

Basic Statistics in R

Created by Michelle Evans

Presented by
Santino Andry & Angelo Andrianiana

Basic Statistics Topics

1. Importing Data (*Importation des données*)
2. Data Visualization and Exploration (*Exploration et visualisation des données*)
3. Verifying Model Assumptions (*Vérification des hypothèses de modèles*)
4. Conducting Correlations (*Analyse de correlations*)
5. Comparing Data Between Groups (*Comparaison de données entre groupes*)
 - a. Parametric (*Paramétrique*)
 - b. Non-Parametric (*Non-Paramétrique*)

Importing Data - Creating an environment for your project

1. Using an R Project (*Utilisation des Rproject*)
2. Folder structure (*Structure des dossiers*)
3. Using reproducible research documents (*Utilisation des documents reproducible*)
(Rmd, quarto)

Create a folder structure

1. Scripts: all your .R files go here

Tous les fichiers .R sont ici

1. Data: All of your data goes here. It is best to make two subdirectories: 'raw' and 'clean'

Les données sont ici. Le meilleur pratique est de créer deux sous-dossiers: 'brut' et 'nettoyé'

1. Results: Results of your analysis will go here. This includes tables of summary statistics, figures, and results of statistical tests

Les résultats des analyses sont ici. Cela inclut les tableaux des statistiques sommaires, les figures, et les résultats des analyses

Name	Size	Modifié
data	0 items	17:00
results	0 items	17:00
scripts	0 items	17:00
.Rproj.user	2 items	17:00
E2M2.Rproj	205 bytes	17:00

Open a quarto document (*ouvrir un document quarto*)

Source

```
## Quarto

Quarto enables you to weave together content and executable code into a
document. To learn more about Quarto see <https://quarto.org>.

## Running Code

When you click the **Render** button a document will be generated that includes
both content and the output of embedded code. You can embed code like this:

```{r}
1 + 1
...

You can add options to executable code like this

```{r}
#| echo: false
2 * 2
...

The 'echo: false' option disables the printing of code (only output is displayed)
```

Visual

Quarto enables you to weave together content and executable code into a document. To learn more about Quarto see <https://quarto.org>.

Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```
{r}
1 + 1
```

You can add options to executable code like this

```
{r}
#| echo: false
2 * 2
```

Comments

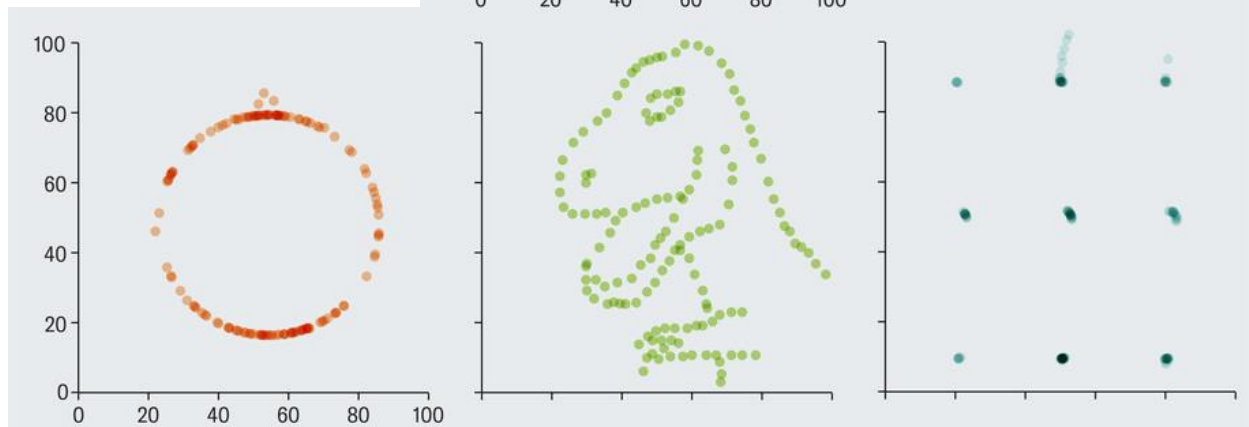
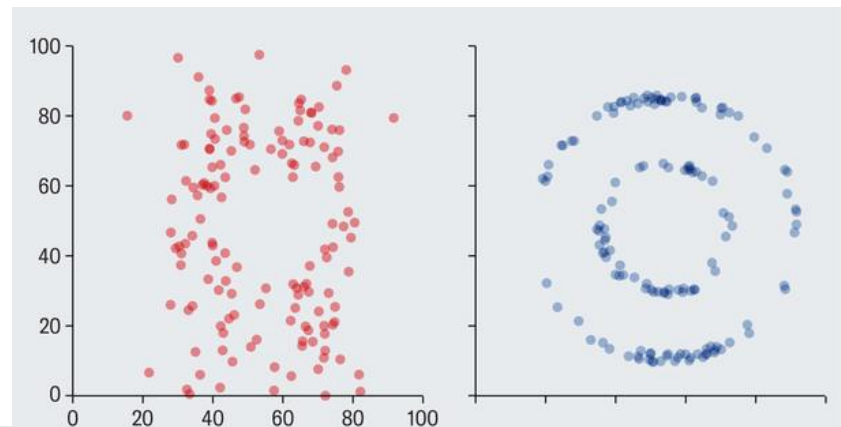
Code chunk

Basic Statistics Topics

1. Importing Data (*Importation des données*)
2. Data Visualization and Exploration (*Exploration et visualisation des données*)
3. Verifying Model Assumptions (Vérification des hypothèses de modèles)
4. Conducting Correlations (*Analyse de correlations*)
5. Comparing Data Between Groups (*Comparaison de données entre groupes*)
 - a. Parametric (*Paramétrique*)
 - b. Non-Parametric (*Non-Paramétrique*)

Which dataset has:

- The highest mean?
- The largest standard deviation?
- The strongest correlation?



Quelle base de données a:

- *Les moyennes la plus haut?*
- *Les étart-types le plus large?*
- *Les coefficients de corrélations le plus fort?*

They are all the same!
Ils sont plus les mêmes!

Summary statistics do
not tell us the whole
story

*Les statistiques sommaires
ne dites pas l'histoire
complet*

All of the following 13 graphs have the same summary statistics:

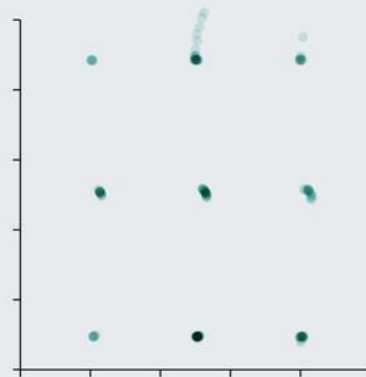
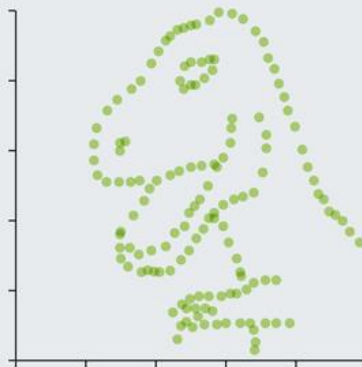
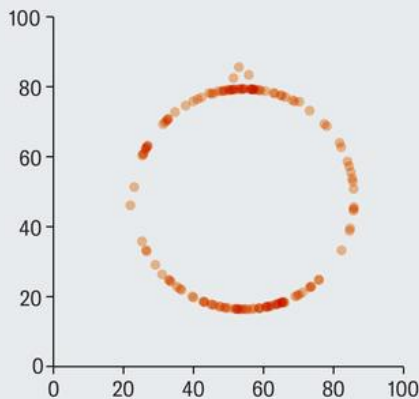
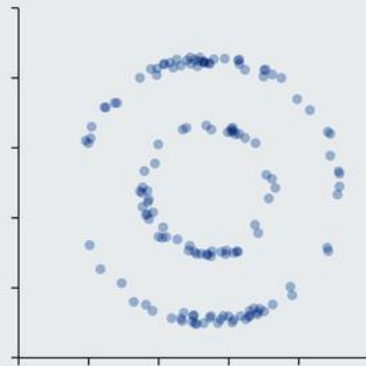
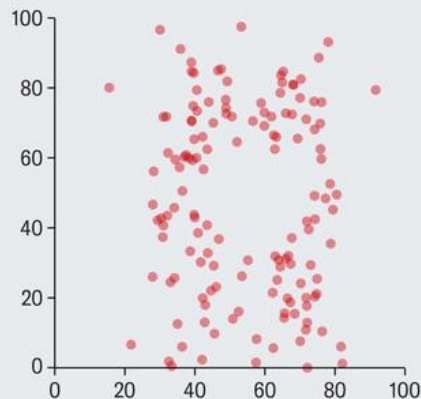
Mean, x-axis = 54.26

Mean, y-axis = 47.83

Standard deviation, x-axis = 16.76

Standard deviation, y-axis = 26.93

Correlation = -0.06



Why do we visualize data first?

- Check for missing data and outliers

Vérifier qu'il n'y a pas des données manquants ou aberrants

- Better understand the distribution of our data

Mieux comprendre la distribution de nos données

- Explore associations and covariance between variables in the dataset

Explorer les liens et covariances entre variables dans la base de données






- Ensure our dataset meets the assumptions of the statistical test we want to perform (e.g. normality)

Confirmer que notre base de données correspond aux hypothèse de l'analyse statistique que nous voulons faire

Some methods for data exploration and visualization

- 'Head' and 'summary'
- The skimr package
- Descriptive statistics (mean, mode, frequency)
- Scatter plots and boxplots between two variables
- Histograms and density plots of variable distributions
- Heatplots and scatterplots to investigate covariance between multiple pairs of variables at once

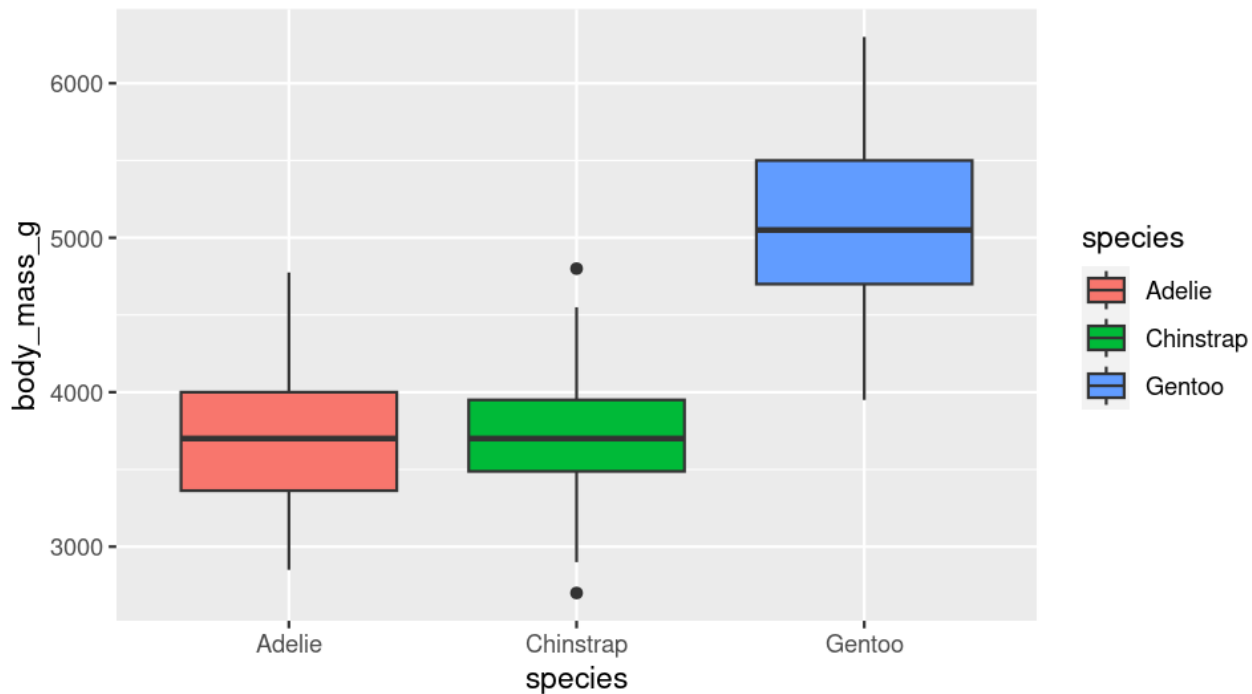
Using the skimr package: skim(data)

Variable type: character											
	skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace			
1	species	0	1	6	9	0	3	0			
2	island	0	1	5	9	0	3	0			
3	sex	0	1	4	6	0	2	0			
Variable type: numeric											
	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
1	bill_length_mm	0	1	44.0	5.47	32.1	39.5	44.5	48.6	59.6	
2	bill_depth_mm	0	1	17.2	1.97	13.1	15.6	17.3	18.7	21.5	
3	flipper_length_mm	0	1	201.	14.0	172	190	197	213	231	
4	body_mass_g	0	1	4207.	805.	2700	3550	4050	4775	6300	
5	year	0	1	2008.	0.813	2007	2007	2008	2009	2009	

Easy way to get a first look at missing data and distributions

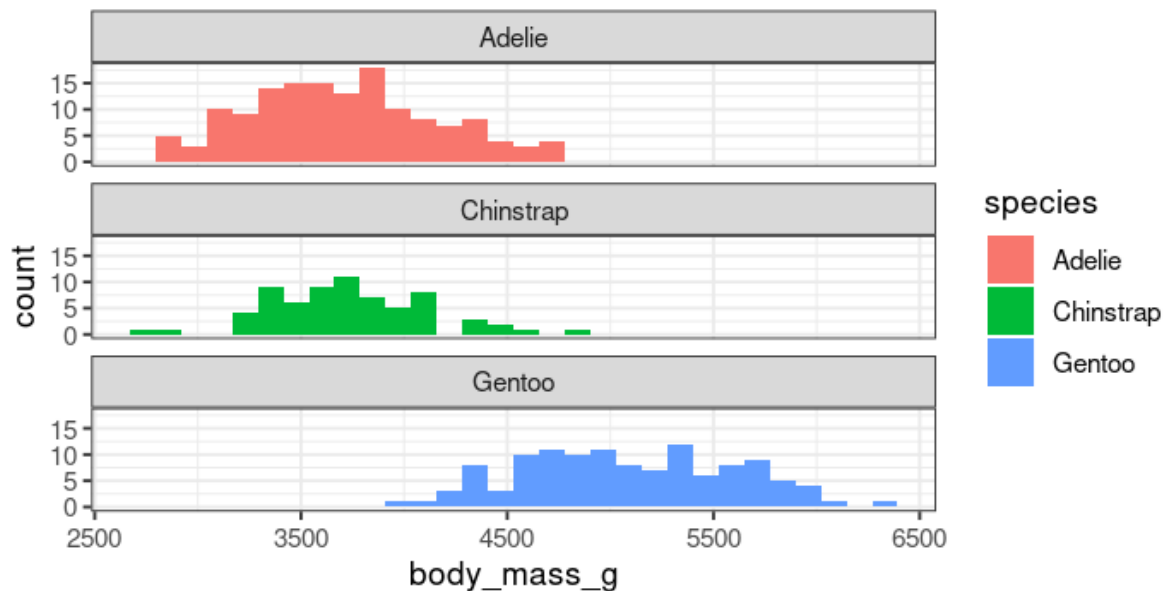
Using boxplots to explore differences between groups

```
ggplot(penguins, aes(x = species, y = body_mass_g, fill = species)) +  
  geom_boxplot()  
#> 
```



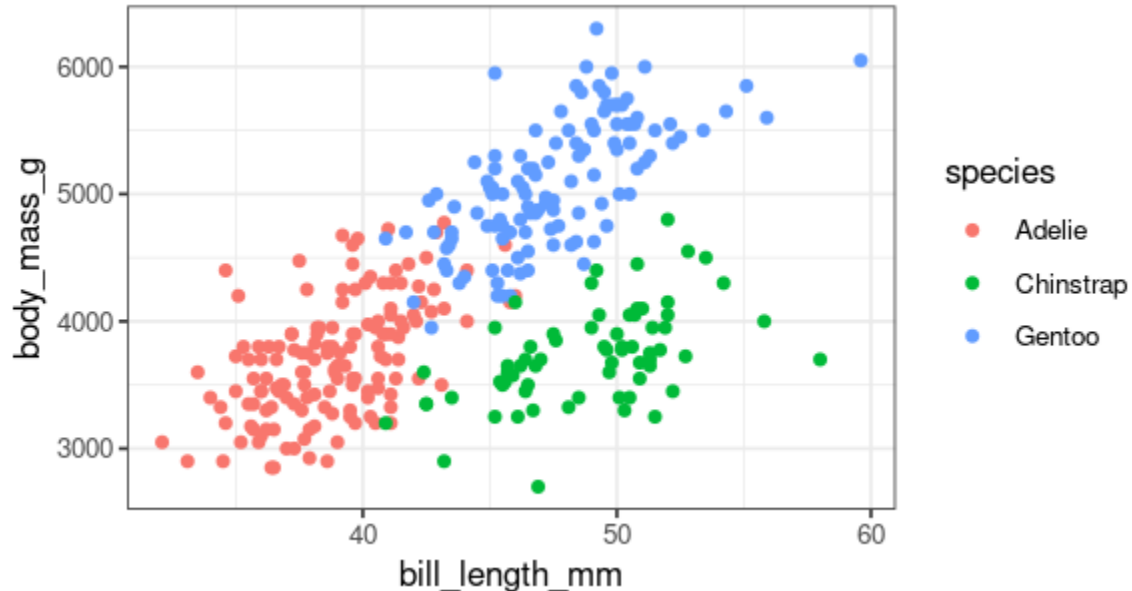
Histograms allow us to explore the distributions of variables

```
```{r}  
ggplot(penguins, aes(x= body_mass_g, fill = species, group = species)) +
 geom_histogram() +
 facet_wrap(~species, nrow = 3)
```
```



Scatterplots are used to explore the relationship between two continuous variables

```
ggplot(penguins, aes(x = bill_length_mm, y = body_mass_g, color = species)) +  
  geom_point()
```



Basic Statistics Topics

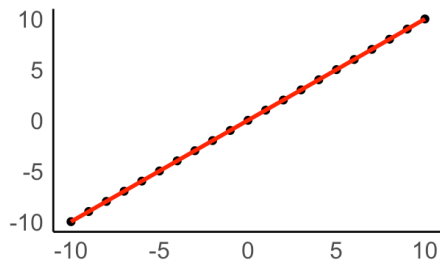
1. Importing Data (*Importation des données*)
2. Data Visualization and Exploration (*Exploration et visualisation des données*)
3. Verifying Model Assumptions (*Vérification des hypothèses de modèles*)
4. Conducting Correlations (*Analyse de correlations*)
5. Comparing Data Between Groups (*Comparaison de données entre groupes*)
 - a. Parametric (*Paramétrique*)
 - b. Non-Parametric (*Non-Paramétrique*)

What assumptions do we need to consider for parametric tests?

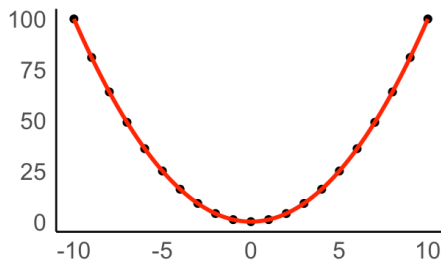
Linearity : The association between two variables is linear

Linéarité: L'association entre deux variables est linéaire

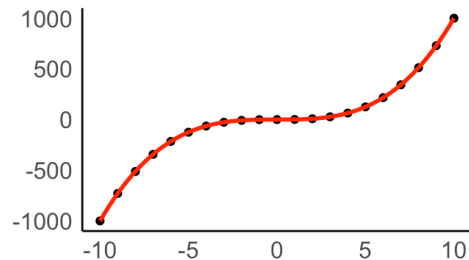
$$y = x$$



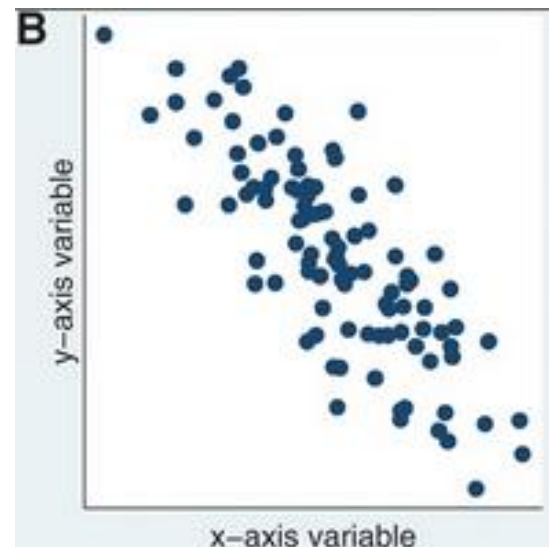
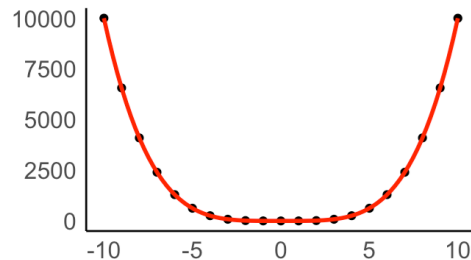
$$y = x^2$$



$$y = x^3$$



$$y = x^4$$

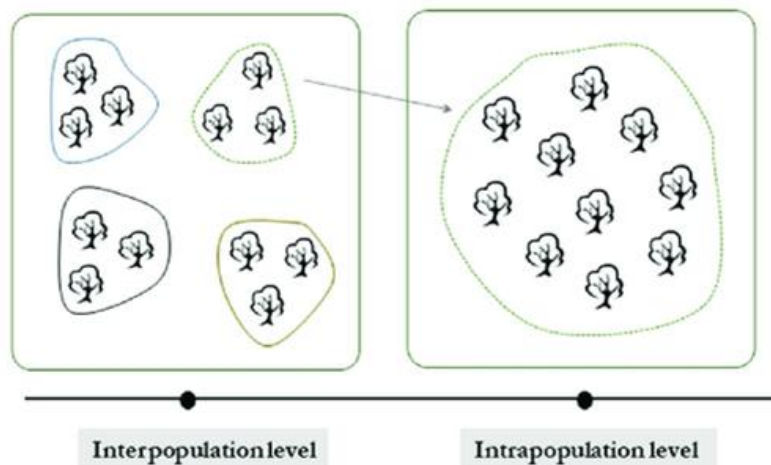


What assumptions do we need to consider for parametric tests?

Linearity : The relation between two variables is linear

Independence : Each observation is independent of all other observations

Chaque observation est indépendante des autres



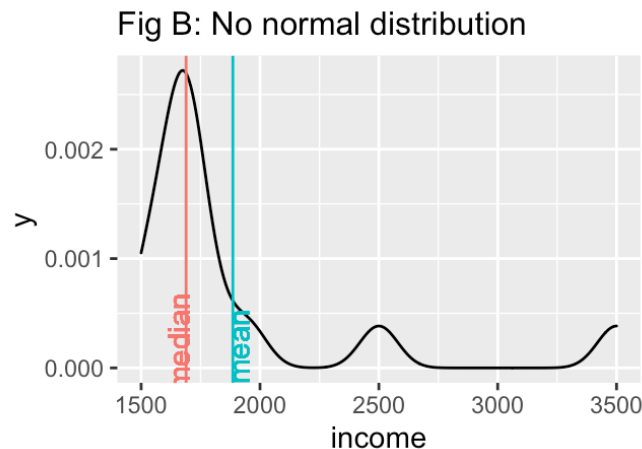
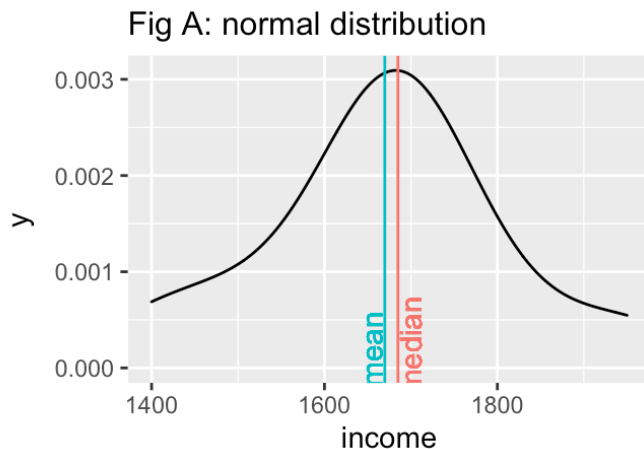
What assumptions do we need to consider for parametric tests?

Linearity : The relation between two variables is linear

Independence : Each observation is independent of all other observations

Normality: The distribution of the data must be normal

La distribution des données doit être normal



What assumptions do we need to consider for parametric tests?

Linearity : The relation between two variables is linear

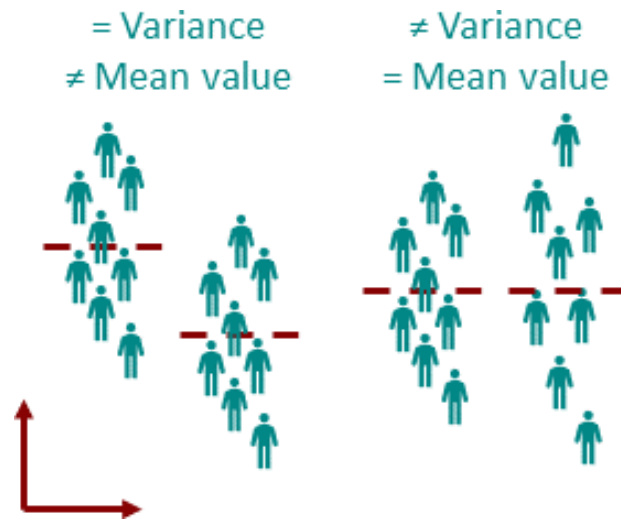
Independence : Each observation is independent of all other observations

Normality: The distribution of the data must be normal

Homogeneity of variance: The variance of subsets

or groups of data should be equal

Les variations des groupes de données doivent être égal



How to test each assumption

| Assumption | Visualization | Statistical Test |
|-------------------|---|--|
| Linearity | Scatterplot | |
| Independent | None, this assumption depends on how the data was collected | |
| Normality | Histogram or density plot | <u>Shapiro-Wilk</u>
shapiro.test(variable) |
| Variance Equality | Boxplot by group | <u>Levene's Test</u>
car::leveneTest(variable ~ group, data = data) |

Basic Statistics Topics

1. Importing Data (*Importation des données*)
2. Data Visualization and Exploration (*Exploration et visualisation des données*)
3. Verifying Model Assumptions (*Vérification des hypothèses de modèles*)
4. Conducting Correlations (*Analyse de correlations*)
5. Comparing Data Between Groups (*Comparaison de données entre groupes*)
 - a. Parametric (*Paramétrique*)
 - b. Non-Parametric (*Non-Paramétrique*)

What are correlations?

Correlations describe the relationship between two variables

Les corrélations décrivent les relations entre deux variables

The variables must be continuous (or at least numeric)

Les variables doivent être continuous (numerique)

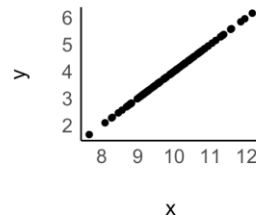
Range between -1 (perfectly negatively correlated) and +1 (perfectly positively correlated)

Les valeurs sont entre -1 (correlation negatif parfait) et +1 (correlation positif parfait)

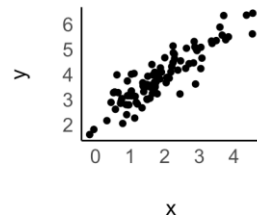
Can be visualized via scatterplots

Peuvent être visualisé avec des scatterplots

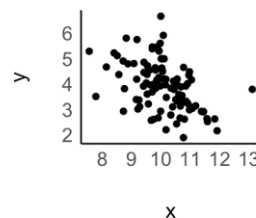
$r = 1$



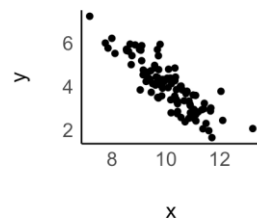
$r = 0.89$



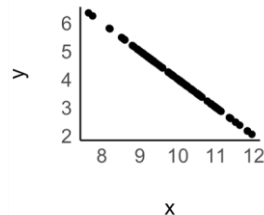
$r = -0.42$



$r = -0.75$



$r = -1$



Two most common types of correlations

Pearson's Correlation

Normally-distributed data

Distribution normal

Linear relationship

Association linéaire

Both variables numeric

Les deux variables doivent être numériques

“Mean” - based

Basé sur la moyenne

Spearman's Correlation

Does not require normally distributed data

N'exige pas les données avec une distribution normal

Correlation is based on rank, not linear relationship

La corrélation est basé sur leur ordre, pas une association linéaire

Both variables numeric

Les deux variables doivent être numériques

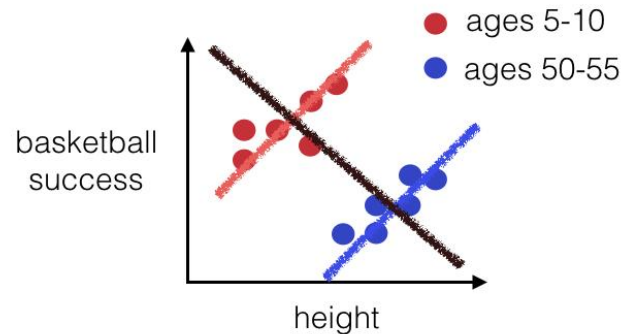
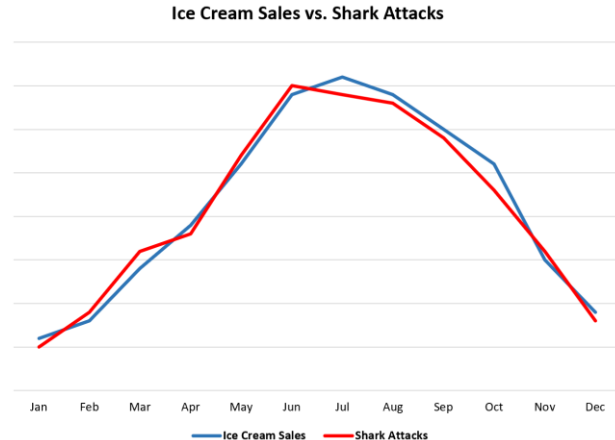
“Median” -based

Basé sur la médiane

Correlations are not causal, they only show associations between variables and may be spurious
Les corrélations ne sont pas causales, elles montrent seulement des associations entre les variables et peuvent être fallacieuses

Correlations may differ depending on what subset of the data they are done on, known as Simpson's Paradox

Les corrélations peuvent différer en fonction du sous-ensemble de données sur lequel elles sont effectuées, ce qui est connu sous le nom de paradoxe de Simpson

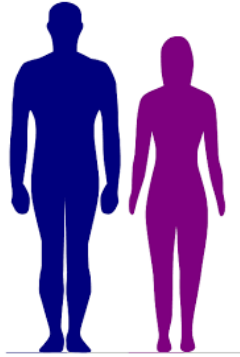


Basic Statistics Topics

1. Importing Data (*Importation des données*)
2. Data Visualization and Exploration (*Exploration et visualisation des données*)
3. Verifying Model Assumptions (*Vérification des hypothèses de modèles*)
4. Conducting Correlations (*Analyse de correlations*)
5. Comparing Data Between Groups (*Comparaison de données entre groupes*)
 - a. Parametric (*Paramétrique*)
 - b. Non-Parametric (*Non-Paramétrique*)

What are some examples of comparisons between groups?

What are some examples of comparisons between groups?

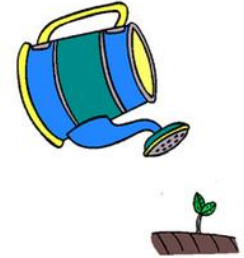


Demographic groups

Experimental group
Bio-fertilizer 'x' is sprayed

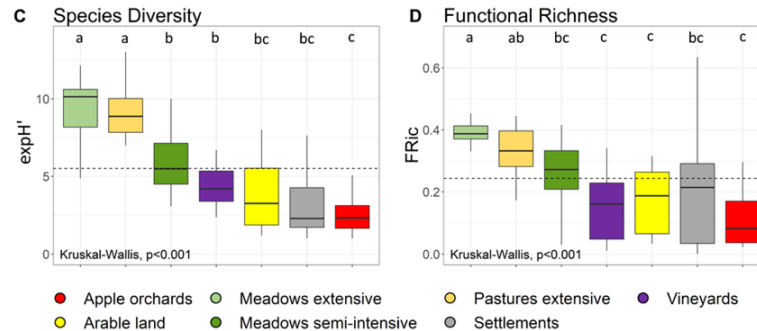


Control group
Bio-fertilizer 'x' is not sprayed

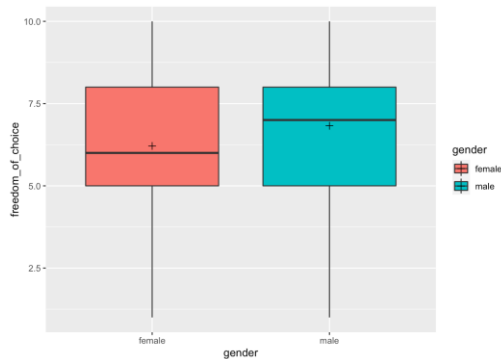


Control vs. treatment

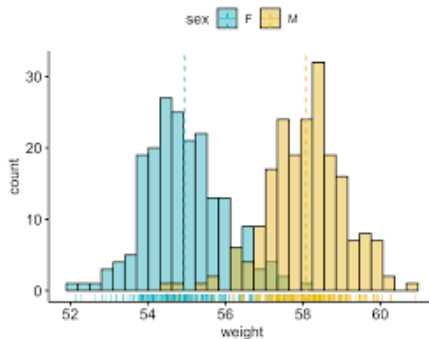
Landcover types



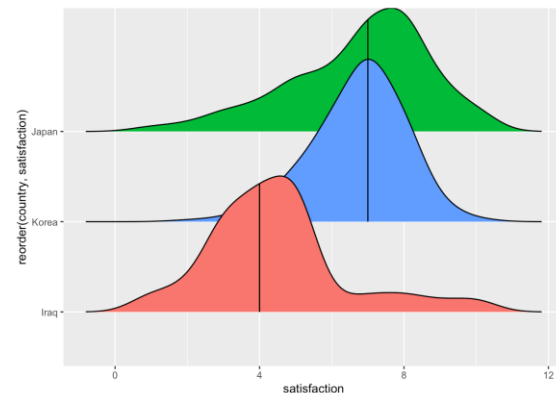
How to visualize comparisons between groups?



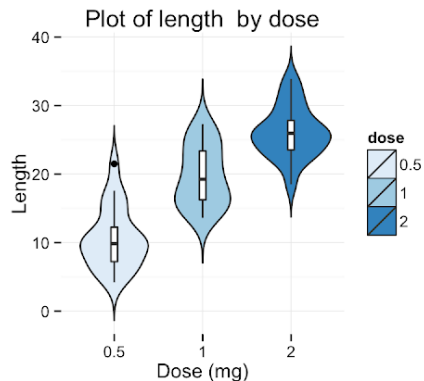
`geom_boxplot()`



`geom_histogram()`



`geom_density()`



`geom_violin()`

Look at packages 'ggdist' and 'ggbeeswarm' for more ways to plot distributions!

Choosing a test

Are the groups big enough to be compared, i.e. are they comparable?

Est-ce que mes groups sont assez large pour être comparé?

Is my data parametric or non-parametric?

Mes donnés ont-elles paramétrique ou non-paramétriques?

How many groups do I wish to compare?

Combien de groupes veux-je comparer?

Y = continuous
X = factor/group
*for unpaired data only

of groups

1

2

3 or more

Parametric

One-sided t-test

Two-sided T-test

Welch's T-test
(if unequal variance)

ANOVA

Non
parametric

Wilcoxon signed-rank

Mann-Whitney U

Kruskal-Wallis

Parametric Data: Use of T-tests

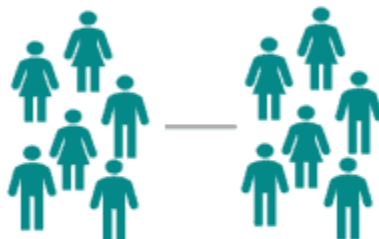
One sample t-Test



Is there a **difference** between a **group** and the **population**

One-sided t-test
(very rare in practice)

Independent samples t-Test



Is there a **difference** between **two groups**

Two-sided t-test
Unpaired t-test

Paired samples t-Test

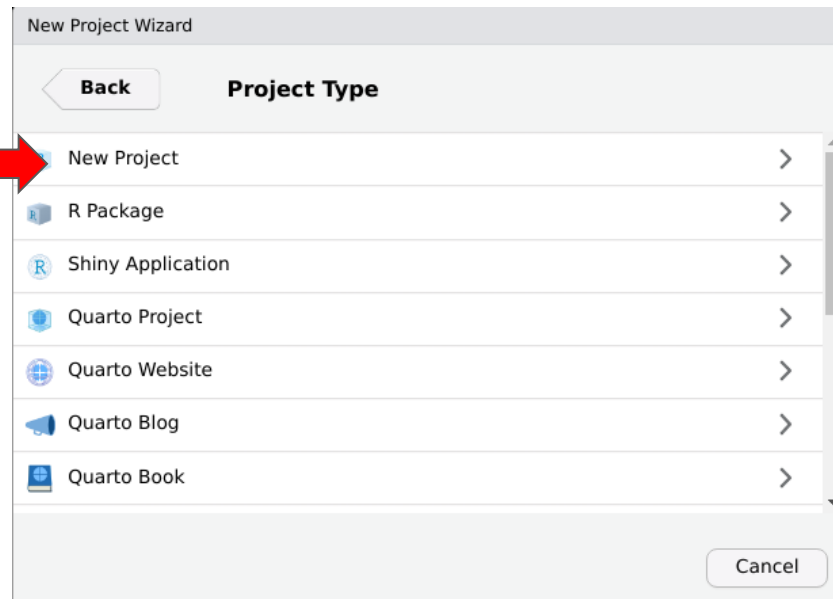
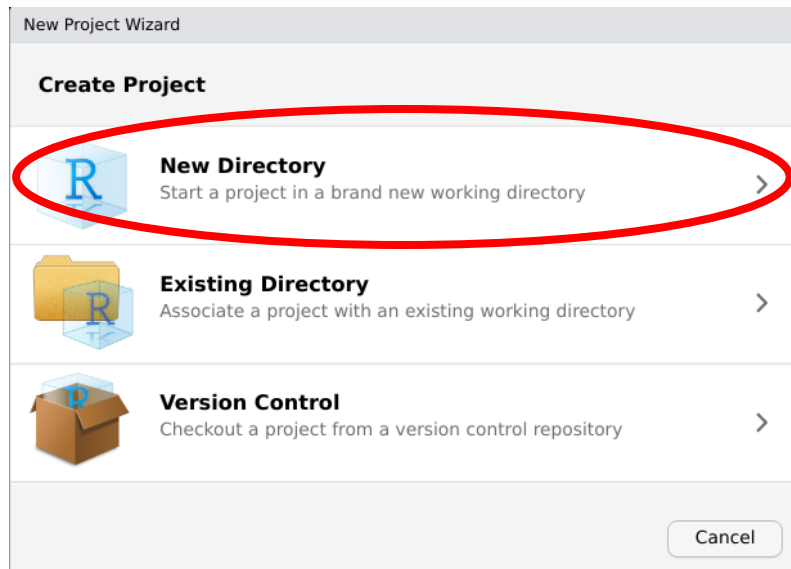


Is there a **difference** in a **group** between **two points in time**

R Practice!

Create a new R Project


File > New Project...



Create a new R Project

New Project Wizard

Back **Create New Project**




Directory name:
E2M2

Create project as subdirectory of:
~ **Browse...**

☐ Create a git repository
☐ Use renv with this project

☐ Open in new session


Create Project **Cancel**



Working directory = RProject directory

New Project Wizard

Back **Create New Project**




Directory name:
E2M2

Create project as subdirectory of:
~/Dropbox/Scholarship/conferences/2024/E2 **Browse...**

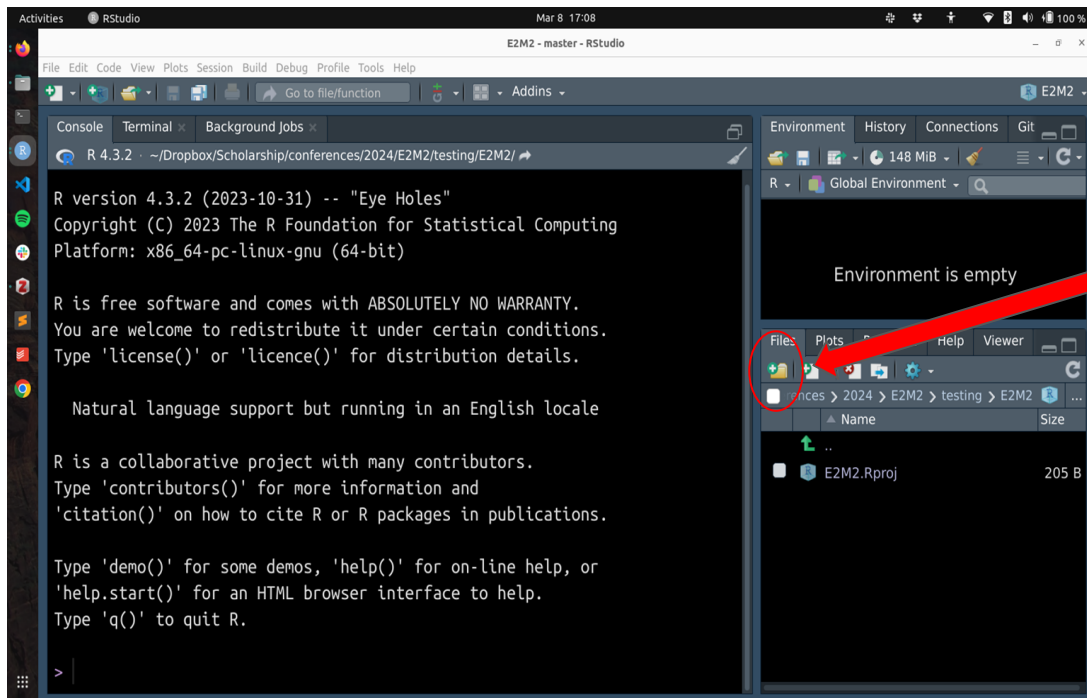
☐ Create a git repository
☐ Use renv with this project

☐ Open in new session

Create Project **Cancel**



Create a folder structure



Create three folders

1. scripts
2. data
3. results

Create a folder structure

1. Scripts: all your .R files go here

Tous les fichiers .R sont ici

1. Data: All of your data goes here. It is best to make two subdirectories: 'raw' and 'clean'

Les données sont ici. Le meilleur pratique est de créer deux sous-dossiers: 'brut' et 'nettoyé'

1. Results: Results of your analysis will go here. This includes tables of summary statistics, figures, and results of statistical tests

Les résultats des analyses sont ici. Cela inclut les tableaux des statistiques sommaires, les figures, et les résultats des analyses

| Name | Size | Modifié |
|-------------|-----------|---------|
| data | 0 items | 17:00 |
| results | 0 items | 17:00 |
| scripts | 0 items | 17:00 |
| .Rproj.user | 2 items | 17:00 |
| E2M2.Rproj | 205 bytes | 17:00 |

Sort your files into the proper folders

Use the File Manager on your computer to move the files for this lesson into the proper folders:

Utiliser le File Manager sur ton Desktop pour placer les fichiers pour cette leçon dans les dossiers correctes

- All .csv files go into data *(tous les fichiers .csv sont les données)*
- All .R, .qmd, or .Rmd files go into scripts *(tous les fichiers .R, .qmd, or .Rmd sont des scripts)*

Do you see those files via the file explorer on RStudio?

Voyez-vous ces fichiers dans le file explorer de RStudio?

Open a quarto document (*ouvrir un document quarto*)

basic-statistics-tutorial.qmd

Source

```
## Quarto

Quarto enables you to weave together content and executable code into a
document. To learn more about Quarto see <https://quarto.org>.

## Running Code

When you click the Render button a document will be generated that includes
both content and the output of embedded code. You can embed code like this:

```{r}
1 + 1
```

You can add options to executable code like this

```{r}
#| echo: false
2 * 2
```
```

The `echo: false` option disables the printing of code (only output is displayed)

Visual

Quarto enables you to weave together content and executable code into a document. To learn more about Quarto see <https://quarto.org>.

Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```
{r}
1 + 1
```

You can add options to executable code like this

```
{r}
#| echo: false
2 * 2
```





Comments

Code chunk

Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: <https://lindeloev.github.io/tests-as-linear>

| Simple regression: $\text{lm}(y \sim 1 + x)$ | Common name | Built-in function in R | Equivalent linear model in R | Exact? | The linear model in words | Icon |
|--|---|--|---|--|---|---|
| | y is independent of x
P: One-sample t-test
N: Wilcoxon signed-rank | t.test(y)
wilcox.test(y) | $\text{lm}(y \sim 1)$
$\text{lm}(\text{signed_rank}(y) \sim 1)$ | ✓
for N >14 | One number (intercept, i.e., the mean) predicts y .
- (Same, but it predicts the <i>signed rank</i> of y .) |  |
| | P: Paired-sample t-test
N: Wilcoxon matched pairs | t.test(y1, y2, paired=TRUE)
wilcox.test(y1, y2, paired=TRUE) | $\text{lm}(y_2 - y_1 \sim 1)$
$\text{lm}(\text{signed_rank}(y_2 - y_1) \sim 1)$ | ✓
for N >14 | One intercept predicts the pairwise y₂-y₁ differences.
- (Same, but it predicts the <i>signed rank</i> of y₂-y₁ .) |  |
| | y ~ continuous x
P: Pearson correlation
N: Spearman correlation | cor.test(x, y, method='Pearson')
cor.test(x, y, method='Spearman') | $\text{lm}(y \sim 1 + x)$
$\text{lm}(\text{rank}(y) \sim 1 + \text{rank}(x))$ | ✓
for N >10 | One intercept plus x multiplied by a number (slope) predicts y .
- (Same, but with <i>ranked x</i> and y) |  |
| | y ~ discrete x
P: Two-sample t-test
P: Welch's t-test
N: Mann-Whitney U | t.test(y1, y2, var.equal=TRUE)
t.test(y1, y2, var.equal=FALSE)
wilcox.test(y1, y2) | $\text{lm}(y \sim 1 + G_2)^A$
$\text{gls}(y \sim 1 + G_2, \text{weights}=\dots^B)^A$
$\text{lm}(\text{signed_rank}(y) \sim 1 + G_2)^A$ | ✓
✓
for N >11 | An intercept for group 1 (plus a difference if group 2) predicts y .
- (Same, but with one variance <i>per group</i> instead of one common.)
- (Same, but it predicts the <i>signed rank</i> of y .) |  |

All the tests we did today can also be thought of as linear regression models, which we will learn about throughout the week.

Tous les analyses que nous avons fait aujourd'hui sont aussi les modèles linéaires, sur lesquels nous allons apprendre pendant cette semaine

Extra Resources

R for non-programmers: https://bookdown.org/daniel_dauber_io/r4np_book/

Basic Statistics as Linear Models: <https://lindeloev.github.io/tests-as-linear/>

Collection of easystats packages: <https://easystats.github.io/easystats/>

Extra slides!

Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: <https://lindeloef.github.io/tests-as-linear>

| | Common name | Built-in function in R | Equivalent linear model in R | Exact? | The linear model in words | Icon |
|--|---|--|---|------------------------|---|-----------------------|
| Simple regression: $\text{lm}(y \sim 1 + x)$ | y is independent of x
P: One-sample t-test
N: Wilcoxon signed-rank | t.test(y)
wilcox.test(y) | $\text{lm}(y \sim 1)$
$\text{lm}(\text{signed_rank}(y) \sim 1)$ | ✓
for $N > 14$ | One number (intercept, i.e., the mean) predicts y .
- (Same, but it predicts the <i>signed rank</i> of y .) | |
| | P: Paired-sample t-test
N: Wilcoxon matched pairs | t.test(y1, y2, paired=TRUE)
wilcox.test(y1, y2, paired=TRUE) | $\text{lm}(y_2 - y_1 \sim 1)$
$\text{lm}(\text{signed_rank}(y_2 - y_1) \sim 1)$ | ✓
for $N > 14$ | One intercept predicts the pairwise y ₂ - y ₁ differences.
- (Same, but it predicts the <i>signed rank</i> of y ₂ - y ₁ .) | |
| | y ~ continuous x
P: Pearson correlation
N: Spearman correlation | cor.test(x, y, method='Pearson')
cor.test(x, y, method='Spearman') | $\text{lm}(y \sim 1 + x)$
$\text{lm}(\text{rank}(y) \sim 1 + \text{rank}(x))$ | ✓
for $N > 10$ | One intercept plus x multiplied by a number (slope) predicts y .
- (Same, but with <i>ranked x</i> and y) | |
| | y ~ discrete x
P: Two-sample t-test
P: Welch's t-test
N: Mann-Whitney U | t.test(y1, y2, var.equal=TRUE)
t.test(y1, y2, var.equal=FALSE)
wilcox.test(y1, y2) | $\text{lm}(y \sim 1 + G_1)^4$
$\text{gls}(y \sim 1 + G_2, \text{weights} = \dots)^4$
$\text{lm}(\text{signed_rank}(y) \sim 1 + G_2)^4$ | ✓
✓
for $N > 11$ | An intercept for group 1 (plus a difference if group 2) predicts y .
- (Same, but with one variance <i>per group</i> instead of one common.)
- (Same, but it predicts the <i>signed rank</i> of y .) | |
| Multiple regression: $\text{lm}(y \sim 1 + x_1 + x_2 + \dots)$ | P: One-way ANOVA
N: Kruskal-Wallis | aov(y ~ group)
kruskal.test(y ~ group) | $\text{lm}(y \sim 1 + G_1 + G_2 + \dots + G_N)^4$
$\text{lm}(\text{rank}(y) \sim 1 + G_1 + G_2 + \dots + G_N)^4$ | ✓
for $N > 11$ | An intercept for group 1 (plus a difference if group $\neq 1$) predicts y .
- (Same, but it predicts the <i>rank</i> of y .) | |
| | P: One-way ANCOVA | aov(y ~ group + x) | $\text{lm}(y \sim 1 + G_1 + G_2 + \dots + G_N + x)^4$ | ✓ | - (Same, but plus a slope on x .)
<i>Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.</i> | |
| | P: Two-way ANOVA | aov(y ~ group * sex) | $\text{lm}(y \sim 1 + G_1 + G_2 + \dots + G_N + S_2 + S_3 + \dots + S_K + G_2 * S_2 + G_3 * S_3 + \dots + G_N * S_K)^4$ | ✓ | Interaction term: changing sex changes the y ~ group parameters.
<i>Note: $G_{2:N,K}$ is an indicator (0 or 1) for each non-intercept levels of the group variable. Similarly for $S_{2:N,K}$ for sex. The first line (with G_1) is main effect of group, the second (with S_2) for sex and the third is the group * sex interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.</i> | [Coming] |
| | Counts ~ discrete x
N: Chi-square test | chisq.test(groupXsex_table) | Equivalent log-linear model
$\text{glm}(y \sim 1 + G_1 + G_2 + G_3 + \dots + G_N + S_2 + S_3 + \dots + S_K + G_2 * S_2 + G_3 * S_3 + \dots + G_N * S_K, \text{family} = \dots)^4$ | ✓ | Interaction term: (Same as Two-way ANOVA.)
<i>Note: Run glm using the following arguments: <code>glm(model, family=poisson())</code> As linear-model, the Chi-square test is $\log(y) = \log(N) + \log(\alpha) + \log(\beta) + \log(\alpha\beta)$ where α_i and β_j are proportions. See more info in the accompanying notebook.</i> | Same as Two-way ANOVA |
| | N: Goodness of fit | chisq.test(y) | $\text{glm}(y \sim 1 + G_1 + G_2 + \dots + G_N, \text{family} = \dots)^4$ | ✓ | (Same as One-way ANOVA and see Chi-Square note.) | 1W-ANOVA |

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation $y \sim 1 + x$ is R shorthand for $y = 1 \cdot b + a \cdot x$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. `lm` requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at <https://lindeloef.github.io/tests-as-linear>.

^A See the note to the two-way ANOVA for explanation of the notation.

^B Same model, but with one variance per group: `gls(value ~ 1 + G2, weights = varIdent(form = ~1|group), method="ML")`.



Parametric Data: Use of ANOVAs

t-test
for independent samples

Group 1



Group 2



One-factor ANOVA
without repeated measures

Group 1



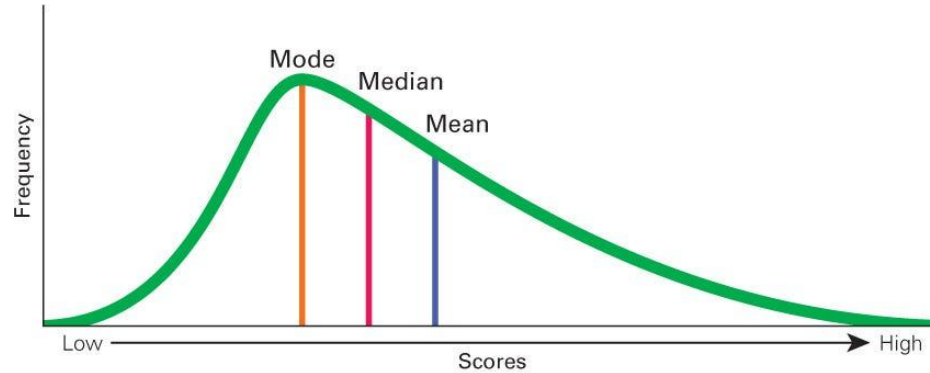
Group 2



Group 3



Difference between right-skewed and left-skewed distributions



(a) Right-skewed distribution



(b) Left-skewed distribution