# Analysis of Quantitative data
# Introduction

Anne Segonds-Pichon
v2020-12

# Outline of this section

- Assumptions for parametric data
- Comparing two means: **Student's *t*-test**
- Comparing more than 2 means
  - One factor: **One-way ANOVA**
  - Two factors: **Two-way ANOVA**
- Relationship between 2 continuous variables: **Correlation**

# Introduction

- **Key concepts to always keep in mind**
  - Null hypothesis and error types
  - Statistics inference
  - Signal-to-noise ratio

# The null hypothesis and the error types

- The null hypothesis ($H_0$): $H_0$ = no effect
  - e.g. no difference between 2 genotypes
- The aim of a statistical test is to reject or not $H_0$.

| Statistical decision | True state of $H_0$ | |
|---|---|---|
| | **$H_0$ True (no effect)** | **$H_0$ False (effect)** |
| **Reject $H_0$** | **Type I error α**<br>**False Positive** | Correct<br>**True Positive** |
| **Do not reject $H_0$** | Correct<br>**True Negative** | **Type II error β**<br>**False Negative** |

- Traditionally, a test or a difference is said to be "**significant**" if the probability of type I error is: **α =< 0.05**
- **High specificity** = low **False Positives** = low **Type I error**
- **High sensitivity** = low **False Negatives** = low **Type II error**

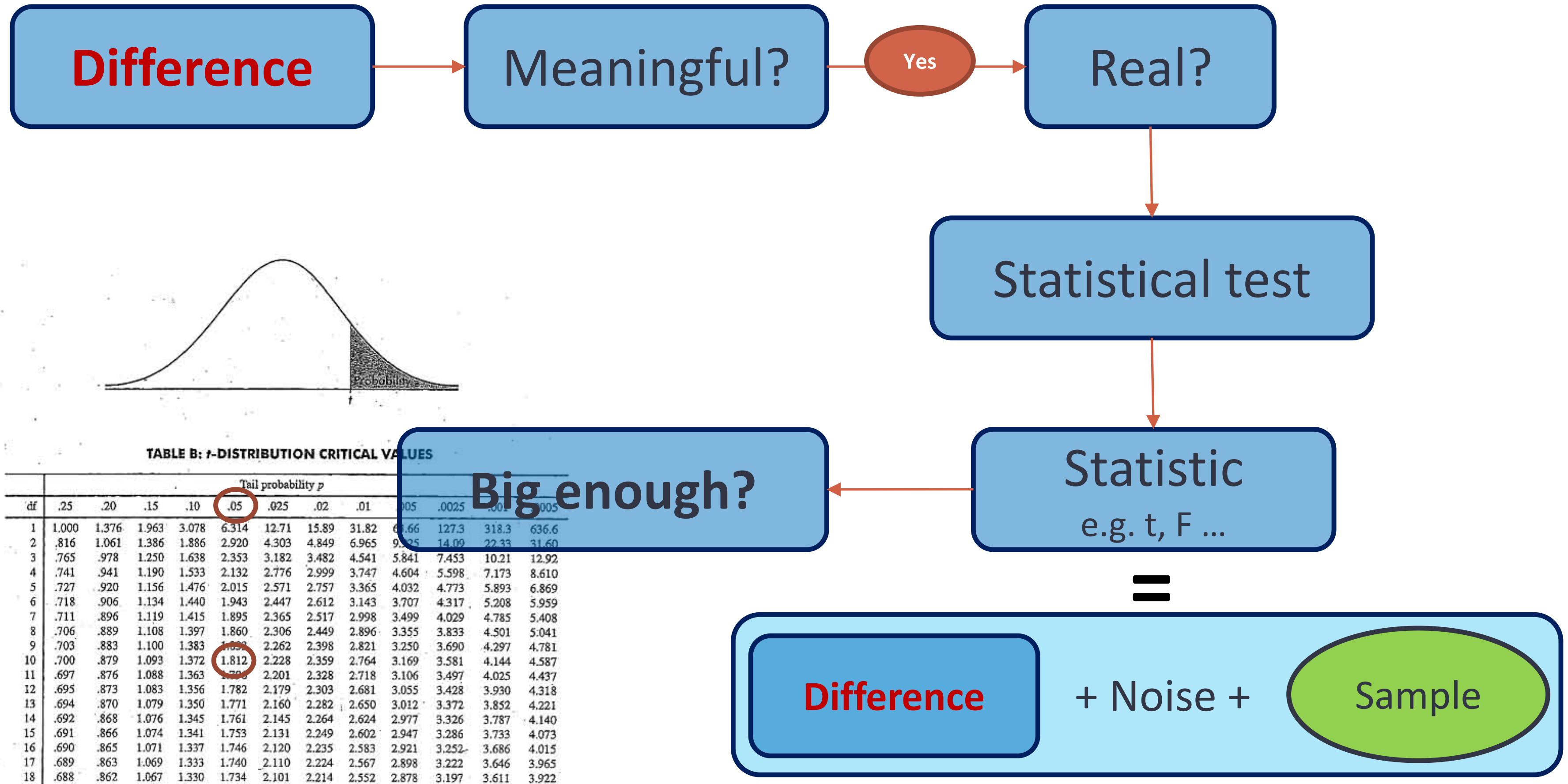
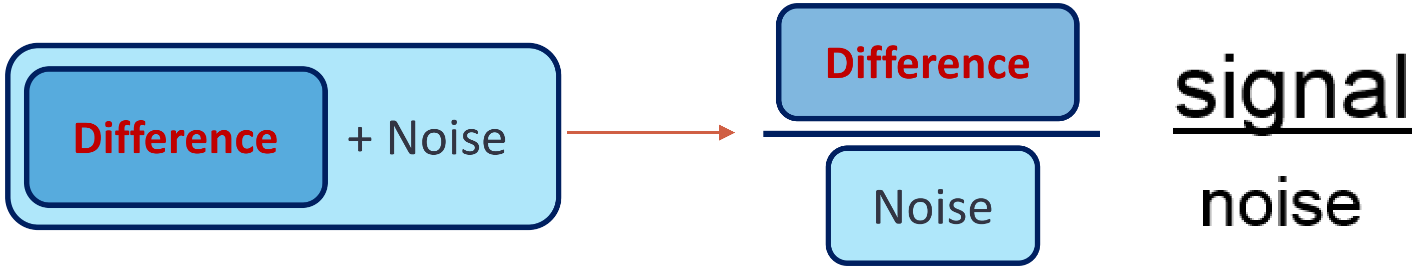




TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

| | Tail probability *p* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| df | .25 | .20 | .15 | .10 | .05 | .025 | .02 | .01 | .005 | .0025 | .001 | .0005 |
| 1 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 15.89 | 31.82 | [illegible] | 127.3 | 318.3 | 636.6 |
| 2 | .816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 4.849 | 6.965 | [illegible] | 14.09 | 22.33 | 31.60 |
| 3 | .765 | .978 | 1.250 | 1.638 | 2.353 | 3.182 | 3.482 | 4.541 | 5.841 | 7.453 | 10.21 | 12.92 |
| 4 | .741 | .941 | 1.190 | 1.533 | 2.132 | 2.776 | 2.999 | 3.747 | 4.604 | 5.598 | 7.173 | 8.610 |
| 5 | .727 | .920 | 1.156 | 1.476 | 2.015 | 2.571 | 2.757 | 3.365 | 4.032 | 4.773 | 5.893 | 6.869 |
| 6 | .718 | .906 | 1.134 | 1.440 | 1.943 | 2.447 | 2.612 | 3.143 | 3.707 | 4.317 | 5.208 | 5.959 |
| 7 | .711 | .896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.517 | 2.998 | 3.499 | 4.029 | 4.785 | 5.408 |
| 8 | .706 | .889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.449 | 2.896 | 3.355 | 3.833 | 4.501 | 5.041 |
| 9 | .703 | .883 | 1.100 | 1.383 | [illegible] | 2.262 | 2.398 | 2.821 | 3.250 | 3.690 | 4.297 | 4.781 |
| 10 | .700 | .879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.359 | 2.764 | 3.169 | 3.581 | 4.144 | 4.587 |
| 11 | .697 | .876 | 1.088 | 1.363 | [illegible] | 2.201 | 2.328 | 2.718 | 3.106 | 3.497 | 4.025 | 4.437 |
| 12 | .695 | .873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.303 | 2.681 | 3.055 | 3.428 | 3.930 | 4.318 |
| 13 | .694 | .870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.282 | 2.650 | 3.012 | 3.372 | 3.852 | 4.221 |
| 14 | .692 | .868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.264 | 2.624 | 2.977 | 3.326 | 3.787 | 4.140 |
| 15 | .691 | .866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.249 | 2.602 | 2.947 | 3.286 | 3.733 | 4.073 |
| 16 | .690 | .865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.235 | 2.583 | 2.921 | 3.252 | 3.686 | 4.015 |
| 17 | .689 | .863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.224 | 2.567 | 2.898 | 3.222 | 3.646 | 3.965 |
| 18 | .688 | .862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.214 | 2.552 | 2.878 | 3.197 | 3.611 | 3.922 |
| 19 | .688 | .861 | 1.066 | 1.328 | [illegible] | 2.093 | 2.205 | 2.539 | 2.861 | 3.174 | 3.579 | 3.883 |
| 20 | .687 | .860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.197 | 2.528 | 2.845 | 3.153 | 3.552 | 3.850 |
| 21 | .686 | .859 | 1.063 | 1.323 | [illegible] | 2.080 | 2.189 | 2.518 | 2.831 | 3.135 | 3.527 | 3.819 |
| 22 | .686 | .858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.183 | 2.508 | 2.819 | 3.119 | 3.505 | 3.792 |

# Signal-to-noise ratio

- Stats are all about understanding and controlling variation.




signal / noise — If the **noise is low** then the **signal is detectable** …
= statistical significance

signal / noise — … but if the **noise** (i.e. interindividual variation) **is large** then the **same signal will not be detected**
= no statistical significance

- In a statistical test, the ratio of signal to noise determines the significance.

# Analysis of Quantitative Data

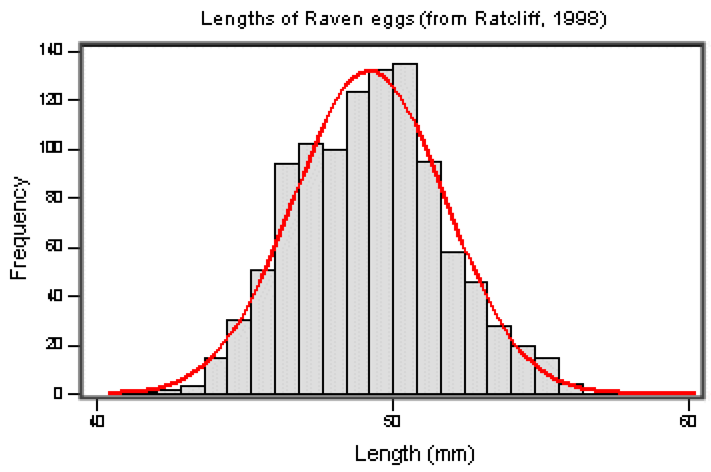
- Choose the correct statistical test to answer your question:
  - They are 2 types of statistical tests:
    - **Parametric tests** with 4 assumptions to be met by the data,
    - **Non-parametric tests** with no or few assumptions (e.g. Mann-Whitney test) and/or for qualitative data (e.g. Fisher's exact and $\chi^2$ tests).

# Assumptions of Parametric Data

- All parametric tests have 4 basic assumptions that must be met for the test to be accurate.

## ***First assumption: Normally distributed data***

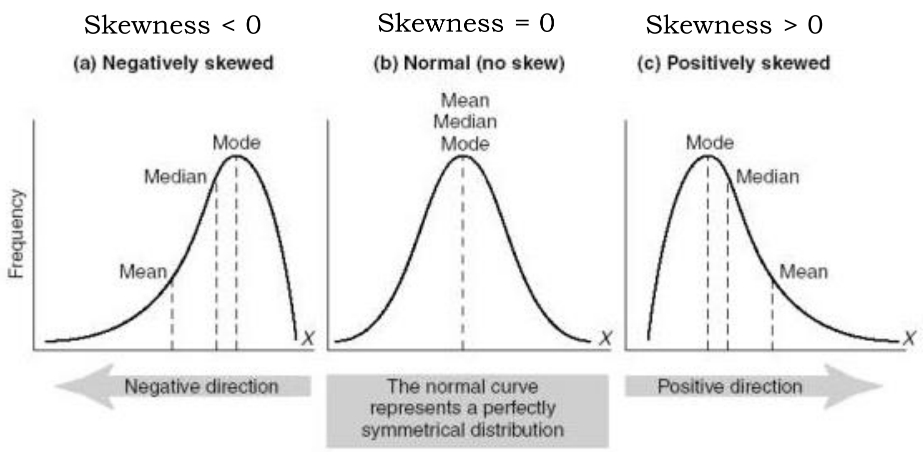
– Normal shape, bell shape, Gaussian shape

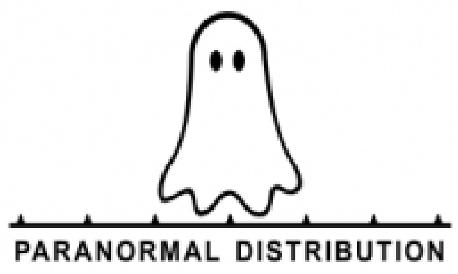



- Transformations can be made to make data suitable for parametric analysis.

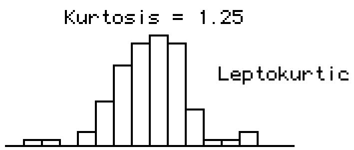
# Assumptions of Parametric Data

- Frequent departures from normality:
  - Skewness: lack of symmetry of a distribution

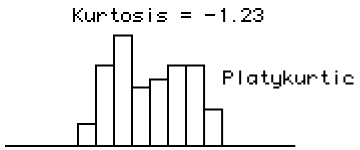


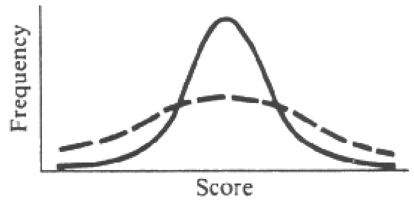

  - Kurtosis: measure of the degree of 'peakedness' in the distribution
    - The two distributions below have the same variance approximately the same skew, but differ markedly in kurtosis.




More peaked distribution: kurtosis > 0




Flatter distribution: kurtosis < 0




(e) Platykurtic and leptokurtic

# Assumptions of Parametric Data

***Second assumption: Homoscedasticity (Homogeneity in variance)***

- The variance should not change systematically throughout the data

***Third assumption: Interval data (linearity)***

- The distance between points of the scale should be equal at all parts along the scale.

***Fourth assumption: Independence***

- Data from different subjects are independent
  - Values corresponding to one subject do not influence the values corresponding to another subject.
  - Important in repeated measures experiments

# Analysis of Quantitative Data

- **Is there a difference between my groups regarding the variable I am measuring?**
  - e.g. are the mice in the group A heavier than those in group B?
    - Tests with 2 groups:
      - Parametric: **Student's *t*-test**
      - Non parametric: **Mann-Whitney/Wilcoxon rank sum test**
    - Tests with more than 2 groups:
      - Parametric: **Analysis of variance (one-way and two-way ANOVA)**
      - Non parametric: **Kruskal Wallis (one-way ANOVA equivalent)**
- **Is there a relationship between my 2 (continuous) variables?**
  - e.g. is there a relationship between the daily intake in calories and an increase in body weight?
    - Test: **Correlation** (parametric or non-parametric)