PubH 7405:
REGRESSION ANALYSIS

Review #3:
ONE-FACTOR EXPERIMENT DESIGN

HOW DOES SOCIETY DEAL WITH UNCERTAINTIES ?

- **We form Assumption/Hypothesis: “Every person is innocent until proven guilty” (written in our Constitution),**
- **We gather data: Evidence against Hypothesis- not against the suspect, then**
- **We decide whether Hypothesis should be rejected (If it is, the verdict is “Guilty”)**

HOW DOES SCIENCE DEAL WITH UNCERTAINTIES ?

- **We form Assumption/Hypothesis: From experience & observations (The process leads to the so-called research questions)**
- **We gather data: Experiments & Trials, Surveys, Medical Records Abstractions.**
- **We make decision by performing DATA ANALYSIS, the "core" area of Biostatistics – the core part, not the only part.**

ELEMENTS OF GOOD RESEARCH

- **A good RESEARCH QUESTION with well-defined objectives & endpoints,**
- **A thorough INVESTIGATION, lots of data**
- **An efficient PRESENTATION: data organization & summarization, and**
- **A proper STATISTICAL INFERENCE (the process & methods of drawing conclusions)**

AREAS OF BIOSTATISTICS

Research is a three-step process:

(1) Sampling/design: Find a way or ways to collect data (going from population to sample).

(2) Descriptive statistics: Learn to organize, summarize and present data which can shed light on the research question (investigating sample).

(3) Inferential statistics: Generalize what we learn from the sample or samples to the target population and answer the research question (going from sample to population).

THE IMPORTANT PHASE

Just as in the case of "Trial by Jury", the most important stage of the "Research Process" is the DESIGN: How & How Much data are collected! Also, It dictates how data should be analyzed. May be it's not the question of "how" to collect your data but the decision on "when to do what"!

Designed (one-factor) experiments are conducted to "demonstrate" a cause-and-effect relation between an explanatory factor (or predictor) and a response variable. The demonstration of a cause-and-effect relationship is accomplished, to put it in a simple way, by altering the level of the explanatory factor (i.e. "designed") and observing the effect of the changes (i.e. designed values of predictor X) on the response variable Y. Designed experiments are often used as "comparative" in natures; that is comparing responses from different levels of the predictor.

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.

Pay attention to which/what is the Explanatory variable (Predictor), Factor levels or treatment arms, Experimental units, and Outcome/Response variable.

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 balances the characteristics that we know and measurable and characteristics that we are not aware or hard to quantify.

Designed experiments are conducted to "demonstrate" a cause-and-effect relation between one or more explanatory factors (or predictors) and a response variable.

Different ways to show case the relationship form different "designs".

The simplest form of designed experiments is the "completely randomized design" where treatments are randomly assigned to the experimental units – regardless of their characteristics. This design is most useful when the experimental units are relatively homogeneous with respect to known confounders.

A confounder is a factor which may be related to the treatment or the outcome even the factor itself may not be under investigation. A study may involve one or several confounders. In a clinical trial, the primary outcome could be SBP reduction and the baseline SBP is a potential confounder. Patients' age may be another one. In theory, values of confounders may have been balanced out between study groups because patients were randomized. But it is not guaranteed; especially if the sample size is not very large.

If confounder or confounders are known apriori, heterogeneous experimental units are divided into homogeneous "block"; and randomizations of treatments are carried out within each block. The result would be a "randomized complete block design". This is the type of data you see in Two-way ANOVA (One factor is Treatment, the other is Block)

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.

Some other factors might affect the numbers of colds contracted by a child: age, gender, etc… Let say we focus on gender.

THE CHOICES

- We could perform complete randomization – disregard the gender of the child, and put Gender into the analysis as a covariate; or
- We could randomize boys and girls separately; at the end the proportions of boys in the two groups are similar and there would be no need for adjustment.
- The first approach is a complete randomized design; the second is a randomized complete block design.
- Similarly, we could block using “age groups”.

It might involve two blocking factors, and you would end up with Three-way ANOVA.

The term “treatment” may also mean different things; a treatment could be a factor or it could be multifactor. For example, let consider two different aspects of a drug regiment: Dose (Low, High) and Administration mode (say, one tablet a day or two tablets every other day). We could combine these two aspects to form 4 combinations; then treating them as 4 treatments and apply a completely randomize design. We call it a (balanced) Factorial Design; the analysis is similar to that of a randomized complete block design.

THE ROLE OF STUDY DESIGN

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

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

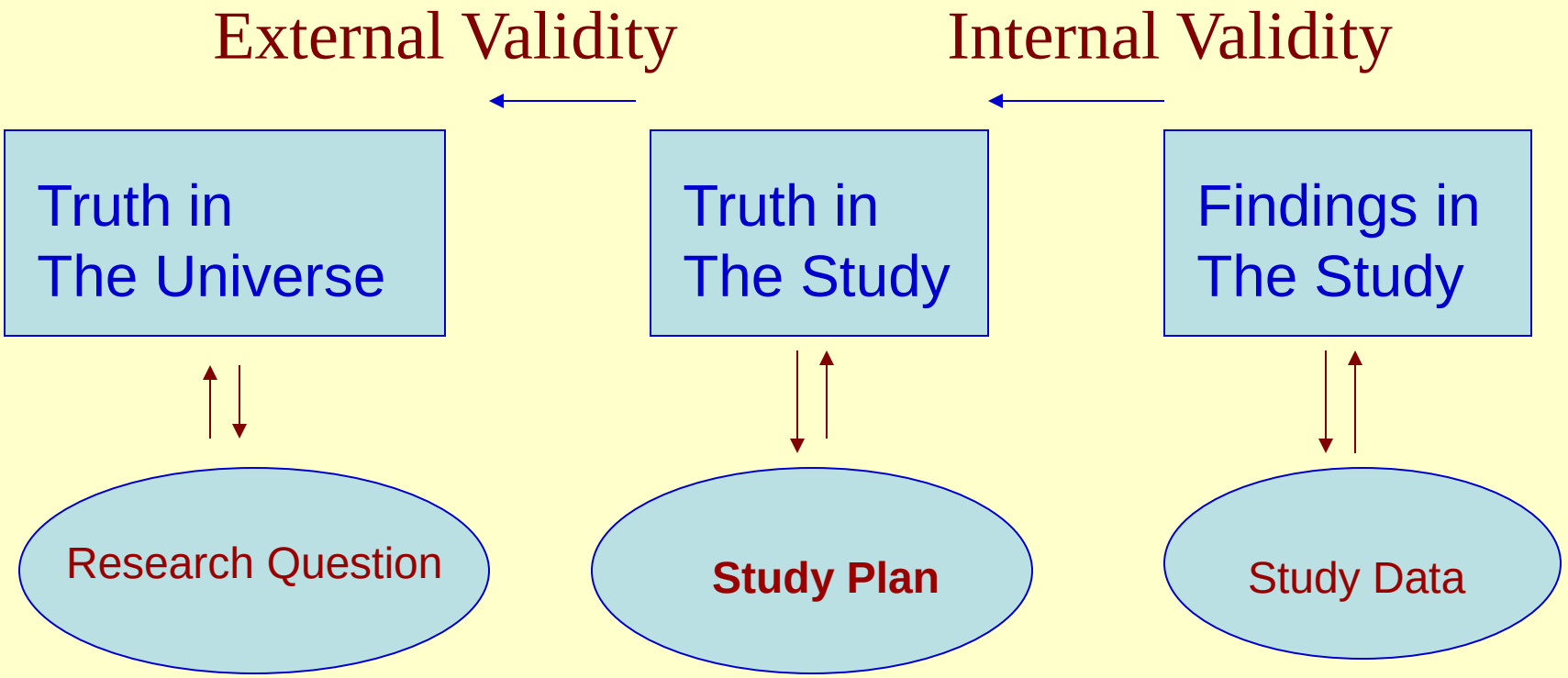
The last component, 'experimental error", includes not only error specific to the experimental process but also includes "subject effect" (age, gender, etc…). Sometimes these subject effects are large making it difficult to assess "treatment effect".

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 popular form in biomedical research.

Today's lecture is limited to the simplest form of designed experiments, the "completely randomized design". The aim is to review some statistical methods you previous learned and strengthen them by relating the solutions to regression analysis.

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).

Statistical contributions involve both Internal Validity (for example, helping to select a sensitive "endpoint" or decide what to do with missing data) and External Validity (helping to choose a proper design and an adequate sample size).

THE BASIC ISSUE IN RESEARCH

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.**

DATA ANALYSIS METHODS

Basic data analysis includes:

Two-sample t-test: to compare two population means (two-by-two Chi-square is a special case);

One-way ANOVA (Analysis Of Variance) to compare several population means;

Two-sample t-test is a popular statistical method for comparing two population means using two independent samples; and One-way Analysis of Variance (ANOVA) extends two-sample t-test to compare means of more than 2 independent samples.

Analysis A: COMPARISON OF TWO **POPULATION MEANS**

- **In this type of problems, we have two independent samples ($n_1, \bar{y}_1, s_1^2$) and ($n_2, \bar{y}_2, s_2^2$); the n's being the sample sizes, $\bar{y}$ the sample means, and s^2 the sample variances (the s are standard deviations).**
- **Often called the "two-sample problem"**
- **Considered as samples with population means μ_1 and μ_2**
- **The aim is to compare the two population means.**
- **("Y" is the "response", a measure of interest)**

#1: TWO-SAMPLE t-TEST

- The Null Hypothesis considered is H_0: $\mu_1 = \mu_2$ or equivalently, H_0: $\mu_2 - \mu_1 = 0$.
- The assumptions are:
  - ❖ Independent observations
  - ❖ Two Normal Distributions
  - ❖ Variances are equal
- (Normal assumption may be dropped if sample sizes are large – due to Central Limit Theorem)

$$t = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n_i - 1}; \quad s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

$$\mathbf{t} = \frac{\bar{\mathbf{y}}_2 - \bar{\mathbf{y}}_1}{\mathbf{s_p}\sqrt{\frac{1}{\mathbf{n_1}} + \frac{1}{\mathbf{n_2}}}}$$

#2: REGRESSION APPROACH

- **Pool the data, treat Y as dependent variable**
- **Binary independent variable (X=0/1; for group1/2)**
- **The assumptions (on Y; Regression of Y on X):**
  - ❖ **Independent observations**
  - ❖ **Normal Distribution for Y**
  - ❖ **Constant Variance**
- **Same as assumptions of the two-sample t-test**

Normal Error Regession Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

Independent Variable X is binary :

$$\mathbf{E(Y \mid X = x) = \beta_0 + \beta_1 x}$$

$$E(Y \mid x = 0; \text{group } 1) = \beta_0$$

$$E(Y \mid x = 1; \text{group } 2) = \beta_0 + \beta_1$$

$$E(Y \mid x = 1) - E(Y \mid x = 0) = \beta_1$$

Same Null Hypothesis :

$$(\mu_2 = \mu_1) \Leftrightarrow (\mu_2 - \mu_1 = 0) \Leftrightarrow (\beta_1 = 0)$$

EQUIVALENCY

- **Same Assumptions**
- **Same Null Hypothesis**
- **In order to prove that they are the same t-test, at df = (n-2); $n = n_1 + n_2$ we prove that they have the same test statistic too.**
- **A sketch of the proof is as follows**

Under the "Normal Error Regression Model":

$$Y = \beta_0 + \beta_1 x + \varepsilon; x = 0/1$$

$$\varepsilon \in N(0, \sigma^2)$$

We have for the estimated slope b_1:

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$\sigma^2(b_1) = \frac{\sigma^2}{\sum (x - \bar{x})^2}$$

$$\bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2}$$

$$\bar{x} = \frac{n_2}{n_1 + n_2}$$

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$= \frac{n_1 (0 - \frac{n_2}{n_1 + n_2})(\bar{y}_1 - \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2}) + n_2 (1 - \frac{n_2}{n_1 + n_2})(\bar{y}_2 - \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2})}{n_1 (0 - \frac{n_2}{n_1 + n_2})^2 + n_2 (1 - \frac{n_2}{n_1 + n_2})^2}$$

$$= \bar{y}_2 - \bar{y}_1$$

$$= \text{Difference of 2 sample means}$$

Since X = 0/1 for group 1/2; group 1 serves as “baseline”.

$$MSE = s_p^2$$

$$\frac{1}{\sum (x - \bar{x})^2} = \frac{1}{n_1} + \frac{1}{n_2}$$

$$SE(b_1) = \sqrt{\frac{MSE}{\sum (x - \bar{x})^2}}$$

$$= s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$t = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$= \frac{\bar{y}_2 - \bar{y}_1}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Two-sample t-test

$$b_1 = \bar{y}_2 - \bar{y}_1$$

$$SE(b_1) = \sqrt{\frac{MSE}{\sum (x-\bar{x})^2}}$$

$$= s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Regression's test for independence

$$t = \frac{b_1}{SE(b_1)}$$

Conclusion:

We can do regression with a binary independent variable; the results are equivalent to those of the two-sample t-test to compare two population means.

Analysis B: COMPARISON OF **POPULATION SEVERAL MEANS**

- **Suppose we want to know whether there are differences in the means of more than two independent groups. For example, do families of different ethnic groups have different income levels?**
- **Two Questions here: How to measure the "difference"? and how to decide of the observed difference is real ? (i.e. statistically significant).**

QUESTION #1:

How do we measure the "difference" among several means? By subtraction? Which from which? If not, then what else can we "measure" differences? That is what should we use instead of difference ($\overline{x}_2$-$\overline{x}_1$) (or their ratio)?

QUESTION #2:

How do we decide if the "difference" among several sample means is large enough to conclude that the population means are different? That is what do we use instead of the t-test?

ONE-WAY “ANOVA”

- What is needed is a different way to summarize the differences between several means and a method of simultaneously comparing these means in one step. This method is ANOVA or One-way ANOVA, for ``ANalysis Of VAriance".
- If that “one-step” test, the ANOVA F-test, is significant indicating that some pair(s) of means are different then we can start looking for that/those pairs- with “allowance” for multiple comparisons.

COMPONENTS OF TOTAL VARIATION

- **The total variation in the combined sample can be decomposed into two components as follows: $(x_{ij}-\bar{x}) = (x_{ij}-\bar{x}_i) + (\bar{x}_i-\bar{x})$:**

(1) The first term reflects the variation within the samples; the following sum is called the "within sum of squares": $SSW = \sum_{i,j}(x_{ij}-\bar{x}_i)^2 = \sum_i (n_i - 1)s_i^2$

(2) The difference, SSB=SST-SSW, is called the "between sum of squares" which measures the differences between samples:

$$SST = \sum_{i,j}(x_{ij}-\bar{x})^2$$

$$SSB = \sum_{i,j}(\bar{x}_i-\bar{x})^2 = \sum_i n_i(\bar{x}_i-\bar{x})^2$$

ANOVA: ANALYSIS OF VARIANCE

- **SST measures the "total variation" in the combined sample with (n-1) degrees of freedom, $n=\Sigma n_i$ is the total size. It is decomposed into: SST=SSW+SSB**
- **SSW measures the variation within samples with $\Sigma(n_i-1)=(n-k)$ degrees of freedom, and**
- **SSB measures the variation between sample means with (k-1) degrees of freedom; k=# of groups**

ANSWER #1:

SSB measures the variation, or difference, between sample means:

$$SSB = \sum_{i,j} (\bar{x}_i - \bar{x})^2 = \sum_{i} n_i (\bar{x}_i - \bar{x})^2$$

(which is a concept similar to the "variance": variation among sample means)

“ANOVA” TABLE

- **The breakdowns of the total sum of squares and its associated degree of freedom are displayed in the form of an “analysis of variance table” (ANOVA table) as follows:**

Source of Variation	SS	df	MS	F ratio	p-val
Between samples	SSB	k-1	MSB	MSB/MSW	
Within samples	SSW	n-k	MSW		
Total	SST	n-1			

- **MSW is a natural extension of the pooled estimate s_p^2 as used in the two-sample t-test; It is a measure of the average variation within the k samples.**

APPROACH TO QUESTION #2:

COMPARE the "average gap/difference" between sample means (MSB) to the average gap/difference between measurements in samples (MSW):

Use F=MSB/MSW

THE “F” TEST

- **The test statistic F for the Analysis of Variance compares MSB (the average variation between the k sample means) and MSE (the average variation within the k samples), a value near 1 supports the null hypothesis of no differences between the k population means.**
- **If we apply the new method of “One-way ANOVA” to compare the means of two groups, the result is identical to that of a two-sided two-sample t-test**

ANOVA ASSUMPTIONS

- **The Null Hypothesis considered is H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$**
- **The assumptions are:**

❖ **Independent observations**

❖ **k Normal Distributions**

❖ **Variances are equal**

- **(Normal assumption may be dropped if sample sizes are large – due to Central Limit Theorem)**

Alternative Approach:

To compare several means from a completely randomized design, we can do regression with a categorical independent variable – using (k-1) "dummy/indicator variables". The results are equivalent to those of the one-way Analysis of Variance (one-way ANOVA). This is "Multiple Regression"; main topic in the second half of this course.

A NEWER POSSIBLE PROBLEM:

Randomization is very crucial because it helps to balanced out the groups. However, even with randomization, groups or arms of a one-factor experiment design might still be unbalanced with respect to some confounder or confounders; without proper adjustment, results might be misleading.

A confounder is a factor which may be related to the treatment or the outcome even the factor itself may not be under investigation. A study may involve one or several confounders. In the above clinical trial example, the primary outcome is SBP reduction and the baseline SBP is a potential confounder. Patients' age may be another one. In theory, values of confounders may have been balanced out between study groups because patients were randomized.

That's theory that study groups are balanced – after a randomization; *in practice, study groups are rarely completely balanced with respect to all factors.* Randomization helps but might not help completely. And this possibility of unbalanced study groups is very real when the sample sizes are not very large – and confounding effect is strong.

We can investigate binary covariates (t-test), we can investigate categorical covariates (One-way ANOVA), and – of course - we can investigate continuous covariates.

Of course, a binary covariate or a categorical covariate can be used in the same model with one or more continuous covariates (confounders); the "combination" forms an interesting case – that's used to be called ANCOVA, Analysis of covariance.

The Analysis of Covariance (ANCOVA) serves the very same main purpose as ANOVA, that is to compares averages or means from different treatments, but it combines the ANOVA method with the Regression method in doing so. The term "ANCOVA" often refers to the Multiple Regression Model without interact terms in which binary/categorical covariate (representing groups) and continuous covariates (confounders) used together.

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

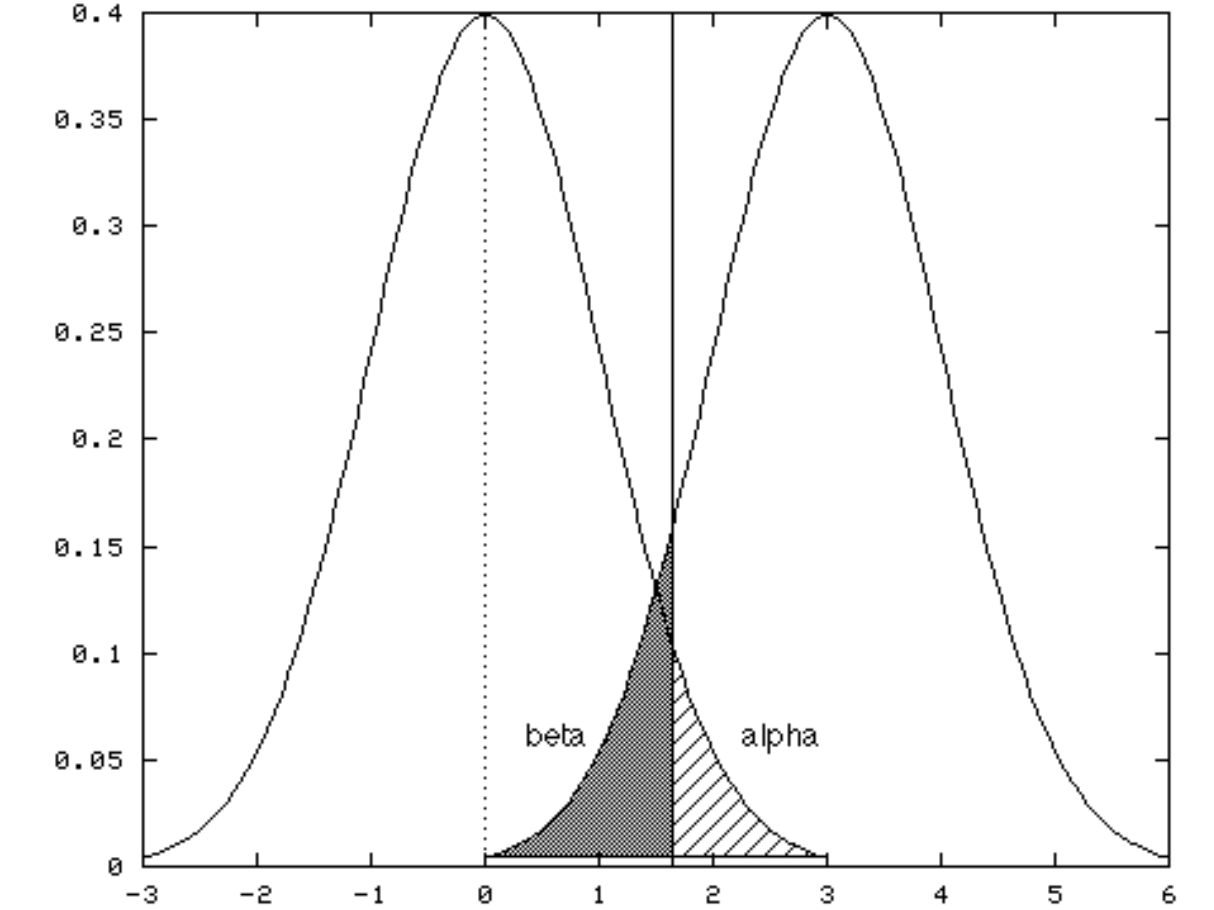
ANALOGIES

- **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 a “One-way ANOVA” problem, if the F-test is significant we can conclude that not all group means are equal but the test does not tell which ones are not the same or how may pairs are not different. We may have to start a series of “pairwise comparisons” using, for example, two-sample t-test.

So, 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 one or two primary comparisons; for example, the "treatment" indicator in clinical trials;
(2) Apply a multiplicity method, such as Benjamini-Hochberg, to all other comparisons

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

DUE AS HOMEWORK

We have data on the conduct of a number of cancer clinical trials from "ClinicalTrials.gov" (File: Minority Enrollment); the aim is to is to investigate potential factors which might affect the enrollment of black patients. There were n=113 trials and the (response) variable under investigation is the percent of black patients ("Black") among those recruited for each trial. To provide possible explanations, we'll investigate 9 possible exploratory (or independent) factors represented by 10 variables: Age (1= under 18, 2 = 18 and above), Gender (1 = Male, 2 = Female, 3 = both), Funder (1 = Government, 2 = Industry, 4 = Combination), Trial Duration (in months), Allocation (1 = Randomized, 2 = Non-randomized), Intervention Model (or Design; 1 = Parallel (multiple arms), 2 = Single group, 3 = Cross-over), Primary Purpose (1 = Therapeutic, 2 = Non-therapeutic), Masking (1 = Open Label, 2 = Double Blind). The final factor, Trial Size, is represented by two variables: Actual enrollment, and Accrual Percentage which expressed accrual as percentage of Planned Accrual.

#3.1 Investigate the role of Allocation, Primary Purpose, and Masking using the two-sample t-test. Define indicator variables and re-investigate the effects of those 2 factors using Simple Linear Regression

#3.2 Investigate the role of Gender and Funder using the method of One-way ANOVA. Define indicator variables and re-investigate the effects of those 2 factor using Multiple Linear Regression (You can skip this exercise if you did not yet have MLR)