

# Biost 517

## Applied Biostatistics I

.....

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

### Lecture 14: Two Sample Inference About Independent Proportions

November 20, 2006

1

© 2002, 2003, 2005 Scott S. Emerson, M.D., Ph.D.

## Lecture Outline

- .....
- Comparing Independent Proportions
    - Large Samples (Uncensored)
      - Chi Squared Test
    - Small Samples (Uncensored)
      - Fisher's Exact Test
      - Adjusted Chi Squared or Fisher's Exact Test

2

## Comparing Independent Proportions

.....

### Large Samples (Uncensored)

3

## Summary Measures

- .....
- Comparing distributions of binary variables across two groups
    - Difference of proportions (most common)
      - Most common
      - Inference based on difference of estimated proportions
    - Ratio of proportions
      - Of most relevance with low probabilities
      - Often actually use odds ratio

4

## Data: Contingency Tables

- The cross classified counts

		Response		Tot
		Yes	No	
Group	0	a	b	$n_0$
	1	c	d	$n_1$
Total		$m_0$	$m_1$	N

5

## Large Sample Distribution

- With totally independent data, we use the Central Limit Theorem
  - Proportions are means
    - Sample proportions are sample means
  - Standard error estimates for each group's estimated proportion based on the mean – variance relationship

6

## Asymptotic Sampling Distn

- Comparing two binomial proportions

Suppose independent  $X_i \sim B(1, p_0)$   $\sum_{i=1}^{n_0} X_i = X \sim B(n_0, p_0)$

and  $Y_i \sim B(1, p_1)$   $\sum_{i=1}^{n_1} Y_i = Y \sim B(n_1, p_1)$

We want to make inference about  $\Delta = p_1 - p_0$

$$\hat{p}_0 = \frac{X}{n_0} \sim N\left(p_0, \frac{p_0(1-p_0)}{n_0}\right) \quad \hat{p}_1 = \frac{Y}{n_1} \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$$

$$\hat{\Delta} = \hat{p}_1 - \hat{p}_0 \sim N\left(p_1 - p_0, \frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}\right)$$

7

## Asymptotic Confidence Intervals

- Confidence interval for difference between two binomial proportions

We want to make inference about  $\Delta = p_1 - p_0$

100(1- $\alpha$ )% confidence interval is

$$\left(\hat{\Delta} - z_{1-\alpha/2} \times se(\hat{\Delta}), \hat{\Delta} + z_{1-\alpha/2} \times se(\hat{\Delta})\right)$$

$$\hat{\Delta} = \hat{p}_1 - \hat{p}_0$$

$$\text{Estimate } se(\hat{\Delta}) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_0}}$$

8

## Asymptotic Hypothesis Tests

- Test statistic for difference between two binomial proportions

Suppose we want to test  $H_0 : \Delta = p_1 - p_0 = 0$

Under the null hypothesis,  $Z = \frac{\hat{\Delta}}{se(\hat{\Delta})} \sim N(0,1)$

Under  $H_0$  the entire distributions are equal so

$$\text{Estimate } se(\hat{\Delta}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_0}} \quad \text{with } \hat{p} = \frac{X+Y}{n_0+n_1}$$

9

## Goodness of Fit Test

- An alternative derivation of the asymptotic test of binomial proportions as a special case of the “goodness of fit test”
  - For contingency tables of arbitrary size
    - “R by C” table
    - E.g., tumor grade by bone scan score in PSA data set

10

## Goodness of Fit Test

- Given categorical random variables, we often have scientific questions that relate to the frequency distribution for the measurements
  - Examples:
    - Checking for Poisson, normal, etc.
    - Distribution of phenotypes to agree with genetic theory
    - Comparing distributions of categorical data across populations
    - Independence of random variables

11

```
. tabulate grade bss, row col cell
```

grade	bss			Total
	1	2	3	
1	1	3	6	10
	10.00	30.00	60.00	100.00
	20.00	27.27	25.00	25.00
	2.50	7.50	15.00	25.00
2	2	4	9	15
	13.33	26.67	60.00	100.00
	40.00	36.36	37.50	37.50
	5.00	10.00	22.50	37.50
3	2	4	9	15
	13.33	26.67	60.00	100.00
	40.00	36.36	37.50	37.50
	5.00	10.00	22.50	37.50
Total	5	11	24	40
	12.50	27.50	60.00	100.00
	100.00	100.00	100.00	100.00
	12.50	27.50	60.00	100.00

## Aside: Sampling Nomenclature

---

- We often characterize the sampling scheme according to the total counts that were fixed by design
  - Poisson sampling: none
  - Multinomial sampling: total counts
  - Binomial sampling: either row or column totals
  - Hypergeometric sampling: both row and column totals

13

## Basic Idea

---

- We compare the observed counts in each cell to the number we might have expected
  - “Observed – Expected”
  - Counts, NOT proportions

14

## Sampling Distribution

---

- The test to see whether the data fits the hypotheses (hence “goodness of fit”)
  - Observed counts in each cell are assumed to be Poisson random variables
    - Standard error is the square root of the mean
  - Test statistic based on sum of Z scores for each cell
    - Actually done as squared Z scores
    - Sum of squared normals has a chi squared distn

15

## Test Statistic

---

- Pearson’s chi-square goodness of fit statistic in the general case
  - Assuming only one “constraint” on the cells

Given  $K$  categories, and for the  $i$ th cell

$O_i$  the observed count and

$E_i$  the expected count (typically  $E_i = p_i N$ )

$$\chi^2 = \sum_{i=1}^K \frac{(O_i - E_i)^2}{E_i} \stackrel{H_0}{\sim} \chi_{K-1}^2$$

16

## Determining the “Expected”

- The scientific null hypothesis usually specifies the probability that a random observation would belong in each cell
  - Most often:
    - Testing independence of variables
      - Probabilities in each cell are expected to be the product of the marginal distributions
  - Other uses
    - Testing “goodness of fit” to distributions
    - Testing agreement with genetic hypotheses

17

## Test for Independence

- Chi-square test for independence (or association)

$R \times C$  table for two categorical variables, in the  $(r, c)$ th cell

$O_{rc}$  the observed count and

$E_{rc} = \hat{p}_r \hat{p}_c N$  the expected count

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \underset{H_0}{\sim} \chi_{(R-1) \times (C-1)}^2$$

where  $\hat{p}_r$  and  $\hat{p}_c$  are the estimated proportions are for the row and column margins, respectively

3

```
. tabulate grade bss, row col cell
```

grade	bss			Total
	1	2	3	
1	1	3	6	10
	10.00	30.00	60.00	100.00
	20.00	27.27	25.00	25.00
	2.50	7.50	15.00	25.00
2	2	4	9	15
	13.33	26.67	60.00	100.00
	40.00	36.36	37.50	37.50
	5.00	10.00	22.50	37.50
3	2	4	9	15
	13.33	26.67	60.00	100.00
	40.00	36.36	37.50	37.50
	5.00	10.00	22.50	37.50
Total	5	11	24	40
	12.50	27.50	60.00	100.00
	100.00	100.00	100.00	100.00
	12.50	27.50	60.00	100.00

```
. tabulate grade bss, row col exp
```

grade	bss			Total
	1	2	3	
1	1	3	6	10
	1.3	2.8	6.0	10.0
	10.00	30.00	60.00	100.00
	20.00	27.27	25.00	25.00
2	2	4	9	15
	1.9	4.1	9.0	15.0
	13.33	26.67	60.00	100.00
	40.00	36.36	37.50	37.50
3	2	4	9	15
	1.9	4.1	9.0	15.0
	13.33	26.67	60.00	100.00
	40.00	36.36	37.50	37.50
Total	5	11	24	40
	5.0	11.0	24.0	40.0
	12.50	27.50	60.00	100.00
	100.00	100.00	100.00	100.00
	12.50	27.50	60.00	100.00

## In 2 x 2 Contingency Tables

.....

- The chi squared test and the Z test comparing binomial proportions are exactly the same test
  - The chi squared statistic is just the square of the Z statistic
  - The chi squared test P value will be the same as the two-sided P value for the Z test
  - Most software packages have a tendency to tell you the value of the chi squared statistic

21

## Elevator Statistics

.....

- Well, nearly elevator statistics

	+	-		
0	a	b		$n_0$
1	c	d		$n_1$
	$m_0$		$m_1$	$N$

$$\chi^2 = \frac{(ad - bc)^2 N}{n_0 n_1 m_0 m_1} \quad Z = (ad - bc) \sqrt{\frac{N}{n_0 n_1 m_0 m_1}}$$

22

## Need for Large Samples

.....

- Note that because the goodness of fit test is relying on asymptotic properties, it is only valid in “large” samples
  - A commonly used rule of thumb is that the expected counts be greater than 5 in the vast majority of the cells

23

## Other Large Sample Tests

.....

- “Likelihood Ratio Test”
  - Also has chi squared distribution in large samples
    - But not the same statistic as “chi squared statistic”
  - Good large sample properties
    - Often most powerful
  - Less commonly used in 2 x 2 tables
    - We will use it more often in logistic regression

24

## Stata Commands: CI

- `"cs respvar groupvar, level(#)"`
  - Both variables must be coded as 0 and 1
  - Response will be called "cases" and "noncases"
    - CI can be found under "Risk difference"
    - Chi squared statistic and two-sided P value
- `"tabulate respvar groupvar, row col chi2 lr"`
  - Row and column percentages
  - Chi squared and likelihood ratio P values

25

## Ex: Stata Commands

- Example: Hepatomegaly by treatment group in PBC data set

```
. g tx= 2 - treatmnt
. cs hepmeq tx
```

	tx		
	Exposed	Unexposed	Total
Cases	73	87	160
Noncases	84	66	150
Total	157	153	310
Risk	.4649682	.5686275	.516129

26

## Ex: Stata Commands (cont.)

- Example: Hepatomegaly by treatment group in PBC data set (cont.)

	Point estimate	[95% Conf. Interval]	
Risk diff	-.1037	-.2143	.0070
Risk ratio	.8177	.6580	1.0161
Prev frac ex	.1823	-.0161	.3420
Prev frac pop	.0923		

chi2(1) = 3.33 Pr>chi2 = 0.0679

27

## Interpretation

- With 95% confidence, the true difference in the prevalence of hepatomegaly at baseline is between .007 higher in treatment group and .214 lower in treatment group.
  - (What are these populations? At baseline, we only have samples that are treated and control. Both groups were drawn from the same population.)
- Based on two sided P = .0679, we cannot reject the null hypothesis of equality

28

## Caveat: Sad Fact of Life

- Different variance estimates are typically used for CI and hypothesis tests
  - We can see disagreement between the conclusion reached by CI and P value
    - The P value might be less than .05, but the CI contain 0
    - The P value might be greater than .05, but the CI exclude 0

29

## Comparison to t Test

- Hepatomegaly by treatment group using two sample t test with unequal variances
  - Z test uses the standard normal distribution and does not use the sample variance

```
.ttest hepmeq, by(tx) unequal
Two-sample t test; unequal variances 0: N obs= 157
                                      1: N obs= 153
```

Variable	Mean	St Err	t	P> t	[95% CI]	
0	.465	.0399	11.6	0.0000	.386	.544
1	.569	.0402	14.2	0.0000	.489	.648
diff	- .104	.0566	-1.83	0.0682	-.215	.008

30

## Comparison to t Test (cont)

```
Satterthwaite's degrees of freedom: 307.8897
Ho: mean(0) - mean(1) = diff = 0
Ha: diff < 0      Ha: diff ~= 0      Ha: diff > 0
t = -1.8300      t = -1.8300      t = -1.8300
P < t= 0.0341    P > |t|= 0.0682    P > t= 0.9659
```

- From “standard” analysis
  - CI: -0.2143, 0.0070; P= 0.0679
- From nonstandard t test based analysis
  - CI: -0.215, 0.008; P= 0.0682

31

## Yates Correction

- Historically, a “continuity correction” to the chi squared test to try to avoid its anti-conservatism in small samples
  - All that was achieved was getting a test that behaves as poorly as the Fisher’s exact test
  - I heartily disrecommend use of the continuity correction when comparing two samples
    - (There is a continuity correction used in one sample Z tests that is useful, but exact distributions are even better)

32



## Comparing Independent Proportions

.....  
Small Samples (Uncensored)

33

## Small Sample Distribution

- .....
- The exact distribution for the difference in two proportions can not be determined in general, because of the mean – variance relationship
    - We need to know the value of the two proportions being compared in order to find the exact distribution of the difference

34

## Small Sample CI

- .....
- We have no way of obtaining exact CI for the difference in proportions
    - We could consider all possible values of the two proportions, and see whether a test would reject each combination
      - But the resulting joint confidence interval would not always give the same decision for equal differences
        - E.g, it might reject .10 and .20, but not .40 and .50

35

## Small Sample Tests

- .....
- We can, however, describe the exact distribution of the data under the null hypothesis conditional on all the “margins” of a contingency table
    - A “permutation” distribution
      - We imagine randomly assigning observations between the groups

36

## Permutation Idea

- Randomly permute  $m_0$  positives and  $m_1$  negatives
- Call the first  $n_0$  “group 0” and the last  $n_1$  “group 1”
- Repeat many times and see how often “group 0” has  $a$  or more positives

		Response		
		+	-	
Group	0	a	b	n <sub>0</sub>
	1	c	d	n <sub>1</sub>
		m <sub>0</sub> m <sub>1</sub>		N

Condition on values of  $m_0, m_1, n_0, n_1$

37

## Permutation Tests

- I usually object to permutation distributions except as a last resort
  - They test equality of distributions, not just equality of the population parameter
    - Usually they are not, however, guaranteed to detect arbitrary differences between distributions even in infinite samples

38

## Permutation with Binary Data

- However, with binary data, distributions are different if and only if the proportions are different
  - Hence permutation tests are okay for testing
  - But still, we have no confidence intervals because we have not quantified alternatives

39

## Small Sample Tests

- Conditioning on the margins
  - Often one margin is fixed by design
    - Cohort studies sample by exposure
    - Case-control studies sample by disease
  - In any case, it can be shown that none of the margin totals contribute information about the difference in proportions

40

## Fisher's Exact Test

- Probability of more extreme contingency tables with the same marginal totals
  - Probabilities by hypergeometric distribution
    - (Use a comp

		Response		
		+	-	
Group	0	$a-k$	$b+k$	$n_0$
	1	$c+k$	$d-k$	$n_1$
		$m_0$	$m_1$	$N$

Consider all possible values of  $k$

41

## Stata Commands

- The Fisher's exact test P values are given by several commands
  - “cs respvar groupvar, exact”
  - “tabulate respvar groupvar, exact”
    - One-sided and two-sided P values are provided

42

## Stata Example

- Example: Hepatomegaly by treatment group (cont.)

```
. cs hepmeq tx, exact
```

	Pt. Est.	[95% CI]	
Risk diff	.1036593	-.007	.214
Risk ratio	1.222938	.984	1.520
Attr fr ex	.1822974	-.016	.342
Attr fr po	.0991242		

1-sided Fisher's exact P = 0.0433

2-sided Fisher's exact P = 0.0704

43

## Comments

- Fisher's exact test does not turn out to be an exact test in practice
  - A problem is posed by the discrete nature of the data
    - To achieve the desired level .05 two-sided test, we would sometimes have to reject the null when both groups had 0 successes
      - With some results flip a biased coin to decide whether significant
      - Few people are willing to do this

44

## Problem

- We then face a dilemma
  - The chi squared test (Z test for proportions) may be anti-conservative in small samples
    - Generally so long as all cell counts in the contingency table are expected to be greater than 5 under the null hypothesis, we are OK
  - The Fisher's exact test is too conservative

45

## Alternatives

- Great improvements in statistical power obtained by modifying either of those tests to achieve as close to the nominal type I error without exceeding
  - Several statistical packages provide such modified tests (e.g., StatExact)
  - Stata does not

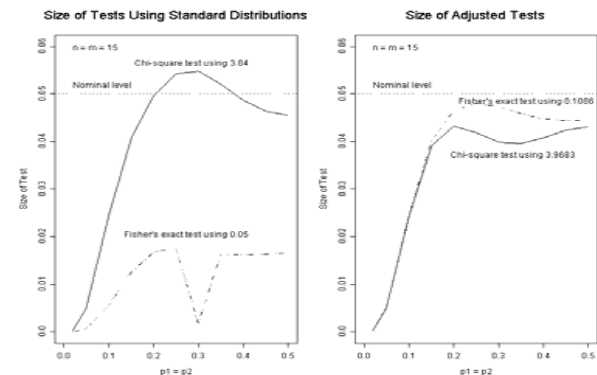
46

## Modifications

- Basic idea
  - Use the statistic
  - Don't presume the classical distribution
    - Don't assume chi squared statistic has chi square distribution
    - Don't assume Fisher's Exact P value has uniform distribution
  - Consider all possible values of  $p$  common to both groups, and use exact distribution
    - Then take worst case

47

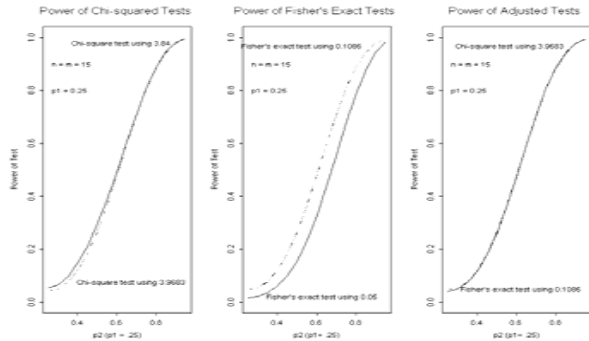
## True Type I Error by Common $p$



48

## Gains in Power

- Power of unadjusted, adjusted level .05



## General Comments

- It is generally immaterial whether the Fisher's exact test P value or the chi square statistic or likelihood ratio statistic is used as the basis for the exact test
- In any case, the critical value is dependent upon the sample sizes
- Using this approach, substantial improvement in power is obtained in low sample sizes
- I strongly recommend its use when confronted with small samples in real life

50

## Inference About Odds Ratios

## Odds of Exceeding a Threshold

- Previously: inference based on the probability of exceeding a threshold
  - Sometimes it is more convenient to discuss the odds of exceeding a threshold
    - $\text{odds} = \text{prob} / (1 - \text{prob})$
  - In one and two sample problems, inference about the odds is easily obtained from inference about the probability (proportion)
    - And the proportion is more easily understood

51

52

## Advantage of Using Odds

.....

- Avoiding effect modification
  - When adjusting for confounders or precision variables, it is intuitively unlikely that differences in proportions will be the same across all subgroups
    - Proportions must be between 0 and 1
    - Odds can be between 0 and infinity
      - (log odds can be between negative infinity and infinity)
- Case – control studies

53

## Case – Control Studies

.....

- When outcome event is rare
  - Case – control sampling is efficient
  - Odds is very close to the probability (1 – prob is approximately equal to prob)
  - Hence, the odds ratio is approximately the risk ratio

54

## Mathematics Based Logic

.....

- The odds ratio is independent of the conditional probability being estimated
  - Cohort studies:  $\Pr(\text{Disease} | \text{Exposed})$
  - Case-control studies:  $\Pr(\text{Exposed} | \text{Disease})$
- Can consider Odds Ratio for exposure based on disease from case-control study
  - Equal to Odds Ratio for disease by exposure
  - For rare disease, this is approximately ratio of disease probability

55

## Odds of Exceeding a Threshold

.....

- Inference about the odds is usually made in the context of the chi squared test
  - In Stata, we can obtain estimates of the odds ratio using
    - `cc casevar expvar`
      - `cc` = case – control
      - Provides odds ratio and chi squared statistic

56

## Looking to the Future

---

- In two sample tests, I think using difference in proportions is best
- When multiple samples or adjusting for covariates we tend to use logistic regression
  - Summary measures based on odds ratio