USING STATISTICS IN RESEARCH

David A. Stephens

Department of Mathematics, Imperial College

d.stephens@imperial.ac.uk
stats.ma.ic.ac.uk/∼das01/StatsShortCourse/

24^{th} March, 2004

Module 5 : 24th March

Power and Sample Size

- Study Design
- Power and Sample size
- Optimal Design
- Protocols

SECTION 1.

THE CONDUCT OF EXPERIMENTAL STUDIES AND TRIALS

Planning of experimental animal or human studies must take into account many considerations

- aims and objectives
- ethics
- design
- data collection
- statistical analysis and reporting

1.1 CLINICAL TRIALS AND STUDIES

A (**clinical) trial** is a designed study comparing the effect and value of treatment/interventions against a control in human subjects.

Typically,

- experimental units - "subjects" - are followed forward in time
- one or more treatments - "interventions"
- involves therapeutic agent, devices, regimens or procedures
- has a control group (similar to the intervention group at start of study)
- the control group is selected to be as similar to the study group as is possible in virtually all respects

The ideal clinical trial includes

- **randomization** of subjects
- **blinding** of subjects and care providers

Randomization allows for the equal allocation of potential effect modifiers and confounders between the two study groups; factors which are possibly unknown or unpredictable at the onset of the study

Blinding attempts to eliminate bias which might be introduced by either the participating subject or care providers

Ethics: Three fundamental ethical principles regarding research:

- respect for animals/persons;
  - for humans, individuals should be treated as autonomous
  - those with diminished autonomy need protection.
- worth and benefit
  - prioritize the well being of the individual
  - benefit for society/class of patients
- justice - treat persons fairly; share the risks/benefits.

Research design issues:

- Randomization
  - may be a problem if the treatment is known (or perceived) to be superior to placebo
  - trial may be unethical
- Placebo control
  - problems of an acceptable placebo
  - deprivation of treatment
- Monitoring of the trial
  - how to handle available data as it accrues
  - monitoring for safety

1.2 ESSENTIAL COMPONENTS

- Review of the scientific background for the study
  - previous animal investigations/laboratory work
  - preliminary evidence from case reports or case series
- Development of specific written hypothesis/hypotheses to be tested
  - *ad hoc* testing for statistical significance is unjustifiable
  - planned comparisons preferred over post hoc
  - multiple comparison issues and control of the Type I error

- What is the basic study design?
  - randomized (controlled) trial
  - non-randomized concurrent controlled study
  - historical controls - non-randomized, non-concurrent
  - crossover designs - subject serves as own control
  - withdrawal studies - assesses response to withdrawal of intervention or a reduction of dosage
  - factorial design - assesses the response to more than one type of intervention

- Study population
  - specific inclusion and exclusion criteria are necessary
  - sample size/power calculations/curves

- Statistical Analysis
  - what is/are the dependent and independent variables?
  - how will bias be controlled?
  - are there specific effect modifiers (**risk factors**) and/or **confounders** which need to be considered ?
  - what measurements are needed
  - how is the validity/accuracy of the measure to be confirmed ?
  - is the proposed sample size practicable ?
  - control of Type I and Type II errors
    - effect size and estimate of variances (signal/noise ratio)
    - If a significant difference exists between groups can it in fact be demonstrated ?
    - does the study have adequate power ?
  - How will attrition/loss to follow-up be handled?

1.3 ERROR AND VALIDITY

1.3.1 SOURCES OF ERROR

1. Random error - handled with the use of statistical tests and methods
2. Systematic error - uncontrolled error which may change the results and/or interpretation of research
3. Specific types of error:
    - **Bias** - any systematic error that results in an incorrect estimate of the association between exposure (intervention) and the risk of disease e.g. selection bias, recall bias, lead time bias

- **Confounding** - when the effect of the exposure (intervention) upon disease is altered by some other unaccounted for factor
  - e.g. in a study of the effect of exercise on the occurrence of coronary artery disease, age could be a confounder
  - Confounding may be adjusted for in the study design or in the final analysis of the data.
  - Controlled by:
    - **Randomization:** assures equal distribution of confounders between study and control groups
    - **Restriction :** subjects are restricted by the levels of a known confounder
    - **Matching:** potential confounding factors are equally distributed between the study groups
    - **Stratification :** (relative) risk estimates are computed for the various levels of potential confounders

- **Effect Modification** - when the association between exposure (intervention) and disease varies by the level of a third factor.
  - This represents an inconsistent distortion or nuisance effect.
  - Cannot adjust for effect modification
  - can compare risk estimates by levels of the effect modifier
  - cannot control for effect modification in the analysis

1.3.2 VALIDITY

- **Internal Validity** Is there in fact a causal relationship between the experimental treatment **(independent variable)** and the observed effect (**dependent variable**)?
- **Validity of Cause**
  - infers that the observed effect is attributable to the specific experimental intervention and not other variables
  - infers that the hypothetical dependent variable is accurately reflected by the measured dependent variable
- **External Validity** : could the observed effect be produced by in other settings, with other populations, at other times...
- **Conclusion Validity** : Are the conclusions reached justifiable on statistical grounds?

THREADS TO VALIDITY

1. Validity of Cause

- psychology
  - being part of a study may cause an increase in the observed/reported magnitude of effect (*Hawthorne effect*)
  - self-fulfilling studies: expectations of the experimenter influence how data are viewed
  - subject apprehension (perceived expectation of response)
- systematic/random variability
  - single (variable) measurement of the outcome
  - multiple measures may improve strength of study
  - aim to reduce standard errors
- weak treatment (small effect size)
- application of intervention of treatment (Integrity)

2. **Validity of Effect**

    - Inadequate theoretical analysis of the variables/concepts studied.
    - Small number of effects measured.

3. **Internal Validity**

    - unexpected systemic change (subject/experimenter based) may explain the observed change.
    - testing on multiple occasions may change the results
    - extreme observations may be only random events.

- selection error or bias
- loss of subjects before the end of the study
  - explanation for losses/dropouts ?
  - ignorable/non-ignorable response
  - missing at random/completely at random.
- introduction of experimental treatment for all patients (compensatory equalization of treatment).
- Subjects who perceive that they are receiving a less desirable treatment may work harder ("*compensatory rivalry*").
- Subjects who perceive that they are receiving a less desirable treatment give up effect ("*resentful demoralization*").

4. **External Validity**

   - Treatment does not generalize
     - to other situation.
     - to other populations.
     - to other experimental/treatment settings.
     - to other time periods.
     - when used in isolation

5. **Statistical Conclusion Validity**

   - Low statistical power
   - Violations of assumptions of statistical tests
   - Multiple testing
   - Low reliability of measures

1.4 RANDOMIZATION

The basic objectives of randomization are to

- eliminate biases due to subject/group assignment
- produce comparable groups
- make statistical analysis more valid
- achieve **balance** in the study group composition

A **randomized** trial differs from an **observational** (sampling, population-based) study as the composition of the study groups are determined by the experimenter

Lack of balance can compromise properties of proposed statistical tests.(for example, power)

> "**The importance of randomization cannot be over stressed. Randomization is necessary for conclusions drawn from the experiment to be correct, unambiguous and defensible**."

`http://www.itl.nist.gov/div898/handbook/pri/section7/pri7.htm`

The objective of balance can be achieved using a number of approaches

Consider the two treatment group (groups **A** and **B**) case:

- simple randomization
  - for fixed sample size n, allocate $n/2$ (selected at random) to each group
  - complete randomization: allocate each individual to group 1 with probability 1/2
- biased randomization
  - may wish to allocate unequal numbers (in accordance with power considerations)
  - allocate with probability p and $1-p$ to the two groups.

- (balanced) permuted block randomization
  - simple randomization does not guarantee a balance over time
  - instead
    - divide study base into K blocks of size $2m$ say
    - (simple) randomize each block with m into each of the two groups
    - maximum imbalance at any time is m
  - for example: let $m = 2$, so that there are $\binom{4}{2} = 6$ possible patterns of allocation

    AABB,ABAB,BAAB,BBAA,BABA,ABBA

    - allocate individuals in blocks of 4, according to one of these six patterns chosen at random.

- stratified randomization
  - may desire to have treatment groups balanced with respect to risk factors/confounders
  - proceed as above for identified strata
- dependent/response dependent randomization
  - can balance the design dynamically (dependent on the current group sizes)
  - can balance the design dependent on response or other external factors

1.5 INFERENCE FOR TYPES OF STUDY

The method of data collection can sometimes influence how the data are subsequently analyzed. Typically, we wish to examine the variability in a **incidence** of the response event with some **exposure** factor, possibly with the presence of **confounding** factors.

In clinical, medical or epidemiological studies, there are two types of study;

- **OBSERVATIONAL** : where the exposure **arises naturally,** and the experimenter attempts to detect differences in response
- **EXPERIMENTAL** : where the exposure is **determined by the experimenter**

The type of study used influences how the data are analyzed.

1.5.1 OBSERVATIONAL STUDIES

Consider the following representation of an observational study; let

- S denote the **inclusion of a subject in the study**,
- E denote **exposure**
- F denote **incidence;** if F occurs, then we observe a **case.**

We will try to examine variation in **incidence** rate across different levels of the **exposure** factor.

Using probability theory

$$P\left(E \cap F \cap S\right) = P(E)P(F|E)P(S|E \cap F).$$

We will use this factorization to deduce estimable quantities from different observational studies that comprise the S "margin" of a $2 \times 2 \times 2$ events table with the recorded number of observations as follows; for the events

	$E \cap S$	$E' \cap S$
$F \cap S$	$E \cap F \cap S$	$E' \cap F \cap S$
$F' \cap S$	$E \cap F' \cap S$	$E' \cap F' \cap S$

and the counts data

	$E \cap S$	$E' \cap S$	TOTAL
$F \cap S$	n_{11}	n_{12}	$n_{1.}$
$F' \cap S$	n_{21}	n_{22}	$n_{2.}$
TOTAL	$n_{.1}$	$n_{.2}$	$n_{..}$

1.5.2 COHORT STUDY

In a **cohort** study, the defining feature is that E and F are **independent** of S so that

$$P(E \cap F \cap S) = P(E)P(F|E)P(S) \qquad \Longrightarrow$$

	E	E'
F	$E \cap F$	$E' \cap F$
F'	$E \cap F'$	$E' \cap F'$

as the S and S' margins are **identical**.

It follows that **all of the following quantities are estimable:**

- **RATES OF EXPOSURE AND INCIDENCE**

$$\theta = P(E) = P\left(E \cap F\right) + P\left(E \cap F'\right)$$

and

$$\phi = P(F) = P\left(E \cap F\right) + P\left(E' \cap F\right)$$

with estimates

$$\widehat{\theta} = \frac{n_{.1}}{n_{..}} \qquad\qquad \widehat{\phi} = \frac{n_{1.}}{n_{..}}$$

- **INCIDENCE RATES IN THE EXPOSED/UNEXPOSED GROUPS**

$$\pi_1 = P(F|E) = \frac{P(E \cap F)}{P(E)}$$

$$\pi_0 = P(F|E') = \frac{P(E' \cap F)}{P(E')}$$

with estimates

$$\widehat{\pi}_1 = \frac{n_{11}}{n_{.1}} \qquad \widehat{\pi}_0 = \frac{n_{12}}{n_{.2}}$$

- **THE RELATIVE RISK**

$$\rho = \frac{\pi_1}{\pi_0} = \frac{P(E \cap F)/P(E)}{P(E' \cap F)/P(E')}$$

with estimate

$$\hat{\rho} = \frac{\widehat{\pi}_1}{\widehat{\pi}_0} = \frac{n_{11}/n_{.1}}{n_{12}/n_{.2}}$$

- **EXPOSURE RATES IN THE CASE AND CONTROL GROUPS**

$$\gamma_1 = P(E|F) = \frac{P(E \cap F)}{P(F)}$$

$$\gamma_0 = P(E|F') = \frac{P(E \cap F')}{P(F')}$$

with estimates

$$\widehat{\gamma}_1 = \frac{n_{11}}{n_{1.}} \qquad \widehat{\gamma}_0 = \frac{n_{21}}{n_{2.}}$$

- **ODDS ON INCIDENCE IN THE EXPOSED AND UNEXPOSED GROUPS**

$$\omega_1 = \frac{\pi_1}{1-\pi_1} = \frac{P(E \cap F)}{P(E \cap F')}$$

$$\omega_0 = \frac{\pi_0}{1-\pi_0} = \frac{P(E' \cap F)}{P(E' \cap F')}$$

with estimates

$$\widehat{\omega}_1 = \frac{\widehat{\pi}_1}{1-\widehat{\pi}_1} = \frac{n_{11}}{n_{21}} \qquad \widehat{\omega}_0 = \frac{\widehat{\pi}_0}{1-\widehat{\pi}_0} = \frac{n_{12}}{n_{22}}$$

- **ODDS ON EXPOSURE IN THE CASE AND CONTROL GROUPS**

$$\Omega_1 = \frac{\gamma_1}{1-\gamma_1} = \frac{P(E \cap F)}{P(E' \cap F)}$$

$$\Omega_0 = \frac{\gamma_0}{1-\gamma_0} = \frac{P(E \cap F')}{P(E' \cap F')}$$

with estimates

$$\widehat{\Omega}_1 = \frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1} = \frac{n_{11}}{n_{12}} \qquad \widehat{\Omega}_0 = \frac{\widehat{\gamma}_0}{1-\widehat{\gamma}_0} = \frac{n_{21}}{n_{22}}$$

- **ODDS RATIO**

$$\psi = \frac{P(F|E)/P(F'|E)}{P(F|E')/P(F'|E')} = \frac{P(E \cap F)/P(E \cap F')}{P(E' \cap F)/P(E' \cap F')} = \frac{\omega_1}{\omega_0} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

or equivalently

$$\psi = \frac{P(E|F)/P(E'|F)}{P(E|F')/P(E'|F')} = \frac{P(E \cap F)/P(E' \cap F)}{P(E \cap F')/P(E' \cap F')} = \frac{\Omega_1}{\Omega_0} = \frac{\gamma_1/(1-\gamma_1)}{\gamma_0/(1-\gamma_0)}$$

with estimate

$$\widehat{\psi} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

1.5.3 CASE-CONTROL STUDY

In a **case-control** study, we look for incidences or **cases** and automatically include them in the study, and then we find a set of controls who do not have the "case response" and include them also. The defining probabilistic feature is that E is **independent** of S **given** F and **given** F', but

$$P(S|E \cap F) = P(S|E' \cap F) \qquad P(S|E \cap F') = P(S|E' \cap F')$$
$$P(E|S \cap F) = P(E|S' \cap F) \qquad P(E|S \cap F') = P(E|S' \cap F')$$

In practice the design proceeds as follows; Our assumption of conditional independence of E and S given F means corresponds to an assumption of no probabilistic dependence between exposure and inclusion in the study.

The case-control study design is perhaps more efficient, but does not allow the full range of inferences to be made.

It can be shown that **only the following quantities are estimable in the absence of other knowledge**

- **EXPOSURE RATES IN THE CASE AND CONTROL GROUPS** with estimates

$$\widehat{\gamma}_1 = \frac{n_{11}}{n_{1.}} \qquad \widehat{\gamma}_0 = \frac{n_{21}}{n_{2.}}$$

- **ODDS ON EXPOSURE IN THE CASE AND CONTROL GROUPS** with estimates

$$\widehat{\Omega}_1 = \frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1} = \frac{n_{11}}{n_{12}} \qquad \widehat{\Omega}_0 = \frac{\widehat{\gamma}_0}{1-\widehat{\gamma}_0} = \frac{n_{21}}{n_{22}}$$

- **ODDS RATIO** with estimate

$$\widehat{\psi} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

EXAMPLE: LIMITATION OF CASE CONTROL STUDIES

An illustration of why case-control studies are limited in their usefulness is presented below; fixing $\gamma_1 = 0.2$ and $\gamma_0 = 0.1$ and changing the size of the CONTROLS group. In Table 1

	$E \cap S$	$E' \cap S$	TOTAL
CASES	20	80	100
CONTROLS	100	900	1000
TOTAL	120	980	1100

and in Table 2

	$E \cap S$	$E' \cap S$	TOTAL
CASES	20	80	100
CONTROLS	500	4500	5000
TOTAL	520	4580	5100

Then clearly if we estimate γ_1 and γ_0, we recover the true values 0.2 and 0.1, and in each case

$$\text{TABLE 1: } \widehat{\psi} = \frac{20 \times 900}{80 \times 100} = \frac{9}{4} \qquad \text{TABLE 2: } \widehat{\psi} = \frac{20 \times 4500}{80 \times 500} = \frac{9}{4}$$

but if we try to estimate, for example π_1 and π_0 in the same way that we would for a cohort study, we get different results from the two tables

$$\text{TABLE 1:} \quad \widehat{\pi}_1 = \frac{20}{120} = \frac{1}{6} \qquad \widehat{\pi}_0 = \frac{80}{980} = \frac{4}{49}$$

$$\text{TABLE 2:} \quad \widehat{\pi}_1 = \frac{20}{520} = \frac{1}{26} \qquad \widehat{\pi}_0 = \frac{80}{4580} = \frac{4}{229}$$

The row totals, corresponding to the total numbers of **cases** and **controls**, $n_{1.}$ and $n_{2.}$, are fixed by the experimenter, and we do **not have a random sample of exposed and unexposed individuals** from the population. In a cohort study, only the total cohort size, $n..$, is fixed.

1.5.4 STANDARD ERRORS FOR EFFECT SIZES

In a 2×2 table analysis, our estimates of key parameters are functions of the counts in the table; these estimates have associated (estimated) standard errors that allow construction of confidence intervals for the parameters, and hence permit hypothesis testing.

Recall the counts data for individuals in the study

	E	E'	TOTAL
F	n_{11}	n_{12}	$n_{1.}$
F'	n_{21}	n_{22}	$n_{2.}$
TOTAL	$n_{.1}$	$n_{.2}$	$n_{..}$

Then we have the following estimates and estimated standard errors for effect sizes; we typically examine such quantities on the (natural) log scale:

- The log **relative-risk**

$$\log \hat{\rho} = \log \frac{\widehat{\pi}_1}{\widehat{\pi}_0} = \log \left(\frac{n_{11}/n_{.1}}{n_{12}/n_{.2}} \right)$$

with **estimated standard error**

$$\sqrt{\left(\frac{1}{n_{11}} - \frac{1}{n_{11}+n_{21}} \right) + \left(\frac{1}{n_{12}} - \frac{1}{n_{12}+n_{22}} \right)}$$

- The log **odds ratio**

$$\log \widehat{\psi} = \log \left(\frac{n_{11}n_{22}}{n_{12}n_{21}} \right)$$

with **estimated standard error**

$$\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}}$$

1.5.5 EXPERIMENTAL STUDIES

Experimental studies are studies where the exposure factor is determined by the experimenter during the study

- treatment/control
- drug/placebo
- dose level $1, 2, 3, ..., K$

An experimental study is a special type of **cohort** study; the most common type of experimental study is a **randomized controlled trial** (as described in previous sections)

SECTION 2.

POWER AND SAMPLE SIZE

General design issues often need to be considered before an experimental study is embarked upon.

- In clinical/animal studies, ethical considerations dictate that the "optimal" number experimental units are considered, and that resources are deployed in an "optimal" fashion.
- Economic forces mitigate against using an expansive study when a smaller one enables the same research hypotheses to be tested.

Data are collected, and hypotheses tested, within a framework of statistical inference and summary; the statistical framework also allows formal assessment of the utility of a study, and allows a statistically optimal study (with respect to a specific hypothesis) to be considered

2.1 STATISTICAL HYPOTHESIS TESTING

Recall the basic components of statistical hypothesis testing: in assessing which of two hypotheses, H_0 and H_1

$$H_0 : \text{NULL HYPOTHESIS}$$
$$H_1 : \text{ALTERNATIVE HYPOTHESIS}$$

is preferable in explaining the observed data, we need to specify, and compute the following quantities

- **TEST STATISTIC**, T
- **NULL DISTRIBUTION**, F_0
- **SIGNIFICANCE LEVEL**, α
- **P-VALUE**, p
- **CRITICAL VALUE(S)/CRITICAL REGION** $\mathcal{R}$

Recall that the **null distribution** is the probability distribution of **test statistic** T **if the null hypothesis**, H_0, is **true**; if t^* is the observed test statistic, lies in the critical region, we **reject** H_0 in favour of H_1, and **do not reject** H_0 otherwise.

The critical region $\mathcal{R}$ is defined via the significance level α by

$$P\left[T \in \mathcal{R} | H_0 \text{ is TRUE}\right] \leq \alpha \tag{1}$$

(where $T \in \mathcal{R}$ means "T takes a value in the set $\mathcal{R}$").

Note that (1) considers only the distribution of T if H_0 is true, and the conditional probability of rejection H_0 in this case.

i.e. it is concerned only with "**false positive**" results**.**

In a classical test of H_0 (null hypothesis) versus H_1 (alternative hypothesis), there are four possible outcomes, two of which are erroneous:

1. Do not reject H_0 when is H_0 true.
2. Reject H_0 when H_0 is not true.
3. Reject H_0 when H_0 is true (**Type I error**).
4. Do not reject H_0 when H_0 is false (**Type II error**).

	Action	
	Do Not Reject H_0	Reject H_0
H_0 True	✓	**Type I Error**
H_0 not True	**Type II Error**	✓

TYPE I : FALSE POSITIVE result

TYPE II : FALSE NEGATIVE result

To construct a test, the distribution of the test statistic under H_0 is used to find a critical region which will ensure that the probability of committing a type I error does not exceed some predetermined significance level α.

Ideally, we would like to make the probability of making any type of error (false positive and false negative) as small as possible. For a finite sample however, this is not achievable, so a pragmatic approach that bounds the probability of a Type I error is adopted.

NOTE: For an infinite sample, we desire that the probabilities of Type I and Type II errors should both be zero.

2.2 POWER CALCULATIONS

The **power**, $1-\beta$, of a statistical test is its ability to **correctly reject the null hypothesis**, or

$$\begin{aligned} 1-\beta &= P\left[\text{Reject } H_0 | H_0 \text{ is not True}\right] = P\left[T \in \mathcal{R} | H_0 \text{ is not True}\right] \\ &= 1 - P\left[\text{Do not Reject } H_0 | H_0 \text{ is not True}\right] \\ &= 1 - P\left[T \notin \mathcal{R} | H_0 \text{ is not True}\right] \end{aligned}$$

so that

$$\beta = P\left[\text{Do not Reject } H_0 | H_0 \text{ is not True}\right] = P\left[T \notin \mathcal{R} | H_0 \text{ is not True}\right]$$

which is based on the distribution of the test statistic under H_1.

This is the first occasion on which we have had to consider the distribution of the test statistic under the alternative hypothesis; as we shall see, in order to consider a sample size or power calculation, we must **explicitly** consider the alternative hypothesis.

Suppose that the hypothesis test concerns a parameter θ that can take values in the parameter space Θ. Suppose that the null and alternative hypotheses partition Θ into two parts, Θ_0 and Θ_1, that is

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_1 \end{aligned}$$

so that, in the simplest case

$$\begin{aligned} H_0 &: \theta = c \\ H_1 &: \theta \neq c \end{aligned}$$

we have $\Theta_0 \equiv \{c\}$, $\Theta_0 \equiv \mathbb{R} \backslash \{c\}$

Under H_1, the probability

$$P\left[\text{Do not Reject } H_0 | H_0 \text{ is not True}\right] = P\left[T \notin \mathcal{R} | \theta \in \Theta_1\right]$$

which we previously defined as β will vary as the true value of θ varies in the set Θ_1, hence we should write β as a function of θ.

EXAMPLE: In a **one-sample test** of a normal mean, we have $X_1, ..., X_n$ as a set of random variables relating to the observed data $x_1, ..., x_n$, and an assumption that

$$X_i \sim N(\mu, \sigma^2)$$

for $i = 1, ..., n$. If σ^2 is known, to perform a two-sided test of equality the hypotheses would be as follows:

$$\begin{aligned} H_0 &: \mu = \theta_0 \\ H_1 &: \mu \neq \theta_0 \end{aligned}$$

The test statistic is

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

and under H_0,

$$Z = \frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

We reject H_0 at significance level α if the z statistic is more extreme than the critical values of the test are

$$\mathcal{R} = \theta_0 \pm C_R \frac{\sigma}{\sqrt{n}} \qquad\qquad C_R = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Now, if H_1 is true, and $\mu = \theta$ for some value θ, then , $X \sim N(\theta, \sigma^2)$, and hence

$$Z = \frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} \sim N\left(\frac{\theta - \theta_0}{\sigma/\sqrt{n}}, 1\right).$$

so the probability that z lies in the critical region if $\mu = \theta$ is

$$\begin{aligned} P\left[T \in \mathcal{R}|\theta\right] &= P\left[Z \leq -C_R|\theta\right] + P\left[Z > C_R|\theta\right] \qquad (2) \\ &= \Phi\left(-C_R - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) + \left(1 - \Phi\left(C_R - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right)\right) \end{aligned}$$

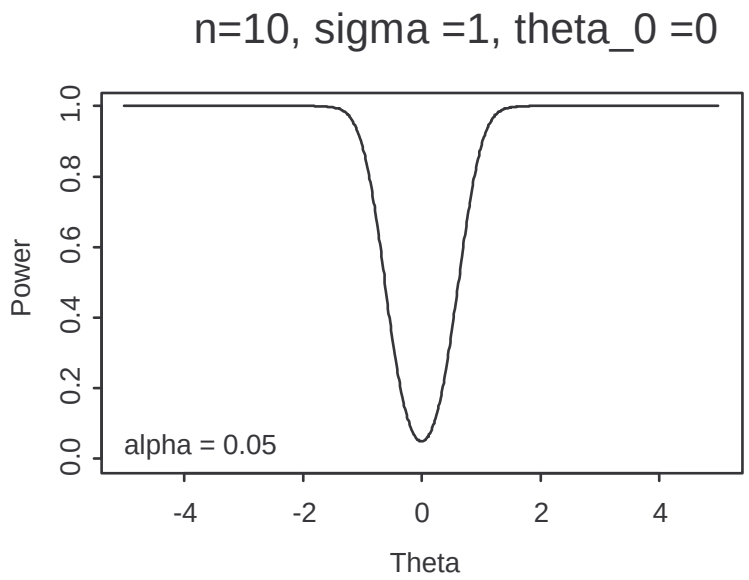
where Φ is the standard normal distribution function.

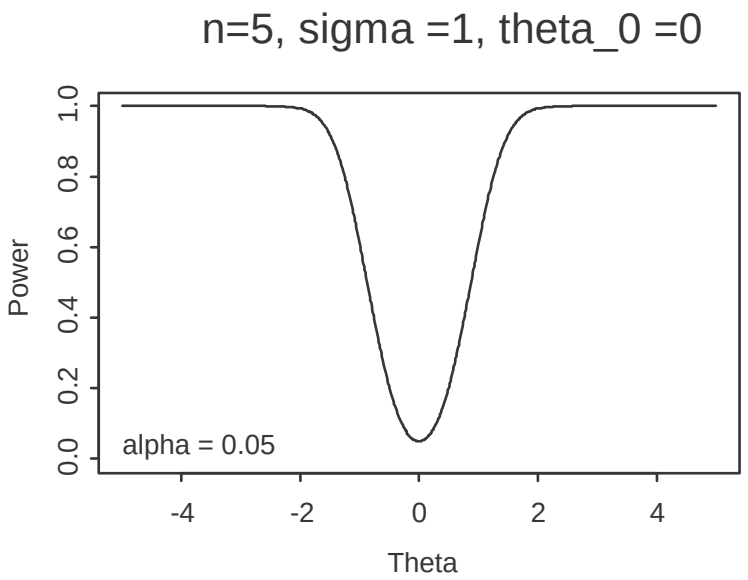
This quantity is the **power function,** $1 - \beta(\theta)$, when μ is actually equal to θ.

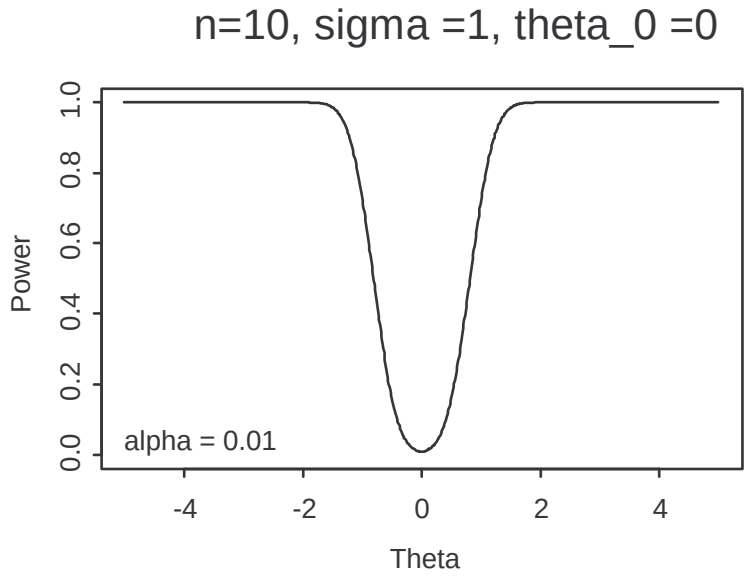
Hence the **probability of a Type II error** when the true is $\beta(\theta)$, where

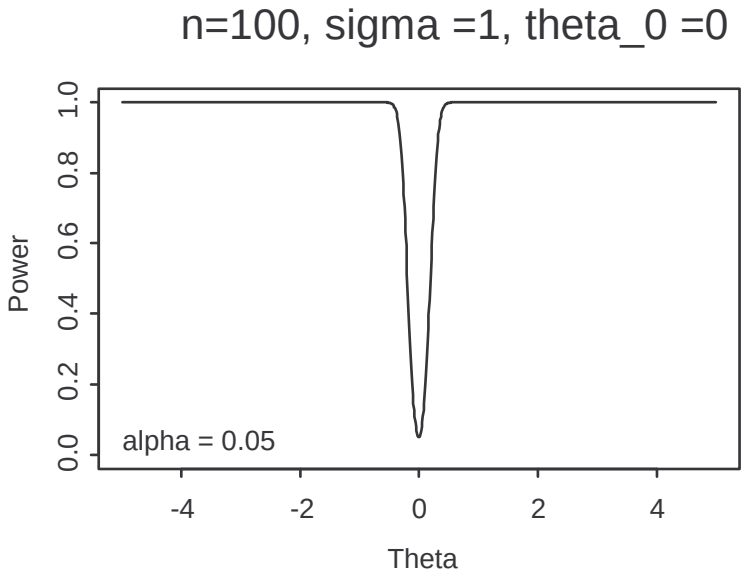
$$\begin{aligned}
\beta(\theta) &= 1 - P[T \in \mathcal{R}|\theta] \\
&= \Phi\left(C_R - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) - \Phi\left(-C_R - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(C_R - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) - \left(1 - \Phi\left(C_R + \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right)\right) \\
&= \Phi\left(C_R - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) + \Phi\left(C_R + \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) - 1
\end{aligned}$$

The plots below illustrate examples of power functions for different choices of σ and n, with $\theta_0 = 0$.


n=10, sigma =1, theta_0 =0
Power
0.0
0.2
0.4
0.6
0.8
1.0
alpha = 0.05
-4
-2
0
2
4
Theta


n=5, sigma =1, theta_0 =0
Power
0.0
0.2
0.4
0.6
0.8
1.0
alpha = 0.05
-4
-2
0
2
4
Theta


n=10, sigma =1, theta_0 =0
Power
0.0
0.2
0.4
0.6
0.8
1.0
alpha = 0.01
-4
-2
0
2
4
Theta


n=100, sigma =1, theta_0 =0
Power
0.0
0.2
0.4
0.6
0.8
1.0
alpha = 0.05
-4
-2
0
2
4
Theta

Thus for fixed α, θ_0, σ and n, we can compute the power function $\beta(\theta)$ as θ varies.

NOTE: The parameters in (2) appear in terms of the ratio

$$\frac{\theta - \theta_0}{\sigma}$$

that is, a **standardized difference** between the hypothesized values of μ under the null and alternative hypotheses.

Similar calculations are available for other of the normal distribution-based tests.

2.2.1 ONE-SIDED TESTS

To perform a one-sided test of the hypotheses

$$\begin{aligned} H_0 &: \mu = \theta_0 \\ H_1 &: \mu < \theta_0 \end{aligned}$$

the power function is

$$1 - \beta(\theta) = P[T \in \mathcal{R}|\theta] = P[Z \leq C_R(\alpha)|\theta] = \Phi\left(C_R(\alpha) - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right)$$

where $C_R(\alpha) = \Phi^{-1}(\alpha)$, with a similar calculation if $H_1 : \mu > \theta_0$

$$1 - \beta(\theta) = P[Z \geq C_R(\alpha)|\theta] = 1 - \Phi\left(C_R(\alpha) - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right)$$

where $C_R(\alpha) = \Phi^{-1}(1 - \alpha)$

2.2.2 UNKNOWN VARIANCE

If σ^2 is unknown, to perform a two-sided test of equality the hypotheses would be as follows:

$$\begin{aligned} H_0 &: \mu = \theta_0 \\ H_1 &: \mu \neq \theta_0 \end{aligned}$$

The test statistic is

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

where s is the sample standard deviation, and under H_0,

$$T = \frac{\overline{X} - \theta_0}{s/\sqrt{n}} \sim Student(n-1).$$

We reject H_0 at significance level α if the t statistic is more extreme than the critical values of the test, with

$$\mathcal{R} = \theta_0 \pm C_R \frac{s}{\sqrt{n}} \qquad\qquad C_R = F_{t_n}^{-1}\left(1 - \frac{\alpha}{2}\right)$$

where $F_{t_k}^{-1}$ is the inverse cdf of the $Student(k)$ distribution

Now, if H_1 is true, and $\mu = \theta$ for some value θ, then

$$\begin{aligned} T &= \frac{\overline{X} - \theta_0}{s/\sqrt{n}} \\ &= \frac{\overline{X} - \theta}{s/\sqrt{n}} + \frac{\theta - \theta_0}{s/\sqrt{n}} = T_0 + \frac{\theta - \theta_0}{s/\sqrt{n}} \end{aligned}$$

where $T_0 \sim Student(n-1)$.

Then the probability that T lies in the critical region is

$$
\begin{aligned}
1-\beta(\theta) &= P[T \in \mathcal{R}|\theta] \qquad (3)\\
&= P\left[\frac{\overline{X}-\theta}{s/\sqrt{n}} + \frac{\theta-\theta_0}{s/\sqrt{n}} \leq -C_R|\theta\right] + P\left[\frac{\overline{X}-\theta}{s/\sqrt{n}} + \frac{\theta-\theta_0}{s/\sqrt{n}} > C_R|\theta\right]\\
&= P\left[\frac{\overline{X}-\theta}{s/\sqrt{n}} \leq -C_R - \frac{\theta-\theta_0}{s/\sqrt{n}}|\theta\right] + P\left[\frac{\overline{X}-\theta}{s/\sqrt{n}} > C - \frac{\theta-\theta_0}{s/\sqrt{n}}|\theta\right]\\
&= F_{t_n}^{-1}\left(-C_R - \frac{\theta-\theta_0}{s/\sqrt{n}}\right) + \left(1 - F_{t_n}^{-1}\left(C_R - \frac{\theta-\theta_0}{s/\sqrt{n}}\right)\right)
\end{aligned}
$$

2.2.3 TWO SAMPLE TESTS

In a two sample problem, if σ^2 is unknown but common for both samples, to perform a test of the hypotheses:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_1 &: \mu_1 - \mu_2 = \delta \end{aligned}$$

The test statistic is

$$T = \frac{\overline{X}_1 - \overline{X}_2}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where s_P is the pooled sample standard deviation, and under H_0,

$$T \sim Student(n_1 + n_2 - 2).$$

We reject H_0 at significance level α if the t statistic is more extreme than the critical values of the test are

$$\mathcal{R} = \pm C_R \frac{s}{\sqrt{n}} \qquad C_R = F^{-1}_{t_{n_1+n_2-2}}\left(1 - \frac{\alpha}{2}\right)$$

Now, if H_1 is true, for the particular value of δ specified

$$\begin{aligned} T &= \frac{\overline{X}_1 - \overline{X}_2}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \delta}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} + \frac{\delta}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = T_0 + \delta_0 \end{aligned}$$

say, where $T_0 \sim Student(n_1 + n_2 - 2)$.

Then the probability that T lies in the critical region is

$$\begin{aligned} 1-\beta(\theta) &= P[T \in \mathcal{R}|\theta] \qquad (4) \\ &= P[T_0+\delta_0 \leq -C_R|\delta] + P[T_0+\delta_0 > C_R|\delta] \\ &= P[T_0+\delta_0 \leq -C_R-\delta_0|\delta] + P[T_0 > C_R-\delta_0|\delta] \\ &= F^{-1}_{t_{n_1+n_2-2}}(-C_R-\delta_0) + \left(1-F^{-1}_{t_{n_1+n_2-2}}(C_R-\delta_0)\right) \end{aligned}$$

and thus the power function is calculable for any combination of α, n_1, n_2 and δ.

SUMMARY: The adequacy of a test to distinguish between two hypotheses is a function of

- The null and alternative hypotheses;
- The target significance level α;
- The desired power to detect H_1 for a specific θ, $\beta(\theta)$;
- The variability within the population(s) under study as measured by σ
- The sample size n (or n_1 and n_2).

Our objective is to find a relationship between the above factors and the sample size that enables us to select a sample size consistent with the desired α and $\beta(\theta)$, typically, we will hypothesize a specific value of θ and compute the corresponding β.

2.2.4 GENERAL POWER CONSIDERATIONS

The principles outlined above can be applied in more complicated situations

- NON-PARAMETRIC TESTS
- NON-NORMAL DATA TESTS
  - Approximate Binomial
  - Exact Binomial
- ONE-WAY/TWO-WAY ANOVA
  - number of groups/cross-categories, K
  - number of observations per category, n_K
  - category levels $\theta_1, ..., \theta_K$
- REPEATED MEASURES

The details of the power calculation are more complicated as the complexity of the experimental procedure increases, but the principles remain the same; we compute

the probability of rejecting a specified null hypothesis when a specific alternative hypothesis corresponds the actual truth

that is, we are obliged to consider both null **and** alternative hypotheses, and their impact on the distribution of the test statistic.

This is fundamentally different from the simple hypothesis testing situation, where we only consider the **null** distribution.

Therefore, a power calculation **necessarily** involves consideration of a specific alternative hypothesis, that is, equivalently, the magnitude of

- $\dfrac{\theta - \theta_0}{\sigma}$ in the Normal sample case with known variance σ^2
- δ if σ^2 is unknown
- $\delta_\pi = \pi_1 - \pi_2$ in a two-sample Binomial problem, and test of

$$\begin{aligned} H_0 &: \quad \pi_1 - \pi_2 = 0 \\ H_1 &: \quad \pi_1 - \pi_2 = \delta_\pi \end{aligned}$$

and so on.

How do we choose these quantities ?

- usually by consideration of a “clinically” or ”experimentally” significant difference, or an “anticipated” effect size..

2.3 EXAMPLES

(see Machin et al, 1997, *Sample Size Tables for Clinical Studies*)

- power/sample size for independent groups of binary, ordered, categorical and continuous data
- paired/repeated measures data
- for equivalence studies
- survival
- observer (inter-rater) agreement

2.4 SIMULATION-BASED CALCULATION

When analytic expressions for the power/Type II error probability are not easily available, we can do approximate power calculations by simulation means

- we formulate the test (null and alternative hypotheses, test statistic) in the usual way
- we repeatedly simulate data under the alternative hypothesis model (for fixed sample size, null model)
- we compute the power/Type II error probability empirically by evaluating the frequency with which the null hypothesis is correctly rejected.

For complicated designs (correlated data, clustered/grouped data), this is often the simplest solution.

2.5 SAMPLE SIZE CALCULATIONS

In all of the above, we have concentrated on computing the **achieved power** for detecting a particular effect (relative effect) in a **fixed** study (perhaps that has already been carried out).

Often it is desirable to reverse the logic and to ask if a certain power β to detect an effect (if it is there) is required for a specified significance level α, how large would sample size n need to be ?

Such a consideration is of strategic importance in study design, and can give insight into the practicability of the proposed study.

Recall the simple concept of standard error in a mean;

$$s.e.\left(\overline{X}\right) = \frac{s}{\sqrt{n}}$$

Clearly as n increases, the standard error decreases. Thus if we wanted a standard error that was no larger than some quantity ϵ, we would have to chose n large enough to ensure this, that is,

$$\frac{s}{\sqrt{n}} \leq \epsilon \Leftrightarrow n \geq \left(\frac{s}{\epsilon}\right)^2$$

This simple idea extends naturally to confidence intervals, and to hypothesis tests, and hence to power assessments.

In the simple case of a single normal sample with known variance, the power equation in (2) can be rearranged to be explicit in one of the other parameters if β is regarded as fixed.

For example, if α, β, θ_0 and θ_1 are fixed, we can rearrange to get a sample size calculation to test for fixed difference $\delta = \theta_1 - \theta_0$

$$n = \frac{\sigma^2 \left(C_R + \Phi^{-1}(1-\beta)\right)^2}{(\theta_1 - \theta_0)^2}$$

or standardized difference $\Delta = \dfrac{|\theta_1 - \theta_0|}{\sigma}$

$$n = \frac{\left(C_R + \Phi^{-1}(1-\beta)\right)^2}{\Delta^2}$$

This idea of rearranging the power calculation to obtain a sample size extends to the general cases described above.

Other issues do need to be considered

- one-sided vs two-sided tests
- in two sample problems, the deployment of the samples to be used
  - equal proportions in the two groups
  - fixed unequal allocation ratio between subjects assigned to the two groups (in observational studies this may be necessary)
- allocation by randomization: exchangeable subjects