Epi Methods Subcommittee Webinar: Skill Refresher: Calculation of Tests of Trends in Proportions and Rates for Public Health

Thursday, March 29at 2pm ET

CSTE

Council of State and Territorial Epidemiologists

Housekeeping

CSTE

- Mute your lines
- The slides and session recording will be archived to the CSTE Webinar Library
- Use the chat box feature to ask the presenter questions
- Complete your evaluation after the webinar
  - Webinar feedback
  - Ideas for future webinars

Announcements

CSTE

- Registration **for the 2018 Annual Conference** opens February 1st
- Next **Spatial Analysis Workgroup Call** will be held on Tuesday, April 17th at 1pm ET, contact Jarrazola@cste.org for more details
- Next **Public Health and Health Care Analytics Workgroup Call** will be held on Thursday, April 12th at 1pm ET, contact Jarrazola@cste.org

Objectives

1. State why trends in proportions are important
2. Calculate trend-related measures using free web applications
3. Calculate trend-related measures using R, that can be expanded for other “real world” production projects.

Presenter

CSTE

Michael Samuel, DrPH, Data Scientist/Senior Epidemiologist at the California Department of Public Health

Refresher on Assessing Trends in Proportions (and Related Issues)

CSTE Training session
Thursday, March 29, 2018

Michael C. Samuel, Dr.P.H.
Data Scientist / Senior Epidemiologist
Fusion Center for Strategic Development and External Relations
California Department of Public Health

Outline

- All materials at: http://www.goo.gl/k9YmXJ
- Last time
  - Some theory and formulas
  - Calculation of simple confidence intervals for proportions, by "hand", in Excel, in Open Epi, and in R
  - Comparing proportions and Rates
- Chi-Squared tests theory
- Chi-Squared trend test
  - What, why and how
- A note on confidence intervals
- Interactive application in R
- Resources

community	cause	pop	deaths		rate
Jefferson	All Causes	6278	169		2691.9
Jefferson	Alzheimer's	6278	9		143.4
Jefferson	Cardiovascular	6278	59		939.8
River	All Causes	4660	43		922.7
River	Alzheimer's	4660	3		64.4
River	Cardiovascular	4660	15		321.9
Rose	All Causes	2312	38		1643.6
Rose	Alzheimer's	2312	2		86.5
Rose	Cardiovascular	2312	15		648.8

community	surveyPop	highSchool		percent at least high school
Jefferson	622	578		92.9%
River	460	216		47.0%
Rose	229	224		97.8%

Number of Sugary Groups Eaten Last	Obese Yes	Obese No	Total	% Obese
1	1	8	9	11.1%
2	2	8	10	20.0%
3	4	7	11	36.4%
4	8	8	16	50.0%
Total	15	31	46	32.6%

Favorite Vegy	Obese Yes	Obese No	Total	% Obese
Broccoli	1	8	9	11.1%
Carrots	2	9	11	18.2%
Lettuce	8	7	15	53.3%
Tomato	4	7	11	36.4%
Total	15	31	46	32.6%

Chi-Squared tests

- Used to assess association (versus independence) in a contingency ("cross tabulation") table
- Assess whether or not observed counts agree with expected counts, based on assumption of independence
- Can be used for
  - "2 by 2" table (two proportions, Relative Risk, Odds Ratio)
  - N x K table
  - 2 by K table with ordered values (trend test)

Chi-Squared Tests

- In general, the higher the chi-square value, the greater the likelihood there is a statistically significant difference between the groups you are comparing
- To know for sure, you need to determine the p-value with your software or in a chi-square table
- Many variants and nuances to the simple tests

General Form/Idea of Chi-Squared Statistic

O_{11}	**O_{12}**	**.**	**.**	**N1. = Row 1 Total**
O_{21}	**O_{ij}**	**.**	**.**	**N2. = Row 2 Total**
.	**.**	**O_{ij}**	**.**	**.**
.	**.**	**.**	**.**	**.**
.	**.**	**.**	**.**	**.**
	.	**.**	**.**	**.**
.		**.**	**O_{64}**	**.**
N.1 = Col	**.**	**.**	**.**	**N.. = Grand Total**

- O_{ij} = observed count
- E_{ij} = Expected count = $\frac{\text{Row i total} * \text{Column j total}}{\text{Grand Total}} = \frac{N_{i.} * N_{.j}}{N_{..}}$
- $\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
- Compare to χ^2 with (number of rows-1)*(number of columns -1) “degrees of freedom”

High School Diploma?

Observed		Yes	No		
Community	Jefferson	578	44	622	92.9%
	River	224	5	229	97.8%

	Yes	No	
Jefferson			622
River			229
	802	49	851

$$\frac{N_{i.} * N_{.j}}{N_{..}} \rightarrow$$

Expected

586.2	35.8
215.8	13.2

Chi-squared contribution:

$\frac{(\mathbf{O}-\mathbf{E})^2}{\mathbf{E}}$	$\frac{(\mathbf{O}-\mathbf{E})^2}{\mathbf{E}}$
$\frac{(\mathbf{O}-\mathbf{E})^2}{\mathbf{E}}$	$\frac{(\mathbf{O}-\mathbf{E})^2}{\mathbf{E}}$

0.11	1.87
0.31	5.08

$\Sigma = 7.38$

http://www.openepi.com

Start | Enter | Results | Examples | Help

Clear | Settings Conf. level=95% | Calculate

Add Stratum | Stratum 1 | Delete Stratum

Open Epi 2 x 2 Table

		Disease		Totals
		(+)	(-)	
Exposure	(+)	578	44	
	(-)	224		
Totals		802		

2 x 2 Table Statistics

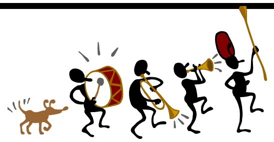
Single Table Analysis

		Disease	
		(+)	(-)
	(+)	578	44622
Exposure	(-)	224	5 229
		802	49851

Chi Square and Exact Measures of Association

Test	Value	p-value(1-tail)	p-value(2-tail)
Uncorrected chi square	7.377	0.003303	0.006605
Yates corrected chi square	6.504	0.005383	0.01077
Mantel-Haenszel chi square	7.369	0.003318	0.006637
Fisher exact		0.003088(P)	0.006176
Mid-P exact		0.001953(P)	0.003906

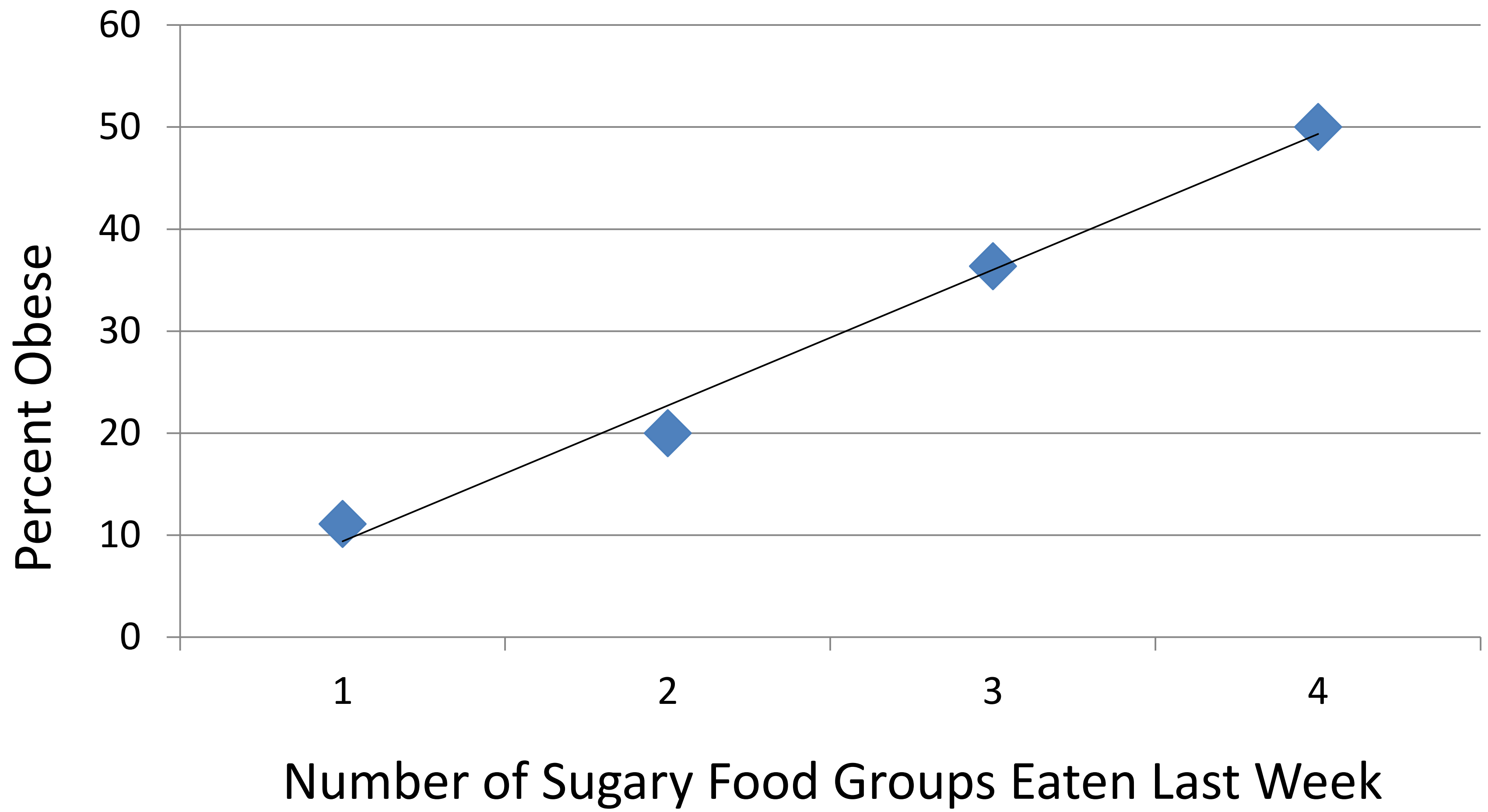
Types of Statistical Tests

Parade of Statistics Guys	
The right test…	***To use when… .***
Pearson chi-square (uncorrected)	Sample size >100 Expected cell counts > 10
Yates chi-square (corrected)	Sample size >30 Expected cell counts ≥ 5
Mantel-Haenszel chi-square	Sample size > 30 Variables are ordinal
Fisher's exact test	Sample size < 30 and/or Expected cell counts < 5

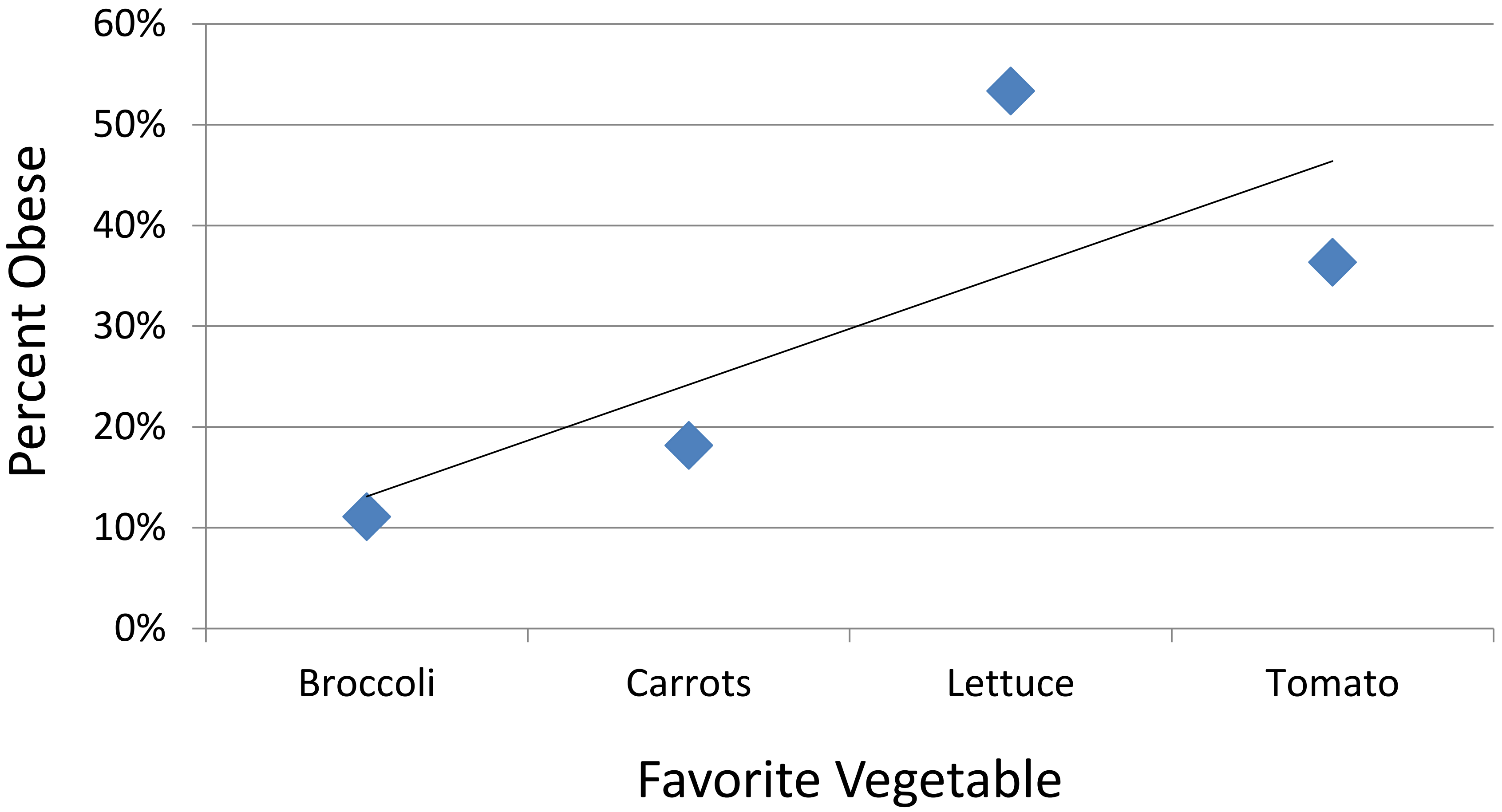
https://sph.unc.edu/nciph/focus/

Number of Sugary Groups Eaten Last	Obese		Total	% Obese
	Yes	No		
1	1	8	9	11.1%
2	2	8	10	20.0%
3	4	7	11	36.4%
4	8	8	16	50.0%
Total	15	31	46	32.6%

Favorite Vegy	Obese		Total	% Obese
	Yes	No		
Broccoli	1	8	9	11.1%
Carrots	2	9	11	18.2%
Lettuce	8	7	15	53.3%
Tomato	4	7	11	36.4%
Total	15	31	46	32.6%



60
50
40
30
20
10
0
Percent Obese
1
2
3
4
Number of Sugary Food Groups Eaten Last Week

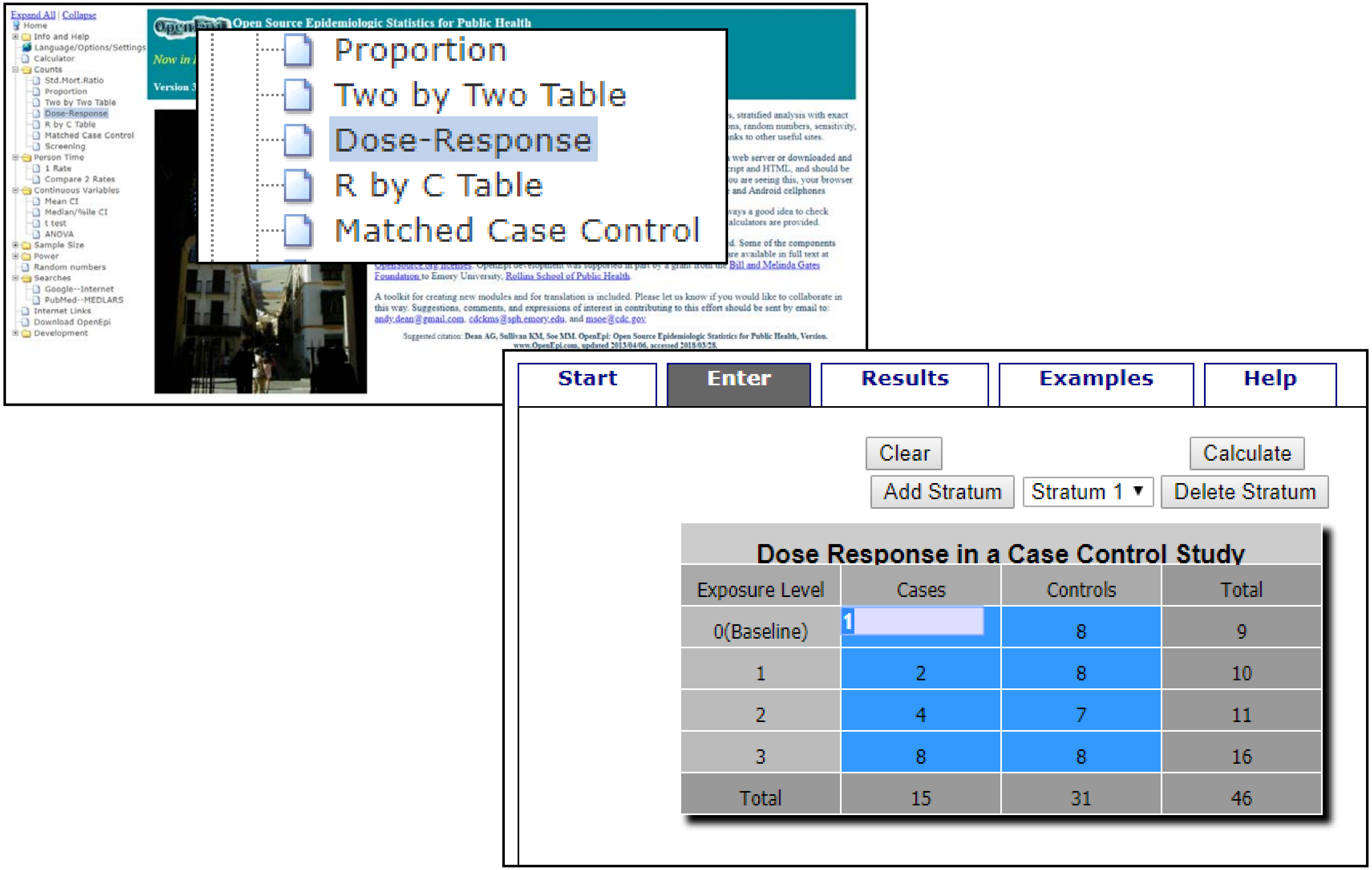

60%
50%
40%
30%
20%
10%
0%
Percent Obese
Broccoli
Carrots
Lettuce
Tomato
Favorite Vegetable

Chi-squared Trend Test

- Test of linear trend in series of proportions
  - or Relative Risks or Odds Ratios
- "Cochran–Armitage test for trend"
- "(Extended) Mantel-Haenszel chi square for linear trend"
- Formula is more complex than general chi-squared
  - http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/8-chi-squared-tests
  - https://en.wikipedia.org/wiki/Cochran%E2%80%93Armitage_test_for_trend

- Algebraically identical to simple test of slope

http://www.openepi.com



Dose Response Analysis

Stratum 1

Exposure Level	Cases	Controls	Total	Odds of Exp.	Odds Ratio
0	1	8	9	0.13	1
1	2	8	10	0.25	2
2	4	7	11	0.57	4.57
3	8	8	16	1	8
Total	15	31	46		

Mantel-Haenszel Summary Odds Ratios and Crude OR for Each Exposure Level

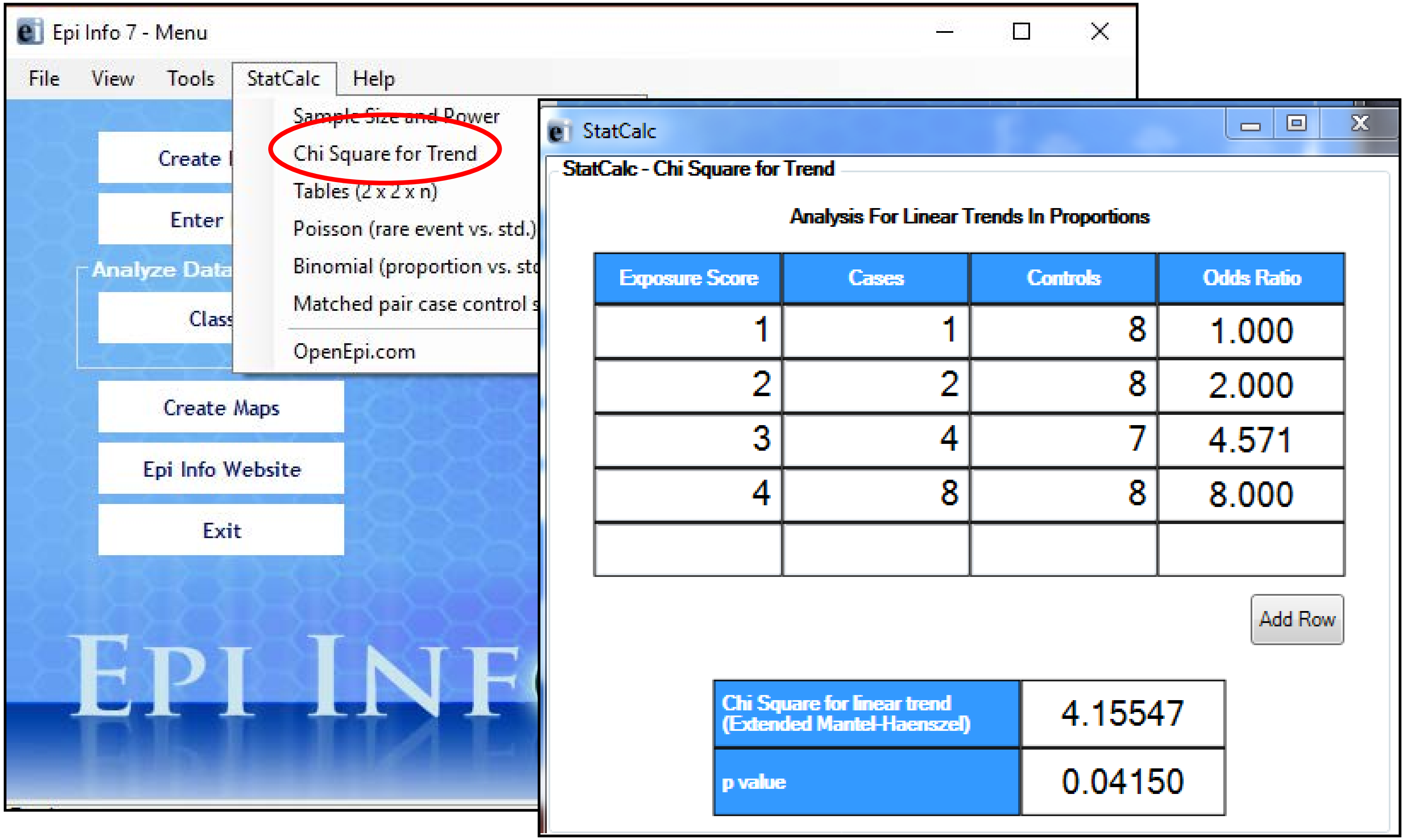
Exposure	**MH Summary OR**	**Crude OR**
Level 0 vs. Level 0:	1	1
Level 1 vs. Level 0:	2	2
Level 2 vs. Level 0:	4.571	4.571
Level 3 vs. Level 0:	8	8

If MH and crude ORs are equal, confounding by the stratifying variable was not present and stratification is unnecessary.

Extended Mantel-Haenszel chi square for linear trend= 4.16
p-value(1 degree of freedom)= 0.04150

Epi Info Stat Calc

https://www.cdc.gov/epiinfo/index.html



Trend Test - SAS

```
proc freq data=data.trendexample order=formatted;

tables grands*ident/ nopercent nocol chisq trend;

run;
```

Table of grands by ident			
grands(grands)	ident(ident)		
Frequency Row Pct	1 Yes	2 No	Total
1	1 11.11	8 88.89	9
2	2 20.00	8 80.00	10
3	4 36.36	7 63.64	11
4	8 50.00	8 50.00	16
Total	15	31	46

Statistics for Table of grands by ident

Statistic	DF	Value	Prob
Chi-Square	3	4.8889	0.1801
Likelihood Ratio Chi-Square	3	5.1980	0.1579
Mantel-Haenszel Chi-Square	1	4.7349	0.0296
Phi Coefficient		0.3260	
Contingency Coefficient		0.3100	
Cramer's V		0.3260	
WARNING: 38% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Sample Size = 46

Cochran-Armitage Trend Test	
Statistic (Z)	2.2000
One-sided Pr > Z	0.0139
Two-sided Pr > |Z|	0.0278

Sample Size = 46

$2.2^2 = 4.84$

Trend Test - SPSS

```
CROSSTABS
  /TABLES=grands BY ident
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

grands * ident Crosstabulation

Count

		ident		Total
		0	1	
grands	1	8	1	9
	2	8	2	10
	3	7	4	11
	4	8	8	16
Total		31	15	46

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.889[a]	3	.180
Likelihood Ratio	5.198	3	.158
Linear-by-Linear Association	4.735	1	.030
N of Valid Cases	46		

a. 3 cells (37.5%) have expected count less than 5. The minimum expected count is 2.93.

R: (at http://www.r-project.org/)

R Studio Interface: http://www.rstudio.com/

Console | Markers

E:/LECTURES/0TREND (lecture)/

```
>
>
>
> x <- c(1,2,4,8)
> N <- c(9,10,11,16)
> prop.trend.test(x,n)

        Chi-squared Test for Trend in Proportions

data:  x out of n ,
 using scores: 1 2 3 4
X-squared = 4.8402, df = 1, p-value = 0.0278

>
```

Sweets and Obesity Example - Differences in X^2 and p-value

Software	X^2 trend	P-value	reference
R	4.84	0.028	?
SAS ("chi")	4.73	0.0296	"Mantel-Haenszel"
SAS ("trend")	4.84 (2.20^2)	0.0278	"Cochran-Armitage"
SPSS	4.735	0.030	?
Open Epi	4.16	0.042	"Extended Mantel Haenszel Chi Square for linear trend" (Schlesselman, JJ. Case-Control Studies: Design, Conduct, Analysis. Oxford Univ. Press, NY, 1982; p.200-206)
Stat Calc	4.16	0.042	"Mantel extension of the Mantel-Haenszel summary odds ratio and chi square"

Reprinted from the JAMA® Journal of the American Medical Association
January 16, 1987 Volume 257

Original Contributions

Sexual Practices and Risk of Infection by the Human Immunodeficiency Virus

The San Francisco Men's Health Study

Warren Winkelstein, Jr, MD, MPH; David M. Lyman, MD, MPH; Nancy Padian, MS, MPH;
Robert Grant, MPH; Michael Samuel; James A. Wiley, PhD; Robert E. Anderson, MD;
William Lang, MI

The San Franc
demiology and
a cohort of 103
probability sa
seropositivity r
homosexual/bis
pants were HIV
male sexual pa
was 17.6%. For
Only receptive
infection. Douc
significantly to r

AMONG homo
large numbers
receptive anal
been the most
risk factors for
viruses associa
immunodeficien
However, all of
studies of risk
infection have b
clinical or volun

From the School of
and Lyman, Ms Padian
and the Survey Resea
of California, Berkeley
Francisco (Drs Anders
ettsial Disease Labor
Health Services, Berk
Research Institute, Ur
cisco (Dr Levy).
Reprint requests to
sity of California, Berk

Table 1.—Association of Number of Male Sexual Partners in Previous Two Years and Human Immunodeficiency Virus (HIV) Serologic Status*

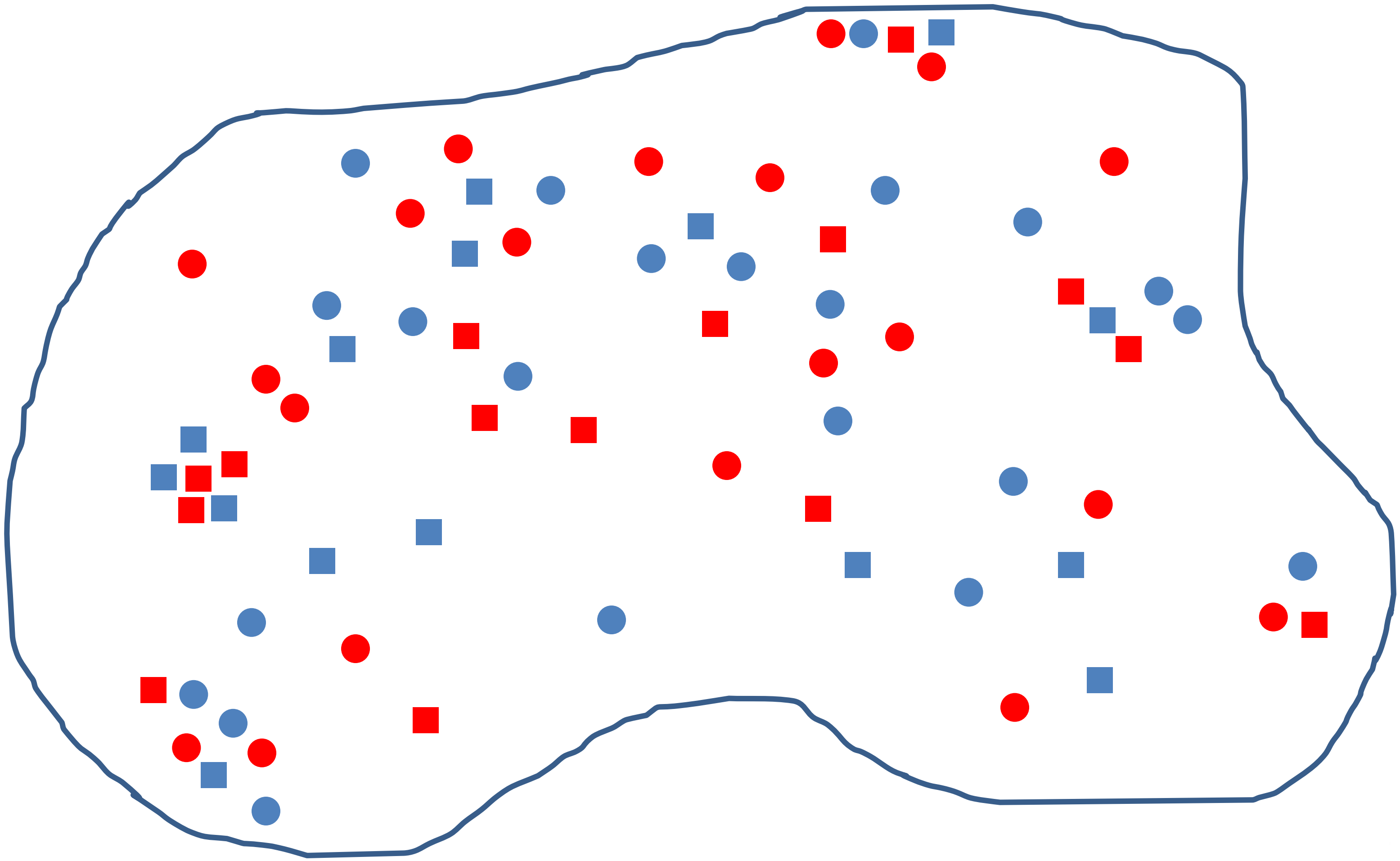
	Study Sample		Population	
No. of Male Partners	No. Examined	% HIV Antibody Positive	% HIV Antibody Positive†	95% Confidence Interval
None	17	17.6	19.2	5.2-41.5
1	66	18.2	17.9	9.5-29.0
2-9	206	31.6	31.9	25.2-39.0
10-49	312	53.8	53.7	47.4-59.6
≥50	195	70.8	70.5	62.7-76.8
Total	**796**	48.5	48.2	44.3-52.0

*Subjects with missing data (n = 13) were excluded. The χ^2 for trend of the association of number of partners and HIV antibody seropositivity in the sample is 86.7, $P < .0001$.

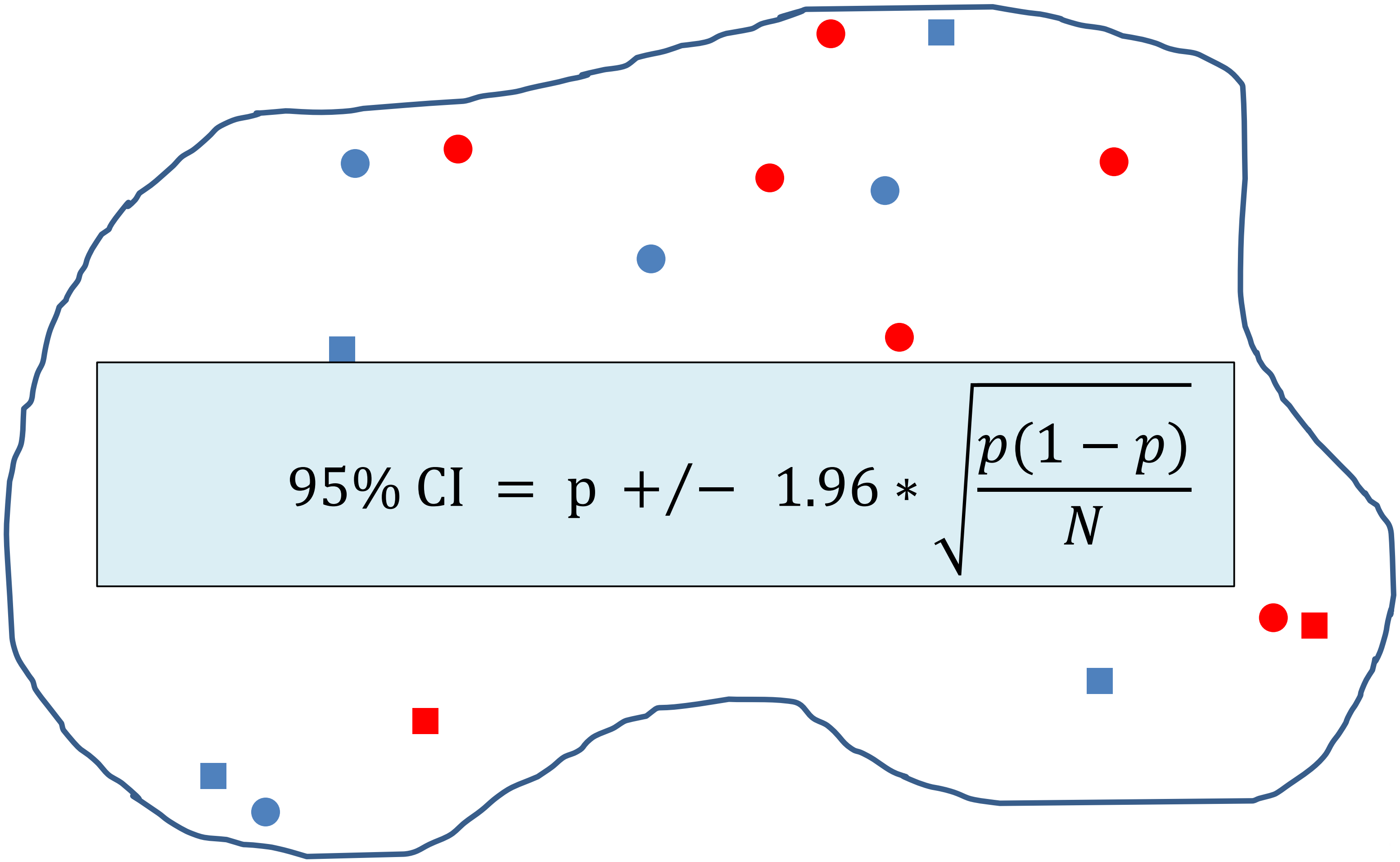
†Weighted for sampling fraction and difference in participation rates between census tracts.

A note on confidence intervals

Population

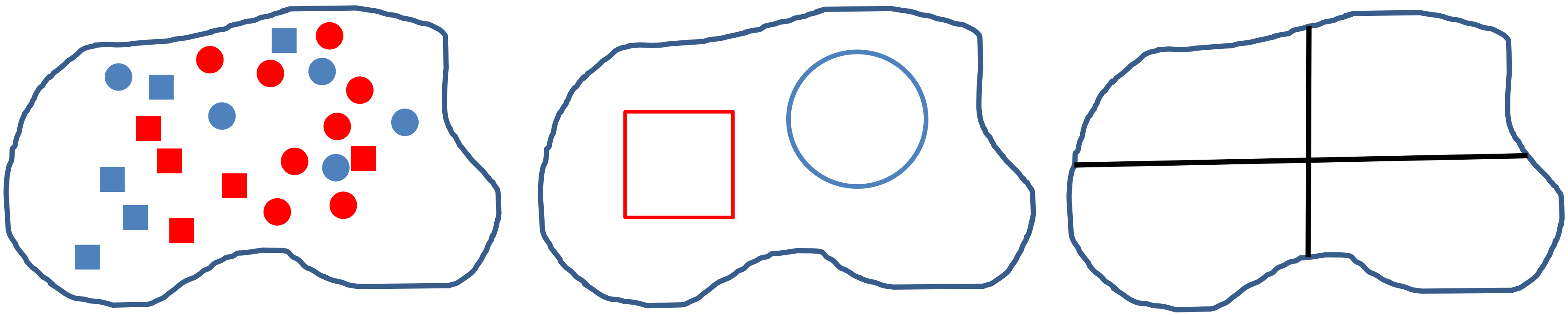


Simple Random Sample



$$95\%\ \mathrm{CI} = \mathrm{p} +/- \ 1.96 * \sqrt{\frac{p(1-p)}{N}}$$

Confidence Interval Methods for Samples



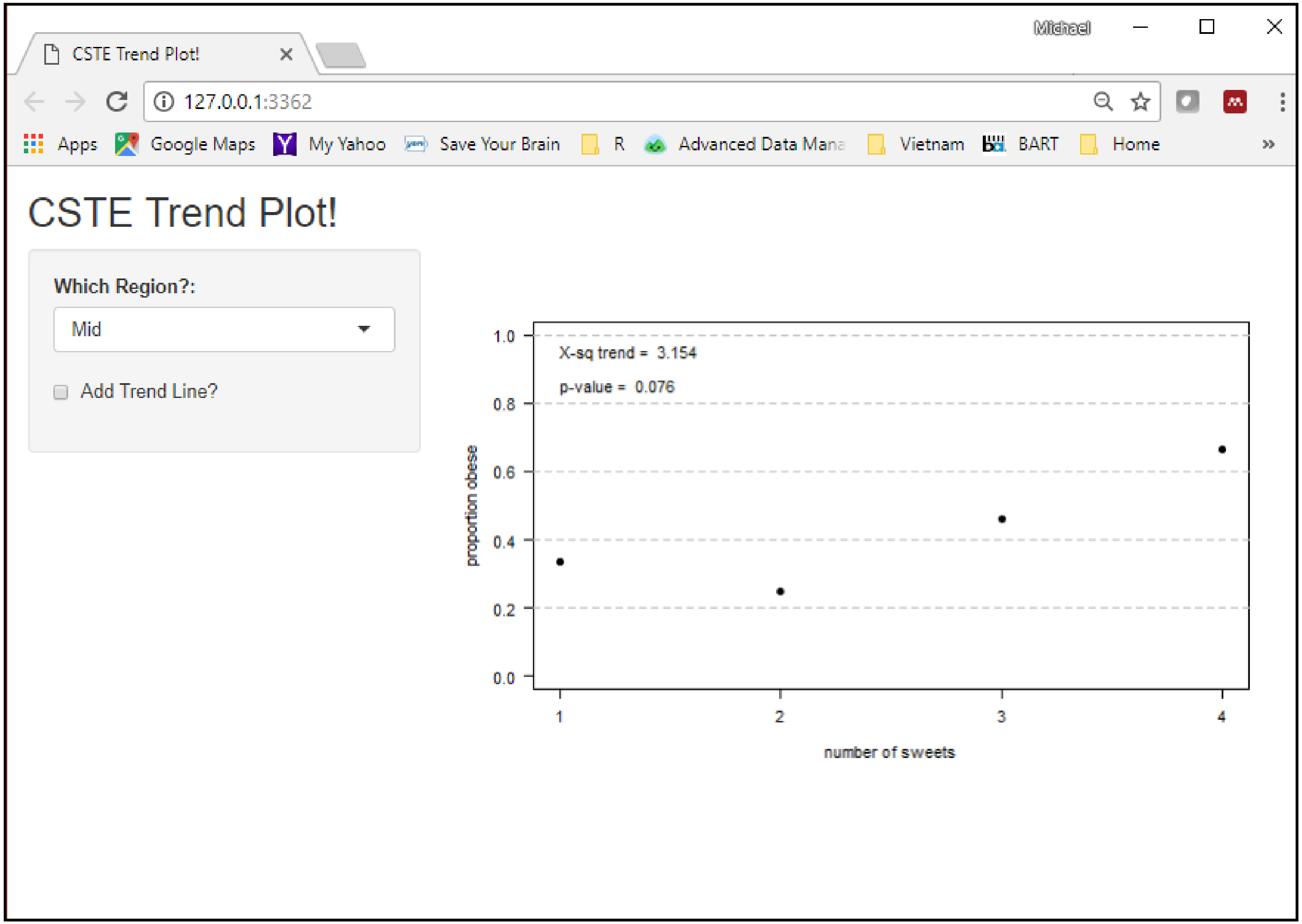
- Based on probability theory, the statistical issues and methods are well worked out, including for confidence intervals for simple random sampling, stratified sampling, cluster sampling and more
- Methods are well worked out for small numbers and other special situations

What theory/methods for population-level data, as we often have in Public Health surveillance?

- No clear theoretical basis upon which we calculate CIs for rates and proportions with population-level data like surveillance data
- Metaphysical argument is that we are observing one "sample" of the unobservable complete world behind us
- Other arguments suggest we know there is some "random" variation in most of things we observe, so therefore justified in using regular probability theory
- (Of note, we rarely take into account the variation in the population denominator estimates (e.g. from census estimates/projections) but these do indeed have documented variability)

What theory/methods for population-level data, as we often have in Public Health surveillance?

- In any case, some expression of the uncertainty and some statistical "tests" of differences in population-level rates and proportions **are essential** for communicating about these data and for making decisions
- Use of the "standard" tests appear to be the best alternative
- Might be reasonable to less precisely define confidence intervals in these situations (?)
  - **"Confidence intervals provide a guidepost for understanding if observed differences are (statistically) meaningful"**



https://phdataviz.shinyapps.io/cstedemo/

Resources

- Slides, spreadsheets, and R code for this presentation
  - www.goo.gl/k9YmXJ
- OpenEpi http://www.openepi.com
  - Web-based open source epidemiologic calculators and statistics for Public Health
- Epi Info Stat Calc https://www.cdc.gov/epiinfo/index.html
- R "home" page https://www.r-project.org/
- UNC FOCUS on Field Epidemiology -- Volume 3 - Data Analysis: Simple Statistical Tests
  - https://sph.unc.edu/nciph/focus/
- BMJ resource on statistical tests
  - http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/8-chi-squared-tests
- Good statistical information and tools
  - http://stattrek.com/
- Statistical Methods for Rates and Proportions, 2nd Edition. Joseph L. Fleiss. Wiley 1981.
  - The classic practical text on rates and proportion; newer 3rd Edition (with Bruce Levin, Myunghee Cho Paik) is likely excellent too

QUESTIONS?

Let's Talk:
Michael C. Samuel, Dr.P.H
Michael.Samuel@cdph.ca.gov
925-285-2926

CDPH California Department of PublicHealth

“Calculation and comparison of rates and proportions are basic skills for epidemiologists. We’ve all done so, at least in grad school, but may be rusty around the edges. This basic webinar will review the fundamental concept of a rate and a proportion; review their calculation, including with confidence intervals; and review the assessment of trends in rates and proportions. We will demonstrate the calculation of these measures using free web-based “calculators” and with a little bit of code using R”

- After the webinar, participants will be able to:
  - State what rates and proportions are and why they are important
  - Calculate rate-related measures using free web applications
  - Calculate rate-related measures using R, that can be expanded for other “real world” production projects