

Power Analysis using nQuery Advisor

Biostatistics
January 2003

Course Outline

1. Introduction to significance testing
2. Power analysis strategy

Examples

3. Two-sample t -test (comparing means)
4. χ^2 - test (comparing proportions)
5. Correlation test

1. Introduction: Significance Testing (*t*-test)

- Null hypothesis : population means are equal
- If sample means are different, this could due to
 - bias (design)
 - chance
 - null hypothesis being false
- Believe null hypothesis unless difference not likely to be due to bias or chance

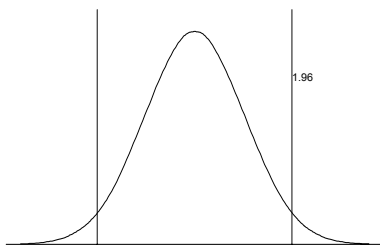
P-Value

- p is the probability of a difference in sample means as large as that observed (or larger) occurring by chance (due to sampling variation) if the null hypothesis is true
- If p is small (less than the level of significance α), we can argue that “such a difference would be unlikely to occur by chance if the null hypothesis were true”
 \Rightarrow we reject the null hypothesis.

Alternative Hypothesis

- Null Hypothesis: $\mu_1 - \mu_2 = 0$ i.e. $\mu_1 = \mu_2$
- Three possible Alternative Hypotheses:
 1. $|\mu_1 - \mu_2| > 0$ i.e. $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$ **two-sided**
 2. $\mu_1 - \mu_2 > 0$ i.e. $\mu_1 > \mu_2$ **one-sided**
 3. $\mu_2 - \mu_1 > 0$ i.e. $\mu_1 < \mu_2$ **one-sided**
- p-values are:
 - For 1., $\Pr(|m_1 - m_2| > 0)$ **two-tailed**
 - For 2., $\Pr(m_1 - m_2 > 0)$ **one-tailed**

One-tailed versus two-tailed



- The p-value for a one-sided test is half that for a two-sided test (for a t -test).
 - Planning to do a one-sided test is only justified if we will interpret a large difference in means in the opposite direction as definitely being due to chance.
- ⇒ Not legitimate if we are willing to accept “surprising” results!

Examples of t-tests

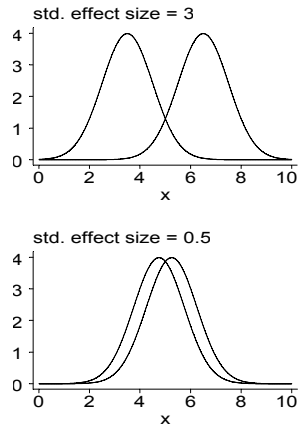
- Two samples of size n , $s=1$, $p=0.05$
 - $n=10$: mean difference = 1.00
 - $n=30$: mean difference = 0.53
 - $n=100$: mean difference = 0.28

⇒ smaller effect size needed to reject null hypothesis if n larger

Power

- The power of a study is the probability that it will give a significant result
- Studies are generally considered worthwhile if the power is at least 80%
 - funding
 - ethical approval

Effect Size



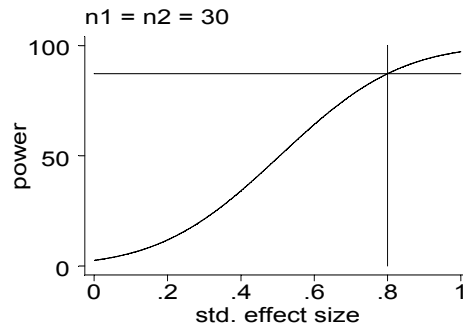
- The power increases with the effect size
- For a t-test, the standardised effect size is $\delta = (\mu_1 - \mu_2) / \sigma$
- δ is the number of standard deviations between the means.

Types of error

		truth	
		true	false
decision	accept	true negative $1 - \alpha$	false negative type II error β
	reject	false positive type I error α	true positive Power $1 - \beta$

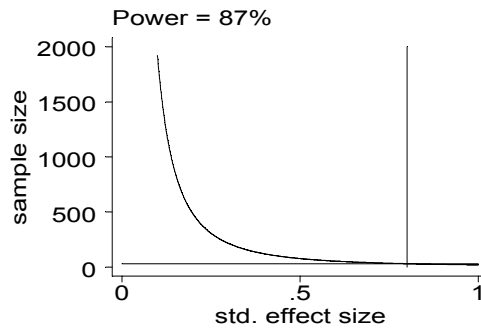
Power and effect size

- For a given sample size, the power increases with effect size.



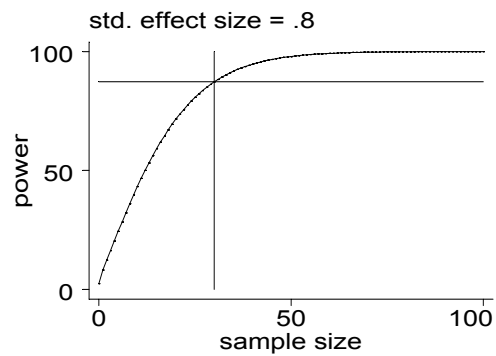
Sample size and effect size

- For a given power, the sample size decreases with effect size.



Power and sample size

- For a given effect size, the power increases with sample size.



Relationship between parameters

	Power $1-\beta$	Effect size δ	Sample size n	Signifi- cance level α
$1-\beta$	•	↑	↑	↑
δ		•	↓	↓
n			•	↓
α				•

The sample size required depends on

- Minimum size of effect considered to be clinically important
- Probability with which we wish to be able to detect the effect
 - Power
- Acceptable probability of a false positive
 - Level of significance

2. Power analysis strategy

What is a power analysis?

A power analysis or sample size calculation can answer questions such as

- Which sample size is required for the study to detect a certain effect size with pre-specified chance?
- Given a certain sample size what was the chance of finding a specific effect?
- What is the smallest effect size that could be detected with the present sample size with a given chance?

Why perform a power analysis?

Practical, ethical and statistical issues are all involved in determining the sample size needed for a study.

- A study with too many subjects may be deemed wasteful of resources and unethical through the unnecessary involvement of extra people. For example, in a treatment trial patients would be subjected to the inferior treatment!
- Studies with samples that are too small will be unlikely to detect clinically important effects. This is also wasteful of resources and unethical.

Steps involved in planning a study

Answering the question ‘what sample size do I need?’ requires five steps.

- Specification of the test parameters,
- selection of a statistical significance test,
- specification of the effect size,
- power requirement,
- sensitivity analysis.

Step 1- Specification of the test parameters

The test problem needs to be quantified.

- Definition of the null hypothesis and alternative hypothesis to be tested in the study, specifically is it a one-sided or two-sided hypothesis?
- Which significance level is required, in other words, under what threshold does the probability of a *type-I-error* have to fall?

Step 2 - Selection of a statistical test

An appropriate statistical test for the problem needs to be identified. This requires knowledge of the

- Study design, in particular it is important to determine whether study units are related, for example because the same subject is observed at several times or twins are observed.
- Response variable - how is the outcome of the study measured? The scale of the outcome measure (interval, ordinal, nominal, binary) affects the set of suitable tests (parametric or non-parametric test, number of categories).

Step 3 - Specification of the effect size

Crucial for power analysis is the size of the effect on the outcome measure that the analysis aims to detect. When testing for differences between two groups this involves

- supplying the minimum difference in population means between the two groups that the study should be able to detect
- and specifying the standard deviation of the outcome measure within the populations
- or specifying only the standardised effect size; here the ratio of the minimum difference divided by the standard deviation

How to choose the minimum difference?

There are two concepts frequently used to determine the minimum difference in population means between the groups.

- The researcher has to specify the minimum *scientifically relevant (clinically significant) difference*. This judgement may involve cost-benefit considerations. Norms can be useful for this purpose.
- The study aims to detect a difference found previously by other researchers. Thus the value is obtained from the scientific literature.

How to specify the standard deviation?

Specifying the standard deviation is often difficult because it cannot be estimated until the study is completed. Approaches used are

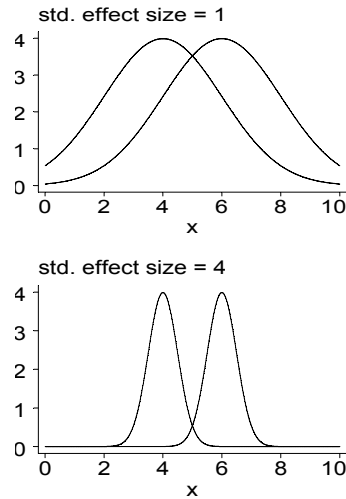
- employment of an estimate from a previous study if previous results are representative for the population(s) under study;
- estimation of the standard deviation from a pilot study;

Interpretation of standardised effect size

The magnitude of the standardised effect size may be interpreted (Cohen, 1977)

- **small if $\delta=0.2$** : difference in mean height between girls aged 15 and 16,
- **medium if $\delta=0.5$** : difference in mean height between girls aged 14 and 18,
- **large if $\delta=0.8$** : difference in mean height between girls aged 12 and 20.

Clinical significance and standardised effect size



In some applications it may be the relative difference between groups that is relevant when judging clinical significance. The expected difference between two groups depends on the population means alone. However, for example, the probability that an individual improves at all after being given a drug treatment relative to a control treatment is a function of both the difference in means and the standard deviation.

Step 4 - Power requirement

The *power* is the probability that a test will reject the null hypothesis when it is false. It is one minus the probability of the *Type-II-error*, failure to reject the null hypothesis when it is false. A high value is required for this 'chance of finding an effect'.

Investigators typically request study powers between 80% and 95%.

Step 5 - Sensitivity analysis

Having specified the test, the effect size and the required power the minimal sample size to achieve these specifications can be determined. However further *sensitivity analysis* is often useful in order to investigate how the required sample size changes when

- the power requirement is increased (reduced),
- the effect size is varied within a plausible range (e.g. for a range of standard deviation estimates),
- another statistical test is used (e.g. a non-parametric alternative).

Variations in power calculations

The steps outlined above result in determination of the sample size, given the effect size and the required power. In a similar manner, steps 3 or 4 can be substituted with the specification of the sample size to determine the minimum detectable effect size or the achieved power respectively.

Basic nQuery commands

To start nQuery advisor, double click its icon. A help system is available from the initial menu. Each sample size table screen provides three different aids to understanding the program (for more details see the manual which also contains tutorials)

- guide cards for each row of each sample size table and many side tables
- yellow tag for each icon
- status line description for each menu icon

3.Example: Two-sample t -test

A clinical trial is considered to investigate the question ‘Does a new drug reduce anemia in elderly women after hip fracture?’. Women who suffered a hip fracture and were treated for it often suffer from anemia. A two-group randomised, double-blind study is planned. Each patient will be randomly assigned the new drug or a placebo. The sample sizes in the two groups will be equal. The primary outcome measure will be the percentage change in hematocrit level from pre-treatment to post-treatment.

Test problem and choice of test

This trial set up can be translated into

- Step 1) the null hypothesis of no difference in change in hematocrit levels between drugs is to be tested against the experimental hypothesis of larger change in hemotocrit levels for the new drug. Since it cannot be ruled out in advance that the new drug has harmful effects, a two-sided alternative hypothesis is specified.
- A significance level of 5% is required.
- Step 2) A two independent samples t -test will be employed to test this hypothesis since the study units are not related and the assumption of normality seemed reasonable.

Two samples t -test in nQuery

To input this information into nQuery

- select **File** menu **New**,
- in the **Study Goal and Design** box set **Goal: Make conclusions using** to **Means**,
- set **Number of groups** to **Two**
- and **Analysis method** to **Test**.
- In the box below, the selection results in three choices of tests. The two-samples t -test is highlighted by default. The choice is accepted by clicking on **OK**
- type in 0.05 for the test's significance level
- type '2' to choose a two-sided test.

Effect size

Step 3) From previous studies it was known that the placebo group showed 0% change while the change in the other group was at least 2%. The common standard deviation of the change in hematocrit for both groups was estimated from the previous studies to be 2%.

- Type '0', '2' and '2' in the rows **Group 1 mean**, **Group 2 mean** and **Common standard deviation** respectively.
- nQuery automatically calculates the **Difference in means** and the **Effect size**.

Power analysis statement

Step 4) The power required of the test was 80%.

- nQuery returns the minimum sample size required in each group under **n per group**. Here 17 patients are needed.
- Press the button '**Create statement**' to describe the result of the sample size calculation. Here the statement 'A sample size of 17 in each group will have 80% power to detect a difference in means of -2 (the difference between a Group 1 mean of 0 and a Group 2 mean of 2) assuming that the common standard deviation is 2 using a two group *t*-test with a 0.05 two-sided significance level.' is returned.

Sensitivity analysis

Step 5) The investigator is concerned that it will not be possible to recruit 17 patients in each group or some could drop out. How would this affect the power of the test?

- Select the column representing the current power analysis.
- On the menu choose **Plot** then **Plot power vs n**. The resulting graph shows the power of the test plotted against the sample size of the test. The graph is arranged so that the required sample size together with smaller and larger sample sizes is shown. Here the power of the test would be reduced to ca. 75% if 15 were considered instead of 17.

Further applications of the two sample t -test

The two sample t -test is the most frequently used test for power calculations. Often test problems can be evaluated as a two-group comparison.

- Test for interaction between a two-level within subject factor and a grouping factor.
- Comparison of several groups.

4. Example: χ^2 - test (comparing proportions)

- When the outcome is binary (only two possible categories A or B can occur) two groups can be compared by the frequency of occurrence of A (or B respectively).
- The test problem is similar to the two-sample t -test except that we are considering proportions rather than means.
- Proportions can be compared between two samples using a χ^2 - test.

Testing proportions

- Null hypothesis: the probability of A is the same in both groups: $\pi_1 = \pi_2$
- Two-sided alternative hypothesis: $\pi_1 \neq \pi_2$
- One-sided alternative hypotheses: $\pi_1 > \pi_2$
or $\pi_1 < \pi_2$
- The standardised effect size is the odds ratio $\psi = \pi_2(1 - \pi_1) / \pi_1(1 - \pi_2)$.

Example: comparing proportions

An investigator is planning a four week clinical trial. A new drug is to be compared against a placebo for its effectiveness in healing duodenal ulcers. After four weeks the proportion healed in each treatment group will be measured

Test problem and choice of test

The study objective can be translated into

- Step 1) the null hypothesis of equal proportions of healing is tested against the two-sided alternative hypothesis of a difference in proportions between the two treatment groups.
- A significance level of 5% is required.
- Step 2) A two-sample χ^2 - test will be employed to test this hypothesis since the samples are not related and the outcome is binary.

Two samples χ^2 -test in nQuery

To input this information into nQuery

- select **File** menu **New**,
- in the **Study Goal and Design** box set **Goal: Make conclusions using** to **Proportions**,
- set **Number of groups** to **Two**
- and **Analysis method** to **Test**.
- In the box below, the selection results in three choices of tests. The χ^2 -test is highlighted by default. The choice is accepted by clicking on **OK**
- type in 0.05 for the test's significance level
- type '2' to choose a two-sided test.

Effect size

Step 3) many previous trials suggest that 45% of those receiving the placebo will be healed at four weeks, and 75% of patients receiving standard H2 blockers will heal by four weeks. The investigator wants to have 95% power of the trial even if the new drug results in the healing of only 65% of the patients

- Type '0.45' and '0.65' in the rows **Group 1 proportion** and **Group 2 proportion**.
- The **Odds ratio** is calculated automatically.

Power analysis statement

Step 4) the power required of the test is 95%

- nQuery returns the minim sample size required in each group under **n per group**. Here 158 subjects per group are needed.
- Press the button **Create statement** to describe the result. nQuery returns ‘A two group χ^2 - test with a 0.05 two-sided significance level will have 95% power to detect the difference between a Group 1 proportion, π_1 , of 0.45 and a Group 2 proportion, π_2 , of 0.65 (odds ratio of 2.27) when the sample size in each group is 158.

5. Example: Correlation test

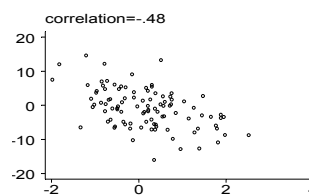
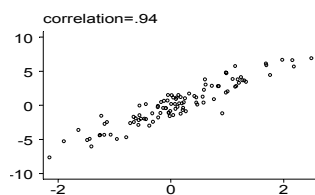
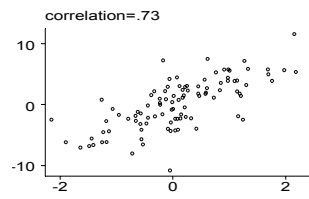
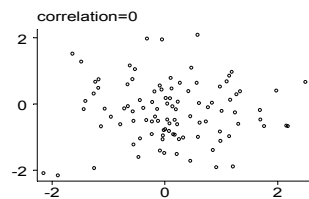
The Pearson correlation, r , is a measure of association between two continuous variables

- $-1 \leq r \leq 1$
- a positive (negative) correlation implies that variable 2 increases (decreases) when variable 1 increases.
- r^2 is the percentage of variance in variable 1 which is explained by variable 2 (or vice versa)

Testing Correlation

- **Null hypothesis:** correlation in the population, ρ , is zero.
- **Two-sided alternative hypothesis:** $|\rho|>0$
- The test of zero correlation is equivalent to the test of zero regression coefficient.
- The standardised effect size is ρ , the correlation itself.

Effect sizes



Example: Correlation via Regression

- A cross-sectional study is planned to investigate the association between depression as measured by the BDI and GP consultation rates (number of consultations in previous year) .
- It is planned to obtain a sample of patients from a GP practice.

Test problem and choice of test

The study objective can be translated into

- Step 1) the null hypothesis of no correlation between depression and consultation rates is to be tested against the alternative hypothesis of a positive correlation . Although we are quite sure that the true correlation is not negative, we decide to carry out the more conservative and more widely used two-tailed test.
- A significance level of 5% is required.
- Step 2) A test of zero regression coefficient is required.

Regression test in nQuery

To input this information into nQuery

- select **File** menu **New**,
- in the **Study Goal and Design** box set **Goal: Make conclusions using** to **Regression**
- set **Number of groups** to **One** and **Analysis method** to **Test**.
- In the box below, select linear regression with one independent variable.
- type in 0.05 for the test's significance level
- type '2' to choose a two-sided test.

Effect size

Step 3) In a paper, the correlation between GHQ and consultation rate is given as 0.24.

- We believe that the correlation is higher for BDI, for example 0.4.
- Type 0.4 in the row **Correlation**.

Power analysis statement

Step 4) The power required of the test is 80%.

- nQuery returns the minimum sample size required in each group under **n**. Here 44 patients are needed.
- Press the button '**Create statement**' to describe the result of the sample size calculation. Here the statement 'When the sample size is 44, the linear regression test of $\rho=0$ ($\alpha= 0.050$ two-sided) for one normally distributed covariate x will have 80% power to detect a ρ of 0.400.' is returned.

Sensitivity analysis

Step 5) We are not all that sure about the correlation.

- Try several correlations in the range from 0.2 to 0.6.
- If you use enough subjects to have an 80% chance of detecting a correlation on 0.4, what are your chances of detecting a correlation of 0.6?
- How high would the correlation have to be for you to have 80% power with a sample size of 30?

Further applications of the regression test

Several other problems can be translated into a simple regression test.

- Multiple regression.
- Correlation between the same variable
 - before and after an event
 - in paired observations, e.g. sisters