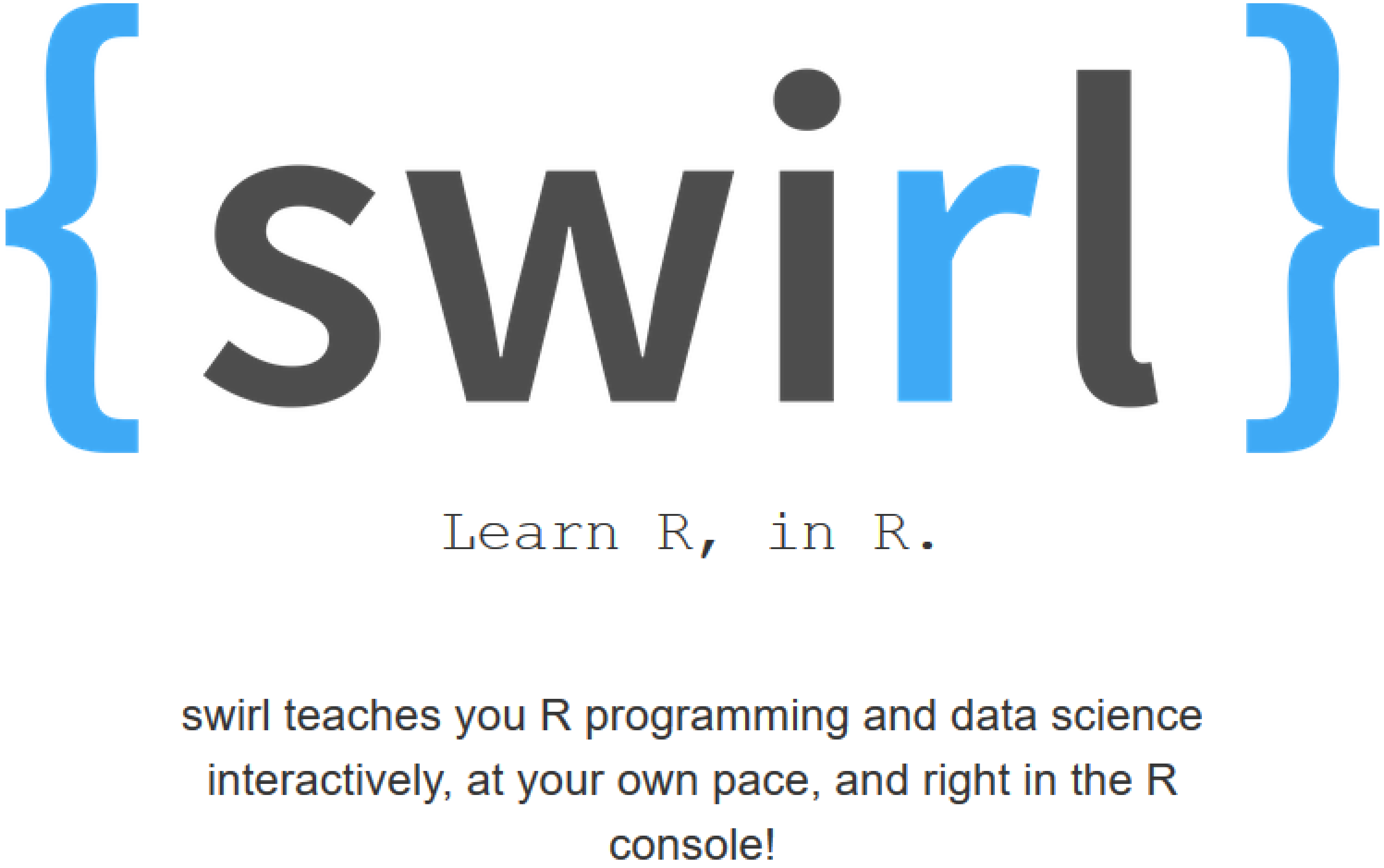
MBP Tech Talks
Introduction to R – Part 2

Department of Medical Biophsyics
University of Toronto

Danton Ivanochko

Swirl

https://swirlstats.com/



Learn R, in R.

swirl teaches you R programming and data science interactively, at your own pace, and right in the R console!

Outline

1. Short lecture
2. Hands-on practice

Data Structures: Review

The data structures the are used in R:

	Homogeneous	Heterogeneous
1d	Atomic vector	List
2d	Matrix	Data frame
nd	Array	

Atomic Vector – A basic data structure of R containing the same type of data. (logical, integer, double and character)

Matrices – A matrix is a rectangular array of numbers or other mathematical objects. We can do operations such as addition and multiplication on a matrix in R.

Lists – Lists store collections of objects when vectors are of same type and length in a matrix.

Data Frames – Generated by combining together multiple vectors, each vector becomes a separate column.

What is my data?? – typeof(), class()

Vectors

The basic data structure in R is the vector. Vectors come in two flavours: atomic vectors and lists. The built-in functions in R have great support for vectors.

```
a <- c(3,2,1)
b <- c(4,5,6)
```

```
> length(a)
[1] 3

> a + 1
[1] 4 3 2

> a + b
[1] 7 7 7

> a == 3
[1]  TRUE FALSE FALSE

> summary(a)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    1.0     1.5     2.0     2.0     2.5     3.0 
```

Adapted from Petr who adopted it from Mehran

Data Frame
“progeny of a matrix and a list”

A data frame is more general than a matrix, in that different columns can have different modes (numeric, character, logical, factor, etc.). Under the hood, a data frame is a list of equal-length vectors.

```
d <- c(1,2,3,4)
e <- c("red", "white", "red", NA)
f <- c(TRUE,TRUE,TRUE,FALSE)

my_df <- data.frame(d,e,f)

colnames(my_df) <- c("ID","Color","Passed")
```

```
> my_df
  ID Colour Passed
1  1    red   TRUE
2  2  white   TRUE
3  3    red   TRUE
4  4   <NA>  FALSE
```

Data Frame
“progeny of a matrix and a list”

There are several methods to subset a data frame

```
> my_df
  ID Colour Passed
1  1    red   TRUE
2  2  white   TRUE
3  3    red   TRUE
4  4   <NA>  FALSE
```

```
my_df[,1:2] # columns 1 and 2 of data frame

my_df[c(1,3,4),] # rows 1, 3 and 4 of data frame
my_df[-2,] # also rows 1, 3 and 4 of data frame

my_df[,c("ID","Passed")] # columns ID and Age from data frame

my_df$Colour # variable x1 in the data frame
```

What is a factor?

A factor is an attribute of categorical (A.K.A. nominal) variables

Unfortunately, most data loading functions in R automatically convert character vectors to factors. This is suboptimal, because there’s no way for those functions to know the set of all possible levels or their optimal order.

Instead, use the argument **stringsAsFactors = FALSE** to suppress this behaviour, and then manually convert character vectors to factors using your knowledge of the data.

Adapted from Petr who adopted it from Mehran

What is a factor?

A factor is an attribute of categorical (A.K.A. nominal) variables.
Factors are made up of two parts: a vector of integers, and a vector of corresponding levels/labels.

```
a <- data.frame(col1 = c(1,2,3), col2 = c("a","b","c"))
b <- data.frame(col1 = c(1,2,3), col2 = c("a","b","c"),stringsAsFactors = FALSE)
```

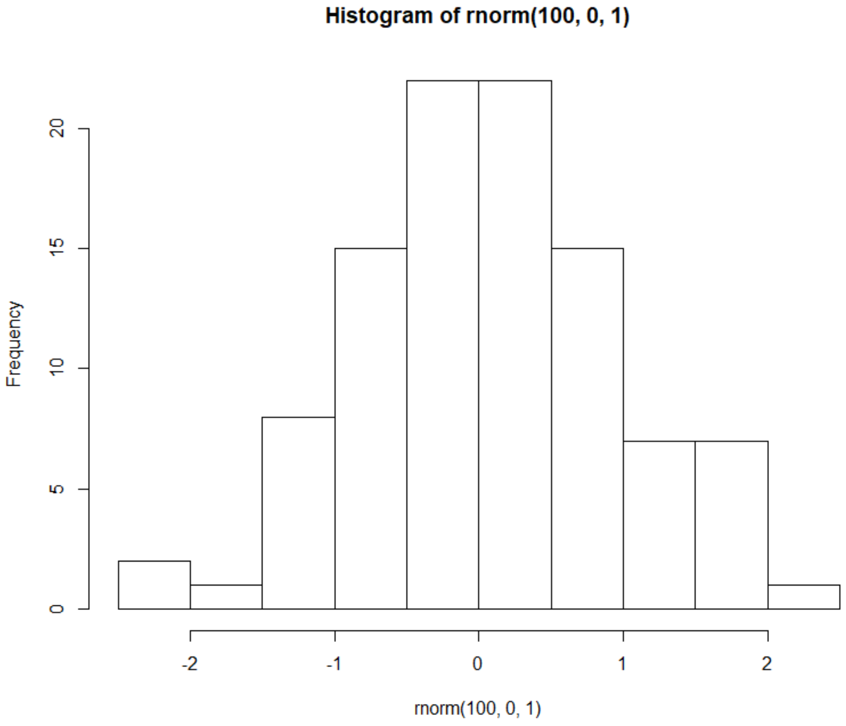
```
> a$col2
[1] a b c
Levels: a b c
> class(a$col2)
[1] "factor"
> a$col1
[1] 1 2 3
> class(a$col1)
[1] "numeric"
```

```
> class(b$col2)
[1] "character"
```

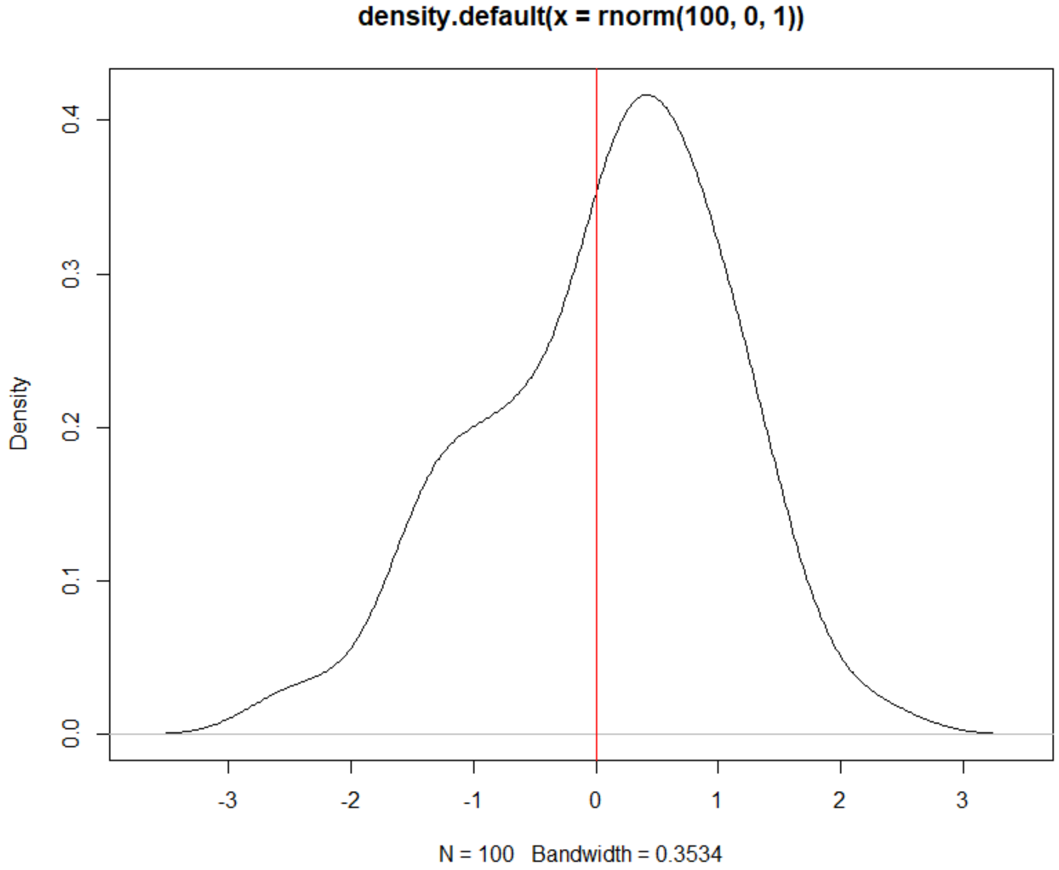
Plotting

R provides many built in plots, and they have the bonus of looking pretty alright!

```
# Histogram
> hist(rnorm(100,0,1))
```



```
# Density plot
> plot(density(rnorm(10,0,1)))
> abline(v=0, col="red")
```

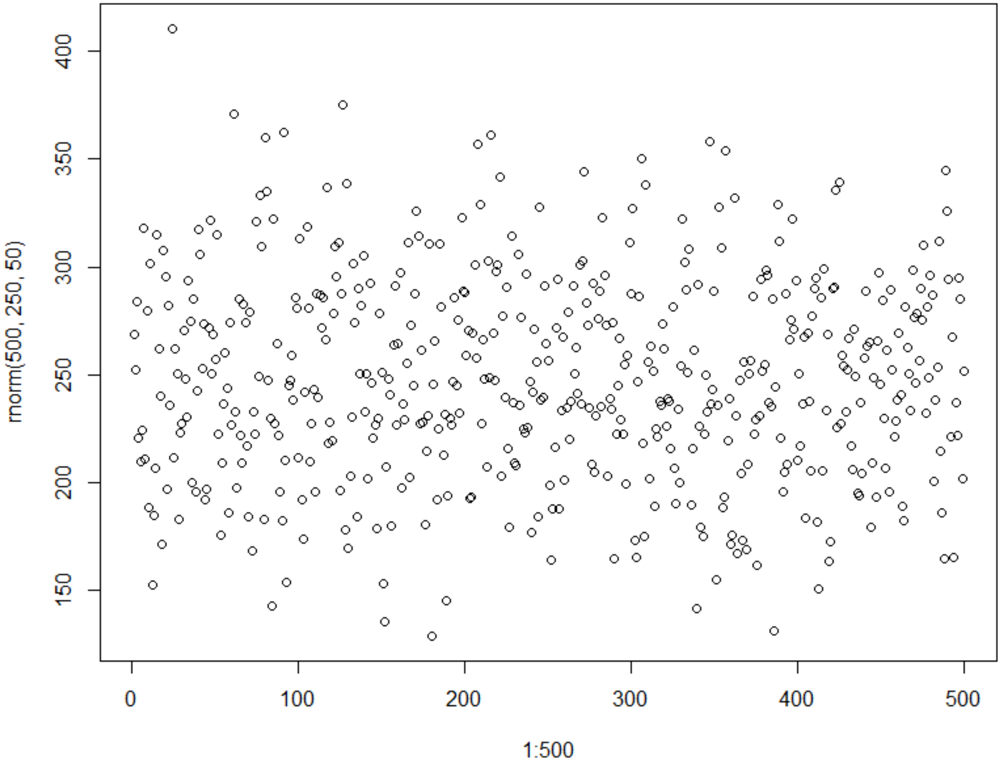


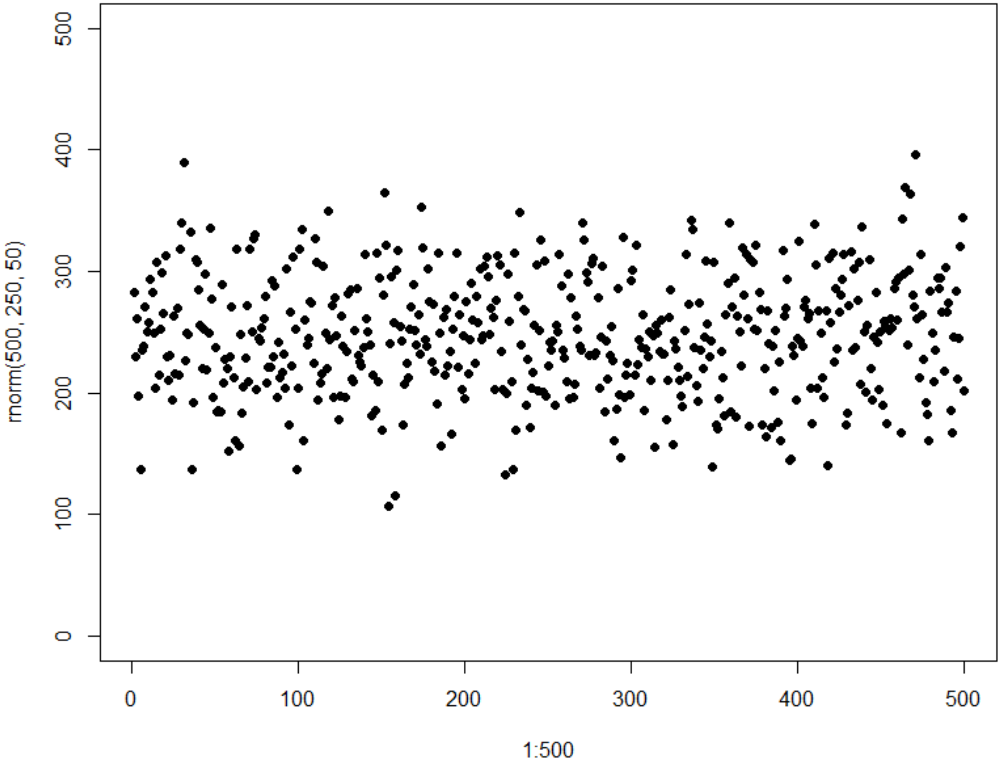
Plotting

R provides many built in plots, and they have the bonus of looking pretty alright!

```
# Scatter plot
> plot(x=1:500, y=rnorm(500,250,50))
```

```
# A nicer scatter plot
> plot(x=1:500, y=rnorm(500,250,50),
+       ylim = c(0,500), pch = 19)
```

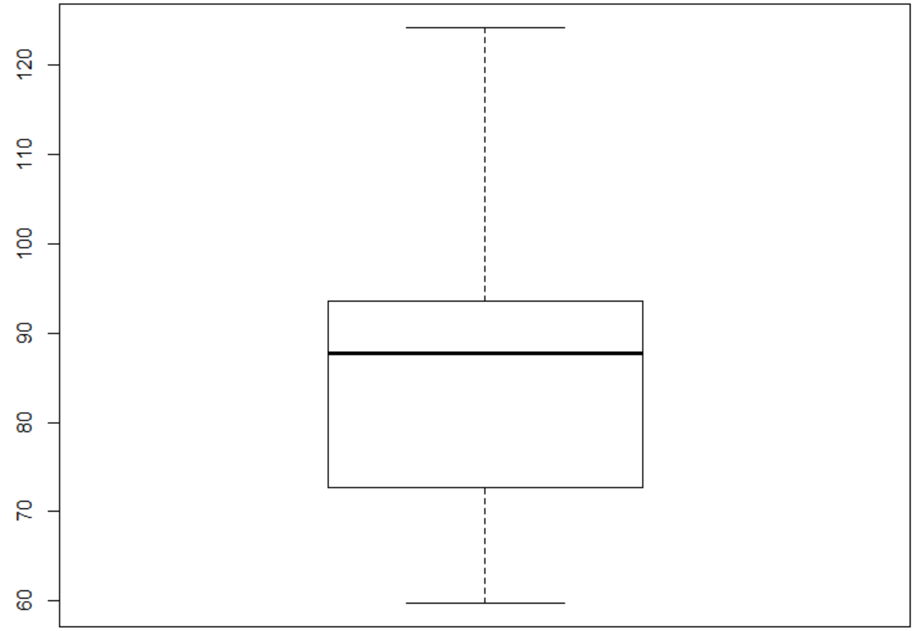




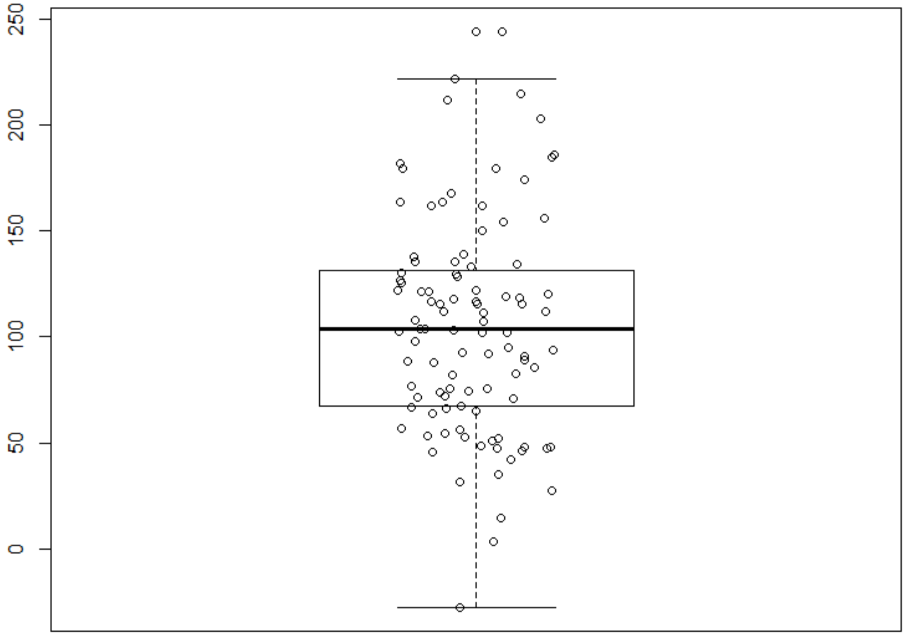
Plotting

R provides many built in plots, and they have the bonus of looking pretty alright!

```
# Boxplot
> boxplot(rnorm(10,100,50))
```



```
# Boxplot with dots
x <- rnorm(100,100,50)
> boxplot(x)
> stripchart(x, add = TRUE,
+               vertical = TRUE, method =
+               "jitter", pch = 21)
```



Stats in R: The p-value

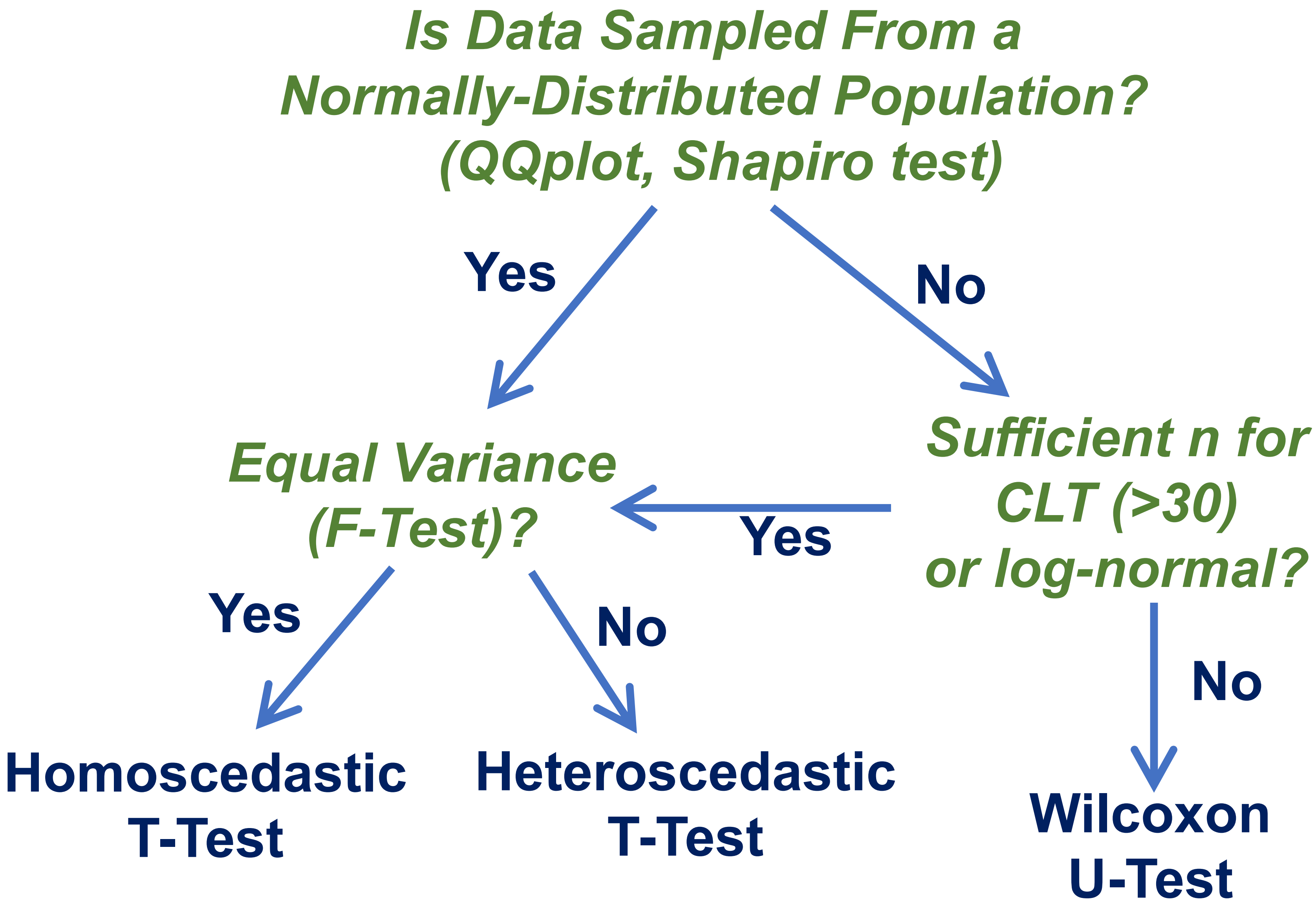
The P stands for *probability*
It measures how likely it is that any observed difference between groups is due to chance.
In other words, the p-value is th probability of observing a value as extreme or more extreme by chance alone.

Most people use a p-value = 0.05 (but this is ***completely*** arbitrary)

It is a measure of how much evidence we have against an alternative hypothesis
(e.g. Null hypothesis: 2 groups are same, Null hypothesis: 2 groups are different).

If the p-value is less than (or equal to) α, then the null hypothesis is rejected in favor of the alternative hypothesis. And, if the p-value is greater than α, cis not rejected.

Flow-Chart For Comparing Continuous Data



Is Data Sampled From a Normally-Distributed Population? (QQplot, Shapiro test)

Yes

No

Equal Variance (F-Test)?

Yes

Sufficient n for CLT (>30) or log-normal?

Yes

No

No

Homoscedastic T-Test

Heteroscedastic T-Test

Wilcoxon U-Test

Adapted from BioStats model 2016 - Brendan Innes and Paul Boutros

Assessing normality and variance

The QQ-plot (qqnorm() and qqline())
The quantile-quantile plot visualizes whether data deviates from normal distribution.
If data is normally distributed, we should observe a straight line. Compare to the qqline.

Shapiro-Wilk normality test (shapiro.test())
shapiro.test tests the Null hypothesis that "the samples come from a Normal distribution" against the alternative hypothesis "the samples do not come from a Normal distribution".

F-test (var.test())
When you want to perform a two samples t-test to check the equality of the variances of the two samples.
Homoscedastic and heteroscedastic imply equal and unequal variance, respectively.

T-test (t.test())
Compute a p-value if both samples are on a normal distributions

Mann-Whitney and Wilcoxon test (wilcox.test())
Non-parametric test to compute a p-value that can apply to either non-normal data and normal data (with less statistical power).

Learning by doing

We have some fake data to do a practice analysis on.

This will be fairly simple stuff that you may have done using excel in the past.
(e.g. process data, calculate averages, do some t-tests and make some graphs)

The experiment:

- 19 mice xenograft tumors models were split into a drug treated (n=9) or untreated (n=10)
- at endpoint, dimensions of tumors were measured and RT-qPCR experiment was done to examine the expression of 3 genes and a control gene

qPCR crash course!!!
CT = cycle threshold, how many reaction doublings before a gene was detected
2-∆∆ct method to display graphs