

Introduction to R

Lecture 2

September 21st, 2020

Dr. Kristin Eccles

kristin.eccles@utoronto.ca

 @kristineccles

https://github.com/kristineccles/intro_to_r_2020

Overview

—Crash course in statistics

- Probability distribution and the p-value
- Descriptive statistics
- Parametric Tests
 - Difference of means: T-Test and ANOVA
 - Relational Statistics: correlations, linear regression
 - PCA

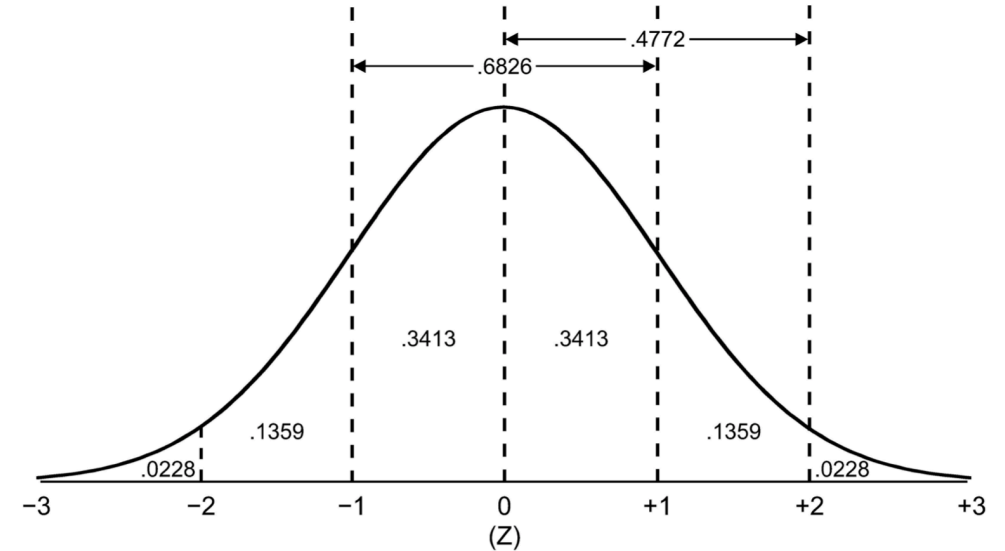
Steps of Statistical Analysis

| | |
|--------|---|
| Step 1 | State the null and alternative hypothesis |
| Step 2 | Select the appropriate statistical test |
| Step 3 | Select level of significance |
| Step 4 | Delineate regions of rejection and non-rejection of the null (α) |
| Step 5 | Calculate test statistic |
| Step 6 | Make decision regarding null and alternative hypothesis |

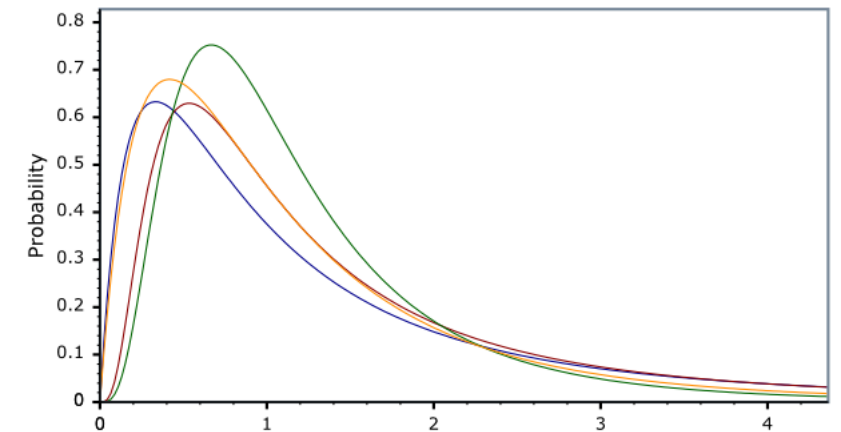
Statistics

- **Descriptive** statistics uses the data to provide descriptions of the population, either through numerical calculations, graphs, or tables.
- **Inferential** statistics makes inferences **and** predictions about a population based on a sample of data taken from the population in question.
 - Based on probability distributions
 - Probability theory deals with random events
 - Random/probabilistic independent sample(s)

T/z-distribution



F-distribution



Probability Distributions

- The area under the curve gives the probability of finding a given value in that range
- Total area under the curve is = 1 (probability is 1)
- The normal distribution
 - What the probability is of getting an extreme value (the tail)
 - 68% of the values are within 1 s.d. of the mean
 - Or, there is a probability of 0.68 that a given value is within 1 s.d. of the mean
 - **95% of the area under the curve is within 1.96 s.d. of the mean**
 - 99.7% of the area is within 3 standard deviations of the mean
- Empirical probability relies on the **law of large numbers**: the relative frequency will eventually converge on the true probability

P-value

- Evaluates the validity of our prediction using a statistical test
 - This will say if the result is statistically significant or not
- Two hypotheses:
 - H_0 – the null hypothesis
 - Status quo, no change, nothing special going on, conservative
 - H_A (aka H_1) – the alternate hypothesis
 - Your hypothesis – there is a difference/change.
- They are mutually exclusive and exhaustive

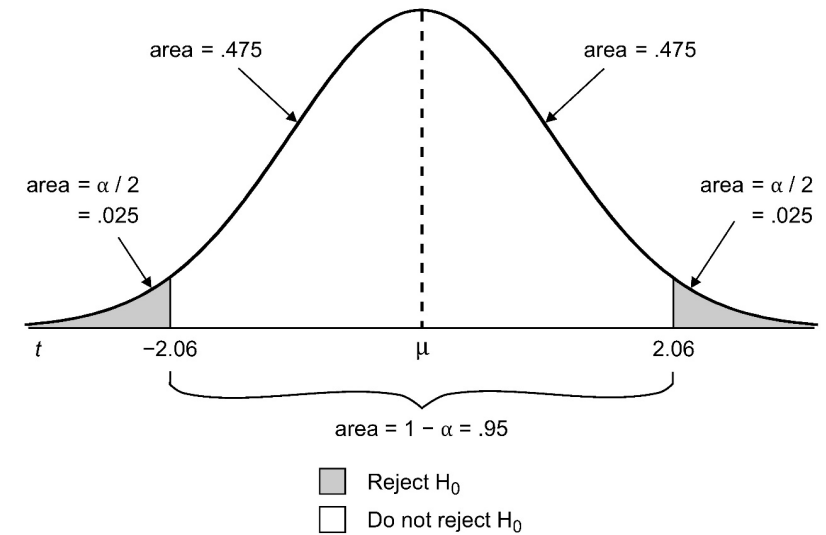
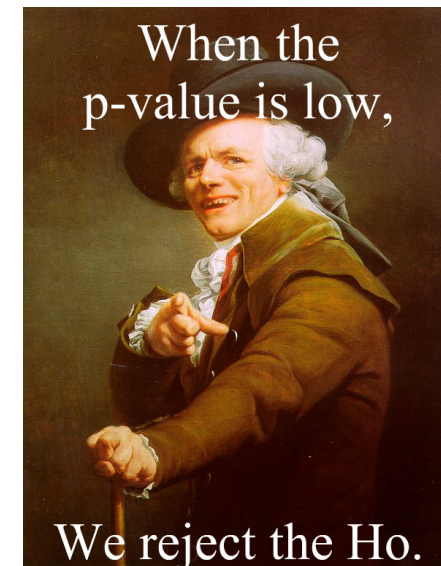


FIGURE 9.2

Normal Distribution Values Associated with a Significance Level (α) = .05: Two-Tailed Case



Descriptive Statistics

- Characteristics of a normal (Gaussian) distribution:
 - Mean, median, mode are all the same & in the middle of the distribution
- Symmetrical, smooth, continuous distribution
- Is my data normal?
 - Make a histogram
 - Look at mean, median, and modal values
 - Check the shape of the distribution:
 - Skewness and Kurtosis

summary(df)

- no st.dev in summary
- sd(df)

library(psych)

- **describe(df)**

Unpaired T-Test

- Objective: Test if there is a difference between the **mean** of two groups
 - Boxplots show the median
- Data type: one continuous and one categorical variable
- Assumptions: variable normality, equal variance between the groups
 - For unequal variance use Welch's T-Test

ANOVA

- Test: If there is a difference between the **mean** of 3 or more groups
 - This only tells you if there is a difference NOT where the difference is
- Data: Three or more groups and a continuous variable
- Assumptions (F-test): variable normality, equal variance between the groups (ish)

ANOVA Post-hoc test (Tukey)

- Test: Multiple comparison to determine the differences in means after an ANOVA
- Data: Three or more groups and a continuous variable
- Assumptions (t-test): variable normality, equal variance between the groups
- Note the p adj: This p-value is adjusted for multiple comparison
 - By default R uses Bonferroni (α / n), where n = number of comparisons

Correlation (Pearson Product-Moment)

- Test: Association between variables
 - correlation coefficient ρ (Greek letter rho)
- Data: Two continuous variables
 - Pearson product-moment correlation
 - Spearman rank and Kendall Tau are the non-parametric version
 - Contingency analysis (categorical data)
- Assumptions (t-test): variable normality, linear association between variables

Linear Regression (Univariate)

- Test: the slope coefficient is significant
- Data: two continuous variables
 - Generalized linear model is used for non-parametric (e.g. logit)
- Assumptions (t -test): , Test statistic: t , Assumptions: Variable normality , Random/probabilistic sample of paired variables, Variables have a linear association
- Test: Goodness of fit (R^2)
 - test that the model predicts a significant amount of the variance in y
 - Coefficient of determination (not the same as p)
- Assumptions (F -test): Variable normality ,Random/probabilistic sample of paired variables, Variables have a linear association

Assumptions of the residuals

- $e = y - \hat{Y}$
- In linear regression we have assumptions on the residuals too
 - Normality (Shapiro-Wilk)
 - Linearity (RESET test)
 - Lack of serial autocorrelation (Durbin-Watson test)
 - Homogeneity (Breusch-Pagan test)
- Can use individual test or `plot(lm1)`

Readings a Regression Output

Call:

```
lm(formula = y ~ x, data = dataset)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|-------|--------|
| -212.04 | -62.52 | 30.87 | 86.77 | 121.88 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|------------|
| (Intercept) | 28.149 | 180.756 | 0.156 | 0.88010 |
| x | 39.057 | 9.172 | 4.258 | 0.00277 ** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 121.8 on 8 deg. of freedom

Multiple R-squared: 0.69, Adjusted R-squared: 0.65

F-statistic: 18.13 on 1 and 8 DF, p-value: 0.002767

- What you asked for:
 - the formula here says predict 'y' from 'x' with data from 'dataset'
- Residuals:
 - 5 number summary of the residuals

Reading a Regression Output

Call:

```
lm(formula = y ~ x, data = dataset)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|-------|--------|
| -212.04 | -62.52 | 30.87 | 86.77 | 121.88 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|------------|
| (Intercept) | 28.149 | 180.756 | 0.156 | 0.88010 |
| x | 39.057 | 9.172 | 4.258 | 0.00277 ** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 121.8 on 8 deg. of freedom

Multiple R-squared: 0.69, Adjusted R-squared: 0.65

F-statistic: 18.13 on 1 and 8 DF, p-value: 0.002767

Model coefficients (aka parameters)

- cols are the coefficient **estimates**, the **standard error** of the coefficient, the ***t* value** for the coefficient & the ***p*-value** for this *t* value
- y-intercept (intercept)
- slope (Estimate)
- legend explaining level of significance codes ($\alpha=0.05$)

Reading a Regression Output

Call:

```
lm(formula = y ~ x, data = dataset)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|-------|--------|
| -212.04 | -62.52 | 30.87 | 86.77 | 121.88 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|------------|
| (Intercept) | 28.149 | 180.756 | 0.156 | 0.88010 |
| x | 39.057 | 9.172 | 4.258 | 0.00277 ** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 121.8 on 8 deg. of freedom
Multiple R-squared: 0.69, Adjusted R-squared: 0.65
F-statistic: 18.13 on 1 and 8 DF, p-value: 0.002767

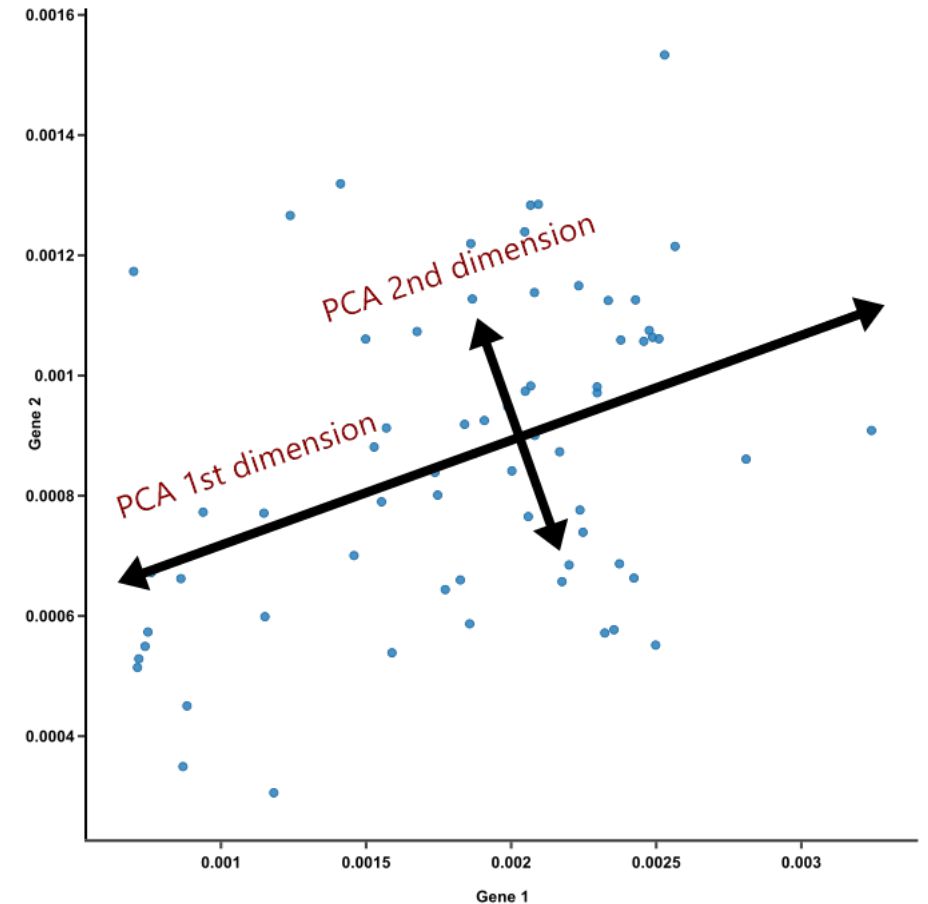
- Overall model performance
- Standard error of the residuals (SSE)
- R^2 and the adjusted R^2
- Use the adj. R^2 since it is more conservative
- F statistic and the p-value (prob. of getting a bigger F)

Multivariate

- $\text{lm}(y \sim v1 + v2 + v3\dots)$
- Assumptions and test the same as univariate linear regression
- One additional
 - Multicollinearity
 - Cannot have a high correlation between independent variables
 - $> +/- 0.70$ it typically the threshold
 - Correlation matrices are helpful
 - Can test this using a variance inflation factor (VIF)
 - >5

Principal Component Analysis (PCA)

- Test:
 - Clustering
 - A way to summarize complex real-valued data with a single categorical variable
 - Dimensionality reduction
 - A way to simplify complex high-dimensional data
- Data: ratio of 5 cases: 1 variable, variables must be of similar magnitude, and must not have outliers
 - Scale and center
- In R **prcomp** and **princomp** are used



PCA Outputs

- Variance explained

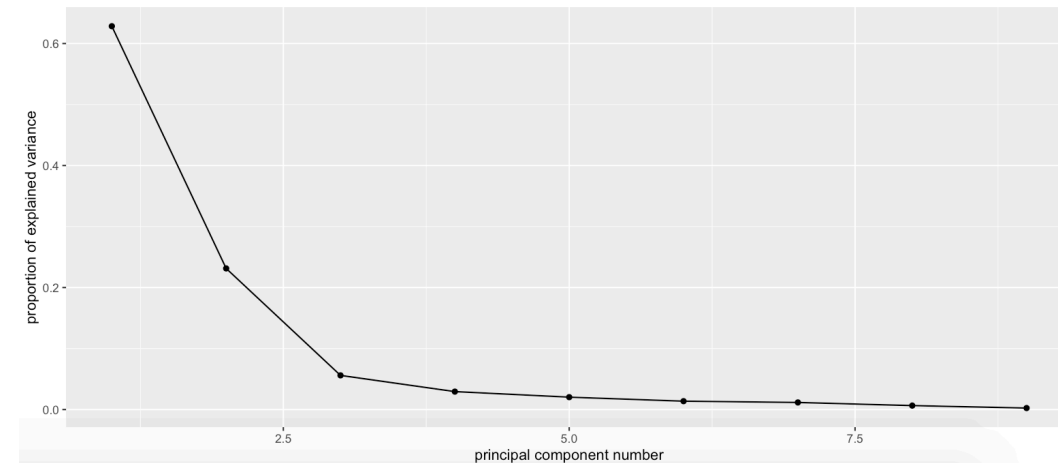
- Summary(pca1)
- # of variables = # of PCs

Importance of components:

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 |
|------------------------|--------|--------|---------|---------|---------|---------|---------|--------|
| Standard deviation | 2.3782 | 1.4429 | 0.71008 | 0.51481 | 0.42797 | 0.35184 | 0.32413 | 0.2419 |
| Proportion of Variance | 0.6284 | 0.2313 | 0.05602 | 0.02945 | 0.02035 | 0.01375 | 0.01167 | 0.0065 |
| Cumulative Proportion | 0.6284 | 0.8598 | 0.91581 | 0.94525 | 0.96560 | 0.97936 | 0.99103 | 0.9975 |

- Scree plot

- Displays how much variation each principal component captures from the data
- Used to determine the number of factors to retain components in an analysis



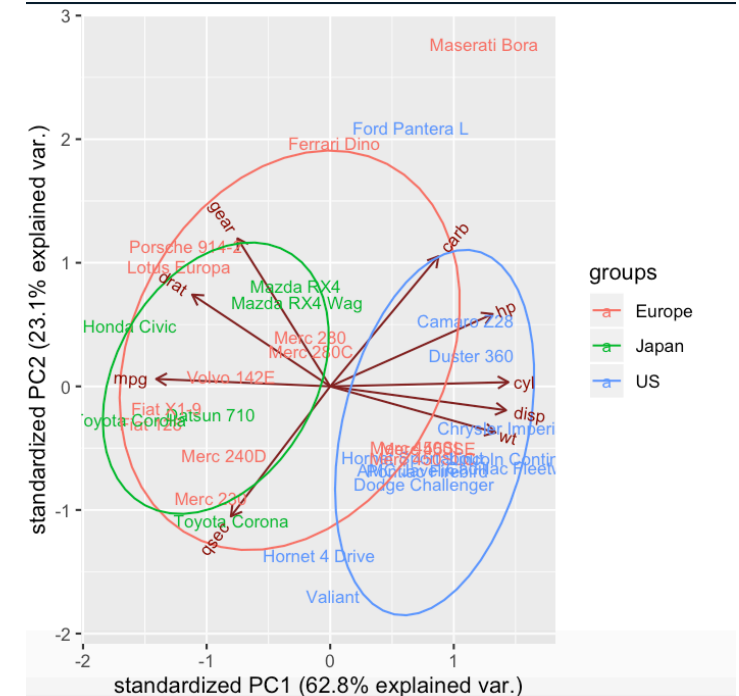
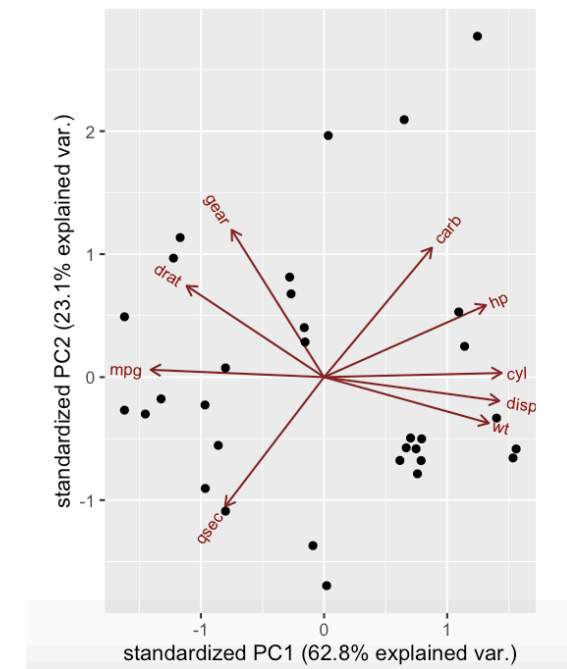
PCA Outputs

■ Biplot

- PCA biplot = PCA score plot (points) + loading plot (arrows)
- Biplots are a type of exploratory **graph** used in statistics, a generalization of the simple two-variable scatterplot.
- A loading plot shows how strongly each characteristic influences a principal component.
- Shows how samples cluster based on their similarity

■ Eigenvalues (scores)

■ Eigenvectors (loadings)



- Loadings (eigenvectors)

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|----------------|-----------|-------------|------------|--------------|--------------|--------------|
| length | 0.3832508 | 0.03786529 | -0.5932799 | -0.089331673 | 0.040512600 | 0.699651086 |
| diameter | 0.3835732 | 0.06532324 | -0.5853661 | -0.008285814 | 0.008517628 | -0.711025627 |
| height | 0.3481438 | 0.86683603 | 0.3148764 | -0.165564868 | -0.027110424 | 0.009841283 |
| whole_weight | 0.3906735 | -0.23327117 | 0.2308252 | 0.052280164 | -0.110183954 | -0.021653298 |
| shucked_weight | 0.3781883 | -0.34801069 | 0.2315678 | -0.496179039 | -0.545339050 | -0.011030516 |
| gut_weight | 0.3815134 | -0.25290295 | 0.2702527 | -0.140972073 | 0.809328460 | -0.023996063 |
| shell_weight | 0.3789217 | -0.05837478 | 0.1621047 | 0.834110000 | -0.181668556 | 0.060561675 |

- Scores (eigenvalues)
- Scores can be mapped (to look for spatial clustering) and can be used in linear regression
 - Can be difficult to interpret

What is reproducible analysis?

- One particular form:
 - code transforms raw data and meta-data into processed data,
 - code runs analyses on the data, and
 - code incorporates analyses into a report
- Ideally, the process involves a one-click build
 - Knitr: an engine for dynamic report generation with R.
- Public sharing of document, code, and data is optional, but forms part of gold standard of scientific openness
 - Journals are now requiring this

intro_to_r_lecture2_excercise.R

kristineccles

2020-01-12

```
#####  
# Introduction to R  
# Lecture 2- Statistics  
# By: Kristin Eccles  
# Written in R 3.6.2  
#####  
  
# Install Libraries  
# only need to run this once  
#install.packages(c("psych", "car", "stats", "corrplot", "factoextra", "lmtest", "devtools"))  
  
# Load Libraries  
library(ggplot2)  
library(psych) # describe and mutli.hist  
  
##  
## Attaching package: 'psych'  
  
## The following objects are masked from 'package:ggplot2':  
##  
##    %+%, alpha  
library(car) #stats  
  
## Loading required package: carData  
  
##  
## Attaching package: 'car'  
  
## The following object is masked from 'package:psych':  
##
```


Reproducible analysis in R

- Combine R and plain text file format to produce documents (e.g., pdfs, HTML documents, etc.)

Rmarkdown

- General workflow
 - Create an RMarkdown File
 - either use Rstudio File – New File – RMarkdown or just create a file with an empty text file with the .rmd extension
 - Optionally add a header
 - Options to specify output format (see RMarkdown cheat sheet)
 - Write the main document in Markdown
 - Embed R code chunks
 - R code chunks can be customised to control output

3. Markdown Next, write your report in plain text. Use markdown syntax to describe how to format text in the final report.

| syntax | becomes |
|---|--|
| Plain text End a line with two spaces to start a new paragraph. *italics* and _italics_ **bold** and __bold__ superscript^2^ --strikethrough-- [link] (www.rstudio.com) | Plain text End a line with two spaces to start a new paragraph. <i>italics</i> and <i>italics</i> bold and bold superscript ² strikethrough link |
| # Header 1 | Header 1 |
| ## Header 2 | Header 2 |
| ### Header 3 | Header 3 |
| #### Header 4 | Header 4 |
| ##### Header 5 | Header 5 |
| ##### Header 6 | Header 6 |
| endash: -- emdash: --- ellipsis: ... inline equation: $\$A = \pi r^2\$$ image: | endash: – emdash: — ellipsis: … inline equation: $\Lambda = \pi * r^2$ image:  |
| horizontal rule (or slide break): *** | horizontal rule (or slide break): <hr/> |
| > block quote * unordered list * item 2 + sub-item 1 + sub-item 2 1. ordered list 2. item 2 + sub-item 1 + sub-item 2 | > block quote • unordered list • item 2 ◦ sub-item 1 ◦ sub-item 2 1. ordered list 2. item 2 ◦ sub-item 1 ◦ sub-item 2 |
| Table Header Second Header ----- Table Cell Cell 2 Cell 3 Cell 4 | Table Header Second Header ----- Table Cell Cell 2 Cell 3 Cell 4 |

See RMarkdown Cheat Sheet for more info

Reproducible Science is good. Replicated Science is better.

ReScience C is an open-access peer-reviewed journal that targets computational research and encourages the explicit [replication](#) of already published research, promoting new and open-source implementations in order to ensure that the original research is [reproducible](#).

To achieve this goal, the whole publishing chain is radically different from other traditional scientific journals. ReScience C lives on [GitHub](#) where each new implementation of a computational study is made available together with comments, explanations and tests. Each submission takes the form of an issue that is publicly reviewed and tested in order to guarantee that any researcher can re-use it. If you ever replicated computational results (or failed at) from the literature in your research, ReScience C is the perfect place to publish your new implementation.

ReScience C is collaborative and open by design. Everything can be forked and modified. Don't hesitate to [write a submission](#), [join us](#) and to [become a reviewer](#).