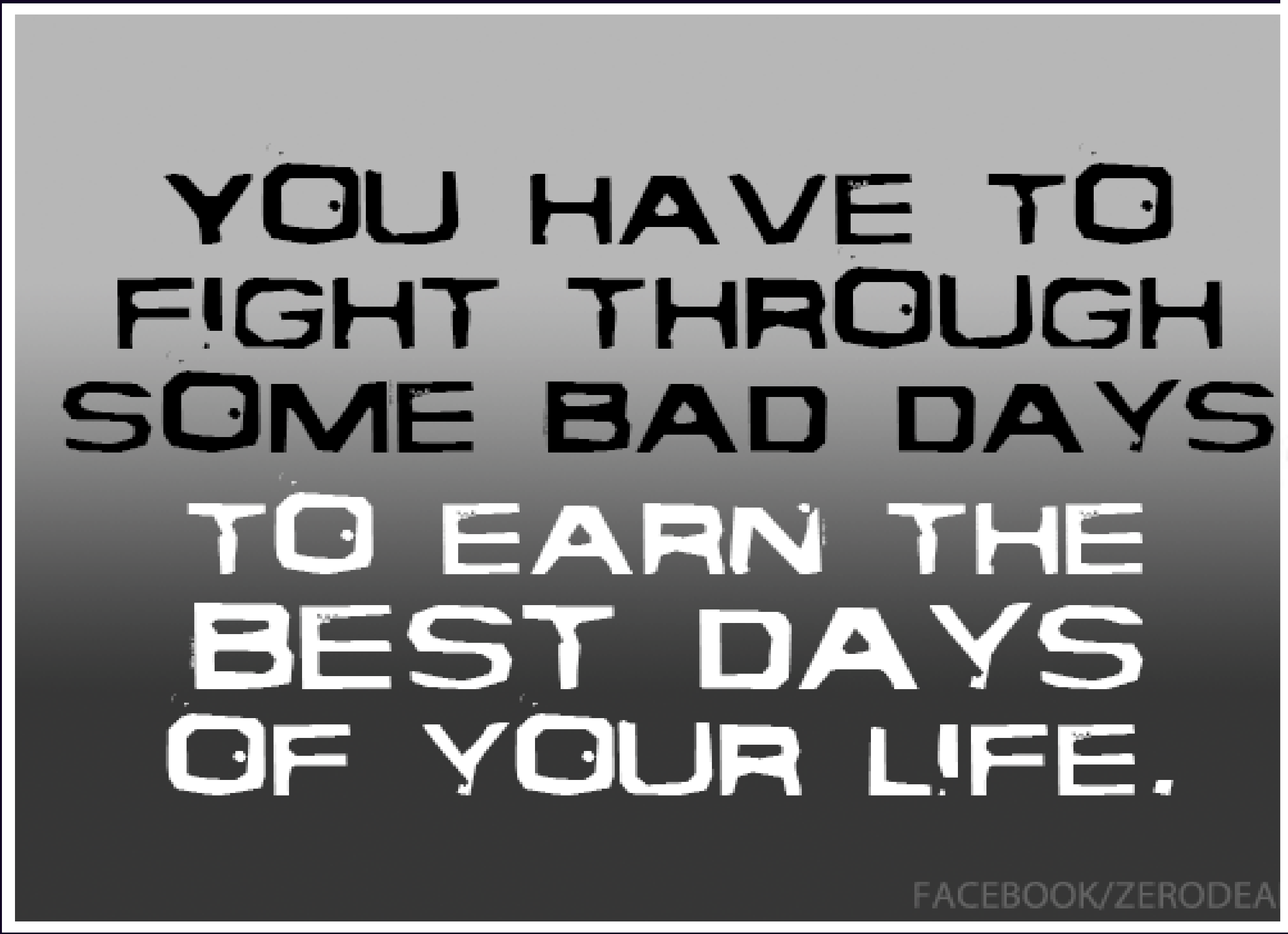
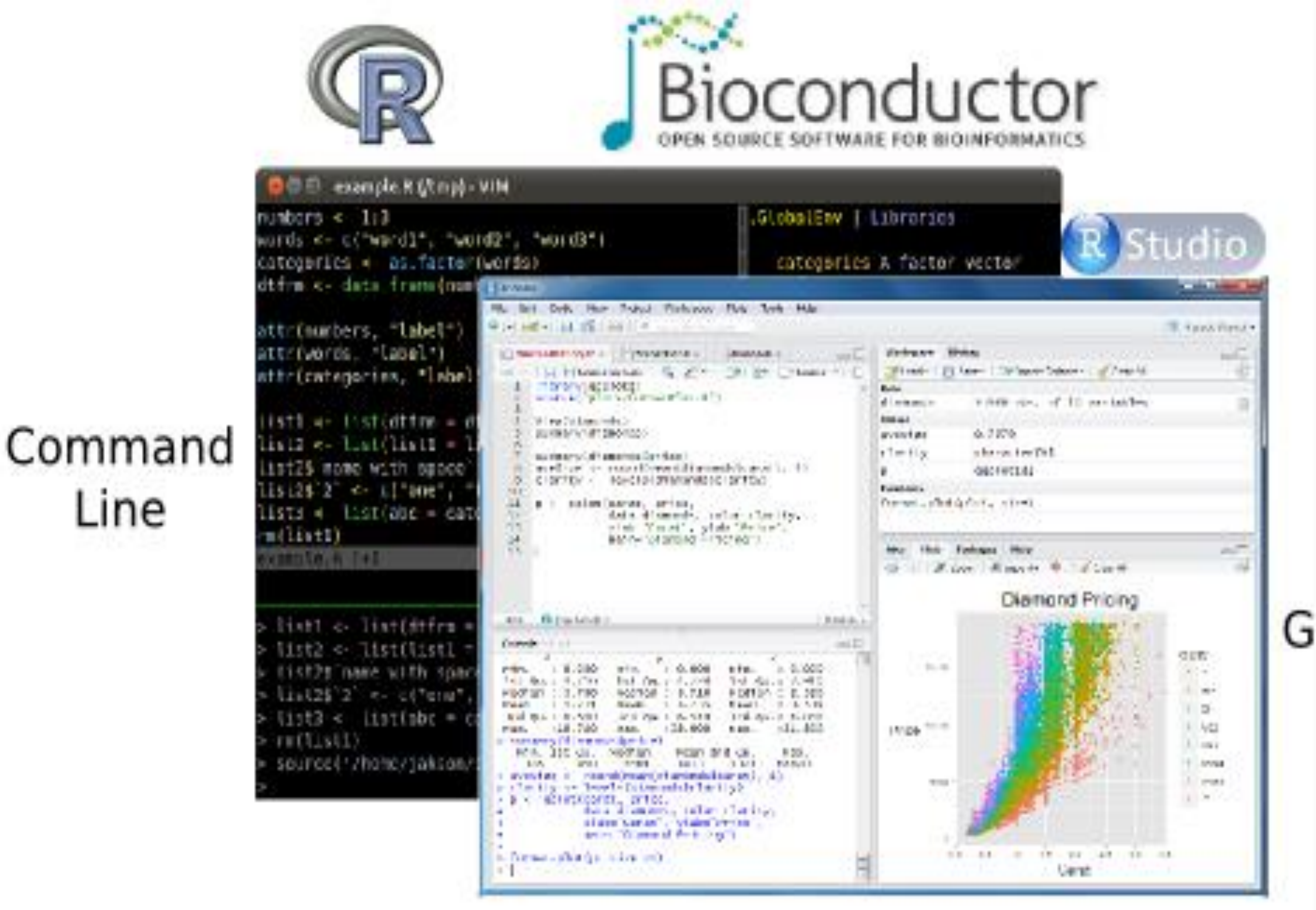


YOU HAVE TO FIGHT THROUGH SOME BAD DAYS TO EARN THE BEST DAYS OF YOUR LIFE.

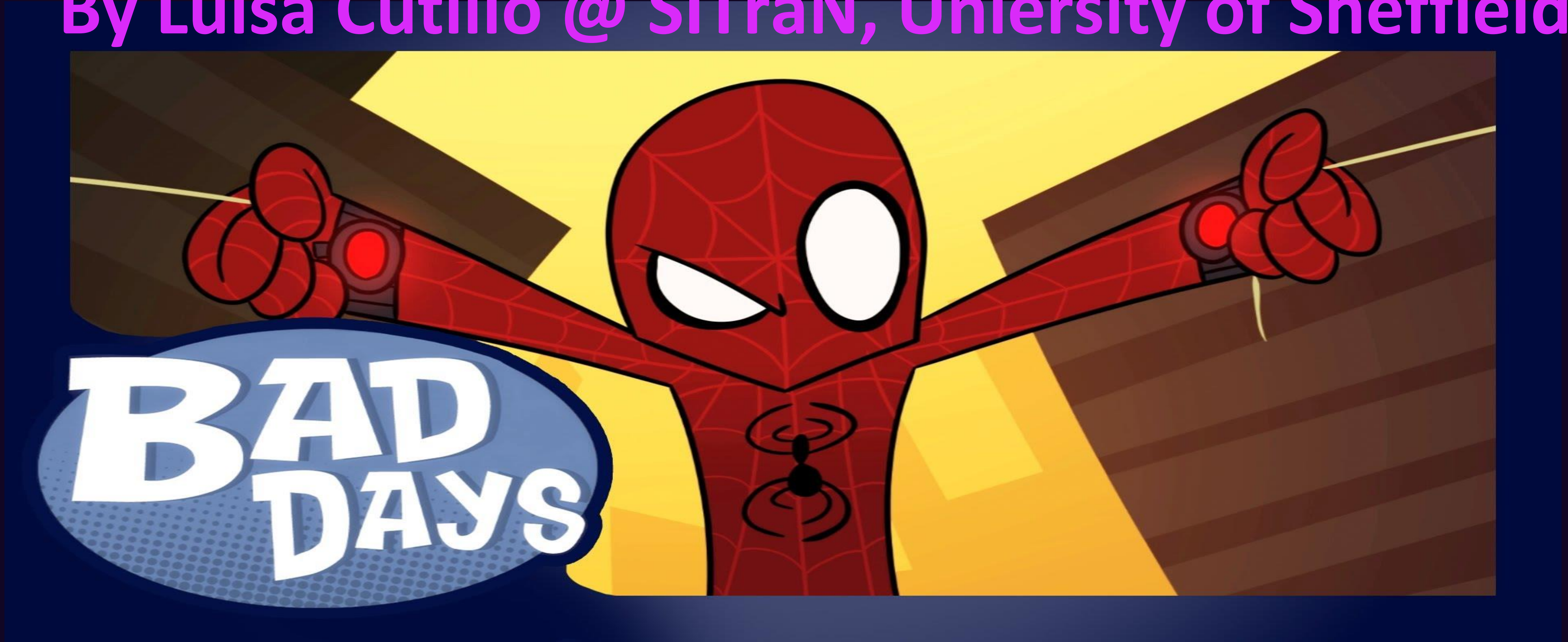
FACEBOOK/ZERODEA



Wed, 4 May at 13:00, Sheffield

BAD Days: Bioinformatics Awareness Day – 1

By Luisa Cutillo @ SITraN, Uniersity of Sheffield



Module I Outline

- A brief introduction about Data Science
- R and Exploratory Data Analysis
- Hands on data: produce good graphics!
- **Tutorial** Tutorial.R
- **Tutorial** Comments.R

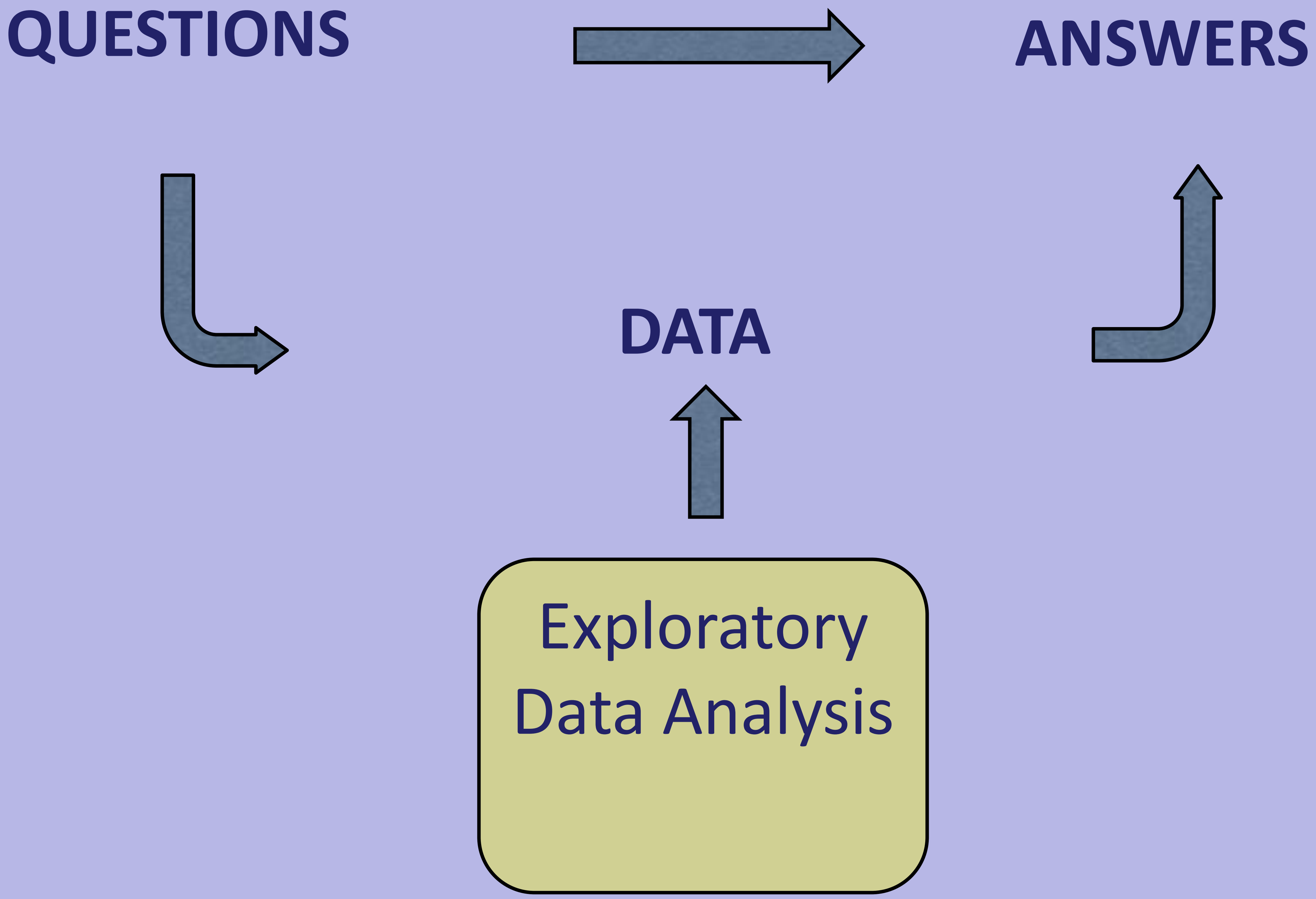
Are you a Data Scientist?

IDENTIKIT!

DS Identikit, do you:

- find and interpret data sources?
- manage large amounts of data to create visualizations to aid in understanding data?
- show the data insights/findings and produce answers in days rather than months?

CONGRATULATIONS! You are in the right place.



QUESTIONS

ANSWERS

DATA

Exploratory Data Analysis

EDA: Exploratory Data Analysis (Tukey, 1961)

Objectives:

- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which statistical inference will be based
- Support the selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments

EDA: A CLEAR PICTURE is worth then a thousand words!

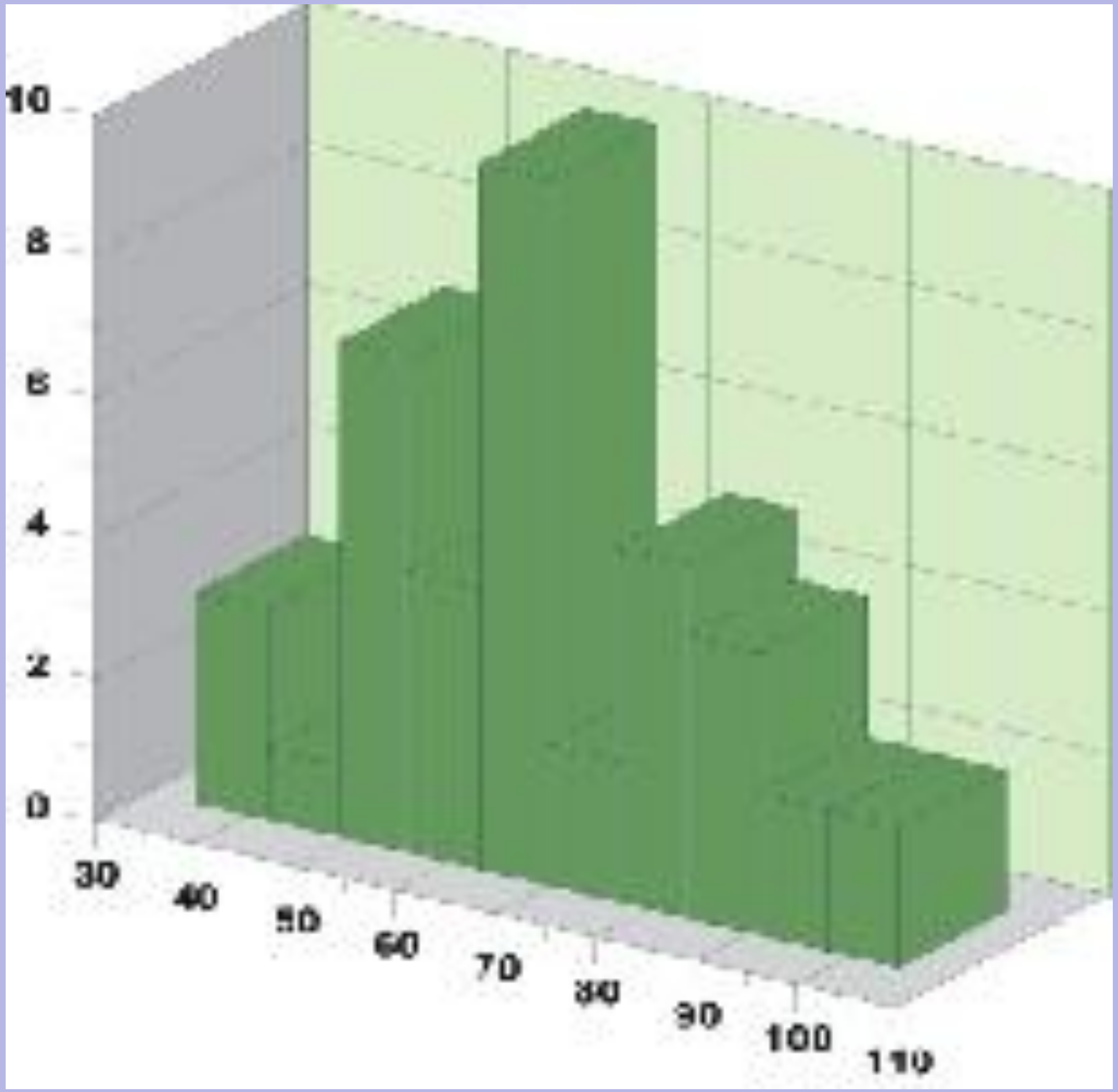
- Mostly graphical
- Plotting the raw data (histograms, scatterplots, etc.)
- Plotting simple statistics such as means, standard deviations, medians, box plots, etc.

How to make a CLEAR graph?

- Avoid 3-D graphics
- Don’t show too much information on the same graph (color, patterns, etc)
- Stay away from Excel, Excel is not a statistics package! It can induce errors (“copy-paste”)!

USE R: it provides a great environment for EDA with good graphics capabilities!!!

Is this a BAD PLOT?



- What's wrong with this one?

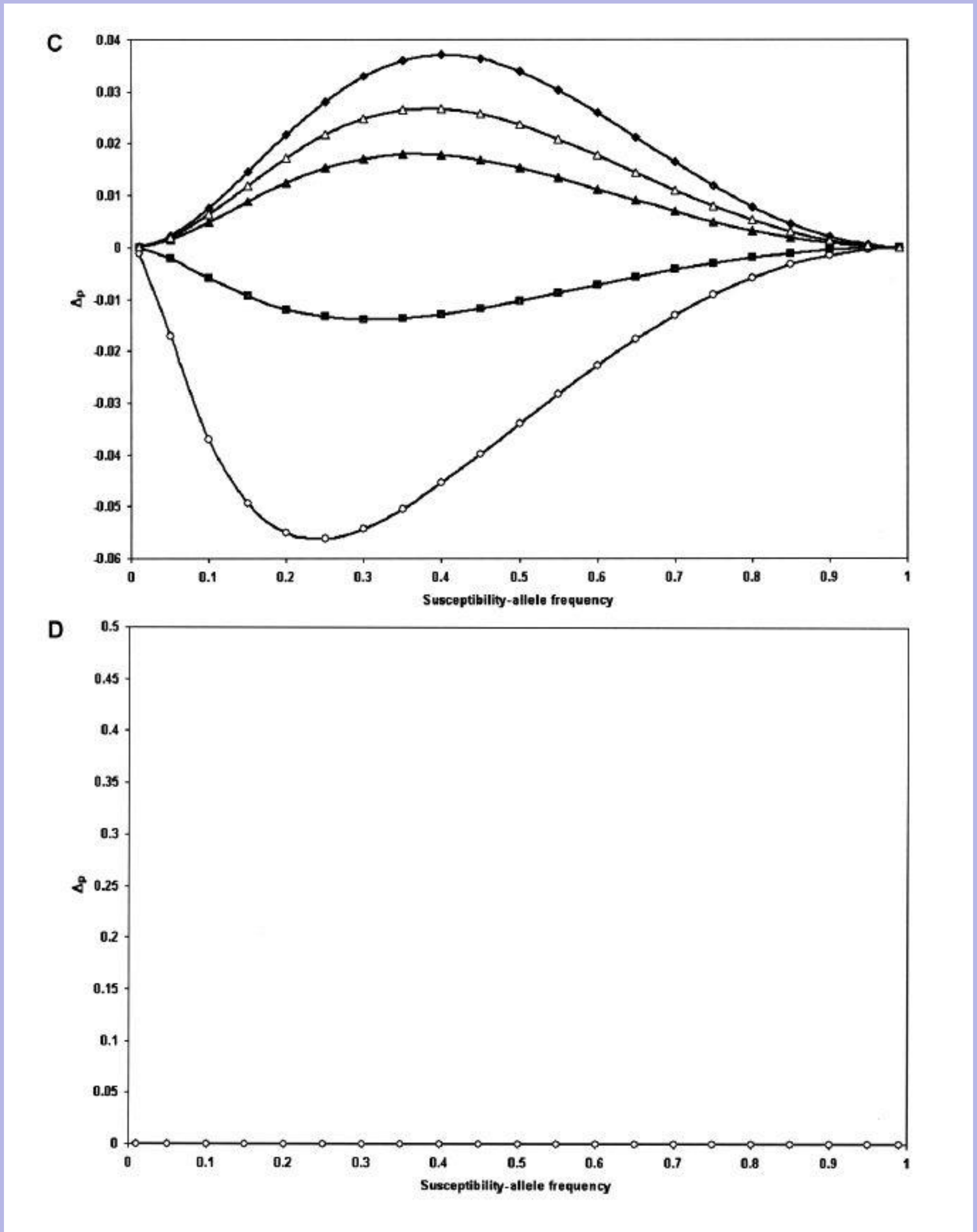
- What should have been done?

Do you want to see few BAD PLOTS?

Try

http://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005). American Journal of Human Genetics 76:967-986, Figure 1



What's wrong with this one?

If you want your graph to be published in large format, it seems that the American Journal of Human Genetics is the place to which you should aim. This figure spans two pages. Panel D is most interesting; it takes a while to identify that there is any information there at all.

What should have been done?

Panel D could have been discarded completely!

K Roeder (1994). Statistical Science 9(2):222-278, Figure 9

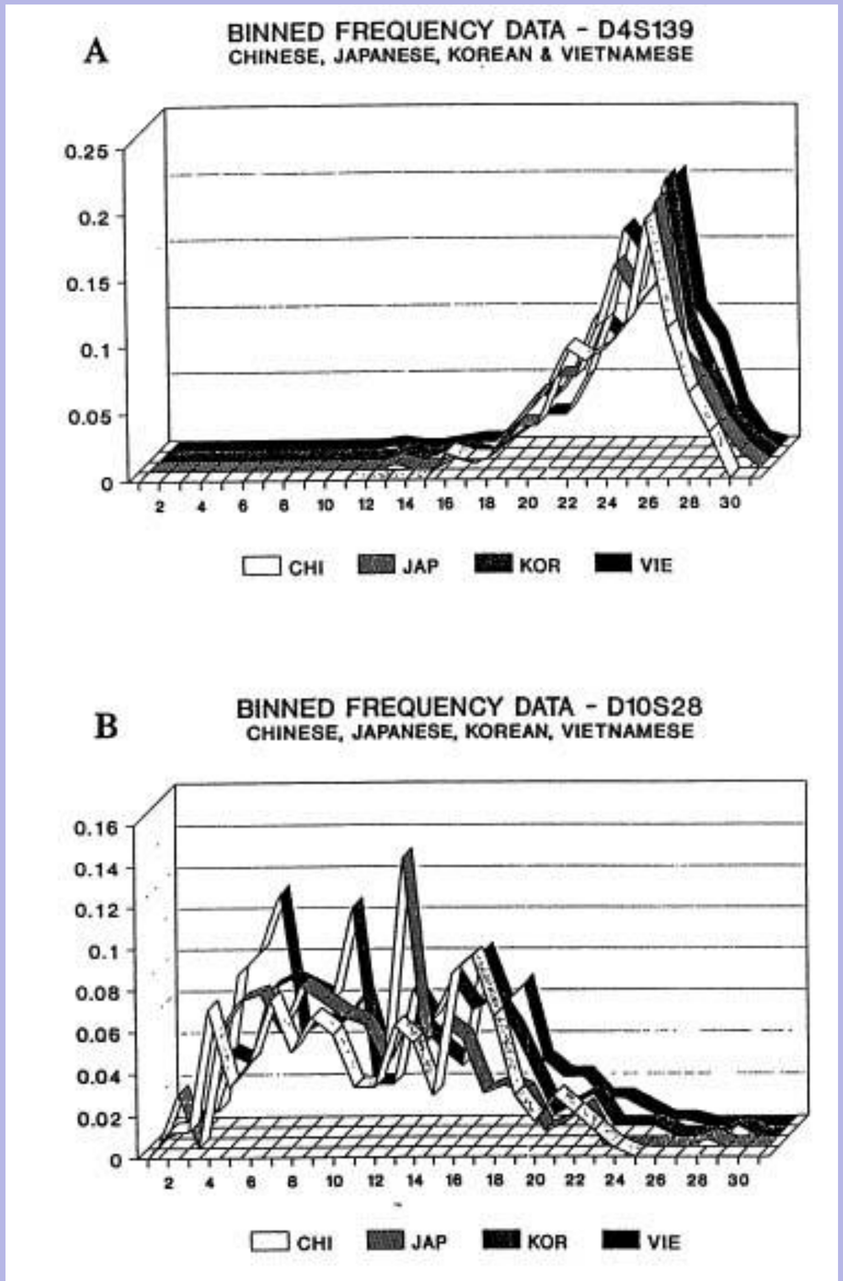


FIG. 4. *Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the 30 bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are* 103, 125, 93 *and* 215 *for* D4S139 *and* 120, 137, 100 *and* 193 *for* D10S28. *The horizontal axis is the bin number; bins are not of equal length.*

What's wrong with this one?

The 3-dimensional rendering of the curves is unnecessary.

What should have been done?

Displaying multiple curves simultaneously and ensure that the individual curves may be seen is challenging. Colors would be nice, but if color is not allowed, four different line types (solid, dashed, dotted, dash-dotted) might work.

Hummer BT, Li XL, Hassel BA (2001) Role for p53 in gene induction by double-stranded RNA. J Virol 75:7774-7777, Figure 4

What's wrong with this one?

The bars and little antennae represent just three data points each!

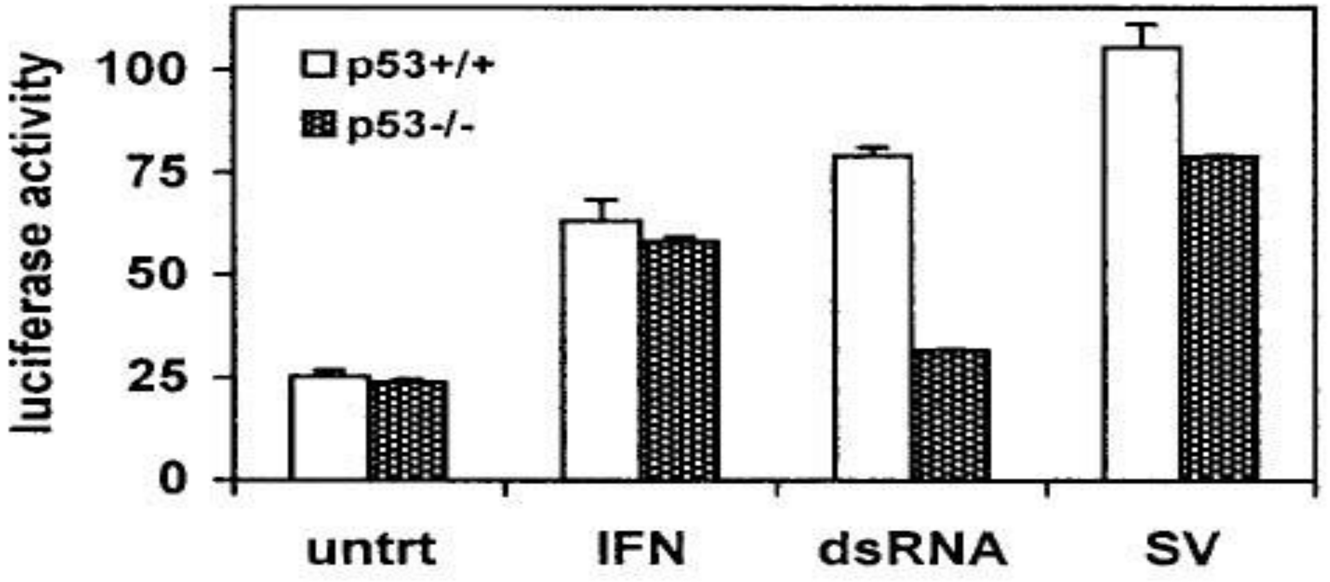


FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNA, and virus. Cells (6 × 10^5 HCT 116) were seeded in 32-mm plates and allowed to attach overnight. Cells were transfected with 500 ng of pGL3/ISG15-Luc, 50 ng of pRL null (Promega), and 450 ng of pcDNA3 for carrier DNA by using Lipofectamine Plus (Life Technologies) following the manufacturer's instructions. Twenty-four hours posttransfection, the medium was aspirated and replaced with medium containing either 1,000 U of IFN-α/ml, 50 μg of dsRNA/ml, or Sendai virus (multiplicity of infection, 10). Cells were incubated for 12 h and then lysed, and luciferase assays were performed. Luciferase activity was assessed on 20 μl of each lysate as directed by the supplier (Dual Luciferase Kit, Promega) using a TD 20/20 luminometer (Turner Designs). Luciferase activity is presented as the ratio of firefly activity to renilla activity to control for differences in transfection efficiency. Each data point is the mean of triplicate samples ± the standard error; the data presented are representative of four independent experiments.

Want more?

What should have been done?

With just three data points in each group, why not just show the data as dots? You could also include line segments at the averages and even confidence intervals...all this in the same amount of space and with less ink.

Why R?

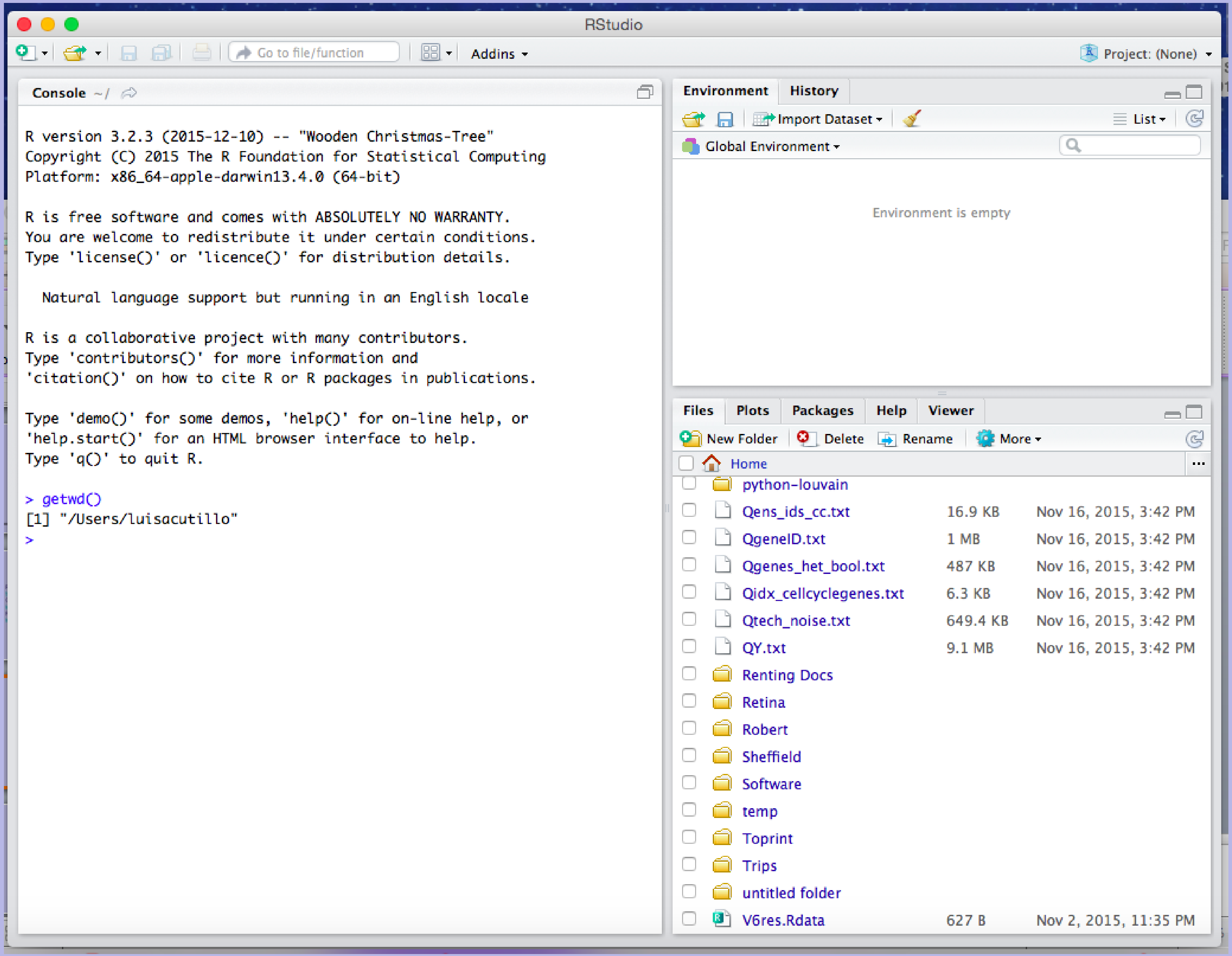
- R is Open source (http://www.r-project.org) language and environment for statistical computing and graphics
- Provide many statistical techniques
- R provides a great environment for EDA with great graphics capabilities
- Highly extensible (e.g. CRAN, Bioconductor)

Useful links and material

- R pacakge documentation for you research purposes: http://www.rdocumentation.org/
- R color charts: http://research.stowers-institute.org/efg/R/Color/Chart/index.htm
- Good page for R graphics and others: http://www.statmethods.net/advgraphs/index.html

- **Books used for this presentation:** https://cran.r-project.org/doc/contrib/Krijnen-IntroBioInfStatistics.pdf http://www.biostathandbook.com http://Rcompanion.org/

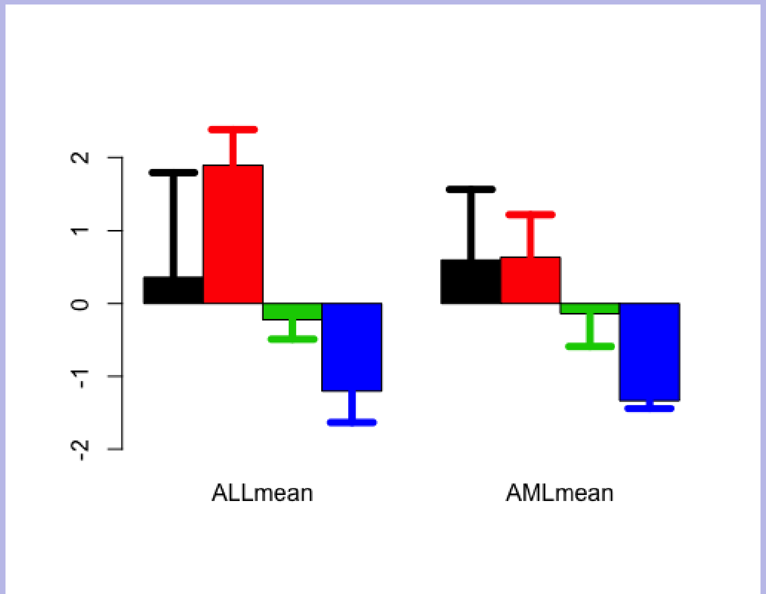
R and Rstudio



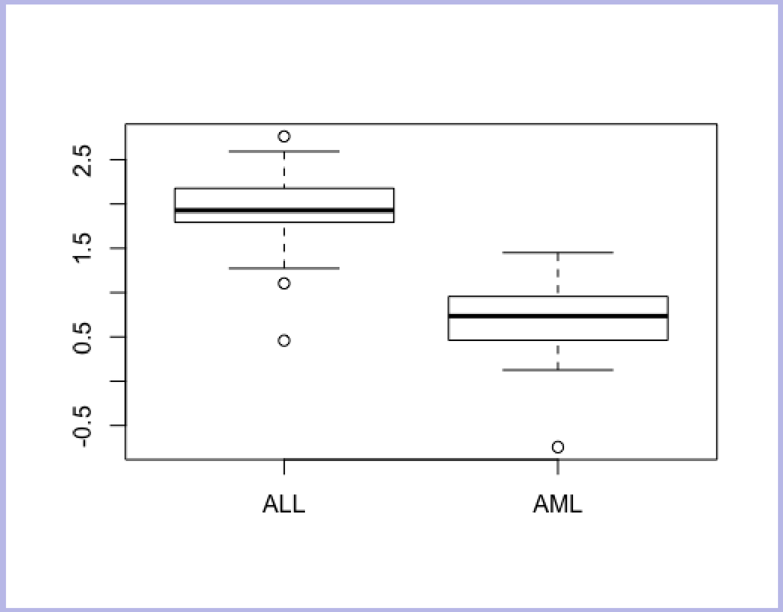
```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> getwd()
[1] "/Users/luisacutillo"
>
```

python-louvain		
Qens_ids_cc.txt	16.9 KB	Nov 16, 2015, 3:42 PM
QgeneID.txt	1 MB	Nov 16, 2015, 3:42 PM
Qgenes_het_bool.txt	487 KB	Nov 16, 2015, 3:42 PM
Qidx_cellcyclegenes.txt	6.3 KB	Nov 16, 2015, 3:42 PM
Qtech_noise.txt	649.4 KB	Nov 16, 2015, 3:42 PM
QY.txt	9.1 MB	Nov 16, 2015, 3:42 PM
Renting Docs		
Retina		
Robert		
Sheffield		
Software		
temp		
Toprint		
Trips		
untitled folder		
V6res.Rdata	627 B	Nov 2, 2015, 11:35 PM

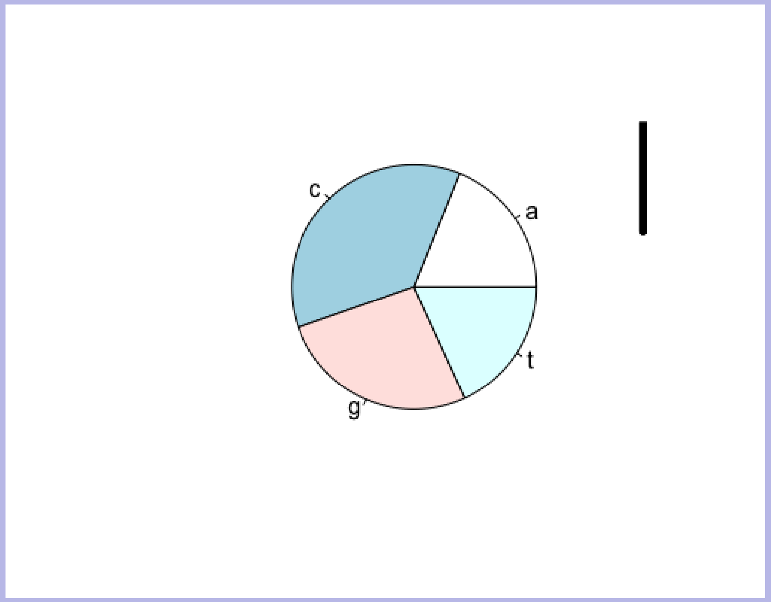
AIM:



Bar Plots

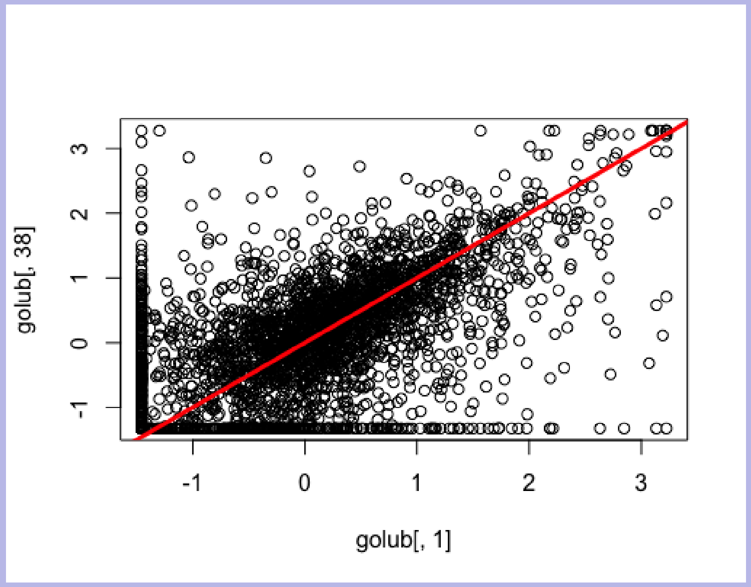


Box Plots

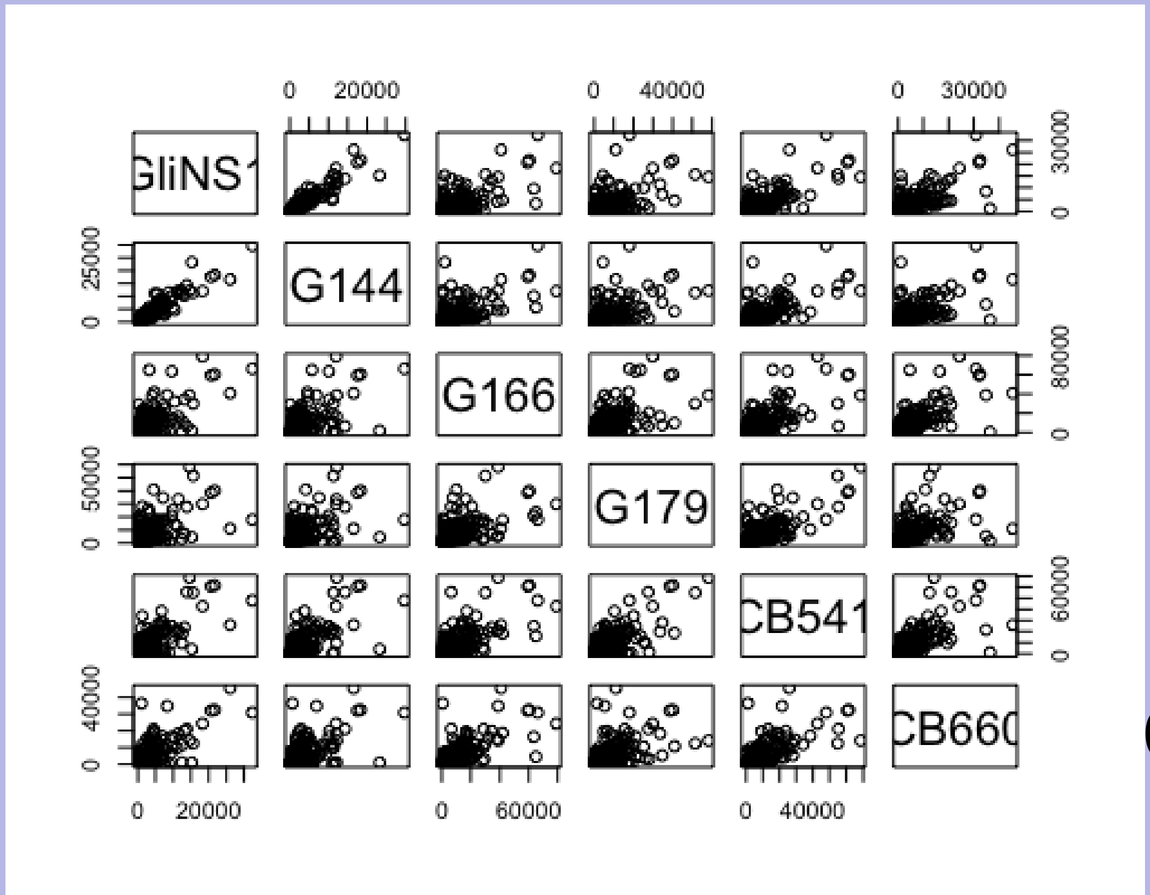


Pie Plots

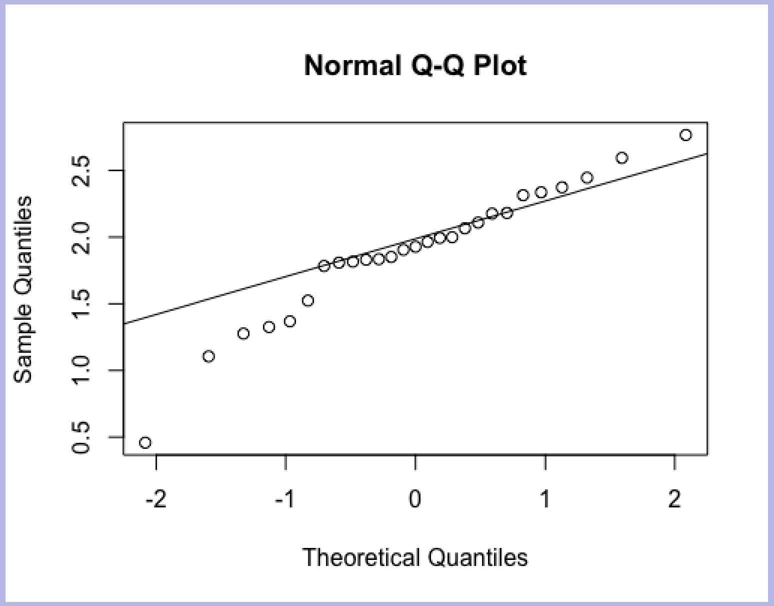
AIM:



Scatter Plots



Trellis Plots



Quantile-Quantile Plots

Instructions to follow the **Tutorial**

- Download the Tutorial.R
- Open Rstudio (or just R)

Few more **Comments**

- R script Comments.R

Probability distributions

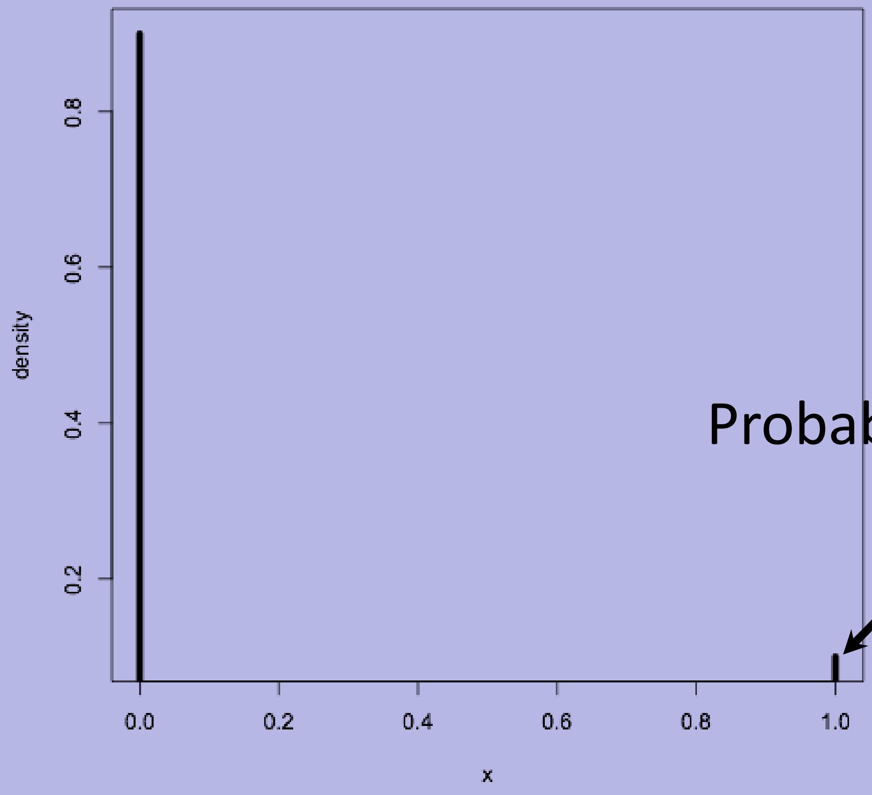
Can be either discrete or continuous (uniform, bernoulli, normal, etc)

Defined by a density function, *p*(x) or *f*(x)

Bernoulli distribution Be(p)

Flip a coin (T=0, H=1). Probability of H is .1.

```
x<-0:1
f<-dbinom(x, size=1, prob=.1)
plot(x,f,xlab="x",ylab="density",type="h",lwd=5)
```



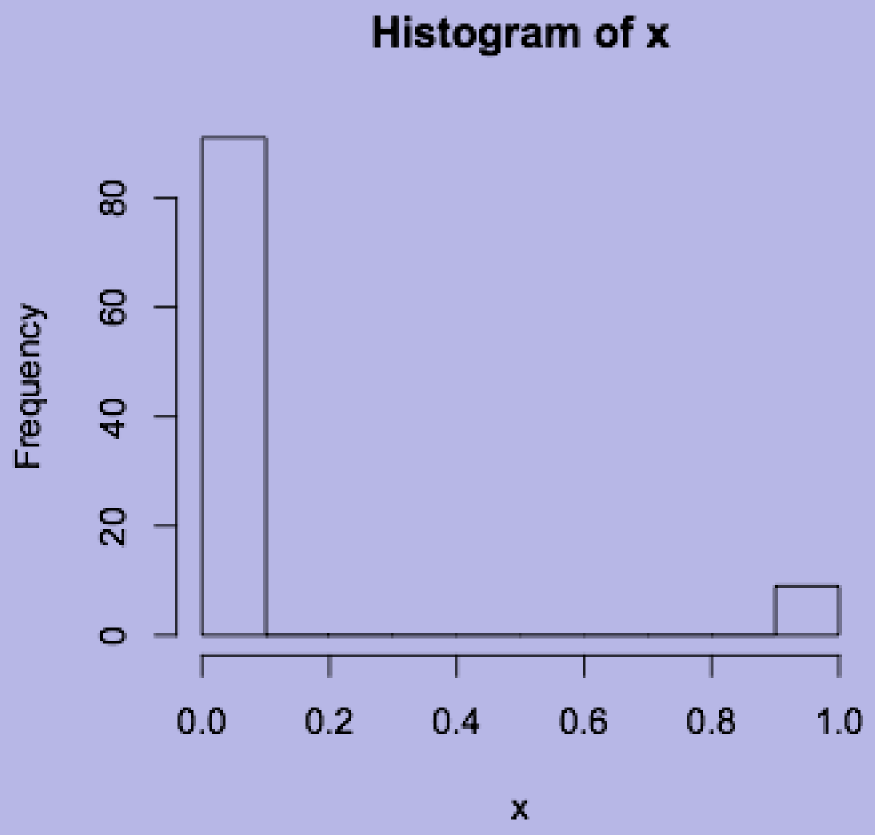
Probability of having a 1

Probability distributions

Random sampling

Generate 100 observations from a Be(.1)

```
set.seed(100)
x<-rbinom(100, size=1, prob=.1)
hist(x)
```

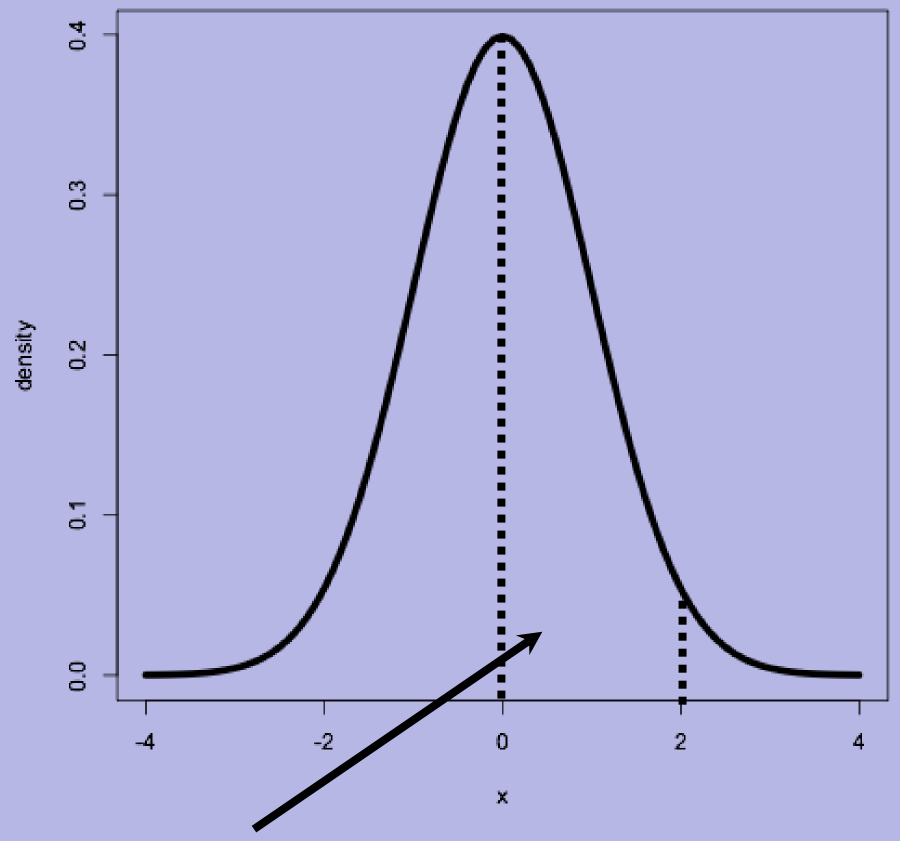


Probability distributions

Normal distribution $N(\mu,\sigma^2)$

μ is the mean and σ^2 is the variance

```
x<-seq(-4,4,.1)
f<-dnorm(x, mean=0, sd=1)
plot(x,f,xlab="x",ylab="density",lwd=5,type="l")
```



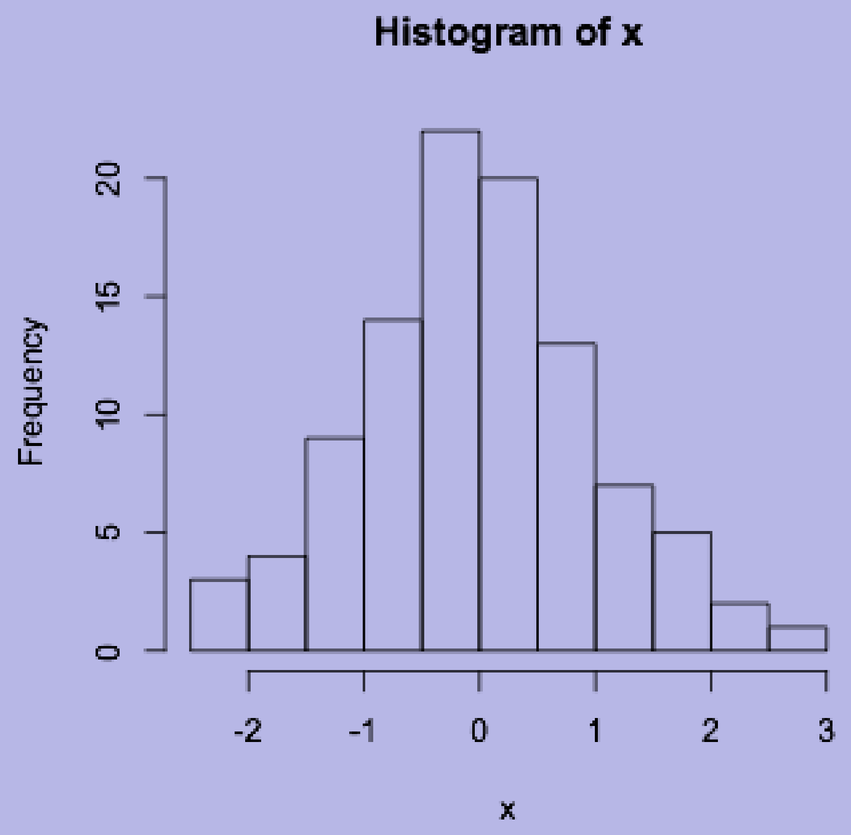
Area under the curve is the prob of having an observation beween 0 and 2.

Probability distributions

Random sampling

Generate 100 observations from a N(0,1)

```
set.seed(100)
x<-rnorm(100, mean=0, sd=1)
hist(x)
```



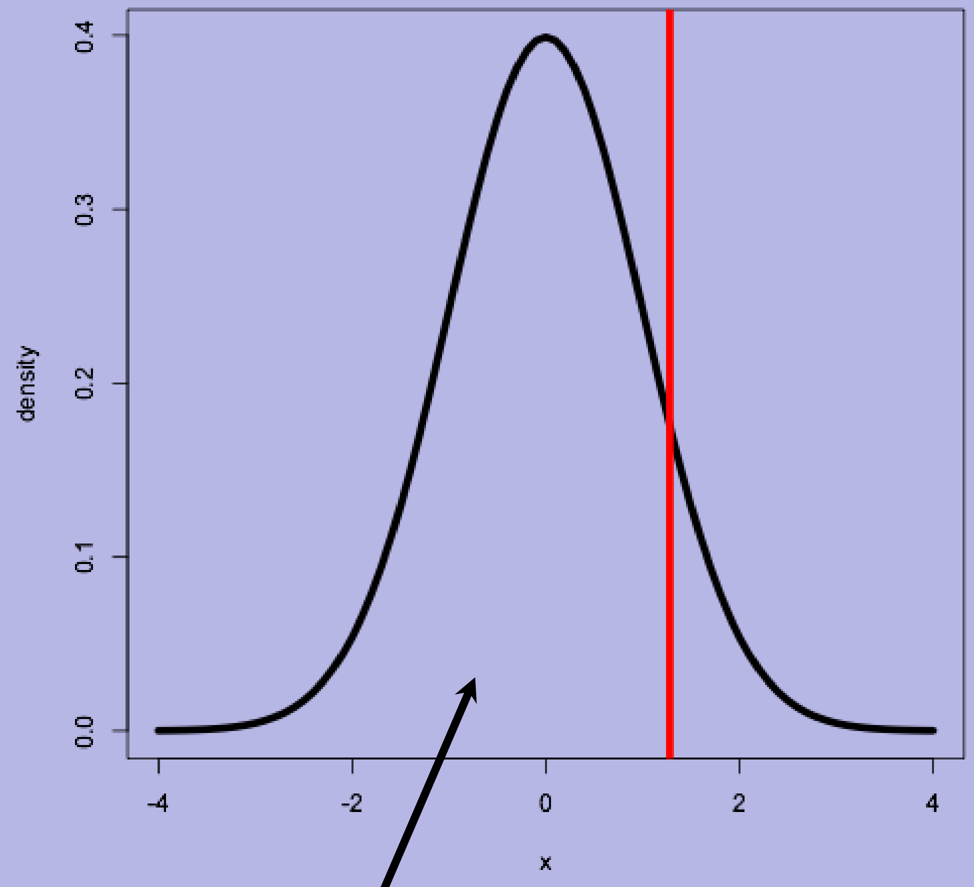
Histograms can be used to estimate densities!

Quantiles

(Theoritical) Quantiles: The *p*-quantile is the value with the property that there is a probability *p* of getting a value less than or equal to it.

```
q90<-qnorm(.90, mean = 0, sd = 1)
x<-seq(-4,4,.1)
f<-dnorm(x, mean=0, sd=1)
plot(x,f,xlab="x",ylab="density",type="l",lwd=5)
abline(v=q90,col=2,lwd=5)
```

The 50% quantile is called the median



90% of the prob. (area under the curve) is on the left of red vertical line.

Descriptive Statistics

Empirical Quantiles: The *p*-quantile is the value with the property that *p*% of the observations are less than or equal to it.

Empirical quantiles can easily be obtained in R.

```
set.seed(100)
x<-rnorm(100, mean=0, sd=1)
quantile(x)
```

```
        0%        25%        50%        75%       100% -2.2719255 -0.6088466 -0.0594199  0.6558911  2.5819589
```

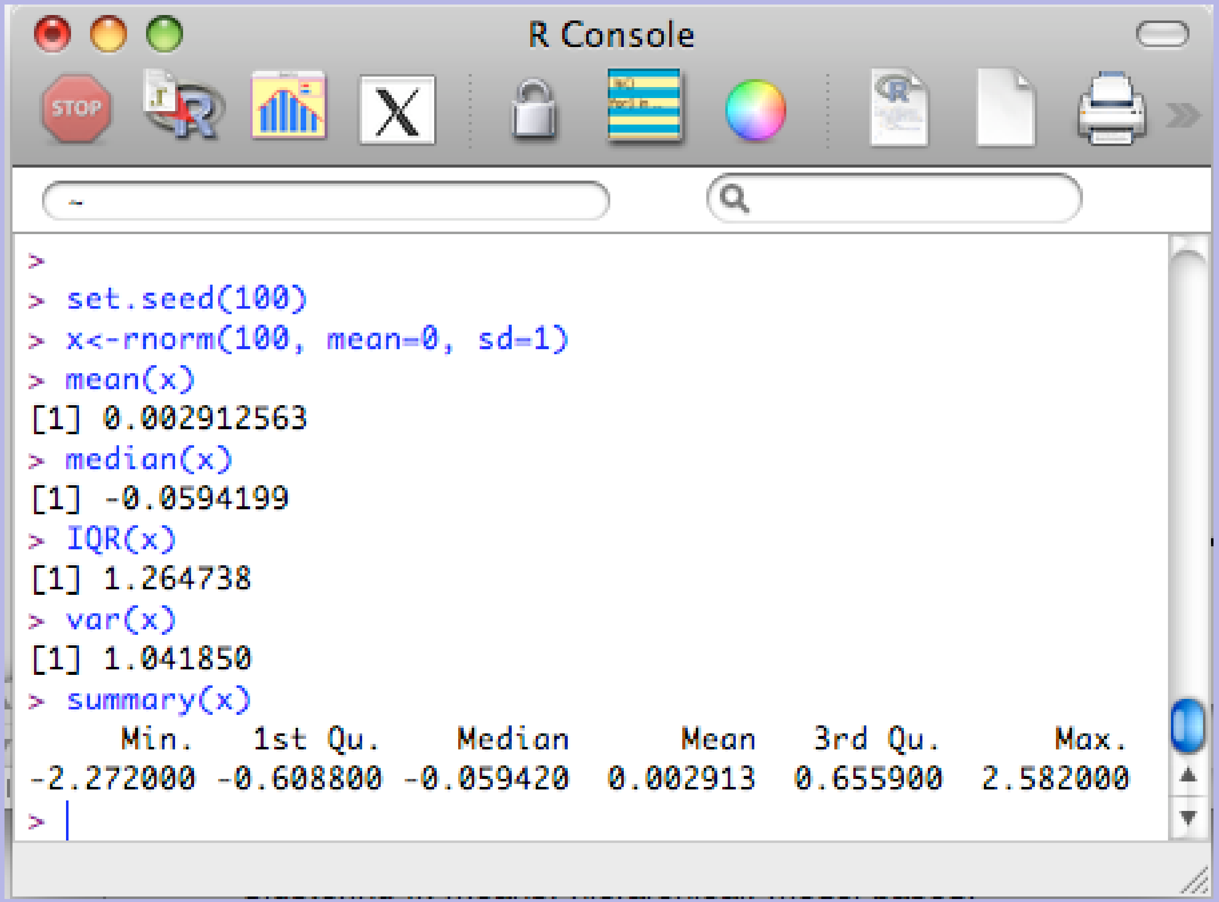
```
quantile(x,probs=c(.1,.2,.9))
```

```
       10%        20%        90% -1.1744996 -0.8267067  1.3834892
```

Descriptive Statistics

We often need to quickly 'quantify' a data set, and this can be done using a set of summary statistics (mean, median, variance, standard deviation)

```
set.seed(100)
x<-rnorm(100, mean=0, sd=1)
mean(x)
median(x)
IQR(x)
var(x)
summary(x)
```

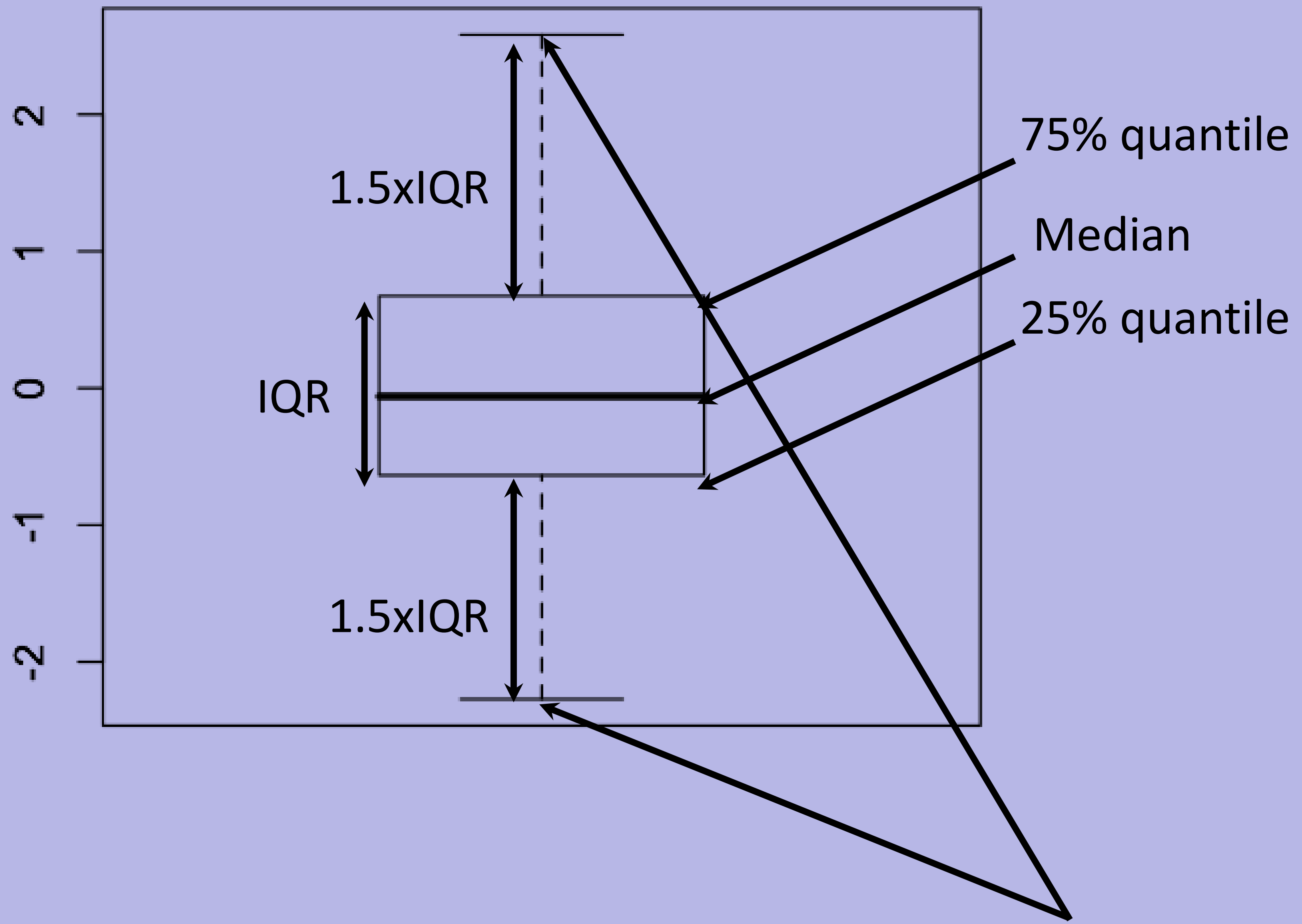


```
>
> set.seed(100)
> x<-rnorm(100, mean=0, sd=1)
> mean(x)
[1] 0.002912563
> median(x)
[1] -0.0594199
> IQR(x)
[1] 1.264738
> var(x)
[1] 1.041850
> summary(x)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
-2.272000 -0.608800 -0.059420  0.002913  0.655900  2.582000 
> 
```

'summary' can be used for almost any R object!
R is object oriented (methods/classes).

Descriptive Statistics - Box-plot

```
set.seed(100)
x<-rnorm(100, mean=0, sd=1)
boxplot(x)
```



75% quantile

Median

25% quantile

1.5xIQR

IQR

1.5xIQR

IQR= 75% quantile -25% quantile= Inter Quantile Range

Everything above or below are considered outliers

QQ-plot

- Many statistical methods make some assumption about the distribution of the data (e.g. Normal).

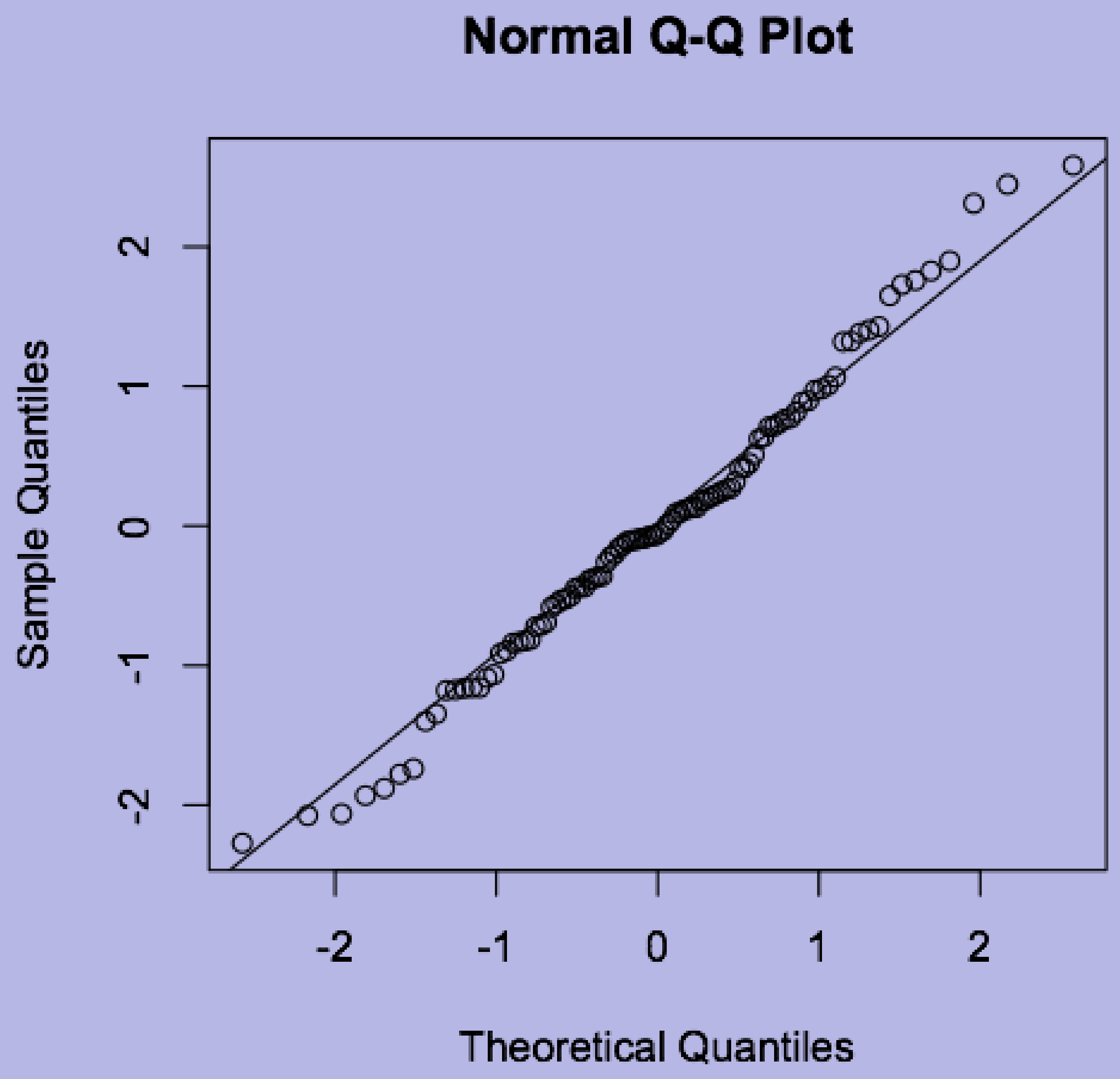
- The quantile-quantile plot provides a way to visually verify such assumptions.

- The QQ-plot shows the theoretical quantiles versus the empirical quantiles. If the distribution assumed (theoretical one) is indeed the correct one, we should observe a straight line.

QQ-plot

```
set.seed(100)
x<-rnorm(100, mean=0, sd=1)
qqnorm(x)
qqline(x)
```

Only valid for the normal distribution!

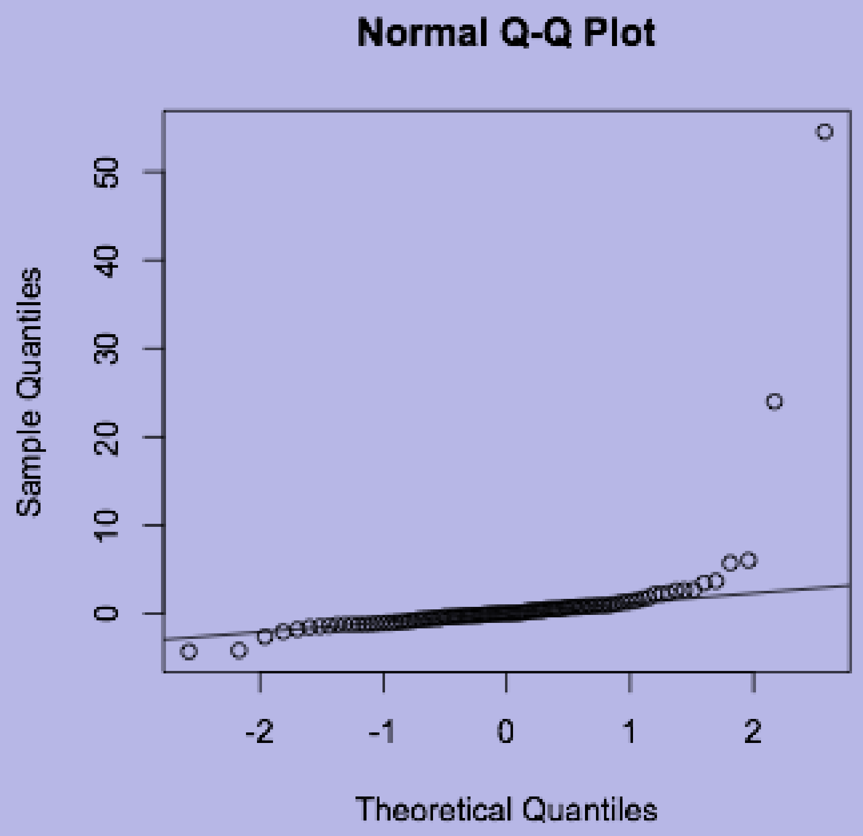

Normal Q-Q Plot

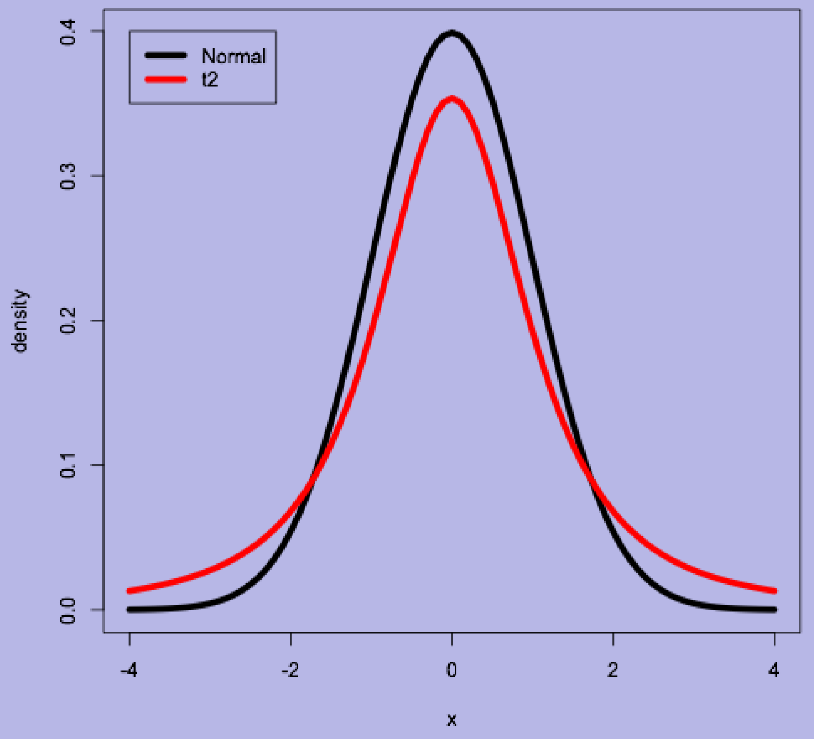
QQ-plot

```
set.seed(100)
x<-rt(100,df=2)
qqnorm(x)
qqline(x)
```

Clearly the *t* distribution with two degrees of freedom is different from the Normal!

```
x<-seq(-4,4,.1)
f1<-dnorm(x, mean=0, sd=1)
f2<-dt(x, df=2)
plot(x,f1,xlab="x",ylab="density",lwd=5,type="l")
lines(x,f2,xlab="x",ylab="density",lwd=5,col=2)
```

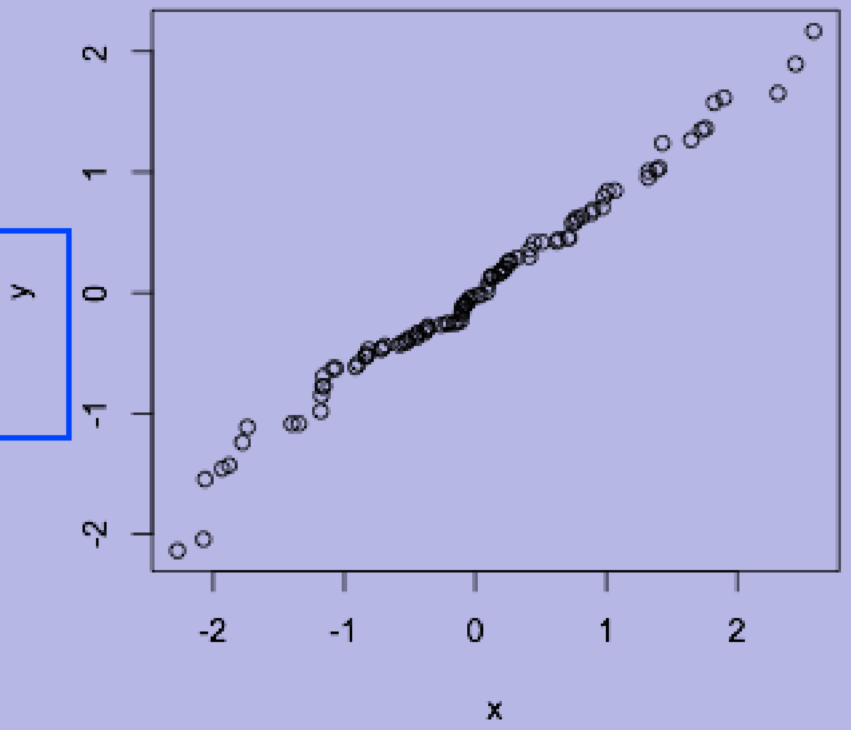




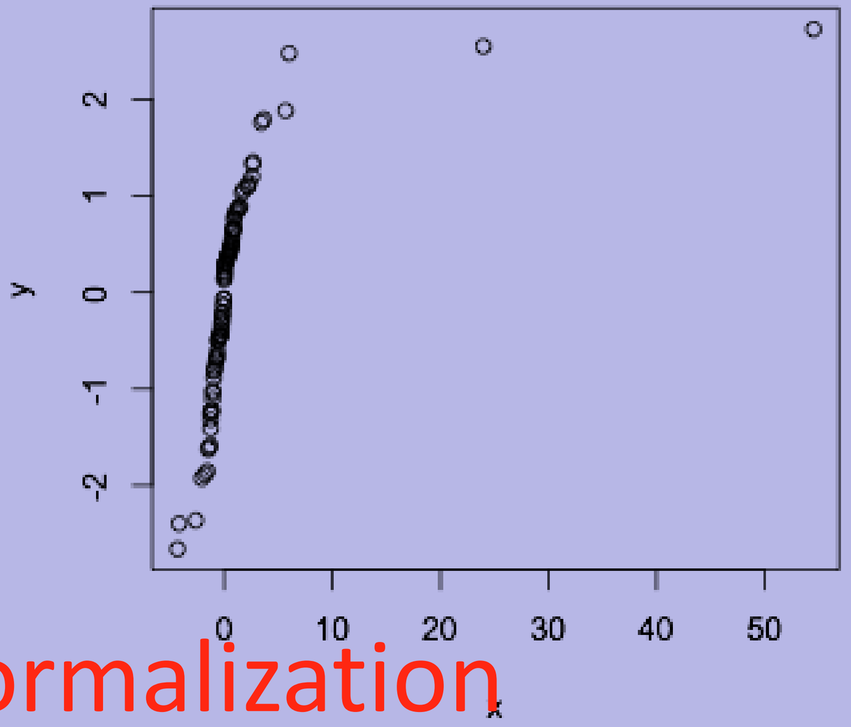
QQ-plot

Comparing two samples

```
set.seed(100)
x<-rnorm(100, mean=0, sd=1)
y<-rnorm(100, mean=0, sd=1)
qqplot(x,y)
```



```
set.seed(100)
x<-rt(100, df=2)
y<-rnorm(100, mean=0, sd=1)
qqplot(x,y)
```



Ex: Try with different values of df.

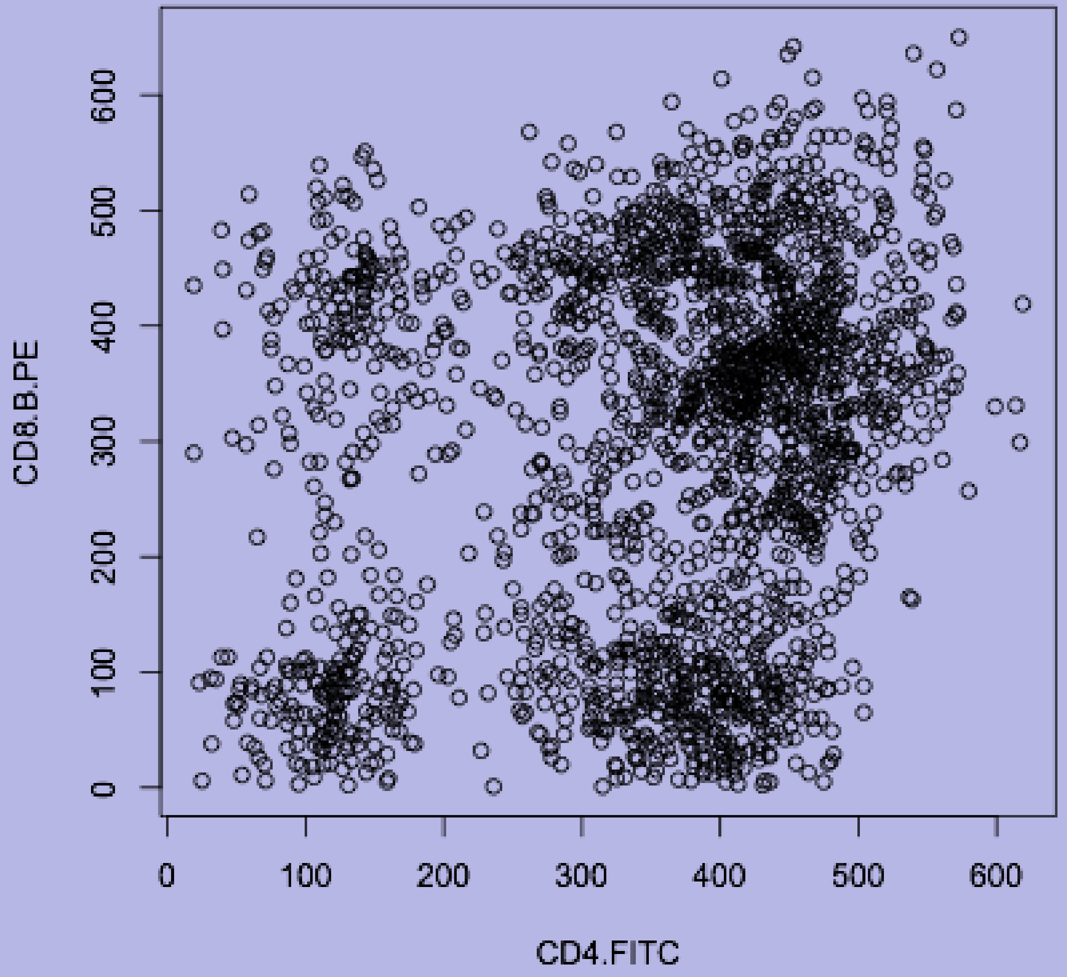
Main idea behind quantile normalization

Scatter plots

Biological data sets often contain several variables
So they are multivariate.

Scatter plots allow us to look at two variables at a time.

What can you tell about this data?



This can be used
to assess independence!

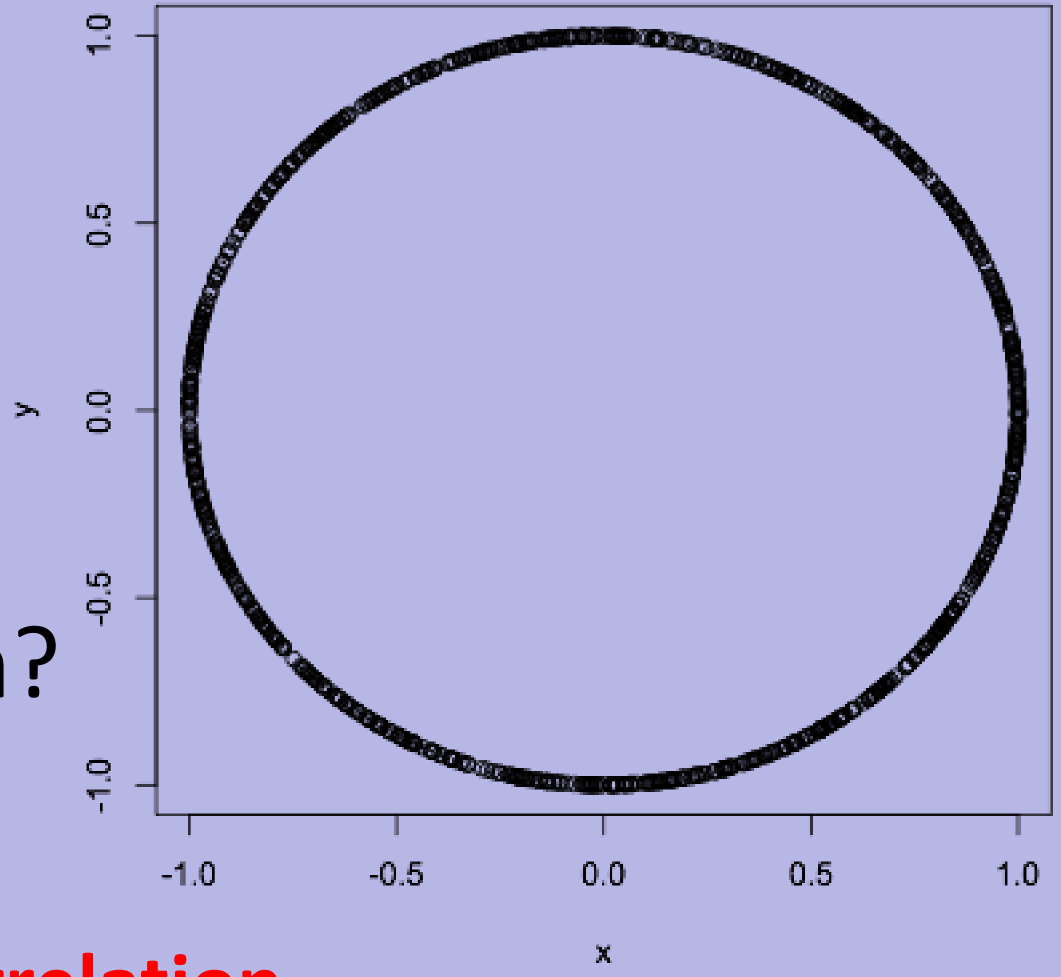
Scatter plots vs. correlations

Note that in the previous example, the correlation between The two variables was 0.23!

Correlation is only good for linear dependence.

```
# Quick comment on correlation
set.seed(100)
theta<-runif(1000,0,2*pi)
x<-cos(theta)
y<-sin(theta)
cor(x,y)
plot(x,y)
```

What is the correlation?



Independence -> Non Correlation

Not viceversa!!!!

Module II Outline:

- Hypothesis Testing Methodology
- p-Value Approach to Hypothesis Testing
- Comparative Statistics examples (T-Test, Anova and Statistics for frequency data)
- Tutorials: Ttest.R, Frequancy.R, Anova.R
- Multiple Hypotesis Testing (FDR)
- Application to Golub dataset: MulHypTestOnGolub.R

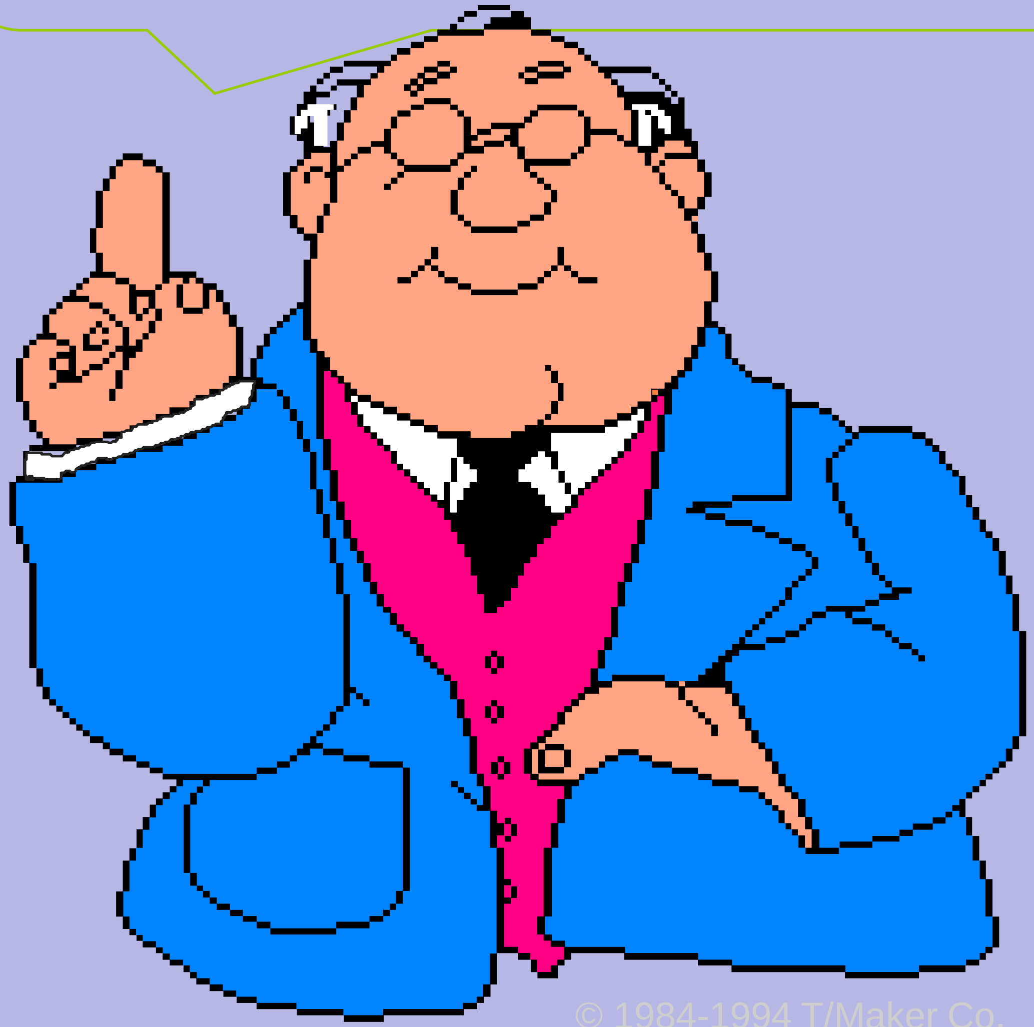
What is a Hypothesis

Of a test?

- *A hypothesis is an assumption about the population parameter.*
  - A parameter is a characteristic of the population, like its mean or variance.
  - The parameter must be identified before analysis.

I assume the mean AGE of this class is 50!!!

Am I correct? TEST IT!



© 1984-1994 T/Maker Co.

The Null Hypothesis, H_0

- States the Assumption (numerical) to be tested

e.g. Our class mean age is 50 (H_0: μ=50)

- Begin with the assumption that the null hypothesis is TRUE.

(Similar to the notion of innocent until proven guilty)

The Null Hypothesis may or may not be rejected,but our aim is to REJECT the null hypothesis!

The Alternative Hypothesis, H_1

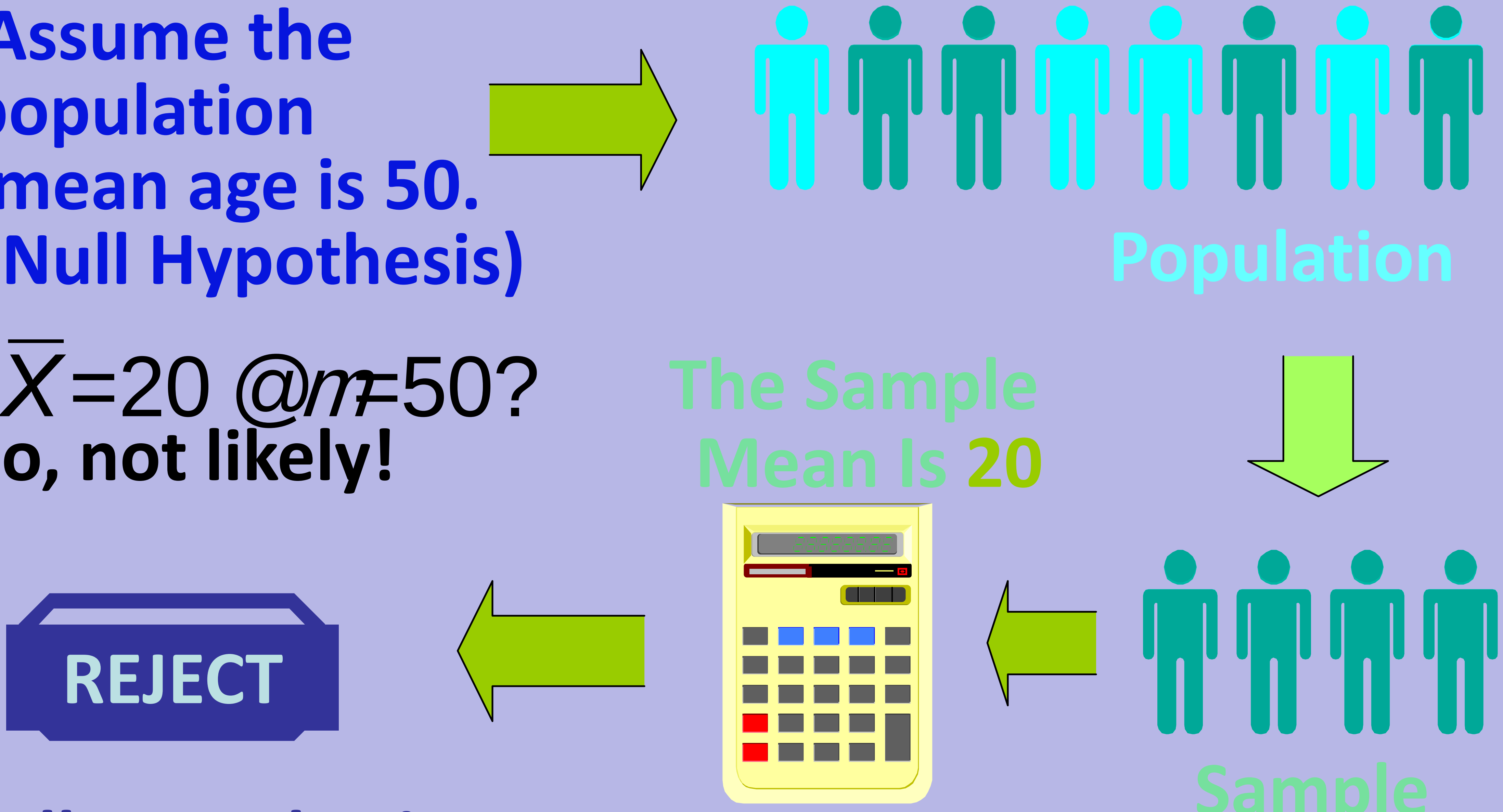
- Is the opposite of the null hypothesis
  e.g. The average age of our class is different from 50 (H_1: $\mu \neq 50$)

- Is generally the hypothesis that is believed to be true by the researcher!

Identify the Problem

- Steps:
  - State the Null Hypothesis
  - State its opposite, the Alternative Hypothesis
    - Hypotheses are mutually exclusive & exhaustive
    - Sometimes it is easier to form the alternative hypothesis first.

Hypothesis Testing Process



Assume the population mean age is 50. (Null Hypothesis)

Population

Is $\bar{X}=20$ @m=50?

No, not likely!

The Sample Mean Is 20

Sample

REJECT

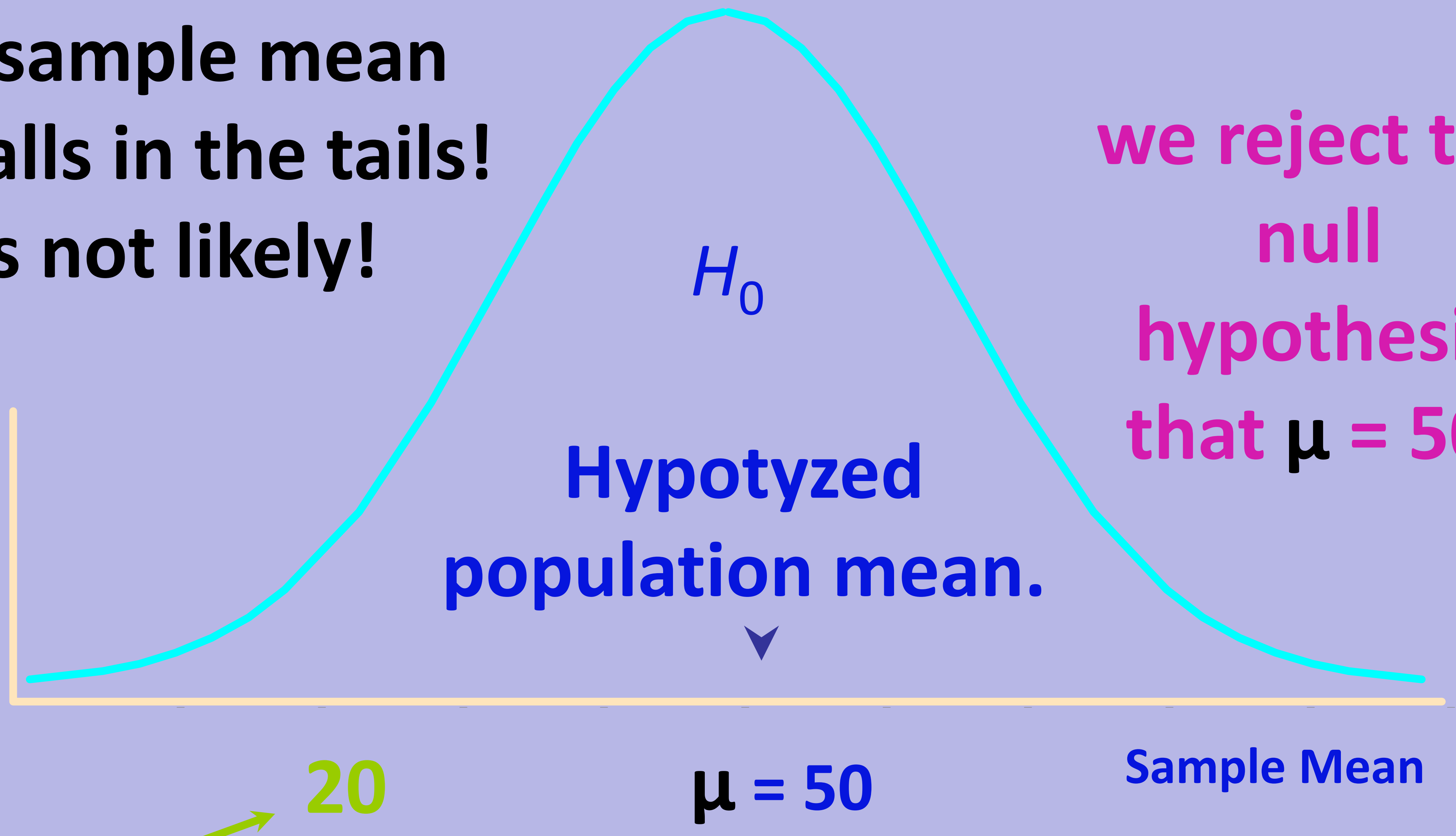
Null Hypothesis

Reason for Rejecting H_0

Sampling Distribution

Our sample mean (20) falls in the tails! It's not likely!

we reject the null hypothesis that μ = 50.



Observed population mean

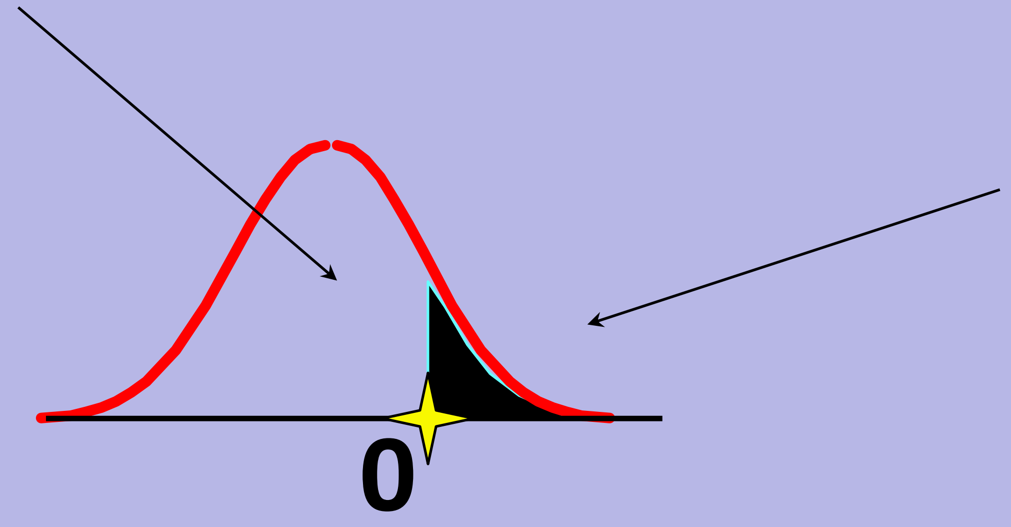
Level of Significance, α

- Defines the Rejection region

- Typical value of *α* is 0.05. It Provides the Critical Value(s) of the Test

Rejection Regions

✦ Critical Value



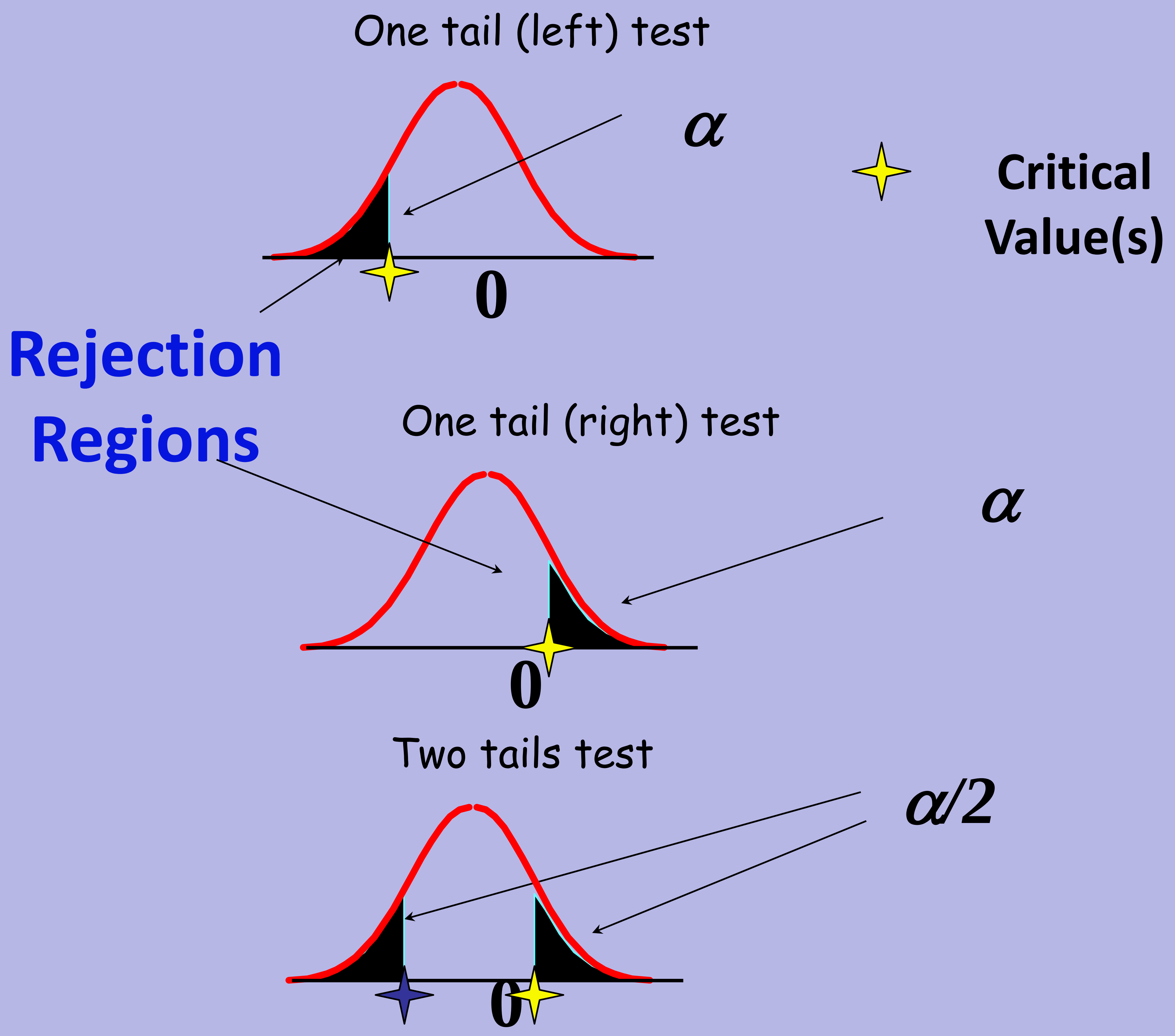
α "Area" of the Rejection region

Level of Significance, α and the Rejection Region

H_0: μ ⍰ 0
H_1: μ < 0

H_0: μ ⍰ 0
H_1: μ > 0

H_0: μ = 0
H_1: μ ≠ 0



One tail (left) test

One tail (right) test

Two tails test

Rejection Regions

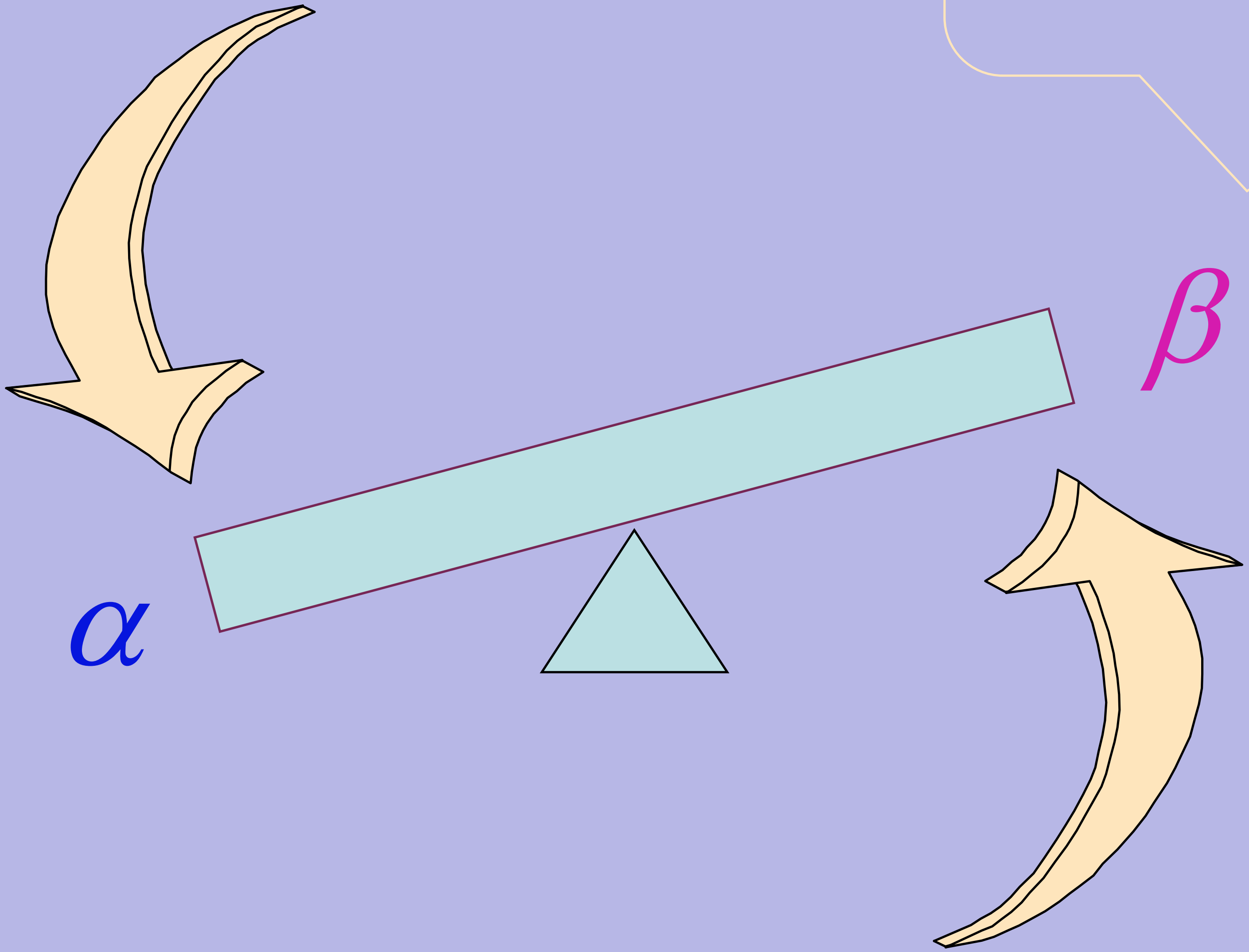
Critical Value(s)

Errors in Making Decisions

- Type I Error
  - Reject Null Hypothesis when it is True (“False Positive”)
  - Has Serious Consequences
  - Probability of Type I Error Is α
    - Called Level of Significance
- Type II Error
  - Do Not Reject Null Hypothesis when it is False (“False Negative”)
  - Probability of Type II Error Is β (Power $1 - \beta$)

α & β Have an Inverse Relationship

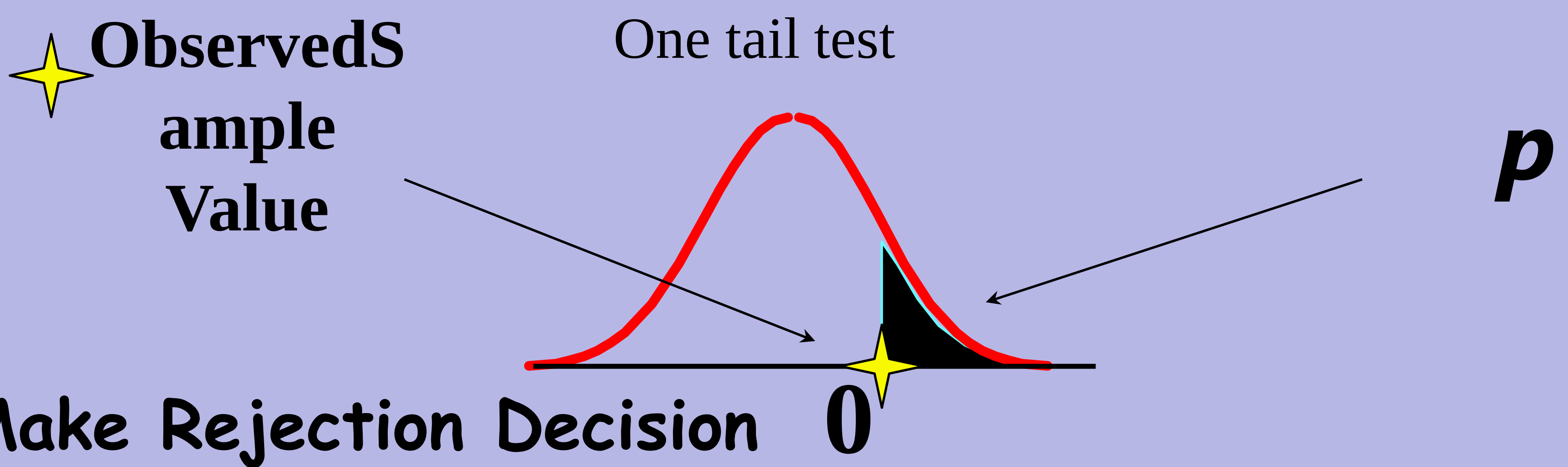
Reduce probability of one error and the other one goes up.



One possibility: Increase the sample size!!!!

What is the *p* Value and how to use it in a Test?

- The p-value is the Probability of Obtaining a Test Statistic (under H_0) more Extreme (≤ or ≥) than the observed Sample Value



- Used to Make Rejection Decision

  - If *p* value < α ⍰⍰⍰⍰ Reject H_0 ⍰⍰⍰⍰ SUCCESS

  - If *p* value ≥ α⍰ ⍰⍰⍰⍰ Do Not Reject H_0 ⍰⍰⍰ FAILURE

Random variables: am I observing continuous or discrete data???

Roughly speaking

a “random” variable is a quantity whose values are “random” and to which a probability distribution is assigned

(e.g. a fair dice outcomes have same chance of coming up at each throw) ;

THE DIFFERENCE BETWEEN CONTINUOUS AND DISCRETE VARIBLES IS FUNDAMENTAL IN CHOOSING THE KIND OF TEST STATISTICS!

Discrete R.V.

If the r.v. X values belongs to a finite set
{x1 ,x2,…, xn}
then X is called DISCRETE (**usually counts**)

As example the flipping of a coin, the number of red cells counted in an image, the number of success in 100 trials…are observations of a discrete variable!

Continuous R.V.

A ***continuous random variable*** is a r.v. which takes an infinite number of possible values.

Continuous random variables are usually measurements. Examples include height, weight, the amount of sugar in an orange, the time required to run a mile,the fluorescence intensity in a microarray, etc.

(A continuous random variable is not defined at specific values
bat over intervals of values)

Which test to use?

First of all you should choose a summary **SAMPLE STATISTIC**!

As. Example:

T-test!

SAMPLE MEAN

$$\overline{X} = \frac{1}{n}\sum_i x_i$$

SAMPLE VARIANCE

$$S^2 = \frac{1}{n-1}\sum_i (x_i - \overline{X})^2$$

SAMPLE COVARIANCE

$$cov(X,Y) = \frac{1}{n-1}\sum_i (x_i - \overline{X})(y_i - \overline{Y})$$

SAMPLE CORRELATION

$$corr(X,Y) = \frac{\sum_i (x_i - \overline{X})(y_i - \overline{Y})}{(n-1)S_x S_y}$$

paired t-Test: s Unknown

(rigth and left eye)

- Assumptions
  - Population is normally distributed
  - If not normal, only slightly skewed & a large sample taken (Central limit theorem applies)
- Parametric test procedure (sample stat. is the sample mean!)
- t test statistic, with n-1 degrees of freedom

$$t = \frac{\bar{X} - m}{S/\sqrt{n}}$$

Rejection Region

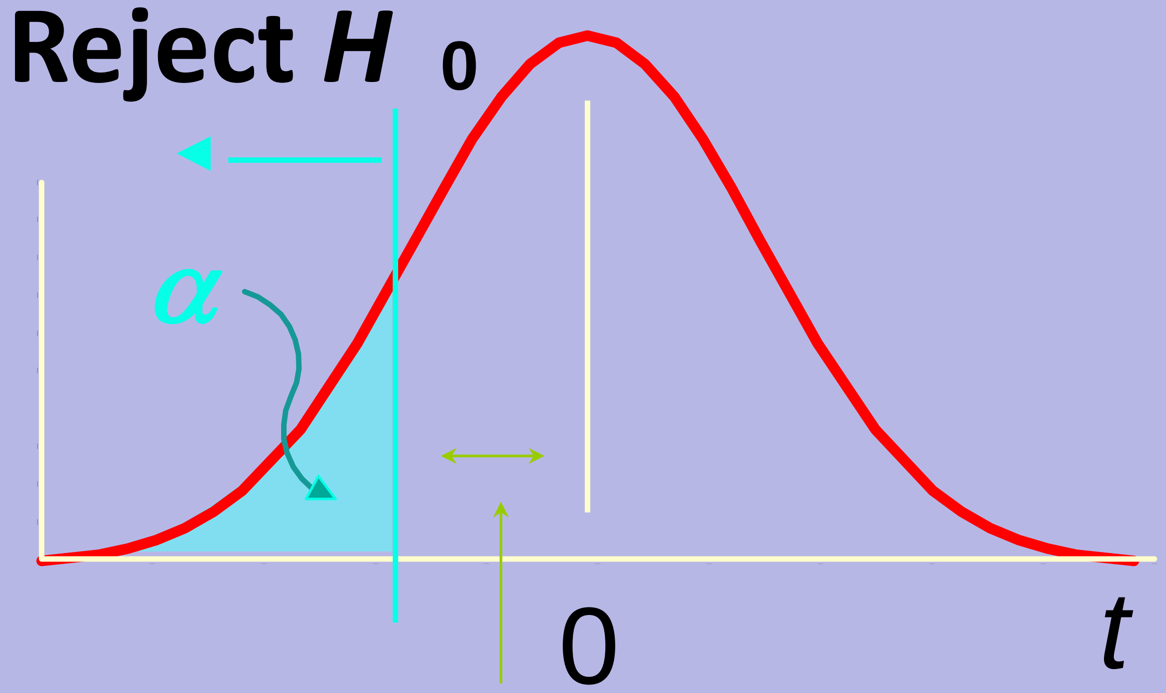
(one tail)

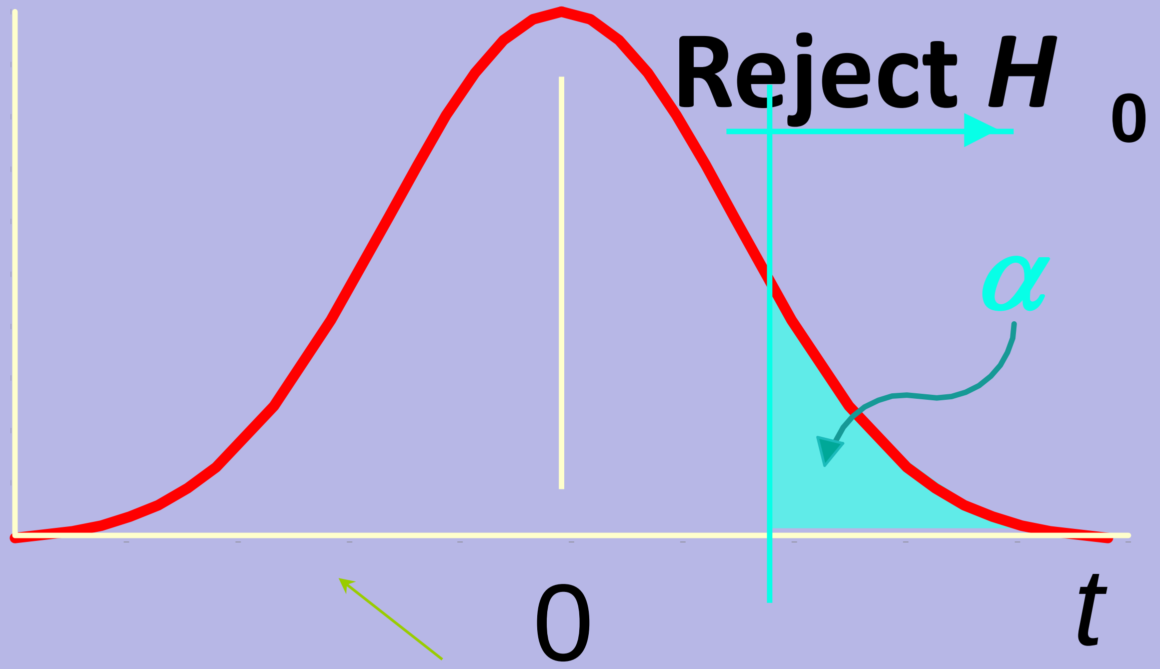
H_0: $\mu \quad \overline{0}$

H_1: $\mu < 0$

H_0: $\mu \quad \overline{0}$

H_1: $\mu > 0$





Must Be *Significantly* below $\mu = 0$

Small values don’t contradict H_0 ->Don’t Reject H_0!

Unpaired T-test

$$t = \frac{\bar{X} - \bar{Y}}{S / \sqrt{n}}$$

- The two sample observations are not coupled
- Not necessary equal sample numbers
- You may distinguish between equal and unequal sample variance

In few words the other tests:

- If you want to compare more then two populations means when you observe 1 characteristic: **one way Anova Test**
- If you want to compare more then two populations means when you observe 2 characteristic: **two way Anova Test**
- If you want to compare two populations variance: **F-test**
- If you want to compare two populations proportions: **Chi-square test**

Remarks

- If you have counts…or few data YOU ARE NOT ALLOWED TO USE T-TEST!!!
- Any test is build upon conjecture about the shape of the null distributions
- If you just want to have a summary about your data, then use the descriptive statistic

Ttest

- Example dataset: Gene expression data (3051 genes and 38 tumor mRNA samples) from the leukemia microarray study of Golub et al. (1999). Pre-processing was done as described in Dudoit et al. (2002).
- Tutorial: Ttest.R

Anova

Use one-way anova when you have one nominal variable and one measurement variable; the nominal variable divides the measurements into two or more groups.

- It tests whether the means of the measurement variable are the same for the different groups
- Tutorial: Anova.R

Statistics for frequency data

- **Sometimes in biology results are not measurements but counts (or frequencies)!**e.g. counts of different phenotypes, counts of cell types ...

- **Task**: Compare frequency data in different categories with some expected data

- You are NOT ALLOWED to perform a t test! Instead you do a **Chi-squared test**;

Statistics for frequency data

- Three different uses:

  - Expected calculated from theory: you test if your observed data agree with the theory. E.g. Mendel theory can be used to predict frequencies of different phenotypes: we expect a genetic cross to be 3:1 ratio of red and white flowers.(P>5% data agree with theory)

  - Expected calculated assuming that the counts in all the categories should be the same: you test whether there is a difference between the observed sets. (P<5% data significantly different from each other)

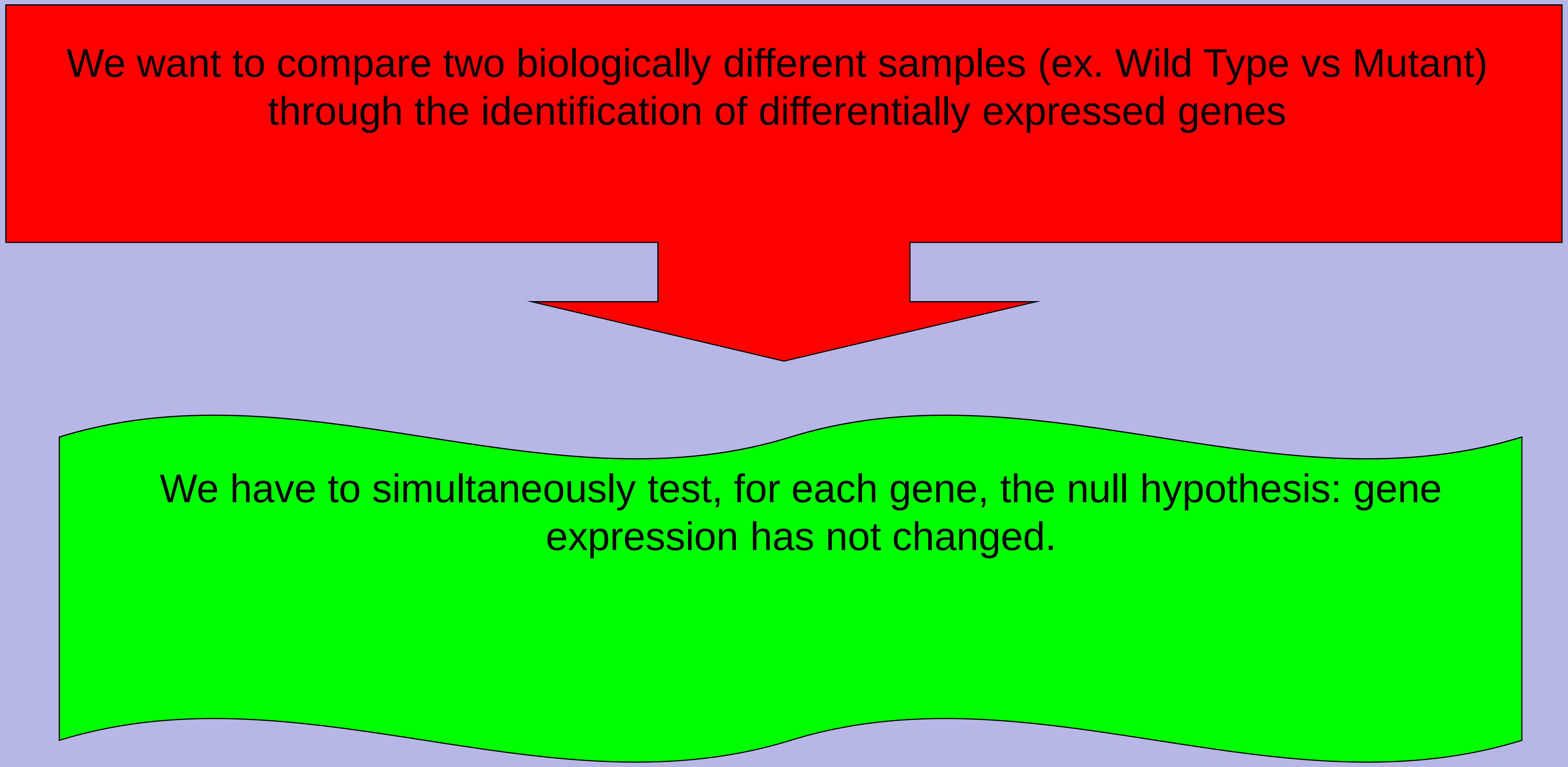
  - Investigate association between frequency data in two separate groups. Expected calculated assuming counts in one group are not affected by counts in the other. (P<5% there is a significant association). Data are set in a *contingency* table. For each cell the expected data is:

    E=(column total x row total)/grand total

Statistics for frequency data

- **chi-square test of goodness-of-fit**: when you have **one nominal variable**, you want to see whether the number of observations in each category fits a theoretical expectation, and the sample size is large.
- **Exact test of GoF**: same but the sample size is small.
- **Chi-square test of independence**: when you have **two nominal variables** and you want to see whether the proportions of one variable are different for different values of the other variable. (sample size is large).
- **Fisher Exact test of independence:** same but sample size is small.
- **Tutorial:** **Frequancy.R**

MICROARRAY Hypothesis Testing as an example of Multiple Hypothesis Testing



We want to compare two biologically different samples (ex. Wild Type vs Mutant) through the identification of differentially expressed genes

We have to simultaneously test, for each gene, the null hypothesis: gene expression has not changed.

For each gene j the test is expressed in term of a Statistic and a p-value

Null Hypothesis
H0 for each gene j:
μj(WT)=μj(KO)

H_{0j}

Which is the test to use in this case?

For each gene the test is expressed in term of a Statistic and a p-value

Null Hypothesis
Ho: μj(WT)=μj(KO) H_{0j}

T-statistic on gene j --> p-value

p-value $p_j = Pr(|T_j| > |t_j| \mid H_{0j}$ **Is true** $), j = 1, \ldots, m$

$p_j \leq Tr$ (α)

Reminder:
The p-value is the probability of finding a false positive (probability of type I error) that is the probability of finding out a differentially expressed gene that actually is not!!!

Ex. If α=0.01 and p<α, then 0.01 represent the probability that the gene detected is a false positive.

Problems in controlling the errors…

Assume that a chip experiment reveals the expression level of m = 20.000 genes relatively to two different biological conditions.

We want to test, simultaneously for each gene, the null hypothesis that the gene is not differentially expressed against the alternative that it is.

If we test each of the m hypothesis at level $p<\alpha=0.01$, we would expect about 200 false positive!!!

Multiple error controlling procedures: Bonferroni

$p_j \leq Tr$ → $Tr = \frac{\alpha}{m}$

Bonferroni
Correction
(FWER)

In practice for each gene you have to compute a new p-value

pj<Tr=α/m ----> pj*m<α ---> Pbonf<α

and you should retrieve all the genes for which Pbonf=pj*m <α

Multiple error controlling procedures: Benjamini - Hockberg

Consider the p-values sorted in ascending order:

$p(1)<p(2)<\ldots <p(m)$

For the j-st gene the new pBH is $p(j)*m/j$

Detect all the genes whose sorted p-value is s.t.

$$p(j)*m/j< \alpha$$

In practice for the j-st gene you have to compute a new p-value

$$Pcorrect(j)=p(j)*m/j$$

and you should retreive all the genes for which

$$Pcorrect<\alpha$$

How to perform the multiple tests in R?

- R has built in methods to adjust a series of p-values either to control the family-wise error rate or to control the false discovery rate.

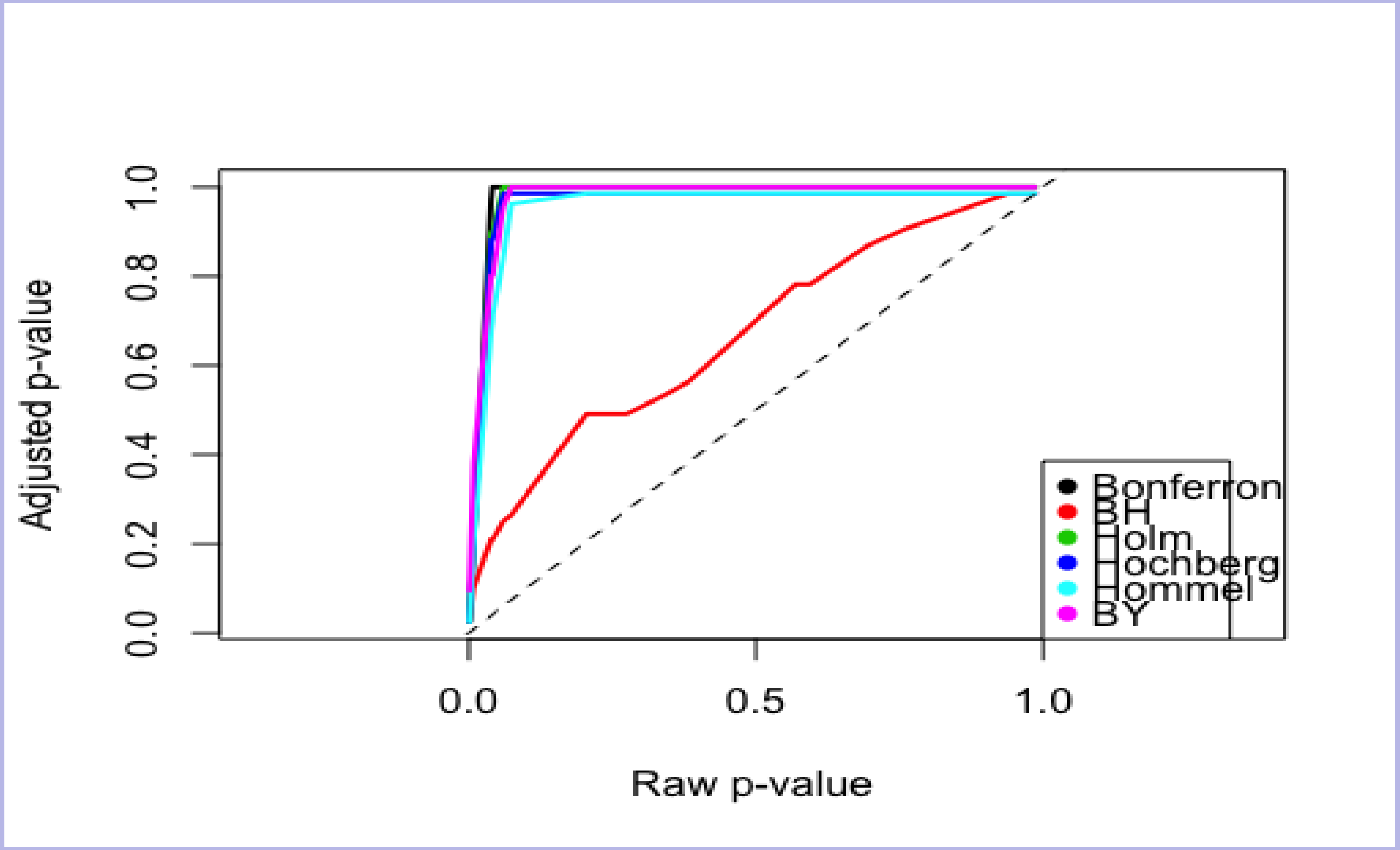
- The methods **Holm, Hochberg, Hommel, and Bonferroni** control the **family-wise error rate**. (they control the probability of even one false discovery, and so are all relatively **strong conservative**).

- The methods **BH** (Benjamini–Hochberg, which is the same as FDR in R) and BY control the **false discovery rate**. These methods attempt to control the expected proportion of false discoveries.

Applications

- Apply multiple hypothesis testing to GOLUB dataset.

- Tutorial MulHypTestOnGolub.R


Adjusted p-value
Raw p-value
● Bonferron
● BH
● Holm
● Hochberg
● Hommel
● BY