Module 1: Review of Analysis of Variance (ANOVA) and CTT

05/10/19

Review of ANOVA concepts

- Basic terms in ANOVA
- ANOVA models
- Fix and random effects
- Nested design

What is ANOVA?

ANOVA is a procedure to compare means by separating the variance associated with different factors

- Variation in scores
- Measuring variation
- Partitioning variation
- Analysis of variation and hypothesis testing

Variation in scores

- Why scores differ/vary?*
  - We purposely did something to create variation (treatment)
  - Individual differences
  - Human errors
  - Other factors
- Scores differences
  - There are differences (variability) among scores we expect to happen, but others are unexpected.
  - We can always attribute some portion of the differences observed among treatments to chance factors (e.g., human errors and individual differences).

*Based on Mathew Mitchell

Measuring variation

- You already know about Variance and SD…
- What measure of variation is calculated in ANOVA?
  - The *sum of squares*, denoted as SS
  - The sum of the squares of all deviation scores, e.g. calculating the SS total of the scores

					Row Totals	
A set of data	5	5	4	6	20	
	7	3	6	4	20	
	3	7	5	5	20	
Column Totals	15	15	15	15	60	Grand Total

Partitioning variation

- Again, calculating the SS for the raw and column
  - If there is a variation in columns or rows, the corresponding sum of squares is greater than 0.
- What does this have to do with SStotal?
  - To say that there is row or column variation in a matrix means that a score's deviation from the grand mean may partly due to the *column* to which it belongs (e.g., because that column tends to be higher than the grand mean), the *row* to which it belongs, or both.
  - This makes sense when rows and columns are referred to as *sources of variation* among scores (usually we call them as treatment effects).

Now, Analysis of Variation (ANOVA)

- What does this have to do with ANOVA?
  - SS can be used to estimate *systematic* variation due to treatment effects and *unsystematic* variation attributable to experimental error.
  - The relationship between treatment effects and experimental error can help you decide about whether H_{Null} is tenable.
- How?
  - If we form a *ratio* of these two estimates!

7

Examples of ANOVA models

- Oneway anova
- Random block anova
- Split-block anova
- Factorial anova
- ….

8

Analyzing cases (I)

Read each case below and determine its design and dependent variable.

- In a clinical experiment on the influence of dietary minerals on blood pressure, each of the 50 rats received three diets prepared with varying amounts of calcium (i.e., high, medium, and low) and with all other ingredients the same. Each diet lasted for 15 days. At the end of each 15-day period the blood pressure was measured.
- In a clinical experiment on the influence of dietary minerals on blood pressure, rats randomly receive one of the diets prepared with varying amounts of calcium and varying amounts of magnesium, but with all other ingredients of the diets the same. High, normal, and low values for each of the two minerals were selected for the study.

9

Analyzing cases (II)

- In order to compare three versions of reading tests, 300 high-school students were randomly assigned to take one of the versions: 100 students for version A, 100 students for version B, and 100 students for version C. The researcher compared the scores across the groups to determine the equivalency of the test versions.
- Sixty college students are introduced to a set of vocabulary words presented in the context of a lecture on physics, math, and history. Each student received all of the three lectures (videotapes in a random order). A vocabulary test with 60 words is administered after each lecture. The purpose of the experiment is to determine which content of the lecture produces better performance on the vocabulary test scores.

10

Analyzing cases (III)

- Researchers designed a study to investigate how test format affects students' math test scores. 80 10^{th} grade students, 40 low- and 40 high-performing, took a math test, including both short answer and multiple-choice items.
- Three types of treatments are evaluated for depressive patients: drugs, peer counseling, and yoga. Patients are matched based on their personality characteristics and level of depression. Eight blocks of 3 matched patients are formed. Patients within each block are randomly assigned to one of the three treatments for 3 months. At the end of the treatment, patients are evaluated with a clinical instrument to determine the level of depression.

11

Analyzing cases (IV)

- A study was conducted to compare different advertisement strategies. Subjects were randomly assigned to either text or video group. Each group reviewed ads varying in length (short or long) and evaluated these ads on a rating scale.
- In a learning opportunity study, researchers wanted to find out how teachers' teaching experience affects their students' performance across gender and grades. Teachers were categories as newcomer (1-2 years teaching experience) and veteran (more than 10 years). The experiment lasted for a quarter. Students at grades 3, 5, and 7 received pretest-and-posttest.

12

Random and fixed effects (i)

- Levels/treatments of the independent variable can be obtained in two ways:
  - Deliberately select them OR randomly sample them

13

Random and fixed effects (ii)

- Why & when to have random and fixed effects?
  - Characteristic of the variables
    - When there are small number of potential levels, that variable is always treated as a fixed variable.
    - When the number of treatment levels is small relative to all possible levels, then the variable is treated as a random variable (e.g., *subjects* are always treated as if they were sampled at random from a large population)
  - Theoretical inference – theoretically justified to be generalized to the same levels or a larger set of levels of IV

14

A one-way ANOVA example

- Fixed and random effects

15

Comparison

Fixed effect	Random effect
• Assuming the independent variable is fixed.	• Assuming the independent variable is random.
• The levels of the IV are thought to represent all (or most) possible values which researchers wish to generalize.	• The levels of the IV are thought to be a small set of the possible values which researchers wish to generalize.
• Probably produce smaller standard errors (more powerful).	• Probably produce larger standard errors (less powerful).

16

Random and fixed effects

- Another way to think about effects -- Replications

Fixed

Treatment Levels

Original study		a_1	a_2	a_3	a_4
Replication study	1	a_1	a_2	a_3	a_4
	2	a_1	a_2	a_3	a_4
	3	a_1	a_2	a_3	a_4

- Levels of the factor are selected arbitrarily and systematically **OR**
- Levels of the factor represent the entire population of treatment conditions in which the researcher has interest
- Statistical generalizations are limited to the treatment effects observed with these particular treatment conditions

Random

Treatment Levels

Original study		a_5	a_{15}	a_{38}	a_{65}
Replication study	1	a_8	a_{26}	a_{43}	a_{58}
	2	a_3	a_{23}	a_{35}	a_{69}
	3	a_9	a_{21}	a_{30}	a_{74}

- Levels of the factor are selected randomly or unsystematically from a larger pool of possible levels
- Levels are assumed to represent a random sample from the larger population of treatment conditions
- Statistical generalizations extends beyond those levels included in the experiment

So…
random and fixed variables

- Fixed variable
  - Assumed to be measured without measurement error.
  - The variable used in the study contains all or most values of the variable as it appears in the population it is drawn from.
  - The same values will be generalized to the population or other studies.
- Random variable
  - Assumed to be measured with measurement error.
  - The number of values in the study is small relative to the values of the variable as it appears in the population it is drawn from.
  - The values are intended to be generalized to a much large population of possible values with a certain probability distribution.

18

Nested design

- For a nested design, the factors are not independent to each other. That is, the levels of factor vary depending on the level of other factor.
- Nested design is used when:
  - Factors are dependent
  - Some treatments by nature are hieratically nested

19

Review of Classic Test Theory (CTT)

20

Main topics

- Assumptions in the classic test theory
- Conceptual understanding around applying the CCT

21

What is measurement error?

Scenario: We used an attitude survey measured students last Monday. Then we used the same tool to students again next week.

- What would you predict about students' scores?

22

The classical expression

$$X = T + E$$

X, observed scores – a variable, the observed test scores.

T, true score – a constant, the expected value or mean of the theoretical distribution of observed scores from a random variable that would be found in repeated independent testings of examinees with the same test.

For example, Annie's true score can be interpreted as the average of the observed scores obtained over an *infinite* number of repeated tests with the *same* test.

E, measurement error – a variable, the discrepancy between the observed and true scores. It refers to **unsystematic** or **random** errors. In CTT, the true score may include the systematic errors.

Core assumptions (I): if we look at a particular student j

- Observed score

  X_j

- True score

  $T_j = \overline{X_j}$

- Measurement error

  $E_j = X_j - T_j$

24

Core assumptions (II): if we look at parallel tests

- For two parallel tests, each examinee has the same true score on both forms and the error variance for the two forms are equal

  If T=T', and Var (E) = Var (E') , then the tests are called parallel tests.

25

Introduction of an example

This example allows us to review the two core assumptions on two previous slides. On next slide, you will be given some information about the scores on two parallel tests.

With the known information, can you figure out other scores? Why?

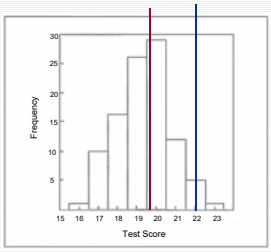
26

	Test A			Test A'		
	X	= T	+ E	X'	= T'	+ E'
Ada	9		1			
Jose	11				12	3
Taylor			4		9	
June	7					-2
...						
Total Variance	100					20

27

Another example – if Ada is tested for infinite times

Mean A score at a given time



Mean = 19.3
SD = 1.56

28

Applications of CTT

Focusing on the consistency of examinees' scores across items, tests, time, raters etc.

Question	Approach
Examining the consistency across the items in one test	Internal consistency, such as Cronbach's alpha coefficient
Examining the consistency between different forms of test	Reliability for parallel or equivalent forms
Examining the consistency between test and retest	Test-retest reliability
Examining the consistency between raters	Inter-rater reliability
Estimating the measurement errors caused by raters, forms, and/or items	

29

One example related to reliability

- We are interested in measuring the quality of teachers' feedback practices based on videotaped lessons and homework. Specifically we would like to explore the following measurement issues:
  - Is the quality same across written and oral?
  - Do raters code the videos (or student work) consistently?
  - How many pieces of videos (or student work) need to be sampled?
  - How should the lessons (or student work) be sampled, consecutive lessons or randomly from a unit?

30

Analyzing the example

- For each measurement issue, identify if it is a validity or reliability issue and describe how you would run the analysis.

31

Weaknesses of CTT

- Capture only a single source of the random measurement error
- Estimate the reliability for relative interpretations of test scores

32

G theory as extension of CTT

- Conceptualize the different sources of measurement errors as facets
- Capture both the systematic and unsystematic errors
- Determine the reliability for relative and absolute interpretations
- Implement the idea of ANOVA

33