BINUS UNIVERSITY

GREATER JAKARTA • BANDUNG • MALANG

EMPOWERING SOCIETY BUILDING THE NATION
YEARS OF BINUS

People
Innovation
Excellence

STAT6171001
Basic Statistics

Research Design
Session 6

Raymond Bahana
rbahana@binus.edu

Session Learning Outcomes

Upon completion of this session, students are expected to be able to

- LO 2. Analyze a problem by using the basic concept of descriptive and inferential statistics
- LO 3. Design a descriptive and inferential statistics solution to meet a given set of computing requirements in the context of computer science

Topic

- Research Design

Research Design

Scientific Method

Scientific Method

- It depends on empirical data
- Empirical evidence is information acquired by observation or experimentation. Scientists record and analyze this data.
- The process is a central part of the scientific method, leading to the proving or disproving of a hypothesis and our better understanding of the world as a result.
- To be a proper scientific investigation, the data must be collected systematically
- However, scientific investigation does not necessarily require experimentation

Scientific Method (Cont.)

- Theories and explanations are very important in science
- Theories in science can never be proved since one can never be 100% certain that a new empirical finding inconsistent with the theory will never be found.

- If a hypothesis is disconfirmed, then the theory from which the hypothesis was deduced is incorrect.
- If a hypothesis derived from a theory is confirmed, then the theory has survived a test, and it becomes more useful and better thought of by the researchers in the field

Scientific Method (Cont.)

- A key difference between scientific explanations and faith-based explanations is simply that faith-based explanations are based on faith and do not need to be testable

- An important attribute of a good scientific theory is that it is parsimonious.
- Parsimonious means the simplest model/theory with the least assumptions and variables but with greatest explanatory power.

Measurement

A good measurement scale should be both **reliable** and **valid**

Reliability

- Get at least approximately the same result if you measure something more than once
- A common way to define reliability is the correlation between parallel forms of a test.

Validity

- Validity is how researchers talk about the extent that results represent reality.
- Research methods, quantitative or qualitative, are methods of studying real phenomenon – validity refers to how much of that phenomenon they measure vs. how much “noise,” or unrelated information, is captured by the results.
- Research validity is broadly gathered into two groups: internal and external.

Validity

- The most common types of validity:
  - **Face validity**: A test that appears valid simply because of the appropriateness or relativity of the testing method, included information, or tools used.
  - **Content validity**: The determination that the measure used in research covers the full domain of the content.
  - **Construct validity**: The assessment of the suitability of the measurement tool to measure the activity being studied.

Validity

- The most common types of validity:
  - **Internal validity**: The assessment of how your research environment affects measurement results. This is where other factors can't explain the extent of an observed cause-and-effect response.
  - **External validity**: The extent to which the study will be accurate beyond the sample and the level to which it can be generalized in other settings, populations, and measures.
  - **Statistical conclusion validity**: The determination of whether a relationship exists between procedures and outcomes
  - **Criterion-related validity**: A measurement of the quality of your testing methods against a criterion measure (like a "gold standard" test) that is measured at the same time.

Face Validity

- A test's face validity refers to whether the test appears to measure what it is supposed to measure
- Example: a researcher may create a questionnaire that aims to measure depression levels in individuals. A colleague may then look over the questions and deem the questionnaire to be valid purely on face value. In other words, on its surface the questionnaire seems to be constructed in such a way that it’s a good tool to use to measure depression levels.
- Face validity is the **most informal** and **subjective** way to measure the validity of a test.

Face Validity

- How to Measure Face Validity
- In practice, we often measure face validity by asking multiple people to rate the validity of a test using a Likert scale.
- For example, the potential responses could be:
  1. The test is completely appropriate for measuring a certain construct.
  2. The test is mostly appropriate.
  3. The test is somewhat appropriate.
  4. The test is neither appropriate nor inappropriate.
  5. The test is somewhat inappropriate.
  6. The test is mostly inappropriate.
  7. The test is completely inappropriate.

Face Validity

- There are three potential groups of people who could provide ratings for the face validity of a test:
  1. People who take the test.
  2. People who work with the test in some way.
  3. Members of the general public who are interested in the test.
- Face validity is a highly informal way to measure validity, but it can be useful for quickly ruling out sub-par research practices and techniques.

Content validity

- The term content validity refers to how well a survey or test measures the construct that it sets out to measure.
- For example, suppose a professor wants to test the overall knowledge of his students in the subject of elementary statistics.
- His test would have content validity if:
  - The test covers every topic of elementary statistics that he taught in the class.
  - The test does not cover unrelated topics such as history, economics, biology, etc.
- A test lacks content validity if it doesn't cover all aspects of a construct, it sets out to measure or if it covers topics that are unrelated to the construct in any way.

Example Content validity

- In practice, content validity is often used to assess the validity of tests that assess content knowledge.
- Example 1: Statistics Final Exam
  - A final exam at the end of a semester for a statistics course will have content validity if it covers every topic discussed in the course and excludes all other irrelevant topics.
- Example 2: Pilot's License
  - An exam that tests whether or not individuals have enough knowledge to acquire their pilot's license will have content validity if it includes questions that cover every possible topic discussed in a pilot's course and exclude all other questions that aren't relevant for the license.

Construct Validity

- Construct validity is more difficult to define.
- In general, a test has construct validity if its pattern of correlations with other measures is in line with the construct it is purporting to measure
- Construct validity can be established by showing a test has both convergent and divergent validity (or discriminant validity)
- Convergent validity shows whether a test that is designed to assess a particular construct correlates with other tests that assess the same construct.
- A French vocabulary test would have high convergent validity if candidates who took the test would get similar scores on different French vocabulary tests.

Construct Validity

- Discriminant validity shows whether a test that is designed to measure a particular construct does not correlate with tests that measure different constructs.
- This is based on the idea that we wouldn't expect to see the same results from two tests that are meant to measure different things (e.g. a math test vs a spelling test).

Internal Validity

- Internal validity, also known as cause-and-effect validity, refers to how well the independent variables in a study were identified and controlled for.
- In other words, results in research occur due to the manipulation of the independent variable—not due to other factors.
- Internal validity measures if your research is sound.
- It’s the degree of confidence with which you can say the results from a study are conclusive and cannot be influenced or changed by other factors.
- No study is ever fully internally valid or invalid.

Example Internal Validity

- Consider a website test that tries to establish whether a hero **image** or **video** performs better for first-time visitors.
- If you establish internal validity, you can confidently use the results of your study to determine which option performs better for future use.
- In this scenario, several factors affect internal validity.
- For example, users who see the hero image may be different demographically from those seeing the video.
- They might visit the website at a different time or already have pre-existing brand awareness compared to the other group.

External Validity

- Internal validity means the results of the study are valid to the specific individuals or items that were a part of the study itself.
- Whether or not the results can generalize to other populations, settings, or situations depends on the study's external validity.
- In other words, external validity is about how well the study applies to the real world.

Statistical Conclusion Validity

- Statistical conclusion validity is the degree to which conclusions about the relationship among variables based on the data are correct or "reasonable".
- This began as being solely about whether the statistical conclusion about the relationship of the variables was correct, but now there is a movement towards moving to "reasonable" conclusions that use: quantitative, statistical, and qualitative data.
- Conclusion validity is only concerned with the question: Based on the data, is there a relationship or isn’t there?

Statistical Conclusion Validity

- Fundamentally, two types of errors can occur:
- type I (finding a difference or correlation when none exists)
- type II (finding no difference or correlation when one exists).

Criterion-Related Validity

- Criterion validity or concrete validity refers to a method of testing the correlation of a variable to a concrete outcome.
- Higher education institutions and employers use criterion validity testing to model an applicant's potential performance.
- Some organizations also use it to model retention rates.
- A properly designed test can predict future outcomes or performance when research has documented a strong correlation between two variables.
- The correlation coefficient, which ranges from -1.0 to +1.0, demonstrates the strength of the two correlated variables.

Types of Criterion Validity

- Two types of criterion validity exist:
  - **Predictive validity**: models the likelihood of an outcome
  - **Concurrent validity**: confirms whether one measure is equal or better than another accepted measure when testing the same thing at the same time

Predictive Validity

- Predictive validity (or empirical validity) refers to a test's ability to predict a relevant behaviour or refers to the extent that it’s valid to use the score on some scale or test to predict the value of some other variable in the future.
- **Example**, we might want to know how well some college entrance exam is able to predict the first semester GPA of students.
- To determine if predictive validity exists, we could use the following process:
  - Administer the college entrance exam to 1,000 high school seniors.
  - One year later, collect data on the first semester GPA of the same 1,000 students.
  - Calculate the correlation between the scores on the entrance exam and the first semester GPA.

Predictive Validity

- If there is a high correlation between scores on the entrance exam and the first semester GPA, it's likely that there is predictive validity between these two variables.
- In other words, the score that a student receives on this particular college entrance exam is predictive of the GPA they're likely to receive during their first semester in college.

Concurrent Validity

- A test that uses concurrent validity tests the same criterion as another test.
- To ensure the accuracy of your test, you administer it as well as an already accepted test, which is scientifically proven to measure the same construct.
- By comparing the results of both tests, you can determine whether the one you have developed accurately measures the variable you're interested in.
- This type of criterion validity test is used in the fields of: Social science, Psychology and Education

Example Concurrent Validity

- If a psychologist developed a new psychological test for measuring depression called the Winters Depression Inventory, they'd need to test its validity before using it in a clinical setting.
- They'd recruit non-patients to take both inventories—their new one and a commonly accepted, such as the Beck Depression Inventory.
- The sample group would take both inventories under controlled conditions.
- They would then compare the 2 test results for each member of the sample group.

Example Concurrent Validity

- This process determines that the test they developed measures the same criterion at least as well as the accepted gold standard.
- In statistical terminology, when the results of the sample population's WDI and BDI match or are close to each other, they're said to have a high positive correlation.
- In this scenario, the psychologist has established the concurrent validity of the two inventories.

Basics of Data Collection

Basics of Data Collection

- Most statistical analyses require that your data be in numerical rather than verbal form
- Verbal form must be coded (represented by numbers)

Student Name	Hair Color	Gender	Major	Height	Computer Experience
Norma	Brown	Female	Psychology	5'4"	Lots
Amber	Blonde	Female	Social Science	5'7"	Very little
Paul	Blonde	Male	History	6'1"	Moderate
Christopher	Black	Male	Biology	5'10"	Lots
Sonya	Brown	Female	Psychology	5'4"	Little

Example Data

Basics of Data Collection (Cont.)

Student Name	Hair Color	Gender	Major	Height	Computer Experience
Norma	Brown	Female	Psychology	5'4"	Lots
Amber	Blonde	Female	Social Science	5'7"	Very little
Paul	Blonde	Male	History	6'1"	Moderate
Christopher	Black	Male	Biology	5'10"	Lots
Sonya	Brown	Female	Psychology	5'4"	Little

Example Data

One solution is to change all the numbers to inches.
So, 5'4" becomes (5 x 12) + 4 = 64,

Basics of Data Collection (Cont.)

Student Name	Hair Color	Gender	Major	Height	Computer Experience
Norma	Brown	Female	Psychology	5'4"	Lots
Amber	Blonde	Female	Social Science	5'7"	Very little
Paul	Blonde	Male	History	6'1"	Moderate
Christopher	Black	Male	Biology	5'10"	Lots
Sonya	Brown	Female	Psychology	5'4"	Little

Example Data

1	2	3	4	5
Very Little	Little	Moderate	Lots	Very Lots

Conversion of verbal descriptions to numbers

Sampling Bias

Sampling Bias

- Sampling bias refers to the method of sampling, not the sample itself
- No guarantee that random sampling will result in a sample representative of the population

Self-Selection Bias

- Self-selection bias can result when the non-random component occurs after the potential subject has enlisted in the experiment
- An online survey about computer use is likely to attract people more interested in technology than is typical.
- In this example, people who “self-select” themselves for the experiment are likely to differ in important ways from the population the experimenter wishes to draw conclusions about.

Undercoverage Bias

- A common type of sampling bias is to sample too few observations from a segment of the population.
- Often happens when a large significant entity goes unselected or has zero chance of getting in your representing sample.

Survivorship Bias

- The observations recorded at the end of the investigation are a non-random set of those present at the beginning of the investigation
- Survivorship bias (or survivor bias) is a cognitive fallacy (a mistaken belief) in which, when looking at a given group, you focus only on examples of successful individuals (the “survivors”) in the selection process rather than the group as a whole (including the “non-survivors”).

Experimental Designs

Experimental Designs

There are many ways an experiment can be designed:

- Basic experimental designs
  - Between-Subjects Designs
  - Within-Subjects Designs
  - Complex Designs

Between-Subjects Designs

- A between-subjects study design is one way that researchers can assign test participants to different treatment groups.
- In this design, each subject is assigned to only one treatment condition, and researchers will compare group differences between participants in these various conditions.
- The goal of a between-subjects study design is to enable researchers to determine if one treatment condition is superior to another.
- Because each subject is assigned to only one condition, this type of design requires a large sample. Thus, these studies also require more resources and budgeting to recruit participants and administer the experiments.

Within-Subjects Designs

- A within-subject design is a type of experimental design in which all participants are exposed to every treatment or condition. It is also known as a repeated measures design.
- One of the most significant benefits of this type of experimental design is that it does not require a large pool of participants.
- Fatigue is another potential drawback of using a within-subject design.
- Participants may become exhausted, bored, or less motivated after taking part in multiple treatments or tests.

Complex Designs

- Designs can contain combinations of between-subject and within-subject variables.
- For example, the “Weapons and Aggression” case study has one between-subject variable (gender) and two within-subject variables (the type of priming word - the act of making something ready - and the type of word to be responded to).

References

- Witte, R.S.&Witte, J.S. (2017). Statistics (11th ed.). Wiley. ISBN: 978-1119386056.
- Lane, D.M., Scott, D., Hebl, M., Guerra, R., Osherson, D.& Zimmer, H. (2003). Introduction to Statistics. Online edition at https://open.umn.edu/opentextbooks/textbooks/459
- Levine, D.M., Stephan, D.F. & Szabat, K.A. (2017). Statistics for Managers Using Microsoft Excel (8th ed.). Pearson. ISBN: 978-0134566672

Thank you