USING STATISTICS IN RESEARCH

David A. Stephens

Department of Mathematics, Imperial College

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

July 9, 2003

WEEK 5

POWER AND SAMPLE SIZE,

ASSOCIATION AND AGREEMENT MEASURES

&

SURVIVAL ANALYSIS

SECTION 8.

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

8.1 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.

8.2 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

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

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

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

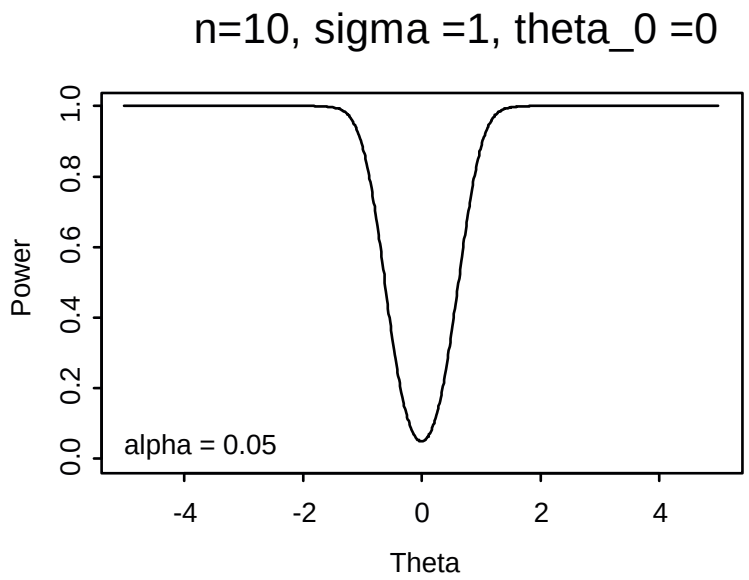
where Φ is the standard normal distribution function.

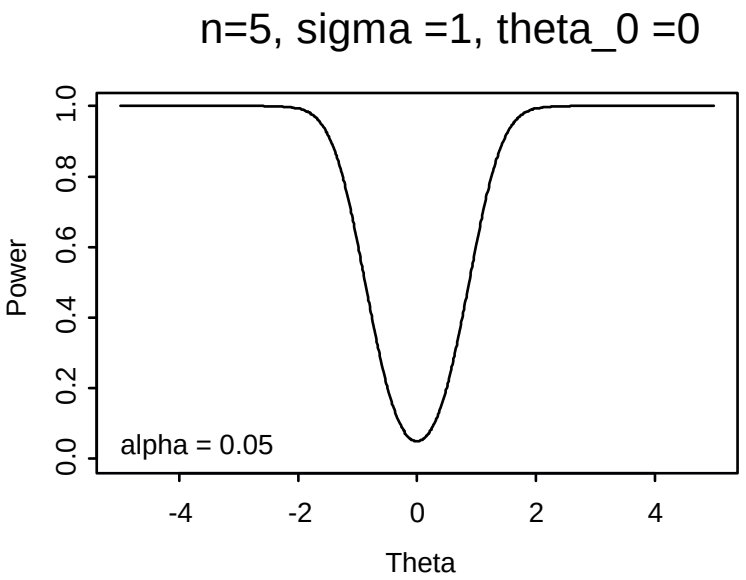
This quantity is the **power function,** $1 - \beta(\theta)$, when μ is actually equal to θ.

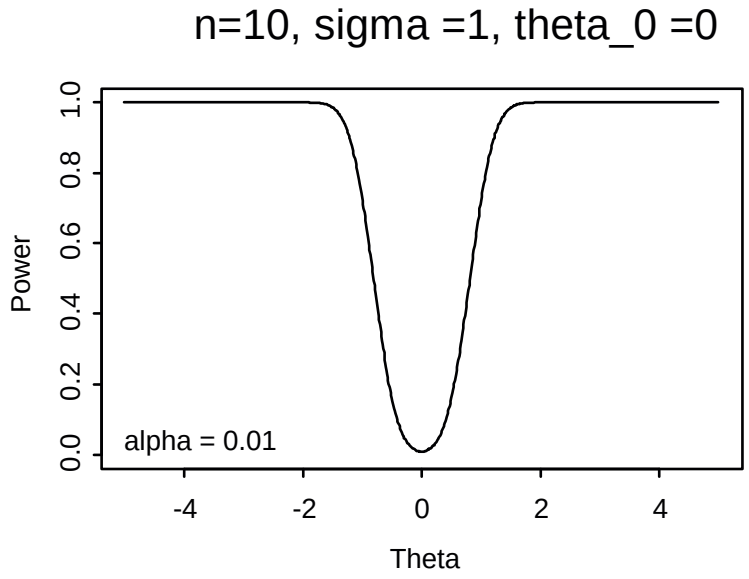
Hence the **probability of a Type II error** when the true is $\beta(\theta)$, where

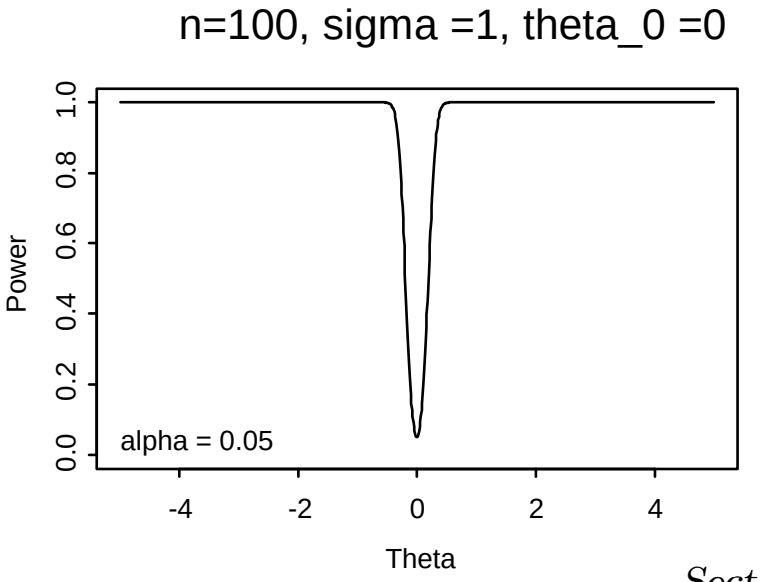
$$\begin{aligned}
\beta(\theta) &= 1 - P[T \in \mathcal{R}|\theta] \\
&= \Phi\left(C_R - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) - \Phi\left(-C_R - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(C_R - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) - \left(1 - \Phi\left(C_R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

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.

8.2.1 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 $H_1 : \mu > \theta_0$

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

8.2.2 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

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

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

Then the probability that T lies in the critical region is

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

8.2.3 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.

8.2.4 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..

8.3 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

8.4 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.

8.5 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

SECTION 9.

OBSERVER AGREEMENT

Assessing the results of diagnostic procedures and the effects of therapies often involves subjective judgements. Observer agreement studies are conducted to investigate the level of consensus on such assessments.

Typically, several observers make assessments on each of a series of subjects and these assessments are compared.

An important consideration for study design is the presence of both **within-observer** and **between-observer** variation. The apparent disagreement between observers may be due to either one of these components or both. It is important to distinguish between them, as any action taken to reduce disagreement will depend on which type of variation dominates. To do this, we require observations repeated by the same observer.

We might consider any of the following types of observer agreement studies

- studies with binary assessments and designs;
- where each of two observers assesses all subjects once,
- where each observer assesses all subjects twice
- where each observer assesses a proportion of the subjects once and the remainder twice.

and so on.

NOTE: Sample-size calculations are conventionally based upon hypothesis-testing theory Observer-agreement studies, however, are designed to estimate the level of observer agreement. Moreover, unlike clinical trials, there are no obvious hypotheses to test. The hypothesis of perfect agreement between observers is unrealistic and the hypothesis of agreement purely by chance is also unrealistic in most circumstances.

Rejection of such a hypothesis does not provide useful information since the investigator needs to know more than the fact that the observed level of agreement is unlikely to be due to chance.

9.1 CONTINUOUS MEASUREMENTS

9.1.1 THE INTRACLASS CORRELATION

The Intraclass Correlation (ICC) assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. The theoretical formula for the ICC is:

$$\rho = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_\epsilon^2}$$

where

- σ_S^2 is the **between subjects** variability
- σ_ϵ^2 is the **within subjects** variability

These quantities are directly estimable from ANOVA analyses.

In a one-way ANOVA with K groups, we have the ANOVA table

Source	D.F.	Sum of squares	Mean square
Between Samples	$K-1$	FSS	$FSS/(K-1)$
Within Samples	$n-K$	RSS	$RSS/(n-K)$
Total	$n-1$	TSS	

where

$$TSS = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\overline{y}_{..})^2 \quad RSS = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\overline{y}_{k})^2 \quad FSS = \sum_{k=1}^{K} n_k(\overline{y}_{k}-\overline{y}_{..})^2$$

Then

$$\widehat{\sigma}^2_{\epsilon} = \frac{RSS}{n-K} \qquad \widehat{\sigma}^2_{S} = \frac{FSS/(K-1) - RSS/(n-K)}{K}$$

9.1.2 DIFFERENT TYPES OF ICC

In their paper, Shrout and Fleiss (1979) describe three classes of ICC for reliability, which they term Case 1, Case 2 and Case 3. Each Case applies to a different rater agreement study design.

- **Case 1:** Raters for each subject are selected at random
  - This case has a pool of raters. For each subject, one randomly samples from the rater pool k different raters to rate this subject. Therefore the raters who rate one subject are not necessarily the same as those who rate another. This design corresponds to a one-way ANOVA in which Subject is a random effect, and Rater is viewed as measurement error.

- **Case 2:** The same raters rate each case. These are a random sample.
  - The same set of k raters rate each subject. This corresponds to a fully-crossed (rater $\times$ subject) two-way ANOVA design in which both Subject and Rater are separate effects.
  - In Case 2, Rater is considered a **random effect**; this means the k raters in the study are considered a random sample from a population of potential raters.
  - The Case 2 ICC estimates the reliability of the larger population of raters.

- **Case 3:** The same raters rate each case. These are the only raters.
  - This is similar to Case 2; a fully-crossed, two-way ANOVA design. But here one estimates the ICC that applies only to the k raters in the study. Since this does not permit generalization to other raters, the Case 3 ICC is not often used.

Shrout and Fleiss (1981) also show that for each of the three Cases above, one can use the ICC in two ways:

- To estimate the reliability of a single rating, or
- To estimate the reliability of a mean of several ratings.

For each of the Cases, then, there are two forms, producing a total of 6 different versions of the ICC.

9.2 DISCRETE MEASUREMENTS

9.2.1 THE KAPPA STATISTIC

Options for discrete data observer agreement analysis are rather more limited; one simple measure of agreement between two raters is the **Kappa Statistic.**

For a $K \times K$ table of results for the observer assessments of two observers on a categorical scale,let n_{ij} is the number of times rater 1 accords a measure i whilst Rater 2 accords a measure j, for $i, j = 1, ..., K$.

- "Considerable Agreement": Diagonal elements "large"
- "Low Agreement": Off-diagonal elements "large"

EXAMPLE: Assessment of xeromammograms by two radiologists

	Radiologist B				
Radiologist A	Normal	Benign	Suspected	Cancer	Total
Normal	21	12	0	0	33
Benign	4	17	1	0	22
Suspected	3	9	15	2	29
Cancer	0	0	0	1	1
Total	28	38	16	3	85

Proportion of Agreements:

$$p_A = \frac{n_A}{n} = \frac{(21 + 17 + 15 + 1)}{85} = 0.64$$

However, this does not take into account/quantify the probability of "chance" agreements; this can be measured by the expected number of chance agreements

$$\widehat{n}_A = \frac{33 \times 28}{85} + \frac{22 \times 38}{85} + \frac{29 \times 16}{85} + \frac{1 \times 3}{85} = 26.2$$

which gives a proportion

$$\widehat{p}_A = \frac{\widehat{n}_A}{n} = \frac{26.2}{85} = 0.31$$

Hence the "excess agreement" in the observed data is

$$\kappa = \frac{p_A - \widehat{p}_A}{1 - \widehat{p}_A}$$

which is termed the **Kappa Statistic**. Guidelines for interpretation of κ are

- $\kappa \leq 0.20 \Longrightarrow$ Poor Agreement
- $0.20 < \kappa \leq 0.40 \Longrightarrow$ Fair Agreement
- $0.40 < \kappa \leq 0.60 \Longrightarrow$ Moderate Agreement
- $0.60 < \kappa \leq 0.80 \Longrightarrow$ Good Agreement
- $0.80 < \kappa \leq 1.00 \Longrightarrow$ Very Good Agreement

Standard errors for κ are also available

9.2.2 WEIGHTED KAPPA

A weighted version of the Kappa Statistic can be used to reflect the ordinal nature of many observation scales (e.g. Normal⟶Benign⟶Suspected⟶Cancer)

Each off-diagonal element in the agreement table is given a weight reflecting how "severe" the disagreement is; usually the weights are proportional to the distance from the diagonal. This gives a **weighted kappa**, κ_W

$$\kappa_W = \frac{p_A^{(W)} - \widehat{p}_A^{(W)}}{1 - \widehat{p}_A^{(W)}}$$

where

$$p_A^{(W)} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij} n_{ij} \qquad \widehat{p}_A^{(W)} = \frac{1}{n^2}\sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij} n_{i.} n_{.j}$$

where $n_{i.}, n_{.j}$ are the row an column totals for row i and column j respectively.

SECTION 10.

SURVIVAL ANALYSIS

Survival (or lifetime, or time-to-event) analysis is a special type of regression modelling that explains the observed variability in a **response** variable Y via consideration of **predictors** $X = (X_1, ..., X_K)$. The principal difference between survival analysis and conventional regression is that account is taken of potential **censoring** in the response variable

- we may observe some actual responses (survival, failure) times, but also some censored responses where we do not observe an actual failure but rather only that the failure occurs after a **censoring time** (the end of study) – this is called **right-censoring**

- the response data is thus bivariate (Y, Z) where Y is the time at which the response is measured, and

$$Z = \begin{cases} 1 & \text{Failure is observed} \\ 0 & \text{Censored} \end{cases}$$

- occasionally, we observe **left-censoring** or **interval-censoring**

The potential presence of censoring fundamentally changes how we view the modelling process; previously we have looked at probability densities and Expected responses.

We now take an alternative view, and examine **survivor** and **hazard** functions.

10.1 THE SURVIVOR FUNCTION

The probability **density function** for response variable Y is f_Y, and the expectation, likelihood function and so on that are required for regression modelling are formed from f_Y. The **distribution function** F_Y is

$$F_Y(y) = P[Y \leq y] = \int_0^y f_Y(t)\, dt$$

In conventional regression modelling, the likelihood contribution for data point i with response y_i is $f_Y(y_i)$. For right-censored data with censoring at y_i, however, the likelihood contribution is

$$P[Y > y_i] = 1 - F_Y(y_i)$$

(i.e. we have "observed" that $Y_i > y_i$, the survival was at least y_i). This motivates consideration of the **survivor (reliability) function**

$$S_Y(y) = 1 - F_Y(y)$$

The likelihood function is thus

$$\left\{ \prod_{i:Z_i=1} f_Y(y_i) \right\} \times \left\{ \prod_{i:Z_i=0} S_Y(y_i) \right\}$$

that is

LIKELIHOOD FOR UNCENSORED DATA

$\times$

LIKELIHOOD FOR CENSORED DATA

and the role of the predictors can be introduced via the parameters of f_Y and F_Y.

10.2 THE HAZARD FUNCTION

As a further alternative method of specification, we consider the **hazard function**

$$h_Y(y) = P\left[\text{Failure at } y | \text{Survival } \geq y\right]$$

$$= \frac{f_Y(y)}{S_Y(y)}$$

and the **integrated hazard**

$$H_Y(y) = \int_0^y h_Y(t)\, dt$$

and it can be shown that

$$S_Y(y) = \exp\left\{-H_Y(y)\right\}$$

10.3 THE KAPLAN-MEIER CURVE

The **Kaplan-Meier curve** is a non-parametric estimate of the survivor function; it takes into account the censored data and produces a decreasing "step-function" curve, where the downward steps take place at the times of the failures, giving the estimated survival function at the jth failure/censoring time as

$$\widehat{S}_j = \prod_{i=1}^{j} \left(1 - \frac{z_i}{n-i+1}\right)$$

This curve can be used to report an estimated survival probability at a given time (1 year, 5 years etc.).

Standard errors for these estimated survival probabilities are also available.

10.4 THE COX REGRESSION MODEL

The **Cox** (or **Proportional Hazards**) model provides a simple way of introducing the influence of predictors into the survival model. The basic components are a **baseline hazard** function, h_0 and a linear predictor and (positive) link function g (similar to the GLM modelling of previous chapters). Then for an experimental unit with observed predictor values $X_1 = x_1$, $X_2 = x_2, ..., X_K = x_K$, the hazard function takes the form

$$h_Y(y; x) = g(x^T \beta) h_0(y)$$

that is, the hazard is modified in a multiplicative fashion by the linked-linear predictor.

Typically, g is the exponential function.

From the previously established relationships,

$$S_Y(y;x) = \exp\left\{-\int_0^y h_Y(t)\ dt\right\} = \exp\left\{-\int_0^y g(x^T\beta)h_0(y)\ dt\right\}$$

If a coefficient β_k is positive, the hazard is **increased**, and the expected failure time **decreased.**

The relevance/significance of a particular predictor is assessed using a **Wald** test based on the magnitude of

$$\frac{\widehat{\beta}}{s.e.\left(\widehat{\beta}\right)}$$

10.5 THE ACCELERATED LIFE MODEL

The **Accelerated Life** model provides another way of introducing the influence of predictors into the survival model. The basic components now are a **baseline survivor** function, S_0 and a linear predictor and (positive) link function g as above. Then for an experimental unit with observed predictor values $X_1 = x_1$, $X_2 = x_2, ..., X_K = x_K$, the survivor function takes the form

$$S_Y(y; x) = S_0(g(x^T\beta)y)$$

that is, the time scale is modified in a multiplicative fashion by the linked-linear predictor.

Again, typically, g is the exponential function. This model allows direct modelling of the influence of predictors on survival.

10.6 THE LOG-RANK TEST

The **log-rank** test is a standard test for significant differences between two (or more) survivor functions that differ because of the influence of the different levels of a discrete predictor.

$$\begin{aligned} H_0 &: S_1 = S_2 \\ H_1 &: S_1 \neq S_2 \end{aligned}$$

It is a non-parametric test based on ranks of samples for the two or more subgroups.

Asymptotic or exact versions of the test can be carried out; SPSS and other packages give further alternatives.

10.7 PARAMETRIC MODELLING

It is possible to fit and compare **parametric** survival models to such data. Parametric densities, survivor functions, hazards etc. can be readily used in the formation of a likelihood, potentially within the proportional hazards/accelerated life framework.

Typical models used are

- Weibull
- Gamma
- Log-Logistic
- Log-Normal
- Pareto