# Foundations II Institute: The Advanced Practice of Institutional Research

# Intermediate / Advanced Research Design & Statistics

## Robert Ploutz-Snyder, Ph.D.

Biostatistician NASA JSC
USRA / Division of Space Life Sciences,
Research Associate Professor of Medicine
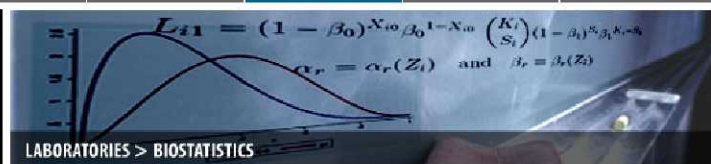SUNY Upstate Medical University

Association for Institutional Research

Foundations II Institute 2009

Association for Institutional Research

# Introductions…

# Previously at SUNY Cortland



- Built and Directed the Office of Institutional Research & Assessment
  - 2/1999 – 12/2001

- Three Primary Functions of the Office

  - **Institutional Research**
    - Ex. Enrollment Management
    - Ex. Salary Studies

  - **Institutional Assessment**
    - General Education
    - Program Majors
    - Administrative Units

  - **Institutional Data Warehousing & Reporting**
    - SUNY/NYS
    - Middle States & Other Accrediting Bodies
    - The Usual Hodgepodge of others…

# More Importantly, *Who are YOU*?

- How many Directors of IR Offices?
  - New Directors?
- How many Associate/Assistant Directors?
  - New to your position?
- How many IR Analysts with 5+ years experience?
- How many IR Analysts with less than 5 years experience?
- Other??

# Purpose of This Module

- To provide Institutional Researchers with an understanding of the principles of advanced research design and the intermediate/advanced statistical procedures consistent with such designs

# You Will Learn How To Use

- Independent Measures Analysis of Variance (ANOVA)
- Repeated Measures ANOVA/MANOVA Analysis of Covariance (ANCOVA)
  - ANOVA with Covariates
- Simple and Multiple Regression
  - Block Regression
  - Forwards, Backwards, Stepwise Regression

# You Will Also be Exposed To

- Exploratory Factory Analysis
  - Principal-Axis Factoring
    - With Varimax Rotation
- Time Series Regression

# I Assume That

- This isn't your first course in statistics?
- That you have the basics covered
  - Foundations I Level Stats
- That you have SPSS loaded on your laptop machines
- That you are interested and motivated to learn!

# Format of This Module

- Hands-On!
  - I'll talk about the statistical tests, assumptions, theory first
  - Then we'll walk through analyses together
    - Using SPSS
  - We'll pay a lot of attention to
    - Analytic choices that you have
    - Interpreting the output
    - Presenting the results to your constituents

# Quick Review:
# Gaussian Distribution Function

- A.K.A. The "Normal Distribution"
- A.K.A. The "Bell-Shaped Curve"
- Has known probabilities associated with it,
- Thus all Parametric Statistics are based on the Gaussian Distribution

Gaussian or "normal" distribution

$f_g(x)$

.0214

.00135    .1359   .3413   .3413   .1359

.0214

.00135

$-3\sigma$   $-2\sigma$   $-\sigma$   0   $\sigma$   $2\sigma$   $3\sigma$

x

$$f_g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\bar{x})^2/2\sigma^2}$$

Where $\bar{x}$ = mean, and $\sigma$ = standard deviation
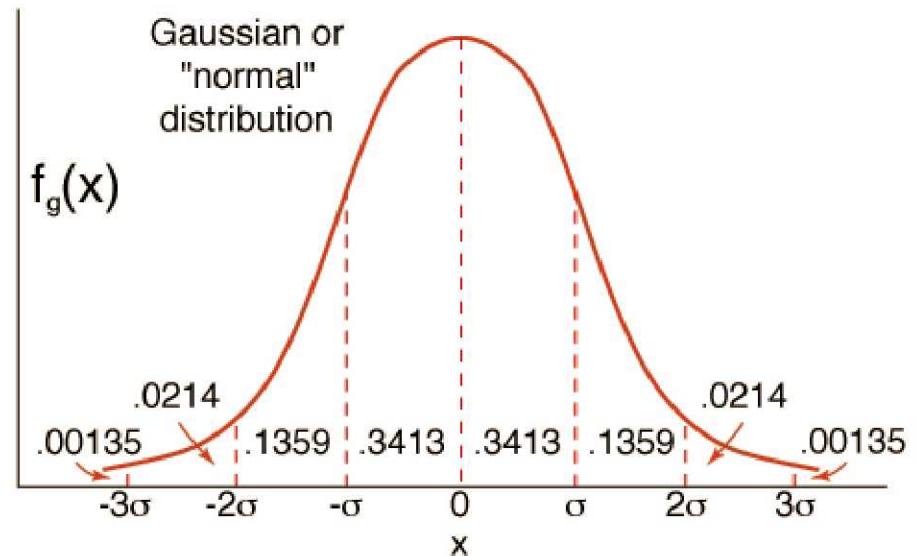
# Quick Review: Gaussian Distribution Function

- About 68% of all scores fall within 1 SD unit from the mean.



Gaussian or "normal" distribution

$f_g(x)$

68%

.0214

.00135        .1359                    .1359        .0214        .00135

-3σ    -2σ    -σ    0    σ    2σ    3σ

x

# Quick Review: Gaussian Distribution Function

- About 68% of all scores fall within 1 SD unit from the mean.

- About 95% of all scores fall within 2 SD units from the mean.

Gaussian or "normal" distribution

$f_g(x)$

95%

.0214

.00135

.0214

.00135

-3σ  -2σ  -σ  0  σ  2σ  3σ

x
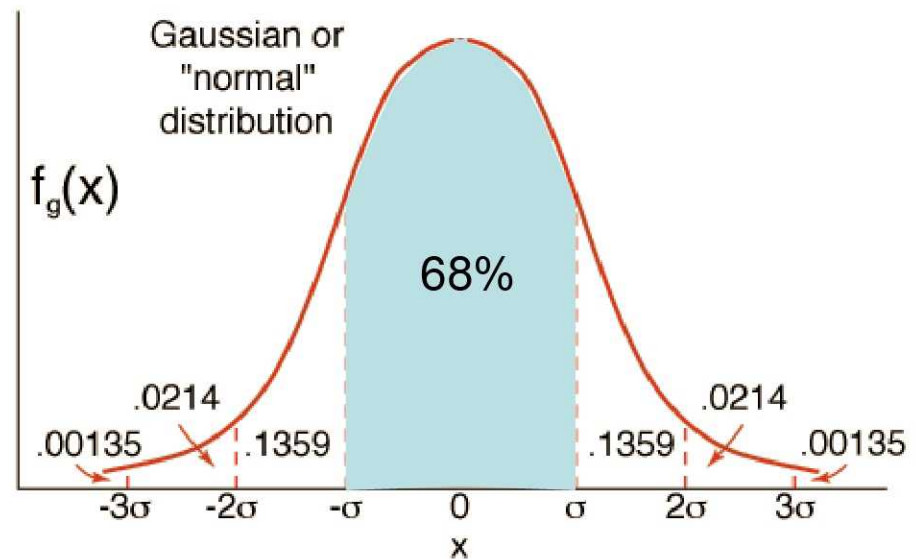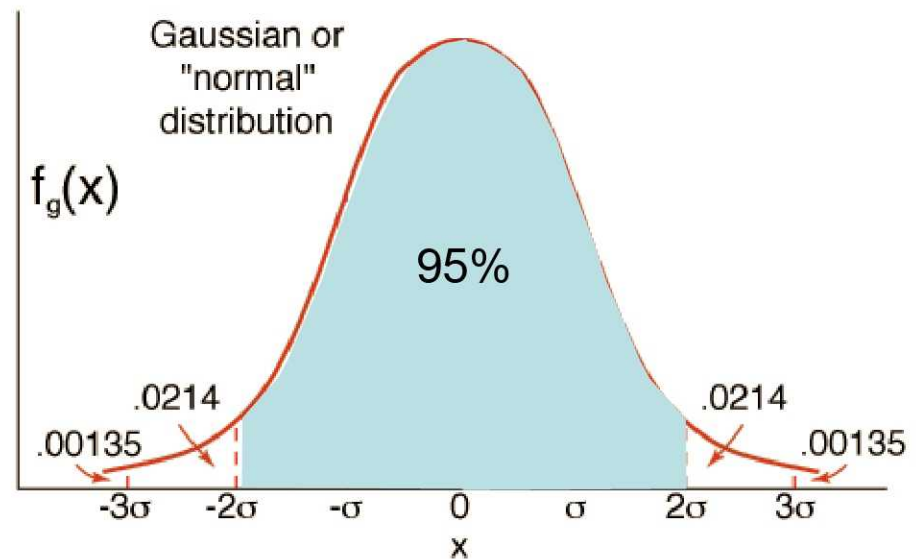
# Quick Review: Gaussian Distribution Function

- About 68% of all scores fall within 1 SD unit from the mean.
- About 95% of all scores fall within 2 SD units from the mean.
- About 99% of all scores fall within 3 SD units from the mean.



Gaussian or "normal" distribution

$f_g(x)$

99%

.0214    .0214
.00135    .00135

-3σ   -2σ   -σ   0   σ   2σ   3σ
x
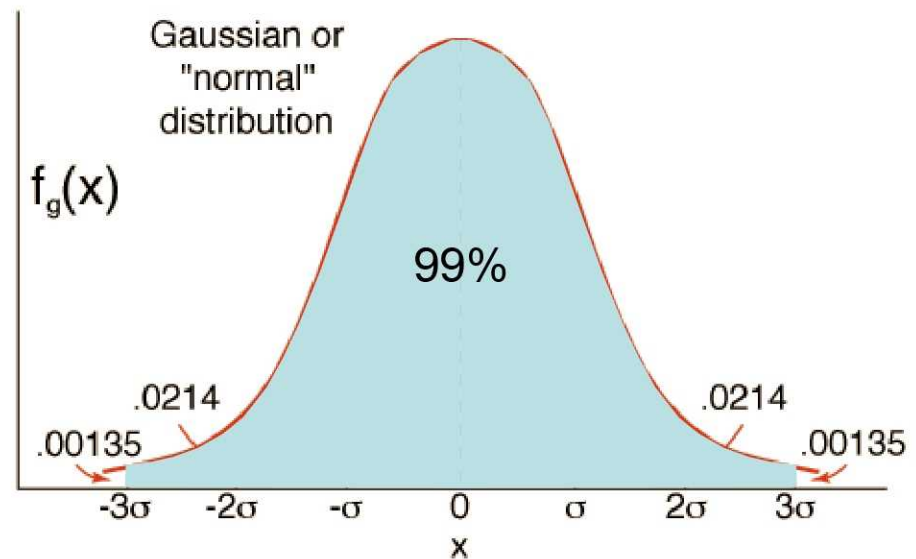
# Central Limit Theorem

- States that for any population with mean μ and standard deviation $\sigma$ , the distribution of sample means with sample size *n* will approach a *normal distribution* with μ and SD of $\sigma/\sqrt{n}$ as n approaches infinity.
- REGARDLESS of the shape of the distribution in the population.
- By the time sample sizes hit around 30, sampling distribution of means is close to normal.

# Demo of central limit theorem.
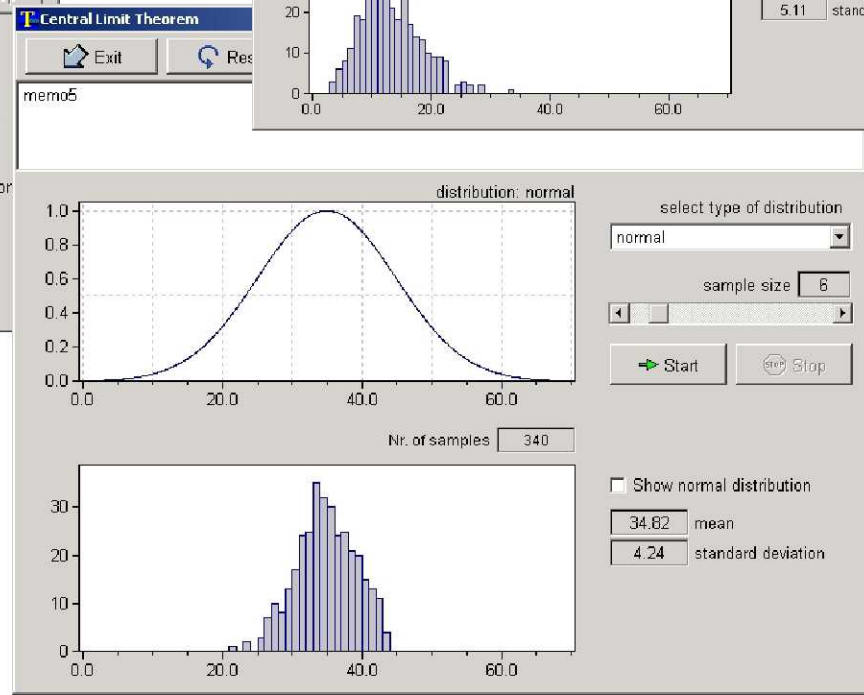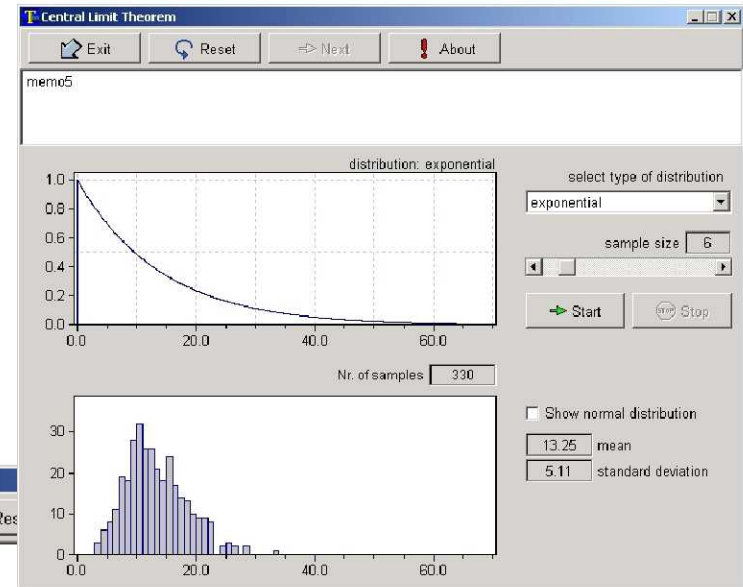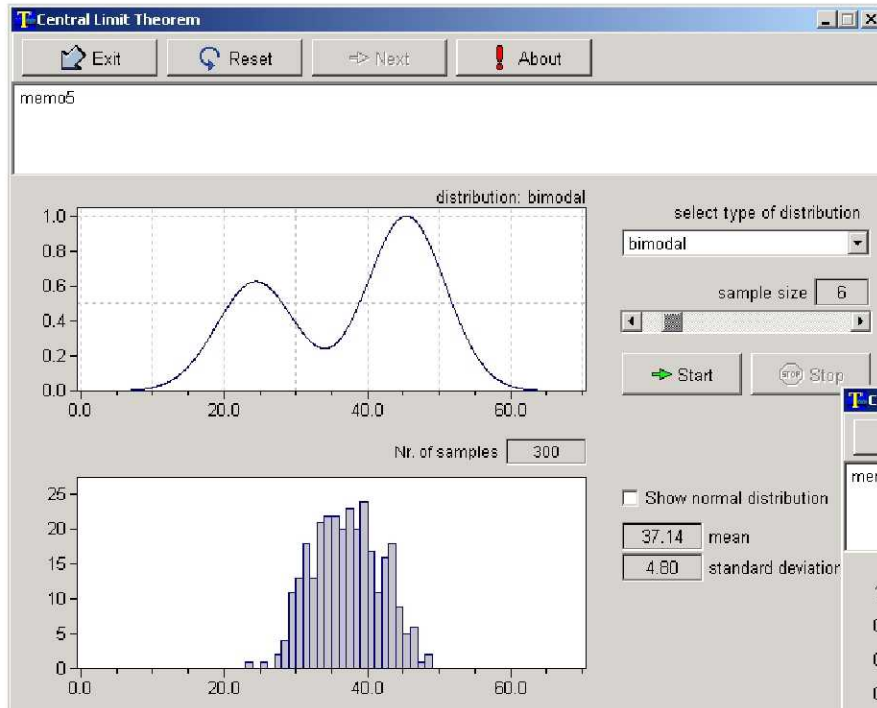
# Thus…

- Since we know so much about the Normal Distribution
- And we know that sample summaries (means or otherwise) tend to follow that distribution
  - Even data collected from non-normal samples
  - Especially so with large sample size (big-n)
- We can usually apply our knowledge of the normal distribution to statistical comparisons, estimates, and probability
  - As long as we do some preliminary screening…

# Ex. You may recall…

- That we can compare a person's score to their population with the "Z-Score"
  - Z is a "standardization"
    - Mean = 0
    - SD = 1
    - Probability tables tell us percentiles, probabilities of being "that far" away from the mean.

# Z-score quick review

- A student takes a standardized test and scores XXX.

- Can compare their score to the population of all test-takers during that time, given population mean and standard deviation as:

$$Z = \frac{X - \mu}{\sigma}, \text{ where}$$

$\mu$ is population mean and $\sigma$ is population std dev

# Z-score quick review

- With their Z-score, we can glean
  - Their percentile rank  [p(lower)]
  - Probability of scoring higher than them
  - Other relevant probabilities.



Gaussian or "normal" distribution

$f_g(x)$

.0214

.00135 | .1359 | .3413 | .3413 | .1359 | .0214 | .00135

-3σ   -2σ   -σ   0   σ   2σ   3σ

x
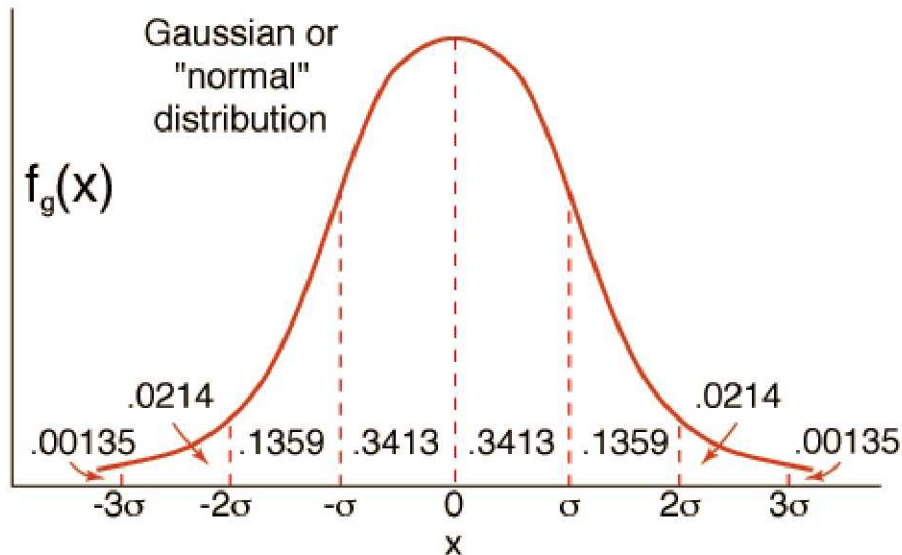
# T-statistic for Comparing Sample to Population

- Where we don't know SD of the population, but we have sample data
  - Thus sample mean and sd
- And we know from CLT that we can estimate population SD by SE

$$SE = s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

# T-statistic for Comparing Sample to Population

- Given our sample data, we can calculate
  - Sample mean
  - Sample SD (s)
- Given SE formula

$$SE = s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- We can calculate Confidence Intervals on this Estimate also
- And with t-tables…p-values

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

# Moving to the t-test for comparing two samples

- Used for comparing two samples collected randomly from two populations

- Fairly simple modifications of the t-statistic comparing sample mean to population mean

Population I

Population 2

Sample 1
X=0
X=2
X=4

$\overline{x} = 2$

$s = 2$

Sample 2
X=4
X=6
X=8

$\overline{x} = 6$

$s = 2$

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{s_{\overline{X}_1 - \overline{X}_2}} \quad \text{where} \quad s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$\text{and} \quad s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2}$$

# Dissect the formula: $s_{\overline{X}_1 - \overline{X}_2}$

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{s_{\overline{X}_1 - \overline{X}_2}} \text{ where } s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$\text{and } s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2}$$

# Dissect the formula: Numerator

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{s_{\overline{X}_1 - \overline{X}_2}}$$

The difference between two sample means

# Dissect the formula: Denominator

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{s_{\overline{X}_1 - \overline{X}_2}}$$

The difference between two sample means

Divided by some measure of standard error of the differences

# Dissect the formula: Question?

$$t = \frac{\left( \overline{X}_1 - \overline{X}_2 \right)}{s_{\overline{X}_1 - \overline{X}_2}}$$

The difference between two sample means

Divided by some measure of standard error of the differences

Are the differences that I see between my two means unusual, given variability among other sample means of this size?

# T-tests on the Computer:

- Software gives us t-score and a p-value
- Allowing us to test hypotheses that the two samples come from the same population (or not)
- And describe the magnitude of the differences (confidence intervals)
- Ex. t = 4.87, p<.001
  - $H_{null}$: Two samples are from same population
  - $H_{alt}$: Two samples are from different populations
- Reject the Null (alpha < .05) & Report the magnitude of the differences

# Virtues of the t-test

- EVERYONE seems to understand it!

- With CLT, it's easy to apply to lots of different data scenarios

- There are other versions that make it very flexible

  – Formula for "Repeated Measures" designs

  – Formula for problems associated with non-normality and/or variance heterogeneity

# Limitations of t-tests

- Alpha risk is .05 for each t-test
  - Probability of falsely rejecting the null, and concluding that there is a difference, when it's really due to chance.
  - So comparing 3, 4, 5 or more groups is quite problematic!
- With large samples, as with ANY statistical test, "significance" does not necessarily indicate a *meaningful* difference.

# Comparing Three Groups

Group I

Group 2

Group 3

# Comparing Three Groups



Group I

Group 2

Group 3

T-test number 1
Alpha risk = .05
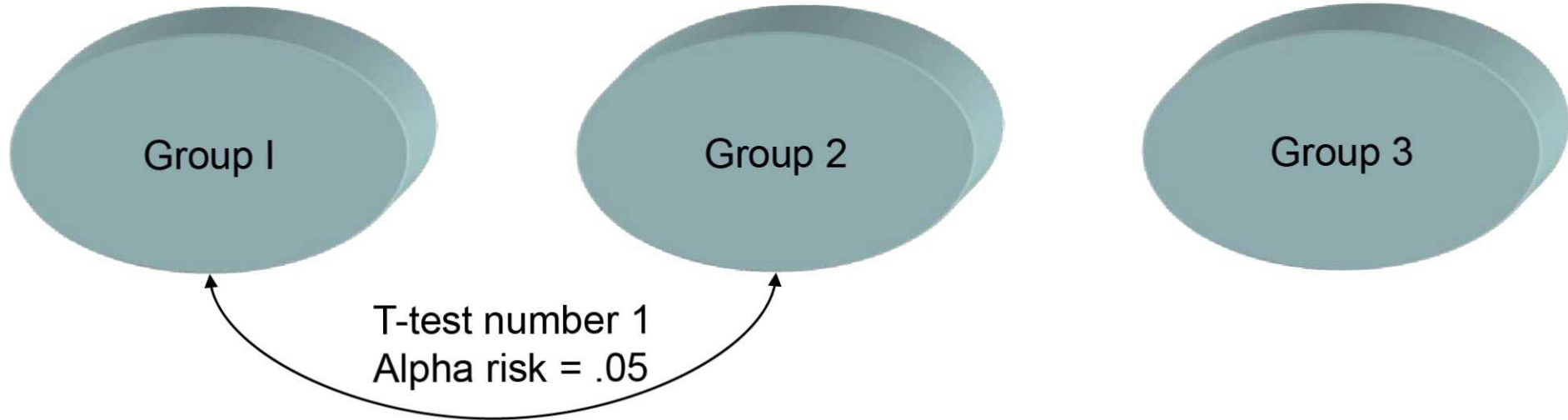
# Comparing Three Groups

# Comparing Three Groups

Group I

Group 2

Group 3

T-test number 1
Alpha risk = .05

T-test number 2
Alpha risk = .05

T-test number 3
Alpha risk = .05

# Comparing Three Groups

Group I

Group 2

Group 3

T-test number 1
Alpha risk = .05

T-test number 2
Alpha risk = .05

T-test number 3
Alpha risk = .05

Total Alpha Risk ~ .15
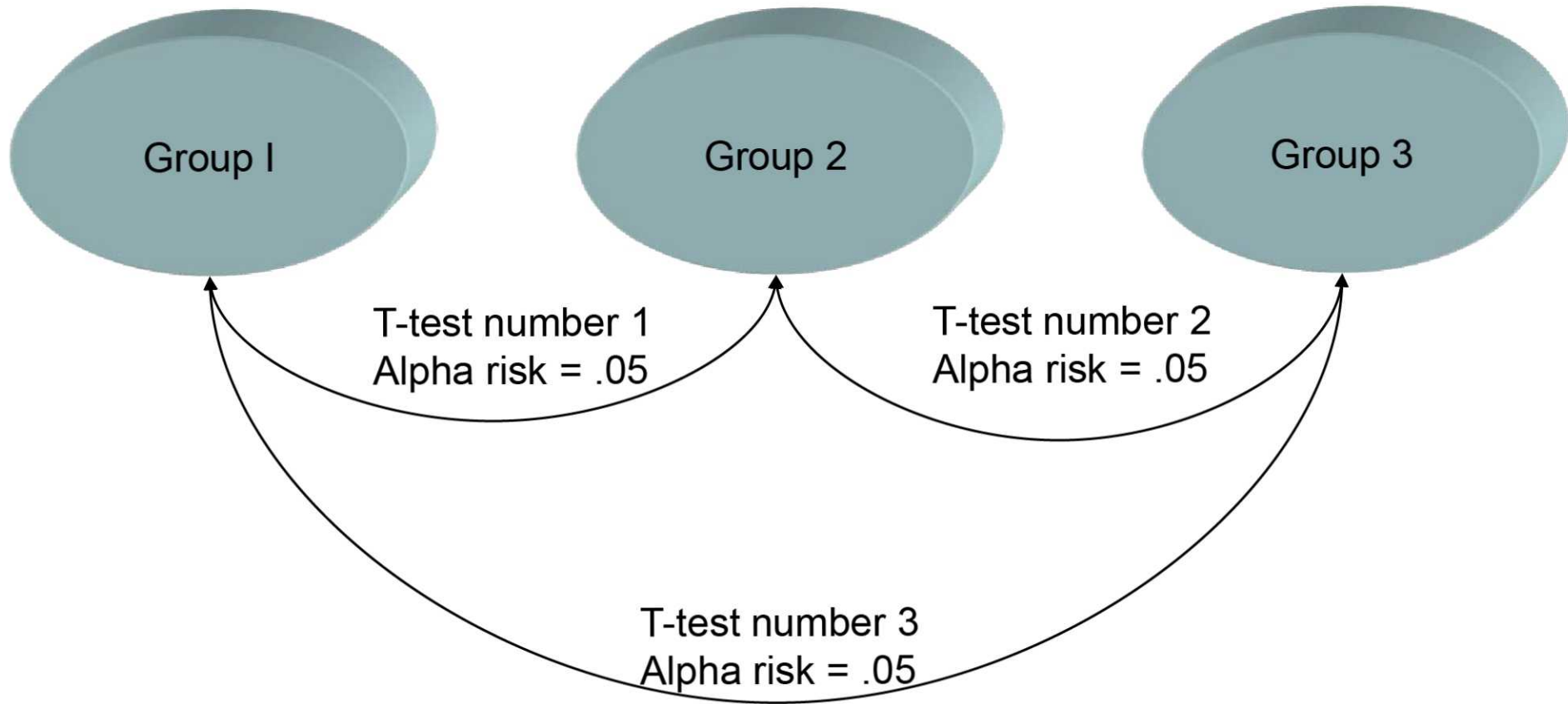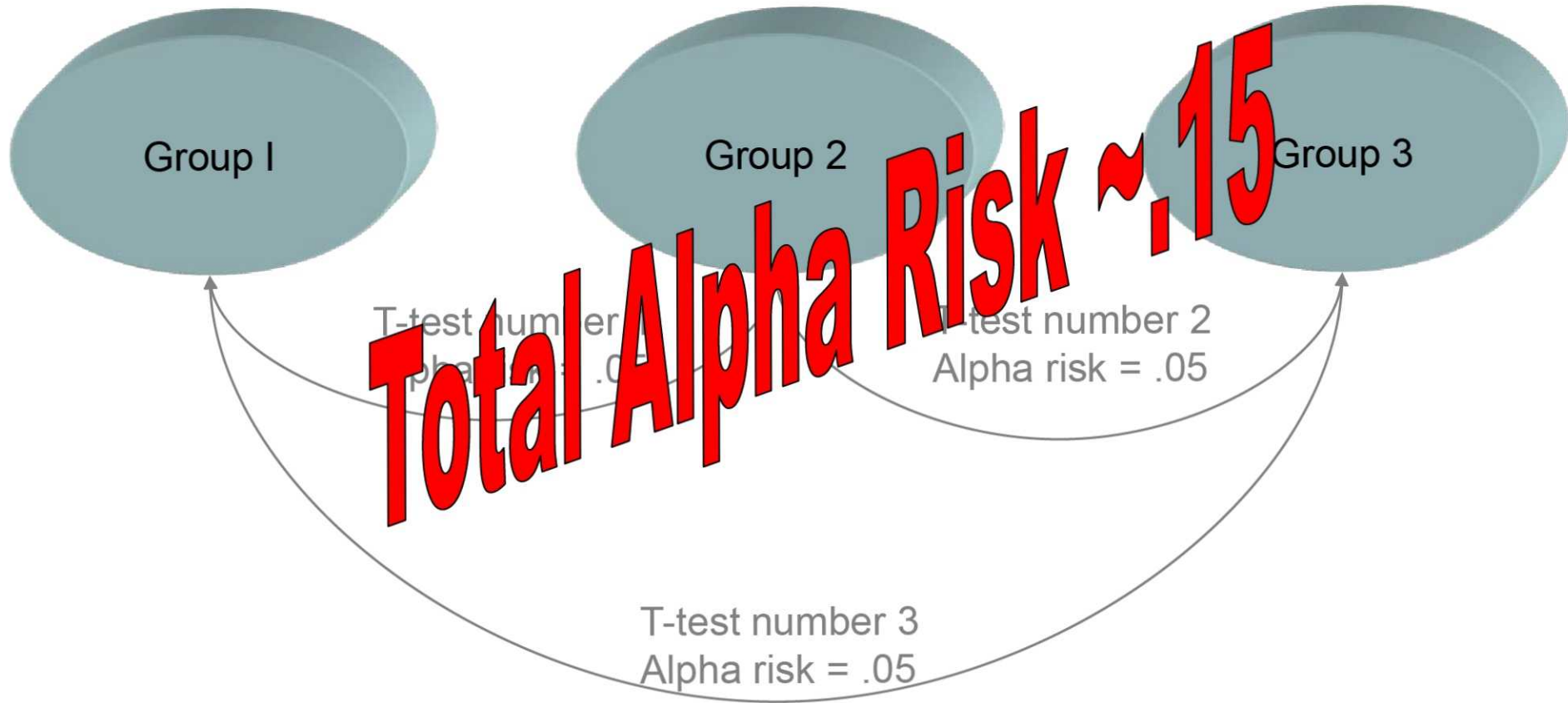
# Analysis of Variance (ANOVA)

- Can compare unlimited number of groups or occurrences, and still keep alpha risk = .05

- Able to take multiple grouping (or time) factors into account and determine their independent and combined effects

- Can examine "trends" in data, and can test specific (often complex) hypotheses

- The analytic focus is on variance, but the interpretation falls back to means—thus results become intuitive

# Assumptions Required of ANOVA

- Data collected randomly from the population, with roughly equal n per cell
  - And sufficiently large n (n>30, common r-o-t)
- Data measured on interval or ratio scale, and is normally distributed
- Homogeneity of variance across groups
- Sphericity for RM designs—variance of the differences between means for any pair of groups is equal to any other pair

# Assumption of Randomly Collected Data with Sufficiently Large n

- In IR, we don't always "randomly select"
  - But can we assume that "today's" data is a random representation of "recent years?"
  - Or can we START randomly selecting a subset of your populations for research?
- How big is big enough?
  - Rule of Thumb… at least 30 per group
  - More is better
    - Cautions about overpowered studies…
  - But BALANCE is critical!!
    - Rule of thumb—smallest group should not be less than 1/3rd the size of the largest group.

# Assumption of Interval or Ratio Scale & Normality

- The "bell-shaped" curve—assumption of all parametric statistics

- Studies show that ANOVA is robust to violations of this, but only if sample size is substantially large, and Homogeneity is met



Gaussian or "normal" distribution

$f_g(x)$

.0214     .0214
.00135   .1359  .3413  .3413  .1359  .00135

-3σ  -2σ  -σ  0  σ  2σ  3σ

x
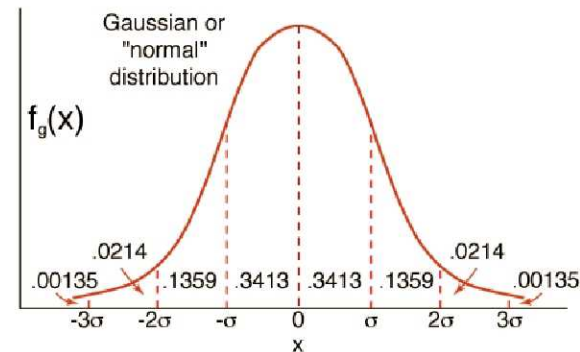
# Assumption of Homogeneity of Variance Across Groups

- Variance on the dependant variable should be similar across groups
  - Why?
- Because we're examining VARIANCE in ANOVA, and so we need for variance in each group to be roughly similar before we can conclude that any differences that we find are attributable to *group* differences (not mere variability differences).
- Even in Means Comparisons (ex.t-tests), since Means are highly affected by variability, we need for variability to be similar in our groups so that differences that we find can be attributed to true group differences, and not merely by variability differences between our groups.

# More on Homogeneity of Variance

- If distributions are normal in one, then should be for all
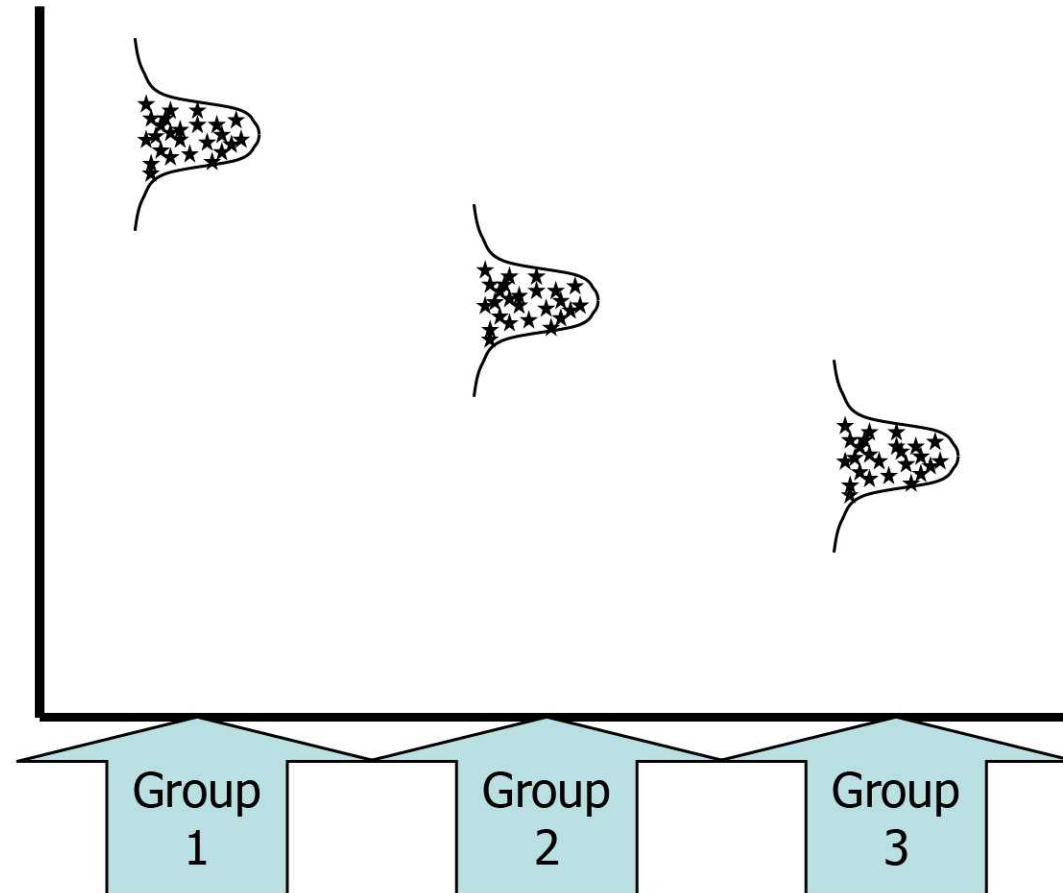
# More on Homogeneity of Variance

- If distributions in 1 group is leptokurtotic (tall and skinny), then it should be for all other groups



Group 1    Group 2    Group 3
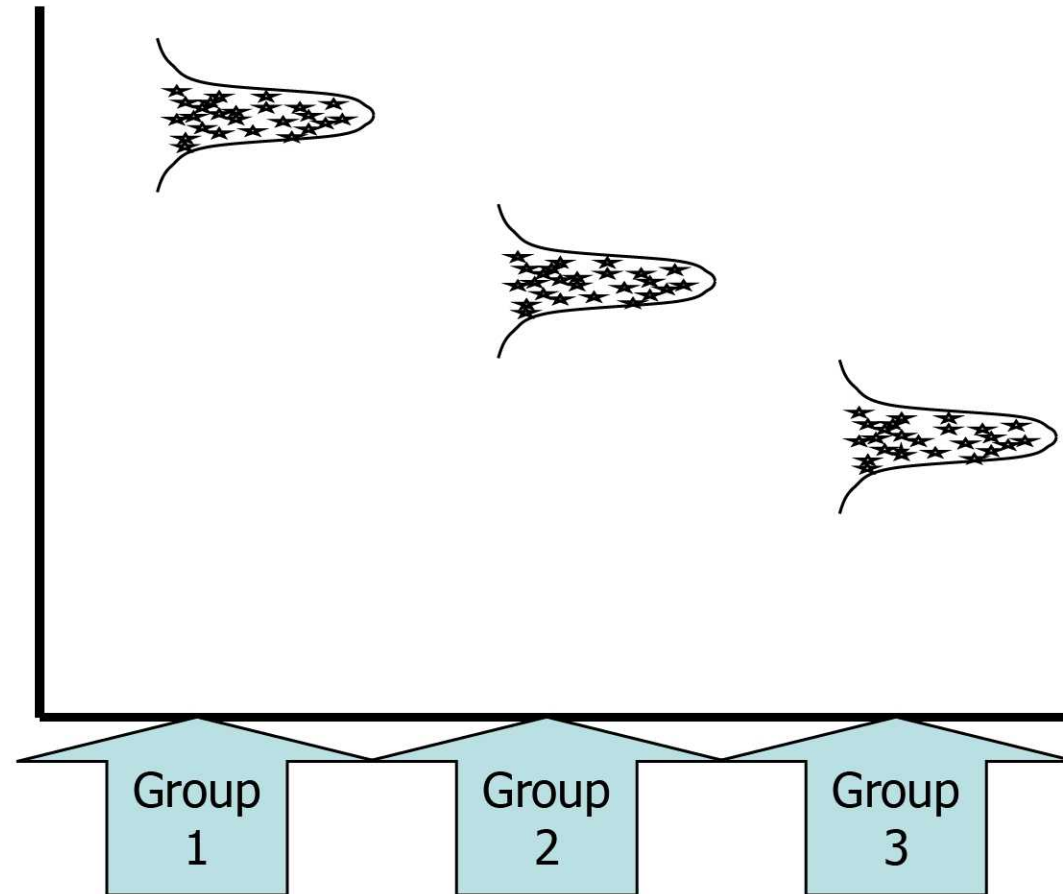
# More on Homogeneity of Variance

- If distributions in 1 group is platykurtotic (short & fat) then it should be for all other groups

Group 1
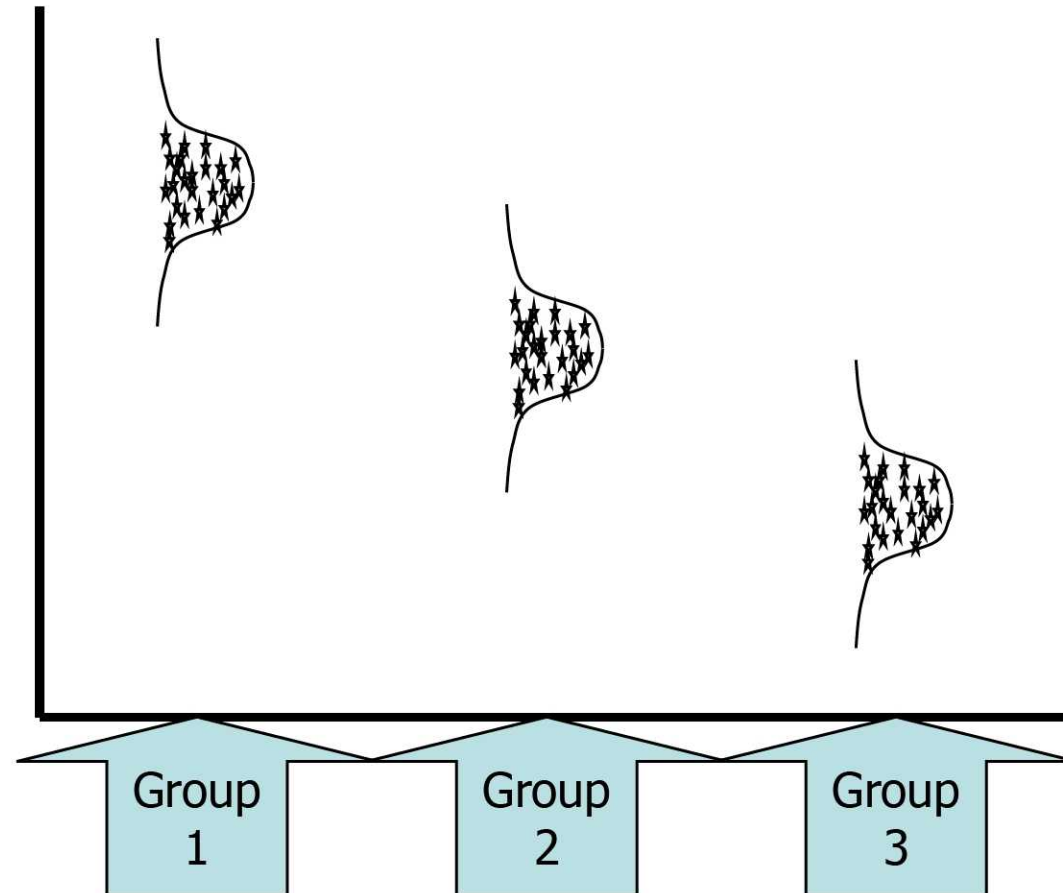
Group 2

Group 3

# More on Homogeneity of Variance

- Any Miss-Match is a Problem
  - Because we might interpret a statistical differences to real group differences, when it's actually due to heterogeneity of variance

- …Thankfully SPSS will test this assumption for us (stay tuned)
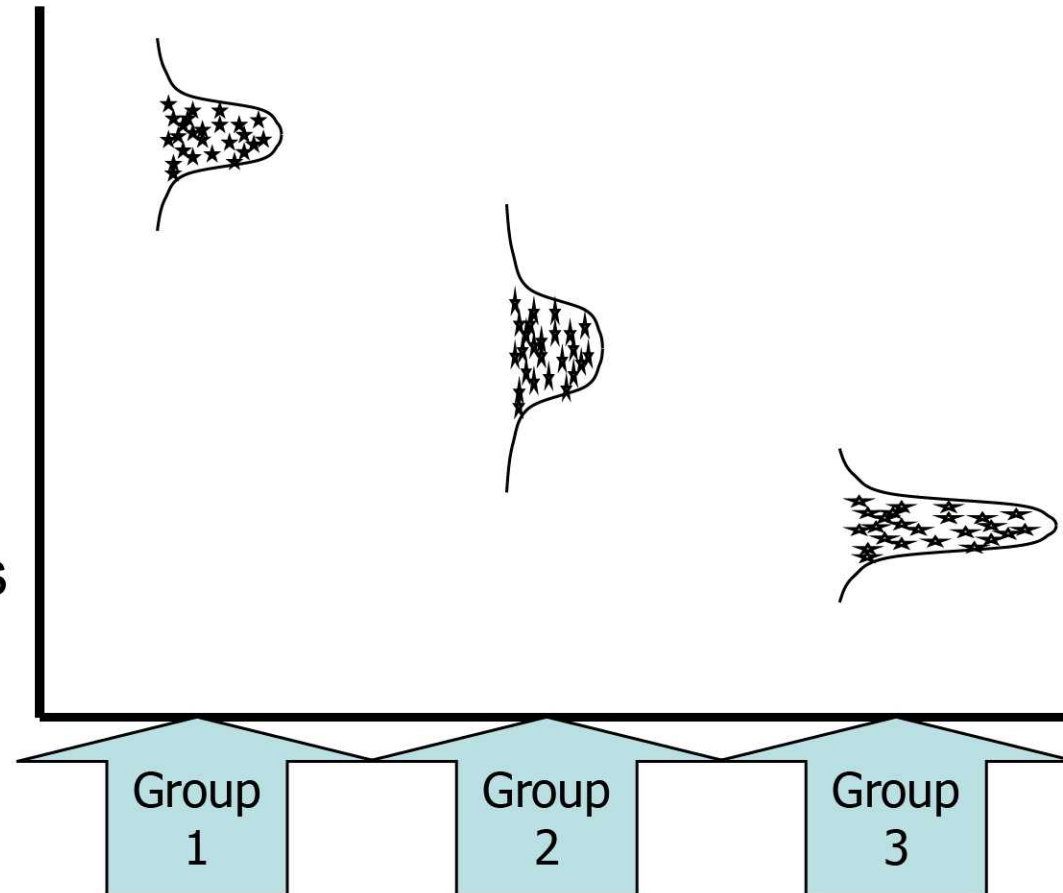
# What about skewed data?

- Positive or negative skews in the data can wreak havoc with statistical analysis
  - Thus always recommend thorough data screening
  - Identify outliers—data entry errors?
  - Consider data transformations if necessary
    - More on this later

# Two General Types of ANOVA

- **Independent Measures ANOVA (IM-ANOVA)**
  - Data are collected from separate groups of subjects, and comparisons among *groups* are desired
    - Student GPA by MAJOR
    - Faculty Salaries by DEPARTMENT

- **Repeated Measures ANOVA (RM-ANOVA)**
  - Data are collected from the same group of subjects on multiple occasions/times, and comparisons of *occasions* are desired.
    - Longitudinal Studies
    - Student Opinions as Fresh, Soph, Jr, Sr
    - Alumni Donations after 1, 3, 5, 7 years post-graduation

# IM & RM Designs…



Repeated Measures Designs
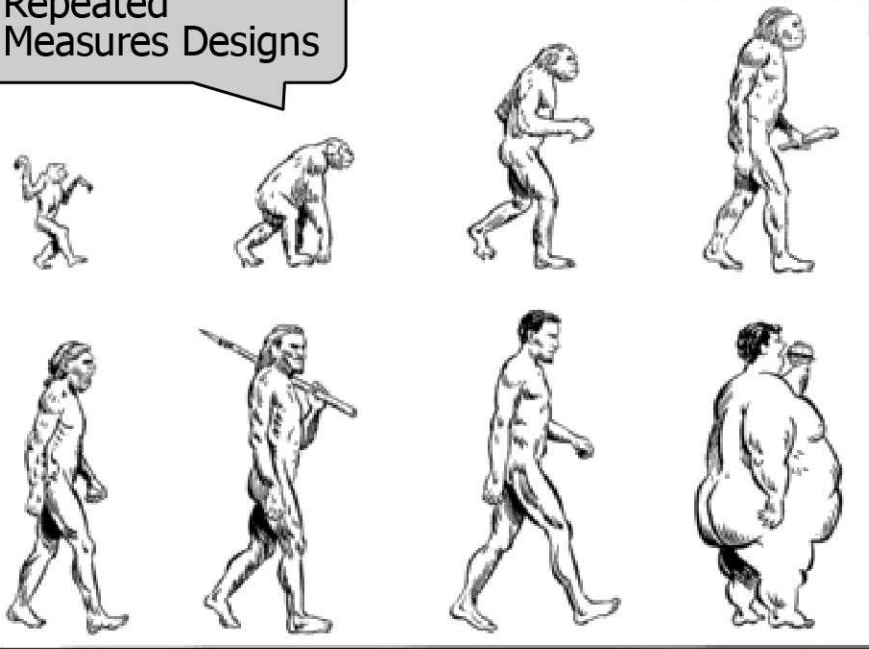
Independent Groups Designs

CONTROL GROUP

OUT OF CONTROL GROUP

# One-Way IM-ANOVA

- ## For comparing two or more populations
  - ### Where sample data have been collected
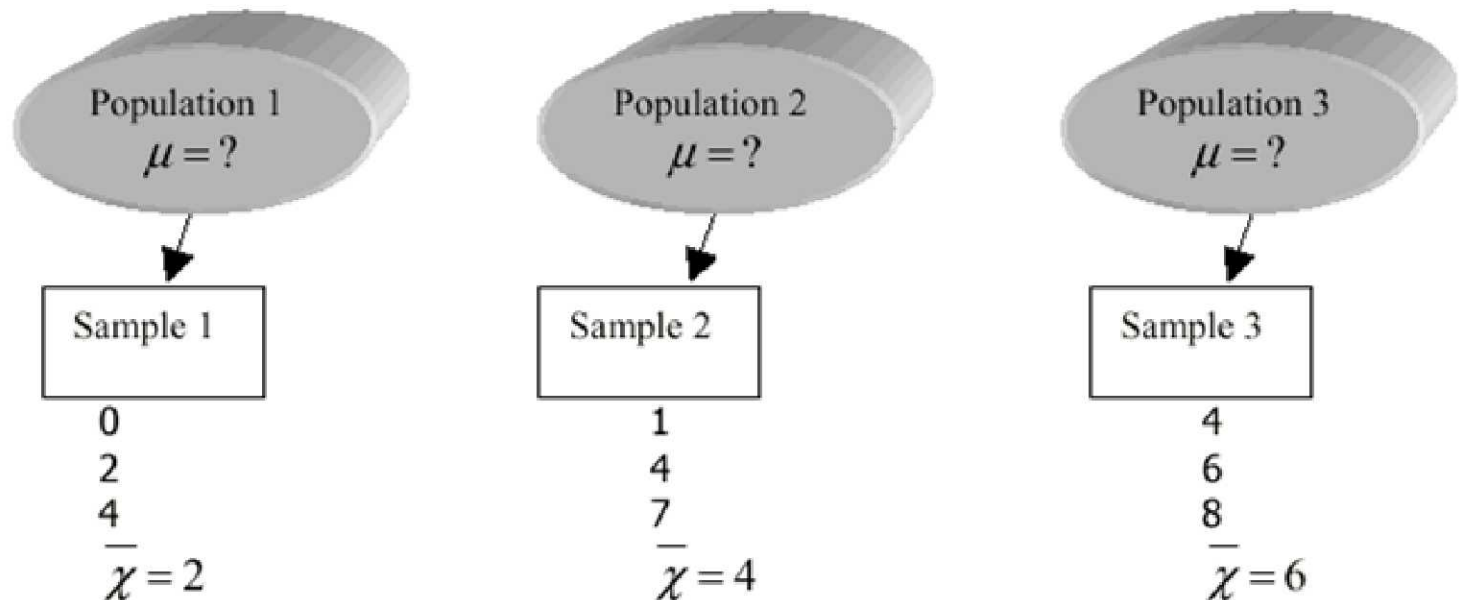
# ANOVA: What's in a Name?
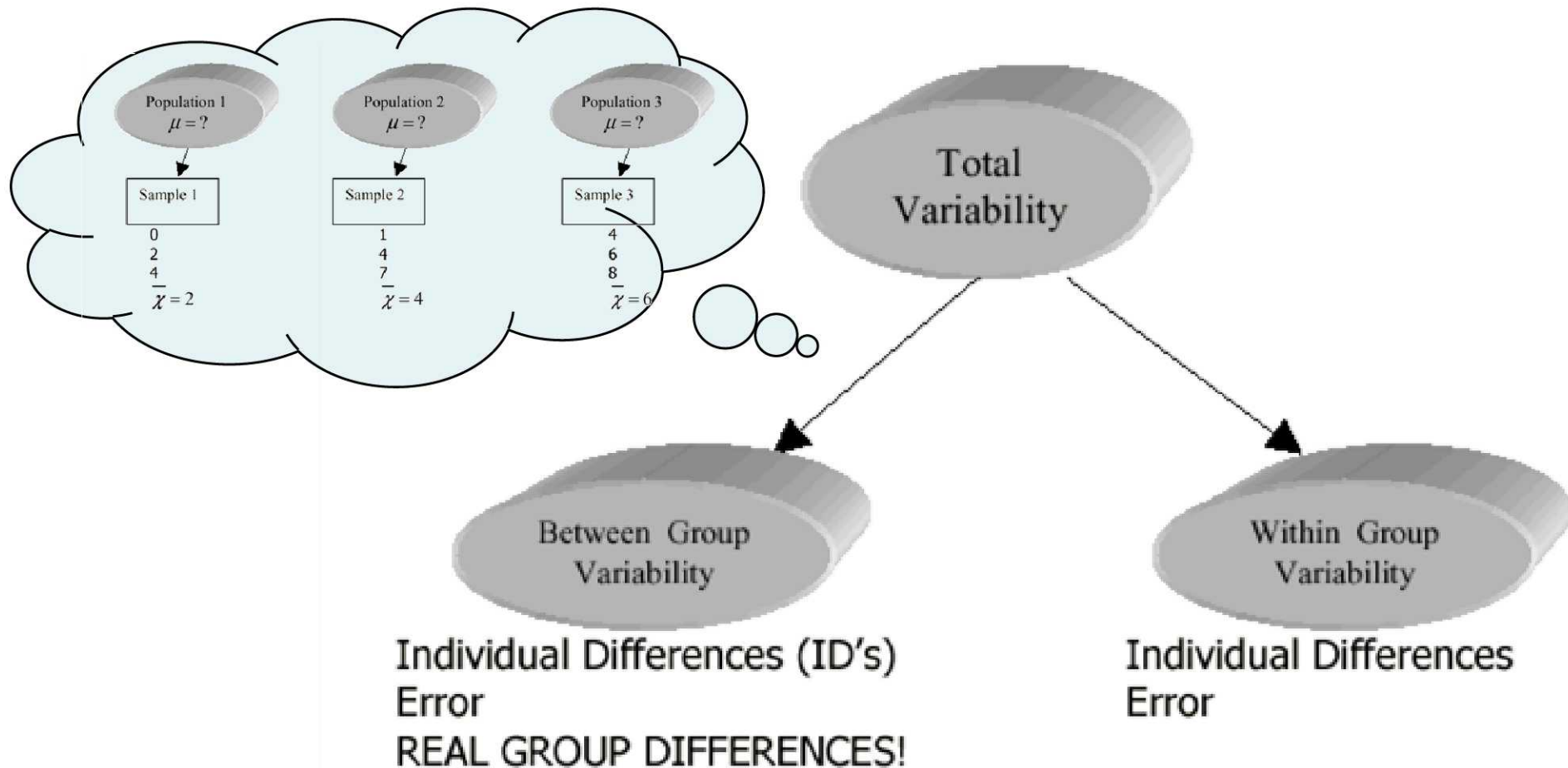
# Analysis of Variance F-Ratio

- ANOVA is truly an analysis of a measure of *variability*, called "*variance*"
  - Within-Groups Variability
  - Between-Groups Variability
- We Evaluate an "F-Ratio" Representing the Ratio of B/T over W/I:

$$F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{\text{ID's} + \text{error} + \text{group differences}}{\text{ID's} + \text{error}}$$

# Recall your Simple Algebra…

- If the same quantity exists in the Numerator and Denominator of a fraction, they "cancel each other out"

**Assuming homogeneity of variance**
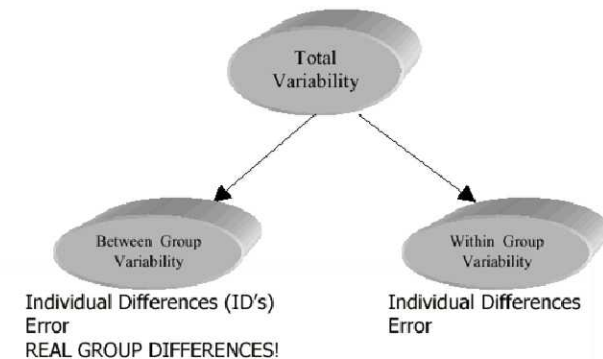
Total Variability

Between Group Variability

Within Group Variability

Individual Differences (ID's) Error
REAL GROUP DIFFERENCES!

Individual Differences Error

**The F-Ratio**

$$F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{\text{ID's} + \text{error} + \text{group differences}}{\text{ID's} + \text{error}}$$

# Recall your Simple Algebra...

- If the same quantity exists in the Numerator and Denominator of a fraction, they "cancel each other out"

  - Leaving us with a number (F) that represents Group Differences!

**The F-Ratio**

$$F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{\text{ID's} + \text{error} + \text{group differences}}{\text{ID's} + \text{error}}$$

# Analysis of Variance F-Ratio



- If F=1…

- As F increases…

- How do you know if F is "big enough" to considered significant?

  – How do you know a t-test is significant??
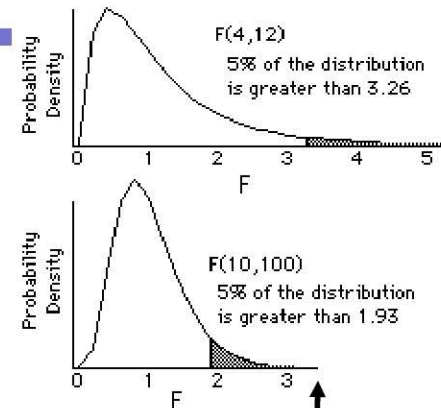
**The F-Ratio**

$$F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{\text{ID's} + \text{error} + \text{group differences}}{\text{ID's} + \text{error}}$$

# Confidence Intervals with the F-test

- CI's for comparing two groups are straightforward and intuitive
- CI's for "Omnibus" differences are less so
  - Effect size calculations exist, but non-intuitive to statistically naieve
- Stay tuned for discussions about post-hoc tests, and how they can sometimes help
- Plots will also be very informative

# IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F
- Total, Between Groups, Within Groups

# IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-t[ ]

  Sum of Squared Deviations from the Mean

  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F

- Total, Between Groups, Within Groups

# IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F

- Total, Between Groups, Within Groups

> Like in a t-test, each F-test has df values for significance testing

# IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F

- Total, Between Groups, Within Groups

MS is the Variance Statistic for ANOVA—calculated with SS & df

# IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)

    The "F" statistic is another word for the F-ratio

  - Probability values associated with F
- Total, Between Groups, Within Groups

# IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F

...and p values tell us the significance level of the ratio

- Total, Between Groups, Within Groups

# This is what it looks like…

|  | df | SS | MS | F | p |
|---|---|---|---|---|---|
| Between Groups | ## | ## | ### |  |  |
| Within Groups (error) | ## | ## | ### | #.# | .## |

# This is where it comes from (Independent Measures Designs)

$$SS_{total} = \sum_{i=1}^{N} \left(x_i - \overline{G}\right)^2 = \sum X^2 - \frac{\left(\sum x\right)^2}{N}$$

$$SS_{between} = \sum_{k=1}^{k} n_k \left(\overline{X}_k - \overline{G}\right)^2$$

$$SS_{within} = \sum SS_{\text{inside each group}}$$

$$df_{total} = N - 1$$

$$df_{between} = k - 1$$

$$df_{within} = N - k$$

# This is where it comes from (Independent Measures Designs)

$$MS_{total} = \frac{SS_{total}}{df_{total}}$$

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$MS_{within} = \frac{SS_{within}}{df_{within}}$$

$$F = \frac{MS_{between}}{MS_{within}}$$

F-tables provide a <u>p</u> value for a given F-statistic, using $df_{between}$ (numerator) and $df_{within}$ (denominator).

# Example 1

- Compare Faculty Ratings Across 3 Departments
  - History
  - Psychology
  - Math

- Simplest of ANOVA Models, with ONE Independent Factor (department)

# The Data:

# One-Way Point-n-Click:



Syntax method:
ONEWAY
 eval BY dept
 /STATISTICS DESCRIPTIVES HOMOGENEITY
 /PLOT MEANS
 /MISSING ANALYSIS
 /POSTHOC = TUKEY BONFERRONI SIDAK GH ALPHA(.05).

# One-Way ANOVA Output

File   Edit   View   Data   Transform   Insert   Format   Analyze   Graphs   Utilities   Add-ons   Window   Help

Output
- Oneway
  - Title
  - Notes
  - SPSS Text
  - Descriptives
  - SPSS Text
  - Test of Homogeneity of Variance
  - SPSS Text
  - ANOVA
  - SPSS Text
- Post Hoc Tests
  - Title
  - Multiple Comparisons
  - SPSS Text
  - Homogeneous Subsets
    - Title
    - Student eval of Intro cour
- Means Plots
  - Title
  - Student eval of Intro course
  - SPSS Text
  - Student eval of Intro course
- Univariate Analysis of Variance

## Oneway--IR Example #1:  Oneway Independant-Measures ANOVA

**Notes**

| Output Created | | 12-JUL-2005 15:24:11 |
|---|---|---|
| Comments | | |
| Input | Data | U:\AIR Stat Institutes & book chapters\AIR Stat Institute 2005\SPSSStuff\IREX1.sav |
| | Filter | <none> |
| | Weight | <none> |
| | Split File | <none> |
| | N of Rows in Working Data File | 253 |
| Missing Value Handling | Definition of Missing | User-defined missing values are treated as missing. |
| | Cases Used | Statistics for each analysis are based on cases with no missing data for any variable in the analysis. |
| Syntax | | ONEWAY eval BY dept /STATISTICS DESCRIPTIVES HOMOGENEITY /PLOT MEANS /MISSING ANALYSIS /POSTHOC = TUKEY ALPHA(.05). |
| Resources | Elapsed Time | 0:00:00.20 |

NOTE The bold face, Red font above.  SPSS Normally "reduces" the Notes output, but it's handy to open it back up if you forget the exact syntax that was used to execute a statistical procedure.  Even if you don't always run stats from syntax, it's nice to see that you COULD duplicate an earlier analysis perfectly by copying the syntax from the Notes output.  NOTE Also that the file location and name are shown here too...

**Descriptives**

Student eval of Intro course

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| Psych 101 | 86 | 6.65 | 1.713 | .185 | 6.28 | 7.02 | 4 | 9 |

# Two-Factor ANOVAS

- ## What if you want to compare 2+ groups on MORE THAN one factor?

  – Effect of students' gender *and* race/ethnicity on performance?

  – Effect of students' major *and* high school on cGPA?

  – Effect of faculty members' level *and* age on job satisfaction ratings

# Two-Factor ANOVA Effects

- ## Main Effects
  - One per factor…an F-statistic evaluating the impact of each factor in the model
    - Gender effect on performance (M/F diffs?)
    - Race/ethnicity effect on performance

- ## Interaction Effects
  - One per interaction… an F-statistic evaluating how two (or more) factors interact with one another to affect the outcome
    - Gender "by" Race/Ethnicity interactive effects on performance
    - More complex…often more interesting!

# Example 2

- Compare first term GPA by Major and Citizenship
  - U.S. versus non-U.S.
  - Three Majors—Math, Business, US History
- 2 x 3 ANOVA

# The Data

# Repeated Measures ANOVA

- Only one sample
- Differences based on time, or condition
- Using the SAME subjects time after time
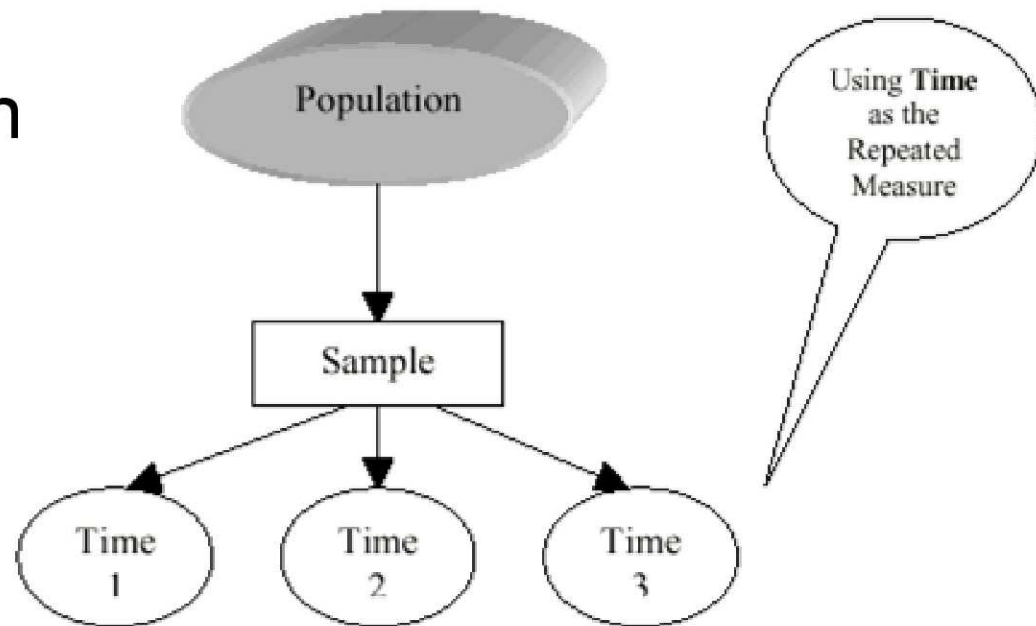
# Repeated Measures ANOVA

- Same people… no "individual differences" in the F-ratio

- More powerful statistics

➢ **The F-Ratio**

$$F = \frac{\text{variability between conditions/times}}{\text{variability within the sample}} = \frac{\text{error} + \text{condition/time differences}}{\text{error}}$$
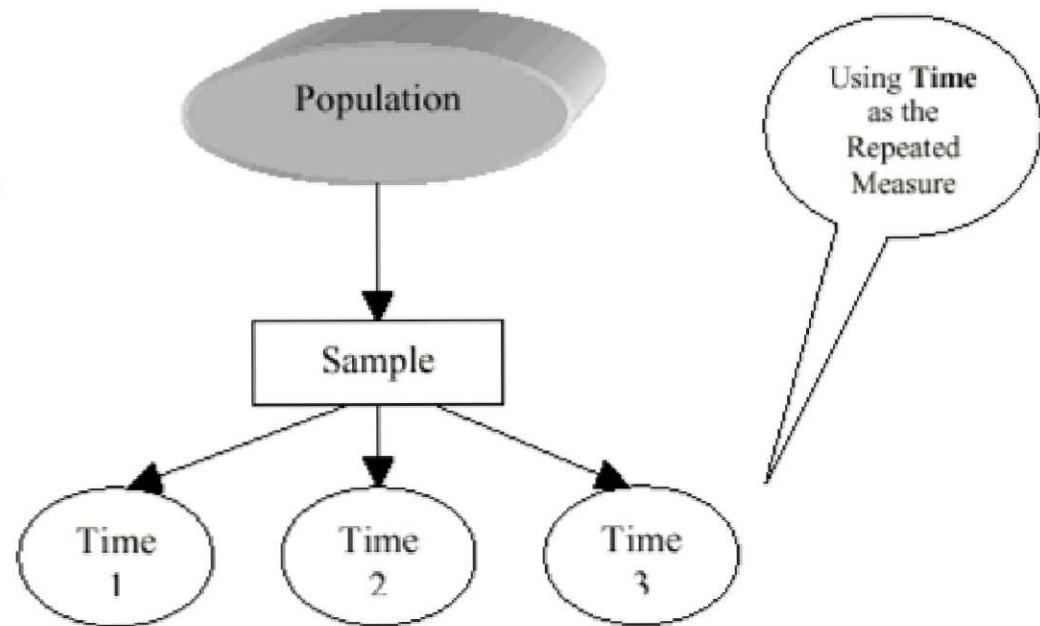
# Repeated Measures ANOVA

- Same people… no "individual differences" in the F-ratio

- More powerful statistics

➢ **The F-Ratio**



$$F = \frac{\text{variability between conditions/times}}{\text{variability within the sample}} = \frac{\text{error + condition/time differences}}{\text{error}}$$
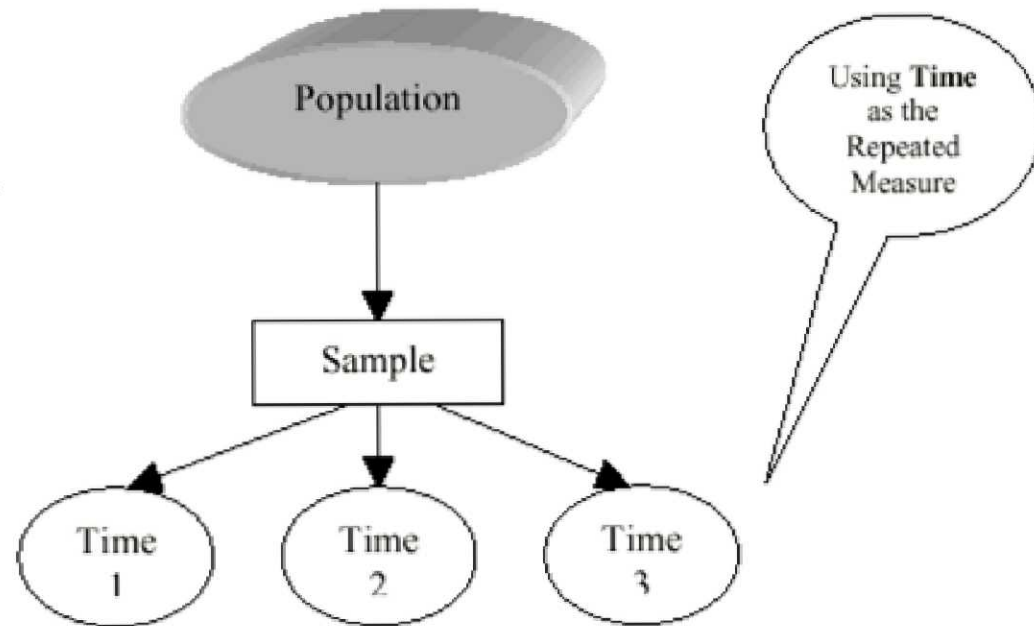
# RM-ANOVA Summary Tables

- Same Concept as IM Table, but now
  - Instead of "Between Groups" effects, we have "Between Treatments" effects
  - And also "Within Treatments"
    - Consist of subject differences (b/t subjects)
    - And error
- One Group measured several times, thus we partition "within group" variability into that which is due to individual differences, and error.

# This is where it comes from (Repeated Measures Designs)

$$SS_{total} = \text{Same as IM Anova}$$

$$SS_{between} = \text{Same as IM Anova}$$

$$SS_{within} = \text{Same as IM Anova}$$

$$SS_{b/t\ subjects} = \sum \frac{\left(\text{each person's total across treatments}\right)^2}{k} - \frac{\left(\sum X\right)^2}{N}$$

$$SS_{error} = SS_{within} - SS_{b/t\ subjects}$$

$$df_{total} = N - 1$$

$$df_{between} = k - 1$$

$$df_{within} = N - k$$

$$df_{b/t\ subjects} = n - 1$$

$$df_{error} = (N - k) - (n - 1)$$

# This is where it comes from (Repeated Measures Designs)

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$MS_{error} = \frac{SS_{error}}{df_{error}}$$

$$F = \frac{MS_{between}}{MS_{error}}$$

F-tables provide a $\underline{p}$ value for a given F-statistic, using $df_{between}$ (numerator) and $df_{error}$ (denominator).

# Example 3

- Compare student satisfaction ratings over time (four time points)
  - Freshman
  - Sophomore
  - Junior
  - Senior
- Same students…different *times*

# Using Covariates in ANOVA

- Sometimes the apparent effects of one factor, can be "explained" by the effects of some other factor—called a covariate

- There is a significant relationship between panty hose wearing behavior and a form of cancer…any ideas what kind of cancer?

# Using Covariates in ANOVA
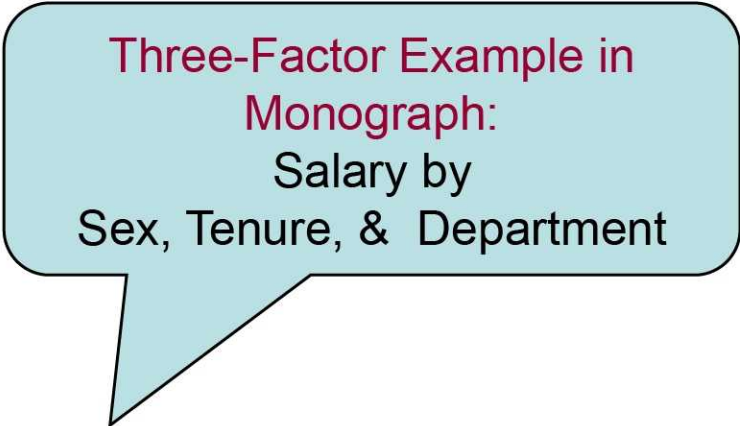
- Let's look at the Faculty Salaries from a fictitious Biology Department:
  - Comparing Salaries by Gender and Tenure Status
    - A simple 2-factor ANOVA
  - Then considering AGE as a possible covariate in our model
    - In-other-words, covary-out the effect of age and see if the salary differences remain…

# Example 4

- Describe effects of sex and tenure on faculty salaries for Biology faculty
  - Sex (male, female)
  - Tenure (tenured, untenured)
  - Use Age as a covariate
- 2 (sex) x 2 (tenure) Model
  - With a covariate

# Related Advanced Topics

- ANOVA can handle multiple factors
  - More than most humans can understand!
- Even just three factors can produce SEVEN effects!
  - 3-way interaction
  - A*B 2-way interaction
  - A*C 2-way interaction
  - B*C 2-way interaction
  - A main effect
  - B main effect
  - C main effect
    - Care to interpret that?
- Four factors = 15 effects
- Five factors = 31 effects!!
- $2^n - 1$

Three-Factor Example in Monograph:
Salary by
Sex, Tenure, &  Department

# Related Advanced Topics

- Mixed-Model ANOVAs
- It is possible to consider both types of factors in a single model
  - Student satisfaction over time and by major
  - Student performance over time and by teaching modality

Mixed-Model Quick Demo:
Intellectual Growth
*by* College *over* time

# Related Advanced Topics

- ## Simple Non-Parametrics
  - Chi-Square (best for 2x2's; purely categorical outcomes)
  - Mann-Whitney U (good for 2 groups; rank or ordinal data)
  - Kruskal-Wallis H (if >2 groups, like Oneway ANOVA)

- ## Loglinear ANOVA
  - Useful for more complicated multi-factor designs
    - Recommend additional training & reading
      - Allan Agresti's Text is a good one

# New Advances

- Hierarchical Modeling (a.k.a. HLM, MLM)
  - Maximum Likelihood based, for using random (i.e. not fixed) factors
  - Assessing impact of "layers" of grouping factors
    - Observations within person
    - Person within group
    - Group within larger group
  - Much better at accommodating for missing data
  - Many variations for different distributions
  - Unfortunately, SPSS is not well-suited

# Day 2: The saga continues…

# Shift Gears: Predicting "Y" from "X" (or several "Xs"

- ## Z-tests, T-tests, ANOVA
  - All for *comparing* groups, or observations over time

- ## Now we'll shift gears and talk about Regression Analysis

# Data Relationships with Two Variables

- Data Relationships with Two Variables
- Getting a Visual on relationships
  - Quantify relationships (Pearson's r)
    - Strength
    - Significance
  - Background leading into Simple Regression

# Are Two Variables Related?

- SES and GPA?
- High School GPA and First Term College GPA?
- SAT Scores and GPA?
- ACT Scores and GPA?
- SAT and ACT Scores?

- We're looking at PAIRS of variables
- And we're assessing relationships (non-causal)

# NOTE THAT

- We are *not* comparing means from 2 different groups here!

- We *are* trying to see if there is a relationship among two different variables
  - Usually continuous in scale

# We could look at a table of numbers?

**Any relationship between X and Y here?**

| G | H | I | J | K | L |
|---|---|---|---|---|---|
| Y | Y | X | Y | X | Y |
|  | 10.032 | 2 | 1.8963 | 2 | 2.4707 |
|  | 1.5777 | 3 | 3.579075 | 0 | 0.577854 |
|  | 1.755798 | 1 | 0.869 | 6 | 29.882 |
|  | 22.341 | 3 | 7.0433 | 1 | 1.678363 |
|  | 3.3259 | 2 | 4.1348 | 3 | 3.9921 |
|  | 7.4065 | 1 | 1.276016 | 4 | 8.705296 |
|  | 2.515718 | 3 | 5.310606 | 3 | 5.6881 |
|  | 2.520029 | 6 | 42.308 | 2 | 2.9196 |
|  | 5.0723 | 2 | 4.1634 | 2 | 2.2384 |
|  | 1.106803 | 6 | 53.8 | 1 | 4.2645 |
|  | 4.319 | 0 | 0.309287 | 6 | 23.53278 |
|  | 1.9365 | 3 | 5.996915 | 0 | 0.412273 |
|  | 10.885 | 3 | 5.088329 | 0 | 0.246107 |
|  | 5.8534 | 6 | 25.69648 | 0 | 0.710681 |
| 3 | 4.294138 | 2 | 4.8349 | 4 | 13.59942 | 4 | 7.300787 |
| 2 |  | 5 | 18.4741 | 4 | 10.987 |
| 2 | 1.6105 | 3 | 3.681995 | 3 | 3.957562 |
| 6 | 1.7175 | 5 | 31.549 | 2 | 2.8526 |
| 2 |  | 2 | 5.5092 | 2 | 1.9853 |
| 5 | 21.36 | 4 | 17.05 | 6 | 27.397 |
| 4 | 10.42 | 0 | 0.434728 | 4 | 13.58486 |
| 0 | 0.729 | 2 | 2.1139 | 1 | 1.1956 |
| 5 | 18.693 |  | 3.3516 | 3 | 5.3035 |
| 0 | 0.281206 |  | 1.9676 | 4 | 11.87442 |
| 1 | 2.2576 | 3 | 7.764669 | 1 | 1.446295 |
| 2 | 1.6274 | 2 | 2.6187 | 3 | 4.9911 |
| 4 | 11.18872 | 3 | 5.230134 | 3 | 4.310958 |
| 4 | 10.56362 | 2 | 1.4808 | 3 | 5.160213 |
| 6 | 51.927 | 5 | 30.4 | 4 | 16.924 |
| 2 | 2.7903 | 6 | 58.879 | 4 | 4.4573 |

| | | | | | |
|---|---|---|---|---|---|
| 1 | 1.805142 | | | 2 | |
| 1 | 2.1398 | 1 | 1.6105 | 2 | |
| 5 | 31.695 | 2 | 1.7175 | 6 | |
| 4 | 8.502829 | 4 | 10.072 | 2 | |
| 4 | 15.6 | 3 | 5.245146 | 5 | 21.36 |
| 1 | 1.5995 | 1 | 0.878264 | 4 | 10.42 |
| 1 | 1.985864 | 3 | 7.138 | 0 | 0.729 |
| 5 | 25.596 | 4 | 10.662 | 5 | 18.693 |
| 5 | 21.10395 | 5 | 20.30073 | 0 | 0.281206 |
| 4 | 14.58116 | 1 | 0.8408 | 1 | 2.2576 |
| 6 | 66.712 | 5 | 15.97328 | 2 | 1.6274 |
| 1 | 1.898284 | 0 | 0.78812 | 4 | 11.18872 |
| 3 | 5.746124 | 6 | 20.283 | 4 | 10.56362 |
| 3 | 4.671579 | 2 | 2.7033 | 6 | 51.927 |
| | | 5 | 18.27308 | 2 | 2.7903 |

# Scatterplots help us visualize…

- Two continuous variables, X and Y
- Scatterplot shows association of X and Y
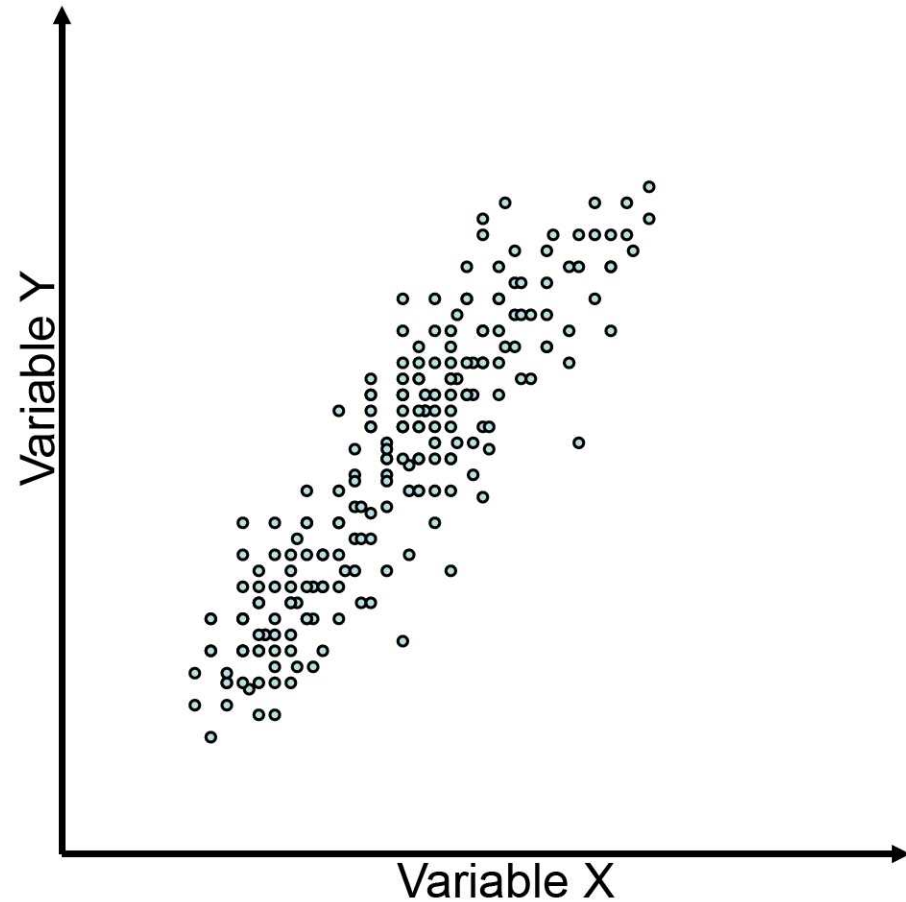
# Recall basic Correlational Analysis

- Two continuous variables, X and Y
- Scatterplot shows association of X and Y
- Dots represent observations (usually people)
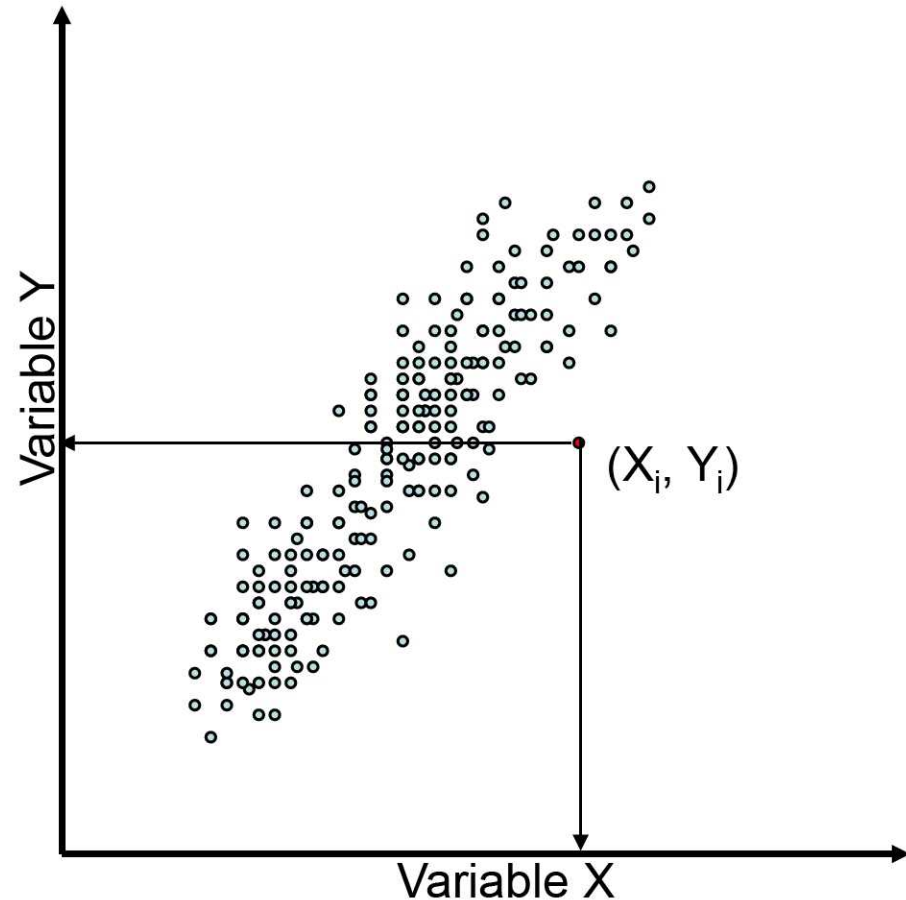
# *Quantifying* the Relationship: Recall Linear Correlation

- Measures the direction and strength of the linear relationship b/t two variables.

$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

  - Sign tells you what?
  - R-value tells you what?
  - P-value tells you what?
  - *Causality*?

# Scatter Plots… Footballs, Basketballs, Directionality
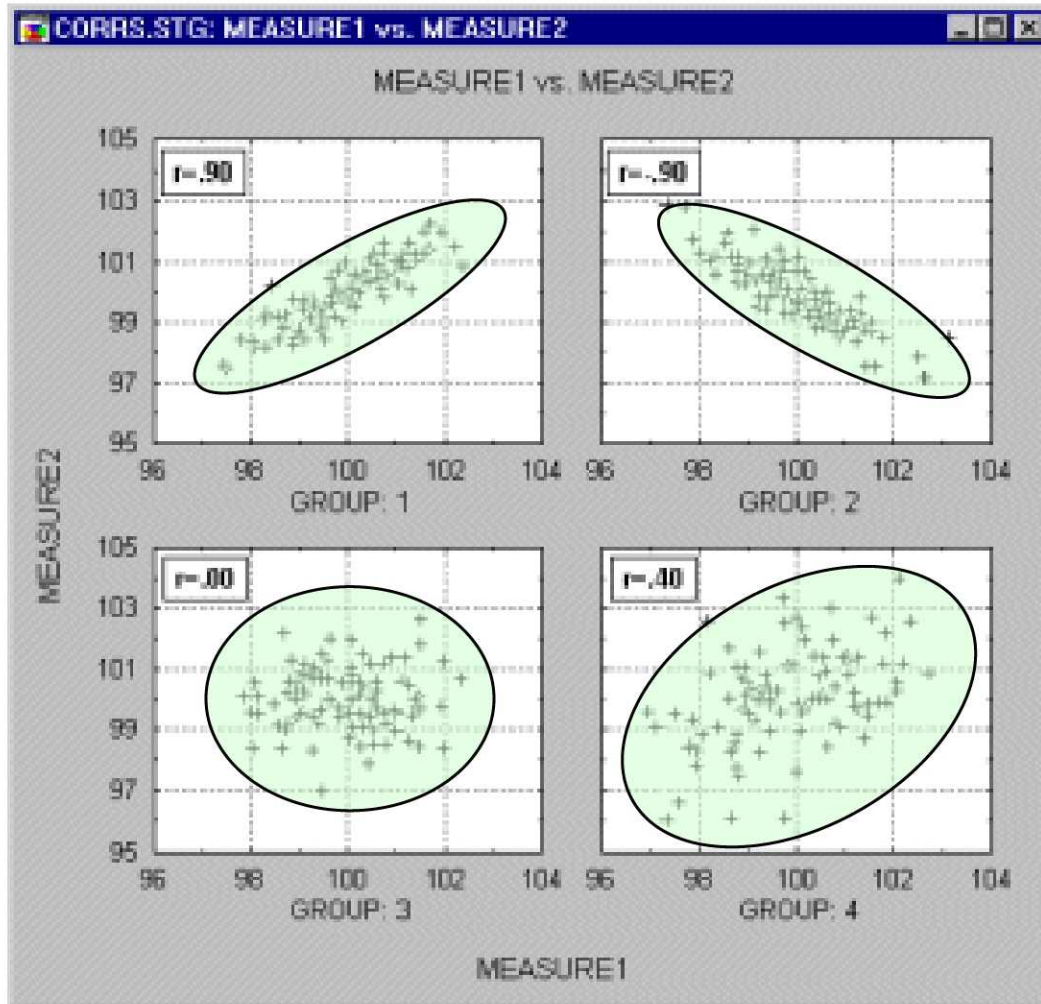
# Correlation Analysis Quantifies our Understanding
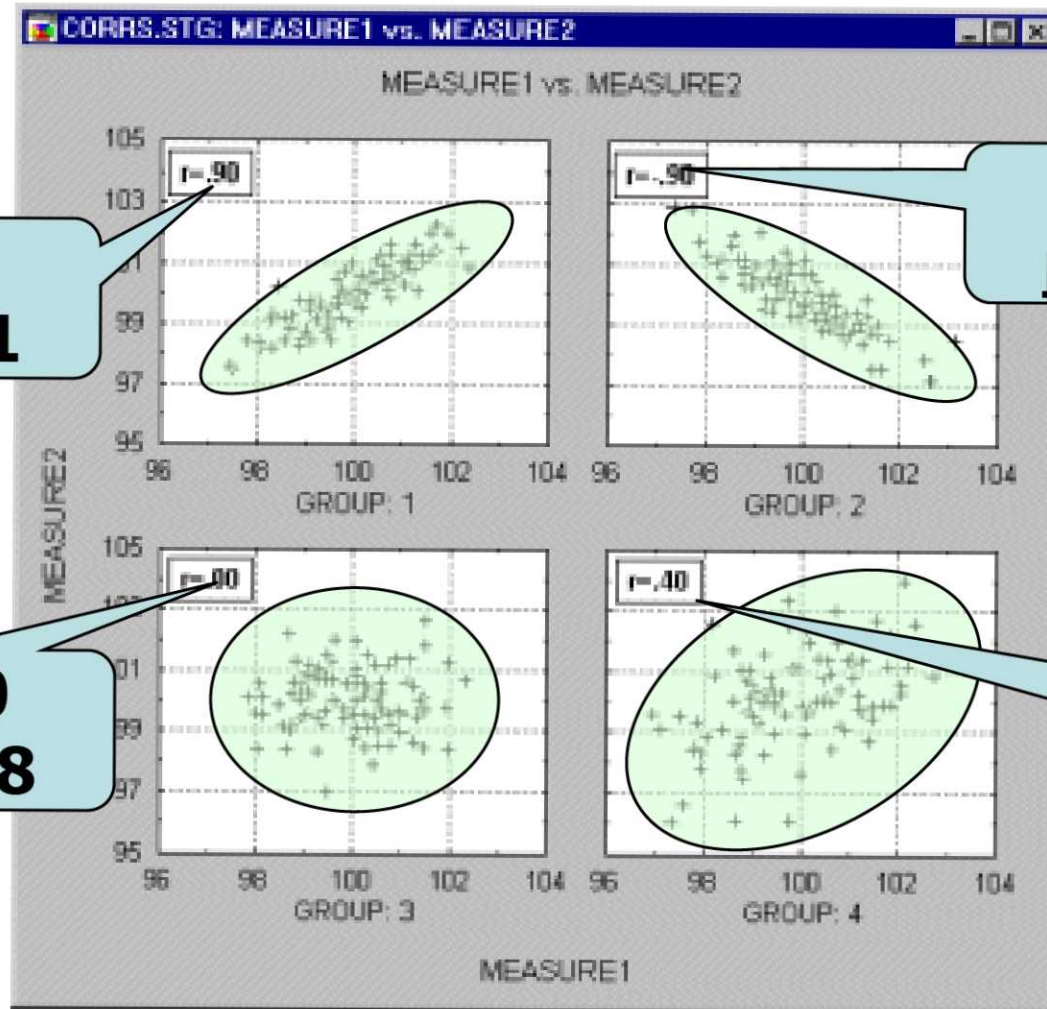
# Ordinary Least Squares Regression

- Using *Simple Regression* to Describe a Linear Relationship

- Regression as a Descriptive Tool

- Regression as a Forecasting/Prediction Tool

- Making Inferences from Simple Regression

# Simple Linear Regression

- Statistic used to describe linear relationships among variables:

$$\hat{y} = b_0 + b_1 x$$

Just like HS Algebra:
Y=m(X)+b

$\hat{y}$ = the dependant (outcome) variable

$x$ = the independent (predictor, explanatory) varible

$b_0$ = the y - intercept of graph of all (x, y) pairs

$b_1$ = the slope of the line

# Consider a simple example of a Height/Age Scatterplot...

- Each green dot represents a person in the sample

**This person's Height**

**Y=Height (cm)**

$(x_{47}, y_{47})$

**This person's Age**

**X=Age (mo.)**
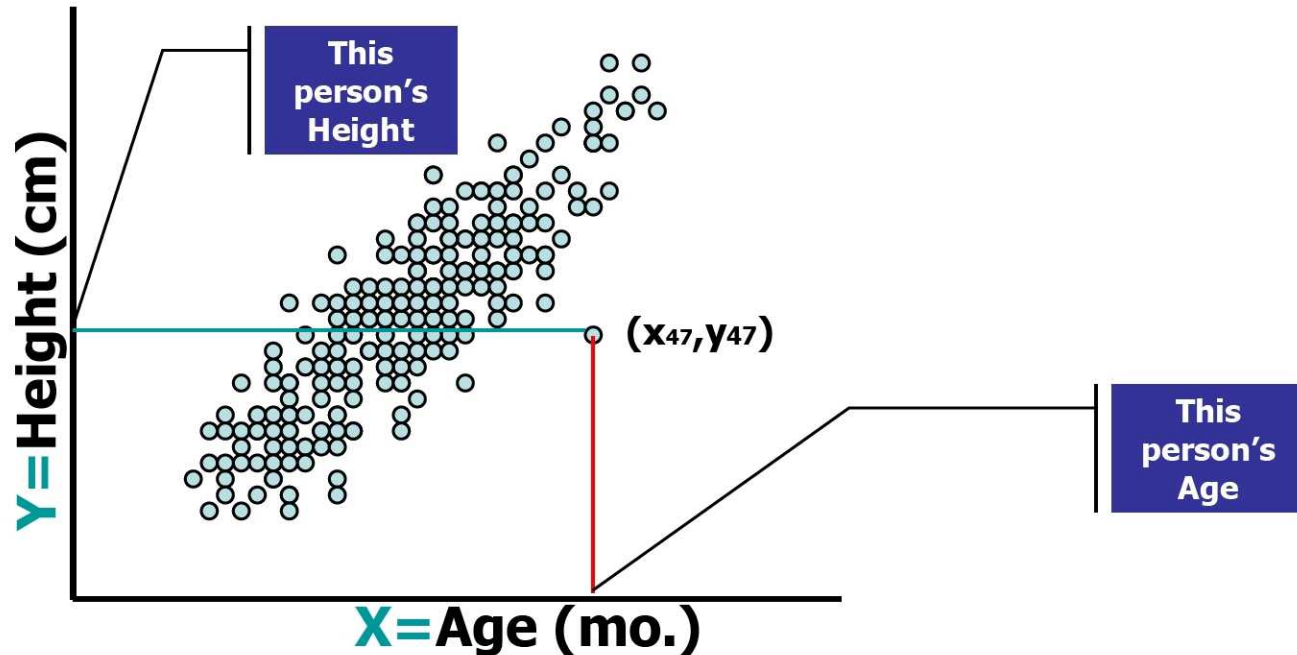
# The OLS Regression Equation

- A "best-fitting" line is calculated from the sample data

- Can be used to predict Y, given X

$$\hat{y}_{47} = a + 24x_{47}$$

$$\text{Error} = y - \hat{y}$$
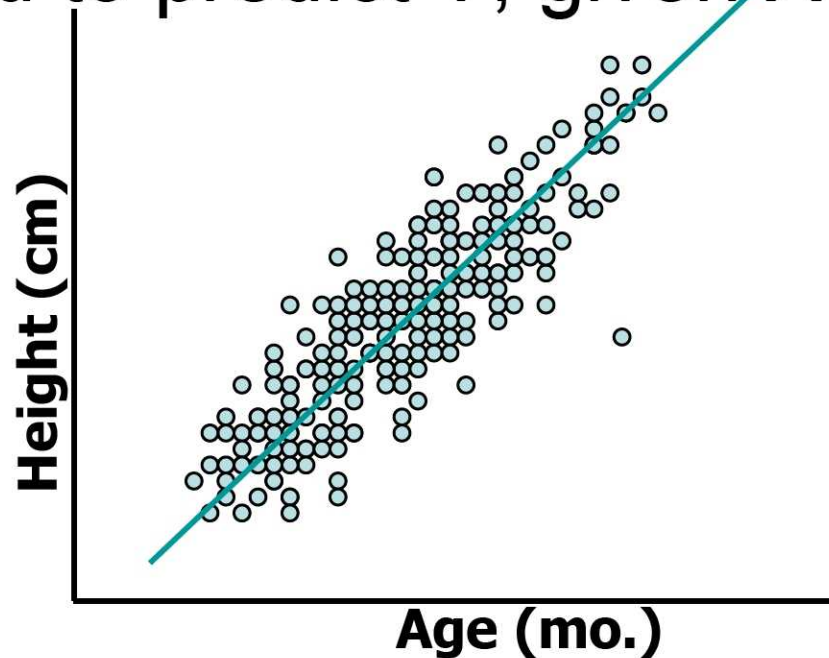


Height (cm) vs. Age (mo.)

# OLS Regression Lines Aren't Perfect!

- Some Predictions are too high:

$$\hat{y}_{47} = a + 24x_{47}$$

$$\mathrm{Error} = y - \hat{y}$$



Height (cm)

Predicted height (given age)

Negativ
error

Actual height (and age)

Age (mo.)
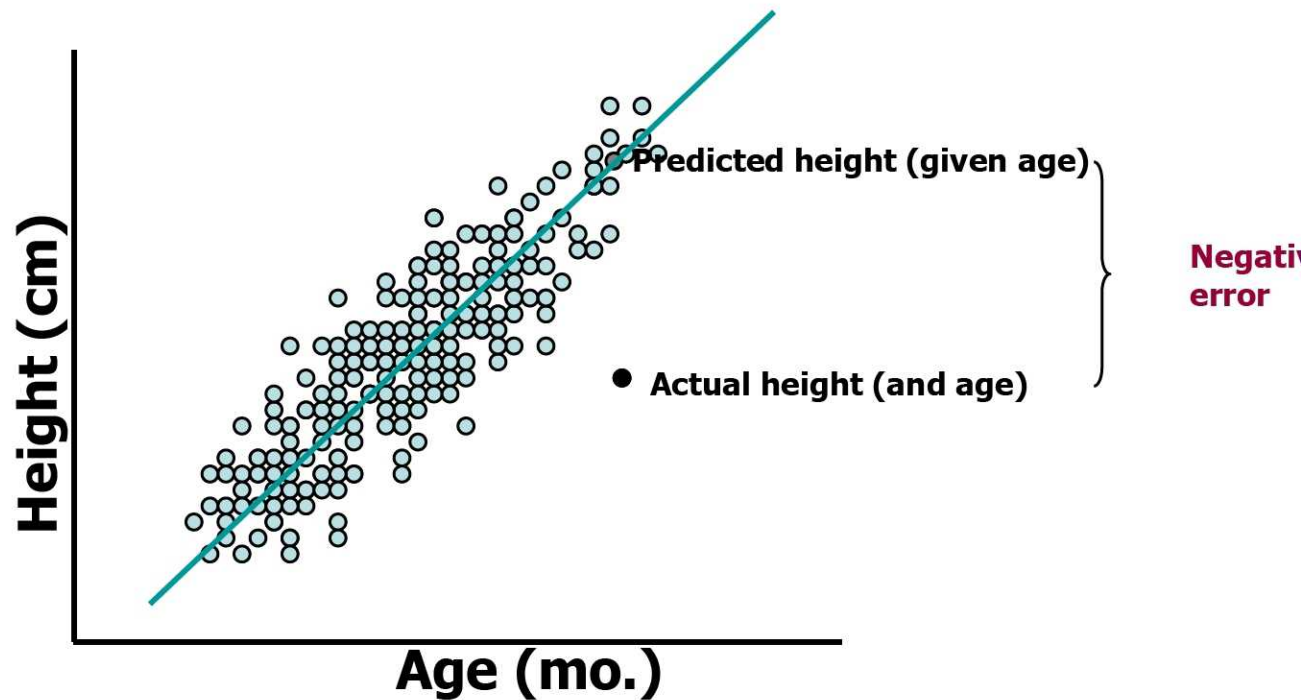
# OLS Regression Lines Aren't Perfect!

- Some Predictions are too low:

$$\hat{y}_{47} = a + 24x_{47}$$

$$\text{Error} = y - \hat{y}$$
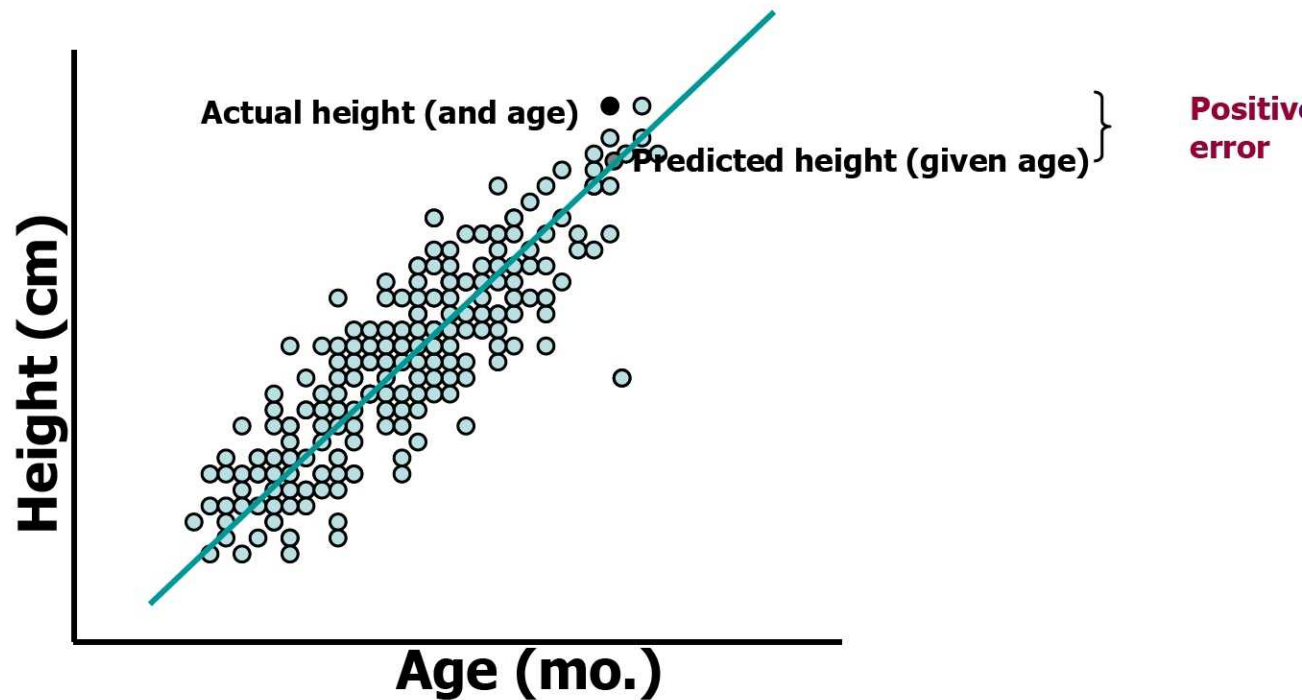


Actual height (and age)

Predicted height (given age)

Height (cm)

Age (mo.)

Positive error

# OLS Regression Lines Aren't Perfect!

- Some Predictions are too high:

$$\hat{y}_{47} = a + 24x_{47}$$

$$\text{Error} = y - \hat{y}$$

Actual height (and age)

Predicted height (given age)

Negative error

Height (cm)

Age (mo.)
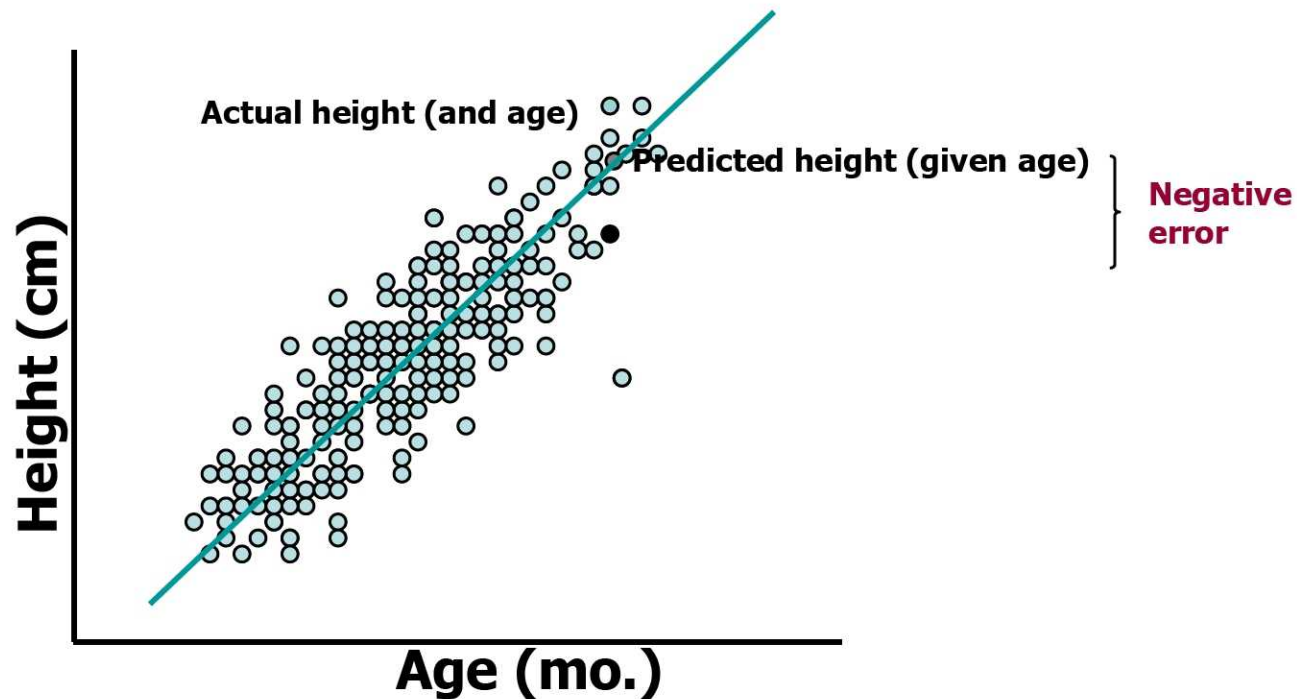
# Deriving the "Best-Fitting" Line

- Getting the best line is KEY!

- One Approach would be a line that Minimizes the Sum of the Errors:

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)$$
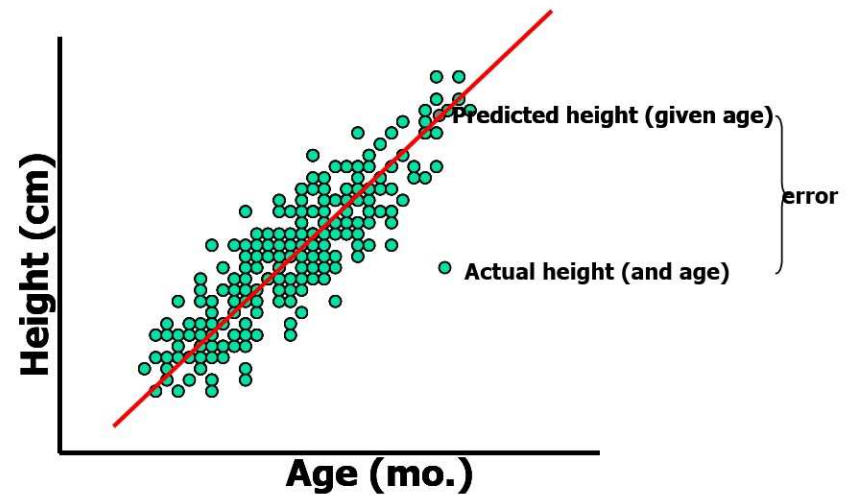
# Why not minimize $\sum_{i=1}^{n}(y_i - \hat{y}_i)$ ?

- Because positive errors cancel out negative errors when minimizing sum of errors…

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$$

- ..We need a way to consider the sign of the error into the minimization function.

# …And Also…

- **_All_** lines minimizing Sum of Errors pass through the point $(\bar{x}, \bar{y})$

$$\sum_{i=1}^{n} \left( y_i - \hat{y}_i \right) = 0$$

Height (cm)

$(\bar{x}, \bar{y})$

Age (mo.)
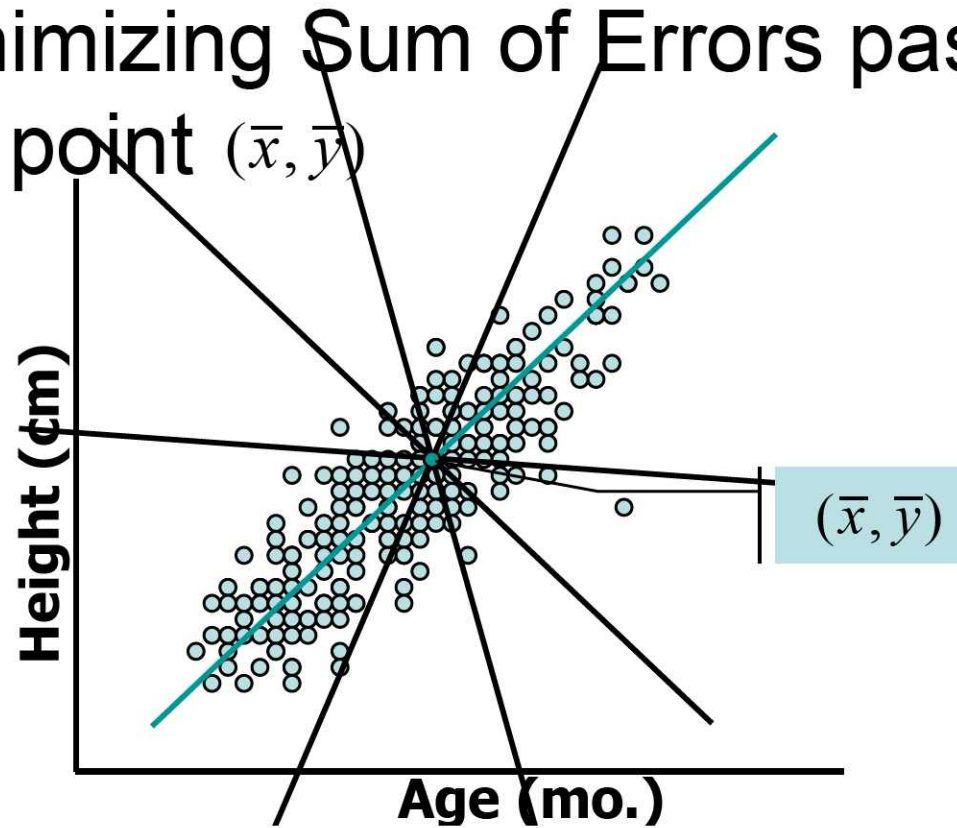
- …Infinite number of lines meet this criteria, but only 1 is really "best"

# Another Approach?

- Minimize the Absolute Value of Errors?

$$\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|$$

Least Absolute Value (LAV)

■ No unique LAV line, plus very complicated calculations

# Third Approach?

- Minimize the Squared errors?

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Least Squares Regression (LS)

# Is LS Minimization Possible?

- Yes!

For a Linear Equation $\hat{y}_i = b_0 + b_1 x$:    ...simpler formula

$$b_1 (\text{slope}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2}$$

$$b_0 (\text{intercept}) = \bar{y} - b_i \bar{x}$$

# …a NOTE about formulae

- Hopefully you'll be able to recognize what they're doing.

- But don't worry about memorizing them!

- SPSS (or whatever your choice of software) will calculate the components of the OLS Regression Line for you anyway!

# Minimizing the SQUARED Errors

- Takes care of the problem of positive errors "canceling out" negative ones

- Computationally simple enough for hand or computer calculations

- Creates a unique "line of best fit"

# Let's do one by hand…

| i | $x_i$ | $y_i$ |
|---|-------|-------|
| 1 | 1 | 3 |
| 2 | 2 | 2 |
| 3 | 3 | 8 |
| 4 | 4 | 8 |
| 5 | 5 | 11 |
| 6 | 6 | 13 |

# Let's do one by hand…

| i | $x_i$ | $y_i$ |
|---|-------|-------|
| 1 | 1 | 3 |
| 2 | 2 | 2 |
| 3 | 3 | 8 |
| 4 | 4 | 8 |
| 5 | 5 | 11 |
| 6 | 6 | 13 |

$$b_i = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2}$$

**Start with formula for $b_i$**

# Let's do one by hand…

| i | $x_i$ | $y_i$ |
|---|-------|-------|
| 1 | 1 | 3 |
| 2 | 2 | 2 |
| 3 | 3 | 8 |
| 4 | 4 | 8 |
| 5 | 5 | 11 |
| 6 | 6 | 13 |
| sum | 21 | 45 |

$$b_i = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2}$$

The $b_i$ formula requires E(x) and E(y)

# Let's do one by hand…

| i | $x_i$ | $y_i$ | $x_iy_i$ | $x^2$ |
|---|---|---|---|---|
| 1 | 1 | 3 | 3 | 1 |
| 2 | 2 | 2 | 4 | 4 |
| 3 | 3 | 8 | 24 | 9 |
| 4 | 4 | 8 | 32 | 16 |
| 5 | 5 | 11 | 55 | 25 |
| 6 | 6 | 13 | 78 | 36 |
| sum | 21 | 45 | 196 | 91 |

$$b_i = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

**The $b_i$ formula requires these two columns also…**

# Let's do one by hand…

| i | $x_i$ | $y_i$ | $x_iy_i$ | $x^2$ |
|---|-------|-------|----------|-------|
| 1 | 1 | 3 | 3 | 1 |
| 2 | 2 | 2 | 4 | 4 |
| 3 | 3 | 8 | 24 | 9 |
| 4 | 4 | 8 | 32 | 16 |
| 5 | 5 | 11 | 55 | 25 |
| 6 | 6 | 13 | 78 | 36 |
| sum | 21 | 45 | 196 | |

$$b_i = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_0 = \bar{y} - b_i\bar{x}$$

$$\bar{x} = \frac{21}{6} = 3.5$$

$$\bar{y} = \frac{45}{6} = 7.5$$

**…and $b_0$ requires means of x and y**

# Using the formula…

For a Linear Equation $\hat{y}_i = b_0 + b_1 x$ :

$$b_1 = \frac{\sum\limits_{i=1}^{n} x_i y_i - \frac{1}{n} \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{\sum\limits_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum\limits_{i=1}^{n} x_i \right)^2} = \frac{196 - \frac{1}{6}(21)(45)}{91 - \frac{1}{6}(21)^2} = \frac{38.5}{17.5} = 2.2$$

$$b_0 = \overline{y} - b_i \overline{x} = 7.5 - 2.2(3.5) = -0.2$$

# Resulting in one, Unique LS Regression Line:

$$\hat{y} = -0.2 + 2.2x$$

- What does this line do for us?
  - Given x=15…. Predict y. — **32.8**
- What can you say about the errors that you will make in your prediction?

**We've minimized our error in prediction**

# Using LS Regression to Describe Relationships

- Can be used to establish the *weight* assigned to various *factors* that may (or may not) predict some *outcome*
  - Ex. Assess the value of knowing an incoming students' SAT Verbal on their first-term GPA, based on a sample of historical student application/enrollment data

$$GPA_{firstterm} = 1.2 + .312(SAT_{verbal})$$

# Using LS Regression to Describe Relationships

- Can be used to establish the *weight* assigned to various *factors* that may (or may not) predict some *outcome*
  - Ex. Assess the value of knowing an incoming students' SAT Verbal on their first-term GPA, based on a sample of historical student application/enrollment data

$$GPA_{firstterm} = 1.2 + .312(SAT_{verbal})$$

# Let's run a Simple Regression using SPSS

**Description of Dataset NH SAT A630**

The dataset contains annual information on the set of New Hampshire students who have taken the Scholastic Aptitude Test (SAT) each year from 1976 through 1998.  The variables in the dataset are defined as follows:

- YEAR
- TOTAL            = Total number of SAT takers in NH
- SATV            = Average SAT-Verbal score
- SATM            = Average SAT-Math score
- PCTDOCT     = Percentage of SAT takers planning on pursuing a doctorate degree
- UNH             = Number of SAT takers sending test scores to the U New Hampshire
- CPI              = Consumer Price Index (1983 = 100)
- UNHPCT      = % of total SAT takers sending test scores to U New Hampshire
- LUNHPCT     = Natural logarithm of UNHPCT (=LN(UNHPCT))
- Resident       = Resident tuition rate at U New Hampshire (in $1000s)
- Nonres        = Nonresident tuition rate at U New Hampshire (in $1000s)
- Private         = Average private tuition rate in New England (in $1000s)
- Income        = Median family income (in $1000s)
- Lres            = Natural logarithm of Resident (=LN(Resident))
- Lprivate       = Natural logarithm of Private (=LN(Private))
- Lincome      = Natural logarithm of Income (=LN(Income))
- Lnonres      = Natural logarithm of Nonres (=LN(Nonres))

# Let's run a Simple Regression using SPSS

- Let's predict the number of SAT scores we might expect to be sent to UNH, so our Enrollment Management Office can do some strategic planning.
  - Y = UNH
- Let's use only 1 predictor—the total number of SAT test-takers
  - X1 = TOTAL
    - Ok.. This is a little boring, but bare with me!

# SPSS Demo

- Simple Regression—used to describe the relationship between total number of SAT test takers in NH, and the number of SAT scores sent to NH Admissions.

# Moving Beyond Describing Relationships…

- Can we use Regression to "go beyond" what sample data suggest?

- Can we make *inferences* about how two variables *in the population* might be related based on observations of *sample* data?

# Why would we want to do this?

- Why bother when most IR offices have access to your local "population" of students (ex. All current incoming freshmen)?
  - i.e. you don't randomly sample your student data warehouse… you tap all data you have?!
- Because you want to make inferences about the whole incoming class based on "today's" data
- Because you want to make inferences about next semester, next year, etc…
- Because you want to do some strategic planning that could benefit by this kind of analysis

# Statistical Inferences require Assumptions about the Population

- Given Population with X and Y

$$\mu_{y|x} = \text{conditional mean of } y, \text{given } x,$$

$$\mu_{y|x} = \beta_0 + \beta_1 x \text{ where}$$

$$\beta_0 = \text{population } y\text{-intercept, and}$$

$$\beta_1 = \text{slope of the population regression line.}$$

# We assume Population X/Y Relationship is Linear
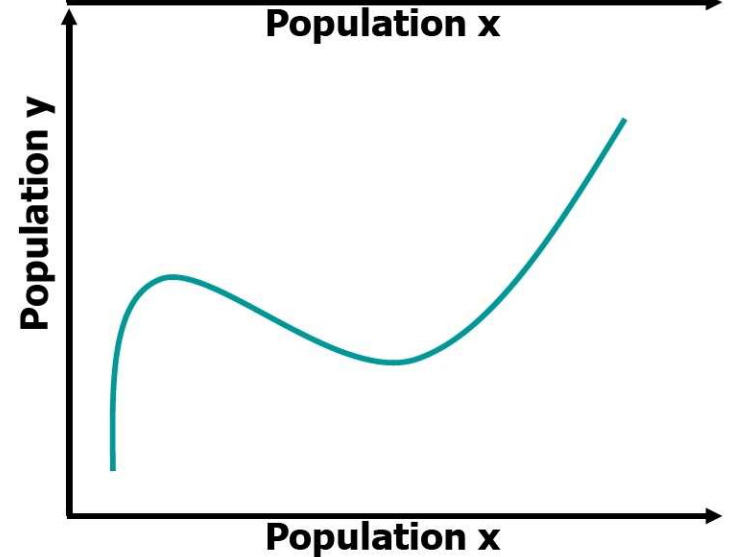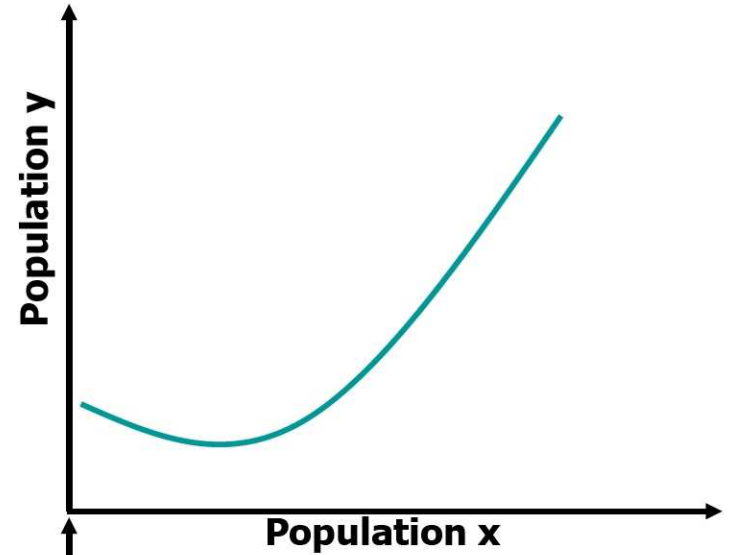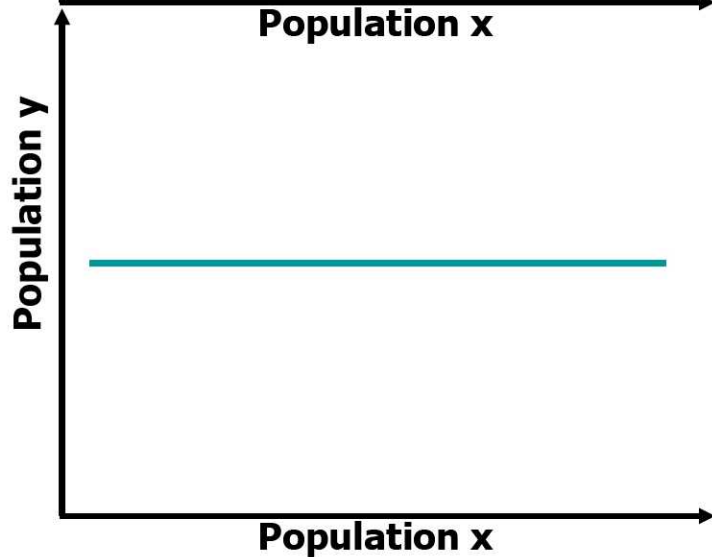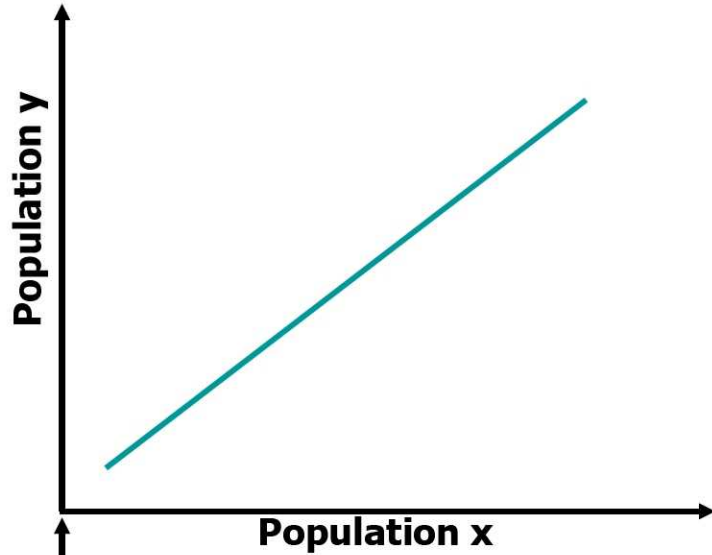
# We assume Population X/Y Relationship is Linear
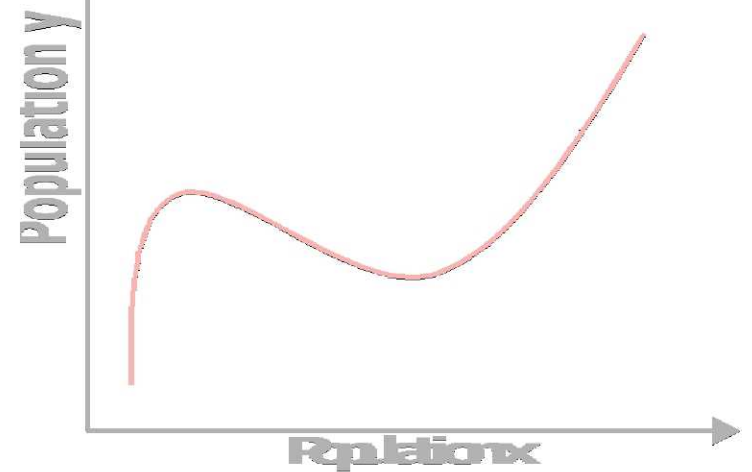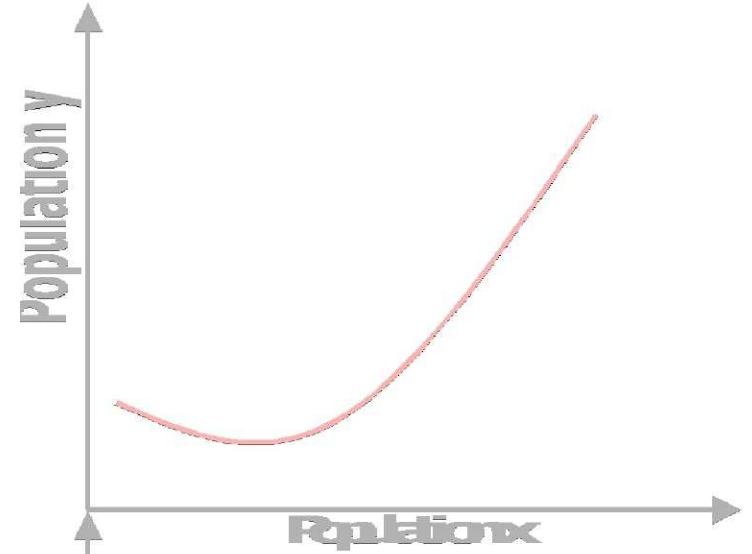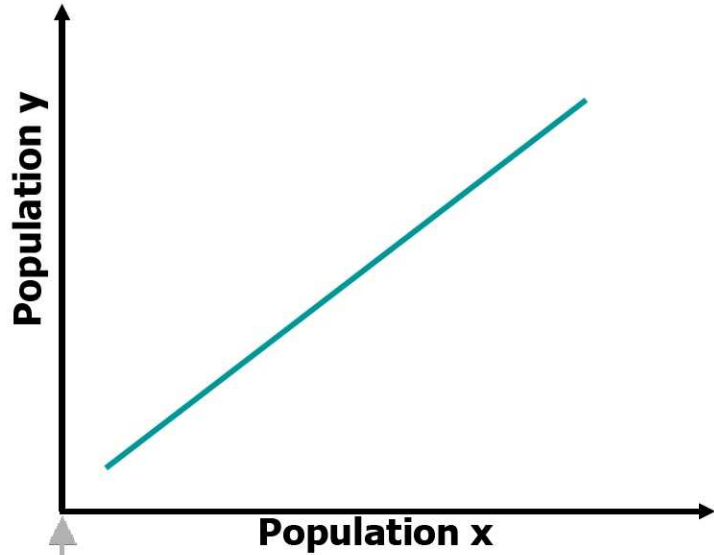
# RE: The Population Conditional Mean…
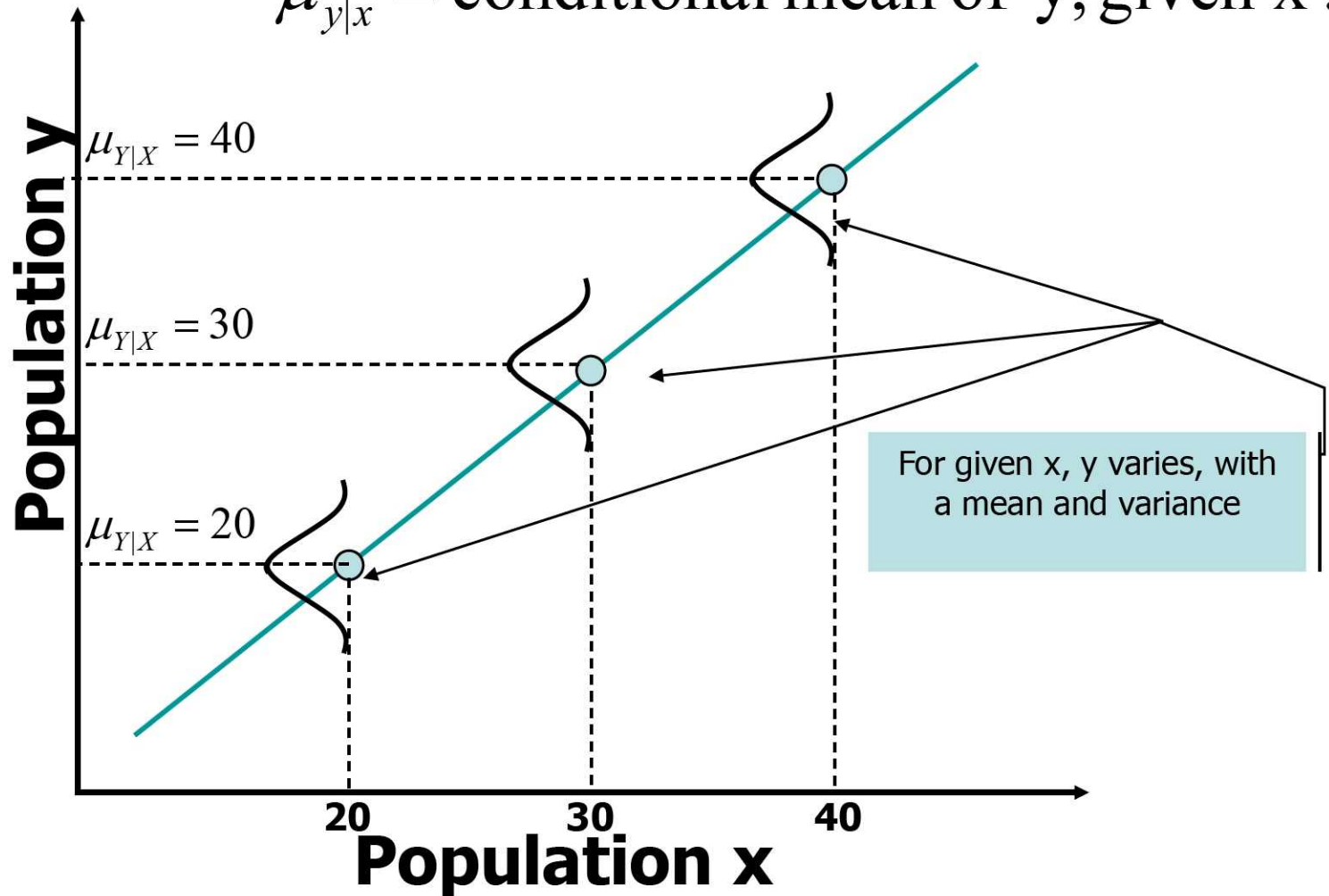
$$\mu_{y|x} = \text{conditional mean of } y, \text{given x}:$$



For given x, y varies, with a mean and variance

# RE: The Population Conditional Mean…

$$\mu_{y|x} = \text{conditional mean of } y, \text{given } x :$$



Population y

$\mu_{Y|X} = 40$

$\mu_{Y|X} = 30$

$\mu_{Y|X} = 20$

For given x, y varies, with a mean and variance **…leading to Prediction Error…or** *Disturbances*

20      30      40

**Population x**

# RE: The Population Conditional Mean…

$$\mu_{y|x} = \text{conditional mean of } y, \text{given } x :$$



**Population y**

$\mu_{Y|X} = 40$

$\mu_{Y|X} = 30$

$\mu_{Y|X} = 20$

For given x, y varies, with a mean and variance …leading to Prediction Error…or *Disturbances*…**adding a term to our regression equation:**

20    30    40
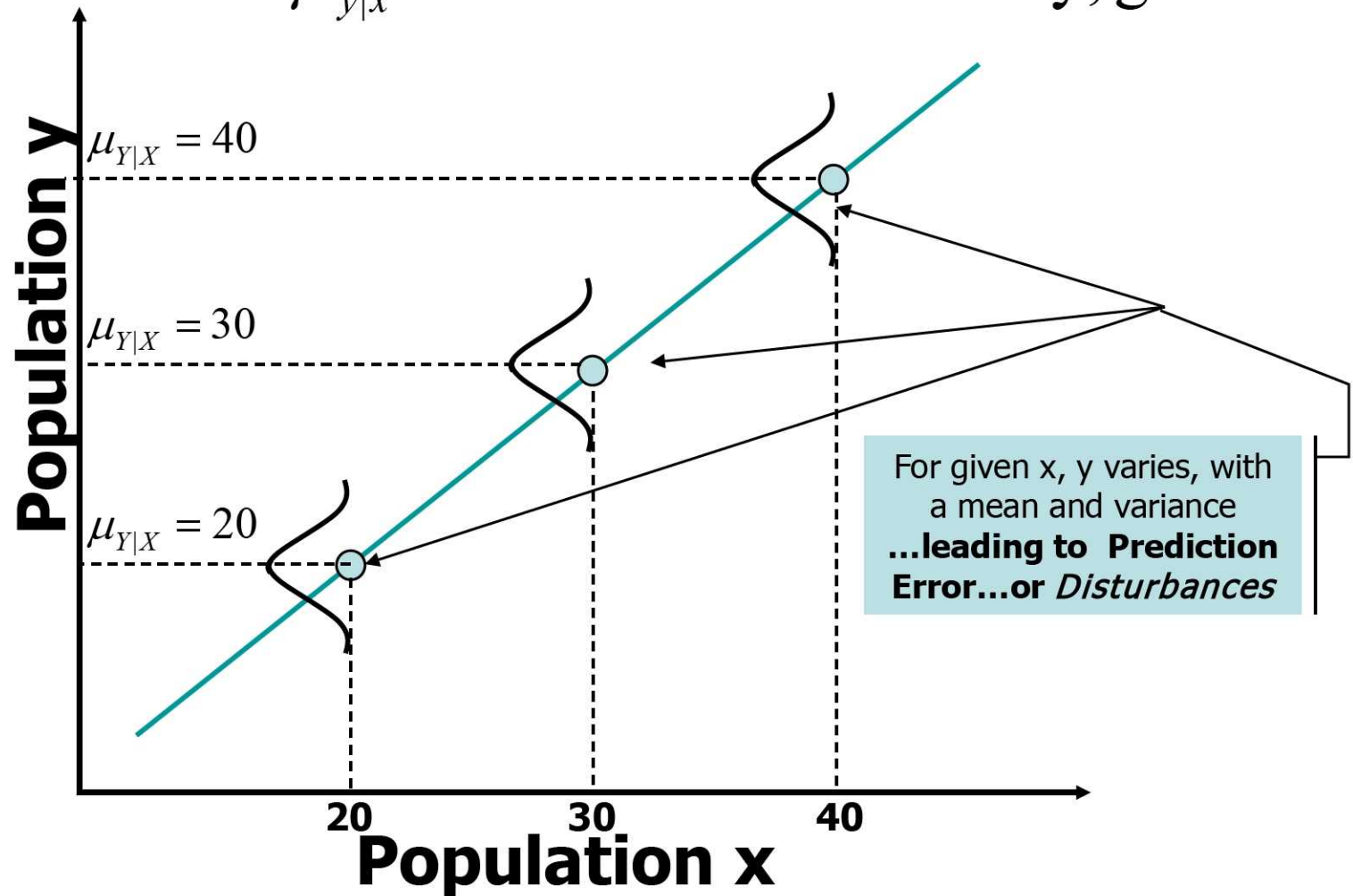
**Population x**
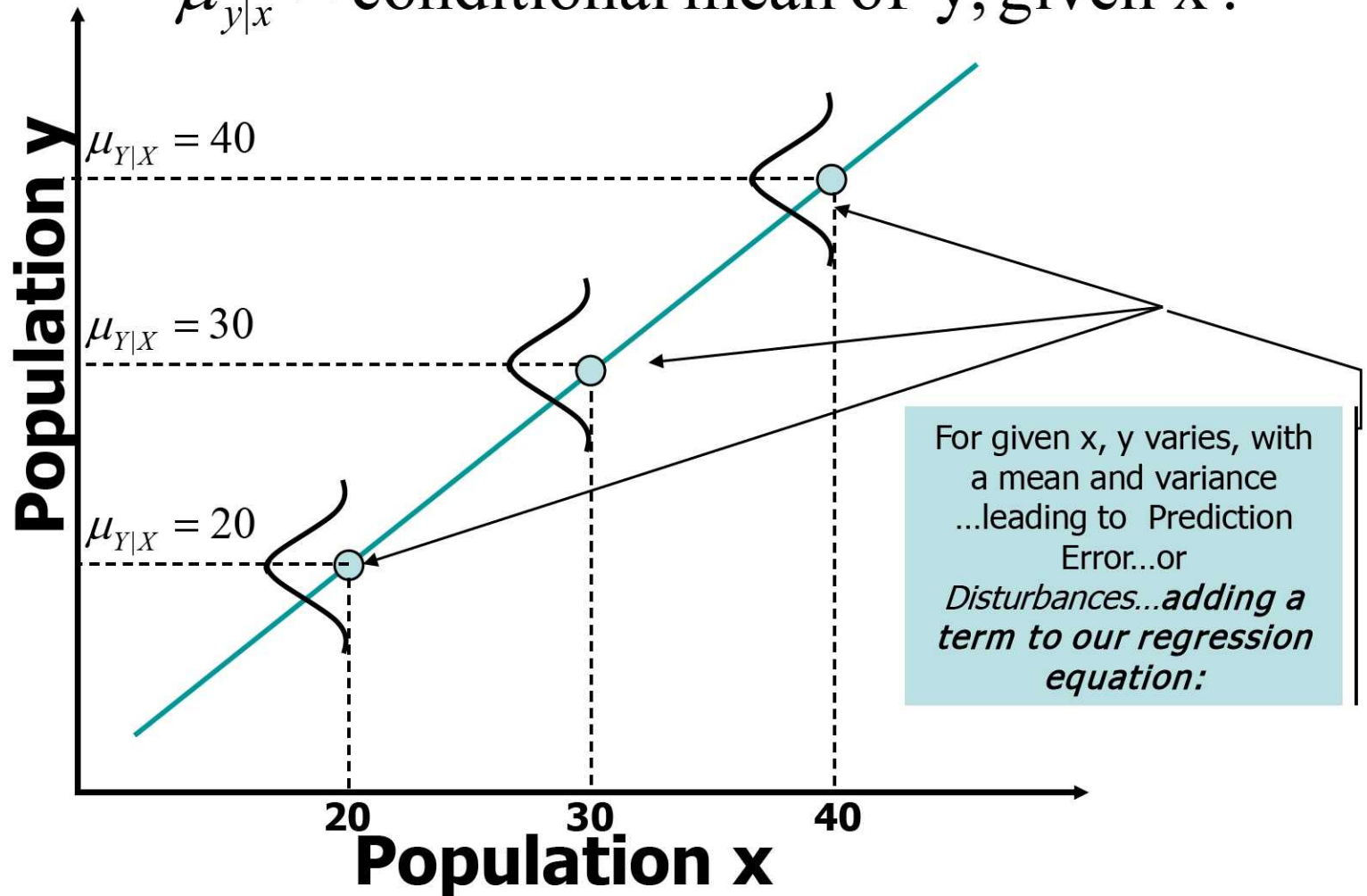
# Resultant (*Inferential*) Regression Equation

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- Where $e_i$ represents the difference between TRUE y and the conditional mean of all y|x
  - *Disturbances, or Error Variance*
  - If our equation were perfect, $e_i = 0$
    - What type of relationship would have to occur for this to happen?

# Leading up to some ASSUMPTIONS for Inferential Regression

1. Expected Value of disturbances = 0 $(E(e_i)=0)$

   a) i.e. The population regression is linear

2. Homogeneity of Variance for $e_i$

3. $e_i$ are normally distributed

4. $e_i$ are independent

# Inferences About $\beta_0$ and $\beta_1$

- $b_0$ and $b_1$ are point estimates of $\beta_0$ and $\beta_1$
  - I.e. they are calculated from samples, thus they are *statistics*
  - Thus they are random variables with probability distributions (sampling distributions)
- **Just as sample means are unbiased estimates as population means…**
  - $b_0$ and $b_1$ are unbiased estimates of $\beta_0$ and $\beta_1$
  - We assume that their sampling distributions are normally distributed
  - …and as n increases, $b_0$ and $b_1$ become closer and closer to $\beta_0$ and $\beta_1$
  - Of all possible estimators of $\beta_1$, $b_1$ has the desirable feature of having smaller sampling errors than any other unbiased estimator.

# In other words…

- $b_0$ and $b_1$ are *unbiased* Estimators
  - Mean of sampling distribution = Pop mean
- $b_0$ and $b_1$ are *Consistent* Estimators
  - As n increases, estimator approaches parameter
- $b_0$ and $b_1$ are a *Minimum Variance* Estimators
  - While there are other unbiased estimators "out there," $b_0$ and $b_1$ have the smallest variance

# GIVEN all of these assumptions

- We can make hypotheses about $\beta_0$ and $\beta_1$, and use $b_0$ and $b_1$ to test our hypotheses
  - like we can about means or other statistics.
- But we're missing one remaining component of the Regression Equation.
  - Variance around the regression line, or Error Variance
  - We know our predictions aren't perfect, but we need to quantify the "error in prediction"

# Estimating Variance around the Regression Line

- The estimate of Variance @ Regression:

$$\sigma_e^2 = s_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2} = MSE$$

- SSE = sum of squared errors, adjusted by n-2 df to incorporate sample size.
  - Df=(sample size – num of coefficients)
    - We're estimating two coefficients, $b_0$ and $b_1$
- MSE=Mean Square Error = any SS/df
- Square Root of $s_e^2$ is termed "*Stand Error of Regression*"

# Back to our SPSS Example…

- Note the SSR, SSE and MS terms in the ANOVA Summary Table
- Use ANOVA Table to Evaluate Significance of the Regression Equation
  - More important later, when we discuss *multiple* regression
- Interpret $R^2$

# Hypothesis testing in Regression

- We know that the LS estimates of $b_0$, & $b_1$ are unbiased estimators of the Population coefficients $ß_k$

- We can calculate Confidence Interval Estimates (& significance values) of $ß_k$

- And we test hypotheses that the population beta weights ($ß_k$) are significantly greater than some hypothesized value (usually 0)
  - Reject the notion that $ß_k$ = *hypothesized value* if observed t-score probability < alpha

# The Theory Behind It?

- Create Null and Alternative Hypotheses:

$$H_0 : \beta_k = \beta_k^* \text{ (usually 0)}$$

$$H_a : \beta_k \neq \beta_k^*$$

- Calculate t-score(s) for coefficients

$$t = \frac{b_k - \beta_k^*}{s_{b_k}} = \frac{b_k}{s_{b_k}} \text{ (when } \beta_k^* = 0)$$

- Compare t-value to critical t (given alpha)

$$\text{Reject } H_0 \text{ if } t > t_{\alpha/2} \text{ or } t < t_{\alpha/2}$$

$$\text{Accept } H_0 \text{ if } -t_{\alpha/2} \leq t \leq t_{\alpha/2}$$

# A Closer look at t-formula $H_0 : \beta_k = \beta_k^*$ (usually 0)

$$H_a : \beta_k \neq \beta_k^*$$

- ## Calculate t-score(s) for coefficients

$$t = \frac{b_k - \beta_k}{s_{b_k}} = \frac{b_k}{s_{b_k}} \text{ (when } \beta_k^* = 0)$$

**SD of the sampling distribution of $b_k$**

**When null is true, t-should be (large or small)?  Why??**
**When Alternative is true, t-should be (large or small)?**

# A Closer look at t-formula

- Calculate t-score(s) for coefficients

$$t = \frac{b_k - \beta_k^*}{s_{b_k}} = \frac{b_k}{s_{b_k}} \text{ (when } \beta_k^* = 0)$$

**NOTE that we are NOT comparing two sample means, even though we're using a t-test! We are comparing our observed $b_k$ to a constant that we choose ($\beta_k^*$, which us usually 0).**

# The "Typical" Situation

- Testing Hypothesis that the coefficients are significantly higher or lower than 0

$$t = \frac{b_k - \beta_k^*}{s_{b_k}} = \frac{b_k - 0}{s_{b_k}} = \frac{b_k}{s_{b_k}}$$

- If $b_1$ coefficient = 0, what would that mean in terms of the Regression Equation?

$$\hat{y} = b_0 + b_1 x$$

# If Null is Accepted

$$H_0 : \beta_k = \beta_k^* \text{ (usually 0)}$$
$$H_a : \beta_k \neq \beta_k^*$$

- ## Small t-statistic, p-value is *not < alpha.*
  - *"X does not appear to be linearly related to Y"*
    - I.e. x doesn't help you predict y

# If Null is Rejected

$$H_0 : \beta_k = \beta_k^* \ (\text{usually } 0)$$

$$H_a : \beta_k \neq \beta_k^*$$

- Large t-statistic, p-value is < *alpha.*
  - *"There is evidence that y and $x_k$ are linearly related, and that $x_k$ helps explain some of the variation in y (not accounted for by the other explanatory variables)."*
    - I.e. x is helpful in predicting y
    - Parentheses used in *multiple* regression… coming soon!

# How will we do this?

- SPSS will calculate:
  - Some diagnostics that help us evaluate our assumptions about the data
  - Estimates of the coefficients
  - SD for the estimates of coefficients
  - T-scores comparing estimate to 0
  - p-values associated with the T-test
- Humans will
  - Interpret the output in plain English!
  - Then explain it to constituents so that *they* can understand it too!

# Back to our SPSS Example

- Find the Table of Coefficients
  - T-values, p-values
  - Can you construct the linear equation?
- Evaluate when our model is least & most accurate
  - Plot actual vs. predicted values
  - Calculate r, $r^2$ & Interpret

# Multiple Regression Analysis

# Simple vs. Multiple Regression

- "Simple" refers to predicting a single outcome (y) from a *single* predictor (x)

- "Multiple" refers to predicting a single outcome (y) from *two or more* predictors (x1, x2, x3)

  - Still assuming a linear relationship
    - But there are ways to "coax" linearity if it's not already there…

# Multiple Regression Examples

- **Ex.** Predict *Faculty Salary* from Age, Department, Years as Faculty Member and Gender

- **Ex.** Predict *Student Performance on GE Outcomes* from Cumulative GPA, College Major, and Gender

# Multiple Regression

- We're still talking about *Linear* relationships

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 \ldots + b_k x_k$$

- Still using method of Least Squares to develop the equation

- Still estimating regression coefficients (betas)

# Multiple Regression

- But Graphing the equation results in a plane, or more complex geometric shape, not a line, even though the relationships are still linear…
  - 3-D graphing?
  - …or 2-D graphing?

# Multiple Regression Plane (example w/2 predictors)



**Predicting Sales from Price and Advertising**

Linear Regression with
95.00% Mean Prediction Interval

**Sales = 8.31 + -0.09 * price + 0.38 * advertising**
**R-Square = 0.45**
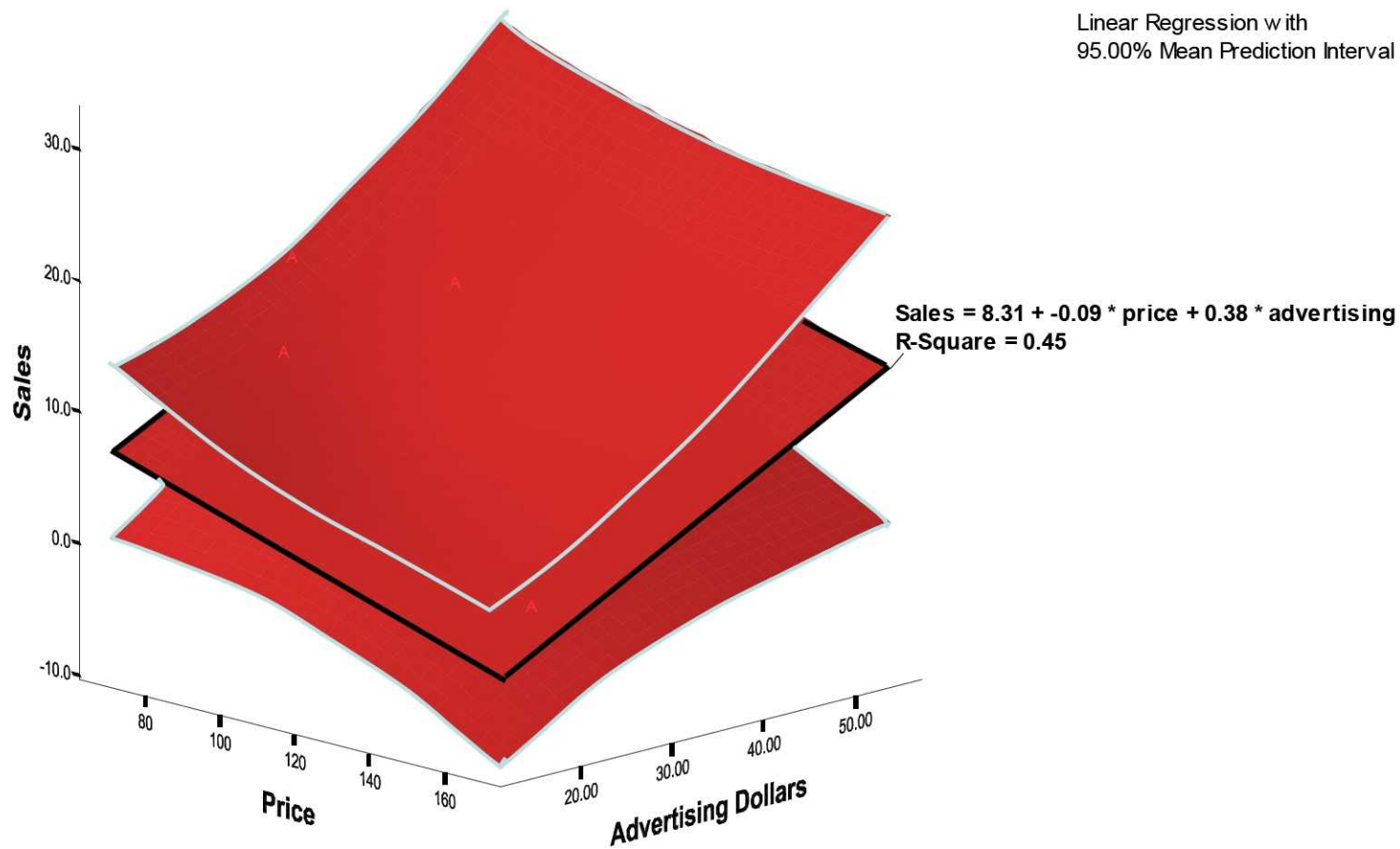
# Assumptions for the Population Multiple Regression

*SAME AS Simple Regression!*

1. Expected value of disturbances is zero : $E(e_i) = 0$

2. Variance of each $e_i$ is equal to $\sigma_e^2$

    (i.e. each disturbance along the regression

    line has equal variance regardless of value of x)

3. The $e_i$ are normally distributed.

4. The $e_i$ are independent.  (Be careful RE longitudinal

    data... usually not independant.)

# Assumptions for the Population Multiple Regression

1. Expected value of disturbances is zero : $E(e_i) = 0$

2. Variance of each $e_i$ is equal to $\sigma_e^2$

   (i.e. each disturbance along the regression

   line has equal variance regardless of value of x)

3. The $e_i$ are normally distributed.

4. The $e_i$ are independent. (Be careful RE l[    ]

   data... usually not independant.)

5. Predictors themselves are independant

One new assumption, because we have multiple predictors...

# Assessing the FIT of a Multiple Regression Line

- In Simple Regression, we mainly focused on discussing the significance of the regression coefficients.

- In Multiple Regression, we must also pay attention to the *overall* Regression Equation?

  - Is it any good at predicting?
  - How do we know?

# Assessing the FIT of a Multiple Regression Line

- ## With Simple Regression, we didn't pay much attention to this.

  - If *the coefficient* was significant, that implied that the equation itself was too.

- ## With Multiple Regression, we must first evaluate the overall equation before diving deeper.

  - Then determine *which, if any* coefficients are significant.

# Recall Hypothesis Testing with Simple Regression…

$$\hat{y} = b_0 + b_1 x$$

$$H_0 : \beta_k = \beta_k^* \text{ (usually 0)}$$
$$H_a : \beta_k \neq \beta_k^*$$

$$t = \frac{b_k - \beta_k^*}{s_{b_k}} = \frac{b_k}{s_{b_k}} \text{ (when } \beta_k^* = 0)$$

$$\text{Reject } H_0 \text{ if } t > t_{\alpha/2} \text{ or } t < t_{\alpha/2}$$
$$\text{Accept } H_0 \text{ if } -t_{\alpha/2} \leq t \leq t_{\alpha/2}$$

# Hypothesis Testing with Multiple-Regression

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 ... b_k x_k$$

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

$$H_a : \text{At least one coefficient is not equal to zero}$$

*If reject*

$$H_0 : \beta_k = \beta_k^* \ (\text{usually } 0)$$

$$H_a : \beta_k \neq \beta_k^*$$

$$\text{Reject H}_0 \text{ if } F > F(\alpha; K, n - K - 1)$$

$$\text{Accept H}_0 \text{ if } F \leq F(\alpha; K, n - K - 1)$$

$$t = \frac{b_k}{s_{b_k}}$$

$$\text{Reject H}_0 \text{ if } t > t_{\alpha/2} \text{ or } t < t_{\alpha/2}$$

$$\text{Accept H}_0 \text{ if } -t_{\alpha/2} \leq t \leq t_{\alpha/2}$$

# The ANOVA Summary Table

- Evaluation of the overall "Fit" of our Regression Equation

- Do all coefficients= 0 (null hyp) or is at least one of them ≠ 0 (alt hyp).

- The results of the ANOVA are found in an ANOVA Summary Table...

$$\text{Reject H}_0 \text{ if } F > F(\alpha; K, n - K - 1)$$
$$\text{Accept H}_0 \text{ if } F \leq F(\alpha; K, n - K - 1)$$

# The ANOVA Summary Table

| Source | DF | SS | MS | F | $\underline{p}$ |
|---|---|---|---|---|---|
| Regression | K | SSR | SSR/K | MSR/MSE | p-value |
| Residual Error | n-k-1 | SSE | SSE/(n-K-1) | | |
| Total | n-1 | SST | | | |

**Sums of Squares**

# ANOVA SS and MS terms…

- The F-Ratio (MSR/MSE), and the associated p-value, tell us whether or not our regression equation is predicting a "significant" amount of the variance in y from knowledge of $x_1$, $x_2$,…$x_k$.
  - If p=value < 0.05 (traditionally), equation is said to be "significant"

# The $R^2$ Term
## (Generated from SS Terms)

- The variation in y:  $SST = SSE + SSR$
  - "Total SS=Regression SS+ Error SS"
- $R^2$ = the ratio of explained-to-total variance (SS) is an evaluation of the overall regression.

$$R^2 = \frac{SSR}{SST}$$

- I.e. "percent of variance accounted for"

# ANOVA p-value VS. Multiple-$R^2$

- ANOVA p-value tells us whether we can account for a significant proportion of variance in Y, by knowledge of all of the predictors ($X_1$, $X_2$... $X_k$).
  - F=MSR/MSE... associated with a p-value

- Multiple-$R^2$ tells is an estimate of how much variance we can account for in Y by knowledge of all of the predictors ($X_1$, $X_2$... $X_k$).
  - $R^2$=SSR/SST

# The Multiple R-Square

- The "Multiple-$R^2$" value is very similar to the correlation coefficient.

$$R^2 = \frac{SSR}{SST} = \left(1 - \frac{SSE}{SST}\right)$$

- But in multiple-regression it has a flaw…
  - It doesn't decrease as new predictors are added, even if they are "useless" additions.

# The Adjusted Multiple-R-Square

- We need to "adjust" the $R^2$ value to correct for the addition of more predictors

$$R^2 = 1 - \frac{SSE}{SST}$$

Num predictors

$$R_{adj}^{\ 2} = 1 - \frac{SSE/(n - K - 1)}{SST/(n - 1)}$$

- Note how the SS in numerator and denominator are adjusted for their df?

# The Adjusted Multiple-R-Square

- This "adjustment" results in an adjusted R-square value that compensates for the number of predictors in the model.
  - No longer represents "the percent of variance accounted for"
  - But CAN BE used to compare different multiple-regression models
    - More on this later…

# Let's Give it a Try, eh?

- SPSS Example of Multiple Regression
  - One Outcome Variable (number SAT scores sent to NH)
  - Several Predictors (Total SAT Takers, other predictors)
- What's the $R^2$? Adjusted $R^2$?
- Is the equation itself any good? (i.e. can it account for sig. prop. of Y variance?)
- Which, if any, of the predictors are useful?
  - Interpret

# Steps We'll Take…

1. F-test for overall fit of regression

2. If F-test is significant, examine the t-tests for each of the coefficients.

3. Report the total percent variation in y explained by the x predictors

4. Examine the Adjusted R2

# SPSS Example

- Multiple Regression predicting Number of SAT scores sent to UNH, by
  - Total number of test takers
  - Average SAT Verbal Score
  - Average SAT Math Score

# Statistical Variable Selection Techniques

# Strategies when theory cannot guide you…

- Thus far, the theory has been our guide on choosing predictors to consider
  - Theory is always the best strategy!!

- Sometimes you may be on a "data mining" mission…

- There are techniques that can help you
  - With Rob's strong dose of caution!

# Statistical Strategies

- **Selection algorithms**: rules for deciding when to drop or add variables
  1. Backwards Elimination
  2. Forward Selection
  3. Stepwise Regression
  4. Run All Possible Models

# Words of Caution

- None guarantee you get the right model because they do not check assumptions or search for omitted factors like curvature.

- None have the ability to use a researcher's knowledge about the situation being analyzed.

- Many among the scientific community do not respect statistical selection strategies like these because they are not grounded in theory, and they capitalize on sample variance relationships that may not exist in the population…

# Backwards Elimination

- Start with all variables in the equation.
- Examine the variables in the model for significance and identify the least significant one.
- Remove this variable if it does not meet some minimum significance level.
- Run a new regression and repeat until all remaining variables are significant.

# Forward Selection

- At each stage, it looks at the $x$ variables not in the current equation and tests to see if they will be significant if they are added. (I.e. a significant partial-F statistics would result.)

- In the first stage, the $x$ with the highest correlation with $y$ is added.

- At later stages it is much harder to see how the next $x$ is selected.

# Stepwise Regression

- A limitation with the backwards procedure is that a variable that gets eliminated is never considered again.

- With forward selection, variables entering stay in, even if they lose significance.

- Stepwise regression corrects these flaws. A variable entering can later leave. A variable eliminated can later go back in.

# Stepwise…

- Begins like Forward (Chooses best predictor, adds it, tests for sig Partial-F, Keeps if pass criteria)
- Next look at remaining predictors, choose "best" one (Highest Partial-F), and includes it.
- Then behaves like Backwards… Potentially REMOVING one of the variables already included if it's not necessary.
- Then adds a new one…
- Then tests to remove any of those included…
- Until finished.

# Stepwise…

- Ultimately, all variables in the equation are adding significantly, and none of the ones eliminated would (according to criteria we establish ahead of time)

- It is possible to add a variable, remove it later, then add again at a later step!

# Example of using Stepwise with SPSS

- Consider parameters for adding/removing variables
  - "Alpha to Remove"
    - maximum p-value a variable can have and stay in the equation
  - "Alpha to Enter"
    - minimum p-vale a variable needs to enter the equation

- Often we use values like .15 or .20 because this encourages the procedures to look at models with more variables.

# SPSS Example

- Predict the number of SAT scores sent to NH
  - Using Stepwise Selection Technique
    - Including all (untransformed) predictors in the data set

# All Possible Models?

- If reasonable, this is likely a better solution than Stepwise, but…
  - Some software (SPSS) cannot easily accommodate
  - Can be unreasonable if there are many potential predictors
  - Still not as good as theory
    - ex. what if a non-linear transformation is really the driver?
- Model selection usually based on Adjusted $R^2$ for OLS regression

# Related Advanced Topics

- Regression with Bivariate or Multinomial Outcomes
  - Logit Procedure
    - Quite different in output and interpretation
  - Logistic Regression
    - Useful for >2 categories to the outcome variable
- Regression for Ordinal outcomes
  - Poisson, Negative Binomial, Others
- Hierarchical (nested) Regressive models
  - "Block" Regression in SPSS
  - Used to compare models with increasing complexity…

# New Advances

- Hierarchical Linear Modeling (HLM)
  - A.k.a. MLM, Mixed Modeling
- Allows for nesting to be considered
- Allows both fixed and random effects
- Allows for time-dependant covariates
- Allows for group-level effects modeling

# Time Series Regression

## Modeling Known (or easily observable) Patterns in Time-Series Data

# Time Series Regression

- Some data are linear in their relationship, but have "cycles" that we'd like to capture also
  - Ex. New Admits over time
    - Where predictable cycles exist among Fall, Spring, Summer terms

# Consider this time series plot:
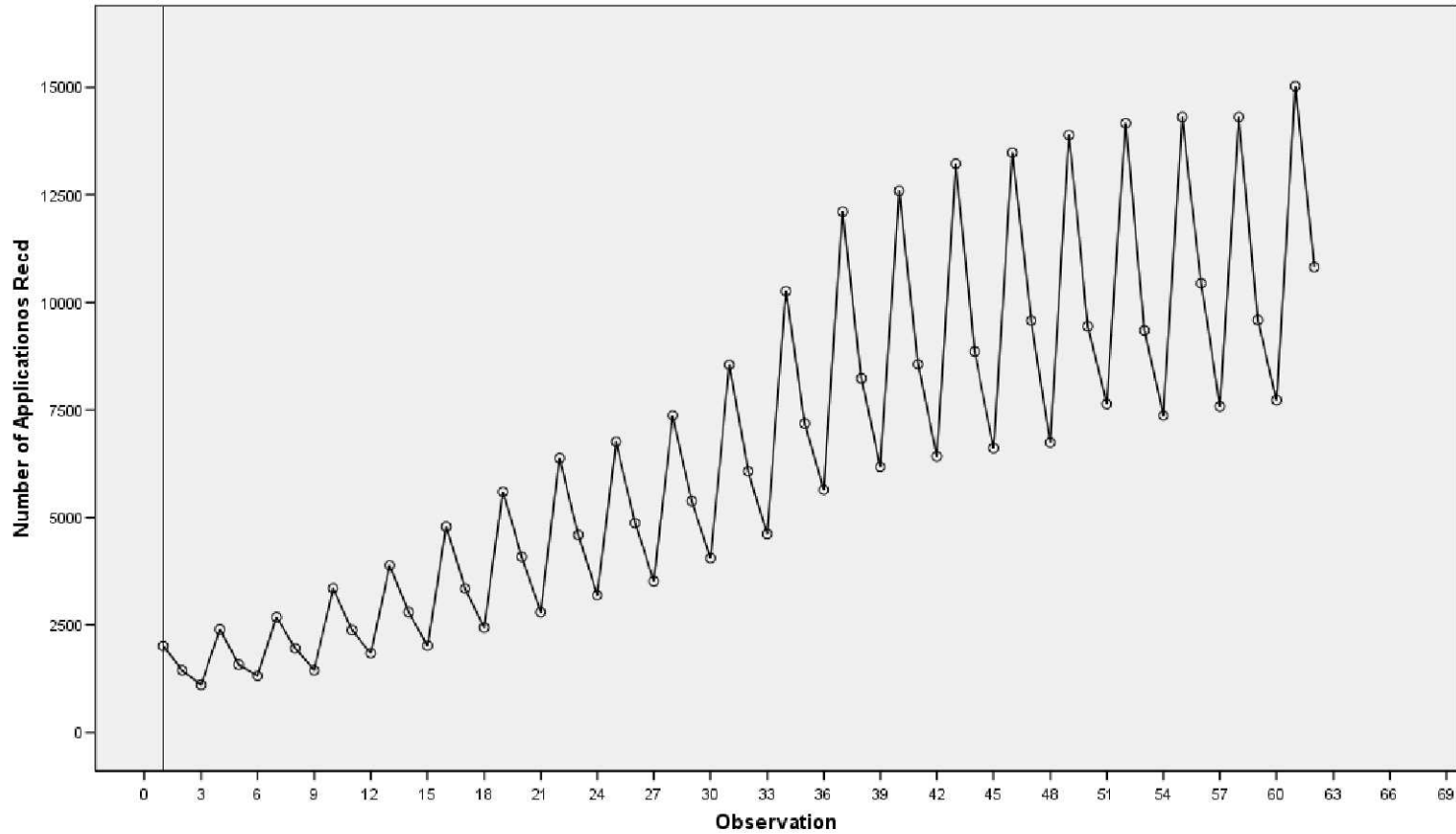
# Consider this time series plot:
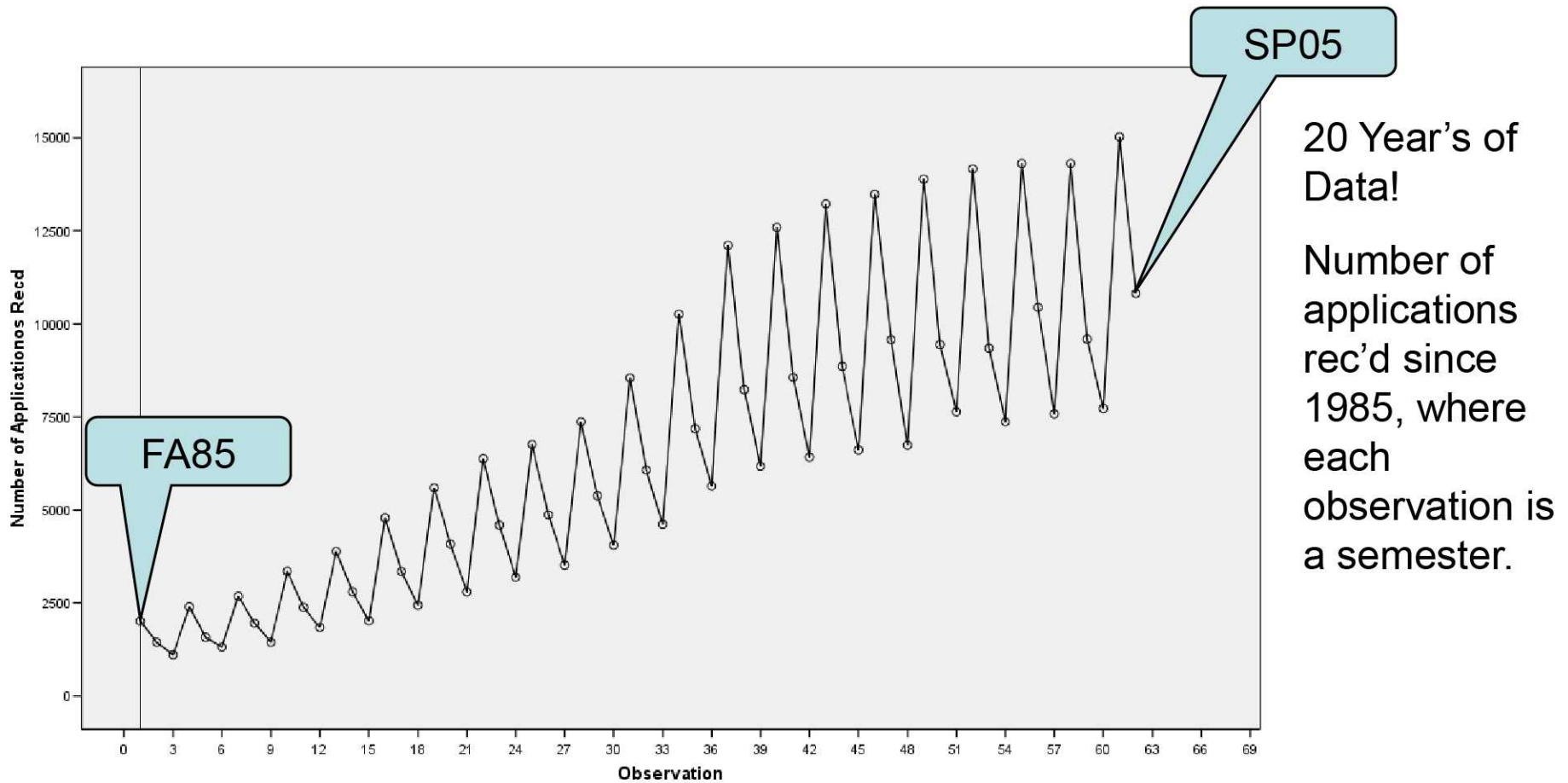
# Consider this time series plot:
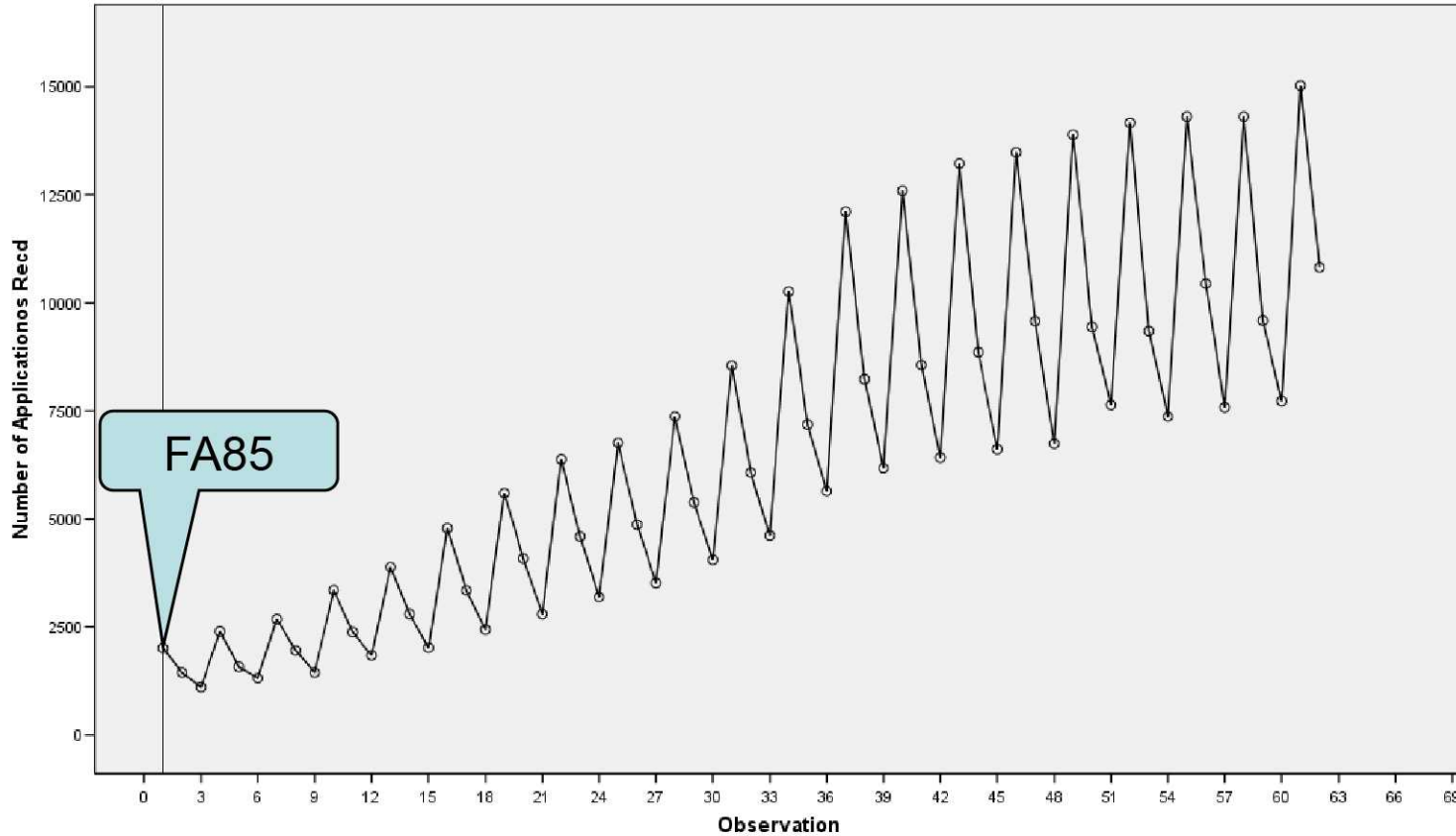


Beginning with Fall, 1985

# Consider this time series plot:
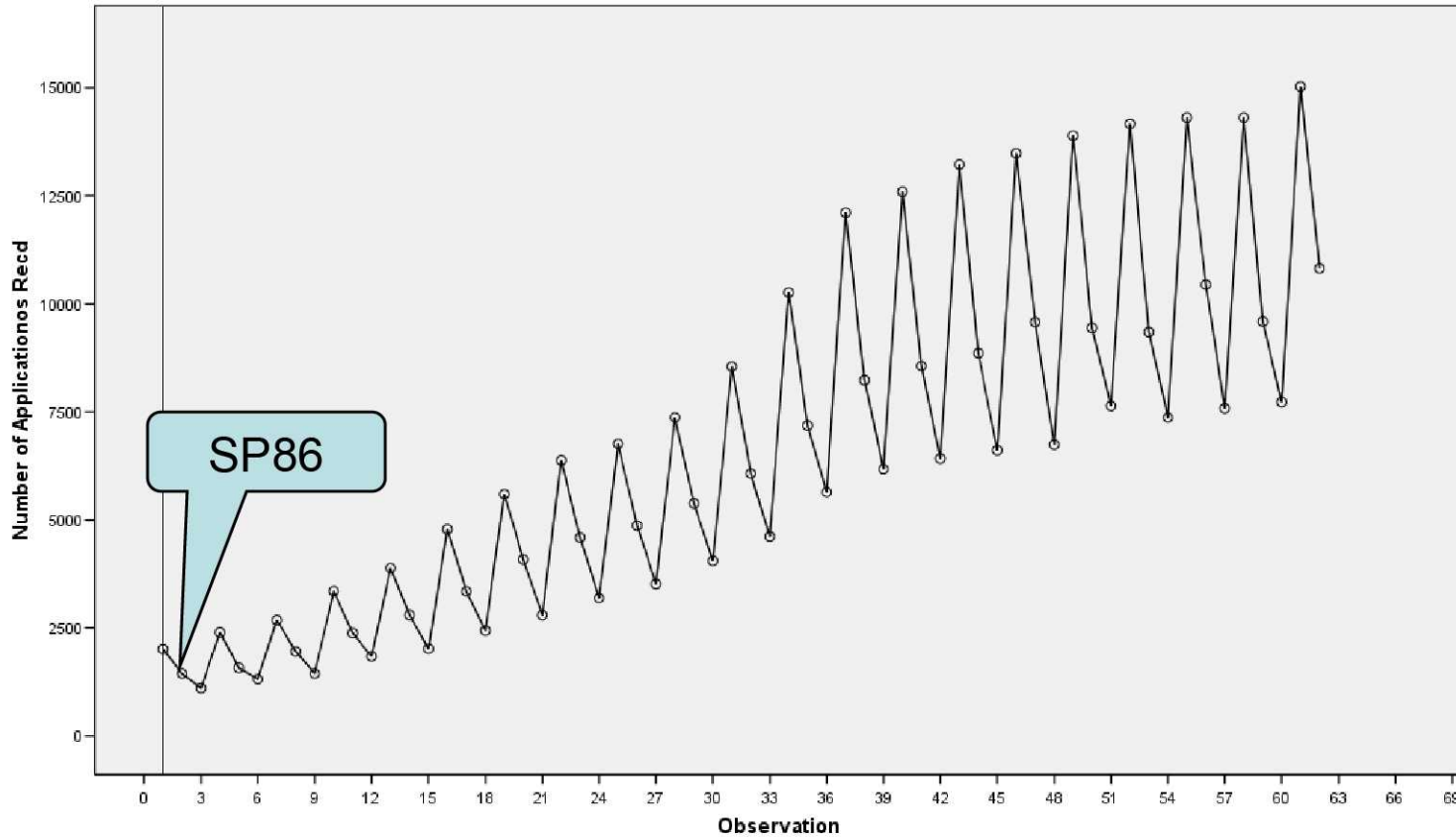


Then Spring, 1986

# Consider this time series plot:
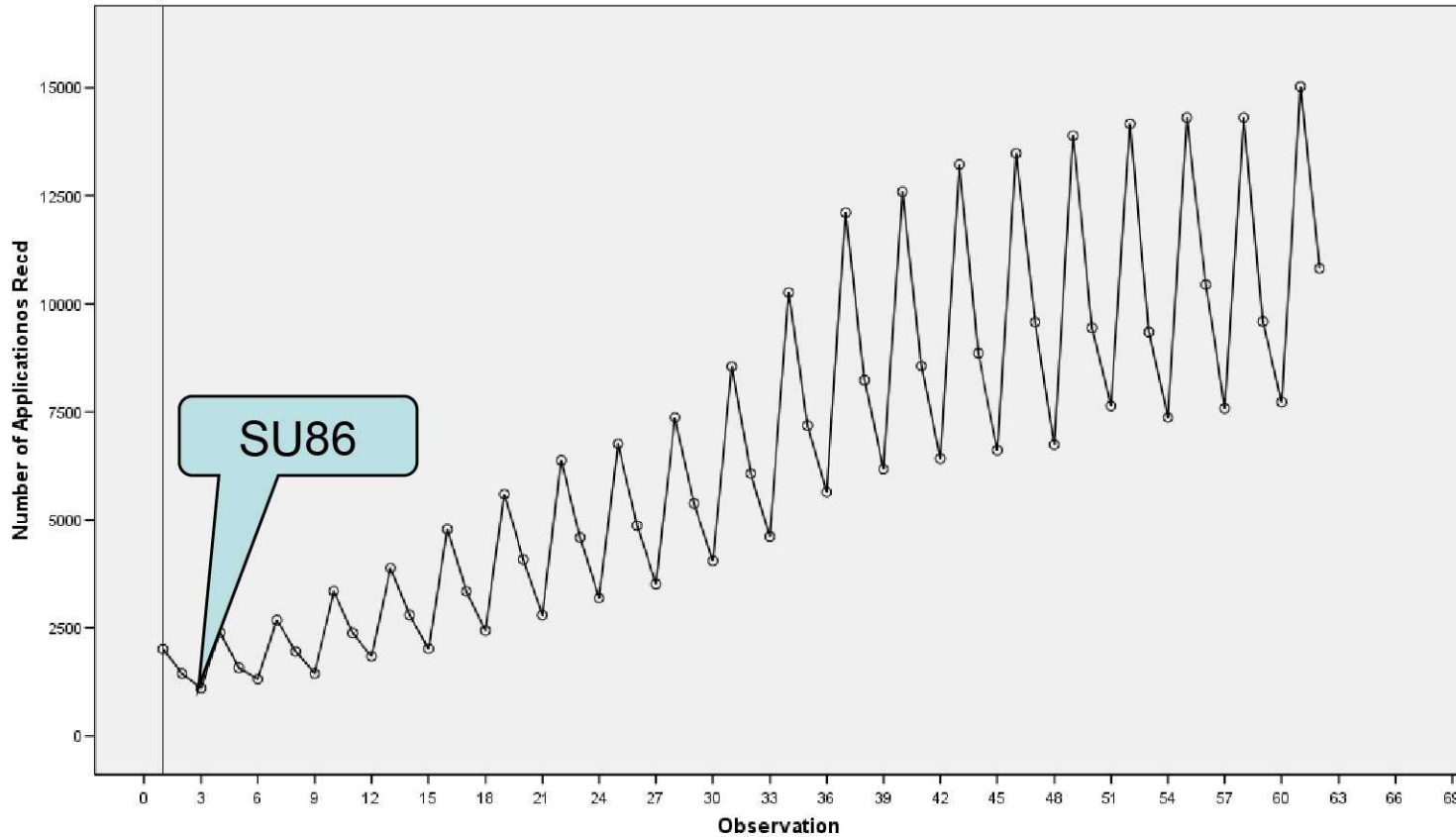


Followed by Summer, 1986….

And so on.

# We could fit a linear regression to the data…$R^2$ = .71



**Model Summary**[b]

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .841[a] | .707 | .702 | 2161.135 |

a. Predictors: (Constant), Observation

b. Dependent Variable: Number of Applicationos Recd

R Sq Linear = 0.707
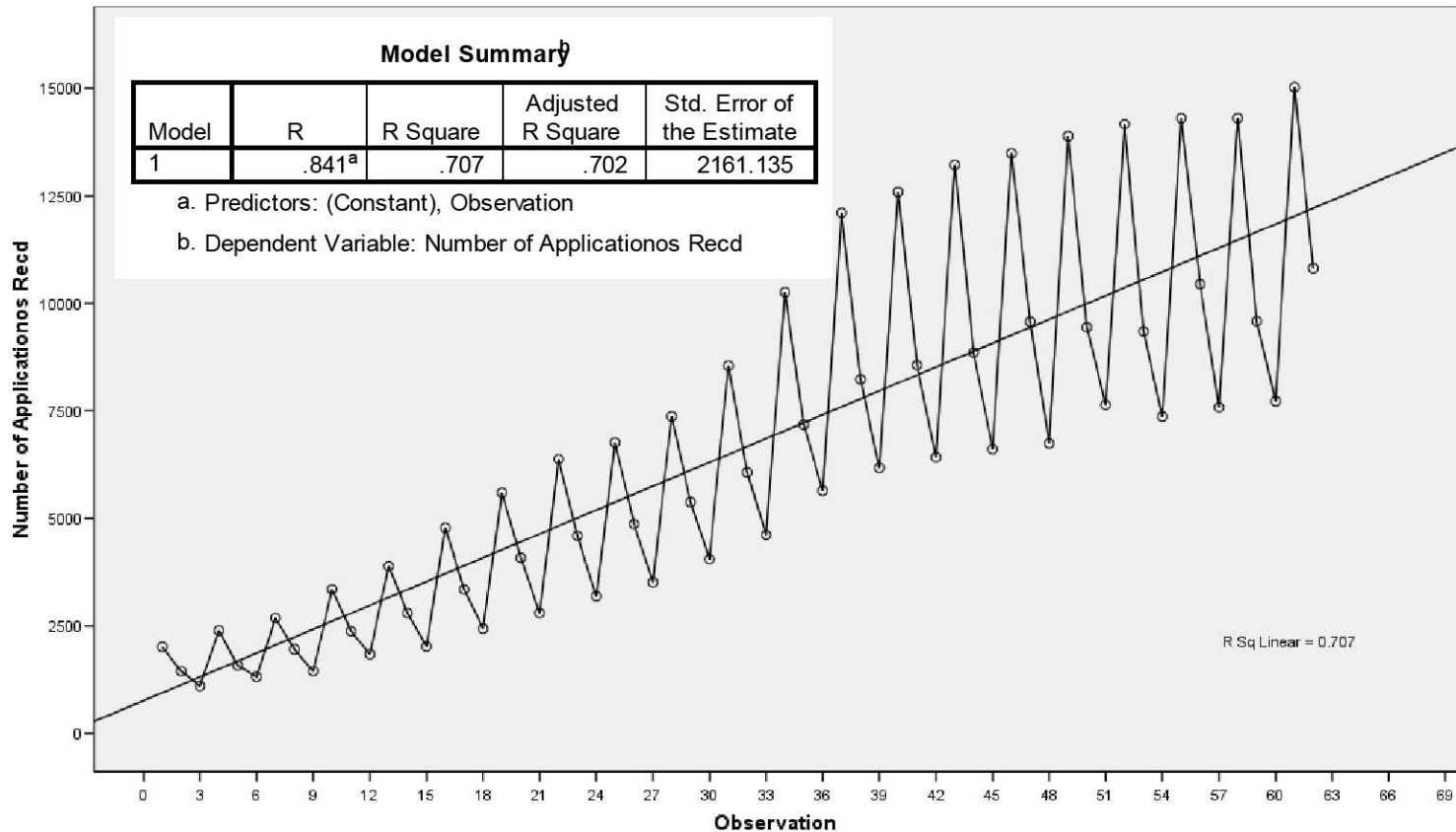
# But if we could "capture" the cycle of FA/SP/SU wouldn't that be better?
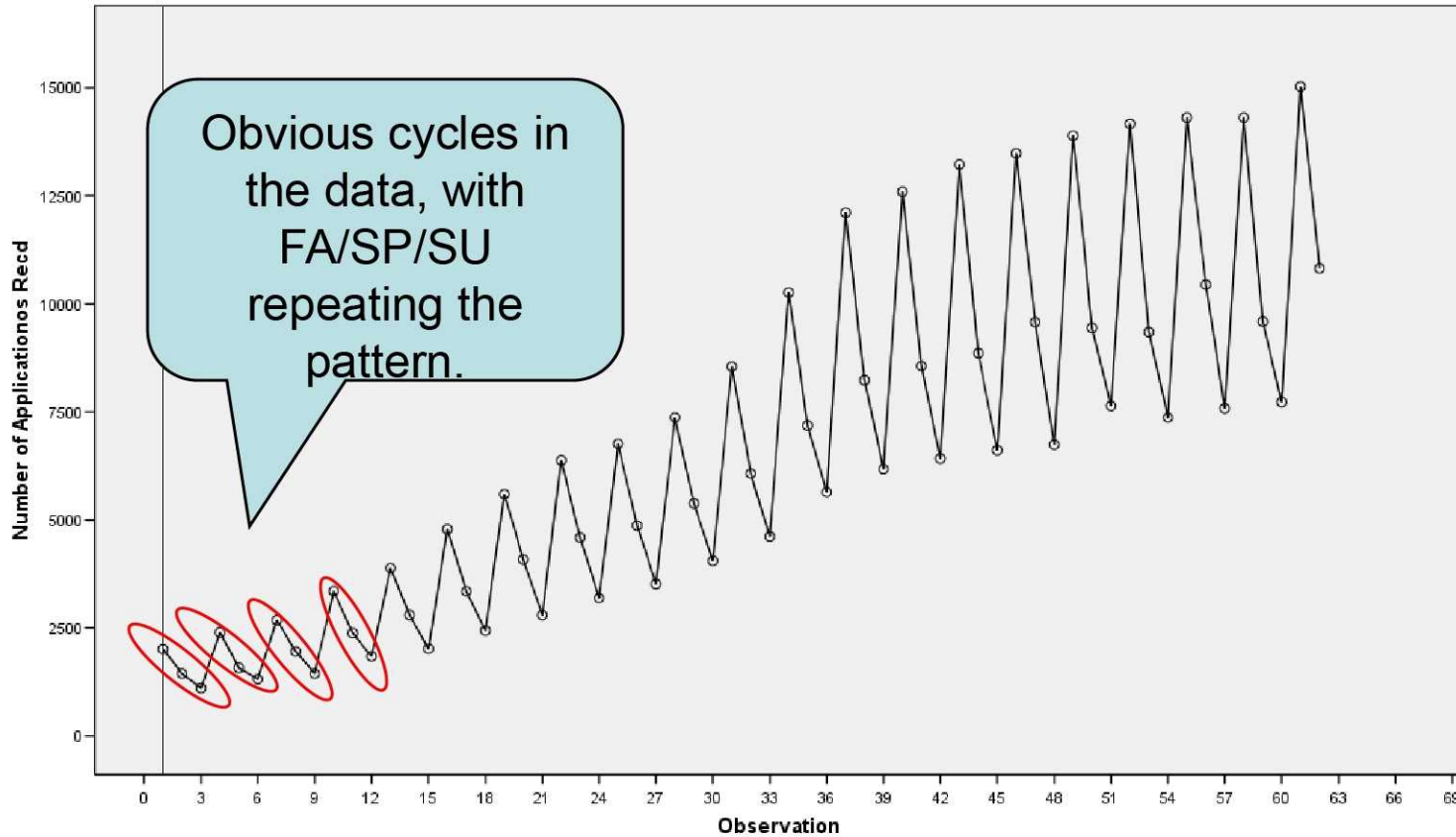
# How?

- Include "dummy" predictors indicating whether data is from a  Fall, Spring, or Summer term
  - Use 2 of the 3, leaving one as "reference"
- Run Multiple Regression:

# Comparison of the two models:

## Simple Regression Model

### Model Summary[b]

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .841[a] | .707 | .702 | 2161.135 |

a. Predictors: (Constant), Observation

b. Dependent Variable: Number of Applicationos Recd

### ANOVA[b]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 6.8E+008 | 1 | 677290217.4 | 145.014 | .000[a] |
| | Residual | 2.8E+008 | 60 | 4670502.918 | | |
| | Total | 9.6E+008 | 61 | | | |

a. Predictors: (Constant), Observation

b. Dependent Variable: Number of Applicationos Recd

### Coefficients[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | 761.925 | 555.637 | | 1.371 | .175 |
| | Observation | 184.692 | 15.337 | .841 | 12.042 | .000 |

a. Dependent Variable: Number of Applicationos Recd

## Time Series Regression with Seasonal Predictors

### Model Summary[b]

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .964[a] | .930 | .926 | 1075.387 |

a. Predictors: (Constant), Spring, Observation, Fall

b. Dependent Variable: Number of Applicationos Recd

### ANOVA[b]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 8.9E+008 | 3 | 296815296.0 | 256.659 | .000[a] |
| | Residual | 67074505 | 58 | 1156456.975 | | |
| | Total | 9.6E+008 | 61 | | | |

a. Predictors: (Constant), Spring, Observation, Fall

b. Dependent Variable: Number of Applicationos Recd

### Coefficients[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | -1352.942 | 340.067 | | -3.978 | .000 |
| | Observation | 186.214 | 7.634 | .848 | 24.393 | .000 |
| | Fall | 4490.993 | 336.016 | .541 | 13.365 | .000 |
| | Spring | 1611.276 | 336.016 | .194 | 4.795 | .000 |

a. Dependent Variable: Number of Applicationos Recd

# Is there anything else we could use in our data???
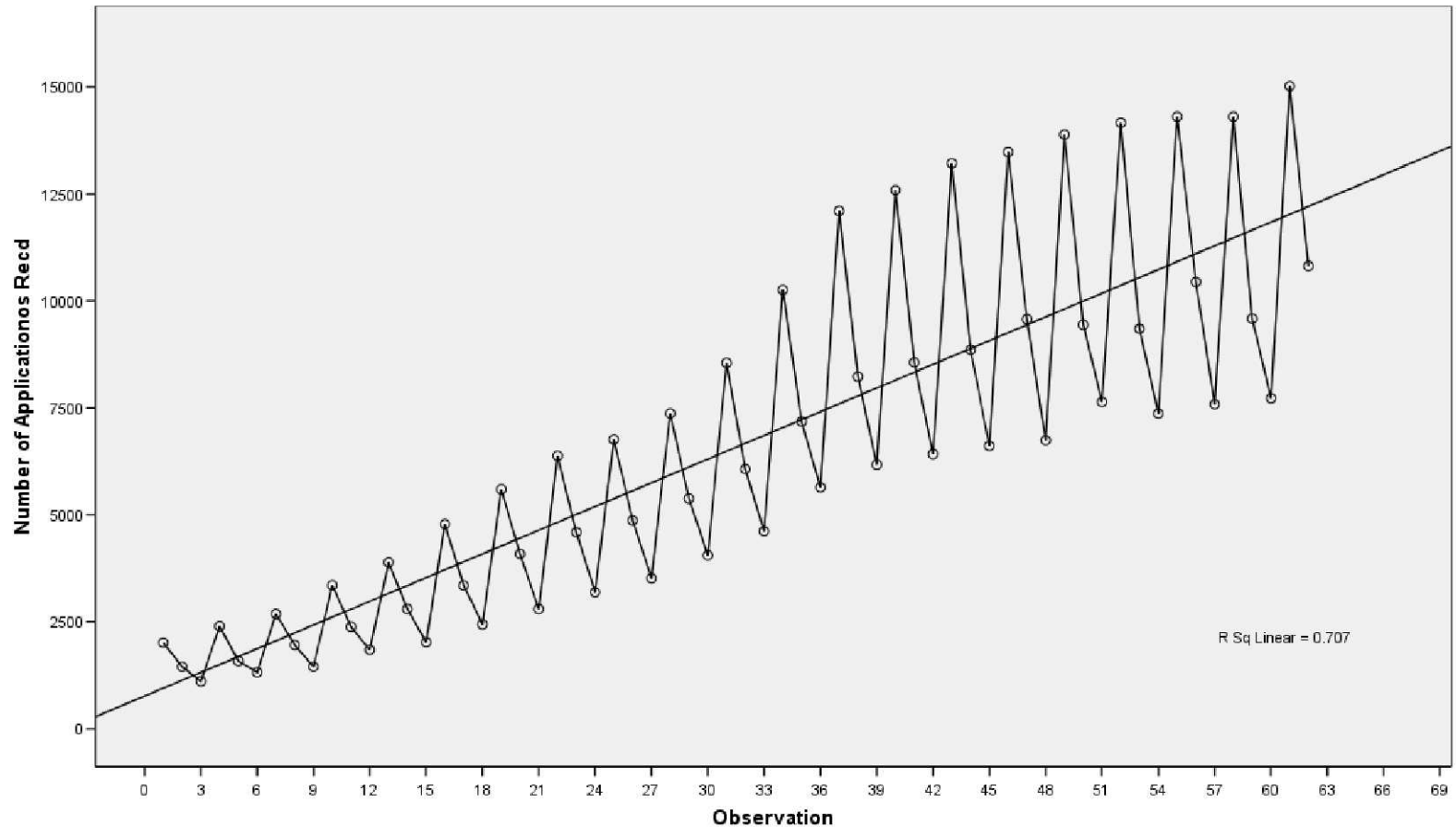
# Is there anything else we could use in our data???
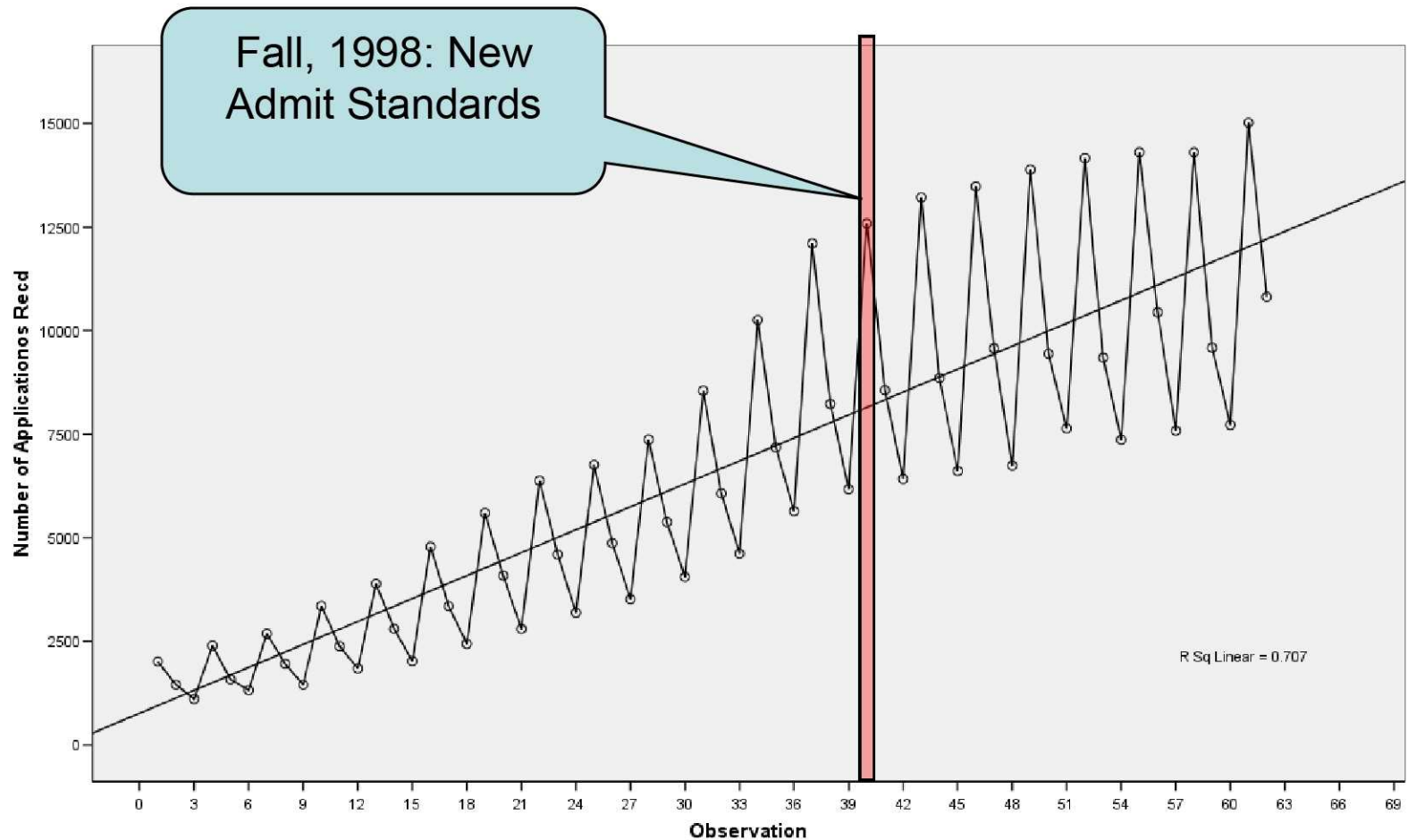
# Modeling Policy Changes…

- ## Simply add a dummy predictor that captures the policy change!
  - *…Though it may be tempting, please make no assumptions about the fact that "policy change" and "dummy" are in the same sentence.* ☺

**Variables Entered/Removed**[b]

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | New Admission Standards (FA98), Spring, Fall, Observation[a] | . | Enter |

a. All requested variables entered.

b. Dependent Variable: Number of Applicationos Recd

# Results?

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .965[a] | .931 | .926 | 1080.052 |

a. Predictors: (Constant), New Admission Standards (FA98), Spring, Fall, Observation

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 8.9E+008 | 4 | 222757289.6 | 190.960 | .000[a] |
| | Residual | 66491234 | 57 | 1166512.879 | | |
| | Total | 9.6E+008 | 61 | | | |

a. Predictors: (Constant), New Admission Standards (FA98), Spring, Fall, Observation

b. Dependent Variable: Number of Applicationos Recd

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | -1219.940 | 389.910 | | -3.129 | .003 |
| | Observation | 177.909 | 14.027 | .810 | 12.684 | .000 |
| | Fall | 4475.465 | 338.188 | .539 | 13.234 | .000 |
| | Spring | 1604.053 | 337.628 | .193 | 4.751 | .000 |
| | New Admission Standards (FA98) | 367.505 | 519.725 | .045 | .707 | .482 |

a. Dependent Variable: Number of Applicationos Recd

In this case, the policy change did not significantly impact admits after modeling time and cycles of semesters…

# Results?

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .965[a] | .931 | .926 | 1080.052 |

a. Predictors: (Constant), New Admission Standards (FA98), Spring, Fall, Observation

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 8.9E+008 | 4 | 222757289.6 | 190.960 | .000[a] |
| | Residual | 66491234 | 57 | 1166512.879 | | |
| | Total | 9.6E+008 | 61 | | | |

a. Predictors: (Constant), New Admission Standards (FA98), Spring, Fall, Observation

b. Dependent Variable: Number of Applicationos Recd

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | -1219.940 | 389.910 | | -3.129 | .003 |
| | Observation | 177.909 | 14.027 | .810 | 12.684 | .000 |
| | Fall | 4475.465 | 338.188 | .539 | 13.234 | .000 |
| | Spring | 1604.053 | 337.628 | .193 | 4.751 | .000 |
| | New Admission Standards (FA98) | 367.505 | 519.725 | .045 | .707 | .482 |

a. Dependent Variable: Number of Applicationos Recd

And the adjusted R2 is similar to the simpler model that did not include the policy change predictors.

# Exploratory Factor Analysis

# Principal –Axis Exploratory Factor Analysis

- Common in survey research
- Useful for "discovery" of underlying constructs
- Useful as a strategy for condensing data
- Useful as a strategy to approximate continuous data from ordinal data elements
  - Combining several Likert-Scaled items into one construct score that behaves "normally"

# What EFA can do?

- Purpose is to discover simple patterns among variables
- If patterns are found, we call them "factors," or "constructs—hence the name
- EX: Is intelligence uni-dimensional, or multi-dimensional?
  - Hint: Consider College Board Exams…
    - Verbal, Math, Logic

# EFA Graphic (after solution has been rendered)
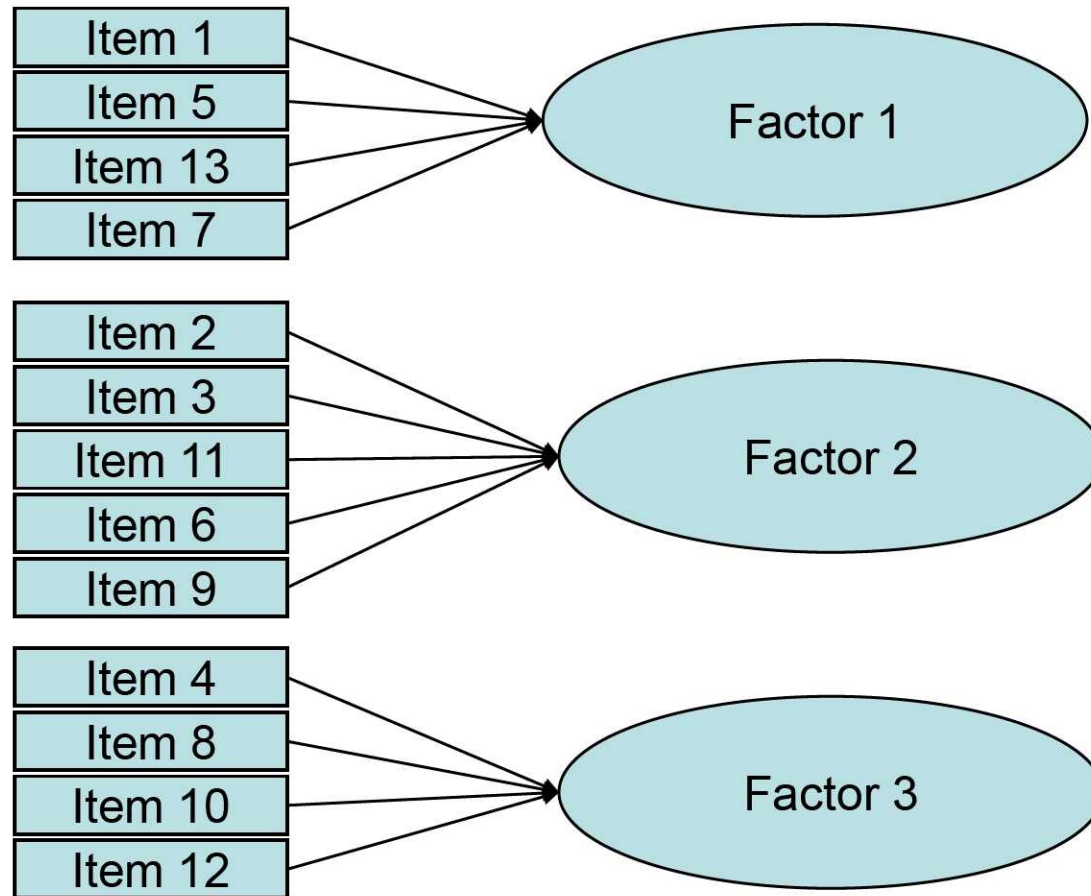
# Goals of EFA

- To better understand the underlying constructs.
  - Thus inferences are made about the constructs, not individual items
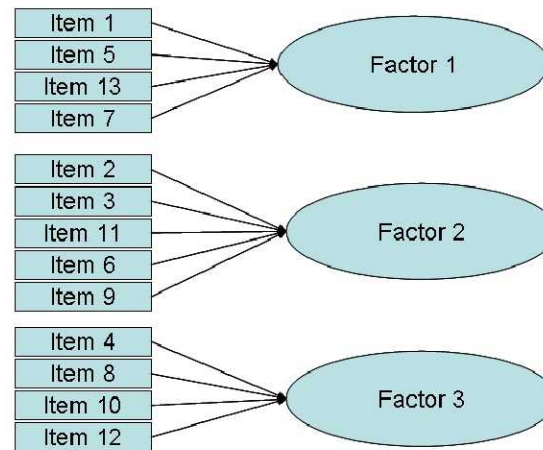
# How it works
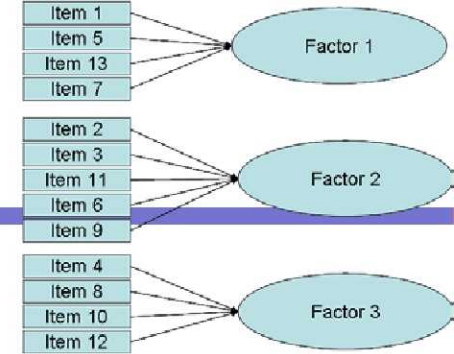


- Begins with a simple correlation matrix
  - In fact, you don't need raw data in some cases (rotation)
- EFA attempts to "categorize" variables according to how similar/dissimilar they are to other variables
  - By calculating factor loadings….
- Goal is to produce the minimum number of factors that adequately explains the data

# Some Details

- As always, there are options w/how to run EFA, and there are detailed references available

- "Rotation" options to simplify our understanding of factor structure

  – Orthogonol (more complex structure, but more independence of factors)

  – Oblique (simpler structure, but factors may correlate more)

# Rotation

- Procedure (ex. Varimax) that searches for linear combinations (i.e. rotations) of original factors so that the variance of the loadings is maximized

# Varimax Rotation

- Varimax Rotation is probably the most popular choice for EFA (Kaiser, 1958)
  - Each factor should have small number of items loading heavily on it
    - Each variable should load mostly onto only one factor
    - Thus simplifying our understanding of underlying constructs

# Wine Tasting Example

- Factor Rotations in Factor Analyses
  - Herve Abdi, University fo Texas at Dallas
- Five Wines are Rated by Seven Questions

Table 1: An (artificial) example for PCA and rotation. Five wines are described by seven variables.

|  | Hedonic | For meat | For dessert | Price | Sugar | Alcohol | Acidity |
|---|---|---|---|---|---|---|---|
| Wine 1 | 14 | 7 | 8 | 7 | 7 | 13 | 7 |
| Wine 2 | 10 | 7 | 6 | 4 | 3 | 14 | 7 |
| Wine 3 | 8 | 5 | 5 | 10 | 5 | 12 | 5 |
| Wine 4 | 2 | 4 | 7 | 16 | 7 | 11 | 3 |
| Wine 5 | 6 | 2 | 4 | 13 | 3 | 10 | 3 |

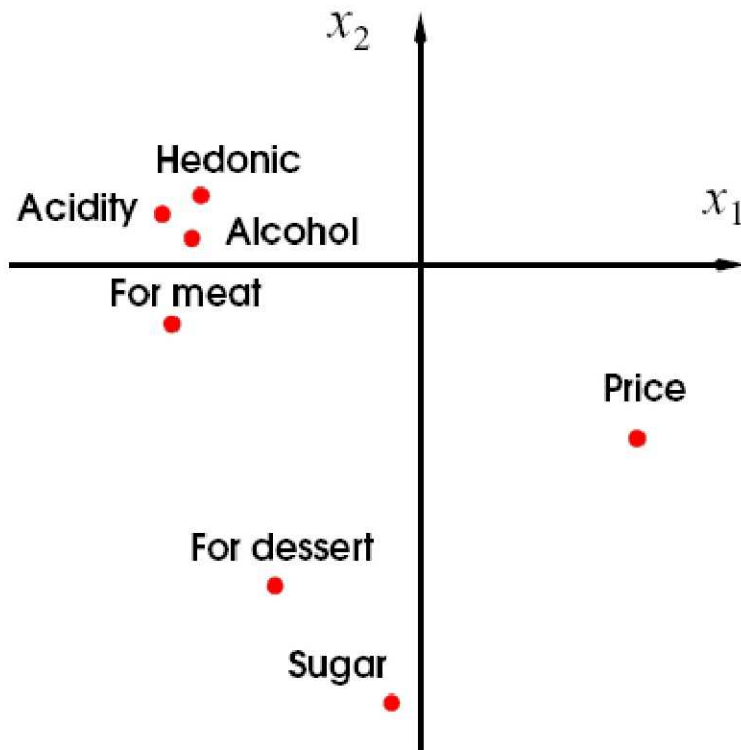# Unrotated Two-Factor Solution: See any patterns?



Table 2: Wine example: Original loadings of the seven variables on the first two components.

|  | Hedonic | For meat | For dessert | Price | Sugar | Alcohol | Acidity |
|---|---|---|---|---|---|---|---|
| Factor 1 | −0.3965 | −0.4454 | −0.2646 | 0.4160 | −0.0485 | −0.4385 | −0.4547 |
| Factor 2 | 0.1149 | −0.1090 | −0.5854 | −0.3111 | −0.7245 | 0.0555 | 0.0865 |

# Varimax Rotated Solution: See any patterns?
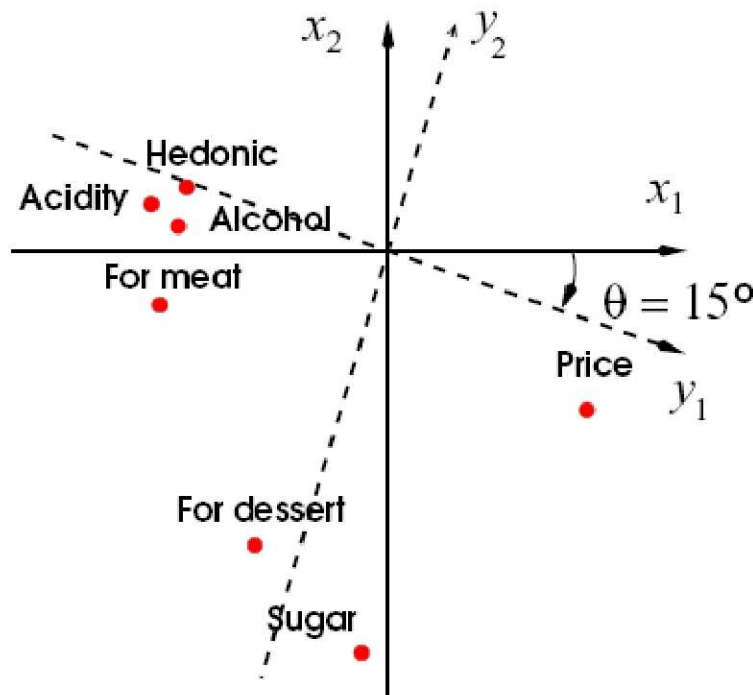


Table 3: Wine example: Loadings, after VARIMAX rotation, of the seven variables on the first two components.

| | Hedonic | For meat | For dessert | Price | Sugar | Alcohol | Acidity |
|---|---|---|---|---|---|---|---|
| Factor 1 | −0.4125 | −0.4057 | −0.1147 | 0.4790 | 0.1286 | −0.4389 | −0.4620 |
| Factor 2 | 0.0153 | −0.2138 | −0.6321 | −0.2010 | −0.7146 | −0.0525 | −0.0264 |

- **Seems likely that there are two dimensions**
  - One factor of sweetness
  - The other linked to price and complex taste qualities

# Next Example using SPSS for EFA

- Data set provided by Mary Ann Coughlin
  - Editor & Co-Author of AIR Intermediate/Advanced Statistics in IR Monograph
  - Twenty-Seven item questionnaire of graduates from a small liberal arts college.
  - Asking about things they gained from their education

# Next Example using SPSS for EFA

# How would you summarize the data?

- Table of Means?
- Histograms?

# Table of Descriptive Statistics:

**Descriptive Statistics**

| | N | Range | Minimum | Maximum | Mean | | Std. |
|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic |
| Write effectively | 621 | 3 | 0 | 3 | 2.46 | .026 | .648 |
| Communicate well orally | 618 | 3 | 0 | 3 | 2.21 | .031 | .779 |
| Acquire new skills and knowledge on my own | 614 | 3 | 0 | 3 | 2.45 | .029 | .714 |
| Think analytically and logically | 616 | 3 | 0 | 3 | 2.40 | .029 | .716 |
| Formulate creative / original ideas and solutions | 611 | 3 | 0 | 3 | 2.15 | .032 | .803 |
| Evaluate and choose between alternative courses | 608 | 3 | 0 | 3 | 2.01 | .033 | .816 |
| Lead and supervise tasks and groups of people | 611 | 3 | 0 | 3 | 1.89 | .038 | .949 |
| Relate well to people of different races, nations | 615 | 3 | 0 | 3 | 2.23 | .035 | .873 |
| Function effectively as a member of a team | 610 | 3 | 0 | 3 | 1.91 | .035 | .860 |
| Use computers for basic tasks (word processing) | 611 | 3 | 0 | 3 | 2.02 | .042 | 1.045 |
| Use computers for complex tasks (graphing) | 610 | 3 | 0 | 3 | 1.00 | .045 | 1.115 |
| Place current problems in historical prospective | 609 | 3 | 0 | 3 | 2.23 | .034 | .835 |
| Identify moral and ethical issues | 612 | 3 | 0 | 3 | 2.09 | .034 | .842 |
| Understand myself, my abilities, interests | 614 | 3 | 0 | 3 | 2.41 | .032 | .800 |
| Function independently without supervision | 612 | 3 | 0 | 3 | 2.26 | .035 | .870 |
| Gain in-depth knowlegde of a field | 614 | 3 | 0 | 3 | 2.38 | .030 | .744 |
| Plan and execute complex projects | 604 | 3 | 0 | 3 | 2.11 | .033 | .809 |
| Read or speack a foreign language | 609 | 3 | 0 | 3 | 1.38 | .048 | 1.196 |
| Appreciate art, literature, music, drama | 611 | 3 | 0 | 3 | 2.12 | .036 | .899 |
| Acquire broad knowledge in the Arts and Sciences | 614 | 3 | 0 | 3 | 2.19 | .032 | .796 |
| Develop feminist awarenenss | 613 | 3 | 0 | 3 | 2.56 | .029 | .710 |
| Develop awareness of social problems | 610 | 3 | 0 | 3 | 2.36 | .030 | .747 |
| Develop self-esteem /self-confidence | 614 | 3 | 0 | 3 | 2.35 | .035 | .859 |
| Form close friendships | 617 | 3 | 0 | 3 | 2.46 | .032 | .803 |
| Establish a course of action to accomplish goals | 614 | 3 | 0 | 3 | 2.13 | .031 | .780 |
| Synthesize and integrate ideas and information | 612 | 3 | 0 | 3 | 2.30 | .029 | .720 |
| Understand the role of science and technology | 614 | 3 | 0 | 3 | 1.55 | .039 | .956 |
| Valid N (listwise) | 537 | | | | | | |

# Bucket-o-Histograms
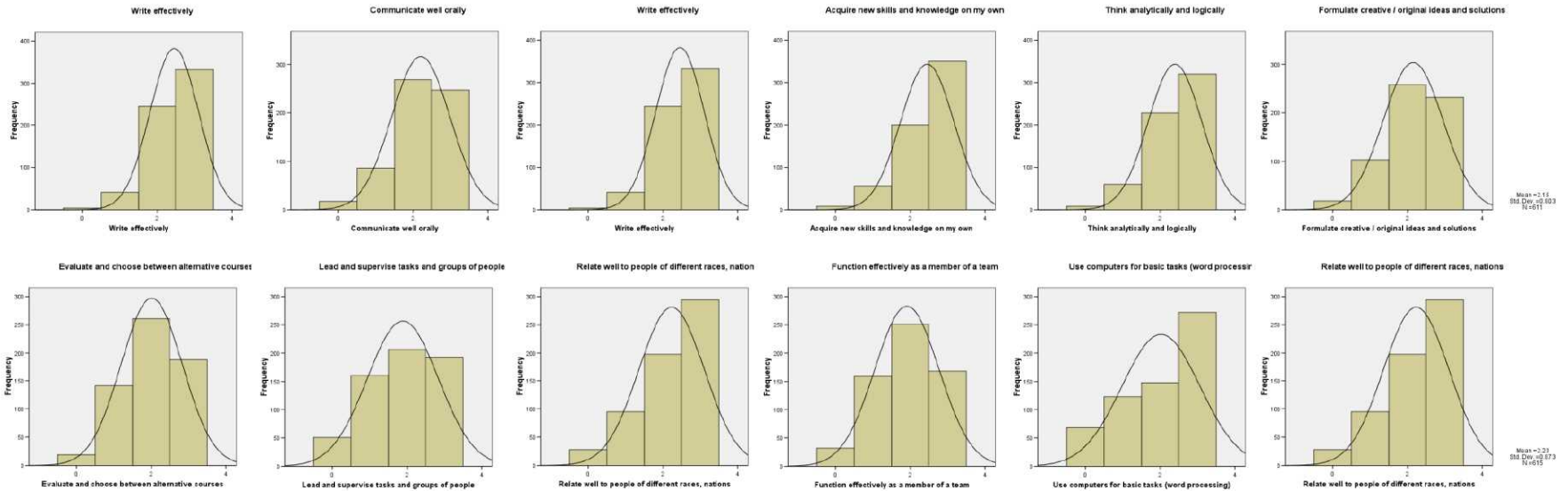


Etc…..

# What's the Big Picture?

- Simple to get the "take-home" message from our graduates?
  - Generally happy
  - Simple Descriptives can tell us that.
- But what's the "big picture" of the benefits of a college education from our college, in the eyes of our graduates?
  - This kind of question is ripe for EFA

# SPSS Case Study

- Run Exploratory Factor Analysis
  - Principal-Axis Method
  - With Varimax Rotation
- Interpret the Results!

# Six Underlying Constructs Resulting from our Rotated Factor Structure

**Rotated Factor Matrix[a]**

| | Factor | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Think analytically and logically | **.643** | .150 | .091 | .006 | .118 | -.026 |
| Formulate creative / original ideas and solutions | **.625** | .146 | .145 | .199 | .031 | .124 |
| Synthesize and integrate ideas and information | **.587** | .202 | .217 | .113 | .208 | .072 |
| Acquire new skills and knowledge on my own | **.581** | .070 | .141 | .090 | .095 | .146 |
| Plan and execute complex projects | **.537** | .042 | .127 | .148 | .230 | .168 |
| Write effectively | **.512** | .263 | .094 | .027 | -.007 | .142 |
| Establish a course of action to accomplish goals | **.496** | .195 | .386 | .223 | .204 | .112 |
| Evaluate and choose between alternative courses | **.485** | .175 | .176 | .397 | .087 | .067 |
| Communicate well orally | **.460** | .119 | .265 | .207 | .037 | .168 |
| Gain in-depth knowlegde of a field | **.442** | .074 | .080 | .039 | .160 | .117 |
| Develop awareness of social problems | .158 | **.732** | .233 | .091 | .102 | .099 |
| Develop feminist awarenenss | .082 | **.560** | .148 | .025 | -.037 | .158 |
| Identify moral and ethical issues | .288 | **.542** | .168 | .185 | .098 | .126 |
| Place current problems in historical prospective | .303 | **.433** | .040 | .166 | -.063 | .198 |
| Form close friendships | .122 | .163 | **.569** | .143 | .077 | .084 |
| Understand myself, my abilities, interests | .328 | .275 | **.547** | .159 | -.056 | .156 |
| Develop self-esteem /self-confidence | .370 | .313 | **.543** | .165 | .096 | .162 |
| Function independently without supervision | .372 | .100 | **.455** | .170 | .165 | .189 |
| Relate well to people of different races, nations | .125 | .335 | **.337** | .307 | .165 | .070 |
| Lead and supervise tasks and groups of people | .178 | .089 | .189 | **.724** | .149 | .048 |
| Function effectively as a member of a team | .160 | .190 | .208 | **.567** | .240 | .038 |
| Use computers for complex tasks (graphing) | .082 | -.034 | -.024 | .112 | **.601** | -.030 |
| Understand the role of science and technology | .380 | -.013 | .089 | .138 | **.571** | .164 |
| Use computers for basic tasks (word processing) | .122 | .118 | .247 | .108 | **.402** | .053 |
| Appreciate art, literature, music, drama | .178 | .248 | .141 | .088 | -.054 | **.650** |
| Acquire broad knowledge in the Arts and Sciences | .269 | .124 | .064 | .050 | .277 | **.413** |
| Read or speack a foreign language | .076 | .079 | .078 | .003 | .038 | **.402** |

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

# Six Underlying Constructs Resulting from our Rotated Factor Structure

**Rotated Factor Matrix[a]**

|  | Factor | | | | | |
|---|---|---|---|---|---|---|
|  | 1 | 2 | 3 | 4 | 5 | 6 |
| Think analytically and logically | **.643** | .150 | .091 | .006 | .118 | -.026 |
| Formulate creative / original ideas and solutions | **.625** | .146 | .145 | .199 | .031 | .124 |
| Synthesize and integrate ideas and information | **.587** | .202 | .217 | .113 | .208 | .072 |
| Acquire new skills and knowledge on my own | **.581** | .070 | .141 | .090 | .095 | .146 |
| Plan and execute complex projects | **.537** | .042 | .127 | .148 | .230 | .168 |
| Write effectively | **.512** | .263 | .094 | .027 | -.007 | .142 |
| Establish a course of action to accomplish goals | **.496** | .195 | .386 | .223 | .204 | .112 |
| Evaluate and choose between alternative courses | **.485** | .175 | .176 | .397 | .087 | .067 |
| Communicate well orally | **.460** | .119 | .265 | .207 | .037 | .168 |
| Gain in-depth knowlegde of a field | **.442** | .074 | .080 | .039 | .160 | .117 |
| Develop awareness of social problems | .158 | **.732** | .233 | .091 | .102 | .099 |
| Develop feminist awarenenss | .082 | **.560** | .148 | .025 | -.037 | .158 |
| Identify moral and ethical issues | .288 | **.542** | .168 | .185 | .098 | .126 |
| Place current problems in historical prospective | .303 | **.433** | .040 | .166 | -.063 | .198 |
| Form close friendships | .122 | .163 | **.569** | .143 | .077 | .084 |
| Understand myself, my abilities, interests | .328 | .275 | **.547** | .159 | -.056 | .156 |
| Develop self-esteem /self-confidence | .370 | .313 | **.543** | .165 | .096 | .162 |
| Function independently without supervision | .372 | .100 | **.455** | .170 | .165 | .189 |
| Relate well to people of different races, nations | .125 | .335 | **.337** | .307 | .165 | .070 |
| Lead and supervise tasks and groups of people | .178 | .089 | .189 | **.724** | .149 | .048 |
| Function effectively as a member of a team | .160 | .190 | .208 | **.567** | .240 | .038 |
| Use computers for complex tasks (graphing) | .082 | -.034 | -.024 | .112 | **.601** | -.030 |
| Understand the role of science and technology | .380 | -.013 | .089 | .138 | **.571** | .164 |
| Use computers for basic tasks (word processing) | .122 | .118 | .247 | .108 | **.402** | .053 |
| Appreciate art, literature, music, drama | .178 | .248 | .141 | .088 | -.054 | **.650** |
| Acquire broad knowledge in the Arts and Sciences | .269 | .124 | .064 | .050 | .277 | **.413** |
| Read or speack a foreign language | .076 | .079 | .078 | .003 | .038 | **.402** |

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

# Six Underlying Constructs Resulting from our Rotated Factor Structure

**Rotated Factor Matrix[a]**

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Think analytically and logically | **.643** | | .091 | .006 | .118 | -.026 |
| Formulate creative / original ideas and solutions | **.625** | | .145 | .199 | .031 | .124 |
| Synthesize and integrate ideas and information | **.587** | .02 | .217 | .113 | .208 | .072 |
| Acquire new skills and knowledge on my own | **.581** | .070 | .141 | .090 | .095 | .146 |
| Plan and execute complex projects | **.537** | .042 | .127 | .148 | .230 | .168 |
| Write effectively | **.512** | .263 | | | | |
| Establish a course of action to accomplish goals | **.496** | .195 | | | | |
| Evaluate and choose between alternative courses | **.485** | .175 | | .397 | .087 | .067 |
| Communicate well orally | **.460** | .119 | | .207 | .037 | .168 |
| Gain in-depth knowlegde of a field | **.442** | .074 | .080 | .039 | .160 | .117 |
| Develop awareness of social problems | .158 | **.732** | .233 | | | |
| Develop feminist awarenenss | .082 | **.560** | .148 | | | |
| Identify moral and ethical issues | .288 | **.542** | .168 | | | |
| Place current problems in historical prospective | .303 | **.433** | .040 | | | |
| Form close friendships | .122 | .163 | **.569** | | .077 | .084 |
| Understand myself, my abilities, interests | .328 | .275 | **.547** | .159 | | |
| Develop self-esteem /self-confidence | .370 | .313 | **.543** | .165 | | |
| Function independently without supervision | .372 | .100 | **.455** | .170 | | .189 |
| Relate well to people of different races, nati | | | | .307 | .165 | .070 |
| Lead and supervise tasks and groups of pe | | | | **.724** | .149 | .048 |
| Function effectively as a member of a team | .160 | .190 | | **.567** | .240 | .038 |
| Use computers for complex tasks (graphing) | .082 | -.034 | -.024 | .112 | **.601** | -.030 |
| Understand the role of science and technology | .380 | -.013 | .089 | .138 | **.571** | .164 |
| Use computers for basic tasks (word processing) | .122 | .118 | .247 | .108 | **.402** | .053 |
| Appreciate art, literature, music, drama | .178 | .248 | .141 | .088 | -.054 | **.650** |
| Acquire broad knowledge in the Arts and Sciences | .269 | .124 | .064 | .050 | .277 | **.413** |
| Read or speack a foreign language | | | | | .038 | **.402** |

Callouts: Knowledge Gains · Social/Moral Reasoning · Self-Awareness/Independence · Leadership/Teamwork · Technology · Arts & Humanities

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

# Advanced Topics Related to EFA

- *Confirmatory* Factory Analysis
  - Testing hypotheses about factor structure
- Structural Equation Modeling
  - Testing structure on the "predictor" and "outcome" side of the equations
  - Path-Analysis-Like model fitting
  - Mediator/Moderator effects testing
  - Just to name a few examples!