Statistical Tools for Wet-Lab Researchers

BY OLIVIA TONG, BSC (HONS) – AKA. NOT AN EXPERT

Version 1 – Summer 2016

Introduction

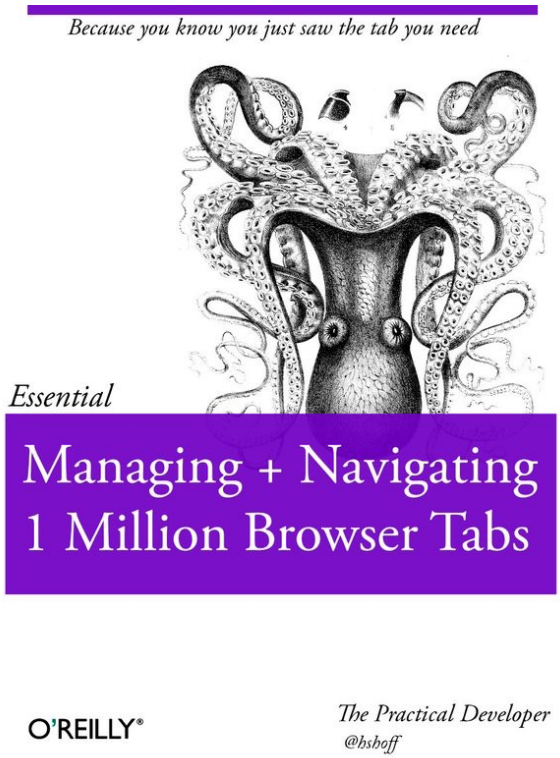


Figure 1 Skills that you need to acquire in the near future

Welcome! If you are reading this manual, you are probably stuck or having trouble with statistical analysis for biology and health care data. This manual is created to facilitate data analysis and interpretation for wet-lab researchers in Dr. Henry Kwok's lab, and the manual is produced as a by-product of the lessons. In other words, I view this manual as a textbook for the course called "Statistical Tools for Wet-lab researchers". The manual is by no means comprehensive, and it is written with the intention to provide you tools to tackle graphs and data. I will also try to provide references and future readings at the end of each section.

I find statistics boring, and I do not enjoy learning theories. Hence, this manual will be taking a practical approach, and I try to emphasize hands-on training. For each statistical concepts or models that we look at, I will pair it up with an R script. Yes, this manual requires the use of programming language R and the software R studio. The end goal of this manual is not to enable you to be an expert in programming, in fact, I think you will still be quite junior by the end of this manual/ course. The ultimate goal is to teach you how to use scripts and tools available to analyze and interpret your data, and the **target audience is the end-users**. If you are looking for a resource for R package developers, then you are at the wrong place.

As with any programming language, you do not need to know everything before you start doing something. Here is my advice:
a) learn as you go - to be honest, I am not an expert either, and I just solve problems as they come to me.
b) start small - focus on learning the visual outputs and the big concepts instead of the details. You can worry about those later.
c) do not be afraid of error messages - they are good stuff that directs you - and one day, you will understand why you need the skills in Fig 1.

Good luck and have fun with your data!

Olivia

@ I mostly compile the information of this manual from different sources, and I do not own the R scripts - I only recreate them to fit in the purpose of this manual. I give credit where credit is due.

Table of Content

Lesson 1 - Intro to R

Install R and RStudio IDE

Install R:
- Install from CRAN – the Comprehensive R Archive Network. Download the version that works for your operating system, such as Mac / Windows
https://cran.rstudio.com/

Install RStudio IDE (aka. Integrated development environment):
- Install the open source edition for RStudio Desktop
https://www.rstudio.com/products/rstudio/

Upgrade R for updates on CRAN:
- Type in the following command in the console

```
update.packages(ask = FALSE, checkBuilt = TRUE)
```

Using R packages

R is an extensive system and people share developed codes as a package via CRAN and GitHub. In the field of Bioinformatics, R is particularly popular because of the Bioconductor project. Bioconductor is a type of R packages, and it provides tools for the analysis and comprehension of high-throughput genomic data. This section will discuss the installation of *dplyr*, *ggplot*, *Bioconductor* packages.

1) dplyr: a package for data manipulation, and it focuses on data.frames and related structures.

```
install.packages("dplyr", dependencies = TRUE)   # to install the package
library(dplyr)  # to load the package
```

If you write `install.packages("dplyr")`, it works fine too. However, with `dependencies = TRUE`, you are installing uninstalled packages which these packages depend on or link to.

Features:
a) Use `filter()` to subset data row-wise
b) Use `select()` to subset the data on variables or columns
c) New pipeline operator
`%>%`

Script Example:

```
filename <- read.csv("msleep_ggplot2.csv")       # read file
Primates<- filter(filename, order == "Primates")  # keep only Primates
PrimatesVals <-select(Primates, sleep_total)
                                    # select only the column <sleep_total> for the primates

# dplyr package for a "pipe"
PrimatesVals <- filter(filename, order == "Primates") %>% select(sleep_total)
```

2) ggplot: a package to plot data for R
   install.packages("ggplot2", dependencies = TRUE) # to install the package
   library(ggplot2) # to load the package

3) Bioconductor: open source software for bioinformatics
   - To install core packages:
     source("https://bioconductor.org/biocLite.R")
     biocLite()

     Install specific packages, eg "GenomicFeatures" and "GEOquery", with
     biocLite(c("GenomicFeatures", "GEOquery"))

     For finding Bioconductor packages and update installed packages, please refer to the official website:
     http://bioconductor.org/install/

Sample Heatmap Script

-This heatmap script is used to cultivate interest for using R.
- Supplemental scripts are under modification to avoid plagiarism, and it is important for me to credit those who created the original script. Hence, this script is currently only available for demonstration during lesson. If time allows, I will recreate a demo script.
Some handy lines:
# Sort data
nba<-nba[order(nba$PTS),] #sorted by points per game from least to greatest

Future readings

For keeners, if you have lots of time and would like to be a pro in R, here's something you can try:

Swirl – It teaches R programming and data science interactively
http://swirlstats.com/

That aside, I believe there are lots of tutorials online regarding **how to use R**, please feel free to follow them, such as: https://www.zoology.ubc.ca/~schluter/R/calculate/
Aside from this, I do not have any specific good ones in mind, but will add to this document if I encounter any in future.

For those interested to learn keyboard shortcuts, here's a good reference:
https://support.rstudio.com/hc/en-us/articles/200711853-Keyboard-Shortcuts

Appendix A: Programming 101 Basics: Variables and Data Types

Variables - the dynamic information is stored

Eg) when you type your name into a web form and send it, your name is a variable

Types of variables:

A. Character (char) - eg. 4, *, x
B. String: eg. Your name
C. Integer (int) - a whole number, eg. 65
D. Floating-point number (float) - a number that may have digits after the decimal place. eg. 65.00
E. Boolean (bool): a variable to represent true or false
F. Array: lists of other variables
   eg) 1, 2, 3, 4, 5 might be stored as an array (of length 5)

Appendix B: Mathematics notation

1. **Indexing**:
   x <- 1:5

   For 5 numbers, they can be represented like x_1, x_2, x_3, x_4, x_5.
   We use dots to simplify this x_1, … x_5, and indexing to simplify even more x_1, $i=1,\ldots,5$. If we want to describe a procedure for a list of any size n, we write x_i, $i=1,\ldots,n$.

   Double indexing when we have several measurements (such as blood pressure, weight, height, age, cholesterol level) for 100 individuals. For example: x_{ij}, $i=1, \ldots, 100, j=1, \ldots,5$

2. **Summation**:
   n < - 1000
   x < - 1:n
   S <- sum(x)

   $$S = \sum_{i=1}^{n} x_i$$

3. **Greek letters**:
   Unknown average - μ
   Standard derivation - σ
   Measurement error or unexplained random variability- ε
   Effect sizes (e.g. Effect of a diet on weight) - β

   Index with Greek Letters to indicate different groups
   - For example, if we have one set of numbers denoted with x and another with y we may use μ_x and μ_y to denote their averages.

4. **Infinity**:
   Asymptotic results: approximation that gets better and better as the number of data points we consider gets larger and larger, and the perfect approximation occurring when the number of data points is ∞. In practice, this is impossible to achieve.

This concept, however, should be applied. Asymptotic results are results that become better and better as some number increases, and we can pick a number so that a computer cannot tell the difference between approximation and the real number.

For example -
`onethird <- function(n) sum(3/10^c(1:n))`
`1/3 - onethird(4)`
[1] 3.333333e-05

`1/3 - onethird(10)`
[1] 3.333334e-11

`1/3 - onethird(16)`
[1] 0
In this example, 16 is practically infinity.

Appendix C: Shortcuts

Default panels:

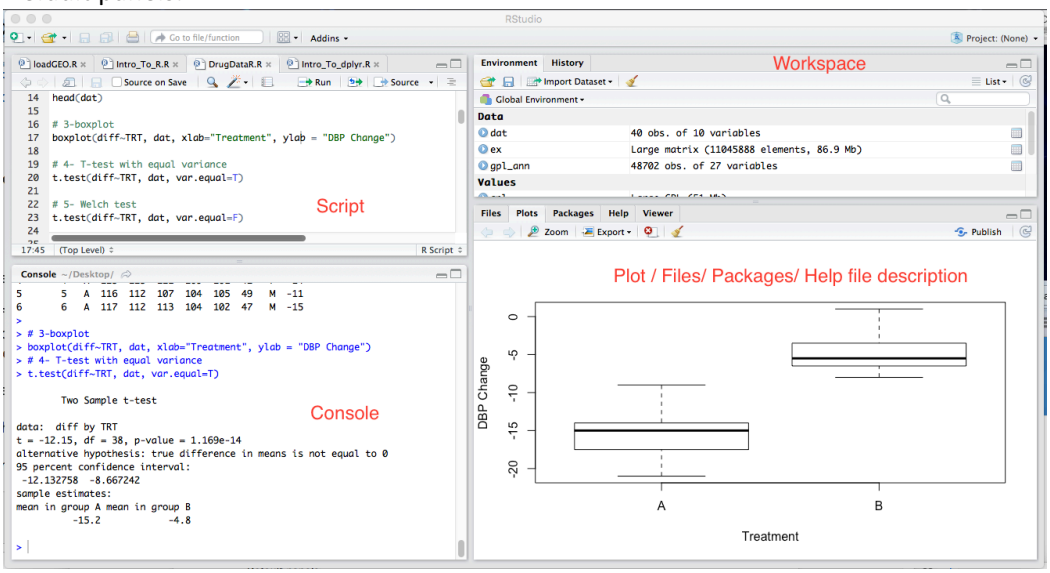


Figure 2 Default panels of R Studio

Type in console:

1. Current directory = `getwd()`
2. Set directory = `setwd("/Users/xxx")`
3. Remove everything = `rm(list = ls())`
4. Remove the object named x = `rm(x)`
5. Quit R = `q()`
6. Get help = `?function`, e.g. for function lm, type `?lm`

Keyboard shortcuts

7. Clear Console = control + L
8. Run script line by line (without copying to console) = control + enter

Lesson 2 – Inference –Part 1

The next two lessons introduce the essential statistical concepts to understand p-values and confidence intervals. These terms are very common in scientific literature.

For example, you often see statement like this:

High-carbs diet-fed mice (+ 2.2 ± 0.2g) than in the normal diet-fed mice (+ 0.1 ± 0.2g; $p < 0.001$)

We will learn what the meanings of these terms and how to compute these values in R.

Random variables and Null Hypothesis

Mouse 1 is 24.02 grams and mouse 25 is 34.29 grams, but they are following the same diet. This is called **variability**, and as a scientist, you are probably familiar with this. In general, we get an average of each group to see the effect.

If we repeat the experiment, we get 25 new mice and after randomly assigning them to different groups, we get a different mean. Every time we repeat the experiment, we get a different value, this type of quantity is called **random variables**.

R programming tips:

After loading your data into R, your data is in the **data frame**. To apply statistics such as mean and variance, you will need to change your data into **numeric**. The following sample script used the data file: msleep_ggplot2.csv. This consists of a data frame with 83 rows and 11 variables, and it has sleep times and weights taken from *V. M. Savage and G. B. West. A quantitative, theoretical framework for understanding mammalian sleep. Proceedings of the National Academy of Sciences, 104 (3):1051-1056, 2007.* This file can be obtained from the supplemental package.

Please put the data file in the **same working directory**.

For example:

```
filename <- read.csv("msleep_ggplot2.csv")      # load data
head(filename)
class(filename)      # test what type of object is returned

# To calculate the average amount of sleep for primates
## To obtain a numeric vector – unlist
   ## filter = keep only primates; select = select only that column; unlist = change data into numeric
PrimatesVals <- filter(filename, order =="Primates") %>% select(sleep_total) %>% unlist

class(PrimatesVals)    # check data type

mean(PrimateVals)   # average of the amount of sleep for primates
```

I randomly generated two datasets: JapFemalesWeightdata.csv and JapFemalesControlData.csv.

Consider the scenario, you are a researcher in Japan and you would like to conduct an experiment to see **the effect of a newly designed exercise program**. As a good researcher, you want to design an experiment that avoids most of the confounding variables.

For this study, the **dependent variable** is Bodyweight, and the **independent variable** is the newly designed exercise program. We also decided to choose Japanese females to reduce potential confounding effect.

The dataset JapFemalesControlData.csv contains data of the bodyweight of Japanese females, i.e. the **population** of this study is Japanese females. We then **sampled** 25 females, and **randomly assigned** into two groups – exercise group and regular group. The sampled data is stored in JapFemalesWeightdata.csv
The exercise group is known as the treatment group, and the regular group is known as the control group.

Disclaimer: To illustrate statistical concepts, the data are randomly generated using random-number generator.

Terms re-visit:

a. Independent variables (IV): feature of the study used to predict or explain the behavior; and there are two types of variables – manipulated and measured

b. Dependent variable / Criterion (DV): characteristic the researcher is accounting for or predicting

c. Population: all members of a defined group that we are studying or collecting information on for data driven decisions.

d. Sample: a set of data collected, or selected from a statistical population by a defined procedure – a part of the population is called a sample.

e. Random assignment: an experimental technique for assigning human participants or animal subjects to different groups in an experiment using randomization, such as by chance procedure (e.g. flipping a coin), or a random number generator. This ensures each participant or subject has an equal chance of being placed in any group.

Please
1) Put the two datasets in your desired destination – you will need to use that folder as your current working directory.
2) run the script "L2_RV_NullHypothesis.R" from supplemental packages.

Terms Revisit … continue:
Null hypothesis (H_0) states that nothing of consequence is apparent in the data distribution, and the data corresponds to our expectation. We did not learn anything new. (*Definition from Exploratory Data Analysis of Biological Data using R, Toronto, May 12 -22, 2015)*

> In our previous example, the null hypothesis refers to **the distribution of the difference** in the mean of the female bodyweights when the null hypothesis is true. We ran a Monte Carlo simulation, and obtained the outcomes of 10,000 simulations of the random variables under the null hypothesis. Further, we showed that if we know the null distribution, we can compute the p-value.

Alternative hypothesis (H_1) states that some effect is apparent in the data distribution. The data is different from our expectation, and we need to account for something new. Not in all cases will this result in a new model, but a new model always begins with the observation that the old model is adequate. (*Definition from Exploratory data Analysis of Biological Data using R, Toronto, May 12 -22, 2015)*

Probability Distribution

Probability Distribution = Describe possible outcomes of random variables

$$\Pr(a \leq x \leq b) = F(b) - F(a) \quad (1)$$

In reality, we do not have a fixed list of numbers because we cannot observe all possible outcomes of random variables. Hence, we describe probabilities. From the population, we randomly pick a bodyweight, then the probability of this bodyweight falls between a and b is described in equation 1. This equation describes the probability of random variable.

Normal distribution

- Bell curve, known as normal distribution or Gaussian distribution
- Math formula to approximate the proportion of values or outcomes in any given interval:

$$\Pr(a < x < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx$$

In R, pnorm sets a to – infinity, and takes b as an argument.

We can compute the proportion of values below value x with pnorm(x, mu, sigma).

Pnorm(x):
- returns the probability that a random variable following the standard normal distribution falls below x
-For probability that is larger than x, 1-pnorm(x) is used.

Appendix A: Formulae

Mean:

$$\mu_X = \frac{1}{m}\sum_{i=1}^{m} x_i \text{ and } \mu_Y = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Variance:

$$\sigma_X^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_X)^2 \text{ and } \sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_Y)^2$$

These are called population parameters.

Sample estimates:

$$\bar{X} = \frac{1}{M}\sum_{i=1}^{M} X_i \text{ and } \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Lesson 3 – Inference – Part 2

Central Limit theorem

Central limit theorem is the most fundamental concepts in statistics and math. This theorem is a statistical model that describes the distribution of our data.

Central limit theorem suggests that **our sample average follows a normal distribution**. To apply CLT, **large samples (>30 in general)** are required.

When the sample size is larger, the **average** of a random sample follows a normal distribution centered at the **population average**, with a standard derivation equal to the **population standard derivation**. This relies on asymptotic results, i.e. large samples.

The normal distribution is centered at zero, i.e. there is no difference, and a standard derivation of 1.

Hence, central Limit theorem tells us that when the samples are large, the random variable is normally distributed with a mean 0 and SD 1, and that we can compute p-values using the function pnorm.

Exercise of Central Limit Theorem

A list of numbers is normally distributed and what proportion of these numbers are within 2 standard derivations away from the average of the list?

```
pnorm(2)-pnorm(-2)
```

Introduction of Test Types

Before we move on to look at other statistical models, it is important for us to have a general idea of hypothesis testing and the different types of test.

Hypothesis testing is confirmatory data analysis, in contrast to exploratory data analysis.
We use hypothesis testing to answer the following questions:
Is a particular sample a part of the distribution, or is it an outlier? Can two sets of samples have been drawn from the same distribution?

Hypothesis Testing Principles:

a) Observation
b) Model of the data
c) Check the probability that the model of your data would contain your observation

P-values: A measure of how much evidence we have against the alternative hypothesis. P-value characterizes where an observation lies with reference to the distribution of our statistics under the null hypothesis.

One- and two-sided tests:
A **one-sided alternative test** establishes the direction of the association between the predictor and the outcome, such as the prevalence of breast cancer is higher in Caucasian women than women of other races. This is one-sided test.
A **two-sided alternative test** establishes only that an association exists without specifying the direction, such as the prevalence of breast cancer in Caucasian women is different from women fo other races. This is two-sided hypothesis tests.

Test types:
There are large variety of types of hypotheses and the proper application of tests can be confusing. Here we name a few common types of test:

1. **Z-test**: compares a sample mean with a normal distribution – known population parameter

2. **T-test**: compares a sample mean with a t-distribution, and this relaxes the requirements on normality for the sample – not known population parameter, and thus do not know shape of the sampling distribution

3. **One-sample t-test**: compares a sample with the mean of a population – applies to n observations that are **independent** and **normally distributed** with equal variance about a mean.

4. **Two-sample t-test**: compares the two samples means with each other – assume the data are **independent (even between groups)** and **normally distributed**, and the variance is the same. **If variances are not equal in the two groups, use Welch's t-test (R default)**. According to central limit theorem, if the sample size is large enough, the t-test will work fine. Please see Example L3-3.
   e.g. For gene expression, is the mean expression level under condition 1 different from the mean expression level under condition 2?

5. **Paired samples t-tests**: compare matched pairs of observations with each other, we look at whether their difference is significant. The two groups are not independent.

6. **Non-parametric tests**: applied if we have no reasonable model from which to derive a distribution for the null hypothesis, i.e. if data not normal, such as Wilcoxon and Mann-Whitney tests.

7. **Chi-squared tests**: analyze how likely the observed distribution is due to chance

8. **Fisher's exact test**: analyze categorical data that result from classifying objects

9. **F-tests**: analyze the variance of populations (ANOVA)

(Modified from Exploratory Data Analysis of Biological Data using R, Toronto, May 12 -22, 2015 & Zou, K. H., Fielding, J. R., Silverman, S. G., & Tempany, C. M. (2003). Hypothesis Testing I: Proportions 1. Radiology, 226(3), 609-613.)

T-distribution and T-tests

We will first discuss the theories and how we can compute T-tests by hand, then we will learn how results can be obtained by R using R scripts.

T-test is used to evaluate the difference between two means from two independent samples or between two samples in which the observations of the second sample are not independent of those in the first sample. T-tests use the t-distribution to determine the P value. **T-distribution** is similar to normal distribution, but that t-distribution compensates for small sample sizes.

Previously in the *Introduction of Test Types* section, we emphasized the importance of normal distribution. If we have a large sample size (>30 in general) and the data are normally distributed, central limit theorem applies and we can evaluate the result by T-test.

What if our data is not normally distributed?

We can apply still t-tests when the sample data is reasonably **symmetrically distributed about the sample mean**. In other words, when the population is not normal, and the sample is small, you can still apply t-tests if your sample data is symmetrically distributed.

So how do you know if the data are symmetrically distributed about the sample mean?
You can do this by checking:

- Boxplot, i.e. median is in the center of the box and the whiskers extend equally in each direction
- Histogram looks symmetrical
- Mean approximately equals to the median

Example L3-1 - Research Design: Racing speed - **Single Sample, Z-test**

You would like to test the effect of ingested estrogen on racing speed of mice. You ingested estrogen to 50 mice and compared the racing speed of those ingested with the average racing speed of mice in that cohort.

	Population	**Sampling Distribution of the mean**	**Sample (n = 50)**
Data	Scores (500Xs)	Means	Scores (50)
Mean	μ = 90s	μ_x = 90s	M = 85.8s
SD	σ= 10.2	σ_x (SE of the mean) = σ / $\sqrt{N}$ =1.4425	SD = 8.00s
Shape	Normal Distribution	Normal Distribution	NA

Step2: Select the a priori criteria: p(alpha) = $0.05_{(2\text{-tailed})}$ z(crit)= +/- 1.96

Step3: Z-statistics: $Zx = (M - \mu_x) / \sigma_x = (85.8-90)/ 1.4425 = -2.9116$

Step4: From the z-score table, z = -2.91 = 0.0016

Step5: p(obs) = 0.0016*2 = 0.0032 & Zx = 2.91, p< 0.004
p(obs) < p(alpha) = 0.0032 < 0.05 OR p(obs) < p(alpha/2) = 0.0016< 0.025

&

z (obs) > z(crit) = -2.91 > -1.96

Step6: Statistical decision: Reject the null hypothesis

However, this is obviously a flawed study because you only have one single sample research.

Example L3-2 – Research Design: Exam scores - **Single sample (One Sample T-test)**

Test scores for 9 students: 95, 60, 80, 70, 75, 65, 90, 60, 80

	Population	**Sampling Distribution of the mean for df = 8 (σ not known)**	**Sample (N=9)**
Data	Scores (X)	Means	Scores (N=25)
Mean	$\mu = 72$	$\mu_x = 72$	M = 675/9 = 75.0
SD	σ= ?	$S_x = S/\sqrt{N} = 12.50/\sqrt{9} = 4.166667$	S = 12.50

** Variance = Sum of Squares/ (N-1) = 1250/8 = 156.25

Step2: Select the a priori criteria: p(alpha) = $0.05_{(2\text{-tailed})}$ t(crit)= +/- 2.306 (2-tailed)

Step3: t-statistics: t(obs) = $(M - \mu_x) / S_x$ = (75-72)/ 4.16667 = 0.7199994
t(8) = 0.720

Step5: t (obs) > t(crit) = reject null; t(obs) < t(crit) = 0.7199994 < 2.306 = retain the null hypothesis

Step6: Statistical decision: Retain the null hypothesis

Example L3-3 – Research Design: Exercise vs control program–

Two sample t-test – equal variances and unequal variances

Subjects are randomly assigned to the two groups, and their performance after the program is evaluated in terms of test scores.

Note that the Two-sample T-test with unequal variances is equal to Welch's t-test (R default). Welch's t-test is also preferred for unequal sample sizes.

Please
1) Put the the dataset "Ttestequal.csv" in your desired destination – you will need to use that folder as your current working directory.
2) Run the script "L3_supp_t_test.R" from supplemental packages.

From the R script, you will learn the two tests that assess the sample variances of the two groups.
In statistical terms, these tests are used to verify the homoscedasticity (homogeneity of variances). For both tests, if we obtain a p value >0.05, we can assume the two variances are homogenous. The two tests are:

1) **F ratio**
   F ratio of group variances, also known as Hartley's test, assesses the variance of the two groups assuming the distributions are normal. F ratio of group variances uses the raw data.

$$F = \frac{s_2^2}{s_1^2}$$

```
var.test(control, treatment)
```

2) **Levene's Test**

The test is a function of the residuals and means within each group to verify the homoscedasticity. Brown-Forsythe test is a modified version of Levene's test, and uses the medians within group and this is recommended when normality may be suspect.

$$F = \frac{MS_{b/t-levels}}{MS_{w/i-levels}}$$

Levene's test uses transformed data, i.e. the absolute values of the deviations from the mean, and is more robust compared to the F ratio of group variances. Levene's Test does not require same sample size, and is also the default test used by SPSS to assess group variances.

```
library(car)
my.data = stack(list(treatment=treatment, control=control)) # get the data into 'stacked' format
leveneTest(values~ind, my.data)
```

Example L3-4 – Research Design: Exercise vs control program– **Paired Sample t-test**

Paired samples are not independent of one another, and they are also called matched samples or repeated samples. Before we look into an example, it is important for us to understand the two terminologies - independent and paired samples.

For example, you want to test the effect of exercise training program on a health performance test. There are two approaches:
1) **Independent samples**: You have 24 people and assigned 12 to participate in the exercise program, and another 12 to participate in control program. Then you give each of the 24 people a health performance test and compare the results.
2) **Paired samples**: You take a sample of 12 people and have each person participate in the exercise program and take the health performance test; then you have the same people to participate in the control program, and again take the health performance test. The results are then compared.

Paired samples tend to require less subjects, and you have greater control over confounding variables. However, there may be order effects and that the order in which people take the tests influence the result and counterbalancing may be required.

In terms of **cancer research**:
For example, you want to find out whether the new chemotherapy affects tumor size, you can have the same eight patients represented in two groups. One group represented by the patient sample before therapy and the second group represented the patient sample after therapy. Say, before therapy, mean tumor size was 5.88 cm, and the mean tumor size after therapy was 5.20 cm. Paired-sample t test can be used to test the null hypothesis if the mean difference in tumor size between the two groups is statistically significance – with the assumption that the datasets are normally distributed.
Now please run the script "L3_supp_t_test.R" from supplemental packages.

Lastly, as a **concept illustration of t-distribution, CLT and t-test**, please:
1) Put the the dataset "JapFemalesWeightdata.csv" in your desired destination – you will need to use that folder as your current working directory.
2) run the script "L3_t-distribution.R" from supplemental packages.

Lesson 4 – Inference – Part 3

Confidence interval includes information about estimated effect size and the uncertainty associated with this estimate, i.e. confidence interval summarizes the variability of the random variable.

For example, a 95% confidence interval is a random interval with a 95% probability of falling on the parameter we are estimating. In other words, 95% of random intervals will contain the true, fixed value; and in 5% of the cases, we fail to cover the true value. The confidence interval does not include 0, and this implies the interval is either bigger or smaller than 0.

If the distribution is not normal, the intervals are different. T-distribution has fatter tails, and hence larger intervals and cover true value more frequently. For instance:

```
qnorm(1-0.05/2)   # normal distribution
## 1.959964
qt(1-0.05/2, df=23) # t-distribution
## 2.068658
```

If a 95% or 99% confidence interval does not include 0, then the p-value must be smaller than 0.05 or 0.01 respectively. To apply the concept of t-test:

```
t.test(treatment, control, conf.level = 0.95)$conf.int
## [1] -6.08463229  0.04296563
## attr(,"conf.level")
## [1] 0.95
```

Power calculations

Type of Error

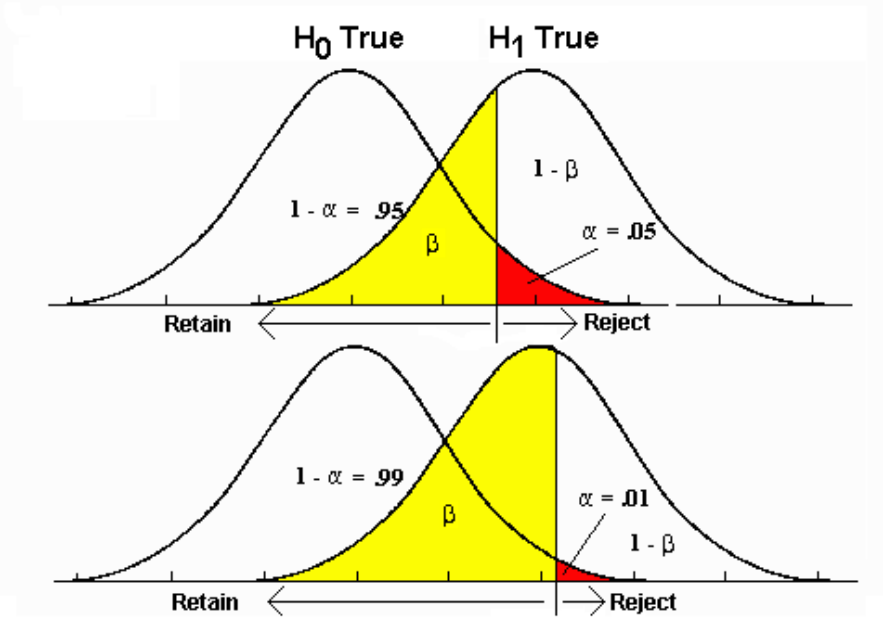


Figure 3 From Introductory Statistics: Concepts, Models and Applications by David W. Stockburger

		TRUTH	
		DIFFERENCE	NO DIFFERENCE
RESULTS OF THE STUDY	DIFFERENCE	CORRECT CONCLUSION (1-β)	FALSE POSITIVE (α error or Type-I error)
	NO DIFFERENCE	FALSE NEGATIVE (β error or Type-II error)	CORRECT CONCLUSION (1-α)

Figure 4 Types of Errors. From Lochner, H. V., Bhandari, M., & Tornetta, P. (2001). Type-II error rates (beta errors) of randomized trials in orthopaedic trauma. J Bone Joint Surg Am, 83(11), 1650-1655.

Type I error is defined as rejecting the null when the null is true, and this is also called false positive. Type II error is defined as not rejecting the null when the null is false, and this is also called false negative.

Arbitrary cut-offs

Although most journals frequently insist the results are significant at the 0.01 or 0.05 levels, there is nothing special about these numbers. In fact, you can choose any p-values and confidence intervals as long as the values are meaningful and informative.

Power

Power is the probability one rejects the null hypothesis when the alternative hypothesis is true, also known as 1-Type II error. Power is affected by alpha, and it depends on the standard error of the estimates – which in turns affects the sample sizes and population standard derivations.

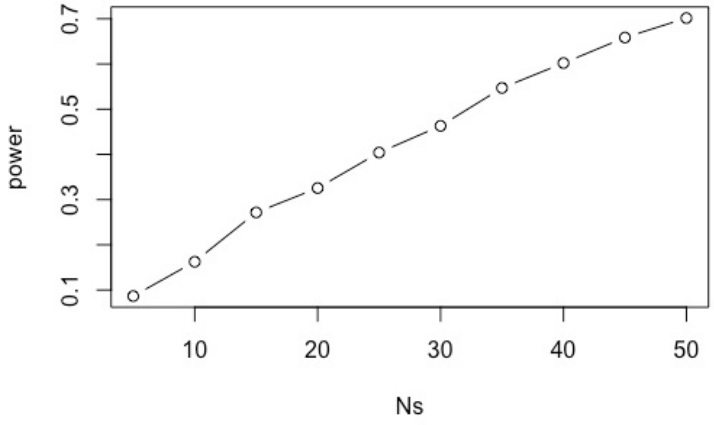


Figure 5 Power increases with N

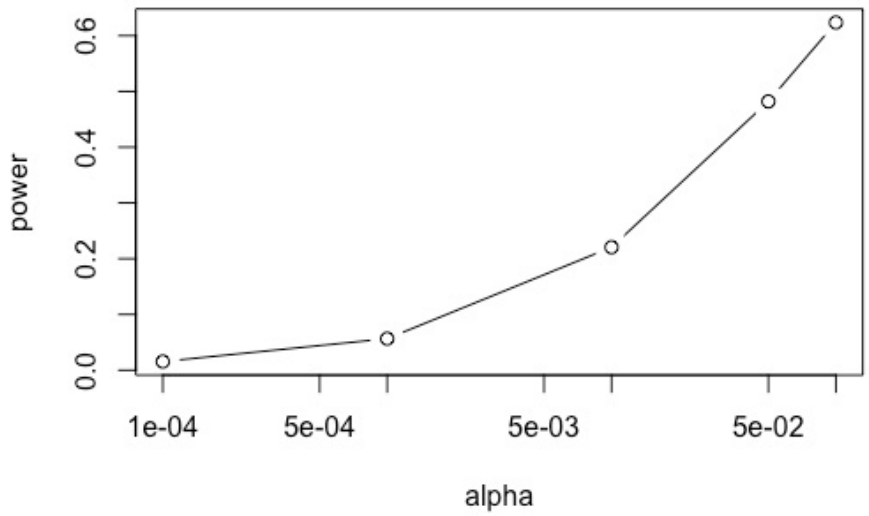


Figure 6 Power varies at the level of alpha we reject

If we want to decrease the chance of type I error, we will have less power. It is a trade-off between the two types of error, and there is no "right" power or alpha level.

We are not interested to have a very small p-values, and having a very small p-value simply means we sampled more subjects than was necessary. Indeed, p-values become smaller as we increase sample size because the numerator of the t-statistic is √N.

Power calculation in R:
power.t.test(n=5, delta=1, sd=2, alternative = "two.sided ", type= "two.sample ")

Hence, it is better to report the effect size with a confidence interval or statistic with meaningful scale. **Cohen's d** is a statistical parameter that reports the difference between two means divided by a standard deviation for the data. It is affected by treatment effect and variability. Please see future readings for details.

Cohen's effect size:
0.2 = small effect size
0.5 = medium effect size
0.8 = large effect size

Power varies directly with 1) sample size, 2) alpha, and 3) magnitude of the effect.
Power varies inversely with 1) beta, and 2) variability.

To illustrate how power is calculated, please
1) Put the dataset "JapFemalesPopulationData.csv" in your desired destination – you will need to use that folder as your current working directory.
2) Run the script "L4.1_power.R" from supplemental packages.

Future reading

Cohen's d effect size – interactive visualization: http://rpsychologist.com/d3/cohend/

Permutation tests

Permutation test allows us to see all of the possible alternative treatment assignments. It gives a simple way to compute the sampling distribution for any test statistic, under the sharp null hypothesis that a set of variants has no effect on the outcome. This is done by reshuffle the data and re-compute the mean. In practice, permutation test is used when we do not want to rely on the assumptions for the normal or t-distribution.

Rationale:
If the null hypothesis is true, change of exposure would have no effect on the outcome. Randomly shuffling the exposures can make up as many datasets as we want, and the shuffled datasets should reflect the real data, and there should be no difference.

To illustrate permutation test, please
1) Put the dataset "JapFemalesPopulationData.csv" in your desired destination – you will need to use that folder as your current working directory.
2) Run the script "L4.2_permutationtest.R" from supplemental packages.

Association tests: Fisher test, Chi square - Test for nominal variables

This section discusses how we handle data that is binary, category and ordinal, such as genotype (values are AA, Aa, or aa). These data are usually 0 (control) or 1 (cases), and it is pretty obvious that they are neither normal distribution or t-distribution. Central limit theorem can apply if the sample size is large, if not, we need to use association tests.

Fisher's Exact Test
Fisher's exact test is based on a famous story – Lady Tasting tea (please see Wikipedia for detail of the story). Fisher's exact test is usually applied when you have a small sample size (arguably below 5 expected values, but some statisticians suggest that as long as the **total sample size is smaller than 1000**, the difference between Fisher's exact test and chi-square test is trivial (McDonald, 2014). The null hypothesis is that there is no relationship between the two variables. The alternate hypothesis is that there is a relationship, but this do not tell you which direction does the relationship go. One caveat is Fisher's exact test requires the row totals and/ or column totals to be fixed, which means you need to know the total number before doing the experiment.

The question: if the observed phenomenon is by chance, or is there any effect? If we have 4 green and 4 red balls, and we pick 4 balls out of a vase. The null hypothesis suggests each ball has the same chance. The chance of observing 3 or more correct balls, under the null hypothesis, is the probability of picking 3 correct balls + probability of picking all 4 correct balls.

For example: Racial differences and 10-year breast cancer survival

	Yes	**No**	*Row Total*
Caucasian	A	B	A + B
Asian	C	D	C + D
Column Total	A + C	B + D	A + B + C + D = n

Probability of any set of values:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\;b!\;c!\;d!\;n!}$$

Example 4-1 **Fisher's Exact Test**

Back to the Lady Tasting Tea story: you are interested to find out what's the probability the lady correctly identified 3 or more cups of tea.

For the probability of picking 3 cups of tea correctly:

	Guessed Before	Guessed After	*Row Total*
Poured Before	3	1	4
Poured After	1	3	4
Column Total	4	4	3+1+1+3 = 8

$$\text{Probability (P)} = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{4!4!4!4!}{3!1!3!1!8!} = \frac{16}{70}$$

For the probability of picking 4 cups of tea correctly:

	Guessed Before	Guessed After	*Row Total*
Poured Before	4	0	4
Poured After	0	4	4
Column Total	4	4	4+0+0+4 = 8

$$\text{Probability (P)} = \frac{\binom{4}{4}\binom{0}{4}}{\binom{8}{4}} = \frac{4!4!4!4!}{4!4!8!} = \frac{1}{70}$$

So the probability of picking 3 or more cups of tea correctly = $\frac{16}{70} + \frac{1}{70} = \frac{17}{70}$

This value is the p-value.
Similar to t-test, Fisher's test can be run in R with one line of code: `fisher.test(dat, alternative="greater")`

Chi-square test

Chi-square test, also known as a "goodness of fit" statistic. It is designed to analyze categorical data, and it measures how well the observed distribution of data fits with the expected distribution if the variables are independent. Chi-squared test assume the observations are independent of one another. Similar to Fisher's Exact test, the null hypothesis is that there is no relationship between the two variables. The alternate hypothesis is that there is a relationship, but this do not tell you which direction does the relationship go. Chi-square test does provide evidence that a relationship does exist between the two variables.

However, chi-squared test is asymptotic and that it is accurate when we have large sample sizes, but not so for small sample sizes.

To sum up, Chi-square test has two purposes:

1) test probability of independence between two variables (Chi-square Test of independence)
2) test for equality of proportions between two or more groups (Chi-square Test of Homogeneity)

**If a radiologist wishes to compare two proportions, such as the sensitivity of two tests, this is the correct test to apply and not t-tests.

Example 4-2 **Chi-square Test**

	Positive MRI result	Negative MRI result	*Row Total*
Patients	46	71	117
Control	37	83	120
Column Total	83	154	= 117+120 = 237

a. Calculate the expected value for each cell of the table

$$\frac{Row\ Total\ x\ Column\ Total}{Total}$$

	Positive MRI result	Negative MRI result	*Row Total*
Patients	46 (40.97)	71 (76.02)	117
Control	37 (42.03)	83 (77.97)	120
Column Total	83	154	= 117+120 = 237

For example, for the cell of (positive MRI result x Patients):
= (83 x 117) / 237 = 40.97

b. Calculate Chi-square statistic

$$Chi - squared = Sum\ of\ \frac{(Observed\ x\ frequency - expected\ x\ frequency)^2}{(Expected\ x\ Frequency)}$$

$$X^2 = \frac{(46-40.97)^2}{40.97} + \frac{(37-42.03)^2}{42.03} + \frac{(71-76.03)^2}{76.03} + \frac{(83-77.97)^2}{77.97} = 1.87$$

c. Assess significance level: df = (2-1) x (2-1) = 1; look up the chi-square table and compare the chi-square value and the degree of table, you will obtain the p-value. In this case, we retain the null hypothesis and that the two variables are independent.

Chi-square test can be run in R with one line of code: `chisq.test(dat)`

For illustration of Fisher's test and Chi-square test, please
1) Put the dataset "chisquaredata.csv" in your desired destination – you will need to use that folder as your current working directory.
2) Run the script "L4.3_associationtest.R" from supplemental packages.

As a note, for both chi-square and fisher's exact test, having **fewer numbers of degrees of freedom will increase the power of the test**. Hence, if you have a large number of categories and the total sample size is small enough to do a Fisher's exact test, it is better to pool the rarer categories together and do a chi-square test instead of having a large number of categories.

Reference

McDonald, J. H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland.

Lesson 5 – Exploratory data analysis

Exploratory data analysis: Data visualization and Correlations

In biomedical sciences, we are interested in the relationship between two or more variables. Visualization is a good way to understand the relationship of the data. We will first learn how to create plots for data.

Example 5-1: Boxplots

It is important to remember that the middle line in the boxplot shows the median, not mean.

Using the R dataset InsectSprays to visualize the effect of different insecticides

```
head(InsectSprays)   # look at the data
 ## boxplot method 1: using split- boxplot(split(values, factor)
boxplot(split(InsectSprays$count,InsectSprays$spray))
## boxplot method 2: using a formula- boxplot(values~factor)
boxplot(InsectSprays$count~InsectSprays$spray)
```

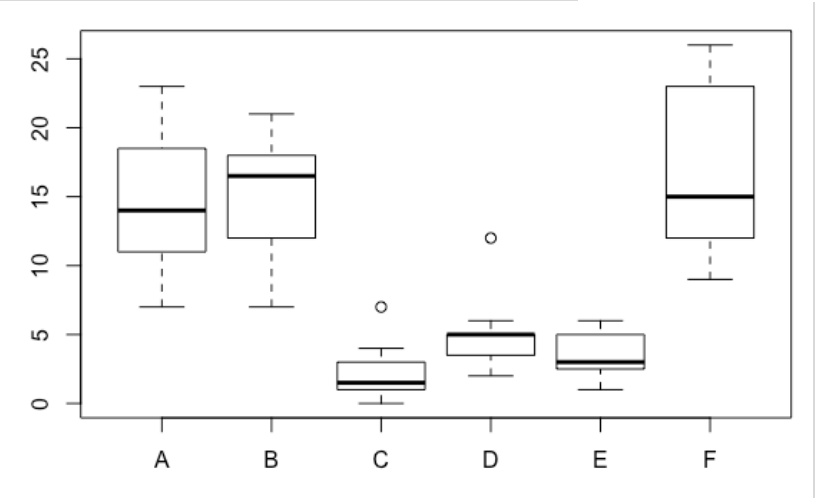


Figure 7 Boxplot of dataset InsectSprays

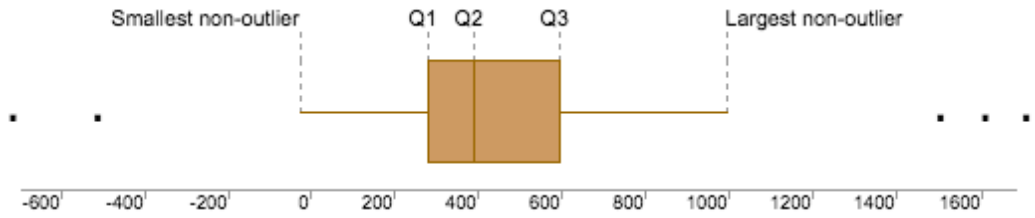


Figure 8 Boxplot basics

Boxplot splits data into quartiles, Q1 is first quartile, and Q3 is the third quartile. Q2 is the median of the dataset. The outliers are plotted separately on the chart.

It is very common for us to use histograms and boxplots together to examine the relationship of our data.

Example 5-2 Histograms and Boxplots

```
library(UsingR)
library(rafalib)
data("nym.2002")
mypar(1,3)
males <- filter(nym.2002, gender=="Male") %>% select(time) %>% unlist
females <- filter(nym.2002, gender=="Female") %>% select(time) %>% unlist
boxplot(females, males)
hist(females,xlim=c(range( nym.2002$time)))
hist(males,xlim=c(range( nym.2002$time)))
```

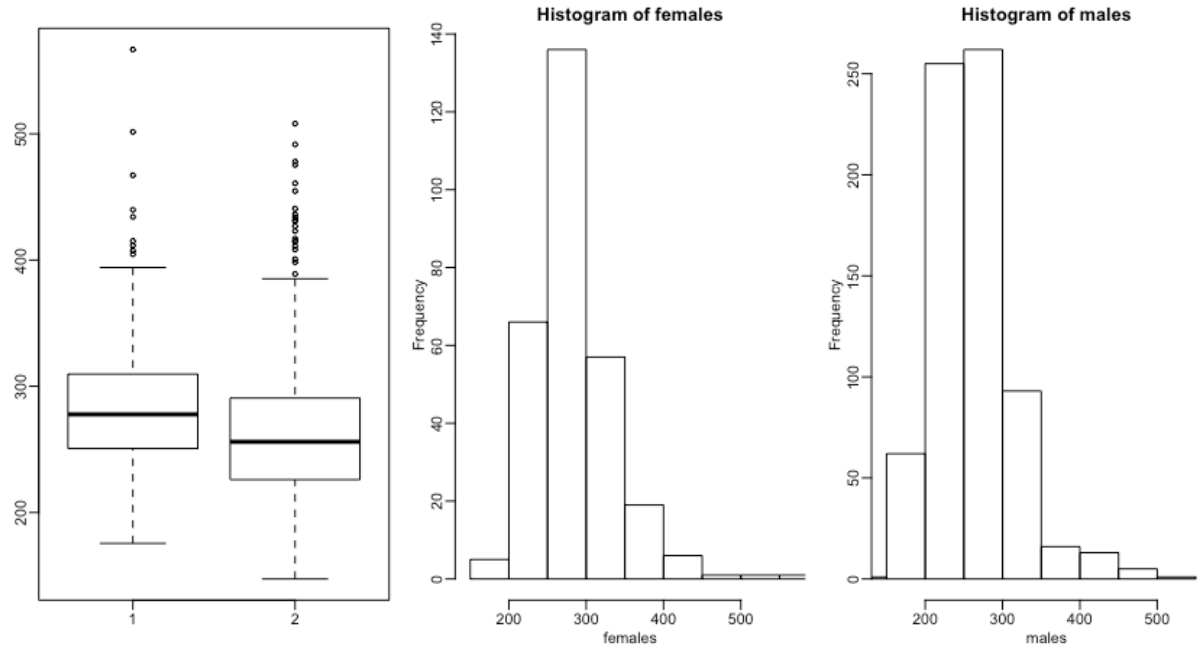


Figure 9 Boxplot and histograms of datasets nym.2002

When one variable depends on the other, the variables are correlated. In R, the function cov() measures covariance and cor() measures the **Pearson coefficient of correlation**. A scatterplot is a good visualization tool. Pearson coefficient of correlation is for linear data, and that is the correlation between variables that follow normal (Gaussian) distribution. For values that do not follow Gaussian distributions, spearman's rho test and Kendall's tau test can be used.

Spearman correlation is mostly used in place of usual linear correlation when 1) the values are integer-valued scores, 2) it has a moderate number of possible scores, 3) assumptions about bivariate relationships are violated. Spearman measures the degree of association between two variables, while Kendall's tau measures the strength of dependence between two variables.

Example 5-3: Spearman's rank correlation coefficient and Kendall's tau test

For Pearson's correlation

```
a <- c(1,3,5,7,9,8,8,9,7,5)
b <- c(2,4,6,8,9,5,5,4,6,2)
cor.test(a, b)
## Pearson's product-moment correlation
## data:  a and b
## t = 2.1244, df = 8, p-value = 0.06637
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04673472  0.89265109
## sample estimates:
##      cor
## 0.6005634
```

For Spearman's test, both variables need to have the same length.

```
a <- c(1,3,5,7,9,8,8,9,7,5)
b <- c(2,4,6,8,9,5,5,4,6,2)
cor.test(a, b, method="spearman")
## Spearman's rank correlation rho
## data:  a and b
## S = 85.062, p-value = 0.1559
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##     rho
## 0.484472
```

The *p-value >0.05* suggests the null hypothesis is retained, and the value of the rho calculated is statistical significance.

For Kendall tau rank correlation coefficient:

```
a <- c(1,3,5,7,9,8,8,9,7,5)
b <- c(2,4,6,8,9,5,5,4,6,2)
cor.test(a, b, method="kendall")
##         Kendall's rank correlation tau
## data:  a and b
## z = 1.2923, p-value = 0.1962
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.3414634
```

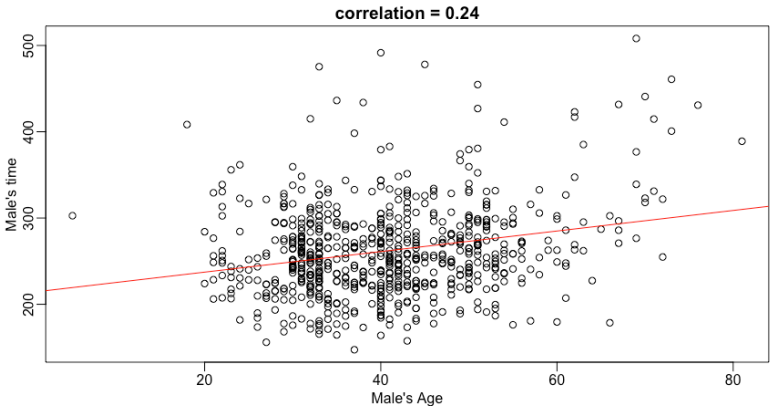
To view the scatterplot

```
plot(a,b)
abline(0,cor(x,y))
```

Example 5-4: Scatterplots and Pearson correlation

We continue using the data from our previous example.

```
library(UsingR)
library(rafalib)
library(dplyr)
data("nym.2002")
males_Age <- filter(nym.2002,
gender=="Male") %>%
select(age) %>% unlist
males_Time<- filter(nym.2002,
gender=="Male") %>%
select(time) %>% unlist
cor(males_Age,males_Time)
## 0.2432273
```




```
plot(males_Age,males_Time,xlab="Male's Age",ylab="Male's time",main=paste("correlation
=",signif(cor(males_Age,males_Time),2)))
model <- lm(males_Time~males_Age, data = nym.2002)
abline(model, col = "red")
```

The scatterplot shows a general trend and the relationship is summarized by the correlation coefficient.

Example 5-5: Stratification

Continue from our previous example, if we are interested to know the times stratified by age groups (such as 20-25, 25-30, etc), we can apply stratification to our boxplots. Stratification followed by boxplots allows us to see the distribution of each group.

```
mypar(2,2)
plot(Females_Age,females_Time)
plot(males_Age,males_Time)
group <- floor(Females_Age/5)*5
boxplot(females_Time~group, ylab ="time", xlab="age")
group <- floor(males_Age/5)*5
boxplot(males_Time~group,ylab ="time", xlab="age")
## if we are interested in the the average time of 72 years old male
print(mean(males_Time[round(males_Age) == 25]))
```

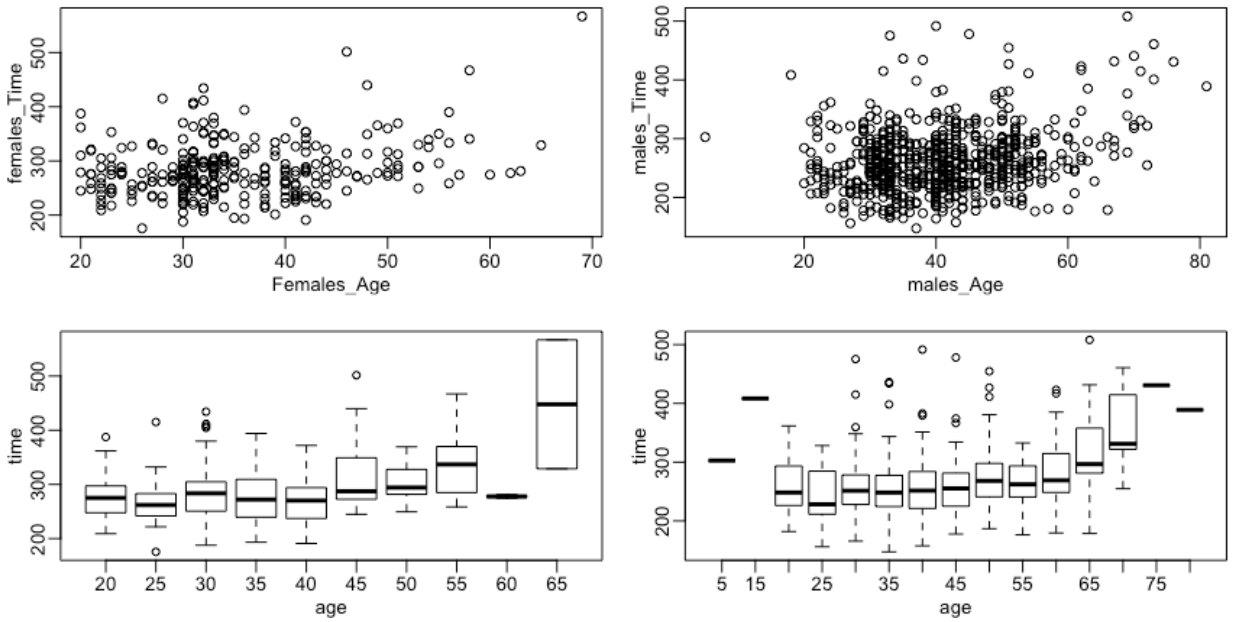


Figure 11 Stratification results

There are many differences are quantified with fold changes. If we ask whether one gene is different from the cancer sample versus the normal one, we can quantify this with a ratio – such as how many times bigger is the expression in this sample compared to the other one. Ratios can be illustrated by logs. Log ratios are symmetric around 0.

Example 5-6: Log Ratios Plot

Continue from the previous example:

```
time<- filter(nym.2002) %>% select(time) %>% unlist
mypar(1,2)
plot(age, time/median(time), ylim=c(1/4,4), main=paste("scatterplot"))
abline(h=c(1/2,1,2))
plot(age, log2(time/median(time)),ylim=c(-2,2), main = paste("log ratios plot"))
abline(h=-1:1)
```

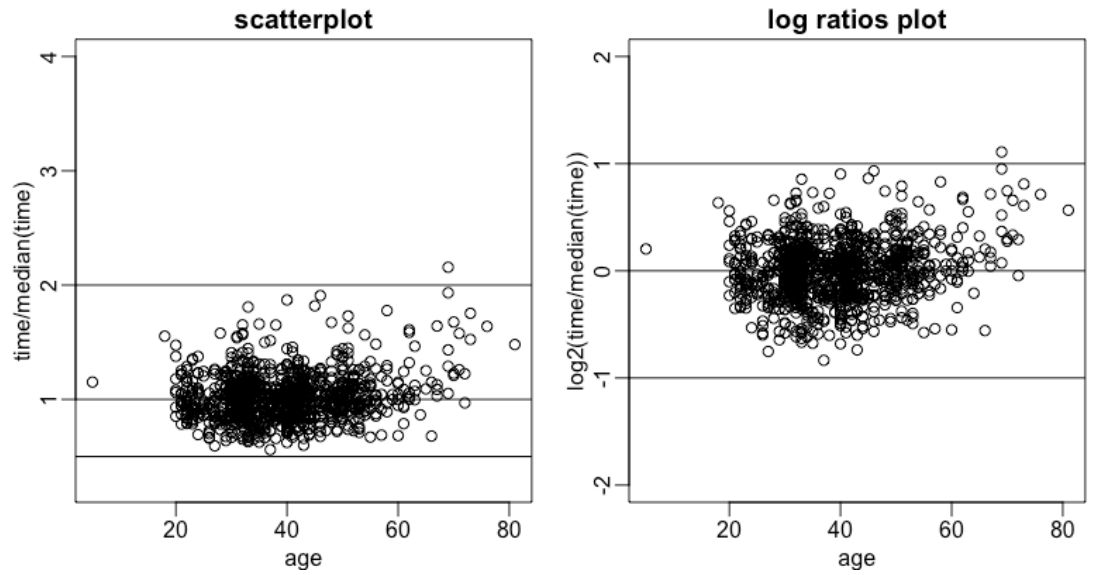


Figure 12 Scatterplot vs Log Ratio Plot

The left plot is the plot of the ratio of times to the median time, with horizontal lines at twice as fast as and twice as slow as median time. The right plot is the plot of the log2 ratio of times to the median time.

Suppose you have groups of data – before and after treatment, then you transform your data into ratio. In ratio data, values greater than 1 implies treated samples have a higher performance, and values below 1 implies the treatment has no effect.

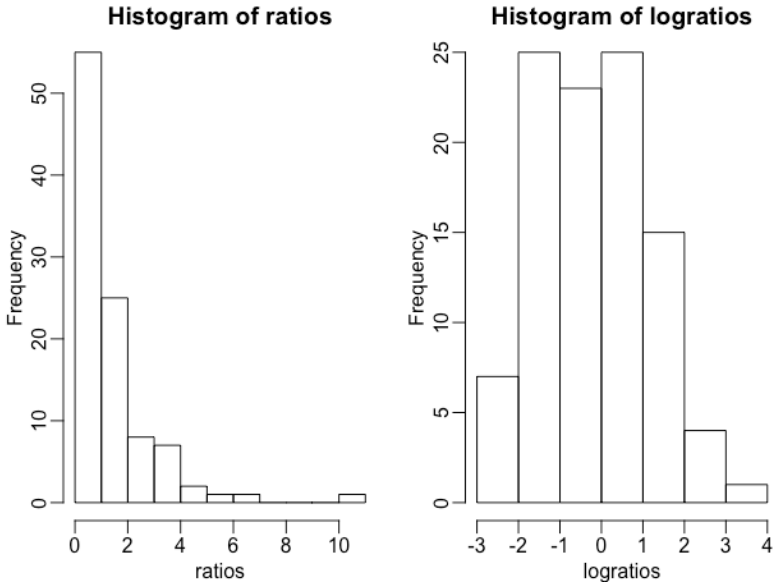


Figure 13 Histograms of ratios and logratios

The ratios are not symmetrical around 1, but log resolves the problem. Log is symmetric:

$$log\frac{x}{y} = \log(x) - \log(y) = -(\log(y) - \log(x)) = -\log\frac{y}{x}$$

In life science, log transformation is common because we are interested to quantify the results and fold changes are a good way to present data.

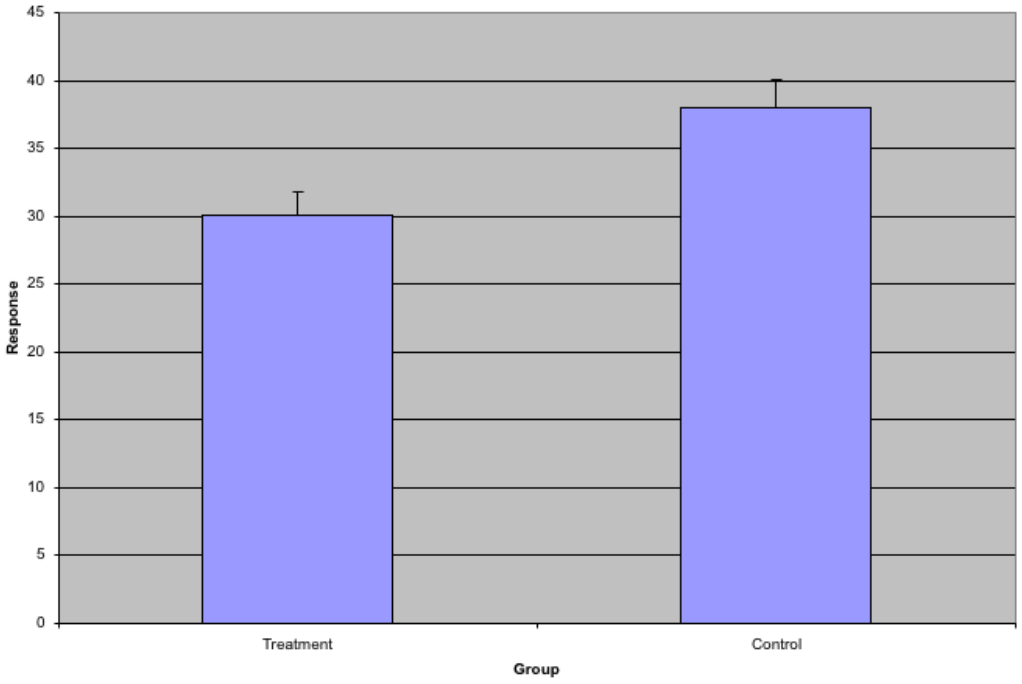
Plots to avoid

This section is devoted to discuss bad ways to display data based on the default plots provided by Microsoft Excel, and it is based on the talk by Dr. Karl W. Broman titled "How to display data badly".

The end goal of data graphics is to display data accurately and clearly.

1) **Bar plots for data summaries**

See also future reading:
Create bar plot in R: http://www.theanalysisfactor.com/r-11-bar-charts/



This bar plot created by Microsoft Excel only shows the group means and the standard errors. It also starts form zero, which takes up unnecessary space. A boxplot would have been more informative.

```
# load data
library("downloader")
filename <- "fig1.RData"
url <- "https://github.com/kbroman/Talk_Graphs/raw/master/R/fig1.RData"
if (!file.exists(filename)) download(url,filename)
load(filename)

# create boxplots
library(rafalib)
mypar(1)
file <- list(Treatment=x,Control=y)
boxplot(file,xlab="Group",ylab="Response",cex=0)
stripchart(file,vertical=TRUE,method="jitter",pch=16,add=TRUE,col=1)
```

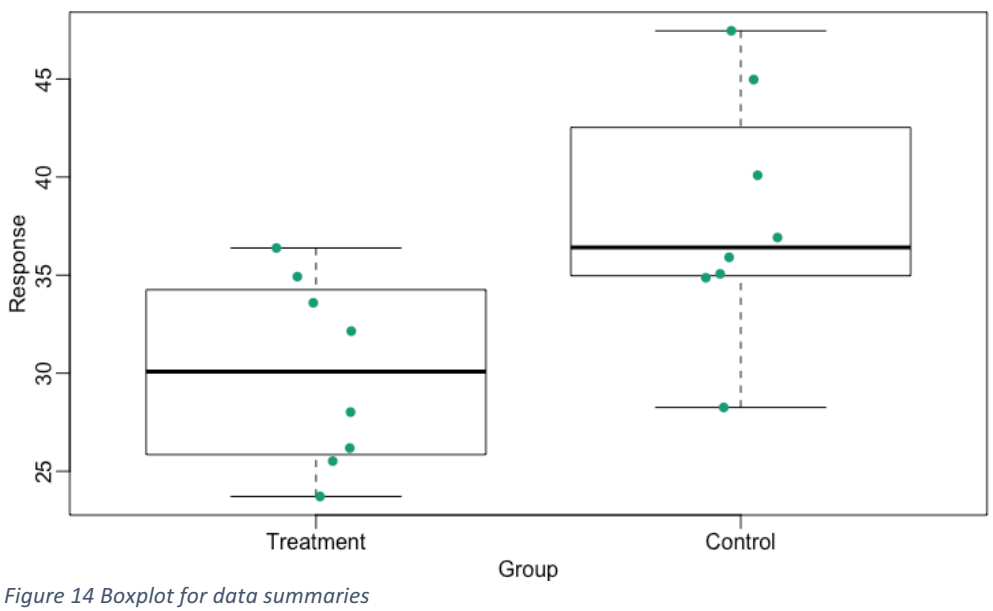
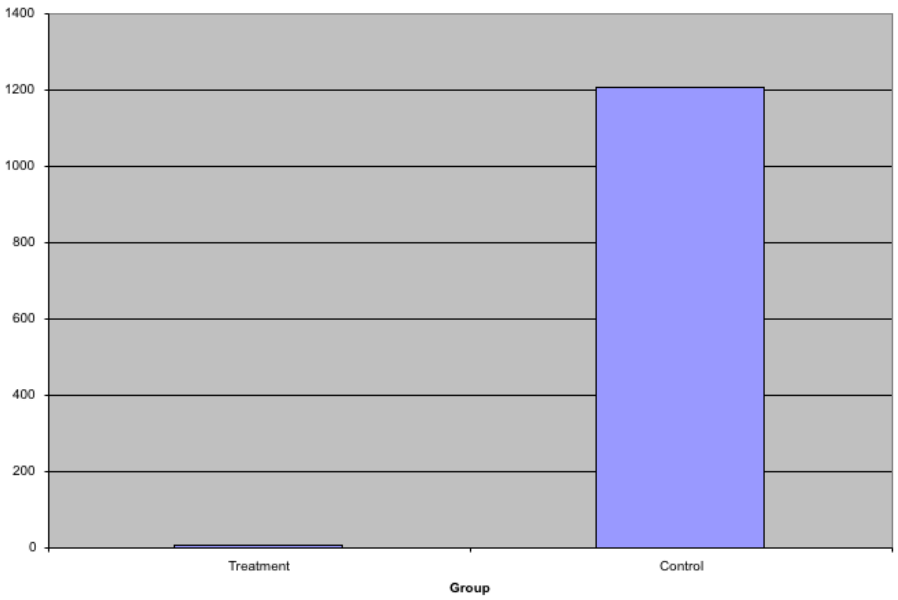


Figure 14 Boxplot for data summaries

Boxplot shows you the center (the median), spread, range and the points themselves. Bar plot only shows the mean and standard error, but standard error is influenced by sample size and not the spread of the data. Moreover, bar plots also magnify the effect of outliers and large tails.

If you have two groups of data (control and experimental), and the data has outliers and very large tails. Graph generated by default plot of Microsoft Excel will result in the following graphs:



This is how boxplots and log ratio plots are handy.

```
# load data
library("downloader")
url <- "https://github.com/kbroman/Talk_Graphs/raw/master/R/fig3.RData"
filename <- "fig3.RData"
if (!file.exists(filename)) download(url,filename)
load(filename)

# create boxplots
library(rafalib)
mypar(1,2)
file <- list(Treatment=x,Control=y)
boxplot(file,xlab="Group",ylab="Response",cex=0)
stripchart(file,vertical=TRUE,method="jitter",pch=16,add=TRUE,col=1)
boxplot(file,xlab="Group",ylab="Response",log="y",cex=0)
stripchart(file,vertical=TRUE,method="jitter",pch=16,add=TRUE,col=1)
```

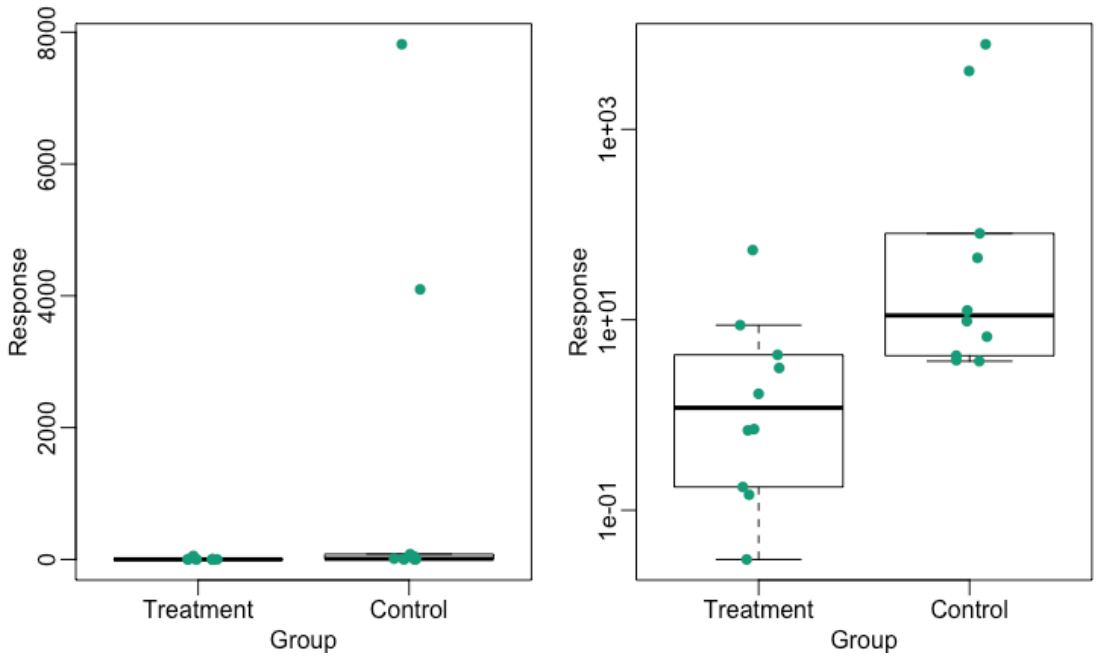


Figure 15 Boxplot and log data plot

In addition, for any data, it is clearer to round the value instead of reporting to 0.000001 cm. This can be done by: `round(x, 1)`. In which x equals to data, 1 equals to decimal, eg. 15.1.

2) Scatterplot

To determine the relationship between two variables, correlations and plots are used. However, having a regression line is a bad way to display data and it is more informative to show the scatter.

```
# load data
library("downloader")
url <- "https://github.com/kbroman/Talk_Graphs/raw/master/R/fig4.RData"
filename <- "fig4.RData"
if (!file.exists(filename)) download(url,filename)
load(filename)

# regression line
library(rafalib)
mypar(1,2)
plot(x,y,lwd=2,type="n")
fit <- lm(y~x)
abline(fit$coef,lwd=2)
## display the words – not relevant to calculation/ graph generation
b <- round(fit$coef,4)
text(78, 200, paste("y =", b[1], "+", b[2], "x"), adj=c(0,0.5))
rho <- round(cor(x,y),4)
text(78, 187,expression(paste(rho," = 0.8567")),adj=c(0,0.5))8

# scatterplots with regression line
plot(x,y,lwd=2)
fit <- lm(y~x)
abline(fit$coef,lwd=2)
```

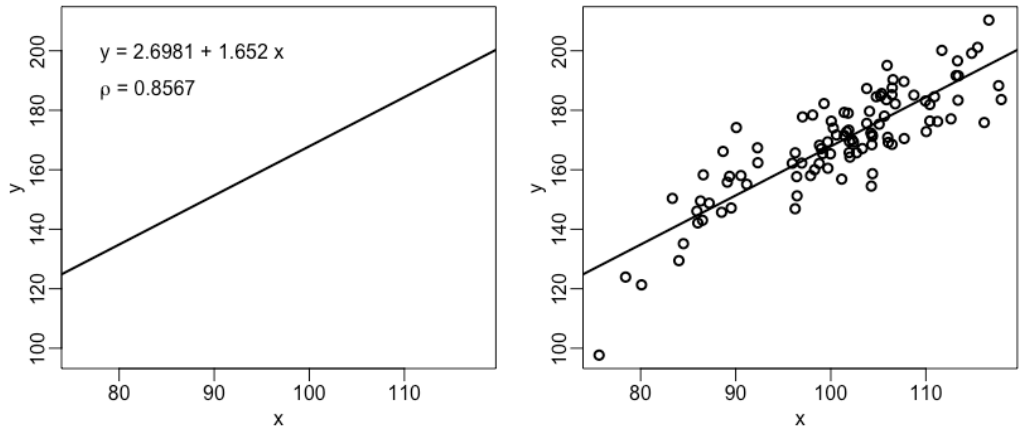
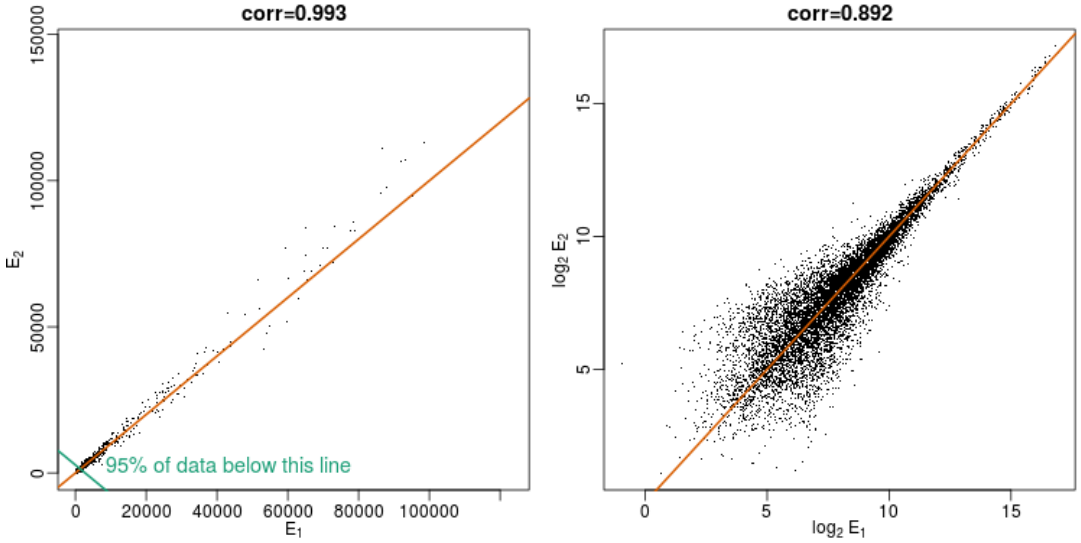
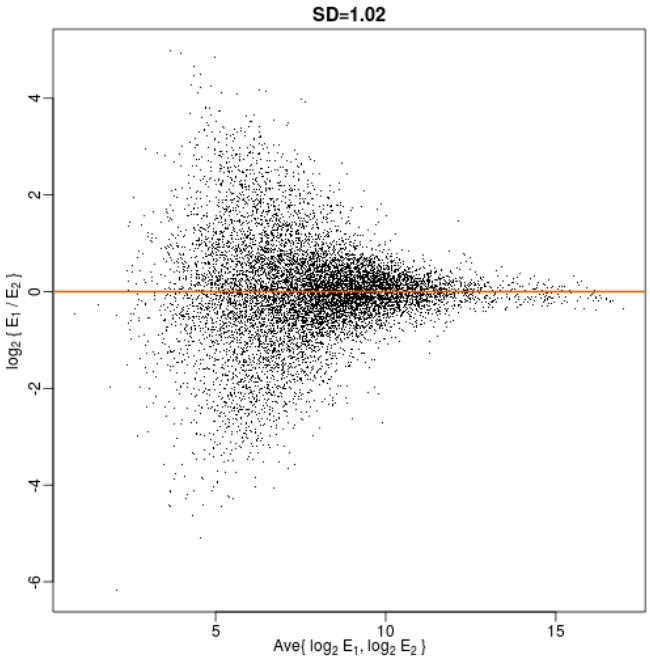


Figure 16 Regression line and Scatterplots

3) High correlation does not mean reproducibility



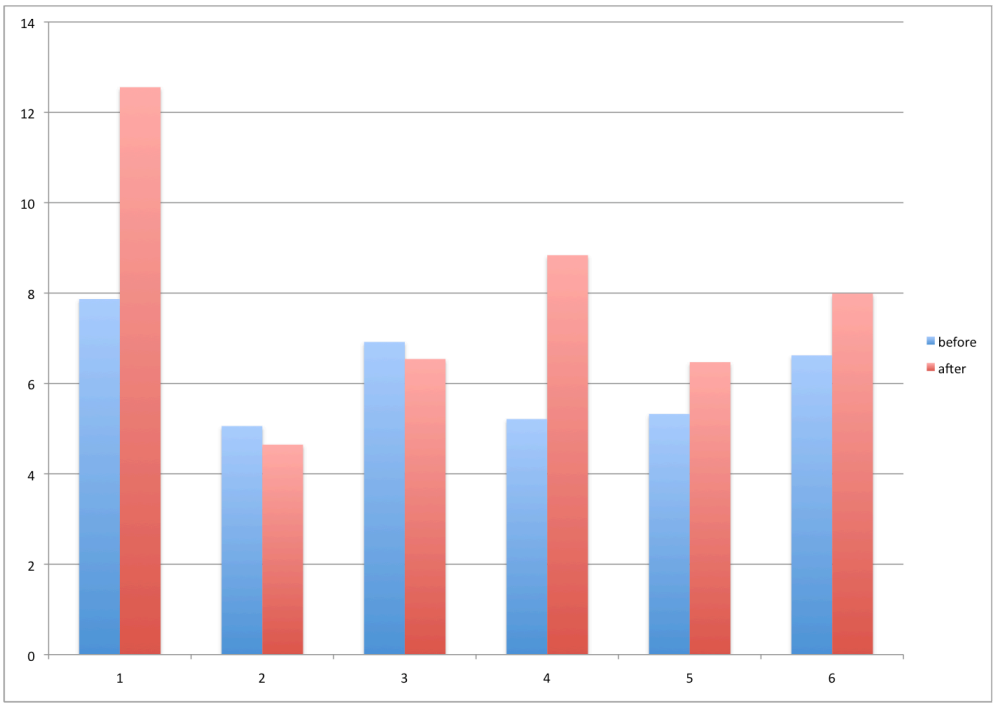
The correlation is different with and without log-scale. If we want to study the differences, we should plot the different in log scale versus the average:



The y-axis is the difference between two samples on the log scale, and the x-axis is the average of the samples on the log scale. This plot shows the difference in the log scale between the two replicated measures is around 1. When measurements are same, on average, we will observe 2-fold difference. We can then compare the variability to the differences we intend to detect.

As mentioned previously, pearson correlation can only be applied to bivariate normal data. Most gene expression data have a fat right tail distribution, and that violates the bivariate normal data assumption. Based on the correlation formula, correlation does not detect cases that are not reproducible because it looks at average changes. This can be overcome by a distance metric, which is beyond the scope of this manual.

4) Barplots for paired data



This plot shows the outcomes before and after a treatment. There are other ways to display data that illustrate an increase after treatment. The first way is to make a scatter plot, and the second way is to plot the differences against the before values.

```
set.seed(1)
before <- runif(6, 5, 8)
after <- rnorm(6, before*1.05, 2)
li <- range(c(before, after))
ymx <- max(abs(after-before))

mypar(1,2)
plot(before, after, xlab="Before", ylab="After", ylim=li, xlim=li)
abline(0,1, lty=2, col=1)
 plot(before, after-before, xlab="Before", ylim=c(-ymx, ymx), ylab="Change (After - Before)",
lwd=2)
abline(h=0, lty=2, col=1)
```

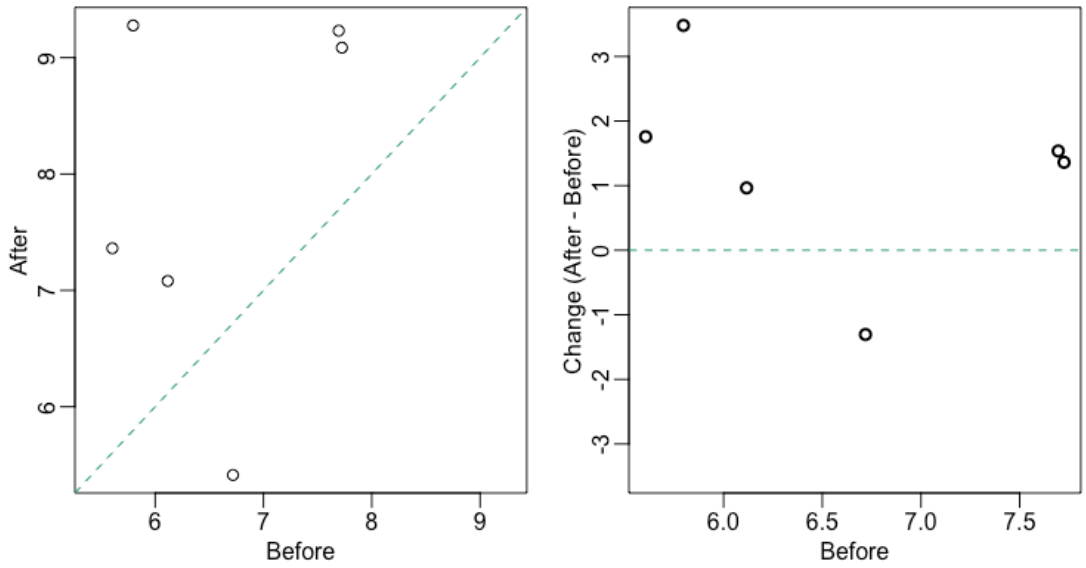


Figure 17 Scatter plot for paired data

The dashed line on the right plot is the identity line.

```
z <- rep(c(0,1), rep(6,2))
mypar(1,2)
difference <- c(before, after)
plot(z, difference, xaxt="n", ylab="Response", xlab="", xlim=c(-0.5, 1.5))
axis(side=1, at=c(0,1), c("Before","After"))
segments(rep(0,6), before, rep(1,6), after, col=1)
boxplot(before,after,names=c("Before","After"),ylab="Response")
```

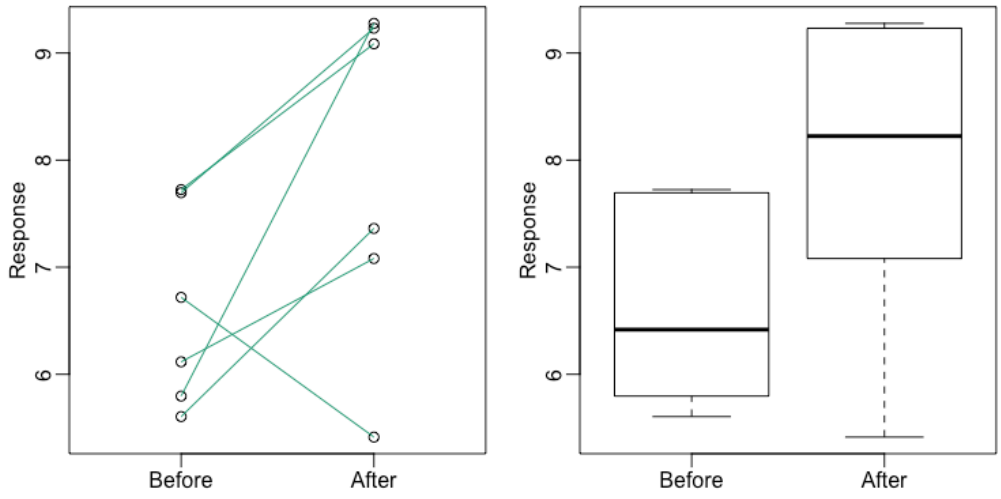
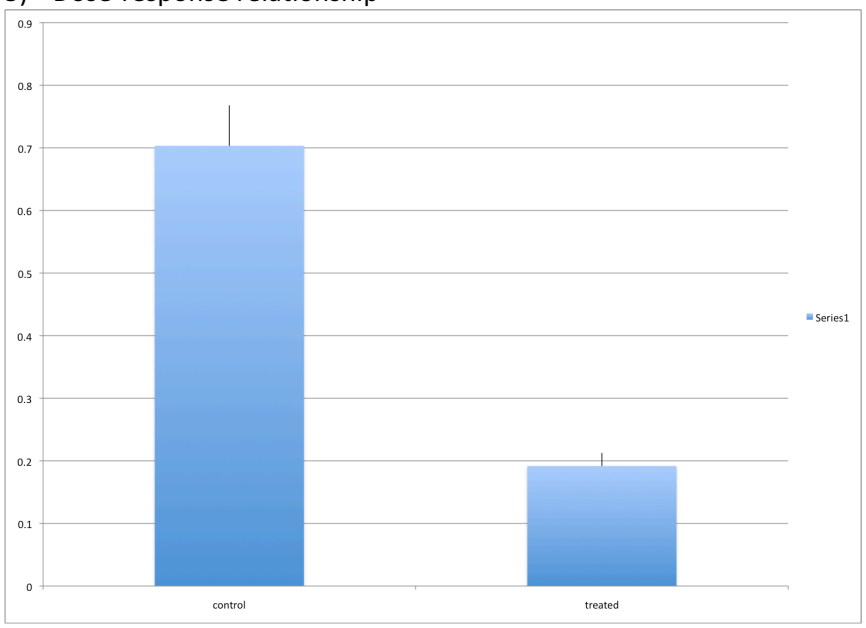


Figure 18 Line plots and box plots

Line plots are another option, but harder to visualize compare to the previous two. Boxplots lose the paired information, but does show the increase.

5) Dose-response relationship



This plot intends to show the dose-response relationship between the two groups: treatment and control. A better way is to show a line plot.

```
## load data
library(downloader)
filename <- "fig8dat.csv"
url <- "https://github.com/kbroman/Talk_Graphs/raw/master/R/fig8dat.csv"
if (!file.exists(filename)) download(url, filename)
x <- read.table(filename, sep=",", header=TRUE)

plot(x[,1],x[,2],xlab="log Dose",ylab="Proportion survived",ylim=c(0,1), type="l",lwd=2,col=1)
lines(x[,1],x[,3],lwd=2,col=2)
legend(1,0.4,c("control","treated"),lwd=2, col=1:3)
```

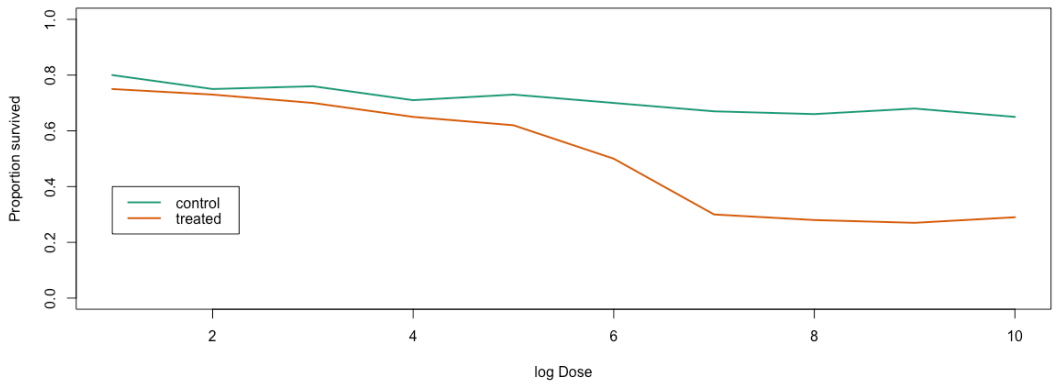


Figure 19 Dose-response curve

Lesson 6 –Normality tests and Non-parametric tests

Normality tests

Before we start learning non-parametric tests, I have modified a script to test normality. Normality tests answer: **is there convincing evidence of any derivation from the normal (Gaussian) distribution?** Normality tests are NOT used to decide when to abandon T-tests, and instead analyze transformed data or use a rank-based non-parametric test. For practical usage, scientists, however, tend to use normality tests as a guideline for choosing which test to apply. The main thing to consider when it comes to which test to apply is whether the errors are normally distributed. This is why the use of normality tests is controversial and there are lots of arguments regarding whether it is useful to use such test.

Furthermore, in this information explosion age, formal normality tests reject on the large sample sizes that most data scientists deal with on a daily basis. This is because every data has some degree of randomness. I believe for most wet-lab researchers in this lab, you will be dealing with small datasets. Normality tests in general work pretty well for small datasets. HOWEVER, and that is a big however, Dr. Ian Fellows (*Normality Tests don't do what you think they do*) has suggested that somehow normality tests have some degree of false positives and false negatives. So please do not be surprised if someone comes up to you and questions your use of normality tests.

I personally take normality tests with a grain of salt, and I consider them as a way of giving me another piece of evidence to back up my decision of using one test over the other. Therefore, I would recommend using the normality test script loosely to check if the distribution is normally distributed, and it is helpful to visualize outliers.

There are a few normality tests, such as Shapiro-Wilk test, Kolmogorov-Smirnov, Jarque-Bera, D'Agostino, and Anderson-Darling. One approach, if you have time, is to use more than one test and check for agreement. The default tests of normality in SPSS are Shapiro-Wilk test and Kolmogorov-Smirnov test. The Shapiro-Wilk test is more appropriate for small sample sizes (<50 samples), but can handle sample sizes as large as 2000, and it is used to test one group of sample. For two-sample data, Anderson-Darling and Kolmogorov-Smirnov tests are used. Anderson-Darling test is more sensitive to the tails of distribution, whereas Kolmogorov-Smirnov test is more aware of the center of the distribution. So long story short, Anderson-Darling test is assumed to be more powerful than Kolmogorov-Smirnov test.

The script is originally created by Todd Connelly (*Assumption Checking – Part 1*) for one-sample, and I modified that to apply for two-sample using both Kolmogorov-Smirnov test and Anderson-Darling test. The default of the script is to perform Kolmogorov-Smirnov test, but I show the results of Anderson-Darling test too. **If the results from Kolmogorov-Smirnov test and Anderson-Darling test conflicts, I will go with non-parametric test.** As usual, if the p-values are greater than 0.05, we can assume normality and t-tests can be applied. The script is in the supplemental package called "NormalityTest.R".

Robust summaries

Outliers skew our data and throw off our analysis, and mean and variance are not useful in this case. With outliers, it is recommended to use **median** and **median absolute deviation**.

Median absolute average (MAD) is defined by computing the differences between each point and the median. MAD is the equivalent of standard derivation.

$$1.4826 \times \text{median}\{|X_i - \text{median}(X_i)|\}$$

ie. Mean -> median

standard derivation -> median absolute derivation

```
x = c(rnorm(200, 0, 1)) # real distribution
median(x)
mad(x)
```

In general, if we know there will be outliers, we use median and MAD over mean and standard deviation. However, these robust statistics are less powerful than the non-robust ones. This is also the case for Pearson correlation versus Spearman correlation, and that Spearman correlation is more robust but has less power.

Optional Reading:
Spearman Correlation - http://www.professorserna.com/Free-Videos/STATS/STATSII/The-Spearman-Test-for-Correlation/The-Spearman-Test-for-Correlation.php

Rank Tests

When assumptions are met, parametric tests have higher testing power than nonparametric tests. We will discuss three commonly used non-parametric tests: the Mann-Whitney U test (aka Wilcoxon rank sum test), the Wilcoxon signed rank test, and the sign test.

Wilcoxon-Mann-Whitney test is a non-parametric test that is analogous to independent samples of t-test, and Wilcoxon-Mann-Whitney test can be used to compare ordinal data. The null hypothesis is that there is no difference between the two variables, and more precisely, there is no location difference between the two population distributions. This implies under the null hypothesis, the medians of the two populations are the same. Wilcoxon signed rank test or the sign test is used to replace paired t test.

Wilcoxon Rank Sum Test

We first discuss Wilcoxon Rank Sum test, also known as Mann-Whitney U test. This is a special type of permutation test. Basically, we combine all the data, turn the values into ranks, and separate them back into their groups. We then compute the sum or average rank to perform a test. Wilcoxon is generally used to examine the relationship between a numeric outcome variable and a 2-level categorical variable when groups are independent.

Assumptions:

1. Independent samples
2. Continuous variable

3. Equal variances with unequal sample sizes, variance is not an issue for same sample size
4. Two distributions have same or different shapes (details explained below)

For Wilcoxon Rank Sum test, if we have **the additional assumption that the distributions of scores for both groups of your independent variable have the same shape**, the null hypothesis suggests that there is no difference in the medians and that the total rank of one sample is close to the total rank of the other sample. The alternative hypothesis suggests the population medians are not equal, and that all the ranks of one sample are smaller than the ranks of the other. This reflects the shifted trend of the location between the two populations. This is obviously not the same for distributions with different shapes.

In essence, if the two distributions have **different shape**, the Wilcoxon Rank Sum test is used to **determine whether there are differences in the distributions of your two groups**. If the two distributions have the **same shape**, the Wilcoxon Rank Sum test is used to **determine whether there are differences in the medians of your two groups**.

There is, however, a significant downside of Wilcoxon-Mann-Whitney test. Once two groups show complete separation, points from higher group are always above points from the lower group and the statistic would not change regardless of the magnitude of the differences. In the similar vein, a p-value has a minimum value regardless of how far apart the groups are. For small sample sizes, the p-value cannot be very small even though the difference is huge. Moreover, **when the sample sizes are unequal and homogeneity of variance is not met, Welch T-test is better**; but if the datasets are not normally distributed with unequal variances, then a data transformation is recommended (we will not cover in this lesson). In essence, the power of Wilcoxon test is less powerful than the t-test.

If you are interested to learn how to do Wilcoxon Rank Sum test by hand, you can either google to find more information, or watch the following videos:
For samples smaller than 10, Wilcoxon Rank Sum Test Part I
http://www.professorserna.com/Free-Videos/STATS/STATSII/The-Wilcoxon-Rank-Sum-Test/The-Wilcoxon-Rank-Sum-Test.php

For both samples greater than 10 or when observations have the same value, the normal approximation is used and Wilcoxon Rank Sum Test is a Z test: Wilcoxon Rank Sum Test Part II
http://www.professorserna.com/Free-Videos/STATS/STATSII/The-Wilcoxon-Rank-Sum-Test-Part-II/The-Wilcoxon-Rank-Sum-Test-Part-II.php

To learn how to do Wilcoxon in R, section one of the script will teach you one data manipulation technique – reshape and another way to use plot, and section two of the script will teach Wilcoxon Sum Rank Test. The script will also illustrate the downside of Wilcoxon-Mann-Whitney test. Please
1) Put the the dataset "wilcoxonSum.csv" in your desired destination – you will need to use that folder as your current working directory. This data comes from Applegate, Tello, & Ying (2003).
2) Run the script "L6_wilcoxonSum.R" from supplemental packages.

Additional example: Wilcoxon Rank-Sum Test in R by UBC senior instructor - Mike Marin
https://www.youtube.com/watch?v=KroKhtCD9eE

Sign Test

A sign test decides whether a binomial distribution has the equal chance of success and failure. While we need to make additional assumption to say Wilcoxon rank sum test is a test of medians, sign test is the actual location-estimate test, and it tests the median averages of within-sample.

Sign test has less power compared to paired t test or Wilcoxon sign rank test. This is because sign test throw information away, but sign test outperforms Wilcoxon sign rank test for heavy-tailed distributions.

Watch from 4:51min to 14:36min, alternatively you can watch the entire video which is very helpful for understanding.
http://www.professorserna.com/Free-Videos/STATS/STATSII/The-Sign-Test/The-Sign-Test.php

In R, binom.test(x,y). For p-values greater than 0.05, we retain null hypothesis and that the two samples are equal / no difference.

Wilcoxon Signed Rank Test

Wilcoxon signed rank test is analogous to the paired two-sample t-test and it is used for data that are matched and from repeated observations of the same subject. Wilcoxon signed rank test accounts for the magnitude of difference within a case.

Assumptions:

1. Dependent samples – differences between a before and after measurement
2. Independence – paired observations are randomly and independently drawn
3. Continuous dependent variable

In R, wilcox.test(x,y, paired=TRUE). For p-values greater than 0.05, we retain null hypothesis and that the two samples are equal / no difference.

Again, for additional example: Wilcoxon Signed Rank Test in R by UBC senior instructor - Mike Marin
https://www.youtube.com/watch?v=zM8OZUM5I4Y

Closing remark

There are other non-parametric tests, but in essence, I will recommend Wilcoxon Rank Sum Test and Wilcoxon Sign Rank test over other non-parametric tests for two-sample comparisons.

For further reading:

Non-parametric test: Wilcoxon rank sum test
http://stats.stackexchange.com/questions/67204/what-exactly-does-a-non-parametric-test-accomplish-what-do-you-do-with-the-res/67210#67210

Wilcoxon Rank Sum test violations

http://www.basic.northwestern.edu/statguidefiles/rank_sum_ass_viol.html

Reference

[1] Applegate, K. E., Tello, R., & Ying, J. (2003). Hypothesis Testing III: Counts and Medians 1. *Radiology*, *228*(3), 603-608.

[2] Hart, A. (2001). Mann-Whitney test is not just a test of medians: differences in spread can be important. *British Medical Journal*, *323*(7309), 391.

Lesson 7 – Linear Models

Introduction to regression
Introduction to Linear models
Standard errors
Interactions and contrasts: ANOVA and F-test

- Kruskal-Wallis Test

Checking (G)LM model assumptions in R

Lesson 8 – Clustering Analysis

Multiple comparisons

- Benjamini-Hochberg Procedure: Handbook of biological statistics by John H. McDonald

Rank transformation approach: modified Bonferroni procedures such as Hochberg or Westfall procedures