



2. Tests of statistical significance

Pierre Legendre

Département de sciences biologiques

Université de Montréal

<http://www.NumericalEcology.com/>



1. Tests of significance

A test of significance is a method of inference used by researchers to determine if a **statistic** computed from a sample of data is compatible with the expected value of the **parameter**, corresponding to the null hypothesis to be tested, in the statistical population from which the sample has been drawn.

The method was proposed by Ronald Fisher in his 1925 book *Statistical Methods for Research Workers* and in later papers, and perfected by later workers, including Jerzy Neyman and Egon Pearson.

The statistic computed from the sample is called a *test statistic* or *test criterion*.

The value of the parameter in the statistical population is stated in the null hypothesis of the test (H_0).

H_0 is determined by some theory or hypothesis of the researcher. It translates the biological “no effect” or “no difference” hypothesis into numerical terms.

H_0 often negates the existence of the phenomenon that the scientist is hoping to evidence.



2. Elements of a test of significance

Elements of a classical statistical test in the Fisherian approach –

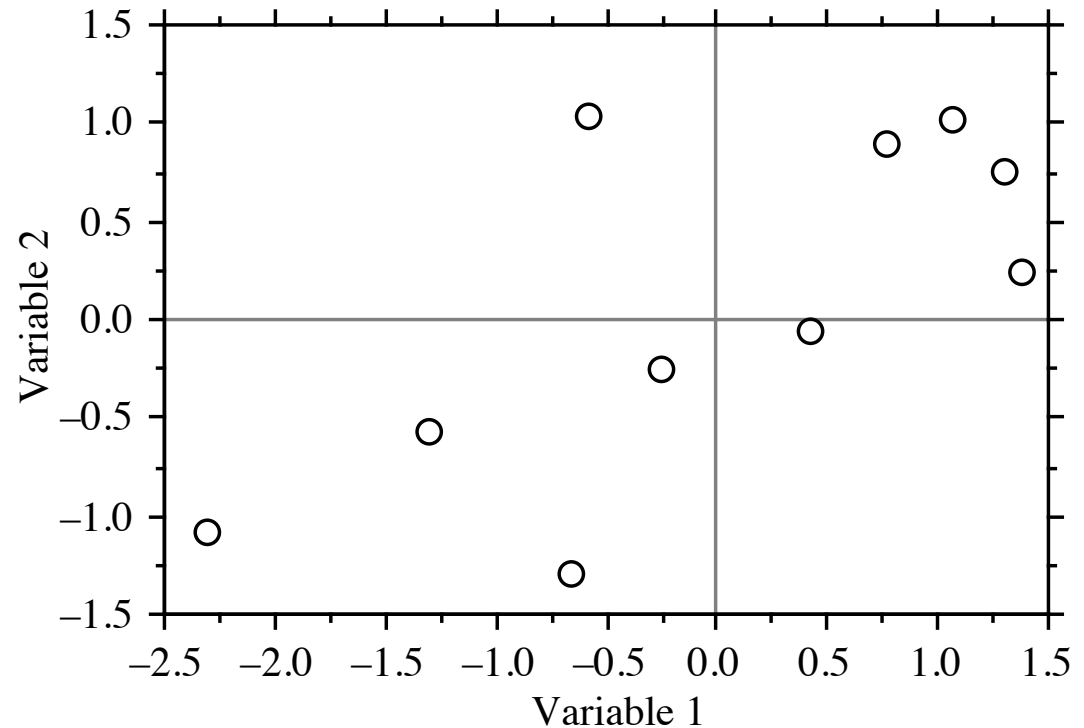
1. A question in the application field
2. Data
3. One or several statistics that can be computed
4. Hypotheses: H_0 , H_1 (3 types)
5. A significance level α
6. A testing method –
 - **parametric**: refer to published tables (or an R function) to find the p-value, under some distribution assumption
 - **permutational**: when the assumption (e.g. normality) is not met
7. Statistical decision, with respect to the significance level

Example – Test of a correlation coefficient

1. Question: are these two variables correlated?

2. Data:

X_1	X_2
-2.31	-1.08
1.06	1.03
0.76	0.90
1.38	0.24
-0.26	-0.24
1.29	0.76
-1.31	-0.57
0.41	-0.05
-0.67	-1.28
-0.58	1.04



3. Statistics: we could use –

- Pearson's $r = 0.70156$
- Spearman's $r = 0.60000$
- Kendall's $\tau = 0.42222$

Example – Test of a correlation coefficient

4. Statistical hypotheses –

Null hypothesis $H_0 : \rho = 0$

Alternative hypothesis $H_1 :$

Alternative-1: $\rho \neq 0$ Two-tailed test

Alternative-2: $\rho < 0$ One-tailed test in the left tail

Alternative-3: $\rho > 0$ One-tailed test in the right tail

The choice of a 2-tailed or 1-tailed test depends on the ecological question.

5. Significance level α : usually one of the values $\{0.05, 0.01, 0.001\}$

We will use $\alpha = 0.05$.

Example – Test of a correlation coefficient

6. Testing method

6.1. Parametric, if the assumption of bivariate normality is met

$$r = 0.70156, t = 2.78456, n = 10, v = (n - 2) = 8$$

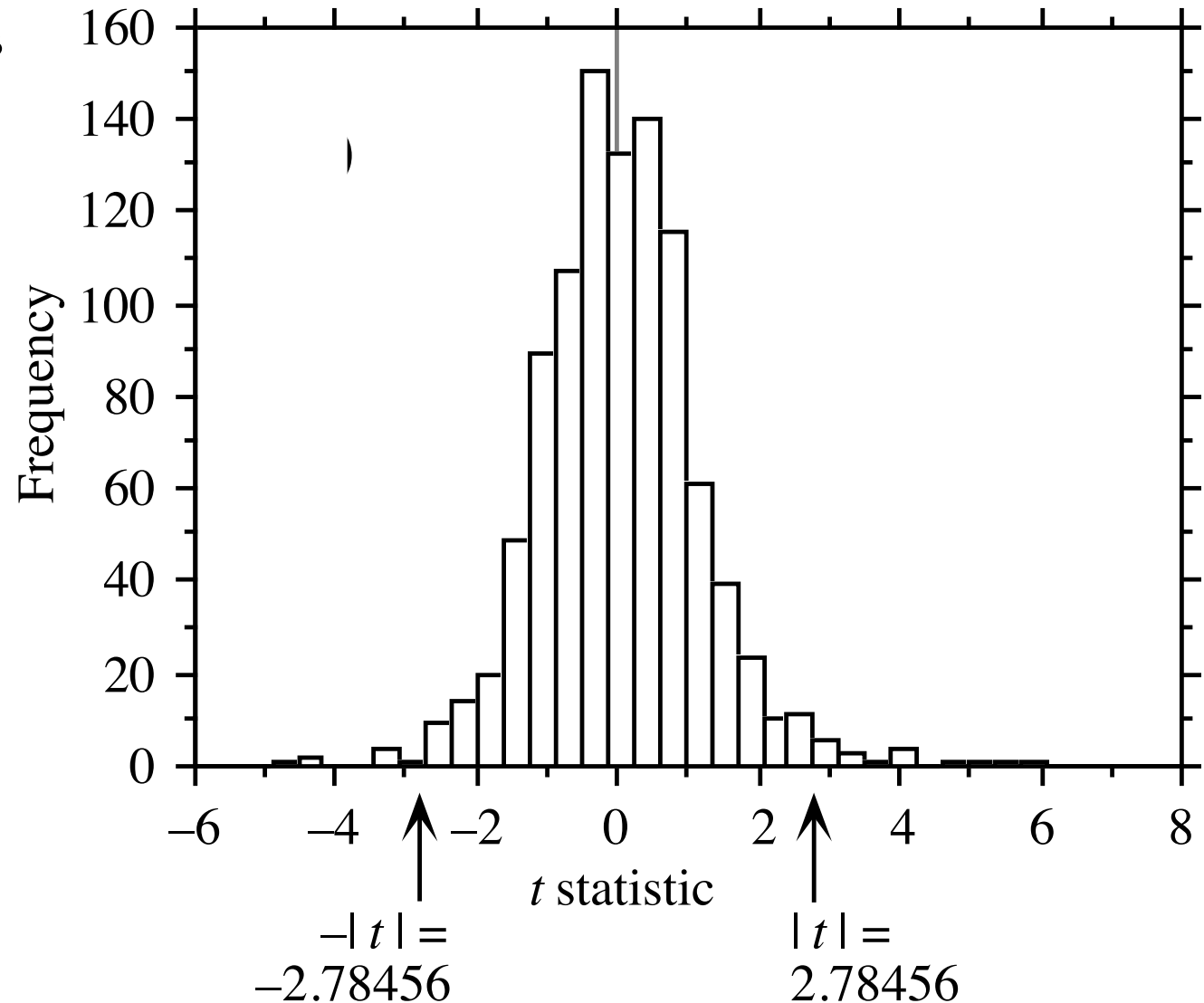
$$\text{prob (one-tailed)} = 0.0119, \text{prob (two-tailed)} = 0.0238$$

6.2. Permutational: we can permute x_1 , x_2 , or both variables.

	r	t
Reference (true) value	0.70156	2.78456
Permute x_2 , #1	0.11796	0.33598
Permute x_2 , #2	-0.23193	-0.67439
Permute x_2 , #3	0.30179	0.89532
...		
Permute x_2 , #10	-0.41093	-1.27491
...		
Permute x_2 , #100	0.07330	0.20789

The permutational method can always be used, including when the assumption of bivariate normality is met.

Assemble results
in a histogram:



Summarize the results of the permutations in a small table – [t.ref](#)

	$t^* < - t $	$t^* = - t $	$- t < t^* < t $	$t^* = t $	$t^* > t $
<i>Statistic t</i>	8	0	974	1 [†]	17

[†] This count corresponds to the observed t value that was added to the reference distribution.

Compute permutational p-values –

One-tailed test, right tail:

What proportion of the [t.perm](#) is larger than or equal to [t.ref](#) ?

$$p = (1 + 17)/1000 = 18/1000 = 0.018$$

One-tailed test, left tail:

What proportion of the [t.perm](#) is smaller than or equal to [t.ref](#) ?

$$p = (8 + 0 + 974 + 1)/1000 = 983/1000 = 0.983$$

Two-tailed test:

What proportion is equal to or more extreme than [t.ref](#) in both tails?

$$p = (8 + 0 + 1 + 17)/1000 = 26/1000 = 0.026$$

Compare the parametric and permutational p-values –

	Parametric	Permutational
Two-tailed test:	0.0238	0.026
One-tailed test, right tail:	0.0119	0.018
One-tailed test, left tail:	0.9881	0.943

7. Statistical decision –

At the $\alpha = 0.05$ significance level, the 2-tailed test and the 1-tailed test in the right tail are both significant.

The results show that a 1-tailed test is more powerful than a 2-tailed test, because the 1-tailed p-value is always smaller than the corresponding 2-tailed p-value.

(Power of a statistical test: capacity to reject H_0 when H_0 is false.)



3. Comparison of testing methods

Which method produces the best estimate of the p-value?

The permutational method is the one producing unbiased estimates of p-values in all cases.

Parametric tests of significance mimic the results of permutation tests when the data meet the distribution assumption, e.g. normality. For correlation analysis, the data must be bivariate normal. Advantage: parametric tests do not involve heavy computations.

With small sample sizes, the assumption of bivariate normality rarely holds.

Make sure that a sufficiently large number of permutations has been run to ensure adequate numerical accuracy of the answer.

How accurate are the permutational estimates of p-values? Do they vary a lot from one run to another?

What is the effect of increasing n. permutations on p-values?

p-values in two-tailed tests (example data), each repeated five times

99 perm \Rightarrow 0.02, 0.03, 0.04, 0.07, 0.03

999 perm \Rightarrow 0.032, 0.027, 0.034, 0.027, 0.026

9999 perm \Rightarrow 0.0250, 0.0285, 0.0260, 0.0282, 0.0248

99999 perm \Rightarrow 0.02555, 0.02585, 0.02568, 0.02504, 0.02564

The n. permutations determines the n. of decimal places of p-values.

One-tailed tests, left tail

99 perm \Rightarrow 0.98

999 perm \Rightarrow 0.979

9999 perm \Rightarrow 0.9868

99999 perm \Rightarrow 0.98578

One-tailed tests, right tail

99 perm \Rightarrow 0.02

999 perm \Rightarrow 0.015

9999 perm \Rightarrow 0.0140

99999 perm \Rightarrow 0.01447

=> If several tests are run together in a study, use a correction for multiple testing (Bonferroni, Holm, Hochberg, etc.). Applying the correction will increase the p-values.

With a Bonferroni correction for example, the corrected p-value is
$$p * (\text{number of simultaneous tests})$$

For some of the p-values to remain significant after a correction for multiple tests, it is often necessary to compute p-values with many decimal places in the first place. This is done by increasing the number of permutations.

Example – A p-value of 0.01 obtained from a permutation test with 99 permutations will become 0.20 after a Bonferroni correction involving 20 simultaneous tests.

Running the test with 999 or 9999 permutations will produce a p-value with more decimals. It could be lower than 0.010 and could well be significant at the 0.05 level after Bonferroni correction.



4. Types of permutation tests

- The example developed above was an example of an unrestricted, sampled permutation test, where we sampled at random from the [most often very large] set of the possible permutations.
- Complete permutation test. Possible only for very small samples, because the number of permutations of n values is $n!$

Examples: $7! = 5040$, $8! = 40320$, $9! = 362880$, $10! = 3628800$

- Restricted permutation tests. Example: in anova for 2 or several crossed factors, to test the significance of one of the factors, the values are permuted separately in each class of the other factor.
 - Restricted permutations in a circle for time series, or on a torus for spatial data, allows one to preserve [most of] the autocorrelation among the observations.
 - Permutation of the residuals of some model instead of permutation of the original data. In multiple regression and canonical analysis.
-

=> Different hypotheses call for different permutation methods.

Examples –

- Correlation (example presented above):

Permute the elements of one of the two vectors at random.

- Simple linear regression: as for correlation, above.
- Simple t -test of difference between two sample means:

Permute the values at random in the joint vector containing the 2 groups, then redivide the values into 2 groups.

- Paired t -test between the means of related observations:

Place the 2 vectors side by side. Permute at random the two related observations about each object.



5. Effect of lack of normality

A test of significance has a **correct type I error rate** at (say) significance level $\alpha = 5\%$ if it produces significant results in about 5% of the cases when the null hypothesis is true. The result is called an “error” because the statistical conclusion is incorrect in these cases.

Parametric tests of significance produce correct type I error rates if the data meet the distribution assumptions of the test (the *normality assumption* in most cases).

For data that do not meet the assumptions of the test,

- tests with an error rate lower than alpha have lower power than tests that have correct error rates.
- tests with an error rate higher than alpha are invalid. Do not use them.

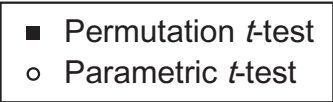
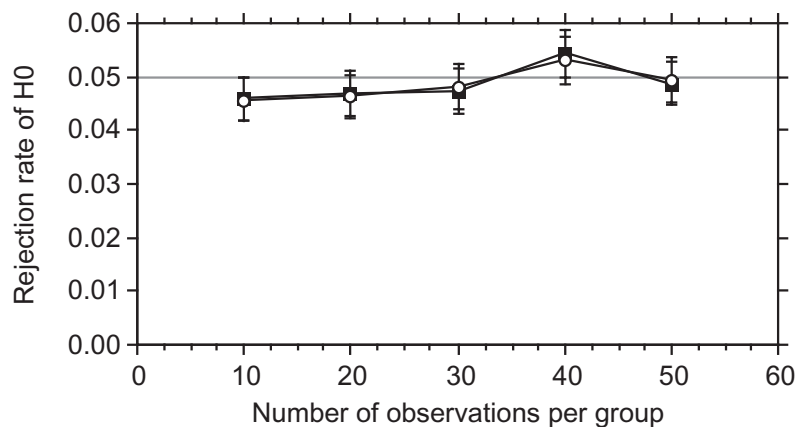
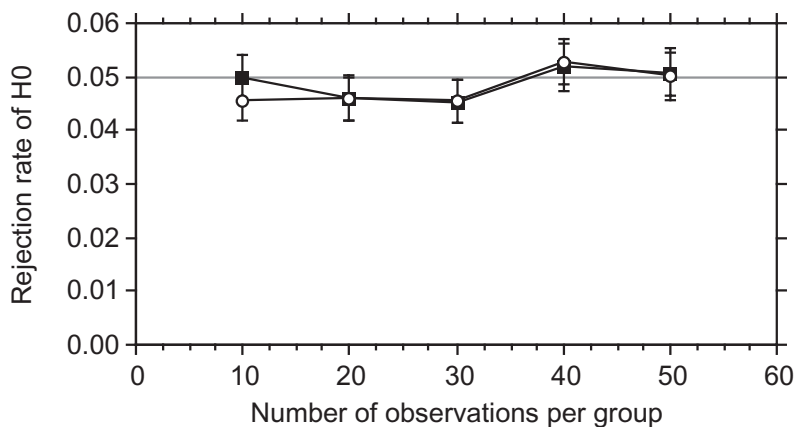
Simulation studies show that, in many cases, permutation tests have correct empirical error rates.

1. Effect of lack of normality on the t -test of difference between 2 sample means.

One-tailed tests

Two-tailed tests

Random normal data

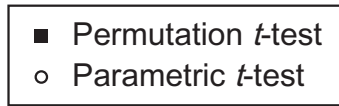
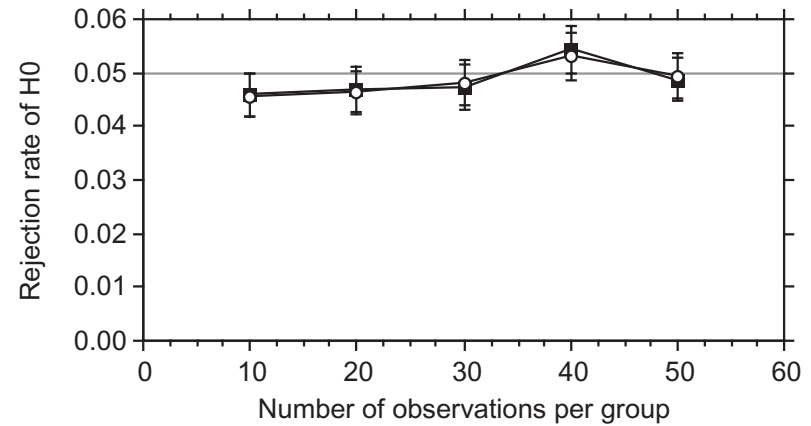
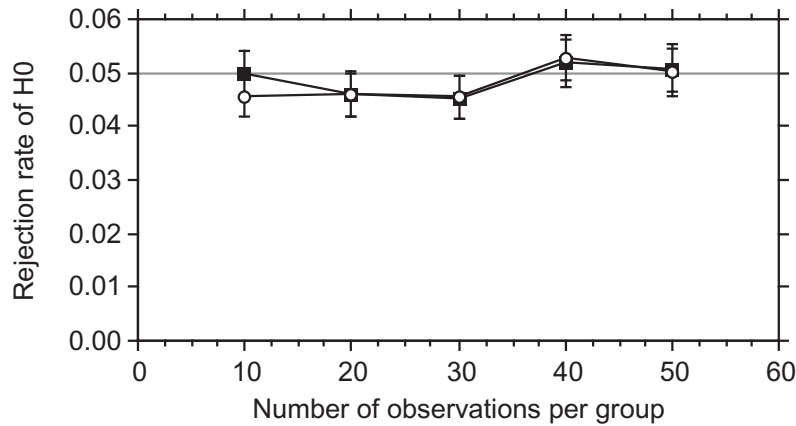


1. Effect of lack of normality on the t -test of difference between 2 sample means.

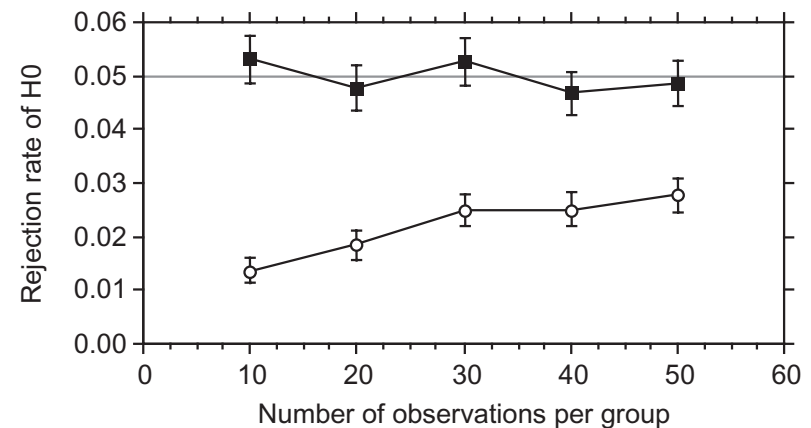
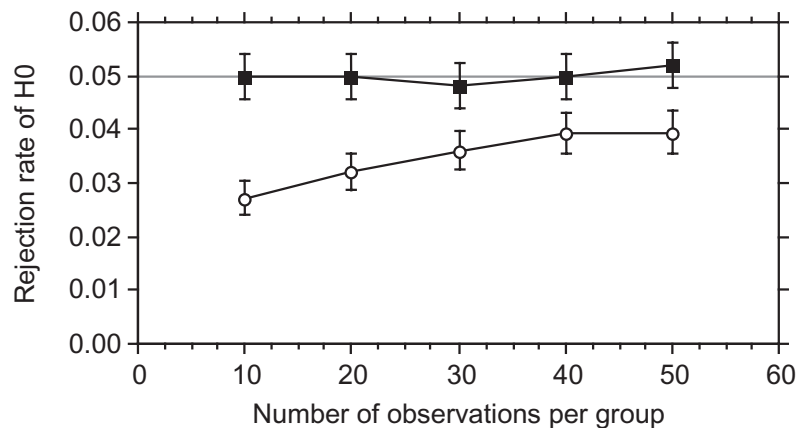
One-tailed tests

Two-tailed tests

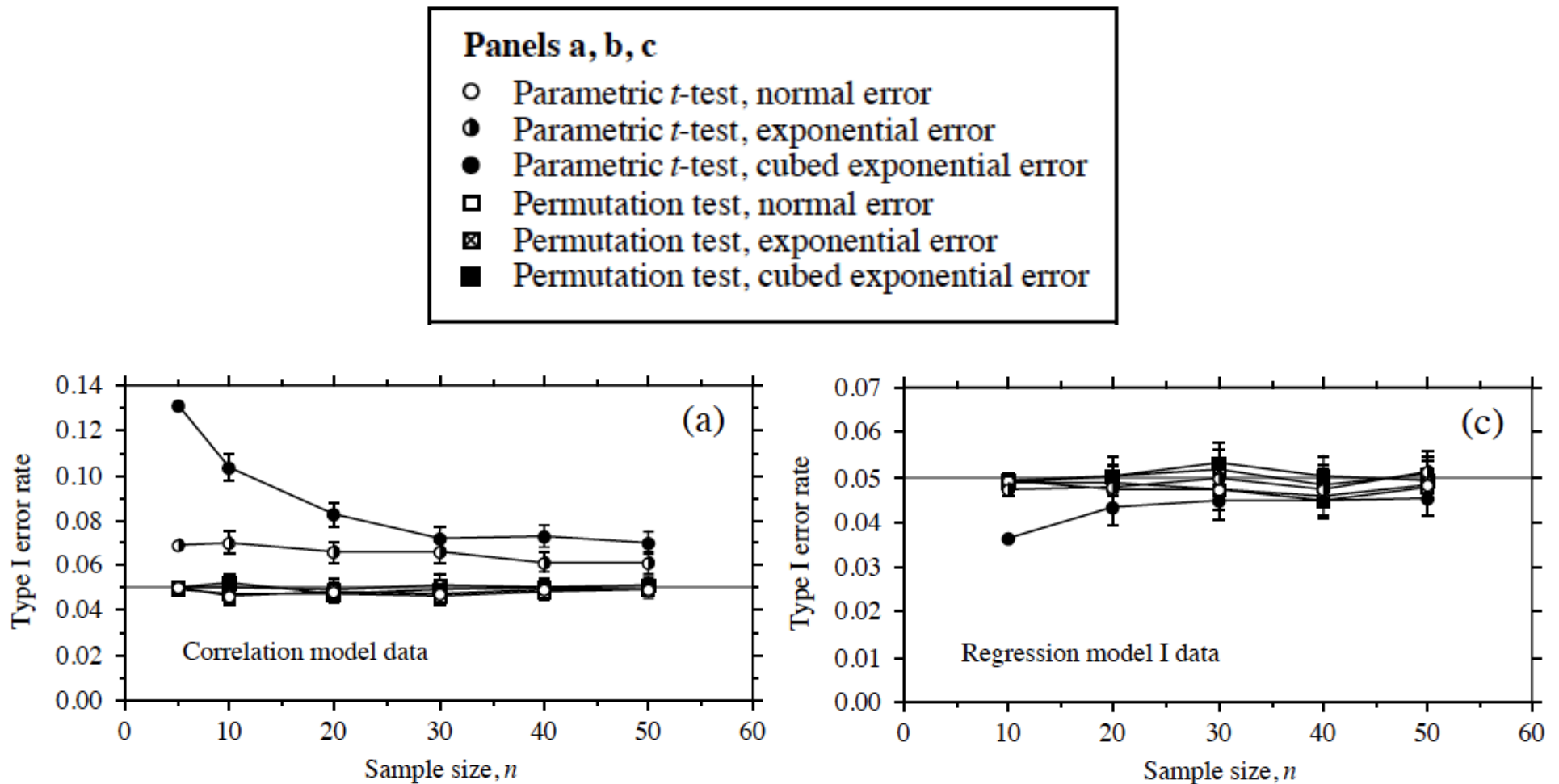
Random normal data



Random exponential data, cubed (highly asymmetric distribution)



2. Effect of lack of normality on the t -tests of correlation coefficients.



Rejection rates after 10000 simulations, one-tailed tests, 999 permutations for each test, $\alpha = 0.05$. Error bars: 95% confidence intervals. (a) Two random variables. (b) x fixed, y random. Panels from Legendre (2000), Fig. 1 .



6. Demo functions, correlation analysis

The *corPerm.R* file contains –

- function `corPerm1()` – This function computes a two-tailed permutation test of a Pearson correlation coefficient between two data vectors. Test statistic is r .
- function `corPerm2()` – This function computes a two-tailed permutation test of a Pearson correlation coeff. between two data vectors. The test statistic is t .
- function `corPerm3()` – This function computes a permutation test of a Pearson correlation coefficient between two data vectors. The test statistic is t .

One-tailed test in the left tail: $\text{tail} = -1$

One-tailed test in the right tail: $\text{tail} = 1$

Two-tailed test: $\text{tail} = 2$



7. References

- Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. *Journal of Statistical Computation and Simulation* 67: 37-73.
- Legendre, P. & L. Legendre. 2012. *Numerical ecology, 3rd English edition*. Elsevier Science BV, Amsterdam. xvi + 990 pp. ISBN-13: 978-0444538680.



End of section