

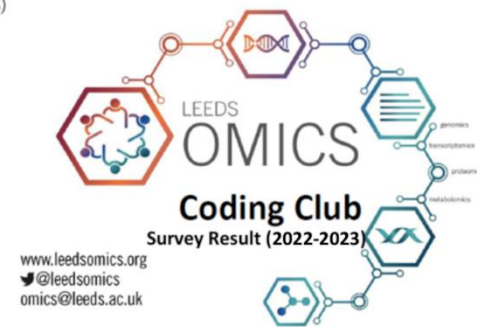
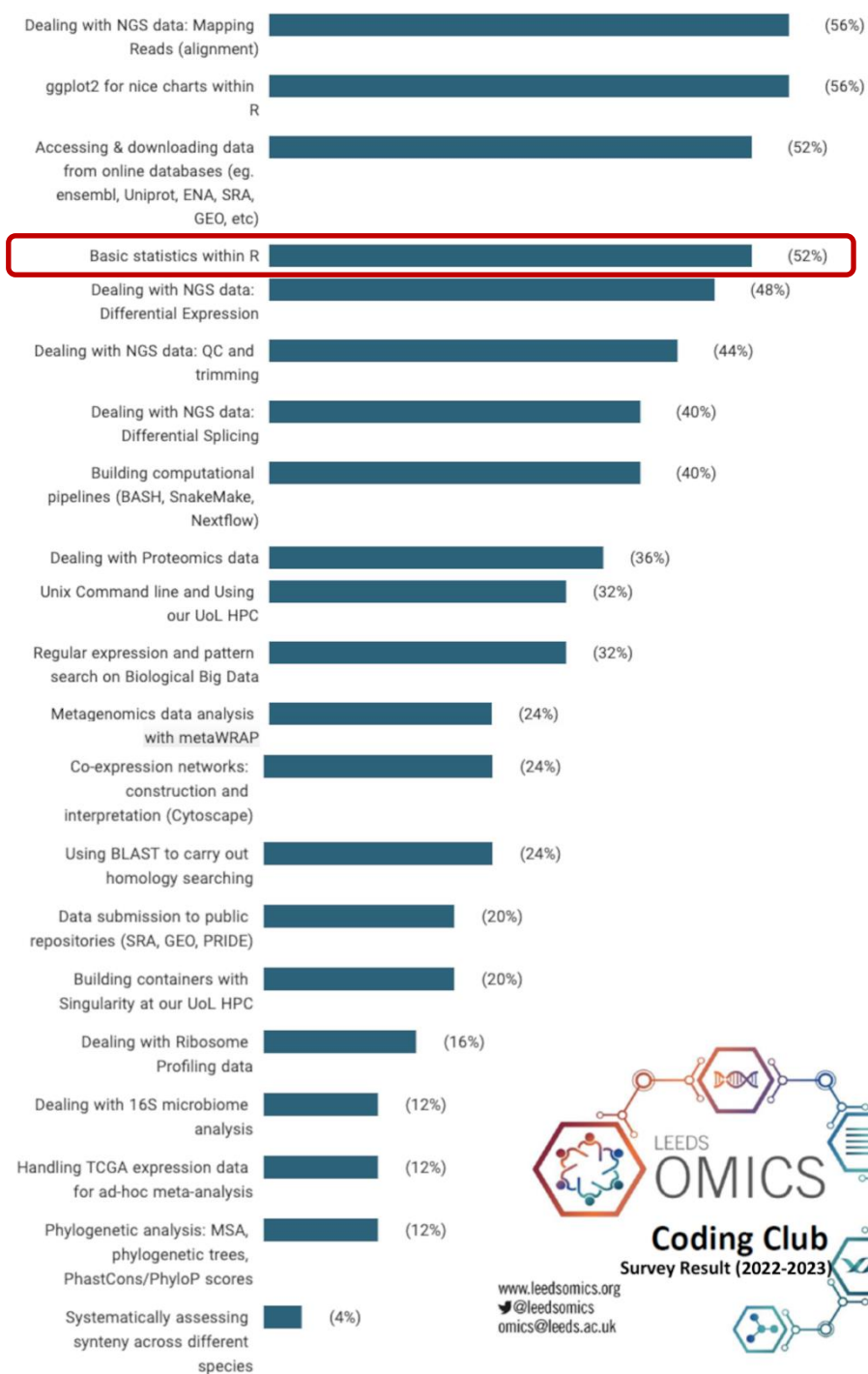
www.leedsomics.org
@leedsomics
omics@leeds.ac.uk

Basic Statistics within R

Club Moderators: Elton Vasconcelos and Euan McDonnell

Topics to be addressed on the 2022-23 season - Survey Result

4th 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	Command	Purpose
<code>help()</code>	Obtain documentation for a given R command	<code>plot()</code>	Produces a scatterplot
<code>example()</code>	View some examples on the use of a command	<code>xyplot()</code>	Lattice command for producing a scatterplot
<code>c()</code> , <code>scan()</code>	Enter data manually to a vector in R	<code>lm()</code>	Determine the least-squares regression line
<code>seq()</code>	Make arithmetic progression vector	<code>anova()</code>	Analysis of variance (can use on results of <code>lm()</code>)
<code>rep()</code>	Make vector of repeated values	<code>predict()</code>	Obtain predicted values from linear model
<code>data()</code>	Load (often into a data.frame) built-in dataset	<code>nls()</code>	estimate parameters of a nonlinear model
<code>View()</code>	View dataset in a spreadsheet-type format	<code>residuals()</code>	gives (observed - predicted) for a model fit to data
<code>str()</code>	Display internal structure of an R object	<code>sample()</code>	take a sample from a vector of data
<code>read.csv()</code> , <code>read.table()</code>	Load into a data.frame an existing data file	<code>replicate()</code>	repeat some process a set number of times
<code>library()</code> , <code>require()</code>	Make available an R add-on package	<code>cumsum()</code>	produce running total of values for input vector
<code>dim()</code>	See dimensions (# of rows/cols) of data.frame	<code>ecdf()</code>	builds empirical cumulative distribution function
<code>length()</code>	Give length of a vector	<code>dbinom()</code> , etc.	tools for binomial distributions
<code>ls()</code>	Lists memory contents	<code>dpois()</code> , etc.	tools for Poisson distributions
<code>rm()</code>	Removes an item from memory	<code>pnorm()</code> , etc.	tools for normal distributions
<code>names()</code>	Lists names of variables in a data.frame	<code>qt()</code> , etc.	tools for student <i>t</i> distributions
<code>hist()</code>	Command for producing a histogram	<code>pchisq()</code> , etc.	tools for chi-square distributions
<code>histogram()</code>	Lattice command for producing a histogram	<code>binom.test()</code>	hypothesis test and confidence interval for 1 proportion
<code>stem()</code>	Make a stem plot	<code>prop.test()</code>	inference for 1 proportion using normal approx.
<code>table()</code>	List all values of a variable with frequencies	<code>chisq.test()</code>	carries out a chi-square test
<code>xtabs()</code>	Cross-tabulation tables using formulas	<code>fisher.test()</code>	Fisher test for contingency table
<code>mosaicplot()</code>	Make a mosaic plot	<code>t.test()</code>	student <i>t</i> test for inference on population mean
<code>cut()</code>	Groups values of a variable into larger bins	<code>qqnorm()</code> , <code>qqline()</code>	tools for checking normality
<code>mean()</code> , <code>median()</code>	Identify “center” of distribution	<code>addmargins()</code>	adds marginal sums to an existing table
<code>by()</code>	apply function to a column split by factors	<code>prop.table()</code>	compute proportions from a contingency table
<code>summary()</code>	Display 5-number summary and mean	<code>par()</code>	query and edit graphical settings
<code>var()</code> , <code>sd()</code>	Find variance, sd of values in vector	<code>power.t.test()</code>	power calculations for 1- and 2-sample <i>t</i>
<code>sum()</code>	Add up all values in a vector	<code>anova()</code>	compute analysis of variance table for fitted model
<code>quantile()</code>	Find the position of a quantile in a dataset		
<code>barplot()</code>	Produces a bar graph		
<code>barchart()</code>	Lattice command for producing bar graphs		
<code>boxplot()</code>	Produces a boxplot		
<code>bwplot()</code>	Lattice command for producing boxplots		

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

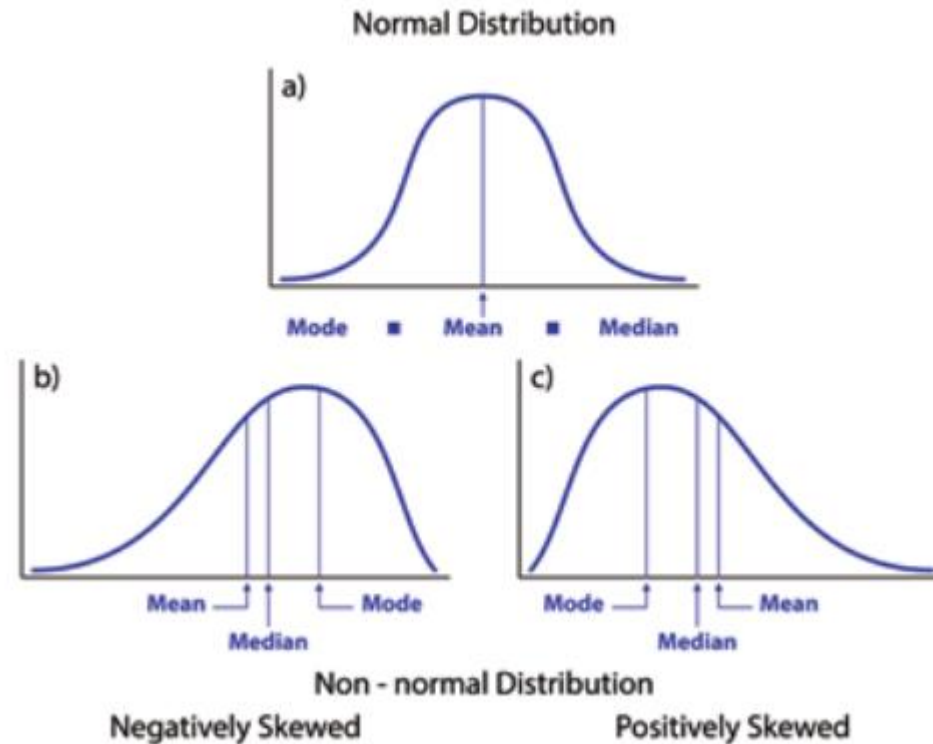
http://www.math.umt.edu/olear/stat458/Rseminar_2.pdf



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



Which test to use?

TYPE OF VARIABLE	Continuous						Categorical					
	Normal distribution				Skewed distribution							
DISTRIBUTION	2 groups		>2 groups		1 group		2 groups		>2 groups		1 group	
No of GROUPS	Independent		Dependent		Independent		Dependent		Independent		Dependent	
Independence between GROUPS	1 group	Independent	Dependent	Independent	Dependent	1 group	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

Summary of some basic statistical tests in R

R function	Test	Parametric (p) or Non-Parametric (np)	Purpose
<code>t.test</code>	T-test	p	pairwise
<code>wilcox.test</code>	Mann-Whitney-Wilcoxon	np	pairwise
<code>aov</code>	ANOVA	p	>2 groups
<code>kruskal.test</code>	Kruskal-Wallis	np	>2 groups
<code>chisq.test</code>	Chi-Square	np	Categorical: 2x2 contingency table
<code>fisher.test</code>	Fisher Exact Test	np	Categorical: 2x2 contingency table
<code>binom.test</code>	Binomial Test	np	#successes and failures on n trials
<code>ks.test</code>	Kolmogorov-Smirnov	np	Equality of cumulative distributions
<code>cor.test(x, y, method = "pearson")</code>	Pearson Correlation (r)	p	Correlation coefficient from pairs of numerical vectors
<code>cor.test(x, y, method = "spearman")</code>	Spearman Correlation (ρ)	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
0.0317461490713851    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
```

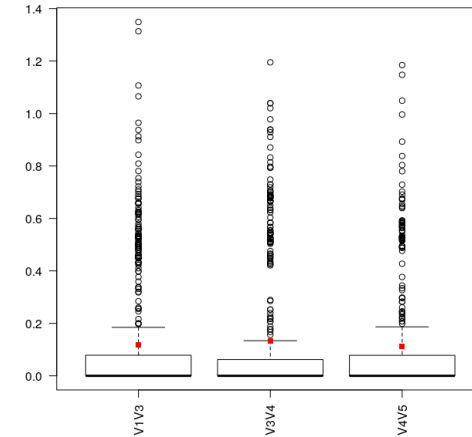
← Input file

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

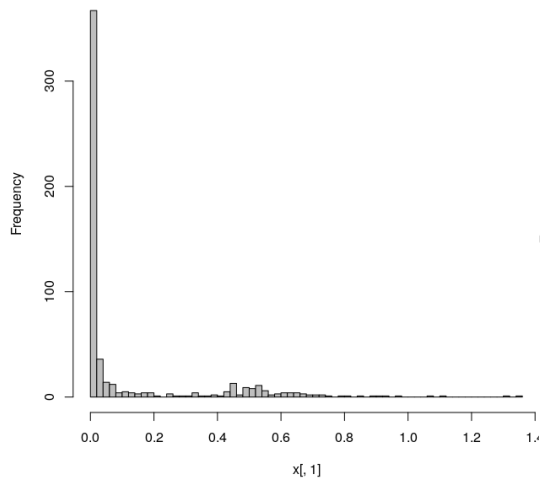
diversity() function from the "vegan" R package

Checking entropy values' distribution with both boxplot and histograms

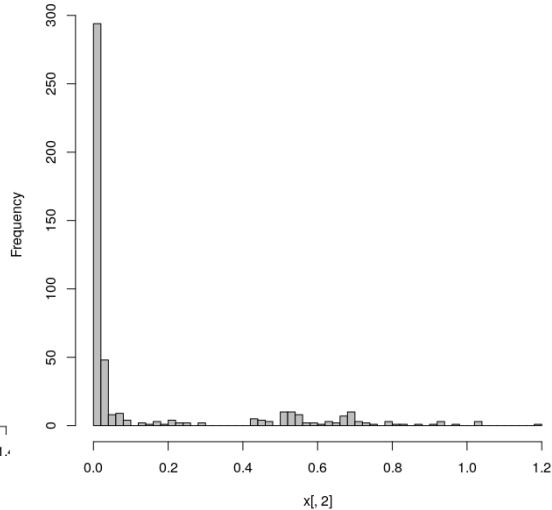
```
x = read.delim("all-entropies.tab")
dim(x)
head(x)
allVmeans = NULL
for(i in 1:3){ allVmeans = as.vector(c(allVmeans, mean(x[,i], na.rm = T))) }
boxplot(x, las=2)
points(allVmeans, pch = 15, col = "red")
hist(x[,1], col = "gray", breaks = 50)
hist(x[,2], col = "gray", breaks = 50)
hist(x[,3], col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



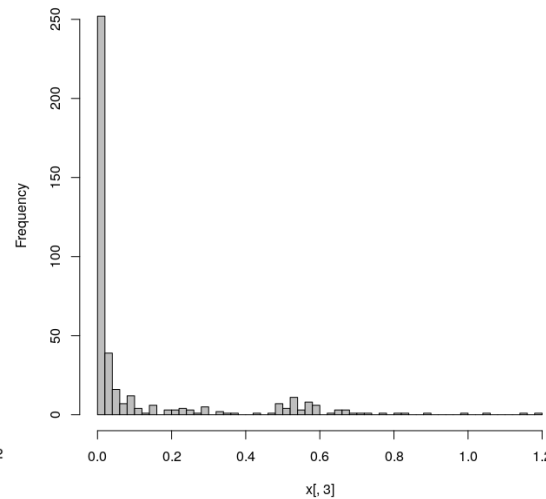
Histogram of x[, 1]



Histogram of x[, 2]



Histogram of x[, 3]



Shapiro-Wilk Test of normality
`shapiro.test()`
 $p\text{-value} > 0.05 = \text{normal distribution}$
 $p\text{-value} < 0.05 = \text{non-normal distribution}$

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!