

# Power Analysis

(Stephane Pauquet)

---

## CONTENTS

- 1/ Introduction
  - 2/ Statistical significance
  - 3/ Factors affecting Power
  - 4/ Power of test
  - 5/ The effect size index
  - 6/ Failures of the assumptions
  - 7/ Significance testing
  - 8/ The use of tables for significance testing
  - 9/ Implications of ignoring Power
  - 10/ References
  - 11/ Sample Power Table
- 

## 1/ Introduction

The power of a statistical test is the probability that it will yield significant results (Cohen, 1969). Scientists strongly wish for statistical significance, a concept which surprisingly is not well understood. Because scientists have a crude concept of statistical significance and an even more primitive concept of statistical power, tests are often conducted under conditions where the hypothesis under scrutiny (the null hypothesis) has very low chances of being rejected, even if it is actually wrong.

Power analysis, which can be performed *a priori* or *a posteriori* to the collection of data, is used to assess the likelihood inherent in the design of a test to reject null hypothesis. A test found to have low power *a priori* should lead to a new, if not completely different experimental design, or to changes in the constraints such as the significance criterion. A test found *a posteriori* to have low power should either convince the scientist to perform the experiment again with a larger sample size, or to at least to reflect on what inferences, if any, can be drawn from this experiment.

However, those types of Power analyses should be performed respecting the same basic principles as for any other statistical test, notably that of defining the precise objectives of the investigation prior to any experiment. For example, testing for Power a

*posteriori* is irrelevant when using the effect size (see below) observed in the data rather than postulating the effect size that one wanted to detect.

---

## 2/ Statistical significance

Since samples only approximate population characteristics, a significance criterion  $\alpha$  for the observed values is set.  $\alpha$  serves as a "standard of proof" that the phenomenon under study exists, or equivalently, "standard of disproof" for the null hypothesis stating that the phenomenon does not exist. From there, we define:

**a type I error (=  $\alpha$ )** as the probability of rejecting the null hypothesis ( $H_0$ ) when in fact it is true.  $\alpha$  is traditionally set very low.

**a type II error (=  $\beta$ )** as the probability of failing to reject  $H_0$  when in fact the null hypothesis is false.  $\beta$  is consequently usually high and often unknown.

This leads us to the definition of *Power* (=  $1 - \beta$ ) as being the probability of correctly rejecting  $H_0$ , the null hypothesis.

---

## 3/ Factors affecting Power

**Power increases with:**

- sample size ( $n$ ): sample reliability always depends upon its size (the smaller the sample, the larger the error). Thus, it is intuitively obvious that increases in sample size will increase statistical power.
- effect size (ES, or  $d$  when standardized). The degree to which a specified alternative hypothesis deviates from the null hypothesis.
- higher  $\alpha$  level: the lower its level, the lower the power;
- observational variability: the lower the variance and standard deviation, the greater the power;

...and the degree to which the data meet the assumptions of the statistical method applied.

**The directionality of the significance criterion also bears on the power:**

Two-tailed tests have less power than one-tailed tests, provided that the sample result is in the direction predicted:

$$\text{Power}[\text{one-tailed test } -\alpha \text{ level}] = \text{Power}[\text{two-tailed test } - 2(\alpha \text{ level})]$$

For example, if a one-tailed test is conducted at an  $\alpha = 0.05$  level and yields results in the right direction, then it will have equal power as if the test had been conducted as a two-tailed test for an  $\alpha$  of 0.10.

---

## 4/ Power of test

**Power is a function  $\alpha$  of,  $n$  and ES:**

- If these are determined, Power is obtained by simple computation, or by using Power tables;
- Power Tables can be used to determine the power of a test, as well as to yield the value of any of these 4 parameters, knowing the 3 others.

**Assumptions:**

The Power tables are designed to yield values for the  $t$ -test (difference between the means of two *independent samples* of equal size drawn from a *normal population* having *equal variances*). Thus the primary assumptions are:

- $n_1 = n_2$
- $\sigma_1 = \sigma_2$

But we will see later that the analysis is robust to certain violations of these assumptions.

**Application:**

By rearranging the appropriate specific form of the following equation:  $\alpha$

$$\text{Power} = \text{some function of } (ES, n, \text{sigma and } \alpha)$$

...it is possible to solve it for any one of its terms:

- "detectable" effect size
  - variance
  - sample size
  - probability of type I error
- 

## 5/ The effect size index

The effect size ES is the degree of departure of the alternative hypothesis from the null hypothesis. It is indexed to obtain a "pure number",  $d$ , free of the original measurement unit (standardization procedure), by dividing the difference of the observed results by their standard deviation:

$$d = \frac{m1 - m2}{\sigma}$$

- for a one-tailed test:

$$d = \frac{|m1 - m2|}{\sigma}$$

- for a two-tailed test:

( $m1$ ,  $m2$  expressed in measurement units)

This has led to a categorization between "small", "medium" and "large"  $d$  values, for instance (Cohen, 1977):

- 0.2 : ES is small relative to uncontrollable extraneous variables ("noise"). This is assumed when the phenomena under study are not under good experimental or measurement control, or both;

- 0.5 : conceived as an effect size "large enough to be visible to the naked eye" (experimentally perceptible), and;

- 0.8 : large effect size (as an example, it is represented by the mean IQ difference estimated between holders of the Ph.D. degree and typical college freshmen...).

Note that only previous practice and knowledge in a particular field can serve for the setting of these otherwise arbitrary values.

---

## 6/ Failures of the assumptions

**case 1:**  $n_1 \neq n_2$ ;  $\sigma_1 = \sigma_2$

The power tables will yield useful approximate values, if the harmonic

mean,  $n'$  is computed instead of the arithmetic mean : 
$$n' = \frac{2n_1n_2}{n_1 + n_2}$$

**case 2:**  $n_1 = n_2$ ;  $\sigma_1 \neq \sigma_2$

The tables are still valid using  $\sigma' = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}$  instead of the standard deviation.

(unless there is a big difference between  $s_1$  and  $s_2$ ,  $s'$  will not differ greatly from their arithmetic mean);

**case 3:** one sample of  $n$  observations

The  $t$ -test can also be used testing  $H_0: m = c$  instead of an alternative hypothesis,  $c$

being a specified value relevant to some theory under consideration:

This requires the computation of  $d_3' = \frac{m - c}{\sigma}$  (conceptually, no change) and still

another calculation ( $d = d_3' \sqrt{2}$ ) in order to consult the tables, because  $c$  is a hypothetical parameter, thus without sampling error, whereas in a two - sample test, each mean contributes to sampling error.

The relevant  $t$ -test will be based on  $(n - 1)$  degrees of freedom (vs.  $[2(n - 1)]$  df in a two - sample test). This approach is problematic only in the case of very small samples.

**case 4:** situation where the data set consists of one sample of  $n$  differences between paired observations (matched into  $n$  pairs  $X, Y$ )

$$Z = X - Y; mZ = mX - mY; d_z' = \frac{mZ}{\sigma_Z} \rightarrow d_4' = \frac{mX - mY}{\sigma}$$

$$d = \frac{d_4'}{\sqrt{1-r}}$$

## 7/ Significance testing

...Is an appraisal of the research results: the second column of the power tables yields the value of the significance criterion for the parameter under study (effect size  $d$  or correlation coefficient  $r$ ).

A "**significance at the  $\alpha$  level**" is attributed when the observed effect size ( $d$ ) or correlation coefficient ( $r$ ) equals or exceeds this criterion.

Advantage: the significance decision can be made without computation, in a "quick check" of the significance of results.

## 8/ The use of tables for significance testing

Provision has been made in the Power tables (see Below) to facilitate significance testing:

**Cases 1 and 2** (see above): Calculate the standard mean difference for the sample:  $ds$

$$ds = \frac{\bar{X}_A - \bar{X}_B}{S}$$

(with  $\bar{X}_A + \bar{X}_B$  being the sample means and  $S$  the *pooled within sample estimate* of the population standard deviation):

$$S = \sqrt{\frac{\sum (X_A - \bar{X}_A)^2 + \sum (X_B - \bar{X}_B)^2}{n_A + n_B - 2}}$$

(samples need to be independent, but can be unequal)

$$ds = t \sqrt{\frac{n_A + n_B}{n_A n_B}}$$

Note:  $ds^*$  is related to the  $t$ -statistic by:

(\*in case 1,  $ds$  simplifies to:  $ds = t \sqrt{\frac{2}{n}}$ )

The value of  $ds$  necessary for significance is called  $dc$ , i.e. the criterion value of  $ds$ . In entering the tables, the value of  $n$  to be used is the harmonic mean of  $na$  and  $nb$ , as defined above.

**Case 3:** Same as above. Provided that the sample sizes are approximately equal, the validity of the t-test is hardly affected (thus, valid under "nonextreme" conditions).

**Case 4** (one sample of  $n$  differences between paired observations:  $X - Y = Z$ ): Some transformations are required:

Compute:  $d's = \frac{\bar{X} - c}{s}$  ( $d's$  indicating a one - sample test) and use  $d'c$  instead of  $dc$ :

$$d'c = dc \sqrt{\frac{1}{2}}, \text{ or } 0.707dc.$$

## 9/ Implications of ignoring Power

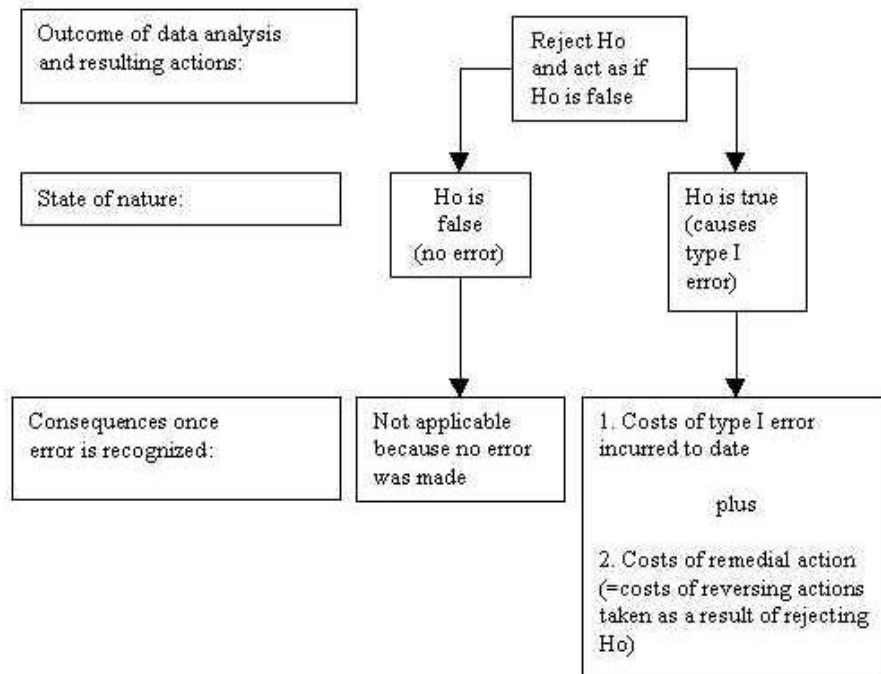
Low-power impact assessment experiments can generate costly type II errors, resulting in the implementation of inappropriate actions (e.g. depletion of aquatic resources, Peterman, 1990);

Low power often results in inefficient and/or unadapted experiments, and could thus lead to miss opportunities to increase understanding of the processes under study, and;

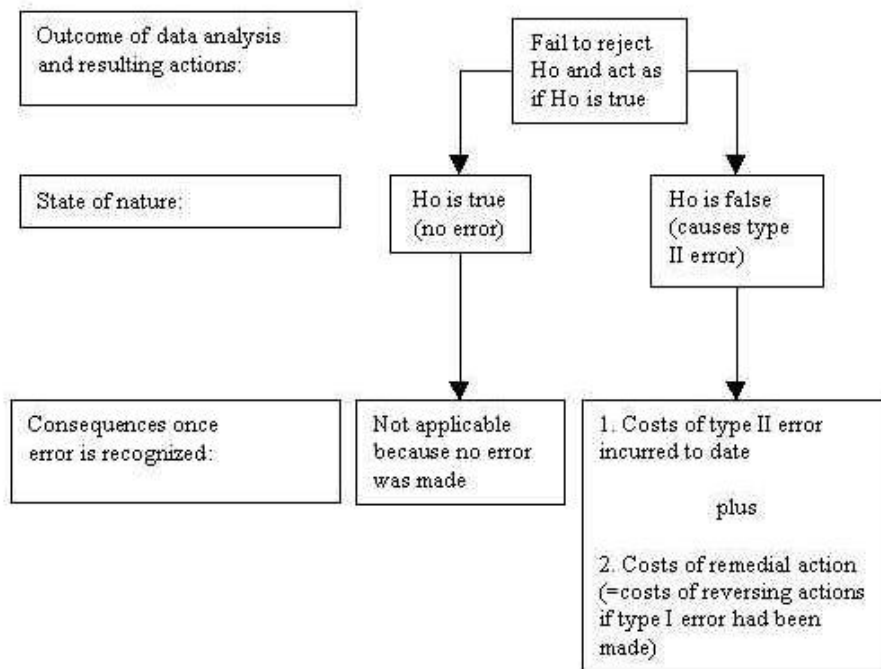
The cost of type II errors can exceed that of type I errors (whereas traditional significance standards usually implicitly assume the opposite).

Taking action as the result of a statistical analysis that fails or not to reject the null hypothesis can lead to the two following sequences of events (Peterman, 1990):

## 1) Hypothetical sequence of events following the rejection of $H_0$



## 2) Hypothetical sequence of events following the failure to reject $H_0$



The total cost of the decision path that reflects acting if the  $H_0$  were true will likely be larger than the path that assumes the  $H_0$  to be false (compare Figs. (1) and (2)).

## 10/ References

Cohen, J. 1977. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press: New-York.

Peterman, R. M. 1990. *Statistical Power Analysis can Improve Fisheries Research and Management*. Canadian Journal of Aquatic Sciences 47:2-15.

Phillips, P. C. 1998. *Designing Experiments to maximize the Power of detecting correlations*. Evolution 52 (1):251-255.

## 11/ Power Table

**Power of t test of  $m_1 = m_2$  at  $\alpha = .01$**

d												
n	dc	.10	.20	.30	.40	.50	.60	.70	.80	1.00	1.20	1.40
8	1.31	02	03	04	05	08	12	14	19	30	43	57
9	1.22	02	03	04	06	09	13	16	22	35	49	63
10	1.14	02	03	04	07	10	14	18	25	40	55	70
11	1.08	02	03	05	07	11	15	21	28	45	61	76
12	1.02	02	03	05	08	12	17	23	31	49	66	81
13	.98	02	03	05	08	13	19	26	34	53	71	85
14	.94	02	03	06	09	14	20	28	38	57	75	88
15	.90	02	04	06	10	15	22	31	41	61	79	90
16	.87	02	04	06	10	16	24	34	44	64	82	92
17	.84	02	04	07	11	18	26	36	47	68	85	94
18	.81	02	04	07	12	19	27	38	49	71	87	95

19	.79	02	04	07	13	20	29	40	51	74	89	96
20	.77	02	04	08	13	21	30	42	54	76	91	97
21	.75	02	05	08	14	22	32	44	56	79	93	98
22	.73	02	05	08	15	23	34	46	59	81	94	98
23	.71	02	05	09	15	24	36	48	61	83	95	99
24	.70	02	05	09	16	25	37	50	64	85	95	99
25	.68	02	05	10	17	27	39	53	66	87	96	99
26	.67	02	05	10	17	28	41	55	68	89	97	99
27	.65	02	05	10	18	29	42	57	70	90	97	*
28	.64	02	05	11	19	30	44	59	72	91	98	
29	.63	02	06	11	19	31	46	60	74	92	98	
30	.62	03	06	11	20	32	48	62	75	93	99	
31	.61	03	06	12	21	34	50	64	77	94	99	
32	.60	03	06	12	22	35	51	66	79	94	99	
33	.59	03	06	13	22	36	52	67	80	95	99	
34	.58	03	06	13	23	37	53	69	81	95	99	
35	.57	03	07	13	24	38	55	70	83	96	*	
36	.56	03	07	14	25	40	56	72	84	96		
37	.55	03	07	14	26	41	58	73	85	97		
38	.55	03	07	15	26	42	60	75	86	97		
39	.54	03	07	15	27	43	61	76	87	98		
40	.53	03	07	15	28	45	62	78	88	98		
42	.52	03	08	16	30	47	64	80	90	98		
44	.51	03	08	17	31	49	67	82	91	99		
46	.49	03	08	18	33	51	69	83	93	99		
48	.48	03	08	19	34	53	71	85	94	99		

\* Power values below this point are greater than .995.

(Source: Cohen, 1977)