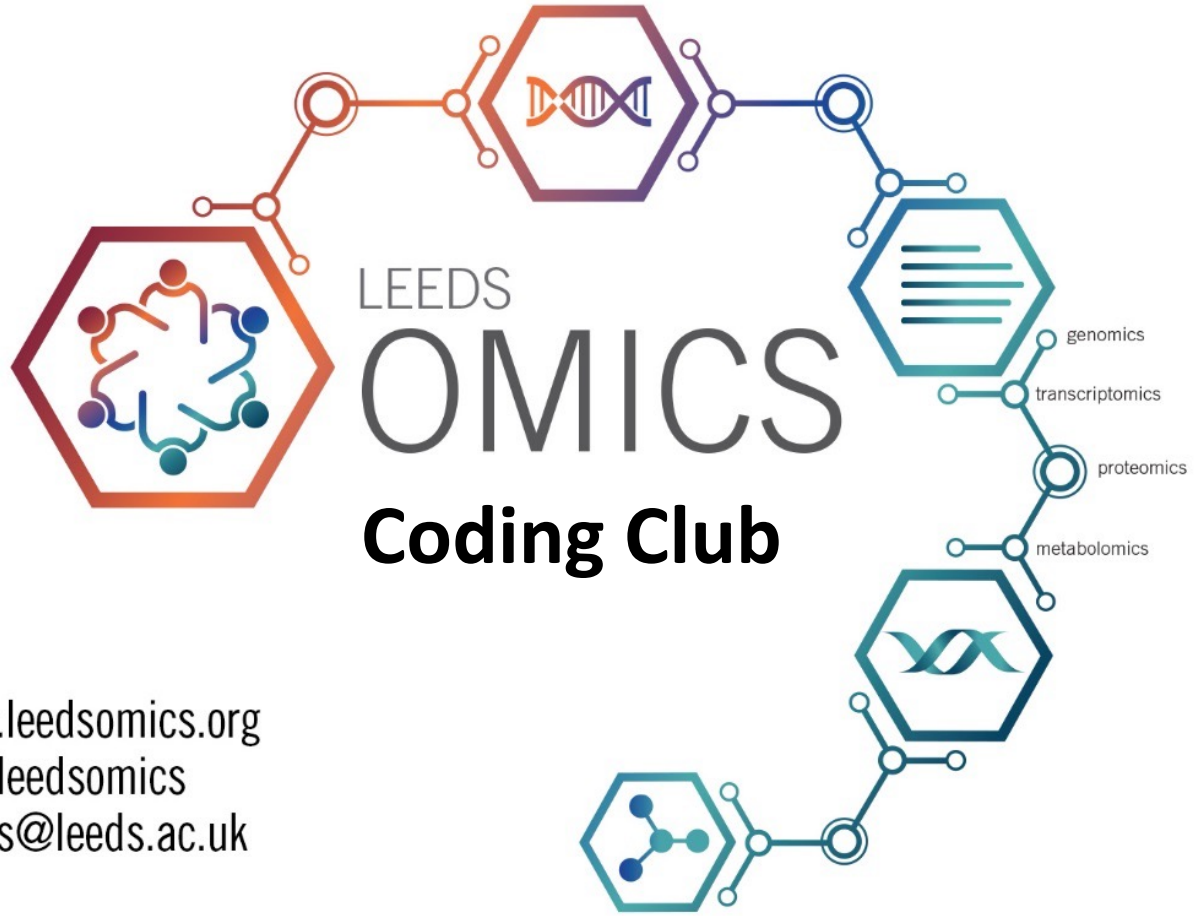



**Coding Club**

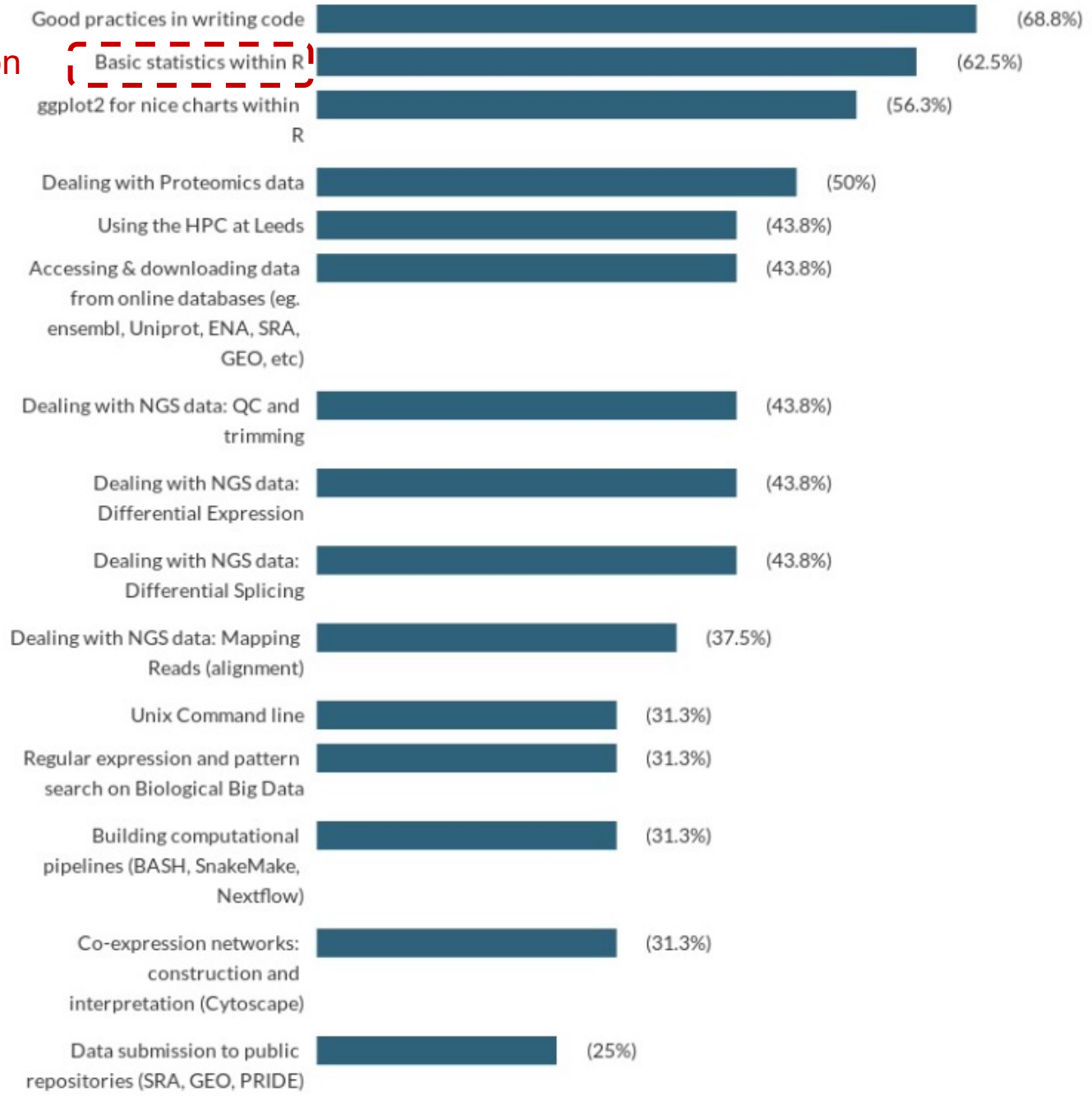
www.leedsomics.org
@leedsomics
omics@leeds.ac.uk

# Basic Statistics within R

**Club Moderators:** Elton Vasconcelos and Euan McDonnell

# Topics to be addressed - **Survey Result (2021-22)**

2nd session



# R has its own command line environment

## Table of Useful R commands

https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

| Command | Purpose |
|---|---|
| help() | Obtain documentation for a given R command |
| example() | View some examples on the use of a command |
| c(), scan() | Enter data manually to a vector in R |
| seq() | Make arithmetic progression vector |
| rep() | Make vector of repeated values |
| data() | Load (often into a data.frame) built-in dataset |
| View() | View dataset in a spreadsheet-type format |
| str() | Display internal structure of an R object |
| read.csv(), read.table() | Load into a data.frame an existing data file |
| library(), require() | Make available an R add-on package |
| dim() | See dimensions (# of rows/cols) of data.frame |
| length() | Give length of a vector |
| ls() | Lists memory contents |
| rm() | Removes an item from memory |
| names() | Lists names of variables in a data.frame |
| hist() | Command for producing a histogram |
| histogram() | Lattice command for producing a histogram |
| stem() | Make a stem plot |
| table() | List all values of a variable with frequencies |
| xtabs() | Cross-tabulation tables using formulas |
| mosaicplot() | Make a mosaic plot |
| cut() | Groups values of a variable into larger bins |
| mean(), median() | Identify "center" of distribution |
| by() | apply function to a column split by factors |
| summary() | Display 5-number summary and mean |
| var(), sd() | Find variance, sd of values in vector |
| sum() | Add up all values in a vector |
| quantile() | Find the position of a quantile in a dataset |
| barplot() | Produces a bar graph |
| barchart() | Lattice command for producing bar graphs |
| boxplot() | Produces a boxplot |
| bwplot() | Lattice command for producing boxplots |

| Command | Purpose |
|---|---|
| plot() | Produces a scatterplot |
| xyplot() | Lattice command for producing a scatterplot |
| lm() | Determine the least-squares regression line |
| anova() | Analysis of variance (can use on results of lm()) |
| predict() | Obtain predicted values from linear model |
| nls() | estimate parameters of a nonlinear model |
| residuals() | gives (observed - predicted) for a model fit to data |
| sample() | take a sample from a vector of data |
| replicate() | repeat some process a set number of times |
| cumsum() | produce running total of values for input vector |
| ecdf() | builds empirical cumulative distribution function |
| dbinom(), etc. | tools for binomial distributions |
| dpois(), etc. | tools for Poisson distributions |
| pnorm(), etc. | tools for normal distributions |
| qt(), etc. | tools for student $t$ distributions |
| pchisq(), etc. | tools for chi-square distributions |
| binom.test() | hypothesis test and confidence interval for 1 proportion |
| prop.test() | inference for 1 proportion using normal approx. |
| chisq.test() | carries out a chi-square test |
| fisher.test() | Fisher test for contingency table |
| t.test() | student t test for inference on population mean |
| qqnorm(), qqline() | tools for checking normality |
| addmargins() | adds marginal sums to an existing table |
| prop.table() | compute proportions from a contingency table |
| par() | query and edit graphical settings |
| power.t.test() | power calculations for 1- and 2-sample $t$ |
| anova() | compute analysis of variance table for fitted model |

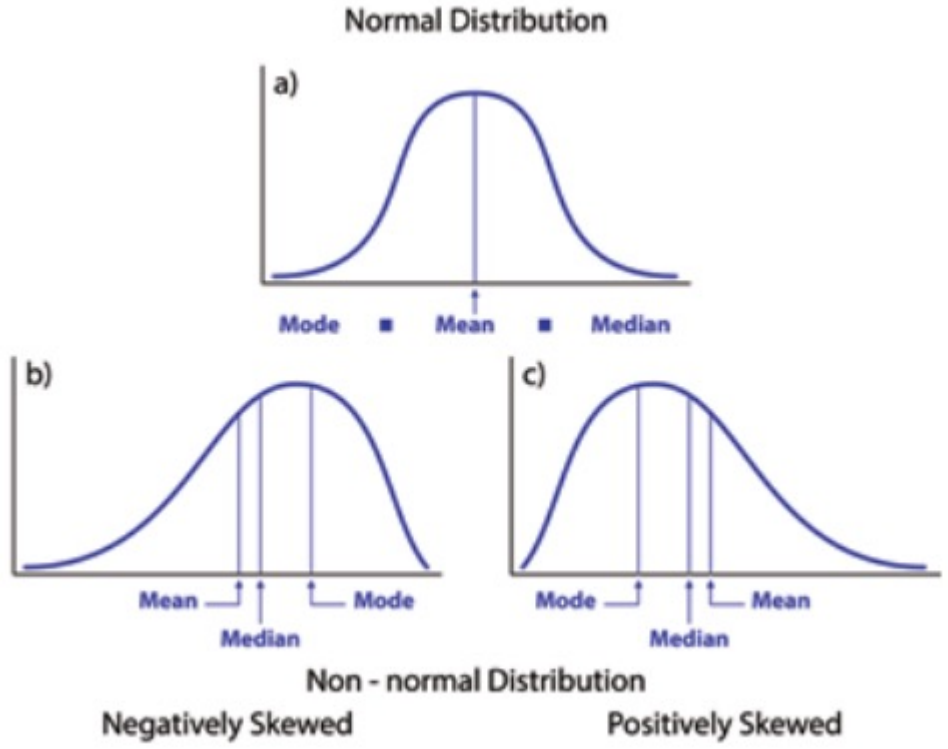
https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

http://www.math.umt.edu/olear/stat458/Rseminar_2.pdf

→ Other useful R material for beginners

# Common practice prior running any statistics on your data

**Check the distribution of whatever data values your samples have**




https://heljves.com/gallery/vol_1_issue_1_2019_8.pdf (*Antonopoulos & Kakisis, 2019*)

# Which test to use?

| TYPE OF VARIABLE | Continuous | | | | | | | | | Categorical | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DISTRIBUTION | Normal distribution | | | | | Skewed distribution | | | | | |
| No of GROUPS | 1 group | 2 groups | | >2 groups | | 1 group | 2 groups | | >2 groups | | |
| Independence between GROUPS | | Independent | Dependent | Independent | Dependent | | Independent | Dependent | Independent | Dependent | Independent | Dependent |
| STATISTICAL TEST | 1-sample t-test | Independent sample t-test | Paired sample t-test | ANOVA | Repeated-measures ANOVA | Sign test, Wilcoxon signed ranks test | Mann-Whitney U test | Sign test, Wilcoxon signed ranks test | Kruskal-Wallis H test | Friedman test | Chi-square test, Fischer exact test | McNemar's test |
| EXAMPLE | Mean age of all patients with AAA | Mean age of AAA patients treated with OSR vs. EVAR | Difference in aortic diameter in patients before vs. after EVAR | Mean age of AAA patients treated with OSR vs. EVAR vs. conservative treatment | Difference in aortic diameter in patients before vs. after 6 months vs. after 1 year of EVAR | Mean age of all patients with AAA | Mean age of AAA patients treated with OSR vs. EVAR | Difference in aortic diameter in patients before vs. after EVAR | Mean age of AAA patients treated with OSR vs. EVAR vs. conservative treatment | Difference in aortic diameter in patients before vs. after 6 months vs. after 1 year of EVAR | Difference in males/ females among patients treated with OSR vs. EVAR | Difference in number of patients with excluded/ non-excluded aneurysmal sac after 6 months vs. after 1 year of EVAR |

**Table 1**. A statistical algorithm depicting which test to use, based on the type of variable, data distribution, number of groups and independence between groups

**Abbreviations:** AAA: Abdominal Aortic Aneurysm, ANOVA: Analysis of Variance, EVAR: Endovascular Aortic Repair, OSR: Open Surgical Repair

https://heljves.com/gallery/vol_1_issue_1_2019_8.pdf (*Antonopoulos & Kakisis, 2019*)

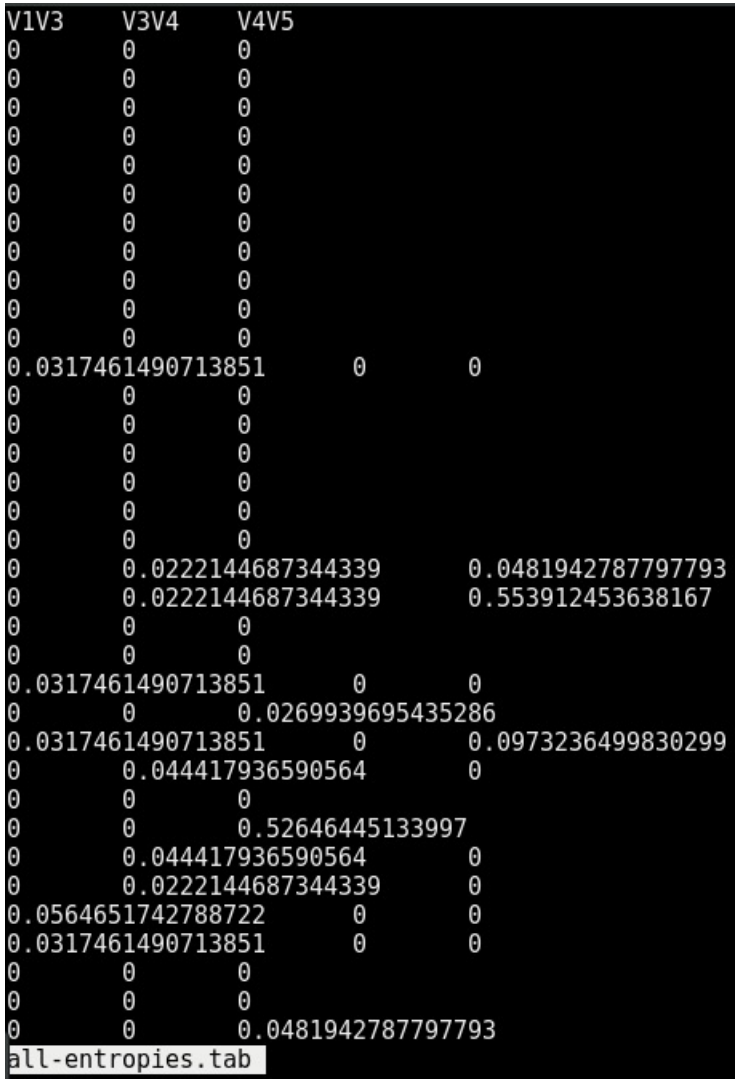
# Summary of some basic statistical tests in R

| R function | Test | Parametric (p) or Non-Parametric (np) | Purpose |
|---|---|---|---|
| t.test | T-test | p | pairwise |
| wilcox.test | Mann-Whitney-Wilcoxon | np | pairwise |
| aov | ANOVA | p | >2 groups |
| kruskal.test | Kruskal-Wallis | np | >2 groups |
| chisq.test | Chi-Square | np | Categorical: 2x2 contingency table |
| fisher.test | Fisher Exact Test | np | Categorical: 2x2 contingency table |
| binom.test | Binomial Test | np | #successes and failures on *n* trials |
| ks.test | Kolmogorov-Smirnov | np | Equality of cumulative distributions |
| cor.test(x, y, method = “pearson”) | Pearson Correlation (*r*) | p | Correlation coefficient from pairs of numerical vectors |
| cor.test(x, y, method = “spearman”) | Spearman Correlation (*rho*) | np | Correlation coefficient from pairs of numerical vectors |

Type ? followed by the function name within R or visit https://www.statmethods.net/stats/index.html for more details

# Practical example on an *Omics* context

→ Are there differences between variable regions from the 16S rRNA evolutionary marker gene regarding discriminatory power for taxonomic classification of vector-borne bacterial pathogens (VBPs)?



```
V1V3    V3V4    V4V5
0       0       0
0       0       0
0       0       0
0       0       0
0       0       0
0       0       0
0       0       0
0       0       0
0       0       0
0       0       0
0       0       0
0       0       0
0.0317461490713851      0       0
0       0       0
0       0       0
0       0       0
0       0       0
0       0       0
0       0       0
0       0.0222144687344339      0.0481942787797793
0       0.0222144687344339      0.553912453638167
0       0       0
0       0       0
0.0317461490713851      0       0
0       0       0.0269939695435286
0.0317461490713851      0       0.0973236499830299
0       0.044417936590564       0
0       0       0
0       0       0.52646445133997
0       0.044417936590564       0
0       0.0222144687344339      0
0.0564651742788722      0       0
0.0317461490713851      0       0
0       0       0
0       0       0
0       0       0.0481942787797793
all-entropies.tab
```

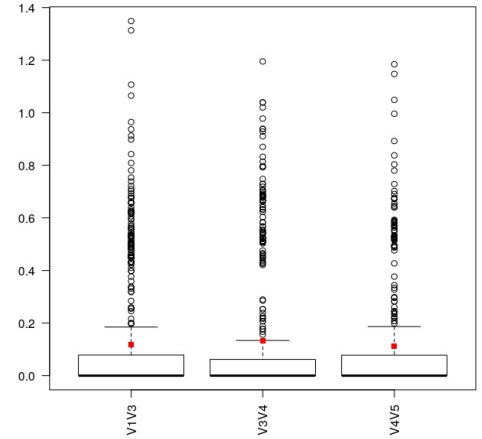
**← Input file**

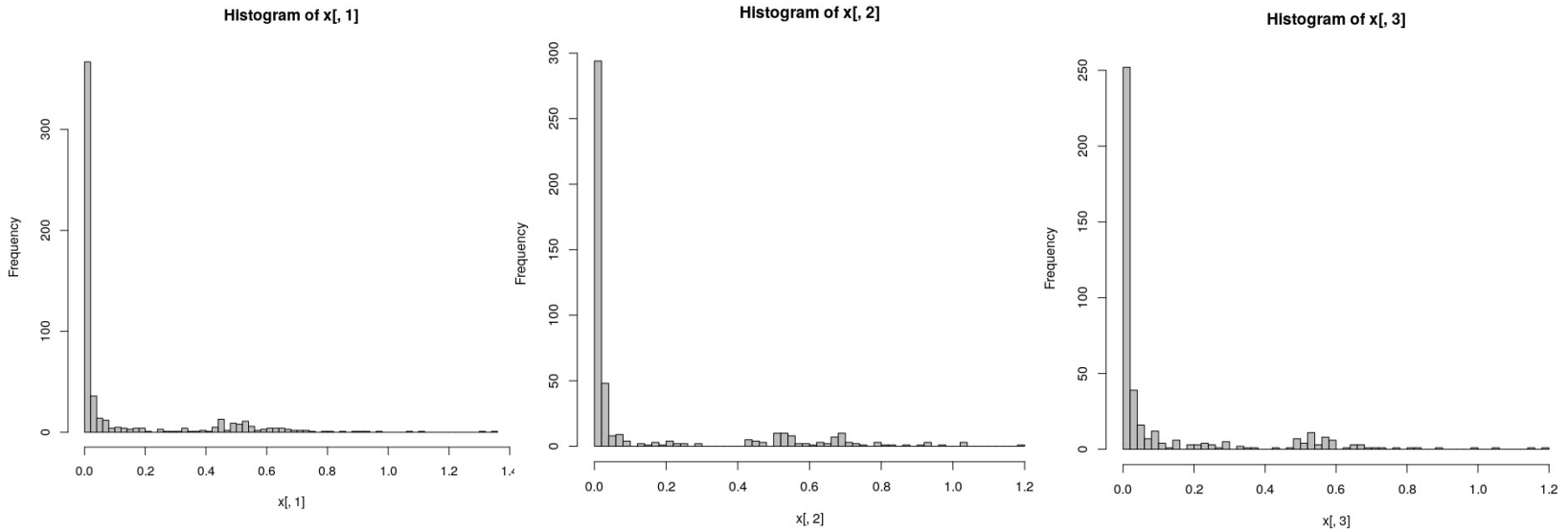
Shannon entropy values per individual position of a 320 VBPs-containing multiple sequence alignment (MSA) from three 16S variable regions

`diversity()` function from the ”vegan” R package

# Checking entropy values' distribution with both boxplot and histograms

```
x = read.delim("all-entropies.tab")
dim(x)
head(x)
allVmeans = NULL
for(i in 1:3){ allVmeans = as.vector(c(allVmeans, mean(x[,i], na.rm = T))) }
boxplot(x, las=2)
points(allVmeans, pch = 15, col = "red")
hist(x[,1], col = "gray", breaks = 50)
hist(x[,2], col = "gray", breaks = 50)
hist(x[,3], col = "gray", breaks = 50)
```




Shapiro-Wilk Test of normality

`shapiro.test()`

*p*-value > 0.05 = normal distribution

*p*-value < 0.05 = non-normal distribution

## Opting for non-parametric tests

```
> kruskal.test(x)

        Kruskal-Wallis rank sum test

data:  x
Kruskal-Wallis chi-squared = 1.2172, df = 2, p-value = 0.5441
```

```
> wilcox.test(x[,1], x[,2])$p.value
[1] 0.4414327
> wilcox.test(x[,1], x[,3])$p.value
[1] 0.3318844
> wilcox.test(x[,2], x[,3])$p.value
[1] 0.5765521
```

Answer:

No, the three assessed 16S variable regions have the same discriminatory power for the 320 VBP species/strains under study.

Bring your issues on!