

NIH StrokeNet Professional Development Seminar – August 2020

INTRODUCTION TO BIOSTATISTICS FOR CLINICAL RESEARCH

Jordan J. Elm, PhD

Department of Public Health Sciences

Medical University of South Carolina

CONFLICT OF INTEREST / DISCLAIMER

I am contact PI of the StrokeNet National Data Management Center (NDMC) in Charleston, SC.

Other grants from NIH

OBJECTIVES

Provide an introduction to basics of biostatistics as applied to clinical research

- Estimation and Hypothesis Testing
- Basic Overview of Common Analyses
- Sample Size Considerations
- Important topics (in brief)



ESTIMATION AND HYPOTHESIS TESTING

POPULATION:

A population is the entire group that we wish to study.

Notes:

Populations are generally very large. Frequently viewed as infinite.

Can also be called study population, reference population or target population.

A POPULATION HAS PARAMETERS:

The population has characteristics that we want (need) to know:

- a) Proportion (p) who experience DLTs
- b) Proportion who will respond favorably to an intervention
- c) Mean (μ) hematoma expansion volume on DWI

These characteristics are called parameters.

99.99% of the time population parameters are unknown!

A SAMPLE HAS STATISTICS:

A **sample** is a representative group drawn from the population.

We use statistics to make estimates about population parameters by using analogous values computed from a sample.

- Proportion of sample who experience DLTs.
- Proportion of sample who respond.
- Sample mean volume.

These sample summary values (descriptive values) are called statistics.

PARAMETERS VS STATISTICS:

The distinction between statistics and parameters is essential to the understanding of statistical inference.

- We use different symbols to represent each
- Parameters are constants, while sample statistics are random variables.
 - The values of parameters do not change from sample to sample, whereas, statistics change whenever the population is resampled.

STATISTICAL INFERENCE:

Statistical inference is inference about a population from a random sample drawn from it.

It includes:

- Point estimation
- Interval estimation
- Hypothesis testing



ESTIMATION

Point estimates provide a single estimate of the parameter (e.g. mean, proportion, odds ratio, RR).

Interval estimates (**Confidence Intervals**) provide a range of values that seeks to capture the parameter.

"We can be 95% confident that the proportion of ischemic stroke patients who have a 90 day mRS < 2 is between 5.1% and 15.3%."

HYPOTHESIS TESTING:

H_0

H_A

Hypothesis testing provides a framework for drawing conclusions on an objective basis rather than on a subjective basis by simply looking at the data.

“There is enough statistical evidence to conclude that the mean normal body temperature of adults is lower than 98.6 degrees F.”

COURT ROOM ANALOGY

In the US court system, we assume that the accused is innocent until proven guilty.

Two competing hypotheses

Null H_0 : Defendant is not guilty (innocent)

Alternative H_A : Defendant is guilty

The jury examines the evidence.**

If there is enough evidence, we reject the null.

In statistics, the **data are the evidence.



COURT ROOM EXAMPLE:

The jury then makes a decision based on the available evidence (**data**):

If the jury finds sufficient evidence — beyond a reasonable doubt — the jury **rejects the null hypothesis** and deems the defendant guilty. **We behave as if the defendant is guilty.**

If there is insufficient evidence, then the jury **does not reject the null hypothesis**. **We behave as if the defendant is innocent.**

In statistics, we always make one of two decisions. We either "reject the null hypothesis" or we "fail to reject the null hypothesis."

ERRORS IN HYPOTHESIS TESTING:

When testing a hypothesis, 1 of 2 decisions can be made:

- Reject H_0
- Fail to reject H_0

		Truth	
		H_0 true	H_0 false
Decision	Fail to Reject (Accept) H_0	OK	ERROR Type II error " β "
	Reject H_0	ERROR Type I error " α "	OK

TYPE I ERROR:

The probability of a type I error is the probability of rejecting the null hypothesis when it is true.

We generally use α to denote probability of a type one error:

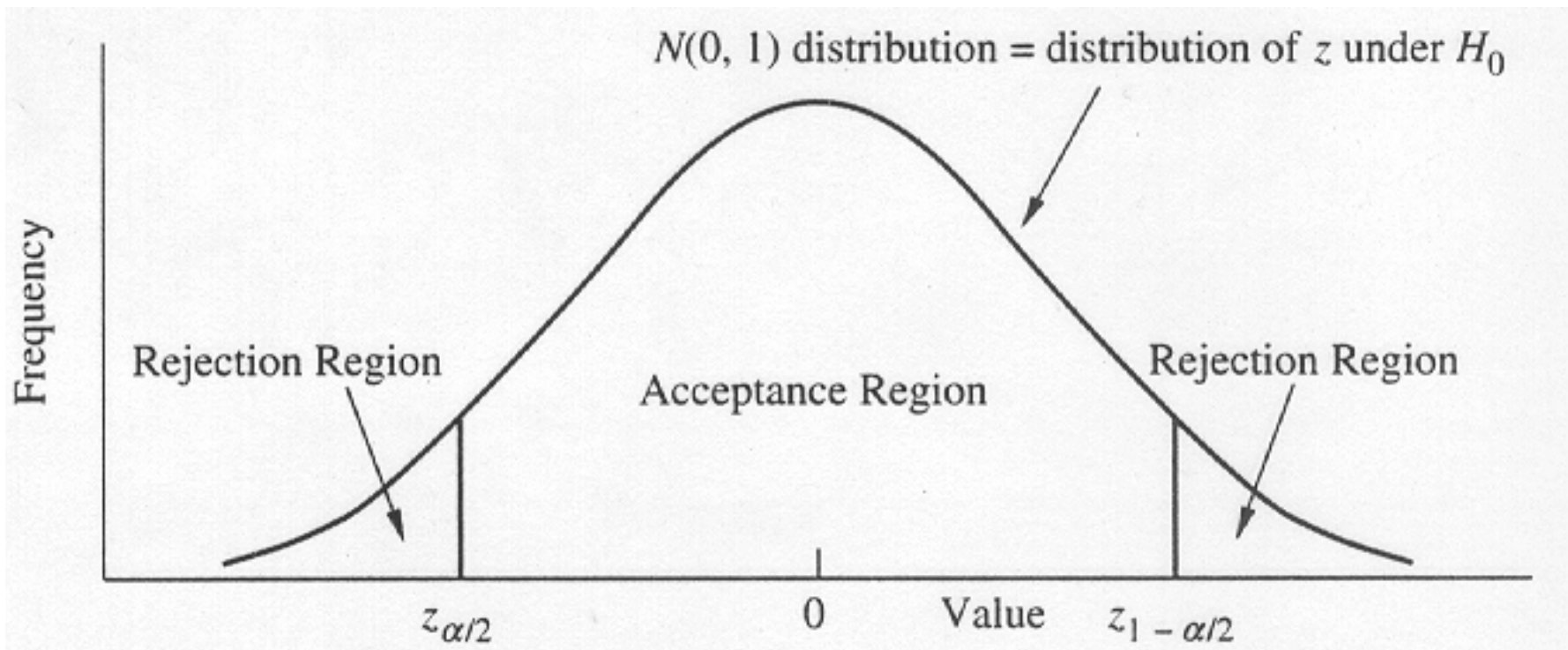
$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

This is called the significance level of a test.

STATISTICAL SIGNIFICANCE

- Hypothesis testing provides a framework for making decisions on an objective basis rather than on a subjective basis by simply looking at the data.
- p-value probability of observing data at least as extreme as that which you have actually observed, assuming that the null hypothesis is true.

NORMAL PROBABILITY CURVE:



TYPE II ERROR AND POWER:

Why should we be concerned about power?

- The power of a test tells us how likely we are to find a significant difference given that the **alternative hypothesis is true**, i.e. given that the true mean μ is different from μ_0 .
- If the power is too low, then we have little chance of finding a significant difference even if the true mean is not equal to μ_0 .

CHOOSING α CAREFULLY:

Because α is chosen by the investigator, it is under his control and is known.

- Thus when you reject H_0 , you know the probability of a Type I error.
- α is chosen a priori (usually set at two-sided 0.05 or 0.01, but could be 0.10 if well justified)

So why not make α very, very small?

This may be the solution in some cases, however, reduction in the α level without increasing your sample size will always increase the probability of a Type II error.

α AND β AND STATISTICAL CONCLUSIONS:

If we reject H_0 we may have made a Type I error, and if we fail to reject we may have made a Type II error.

Because we have these two types of error and one is potentially possible in any decision, we **NEVER** say that we have proved that H_0 is true or that H_0 is false.

- Proof implies that there is no possibility for error.

Instead we say that the data support or fail to support the null hypothesis (i.e. reject or fail to reject H_0 , respectively.)

STATISTICAL VS CLINICAL SIGNIFICANCE:

The investigator must distinguish between results that are statistically significant and results that are clinically significant.

- Very small differences can become statistically significant. However, very small differences may not have clinical meaning.

Statistical significance does not imply clinical significance.

BRIEF OVERVIEW OF COMMON ANALYSES

Analysis depends on type of measurement:

- Continuous measurement ($^{\circ}$ F temperature) or a Rating Scale (e.g. NIHSS 0, 1, 2,24)
- Nominal (low, medium, high) or Ordinal (mRS 0, 1, 2, 3, 4, 5, 6)
- Binary (yes/no)
- Time to event (yes/no over varying follow-up)

CLINICAL TRIAL

Estimate treatment effect

- **Continuous/Interval Measure** (Blood Pressure, Rating Scale)
 - Differences between means (averages)
- **Binary Proportion** (Adverse Event, mRS \leq 2)
 - Odds ratio (OR) [$\{p_1 / (1 - p_1)\} / \{p_0 / (1 - p_0)\}$]
 - Absolute risk reduction [$p_1 - p_0$]
 - Relative risk (RR) [p_1 / p_0]
 - Relative risk reduction (RRR) [$1 - (p_1 / p_0)$]
- **Time to Event** (death, recurrent stroke)
 - Hazard ratio (HR) (similar to relative risk)

WHAT IS AN ODDS RATIO?

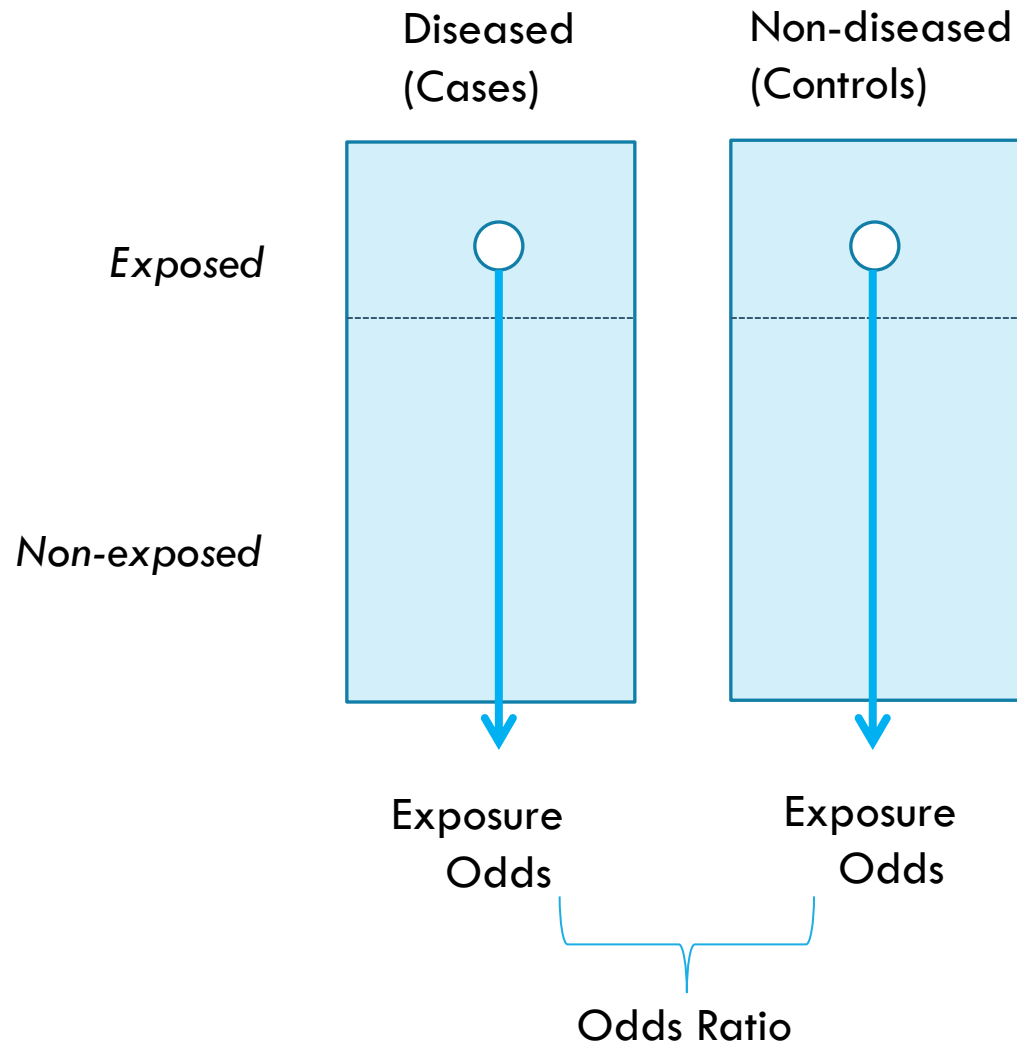
....LETS START WITH THE “ODDS”

The probability that an event will occur is the fraction of times you expect to see that event in many trials. Probabilities always range between 0 and 1.

The odds are defined as the probability that the event will occur divided by the probability that the event will not occur.

If the horse runs 100 races and wins 80, the probability of winning is $80/100 = 0.80$ or 80%, and the odds of winning are $80/20 = 4$ to 1.

ANALYTIC APPROACH



MEASURE RISK

	Cases	Controls	
Exposed	a	b	a + b
Unexposed	c	d	c + d
	a + c	b + d	

Odds Ratio: $a/c \div b/d \approx$ Relative Risk

EXAMPLE

	Movement Disorder Cases	Spousal Controls	
<i>Fragile X Gene</i> Carriers (Exposed)	14	7	23
Non carriers Unexposed	338	267	605
	355	273	

Odds Ratio: $a/c \div b/d \approx$ Relative Risk

$$\text{OR: } 14/338 \div 7/267 = 1.6$$

FIXED COHORT ANALYSIS

		Disease		
		+	-	
Exposure	+	a 40	b 160	Risk=a/(a+b)
	-	c 40	d 760	Risk=c/(c+d)

$$\text{Relative Risk} = \frac{a/(a+b)}{c/(c+d)} = \frac{0.2}{0.05} = 4$$

DYNAMIC COHORT ANALYSIS

		Disease		Time at risk
		+	-	
Exposure	+	a 40	b 160	Risk = a / 100 Person-Years 1800 Person-Years
	-	c 40	d 760	Risk = c / 100 Person-Years 3600 Person-Years

$$\text{Relative Risk} = \frac{a / (100 \text{ P-Y})}{c / (100 \text{ P-Y})} = 2.2 / 1.1 = 2$$

TIME TO EVENT (OR SURVIVAL) ANALYSIS

We can also compare the time to event between treatment groups (or exposed and unexposed) groups.

This is known as a survival analysis, even though the event or outcome might not always be “death”. This is the standard name for an analysis that takes into account time to event.

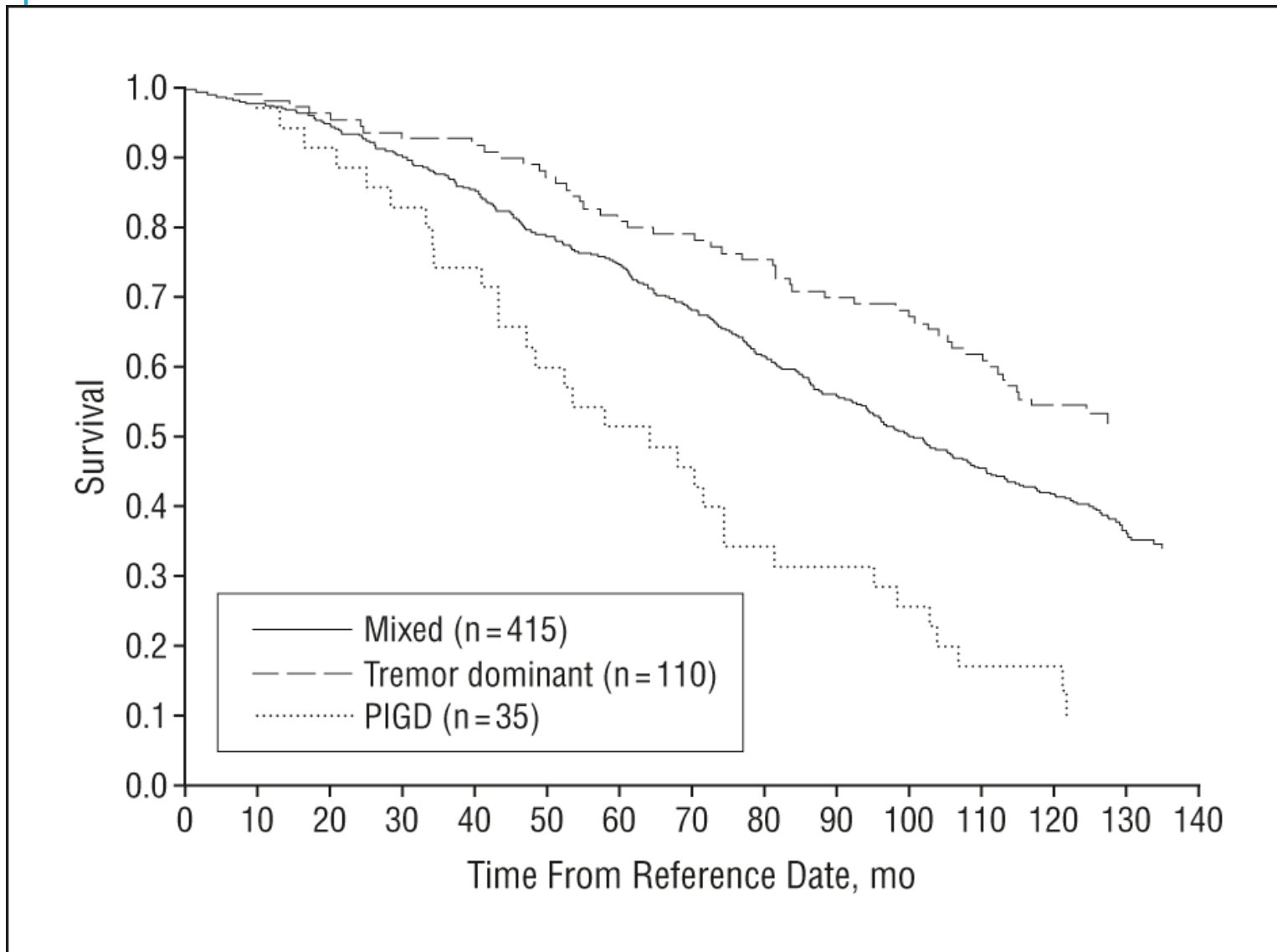
Proportion surviving at a specific time point (2 years)

Median survival: half of the patients in the treatment group have survived for 2246 days (median survival rate) compared to 906 days in the control group.

Cox proportional hazard model) → HR

This method is good when disease onset may take some time. Recurring cancer or prevention trials in Stroke.... Recurrent stroke events ...realistically we need to stop the study after a certain amount of follow-up, but we know that many people would have eventually gotten cancer had we followed them up for longer. These people are said to be “censored” at the end of the study (we know they didn’t have cancer as of the end of the study, but we don’t know their true time to cancer).

KAPLAN-MEIER PLOT OF TIME TO DEATH FOR CLINICAL SUBTYPE





SAMPLE SIZE

WHY WORRY ABOUT POWER/SAMPLE SIZE?

Provides assurance that the trial has a reasonable probability of being conclusive

Allows one to determine the sample size **necessary**, so that resources are efficiently allocated

Ethical Issues

- Study too large implies some subjects needlessly exposed, resources needlessly spent
- Study too small implies potential for misleading conclusions, unnecessary experimentation

SAMPLE SIZE CALCULATIONS

- α (Type I error)
- β (Type II error)
- σ (variance of outcome)
- Δ (clinically relevant difference)

$$\text{sample size} = \frac{(Z_{1-\beta} + Z_{1-\alpha/2})^2 (\text{variance})}{(\text{effect size})^2}$$

VARIABILITY

Is the outcome continuous or categorical?

Continuous

- Need estimate of standard deviation/variance
- based on relevant clinical literature or a range of plausible values

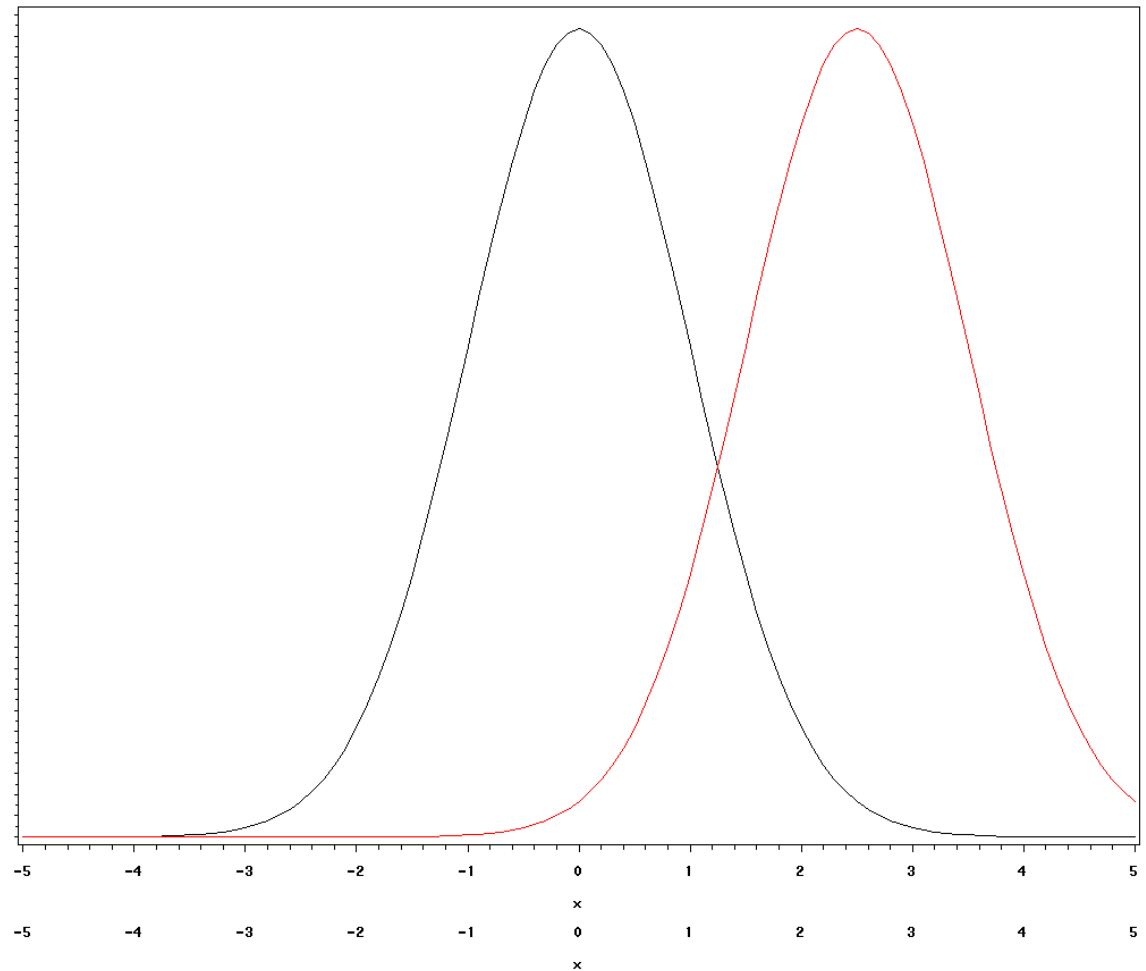
Dichotomous

- Need estimate of control proportion

MINIMUM SCIENTIFICALLY IMPORTANT DIFFERENCE

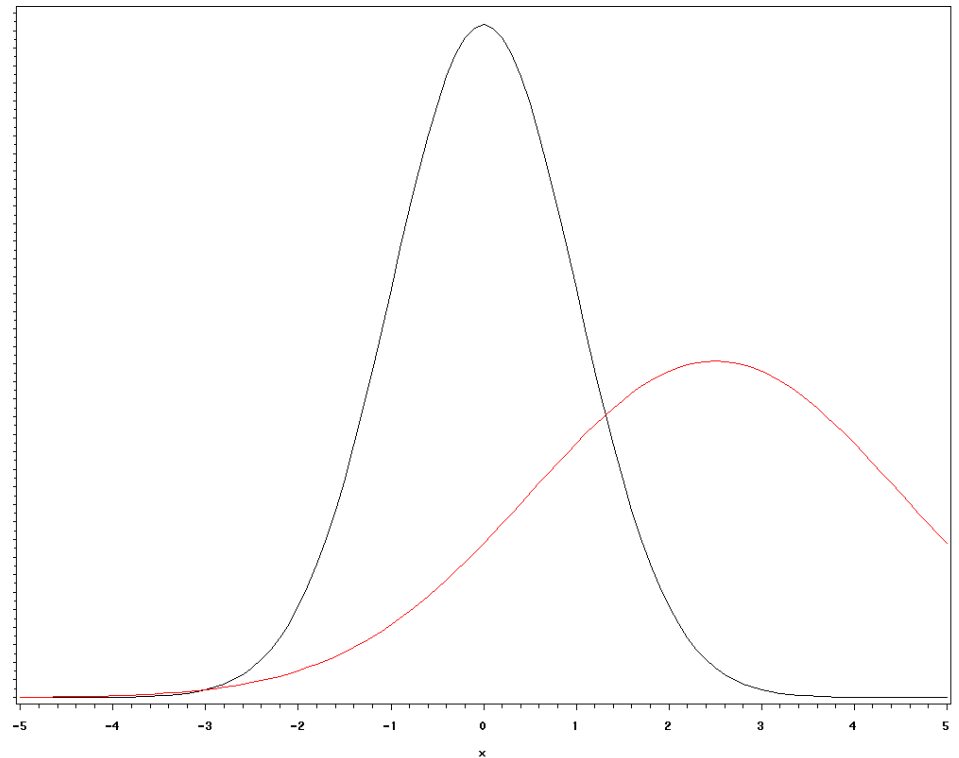
the smallest difference
which would change in
clinical practice

“Larger the difference,
smaller the sample size”



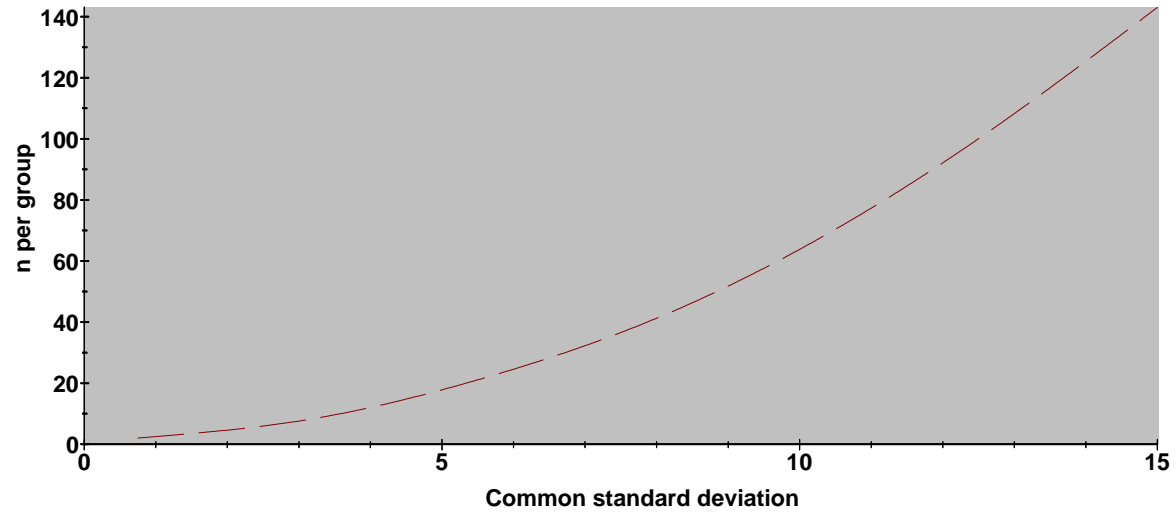
VARIABILITY

“Larger the difference, smaller the sample size” ignores contribution of variability



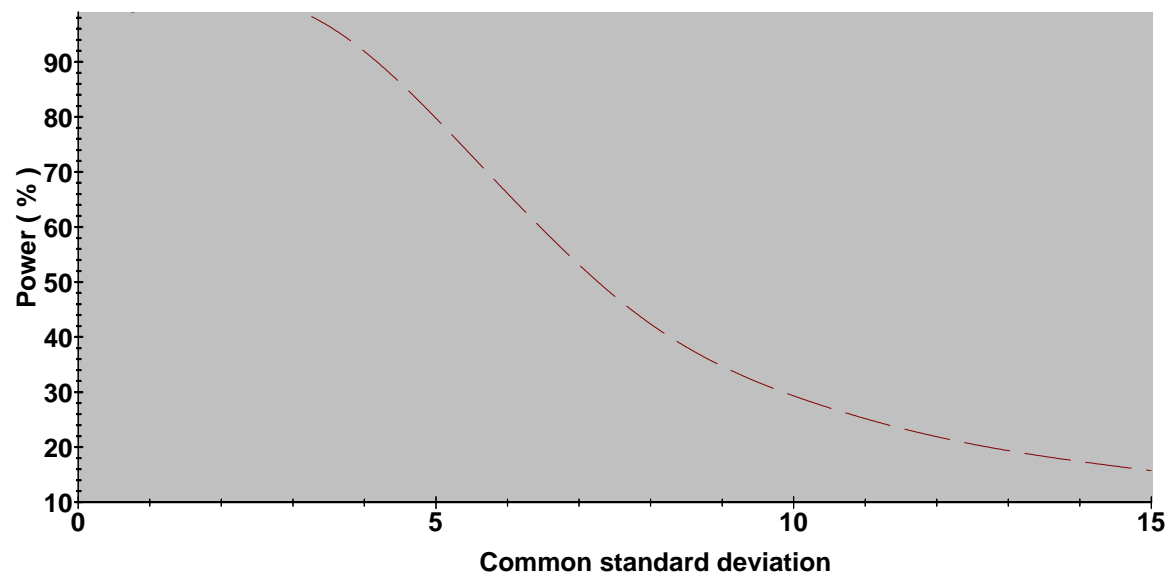
Two group t-test of equal means (equal n's)

--- 80% power, MCID 5 units

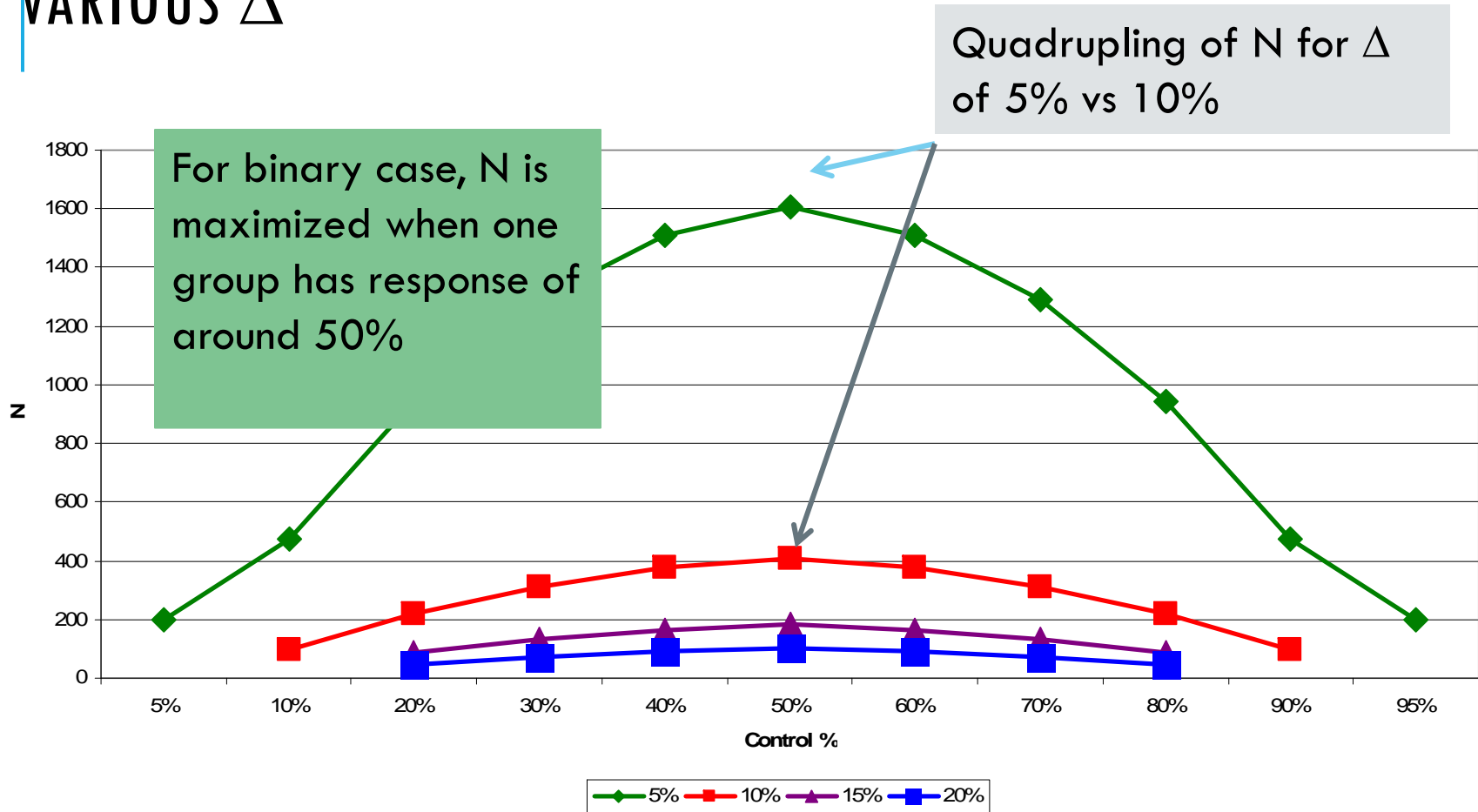


Two group t-test of equal means (equal n's)

--- 80% power, MCID 5 units



N PER GROUP BY CONTROL GROUP % GOOD OUTCOME FOR VARIOUS Δ



Assume 80% power with 2-sided $\alpha=0.05$

ADDITIONAL FACTORS TO CONSIDER FOR TIME-TO-EVENT ANALYSIS

- Number of events of interest
- Study duration and follow-up period
- Subject accrual and lost-to-follow-up rates
- Proportion of censoring

Good reference: Lachin, *Controlled Clinical Trials* 2:93-113, 1981

SAMPLE SIZE ISSUES: MULTIPLICITY



CAUSES OF MULTIPLICITY

- Multiple treatments (e.g., 2 doses + control)
- Multiple outcomes (e.g., efficacy + safety)
- Repeated measures (e.g., Day 1, 7, 30, 90)
- Subgroup analyses (e.g., mild, mod, severe cases)
- Multiple looks (*i.e.*, interim analyses)

SAMPLE SIZE ISSUES: ADJUSTMENTS FOR POTENTIAL MISSING OUTCOME DATA AND NONCOMPLIANCE

INTENT-TO-TREAT (ITT) PRINCIPLE

- Comparison of treatment policies
- Subjects' data are analyzed in the group to which they were randomized regardless of their compliance with the protocol
- Preservation of the benefits of randomization
- Most Phase II/III studies analyzed according to the ITT principle

WERE ALL PARTICIPANTS ANALYZED IN THE GROUPS TO WHICH THEY WERE RANDOMIZED?

“Excluding randomized participants or observed outcomes from analysis and subgrouping on the basis of outcome or response variables can lead to biased results of unknown magnitude or direction”

Friedman LM, Furberg CD, DeMets DL. Fundamentals of Clinical Trials, 3rd Edition. New York: Springer-Verlag, 1998, p. 284.

MISSING OUTCOME DATA

- Subject became lost-to-follow-up
- Subject withdrew consent
- Subject died
- No other reason should exist for missing outcome data!

NONCOMPLIANCE (PROTOCOL VIOLATIONS)

- Subject became lost-to-follow-up
- Subject withdrew consent
- Subject had not met eligibility criteria
- Subject/investigator did not comply with treatment regimen
- Crossover in treatment allocation

ANALYSIS EXCLUDING MISSING OUTCOME/ NONCOMPLIANCE CASES

If $d \times 100\%$ of subjects is anticipated not to complete the protocol, and their outcome is unknown or not imputed, then divide the calculated N by $(1-d)$ to get the adjusted (inflated) N

EXAMPLES

- If 10% of recruited subjects are anticipated to drop out or become ineligible during a run-in period, then required $N = (\text{estimated } N) / 0.90$.
- If plan to do per-protocol analysis and expect that 5% of subjects during follow-up will drop out, then required $N = (\text{estimated } N) / 0.95$

ADJUSTMENT FOR ITT ANALYSIS

- If $r_1 \times 100\%$ of the patients is expected to “switch” from intervention to control and $r_2 \times 100\%$ of the patients is expected to “switch” from control to intervention, then multiply the calculated N by the inflation factor: $IF = 1 / (1 - r_1 - r_2)^2$
- The IF is to compensate for the dilution of the difference in the treatment effect, i.e., the actual difference may be smaller than what was estimated prior to the study initiation.

ITT EXAMPLE

Suppose for a study using weight change outcome:

Tx Grp	N	Est μ	Drop out	σ
A	63	30 lbs	15%	20
B	63	20 lbs	25%	20

So, $\Delta = \mu_A - \mu_B = 10$ with planned total $N=126$ and power of 80%

ITT EXAMPLE (CONT'D)

With the drop in/out, the observed $\Delta = \Delta'$:

$$\Delta' = \frac{[(30 \times 0.85) + (20 \times 0.15)] - [(30 \times 0.25) + (20 \times 0.75)]}{30 + 20} = 6$$

< original planned Δ of 10

$$IF = 1/[(1 - r_1 - r_2)^2] = 1/[(1 - 0.15 - 0.25)^2] = 2.78$$

$$\text{New N under ITT: } N' = 126 \times 2.78 = \mathbf{350}$$

DISCUSSION 1

If you claim to conduct an intention-to-treat analysis and a randomized subject stops taking the assigned treatment due to an adverse event, do you follow that person according to the protocol or do you do their final assessments at that point and remove them from the study?

STATISTICAL CONSIDERATIONS

Were the Groups Comparable at the Start of the Study?

Were All Participants Accounted for at the end of Follow-up?

How complete was the follow-up?

- **Impute Missing data**



Excellent health statistics - smokers are less likely to die of age related illnesses.'

HANDLING MISSING DATA

Impute missing data

- Single point imputation (LOCF, Worse case, best case, mean imputation)
- Multiple imputation (Using a modelling approach repeatedly impute the missing cases (e.g. 20 times, perform the test, and summarize the findings across imputed datasets)

PRE-SPECIFIED STATISTICAL ANALYSIS PLAN

Avoid of Statistician Bias

Sample Size/Power/Study Design should be in agreement.

State error rates, approach to deal with multiplicity.

Randomization plan

Baseline comparisons

Missing data

Analysis Samples, ITT/Per Protocol

Plans for Interim Analyses

Pre-specify model building approach and baseline covariates/confounders to be adjusted

Prioritization of outcomes

- Primary vs. secondary vs. exploratory outcomes (Standard definitions)