Statistics in biomedical research, 1^{st} session: Why do we need statistics? Generalities and introduction to the t-test.

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

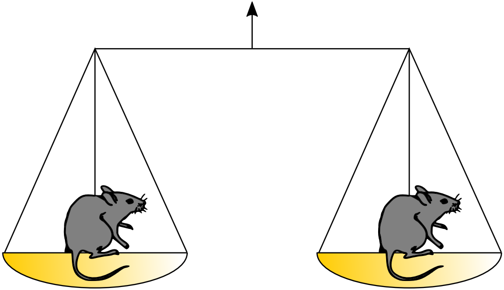
IGH (UMR 9002 CNRS et université de Montpellier)

November 16, 2023

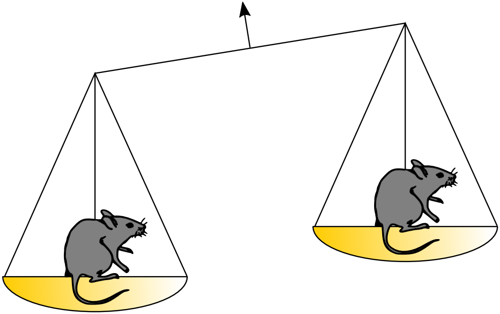
This slideshow is accessible at:
http://www.igh.cnrs.fr/equip/Seitz/en_Stats1.pdf

Why do we need statistics ?

Why do we need statistics ?



Why do we need statistics ?

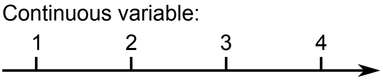


Continuous and discontinuous variables

Example of a continuous variable: mouse body mass (can take any intermediary value).

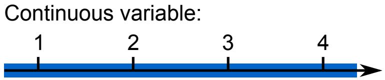
Continuous and discontinuous variables

Example of a continuous variable: mouse body mass (can take any intermediary value).



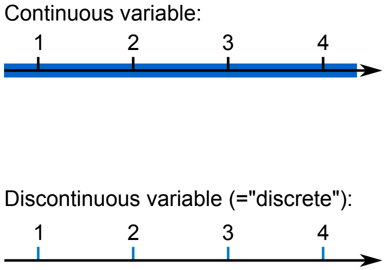
Continuous and discontinuous variables

Example of a continuous variable: mouse body mass (can take any intermediary value).



Continuous and discontinuous variables

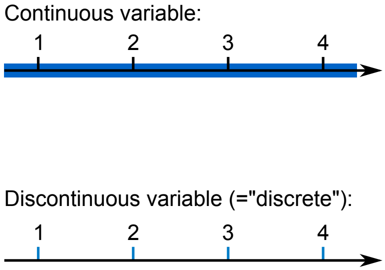
Example of a continuous variable: mouse body mass (can take any intermediary value).



Continuous and discontinuous variables

Example of a continuous variable: mouse body mass (can take any intermediary value).

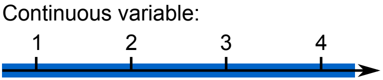
Example of a discontinuous (=”discrete”) variable: number of progeny at birth (can only take specific values).

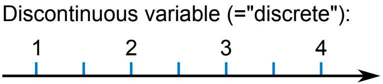


Continuous and discontinuous variables

Example of a continuous variable: mouse body mass (can take any intermediary value).

Example of a discontinuous (="discrete") variable: number of progeny at birth (can only take specific values).





Continuous and discontinuous variables

Continuous variable: two perfectly precise measurements will never give the exact same result (with all the decimals being identical).

Continuous and discontinuous variables

Continuous variable: two perfectly precise measurements will never give the exact same result (with all the decimals being identical).

Implication: measurement in "Condition #1" vs. "Condition #2": I know beforehand that there will be a difference anyway!

Experimental irreproducibility

Sources of experimental irreproducibility:

Experimental irreproducibility

Sources of experimental irreproducibility:

- Imperfection in measurement device.

Experimental irreproducibility

Sources of experimental irreproducibility:

- Imperfection in measurement device.
- A real difference between the measured objects, which I decide to consider un-important.

Experimental irreproducibility

Sources of experimental irreproducibility:

- Imperfection in measurement device.
- A real difference between the measured objects, which I decide to consider un-important.

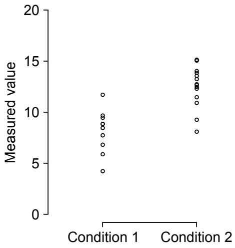
$\longrightarrow$ we have to measure several replicates, then compare intra-group variability to inter-group variability.

Experimental irreproducibility

Sources of experimental irreproducibility:

- Imperfection in measurement device.
- A real difference between the measured objects, which I decide to consider un-important.

$\longrightarrow$ we have to measure several replicates, then compare intra-group variability to inter-group variability.

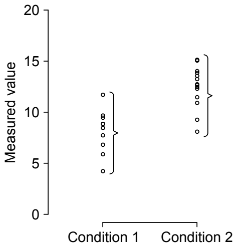


Experimental irreproducibility

Sources of experimental irreproducibility:

- Imperfection in measurement device.
- A real difference between the measured objects, which I decide to consider un-important.

$\longrightarrow$ we have to measure several replicates, then compare intra-group variability to inter-group variability.



Experimental irreproducibility

Sources of experimental irreproducibility:

- Imperfection in measurement device.
- A real difference between the measured objects, which I decide to consider un-important.

$\longrightarrow$ we have to measure several replicates, then compare intra-group variability to inter-group variability.

In biology: “biological replicates” (distinct biological objects), “technical replicates” (several measurements on the same biological object).

Experimental irreproducibility

Sources of experimental irreproducibility:

- Imperfection in measurement device.
- A real difference between the measured objects, which I decide to consider un-important.

$\longrightarrow$ we have to measure several replicates, then compare intra-group variability to inter-group variability.

In biology: "biological replicates" (distinct biological objects), "technical replicates" (several measurements on the same biological object; do not capture biological variability).

Continuous and discontinuous variables

Continuous variable: two perfectly precise measurements will never give the exact same result (with all the decimals being identical).

Continuous and discontinuous variables

Continuous variable: two perfectly precise measurements will never give the exact same result (with all the decimals being identical).

Implication: measurement in "Condition #1" vs. "Condition #2": I know beforehand that there will be a difference anyway!

Discontinuous variable: measured values can be absolutely identical.

Continuous and discontinuous variables

Continuous variable: two perfectly precise measurements will never give the exact same result (with all the decimals being identical).

Implication: measurement in “Condition #1” vs. “Condition #2”: I know beforehand that there will be a difference anyway!

Discontinuous variable: measured values can be absolutely identical.

... but sources of irreproducibility still imply replicating measurements, and comparing intra- to inter-group variability.

The p-value

The t-test compares two datasets (example: the same measurement, performed in two experimental conditions, with replicates).

The *p*-value

The t-test compares two datasets (example: the same measurement, performed in two experimental conditions, with replicates).

Its *p*-value is the probability that the two sampled populations have the same mean.

The *p*-value

The t-test compares two datasets (example: the same measurement, performed in two experimental conditions, with replicates).

Its *p*-value is the probability that the two sampled populations have the same mean.

Danger ! It is very easy to misunderstand that definition!

The p-value

The t-test compares two datasets (example: the same measurement, performed in two experimental conditions, with replicates).

Its p-value is the probability that the two sampled populations have the same mean.

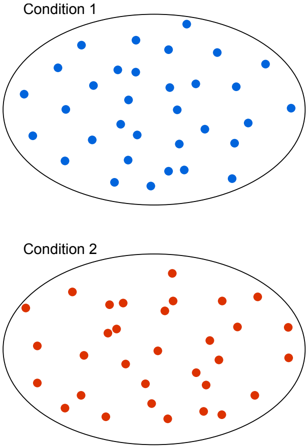
Danger ! It is very easy to misunderstand that definition!

The *p*-value

Another way to perceive the experiment:

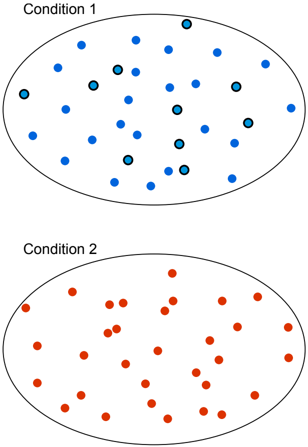
The *p*-value

Another way to perceive the experiment:



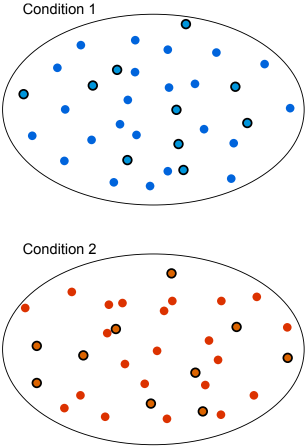
The *p*-value

Another way to perceive the experiment:



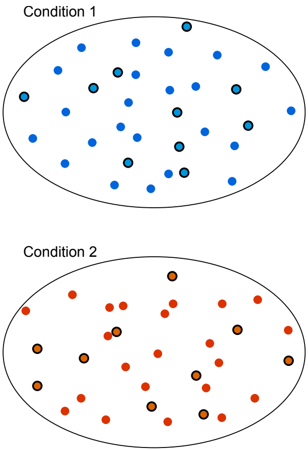
The *p*-value

Another way to perceive the experiment:



The *p*-value

Another way to perceive the experiment:



Random picking from an infinite population.

The p-value

The t-test compares two datasets (example: the same measurement, performed in two experimental conditions, with replicates).

Its p-value is the probability that the two sampled populations have the same mean.

The *p*-value

The t-test compares two datasets (example: the same measurement, performed in two experimental conditions, with replicates).

Its *p*-value is the probability that the two sampled populations have the same mean.

The "sampled population" is that theoretical, infinite population, where replicates have been randomly picked.

The p-value

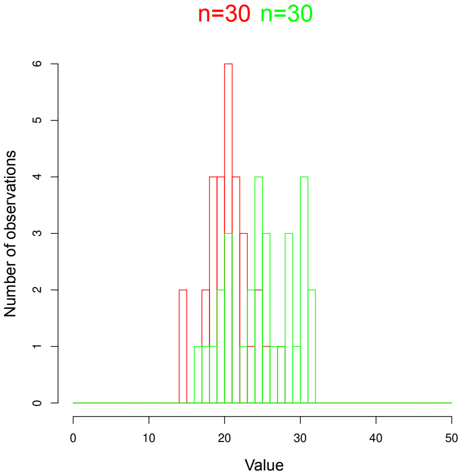
The t-test compares two datasets (example: the same measurement, performed in two experimental conditions, with replicates).

Its p-value is the probability that the two sampled populations have the same mean.

The "sampled population" is that theoretical, infinite population, where replicates have been randomly picked.

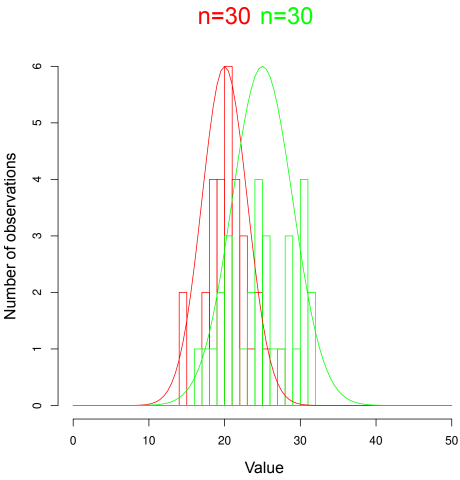
$\longrightarrow$ The t-test permits concluding on the equality of means of infinite cohorts (inaccessible to experimentation).

The *p*-value



R commands used to generate that graph: [link]

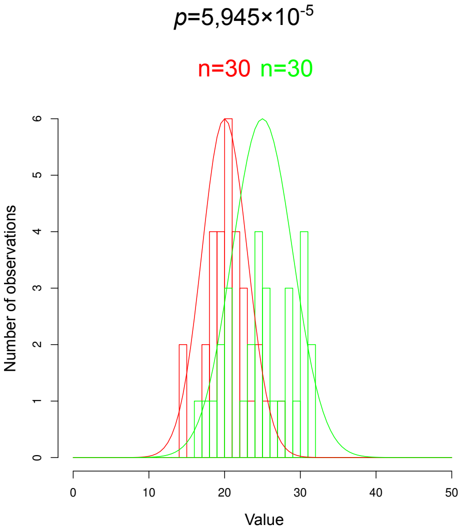
The *p*-value



R commands used to generate that graph: [link]

H. Seitz

The *p*-value



R commands used to generate that graph: [link]

Vocabulary in descriptive statistics

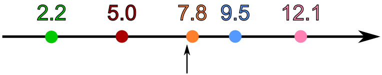
Vocabulary in descriptive statistics

Vocabulary in descriptive statistics

- The mean (implicitly: "arithmetic" mean): sum of values, divided by the number of values.

Vocabulary in descriptive statistics

- The mean (implicitly: “arithmetic” mean): sum of values, divided by the number of values.



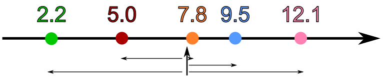
$$m = \frac{2.2+5.0+7.8+9.5+12.1}{5} = 7.32$$

Vocabulary in descriptive statistics

- The mean (implicitly: “arithmetic” mean): sum of values, divided by the number of values.
- Variance: mean of the squared distance to the mean.

Vocabulary in descriptive statistics

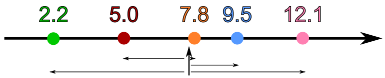
- The mean (implicitly: "arithmetic" mean): sum of values, divided by the number of values.
- Variance: mean of the squared distance to the mean.



$$v_{pop.} = \frac{(7.32-2.2)^2+(7.32-5.0)^2+(7.8-7.32)^2+(9.5-7.32)^2+(12.1-7.32)^2}{5} = 11.8856$$

Vocabulary in descriptive statistics

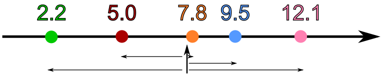
- The mean (implicitly: "arithmetic" mean): sum of values, divided by the number of values.
- Variance: mean of the squared distance to the mean. Sample variance: $\frac{n}{n-1} \times$ population variance, for a sample of n values.



$$v_{pop.} = \frac{(7.32-2.2)^2+(7.32-5.0)^2+(7.8-7.32)^2+(9.5-7.32)^2+(12.1-7.32)^2}{5} = 11.8856$$

Vocabulary in descriptive statistics

- The mean (implicitly: "arithmetic" mean): sum of values, divided by the number of values.
- Variance: mean of the squared distance to the mean. Sample variance: $\frac{n}{n-1} \times$ population variance, for a sample of n values.



$$v_{pop.} = \frac{(7.32-2.2)^2+(7.32-5.0)^2+(7.8-7.32)^2+(9.5-7.32)^2+(12.1-7.32)^2}{5} = 11.8856$$

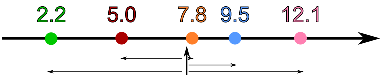
$$v_{samp.} = \frac{(7.32-2.2)^2+(7.32-5.0)^2+(7.8-7.32)^2+(9.5-7.32)^2+(12.1-7.32)^2}{4} = 14.857$$

Vocabulary in descriptive statistics

- The mean (implicitly: “arithmetic” mean): sum of values, divided by the number of values.
- Variance: mean of the squared distance to the mean. Sample variance: $\frac{n}{n-1}\times$ population variance, for a sample of n values.
- Standard deviation: squareroot of variance (hence: homogeneous to the mean).

Vocabulary in descriptive statistics

- The mean (implicitly: “arithmetic” mean): sum of values, divided by the number of values.
- Variance: mean of the squared distance to the mean. Sample variance: $\frac{n}{n-1}\times$ population variance, for a sample of n values.
- Standard deviation: squareroot of variance (hence: homogeneous to the mean).



$$\sigma_{pop.} = \sqrt{V_{pop.}} = 3.44755$$

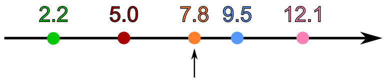
$$\sigma_{samp.} = \sqrt{V_{samp.}} = 3.854478$$

Vocabulary in descriptive statistics

- Median: in a list of sorted (*e.g.*, increasing) values: the value in the middle (if *ex aequo*: their mean).

Vocabulary in descriptive statistics

- Median: in a list of sorted (*e.g.*, increasing) values: the value in the middle (if *ex aequo*: their mean).

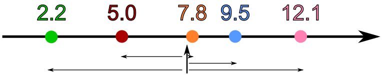


Vocabulary in descriptive statistics

- Median: in a list of sorted (*e.g.*, increasing) values: the value in the middle (if *ex aequo*: their mean).
- Coefficient of variation: standard deviation divided by mean (therefore normalizing by the amplitude of values).

Vocabulary in descriptive statistics

- Median: in a list of sorted (*e.g.*, increasing) values: the value in the middle (if *ex aequo*: their mean).
- Coefficient of variation: standard deviation divided by mean (therefore normalizing by the amplitude of values).



$$CV_{pop.} = \frac{\sigma_{pop.}}{m_{pop.}}$$

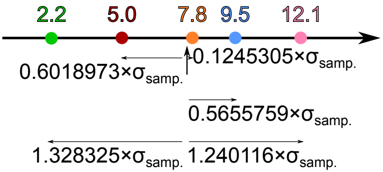
$$CV_{samp.} = \frac{\sigma_{samp.}}{m_{samp.}}$$

Vocabulary in descriptive statistics

- Median: in a list of sorted (*e.g.*, increasing) values: the value in the middle (if *ex aequo*: their mean).
- Coefficient of variation: standard deviation divided by mean (therefore normalizing by the amplitude of values).
- z-score : for each replicate: its distance to the mean (measured in terms of standard deviations).

Vocabulary in descriptive statistics

- Median: in a list of sorted (*e.g.*, increasing) values: the value in the middle (if *ex aequo*: their mean).
- Coefficient of variation: standard deviation divided by mean (therefore normalizing by the amplitude of values).
- z-score : for each replicate: its distance to the mean (measured in terms of standard deviations).

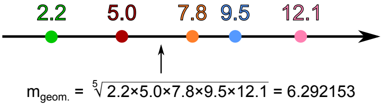


Vocabulary in descriptive statistics

- Geometric mean: n^{th} root of the product of all n values.

Vocabulary in descriptive statistics

- Geometric mean: n^{th} root of the product of all n values.

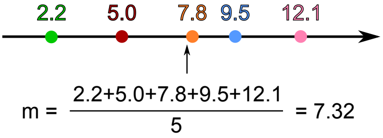


Vocabulary in descriptive statistics

- Geometric mean: n^{th} root of the product of all n values.
- Standard error (implicitly: "of the mean"): standard deviation of a distribution of estimations of the mean (based on distinct sampled datasets).

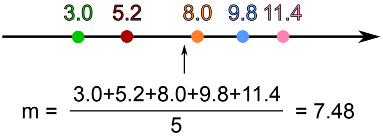
Vocabulary in descriptive statistics

- Geometric mean: n^{th} root of the product of all n values.
- Standard error (implicitly: "of the mean"): standard deviation of a distribution of estimations of the mean (based on distinct sampled datasets).



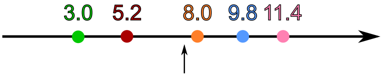
Vocabulary in descriptive statistics

- Geometric mean: n^{th} root of the product of all n values.
- Standard error (implicitly: "of the mean"): standard deviation of a distribution of estimations of the mean (based on distinct sampled datasets).



Vocabulary in descriptive statistics

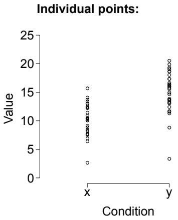
- Geometric mean: n^{th} root of the product of all n values.
- Standard error (implicitly: "of the mean"): standard deviation of a distribution of estimations of the mean (based on distinct sampled datasets). Calling σ the (theoretically identical) standard deviation of each of these sampled datasets, standard error equals: $\frac{\sigma}{\sqrt{n}}$.



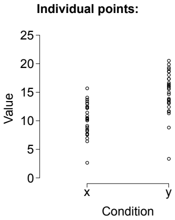
$$m = \frac{3.0+5.2+8.0+9.8+11.4}{5} = 7.48$$

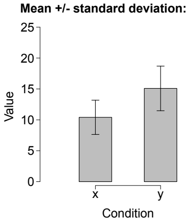
Graphical display

Graphical display

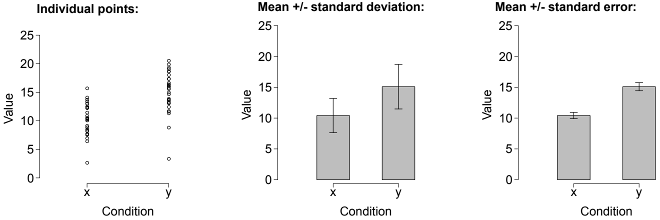


Graphical display



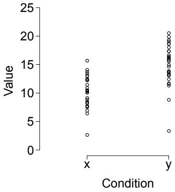


Graphical display

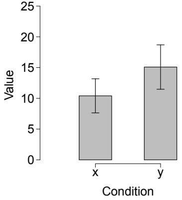


Graphical display

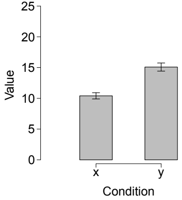
Individual points:



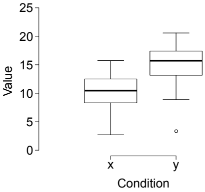
Mean +/- standard deviation:



Mean +/- standard error:

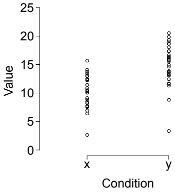


Boxplot:

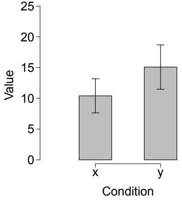


Graphical display

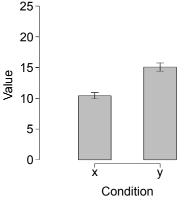
Individual points:



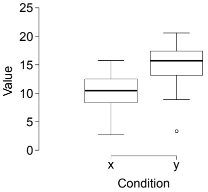
Mean +/- standard deviation:



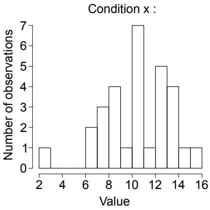
Mean +/- standard error:



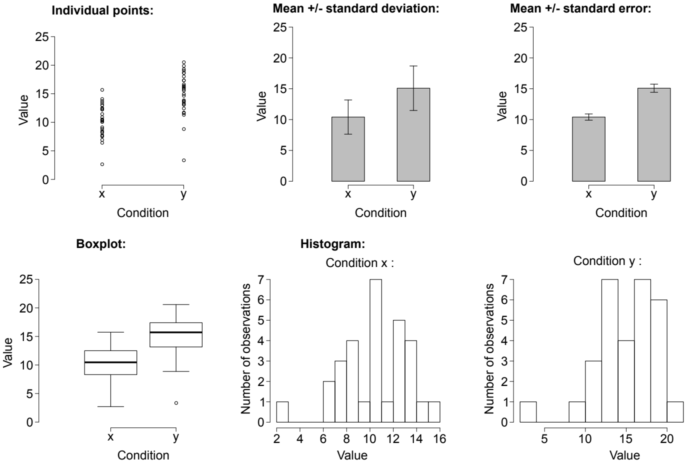
Boxplot:



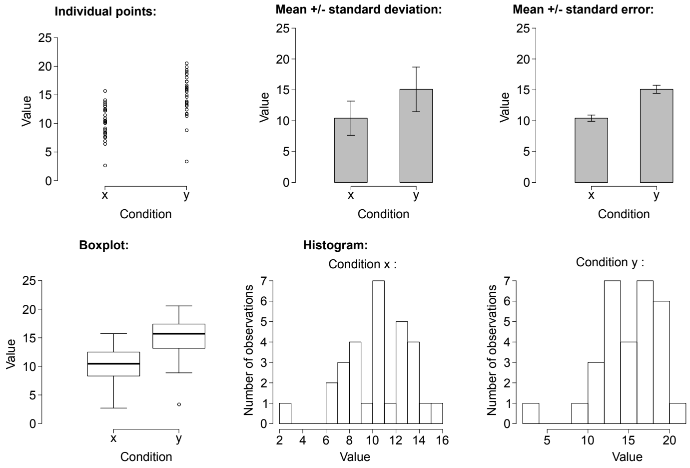
Histogram:



Graphical display



Graphical display



Always explain what error bars represent!