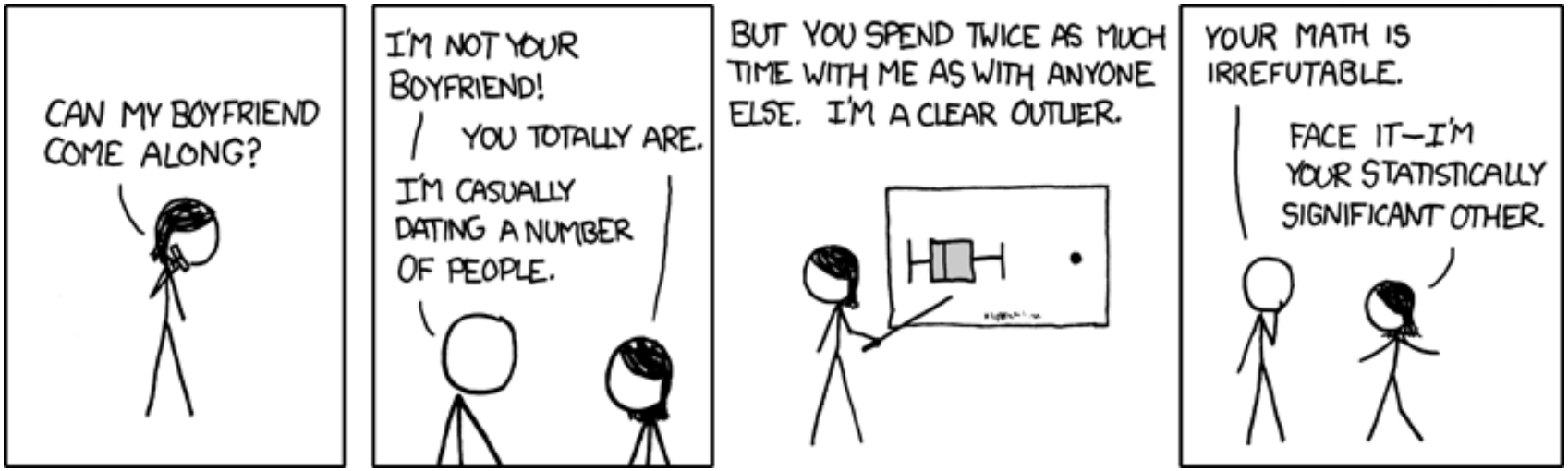
Introduction to data analysis in R

Basic statistics in R

Mahmoud Ahmed

April 2022



https://xkcd.com/539/

What is statistics?

Statistics is the practice concerned with collecting and analyzing data.

This module is about doing basic statistical summarizing and testing in R.

We will only be looking at a few ideas in statistics that will be needed for the upcoming modules. Namely

- Summarizing data
- Measuring variance
- Applying statistical tests
- Plotting data

Creating a toy dataset

The code below creates three variables and collect them in a _data.frame. You can think of this object d as encoding an experimental data. In the following sense

- V1 the sample ID
- V2 measurements
- V3 a grouping variable

```r
# create a data.frame
set.seed(1234)
d <- data.frame(
    V1 = 1:30,
    V2 = rnorm(30, mean = 1, sd = 1),
    V3 = rep(c('A', 'B'), each = 15)
)
```

God spiked the Bees

The code below introduces some bias in the measurements of the group 'B'.

```
# add 2 to V2 if, V3 == 'B'
d$V2 <- ifelse(d$V3 == 'B',
               d$V2 + 1,
               d$V2)
```

We will use summary statistics to explore the bias we introduce between the groups using different statistical and visual tools.

Summary statistics

Summarizing data is a sensible starting point when dealing with many data points. Two useful summary statistics are

- *Averages* are single numbers that describe a list of numbers.
- *Variance* is a measure of dispersion.

Averages

There are more than one way of defining averages.

1. Mean: the sum divided by the length
2. Median: the value separating the higher half from the lower half
3. Mode: the value that appears most often

There are in R functions that correspond to the first two of these statistics.

```
# calculate the mean of V2
mean(d$V2)

## [1] 1.203575

# calculate the median of V2
median(d$V2)

## [1] 1.220129
```

Variance

Variance is a measure of how far a set of numbers is spread out from their average value.

Here is the code to calculate the variance and the stardard deviation.

```
# calculate the variance of V2
var(d$V2)

## [1] 1.11628

# calculate the standard deviation of V2
sd(d$V2)

## [1] 1.056542
```

Did god spike the Bees?

Now we can calculate the difference we introduced to the measurements of the gourp 'B'.

```
# calculate the mean of the groups
A <- mean(d$V2[d$V3 == 'A'])
B <- mean(d$V2[d$V3 == 'B'])

# calculate the difference
B - A

## [1] 1.081744
```

Plots

There are many different to visualize any set of data.

The basic graphs are

- Points
- Bars
- Lines

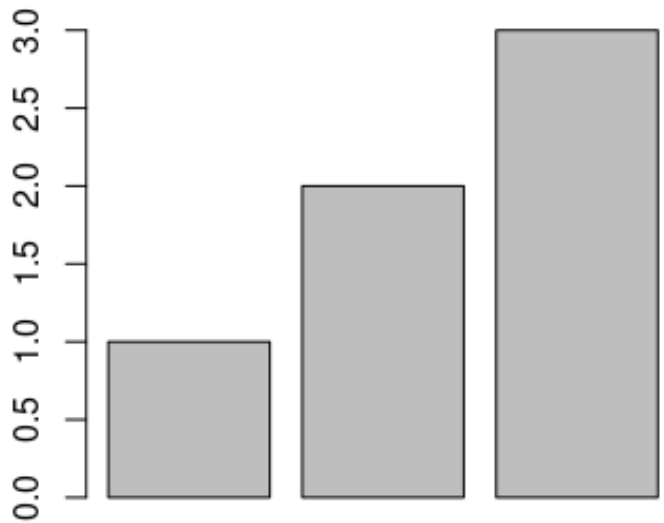
Used in the right way and context they can be very illuminating.

Once your data is large enough, graphs like these will be the only way to look into the data and the relationships between the variables.

Bar plot

The function barplot creates a bar graph. The only required argument you need is called ‘height‘, the rest is optional.

```
# plot 1 to 3
barplot(1:3)
```

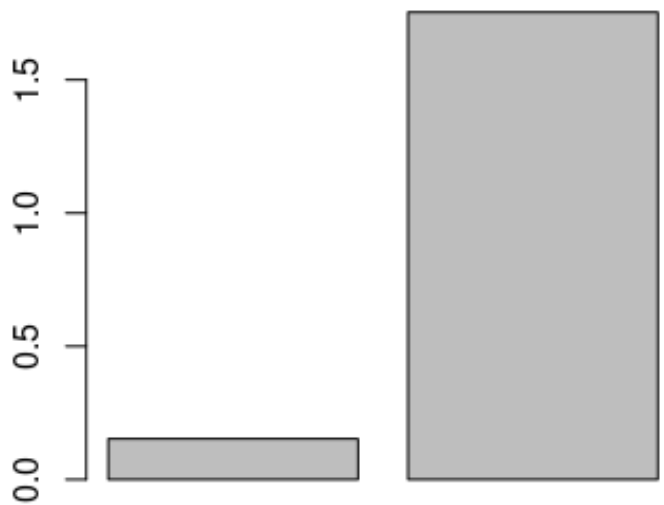


Bar plot

Here are the means of of measurements from each group.

```
# calculate and plot the group means
m <- c(mean(d$V2[d$V3 == 'A']),
       mean(d$V2[d$V3 == 'B']))

barplot(m)
```



Bar plot (Pretty)

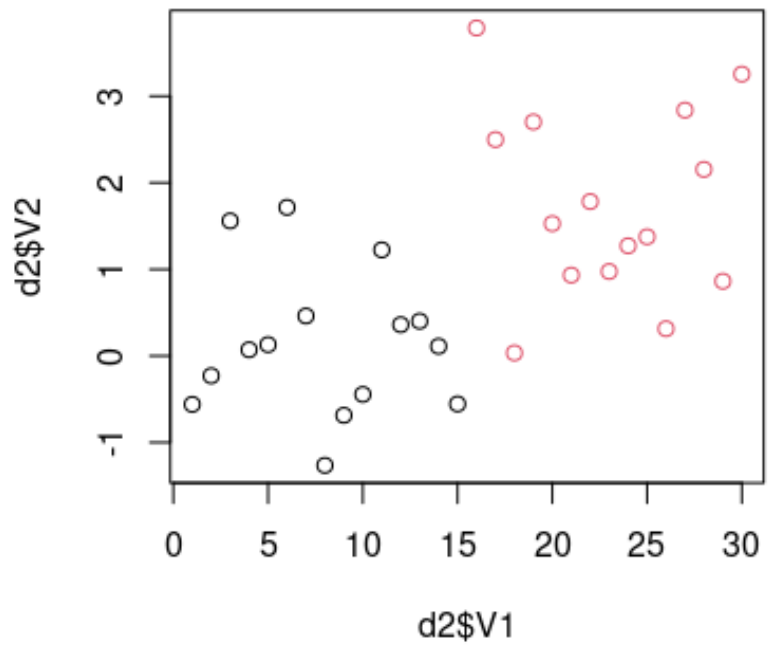
The code below shows some embellishment in the previous graph.

```
# modify the graph
barplot(m, xlab = 'Group', ylab = 'Measurement',
ylim = c(0, 2))
axis(side = 1, at = c(.7, 1.9), labels = c('A', 'B'))
```

Point plot(or scatter plot)

This code plots V1 on the x-axis and V2 on the y-axis and colors the points by V3.

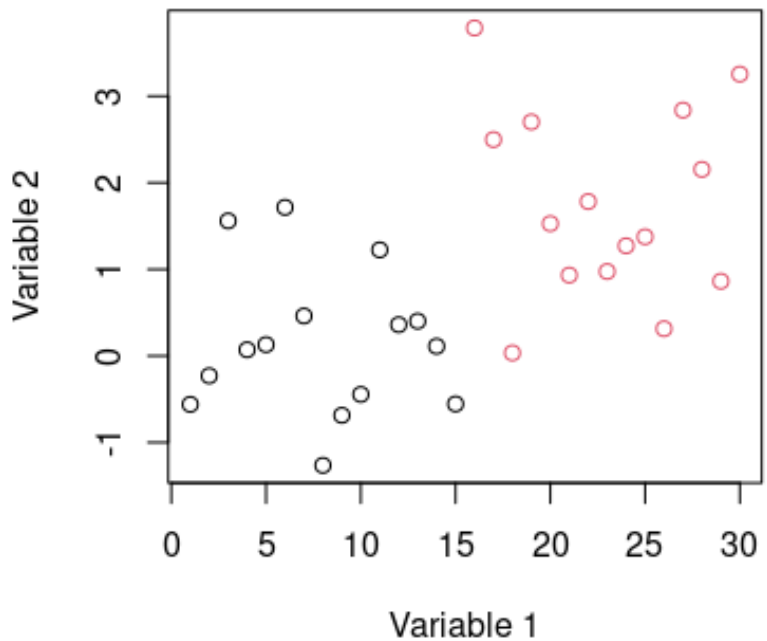
```
# plot all three variables
# as factor is necessary to force A and B to colors
plot(x = d$V1, y = d$V2, col = as.factor(d$V3))
```



Point plot (Pretty)

The code below shows some embellishment in the previous graph.

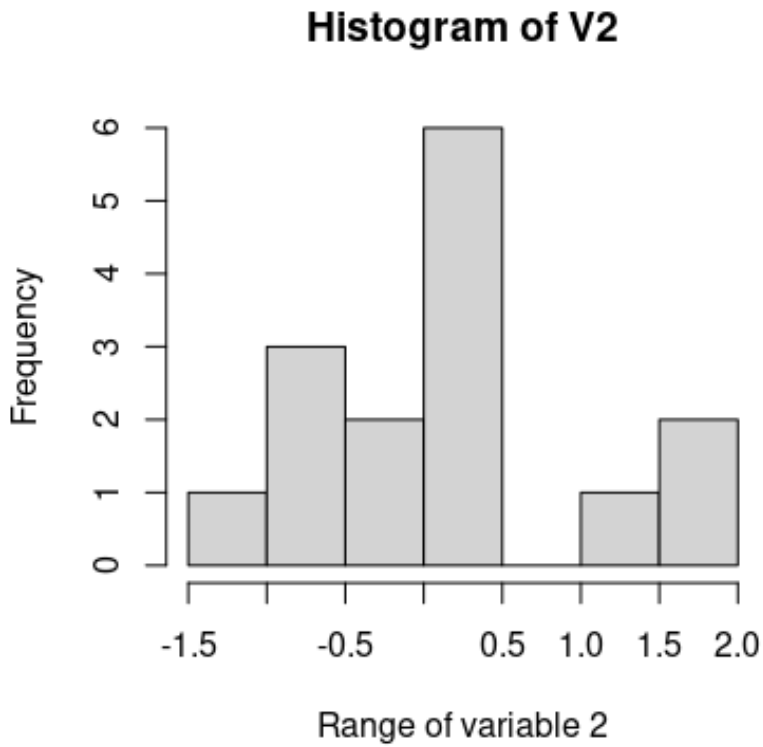
```
# plot all three variables
# as factor is necessary to force A and B to colors
plot(x = d$V1, y = d$V2, col = as.factor(d$V3),
     xlab = 'Variable 1', ylab = 'Variable 2')
```



Histogram plot

A histograms counts the frequency of data points within specified ranges. In other word, it divides the data into ranges called bins and count how many occur in each of these bins

```
# draw a histogram
hist(d$V2, xlab = 'Range of variable 2')
```



Tests

Looking at the mean, variance, and the plots we made, you can easily there is a difference between their means

Is this difference statisticly significant?

One way to answer this kind of question is to use statistical tests.

Choosing which test to use, indeed whether you should use tests at all or whether significance mean anything is beyond the topic at hand.

Tests

This code that applies the a *t*-test to the two groups, A and B.

```
# apply t.test
t.test(d$V2[d$V3 == 'B'], d$V2[d$V3 == 'A'])

##
##  Welch Two Sample t-test
##
## data:  d$V2[d$V3 == "B"] and d$V2[d$V3 == "A"]
## t = 3.2271, df = 27.974, p-value = 0.003181
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  0.3950823 1.7684055
## sample estimates:
## mean of x mean of y
##  1.744447  0.662703
```

Summary

What you learned

- Summary statistics
- Plots
- Tests

What's next

- Practice (Link)
- Homework (Link)
- Module 3: Quantifying mRNA using the pcr package (Link)