# Statistics

## Analysis of Variance (ANOVA)

February 24, 2010

Jobayer Hossain, Ph.D. & Tim Bunnell, Ph.D.

*Nemours Bioinformatics Core Facility*

# Class Objectives -- You will Learn:

- Experimental design terminologies
- Analysis of variance (ANOVA) and the situation it is applicable
- One-way ANOVA
- Multiple comparisons
- How to do ANOVA with SPSS

# Experimental Design Terminology

- An **Experimental Unit** is the entity on which measurement or an observation is made. For example, subjects are experimental units in most clinical studies.
- **Homogeneous Experimental Units**: Units that are as uniform as possible on all characteristics that could affect the response.
- **Randomization** is the process of assigning experimental units randomly to different experimental groups. It is the most reliable method of creating homogeneous treatment groups, without involving potential biases or judgments.

# Experimental Design Terminology

- **Replication** is the repetition of an entire experiment or portion of an experiment under two or more sets of conditions.
- Replication reduces variability in experimental results and increases the significance and the confidence level with which a researcher can draw conclusions about an experimental factor
- To establish the significance of an experimental result, *replication*, the repetition of an experiment on a large group of subjects, is required.
  - If a treatment is truly effective, the average effect of replicated experimental units will reflect it.
  - If it is not effective, then the few members of the experimental units who may have reacted to the treatment will be negated by the large numbers of subjects who were unaffected by it.

# Experimental Design Terminology

- A **Block** is a group of homogeneous experimental units. For example, if an investigator had reason to believe that response of the medication differs significantly with age difference, he might choose to first divide the experimental subjects into age groups, such as under 5 years old, 5-10 years old, and over 10 years old etc to acquire homogeneity of response of the medication within age group.

# Experimental Design Terminology

- A **Factor** is a controllable independent variable that is being investigated to determine its effect on a response. E.g. treatment group is a factor.
- Factors can be **fixed** or **random**
  - **Fixed** -- the factor can take on a finite number of levels and these are the only levels of interest e.g. treatment group in our data set is a fixed factor.
  - **Random** -- the factor can take on a wide range of values and one wants to generalize from specific values to all possible values e.g. subjects in the data set can be a random factor.
- Each specific value of a factor is called a **level**. E.g. treatment group: A, B, and placebo. Then all these are three levels.

# Experimental Design Terminology

- A **covariate** is an independent variable not manipulated by the experimenter but still affecting the response. E.g. in many clinical experiments, the demographic variables such as race, gender, age  may influence the response variable significantly even though these are not the variables of interest of the study. These variables are termed as covariate.
- **Effect** is the change in the average response between two factor levels. That is, factor effect = average response at one level – average response at a second level.

# Experimental Design Terminology

- **Interaction** is the joint factor effects in which the effect of one factor depends on the levels of the other factors.

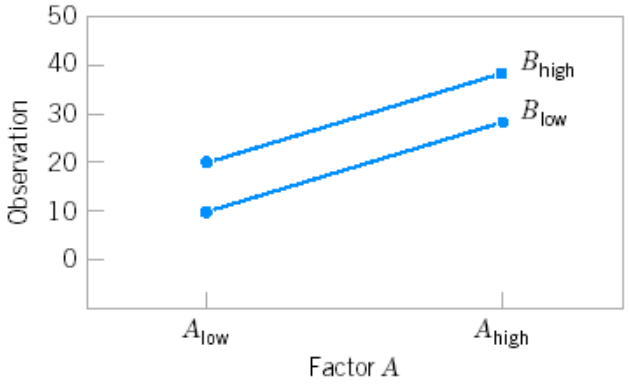



**Figure 7-1** An interaction plot of a factorial experiment, no interaction.

No interaction effect of factor A and B

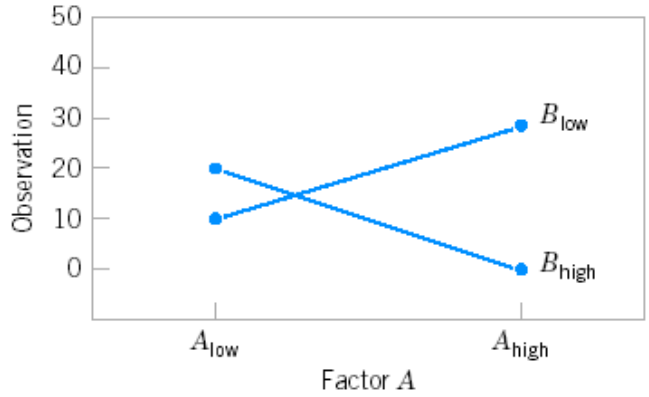



**Figure 7-2** An interaction plot of a factorial experiment, with interaction.

Interaction effect of factor A and B

# The basic ANOVA situation

- Type of variables: Quantitative response and Categorical (factor) predictors (independent variable).
- Main Question: Are mean response measures of different groups equal?
- One categorical variable with only 2 levels (groups):
  - 2-sample t-test
- One categorical variable with more than two levels (groups):
  - One way ANOVA
- Two or more categorical variable, each with at least two or more levels (groups) of each:
  - Factorial ANOVA

# Analysis of variance (ANOVA)

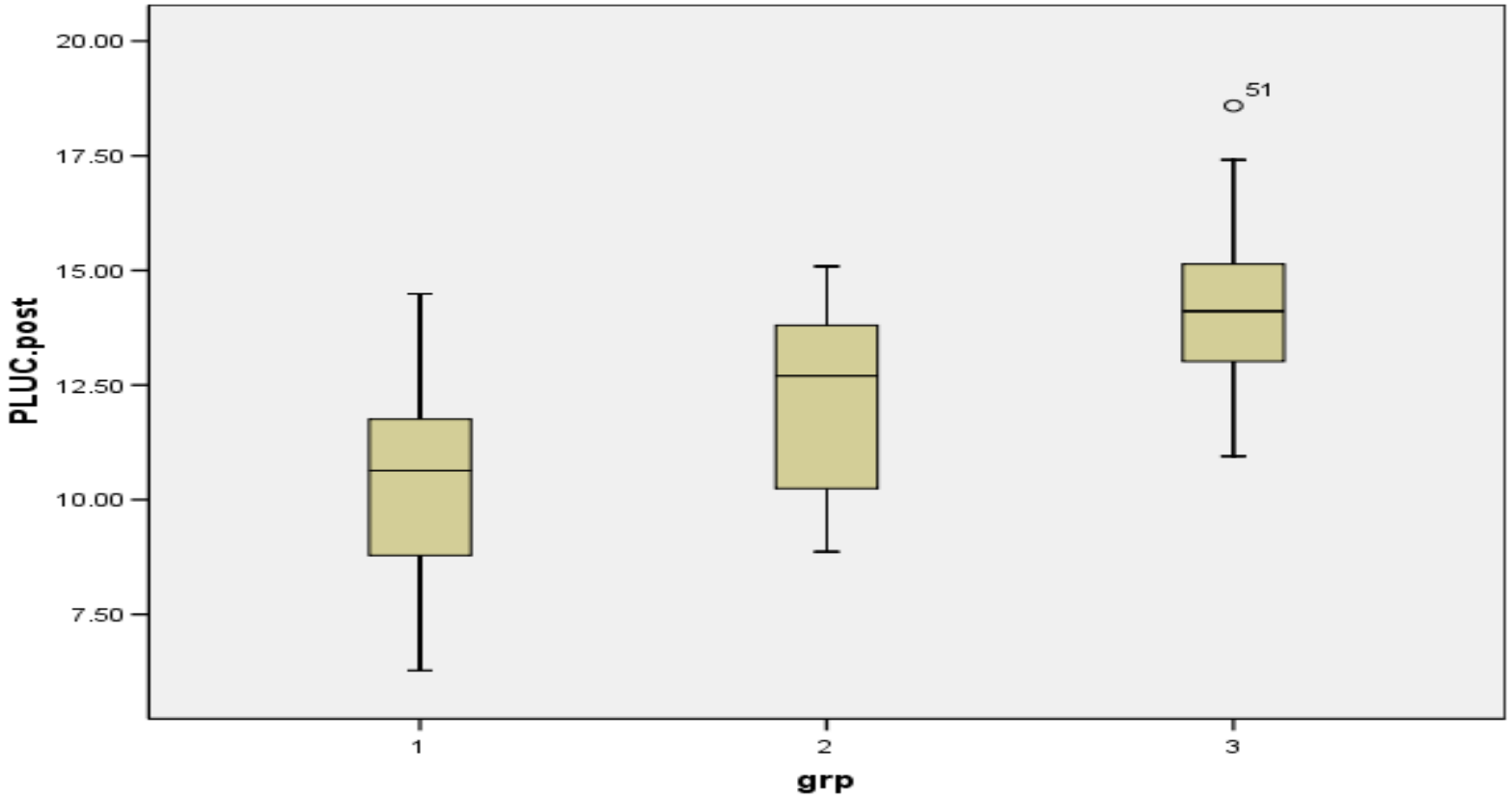
- The **analysis of variance (ANOVA)** is a technique of decomposing the total variability of a response variable into:
  - Variability due to the experimental factor(s) and…
  - Variability due to error (i.e., factors that are not accounted for in the experimental design).
- The basic purpose of ANOVA is to test the equality of several means.
- A **fixed effect model** includes only fixed factors in the model.
- A **random effect model** includes only random factors in the model.
- A **mixed effect model** includes both fixed and random factors in the model.

# Analysis of variance (ANOVA)

- Graphical investigation:
  - side-by-side box plots
  - multiple histograms

# Analysis of variance (ANOVA)

## Graphical Investigation: Side by Side Boxplots



# One-way ANOVA

- One factor of k levels or groups. E.g., 3 treatment groups in our default data
- Total variation of observations (SST) can be split in two components: variation between groups (SSG) and variation within groups (SSE).
- Variation between groups is due to the difference in different groups. E.g. different treatment groups or different doses of the same treatment.
- Variation within groups is the inherent variation among the observations within each group.
- Completely randomized design (CRD) is an example of one-way analysis of variance.

# One-way ANOVA

- Model:
  - $y_{ij} = \mu + a_i + e_{ij}$
  - Where $y_{ij}$ is the $i^{th}$ observation of the $j^{th}$ group
  - $a_i$ is the effect of the $i^{th}$ group
  - $\mu$ is the grand mean and $e_{ij}$ is the error.
- Assumptions:
  - Observations $y_{ij}$ are independent.
  - $e_{ij}$ are normally distributed with mean zero and constant standard deviation.
  - The second assumption implies that response variable for each group is normal (Check using q-q plot, histogram, or test for normality) and standard deviations for all groups are equal (rule of thumb: ratio of largest to smallest are approximately 2:1).

# One-way ANOVA

- Hypothesis:
  - $H_o$: Means of all groups are equal.
  - $H_a$: At least one of them is not equal to other.
    - doesn't say how or which ones differ.
    - Can follow up with "multiple comparisons"
- ANOVA Table for one way classified data

| Sources of Variation | Sum of Squares | df | Mean Sum of Squares | F-Ratio |
|---|---|---|---|---|
| Group | SSG | k-1 | MSG=SSG/k-1 | F=MSG/MSE |
| Error | SSE | n-k | MSE=SSE/n-k | |
| Total | SST | n-1 | | |

Note: Large F means that MSG is large compared to MSE

# One-way ANOVA

- Significance of the differences between the groups depends on
  - the difference in the means
  - the standard deviations of each group
  - the sample sizes
  - A useful web
    www.psych.utah.edu/stat/introstats/anovaflash.html
- ANOVA determines P-value from the F statistic

# Multiple comparisons

- If the F test is significant in ANOVA table, then we intend to find the pairs of groups are significantly different. Following are the commonly used procedures:
  - Fisher's Least Significant Difference (LSD)
  - Tukey's HSD method
  - Bonferroni's method
  - Scheffe's method
  - Dunn's multiple-comparison procedure
  - Dunnett's Procedure

# SPSS demonstration: One Way ANOVA

- Analyze -> Compare Means -> One-Way ANOVA-> select a dependent variable (e.g. PLUC.post) and a Factor (e.g. grp) and then select other output options. For Multiple comparisons, click on Post Hoc and select a suitable test.

# Thank you