

Binary, ch-square, associations

Osman Abdullahi

Analysis of binary data

Objectives

- Present binary data
- Calculate proportions & standard error of the proportion from sample data.
- Use standard error to calculate 95% CI and to test hypothesis on proportions.
- Use Chi-squared test.
- Teach some theory, let you explore the concepts using R.

Binomial distribution

Binary data = Yes/No or 0/1 or Pos/Neg Calculate proportion as number positive/Total in sample

Population proportion is P Sample proportion is p

Assumptions: - Our sample accurately reflects the population from which it is drawn - Our data is drawn from a binomial distribution.

If the distribution of the data is binomial, then we estimate the proportion, p .

Proportion $p = \frac{\text{Number positive}}{\text{Total number in sample}}$

The standard error of the proportion (large sample, normal approx).

Standard error of $p = SE(p) = \sqrt{p(1 - p)/n}$

Summary of SE

- The population proportion is unknown, and fixed. The standard error does not refer to the population proportion p .
- Standard errors are calculated for estimated proportions (p) to show the uncertainty of the estimate.
- The larger the sample size, the smaller the standard error of the estimated proportion.
- Standard errors are used in 2 ways;
 - To calculate 95% confidence limits around our estimate
 - To test hypothesis about our estimate.

95% Confidence Interval for a Proportion–

From our sample we estimated: the proportion positive $p = \frac{(pos)}{(totalinsample)}$

And the standard error of $p = SE(p) = \frac{(p)(1-p)}{n}$ Using the normal approximation we can obtain 95% CI of our estimate :

$$p \pm 1.96(SE)$$

95% CI

The meaning of the 95% CI is we are 95% sure the true proportion P lies is covered by this interval.

95% CI of the sample proportio

The 95% CI of the sample proportion will contain the (unknown) population proportion for 95% of possible samples taken from the population.

Larger sample size gives smaller 95%CI

95% CI for proportion

95% CI for proportion

In R, the command *ci* with option binomial

```
> ci(birthweight2$lbw2==1)
  events total probability          se exact.lower95ci exact.upper95ci
    80   641    0.124805 0.01305387      0.1002059      0.1529111
```

95% CI for the estimate of the proportion within groups

```
> require(data.table)
> data <- data.table(birthweight2)
> data[, data.frame(ci(lbw2==1)), by=sex]
      sex events total probability          se exact.lower95ci exact.upper95ci
1: Female    45   315    0.1428571 0.01971616      0.10614286      0.1864571
2:  Male    35   326    0.1073620 0.01714566      0.07592638      0.1461518
```

Significance testing (1)

Significance testing (1)

Assuming the normal approximation.

Accept or reject the null hypothesis, depending on the value of T.

Test H_0 : Proportion of normal birth weight babies is 90%

Use the R function **prop.test(n,N,p0)**

```
>prop.test(sum(birthweight2$lbw2==0),length(birthweight2$lbw2),p=0.9, correct=F)
```

```
1-sample proportions test without continuity correction
```

```
data: sum(birthweight2$lbw2 == 0) out of length(birthweight2$lbw2), null probability 0.9
```

```
X-squared = 4.3822, df = 1, p-value = 0.03632
```

```
alternative hypothesis: true p is not equal to 0.9
```

```
95 percent confidence interval:
```

```
0.8473534 0.8985664
```

```
sample estimates:
```

```
p
```

```
0.875195
```

Significance testing (2)

```
birthweight2 <- read.csv("birthweight2.csv")
birthweight2$lbw2 <- as.numeric(birthweight2$lbw)
binom.test(sum(birthweight2$lbw2==0),length(birthweight2$lbw2))

##
## Exact binomial test
##
## data:  sum(birthweight2$lbw2 == 0) and length(birthweight2$lbw2)
## number of successes = 0, number of trials = 641, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.0000000000 0.005738355
## sample estimates:
## probability of success
##                                0
```

Summary: Basic tools for the analysis of binary data:

Descriptive: Bar charts, and tabulation of the data

Analytic : Creating 95% CI and hypothesis testing. 1. Assuming approximation, use `prop.test()` 2. Exact methods based on binomial distribution. Use `ci()` and `binom.test()`

Practical 5. Analysing Low birth weight

- Use birthweight2
- Check the variables, and explore the data.
- Look at the variable lbw, it is coded 0=LBW, 1=Normal
- Generate a new variable showing 1=LBW and 0=Normal
- Get the proportion of low birth weight babies and 95% CI.
- Get the proportion of lbw babies (and 95% CI) by sex.
- Test the hypothesis that $p=0.90$ (90% normal BW)
- Test this hypothesis for male babies and female babies separately

Comparing proportions

Objectives:

To estimate differences in proportions, and get 95% CI for the difference.
To test the hypothesis that the proportions are different, there are several ways to do this: - Using a normal approximation (Z-test) - Using chi-squared test (session 3) - Using exact methods (session 3)
Show how to do this in R, with useful options to explore binary data.

Difference in proportions

Difference between two proportions is: $p_1 - p_2$

Standard error of $(p_1 - p_2)$ for the 95% CI

\$\$

$$SE = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

\$\$

Then calculate the 95% CI using the standard method:

$$(p_1 - p_2) \pm 1.96 * SE(p_1 - p_2)$$

Hypothesis testing

Null hypothesis H_0 : Both proportions the same = **overall p**

Calculate overall proportion

$$p = \frac{(r1) + (r2)}{(n1 + n2)}$$

The common proportion will always be between the two proportions

Standard error of \bar{p} to test the null hypothesis.

$$SE(\bar{p}) = \sqrt{\bar{p} \cdot (1 - \bar{p}) \cdot \left(\frac{1}{n1} + \frac{1}{n2}\right)} = \text{thepooledSE}$$

Relationship between significance and 95% CI

95% CI includes zero — H_0 not rejected at 5% level

95% CI does not include zero — H_0 rejected at 5% level

Null hypothesis: $H_0: p_1 = p_2$ or $H_0 : p_1 - p_2 = 0$

The calculation of the standard error of the difference in proportions for the hypothesis test IS DIFFERENT FROM the calculation of the standard error of the difference ($p_1 - p_2$) for the 95% CI.

This is because the hypothesis test assumes there is no difference (the NULL hypothesis), whereas the 95% CI assumes there is a difference (and we want to quantify the uncertainty around the difference).

Session 3: Chi- squared test

see

Comparing proportions - chi-squared test

Comparing two (or more) proportions

– the Chi-squared test uses Expected numbers.

Chi-squared test is valid for any contingency table

Assumptions: sufficient numbers in each cell of the table

- 1 State the null hypothesis: No association between the two variables.
- 2 Calculate the expected numbers for each cell.
- 3 Calculate the Chi-squared statistic from the Observed and Expected numbers
- 4 Test against the chi-squared distribution.
- 5 Obtain the p-value for the data, under H_0

Chi-squared test – the calculations

$$\text{Expected number in each cell} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

Equivalent to the same percentage in each group. Chi-squared statistic:

$$\sum (\text{observed} - \text{expected})^2 / \text{expected}, \chi^2 = \sum (O - E)^2 / E$$

Note the calculation is done for each cell, and then summed up over all cells.

```
mytable <- table(birthweight2$sex,birthweight2$lbw2)
mytable
```

```
##
##           1    2
##   Female 270  45
##   Male   291  35
```

```
summary(mytable)
```

```
## Number of cases in table: 641
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1.8479, df = 1, p-value = 0.174
```

```
chisq.test(birthweight2$sex,birthweight2$lbw2,correct = FALSE)
```

```
##
```

```
## Pearson's Chi-squared test
```

```
##
```

```
## data: birthweight2$sex and birthweight2$lbw2
```

```
## X-squared = 1.8479, df = 1, p-value = 0.174
```

Contingency tables – the exact test

If Chi-squared test not valid then get R to test the null hypothesis H_0 using the Fishers exact test.

```
fisher.test(birthweight2$sex,birthweight2$lbw2)
```

```
##  
## Fisher's Exact Test for Count Data  
##  
## data: birthweight2$sex and birthweight2$lbw2  
## p-value = 0.1895  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.436256 1.187129  
## sample estimates:  
## odds ratio  
## 0.7220222
```

Chi-squared test - for larger tables

Chi-squared test can be used for larger tables, with more categories (eg. agegroups). Same assumptions about expected number

- Tabulate outcome by explanatory factor
- Calculate expected numbers for each cell
- Calculate the test statistic:

$$\chi^2 = \sum (O - E)^2 / E$$

- Calculate the degrees of freedom (d.f.) **d.f. = (number rows - 1) * (number of cols - 1)**
- Test against the Chi squared distribution, and get the p-value under the null hypothesis

$$(H_0)$$

Chi-squared test - for larger tables

Larger tables using R

```
mytable3 <- table(birthweight2$ethnic,birthweight2$lbw2)
mytable3
```

```
##
##      1    2
##  1 230  30
##  2   71  10
##  3  134  25
##  4  126  15
```

```
summary(mytable3)
```

```
## Number of cases in table: 641
## Number of factors: 2
```

Larger tables -many levels of an exposure

For an ordered categorical exposure variable, it is possible to analyse for a trend across exposure levels. Two methods of doing this:

- Chi-squared test for trend.
- Test for trend in odds across the levels.

```
t<- table(birthweight2$lbw2 ,birthweight2$gestwks)
t
```

```
##
##      25  26  28  29  30  31  32  33  34  35  36  37  38  39
##  1    0   0   0   0   0   0   0   5   8  11  30  87 167
##  2    1   1   3   1   3   5   7   7   9   6  11  14   3
```

```
x<-t[2,] # number of low birth weights
n<-apply(t,2,FUN=sum) # total number of births in each gestation
prop.trend.test(x,n) # Trend test; Significant
```

Summary of the comparison of proportions

Using the normal approximation (use Z-test): - $SE(diff)$ for calculating the 95% CI - $SE(p)$ to test H_0 Using Chi-squared to test the null hypothesis. Needs sufficient numbers for each cell (`chisq.test()` , `summary(table())`) If not then use exact methods to test difference – Fishers exact test (`fisher.test()`)

Practical 6. Analysing Low birth weight

- Use `birthweight2`, with outcome low birth weight (`lbw`)
- Ensure you have the variable that shows 1= LBW, 0=Normal
- Tabulate and test if `lbw` differs by sex of baby. What is the difference in proportion `lbw` between the sexes.
- Tabulate the low birth weight by hypertension status of mothers (variable is called `ht`)
- Look at the association between `lbw` and hypertension (`ht`), using the chi-squared test
- Compare the proportion with low birth weight by the ethnic groups. What problem do you see?

Measures of association

Measures of association

Objectives:

- 1 To define risk ratios, odds ratios and other measures of association
- 2 How to get standard errors for risk ratios and odds ratios, and to use these to obtain 95% CI for these measures.
- 3 How to obtain these measures in R
- 4 When the different measures are used.

Measures of association- Prevalence ratio—

$$Prevalence(risk) = \frac{Numberpositive}{Totalnumber}$$

$$Prevalenceratio(riskratio) = \frac{Prevalenceinexposedgroup}{Prevalenceinunexposedgroup}$$

What is the standard error of Risk ratio (RR) ?

Risk ratio (RR)

Risk ratio (RR)

$$RR = (a/(a+c))/(b/(b+d))$$

But what is the standard error (SE)?

The SE is best estimated on the log scale.

	Exposed	Unexposed	Total
Disease	a	b	(a+b)
No disease	c	d	(c+d)
Total	(a+c)	(b+d)	N

It can also be shown that the SE(logRR) can be written as

$$SE \text{ for the } \log(RR) = \sqrt{\{1/a - 1/(a+b) + 1/c - 1/(c+d)\}}$$

Measures of association – odds ratio

Measures of association – odds ratio

Odds = number positive / number negative.

An even more useful measure than risk ratio (RR) is the odds ratio (OR) of infection.

Odds ratio (OR) = $\frac{\text{Odds in exposed group}}{\text{Odds in unexposed group}}$

Odds ratio = $(a/c) / (b/d)$

OR = $(a * d) / (b * c)$

What is the SE
of this measure?

	Exposed	Unexposed	Total
Disease	a	b	(a+b)
No disease	c	d	(c+d)
Total	(a+c)	(b+d)	N

Odds ratio (OR)

Odds ratio (OR)

$$\text{Odds ratio (OR)} = \frac{\text{Odds in exposed group}}{\text{Odds in unexposed group}}$$

$$OR = (a * d) / (b * c)$$

Again the SE is best estimated on the log scale.

It is simpler and easier to use than RR

	Exposed	Unexposed	Total
Disease	a	b	(a+b)
No disease	c	d	(c+d)
Total	(a+c)	(b+d)	N

$$SE \text{ for the } \log(OR) = \sqrt{\{1/a + 1/b + 1/c + 1/d\}}$$

$$95\% \text{ CI} = OR/EF \text{ to } OR \times EF$$

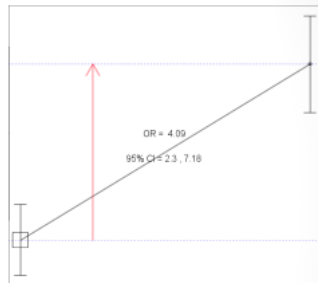
$$\text{Where EF} = \text{Error factor} = \exp(1.96 \times \log(SE))$$

Odds ratio & risk ratios in R

Odds ratio & risk ratios in R

```
> cc(birthweight2$lbw2 , birthweight2$ht2)
      birthweight2$ht2
birthweight2$lbw2  0    1 Total
0                499  62   561
1                 53  27    80
Total            552  89   641
```

```
OR = 4.1
Exact 95% CI = 2.3, 7.18
Chi-squared = 30.17, 1 d.f., P value = 0
Fisher's exact test (2-sided) P value = 0
```



Odds ratio & risk ratios in R

Odds ratio & risk ratios in R

```
> cs(birthweight2$lbw2 , birthweight2$ht2)
```

Outcome	Exposure		
	Non-exposed	Exposed	Total
Negative	499	62	561
Positive	53	27	80
Total	552	89	641

	<u>Rne</u>	<u>Re</u>	<u>Rt</u>
Risk	0.1	0.3	0.12

	Estimate	Lower95ci	Upper95ci
Risk difference (attributable risk)	0.21	0.12	0.27
Risk ratio	3.16	2.07	4.83
<u>Attr. frac. exp. -- (Re-Rne)/Re</u>	0.68		
<u>Attr. frac. pop. -- (Rt-Rne)/Rt*100 %</u>	23.07		
Number needed to harm (NNH) or 1/(risk difference)	4.82	3.74	8.32

Odds ratios and Risk ratios

Standard errors can be obtained on the log scale, and used to obtain 95% CI and to test hypothesis

Several commands in R to obtain odds ratios, and risk ratios.

For `cc`, and `cs` functions, make sure you have the coding right.

Practical 7

- Use the same dataset birthweight2.dta
- Check the Odds ratio for the association between LBW and hypertension
- Look at the association between LBW and gestational age. Divide gestwks into quartiles and analyse as groups, check for trend
- Look at birth weight and maternal age (in groups).
- Finally look at a different outcome, hypertension and age.

Summary

Proportions

Proportions

- Categorical data are presented as proportions or percentages

- SE(p) is =

$$SE(p) = \sqrt{p(1-p)/n}$$

- 95% CI for the proportion is $= \text{prop} \pm 1.96 \times SE(\text{prop})$
=

$$p \pm 1.96 \sqrt{p(1-p)/n}$$

- Significance test for a proportion

$$Z = (p - \pi_0) / SE(\pi)$$

comparing two proportions (1)

Comparing two proportions (1)

- Assume normal approximation to binomial distribution if samples are large
- Difference in two proportions
- 95% CI in difference in proportions
 - *diff in prop $\pm 1.96 \times SE$ (diff in proportions)*
 - $(p_1 - p_2) \pm 1.96 \times SE (p_1 - p_2)$
- Where

$$SE (p_1 - p_2) = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

comparing two proportions (2)

Comparing two proportions (2)

- Null hypothesis is $p_1 = p_2$
- Use a common proportion to calculate pooled SE

- Pooled SE =

$$SE(p_1 - p_2) = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$p = \frac{r_1 + r_2}{n_1 + n_2}$$

- Hypothesis test for difference in proportions
- *P value gives the probability of the observed difference in proportions if the null hypothesis were true*

Chi-square

- 1 State the null hypothesis.
- 2 Calculate the expected numbers if H_0 were true.
- 3 Calculate a test statistic that measures how far the observed numbers are from the expected.
- 4 Compare this test statistic with its theoretical distribution. Calculate the probability that this result (or one more extreme) could have occurred by chance.
- 5 Interpret the result: assess the strength of the evidence against the null hypothesis.

Measures of association—

Measures of association

- Odds ratio (OR) =
$$\frac{\text{Odds in exposed group}}{\text{Odds in unexposed group}}$$
- $OR = ad/bc$
- SE for the $\log(OR) = \sqrt{\{1/a + 1/b + 1/c + 1/d\}}$
- 95% CI = OR/EF to $OR \times EF$
- Where $EF = \text{Error factor} = \exp(1.96 \times \log(SE))$
- In R— use `cc` or `cs`