PubH 7405:
REGRESSION ANALYSIS

DESIGN ISSUES: VALIDITY & SAMPLE SIZE

THE ANATOMY & PHYSIOLOGY OF CLINICAL RESEARCH

We form and/or evaluate a research or research project from/on two different angles or parts: the anatomy and the physiology of research; just like the hardware and software to run a computer operation.

THE ANATOMY PART

- From the anatomy of the research, one can describe/see **what it's made of**; this includes the tangible elements of the study plan: **research question, design, subjects, measurements, sample size calculation**, etc…

❖The **goal** is to create these elements in a form that will make the project **feasible, efficient, and cost-effective.**

THE PHISIOLOGY PART

- From the physiology of the research, one can describe/see **how it works**; first about **what happened in the study sample** and then about **how study findings generalized to people outside the study**.

❖The **goal** is to minimize the errors that threaten conclusions based on these **inferences.**

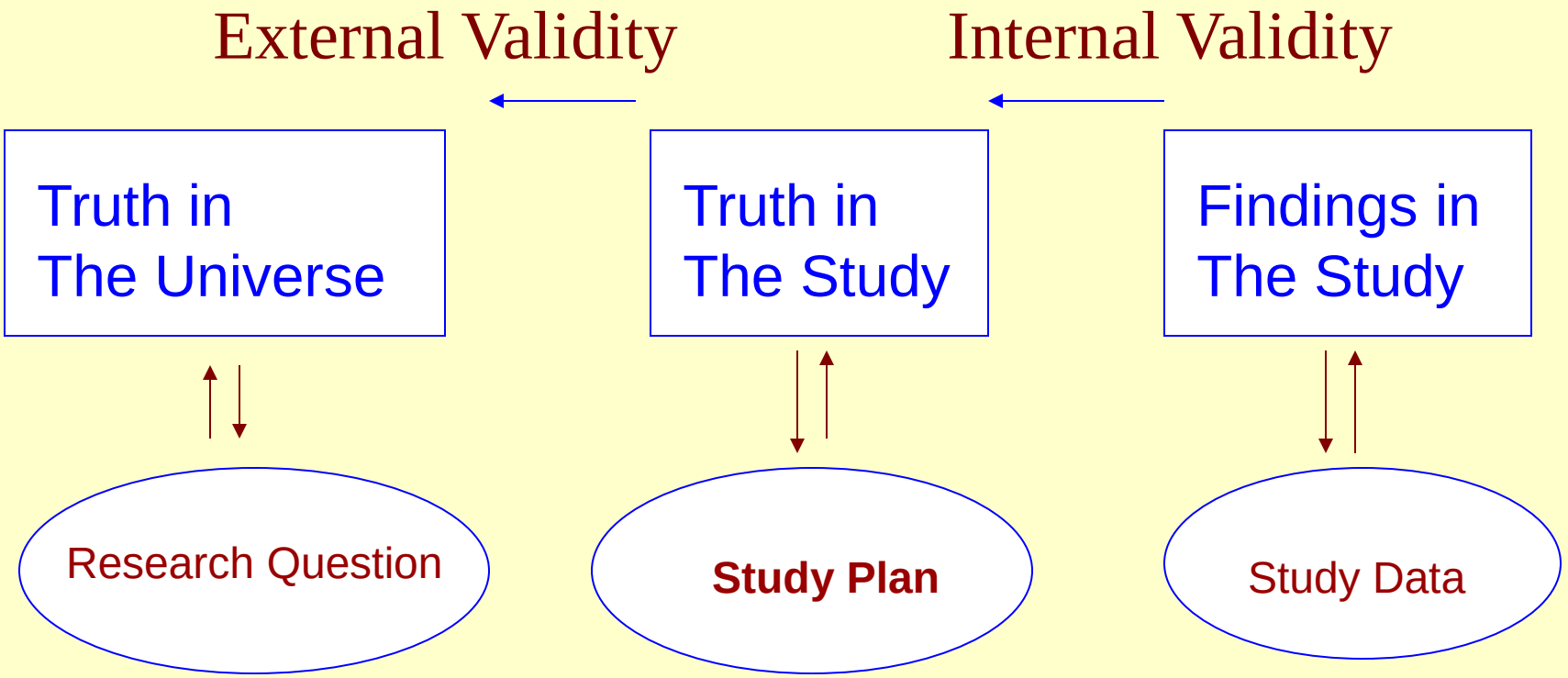
Very briefly, designed experiments are conducted to demonstrate a cause-and-effect relation between one or more explanatory factors (or predictors) and a response variable. The demonstration of a cause-and-effect relationship is accomplished by altering the level or levels of the explanatory factors and observing the effect of the changes (i.e. designed values of predictors X's) on the response variable Y. There is a good reason that designed experiments are often used as "comparative" in natures. Why?

INFERENCES & VALIDITIES

- Two major levels of inferences are involved in interpreting a study

❖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

With the goal of maximizing the validity of the inferences, the investigator reverses the process: (i) designs a study plan in which the choice of the research question, the subjects, and the measurements enhances the External Validity, (ii) is conducive to implementation with a high degree on Internal Validity.

That is to focus on the External Validity first (Design) then Internal Validity (Implementation).

THE BASIC ISSUE

Most of the times, inexperienced researchers mistakenly act like there is an identifiable, existent parent population or populations of subjects. We act as if the sample or samples is/are obtained from the parent population or populations according to a carefully defined technical procedure called random sampling. And we simply compare population means.

This is not true in real-life biomedical studies. The laboratory investigator uses animals in his projects but the animals are not randomly selected from any large population of animals. The clinician, who is attempting to describe the results he has obtained with a particular therapy, cannot say that his patients is a random sample from a parent population of patients.

THE VALUE OF TRIALS

- **Because they are not population-based (there is not an identifiable, existent parent population of subjects for sample selection), biomedical studies – designed experiments are "comparative". That is the validity of the conclusions is based on a comparison.**
- **In a clinical trial, we compare the results from the "treatment group" versus the results from the "placebo group". The validity of the comparison is backed by the randomization.**

COMPARISON OF TWO MEANS

- In many cohort studies, 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.

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

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

COMPARISON OF 2 PROPORTIONS

- In many cohort studies, 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 of the investigation** is $\theta = p_2 - p_1$

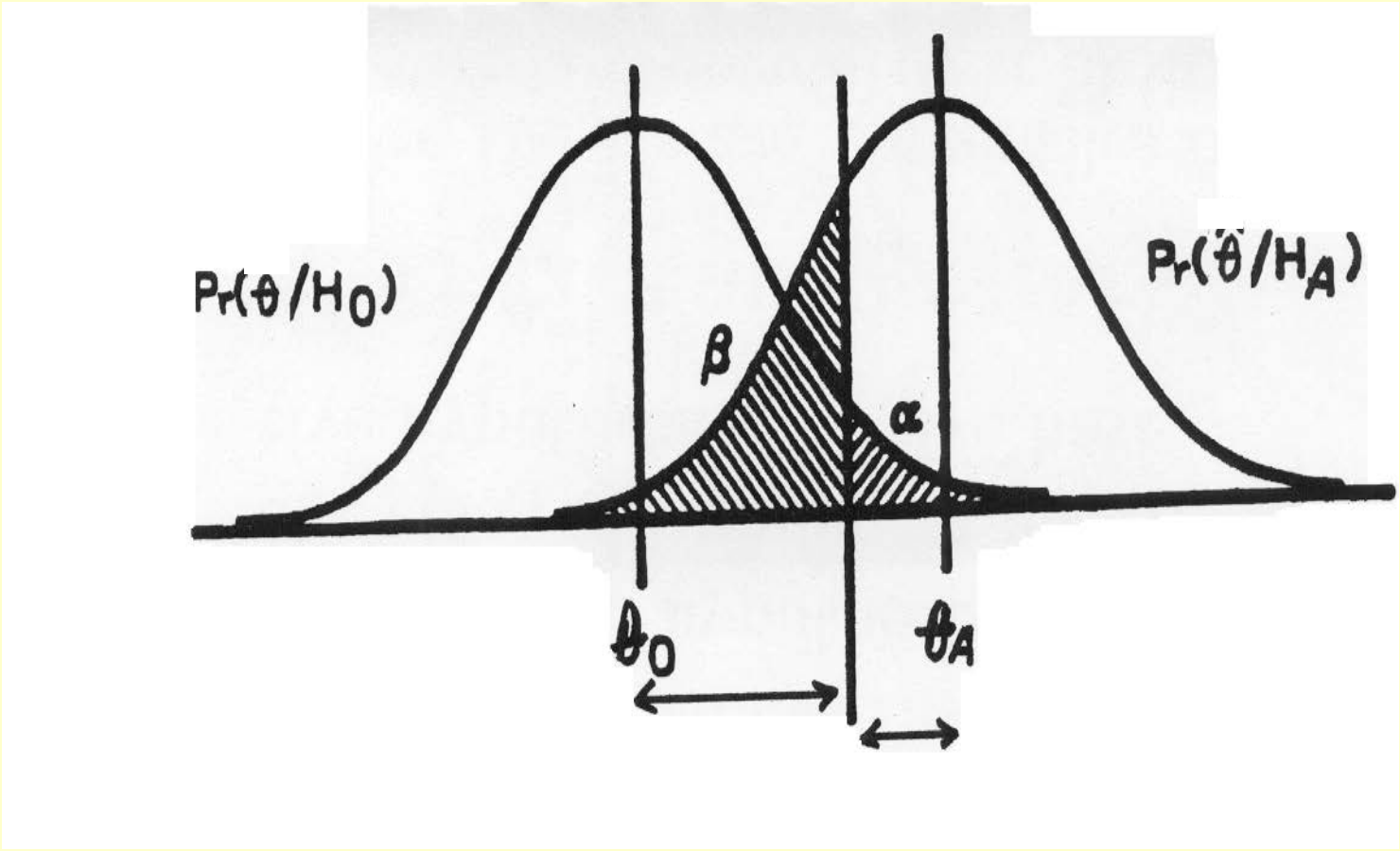
STATISTICAL ISSUES

- Statistics is a way of thinking, thinking about ways to **gather and analyze data.**
- The gathering part (i.e. data collection)comes before the analyzing part; the first thing a statistician or a learner of statistics does when faced with a biomedical project is data collection (followed by data management and data analysis).
- Studies may be inconclusive because they were poorly planned, **not enough data were collected** to accomplished the goals and support the hypotheses.

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**"? For one, it's possible that there isn't sufficient statistical power to detect a relationship even if it exists; **perhaps the sample size is too small.**

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

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

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

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

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

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

RESULTS FOR CORRELATION

- The null hypothesis to be tested is H_0: $\rho = 0$
- The target statistic is Fisher's z
- 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:

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

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

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

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 $\theta = \overline{x_2} - \overline{x_1}$

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

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

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

- Then $d^2 = (z_{1-\alpha} + z_{1-\beta})^2 \Sigma^2$ leads to total sample size:

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

If the two groups are planned to have different sizes, with $n_1 = pN$ and $n_2 = (1-p)N$, $(0<p<1)$; then the total sample size is:

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

NEEDED COMPONENTS

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

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

(2) The desired power $(1-\beta)$. 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 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 $\sigma^2 = 36$.
- 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 $\theta = p_2 - p_1$

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

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

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

- Then $d^2 = (z_{1-\alpha} + z_{1-\beta})^2 \Sigma^2$ leads to total sample size:

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

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 determination is a clinical decision, not a statistical decision).

(4) π is the average proportion $\bar{\pi} = (\pi_2 + \pi_1)/2$; 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:

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

SAME APPROACH

- 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:

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

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 that 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

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

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

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

RESULT

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

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

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

Where p is the percent of subjects with events (Y=1); 0<p<1

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%.

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

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

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

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

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

For one-arm trial, for example in phase II trials, the sample size is determined by controlling the width of the 95% confidence interval.

The Issue of Multiplicity

VARIABILITY & ERRORS

In some medical cases such as infections, the presence or absence of bacteria and viruses – a binary outcome - is easier to confirm; "test decisions" are made correctly.

For a continuous outcome, we have different "distributions" for sub-populations. In efforts to separate them, errors are unavoidable.

And that's also the case of statistical tests of significant: "test statistics" have different distributions under the Null and the Alternative.

ERRORS

In making a decision concerning the Null Hypothesis to compare μ_U versus μ_{NU}, errors are unavoidable. Since a null hypothesis H_0 may be true or false and our possible decisions are whether to reject or not to reject it, there are four possible outcomes combinations. Two of the four outcomes are correct decisions:

(i) not rejecting a true H_0

(ii) rejecting a false H_0

There are also <u>two possible ways to commit an error:</u>

<u>Type I</u>: a true H_0 is rejected

<u>Type II</u>: a false H_0 is not rejected

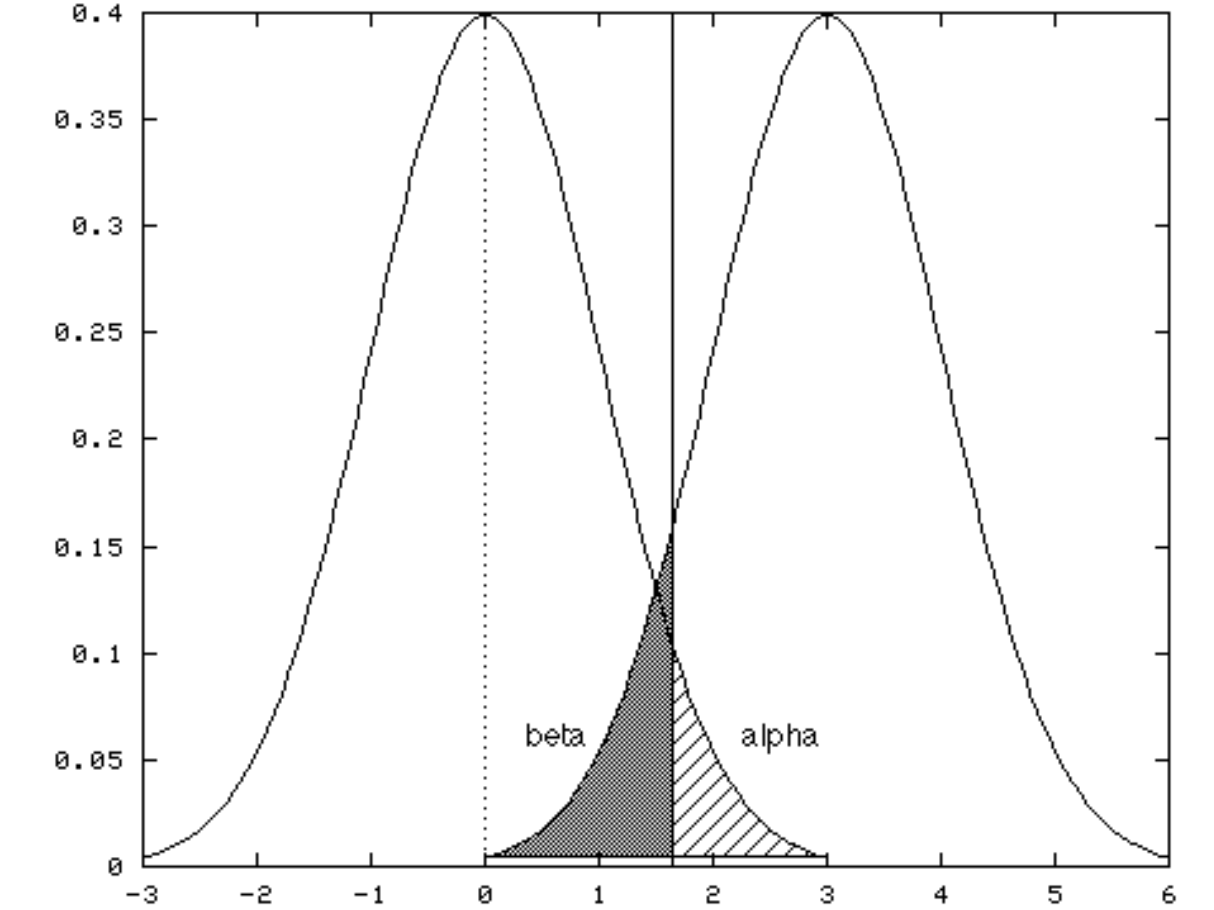
ANALOGIES

- **Type I error: Convicting an innocent man (top priority: to keep the probability of committing this error low – that's in "trial phase")**
- **Type II error: Acquitting a guilty suspect (Type II error is controlled earlier in the process, i.e. making sure to have enough evidence for a conviction by a thorough investigation – in "investigation phase").**

Truth	H_0 not rejected	H_0 is rejected
H_0 is true	Correct Decision	Type I Error
H_0 is false	Type II Error	Correct Decision

α = Pr(Type I Errors)

β = Pr(Type II Errors)



1-β = Statistical Power

The aim of investigators is to keep α and β, the probabilities - in the context of repeated sampling – of types I and II errors respectively, as small as possible. However, resources are limited, this goal requires a compromise because these actions are contradictory; We fix α at some specific conventional level- say .05 or .01 and β is controlled through the use of sample size.

In other words, in research, the control of type I errors lies in the "analysis stage" and the control of type II errors lies in the "design stage", making sure to have a large study to collect enough data.

In practice, we often have to make more than one decision at a time. Multiplicity – or Multiple Decision Problem - occurs when one considers a set of statistical inferences simultaneously . Examples include:

(1) Pairwise differences in ANOVA problem;
(2) Studies with multiple endpoints;
(3) Interim analyses; and, of course,
(4) Subgroup analyses in clinical trials
(5) Multiple regression.

What's the problem?

FAMILYWISE ERROR RATE (FER)

FER = $P(at\ least\ one\ false\ positive\ result)$

= $1-P(zero\ false\ positive\ results)$

= $1-(1-\alpha)^k$

We often want to maintain FER at a pre-determined level, say, the conventional choice of 0.05 or 0.01

EXAMPLES

Number of tests	Probability
1	0.05
2	0.0975
5	0.226
10	0.401
50	0.923

Probability of at least one false significant result (Note: not proportional to number of tests; with 10 tests, it's not (10)(0.05) = 0.50).

BONFERRONI METHOD

(1) N different Null Hypotheses H_{01},…, H_{0N}
(2) Calculate corresponding p-values: p_1,…, p_N
(3) Reject H_{0i} if and only if $p_i < \alpha/N$

e.g.
For 10 comparisons; per comparison, compare p-value to:
adjusted α = 0.05/10 = 0.005

Bonferroni is the most simple, most commonly used method. However:

(1) It is too conservative (low power);
(2) Do not take into account correlation between decisions.

HOLM METHOD

(1) N different Null Hypotheses H_{01},…, H_{0N}
(2) Calculate corresponding p-values: p_1,…, p_N
(3) Order these p-values from smallest to largest,
$p_{(1)} < p_{(2)} < \ldots < p_{(N)}$
(4) **Starting with the smallest p-value:**

(a) If p(1)≥ α/N, testing stops with no statistically significant differences;

(b) If p(1)< α/N, that comparison is deemed significant, and p(2) is then compared with α/(N-1)

(c) If p(2)≥ α/(N-1), testing stops and no further differences are declared significant. Otherwise, that comparison is deemed significant, and p(3) is then compared with α/(N-2) etc…

At the jth step, reject H(j) if p(j)< α/(N-j+1); for example, at the last step, compare the largest p-value p(N) to α.

Holm method is more powerful than Bonferroni's but it's still somewhat conservative because it does not take into account correlation between decisions.

HOCHBERG METHOD

(1) N different Null Hypotheses H_{01},…, H_{0N}
(2) Calculate corresponding p-values: p_1,…, p_N
(3) Order these p-values from smallest to largest,
$p_{(1)} < p_{(2)} < \ldots < p_{(N)}$
(4) **Starting with the largest p-value:**

(a) If p(N)< α, testing stops and declare all comparisons significant at level (*i.e.* reject all Null Hypotheses). Otherwise fail to reject H(N) and go on to the next step
(b) If p(N-1)< α/2, stop & declare H(1), H(2), …, H(N-1) are all significant. Otherwise fail to reject H(N-1) and go on to compare p(N-2) to α/3, etc…
(c) In general, compare p(N-k) to α/(k+1)

Hochberg (also known as Benjamini-Hochberg) method and Holm method are equivalent. They are both sequential but moving in different direction (one like “backward elimination and one “forward selection”). In recent years, Hochberg method becomes increasing more popular and more cited.

Both methods are more powerful than Bonferroni but not take into account correlation between decisions.

EXAMPLE

Suppose we performed N=5 tests of hypothesis simultaneously (or fitted a multiple regression model with 5 predictors) and want to keep the overall type I errors below the conventional level of 0.05.

Let the ordered p-values be:

p(1) = 0.009

p(2) = 0.011

p(3) = 0.015

p(4) = 0.034

p(5) = 0.512

Investigating the ordered p-values:
p(1) = 0.009 vs. 0.05/5 = 0.01
p(2) = 0.011 vs. 0.05/5 = 0.01
p(3) = 0.015 vs. 0.05/5 = 0.01
p(4) = 0.034 vs. 0.05/5 = 0.01
p(5) = 0.512 vs. 0.05/5 = 0.01

Since 0.05/5 = 0.01; by Bonferroni method, only the first test (with p=0.09) is declared significant.
Result: Only one test is significant at the "overall p-value" of 0.05 (Note: 4 p-values are less than 0.05)

Investigating the sequence of ordered p-values:
p(1) = 0.009 vs. 0.05/5 = 0.01 Starting here & move down
p(2) = 0.011 vs. 0.05/4 = 0.0125
p(3) = 0.015 vs. 0.05/3 = 0.0167
p(4) = 0.034 vs. 0.05/2 = 0.025 investigation stops!
p(5) = 0.512

Result: by Holm method, the first three tests (with p=0.09, 0.011, and 0.015) are declared significant at the "overall p-value" of 0.05 (Note: 4 p-values are less than 0.05).

Investigating the sequence of ordered p-values:
p(1) = 0.009
p(2) = 0.011
p(3) = 0.015 vs. 0.05/3 = 0.0167 investigation stops!
p(4) = 0.034 vs. 0.05/2 = 0.025
p(5) = 0.512 vs. 0.05 Starting here & moving up

Result: by Hochberg method, the first three tests (with p=0.09, 0.011, and 0.015) are declared significant at the "overall p-value" of 0.05 (Note: 4 p-values are less than 0.05).

The only way to take into account the correlation between tests is using some "resampling" procedure" which preserve the correlation structure of test statistics, then use PROC MULTTEST in SAS to obtained adjusted p-values. For example, the Westfall and Young method using the Bootstrap resampling (resampling with replacement). Most these newer methods are rather complicated and time consuming, not popular with practitioners.

GUIDELINE FOR MULTIPLE REGRESSION?

(1) Identify a primary predictor (apriori; no more than two?); for example, the "treatment" (indicator variable) in clinical trials;
(2) Apply a multiplicity method, such as Benjamini-Hochberg, to all other predictors

(Note: These are my own recommendation; no formal guidelines exist; most investigators are still overly excited with p-values and "significance")

Due As Homework

23.1 Suppose we want to compare the use of medical care by black and white teenagers. The aim is to compare the proportions of kids without physical check-ups within the last two years. Some recent survey shows that these rates for blacks and whites are 17% and 7% respectively. How large should a total sample be so that it would be able to detect such a 10% difference with a power of 90% using a statistical test at the two-sided level of significance of .01?

23.2 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 lymp 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.