
Clinical Trials

Part 4

2002 VA DISTANCE LEARNING/CYBER SESSION SHORT COURSE IN CLINICAL TRIALS

WILLIAM G. HENDERSON, PH.D.
VISITING PROFESSOR UNIVERSITY OF
COLORADO HEALTH OUTCOMES
PROGRAM

STATISTICAL INFERENCE, SAMPLE SIZE, AND STATISTICAL POWER

- I. STATISTICAL INFERENCE
 - A. DEFINITIONS
 - B. TYPES
 - 1. ESTIMATION
 - 2. HYPOTHESIS TESTING
 - 3. PROS AND CONS
- II. SAMPLE SIZE DETERMINATION
 - A. PURPOSE
 - B. ELEMENTS OF SAMPLE SIZE
 - C. TYPE I AND TYPE II ERROR
 - D. EXAMPLES
 - E. SOME SAMPLE SIZE EQUATIONS
 - F. SAMPLE SIZE PROTOCOL PRESENTATION

DEFINITIONS

- *Statistical Inference* – inference (drawing conclusions) to a general population based on sample data or estimates from that population; broadly, any generalization, prediction, estimate or decision based on a sampling.

4

DEFINITIONS

- *Population* – the total number of persons, entities, objects or units in a defined area or setting for which we have an interest at a particular time.
- *Sample* – a finite part or subset of a defined population, selected or drawn for study and for deriving estimates and drawing inferences related to that population.

5

POPULATION – SAMPLE EXAMPLES

1. Population – All people in the U.S. over age 40
– Sample – All veterans in the U.S. over age 40
2. Population – All veterans in the U.S. over age 40
– Sample – Veterans over age 40 using VA medical centers
3. Population – Men in U.S. with prostate cancer
– Sample – Men with prostate cancer at VA sites participating in a VA prostate cancer trial

6

DEFINITIONS

- *Parameter* – a constant in a mathematical expression that characterizes some population, and whose true value is generally unknown but can be estimated; a descriptive measure computed from the data of a population; summary numbers, such as the mean or variance of a population variable, that are useful in describing the population.

7

DEFINITIONS

- *Estimate* – a numerical value such as a mean, median or variance, based on observed data and serving as an approximation to some true underlying parameter; summary numbers computed from a sample.

8

TYPES OF STATISTICAL INFERENCE

1. Estimation – purpose is to estimate as accurately as possible an unknown population parameter (e.g. μ or σ) or some combination of parameters ($\mu_1 - \mu_2$ or $\sigma_1 - \sigma_2$).
2. Hypothesis testing – purpose is to test specific hypotheses about population parameters or sets of parameters (e.g. $H_0: \mu = C$; $H_0: \mu_1 = \mu_2$; $H_0: \sigma = C$; $H_0: \sigma_1 = \sigma_2$).

9

ESTIMATION – POINT ESTIMATORS

- Point Estimate – the single, best guess of the value of a population parameter; an estimation of a parameter characterized by a single value or point.

10

ESTIMATION – POINT ESTIMATORS

Population Parameter	Point Estimators
P	$\hat{p} = \frac{\text{\# with characteristic}}{n}$
M	$\bar{x} = \sum x_i / n$
S ²	$s^2 = \sum (x_i - \bar{x})^2 / (n-1)$
P ₁ – P ₂	$\hat{P}_1 - \hat{P}_2$
M ₁ – M ₂	$\bar{x}_1 - \bar{x}_2$

11

EXAMPLES OF POPULATION PARAMETERS AND POINT ESTIMATORS

Population Parameter	Point Estimator
P = % of patients deceased or with major morbidity at 1 year following CABG	P = % of patients deceased or with major morbidity at 1 year following CABG in a clinical trial at 20 VAMCs
P1 – P2 = % of patients deceased or with major morbidity at 1 year following off-pump vs. on-pump (traditional) CABG	P1 – P2 = % of patients deceased or with major morbidity at 1 year following off-pump vs. on-pump CABG in a clinical trial at 20 VAMCs

12

EXAMPLES OF POPULATION PATAMETERS AND POINT ESTIMATORS

Population Parameter

M = Mean reduction in DBP from HCTZ in all hypertensives

M1-M2 = Mean reduction in DBP from HCTZ vs. Propranolol in all hypertensives

Point Estimator

\bar{X} = Mean reduction in DBP from HCTZ in hypertensives in a clinical trial at 14 VAMCs

$\bar{X}_1 - \bar{X}_2$ = Mean reduction in DBP from HCTZ vs. Propranolol in hypertensives in a clinical trial at 14 VAMCs

ESTIMATION – CONFIDENCE INTERVAL OR INTERVAL ESTIMATE

- How accurate is the point estimate?
- An interval of values, estimated from observed data presumed to include the parameter of interest (e.g. the population mean) at a specified confidence level (e.g. 95%); an interval of values within which we are fairly confident that the population parameter lies.

ESTIMATION – CONFIDENCE INTERVAL OR INTERVAL ESTIMATE

- Confidence Level (Coefficient) – the amount of confidence that we have that the confidence interval contains the population parameter. Usually 95% is used.

EXAMPLES OF CONFIDENCE INTERVALS

- Difference Between Proportions (P1 - P2)

$$\left[(\hat{p}_1 - \hat{p}_2) \pm Z \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$$

- Difference Between Means (M1 - M2)

$$\left[(\bar{x}_1 - \bar{x}_2) \pm Z \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

- Z = standard normal values:

.90	Z
.95	1.645
.99	1.96
.99	2.57

HYPOTHESIS

- A hypothesis may be defined simply as a statement about one or more populations. The hypothesis is usually concerned with the parameters of the populations about which the statement is made.
- An assumption or statement, not yet proven by experiment or observation, adopted for the sake of testing its soundness.

HYPOTHESIS TESTING

- The purpose of hypothesis testing is to aid the clinician, researcher, or administrator in reaching a decision concerning a population by examining a sample from that population.
- Hypothesis testing is used to obtain a "yes" or "no" answer to our research question.

HYPOTHESIS TESTING

- In hypothesis testing, there are two hypotheses involved and both should be explicitly stated.
- **Null Hypothesis (H0):** In statistics, a hypothesis that asserts that there is no underlying difference in the populations or groups being compared with regard to the factor, trait, characteristic or condition of interest. In general, the null hypothesis is set up for the express purpose of being discredited.
- **Alternative Hypothesis (H1):** A hypothesis stated as an alternative to the null hypothesis.

19

HYPOTHESIS TESTING EXAMPLE

- **Null Hypothesis:** There will be no difference in percent of patients dead or with major morbidity at one year in patients randomized to traditional CABG and those randomized to off-pump CABG.
($H_0: P_1 = P_2$ or $H_0: P_1 - P_2 = 0$)

20

HYPOTHESIS TESTING EXAMPLE

- **Alternative Hypothesis #1:** Patients randomized to off-pump CABG will be less likely to have mortality, morbidity at one year than those randomized to traditional CABG.
($H_A: P_1 < P_2$ or $H_A: P_1 - P_2 < 0$)

21

HYPOTHESIS TESTING EXAMPLE

- Alternative Hypothesis #2: There will be a difference in one-year mortality/morbidity in patients randomized to off-pump CABG and traditional CABG.
(HA: $P1 \neq P2$ or HA: $P1 - P2 \neq 0$)

22

HYPOTHESIS TESTING

- It should be remembered that neither hypothesis testing nor statistical estimation leads to proof of a hypothesis, but merely indicates whether the hypothesis is supported or not by the available data.

23

PROS AND CONS OF ESTIMATION VS HYPOTHESIS TESTING

1. Hypothesis testing is more conventional than estimation. It provides a very specific answer to a research question.
2. Estimation may be a more desirable procedure because instead of just providing a "yes-no" answer to a specific question it can provide more information.
 - a. Estimation provides estimates of value and/or magnitude of differences.
 - b. A confidence interval is actually a simultaneous test of an infinite number of hypotheses (e.g. a 95% CI of (.10, .20) for $P1 - P2$ would accept all H: $P1 - P2 = .10$ to $.20$)

24

SAMPLE SIZE PLANNING

- “A scientific way of making a ballpark estimate.”
- There is more to sample size determination than plugging a significance level and power into some equation!

25

SAMPLE SIZE PLANNING

- Purpose: To ensure that the size of the study is adequate to address the study's primary hypothesis without wasting resources.

26

SAMPLE SIZE PLANNING

- Purpose: A study that is too small will have little chance to answer clearly the study hypothesis. May reject a treatment that is truly effective.

27

SAMPLE SIZE PLANNING

- Purpose: A study that is too large will waste money and resources and could subject patients needlessly to inferior treatments.

28

ELEMENTS OF SAMPLE SIZE

- Primary Hypothesis
- Study Design
- Outcome Measures
- Types of Analyses
- Estimate of expected outcome in control group
- Size of effect to be detected
- Significance Level
- Power
- One-tailed or two-tailed test
- Drop-outs/drop-ins
- Multiple Hypotheses

29

PRIMARY HYPOTHESIS

- Sample size must be related to primary hypothesis
- Hypothesis should be:
 - Simple (break complex ones into simple ones)
 - Specific (fix the subjects, interventions, outcome variable, and statistical test)
 - Specified in advance of the research

30

STUDY DESIGN

- A. Are you doing a comparison study, a modeling study, etc.
- B. Hypothesis Testing
 - Parallel group or completely randomized design
 - Factorial design
 - Cross-over design
 - Number of study arms
 - Repeated measures

31

OUTCOME MEASURES

- A. Discrete (e.g. mortality; better, same, worse)
- B. Continuous (e.g. blood pressure; test scores such as SF-36 or Beck Depression Inventory)
- C. Time to event (e.g. time to death, time to first stroke)

32

ANALYSIS METHOD

- Type of analysis to be used determines the required sample size equation.

- Discrete and Continuous

Predictor/Grouping Variable	Outcome Variable	
Discrete	Discrete Z-statistic or chi-square test or ANOVA	Continuous t-test or ANOVA correlation coefficient
Continuous		

- Time to Event
 - Life Table Methods
 - Kaplan-Meier survival analysis and log-rank test
 - Test of exponential hazard rates
 - Cox regression

33

ESTIMATE OF EXPECTED OUTCOME IN CONTROL GROUP

- From literature or previous research, obtain estimates of the expected results in the control group for the outcome variables being studied.
- Sometimes, best guesses are needed or pilot studies are performed to obtain these estimates.

SIZE OF EFFECT TO BE DETECTED

- How large a difference are we looking for?
 - With a large enough sample size, any size difference can be detected.
 - Should be clinically meaningful.
 - Should be large enough to convince clinicians to change the way that they practice medicine.

SIGNIFICANCE LEVEL – TYPE I ERROR

- The probability of rejecting the null hypothesis with a test of significance when it is true. Also called Type I Error or α error.
- Convention sets Significance Level at 0.05, but can be varied for various situations.

STATISTICAL POWER AND TYPE II ERROR

- Type II Error – The probability of accepting the null hypothesis when it is false. Also called Beta error.
- Statistical Power – The probability of rejecting the null hypothesis when it is false; it is one minus the Type II error (1- Beta).

37

TYPE I AND TYPE II ERRORS

		Decision	
		Accept H_0	Reject H_0
Real Life Situation	H_0 True	Correct Decision	Type I or α error
	H_0 False	Type II or Beta error	Correct Decision

38

ONE OR TWO-TAILED TEST

- A. One-tailed test
 - Association only in one direction (e.g. $H_1: P_1 > P_2$)
 - Requires smaller sample size
- B. Two-tailed test
 - Association in either direction (e.g. $H_1: P_1 \neq P_2$)
 - Larger sample size
- C. In most scientific inquiries, two-tailed test is preferred because the unexpected could happen

39

DROP-OUTS/DROP-INS

- Must adjust sample size for expected drop-out/drop-in rate.
- Estimate rates from literature or previous work.
- Drop-outs are intervention patients who stop intervention; drop-ins are control patients who start intervention

40

DROP-OUTS/DROP-INS

- These tend to lower treatment effect size and increase required sample size
- Drop-outs/drop-ins should be continued in study

41

MULTIPLE HYPOTHESES

- Controversial on how to adjust for more than one major hypothesis.
- For two or three major hypotheses, some would say that significance level must be adjusted to prevent reporting chance findings. Others would say that as long as hypotheses are spelled out in advance of the trial, then there is no need to adjust the significance level.

42

MULTIPLE HYPOTHESES

- For many hypotheses, significance level must be adjusted.
- If you had 100 hypotheses, you could expect five to be significant by chance alone.

43

SAMPLE SIZE EXAMPLE

- Primary Hypothesis: For patients undergoing CABG, there will be no difference in one-year mortality/morbidity between patients randomized to experimental (off-pump CABG) and control (traditional CABG) procedures. ($H_0: P_1 = P_2$)
- Study Design: Two-group comparison; parallel group or completely randomized design.

44

SAMPLE SIZE EXAMPLE

- Outcome Measure: One-year mortality/morbidity (discrete)
- Analysis Method: Continuity-corrected Chi-Square.
- Estimate of Expected Outcome in Control Group: From VA's CICSP database, estimated one-year mortality/morbidity is 8.15%.

45

SAMPLE SIZE EXAMPLE

- Size Effect to be Detected: Planning Committee decided 40% reduction (8% to 4.8%).
- Significance Level: 0.05
- Statistical Power: 0.80
- One or Two-Tailed Test: Two-tail
- Drop-outs: 10%

46

nQUERY 3.0 SAMPLE SIZE DETERMINATIONS

Column	1	2	3	4
Test Significant Level	0.050	0.050	0.050	0.050
1 or 2 Sided Test?	2	2	2	2
Group 1 proportion	0.080	0.080	0.090	0.090
Group 2 proportion	0.048	0.048	0.054	0.054
Odds Ratio	0.580	0.580	0.577	0.577
Power (%)	80	90	80	90
n per group	979	1290	863	1137
% Drop-out	1088	1434	959	1264

47

SAMPLE SIZE PER GROUP FOR LATE PRIMARY OUTCOMES MEASURE FOR TWO INDEPENDENT PROPORTIONS, 2-SIDED TEST, ALPHA=0.05 ADJUSTING FOR 10% DROPOUTS

Control Group	Experimental Group	% Change	Power		
			.80	.85	.90
.07	.0525	25	3400	3871	4508
	.0490	30	2317	2636	3066
	.0455	35	1670	1897	2205
	.0420	40	1254	1423	1651
.08	.0385	45	970	1100	1276
	.0600	25	2948	3357	3909
	.0560	30	2010	2286	3770
	.0520	35	1449	1646	1912
.09	.0480	40	1088	1235	1434
	.0450	45	842	955	1107
	.0675	25	2597	2956	3442
	.0630	30	1771	2014	2342
.10	.0585	35	1277	1451	1686
	.0540	40	959	1098	1264
	.0495	45	744	842	976
	.0750	25	2316	2636	3069
	.0700	30	1580	1797	2090
	.0555	35	1140	1295	1504
	.0540	40	857	971	1127
	.0550	45	664	752	871

48

SAMPLE SIZE FOR DIFFERENCE BETWEEN TWO MEANS

$$N = \frac{4 * S^2 * [Z_{\alpha/2} + Z_{\beta}]^2}{(\bar{X}_1 - \bar{X}_2)^2}$$

where: \bar{X}_1 = mean for group 1
 \bar{X}_2 = mean for group 2

S^2 = pooled variance
 $Z_{\alpha/2}$ = critical value of type I error
 Z_{β} = critical value for power
 N = total sample size

IMPORTANT RELATIONSHIPS

1. If S^2 increases, N increases
2. If α decreases, Z_{α} increases and N increases
3. If β decreases, Z_{β} increases and N increases
4. If $\bar{X}_1 - \bar{X}_2$ decreases, N increases

49

SAMPLE SIZE FOR DIFFERENCE BETWEEN TWO PROPORTIONS

$$N = \frac{4 * \bar{P} * (1 - \bar{P}) * (Z_{\alpha/2} + Z_{(\beta)})^2}{(\bar{P}_1 - \bar{P}_2)^2}$$

where: \bar{P}_1 = proportion for group 1
 \bar{P}_2 = proportion for group 2
 $\bar{P} = (P_1 + P_2)/2$
 $Z_{\alpha/2}$ = critical value for type I error
 Z_{β} = critical value for power
 N = total sample size

50

SAMPLE SIZE CALCULATIONS IN FRIEDMAN, ET AL.

- Dichotomous response variable
 - Two independent samples
 - Paired samples
- Continuous response variable
 - Two independent samples
 - Paired samples
- Repeated measures

51

SAMPLE SIZE CALCULATIONS IN FRIEDMAN, ET AL.

- Time to failure (survival analysis)
- Equivalence tests
- Cluster randomization
 - Continuous response
 - Dichotomous response

52

- Continuous variables generally require smaller sample sizes than dichotomous variables (e.g., differences in mean DBP vs. % of patients whose BP is controlled (DBP <90))
- Time to event analyses generally require smaller sample sizes than dichotomous variables (e.g., survival time vs. % of patients alive at 1 year)

53

EXAMPLE: Antihypertensive drug vs. placebo, 100% Improvement

- Dichotomous variable: % of patients controlled (DBP <90)
30% placebo vs. 60% active drug, alpha = .05, Beta = .10
$$N = 4 \times .45 \times .55 (1.96 + 1.282)^2 / (.3 \times .3)$$
- Continuous variable: Mean reduction in DBP
5 mm Hg + 2.5 vs. 10mm Hg + 5
$$N = 4 \times 15.625 (1.96 + 1.282)^2 / (5 \times 5)$$

54

EXAMPLE: Off-pump vs. On-pump CABG

– % With Mortality/Morbidity at 1 Year vs. Time to Mortal/Morbid Event

- Dichotomous variable: % patients with mortality/morbidity at 1 year

(H0: P1= P2)

8% vs. 4.8%, alpha = .05, Beta = .20, 3 year intake, 1 year follow-up

N = 1958

55

EXAMPLE: Off-pump vs. On-pump CABG

- Time to mortal/morbid event

– Assuming exponential distribution $S(t) = e^{-\lambda t}$

λ = hazard rate or force of mortality (H0: $\lambda_1 = \lambda_2$)

8% vs. 4.8% at 1 year $\lambda_1 = .0834, \lambda_2 = .0492$

N = 765, Using formulas in Friedman, et al, page 115

56

HANDLING DROP-OUTS/DROP-INS

N should be multiplied by $\frac{1}{(1-R_0 - R_I)^2}$

Where R0 = Drop-out rate

RI = Drop-in rate

Example

R0 = .10, RI = .05

Multiplication factor = 1.3841

57

SAMPLE SIZE PROTOCOL PRESENTATION

- Must relate to primary hypothesis.
- Assumptions Used (e.g. effect size, estimates in control group) must be justified.
- Significance level and statistical power must be stated as well as whether tests are one-tailed or two-tailed.

58

SAMPLE SIZE PROTOCOL PRESENTATION

- Appropriate methods used to calculate sample size.
- Sample size calculated accurately.
- Reference(s) to method(s) used given.

59

BIBLIOGRAPHY – STATISTICAL INFERENCE, SAMPLE SIZE, STATISTICAL POWER

- Browner WS, Newman TB, Cummings SR, Hulley SB: Getting ready to estimate sample size: hypotheses and underlying principles. in Hulley SB (ed): Designing Clinical Research. Baltimore, Williams & Wilkins; 1988.
- Browner WS, Black D, Newman TB, Hulley SB: Estimating sample size and power. in Hulley SB (ed): Designing Clinical Research. Baltimore, Williams & Wilkins; 1988.

60

**BIBLIOGRAPHY – STATISTICAL INFERENCE,
SAMPLE SIZE, STATISTICAL POWER**

- Cohen J: Statistical Power Analysis for the Behavioral Sciences, 2nd edition Hillsdale, NJ, Lawrence Erlbaum Associates, Publishers, 1988.
- Donner A: Approaches to sample size estimation in the design of clinical trials – a review. *Statistics in Medicine* 1984;3:199-214.

61

**BIBLIOGRAPHY – STATISTICAL INFERENCE,
SAMPLE SIZE, STATISTICAL POWER**

- Fleiss JL: The design and analysis of clinical experiments. New York, John Wiley and Sons, 1986.
- Hsieh FY: Sample size tables for logistic regression. *Statistics in Medicine* 1989;8:795-802.
- Lachin JM: Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials* 1981;2:93-114.

62

**BIBLIOGRAPHY – STATISTICAL INFERENCE,
SAMPLE SIZE, STATISTICAL POWER**

- Moher D, Dulberg CS, Wells GA: Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994;272:122-124.
- Pocock SJ, et al. Statistical problems in the reporting of clinical trials. A survey of three medical journals. *N Engl J Med* 1987;317:428-432.

63

**BIBLIOGRAPHY – STATISTICAL INFERENCE,
SAMPLE SIZE, STATISTICAL POWER**

- Reed JF, Slaichert W: Statistical proof in inconclusive “negative” trials. Arch Int Med 1981;141:1307-1310.
- Rothman KJ: Significance questing. (Editorial). Ann Int Med 1986;105:445-457.
- Schork AM, Remington RO: The determination of sample size in treatment-control comparisons for chronic disease studies in which dropout or non-adherence is a problem. J Chron Dis 1967;20:233-239.

**BIBLIOGRAPHY – STATISTICAL INFERENCE,
SAMPLE SIZE, STATISTICAL POWER**

- Shuster JJ: Practical Handbook of sample size guidelines for clinical trials. Boca Raton, CRC Press, 1993. (Comes with program for survival analysis samples sizes).
- Simon R: Confidence intervals for reporting results of clinical trials. Ann Int Med 1986;105:429-435.
