Brief Outline of the Class

Sunthud Pornprasertmanit
Chulalongkorn University

Lecture1

What is statistics?
The concept of sample, population, parameter, descriptive statistics, inferential statistics and random sampling
Measurement scale
Measurement error
Concept of central tendency
Concept of dispersion
Objective of descriptive statistics

1) Describe sample
2) Parameter estimate (Estimation)
   - Valid
   - Low dispersion

Frequency distribution, polygon, histogram, stem and leaf display
Normal distribution
Standard score
Sample distribution, Population distribution, Sampling distribution
Central limit theorem
Inferential statistics

1) Hypothesis testing
2) Estimation (Confident interval)

Lecture2

Type I error and Type II error
Power (Practical significance)
Effect size
How effect size in population, sample size (or df), power and type I error related
Latent variable, manifest variable (predictor), variate, composite score and summate scale
Measurement model
Structural model
Mediation and Moderation (conditional relationship)
Variate, Linear combination
Specification error
Model parsimony
Concept of relationship and difference
Introduction to statistical techniques
Dependence techniques

- One sample t-test
- Two sample t-test
  - Independent
  - Dependent
- Correlation
- Simple regression

Author Note
This article was written in November 2006 as the class document in advanced statistics tutoring, Faculty of Psychology, Chulalongkorn University.
Correspondence to Sunthud Pornprasertmanit. Email: psunthud@gmail.com

- Difference in correlation test
- Difference in regression test
- One-way ANOVA
- Multiple comparison and priori contrast
- Repeated-measure ANOVA
- Factorial design in ANOVA
- Randomized block design ANOVA
- Multiple regression
- Multiple regression with dummy variables
- Multiple regression with polynomial terms
- Multiple regression with interaction terms
- Stepwise, forward, backward method
- Hierarchical stepwise regression
- ANCOVA

Lecture3

- Multiple discriminant analysis
- Logistic regression (Linear probability model)
- MANOVA
- Repeated-measure MANOVA
- Factorial design in MANOVA
- Canonical regression
- Conjoint analysis
- Structure equation modeling
- Hierarchical linear model
- Chi-square
- Proportion z-test

Interdependence techniques

- Factor analysis: Q-type, R-type; Exploratory, Confirmatory
- Cluster analysis
- Correspondence analysis
- Multidimensional scaling

Graphical display

Univariate: Histogram, Stem and leaf display
Bivariate: Scatterplot, Boxplot
Multivariate: Profile, Transformation, Iconic displays

How to clean data

Data entry error
Missing data
Outlier

Missing data

How missing data come from?
Impact of missing data
Reduction of sample size
Statistical bias
Understanding the process of missing data (e.g. are missing data random?) and cure their impact (delete case, imputation)
Four-step process in identify missing data
1) Is it ignorable missing data?
a. Ignorable missing data

i. Parameter
ii. Missing data from design that known
iii. Censored data

b. Not ignorable
i. Known and can control
ii. Known but cannot control
iii. Unknown

2) How much missing data?
a. If not much, skip to step 3
b. If much (more than 10%), may delete cases or variables

3) Diagnose the randomness of data

Missing at random (MAR) or Missing at completely random (MCAR); Test by Little's MCAR test or analyse all variables to explore missing data process

4) Imputation method
a. MAR – EM approach
b. MCAR
i. Only valid data; Listwise, pairwise
ii. Replacement value; Hot-cold deck imputation, case substitution, Mean substitution, Regression imputation, EM approach

SPSS – Missing value analysis

<u>Lecture 4</u>

Outlier

Influence of outlier

Effect on result of analysis (Both improve or prevent relationship)

Don't present population

Question that frequently asked: Do the outliers represent population?

Four classes of outliers

1) procedural error
2) extraordinary event: has explanation
3) extraordinary observation: has no explanation
4) unique in their combination (bivariate outlier or multivariate outlier)

Detecting outlier

Univariate: z-score, boxplot

Bivariate: Scatterplot (Confident eclipse), boxplot

Multivariate: Mahalanobis D^2 (Euclidian distance/df) dispersed by t-distribution

If found outlier, should find the process of outlier and decide whether retention or deletion

If deleted, less generalizability.

Maybe analysis both with outlier and without outlier

Testing assumption of statistical analysis

Why are there assumptions of statistical analysis?

For accuracy of statistical analysis

Explore more information from data

Concept of robustness

Assessing individual variable and variate

Four statistical assumptions

1) Normality: univariate normality, multivariate normality

How they affect if violated: made t or F test inaccuracy, make unequal error in prediction

How to detect

(1) graphical display: histogram, normal p-p plot

(2) statistical detection: skewness, kurtosis, Komogorov-Smirnov test

Large sample size reduce detrimental effects of nonnormality

Remedy: choose another statistical technique, data transformation

2) Homoscadasticity; Homogeneity of variance, Homogeneity of variance/covariance matrices

How they are occur: variable type, skewed distribution

How they affect if violated: unequal accuracy of prediction, made statistical testing too liberal or too conservative

How to test

(1) graphical display: scatterdiagram, boxplot

(2) statistical detection: Levene test, Hartley F_{max} test, Box's F test, Box's M test

Remedy: Data transformation, Weighted least square approximation

<u>Lecture 5</u>

3) Linearity

How they affect if violated: reduce true correlation b/w variables

How to detect: scatterdiagram, examine residual, nonlinear relationship

Remedy: Data transformation

4) Absence of correlated errors: No patter of error or unexplained systematic relationship exists in the dependent variable

How they affect if violated: specification error, inaccuracy in prediction, no maximum variance extracted

Example: Time-series data, group that not included in model

How to detect: examine residual

Remedy: correcting the specification error

Data transformation

From theory or data derived

Change in interpretation of results

Maybe analysis both transformed and not transformed

Should not transform dependent variable b/c difficult in interpretation

Dummy variable

Changing from nonmetric variable to metric variable

How to build dummy variable

1) indicator coding
2) effect coding
3) orthogonal coding

Inferential Statistics

T-test

One-sample t-test

Unknown population variance

Sample standard error
Student's t-distribution
Degree of freedom
Assumption of one-sample t-test
Effect size

Dependent t-test
Repeated measure
Match paired or randomized block design
Nested data
Difference score or gain score
Assumption
Effect size
Practice effect or carry over effect

Independent t-test
Each group drawn form different population
Experimental design: IV, DV, EV
Variance explained by IV in DV, Variance explained by covariate and Error variance in research design (Maxmincon principle), measurement (shared variance, unique variance and error variance) and statistics
Method of control nuisance variables
1) hold the nuisance variable constant for all subjects (e.g. experiment in animals, tape record controlling experimenter bias)
2) assign subjects randomly to experimental conditions: random assignment
3) include the variable as one of the factors in experimental design
4) statistical control: partial correlation, analysis of covariance, hierarchical method in multiple regression

Casual inference (X is a cause of Y) carry four requirements
1) X precedes Y in time
2) Some mechanism explained
3) Change in X is accompanied by change in Y
4) Effect X on Y cannot be explained by other variables

Homogeneity of variance
Degree of freedom
Effect size
Variance explained by group variable
Assumption

<u>Lecture 6</u>

ANOVA

One-way ANOVA
Spurious effect from t-test several times
Predicting observed score
- by a value (sample mean is the best value)
- by group mean

Error term and sum of square error
Deviation in observed score is explained by group deviation and deviation within group

SStotal is sum of SSgroup and SSwithin
Null hypothesis
If null hypothesis is true, the methods for estimate population variance can divided in two ways: by MSgroup and MSerror (Central limit theorem)
If null hypothesis is not true, however, MSgroup estimate sum of population variance and between group variance
Degree of freedom
F ratio: F-distribution
Critical value
F test is omnibus test
F is equal to t square
Assumption
Fixed effect and random effect
Intraclass correlation (eta square) and unbiased intraclass correlation (omega square: the more number of group, the less omega square)
(SSBG – (p – 1)MSWG)/(SSBG + p(m(n) – 1)MSWG + MSWG)
Multiple comparisons: control inflated type I error
Fisher's protected t-test (Least significance difference)
Tukey HSD
Bonferreni
Scheffe
Dunnett C
Contrast or priori contrast

Lecture 7

Factorial ANOVA
Example of interaction effect
Why not ANOVA 2 times?
Concept of moderator
General linear model
Explained observed score deviation
Explained SStotal
Degree of freedom
MSa, MSb, MSab, MSe
F-test: main effect, interaction effect
Multiple comparisons: main effect, interaction effect
Eta square
Higher order factorial design

Repeated-measure ANOVA
Dependent measure
Compare between source of measure (e.g. repeated-measure)
Xst = X.. + as + bt + abst + est
Assumption
1. Independent of observation
2. Homogeneity of variance
3. Normality
4. Sphericity – correlation (covariance) between pairs of level are equal (Mauchley's test of sphericity)
Huynh-Feldt; Greenhouse-Geisser

Epsilon describe the degree of departure form sphericity assumption
Multiple comparisons
Trend analysis
Mixed design ANOVA

Lecture 8

Correlation

Scatterplot
Measure of relationship
Linear relationship
Congruent of two-scale standard score (adjust of unit and variance)
$R = 1 - [Sum(z_x - z_y)^2/2(n - 1)]$
Another formula = Covxy/sxsy
Another formula = Sum(zxzy)/(n – 1)
Graphical illustration for covariance]
Pearson product moment correlation
Interpretation of correlation
1) statistical significance (t-test or z-test by Fisher z transformation or parameter estimation by z)
$z = 0.5\ln[(1 + r)/(1 - r)]$
$SEz = 1/\sqrt{n - 3}$
2) direction
3) magnitude
4) coefficient of determination
Reduced form of pearson product moment correlation
Point-biserial correlation
Phi coefficient
Spearman rank correlation
Independent testing for difference between correlation (Moderation)
$SEz1\text{-}z2 = \sqrt{[(1/(n1 - 3)) + (1/(n2 - 3))]}$
More than 2 groups
Zbar = Sum(wjzj)/Sum(wj)
Chi-square = sum(wjz2j) – sum(wj)square(zbar) [df = j – 1]
Wj = nj - 3
Range estimation is inverse varied by sample size
Range restriction
Correlation of attenuation
Heterogeneity of sample
Outlier
Biserial correlation
Part-whole correlation (bias estimate if autocorrelation)
Linearity

Simple Regression

In previous section, predicting observed score by a value and group mean
Regression is generalize method than ANOVA, take some knowledge of variable to predict observed score (In ANOVA, group mean can transform to dummy variables) by transforming that variable to variate
Predicted Y = a + bX

Or $Y = a + bX + e$

Predicted Y must have sum(e) = 0 and least sum(e square) [Ordinary Least square: OLS]

How do you know what line is the best estimate? (minimum SSerror) – *Supplementary reading 1*

Lecture 9

What do a (intercept) and b (slope) mean?

Objective of regression: Explanation casual effect and prediction

Functional and statistical relationship

If transform x and y to standard score, the intercept = 0 and slope = correlation coefficient.

Symmetric statistics and asymmetric statistics

The deviation in Y is explained by X and error, then

SSy or SStotal = SSregression + SSerror

R2 = SSregression/SStotal [Venn-Euler Diagram] [Coefficient of determination]

1 – R2 [Coefficient of indetermination]

R = correlation with predicted Y (derived from X) and observed Y

Standard error of estimate is the estimated population sd of the residuals

If no correlation, b = 0.

B depend on the variation of x.

Multiple R

Null hypothesis: r = 0, b = 0, a = 0

Different in two slope from different population

Precision on estimation or sensitivity of null hypothesis is affected by sample size

The relation between Power, Alpha, Sample size and Effect size in population

Dummy variable

Centering

Regression toward the mean

Sample size: equal to t-test

Assumption

1) Normality
2) Homoscadasticity
3) Linearity
4) Independent of error terms
5) No measurement error: esp. in DV; reduce in predictive effect size

Residual null plot detecting assumption

Objective of prediction

Standard error of predicted value, from formula, show that the nearer mean of X, the more precision in estimation

Concept of fixed independent variable and random dependent variable that affect generalizability

Lecture 10

Multiple Regression Analysis

Use of multiple independent variables

Predicted $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$

Or $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e$

Remark: Matrix analysis derivation in method of OLS, then

$\mathbf{SS_{pp}b = SS_{pc}}$

And $M(Y) = b_0 + b_1M(X_1) + b_2M(X_2) + \ldots + b_nM(X_n)$

Sum(e) = 0 and Sum(e^2) is minimum

What do intercept and slope mean? (Regression coefficient)

Specification error

1) include irrelevant variable: reduce model parsimony (reduce generalizability); mask the effect of useful variables (esp. sequential estimation)
2) not include relevant variable: bias the result (regression estimate) (esp. suppression effect, spurious effect), reduce predictive accuracy

What happens when transform all variables to standard scores? (Standardized regression coefficient)

The deviation in Y is explained by X and error, then

SSy or SStotal = SSregression + SSerror

R2 = SSregression/SStotal [Venn-Euler Diagram] [Coefficient of determination]

Adjusted R2

1 – R2 [Coefficient of indetermination]

R = correlation with predicted Y (derived from X) and observed Y

Standard error of estimate is the estimated population sd of the residuals

What happen when independent variables is uncorrelated?

Redundant of information from those independent variables: How to measure contribution of each variable (Venn-Euler diagram)

1) Simple correlation or zero-order correlation
2) Part correlation or semipartial correlation
3) Partial correlation

How to measure impact on Y: Standardized regression coefficient

Pattern of association between Y and two independent variables

1) direct and indirect effect
2) partial redundancy (sri or betai is smaller than ri) and full redundancy (sri or betai insignificant but ri significant)
3) suppression effect (change direction or increase magnitude in this direction) (when either ry1 or ry2 is less than the product of the other with r12 or when r12 is negative (assuming positive ry1 and ry2))
4) Spurious effect or entirely indirect effect (From full redundancy)

<u>Lecture 11</u>

Hypothesis testing

1) Correlation coefficient (F-test)
2) Regression coefficient (T-test) (can test partial and part correlation)

Standard error of b depend on sample size and its tolerance; the less tolerance, the less chance to reject null hypothesis. Maybe correlation coefficient is significant but all regression coefficient is not significant

The objective for prediction

Standard error and confident interval of each predicted Y from new case

Any discrepancy between the sample estimated regression coefficient and the population regression coefficients will result in larger errors in predicted Y when values are far from their mean that when they are close.

The standard deviation of each residual is affected by

1) standard error of estimate
2) absolute standard score of each IV
3) correlation between independent variables

Is not robust in statistical assumption

Nonnormality – residual is not symmetry (If positively skewed, the residual will be more in those above mean.

Heteroscadasticity – Not equal residual in any value of independent variable

Techniques for interpretation –

1) Loading for regression variate: reflect the variance share for regression variate esp. stepwise procedure, the variable that large contribution may be not include for explanation
2) Zero-order correlation: avoid misinterpretation in full redundancy and suppression effect (For researcher that not draw casual model)

Method for include independent variable for predicting dependent variable

1) Simultaneous analysis (Method Enter)
2) Computer sequential estimation: backward, forward, stepwise
   Problem: atheorical nature, lack of generalizability (In other population, the result may be reversed), multicollinearity consideration, maybe illogical result
   Advantage: maximum predictive accuracy with least IVs.
   Sample size should more than 40 cases for 1 variables
3) All possible combination
4) Hierarchical analysis (Priori sequential of analysis)

Sample size

Power and Sample size

Generalizability + concept of degree of freedom in generalizability

Assumption

1) Normality
2) Homoscadasticity
3) Linearity
4) Independent of error terms
5) No measurement error
6) Multicollinearity: reduce predictive power of each independent variable (more standard error)
   a. Correlation
   b. Tolerance (R_i^2) and Variance inflation factor ($1/\sqrt{R_i^2}$)
   c. Two-part process
      i. Principal component analysis to independent variables
      ii. Found eigenvalues
      iii. Calculate for condition index ($\sqrt{\lambda_i}/\sqrt{\lambda_{\min}}$): threshold is 30 or larger

iv. Find factor loading of each condition index (each eigenvalue: find 2 or more loading that is more than .90

Technique adapted for regression analysis

1) Dummy variable
2) Polynomials (beware of multicollinearity)
3) Interaction effect or moderator (beware of multicollinearity)

Lecture 12

ANCOVA (Only one covariate)

In One-way ANOVA, the variance of X are divided to between-group variance and within-group variance

The objective of ANCOVA is to control the extraneous metric variable (covariate)

The principle of ANCOVA is adjusting group mean that the extraneous variable are equal by the regression principle

$$X_{ij} = \overline{X}.. + (X_{i.} - \overline{X}..) + b(Z_{ij} - \overline{Z}..) + e$$

or $$X_{ij} = A + b_1 D_1 + b_2 D_2 + b_3 (Z_{ij} - \overline{Z}..) + e$$

In One-way ANCOVA, the variance of X partition to between-group variance, control variance and within-group variance (similar to partial correlation)

The variance explained if controlled covariate is SS_{group}/SS_{within}

Because of maximizing the variance explained, the good covariate should correlate with dependent variable and not correlate with independent variable.

Assumption that more than ANOVA: the regression coefficient of dependent variable on covariate is equal in all groups of independent variables.

If it is not equal slope, the difference of adjusted mean is not equal at every value of covariate.

SPSS: Test homogeneity of regression

```
MANOVA
  x6 BY x3(0,1) with x10
  /PRINT=SIGNIF(BRIEF)
  /ANALYSIS = x6
  /METHOD=SEQUENTIAL
  /DESIGN = x10 x3 x10 by x3 .
```

Factorial design with covariate

Unequal n size: Nonorthogonality between main effect and interaction effect: how to deal with

1. Experimental research: unequal n size is not naturalistic: Type III SS – Each main effect and interaction assessed after adjustment is made for all other main effects and interactions, as well as CV (Most conservative)

2. Survey research: unequal n size is naturalistic: Type I SS – Hierarchy of testing effect where main effects are adjusted for each other and for CVs, while interaction are adjusted for main effects, for CVs, and for same- and lower-level interactions.

Choose covariate that correlate with DV and uncorrelated with IV

If uncorrelate with DV, it may add no significant adjustment.

If correlate with IV, it may less variance explained because take off predictive variance.

If more than one covariate, use covariate that low autocorrelation, predictability to DV greatest.

How to create Eta square

Alternative to ANCOVA

1. Gain score (Problem: ceiling effect)
2. Randomized block design
3. Take CV to IV by blocking (Advantage: Nonlinearity, interaction (heterogeneity of regression)) (Disadvantage: Loss of information, unequal n size)

Hierarchical multiple regression analysis

Enter to regression analysis with prespecified sequence

Sequence is specified by logic of research: casual priority (logic of causality) and the removal of confounding variable (spurious relationship)

Select variables that less research relevance to last order: the statistical power is maximal because less effect to degree of freedom.

Select distal cause and next step select proximal cause to find mediating effect of distal cause. (Use when theory ambiguous) e.g. psychological cause → physiological reaction → behavior

Each variable in investigation should be entered only after variables that may be a source of spurious relationship have been entered i.e. gender/weight → force

Unique partitioning of the total variance accounted for by the k IVs may be made.

$$R^2_{y.123} = r^2_{Y.1} + sr^2_{2.1} + sr^2_{3.12}$$

Change sequence, change composition of variance

Significant test for sr value (increment variance)

Like ANCOVA

Technique for calculating path analysis

1. by correlation
   - R = DE + IE + SE + JE
   - Use correlation coefficient by data and by model
   - Examine just identified model
   - Do each regression equation to calculate regression coefficient
   - Plus DE + IE + SE to calculate correlation by model
   - Examine converge correlation matrix between data and model
2. by hierarchical regression analysis
   - Use full model regression (Hierarchical regression analysis) Use X and Last Y in first step
   - Insert variable that first mediated
   - Use regression coefficient in each step to calculate DE, IE and SE
   - Calculate correlation by model

Drawbacks

1. Use separate regression equation: LISREL use simultaneous analysis
2. No significant test of each effect
3. Time consumed
4. No recursive effect and latent variable

Set of multiple independent variables

Why use set of independent variables

1) Structural set – single construct that use two or more variables to represent
2) Functional set – logic of research i.e. demographic variables, group of antecedent variables, independent and control group

Both must theoretical oriented

Simultaneous and hierarchical analyze for sets

$$R^2_{Y.TUVW} = R^2_{YT} + R^2_{Y(U\cdot T)} + R^2_{Y(V\cdot UT)} + R^2_{Y(W\cdot VUT)}$$

e.g. adjusted mean for ANCOVA that has several covariates

Hierarchical analysis is more interpretable than simultaneous

The increment in Y variance accounted for by that set cannot be influenced by Y variance associated with subsequent sets

We are interested in contributed variance of each set: both all variance and unique variance [Venn-Euler diagram]

Part variance (squared part correlation)

Partial variance (squared partial correlation) – Eta square in ANCOVA

Regression coefficient of each variable in each set – beware of interpretation

Regression coefficient reflect the influence of a variable net of the influence of all other variable in equations

Maybe no independent variables is significant (High multicollinearity with in sets)

Maybe negative value in joint contribution to DV of two sets (that is suppression effect)

Incremental R square (F test)

Incremental adjusted R square may be negative

Additional topics in MRA

Less is more

The more variables, the more hypothesis tested

Increase chance of spurious significant

In each MRC, the greater the number of IVs, the lower the power of the test on each IV because 1) reduce df_{error} 2) power reduce from power formula 3) increase multicollinearity then increase SE_b and then power reduce

Increase n not solve this problem because the inflated type I error don't depend on n.

Solving problem: transforming variables to a few latent variables by factor analysis

When use a lot of IVs, it may be suppression effect that create difficulties in interpreting results

Least is last: least relevant variables should take into regression last.

Research question that use correlation and regression

Correlation provide information about interrelationship between variables

Regression provide casual effect of IV to DV

Technique: Use Stepwise+Hierarchical

Hierarchical Linear Model (Multilevel models)

Hierarchical Data (Nested design)

- Organizational research (e.g. workplace/worker, countries/household, classroom/student)
- Developmental research (e.g. observations/person)
- Meta-analysis (e.g. studies/subject)

Differences between cross-design and nested design

How to deal with hierarchical data in conventional research

- Disaggregate higher order variables
    - Violate assumption of independent of observations
    - Reduce variation because repeated value
    - Misestimated precision
- Aggregate lower order variables
    - Throw away within group information
    - Relations between aggregated variable are often stronger and they can be different from the relation b/w nonaggregate varibles (Aggregation bias)
    - Misestimated precision

Adapt from regression analysis; individual in different group can be independent but individual in same group share values on many variables; variance in each group may be alter

History

Variance component model (Method that partitioning variance from nested data)

Different regression model in each group

Intercept and slope are varied across groups → random-coefficient regression models

However random-coefficient regression models cannot incorporate higher level variables

Lindley & Smith (1972) – Bayesian estimation in linear model for dealing with unbalanced data

Dempster, Laird, & Rubin (1977) – EM Algorithm → Covariance estimation

Hierarchical Linear Model

Characteristics of Hierarchical Linear Model

Each level has each relationship among variables

Example: High School and Beyond (HS&B) Survey sample data from 160 schools. Each school has 45 students on average

DV = Math achievement

Student level IV = Socioeconomic status

Level 1 regression equation

$Y_i = \beta_0 + \beta_1 X_i + r_i$ (Overall equation)

$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$ (School j equation)

Meaningful intercept → centering SES by grand mean or group mean

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \overline{X}..) + r_{ij}$$

Centering grand mean: intercept means adjusted Y mean that X = $\overline{X}..$ in j school

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \overline{X}._j) + r_{ij}$$

Centering group mean: intercept means school j mean in Y

If SES is IV, then intercept if centering group mean means math achievement mean in each school but if centering by grand mean means math achievement mean adjusted for SES. (Effective)

Slope means effect of SES on math achievement in each school. (Equitable)

If random school from the population of schools, the effect of level-1 predictor to DV is

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \overline{X}..) + r_{ij}$$

Sources of Y variation in each school are 1) predictor and 2) error variance or variance of r_{ij} that is σ^2

Assumed that $r_{ij} \sim N(0, \sigma^2)$ that is 1) error term is normal distribution 2) mean of error term is equal to zero (Least square estimation) 3) homogeneous variance across schools.

Intercept and slope are varied among schools.

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

And $E(\beta_{0j}) = \gamma_{00}; E(\beta_{1j}) = \gamma_{10}; Var(u_{0j}) = Var(\beta_{0j}) = \tau_{00};$ $Var(u_{1j}) = Var(\beta_{1j}) = \tau_{11}$ and $Cov(u_{0j}, u_{1j}) = Cov(\beta_{0j}, \beta_{1j}) = \tau_{01}$

Remark: $\tau_{11} = 0$ is parallel of regression slope (assumption of ANCOVA)

Assumed that u_{0j} and u_{1j} is random variable (Multivariate normality) with zero means, variances τ_{00} and τ_{11} respectively, and covariance τ_{01}

Matrix form $\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(O, T)$ and $T = \begin{bmatrix} \tau_{00} & \tau_{10} \\ \tau_{01} & \tau_{11} \end{bmatrix}$

$\gamma_{00}, \tau_{00}, \tau_{01}$ meaning depends on centering in level 1 equation

- Group mean centering; γ_{00} means grand mean.
- Grand mean centering; γ_{00} means adjusted grand mean that partial X effect ($X_j = \overline{X}..$)
- No centering; γ_{00} means adjusted grand mean that partial X effect ($X_j = 0$)

But γ_{10}, τ_{11} has the same meaning in any centering

Population between mean and slope is

$$\rho(\beta_{0j}, \beta_{1j}) = \frac{\tau_{01}}{\sqrt{\tau_{00}\tau_{11}}}$$

This correlation explain how attribute of schools correlated (in this context, effective and equitable)

The level-2 predictor may add to explain the variability of intercept (effective) and slope (equitable)

This example School IV = Sector (Catholic = 1, Public = 0)

Level-2 equation

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}\text{j}$$

In these equation, reading γ_{00}, γ_{10} is depend on level-1 centering and level-2 centering

If there are level-2 predictor, the value of γ_{00}, γ_{10} are conditioned.

- No centering, γ_{00}, γ_{10} is conditioned if level-2 predictor equal to 0.
- Grand mean centering, γ_{00}, γ_{10} is conditioned if level-2 predictor is equal to predictor grand mean.

In these equation, assumed that

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(\mathcal{O}, T) \text{ and } T = \begin{bmatrix} \tau_{00} & \tau_{10} \\ \tau_{01} & \tau_{11} \end{bmatrix}$$

In these equations, estimation cannot produce because the outcomes (β_{oj}, β_{1j}) are not observed; therefore, mixed level-1 equation and level-2 equation.

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + \gamma_{10}(X_{ij} - \overline{X}._j) + \gamma_{11} W_j (X_{ij} - \overline{X}._j) + u_{0j} + u_{1j}(X_{ij} - \overline{X}._j) + r_{ij}$$

It is not linear model in OLS; OLS cannot estimate. Instead, iterative maximum likelihood procedures can estimate variance-covariance matrix (**T**), fixed effect (regression coefficient in level-2) and random level-1 coefficient (regression coefficient in level-1). (In this context, regression coefficient include both intercept and slope)

Submodel of HLM

1) One-way ANOVA with random effect
   - Level-1, level-2 and combine equation
   - Level-1 and Level-2 variance
   - Intraclass correlation
   - Error variance of $\overline{Y}._j$ in each group
   - Changing in mean of DV in each group consist of changing group and changing residual in each group
   - Supposed that we known level-1 variance and level-2 variance
   - Then, $Var(\overline{Y}._j) = Var(u_{0j}) + Var(\overline{r}._j)$;
     $\Delta_j = \tau_{00} + V_j = \tau_{00} + (\sigma^2 / n_j)$
   - Use the variation of each group mean to compute grand mean and then confident interval of grand mean
   - Parameter β_{0j} will be predicted by 1) mean of DV in each group (estimate parameter by OLS) and 2) grand mean
   - Because of two ways of prediction, Bayes estimation is an optimal weighted combination of these two.
   - $\beta^*_{0j} = \lambda_j \overline{Y}._j + (1 - \lambda_j)\hat{\gamma}_{00}$
   - $Reliability(\overline{Y}._j) = Var(\beta_{0j}) / Var(\overline{Y}._j)$ or $\lambda_j = \tau_{00} / (\tau_{00} + V_j)$
   - Because of it measures the ration of the true score or parameter variance, relative to the observed score or total variance of the sample mean, it names reliability.
   - The more parameter variance, the more depending on group mean but the more error variance, the more depending on grand mean because the group mean is not reliable.
   - Bayes estimate is biased toward grand mean (Then, it is called shrinkage estimator)
2) Means as outcome regression – Add level-2 predictor
   u_{0j} will be conditioned residual and τ_{00} will be condition variance
3)

Skip…

Exploratory Factor Analysis

Structural model and measurement model

Latent variable and indicator

Factor analysis purpose is to define the underlying structure among the variables in the analysis: grouped highly correlated variables together and make variables to variate that represent factor or underlying structure

Factor = a1x1 + a2x2 + … + anxn

Benefit: reduce variables to factor (underlying structure); therefore, lead to parsimonious model and solve the multicollinearity problem

If there are construct that want to measure but cannot measure by only one indicator, the factor analysis help to synthesize the multiple variables (that interrelated) representing construct to single or multiple factors.

Can summarize details (indicator) to a broader dimension

Exploratory and confirmatory approach (no assumption or confirm priori assumption)

7 Steps to do factor analysis

1) Research problem that match factor analysis – reduce variables to smaller set of factors with minimum loss of information

- Unit of analysis (R factor, Q factor, Cluster analysis)
- Data summarization or data reduction
- Variable selection (Conceptual underpinning; GIGO model)
- Using factor analysis with other multivariate techniques; solve the problem that other analysis may affect if there are a lot of variables

2) Design factor analysis