BIOSTATISTICS METHODS

FOR TRANSLATIONAL & CLINICAL RESEARCH

DESIGNING CLINICAL RESEARCH

Part B: Statistical Issues

Relative Risk & Odds Ratio

Suppose that each subject in a large study, at a particular time, is classified as positive or negative according to some risk factor, and as having or not having a certain disease under investigation. For any such categorization the population may be enumerated in a 2x2 table, as follows:

	Disease		Total
Factor	Yes (+)	No (-)	
Exposed (+)	A	B	A + B
Un-exposed (-)	C	D	C + D
Total	A + C	B + D	N = A + B + C + D

	Disease		Total
Factor	Yes (+)	No (-)	
Exposed (+)	A	B	A + B
Un-exposed (-)	C	D	C + D
Total	A + C	B + D	N = A + B + C + D

The entries A, B, C and D in the table are sizes of the four combinations of disease presence-and-absence and factor presence-and-absence and the number N at the lower right corner of the table is the total population size. The relative risk is

$$RR = \frac{A}{A+B} \div \frac{C}{C+D}$$

$$= \frac{A(C+D)}{C(A+B)}$$

$$RR = \frac{A}{A+B} \div \frac{C}{C+D}$$

$$= \frac{A(C+D)}{C(A+B)}$$

In many situations, the number of subjects classified as disease positive is very small as compared to the number classified as disease negative, that is,

$$C + D \cong D$$

$$A + B \cong B$$

$$RR \cong \frac{AD}{BC}$$

$$= \frac{A/B}{C/D} = \frac{A/C}{B/D}$$

$$RR \cong \frac{AD}{BC}$$

$$= \frac{A/B}{C/D} = \frac{A/C}{B/D}$$

The resulting ratio, AD/BC, is an approximate relative risk, but it is often referred to as *odds ratio* because

- **A/B and C/D are the *odds* in favor of having disease from groups with or without the factor;**
- **A/C and B/D are the odds in favor of having exposed to the factors from groups with or without the disease.**
- **The two odds, A/C and B/D, can be easily estimated in case-control studies, by using sample frequencies, a/c and b/d.**

Modeling a Proportion

Regressiondata are in the form :

$$\{(y_i; x_{1i}, x_{2i}, \cdots, x_{ki})\}_{i=1,\cdots,n}$$

In "regular" Regression Model, we impose the condition that Y is on the continuous scale – We even assume that Y is normally distributed with a constant variance.

We impose the condition that Y is on the continuous scale maybe because of the "normal error model" - not because Y is always on the continuous scale. In a variety of applications, the Dependent Variable of interest may have only two possible outcomes, and therefore can be represented by an Binary or Indicator Variable Y taking on values 0 and 1.

Let:

$\pi = Pr(Y=1)$

Let Y be the Dependent Variable Y taking on values 0 and 1, and:

π = Pr(Y=1)

Y is said to have the "Bernouilli distribution" (Binomial with n = 1). We have:

$$E(Y) = \pi$$

$$Var(Y) = \pi(1-\pi)$$

Consider, for example, a study of Drug Use among middle school kids, as a function of gender and age of kid, family structure (e.g. who is the head of household), and family income. In this study, the dependent variable Y was defined to have two possible outcomes:

(i) Kid uses drug (Y=1), and
(ii) Kid does not use drug (Y=0).

In another example, say, a man has a physical examination; he's concerned: Does he have prostate cancer? The "truth" would be confirmed by a biopsy. But it's a very painful process (at least, could we say if he needs a biopsy?)

In this study, the dependent variable Y was defined to have two possible outcomes:

(i) Man has prostate cancer (Y=1), and
(ii) Man does not have prostate cancer (Y=0).

Possible predictors include PSA level, age, race.

Or in a case-control study, the response variable is the “disease” – a binary indicator, often code as 0/1.

The basic question is: Can we do “regression” when the dependent variable, or “response”, is binary?

For "binary" Dependent Variables, we run into problems with the "Normal Error Model" – The distribution of Y is Bernouilli. However, the "normal" assumption is not very important (i.e. "robust"); effects of any violation is quite minimal – especially if n is large!

The Mean of Y is in well-defined but it has limited range:

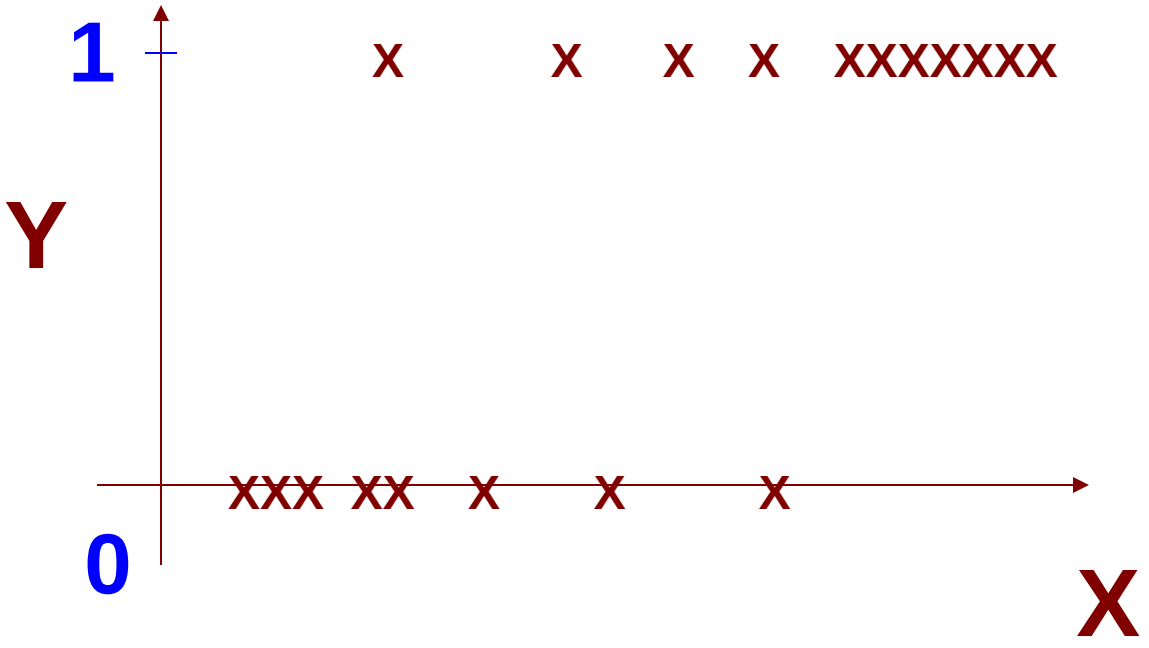
Mean of Y = Pr(Y=1) = π
$0 \leq \pi \leq 1$,

and fitted values may fall outside of (0,1). However, that's a minor problem.

The Variance (around the regression line) is not constant (a model violation that we learn in diagnostics); variance is function of the Mean π of Y (which is a function of predictors):

$$\sigma^2 = \pi(1- \pi)$$

More important, the relationship is not linear. For example, with one predictor X, we usually have:



In other words, We still can focus on "modeling the mean", in this case it is a Probability, π = Pr(Y=1), but the usual linear regression with the "normal error regression model" is definitely not applicable – all assumptions are violated, some may carry severe consequences.

So, what should we do?

We have to get out of that limited range (0,1); we need some transformation of (binary variable) Y

EXAMPLE: Dose-Response

Data in the table show the effect of different concentrations of (nicotine sulphate in a 1% saponin solution) on fruit flies; here X = log(100xDose), just making the numbers easier to read.

Dose(gm/100cc)	# of insects, n	# killed, r	x	p (%)
0.1	47	8	1.000	17.0
0.15	53	14	1.176	26.4
0.2	55	24	1.301	43.6
0.3	52	32	1.477	61.5
0.5	46	38	1.699	82.6
0.7	54	50	1.845	92.6
0.95	52	50	1.978	96.2

Proportion p is an estimate of Probability π

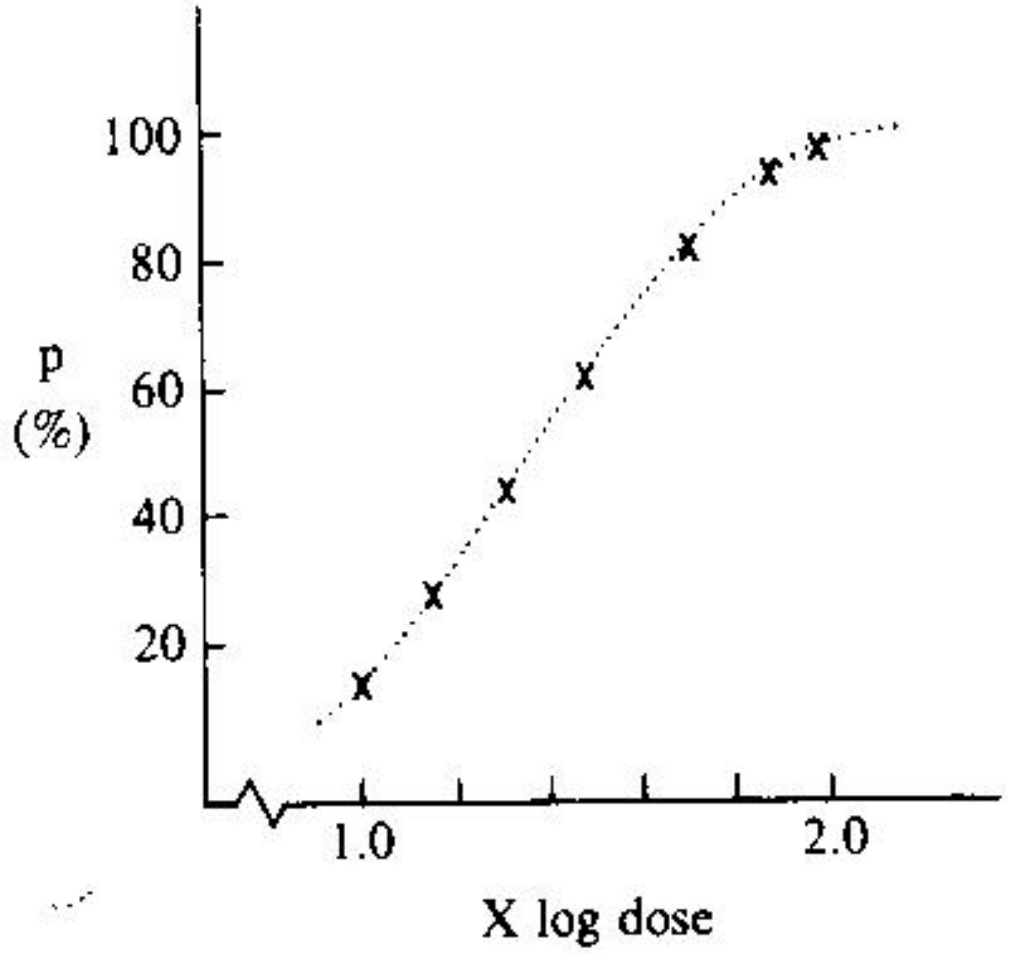
UNDERLYING ASSUMPTION

It is assumed that each subject/fly has its own tolerance to the drug. The amount of the chemical needed to kill an individual fruit fly, called "individual lethal dose" (ILD), cannot be measured - because only one fixed dose is given to a group of n flies (indirect assay)
(1) If that dose is below some particular fly's ILD, the insect survived.
(2) Flies which died are those with ILDs below the given fixed dose.

INTERPRETATION OF DATA

Dose	# n	# killed	X	p(%)
0.1	47	8	1.000	17.0
0.15	53	14	1.176	26.4
0.2	55	24	1.301	43.6
0.3	52	32	1.477	61.5
0.5	46	38	1.699	82.6
0.7	54	50	1.845	92.6
0.95	52	50	1.978	96.2

- **17% (8 out of n=47) of the first group respond to dose of .1gm/100cc (x=1.0); that means 17% of subjects have their ILDs less than .1**
- **26.4% (14 out of n=53) of the 2nd group respond to dose of .15gm/100cc (X=1.176); that means 26.4% of subjects have their ILDs less than .15**
- **we view each dose D (with X = logD) as upper endpoint of an interval and p the cumulative relative frequency.**



A symmetric sigmoid dose-response curve suggests that it be seen as some cumulative distribution function (cdf).

"Empirical evidence", i.e. data, suggest that we view p the cumulative relative frequency. This leads to a "transformation" from "π" to an "upper endpoint", say Y* (which is on the continuous scale) corresponding to that cumulative frequency of some cdf. After this transformation, the regression model is then imposed on Y*, transformed value of π

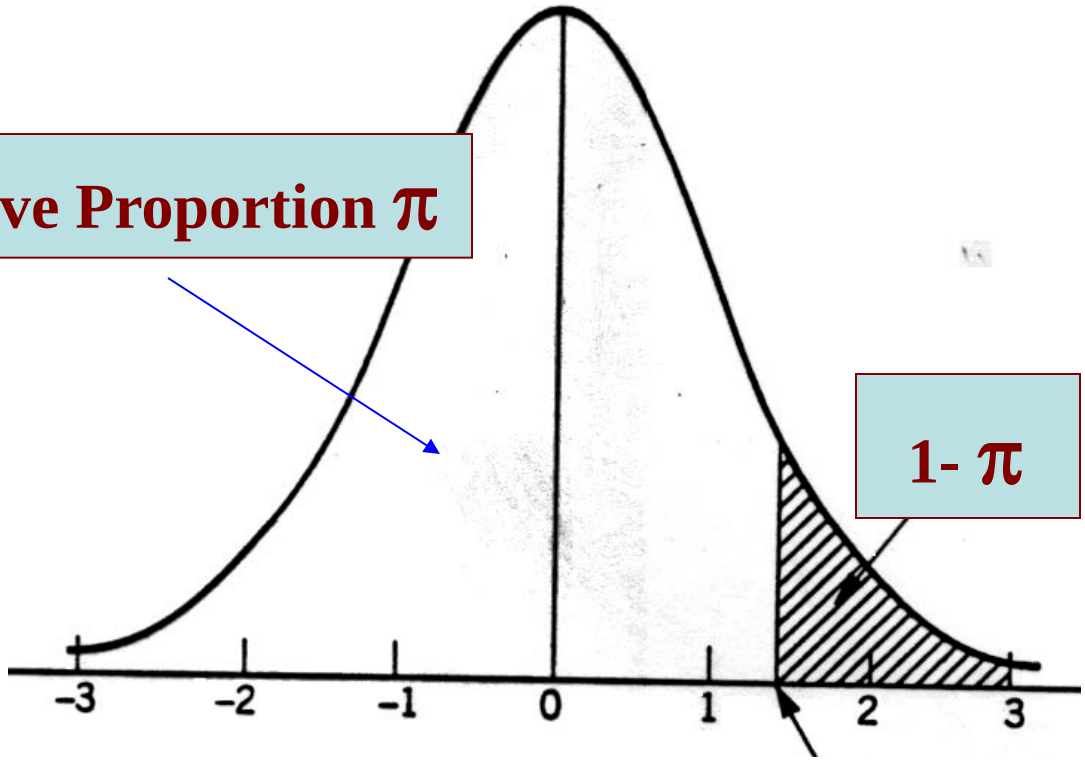
MODELING A PROBABILITY

Let π be the probability "to be modeled" and X a covariate (let consider only one X for simplicity). The first step in the regression modeling process is to obtain "the equivalent deviate Y* of π" using the following transformation:

$$\pi = \int_{-\infty}^{Y^*} f(z)dz \quad \text{or} \quad \int_{Y^*}^{\infty} f(z)dz$$

f(z) is some probability density function.

Cumulative Proportion π



Transformation: π to Y*

which is on a linear scale

As a result, the proportion π has been transformed into a variable Y* on the "linear" or continuous scale with unbounded range. We can use Y* as the dependent variable in a regression model. (We now should only worry about "normality" which is not very important)

The relationship between covariate X (in the example, log of the dose) or covariates X's and Probability π (through Y) is then stipulated by the usual simple linear regression:

$$Y^* = \beta_0 + \beta_1 x$$

or multiple regression:

$$Y^* = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

All we need is a "probability density function" f(.) in order to translate π to Y^* through :

$$\pi = \int_{-\infty}^{Y^*} f(z)dz \quad \text{or} \quad \int_{Y^*}^{\infty} f(z)dz$$

In theory, any probability density function can be used. We can choose one either by its simplicity and/or its extensive scientific supports. And we can check to see if the data fit the model (however, it's practically hard because we need lots of data to tell).

A VERY SIMPLE CHOICE

A possibility is "Unit Exponential Distribution" with density :

$$f(z) = e^{-z};\ z \geq 0$$

Result (for one covariate X) is:

$$\pi = \int_{-\beta_0-\beta_1 x}^{\infty} e^{-z} dz$$

$$= e^{\beta_0+\beta_1 x}; \text{ or}$$

$$\ln\pi = \beta_0 + \beta_1 x$$

That is to model the "log" of the probability as a "linear function" of covariates.

The advantage of the approach of modeling the "log" of the probability as a "linear function" of covariates, is easy interpretation of model parameters, the probability is changed by a multiple constant (i.e."multiplicative model" which is usually plausible)

REGRESSION COEFFICIENTS

- If X_1 is binary (=0/1) representing an exposure, β_1 represents the (log of) the "odds" (of having the event represented by Y) associated with the exposure – adjusted for that of X_2
- If X_1 is on a continuous scale, β_1 represents the (log of) the "odds" (of having the event represented by Y) associated with one unit increase in the value of X_1 - adjusted for X_2

A HISTORICAL CHOICE

Besides the Unit Exponential probability density, one can also use of the Standard Normal density in the transformation of π:

$$\pi = \int_0^{y^*} f(\theta(\theta)$$

"f" is the Standard Normal density :

$$f(\theta(= \frac{1}{\sqrt{2\pi}} \exp(-\frac{\theta^2}{2})$$

This "Probit Transformation" leads to the "Probit Model"; Y* is called the "probit" of π. The word "probit" is a shorten form of the phrase "PROBability unIT" (but it is not a probability), it is a standard normal variate.

The Probit Model was popular in years past and had been used almost exclusively to analyze "bioassays" for many decades. However, there is no closed-form formula for Y* (it's not possible to derive an equation relating π to x without using an integral sign):

$$\pi = \int_0^{\beta_0+\beta_1 x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right) d\theta$$

LOGISTIC TRANSFORMATION

(Standard) Logistic Distribution with density :

$$f(\theta) = \frac{\exp(\theta x}{[1 + \exp(\theta x)]^2}$$

Result is:

$$\pi = \int_{-\infty}^{Y^* = \beta_0 + \beta x} \frac{e^{\theta}}{[1 + e^{\theta}]^2} d\theta$$

$$= \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$= \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}$$

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

$$1 - \pi = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}$$

$$\frac{\pi}{1 - \pi} = e^{\beta_0 + \beta_1 x}$$

$$\log \frac{\pi}{1 - \pi} = \beta_0 + \beta_1 x$$

We refer to this as “Logistic Regression”

Exponential transformation leads to a linear model of "Log of Probability": ln(π); Logistic transformation leads to a linear model of "Log of Odds": ln[π/(1- π)] – also called "logit"

When π is small (rare disease/event), probability and the odds are approximately equal.

$$\text{Odds} = \frac{\pi}{1-\pi}$$

$$\pi = \frac{\text{Odds}}{1+\text{Odds}}$$

Advantages:

(1) Also very simple data transformation:
$Y = \log\{p/(1-p)\}$

(2) The logistic density, with thicker tails as compared to normal curve, may be a better representation of real-life processes (compared to Probit Model which is based on the normal density).

A POPULAR MODEL

- **Although one can use the Standard Normal density in the regression modeling process (or any density function for that purpose),**
- **The Logistic Regression, as a result of choosing Logistic Density remains the most popular choice for a number of reasons: closed form formula for π, easy computing (Proc LOGISTIC)**
- **The most important reasons: interpretation of model parameter and empirical supports!**

REGRESSION COEFFICIENTS

- **If X_1 is binary (=0/1) representing an exposure, β_1 represents the (log of) the "odds ratio" (of having the event represented by Y) associated with the exposure – adjusted for that of X_2**
- **If X_1 is on a continuous scale, β_1 represents the (log of) the "odds ratio" (of having the event represented by Y) associated with one unit increase in the value of X_1 - adjusted for X_2**

Example : Say X_1 is binary

$$\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$x_1 = 0 \text{ (unexposed)} : \ln Odds_{unexposed} = \beta_0 + \beta_2 x_2$$

$$x_1 = 1 \text{ (exposed)} : \ln Odds_{exposed} = \beta_0 + \beta_1 + \beta_2 x_2$$

$$\ln Odds_{exposed} - \ln Odds_{unexposed} = \beta_1$$

$$\frac{Odds_{exposed}}{Odds_{unexposed}} = \text{Odds Ratio} = (e^{\beta_1})$$

β_1 is the odds ratio on the log scale if X is binary

Logistic Regression applies in both prospective and retrospective (case-control) designs. In prospective design, we can calculate/estimate the probability of an event (for specific values of covariates). In retrospective design, we cannot calculate/estimate the probability of events because the "intercept" is meaningless but relationship between event and covariates are valid.

SUPPORTS FOR LOGISTIC MODEL

The fit and the origin of the linear logistic model could be easily traced as follows. When a dose D of an agent is applied to a pharmacological system, the fractions f_a and f_u of the system affected and unaffected satisfy the so-called "median effect principle" (Chou, 1976):

$$\frac{f_a}{f_u} = \left\{ \frac{d}{ED_{50}} \right\}^m$$

where ED_{50} is the "median effective dose" and "m" is a Hill-type coefficient; m = 1 for first-degree or Michaelis-Menten system. The median effect principle has been investigated much very thoroughly in pharmacology. If we set " $\pi = f_a$", the median effect principle and the logistic regression model are completely identical with a slope $\beta_1 = m$.

PARAMETER ESTIMATION: MLE

Model:

$$\pi = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\alpha + \beta x)}$$

Likelihood Function :

$$L = \prod_{i=1}^{n} \Pr(Y_i = y_i)$$

$$= \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

$$= \prod_{i=1}^{n} \frac{\{\exp[\beta_0 + \beta_1 x_i]\}^{y_i}}{1 + \exp[\beta_0 + \beta_1 x_i]}; y_i = 0/1$$

An Application: Quantal Assays

Indirect Assays: Dose fixed, Response random.

Depending on the "measurement scale" for the response (our random variable), we divide indirect assays into two groups:

(1) Quantal assays, where the response is binary: whether or not an event (like the death of the subject) occurs,

(2) Quantitative assays, where measurements for the response are on a continuous scale.

Quantal response assays belong to the class of qualitative indirect assays. They are characterized by experiments in which each level of a stimulus (eg. dose of a drug) is applied to n experimental units; r of them respond and (n-r) do not response.That is "binary" response (yes/no). The group size "n" may vary from dose to dose; in theory, some n could be 1 (so that r = 0 or 1).

THE ASSAY PROCEDURE

- **The usual design consists of a series of dose levels with subjects completely randomized among/to the dose levels. The experiment may include a standard and a test preparations; or maybe just the test.**
- **The dose levels chosen should range from "very low" (few or no subjects would respond) to "rather high" (most or all subjects would respond).**
- **The objective is often to estimate the LD50; the number of observations per preparation depends on the desired level of precision of its estimate – sample size estimation is a very difficult topic.**

The most popular parameter LD50 (for median lethal dose), or ED50 (for median effective dose), or EC50 (for median effective concentration) is the level of the stimulus which result in a response by 50% of individuals in a population.

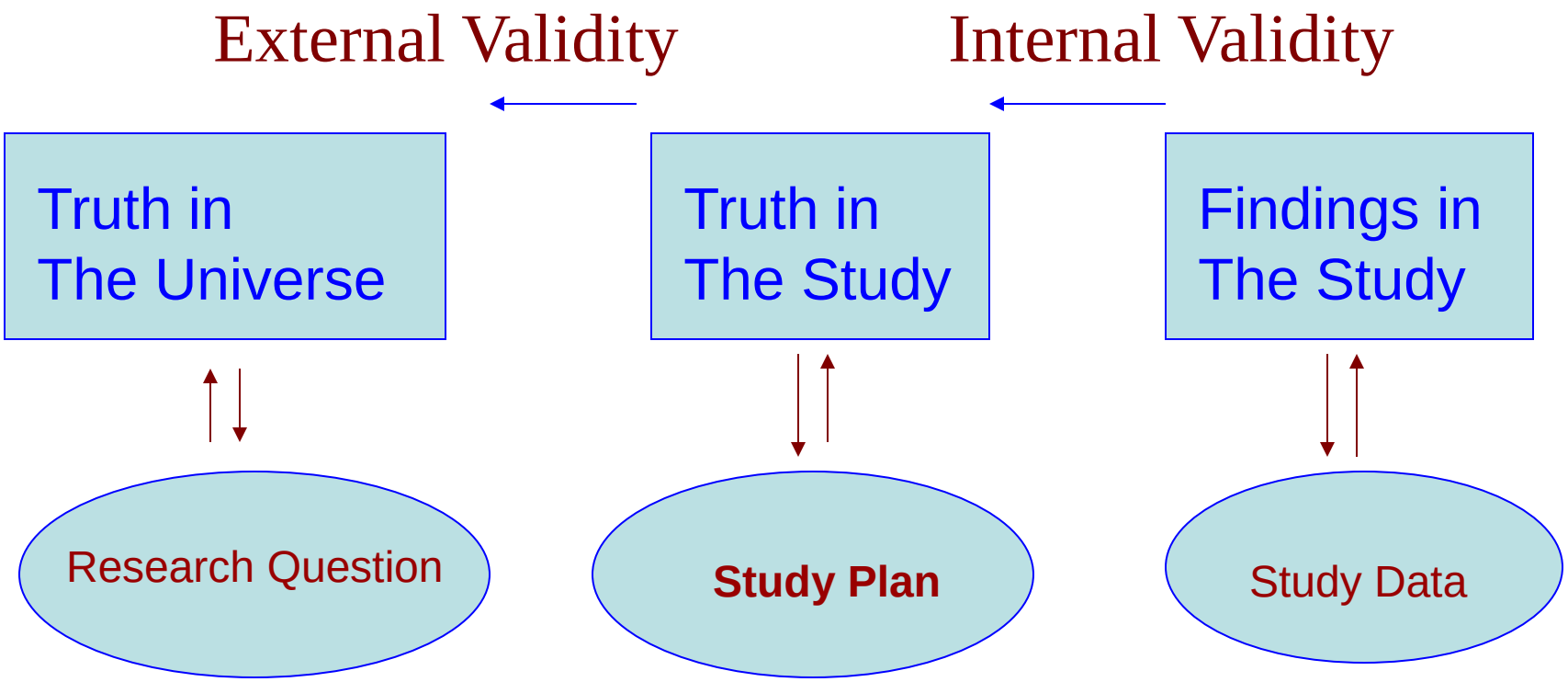
(1) It is a measure of the agent's potency, which could be used to form relative potency.

(2) It is chosen by a statistical reason; for any fixed number of subjects, one would attain greater precision as compared to estimating, say, LD90 or LD10 or any other percentiles.

Inference & Validity

INFERENCES & VALIDITIES

- **Two major levels of inferences are involved in interpreting a study, a clinical trial**
- **The first level concerns Internal validity; the degree to which the investigator draws the correct conclusions about what actually happened in the study.**
- **The second level concerns External Validity (also referred to as generalizability or inference); the degree to which these conclusions could be appropriately applied to people and events outside the study.**


External Validity
Internal Validity
Truth in
The Universe
Truth in
The Study
Findings in
The Study
Research Question
Study Plan
Study Data

A Simple Example:

An experiment on the effect of Vitamin C on the prevention of colds could be simply conducted as follows. A number of n children (the sample size) are randomized; half were each give a 1,000-mg tablet of Vitamin C daily during the test period and form the "experimental group". The remaining half , who made up the "control group" received "placebo" – an identical tablet containing no Vitamin C – also on a daily basis. At the end, the "Number of colds per child" could be chosen as the outcome/response variable, and the means of the two groups are compared.

Assignment of the treatments (factor levels: Vitamin C or Placebo) to the experimental units (children) was performed using a process called "randomization". The purpose of randomization was to "balance" the characteristics of the children in each of the treatment groups, so that the difference in the response variable, the number of cold episodes per child, can be rightly attributed to the effect of the predictor – the difference between Vitamin C and Placebo. Randomization helps to assure Internal Validity.

What about External Validity?

For instance, when looked to establish a relationship but found no statistically significant correlation. On this basis it is *concluded* that there is no relationship between the two factors. How could this conclusion be wrong -- that is, what are the "threats to validity"? A "conclusion" is a generalization from findings of the study to truth in the universe; it involves external validity.

REJECTION

- **Back to a "Statistical Test".**
- **The Null Hypothesis H_0 is a "Theory". The Data are "Reality". When they do not agree, then we have to trust the reality; That's when H_0 is rejected.**
- **How do we tell if Theory and Reality do not agree? When the data show overwhelmingly that it is almost impossible to have the data that we already collected if H_0 is true (that it is possible but with a very small probability).**

ABOUT REJECTIONS

Hypothesis Testing is like Trial by Jury

A very important concept: when a null hypothesis is not rejected it does not necessarily lead to its acceptance, because a "not guilty" verdict is just an indication of "lack of evidence" and "innocence" is just one of its possibilities. That is, when a difference in not statistically significant, there are still two possibilities:

(i)The null hypothesis is true,

(ii)There is not enough evidence from sample data to support its rejection (i.e. not enough data).

Studies may be inconclusive because they were poorly planned, not enough data were collected to accomplished the goals and support the hypotheses.

To assure external validity, we have to assure of adequate sample size.

A TYPICAL SCENARIO

An investigator wants to randomize mice with induced tumors – say, lung tumors - into two groups; mice in one group receive placebo and the others some new agent – the effect of the new agent is to reduce the size/volume of the tumors. And he/she needs help to figure out the sample sizes.

A TYPICAL “PRODUCT”

A statement such as “with 15 mice per group, we would be able to detect – with a statistical power of 80% - a reduction of 40% in tumor volume using a two-sided two-sample t-test at the 5% level”

Where do we get that 40% tumor volume specified in the Alternative Hypothesis? Whose responsibility, investigator's or statistician's?

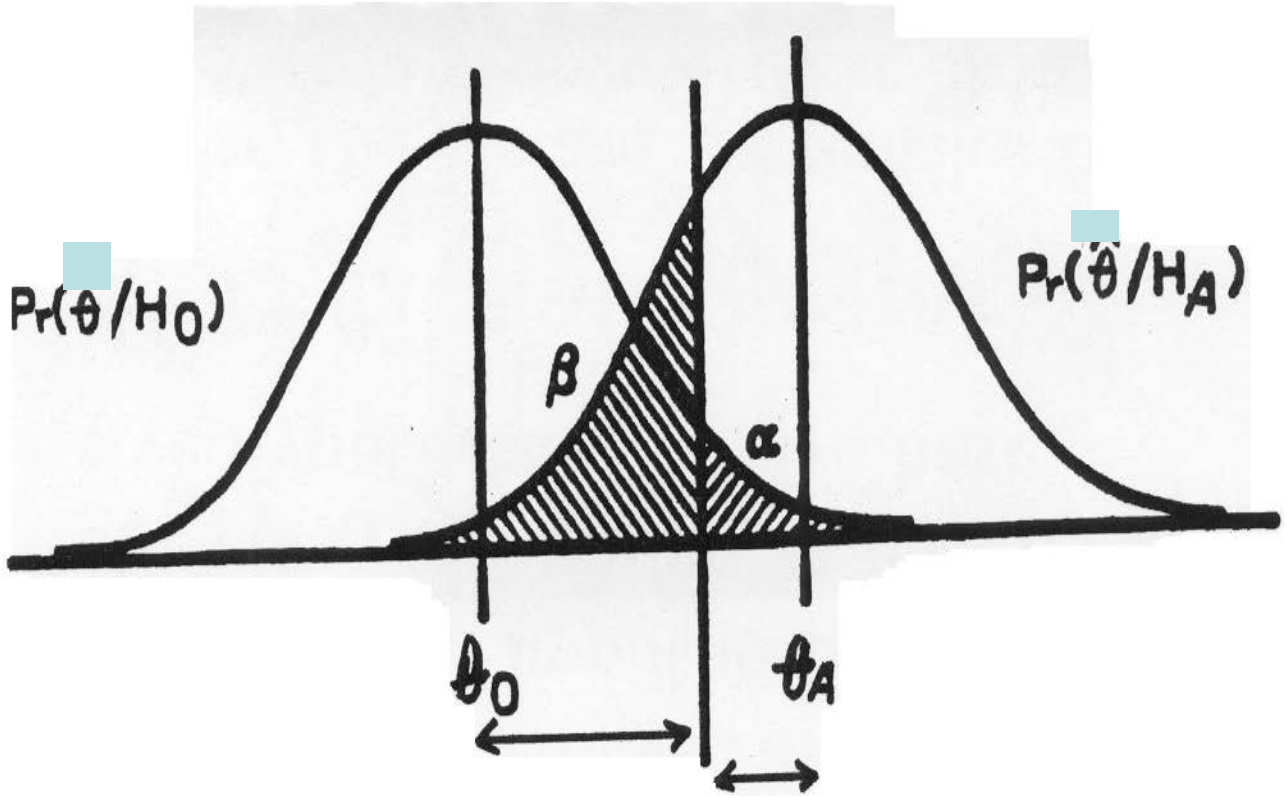
It's easier in basic sciences with animal studies: it's the "dose" – previously determined from a one-arm trial: ED40

It could be a problem when both sides are inexperienced. Sometimes the investigator simply makes up an effect size. Sometimes they take an easy way out: See how much the investigator can afford – say n = 12 per group – then how much we can "see" or detect – a reduction of 27% or 32% etc…

Sample Size Determination

APPROACH TO SAMPLE SIZE

- **The target of the investigation is a statistic θ; for example, the difference of two sample means or two sample proportions.**
- **Consider the statistic θ which often the MLE of some parameter (e.g. the difference of two population means), and assume that it is normally distributed as $N(\theta_0, \Sigma_0^2)$ under the null hypothesis H_0 and as $N(\theta_A, \Sigma_A^2)$ under an alternative hypothesis H_A; usually $\Sigma_0^2 = \Sigma_A^2$ or we can assume this equality for simplification.**



$$| \theta_0 - \theta_A | = z_{1-\alpha}\Sigma_0 + z_{1-\beta}\Sigma_A$$

MAIN RESULT

- **We have:**
  $| \theta_0 - \theta_A | = z_{1-\alpha}\Sigma_0 + z_{1-\beta}\Sigma_A$
  where the z's are percentiles of N(0,1).
- **Or if $\Sigma_0^2 = \Sigma_A^2 = \Sigma$, or if we assume this equality for simplification, then**
  $(\theta_0 - \theta_A)^2 = (z_{1-\alpha} + z_{1-\beta})^2 \Sigma^2$
- **This "the Basic Equation for Sample Size Determination"; and we use $z_{1-\alpha/2}$ if the statistical test is used as two-sided.**

DETECTION OF A CORRELATION

- **The Problem: To confirm certain level of correlation between two continuously measured variables**
- **The Null hypothesis to be tested is H_0: $\rho = \rho_0$, say $\rho = 0$.**
- **The Alternative hypothesis to be tested is H_0: $\rho = \rho_A$, say $\rho = .4$.**
- **The target statistic is Pearson's "r"; indirectly through Fisher's transformation to "z".**

The Coefficient of Correlation ρ between the two random variables X and Y is estimated by the (sample) Coefficient of Correlation r but the sampling distribution of r is far from being normal. Confidence intervals of is by first making the "Fisher's z transformation"; the distribution of z is normal if the sample size is not too small

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

$$z \in \text{Normal}$$

$$E(z) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

$$\sigma^2(z) = \frac{1}{n-3}$$

DETECTION A CORRELATION

- The null hypothesis to be tested is H_0: $\rho = 0$
- The target statistic is Fisher's z: $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$
- Basic parameters are:

$$\theta_0 = \frac{1}{2}\ln\frac{1+0}{1-0} = 0;\ \theta_A = \frac{1}{2}\ln\frac{1+\rho_A}{1-\rho_A};\ \text{and}\ \Sigma^2 = \frac{1}{n-3}$$

- Result: Total required sample size:

$$(\theta_0 - \theta_A)^2 = (z_{1-\alpha} + z_{1-\beta})^2 \Sigma^2$$

$$n = 3 + \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\theta_A^2}$$

RESULTS FOR CORRELATION

- **The null hypothesis to be tested is H_0: $\rho = 0$**
- **The target statistic is Fisher's z:** $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$
- **Basic parameters are:**

$$\theta_0 = \frac{1}{2}\ln\frac{1+0}{1-0} = 0;\ \theta_A = \frac{1}{2}\ln\frac{1+\rho_A}{1-\rho_A};\ \text{and}\ \Sigma^2 = \frac{1}{n-3}$$

- **Result: Total required sample size:**

$$n = 3 + \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\theta_A^2}$$

- **<u>Example</u>: If** $\rho_A = .4 \rightarrow \theta_A = .424$; for 5% 2 – sided & 80% power :

$$n = 3 + \frac{(1.96 + .84)^2}{.424^2} \geq 47$$

NEEDED COMPONENTS

- **This required total sample size is affected by three factors:**

(1) The size α of the test; conventionally, α =.05 is used.

(2) The desired power (1-β). This value is selected by the investigator; a power of 80% or 90% is often used.

$$n = 3 + \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\theta_A^2}$$

(3) The size of the coefficient of correlation to be detected:

$$\theta_A = \frac{1}{2} \ln \frac{1+\rho_A}{1-\rho_A}$$

It is obvious that the planning sample size is not easy; a good solution requires knowledge of the scientific problem, some good idea of what would be the alternative Hypothesis.

COMPARISON OF TWO MEANS

- **The Problem: The endpoint is on a continuous scale; for example, a researcher is studying a drug which is to be used to reduce the cholesterol level in adult males aged 30 and over. Subjects are to be randomized into two groups, one receiving the new drug (group 1), and one a look-alike placebo (group 2). The response variable considered is the change in cholesterol level before and after the intervention.**

❖ **The null hypothesis to be tested is H_0: $\mu_2 - \mu_1 = 0$**

❖ **The target statistic is $\theta = \bar{x}_2 - \bar{x}_1$**

DIFFERENCE OF TWO MEANS

- **The null hypothesis to be tested is H_0: $\mu_1 = \mu_2$**
- **The target statistic is $\theta = \bar{x}_2 - \bar{x}_1$**
- **Basic parameters are: $\theta_0 = 0$, $\theta_A = d$, and**

$$\Sigma^2 = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \sigma^2 \frac{4}{N}$$

Or: $$d^2 = (z_{1-\alpha} + z_{1-\beta})^2 \Sigma^2$$

$$(\theta_0 - \theta_A)^2 = (z_{1-\alpha} + z_{1-\beta})^2 \Sigma^2$$

RESULTS FOR TWO MEANS

- The null hypothesis to be tested is H_0: $\mu_1 = \mu_2$
- The target statistic is $\theta = \bar{x}_2 - \bar{x}_1$
- Basic parameters are: $\theta_0 = 0$, $\theta_A = d$, and

$$\Sigma^2 = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \sigma^2 \frac{4}{N}$$

- Or: $d^2 = (z_{1-\alpha} + z_{1-\beta})^2 \Sigma^2$

$$N = 4(z_{1-\alpha} + z_{1-\beta})^2 \frac{\sigma^2}{d^2}$$

NEEDED COMPONENTS

- **This required total sample size is affected by four factors:**

(1) The size α of the test; conventionally, α =.05 is used.

(2) The desired power (1-β). This value is selected by the investigator; a power of 80% or 90% is often used.

$$N = 4(z_{1-\alpha} + z_{1-\beta})^2 \frac{\sigma^2}{d^2}$$

NEEDED COMPONENTS

(3) The quantity d, called the "minimum clinical significant difference", $d = |\mu_2 - \mu_1|$, (its its determination is a clinical decision, not a statistical decision).

(4) The variance of the population. This variance σ^2 is the only quantity which is difficult to determine. The exact value is unknown; we may use information from similar studies or past studies or use some "upper bound".

$$N = 4(z_{1-\alpha} + z_{1-\beta})^2 \frac{\sigma^2}{d^2}$$

EXAMPLE

- **Specifications: Suppose a researcher is studying a drug which is used to reduce the cholesterol level in adult males aged 30 or over, and wants to test it against a placebo in a balanced randomized study. Suppose also that it is important that a reduction difference of 5 be detected (d=5). We decide to preset α =.05 and want to design a study such that its power to detect a difference between means of 5 is 95% (or β =.05). Also, the variance of cholesterol reduction (with placebo) is known to be about σ^2 = 36.**
- **Result:**

$$N = 4(1.96 + 1.65)^2 \frac{36}{5^2} = 76; \text{ or } 38 \text{ subjects in each group}$$

COMPARISON OF 2 PROPORTIONS

- **The Problem: The endpoint may be on a binary scale. For example, a new vaccine will be tested in which subjects are to be randomized into two groups of equal size: a control (not immunized) group (group 1), and an experimental (immunized) group (group 2). Subjects, in both control and experimental groups, will be challenged by a certain type of bacteria and we wish to compare the infection rates.**

❖ **The null hypothesis to be tested is $H_0: \pi_2 - \pi_1 = 0$**

❖ **The target statistic is $\theta = p_2 - p_1$**

DIFFERENCE OF 2 PROPORTIONS

- **The null hypothesis to be tested is H_0: $\pi_1 = \pi_2$**
- **The target statistic is $\theta = p_2 - p_1$**
- **Basic parameters are: $\theta_0 = 0$, $\theta_A = d$, and approximately**

$$\Sigma^2 = \bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right) = \bar{\pi}(1-\bar{\pi})\frac{4}{N}$$

$$\boxed{(\theta_0 - \theta_A)^2 = (z_{1-\alpha} + z_{1-\beta})^2 \Sigma^2}$$

Or

$$d^2 = (z_{1-\alpha} + z_{1-\beta})^2 \Sigma^2$$

RESULTS FOR 2 PROPORTIONS

- **The null hypothesis to be tested is H_0: $\pi_1 = \pi_2$**
- **The target statistic is $\theta = p_2 - p_1$**
- **Basic parameters are: $\theta_0 = 0$, $\theta_A = d$, and approximately**

$$\Sigma^2 = \bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right) = \bar{\pi}(1-\bar{\pi})\frac{4}{N}$$

Or

$$d^2 = (z_{1-\alpha} + z_{1-\beta})^2 \Sigma^2$$

$$N = 4(z_{1-\alpha} + z_{1-\beta})^2 \frac{\bar{\pi}(1-\bar{\pi})}{d^2}$$

NEEDED COMPONENTS

- **This required total sample size is affected by four factors:**
- **(1) The size α of the test; conventionally, α =.05 is used.**
- **(2) The desired power (1-β). This value is selected by the investigator; a power of 80% or 90% is often used.**

$$N = 4(z_{1-\alpha} + z_{1-\beta})^2 \frac{\bar{\pi}(1-\bar{\pi})}{d^2}$$

NEEDED COMPONENTS

(3) The quantity d, also called the "minimum clinical significant difference", $d = |\pi_2 - \pi_1|$ (its its determination is a clinical decision, not a statistical decision).

(4) π is the average proportion $\pi = (\pi_2 + \pi_1)/2$; It is obvious that the planning sample size is more difficult and a good solution requires knowledge of the scientific problem, some good idea of the magnitude of the proportions themselves.

$$N = 4(z_{1-\alpha} + z_{1-\beta})^2 \frac{\bar{\pi}(1-\bar{\pi})}{d^2}$$

EXAMPLE

- **Specifications: Suppose we wish to conduct a clinical trial of a new therapy where the rate of successes in the control group was known to be about 5%. Further, we consider the new therapy to be superior- cost, risks, and other factors considered- if its rate of successes is about 15%. In addition, We decide to preset α =.05 and want to design a study such that its power to detect the desired difference of 15% vs. 5% is 90% (or β =.10).**
- **Result:**

$$N = 4(1.96 + 1.28)^2 \frac{(.10)(.90)}{(.15 - .05)^2} = 378; \text{ or } 189 \text{ per group}$$

CASE-CONTROL STUDIES

- **Both cohort and case-control- are comparative; the validity of the conclusions is based on a comparison.**
- **In a cohort study, say a clinical trial, we compare the results from the "treatment group" versus the results from the "placebo group".**
- **In a case-control study, we compare the "cases" versus the "controls" with respect to an exposure under investigation ("exposure" could be binary or continuous).**

DIFFERENT FORMULATION

- In a cohort study, for example a two-arm clinical trial, the decision at the end is based on a "difference"; difference of two means or of two proportions. The "size" of the difference is the major criterion for sample size determination.
- In a case-control study, we compare the exposure histories of the two groups. At the end, we do not search for a difference; instead, the alternative hypothesis of a case-control study is postulated in the form of a relative risk. But the two are related.

CASE-CONTROL DESIGN FOR A BINARY RISK FACTOR

- The data analysis maybe similar to that of a Clinical Trial where we want to compare two proportions.
- However in the design stage, the alternative hypothesis is formulated in the form of a relative risk ρ. Since we cannot estimate or investigate "relative risk" using a case-control design, we would treat the given number ρ as an "odds ratio", the ratio of the odds of being exposed by a case divided by the odds of being exposed by a control.

$$\rho = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}$$

CLINICAL SIGNIFICANT DIFFERENCE

- From:

$$\rho = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}$$

- **We solve for the proportion for the cases, and use the previous formula for sample size applies with d = π_1 - π_0:**

$$\pi_1 = \frac{\rho\pi_0}{1+(\rho-1)\pi_0}$$

CASE-CONTROL DESIGN FOR A CONTINUOUS RISK FACTOR

- **Data are analyzed using Logistic Regression**
- **The Model is:**

$$p_x = \Pr(Y = 1 \mid X = x) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x)]}$$

$$\text{Logit} = \ln \frac{p_x}{1 - p_x} = \beta_0 + \beta_1 x$$

- **Key Parameter: β_1 is the log of the Odds Ratio due to one unit increase in the value of X**

BAYES' THEOREM

Recall:

$$\Pr(A \mid B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)} = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B \mid A)\Pr(A) + \mathrm{PR}(B \mid \text{not } A)\Pr(\text{not } A)}$$

For example,

$$\Pr(Y = 1 \mid X = x) = \frac{\Pr(X = x \mid Y = 1)\Pr(Y = 1)}{\Pr(X = x \mid Y = 1)\Pr(Y = 1) + \Pr(X = x \mid Y = 0)\Pr(Y = 0)}$$

$$\Pr(Y = 0 \mid X = x) = \frac{\Pr(X = x \mid Y = 0)\Pr(Y = 0)}{\Pr(X = x \mid Y = 0)\Pr(Y = 0) + \Pr(X = x \mid Y = 1)\Pr(Y = 1)}$$

Take the ratio, denominators are cancelled

APPLICATION TO LOGISTIC MODEL

We use the Bayes' Rule to express the ratio of posterior probabilities as the ratio of prior probabilities times the likelihood ratio:

$$\frac{\Pr(Y=1\mid X=x)}{\Pr(Y=0\mid X=x)}=\frac{\Pr(X=x\mid Y=1)\Pr(Y=1)}{\Pr(X=x\mid Y=0)\Pr(Y=0)}$$

$$\frac{\Pr(Y=1\mid X=x)}{\Pr(Y=0\mid X=x)}=\left\{\frac{\Pr(Y=1)}{\Pr(Y=0)}\right\}\left\{\frac{\Pr(X=x\mid Y=1)}{\Pr(X=x\mid Y=0)}\right\}$$

THE LOGISTIC MODEL

$$\left\{\frac{\Pr(Y=1|X=x)}{\Pr(Y=0|X=x)}\right\} = \left\{\frac{\Pr(Y=1)}{\Pr(Y=0)}\right\}\left\{\frac{\Pr(X=x|Y=1)}{\Pr(X=x|Y=0)}\right\}$$

Taking the log of the left-hand side, we obtain the Logistic Regression Model; On the right-hand side: The ratio of prior probabilities is a constant (with respect to x) and the likelihood ratio is the ratio of two pdf's or two densities.

NORMAL COVARIATE

- **Assume covariate X is normally distributed**

$$\text{Logit} = \text{Constant} + \ln(\text{ratio of densities})$$

$$\text{Logit} = \text{Constant} + \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}\right)x + \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)x^2$$

$$\text{Logit} = \text{Constant} + \left(\frac{\mu_1 - \mu_0}{\sigma^2}\right)x \quad \text{if} \quad \sigma_1^2 = \sigma_0^2 = \sigma^2$$

- **The log of the Odds Ratio associated with "one standard deviation increase in value of X" is:**

$$\ln \rho = \frac{\mu_1 - \mu_0}{\sigma}; \text{ so that } d = (\ln \rho)\sigma$$

RESULT

$$N = 4(z_{1-\alpha} + z_{1-\beta})^2 \frac{\sigma^2}{d^2}$$

$$= 4(z_{1-\alpha} + z_{1-\beta})^2 \frac{\sigma^2}{(\log\rho)^2\sigma^2}$$

$$= \frac{4(z_{1-\alpha} + z_{1-\beta})^2}{(\log\rho)^2}$$

EXAMPLE

Suppose that an investigator is considering to design a case-control study; its aim is to investigate a potential association between coronary heart disease and serum cholesterol level. Suppose further that it is desirable to detect an odds ratio $\rho = 2.0$ for a person with cholesterol level 1 standard deviation above for the mean for his or her age group using a two-sided test with a significance level of 5% and a power of 90%.

$$\alpha = .05 \rightarrow z_{1-\alpha} = 1.96$$

$$\beta = .10 \rightarrow z_{1-\beta} = 1.28$$

$$N = \frac{4}{(\rho)^2}(z_{1-\alpha} + z_{1-\beta})^2$$

$$= \frac{4}{(\ln 2)^2}(1.96 + 1.28)^2$$

$$\cong 62 \text{ subjects; } 31/\text{group}$$

A More Interesting & complete illustration:

CROSS-OVER EXPERIMENT DESIGNS

We are now ready to look at a more comprehensive, more advanced, and more interesting example. I choose the "Cross-over Design" which is rather popular in clinical research but, but one reason or another, it is not covered in first-year programs and books. We will consider both binary and continuous outcomes.

Lung cancer is the leading cause of cancer death in the United States and worldwide. Cigarette smoking causes approximately 90% of lung cancer. Despite anti-smoking campaigns over the past 40 years, over 45 million (22%) adult Americans are still current smokers. The development of a viable chemoprevention strategy targeting current smokers potentially could decrease lung cancer mortality.

Previous studies have shown (1) that the tobacco specific nitrosamine 4(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) is a major lung carcinogen in tobacco smoke, (2) that 2-phenethyl isothiocyanate (PEITC) is a potent inhibitor of NNK-induced lung carcinogenesis in rats and mice. So we applied and got a grant to assess, via a placebo-controlled cross-over design, the effect of PEITC supplementation on changes of NNK metabolism in smokers. We hypothesize that there will be a 30% increase in urinary NNAL plus NNAL-Gluc among PEITC treated subjects (taking more toxin out).

THE ROLE OF STUDY DESIGN

In a "standard" experimental design, a linear model for a continuous response/outcome is:

$$Y = \left[\begin{array}{c}\text{Overall}\\ \text{Constant}\end{array}\right] + \left[\begin{array}{c}\text{Treatment}\\ \text{Effect}\end{array}\right] + \left[\begin{array}{c}\text{Experimental}\\ \text{Error}\end{array}\right]$$

The last component, 'experimental error", includes not only error specific to the experimental process but also includes "subject effect" (age, gender, etc…). Sometimes these subject effects are large making it difficult to assess "treatment effect"; given certain level "hypothesized" treatment effect (e.g. Our 30% total NNAL), it would require a much larger sample size – larger than we could afford.

Blocking (to turn a completely randomized design into a randomized complete block design) would help. But it would only help to "reduce" subject effects, not to "eliminate them: subjects in the same block are only similar, not identical – unless we have "blocks of size one". And that the basic idea of "Cross-over Designs" – a very special form of "repeated measure designs/studies..

In the most simple cross-over design, subjects are randomly divided into two groups (often of equal sign); subjects in both groups/series take both treatments (experimental treatment and placebo/control) but in different "orders".

Of course, "order effects" and "carry-over effects" are possible. And the cross-over designs are not always suitable. They are commonly used when treatment effects are not permanent; for example some treatments of rheumatism.

THE DESIGN

In our "PEITC trial", measurements (urinary total NNAL) will be taken from each subject in the two supplementation sequences as seen in the following diagram:

Group 1: Period #1 (PEITC; A1) – washout – Period #2 (Placebo; B2)
Group 2: Period #1 (Placebo; B1) – washout – Period #2 (PEITC; A2)

The letter is used to denote supplementation or treatment (A for PEITC and B for Placebo) and the number, 1 or 2, denotes the period; e.g. "A1" for PEITC taken in period #1.

The "washout periods" are inserted in order to eliminate possible "carry-over effects" (The half-life of dietary PEITC in vivo is between 2-3 hours, with complete excretion within 1-2 days following ingestion) and, in the modeling and analysis, we will have to find a way cancel "order effects" in order assess "treatment effects". Here and throughout the first part, we consider outcomes on continuous scale.

OUTCOME VARIABLE

From the design:

Group 1: Period #1 (PEITC; A1) – washout – Period #2 (Placebo; B2)

Group 2: Period #1 (Placebo; B1) – washout – Period #2 (PEITC; A2)

Our data analysis will be based on the following "outcome variables" (Treatment - Placebo):

X1 = A1 - B2; and X2 = A2 - B1

This subtraction will cancel the "within-sequence" effects of all subject-specific factors.

This process will result in two independent samples (often with the same or similar sample size if there are no or minimal dropouts or missing data.

Recall the general model:

$$Y = \left[\begin{array}{c}\text{Overall}\\ \text{Constant}\end{array}\right] + \left[\begin{array}{c}\text{Treatment}\\ \text{Effect}\end{array}\right] + \left[\begin{array}{c}\text{Experiment al}\\ \text{Error}\end{array}\right]$$

The subtractions

X1 = A1 - B2; and X2 = A2 - B1

will cancel not only effects of all subject-specific factors; they cancel the "overall constant" as well, leaving only two components/parameters to be modeled and analyzed: the "treatment effect" (difference between PEITC & Placebo) and the "period (or order) effect" difference between period 1 and period 2).

A SIMPLE LINEAR MODEL

Design:

Group 1: Period #1 (PEITC; A1) – washout – Period #2 (Placebo; B2)

Group 2: Period #1 (Placebo; B1) – washout – Period #2 (PEITC; A2)

Outcome variables:

X1 = A1 - B2; and X2 = A2 - B1

There are different ways to express a linear model; a very simple one would be as follows:

X1 is normally distributed as $N(\alpha+\beta,\sigma^2)$

X2 is normally distributed as $N(\alpha-\beta,\sigma^2)$

In this models, α represents the PEITC supplementation effect (α>0 if and only if PEITC increases the total NNAL) and β represents the period effect (β>0 if and only if measurement from period 1 is larger than from period 2).

(1) The X's do not really need to have normal distributions; the robustness comes from the fact that our analysis will be based on the normal distribution of the sample mean – not of the data, and the sample mean would be almost normally distribution for moderate to large sample sizes (Central Limit Theorem).

(2) Among the three parameters, α represents the PEITC effect and is the primary target; we have no interest in the other two (β and σ^2). They are nuisance; we have to handle them properly to make inferences on α valid and efficient.

MODEL/DIMENSION REDUCTION

Design:

Group 1: Period #1 (PEITC; A1) – washout – Period #2 (Placebo; B2)

Group 2: Period #1 (Placebo; B1) – washout – Period #2 (PEITC; A2)

Outcome variables:

X1 = A1 - B2; and X2 = A2 - B1

From the model:

X1 is normally distributed as $N(\alpha+\beta,\sigma^2)$

X2 is normally distributed as $N(\alpha-\beta,\sigma^2)$

Let the sample means and sample variances be defined as usual and n the group size (total sample size is 2n); then, we can easily prove the followings:

INTERMEDIDATE RESULT

$\bar{x}_1$ is distribute d as normal $N(\alpha+\beta, \frac{\sigma^2}{n})$,

$\bar{x}_2$ is distribute d as normal $N(\alpha-\beta, \frac{\sigma^2}{n})$

$$\mathbf{a} = \frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2}{2}$$

a is distribute d as normal $\mathbf{N(\alpha, \frac{\sigma^2}{2n})}$

ESTIMATION OF PARAMETERS

(1) Estimation of Variance. We can pool data from the two sequences to estimate the common variance σ^2 by s_p^2 – the same pooled estimate used in two-sample t-test.

(2) Estimation of Treatment Effect. Parameter α representing the PEITC effect, the difference between PEITC and the placebo, is estimated by a – the average of the two sample means. Its 95 percent confidence interval is given by:

$$a \pm t_{.975} \frac{s_p}{\sqrt{N}}$$

The t-coefficient goes with (N-2) degrees of freedom; without missing data, N = 2n – total number of subjects.

TESTING FOR TREATMENT EFFECT

Testing for PEITC Treatment Effect. Null hypothesis of "no treatment effects" H0: α = 0 is tested using the "t test", with (N-2) degrees of freedom:

$$t = \frac{a}{s_p/\sqrt{N}}$$

It's kind of "one-sample t-test" but we use the degree of freedom associated with s_p. Alternatively, one can frame it as a two-sample t-test comparing X1 versus (-X2) as seen from

X1 is normally distributed as $N(\alpha+\beta,\sigma^2)$

X2 is normally distributed as $N(\alpha-\beta,\sigma^2)$

ISSUES FOR SAMPLE SIZE

The basis for sample size determination is the "statistic" "a" and its distribution:

$$a = \frac{\bar{x}_1 + \bar{x}_2}{2}$$

$$a \text{ is distributed as normal } N(\alpha, \frac{\sigma^2}{2n})$$

from which one obtains distributions under the Null and Alternative Hypotheses.

VARIANCE COMPONENTS

Recall that **X1 and X2 – each is the difference of two total NNAL - on the log scale; and from the model, X1 is distributed as N($\alpha+\beta,\sigma^2$) and X2 as N($\alpha-\beta,\sigma^2$). We would get to the common variance σ^2 by:**

$$\begin{aligned} \mathrm{Var}(A = \ln \mathrm{NNAL}) &= \mathrm{Var}(B) \\ &= \left[\mathrm{CV}(\mathrm{NNAL})\right]^2 \\ \mathrm{Var}(X) &= \mathrm{Var}(A) + \mathrm{Var}(B) - 2\mathrm{Cov}(A, B) \\ &= 2\mathrm{Var}(A)\left[1 - r(A, B)\right] \end{aligned}$$

We need CV(NNAL) and r(A, B):

Two-period crossover designs are often used in clinical trials in order to improved sensitivity of the trial by eliminating individual patient effects. They have been popular in dairy husbandry studies, long-term agricultural experiments, bioavailability and bioequivalence studies, nutrition experiments, arthritic and periodontal studies, and educational and psychological studies – where treatment effects are not permanent.

The response could be quantitative but quite often, e.g. the response is whether or not relief from pain is obtained, the response variable is binary.

THE DESIGN

Recall the following design; the only difference is that, in this case, the four outcomes A1, A2, B1, and B2 are binary – say 1 if positive response and 0 otherwise:

Group 1: Period #1 (Trt A; A1) – washout – Period #2 (Trt B; B2)
Group 2: Period #1 (Trt B; B1) – washout – Period #2 (Trt A; A2)

The washout periods are optional and the group sizes could be different (due to dropouts).

In general, let Y be the outcome or dependent variable taking on values 0 and 1, and:

π = Pr(Y=1)

Y is said to have the "Bernouilli distribution" (Binomial with n = 1). We have:

$$E(Y) = \pi$$

$$Var(Y) = \pi(1-\pi)$$

Studies would involve some independent variables (treatment, order, etc…)

THE LOGISTIC MODELS

The Logistic models for cross-over design are (J. J. Gart, Biometrika 1969):

$$\Pr(A1=1)=\frac{e^{\lambda_i+\alpha+\beta}}{1+e^{\lambda_i+\alpha+\beta}};\Pr(B2=1)=\frac{e^{\lambda_i-\alpha-\beta}}{1+e^{\lambda_i+\alpha+\beta}}$$

$$\Pr(B1=1)=\frac{e^{\lambda_i-\alpha+\beta}}{1+e^{\lambda_i-\alpha+\beta}};\Pr(A2=1)=\frac{e^{\lambda_i+\alpha-\beta}}{1+e^{\lambda_i+\alpha-\beta}}$$

In this models,
(1) λ‘s represent the subjects effects varying from subject to subject;
(2) α represents the PEITC supplementation effect (α>0 if and only if PEITC is more effective) – our main interest - and
(3) β represents the period effect (β>0 if and only if a treatment from period 1 is more effective than from period 2).

In this modeling:

(1) We “code’ binary covariates (Treatment and Order) as (+1/-1) instead of (0,1);

(2) All subject-specific covariates are lumped together with the Intercept.

It was shown by Gart (1969) that optimum inferences about treatment and order effects, regarding subjects effects as nuisance, are based on those subjects with unlike responses in two periods; that are subjects whose pair of outcomes are either (0,1) or (1,0). This is similar to the argument leading to the McNemar Chi-square test.

 # 1: Period 1 (Trt A; A1) – Period 2 (Trt B; B2)
 # 2: Period 1 (Trt B; B1) – Period 2 (Trt A; A2)

The analysis will be conditioned on:

A1+B2 = 1, and

B1+A2 = 1

$$\Pr(A1=1)=\frac{e^{\lambda_i+\alpha+\beta}}{1+e^{\lambda_i+\alpha+\beta}};\Pr(B2=1)=\frac{e^{\lambda_i-\alpha-\beta}}{1+e^{\lambda_i+\alpha+\beta}}$$

$$\Pr(B1=1)=\frac{e^{\lambda_i-\alpha+\beta}}{1+e^{\lambda_i-\alpha+\beta}};\Pr(A2=1)=\frac{e^{\lambda_i+\alpha-\beta}}{1+e^{\lambda_i+\alpha-\beta}}$$

$$\Pr(A1=1\mid A1+B2=1)=\frac{\Pr(A1=1,B2=0)}{\Pr(A1=1,B2=0)+\Pr(A1=0,B2=1)}$$

$$=\frac{\Pr(A1=1)\Pr(B2=0)}{\Pr(A1=1)\Pr(B2=0)+\Pr(A1=0)\Pr(B2=1)}$$

$$\Pr(A1 = 1 \mid A1 + B2 = 1) = \frac{\Pr(A1 = 1, B2 = 0)}{\Pr(A1 = 1, B2 = 0) + \Pr(A1 = 0, B2 = 1)}$$

$$= \frac{\Pr(A1 = 1)\Pr(B2 = 0)}{\Pr(A1 = 1)\Pr(B2 = 0) + \Pr(A1 = 0)\Pr(B2 = 1)}$$

$$= \frac{1}{1 + e^{-2(\alpha+\beta)}}$$

$$\Pr(A2 = 1 \mid B1 + A2 = 1) = \frac{1}{1 + e^{-2(\alpha-\beta)}}$$

Data: Frequencies of subjects with different Outcomes (0,1) and (1,0)

	Treatments (A,B)	Treatments (B,A)
Oucome A=1	y_{a1}	y_{a2}
Oucome B=1	y_{b1}	y_{b2}
Total	n_1	n_2

Notation: y_{ij}; j is the sequence (1if AB and 2 if BA) and i is the treatment with positive outcome.

Note: The same set of data can also be assembled into a different 2x2 Table

	Treatments (A,B)	Treatments (B,A)
1st Outcome=1	y_{a1}	y_{b2}
2nd Outcome =1	y_{b1}	y_{a2}
Total	n_1	n_2

$$\Pr(A1 = 1 \mid A1 + B2 = 1) = \frac{1}{1 + e^{-2(\alpha+\beta)}} = p1$$

$$\Pr(A2 = 1 \mid B1 + A2 = 1) = \frac{1}{1 + e^{-2(\alpha-\beta)}} = p2$$

	Treatments (A,B)	Treatments (B,A)
Oucome A=1	y_{a1}	y_{a2}
Oucome B=1	y_{b1}	y_{b2}
Total	n_1	n_2

Results: With n_1 and n_2 fixed, y_{a1} and y_{a2} are distributed as Binomials $B(n_1,p1)$ and $B(n_2,p2)$

Results: With n_1 and n_2 fixed, y_{a1} and y_{a2} are distributed as Binomials $B(n_1,p1)$ and $B(n_2,p2)$

(Conditional) Likelihood Function:

$$L = \binom{n_1}{y_{a1}} p1^{y_{a1}} (1-p1)^{y_{b1}} \binom{n_2}{y_{a2}} p2^{y_{a2}} (1-p2)^{y_{b2}}$$

$$p1 = \frac{1}{1 + e^{-2(\alpha+\beta)}}$$

$$p2 = \frac{1}{1 + e^{-2(\alpha-\beta)}}$$

$$L = \binom{n_1}{y_{a1}} p1^{y_{a1}} (1-p1)^{y1} \binom{n_2}{y_{a2}} p2^{y_{a2}} (1-p2)^{y2}$$

$$= \frac{\binom{n_1}{y_{a1}}\binom{n_2}{y_{a2}}\left[\exp\left\{-2y_{b1}(\alpha+\beta)+2y_{b2}(\beta-\alpha)\right\}\right]}{\left[1+e^{-2(\alpha+\beta)}\right]^{n_1}\left[1+e^{2(\beta-\alpha)}\right]^{n_2}}$$

TREATMENT EFFECT

$$\hat{\alpha} = a = \frac{1}{4} \ln \frac{y_{a1} y_{a2}}{y_{b1} y_{b2}}$$

$$\hat{\beta} = b = \frac{1}{4} \ln \frac{y_{a1} y_{b2}}{y_{a2} y_{b1}}$$

$$\mathrm{Var}(a) = \mathrm{Var}(b) = \frac{1}{16} \left[\frac{n_{ab}}{y_{a1} y_{b1}} + \frac{n_{ba}}{y_{a2} y_{b2}} \right]$$

SUGGESTED EXERCISES

#1. When a patient is diagnosed as having cancer of the prostate, an important question in deciding on treatment strategy for the patient is whether or not the cancer has spread to the neighboring lymph nodes. The question is so critical in prognosis and treatment that it is customary to operate on the patient (i.e., perform a laparotomy) for the sole purpose of examining the nodes and removing tissue samples to examine under the microscope for evidence of cancer. However, certain variables that can be measured without surgery may be predictive of the nodal involvement; one of which is level of serum acid phosphatase. Suppose an investigator considers to conduct a case-control study to evaluate this possible relationship between nodal involvement (cases) and level of serum acid phosphatase. Suppose further that it is desirable to detect an odd ratio of $\theta = 1.5$ for an individual with a serum acid phosphatase level of one standard deviation above the mean for his/her age group using a two-sided test with a significance level of 5 percent and a power of 80 percent. Find the total sample size needed for using a two-sided test at the .05 level of significance.

#2. With data in this arrangement, what does an independence between "columns" and "rows" mean and how does it fit in with parameter estimates ?

	Treatments (A,B)	Treatments (B,A)
1st Outcome=1	y_{a1}	y_{b2}
2nd Outcome =1	y_{b1}	y_{a2}
Total	n_1	n_2

#3. With data in this arrangement, what does an independence between "columns" and "rows" mean and how does it fit in with parameter estimates ?

	Treatments (A,B)	Treatments (B,A)
Oucome A=1	y_{a1}	y_{a2}
Oucome B=1	y_{b1}	y_{b2}
Total	n_1	n_2

#4. Back to the case of the PEITC trial, finish the sample size calculation with the following assumptions for NNAL: Coefficient of variation is estimated at 0.6, and Intra-subject correlation is estimated at 0.4.