# *Comparison of two samples*

Pierre Legendre, Université de Montréal
August 2009

## 1 - Introduction

This lecture will describe how to compare two groups of observations (samples) to determine if they may possibly have been drawn ($H_0$), or not ($H_1$), from the same statistical population.

For two groups (samples) of observations, the comparison involves two aspects of the distributions: we will first compare the variances, then the means of the groups.

a) First, a *F*-test will be used to compare the variances of two independent samples (R language: *var.test*).

b) To compare the means, different tests are available:

- *t*-test to compare the means of two independent samples (parametric or permutational test) (R language: *t.test*).
- *t*-test to compare the means of two related samples (R language: *t.test*).
- Non-parametric test to compare the medians of two independent samples: Wilcoxon-Mann-Whitney *U* test (R language: *wilcox.test*). [Not described further in this lecture].
- Non-parametric tests to compare the medians of two related samples: sign test, Wilcoxon signed-ranks test (R language: *wilcox.test*). [Not described further in this lecture].

## 2 - Comparing the variances of two independent samples

In parametric statistical tests, we construct a pivotal statistic which is known to be distributed like one of the standard distribution laws when the null hypothesis ($H_0$) is true. Conditions for the application of the test may emerge during that construction.

To compare the variances of two independent samples, we construct a statistic distributed like the $F$ distribution when $H_0$ is true. Here is how.

1. From the study of the confidence interval of the variance, we know that if a sample is drawn at random from a population with <u>normal distribution</u> $N(\mu, \sigma)$, the (pivotal) *chi-square* test statistic

$$X^2 = \frac{(n-1)\, s_x^2}{\sigma^2}$$

follows a $\chi^2$ distribution with $(n - 1)$ degrees of freedom (d.f.). $X^2$ is the capital greek letter *chi-square*; it represents a statistic distributed like $\chi^2$.

2. The samples to be compared are of sizes $n_1$ and $n_2$. We thus have:

$$X_1^2 = \frac{(n_1-1)\, s_{x_1}^2}{\sigma_1^2} \quad \text{and} \quad X_2^2 = \frac{(n_2-1)\, s_{x_2}^2}{\sigma_2^2}$$

3. We also know that the test statistic

$$F = \frac{\chi_1^2 / \nu_1}{\chi_2^2 / \nu_2}$$ follows an $F$ distribution with $\nu_1$ and $\nu_2$ d.f.

4. The ratio of the two $X^2$ statistics described above, with their degrees of freedom, is a new test statistic distributed like $F$:

$$F = \frac{\dfrac{(n_1 - 1)\, s_{x_1}^2}{\sigma_1^2} \Big/ (n_1 - 1)}{\dfrac{(n_2 - 1)\, s_{x_2}^2}{\sigma_2^2} \Big/ (n_2 - 1)} = \frac{s_{x_1}^2}{s_{x_2}^2} \times \frac{\sigma_2^2}{\sigma_1^2}$$

Note: $\dfrac{s_{x_1}^2}{\sigma_1^2} \approx 1$ and $\dfrac{s_{x_2}^2}{\sigma_2^2} \approx 1$ since in each case $s_x^2$ is an unbiased estimator of the population variance, $\sigma^2$.

5. The hypothesis to be tested is H$_0$: $\sigma_1^2 = \sigma_2^2$ (equality of the pop. $\sigma^2$). If that hypothesis is true, it follows that

$$\sigma_2^2 / \sigma_1^2 = 1$$

$$s_{x_1}^2 / s_{x_2}^2 \approx 1$$

The random differences of the latter fraction, $s_{x_1}^2 / s_{x_2}^2$, from the value 1, observed for pairs of real samples, should be distributed like $F$.

Hence, the test statistic $F = \dfrac{s_{x_1}^2}{s_{x_2}^2}$ should follow an $F$ distribution with $(n_1 - 1)$ and $(n_2 - 1)$ d.f. (a) if H$_0$ is true (equality of the population variances) and (b) if the two populations from which the samples have been drawn are normal.

Else, $s^2_{x_1} / s^2_{x_2}$ may be more distant from 1 than expected from the *F* distribution.

In practice,

we subject the data to a test of the null hypothesis ($H_0$) of equality of the population variances

$$H_0\text{:}\ \sigma^2_1 = \sigma^2_2$$

The alternative hypothesis ($H_1$) can be stated in three different ways:

• One-tailed tests: $H_1 = \sigma^2_1 > \sigma^2_2$ or $\sigma^2_1 < \sigma^2_2$

• Two-tailed test: $H_1 = \sigma^2_1 \neq \sigma^2_2$

Example of a case that requires a one-tailed test of hypothesis — $H_1$: The old lab equipment #1 generates results that are more variable than the new equipment #2. If the data support $H_1$, the new equipment is better.

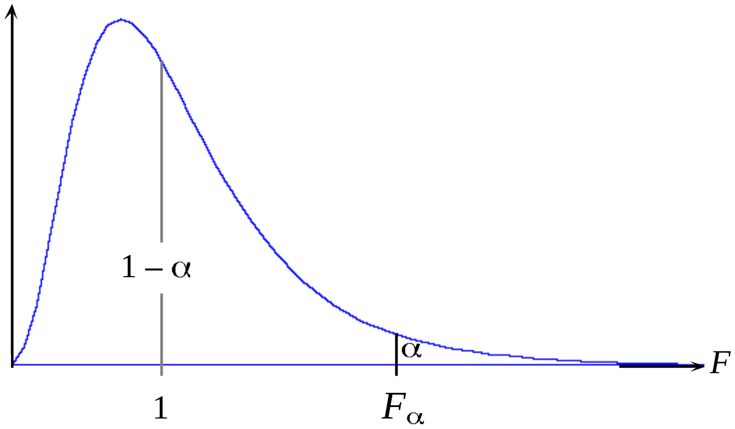
Example of a case that may call for a two-tailed test of hypothesis — Compare the distributions of diameters at breast height (*dbh*) for a dominant species in two sections of a permanent forest plot. Highly different variances would indicate different population histories for that species in the two sections, due for example to logging or forest fires.

R language

The function *var.test* performs an *F*-test to compare the variances of two samples drawn from normal populations.

First case — H$_1$: $\sigma_1^2 > \sigma_2^2$ ; one-tailed test.

If H$_0$ is true, the test statistic $F_{(\nu_1,\nu_2)} = s_{x_1}^2 / s_{x_2}^2$ has a probability (1 – α) to be smaller than the critical value $F_\alpha$.



⇒ Reasoning detail: if H$_1$ is true, the test statistic $F$ is not distributed like the $F$-distribution and its value will be markedly larger than 1.

⇒ If H$_0$ is true, the probability that $F \geq F_\alpha$ is equal to α, which is the significance level selected for the test.

Second case — H$_1$: $\sigma_1^2 < \sigma_2^2$; one-tailed test.

Same reasoning. The test statistic is $F_{(\nu_1,\nu_2)} = s_{x_1}^2 / s_{x_2}^2$. For that test, one should look up the probability in the left-hand tail ($F_{0.95\,(\nu_1,\nu_2)}$) of the $F$ distribution.

⇒ Most $F$ tables only give critical values in the right-hand tail of the distribution. So in the second case, to obtain p-values from tables, one can use the value

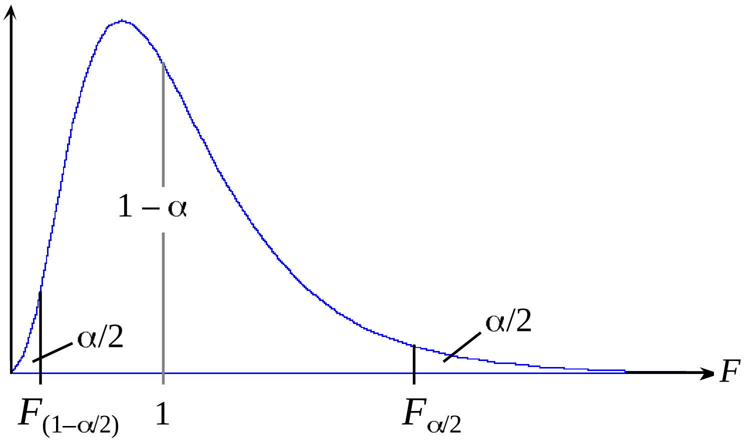
$$F_{(\nu_2,\nu_1)} = s_{x_2}^2 / s_{x_1}^2$$

computed by placing in the numerator the variance $s_x^2$ of the group which, under H$_1$, would have the largest variance $\sigma^2$. This will produce the same p-value since $F_{0.95\,(\nu_1,\nu_2)} = 1 / F_{0.05\,(\nu_2,\nu_1)}$.

Third case — $H_1$: $\sigma_1^2 \neq \sigma_2^2$ ; two-tailed test.

The largest value of $s_x^2$ is placed in the numerator. Assume that $s_{x_1}^2 > s_{x_2}^2$ ; one obtains

$$F = s_{x_1}^2 / s_{x_2}^2 > 1.$$



If $H_0$ is true, then $\Pr(F < F_{\alpha/2}) = (1 - \alpha)$, since the largest of the two $s_x^2$ values has been selected and placed in the numerator.

Application conditions

- The *F*-test is only applicable to quantitative variables.
- It requires that each population from which the observations have been drawn be normally distributed. The normality condition is introduced at step 1 (above) of the reasoning of the construction of the test-statistic.

[Test of normality in the R language: *shapiro.test*.]

Comparison of the variances of several independent samples: see the anova lecture.

## 3 - Compare the means of two independent (i.e. unrelated) samples

Question: while two samples nearly always differ in their mean values, does that indicate that their reference populations differ in their means?

The null hypothesis ($H_0$) states that the two samples come from the same statistical population, or from populations that have the same mean (or the same median in non-parametric tests).

• The parametric *t*-test is valid for normally distributed population data. The *t*-statistic is a pivotal statistic which compares the means while taking the standard deviations into account. There are different types of *t*-tests: one-tailed or two-tailed tests; tests for independent and for related samples.

• The *t*-statistic can also be tested by permutations. That form of test is valid for normal and non-normal population data. Refer to the lecture on permutation tests.

• There are also non-parametric tests for the comparison of independent or related samples; see page 1.

R language

Function *t.test* computes a parametric *t*-test of comparison of the means of two independent or related samples.

# Compare the means of two large samples

Objective: Create a pivotal test-statistic that will be distributed like Student's *t* if $H_0$ is true.

1. Theory: sum of random variables

Let Y be a sum (with + or – signs) of random variables, independent, and normally distributed: $Y = X_1 + X_2 + \ldots + X_n$
or $Y = X_1 - X_2 - \ldots + X_n$, or any other combination.

One can show that Y is normally distributed with mean (expected value)

$E(Y) = E(X_1) - E(X_2) - \ldots + E(X_n)$ [or any other combination]

whereas its variance is the sum of the original variables variances (Morrison 1976):

$$\sigma^2_{(Y)} = \sigma^2_{(X_1)} + \sigma^2_{(X_2)} + \ldots + \sigma^2_{(X_n)}$$

2. Consider now two normally distributed populations with means $\mu_1$ and $\mu_2$ et variances $\sigma_1^2$ and $\sigma_2^2$.

The study of the confidence interval of the mean shows that the random variable "mean" ($M_1$) of a random sample from population 1 is normally distributed with mean $E(M_1) = \mu_1$ (estimated by $\bar{x}_1$) et variance $\sigma^2_{M_1} = \sigma_1^2 / n_1$ (estimated by $s_1^2 / n_1$ ; $\sqrt{s_1^2 / n_1}$ is the unbiased estimator of the standard error of the mean $\mu_1$). Likewise for population 2.

Consider now a new variable *D* which measures the difference between the means of random samples 1 et 2:

$$D = M_1 - M_2 \text{ estimated by } D = \bar{x}_1 - \bar{x}_2$$

Because of point #1, we know that $D$ is normally distributed

with mean $\bar{D} = E(M_1) - E(M_2) = \mu_1 - \mu_2$

and variance $\sigma_D^2 = \sigma_{M_1}^2 + \sigma_{M_2}^2$ estimated by $s_D^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$ .

From that, it follows that the standardized variable (pivotal statistic)

$$t = \frac{D - \bar{D}}{\sigma_D} \text{ estimated by } t = \frac{D - \bar{D}}{\sqrt{s_D^2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

follows a Student $t$-distribution with $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ degrees of freedom (Sokal & Rohlf 1981, p. 227, eq. 9.4). One degree of freedom is lost when computing each of the means $\bar{x}_1$ and $\bar{x}_2$ .

Note: when the sample sizes $n_1$ and $n_2$ are very large, the $t$-distribution tends towards the standard normal distribution $z$. One should always use the $t$-distribution for testing since it is always as precise as or more precise than the $z$-distribution.

## 3. Comparison test

$$H_0: \mu_1 = \mu_2 \Rightarrow t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Because we invoke $H_0$, we can eliminate the unknown values $\mu_1$ and $\mu_2$ from the equation. The value of the $t$-statistic can now be calculated.

The alternative hypothesis ($H_1$) of the test can be formulated in different ways.

3.1. $H_1$: $\mu_1 \neq \mu_2$ (two-tailed test)

In that case, the value of the $t$ test-statistic has a probability $(1 - \alpha)$ to be found in the interval $[-t_{\alpha/2} < t < t_{\alpha/2}]$ if $H_0$ is true:

$$\Pr(-t_{\alpha/2} < t < t_{\alpha/2}) = (1 - \alpha)$$

Decision rules: $H_0$ cannot be rejected if $t$ has a value between $-t_{\alpha/2}$ and $t_{\alpha/2}$. $H_0$ is rejected only if $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$, or if the two-tailed p-value $\leq \alpha$.

3.2. $H_1$: $\mu_1 > \mu_2$ (one-tailed test), i.e. $t > 0$

$$\Pr(t < t_{\alpha}) = (1 - \alpha)$$

Decision rules: $H_0$ is rejected if $t$ is larger than or equal to $t_{\alpha}$, or if the one-tailed p-value in the right-hand tail $\leq \alpha$.

3.3. $H_1$: $\mu_1 < \mu_2$ (one-tailed test), i.e. $t < 0$

$$\Pr(-t_{\alpha} < t) = (1 - \alpha)$$

Decision rules: $H_0$ is rejected if $t$ is smaller than or equal to $-t_{\alpha}$, or if the one-tailed p-value in the left-hand tail $\leq \alpha$.

## Compare the means of two small samples

When the samples are small, each one does not provide a good estimate of the common population variance. The two small-sample variance estimates can be combined to obtain a better estimate of the population variance:

$$s^2_{wm} = \text{weighted mean of the variances}$$

Reminder: under $H_0$, the two samples have been drawn from the same statistical population. The two sample variances thus estimate the same population variance.

The development, presented in textbooks such as Sokal & Rohlf (1981, p. 226), produces the following formulas:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{wm}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{where} \quad s^2_{wm} = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

and $\nu = n_1 + n_2 - 2$. The decision rules are the same as in the large-sample case.

For calculations done by hand, this correction becomes necessary when $n_1$ or $n_2 < 30$. Its effect on the value of the *t*-statistic decreases as $n_1$ and $n_2$ increase. Since it is more precise than the formula of the previous section, this correction is always used in computer programs, except when the Welch correction is applied (next page).

## Conditions of application of the parametric *t*-test

- The variable must be quantitative.

- Samples drawn from populations with normal distributions.

- Samples drawn from populations with equal variances.

- Independence of the observations. This means that the value of an observation should have no influence on the values of another observation. That condition is violated, for example, in the case of observations **autocorrelated** through space or time.

# Violation of the condition of application of the *t*-test

## 1. Normal data, unequal variances

⇒ The *t*-test is actually simultaneously testing two different null hypotheses: the equality of means **and** the equality of variances. This confusion of the hypotheses is called *the Behrens-Fisher problem*.

To be able to only compare means, the ordinary *t*-test assumes that the two independent samples are drawn from populations with equal variances, $\sigma_1^2 = \sigma_2^2$.

That condition has to be tested by an *F*-test of equality of the variances before proceeding with a *t*-test of equality of the means.[*]

What should we do if the *F*-test has rejected the hypothesis of equality of the variances? One can use a *t*-test with Welch (1936) correction. In that test, the *t*-statistic is still estimated by the formula

$$t_{Wc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

That $t_{Wc}$ statistic does not follow a Student *t*-distribution if $H_0$ is true. However, a correct p-value can be obtained from the *t*-distribution with degrees of freedom computed as follows:

[*] In the same way, in analysis of variance (anova), a test of homogeneity of the variances will have to be done before proceeding with anova.

$$\nu_{\text{Welch}} = \frac{\left[ (s_1^2/n_1) + (s_2^2/n_2) \right]^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}$$

That value (Satterthwaite correction) is rounded down to the largest integer not greater than the computed value (Zar, 1999).

If the variances are equal, that formula gives exactly $\nu = n - 2$ if $n_1 = n_2$. Examples:

- $s_1 = 5$, $s_2 = 5$, $n_1 = 10$, $n_2 = 10$ ⇒ $\nu = 18$, $\nu_{\text{Welch}} = 18$
- $s_1 = 1$, $s_2 = 9$, $n_1 = 10$, $n_2 = 10$ ⇒ $\nu = 18$, $\nu_{\text{Welch}} = 9.22$ rounded to 9

**Simulations** carried out by Legendre and Borcard show the following for data with normal distributions (see document: Welch_correction.pdf).

- When the variances are equal or nearly so, the type I error of the parametric and permutational *t*-tests are correct (Figs. 6a, 6e).
- When the variances are unequal and $n_1$ and $n_2$ are small (< 50), the type I error of the parametric and permutation *t*-tests are too high. Both forms of test are invalid under these conditions (Fig. 6a).
- The type I error of the *t*-test with Welch correction is correct when the data are drawn from normal distributions with unequal variances. In that situation, the Welch-corrected *t*-test should be used instead of the standard parametric *t*-test (Figs. 6a, 6e).
- When $n_1$ and $n_2$ are large (≥ 50, Fig. 6e), the type I errors of the parametric and permutational *t*-tests are correct, even when the variances are unequal. It is not necessary to use the Welch-corrected *t*-test in that case. The three forms of test are equally valid.

### 2. Non-normal data, *n* not small, equal variances

For data drawn from highly non-normal data whose variances are equal or nearly so, the *t*-test is too conservative, meaning that it does not reject $H_0$ often enough when compared to the significance level $\alpha$ when $H_0$ is true (Next page, Figure 1). The *t*-test remains valid*, but power to detect a difference in means is reduced when a difference is present in data.

- One can try normalizing data by an appropriate transformation.

- The best overall solution is to test the *t*-statistic by permutation: the type I error of the permutational *t*-test is always correct (next page, Fig. 1) and its power is much higher than that of the parametric *t*-test.

### 3. Non-normal data, small *n*

When the data are drawn from non-normal populations and the groups are small ($n_1$ and $n_2 \le 30$), the permutational *t*-test can be used if the variances are equal. Otherwise, use the non-parametric Wilcoxon-Mann-Whitney *U* test.

### 4. Semi-quantitative variables

For a semi-quantitative variable, the group means do not exist. Use the Wilcoxon-Mann-Whitney *U* test to compare the medians of the groups.

## Alternative computation methods

A two-group *t*-test is equivalent to a *t*-test of the correlation coefficient, or a *t*-test of the linear regression coefficient, between the response variable and a dummy (binary) variable representing the two groups.

* A statistical test is *valid* if its type I error rate is not higher than the significance level $\alpha$ for any value of $\alpha$ (Edgington, 1995).

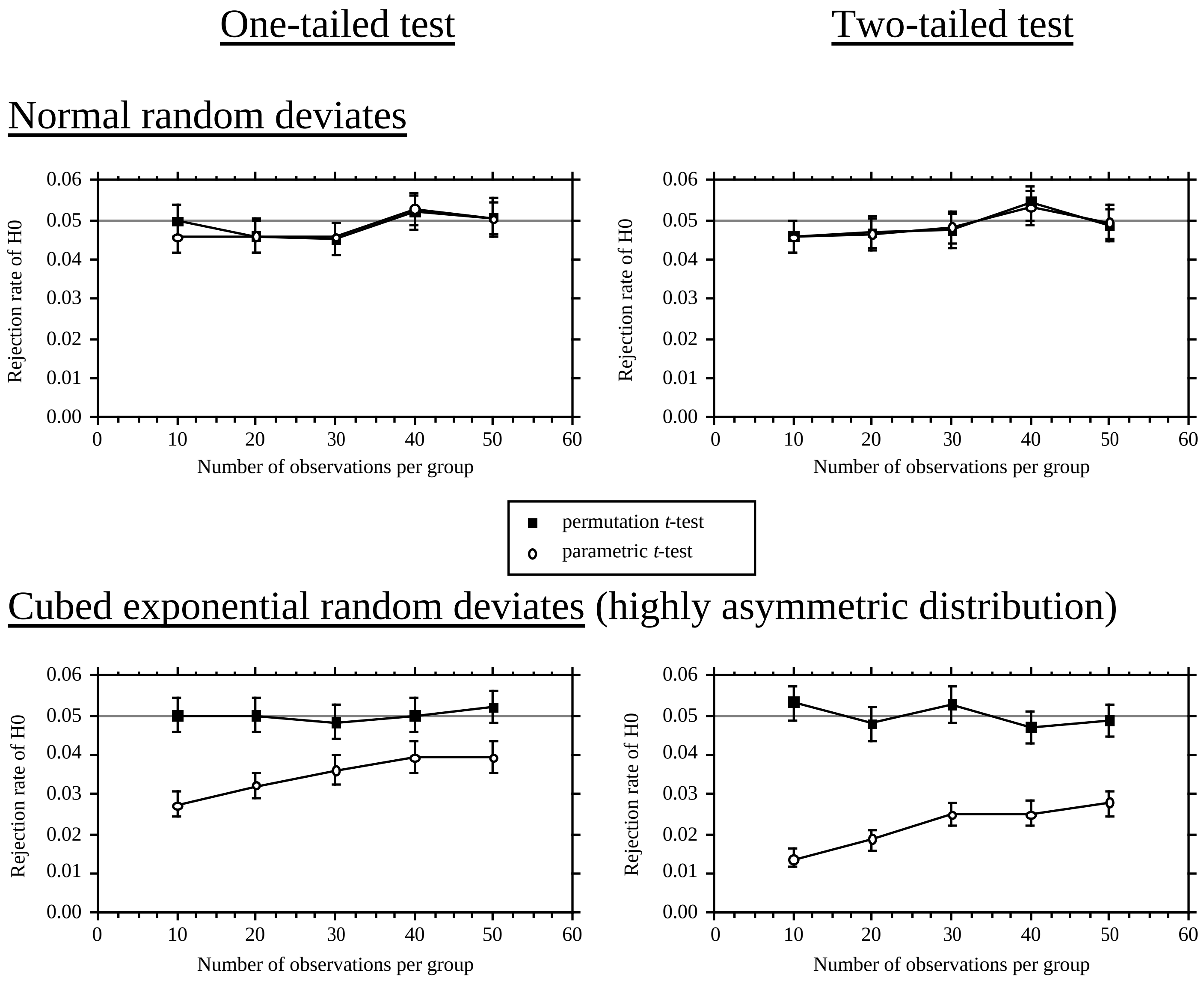


**Figure 1** – Mean and 95% confidence interval of the rejection rate of $H_0$ (ordinate) during tests comparing the means of two groups of simulated observations, as a function of the number of observations per group (abscissa).

*Method*: parametric *t*-tests (open circles) and permutation *t*-tests (black squares), $\alpha = 0.05$. The null hypothesis was always true: the two groups of observations were drawn from the same statistical population in all cases. Each point (circle or square) is the rejection rate after the analysis of 10 000 simulated data sets. 99 permutations done for each permutation test.

Observations

• For normally distributed data, the two tests are equivalent.

• For non-normally distributed data, the type I error of the parametric *t*-test is too low (< 0.05), resulting in a loss of power when a difference in means has to be detected. The type I error of the permutation test is always correct.

## 4 - Compare the means of two related samples

For related samples, the information we possess about the related observations should be used in the test to increase power. Examples of related observations are: the same individuals have been observed or measured before and after a treatment; data have been obtained at the same altitudes (forests), or the same depths (lakes), at two sites. In the *t*-test for unrelated samples presented in the previous section, all observations “before” form a group and all observations “after” form the other group. That group does not use the information of the sampling design about the relatedness of the observations.

In the comparison of the means of related samples, we analyse the differences observed for each of *n* pairs of observations *i*: $d_i = (x_{i1} - x_{i2})$.

• The sample of *n* values $d_i$ has a mean $\bar{d}$ and a standard deviation $s_d$.

• Consider several sets of related samples: the random variable $\bar{d}$ has a mean

$$\mu_{\bar{d}} = \mu_1 - \mu_2$$

(the mean of the differences is equal to the difference of the means) and a standard deviation

$$s_{\bar{d}} = s_d / \sqrt{n}$$

• The test-statistic will be the standardized variable

$$t = (\bar{d} - \mu_{\bar{d}}) / s_{\bar{d}}$$

which follows a Student *t* distribution with $(n - 1)$ degrees of freedom. Note that *n* is the number of pairs of observations.

• From the data, we can compute $\bar{d}$ and $s_{\bar{d}}$, but not $\mu_{\bar{d}}$. We can, however, state the null hypothesis $H_0$: $\mu_1 = \mu_2$ which eliminates the portion of the equation that cannot be estimated from the data. The *test-statistic* used in the test of $H_0$ is:

$$t = \bar{d} / s_{\bar{d}}$$

This *t*-statistic follows a Student *t* distribution with $(n - 1)$ degrees of freedom (d.f.).

<u>Conditions of application</u>: the variable must be quantitative; the distribution of $d_i$ values must be normal in the statistical population; the observations must be independent (i.e. not autocorrelated).

<u>R language</u>: *t.test(x, y, paired=TRUE)*

<u>Example</u> (Zar, 1999, p. 162): do the fore (front) and hind legs of deers have the same lengths?

| Deer<br>*i* | Length of<br>hind legs (cm) | Length of<br>fore legs (cm) | Difference<br>$d_i$ (cm) |
|---|---|---|---|
| 1 | 142 | 138 | 4 |
| 2 | 140 | 136 | 4 |
| 3 | 144 | 147 | –3 |
| 4 | 144 | 139 | 5 |
| 5 | 142 | 143 | –1 |
| 6 | 146 | 141 | 5 |
| 7 | 149 | 143 | 6 |
| 8 | 150 | 145 | 5 |
| 9 | 142 | 136 | 6 |
| 10 | 148 | 146 | 2 |

1. Naive solution: *t*-test for two independent samples

$n = 20$; $\bar{x}_1 - \bar{x}_2 = 3.3$ cm; $\nu = n_1 + n_2 - 2 = 18$; $t = 1.97802$; p = 0.0634 for a two-tailed test.

2. Better solution: *t*-test for two related samples

$n = 10$; $\bar{d} = 3.3$ cm (positive sign); $\nu = n - 1 = 9$; $t = 3.41379$; p = 0.0077 for a two-tailed test.

Biological conclusion: the hind legs are longer than the fore legs, at least in that deer population.

Statistical conclusion: when the data are related, the test for related samples has more power than the test for independent samples.

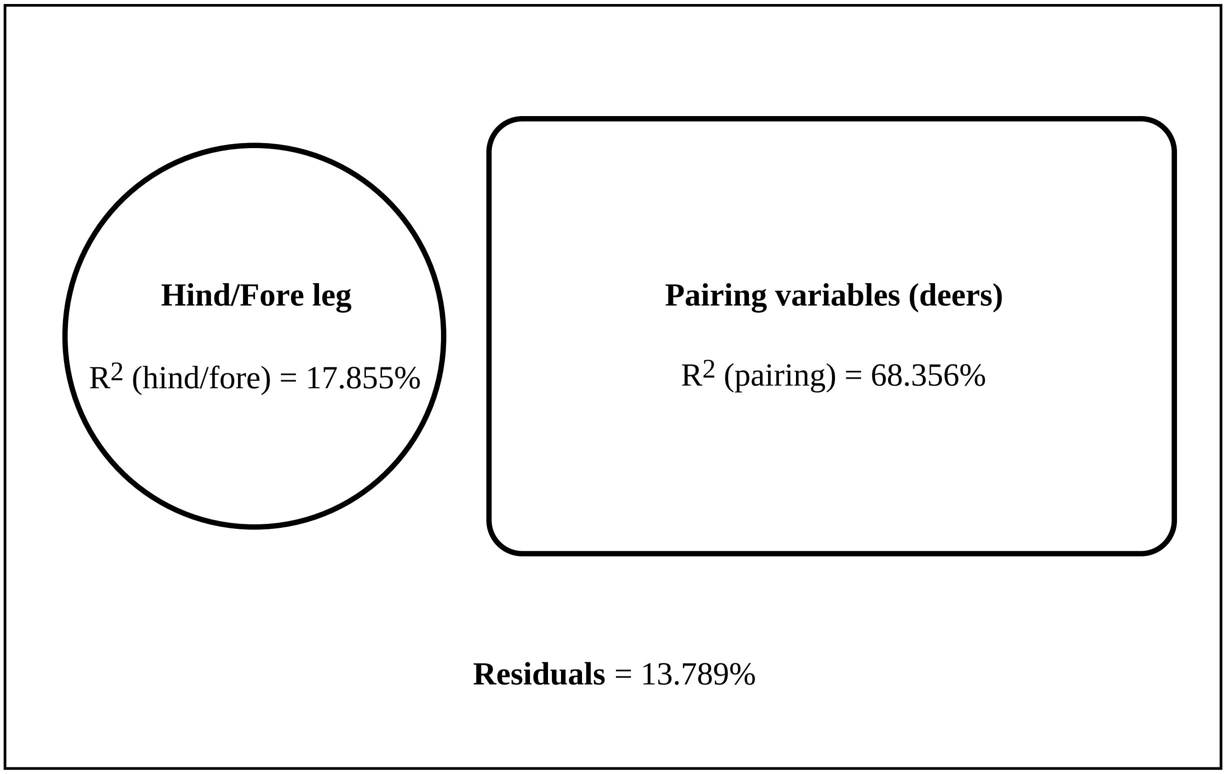
## Alternative computation methods

A two-group *t*-test for related samples is equivalent to a *t*-test of the linear regression coefficient between the response variable and a dummy variable representing the two groups, in the presence of binary or Helmert covariables representing the related observations; see Practicals.

## Advanced topic

Why is the test for paired samples more powerful than the test for independent samples? This can be understood in the context of variation partitioning. Using regressions, one can compute the *R*-squares corresponding to the different factors, as well as the residual *R*-square:

- $R^2$ of the regression of Length on Hind.Fore = 0.1786
- $R^2$ of the regression of Length on the pairing variables = 0.6836
- $R^2$ of regression of Length on Hind.Fore and pairing variables = 0.8621
- So, the residual $R^2$ is 1 – 0.8621 = 0.1379.


**Hind/Fore leg**
R$^2$ (hind/fore) = 17.855%
**Pairing variables (deers)**
R$^2$ (pairing) = 68.356%
**Residuals** = 13.789%

Put these values in the equation of the *F*-statistic in multiple regression:

$$F = \frac{R^2 \text{explained by the factor of interest} / m}{\text{residual} R^2 / (n - 1 - m - q)}$$

*n* = number of observations, *m* = number of explanatory variables (1 in this example), *q* = number of covariables.

1. *t*-test for independent samples ($q = 0$):

$$F = \frac{17.855/1}{(13.789 + 68.356)/18} = 3.91257;\ t = \sqrt{F} = 1.97802;\ \text{p} = 0.0634.$$

2. *t*-test for related samples ($q = 9$):

$$F = \frac{17.855/1}{13.789/9} = 11.65398;\ t = \sqrt{F} = 3.41379;\ \text{p} = 0.0077.$$