Clinical Research Methods: Chapters 4 & 5

Planning the Measurements & Getting Ready to Estimate Sample Size

Dr. Alexander Weber

Assistant Professor, Department of Pediatrics,
Division of Neurology, Faculty of Medicine
Imaging Staff Scientist, BC Children's Hospital Research Institute
University of British Columbia

February 11, 2020

Chapter 4: Planning the Measurements: Precision and Accuracy

Measurement Scale:

Categorical

Not quantifiable measurement; classified by placing into *categories*

- Dichotomous: 0 or 1; ex: dead or alive
- Nominal: unordered categories
- Ordinal: ordered with non-quantifiable intervals

Chapter 4: Planning the Measurements: Precision and Accuracy

Measurement Scale:

Categorical

Not quantifiable measurement; classified by placing into *categories*

- Dichotomous: 0 or 1; ex: dead or alive
- Nominal: unordered categories
- Ordinal: ordered with non-quantifiable intervals

Continuous

Quantified on infinite scale

- Discrete: limited to integers

Measurement Scale

TABLE 4.1 Measurement Scales

Type of Measurement	Characteristics of Variable	Example	Descriptive Statistics	Information Content
Categorical*				
Nominal	Unordered categories	Sex; blood type; vital status	Counts, proportions	Lower
Ordinal	Ordered categories with intervals that are not quantifiable	Degree of pain	In addition to the above: medians	Intermediate
Continuous or ordered discrete†	Ranked spectrum with quantifiable intervals	Weight; number of cigarettes/day	In addition to the above: means, standard deviations	Higher

* Categorical measurements that contain only two classes (e.g., sex) are termed **dichotomous.**

† Continuous variables have an infinite number of values (e.g., weight), whereas discrete variables are limited to integers (e.g., number of cigarettes/day). Discrete variables that are ordered (e.g., arranged in sequence from few to many) and that have a large number of possible values resemble continuous variables for practical purposes of measurement and analysis.

Precision

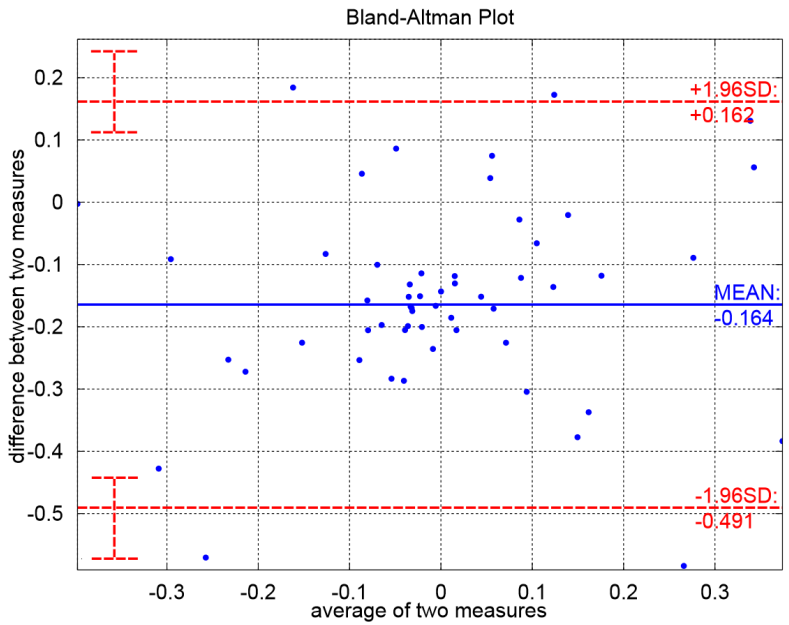
Precision

Degree to which a variable is reproducible, with same value each time

- Function of Random Errors: Chance variability
- Three types of random errors:
  1. Observer Variability: error due to observer (e.g. skill)
  2. Instrument Variability: error due to environmental factors (e.g. temp.; aging mechanical component)
  3. Subject Variability: error is subject biology
- Assess precision by testing reproducibility:
  - within-observer reproducibility
  - between-observer reproducibility

Reproducibility

- Reproducibility of continuous variables = within-subjet standard deviation
- If a Bland-Altman plot shows linear association of within-subject SD vs subject mean, then coefficient of variation is preferred (within subject SD divided by mean)



- For categorical variables, percent agreement and kappa statistic often used

How to Enhance Precision

Five approaches to enhance precision:

1. Standardize measurement method; e.g. written directions
2. Training and certifying the observers
3. Refining the instruments
4. Automating
5. Repetition: repeat the measurements

Strategies for Enhancing Precision

TABLE 4.2 Strategies for Reducing Random Error in Order to Increase Precision, with Illustrations from a Study of Antihypertensive Treatment

Strategy to Reduce Random Error	Source of Random Error	Example of Random Error	Example of Strategy to Prevent the Error
1. Standardizing the measurement methods in an operations manual	Observer	Variation in blood pressure (BP) measurement due to variable rate of cuff deflation (sometimes faster than 2 mm Hg/second and sometimes slower)	Specify that the cuff be deflated at 2 mm Hg/second
	Subject	Variation in BP due to variable length of quiet sitting	Specify that subject sit in a quiet room for 5 minutes before BP measurement
2. Training and certifying the observer	Observer	Variation in BP due to variable observer technique	Train observer in standard techniques
3. Refining the instrument	Instrument and observer	Variation in BP due to digit preference (e.g., the tendency to round number to a multiple of 5)	Design instrument that conceals BP reading until after it has been recorded
4. Automating the instrument	Observer	Variation in BP due to variable observer technique	Use automatic BP measuring device
	Subject	Variation in BP due to emotional reaction to observer by subject	Use automatic BP measuring device
5. Repeating the measurement	Observer, subject, and instrument	All measurements and all sources of variation	Use mean of two or more BP measurements

Accuracy

Accuracy

Degree to which a variable represents what it claims to represent

Validity

Degree to which observed findings lead to correct inferences about real world

Three type of systemic error:

1. Observer bias: distortion: conscious or unconscious
2. Instrument bias: faulty function of instrument
3. Subject bias: distortion by subject

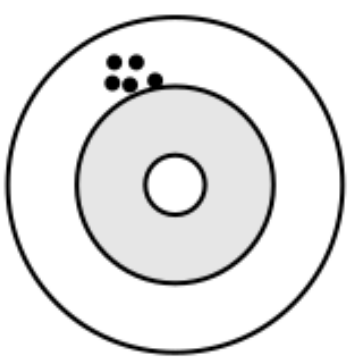
Accuracy best assessed by comparing to gold standard

Precision and Accuracy of Measurements

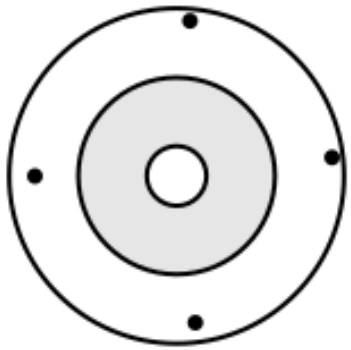
TABLE 4.3 The Precision and Accuracy of Measurements

	Precision	**Accuracy**
Definition	The degree to which a variable has nearly the same value when measured several times	The degree to which a variable actually represents what it is supposed to represent
Best way to assess	Comparison among repeated measures	Comparison with a reference standard
Value to study	Increase power to detect effects	Increase validity of conclusions
Threatened by	Random error (chance) contributed by The observer The subject The instrument	Systematic error (bias) contributed by The observer The subject The instrument

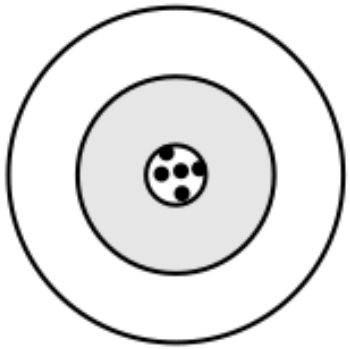
Difference Between Precision and Accuracy



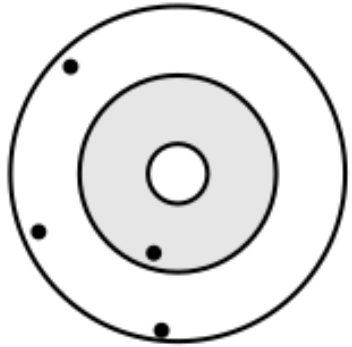
Good precision
Poor accuracy



Poor precision
Good accuracy



Good precision
Good accuracy



Poor precision
Poor accuracy

Validity

Three types:

Content Validity

How well the assessment represents all aspects of the phenomena under study; is the test fully representative of what it aims to measure?

Construct Validity

How well the assessment conforms to theoretical constructs; the degree to which a test measures what it claims, or purports, to be measuring

Criterion Validity

How well assessment correlates to well accepted existing measures (e.g. predictive validity); do the results correspond to a different test of the same thing?

Validity

Content Validity Example

A mathematics teacher develops an end-of-semester algebra test for her class. The test should cover every form of algebra that was taught in the class. If some types of algebra are left out, then the results may not be an accurate indication of students' understanding of the subject. Similarly, if she includes questions that are not related to algebra, the results are no longer a valid measure of algebra knowledge.

Validity

Construct Validity Example

There is no objective, observable entity called "depression" that we can measure directly. But based on existing psychological research and theory, we can measure depression based on a collection of symptoms and indicators, such as low self-confidence and low energy levels.

Validity

Criterion Validity Example

A university professor creates a new test to measure applicants' English writing ability. To assess how well the test really does measure students' writing ability, she finds an existing test that is considered a valid measurement of English writing ability, and compares the results when the same group of students take both tests. If the outcomes are very similar, the new test has a high criterion validity.

Strategies for Enhancing Accuracy

1. Standardize measurement methods
2. Training
3. Refining instruments
4. Automating instruments
5. Making unobstructive measurements: eliminate possibility of conscious bias
6. Calibrate instrument
7. Blinding

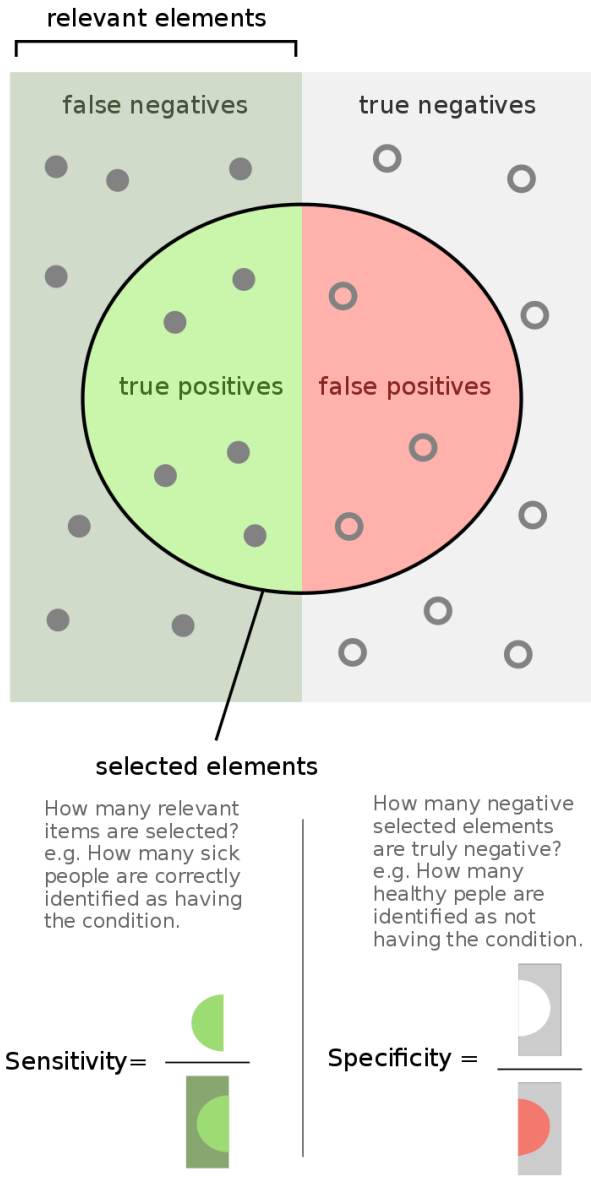
Other features

Individual measurements should be:

- Sensitive
- Specific
- Appropriate
- Objective - reduce involvement of observer
- Produce a range of values

In the aggregate, measurements should be:

- Broad but parsimonious
- Serve research question at moderate cost in time and money



Chapter 5: Getting Ready to Estimate Sample Size: Hypotheses and Underlying Principles

Goal of sample size planning is to estimate the appropriate number of subjects for a given study design

Should estimate early, as calculations often reveal that the research design is actually not feasible, or that different predictor or outcome variables are needed

Hypotheses

Research Hypothesis

Specific version of the research question that summarizes the main elements of the study in a form that establishes the basis for tests of statistical significance

Characteristics of a good hypothesis:

- Simple versus complex: e.g. one predictor and one outcome variable
- Specific versus vague: no ambiguity; clear; concise
- In-Advance versus After-the-Fact: hypothesis should be stated in writing at the outset of the study. Hypotheses that are formed after examining the data can lead to over-interpreting the importance of the findings (false positives)

Types of Hypotheses

Essentially, statistical tests help to estimate the probability that an association observed in a study is due to chance

- **Null Hypothesis:** states that there is no difference between groups, or no association between the predictor and the outcome variable
- **Alternative Hypothesis:** states that there is an association between the predictor and outcome variable. It cannot be tested directly
- **One and two-sided alternative hypotheses:** One-sided specifies the direction of the association between the predictor and outcome variables. A two sided hypothesis states only that an association exists

Example

β-carotene therapy was shown to reduce risk of lung cancer in initial one-side hypothesis test. Later results from well-done trials showed the opposite was in fact true

Nearly all hypotheses should (probably) be two-sided

Underlying Statistical Principles

TABLE 5.1 The Analogy between Jury Decisions and Statistical Tests

Jury Decision	Statistical Test
Innocence: The defendant did not counterfeit money.	**Null hypothesis:** There is no association between dietary carotene and the incidence of colon cancer in the population.
Guilt: The defendant did counterfeit money.	**Alternative hypothesis:** There is an association between dietary carotene and the incidence of colon cancer.
Standard for rejecting innocence: Beyond a reasonable doubt.	**Standard for rejecting null hypothesis:** Level of statistical significance (α).
Correct judgment: Convict a counterfeiter.	**Correct inference:** Conclude that there is an association between dietary carotene and colon cancer when one does exist in the population.
Correct judgment: Acquit an innocent person.	**Correct inference:** Conclude that there is no association between carotene and colon cancer when one does not exist.
Incorrect judgment: Convict an innocent person.	**Incorrect inference (type I error):** Conclude that there is an association between dietary carotene and colon cancer when there actually is none.
Incorrect judgment: Acquit a counterfeiter.	**Incorrect inference (type II error):** Conclude that there is no association between dietary carotene and colon cancer when there actually is one.

Type I and II Errors

Type I Error

False positive; if a null hypothesis is rejected when it is actually true

Type II Error

False negative; if a null hypothesis is accepted when it is actually false

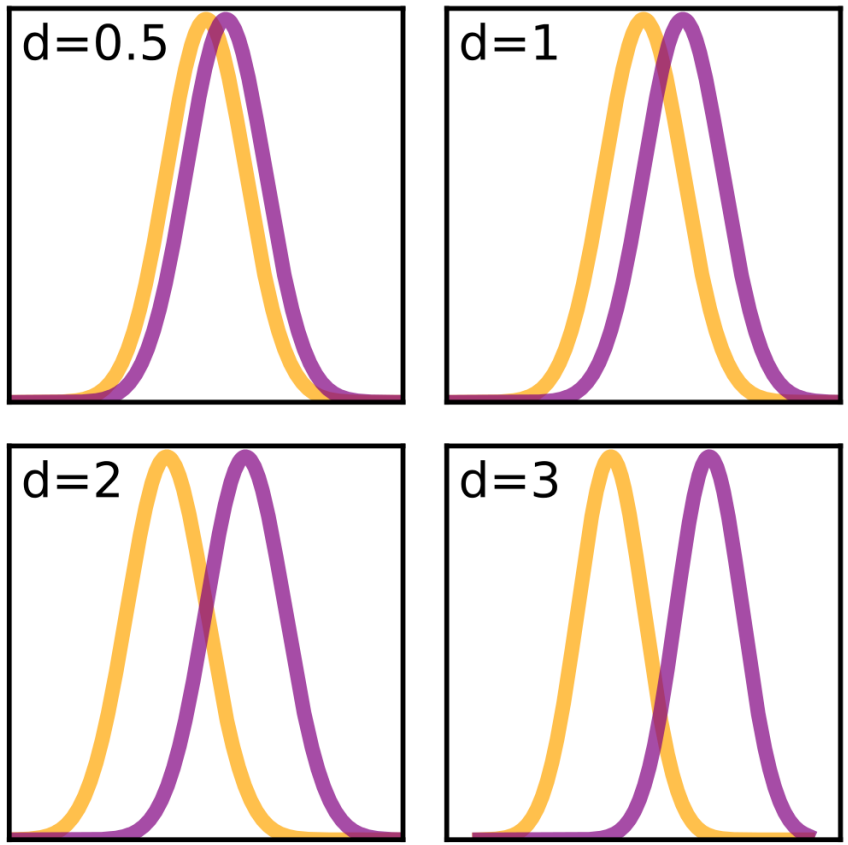
Effect Size

Effect Size

Magnitude of a phenomenon

Cohen's d:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$



Effect Size

Selecting an appropriate effect size is the most difficult aspect of sample size planning

Two options:

1. Look for previous studies that may give an indication
2. Perform a pilot study
3. Choose the smallest effect size that you believe would be clinically meaningful

If there are several hypotheses of similar importance, then the sample size for the study should be based on whichever hypothesis needs the largest sample

α, β and Power

Depending on whether the null hypothesis is true or false in the target population, and assuming that the study is free of bias, four situations are possible

TABLE 5.2 Truth in the Population versus the Results in the Study Sample: The Four Possibilities

	Truth in the Population	
Results in the Study Sample	**Association Between Predictor and Outcome**	**No Association Between Predictor and Outcome**
Reject null hypothesis	Correct	Type I error
Fail to reject null hypothesis	Type II error	Correct

The probability of committing a type I error (rejecting the null hypothesis when it is actually true) is called α (**alpha**)

Another name for α is the **level of statistical significance**

α, β and Power

The probability of making a type II error (failing to reject the null hypothesis when it is actually false) is called β (**beta**)

The quantity $[1 - \beta]$ is called **power**, the probability of correctly rejecting the null hypothesis in the sample if the actual effect in the population is equal to (or greater than) the effect size

Why not just set α and β to zero?

Sample size would then need to include the entire population

Many studies set α at 0.05 and β at 0.20 (a power of 0.80)

P value

P value

probability of obtaining the observed results of a test, assuming that the null hypothesis is correct

A "nonsignificant" result (i.e., one with a P value greater than α) does not mean that there is no association in the population;

it only means that the result observed in the sample is small compared with what could have occurred by chance alone

Multiple and Post Hoc Hypotheses

Multiple Comparisons Problem

When more than one hypothesis is tested in a study, especially if some of those hypotheses were formulated after the data were analyzed (post hoc hypotheses), the likelihood that at least one will achieve statistical significance on the basis of chance alone increases

Example

For example, if 20 independent hypotheses are tested at an α of 0.05, the likelihood is substantial (64%; $[1 - 0.95^{20}]$) that at least one hypothesis will be statistically significant by chance alone.