Biost 536 / Epi 536
Categorical Data Analysis in Epidemiology

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Reporting Associations
(Confidence Intervals)

October 21, 2014

1

Inference About Associations

2

Classical Hypothesis Tests

- Only stated in terms of null hypothesis

One-sided test of greater alternative:

$$H_0 : \theta \le \theta_0 \quad vs \quad H_1 : \theta > \theta_0$$

One-sided test of lesser alternative:

$$H_0 : \theta \ge \theta_0 \quad vs \quad H_1 : \theta < \theta_0$$

Two-sided test:

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta \ne \theta_0$$

3

Classical Conclusions

- Either
  - Reject null hypothesis
    - Because data is atypical of what would be expected when null is true
  - Do not reject null hypothesis
    - Because we cannot say that the data is atypical of what would be expected when null is true
      - Either null is true, or
      - Null is false and we "lack power"

4

My Objections to Classical Approach

- This focus on rejection or non-rejection of the null means that failure to reject the null hypothesis is "noninformative"
  - All we can say is "We don't know."
  - (That is what we said before we started the study)
- I prefer approaches that always allow an interpretation of study results
- I illustrate this with a hypothetical example

5

Reporting Associations

- Hypothetical study to detect an association between Event B and Exposure F
  - Unexposed: 0 of 5 have Event B
    - Estimated incidence rate: 0.000
    - 95% CI for incidence rate: 0.000 – 0.522
  - Exposed: 3 of 5 have Event B
    - Estimated incidence rate: 0.600
    - 95% CI for incidence rate: 0.147 – 0.947
  - Fisher's Exact two-sided P: 0.167
- How would you characterize the presence of an association between these two variables?

6

WRONG Criteria

- Incorrect criteria for stating the existence of a statistically significant association
  - "Because the confidence intervals overlap, there is no association."
  - (We need to use a P value. The use of confidence intervals in this manner is more complicated.)

7

Independent CI and Tests

- Rules for **independent** strata
- IF two independent 95% CI do not overlap
  - THEN we know a statistically significant difference exists (? P less than .006?)
- IF the 95% CI for one stratum contains the point estimate of the other stratum
  - THEN we know the difference is not statistically significant (? P greater than .16?)
- OTHERWISE all bets are off
  - Especially: we cannot reverse the above claims

8

WRONG

- An overstated, purely statistical report
  - "As the P value is greater than 0.05, we conclude that there is no association between exposure F and event B."
- (We should not conclude that there is no association, because we lacked precision to rule out differences that might be of interest.)

9

Scientifically USELESS

- A correctly stated, purely statistical report
  - "As the P value is greater than 0.05, we conclude that there is not sufficient evidence to rule out the possibility of no association between exposure F and event B."
- (Stated correctly, but gives no idea of whether we had ruled out differences that we cared about or we had merely done an abysmal study.)

10

CORRECT and USEFUL

- Scientific estimates and quantification of statistical evidence
  - "Incidence rates of 60% in the exposed (95% CI: 15% - 95%) and 0% in the unexposed (95% CI: 0% - 52%). Unfortunately, the precision was not adequate to demonstrate that such a large difference in incidence rates would be unlikely in the absence of a true association (P = 0.17)."
- (These data are not atypical of setting in which F= female and B= giving birth.)

11

Hypothetical Example

- Clinical trials of new treatments for high blood pressure
- Consider four possible scenarios
  - Measure of treatment effect is the difference in average SBP at the end of six months treatment
  - Scenarios differ in
    - Sample size
    - Variability of blood pressure
    - Treatment effect
  - (The scenarios are not replications of the same experiment or even the same scientific setting)

12

Reporting P values

Study	P value
A	0.1974
B	0.1974
C	0.0099
D	0.0099

13

Point Estimates

Study	SBP Diff	P value
A	27.16	0.1974
B	0.27	0.1974
C	27.16	0.0099
D	0.27	0.0099

14

Confidence Intervals

Study	SBP Diff	95% CI	P value
A	27.16	-14.14, 68.46	0.1974
B	0.27	-0.14, 0.68	0.1974
C	27.16	6.51, 47.81	0.0099
D	0.27	0.06, 0.47	0.0099

15

Interpreting Nonsignificance

- Studies A and B are both “nonsignificant”
- Only study B ruled out clinically important differences
- The results of study A might reasonably have been obtained if the treatment truly lowered SBP by as much as 68 mm Hg

16

Interpreting Significance

- Studies C and D are both statistically significant results
- Only study C demonstrated clinically important differences
- The results of study D are only frequently obtained if the treatment truly lowered SBP by 0.47 mm Hg or less

17

Bottom Line

- If ink is not in short supply, there is no reason not to give point estimates, CI, and P value
- If ink is in short supply, the confidence interval provides most information
  - (but sometimes a confidence interval cannot be easily obtained, because the sampling distribution is unknown under the null)

18

But: Impact of "Three (3.69?) over n"

- The sample size is also important
- The pure statistical fantasy
  - The P value and CI account for the sample size
- The scientific reality
  - We need to be able to judge what proportion of the population might have been missed in our sample
  - There might be "outliers" in the population
  - If they are not in our sample, we will not have correctly estimated the variability of our estimates
- The "Three over n" rule provides some guidance
  - 95% upper confidence bound when observe 0 events: 3 / n
  - 97.5% upper confidence bound when observe 0 events: 3.69 / n
    - This corresponds better to a two-sided 95% CI

19

Full Report of Analysis

Study	n	SBP Diff	95% CI	P value
A	20	27.16	-14.14, 68.46	0.1974
B	20	0.27	-0.14, 0.68	0.1974
C	80	27.16	6.51, 47.81	0.0099
D	80	0.27	0.06, 0.47	0.0099

20

Exceptions

- When reporting associations **or** evidence of effect modification, you should always strive to quantify the magnitude of the effect
  - Associations: a difference or a ratio
  - Effect modification: a difference of differences or a ratio of ratios

- Unfortunately, there are some times this is impossible
  - No parameterization of the alternative hypothesis on a scientifically relevant scale
    - E.g., Wilcoxon rank-sum test
  - Multiple parameters modeling the question of interest
    - E.g., Dummy variables, polynomials, or linear splines modeling some scientific factor
    - E.g., Testing for overall effect when modeling both main effect and interaction

21