PG<2%: A CCGR workshop on research challenges with a small population

Vancouver, February 2, 2015

Session: Quantitative Methods for Small Samples

Susan Bondy

UNIVERSITY OF TORONTO
DALLA LANA SCHOOL OF PUBLIC HEALTH

Approach for this session

hallenges of rare vents

Small numerators

- Estimating prevalence with precision
- Zero cell analyses

Small denominators

- Analysis options with small datasets

Finite target populations

Resources and methods to draw upon

- **Survey** sampling theory and methods of analysis
- **Epidemiologic** sampling designs
- **Statistical analyses** for small samples
  - Classical methods
  - Emerging methods

Estimating rates (usually surveys)

With rare events…

Simple Random Sampling, and

True *representivity* of population surveys

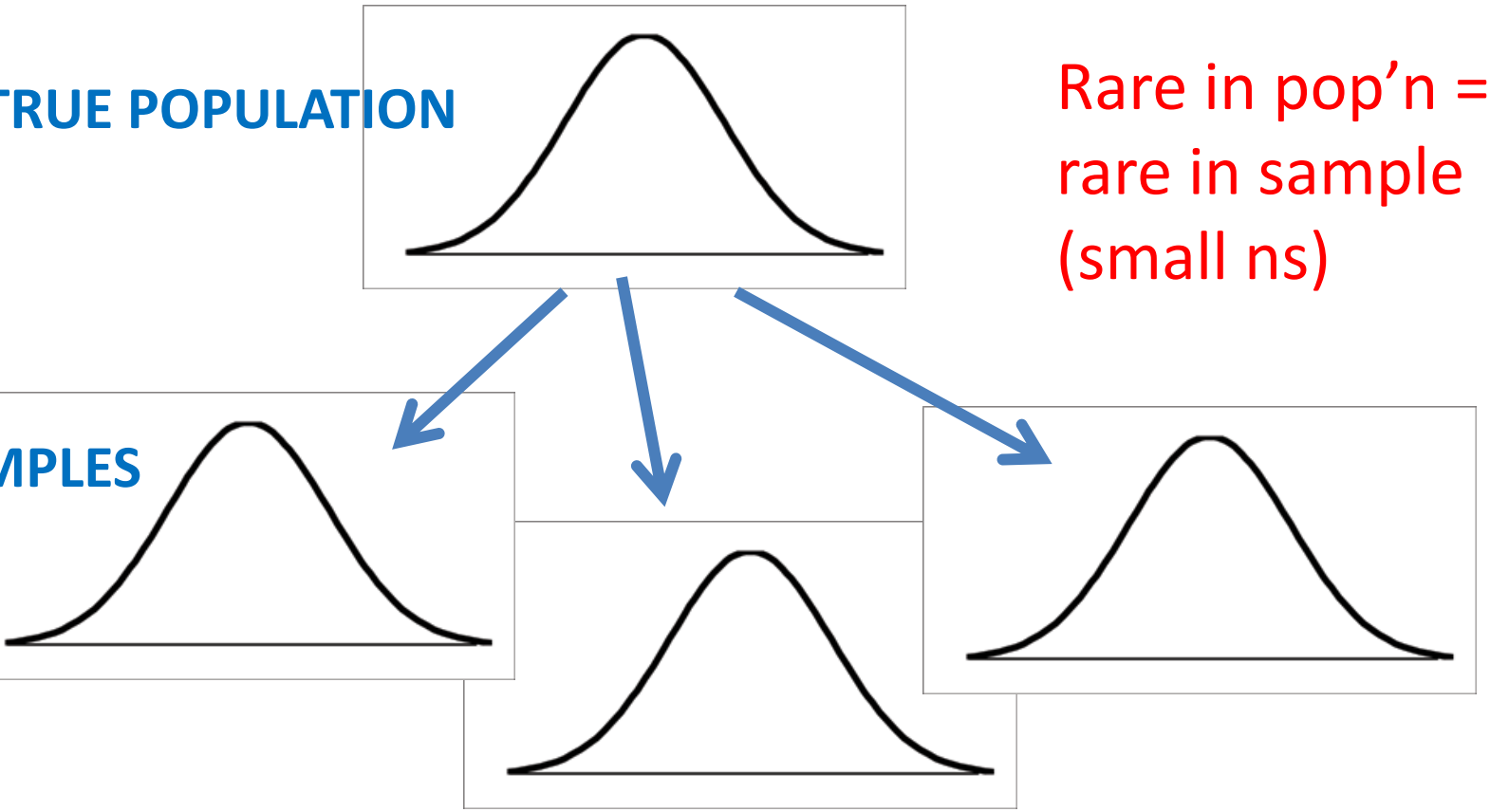
Are not your best friends

Representivity"

he sample contains all the unique subgroups as
the true population; in the same proportions

Truly random samples tend to imitate full population

(with large samples 'representivity' happens, usually)



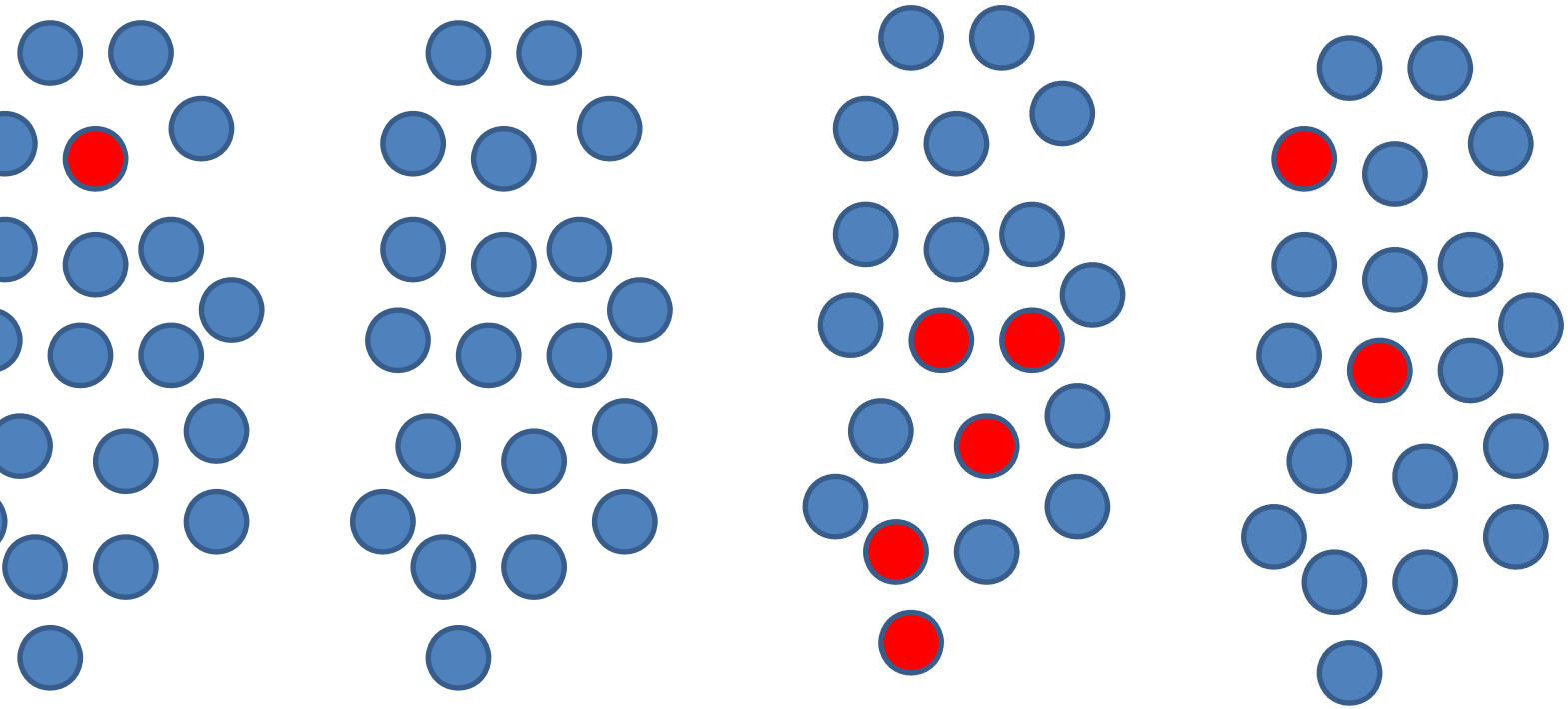
Rare in pop'n = rare in sample (small ns)

stimates of MEAN VALUES are pretty good (precise, reliable)

(low random sampling variability; low Margin of Error - MOE)

BUT, simple chance makes N or % of rare events unreliable

prevalence hard to estimate and report



in 20 0 in 20 5 in 20 2 in 20

tly possible variation by chance in Simple Random Samples (SRS)

Whenever you have any say in the survey or study design,

Do STRATIFIED SAMPLING

Sub-divide target pop into layers (strata);
Sample (at random) within each layer;
Layer by strongest predictors of gambling (PG) you can identify and use

TYPE 1: Explicit stratification,

or 'over-sampling'

Ensure min. N in every subgroup of interest

Higher sampling fraction in smaller groups

- Why it's called 'over-sampling'
- Form of 'quota' sampling, but <u>random</u> within strata

wever, you can't directly oversample PGs

if you already knew who was PG, no need for study

over-sample groups with KNOWN higher gambling ates (e.g., by age, sex, geography)

– Will get more PGs in total

– Don't worry, sampling weights will restore representivity, overall

Stratified sampling
TYPE 2: Probability Proportionate to Size, of stratum (PPS)

Again, use demographic and other factors associated with HIGH AND LOW rates of problem gambling

Random sample in each sub-group (stratum)

- As per previous slide

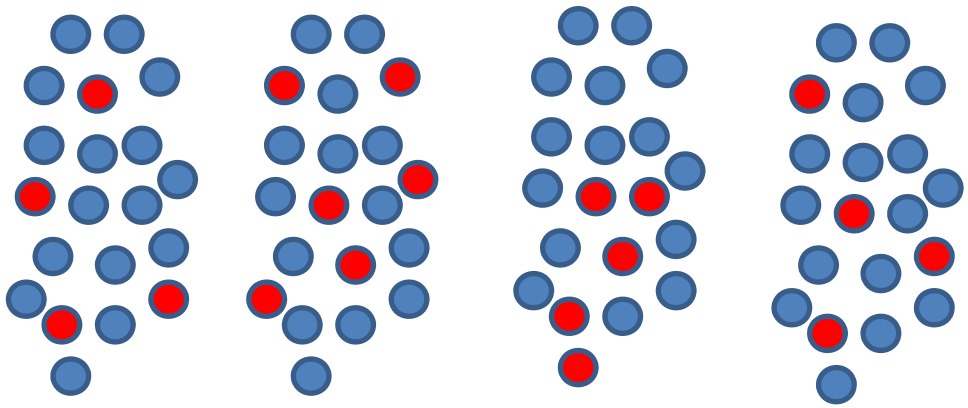
This time FORCE representivity (PPS)

- E.g., a group is 10% of pop'n; must be 10% of sample

Using stratification info. with survey software often gains some precision!

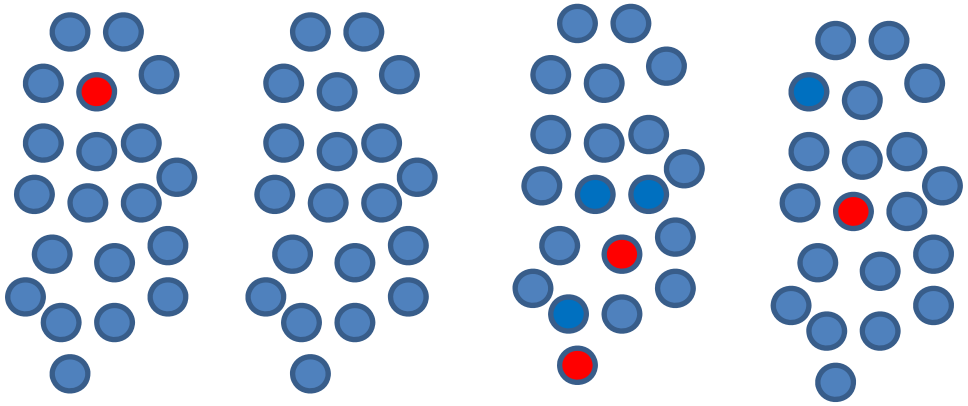
Two demographic or predictor groups:
one high rate, one low rate

risk,
of



Oversampling: boosts N gamblers

isk,
of



USE software for stratified sampling.

Pooled prevalence estimates have more precision.

5/40 6/40 7/40 5/40

pidemiologic designs for rare events

Case-control study

- **Gather cases from richest sources**
  - Representative sample not formally required
  - Rigorously matched comparison group
    - Nothing casual about this, rules apply
    - Selection and analysis designed to test on specified predictors
- **Designed for hypotheses of etiology**
  - Typically, the pop description value is poor
  - Often dismissed, but can be far stronger than reputation

Case-case study

- Groups of frequent gamblers – divided problem or not
- Analysis focused on factors that discriminate between sub-types
- Can be efficient; many other things held constant or more easily matched

But usually you have small numbers; now what?

Analysis methods

Mini-outline:

Parametric methods

Rare events and zero cells

Non-parametric methods

Classical methods for small samples

Emerging methods

Moral of story = Gaining or Salvaging Statistical Power and Precision

Starting with simplest analysis

Remember to review and weigh all your analysis options

Be conservative – but not excessively

Estimating a simple %

Resulting 95%CI, by method used

w data: 0% and 2% PG in 30 observations

thod	0 of 30 observed Reported 95% CI	2 of 30 observed Reported 95% CI
rmal approx. ld	0.0 – 0.0 % (Cannot be used)	1.8 – 21.3%
esti-Coull son	0.0 – 13.5% 0.0 – 11.4%	0.8 – 22.4% **1.8 – 21.3% (rec'd openepi)**
omial exact	0.0 – 11.6%	0.8 - 21.0%
sson exact	0.0 – 12.3%	0.8 – 24.0% (maybe over-conservative)

ration only. Cells fewer than 5 observations may break data disclosure rules.

For simple, small samples
www.openepi.com

Open Source Epidemiologic Statistics for Public Health

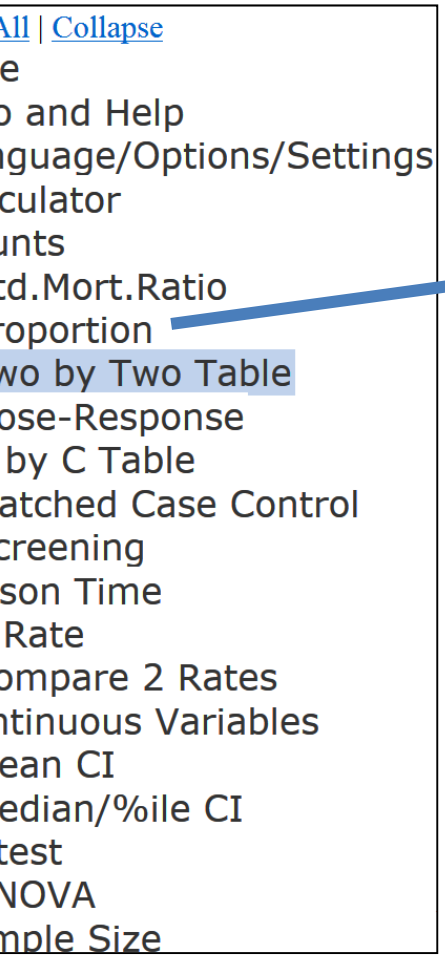
w in English, French, Spanish, Italian, and Portuguese

rsion 3.03 Updated 2014/09/22 ***Try it in a Smartphone browser!***

OpenEpi provides statistics for counts and measurements in descriptive and analytic studies, stratified analysis with exact confidence limits, matched pair and person-time analysis, sample size and power calculations, random numbers, sensitivity, specificity and other evaluation statistics, R x C tables, chi-square for dose-response, and links to other useful sites.

OpenEpi is free and **open source** software for epidemiologic statistics. It can be run from a web server or downloaded and run without a web connection. A server is not required. The programs are written in JavaScript and HTML, and should be compatible with recent Linux, Mac, and PC browsers, regardless of operating system. (If you are seeing this, your browser settings are allowing JavaScript.) The

Multiple methods – annotated with "editor's choice"



Proportion
Confidence Limits for a Single Proportion

This module provides confidence limits for simple (binomial) proportions. Entering a numerator and denominator produces confidence limits calculated by several different methods. The numerator must be smaller than the denominator and both must be positive numbers.

95% Confidence Limits for Proportion 10/11
Multiplier=100
Large population size or sample with replacement.

	Lower CL	Per 100	Upper CL
Proportion as Percent		90.9091	
Mid-P Exact	62.66		99.55
Fisher Exact(Clopper-Pearson)	58.72		99.77
Wald (Normal Approx.)	73.92		100
Modified Wald(Agresti-Coull)	60.1		100
Score(Wilson)*	62.27		98.38
Score with Continuity Correction (Fleiss Quadratic)	57.12		99.52

The npq of 0.9091 is <5. The Wald method is not recommended.

Don’t be too conservative

“Exact” methods

- Sound ‘best’
- Popular, and increasingly available
- With very small n, can even be TOO conservative
- See mid-p, for example of a less conservative modification (see tutorial in www.openepi.com)

For simple estimates and tests

- Free Openepi.com is one easy source for running options and guide to selection and interpretation
- Simple explanations

Finite Population Correction (FPC)

Use with extreme caution, but there MAY be times when this is applicable

What is it?

- Random sampling (normal p-values) assume 'sampling with replacement'
  - Often said, well, sample is small relative to full population
  - IF your sample is 50%+ of whole, tell your software to use FPC and p-values and limits shrink – sometimes dramatically
- Example: you know answer for 10 people – all say yes
- When you have 10 answers, it can only range from 8-10

Finite Population Correction (FPC)

In Science a 'knee-jerk' response against it

- BUT, it DOES have a role in program eval. and quality assurance

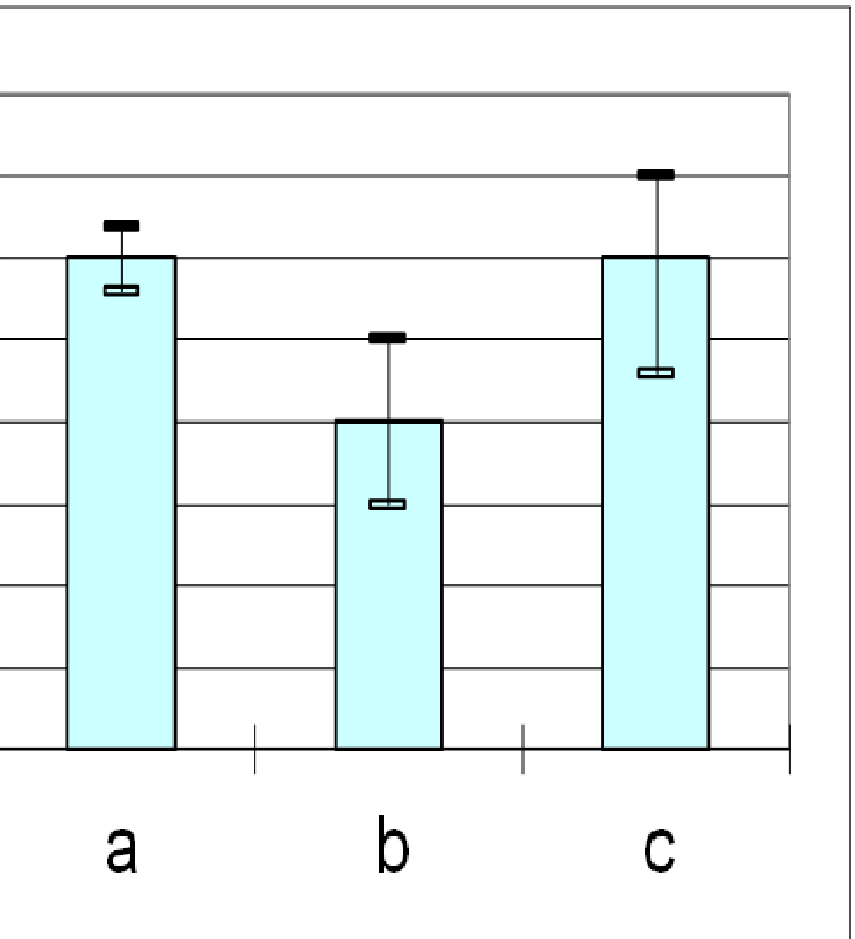
FPC MAY be applicable, IF one statistic is supposed to only apply to a singular, clearly defined total pop

- Note- even full-population health data, not used
- MAY be applicable in program evaluation where you truly want only to describe outcome in very specific groups of people
  - E.g., recidivism with 30 pple, 25 of whom have follow-up

Cannot then apply the confidence limits to OTHER populations

Also too <u>conservative</u>:

Significance judged by non-overlapping intervals

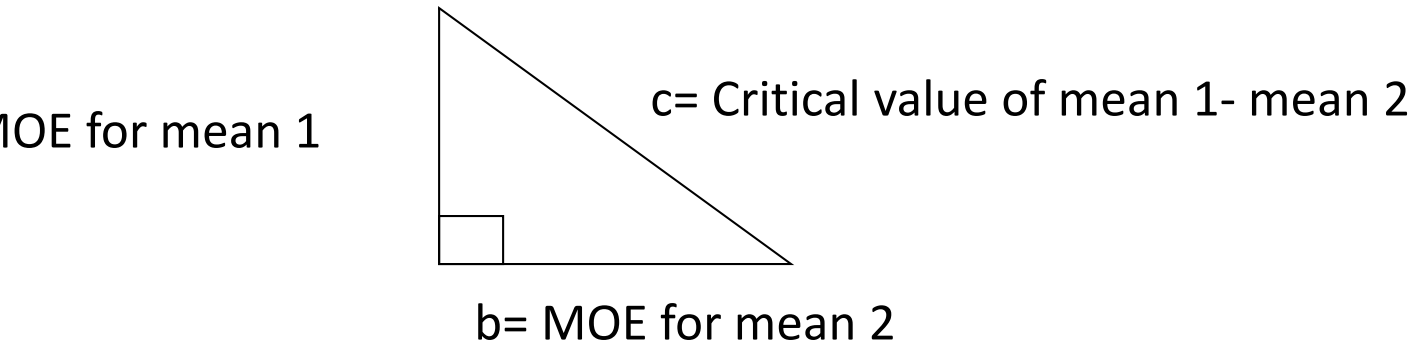


- Only a-b are significantly different, right?
- In truth, b-c are as well.

- Perform the actual contrast test, instead
  - (raw or published data)

Pythagorean theorem

"The square of the [critical diff between 2 means] = the sum of the squares of the two Margins of Error (MOE)"

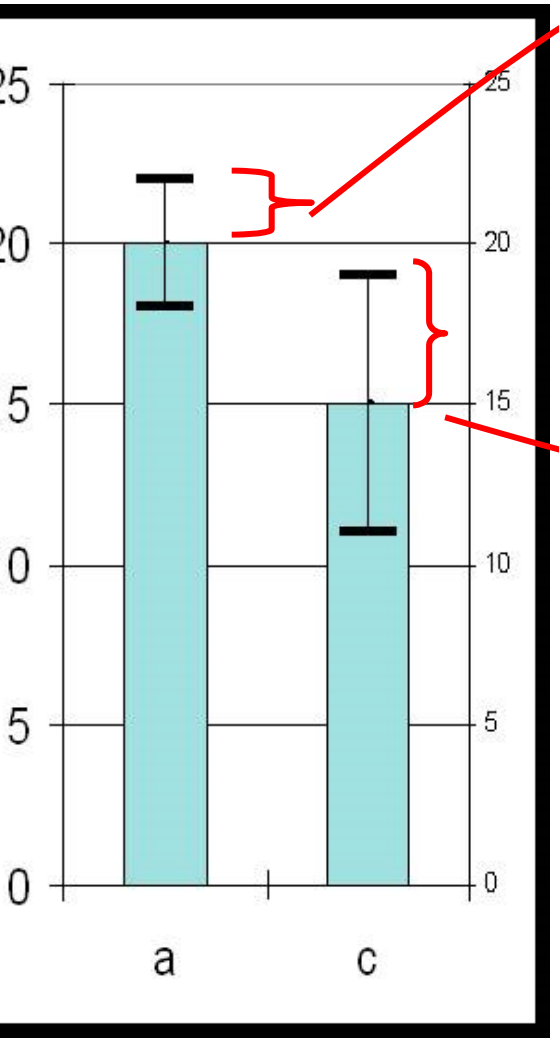


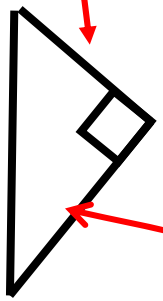
- Length of c

  = square root of ($a^2 + b^2$)

- SE of $diff_{a-b}$ = root of ($se_a^2 + se_b^2$)

For printed CIs, take the corner of a piece of paper and fold it according to the observed confidence limits, like this…

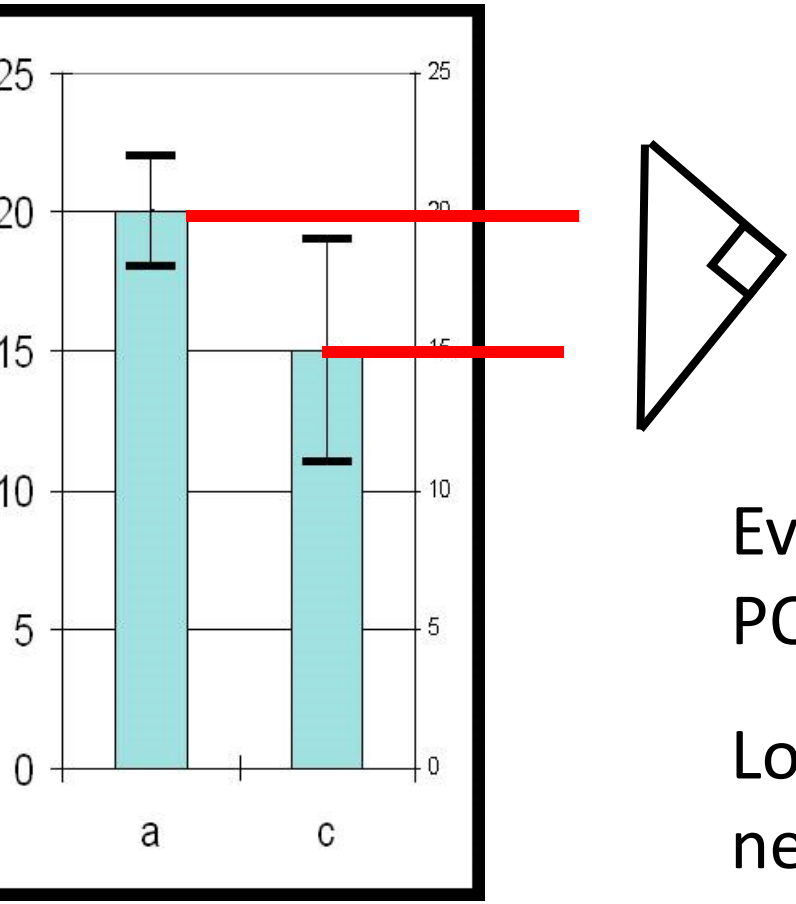




Define a right angle tri-angle.

½ of each 95% CI makes a side

For printed CIs, take the corner of a piece of paper and fold it according to the observed confidence limits, like this…



Evaluate distance BETWEEN POINT ESTIMATES

Long side shows distance needed (greater = signif)

Inefficient data analysis methods

Squandering measurement precision

- Don’t over-use cut-points or collapse values (bins)
  - Continuous & ordinal = more statistical power
  - Dichotomizing creates measurement error, often bias, and inadequate control for covariates

Revisit your bivariate and regression methods

- Y-variable - Regression models for any form!
- X-variables never had to be continuous & Normal
- Software packages and manuals always getting better
  - adding graphs and other options to aid

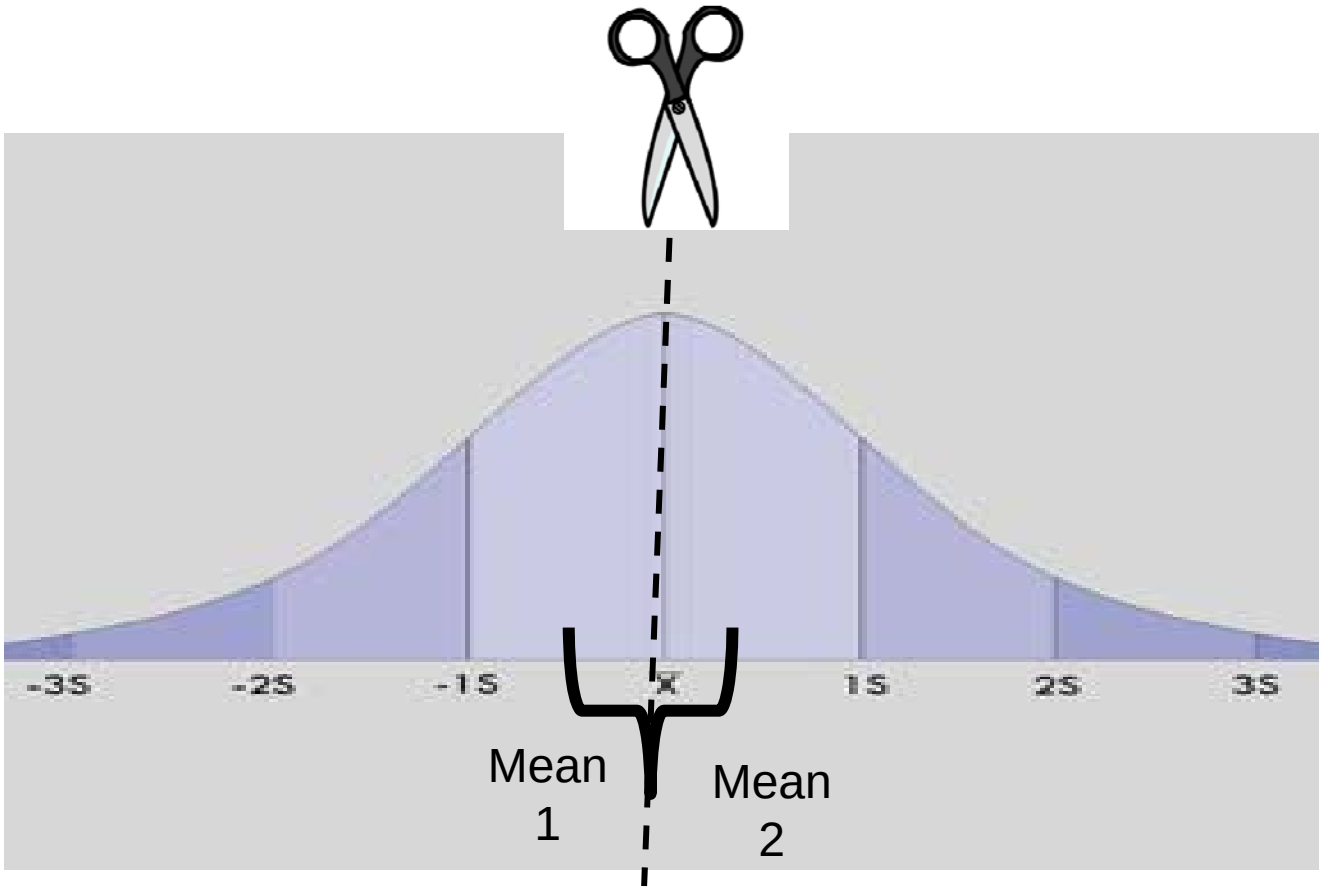
Even simple analyses - Do not always collapse to 2 by 2 tables

Small, simpler datasets,

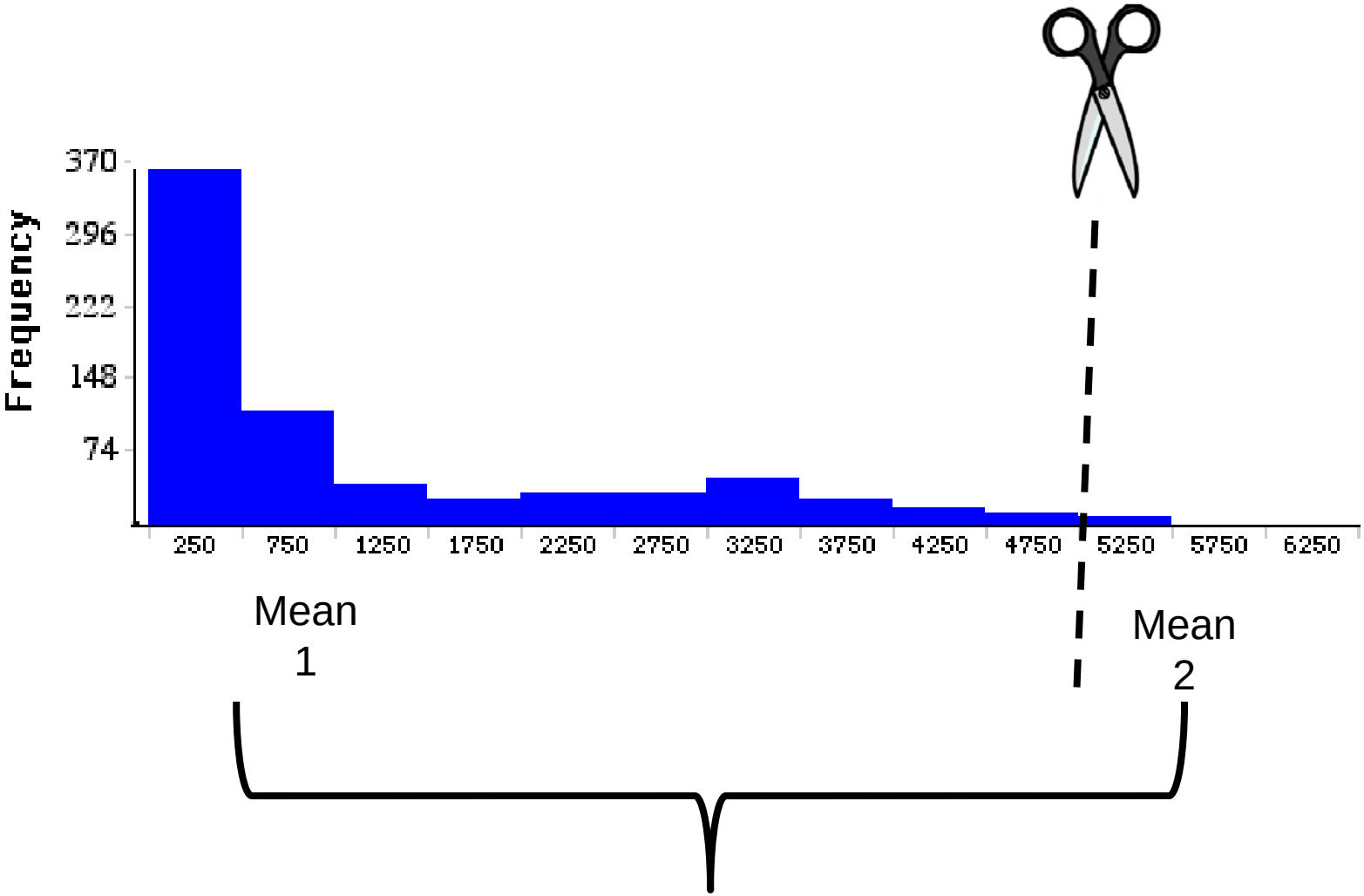
- Broad mix of tools for bivariate tests
  - Spearman correlations, rank order tests, etc.
  - Associations across many categories
  - So, keep more levels

“older” epi or biostats texts

- Many texts simplify all examples to 2 by 2 tables
- More advanced ones tend to emphasize generalized regression methods



Contrasting the groups above v. below the mean may actually be asking to find a tiny difference as



Contrasting groups at opposite ends of a spectrum may make a trivial association seem meaningful. Better is correlation per

Another inefficient analysis method

Sub-setting data to study things like gender-effects

- In simpler reports, all estimates are based on small and smaller and smaller subgroups of respondents
  - Females – Males
  - Ages 18-34, and 35+
  - Females AND under 35, etc.

Greater risks of breaching confidentiality rules

- May have to suppress estimates

Inefficient, statistically

Use current regression methods

Group-specific descriptive statistics

Example: % or mean, just women, under 35, urban

- Select only those people, yields
- Imprecise estimates; may suppressed for privacy (cells < 5)

Instead, use regression-fitted means or proportions (for specified combos of predictors)

= 1*B(sex=F) + 3B(age_10years) + B(interactions) + etc…

“Fitted estimates” with confidence intervals:

- Predicted value of outcome, get 95% CIs too
- More accessible with current software versions

Same for "Interaction" or "Effect Modification" (as per your discipline)

Example:

- Odds Ratio, in men only
- Odds Ratio, in women only

Commonly seen: Select just men, and then just women. . .

Better: Model the <u>conditional</u> effect

- **covariate-level specific OR, RR., etc.**
  - Software tends now to include options
    e.g., SAS logistic command
    - Ask for odds ratio (of interest) by levels of age, sex, etc.
    - Statistician can develop these for you as custom 'linear combinations'

ession books: Agresti, Hosmer, Kleinbaum, Vittinghoff

Regression methods for ‘zero’ cell values

Count-based models (Poisson regression and variants) are not offended by zeros (these are valid random events)

Bootstrap and other ‘re-sampling methods’

General class of statistical methods – designed for when ‘large sample’ methods not appropriate

Bootstrapping now the most popular form of re-sampling estimation

- Replaces Jacknife and others

Your study dataset become the sampling frame

- Carefully-defined random sub-samples drawn
- Repeated 500 to 6000+ times
- Each time your analysis is repeated and results stored
- OUTPUT is evaluated to generate a new “variance showing the variability across an infinite number of random samples”

Bootstrapping

Whatever the problem, the answer is bootstrapping"

Yes, it works when samples get smaller than needed for 'guideline' use of large sample methods (e.g., most regression models)

Yes, there are SOME limits.

- Should have enough data for it to be POSSIBLE to create random sub-samples that are unique, 100-500+ times
- Might work if PGs are quite few, so long as at least lots of matched controls you could select from (and other situations)

Bootstrapping

Now an option you can add to familiar analysis in several packages

- Stata, SAS etc.

There are pitfalls setting it up

Must understand sampling underlying study design and knowledge of sampling theory

- A consulting statistician can probably set you up the first time you use this method (per data set),
- Once you have your 'BS' settings down, can be applied to many forms of analysis

Emerging areas (expanding?)

Small area estimation (Domain analysis)

- Use: In large survey samples, need to estimate rates (etc.,) in local geographical areas (smaller than ideal samples)
  - I.e., small samples within huge sample
- Multi-level modeling and estimate rates for that place, based on characteristics of individuals (there) and characteristics of the setting, relative to other locations type
- Estimating rates in small geographical segments within larger survey data
  - E.g.,

Emerging areas (expanding?)

Bayesian analysis

- Now enormously popular in medical and health services research
- Truly NOT an expert
- A lot depends on validity of priors
- Raised mostly as an area to watch (jump when it's ready for prime time)