Statistics in biomedical research, 3rd session: Multiple hypothesis testing; correlation testing; complex experimental designs.

Hervé Seitz (email: herve.seitz@cnrs.fr)

IGH (UMR 9002 CNRS et université de Montpellier)

November 17, 2023

This slideshow is accessible at:
http://www.igh.cnrs.fr/equip/Seitz/en_Stats3.pdf

Comparison of categorical distributions

Comparing counts per category between several experimental conditions.

Comparison of categorical distributions

Comparing counts per category between several experimental conditions.

	Cond. 1	Cond. 2
G1	19	5
S	4	8
G2	25	36
M	3	1

Comparison of categorical distributions

Comparing counts per category between several experimental conditions.

	Cond. 1	Cond. 2
G1	19	5
S	4	8
G2	25	36
M	3	1

A t-test on each category ?

Comparison of categorical distributions

Comparing counts per category between several experimental conditions.

	Cond. 1	Cond. 2
G1	19	5
S	4	8
G2	25	36
M	3	1

A t-test on each category ? Would require replicates of the counting (which already contains multiple observations).

Comparison of categorical distributions

Comparing counts per category between several experimental conditions.

	Cond. 1	Cond. 2
G1	19	5
S	4	8
G2	25	36
M	3	1

A t-test on each category ? Would require replicates of the counting (which already contains multiple observations). What if a category appears significantly different but not the others ?

Comparison of categorical distributions

Comparing counts per category between several experimental conditions.

	Cond. 1	Cond. 2
G1	19	5
S	4	8
G2	25	36
M	3	1

A t-test on each category ? Would require replicates of the counting (which already contains multiple observations). What if a category appears significantly different but not the others ?

$\longrightarrow$ t-test not adapted here.

Comparison of categorical distributions

To compare count tables: χ^2 test or (more precise): Fisher's exact test.

Comparison of categorical distributions

To compare count tables: χ^2 test or (more precise): Fisher's exact test.

Here: χ^2 test *p*-value=0.005921; Fisher's exact test *p*-value=0.003375.

Comparison of categorical distributions

To compare count tables: χ^2 test or (more precise): Fisher's exact test.

Here: χ^2 test p-value=0.005921; Fisher's exact test p-value=0.003375.

χ^2 test: imprecise for small numbers (less than $\approx$10 observations in at least one category).

Comparison of categorical distributions

To compare count tables: χ^2 test or (more precise): Fisher's exact test.

Here: χ^2 test p-value=0.005921; Fisher's exact test p-value=0.003375.

χ^2 test: imprecise for small numbers (less than $\approx$10 observations in at least one category). Fisher's exact test: calculations can be very long if the number of observations is large (many possible permutation combinations).

Comparison of categorical distributions

To compare count tables: χ^2 test or (more precise): Fisher's exact test.

Here: χ^2 test p-value=0.005921; Fisher's exact test p-value=0.003375.

χ^2 test: imprecise for small numbers (less than $\approx$10 observations in at least one category). Fisher's exact test: calculations can be very long if the number of observations is large (many possible permutation combinations).

Danger ! These tests use raw counting data (no normalization !).

Comparison of categorical distributions

3	2
6	6
5	6

p-value=1

Comparison of categorical distributions

3	2
6	6
5	6

p-value=1

30	20
60	60
50	60

p-value=0.2413

Comparison of categorical distributions

3	2
6	6
5	6

p-value=1

30	20
60	60
50	60

p-value=0.2413

300	200
600	600
500	600

p-value=4.588×10^{-7}

Comparison of categorical distributions

3	2
6	6
5	6

p-value=1

30	20
60	60
50	60

p-value=0.2413

300	200
600	600
500	600

p-value=4.588×10^{-7}

Normalization (*e.g.*, percentage) would lose the information on raw observation number.

Comparison of numerical distributions

Comparing numerical distributions globally (not just their means).

Comparison of numerical distributions

Comparing numerical distributions globally (not just their means).

Kolmogorov-Smirnov test: null hypothesis: the two datasets were sampled from the same distribution (unknown, any shape).

Comparison of numerical distributions

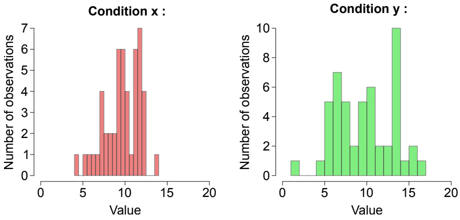
Comparing numerical distributions globally (not just their means).

Kolmogorov-Smirnov test: null hypothesis: the two datasets were sampled from the same distribution (unknown, any shape). Historical version of the test: for continuous variables only.

Comparison of numerical distributions

Comparing numerical distributions globally (not just their means).

Kolmogorov-Smirnov test: null hypothesis: the two datasets were sampled from the same distribution (unknown, any shape).

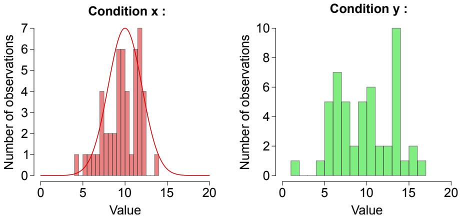


R commands used to generate these graphs: [link].

Comparison of numerical distributions

Comparing numerical distributions globally (not just their means).

Kolmogorov-Smirnov test: null hypothesis: the two datasets were sampled from the same distribution (unknown, any shape).

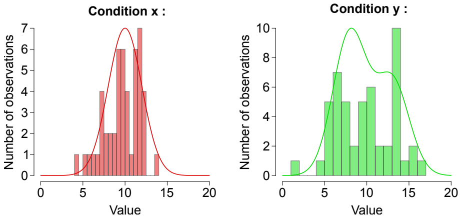


R commands used to generate these graphs: [link].

Comparison of numerical distributions

Comparing numerical distributions globally (not just their means).

Kolmogorov-Smirnov test: null hypothesis: the two datasets were sampled from the same distribution (unknown, any shape).

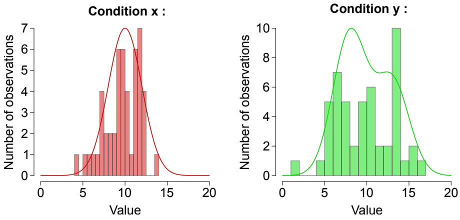


R commands used to generate these graphs: [link].

Comparison of numerical distributions

Comparing numerical distributions globally (not just their means).

Kolmogorov-Smirnov test: null hypothesis: the two datasets were sampled from the same distribution (unknown, any shape).



R commands used to generate these graphs: [link]. *p*-values: t-test: 0.9005; Kolmogorov-Smirnov test: 0.02171.

Comparison of numerical distributions

Comparing numerical distributions globally (not just their means).

Kolmogorov-Smirnov test: null hypothesis: the two datasets were sampled from the same distribution (unknown, any shape).

Comparison of numerical distributions

Comparing numerical distributions globally (not just their means).

Kolmogorov-Smirnov test: null hypothesis: the two datasets were sampled from the same distribution (unknown, any shape).

$\longrightarrow$ More sensitive, but harder to interpret (requires a detailed mechanistic understanding of the process).

Multiple hypothesis testing

Multiple hypothesis testing

Significance threshold of 0.05: expect $\approx$5% false positives.

Multiple hypothesis testing

Significance threshold of 0.05: expect ≈5% false positives.

If you perform many tests (Is there a significant difference between conditions "x" and "y" at day 1 ? At day 2 ? At day 3 ?...)

Multiple hypothesis testing

Significance threshold of 0.05: expect ≈5% false positives.

If you perform many tests (Is there a significant difference between conditions "x" and "y" at day 1 ? At day 2 ? At day 3 ?...) the number of false positives will be $\approx 0.05 \times$number of tests.

Multiple hypothesis testing

Significance threshold of 0.05: expect ≈5% false positives.

If you perform many tests (Is there a significant difference between conditions “x” and “y” at day 1 ? At day 2 ? At day 3 ?...) the number of false positives will be $\approx 0.05 \times$number of tests.

$\longrightarrow$ performing 20,000 tests, you would get ≈1,000 false positives (transcriptomics experiments would always be wrong !).

Multiple hypothesis testing

Significance threshold of 0.05: expect ≈5% false positives.

If you perform many tests (Is there a significant difference between conditions "x" and "y" at day 1 ? At day 2 ? At day 3 ?...) the number of false positives will be $\approx 0.05 \times$number of tests.

$\longrightarrow$ performing 20,000 tests, you would get ≈1,000 false positives (transcriptomics experiments would always be wrong !).

Difference between true and false positives: true positives are reproducible (but: large experiments are hard to reproduce in practice).

Multiple hypothesis testing

Empirical method: making significance threshold more and more stringent if the number of tests increases.

Multiple hypothesis testing

Empirical method: making significance threshold more and more stringent if the number of tests increases.

Bonferroni correction: if n is the number of tested hypotheses, and α is the usual threshold, then rather use α/n as a threshold.

Multiple hypothesis testing

Empirical method: making significance threshold more and more stringent if the number of tests increases.

Bonferroni correction: if n is the number of tested hypotheses, and α is the usual threshold, then rather use α/n as a threshold. Easier to use: multiply every p-value by n (and whenever the product is larger than 1, set it to 1) rather than dividing the threshold.

Multiple hypothesis testing

Empirical method: making significance threshold more and more stringent if the number of tests increases.

Bonferroni correction: if n is the number of tested hypotheses, and α is the usual threshold, then rather use α/n as a threshold. Easier to use: multiply every p-value by n (and whenever the product is larger than 1, set it to 1) rather than dividing the threshold.

Benjamini-Hochberg correction: do not multiply all n p-values by n, but: by an incrementally increasing factor (from 1 to n) in the decreasing list of p-values.

Multiple hypothesis testing

Empirical method: making significance threshold more and more stringent if the number of tests increases.

Bonferroni correction: if n is the number of tested hypotheses, and α is the usual threshold, then rather use α/n as a threshold. Easier to use: multiply every p-value by n (and whenever the product is larger than 1, set it to 1) rather than dividing the threshold.

Benjamini-Hochberg correction: do not multiply all n p-values by n, but: by an incrementally increasing factor (from 1 to n) in the decreasing list of p-values. Less stringent, less false negatives. Completely *ad hoc* principle, but very popular in high-throughput molecular biology.

Multiple hypothesis testing

Empirical method: making significance threshold more and more stringent if the number of tests increases.

Bonferroni correction: if n is the number of tested hypotheses, and α is the usual threshold, then rather use α/n as a threshold. Easier to use: multiply every p-value by n (and whenever the product is larger than 1, set it to 1) rather than dividing the threshold.

Benjamini-Hochberg correction: do not multiply all n p-values by n, but: by an incrementally increasing factor (from 1 to n) in the decreasing list of p-values. Less stringent, less false negatives. Completely *ad hoc* principle, but very popular in high-throughput molecular biology.

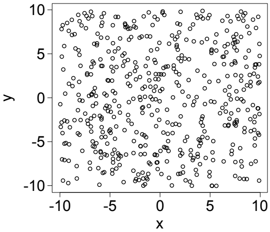
A particular case: multiple t-tests against a common control condition: Dunnett's test.

Correlation tests

Correlation tests

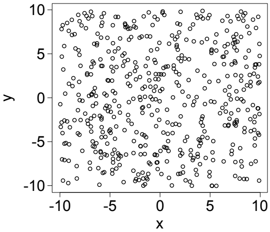
Principle: do two variables tend to co-vary, or do they vary independently ?

Correlation tests



R commands used to generate that graph: [link].

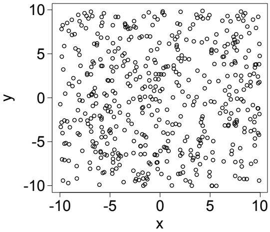
Correlation tests



R commands used to generate that graph: [link].

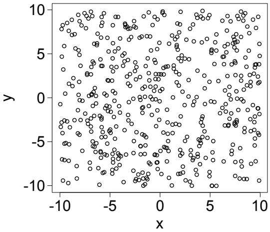
Pearson's coefficient: $r = 0.0352$ (p-value$=0.432$).

Correlation tests



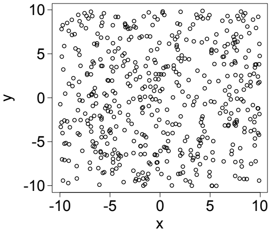
R commands used to generate that graph: [link].
Null hypothesis: correlation coefficient is 0.
Pearson's coefficient: $r = 0.0352$ (p-value$=0.432$).

Correlation tests



R commands used to generate that graph: [link].
Null hypothesis: correlation coefficient is 0.
Pearson's coefficient: $r = 0.0352$ (p-value$=0.432$). $= +1$ for a perfect and increasing linear correlation, and -1 if it is decreasing; intermediary values for imperfect correlation.

Correlation tests



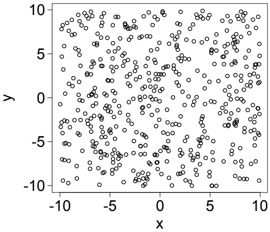
R commands used to generate that graph: [link].

Null hypothesis: correlation coefficient is 0.

Pearson's coefficient: $r = 0.0352$ (p-value=0.432).

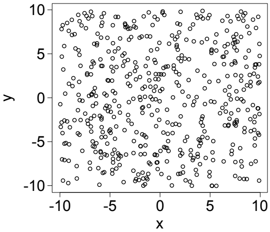
Kendall's coefficient: $\tau = 0.0203$ (p-value=0.4968).

Correlation tests



R commands used to generate that graph: [link].
Null hypothesis: correlation coefficient is 0.
Pearson's coefficient: $r = 0.0352$ (p-value$=0.432$).
Kendall's coefficient: $\tau = 0.0203$ (p-value$=0.4968$). $= +1$ if every point pair varies concordantly, and -1 if they all vary discordantly; intermediary values otherwise.

Correlation tests



R commands used to generate that graph: [link].

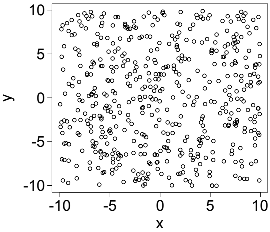
Null hypothesis: correlation coefficient is 0.

Pearson's coefficient: $r = 0.0352$ (p-value=0.432).

Kendall's coefficient: $\tau = 0.0203$ (p-value=0.4968).

Spearman's coefficient: $\rho = 0.0315$ (p-value=0.4827).

Correlation tests



R commands used to generate that graph: [link].

Null hypothesis: correlation coefficient is 0.

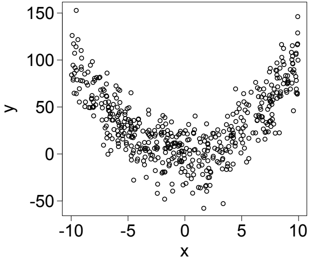
Pearson's coefficient: $r = 0.0352$ (p-value=0.432).

Kendall's coefficient: $\tau = 0.0203$ (p-value=0.4968).

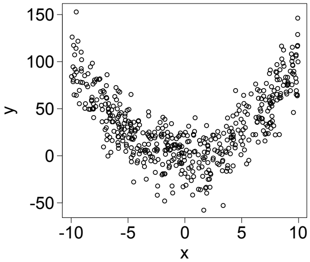
Spearman's coefficient: $\rho = 0.0315$ (p-value=0.4827).

Pearson's coefficient on values' ranks (looks for a monotonous relationship, not necessarily linear).

Correlation tests

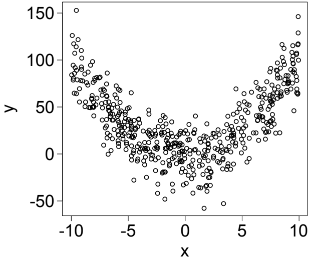


Correlation tests



Pearson's coefficient: $r = 0.0846$ (p-value=0.05882)
Kendall's coefficient: $\tau = 0.0126$ (p-value=0.6732)
Spearman's coefficient: $\rho = 0.0435$ (p-value=0.3313)

Correlation tests



Pearson's coefficient: $r = 0.0846$ (p-value=0.05882)
Kendall's coefficient: $\tau = 0.0126$ (p-value=0.6732)
Spearman's coefficient: $\rho = 0.0435$ (p-value=0.3313)

$\longrightarrow$ Need to have a mathematical model for the response y to x ("is there a correlation between y and x^2 ?").

Correlation tests

A classical trap: correlation does not imply causality.

Correlation tests

A classical trap: correlation does not imply causality.

A is a cause for B, or B is a cause for A ? Are A and B two consequences of the same cause C ? ...

Elaborated experimental design

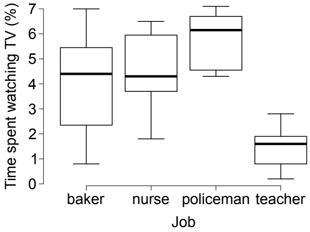
Comparing more than 2 groups.

Elaborated experimental design

Comparing more than 2 groups (*e.g.*, "Between bakers, teachers, policeman, nurses, is there a difference in the time spent watching TV ? ").

Elaborated experimental design

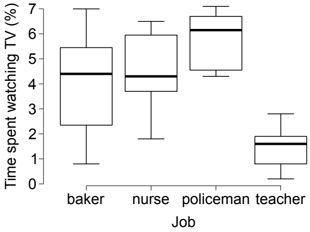
Comparing more than 2 groups.



R commands used to generate that graph: [link].

Elaborated experimental design

Comparing more than 2 groups.

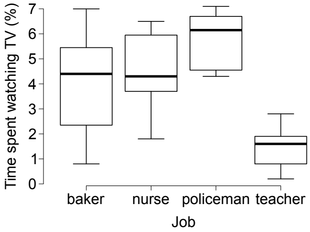


Analysis of variance (ANOVA): conditions: residual normality ($\Longrightarrow$ normality of observations within each group), variance homogeneity, and independence between observations.

R commands used to generate that graph: [link].

Elaborated experimental design

Comparing more than 2 groups.



Analysis of variance (ANOVA): conditions: residual normality ($\Longrightarrow$ normality of observations within each group), variance homogeneity, and independence between observations.
ANOVA p-value$=7.39\times10^{-6}$ $\longrightarrow$ an effet of job (without further detail !).
R commands used to generate that graph: [link].

Elaborated experimental design

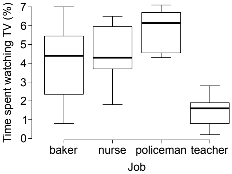
“Post-hoc” tests (here: pairwise t-tests) to identify mutually significantly different groups.

Elaborated experimental design

"Post-hoc" tests (here: pairwise t-tests) to identify mutually significantly different groups.

t-test p-values with Benjamini-Hochberg correction:

	baker	nurse	policeman
nurse	0.47648	-	-
policeman	0.03867	0.12304	-
teacher	0.00290	0.00036	5.2×10^{-6}



Elaborated experimental design

"Post-hoc" tests (here: pairwise t-tests) to identify mutually significantly different groups.

t-test *p*-values with Benjamini-Hochberg correction:

	baker	nurse	policeman
nurse	0.47648	-	-
policeman	0.03867	0.12304	-
teacher	0.00290	0.00036	5.2×10^{-6}

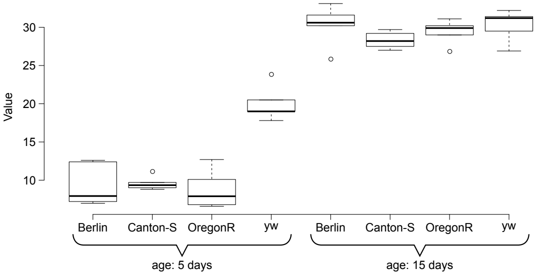
Danger ! Start with ANOVA before engaging into pairwise t-tests (high risk of false positives otherwise: multiple hypothesis testing).

Elaborated experimental design

Several variables simultaneously (*e.g.*, effect of age and *Drosophila* strain on a physiological response).

Elaborated experimental design

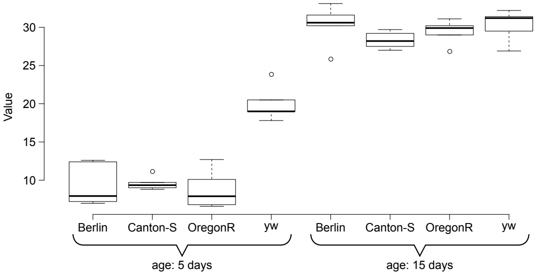
Several variables simultaneously (*e.g.*, effect of age and *Drosophila* strain on a physiological response).



R commands used to generate that graph: [link].

Elaborated experimental design

Several variables simultaneously (*e.g.*, effect of age and *Drosophila* strain on a physiological response).

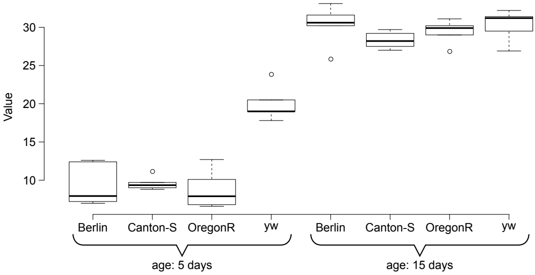


Multidimensional ANOVA (here: two variables $\longrightarrow$ two-way ANOVA).

R commands used to generate that graph: [link].

Elaborated experimental design

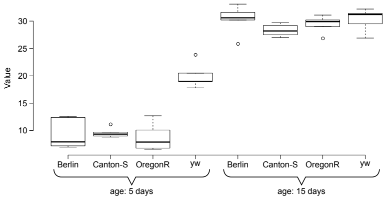
Several variables simultaneously (*e.g.*, effect of age and *Drosophila* strain on a physiological response).



Multidimensional ANOVA (here: two variables $\longrightarrow$ two-way ANOVA).
Same requirements than one-way ANOVA: normality, homoscedasticity, independence.
R commands used to generate that graph: [link].

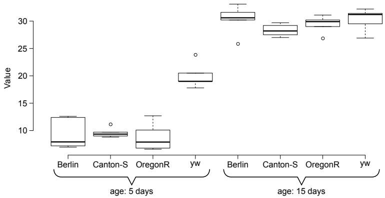
Elaborated experimental design

Two-way ANOVA without interaction: p-values: strain: 1.47×10^{-4}; age: $< 2 \times 10^{-16}$.



Elaborated experimental design

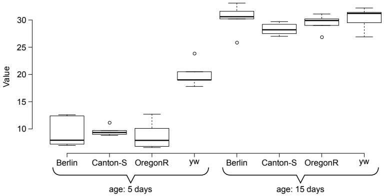
Two-way ANOVA without interaction: *p*-values: strain: 1.47×10^{-4}; age: $<2\times10^{-16}$.



If each variable has an effect, their interaction could have one too.

Elaborated experimental design

Two-way ANOVA without interaction: *p*-values: strain: 1.47×10^{-4}; age: $<2 \times 10^{-16}$.

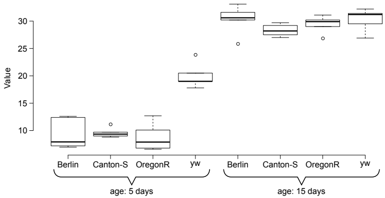


If each variable has an effect, their interaction could have one too.

Two-way ANOVA with interaction: *p*-values: strain: 3.17×10^{-7}; age: $<2 \times 10^{-16}$; their interaction: 6.25×10^{-6}.

Elaborated experimental design

Two-way ANOVA without interaction: p-values: strain: 1.47×10^{-4}; age: $<2\times10^{-16}$.



If each variable has an effect, their interaction could have one too.

Two-way ANOVA with interaction: p-values: strain: 3.17×10^{-7}; age: $<2\times10^{-16}$; their interaction: 6.25×10^{-6}. Interpretation: age has an effect, strain has an effect, and aging affects these various strains differently.

Elaborated experimental design

If conditions of applicability of ANOVA are not met:

Elaborated experimental design

If conditions of applicability of ANOVA are not met:

- A mathematical transformation (ex.: log) could make them being met.

Elaborated experimental design

If conditions of applicability of ANOVA are not met:

- A mathematical transformation (ex.: log) could make them being met.
- Non-parametric alternatives (robust to non-normality and heteroscedasticity) for one-way ANOVA: Kruskal-Wallis test (non-repeated mesurements), Friedman test (repeated measurements on each subject).

Elaborated experimental design

If conditions of applicability of ANOVA are not met:

- A mathematical transformation (ex.: log) could make them being met.
- Non-parametric alternatives (robust to non-normality and heteroscedasticity) for one-way ANOVA: Kruskal-Wallis test (non-repeated mesurements), Friedman test (repeated measurements on each subject).

If variables are not categorical ("job", "*Drosophila* strain") but numerical with more than 2 levels: mathematical models (*e.g.*, linear models) to extract the effect of each variable.

Conclusion

Conclusion

- Basic concepts, generalizable to many statistical tests (p-value, confidence interval, ...).

Conclusion

- Basic concepts, generalizable to many statistical tests (p-value, confidence interval, ...).
- Vocabulary (standard deviation $\neq$ standard error; normality; homoscedasticity; ...).

Conclusion

- Basic concepts, generalizable to many statistical tests (p-value, confidence interval, ...).
- Vocabulary (standard deviation $\neq$ standard error; normality; homoscedasticity; ...).
- $\longrightarrow$ being able to find information by yourself for more complicated cases.

Supplements

Summarized versions of this course, in French:

- Written: first part (published in July 2010 in *Regard sur la biochimie*), second part (published in October 2010 in *Regard sur la biochimie*).
- Video: "Les statistiques en biologie moléculaire ".