Statistics in biomedical research, 2^{nd} session: Classical distributions; generalities about statistical tests.

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

IGH (UMR 9002 CNRS et université de Montpellier)

November 16, 2023

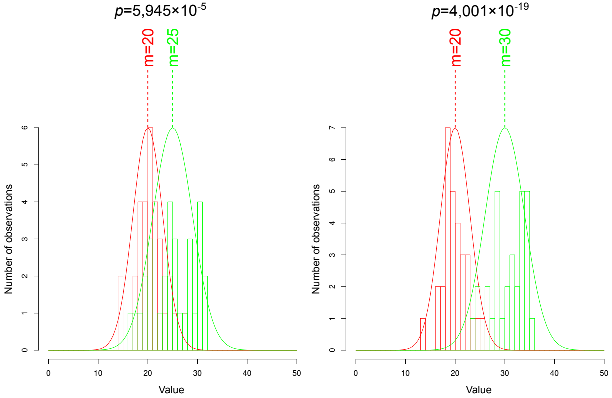
This slideshow is accessible at:
http://www.igh.cnrs.fr/equip/Seitz/en_Stats2.pdf

Factors determining t-test p-value

Everything else being equal::
difference between means $\uparrow \implies p$-value $\downarrow$

Factors determining t-test *p*-value

Everything else being equal::
difference between means $\uparrow \Longrightarrow$ *p*-value $\downarrow$



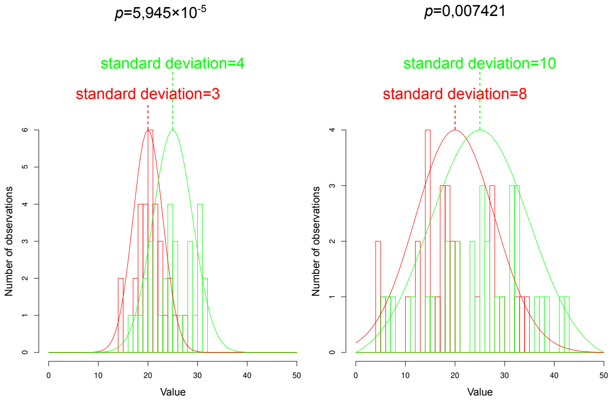
R commands used to generate these graphs: [link]

Factors determining t-test p-value

Everything else being equal::
standard deviations $\downarrow \Longrightarrow$ p-value $\downarrow$

Factors determining t-test *p*-value

Everything else being equal::
standard deviations $\downarrow \implies$ *p*-value $\downarrow$



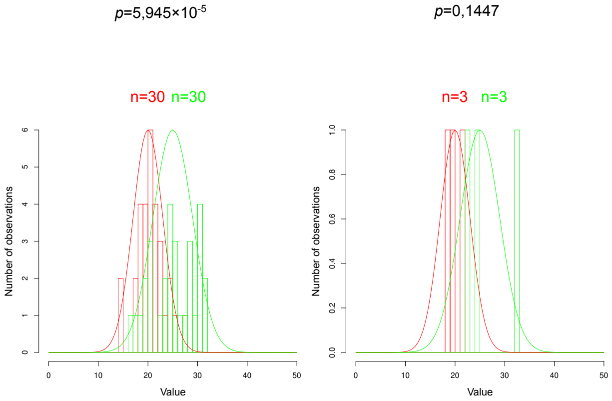
R commands used to generate these graphs: [link]

Factors determining t-test p-value

Everything else being equal::
number of replicates $\uparrow \implies p$-value $\downarrow$

Factors determining t-test p-value

Everything else being equal::
number of replicates $\uparrow \implies p$-value $\downarrow$



R commands used to generate these graphs: [link]

Factors determining t-test p-value

Reconsider the importance of a p-value !

Factors determining t-test p-value

Reconsider the importance of a p-value ! Increasing replicate number will tend to decrease it.

Factors determining t-test p-value

Reconsider the importance of a p-value ! Increasing replicate number will tend to decrease it.

Amplitude of the effect (measured by the fold-change) and its significance (measured by the p-value) are both important.

Factors determining t-test *p*-value

Reconsider the importance of a *p*-value ! Increasing replicate number will tend to decrease it.

Amplitude of the effect (measured by the fold-change) and its significance (measured by the *p*-value) are both important.

The effect is measured by its amplitude; the *p*-value measures how much confidence we can have in that fold-change.

t-test conditions of applicability

t-test conditions of applicability

1^{st} condition: each of the two sampled populations is distributed according to a normal distribution (="Gaussian distribution").

t-test conditions of applicability

1^{st} condition: each of the two sampled populations is distributed according to a normal distribution (="Gaussian distribution").

t-test conditions of applicability

1^{st} condition: each of the two sampled populations is distributed according to a normal distribution (="Gaussian distribution").

Evaluating if our samples (= our replicate datasets) can realistically be sampled from a normally distributed population.

t-test conditions of applicability

1^{st} condition: each of the two sampled populations is distributed according to a normal distribution (="Gaussian distribution").

Evaluating if our samples (= our replicate datasets) can realistically be sampled from a normally distributed population.

The Shapiro-Wilk test

t-test conditions of applicability

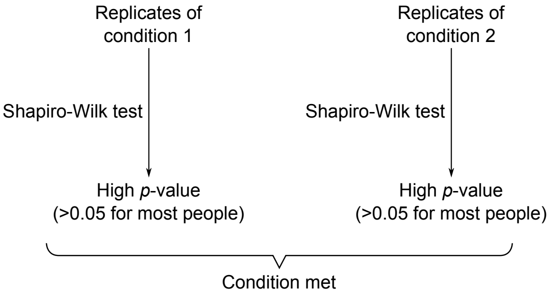
1^{st} condition: each of the two sampled populations is distributed according to a normal distribution (="Gaussian distribution").

Evaluating if our samples (= our replicate datasets) can realistically be sampled from a normally distributed population.

The Shapiro-Wilk test (p-value: probability that the replicate series could have been sampled from a normally distributed population).

t-test conditions of applicability

1^{st} condition: each of the two sampled populations is distributed according to a normal distribution (="Gaussian distribution").



t-test conditions of applicability

t-test conditions of applicability

2^{nd} condition: variances in both conditions are homogeneous ($\approx$ equal)

t-test conditions of applicability

2^{nd} condition: variances in both conditions are homogeneous ($\approx$ equal) ("homoscedasticity" condition).

t-test conditions of applicability

2^{nd} condition: variances in both conditions are homogeneous ($\approx$ equal) (“homoscedasticity” condition).

Levene’s test

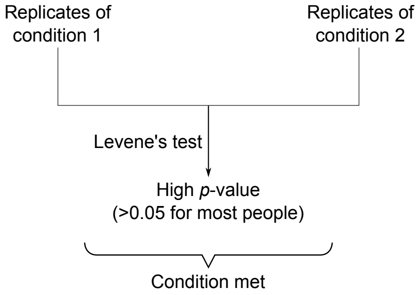
t-test conditions of applicability

2^{nd} condition: variances in both conditions are homogeneous ($\approx$ equal) ("homoscedasticity" condition).

Levene's test (p-value: probability that the two series of replicates were sampled from populations having the same variance).

t-test conditions of applicability

2^{nd} condition: variances in both conditions are homogeneous ($\approx$ equal) ("homoscedasticity" condition).



t-test conditions of applicability

What should we do if at least one of these conditions is not met ?

t-test conditions of applicability

What should we do if at least one of these conditions is not met ?

Homoscedasticity: use Welch's t-test (= "t-test with heterogeneous variances") instead of Student's t-test (= "t-test with homogeneous variances").

t-test conditions of applicability

What should we do if at least one of these conditions is not met ?

Homoscedasticity: use Welch's t-test (= "t-test with heterogeneous variances") instead of Student's t-test (= "t-test with homogeneous variances").

If Student's t-test can be used, then it is preferrable: it is more powerful.

t-test conditions of applicability

What should we do if at least one of these conditions is not met ?

Homoscedasticity: use Welch's t-test (= "t-test with heterogeneous variances") instead of Student's t-test (= "t-test with homogeneous variances").

If Student's t-test can be used, then it is preferrable: it is more powerful (detects significance with better sensitivity).

t-test conditions of applicability

What should we do if at least one of these conditions is not met ?

Homoscedasticity: use Welch's t-test (= "t-test with heterogeneous variances") instead of Student's t-test (= "t-test with homogeneous variances").

If Student's t-test can be used, then it is preferrable: it is more powerful (detects significance with better sensitivity).

Normality: use a mathematical transformation (logarithm, square root, ...; preferably if it makes physical sense).

t-test conditions of applicability

What should we do if at least one of these conditions is not met ?

Homoscedasticity: use Welch's t-test (= "t-test with heterogeneous variances") instead of Student's t-test (= "t-test with homogeneous variances").

If Student's t-test can be used, then it is preferrable: it is more powerful (detects significance with better sensitivity).

Normality: use a mathematical transformation (logarithm, square root, ...; preferably if it makes physical sense). If the log of calues appears normally distributed, then perform the t-test on the log of values.

t-test conditions of applicability

If even a mathematical transformation is not enough: use Wilcoxon's test (= Mann-Whitney test).

t-test conditions of applicability

If even a mathematical transformation is not enough: use Wilcoxon's test (= Mann-Whitney test). Analyze the rank of data points (robust to deviations to normality).

t-test conditions of applicability

If even a mathematical transformation is not enough: use Wilcoxon's test (= Mann-Whitney test). Analyze the rank of data points (robust to deviations to normality).

If data points appear normally distributed: t-test is preferrable (more powerful).

t-test conditions of applicability

If even a mathematical transformation is not enough: use Wilcoxon's test (= Mann-Whitney test). Analyze the rank of data points (robust to deviations to normality).

If data points appear normally distributed: t-test is preferrable (more powerful).

Example: comparing two series of 3 replicates each: by construction, Wilcoxon test p-value cannot be lower than 0.1 whatever the distribution of the two series !

Paired data

Paired data

Is there a particular link between each replicate of series #1 and one (and only one) replicate in series #2 ?

Paired data

Is there a particular link between each replicate of series #1 and one (and only one) replicate in series #2 ?

Example: it is the same individual, analyzed in two conditions (and each replicate is a different individual).

Paired data

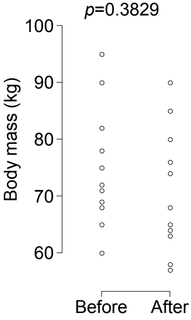
Is there a particular link between each replicate of series #1 and one (and only one) replicate in series #2 ?

Example: it is the same individual, analyzed in two conditions (and each replicate is a different individual).

Implication: the two datasets have the same number of replicates.

Paired data

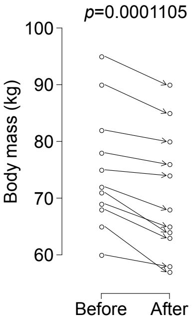
Is there a particular link between each replicate of series #1 and one (and only one) replicate in series #2 ?



R commands used to generate these graphs: [link]

Paired data

Is there a particular link between each replicate of series #1 and one (and only one) replicate in series #2 ?



R commands used to generate these graphs: [link]

“One-tailed” or “two-tailed” t-test ?

“One-tailed” or “two-tailed” t-test ?

Am I only interested in a single sense of variation ?

“One-tailed” or “two-tailed” t-test ?

Am I only interested in a single sense of variation ?

If I am looking for a difference, whatever its orientation: two-tailed (= two-sided) t-test.

"One-tailed" or "two-tailed" t-test ?

Am I only interested in a single sense of variation ?

If I am looking for a difference, whatever its orientation: two-tailed (= two-sided) t-test.

If I only care about one sense of variation (*e.g.*, are mutant mice lighter than wild-type mice ?): one-tailed (= one-sided) t-test.

"One-tailed" or "two-tailed" t-test ?

Am I only interested in a single sense of variation ?

If I am looking for a difference, whatever its orientation: two-tailed (= two-sided) t-test.

If I only care about one sense of variation (*e.g.*, are mutant mice lighter than wild-type mice ?): one-tailed (= one-sided) t-test.

Be careful ! Ignoring one of the two possible orientations is rare in research, and one-tailed t-test tends to give lower p-values.

"One-tailed" or "two-tailed" t-test ?

Am I only interested in a single sense of variation ?

If I am looking for a difference, whatever its orientation: two-tailed (= two-sided) t-test.

If I only care about one sense of variation (*e.g.*, are mutant mice lighter than wild-type mice ?): one-tailed (= one-sided) t-test.

Be careful ! Ignoring one of the two possible orientations is rare in research, and one-tailed t-test tends to give lower p-values.
$\Longrightarrow$ Possible abuse, in order to decrease a p-value in an illegitimate manner.

"One-tailed" or "two-tailed" t-test ?

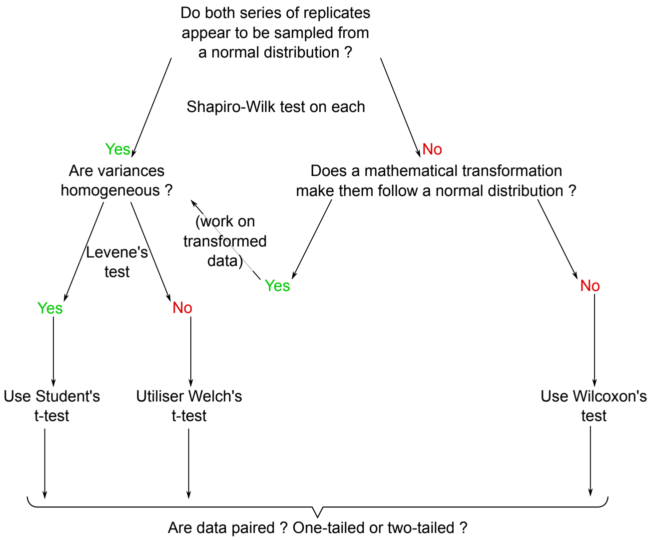
Am I only interested in a single sense of variation ?

If I am looking for a difference, whatever its orientation: two-tailed (= two-sided) t-test.

If I only care about one sense of variation (*e.g.*, are mutant mice lighter than wild-type mice ?): one-tailed (= one-sided) t-test.

Be careful ! Ignoring one of the two possible orientations is rare in research, and one-tailed t-test tends to give lower *p*-values.
$\Longrightarrow$ Possible abuse, in order to decrease a *p*-value in an illegitimate manner. If there is no imperative reason to neglect one sense of variation, then choose the two-tailed t-test.

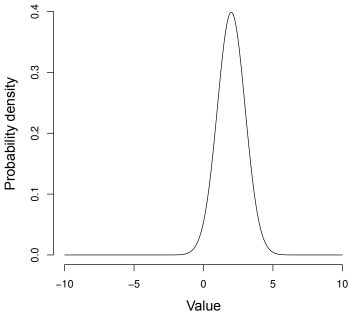
Decision tree for t-test usage



Classical distributions

A few common examples:

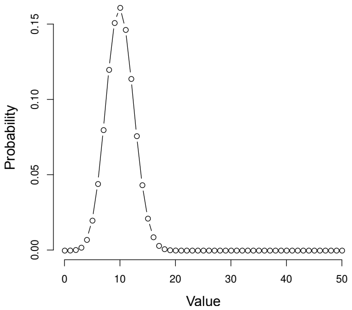
- Normal (= "Gaussian") distribution: continuous, symetrical, from $-\infty$ to $+\infty$.



Classical distributions

A few common examples:

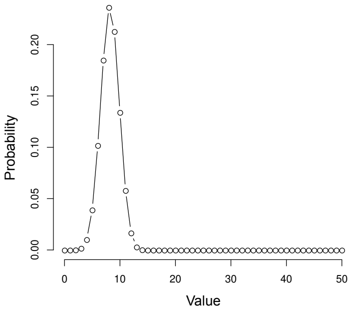
- Normal (= "Gaussian") distribution
- Binomial distribution (probability of getting a given number of successes in independent random drawing events): discontinuous, only defined for positive values.



Classical distributions

A few common examples:

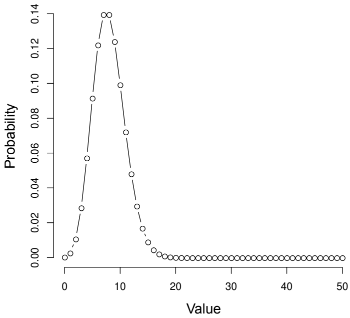
- Normal (= "Gaussian") distribution
- Binomial distribution
- Hypergeometric distribution (idem but draws without replacement).



Classical distributions

A few common examples:

- Normal (= "Gaussian") distribution
- Binomial distribution
- Hypergeometric distribution
- Poisson distribution (probability of getting k successes in a given time interval, if there are on average λ independent successes in such an interval): discontinuous, only defined for positive values.



Generalization: statistical tests

The t-test estimates the probability of a hypothesis (equality of the means of the two sampled populations).

Generalization: statistical tests

The t-test estimates the probability of a hypothesis (equality of the means of the two sampled populations).
$\longrightarrow$ the "null hypothesis".

Generalization: statistical tests

The t-test estimates the probability of a hypothesis (equality of the means of the two sampled populations).
$\longrightarrow$ the "null hypothesis".

Null hypothesis in Shapiro-Wilk's test: the sampled populations follows a normal distribution; of Levene's test: both sampled populations have the same variance.

Generalization: statistical tests

The t-test estimates the probability of a hypothesis (equality of the means of the two sampled populations).
$\longrightarrow$ the "null hypothesis".

Null hypothesis in Shapiro-Wilk's test: the sampled populations follows a normal distribution; of Levene's test: both sampled populations have the same variance.

Measurement of uncertainty: by a probability (the *p*-value) or by a confidence interval ("the difference between means of sampled populations lies in interval [... ; ...] with a probability $\geqslant 0.95$ ").

Generalization: statistical tests

The t-test estimates the probability of a hypothesis (equality of the means of the two sampled populations).
$\longrightarrow$ the “null hypothesis”.

Null hypothesis in Shapiro-Wilk’s test: the sampled populations follows a normal distribution; of Levene’s test: both sampled populations have the same variance.

Measurement of uncertainty: by a probability (the p-value) or by a confidence interval (“the difference between means of sampled populations lies in interval [... ; ...] with a probability $\geqslant 0.95$ ”).

Be careful with the wording ! Say “I did not see any significant difference”, rather than “there is no difference” (it could come out significant with more replicates).

Generalization: statistical tests

For each test: significance threshold determines the expected percentage of false positives.

Generalization: statistical tests

For each test: significance threshold determines the expected percentage of false positives. For example with the t-test with a 0.05 threshold: even samples drawn from populations with the same mean will appear significantly different $\approx$5% of the time.

Generalization: statistical tests

For each test: significance threshold determines the expected percentage of false positives. For example with the t-test with a 0.05 threshold: even samples drawn from populations with the same mean will appear significantly different $\approx$5% of the time.

Vocabulary: a false positive is also called "Type 1 error".

Generalization: statistical tests

For each test: significance threshold determines the expected percentage of false positives. For example with the t-test with a 0.05 threshold: even samples drawn from populations with the same mean will appear significantly different $\approx$5% of the time.

Vocabulary: a false positive is also called “Type 1 error”.

False negative (“Type 2 error”): does not appear significant while the sampled populations have different means.

Generalization: statistical tests

For each test: significance threshold determines the expected percentage of false positives. For example with the t-test with a 0.05 threshold: even samples drawn from populations with the same mean will appear significantly different ≈5% of the time.

Vocabulary: a false positive is also called "Type 1 error".

False negative ("Type 2 error"): does not appear significant while the sampled populations have different means. Its frequency depends on the amplitude of the difference, on each dataset's variance, on the number of replicates, and on the chosen significance threshold.