

Some Basic yet Key statistical concepts

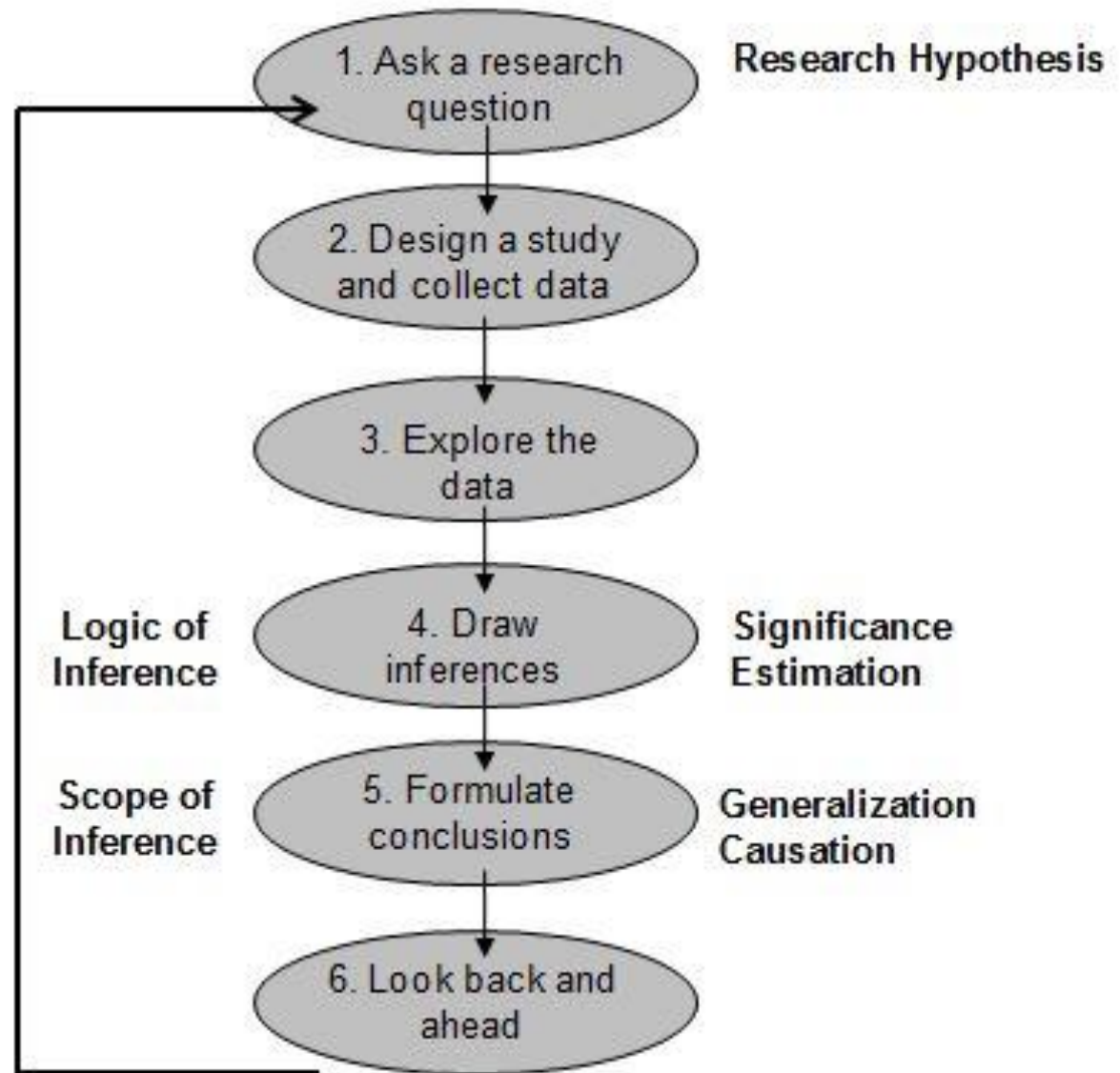
Oct 29, 2025

Nan Xue

Division of Biostatistics

Department of Epidemiology and Population Health

Statistical concepts important across research process

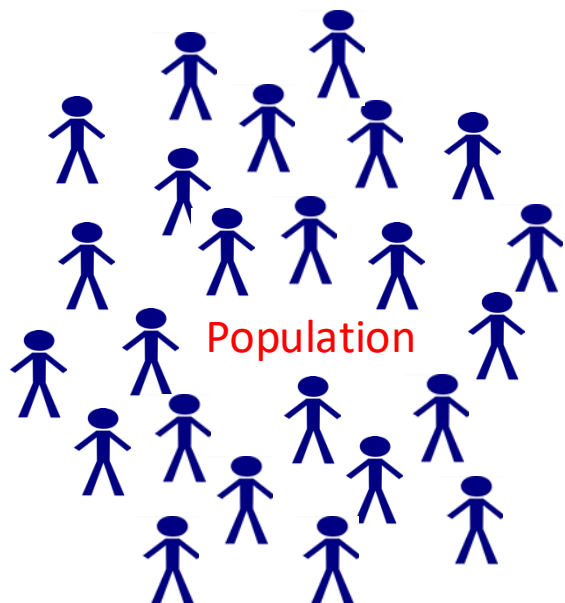


List of Key Stat Concepts

1. Population vs Sample
2. Hypothesis testing
3. Power and Sample Size
4. P-value
5. Confidence intervals
6. Multiple testing and false discovery
7. Bias
8. Confounding
9. Bias and Precision

1. Population vs Sample

We want to know about this:



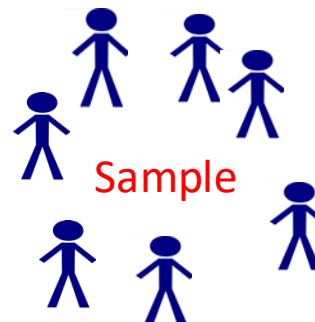
Population



Parameter:

Population Mean μ

But have to work with this:



Sample



Statistic:

Sample mean \bar{x}

Random Selection



Inference



Sample vs Population

- Collect Data from sample
- Make inference on the population
- Key features of sample
 - Representative (not biased)
 - Adequate size (enough information)

2. How to perform a hypothesis test

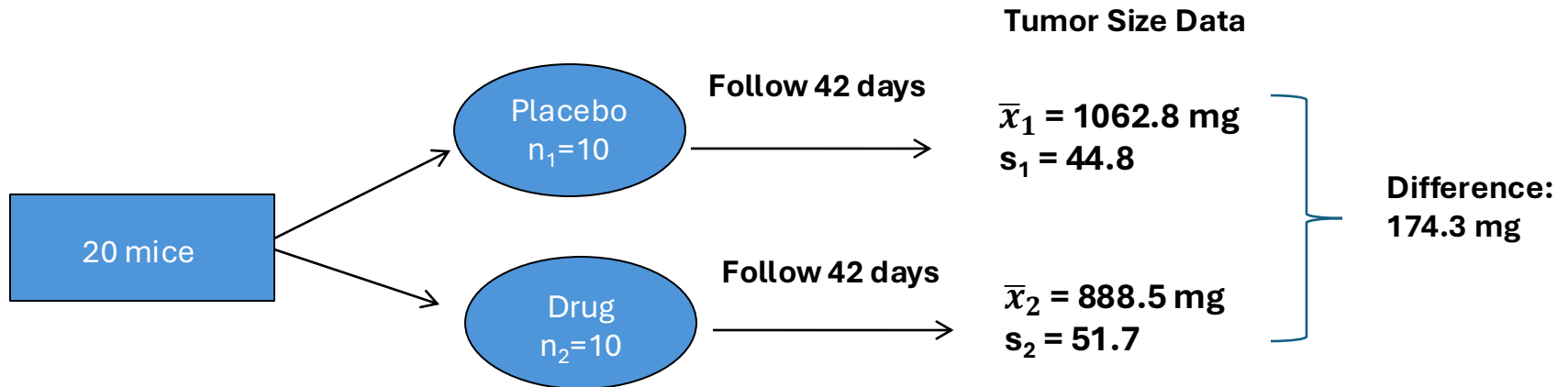
Hypothesis Testing Framework

1. Start with neutral assumption (e.g., no difference between groups)
2. Analyze data (evidence)
3. Compute P-value: How likely is the observed difference happening if truly no difference between groups (null hypothesis)?
4. Assume rare things do not happen: If this probability is VERY unlikely (less than pre-determined significance level α , commonly 0.05); reject the assumption of no difference and conclude treatment has an effect
- 5 Make a decision (reject null or fail to reject)

Example

1. Formulate hypotheses: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

2. Collect data:



3. Combine data into test statistic to evaluate hypothesis

Choose the right statistical test for your outcome and study design

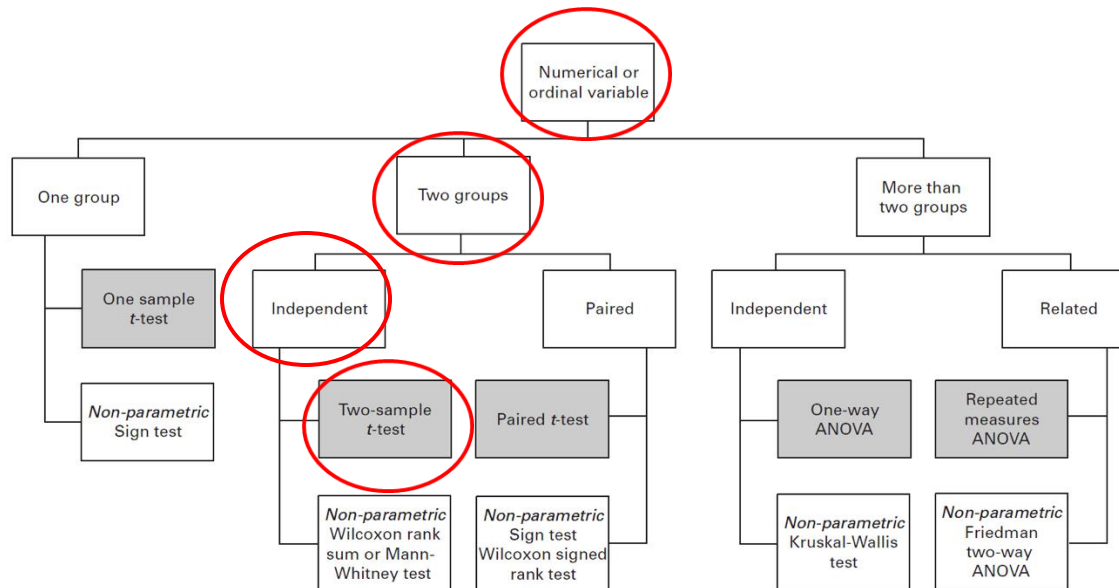


Fig. 4

Flowchart indicating choice of test when the data are numerical (tests in shaded boxes require relevant assumptions to be satisfied) (ANOVA, analysis of variance).

Petrice, A. Statistics in Orthopaedic Papers (2006)

Test Statistic for Comparing Two Means

$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

Test statistic is “signal to noise” ratio weighted by sample size:

$$t = \frac{\sqrt{n}(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_1^2 + s_2^2}}$$

Sample Size

Signal:
Magnitude of effect

Noise:
Degree of variability:

Note: This is the statistic for the T-test when sample sizes are equal in the two groups

Hypothesis Testing

- Mean tumor size at day 42:

$$\begin{array}{ll} \text{Placebo (n=10):} & \bar{x}_1 = 1062.8 \text{ mg; } s_1 = 44.8 \\ \text{Drug (n=10):} & \bar{x}_2 = 888.5 \text{ mg; } s_2 = 51.7 \end{array} \quad \left. \vphantom{\begin{array}{l} \text{Placebo} \\ \text{Drug} \end{array}} \right\} \text{Diff} = 174.3 \text{ mg}$$

- Summarize results into appropriate test statistic for comparing means:

$$t = \frac{\sqrt{n}(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_1^2 + s_2^2}}$$

Note: when n is the same
in two groups

$$t = \frac{\sqrt{10}(174.3)}{\sqrt{(44.8)^2 + (51.7)^2}} = 8.1$$

- $t=8.1$ falls in rejection region (≥ 1.734), reject null

What can go wrong in hypothesis testing



Types of Error

		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Accept H_0	Correct decision	Type II error

How to minimize these errors

- More stringent threshold for significance
(e.g., $\alpha = 0.01$ instead of $\alpha = 0.05$)
- Reduce sources of variability (decrease SD or background “noise”)
- Increase sample size

3. Sample Size and Power



How large does my study need to be?



- Too few: Can't detect effect of interest (won't get statistically significant results)
- Too many: Waste of resources; even small effects that are not biologically important become statistically significant
- Typically want to compute sample size to achieve 80% power
- Use sample size software

What information do you need to determine sample size?



Statistician

Experimental Design

Effect Size

Variability

Type I error

Power

Power and Sample Size

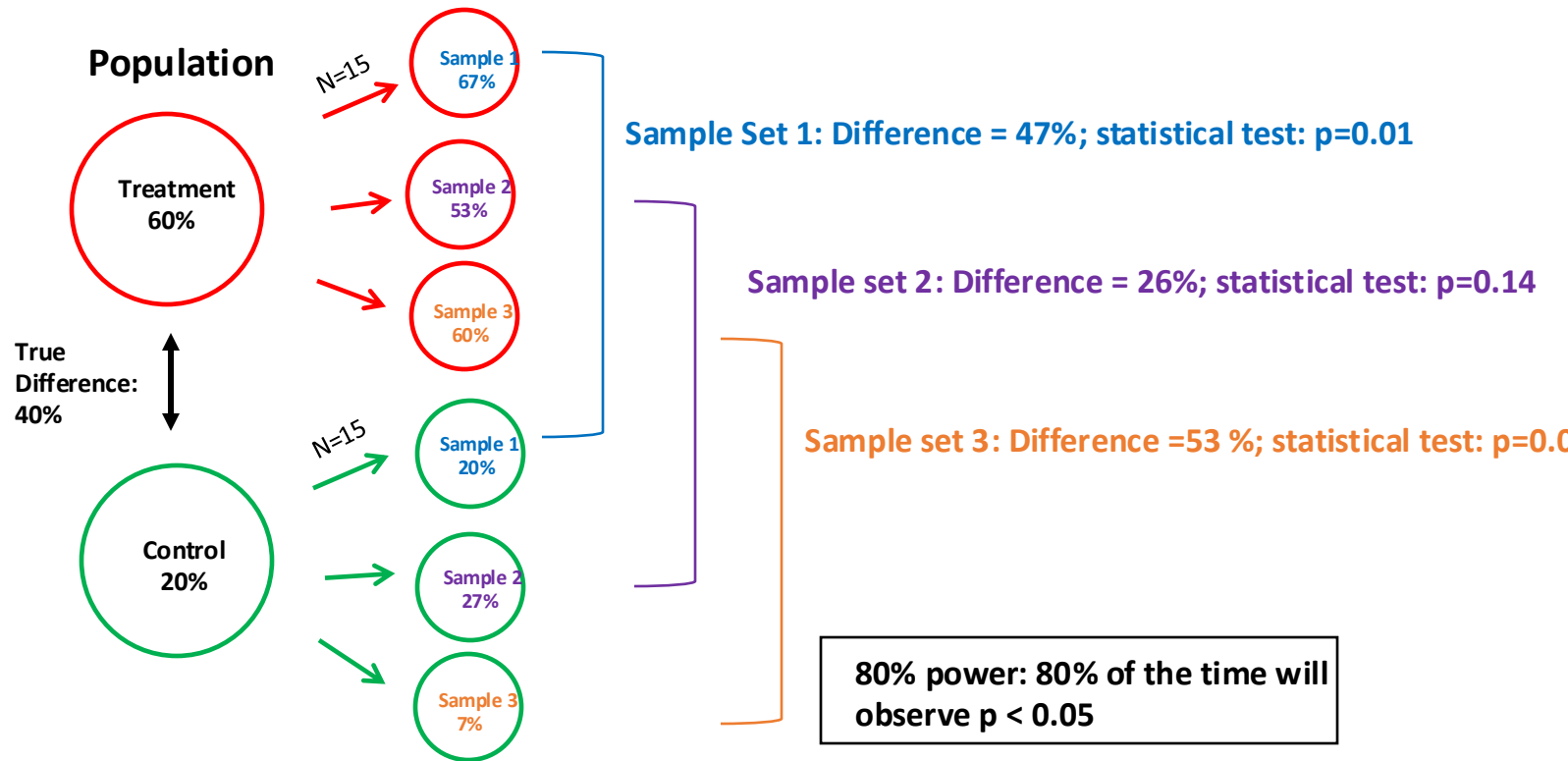
		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Accept H_0	Correct decision	Type II error

Power

- Power = $1 - \Pr(\text{Type II error})$
= $\Pr(\text{correctly concluding effective treatment works})$
- 80% power: Probability concluding that an effective treatment works is 80%.

Example

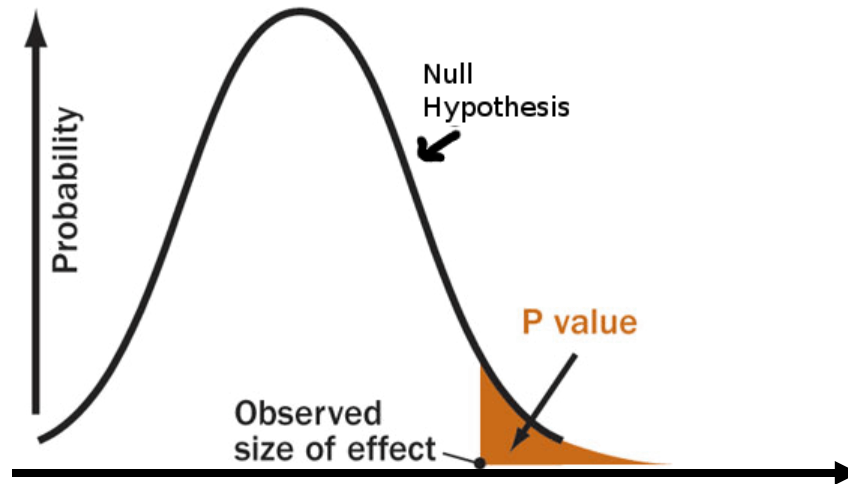
“The study had 80% power with $N=100$ per group and a two-sided Type 1 error rate of 5% to detect a 20% difference in response rates between treatment and control arms (60% in experimental vs. 40% in control).”



How many subjects?

Decreases N	Increases N
↑ Rx effect	↓ Rx effect
↓ Variability	↑ Variability
↑ α	↓ α
↓ Power	↑ Power

4. P-Values



Back to Example

- Mean tumor size at day 42:

Placebo (n=10): $\bar{x}_1 = 1062.8$ mg; $s_1 = 44.8$
Drug (n=10): $\bar{x}_2 = 888.5$ mg; $s_2 = 51.7$ } Diff = 174.3 mg

- Summarize results into appropriate test statistic for comparing means:

$$t = \frac{\sqrt{n} (\bar{x}_1 - \bar{x}_2)}{\sqrt{s_1^2 + s_2^2}}$$

Note: when n is the same
in two groups

$$t = \frac{\sqrt{10} (174.3)}{\sqrt{(44.8)^2 + (51.7)^2}} = 8.1$$

- What is probability of observing t as large as 8.1 by chance if truly no difference between control and drug?

$$\begin{aligned} \Pr(\text{observe } t > 8.1) + \Pr(\text{observe } t < -8.1) = \\ 2 \times (9.76 \times 10^{-7}) = \mathbf{1.95 \times 10^{-6}} \end{aligned}$$

P-value

- Probability of observing a result at least as extreme as what you observed if null hypothesis were true (i.e., no treatment effect).
- High p-value: If treatment groups same, the observed difference is likely result under the null; consistent with no treatment effect.
- Low p-value: What you observed is unlikely result under null;
- Example: Observed $p=1.95 \times 10^{-6}$; conclusion?

P-value

- We statisticians do not think RARE events happen
- RARE event is evidence AGAINST the null hypothesis
- Low p-value leads to rejection of the null

Keep in mind factors that affect P-value

Sample Size

Signal:

Magnitude of effect

$$t = \frac{\sqrt{n} (\bar{x}_1 - \bar{x}_2)}{\sqrt{s_1^2 + s_2^2}}$$

Noise:

Degree of variability:

Six Ways to P-Hack

1. Stop collecting data once $p < .05$
2. Analyze many measures, but report only those with $p < .05$.
3. Collect and analyze many conditions, but only report those with $p < .05$.
4. Use covariates to get $p < .05$.
5. Exclude participants to get $p < .05$.
6. Transform the data to get $p < .05$.

Leif Nelson, UC Berkeley Professor

NEWS ANALYSIS

More Evidence That Nutrition Studies Don't Always Add Up

A Cornell food scientist's downfall could reveal a bigger problem in nutrition research.



Dr. Brian Wansink at the 2013 Discovery Vitality Summit in Johannesburg. Lefty Shivambu/Gallo Images

"There's nothing wrong with having a lot of data and looking at it carefully....The problem is p-hacking."

Andrew Althouse, University of Pittsburgh

Effect of Sample Size on P-value

Marital satisfaction and break-ups differ across on-line and off-line meeting venues

John T. Cacioppo^{a,1}, Stephanie Cacioppo^a, Gian C. Gonzaga^b, Elizabeth L. Ogburn^c, and Tyler J. VanderWeele^c

^aDepartment of Psychology, Center for Cognitive and Social Neuroscience, University of Chicago, Chicago, IL 60637; ^bGestalt Research, Santa Monica, CA 90403; and ^cDepartment of Epidemiology, Harvard University, Boston, MA 02115

Edited by Linda M. Bartoshuk, University of Florida, Gainesville, FL, and approved May 1, 2013 (received for review December 24, 2012)

Marital discord is costly to children, families, and communities. The advent of the Internet, social networking, and on-line dating has affected how people meet future spouses, but little is known about because on-line venues have tended to be treated as a homogenous terrain (2) despite on-line venues having grown in number, variety, and complexity.

- Study of > 19,000 people
- Those who met spouses on-line were less likely to divorce ($p < 0.002$) and more likely to have high marital satisfaction ($p < 0.001$) than those who met off-line

P-value and Effect Size

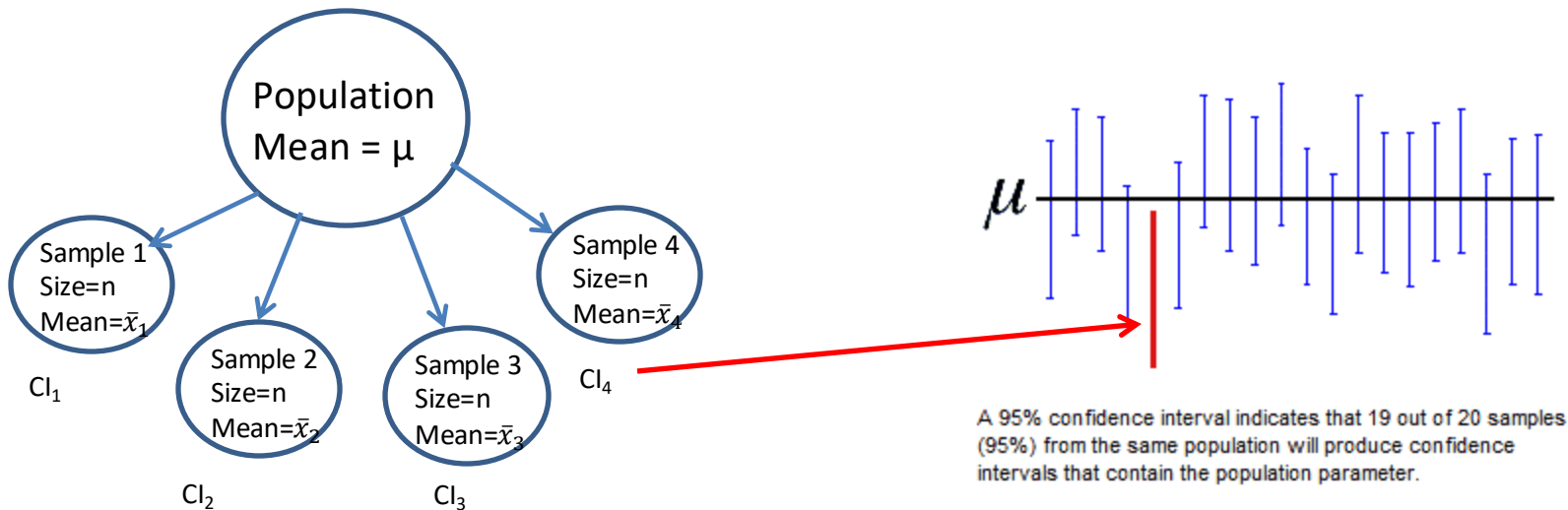
	Off line	On Line	P-value
Divorce rate	7.7%	6.0%	$P < 0.002$
Happiness (7 point scale)	5.5	5.6	$P < 0.001$

5. Confidence interval

95% CI for μ : $\bar{x} \pm 1.96 SE$


Confidence Interval

- 95% Confidence Interval: 95% chance or probability that given interval includes true but unknown population value

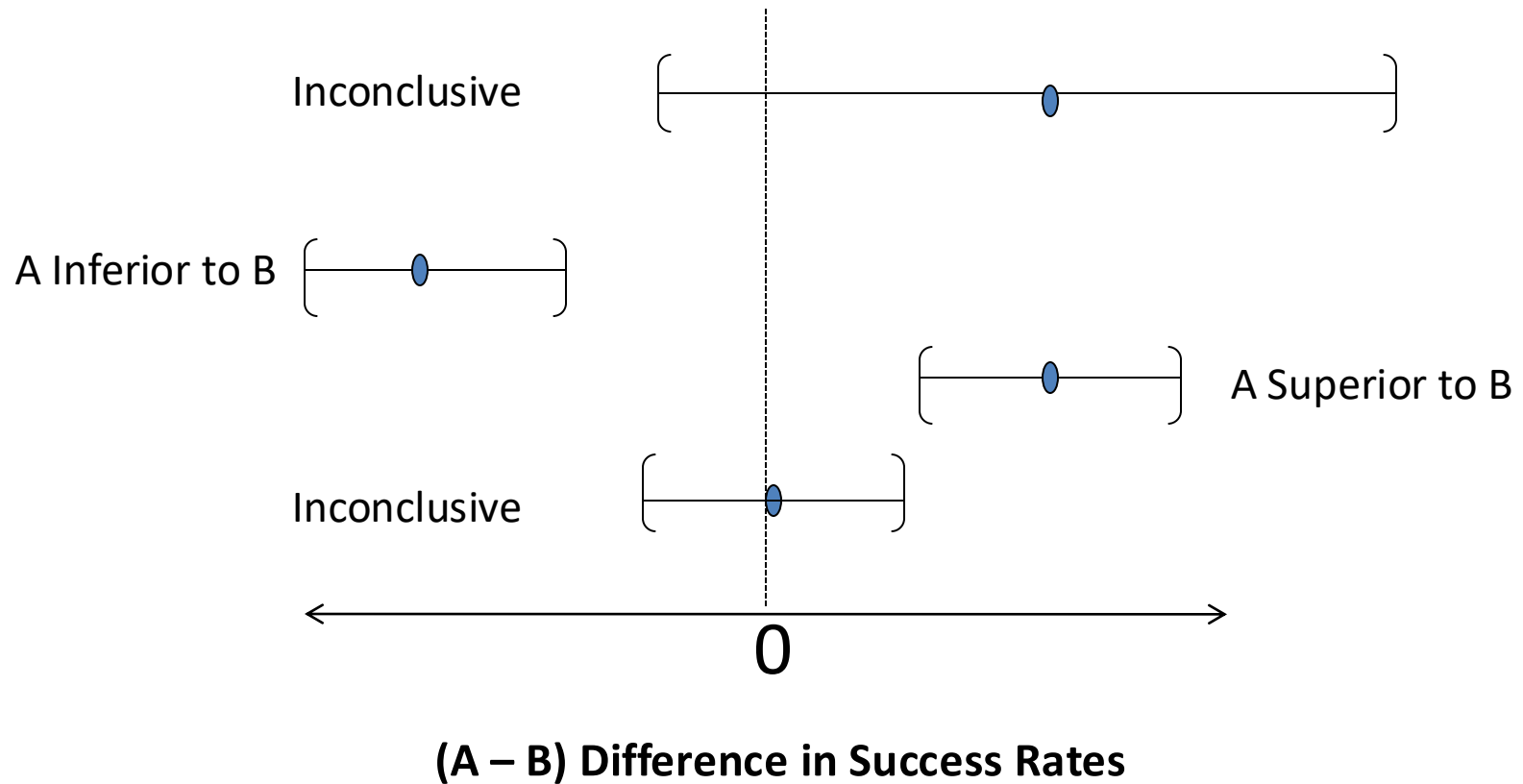


Confidence Interval

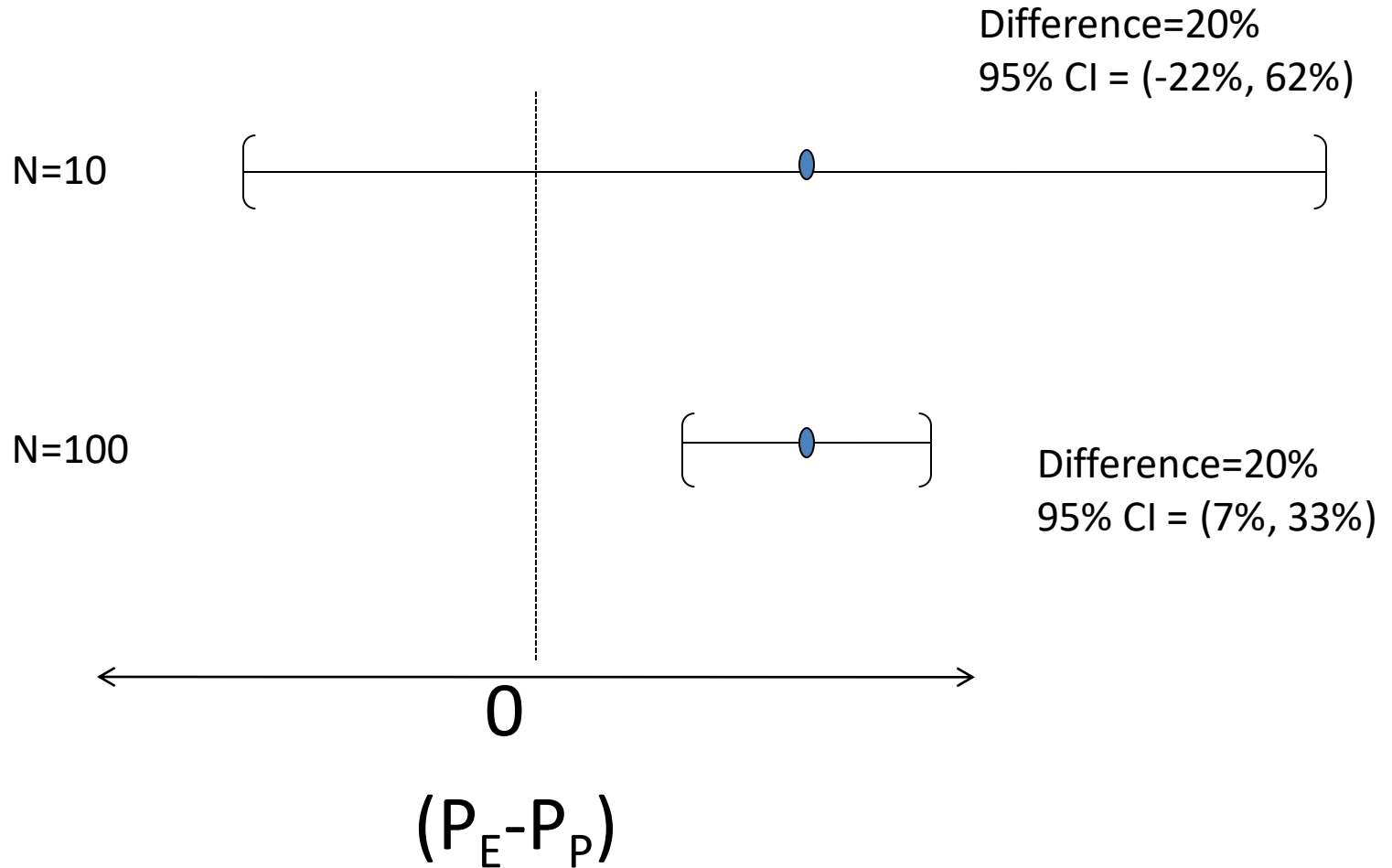
- More informative than p-value
- All values in the CI for population parameter could NOT be rejected at $p < 0.05$ level if these values were tested using sample data

 Values for true population parameter that are “consistent” with your data.

Inference with Confidence Intervals



Impact of Sample Size on Confidence Intervals



6. Multiple testing

Consequences of Multiple Testing

- If test 100 genes for association with a disease using $\alpha = 0.05$, probability of observing at least one false positive result is >99% just by chance even if no genes are truly associated.



False positive rate/Type I error rate increases with number of tests performed

Adjusting for multiple testing

- Bonferroni Correction (very conservative)
 - Adjust p-value threshold: $\alpha^* = 0.05/100 = 0.0005$
 - Only consider significant if $p < 0.0005$
 - Prevents false positives, but may miss real effects
- False Discovery Rate (FDR; more flexible)
 - Controls the **proportion** of false positives among all the “significant” results
 - Common method: **Benjamini-Hochberg**
 - Widely used in high-dimensional fields (e.g., genomics, machine learning)

7. Bias

- Occurs when subjects, specimens, data being compared are not inherently same or not handled similarly, resulting in systematic difference between groups
- Sources of bias:
 - Biological/clinical characteristics
 - Investigator bias
 - Study sample not representative of underlying study population (e.g., EHR)
 - Laboratory error
 - Data entry or coding error
 - Wrong statistical approach

Mechanisms of disease

🔍 Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

Summary

Background New technologies for the detection of early-stage ovarian cancer are urgently needed. Pathological changes within an organ might be reflected in proteomic patterns in serum. We developed a bioinformatics tool and used it to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary.

Introduction

Application of new technologies for detection of ovarian cancer could have an important effect on public health,¹ but to achieve this goal, specific and sensitive molecular markers are essential.²⁻⁵ This need is especially urgent in women who have a high risk of ovarian cancer due to family or personal history of cancer, and for women with

- Blood test based on proteomic patterns for the early detection of ovarian cancer
- SE=100% , SP=95%, PPV= 94%



Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments

Keith A. Baggerly*, Jeffrey S. Morris and Kevin R. Coombes

Department of Biostatistics, U.T. M.D. Anderson Cancer Center, 1515 Holcombe Blvd, Box 447, Houston, TX 77030-4009, USA

Received on July 14, 2003; revised on October 14, 2003; accepted on October 16, 2003
Advance Access publication January 29, 2004

ABSTRACT

Motivation: There has been much interest in using patterns derived from surface-enhanced laser desorption and ionization (SELDI) protein mass spectra from serum to differentiate samples from patients both with and without disease. Such patterns have been used without identification of the underlying proteins responsible. However, there are questions as to the stability of this procedure over multiple experiments.

Results: We compared SELDI proteomic spectra from serum from three experiments by the same group on separating

However, there are questions as to the stability of this procedure over multiple experiments. An illustration of the potential power of the proteomic technique is apparently provided by ovarian cancer. Ovarian cancer is frequently a deadly disease, and its morbidity is strongly linked to our inability to detect the tumors at an early stage. Neither X-rays nor MRIs are able to differentiate between cancers and benign cysts, surgical verification of cancer status is invasive, and gene product assays (such as CA125) have never been shown to be effective in screening programs. A simple, easily applied diagnostic test

“...these concerns suggest that much of the structure uncovered in these experiments could be due to artefacts of sample processing, not the underlying biology of cancer”

Selection Bias



Example: Use Electronic Health Records data to study risk factors for diabetes

- Extract data from patients who had at least one fasting glucose test in past 2 years
- Problem: Patients with higher BMI or family history more likely to be tested; sample overrepresents high-risk individuals
- Inclusion in EHR dataset via testing, diagnosis, utilization is not random; can introduce selection bias
- How to address? Use data from multiple sources; adjust for health care utilization patterns; inverse probability weighting

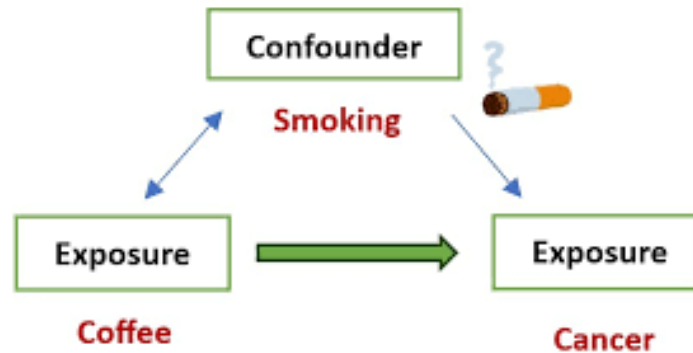
Missing Data



Type	Description	Example	Bias
MCAR	Missing Completely at Random	Equipment failure	No
MAR	Missing At Random (Depends on observed data)	Elderly more likely to skip a survey question	Maybe
MNAR	Missing Not At Random (Depends on unobserved data)	Patients with worse symptoms drop out	Yes

- Missing data can reduce sample size and introduce bias
- Prevention is better than correction (e.g, imputation)
- Assess and report degree of missingness

8. Confounding



Confounding



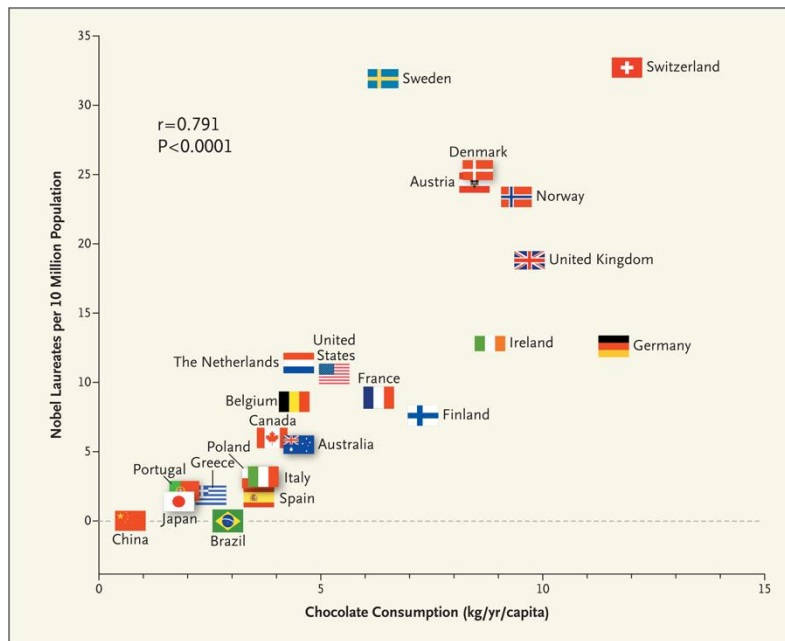
The NEW ENGLAND
JOURNAL of MEDICINE

OCCASIONAL NOTES

f X

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Author: Franz H. Messerli, M.D. [Author Info & Affiliations](#)



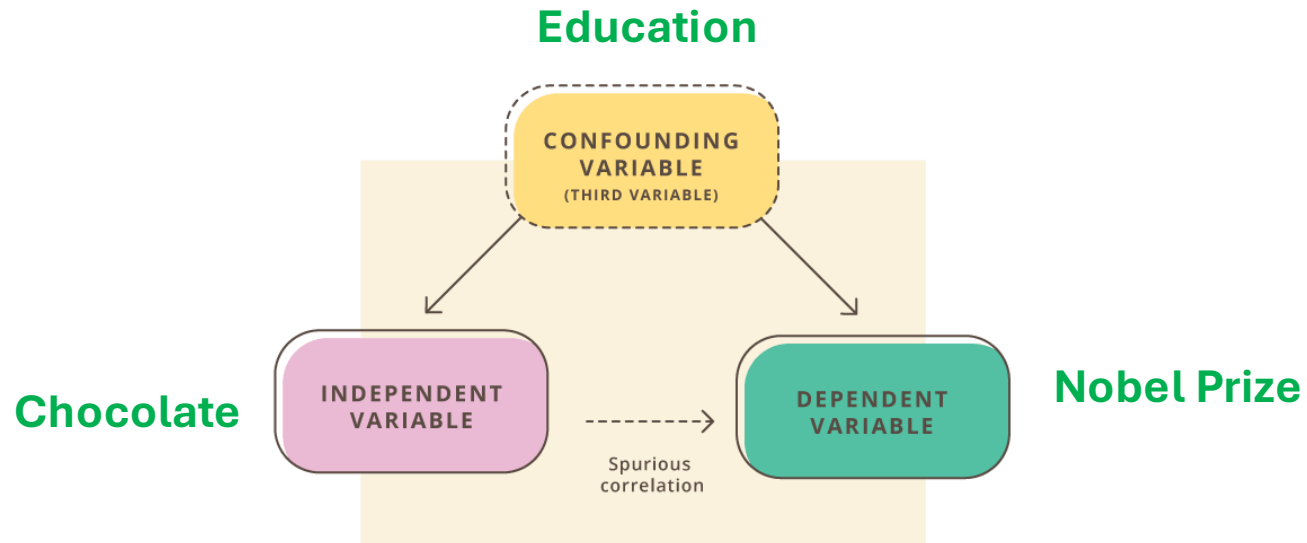
Explanations for finding:

- Author is correct: Chocolate increases cognitive performance
- A third variable (wealth, education) explains relationship, e.g., countries with more resources and education to develop nobel prize winners also happen to consume more chocolate

Conclusions

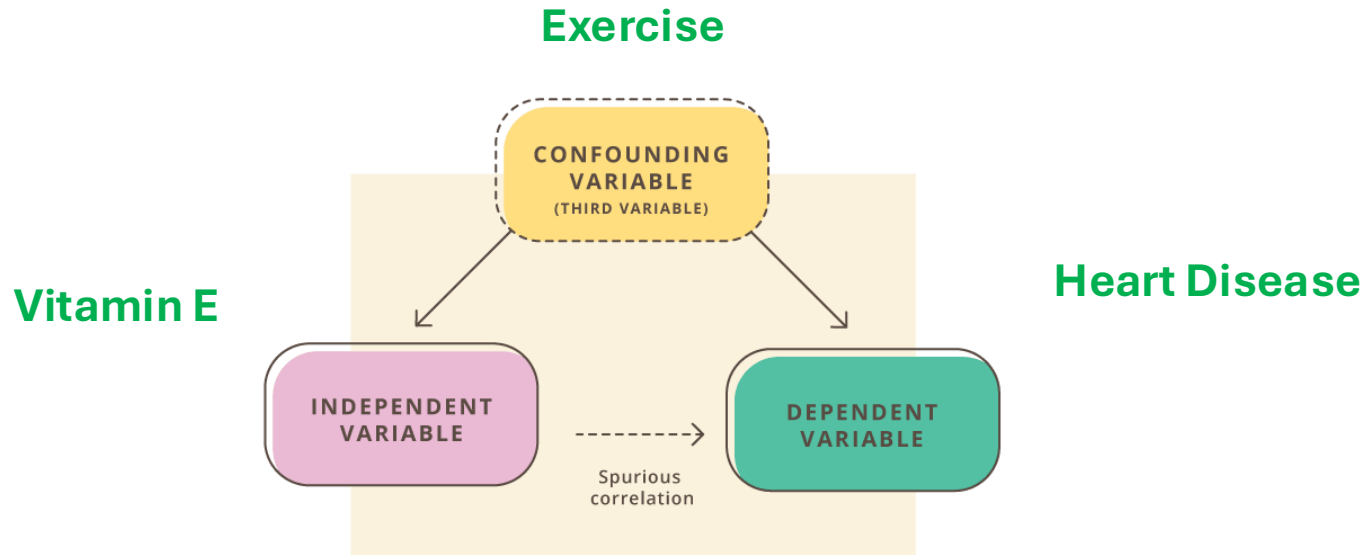
“Chocolate consumption enhances cognitive function, which is a sine qua non for winning the Nobel Prize, and it closely correlates with the number of Nobel laureates in each country”

Confounding



What at first looks like a causal relationship between IV and DV is ultimately spurious. The confounding variable is the hidden explanation.

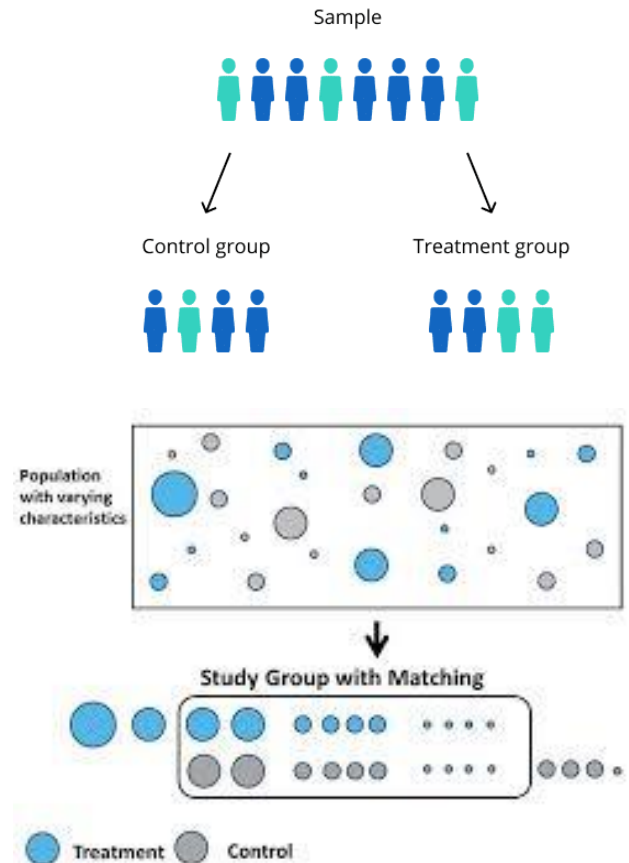
Vitamin E and Heart Disease



What at first looks like a causal relationship between IV and DV is ultimately spurious. The confounding variable is the hidden explanation.

How to minimize bias and confounding

- Sound laboratory practices
- Be aware of selection bias
- Randomization
- Blinding
- Matching
- Statistical models to adjust:
$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$
- Avoid missing data and measurement error



9. Bias and Precision

- Goal: Want to obtain results which are both ACCURATE (unbiased) and PRECISE

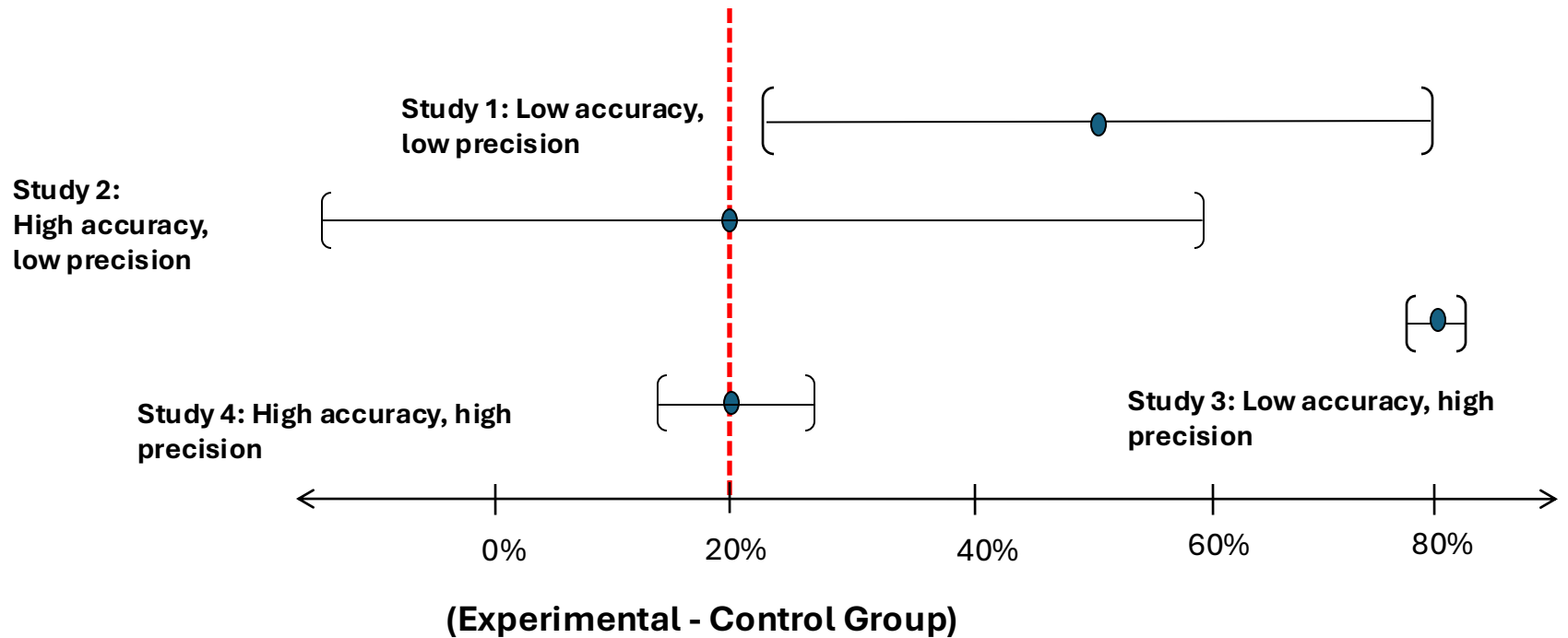


- Low accuracy: Biased results; over or underestimate effects of biological or clinical factors of interest
- Low precision: A lot of uncertainty/inconsistency in results; unable to detect true biological or clinical signal. Results less reproducible.

Recall CI

$$\bar{x} \pm 1.96 SE$$

Results from 4 Studies: 95% CI



How to improve precision

(i.e. get narrower CI)

Example: 95% CI for mean: $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$

- Decrease σ :
 - Technical variability**: Replicate and use average
 - Biological variability**: More homogeneous study population
- Increase sample size (n)

Biological Replicates

Biological Replicates:

- **Definition:** Independent samples that represent biological variation.
- **Purpose:** Estimate how much variability exists across individuals or biological systems.
- **Example:**
 - Measuring gene expression in **three different mice**.
 - Running a drug assay on **cells from three different patients**.
- **Statistical role:** These count as **independent observations** and are used for **statistical inference** (e.g., t-tests, regression).

Technical Replicates



Technical Replicates:

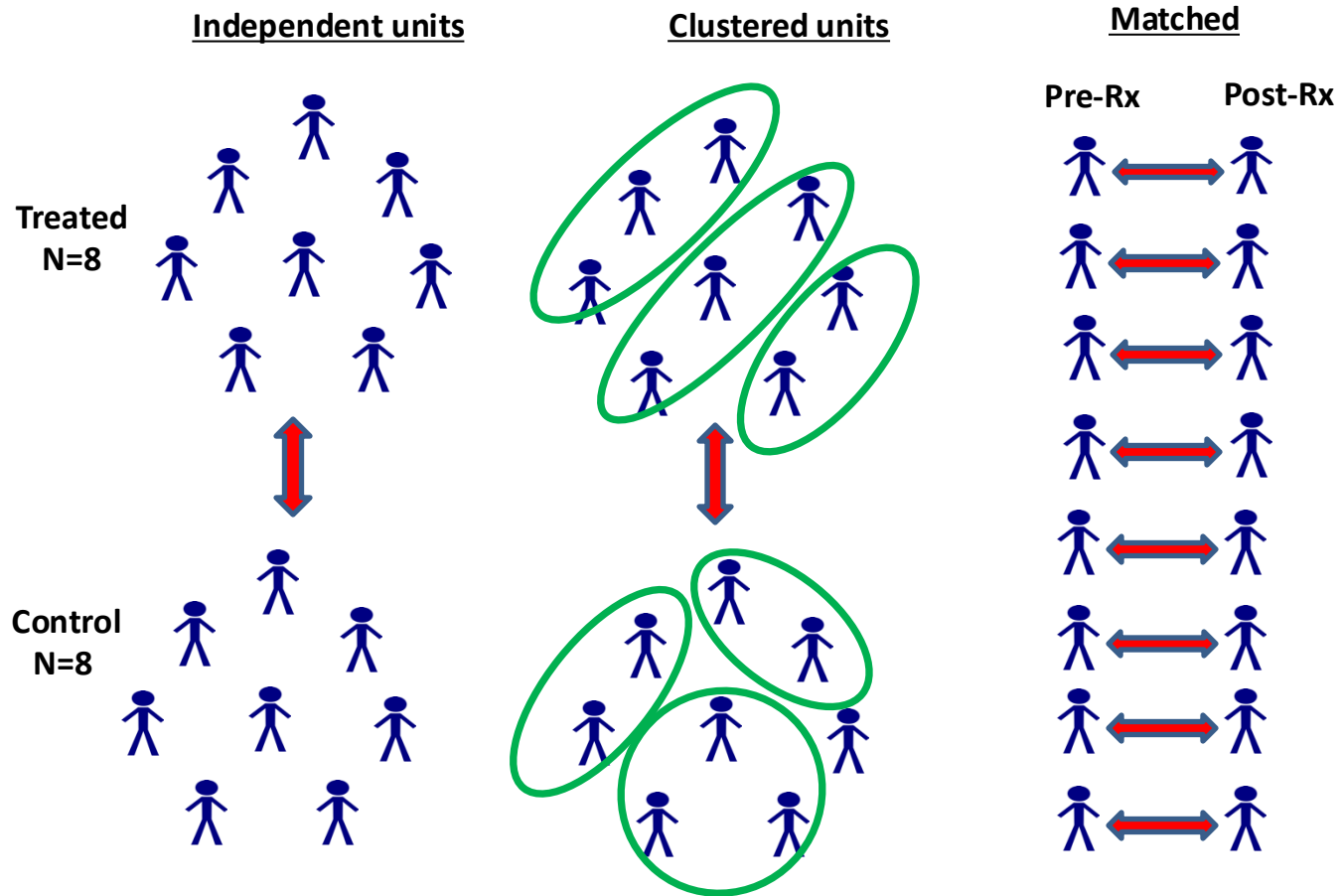
- **Definition:** Repeated measurements of the **same biological sample**.
- **Purpose:** Quantify and minimize **measurement error** or lab technique variability.
- **Example:**
 - Running the **same RNA sample** on the qPCR machine three times.
 - Reading **the same well** multiple times in a plate reader.
- **Statistical role:** Not considered independent; usually **averaged or modeled as random error**.
- **Explain the consequences of confusing them:**
 - Inflated sample size if technical replicates are treated as biological ones
 - False precision and invalid p-values
 - Technical replicates are useful for **QC** and estimating **measurement error**.
 - Mixed-effects models can explicitly model **within-sample vs. between-sample** variability when both types are used.

Biological vs Technical Replicates

The consequences of confusing them:

- Inflated sample size if technical replicates are treated as biological ones
- False precision and invalid p-values
- Technical replicates are useful for **QC** and estimating **measurement error**.
- Mixed-effects models can explicitly model **within-sample vs. between-sample** variability when both types are used.

Experimental Design Affects Efficiency



Bias vs Variance (Precision) Tradeoff in Statistical Modeling

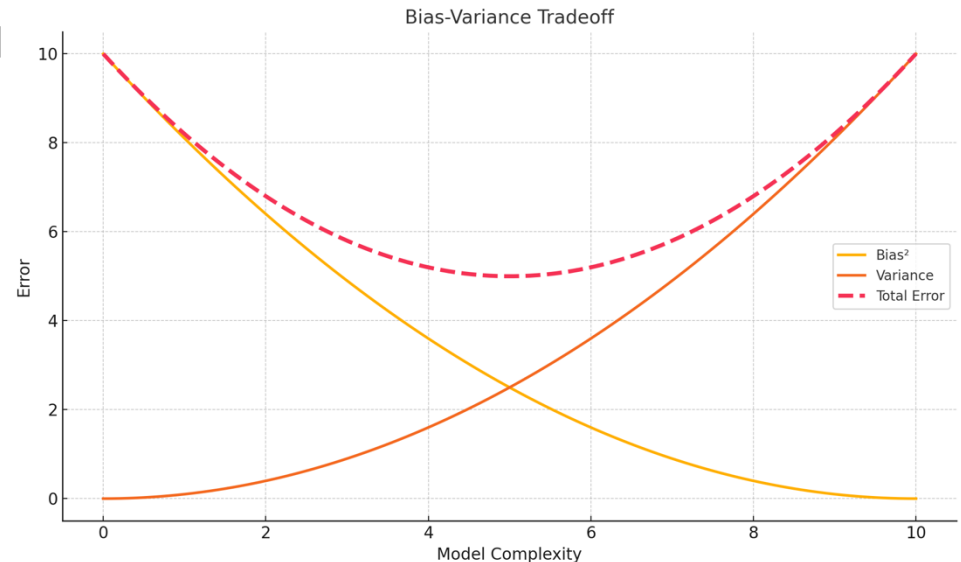
- **Prediction error is a function of bias and variance**

- **High bias (underfitting):**

- Model is too simple (linear model for non-linear data)
- Fails to capture underlying patterns

- **High variance (overfitting):**

- Model is too complex
- Captures noise as if it were signal

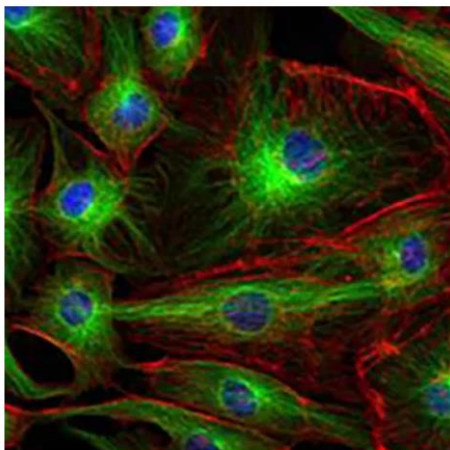


- **Bias** decreases as model complexity increases.
- **Variance** increases with model complexity.
- **Total Error** first decreases, then increases

Take home messages

- Statistical reasoning and methods are important throughout all phases of research: design, conduct, analysis, to produce rigorous and reproducible translational research
- Take advantage of training opportunities in statistical methods and reproducible research at the institution, online
- Study design affects validity and interpretability of statistical results. Rigor and reproducibility are essential for trustworthy science. Biostatistics plays a critical role in study design, analysis, and transparency.
- Collaborate with biostatisticians: TEAM SCIENCE!

AECC Biostatistics Shared Resource



Analytical Imaging Facility

Shared Resource Director: John Condeelis, PhD

The **Analytical Imaging Facility (AIF)** is a comprehensive sample preparation, histopathology, light and electron microscopy, and magnetic resonance imaging shared resource.

[Read More](#)

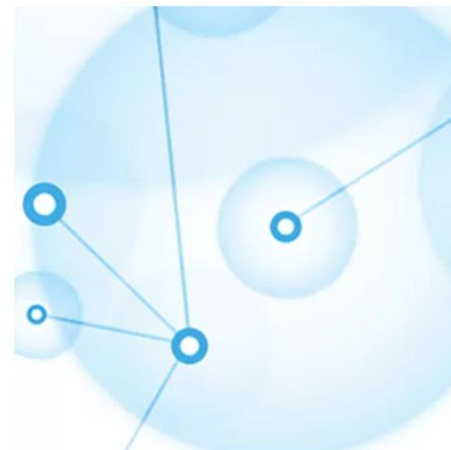


Animal Model Shared Resource

Shared Resource Director: Winfried Edelmann, PhD

The **Animal Model Shared Resource (AMSR)** provides a comprehensive approach for the generation, housing, and analysis of animal models of human cancer.

[Read More](#)



Biostatistics Shared Resource

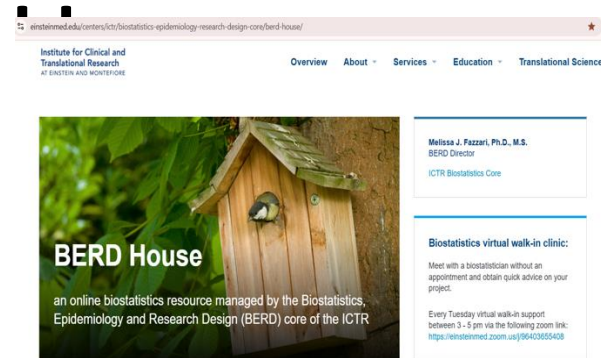
Shared Resource Director: Xiaonan Xue, PhD

The **Biostatistics Shared Resource (BSR)** provides statistical expertise and collaboration on all phases of basic science, translational, clinical and epidemiological research.

[Read More](#)

BERD

<https://einsteinmed.edu/centers/ictr/biostatistics-epidemiology-research-design-core/berd-house/>



Statistics, Machine Learning, and Data Science Training:

Getting Started	OMICS Data Analysis
Statistical Methods	Study Design and Statistical Power
Machine Learning	Journal Club
Data Science 101	Announcements and Upcoming workshops

Virtual Walk-In Stat Consulting Center: Tuesday afternoons 3-5 pm via zoom

Where to go for help

- Virtual Walk-In Stat Consulting Center:
Tuesday afternoons 3-5 pm via zoom
- Clinical Research Training Program
- Courses; online resources