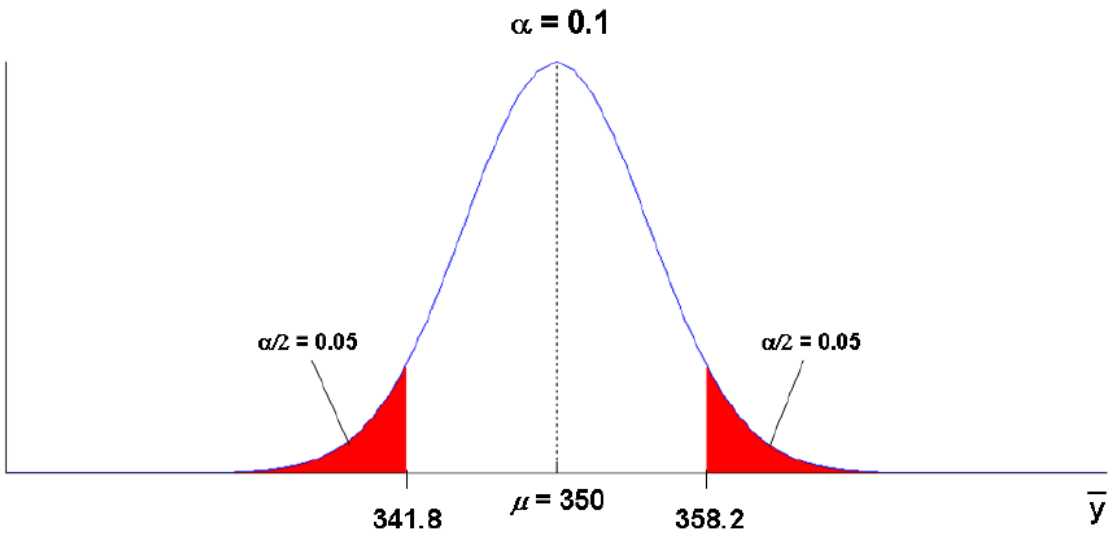


The probability of the population mean to fall in the defined critical area

Significance Level

Significance levels are the probabilities for one of the errors to happen:

- α: Probability of a type I error;
- β: Probability of a type II error;



Statistically, we can change the statistic by changing the sample size and the confidence interval

Things to keep in mind about statistical errors

- Type I errors depends only on the probability distribution of the null hypothesis - easier to control;
- Type II errors depends on the true value of the parameter under study; Harder to specify and control;
- Rejecting H_0 - Strong conclusion;
- Not rejecting H_0 - weak conclusion
  - In particular, Not rejecting H_0 is *NOT* evidence towards H_0;

General Procedure for an hypothesis Test

- Identify the parameter that interests us;
- Define H_0
- Define H_A, decide whether it is one or two sided;
- Determine the confidence level α;
- Determine which statistical test to use;
- Decide the critical region of the test (parameter values);
- Calculate the statistics;
- Reject, or not reject H_0;

Case: Normal Distribution; Known Variance (1)

$$H_0 : \mu = \mu_0$$

$$H_A : \mu \neq \mu_0$$

- Desired significance level: $\alpha = 0.05$
- The sample distribution of $\bar{y}$ is normal, with variance σ^2/n; ($\bar{y}$ is an estimation of μ);
- If H_0 is true, $\bar{y}$ $N(\mu_0, \sigma^4/n)$;

Case: Normal Distribution; Known Variance (2)

- Standardization of the sample mean:

$$Z_o = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$$

- If H_0 is true, Z_0 will fall in $N(0, 1)$
- Probability $(1 - \alpha)$ of Z_0 falling in the interval $-Z_{\alpha/2} + Z_{\alpha/2}$
  - $Z_{\alpha/2}$ is the $100(1 - \alpha/2)$ of a normal distribution.
  - This is for a two-sided test.

Case: Normal Distribution; Known Variance (3)

- If $z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$, then reject H_0 with significance level α
- If $-z_{\alpha/2} < z_0 < z_{\alpha/2}$, then do no reject H_0

Case: Normal Distribution; Known Variance (4)

- Let's assume $\bar{y} = 359.6ml, n = 20, \sigma = 22ml$

$$z_0 = \frac{359.6 - 350}{22/\sqrt{20}} = 1.962$$

- The critical values of the Z test for $\alpha = 0.05$ are 1.96
- Since $z_0 > 1.96$ there is enough evidence to reject the null hypothesis.

```
> library(TeachingDemos)
> samples<-c(373, 346, 357, 322, 382, 365, 390, 349, 366, 402,
  344, 353, 377, 355, 345, 360, 356, 316, 365, 379)
> (r.test(samples, mu=350, sd=22, alternatives, "two.sided",
     conf.level = 0.95))
```

Case: Normal Distribution; Unknown Variance (1)

The Z test had some assumptions about the population (distribution, variance). When those assumptions don't hold, we need to adapt the test accordingly.

$$H_0 : \mu = \mu_0$$

$$H_A : \mu < \mu_0$$

- Desired significance level: $\alpha = 0.01$
- Sample Variance: S^2. Used as an estimator for σ^2
- If H_0 is true, then

$$T_0 = \frac{\bar{y} - \mu_0}{S/\sqrt{n}}$$

follows a t_{n-1} probability distribution (different from N!)

Case: Normal Distribution; Unknow Variance (2)

- Assume $\bar{y} = 343.1$ml, n = 29, s = 18.2ml:

$$t_0 = \frac{343.1 - 350}{18.2/\sqrt{20}} = -1.69$$

- Critical value for the test: $-t_{0.01,19} = -2.54$;
- Because the value for the statistical test is not lower than the critical value, we conclude that the evidence is insufficient to reject H_0, at the 99% significance level.

```
> samples<-c(365,318,341,340,309,368,344,346,330,
  382,346,338,328,334,349,358,365,318,342)
> (t.test(samples, alternative = "less", mu = 350,
  conf.level = 0.99))
```

Our Results so far

- The result of a Hypothesis test, as we have done them so far, will say that:

"The evidence is enough/not enough to reject H_0, with significance level α."

- This is actually not really useful to judge an experiment:
  - Does not offer the exact value of the test statistic (how close is it to α?)
  - Fixes the significance level beforehand.

The P-Value!

P-Value

The minimum significance level that would result in the rejection of H_0 for the given data.

In other words...

P-Value

Probability that the test statistic will assume a value more extreme than the observed, if H_0 is true (probability of a Type I error).

P-Value (2)

- Example: in the last case, for $t_0 = -1.84$, the p-value would be:

$$p = P(t_0 \leq -1.84|H_0) = \int_{\infty}^{-1.84} t_{19} dy = 0.041$$

- The null hypothesis would be rejected for any $\alpha > 0.041$;
- We still need to define an *a priori* desired significance level!

Reality Check Time!

“pvaluecomic.jpg”

P-value and magnitude

Another issue with P values

We can calculate an arbitrarily low P value by choosing a large enough n.

... and this is easy to do in computer sciences!

Example: n = 5000, $\bar{y}$ = 349ml, s = 21ml

- $t_0 = -3.36$
- $p = 3.93x10^{-4}$

P-value and magnitude

So, we need to use estimators for the effect's magnitude, along with tests for statistical significance.

- Simple difference: $\bar{y} - \mu_0$
- Cohen's d:

$$d = \frac{\bar{y} - \mu_0}{s}$$

- Bootstrap estimators;