DISCOVERING STATISTICS USING R

Categorical Data

ANDY FIELD

Aims

- Categorical data
  - One sample problem
  - Two-way contingency tables
  - Odds ratio

Categorical Data

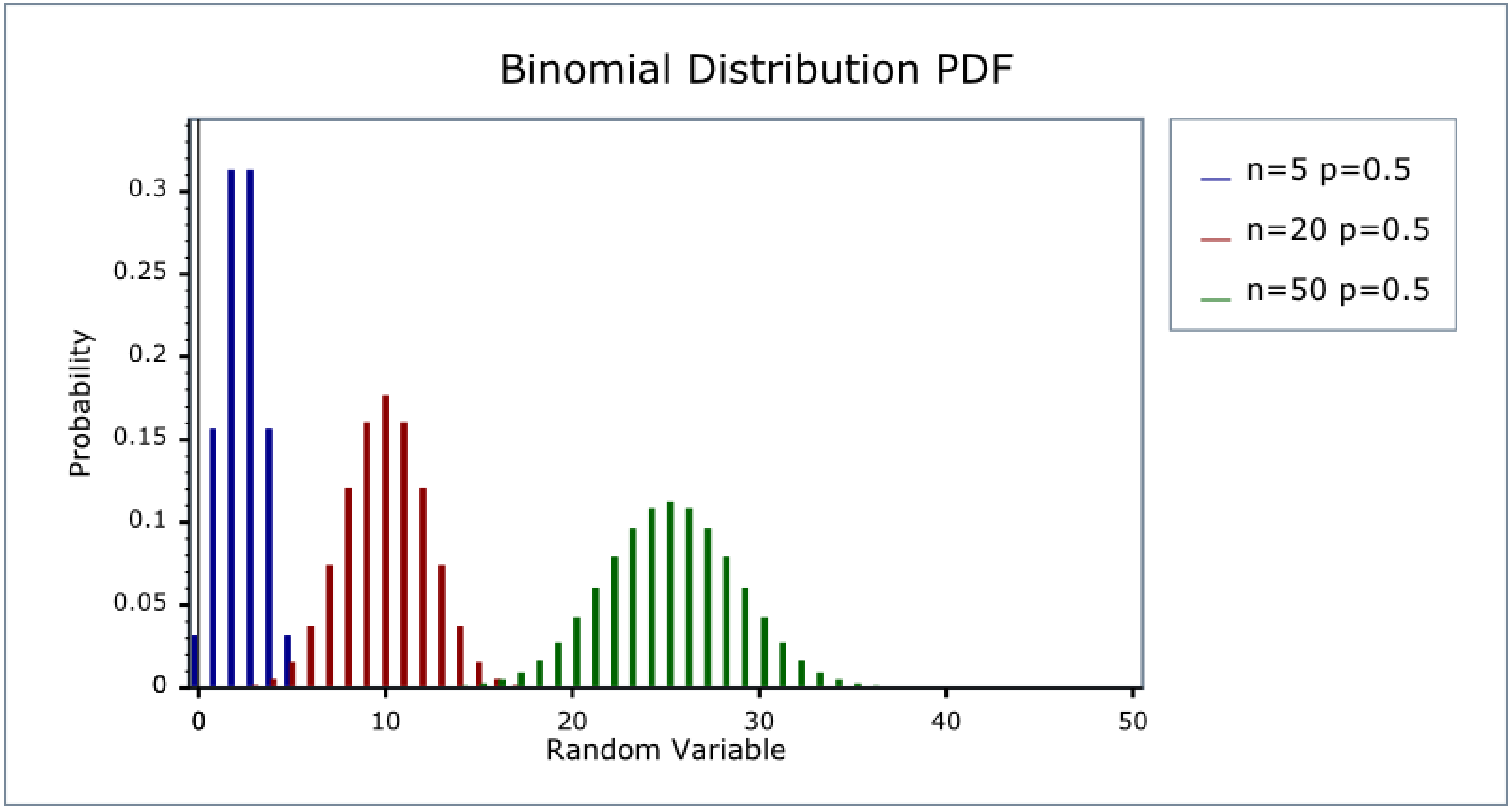
- Sometimes we have data consisting of the frequency of cases falling into unique categories
- Examples:
  - Number of people voting for different politicians
  - Numbers of students who pass or fail their degree in different subject areas
  - Number of patients who are ‘free from diagnosis’ (or not) following a treatment

The binomial distribution

- When working with categorical data, the normal distribution is no longer appropriate
- tests for proportions are based on the binomial distribution.
- the **binomial distribution** gives you the point probabilities of getting k successfull trials out of n trials given a probability π of success.
  - $P(X = k) = \binom{n}{k}\pi^k(1-\pi)^{(n-k)}$
  - $E(X) = n\pi$
  - $Var(X) = n\pi(1-\pi)$

- Eg: if you were rolling fair, 6 sided dice 50 times, the probability of rolling a five 10 times is 11.6%
  - $P(X = 10)$ = dbinom(x=10,size = 50,prob = 1/6) =0.116
  - Expected times that you will roll a 5 =E(X) = $n\pi = 50(1/6) = 8$

Binomial distribution



One sample problem

- Test of proportion
  - If you have one proportion that you would like to test whether it is significantly different from some a priori assumption, you can use prop.test() or binom.test()
  - Eg: test whether a proportion is significantly different from a population where the probability of success is 8/20 :

```
> binom.test( x=17,n=25,p=8/20)
```

One sample problem

- output

```
Exact binomial test

data:  17 and 25
number of successes = 17, number of trials = 25, p-value = 0.006693
alternative hypothesis: true probability of success is not equal to 0.4
95 percent confidence interval:
 0.4649993 0.8505046
sample estimates:
probability of success
                  0.68
```

One sample problem

- Prop.test(x,n)
- Note that by default:
  - the null hypothesis $\pi = .5$ is tested against the two-sided alternative $\pi \neq .5$;
  - a 95% confidence interval for π is calculated
  - both the test and the CI incorporate a continuity correction.
  - Any of these defaults can be changed. The call above is equivalent to

```
> prop.test(x, n, p = .5, alternative="two.sided",
conf.level = 0.95, correct = TRUE)
```

One sample problem

- One categorical variable with more than two possible outcomes

  Eg. A cross was performed between vestigial and sepia flies. We expect a ratio of 9:3:3:1 of
  - flies with normal eyes and wings
  - flies with vestigial wings and normal eyes
  - flies with normal wings and sepia eyes
  - flies with vestigial wings and sepia eyes

- Chi-square test

$$\chi^2 = \sum_{i=1}^{k} \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

One sample problem

- Eg:
  - A cross was performed between vestigial and sepia flies. We expect a ratio of 9:3:3:1 of
    - flies with normal eyes and wings
    - flies with vestigial wings and normal eyes
    - flies with normal wings and sepia eyes
    - flies with vestigial wings and sepia eyes
  - R code:

```
> drosophila=c(52,16,21,3)
> n=sum(drosophila)
> p.expected=c(9/16,3/16,3/16,1/16)
> n.expected=p.expected*n
> chisq.test(drosophila, p = p.expected, rescale.p = TRUE, correct=FALSE)
```

One sample problem

- Outcome

Chi-squared test for given probabilities

data: drosophila
X-squared = 3.3785, df = 3, p-value = 0.3369

Two-way contingency tables

- Analysing two or more categorical variables
  - The mean of a categorical variable is meaningless
    - The numeric values you attach to different categories are arbitrary
    - The mean of those numeric values will depend on how many members each category has.
  - Therefore, we analyse frequencies
- An example
  - Can animals be trained to line-dance with different rewards?
  - Participants: 200 cats
  - Training
    - The animal was trained using either food or affection, not both
  - Dance
    - The animal either learnt to line-dance or it did not
  - Outcome:
    - The number of animals (frequency) that could dance or not in each reward condition
  - We can tabulate these frequencies in a **contingency table**

A Contingency Table

> xtabs(~ Training + Dance, data=catsData)

```
Dance
Training                  Yes   No
  Food as Reward           28   10
  Affection as Reward      48  114
```

Table 18.1: Contingency table showing how many cats will line dance after being trained with different rewards

		Training		
		Food as Reward	Affection as Reward	Total
Could They Dance?	Yes	28	48	76
	No	10	114	124
	Total	38	162	200

Two-way contingency tables

- Comparing two (row) proportions
  - prop.test(cats.matrix, correct=F)

	yes	no	prop
food	28	10	28/38=0.74
affection	48	110	48/162=0.30

```
        2-sample test for equality of proportions without continuity correction

data:  data
X-squared = 25.356, df = 1, p-value = 4.767e-07
alternative hypothesis: two.sided
95 percent confidence interval:
 0.2838731 0.5972186
sample estimates:
   prop 1    prop 2
0.7368421 0.2962963
```

Pearson's Chi-Square Test of independence

- Use to see whether there's a relationship between two categorical variables
  - Compares the frequencies you observe in certain categories to the frequencies you might expect to get in those categories by chance.
- The equation:

$$\chi^2 = \sum \frac{\left(Observed_{ij} - Model_{ij}\right)^2}{Model_{ij}}$$

  - *i* represents the rows in the contingency table and *j* represents the columns.
  - The observed data are the frequencies the contingency table
- The 'model' is based on 'expected frequencies'.
  - Calculated for each of the cells in the contingency table.
  - *n* is the total number of observations (in this case 200).

$$\text{Model}_{ij} = E_{ij} = \frac{\text{Row Total}_i \times \text{Column Total}_j}{n}$$

- Test statistic
  - Checked against a distribution with $(r - 1)(c - 1)$ degrees of freedom.
  - If significant then there is a significant association between the categorical variables in the population.
  - The test distribution is approximate so in small samples use ***Fisher's exact test***.

	yes	no	Row totals
food	28	10	n1.= 38
affection	48	110	n2.= 162
Col totals	n.1=76	n.2=124	

$$\text{Model}_{\text{Food, Yes}} = \frac{\text{RT}_{\text{Yes}} \times \text{CT}_{\text{Food}}}{n} = \frac{76 \times 38}{200} = 14.44$$

$$\text{Model}_{\text{Food, No}} = \frac{\text{RT}_{\text{No}} \times \text{CT}_{\text{Food}}}{n} = \frac{124 \times 38}{200} = 23.56$$

$$\text{Model}_{\text{Affection, Yes}} = \frac{\text{RT}_{\text{Yes}} \times \text{CT}_{\text{Affection}}}{n} = \frac{76 \times 162}{200} = 61.56$$

$$\text{Model}_{\text{Affection, No}} = \frac{\text{RT}_{\text{No}} \times \text{CT}_{\text{Affection}}}{n} = \frac{124 \times 162}{200} = 100.44$$

$$\chi^2 = \frac{(28-14.44)^2}{14.44} + \frac{(10-23.56)^2}{23.56} + \frac{(48-61.56)^2}{61.56} + \frac{(114-100.44)^2}{100.44}$$

$$= \frac{(13.56)^2}{14.44} + \frac{(-13.56)^2}{23.56} + \frac{(-13.568)^2}{61.56} \frac{(13.56)^2}{100.44}$$

$$= 12.73 + 7.80 + 2.99 + 1.83$$

$$= 25.35$$

Likelihood Ratio Statistic

- An alternative to Pearson's chi-square, based on maximum-likelihood theory.
  - The resulting statistic compares observed frequencies with those predicted by the model
  - *i* and *j* are the rows and columns of the contingency table and ln is the natural logarithm

$$G = 2\sum Observed_{ij} ln \frac{Observed_{ij}}{Expected_{ij}}$$

- Test statistic
  - Has a chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom.
  - Preferred to the Pearson's chi-square when samples are small.

Likelihood Ratio Statistic

$$G = 2\left(28 ln\frac{28}{14.44} + 10 ln\frac{10}{23.56} + 48 ln\frac{48}{61.56} + 114 ln\frac{114}{100.44}\right)$$
$$G = 24.94$$

Comparing two proportions: odds ratios

-- The difference π_1-π_2

- If X and Y independent then (π_1-π_2)=0
- however 0.1-0.01=0.09 and 0.5—0.41=0.09

– Relative risk r= π_1/π_2

- If X and Y independent then r=1
- Disadvantage $\pi_1/\pi_2 \neq (1-\pi_1)/(1-\pi_2)$
- Lungcancer example: p_1=688/1338=0.51, p_2=21/80=0.26
  p_1/p_2=1.96 and $(1-p_1)/(1-p_2)$ =0.66

– Odds ratio θ

	Y	
X	success	failure
Group 1	π_1	1-π_1
Group 2	π_2	1-π_2

	lungcancer	
Smoking	cases	controls
Yes	688	650
no	21	59

Odds Ratios

- An *odds ratio* indicates how much more likely, with respect to odds, a certain event occurs in one group relative to its occurrence in another group.
  - Odds: $\Omega=\pi/(1-\pi)$
    - In a 2x2 table:

	Y	
X	success	failure
Group 1	π_1	$1-\pi_1$
Group 2	π_2	$1-\pi_2$

within a row i are the odds of succes vs failure: $\Omega_i=\pi_i/(1-\pi_i)$

  - the odds ratio $\theta=\Omega_1/\Omega_2$
  - If X and Y independent then θ=1

Odds ratio

– Odds ratio θ

Lungcancer example:

p_1=688/1338=0.51, p_2=21/80=0.26

(1-p_1) = 1-0.51 = 650/1338=0.49, (1-p_2) = 1-0.26 = 59/80=0.74

P_1 /(1-p_1) = 688/650 = 1.06

p_2/(1-p_2)= 21/59 = 0.36

Odds ratio = {P_1 /(1-p_1) }/ {P_2 /(1-p_2) }=2.97

The odds for having lungcancer is 2.97 times higher in the smoking group than in the no smoking group

	Y	
X	success	failure
Group 1	π_1	1-π_1
Group 2	π_2	1-π_2

	lungcancer	
Smoking	cases	controls
Yes	688	650
no	21	59

Probability versus Odds of an Outcome

	Outcome		
	No	**Yes**	**Total**
Group A	20	60	80
Group B	10	90	100
Total	30	150	180

Total **Yes** outcomes in Group B ÷ Total outcomes in Group B

Probability of a **Yes** in Group B = **90** ÷ **100** = **0.9**

Probability versus Odds of an Outcome

	Outcome		
	No	**Yes**	**Total**
Group A	20	60	80
Group B	10	90	100
Total	30	150	180

Probability of **Yes** in Group B = 0.90

÷

Probability of **No** in Group B = 0.10

Odds of **Yes** in Group B = **0.90** ÷ **0.10** = **9**

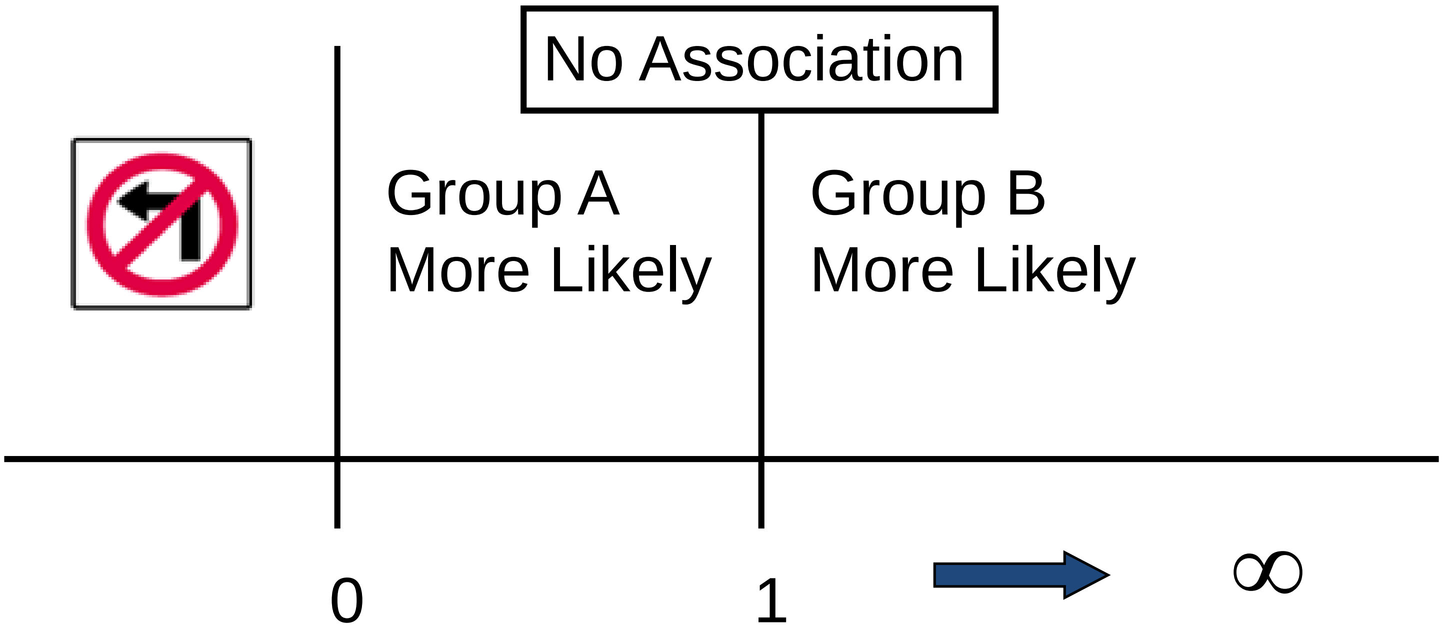
Odds Ratio

	Outcome		
	No	**Yes**	**Total**
Group A	20	60	80
Group B	10	90	100
Total	30	150	180

Odds of Yes in **Group B** = **9** ÷ Odds of Yes in **Group A** = **3**

Odds Ratio, **B** to **A** = **9** ÷ **3** = **3**

Properties of the Odds Ratio, B to A



Important Points

- The chi-square test has two important assumptions:
  - Independence:
    - Each person, item or entity contributes to only one cell of the contingency table.
  - The expected frequencies should be greater than 5.
    - In larger contingency tables up to 20% of expected frequencies can be below 5, but there is a loss of statistical power.
    - Even in larger contingency tables no expected frequencies should be below 1.
    - If you find yourself in this situation consider using Fisher's exact test.
- Proportionally small differences in cell frequencies can result in statistically significant associations between variables if the sample is large enough
  - Look at row and column *percentages* to interpret effects.

Fisher's exact test

- When samples are small, the distributions of χ^2 and G are not well approximated by the chi-squared distribution. In such situation we can perform inference using exact distributions
- We may use exact tests when
  - the row totals n_{i+} and the column totals n_{+j} are both fixed by design of the study

Example - Lady tea tasting

– In a summer tea-party in Cambridge, England, a lady claimed to be able to discern, by taste alone, whether a cup of tea with milk had the tea poured first or the milk poured first. An experiment was performed by Sir R.A. Fisher himself, then and there, to see if her claim is valid. Eight cups of tea are prepared and presented to her in random order. Four had the milk poured first, and four had the tea poured first. The lady tasted each one and rendered her opinion.

Example - Lady tea tasting

```
              Lady_says
Truth    Tea first Milk first
   Tea           3          1
   Milk          1          3
```

– The row totals are fixed by the experimenter. The column totals are fixed by the lady, who knows that four of the cups are "tea first" and four are "milk first."

Entering data as a Contingency Table

```
cats.matrix=matrix(c(28,10,48,114),ncol=2,byrow=TRUE,
dimnames=list(training=c("food reward","affection reward"),
                    dancing=c(“yes”,”no”)))
```

The resulting data look like this:

Training/dancing	yes	no
Food reward	28	10
Affection reward	48	110

Cross tabulation with tests of independence

- Crosstable() {gmodels}
  - For raw data, the function takes the basic form:

CrossTable(predictor, outcome, fisher = TRUE, chisq = TRUE, expected = TRUE, sresid = TRUE, format = "SAS"/"SPSS")

  - and for a contingency table:

CrossTable(contingencyTable, fisher = TRUE, chisq = TRUE, expected = TRUE, sresid = TRUE, format = "SAS"/"SPSS")

Output from the *CrossTable()* Function

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  200

                    | catsData$Dance
  catsData$Training |       Yes |        No | Row Total |
--------------------|-----------|-----------|-----------|
     Food as Reward |        28 |        10 |        38 |
                    |    14.440 |    23.560 |           |
                    |    12.734 |     7.804 |           |
                    |   73.684% |   26.316% |   19.000% |
                    |   36.842% |    8.065% |           |
                    |   14.000% |    5.000% |           |
                    |     3.568 |    -2.794 |           |
--------------------|-----------|-----------|-----------|
Affection as Reward |        48 |       114 |       162 |
                    |    61.560 |   100.440 |           |
                    |     2.987 |     1.831 |           |
                    |   29.630% |   70.370% |   81.000% |
                    |   63.158% |   91.935% |           |
                    |   24.000% |   57.000% |           |
                    |    -1.728 |     1.353 |           |
--------------------|-----------|-----------|-----------|
       Column Total |        76 |       124 |       200 |
                    |   38.000% |   62.000% |           |
--------------------|-----------|-----------|-----------|
```

```
Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  25.35569     d.f. =  1     p =  4.767434e-07

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  23.52028     d.f. =  1     p =  1.236041e-06

Fisher's Exact Test for Count Data
------------------------------------------------------------
Sample estimate odds ratio:  6.579265

Alternative hypothesis: true odds ratio is not equal to 1
p =  1.311709e-06
95% confidence interval:  2.837773 16.42969

Alternative hypothesis: true odds ratio is less than 1
p =  0.9999999
95% confidence interval:  0 14.25436

Alternative hypothesis: true odds ratio is greater than 1
p =  7.7122e-07
95% confidence interval:  3.193221 Inf


       Minimum expected frequency: 14.44
```

DEMO CATEGORICAL DATA

Open the program Ch18_categorical.R

Exercises

- Task 1
  - Say you were studying genetic inheritance, and your theory predicted that 3/4 of the offspring of two pea plants would be giants, and 1/4 would be dwarves. After breeding them, you end up with 682 giants, and 243 dwarves, for a total of 925 offspring. So, 73.7% of the offspring were giants, but is that significantly different from 75%?

Exercises

- Task 2
  - Students have analysed a vial of F1 fruit flies from a cross between wingless red-eyed (apterous) females and winged sepia-eyed (sepia) males. They counted 200 F2 flies: 108 were wild-type, 40 were apterous, 35 sepia and 17 apterous sepia. The expected F2 ratio is 9 wild-type: 3 apterous: 3 sepia: 1 apterous sepia. Determine whether the gene for eye colour and wing length are linked together.

exercises

- Task 3
  - From a microarray study, the researcher found 350 sign differentially expressed genes, of which 7 belonged to the GO category immune response to tumor cell. In the whole genome there are 638118 gene products of which 77 belong to the GO category immune response to tumor cell. Complete the two by two table and perform the correct test for overrepresentation.
  - Hint. One-sided Fisher exact test

	GO: immune	GO: rest	
DE	7		350
Not DE			
	77		638118

Exercises

- Task 4
  - Imagine we have 250 individuals, where some of them have a given disease and the rest do not. We observe that 20% of the individuals that are homozygous for the minor allele (aa) have the disease compared to 10% of the rest. Is there an association between the marker and the disease?

	AA/Aa	aa
control	180	40
cases	20	10