Quantitative Research

Instructor: Nguyen Ngoc Vu, Ph.D.

Primary vs. Secondary data

Primary data

Information collected first hand by researchers

- Surveys
- Interviews
- Focus groups
- Questionnaires

Secondary data

Information already available

- Journals
- Books
- Census Data
- Newspaper articles
- Biographies

Quantitative data

- Deals with numbers.
- Data which can be measured.
- Length, height, area, volume, weight, speed, time, temperature, humidity, sound levels, cost, members, ages, etc.
- **Quantit**ative → **Quantit**y

3B quantitative data

- 48 students
- 42 girls, 6 boys
- 40% live in HCMC
- 25 learn French

Qualitative data

- Deals with descriptions.
- Data can be observed but not measured.
- Beliefs, preferences, colors, textures, smells, tastes, appearance, beauty, etc.
- **Qualit**ative → **Qualit**y

3B qualitative data

- friendly demeanors
- technology savvy
- environmentalists
- positive school spirit
- hard working students

Key concepts

- Population:

  The group to which the researcher would like the results of a study to be generalizable; it includes all individuals with certain specified characteristics.

- Sample:

  The group on which information is obtained.

Probability sampling

Any method of sampling that utilizes some form of *random selection*

Simple Random Sample

Obtaining a genuine random sample is difficult. We usually use Random Number Tables, and use the following procedure:

- Number the population from 0 to n
- Pick a random place I the number table
- Work in a random direction
- Organize numbers into the required number of digits (e.g. if the size of the population is 80, use 2 digits)
- Reject any numbers not applicable (in our example, numbers between 80 and 99)
- Continue until the required number of samples has been collected

Simple Random Sample

- **Advantages:**
  - The sample will be free from bias (i.e. it's random!)
- **Disadvantages:**
  - Difficult to obtain
  - Due to its very randomness, "freak" results can sometimes be obtained that are not representative of the population.

Systematic Sample

- Items are chosen from the population according to a fixed rule, e.g. every 10th house along a street.
- **Advantages:**
  - Can eliminate other sources of bias
- **Disadvantages:**
  - Can introduce bias where the pattern used for the samples coincides with a pattern in the population.

Stratified Sampling

- Population is broken down into categories, and a random sample is taken of each category.
- Proportions of the sample sizes are the same as the proportion of each category to the whole.
- **Advantages:**
  - Yields more accurate results than simple random sampling
  - Can show different tendencies within each category (e.g. men and women)
- **Disadvantages:**
  - Nothing major, hence it's used a lot

Cluster Sampling

- Used when populations can be broken down into many different categories, or clusters
- Rather than taking a sample from each cluster, a random selection of clusters is chosen to represent the whole.
- Within each cluster, a random sample is taken.
- **Advantages:**
  - Less expensive and time consuming than a fully random sample
  - Can show "regional" variations
- **Disadvantages:**
  - Not a genuine random sample
  - Likely to yield a biased result (especially if only a few clusters are sampled)

Non probability sampling

- Convenience sampling: Use who's available.
- Purposive sampling: Selection based on purpose.
- Modal instance sampling: Focus on 'typical' people.
- Expert sampling: Selecting 'experts' for opinion or study.
- Quota sampling: Keep going until the sample size is reached.

Non probability sampling

- Proportionate quota sampling: Balance across groups by population proportion.
- Non-proportionate quota sampling: Study a minimum number in each sub-group.
- Diversity sampling: Seeking variation with a wide net.
- Snowball sampling: Get sampled people to nominate others.
- Judgment sampling: Selecting what seems like a good enough sample.

What is a variable?

- An attribute of a person or an object which “varies” from person to person or from object to object

Functions of variables

- Independent variable
- Dependent variable
- Moderator variable
- Control variable
- Intervening variable

Independent (Experimental, Manipulated, Treatment, Grouping) Variable

- The major variable you hope to investigate
- Selected, manipulated and measured by the researcher

Dependent (Outcome) Variable

- The variable which you observe and measure to determine the effect of the independent variable

Practice: Identify DV & IV

- What is the relation between intelligence and achievement?
- Do students learn more from a supportive teacher or a non-supportive teacher?
- Are students aged 55 and older more likely to drop out of college than students of ages between 30 and 40?
- What is the relationship between grade point average and dropping out of high school?
- How do three counseling techniques—rational-emotive, gestalt, and no-counseling—differ in their effectiveness in decreasing test anxiety in high school juniors?
- What is the relationship among leadership skills, intelligence, and achievement motivation of high school seniors?

Moderator variable

- A special type of independent variable which you may select for study in order to investigate whether it modifies the relationship bet. Dependent & major independent variables.
- Ex: Effect of grammar practice on writing score => Gender/nationality/learning style are moderator variable

Control variable

- A variable which is held constant in order to neutralize the potential effect it may have on the outcomes.
- Ex: Experiment of NPK fertilizers on vegetables' growth => Water/temperature/location are control variables

Intervening variable

Hypothetical internal state that is used to explain relationships between observed variables, such as independent and dependent variables.

- Ex: motivation, tiredness, boredom, preference for the instructor's teaching style, learning style …

Extraneous variable

- Defined as any variable other than the independent variable that could cause a change in the dependent variable
- Extraneous variables are dangerous and may damage a study's validity

- EX: age, gender, family history, education of parents, time of the day …

Scales of measurement

Nominal Scale

- **Nominal**: Not a measure of quantity. Measures identity and difference. People either belong to a group or they do not.
- a.k.a. **categorical**, **taxonic**, **qualitative.**
- Examples:
  - Eye color: blue, brown, green, etc.
  - Biological sex (male or female)
  - Democrat, republican, green, libertarian, etc.
  - Married, single, divorced, widowed

Nominal Scale

- Sometimes numbers are used to designate category membership.
- Example:

  Country of Origin

  1 = United States 3 = Canada

  2 = Mexico 4 = Other

- Here, the numbers do not have numeric implications; they are simply convenient labels.


USA
10

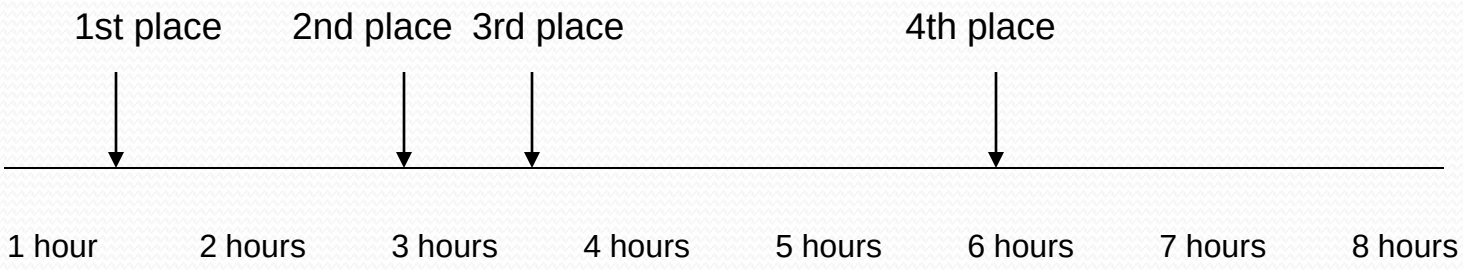

KING
21


Kings
16

Ordinal Scale

- **Ordinal**: Designates an ordering: greater than, less than.
- Does not assume that the intervals between numbers are equal.
- Example:

  finishing place in a race (first place, second place)



Ordinal Scale

- Ranking is also ordinal:
- Example: Rank your food preference where 1 = favorite food and 5 = least favorite:

____ sushi

____ hamburger

____ lau lau

____ chocolate

____ papaya

Interval Scale

- **Interval**: designates an equal-interval ordering.
- The difference in temperature between 20 degrees F and 25 degrees F is the same as the difference between 76 degrees F and 81 degrees F.
- Examples: Temperature in Fahrenheit or Celsius is interval. Common IQ tests are *assumed* to use an interval metric.

Interval Scale

- Likert scale: How do you feel about Stats?

  1 = I'm totally dreading this class!

  2 = I'd rather not take this class.

  3 = I feel neutral about this class.

  4 = I'm interested in this class.

  5 = I'm SO excited to take this class!

Ratio Scale

- **Ratio**: designates an equal-interval ordering with a true zero point (i.e., the zero implies an absence of the thing being measured).
- Examples:

  Measurements of heights of students in this class (Zero means complete lack of height).
  - Someone 6 ft tall is twice as tall as someone 3 feet tall.

Questionnaire

- A set of questions designed to generate the data necessary for accomplishing a research project's objectives

Major Issues

- What should be asked?
- How should each question be phrased?
- In what sequence should the questions be arranged?
- What questionnaire layout will best serve the research objectives?
- How should the questionnaire be pretested? Does the questionnaire need to be revised?

Question Form

- Nonstructured questions
  - Open-ended
- Structured questions
  - Fixed-response

Guidelines for writing good survey questions

Think about the form

- Don't write overly long questions
- Don't write unclear and ambiguous questions
- Don't write negative questions
- Don't write incomplete questions
- Don't write overlapping choices in questions
- Don't write questions across two pages

Think about the meaning

- Don’t write double-barreled questions
- Don’t write loaded questions
- Don’t write leading questions
- Don’t write prestige questions
- Don’t write embarrassing questions
- Don’t write biased questions

Think about the respondents

- Don’t use wrong level of language
- Don’t use questions that respondents may be unable to answer
- Don’t assume that everyone has an answer
- Don’t make respondents answer questions that don’t reply
- Don’t use irrelevant questions
- Don’t write superfluous information into questions

Descriptive statistics

Measures of central tendency

- Mean
- Median
- Mode

Mean

- Sum of the values divided by the number of cases

$$\bar{y} = \frac{\sum y_i}{n}$$

Calculating the mean for high temperatures

Date	High Temperature
2-Jan	59
3-Jan	60
4-Jan	43
5-Jan	42
6-Jan	35
7-Jan	32
8-Jan	32
9-Jan	46
10-Jan	41
11-Jan	52
Sum	442

- Add values

$$\sum y_i = 442$$

- Number of cases

$$n = 10$$

- Calculate mean

$$\bar{y} = \frac{\sum y_i}{n} = \frac{442}{10} = 44.2$$

Median

- The median represents the middle of the ordered sample data
- When the sample size is odd, the median is the middle value
- When the sample size is even, the median is the midpoint/mean of the two middle values

Calculating the median for high temperatures

Date	High Temperature	
7-Jan	32	
8-Jan	32	
6-Jan	35	
10-Jan	41	
5-Jan	42	**<===Middle values**
4-Jan	43	**<===Middle values**
9-Jan	46	
11-Jan	52	
2-Jan	59	
3-Jan	60	

$$median = \frac{42 + 43}{2} = 42.5$$

Mode

- The mode is the value that occurs most frequently
- It is the least useful (and least used) of the three measures of central tendency
- The mode may help to correct false impressions if you know the mean and the median but don't actually see the data.
- A set of data can be bimodal, multimodal or with no mode.

e.g. ***101 99 1 1***

The mean is (101 + 99 + 1 + 1)/4 = 202/4 = 50.5 and the median = (99+1)/2 = 50. But the mode here is ***1. In this case, the mean and median values are misleading.***

Calculating the mode for high temperatures

Date	High Temperature	
2-Jan	59	
3-Jan	60	
4-Jan	43	
5-Jan	42	
6-Jan	35	
7-Jan	32	**<===Mode**
8-Jan	32	**<===Mode**
9-Jan	46	
10-Jan	41	
11-Jan	52	

$$mode = 32$$

Measures of central tendency and levels of measurement

- Mean assumes numerical values and requires interval or ratio data
- Median requires ordering of values and can be used with ratio, interval and ordinal data
- Mode only involves determination of most common value and can be used with ratio, interval, ordinal, and nominal data

Comparison of mean and median

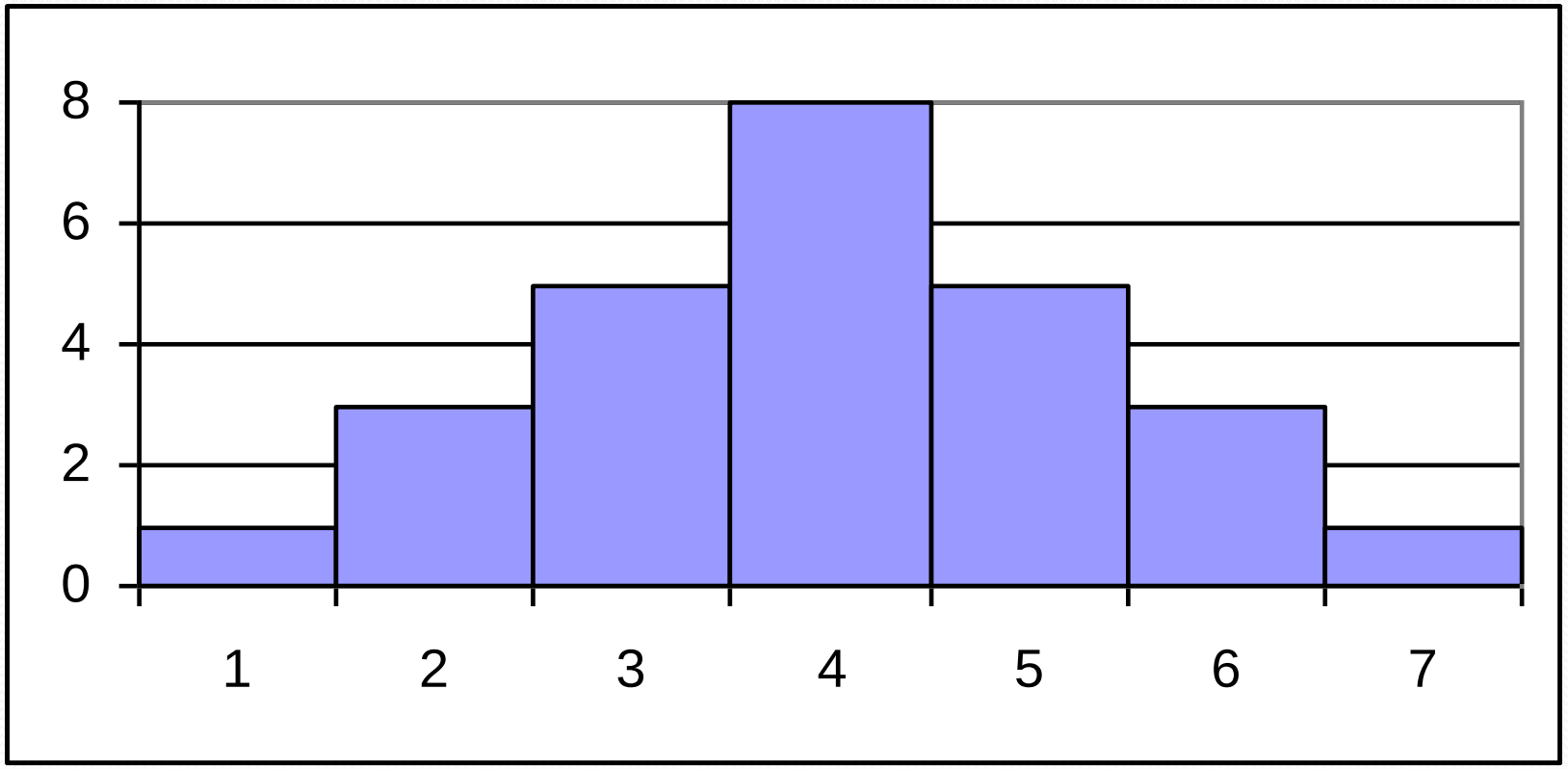
- Mean
  - Uses all of the data
  - Has desirable statistical properties
  - Affected by extreme high or low values (outliers)
  - May not best characterize skewed distributions
- Median
  - Not affected by outliers
  - May better characterize skewed distributions

The mean and median and the distribution of the data

- For symmetric distributions, the mean and the median are the same
- For skewed distributions, the mean lies in the direction of the skew (the longer tail) relative to the median

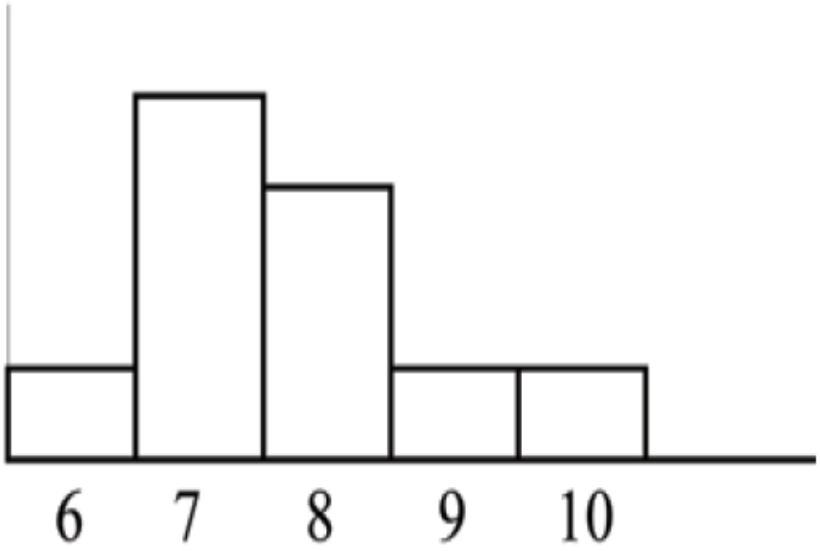
Distribution shapes

Symmetric: bell shaped



Positively skewed

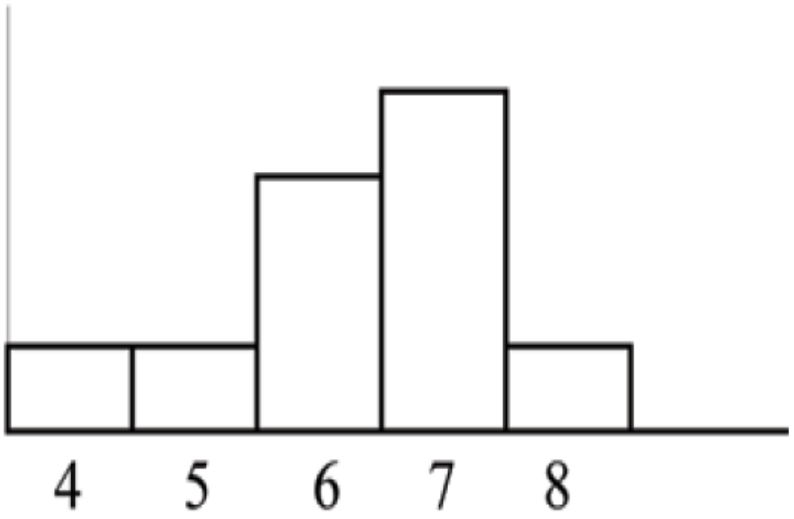
e.g.3 Distribution skewed to the right (Data set: 6 ; 7 ; 7 ; 7 ; 7 ; 7 ; 7 ; 8 ; 8 ; 8 ; 9 ; 10)



The mean is 7.7, the median is 7.5, and the mode is 7. *Notice that the mean is the largest statistic, while the mode is the smallest.* Again, the mean reflects the skewing the most. (Positively skewed)

Negatively skewed

e.g.2 Distribution skewed to the left (Data set: 4 ; 5 ; 6 ; 6 ; 6 ; 7 ; 7 ; 7 ; 7 ; 7 ; 7 ; 8)



The mean is 6.3, the median is 6.5, and the mode is 7. *Notice that the mean is less than the median and they are both less than the mode.* The mean and the median both reflect the skewing, but the mean more so. (Negatively skewed)

Measures of variation

- Range
- Variance and standard deviation
- Interquartile range

Range

- Range is the difference between the minimum and maximum values

Calculating the range for high temperatures

Date	High Temperature	
7-Jan	32	**<===Lowest Value**
8-Jan	32	
6-Jan	35	
10-Jan	41	
5-Jan	42	
4-Jan	43	
9-Jan	46	
11-Jan	52	
2-Jan	59	
3-Jan	60	**<===Highest Value**

$$range = 60 - 32 = 28$$

Variance and standard deviation

- The variance s^2 is the sum of the squared deviations from the mean divided by the number of cases minus 1

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

- The standard deviation s is the square root of the variance

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

Calculating the variance and standard deviation for high temperatures

Date	High Temperature	Difference X - mean	Difference Squared
2-Jan	59	14.80	219.04
3-Jan	60	15.80	249.64
4-Jan	43	-1.20	1.44
5-Jan	42	-2.20	4.84
6-Jan	35	-9.20	84.64
7-Jan	32	-12.20	148.84
8-Jan	32	-12.20	148.84
9-Jan	46	1.80	3.24
10-Jan	41	-3.20	10.24
11-Jan	52	7.80	60.84
Sum	442		931.60
n	10		
Mean	44.2		

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{931.60}{10-1} = 103.51 \qquad s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{103.51} = 10.2$$

Interpretation of standard deviation

- If distribution of data approximately bell shaped, then
  - About 68% of the data fall within one standard deviation of the mean
  - About 95% of the data fall within two standard deviations of the mean
  - Nearly all of the data fall within three standard deviations of the mean

Interquartile range

- Difference between upper (third) and lower (first) quartiles
- Quartiles divide data into four equal groups
  - Lower (first) quartile is 25^{th} percentile
  - Middle (second) quartile is 50^{th} percentile and is the median
  - Upper (third) quartile is 75^{th} percentile

Calculating the interquartile range for high temperatures

Date	High Temperature	
7-Jan	32	
8-Jan	32	
6-Jan	35	**<===Bottom Half Middle Value = First Quartile = 35**
10-Jan	41	
5-Jan	42	**<===Middle Value**
4-Jan	43	**<===Middle Value**
9-Jan	46	
11-Jan	52	**<===Top Half Middle Value = Third Quartile = 52**
2-Jan	59	
3-Jan	60	

Median = Second Quartile = 42.5

$$\textit{interquartile range} = 52 - 35 = 17$$

Interquartile range and outliers

- Value can be considered to be an outlier if it falls more than 1.5 times the interquartile range above the upper quartile or more than 1.5 times the range below the lower quarter
- Example for high temperatures
  - Interquartile range is 17
  - 1.5 times interquartile range is 25.5
  - Outliers would be values
    - Above 52 + 25.5 = 77.5 (none)
    - Below 25 – 25.5 = 9.5 (none)

Comparison of range, standard deviation, and interquartile range

- Sensitivity to extreme values
  - Range – extremely sensitive
  - Standard deviation – very sensitive
  - Interquartile range – not sensitive
- Standard deviation
  - Has desirable statistical properties
  - Suggests numbers of cases in different intervals for bell-shaped distributions

Correlational Research

Correlational Research

- The purpose of correlational research is to discover relationships between two or more variables.
- *Relationship* means that an individuals status on one variable tends to reflect his or her status on the other.

Correlational Research

- Helps us understand related events, conditions, and behaviors.
  - Is there a relationship between educational levels of farmers and crop yields?
- To make predictions of how one variable might predict another
  - Can high school grades be used to predict college grades?

Pearson Product-Moment Correlation

- Used when both the criterion and predictor variable contain continuous interval data such as test scores, years of experience, money, etc.

Examples of when to use the Pearson Correlation

Predictor Variable	Criterion Variable
Years of Experience in Extension	Job Satisfaction score
Family Income	End of Course (EOC) Test Scores
Distance from Krispy Kreme donut shop.	Weight

Suggested interpretation

- -1.0 to -0.7 strong negative association.
- -0.7 to -0.3 weak negative association.
- -0.3 to +0.3 little or no association.
- +0.3 to +0.7 weak positive association.
- +0.7 to +1.0 strong positive association.

Note: level for interpretation may vary

Types of quantitative designs

PRE-EXPERIMENTAL

1. One shot study - difficult to conclude anything - can't really attribute the performance to the treatment.
2. One group pretest-post test - can now see improvement but you don't know why (could be maturation, testing, history etc.)
3. Static group comparison - problem is that the groups are not equivalent at the beginning so hard to conclude the reason for any observed differences.

EXPERIMENTAL

4. Randomized groups design - is like #3 but with random groups. Controls for many of the threats to IV
5. Pretest - post test randomized group design - Useful if interested in the amount of change as a result of treatment. Possible reactive effects of testing is a threat to IV but should be evened out between groups. However reactive effects of testing remain a threat to EV.
6. Solomon four group design - can now evaluate the effects of testing in EV. But you need lots of subjects!

Quasi-experimental

7. Time series designs - can compare rates of changes over time, e.g. between O1 - O4 and O5 - O7 as well as differences between O4 - O5.
8. Reversal design - similar to above, allows insertion of treatment then withdrawal, then treatment.
9. Non-equivalent control groups - we attempt to find similar groups and use as controls. Is similar to design #5 but without the randomization.
10. Ex Post Facto design - Is like #3 but with no control over the treatment the groups received. For example, we could go back and look at records of students in several school districts and perform a statistical analysis.