Statistical Analysis in R

An introduction to Statistics and their use in R

Disclaimer

The purpose of this tutorial is simply to provide an introduction to a variety of topics in R. It is in no way a substitute for a full for-credit course on R, and should students want to pursue R on a more serious level they should look to the university courses on the subject

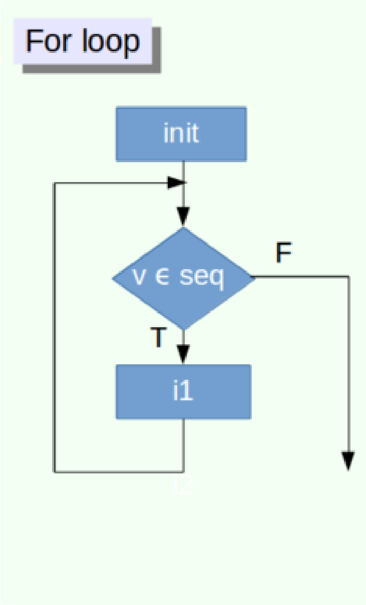
Loops

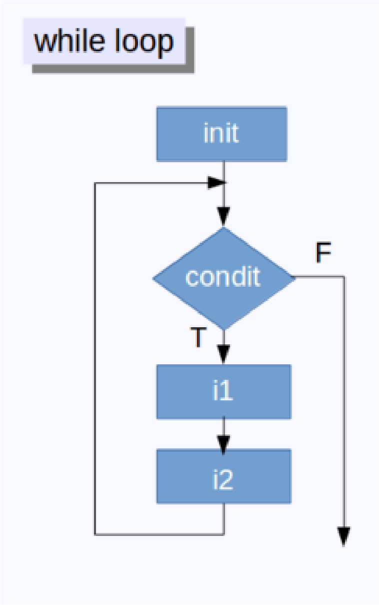
We will start with one of the most important and transferable skills in R: Loop. "Looping" is just the process of repeating a set of instructions for as many times as deemed necessary by our code. We use two main types: "for" loops and "while" loops .

Examples of why we might want to use a loop:

- Count from 1 to 10.
- Go through all the words in a dictionary to see whether they're palindromes.
- For each customer that has an outstanding balance, send out an email reminder that payment is due.

A **for loop** specifies the number of iterations of our pattern by expressing an actual number, ie "for 7 iterations please do this task". Meanwhile a **while loop** does a task until a certain condition is met, ie "while x is less than 7, please keep doing the task".





Let's do some practice:

Print the numbers 1 through 10:

With a for loop:

```
for (i in 1:10) {
      print(i)
      }
```

With a while loop:

```
x<-1
while(x<11){
  print(x)
  xi<-x+1
  x <-xi
}
```

Loops are particularly useful when we want to do many trials of something without having to manually run the code many times.

Probability Distributions

Recall that R is a programming language that was built by statisticians for statisticians, as a result, the probably most powerful part of R is the ability to do complex analysis on a variety of probability distributions. Today we will cover 3 types of probability distribution:

First a note on mean and variance:

You may be most familiar with the simple definition of mean as an average of a set of variables. In statistics though, as distributions become more complicated, we adjust our definition for the mean.

In upper level statistics we have a concept called the **Expected Value**, which is the sum of all the possible values multiplied by the probability of each outcome.

Mathematically speaking:

$$E(X)=\sum_{i=1}^{n} X_i P(X_i)$$

We additionally rely on a concept called the **Variance**, which is the expected value of the squared deviation from the mean. It measures how far the values of a distribution are from the mean.

Mathematically speaking:

$$\operatorname{Var}(X) = \operatorname{E}\left[(X-\mu)^2\right].$$

For most of these distributions, we will talk about them in terms of their variance and mean.

1. **The Binomial Distribution:**

a. Fixed number of n trials.
b. Each trial only has two possible outcomes, either a success or a failure
c. The trials are independent (the outcome of one trial will not affect the outcome of other trials)
d. The probability of success is expressed as p, with the probability of failure expressed as 1-p (sometimes written q)
e. Common examples are a coin flip, where heads is seen as a success and tails is seen as a failure
f. The parameters for this distribution are n = number of trials, p = probability of success, and q = 1-p = probability of failure.
g. The mean of a binomial distribution is n*p, and the variance is n*p*q

R code:

```
rbinom(n, size, prob)
        n = number of observations, size = number of trials, prob = probability of success
rbinom(100, 1, .5)
```

Will produce 1 sample with 100 observations, each observation having a 50% chance of being a success.

2. **The Poisson Distribution**
   a. For scenarios where the average likelihood of an event occurring in a certain period of time is known
   b. If we know the average person walks 2 miles in one day, then what is the likelihood they will walk 10 miles? 0 miles?
   c. The parameter for this distribution is simple **λ** (lambda) which represents the average likelihood for our time interval
   d. The mean of the poisson distribution is **λ** and the variance is also **λ**

R code:

```
rpois(n, lambda)
        n = number of random values to return, lambda = λ
rpois(10, 15)
```

Will produce 10 random values, for which the expected value was 15

3. **The Exponential Distribution**
   a. Like poisson distributions, the average likelihood of an event occurring over a given time period is known.
   b. The variable in question is how long until the next occurrence of an event.
   c. The parameter for this distribution is also **λ** (lambda) which represents the mean waiting time, or the rate
   d. The mean of the poisson distribution is 1/**λ** and the variance is 1/(**λ**^2)

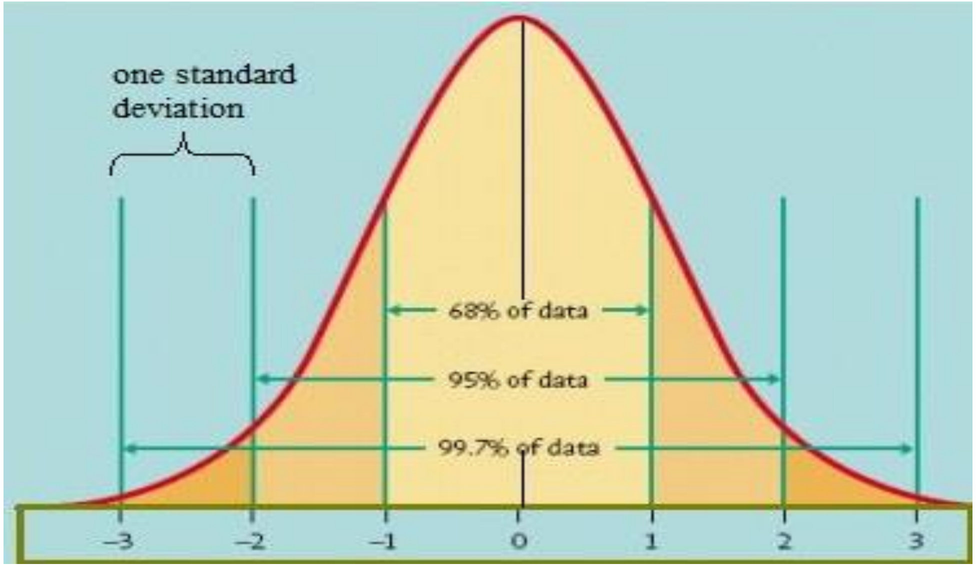
R code:

```
rexp(n , rate)
        n = number of observations, rate = λ
```

rexp(10, 6)

Will produce 10 observations with an expected value of 1/6

4. **The Normal Distribution**
   a. The mean of a normal distribution is ***μ*** (Mu) and **σ**^2 (sigma squared)
   b. The normal distribution follows a normal curve like so:



R code:
rnorm(n, mean, sd)

Regression Analysis

(From Google) A measure of the relation between the mean value of one variable (eg output) and corresponding values of other variables (eg time and cost)

model=lm(faithful)
model

OLS Regression

Ordinary least squares, or linear least squares, estimates the parameters in a regression model by minimizing the sum of the squared residuals. This method draws a line through the data points that minimizes the sum of the squared differences between the observed values and the corresponding fitted values. (http://statisticsbyjim.com/glossary/ordinary-least-squares/)

The following code shows students that they have a wide range of datasets to explore and chose from.

```
library(datasets)
library(help = "datasets")


dat<-iris
View(dat)

reg<- lm(Sepal.Length ~., data = dat)
summary(reg)
```

The output of this regression is a good way to start to look at how variables affect others in data exploration. The intercept is the value we are regressing the other variables on in this case _**Sepal.Length.**_ _In the equation, ~. represents that we are regressing all other variables on Sepal.Length. We could specify variables with the following code:_

```
reg1<- lm(Sepal.Length ~ Sepal.Width + Petal.Length, data = dat)
summary(reg1)
```

In these formulas, lm species the OLS regression being run, and data is assigned to your assigned data

Estimate is going to represent your coefficient. This is the weight of the variable in the overall formula

A p-value of <.05 (‘*’) is considered significant. In this case we reject the null hypothesis (ex: there is no relationship between sepal length and sepal width)

Std. Error is standard deviation of the coefficient

T value is the coefficient divided by the standard error. If you have a t value of >2 or <-2 for a regression with 30 or more observations, it is significant; however, we primarily use p-value for its measure of significance

The R^2 value can be thought of as variation explained by our regression/total variation
The higher the R^2 value, the better
When you have multiple variables in a linear regression, it is important to use adjusted R^2 as your R^2 value
R^2 will always increase with the more variables added into the equation
Adjusted R^2 includes a penalty for extra input variables

**Make sure to tell students to come back for advanced statistical analysis in R which will teach how to control for influence of each variable in an OLS regression**