Power and Significance Considerations in Epidemiology

Neal Simonsen

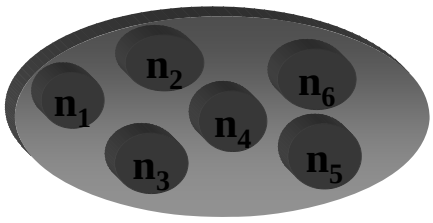
Stanley S. Scott Cancer Center &
School of Public Health

LSU Health Sciences Center

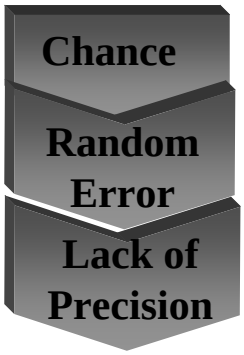
Why Epidemiology is Chancey

- In most epidemiologic studies, it is impossible to evaluate every member of the entire population. Thus, the relationship between exposure and health-related event is judged from observations on sample of the population…

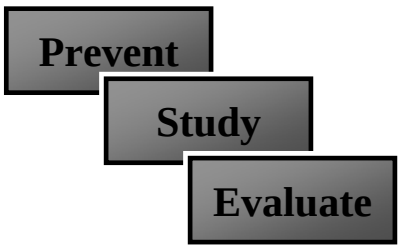
Sampling



N



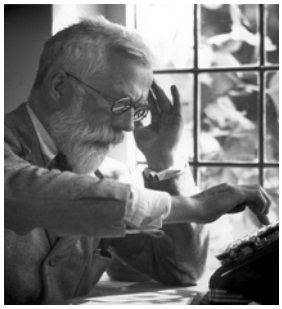
Control of Random Error



Addressing Random Error

- Two tools
  - *p* values
  - Confidence intervals (CIs)
- Notes:
  - *Neither* address systematic error
  - *p* values are frequently misused
  - CIs provide information that p-values do not

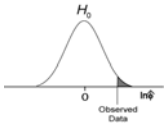
R.A. Fisher: Prototypical p-value proponent



Sterne, J. A C et al. BMJ 2001;322:226-231

Interpretation of *p*-values

- *P*-value = probability of the observed difference if there were no true difference in population
- Small *p*-value gives you evidence against the hypothesis of "no difference"
  - e.g., $p = .01$ says there is a 1% chance the observed difference could have come about randomly (~ is explained chance)
  - When the *p* value is small (say, less than .10, .05, .01) → the observed difference is unlikely due to chance → difference is labeled **statistically significant**

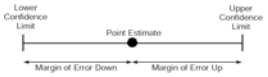


Illustrative Example (*p*-values)

Childhood socioeconomic factors and stroke mortality (Boyd Orr via Galobardes et al., *Epidemiologic Reviews*, 2004, p. 14)

Factor	Hazard Ratio	*p* value
Crowding (persons/room)	< 1.5 = 0.4 1.5 – 2.49 = 1.0 (referent) 2.5 – 3.49 = 0.6 ≥ 3.5 = 1.0	trend $p = 0.53$
Tap water	0.73	$p = 0.53$
Ventilation	very good = 1.0 (referent) fair = 1.7 poor = 1.7	trend $p = 0.08$
Cleanliness	very good= 1.1 fair = 1.0 (referent) poor = 0.5	trend $p = 0.07$

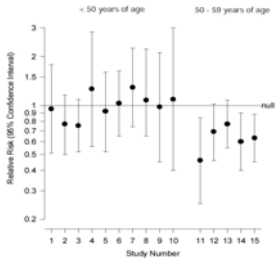
Interpretation of Confidence Intervals



- Locate parameter with "margin of error"
  - e.g., 95% confidence interval for a risk difference might be .10 ± .02
  - This is written (.08, .12), where .08 is the lower confidence limit (LCL) and .12 is the upper confidence limit (UCL)
- Confidence interval width quantifies precision
  - Narrow confidence intervals → precise
  - Wide confidence intervals → imprecise
  - e.g., a 95% CI of (.08, .12) is more precise than one of (.04, .16)
- Confidence interval width is inversely related to sample size
  - Big studies → narrow confidence intervals → precise estimates
  - Small studies → wide confidence intervals → imprecise estimates

Illustrative Example (CIs)

Breast Cancer Mortality and Mammography



Source: Hurley & Kaldor (1992)

- **Power** is the great aphrodisiac.
  - Henry Kissinger 1971

The "Straw Man" Assumption of Typical Statistical Testing?

H_0 is true

Truth, conclusions and error: spawn of the null hypothesis

The study concludes…	The TRUTH is… H_0 True	The TRUTH is… H_0 False
H_0 Accepted	CORRECT	TYPE II (β) Error
H_0 Rejected	TYPE I (α) Error	CORRECT

An alternative framing: Does the new treatment improve outcome?

The study concludes…	The TRUTH is… YES	The TRUTH is… NO
YES	CORRECT	TYPE I (α) Error
NO	TYPE II (β) Error	CORRECT

Alpha, Beta, Power…

- α : What is the chance of saying a treatment works when it really doesn't? (pα)
- β : What is the chance of saying a treatment doesn't work when it really does? (pβ)
- Power: How high is the probability that you will not say a treatment doesn't work when it really does? (1-pβ)
  - I.e.: Probability of not making a β error

- Chance of α error usually predetermined: e.g. 0.05
- Chance of β error depends on several factors:
  - the choices of the investigators
  - the size of the α error
  - the event rate in the comparison group
  - how effective the treatment is (Δ, or "magnitude of effect")
  - the number of patients in the study

"How NOT To" Hints from Ray Hoffman

- I talked with Dr. X and he liked my sample size.
- We used 15 subjects in our last study and had significant differences, so…
- The study aimed at looking into the natural history of respiratory illness in the neonatal intensive care unit. So no power calculation is needed.

Power and Sample Size: "How To"s

- The relationships mean that for a given β and a treatment effect of a given magnitude, one can estimate how many patients are needed for that treatment effect to be statistically significant.
- Chance of a Type I error almost always set at 0.05, but
- Type II errors deemed acceptable often as high as 0.10 or 0.20 (Power of 0.9 or 0.8)

Power Players

- The purpose of the study as well as the question under study will influence power considerations

E.g. Pilot Study…

- To demonstrate feasibility of obtaining samples/data
  - May only need a couple of subjects
- To get a ballpark estimate of the intra-subject variability in a particular test/measurement
  - Often 4-8 subjects suffices (repeated measurements)

Power Players (Cont'd)

Studies to demonstrate an effect and/or efficacy

- More formal power calculations called for
- Best to have estimates of:
  - Type of effect (outcome) under study
  - Magnitude of effect reasonably anticipated
  - Variability of outcome measure(s)
  - Variability of exposure (treatment, risk factor) measures
  - Frequency of outcomes expected (e.g. rate of relapse with conventional treatment from past experience)

Power and Sample Size

- If power is held constant, the greater the treatment effect, the fewer patients are needed. New treatments that only improve over traditional treatments by "small" amounts (e.g. 25%) can require large sample sizes to demonstrate a statistically significant improvement.
- Outcome events, not patients, are the most important determinant: for a given RR, a mortality study with 1000 people, 50 of whom die, is only slightly stronger than a study of 100 people, 50 of whom die.

Example

- Trial between streptokinase and tPA: Investigators wanted to, with a power of 0.9, be able to detect a 15% reduction in mortality, with the baseline estimated at 8%.
  - Small magnitude of effect, small anticipated rate of events
  - End result: 41,000 persons needed for trial.

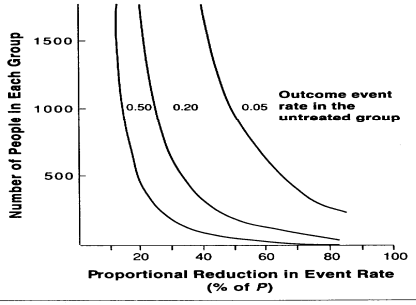


Figure 9.3. The number of people required in each of two treatment groups (of equal size) to have an 80% chance of detecting a difference ($p = 0.05$) in a given outcome rate (P) between treated and untreated patients, for various rates of outcome events in the untreated group. (Calculated from formula in Weiss NS. Clinical epidemiology. The study of the outcome of illness. New York: Oxford University Press, 1986.)

Power Statement: Example

"This study was designed to detect a 25% therapeutic difference between groups, assuming a baseline rate of 10% recurrence in the untreated group. With an alpha level of 0.05 and a power of .9, this resulted in the need for 360 patients per group."

- What is the actual rate of recurrence in the treated group that will result in statistical significance?

Power statement: deciphering the code…

1. "This study was designed to detect a 25% therapeutic difference between groups ...."

- How effective is the new treatment? How much improvement over the standard treatment will occur when the new treatment is used?

2. "...assuming a baseline rate of 10% recurrence in the untreated group...."

- What is the rate of recurrence in the untreated group, based on your best estimate from the available literature ± pilot data?

3. "With an alpha level of 0.05 and a power of .9, this resulted in the need for 360 patients per group."

- We are accepting 0.05 as the chance of an error saying the treatment works when it doesn't,…
- and 0.10 (i.e., 1 – 0.9) as the chance of an error saying the treatment doesn't work when it actually does.
- Given this information, sample size calculations indicate that we need 360 patients per group (= 720 patients overall) to assure a 90% chance of finding a statistically significant result <u>if</u> the improvement really is at least 25% .

So what's significant?

- To improve the 10% baseline recurrence rate by 25%, need to prevent .25 X .1 recurrences, or .025 (2.5%).
- Subtracting those from baseline recurrence of 10%, we are aiming to find a recurrence rate in the treated group of 7.5% or lower.
- If the rate is lower than 7.5%, that means we decreased the baseline rate by greater than 25%.....

Statistical and Clinical Significance

- The scenarios that follow are each discussing a different therapy. Assume that the treatment involved is not prohibitively costly, does not cause an unusual amount of side effects, and is relatively safe.
- Would you use the following treatments in your patients?

Therapy Example One

A new treatment results in a 36% relative decrease in distant metastasis over a five-year period, which is statistically significant. The 95% confidence interval ranges from a 57% decrease to a 9% decrease.

Therapy Example Two

A medical intervention results in a 1.4% (absolute) increase in recurrence-free survival (3.9% vs. 2.5%). The increase is statistically significant. The 95% confidence interval ranges from a 2.5% increase to a 0.3% increase in recurrence-free survival.

Significance

- Statistical significance: is this difference likely to be nonrandom in origin?
- Clinical significance: is this difference likely to be clinically important?
- Both are influenced by sample size:
  - If small, clinically significant differences may not reach statistical significance.
  - If large, clinically insignificant results may be found to be statistically significant.

A Rule of Thumb for Power in Trials

- **Trial outcomes usually described as decreased risk of experiencing an outcome in the treated RELATIVE to the risk in the controls. Here are 3 different studies:**

	Event Rate		
Untreated	**50%**	**20%**	**2%**
Treated	**25%**	**10%**	**1%**
Relative Risk Reduction	**50%**	**50%**	**50%**

All these treatments have the same RRR, but not the same utility. How much do they decrease the absolute risk of an outcome?

	Event Rate		
Untreated	50%	20%	2%
Treated	25%	10%	1%
Relative Risk Reduction	50%	50%	50%
Absolute Risk Reduction	**25%**	**10%**	**1%**

How many patients must be treated to prevent one outcome from occurring? Estimate by taking the inverse of the ARR. This is called the Number Needed to Treat (NNT).

Untreated risk	**50%**	**20%**	**2%**
Treated risk	**25%**	**10%**	**1%**
Relative Risk Reduction	**50%**	**50%**	**50%**
Absolute Risk Reduction	**25%**	**10%**	**1%**
NNT	**4 (1/.25)**	**10 (1/.10)**	**100 (1/.01)**

"We reviewed all 383 RCTs published in JAMA, Lancet, and the New England Journal of Medicine in 1975, 1980, 1985, and 1990…

- **Most trials with negative results did not have large enough sample sizes to detect a 25% or a 50% relative difference.**
- **This result has not changed over time.**
- **Few trials discussed whether the observed differences were clinically important….**
- **The reporting of statistical power and sample size also needs to be improved."**
  - **JAMA 1994 Jul 13;272(2):122-4**

Adequacy of Power: Not Just a Statistical Issue

- Ethical issues arise, especially in clinical studies
- Too large a sample size?
  - Excessively large number of patients and/or duration of study may be unethical: overkill and/or delay
- Too small a sample size?
  - Hopelessly small sample may be unethical: probably too few subjects to show anything even if effects <u>are</u> present

The Kiss again, before we part...

- **Power** is the great aphrodisiac.
  - Henry Kissinger 1971

- The reality…
- Lack of power is a real turn-off for investigators and the research community in general
  - Perhaps we should rephrase…
- "Adequate statistical power is the great aphrodisiac."
  - Makes not just you, but others love your results, too

- Don't get swept away, though…

  you want a meaningful relationship:
- Statistical significance is no substitute for clinical/biological significance
- Statistical significance is no substitute for validity:
  - If the results are driven by bias, no amount of power can assure that you've found a meaningful relationship