Analysis of Categorical Data

Anders Stockmarr
Course developers: Anders Stockmarr, Elisabeth Wreford Andersen

DTU Department of Applied Mathematics and Computer Science
Section for Statistics and Data Analysis
Technical University of Denmark
anst@dtu.dk

January 9th, 2025

Plan for this week

Monday	Statistical inference, and the t-test
Tuesday	Simple and Multiple regression
Wednesday	ANOVA, ANCOVA, and linear models
Thursday	Categorical data, Writing statistical reports, Logistic regression
Friday	Introduction to repeated measures , Principal Component Analysis

Outline

Outline

Categorical Data

- Binary data
  - Yes/No
  - Dead/Alive
- Nominal ("label", several groups)
  - Eye colour: Blue/ Brown / Grey / Green
  - Where do you live: Denmark, Germany, Sweden.
- Ordinal
  - How do you feel today?: Very unhappy, unhappy, OK, happy, very happy.
  - Do you try to eat healthily?: Never, Sometimes, Always
- Interval (does have a numerical distance between values)
  - BMI categories (<25, 25-30, 30+).
  - Annual income groups.

Example: Colour Blind

We have a study of 270 children where we have registered whether they were colour blind or not.

	Colour blind		
	Yes	No	Total
Girls	1	119	120
Boys	6	144	150
Total	7	263	270

Example: Colour Blind

We have a study of 270 children where we have registered whether they were colour blind or not.

	Colour blind		
	Yes	No	Total
Girls	1	119	120
Boys	6	144	150
Total	7	263	270

```
  Boy Colour_blind Count
1 1 6
1 0 144
0 1 1
0 0 119
```

Outcome: Colour blind yes/no
Covariate: Sex boy/girl.

Example: Tables in R

```
colourTab <- xtabs(Count ~ Boy + Colour_blind, data = colour_dat)
```

```
#Print the table
ftable(colourTab)

   Colour_blind   0   1
Boy
0               119   1
1               144   6


#Row totals
margin.table(colourTab,1)

Boy
  0   1
120 150
```

```
#Row percentages
prop.table(colourTab,1)

  Colour_blind
Boy           0           1
  0 0.991666667 0.008333333
  1 0.960000000 0.040000000

#Column total
margin.table(colourTab,2)

Colour_blind
  0   1
263   7
```

Risk Ratio

We want to compare the probability that a boy is colour blind (p_1) with the probability that a girl is colour blind (p_0).

- The probabilities are unknown.
- Variation from random sampling of children for the study.
- Estimate the probabilities

$$\begin{aligned}\hat{p}_1 &= \frac{\text{"}number\ colour\ blind\ boys\text{"}}{\text{"}number\ of\ boys\text{"}} = \frac{6}{150} = 0.04 \\ \hat{p}_0 &= \frac{\text{"}number\ colour\ blind\ girls\text{"}}{\text{"}number\ of\ girls\text{"}} = \frac{1}{120} = 0.0083\end{aligned}$$

Risk Ratio

We want to compare the probability that a boy is colour blind (p_1) with the probability that a girl is colour blind (p_0).

- The probabilities are unknown.
- Variation from random sampling of children for the study.
- Estimate the probabilities

$$
\begin{aligned}
\hat{p}_1 &= \frac{\text{"}number\ colour\ blind\ boys\text{"}}{\text{"}number\ of\ boys\text{"}} = \frac{6}{150} = 0.04 \\
\hat{p}_0 &= \frac{\text{"}number\ colour\ blind\ girls\text{"}}{\text{"}number\ of\ girls\text{"}} = \frac{1}{120} = 0.0083
\end{aligned}
$$

- A measure to compare probabilities
    - Risk Ratio (RR)$=\frac{p_1}{p_0}$

Example: Colour Blind

* How many colour blind children would we expect?

* Assume that the probability of colour blindness is 0.026 (7 out of 270), independent of gender.
  Then we would expect:
  - for 150 boys: 150*0.026=3.9
  - for 120 girls: 120*0.026=3.1

* We observed 6 and 1; we need statistical methods to decide whether this was a coincidence, or whether colour blindness differs for girls and boys.

Binomial Distribution

X=Number of events (colour blind children) out of N, with p= the probability of event.

$$P(X = x) = \binom{N}{x} p^x (1-p)^{N-x}$$

Here p is the unknown parameter (the probability of colour blind). Our best guess at p (the estimate) is the observed proportion of colour blind.

$$\hat{p} = \frac{x}{N} = \frac{7}{270}$$

Binomial Distribution Approximate CI

If N is large then

$$s.e.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

The approximate confidence interval for $\hat{p}$ using a Normal approximation:

$$\hat{p} \pm 1.96 s.e.(\hat{p})$$

Binomial Distribution Approximate CI

If N is large then

$$s.e.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

The approximate confidence interval for $\hat{p}$ using a Normal approximation:

$$\hat{p} \pm 1.96 s.e.(\hat{p})$$

For the girls and boys from the example:

```
library(epitools)
binom.approx(colourTab[,2],margin.table(colourTab,1))

  x n.Boy n.Freq  proportion.Freq  lower.Freq     upper.Freq
0 1     0    120    0.008333333   -0.007931503   0.02459817
1 6     1    150    0.040000000    0.008640576   0.07135942
```

Binomial Distribution 'Exact' CI

Uses the correspondence between test and confidence interval (CI).
The exact CI includes the p's that would be accepted in an 'exact' test.
For the girls and boys from the example:

```
binom.exact(colourTab[ , 2], margin.table(colourTab, 1))

   x n.Boy n.Freq  proportion.Freq        lower       upper
0 1     0     120   0.008333333      0.0002109595 0.04555551
1 6     1     150   0.040000000      0.0148185211 0.08502781
```

Compare p_0 and p_1

1. Risk ratio: $\frac{p_1}{p_0}$.
2. Odds ratio: $\frac{p_1}{1-p_1} / \frac{p_0}{1-p_0}$.

But what are odds? - and why do we need them?

Odds and Probability

- Definition of odds:

$$
\begin{aligned}
Odds(A) &= \frac{Probability(A)}{Probability(A \text{ does not happen})} \\
&= \frac{Probability(A)}{1 - Probability(A)}
\end{aligned}
$$

- Back to probabilities

$$
Probability(A) = \frac{Odds(A)}{1 + Odds(A)}
$$

Odds and Probability

- At the bookmaker: The odds for "SønderjyskE" winning against FCK.
- DanskeSpil 2 November 2014 odds=6.1.

$$Odds(SønderjyskE\, wins) = \frac{P(SønderjyskE\ wins)}{P(SønderjyskE\ does\ not\ win)} = \frac{1}{6.1}$$

- The probability that SønderjyskE wins:

$$\frac{1/6.1}{1 + 1/6.1} = 0.14$$

- (Result: FCK-SønderjydskE 1-1)

Characteristics of Odds

- ✓ Odds are between 0 and infinity.
- ✓ Often log odds (no boundaries).
- ✓ When the probability is 0.5 then odds are 1.
- ✓ Odds are larger than probability.
- ✓ Note: When the probability is small (≤ 0.1) then probability and odds nearly equal.

Odds Ratio

The odds ratio is the ratio between the odds in the two groups.

$$OR = \frac{Odds(group_1)}{Odds(group_0)} = \frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0}$$

	Response	
Group	No	Yes
0 (ref)	a	b
1	c	d

$$OR = \frac{d/(c+d)/c/(c+d)}{b/(a+b)/a(a+b)} = \frac{d/c}{b/a} = \frac{ad}{bc}$$

OR in R

```
library(epitools)
oddsratio(colourTab, method = "wald")

$data
       Colour_blind
Boy        0 1 Total
  0      119 1   120
  1      144 6   150
  Total 263 7   270

$measure
   odds ratio with 95% C.I.
Boy estimate     lower    upper
  0 1.000000        NA       NA
  1 4.958333 0.5887152 41.76055
```

OR in R

```
library(epitools)
oddsratio(colourTab, method = "wald")

$data
       Colour_blind
Boy        0 1 Total
  0      119 1   120
  1      144 6   150
  Total  263 7   270

$measure
   odds ratio with 95% C.I.
Boy estimate     lower    upper
  0 1.000000        NA       NA
  1 4.958333 0.5887152 41.76055
```

Odds for a boy being colour blind are 4.96 (95% CI 0.6 to 41.8) times larger than for girls.

RR in R

Remember the $RR = \frac{p_1}{p_0}$

```
riskratio(colourTab)

$data
       Colour_blind
Boy         0 1 Total
  0       119 1   120
  1       144 6   150
  Total   263 7   270

$measure
   risk ratio with 95% C.I.
Boy estimate     lower    upper
  0      1.0        NA       NA
  1      4.8 0.5858252 39.32914
```

RR in R

Remember the $RR = \frac{p_1}{p_0}$

```
riskratio(colourTab)

$data
       Colour_blind
Boy       0 1 Total
  0     119 1   120
  1     144 6   150
  Total 263 7   270

$measure
   risk ratio with 95% C.I.
Boy estimate     lower    upper
  0      1.0        NA       NA
  1      4.8 0.5858252 39.32914
```

The risk of a boy being colour blind is 4.8 (95% CI 0.6 to 39.3) times larger than for girls.

Odds Ratio and Risk Ratio

- OR varies freely from 0 to infinity.
- RR always between 1 and OR.
- OR is symmetric

$$OR(response = 1) = \frac{1}{OR(response = 0)}$$

- RR is not symmetric

$$RR(response = 1) \neq \frac{1}{RR(response = 0)}$$

- For rare events, $OR \approx RR$.

χ^2 (Chisquare) Test

The Hypothesis: OR=1 or equivalently RR=1.

Observed:

	Response		
Group	No	Yes	Total
0 (ref)	a	b	a+b
1	c	d	c+d
Total	a+c	b+d	N

χ^2 (Chisquare) Test

The Hypothesis: OR=1 or equivalently RR=1.

Observed:

	Response		
Group	No	Yes	Total
0 (ref)	a	b	a+b
1	c	d	c+d
Total	a+c	b+d	N

Expected:

	Response		
Group	No	Yes	Total
0 (ref)	(a+b)(a+c)/N	(a+b)(b+d)/N	a+b
1	(c+d)(a+c)/N	(c+d)(b+d)/N	c+d
Total	a+c	b+d	N

$$\chi^2 = \sum \frac{(Obs - Expected)^2}{Expected}$$

Test Colour Blind

Are the odds of being colour blind the same for boys and girls?
Equivalently is colour blindness independent of sex?

$$H_0: \ p_0 = p_1$$

Use:

- χ^2 test.

```
> chisq.test(colourTab, correct=FALSE)

Pearson's Chi-square test

data:  colourTab
X-squared = 2.6472, df = 1, p-value = 0.1037
```

Test in Example

The Hypothesis: OR=1 or equivalently RR=1.
Observed:

Obs / Expected / Boy	Colour Blind No	Colour Blind Yes	Total
0 (ref)	119	1	120
	116.9	3.1	
1	144	6	150
	146.1	3.9	
Total	263	7	270

OR and Chi2 test in R

```
> epitools::oddsratio(colourTab, method="wald")
$data
       Colour_blind
Boy        0 1 Total
  0      119 1   120
  1      144 6   150
  Total  263 7   270

$measure
   odds ratio with 95% C.I.
Boy estimate     lower    upper
  0 1.000000        NA       NA
  1 4.958333 0.5887152 41.76055

$p.value
   two-sided
Boy midp.exact fisher.exact chi.square
  0         NA           NA         NA
  1  0.1200585    0.1363846  0.1037323
```

OR and Chi2 test in R

```
> epitools::oddsratio(colourTab, method="wald")
$data
       Colour_blind
Boy       0 1 Total
  0     119 1   120
  1     144 6   150
  Total 263 7   270

$measure
   odds ratio with 95% C.I.
Boy estimate     lower    upper
  0 1.000000        NA       NA
  1 4.958333 0.5887152 41.76055

$p.value
   two-sided
Boy midp.exact fisher.exact chi.square
  0         NA           NA         NA
  1  0.1200585    0.1363846  0.1037323
```

The hypothesis of OR=1 is accepted p=0.14 > 0.05, but CI very wide.

Exercise

Identify each variable as nominal, ordinal or interval.

1. UK political party preference (Labour, Conservative, Social Democrat).
2. Depression rating (none, mild, moderate, severe, very severe).
3. Patient survival (in number of months).
4. University location (Lyngby, Copenhagen, Odense, Aarhus, Aalborg).
5. Favorite beverage (water, juice, milk, soft drink, beer, wine).
6. Appraisal of company's inventory level (too low, about right, too high).

Outline

Confounding

Instead of just a risk factor (boy/girl) and an outcome (colour blindness) one might have a third factor.

Example: Two treatments (A and B) for kidney stone. The outcome is success or failure of the treatment. We also have registered whether the stone was small or large.

	Treatment	Stone	Success1	Count
1	A	Small	1	81
2	A	Small	0	6
3	A	Large	1	192
4	A	Large	0	71
5	B	Small	1	234
6	B	Small	0	36
7	B	Large	1	55
8	B	Large	0	25

Tables in R

```
mytable <- xtabs(Count ~ Treatment + Success1 + Stone,
data = kidney)
ftable(mytable)

                 Stone Large Small
Treatment Success1
A         0                71     6
          1               192    81
B         0                25    36
          1                55   234
```

The order of the variables in xtabs is important. First exposure, second outcome, last extra factors.

Ignoring the Size of the Stone

```
Treat_Succ <- margin.table(mytable, 1:2)
oddsratio(Treat_Succ, method = "wald")

$data
         Success1
Treatment   0   1 Total
    A      77 273   350
    B      61 289   350
    Total 138 562   700

$measure
         odds ratio with 95% C.I.
Treatment estimate     lower    upper
        A 1.000000        NA       NA
        B 1.336276 0.9188954 1.943238
```

The odds of success for treatment B are 1.34 times the odds for A.

The Effect of Treatment for Small Stones

```
Small <- mytable[ , , 2]
oddsratio(Small, method = "wald")

$data
         Success1
Treatment  0   1 Total
    A      6  81    87
    B     36 234   270
    Total 42 315   357

$measure
         odds ratio with 95% C.I.
Treatment  estimate     lower    upper
        A 1.0000000        NA       NA
        B 0.4814815 0.1956696 1.184775
```

Treatment A is better for small stones OR=0.48.

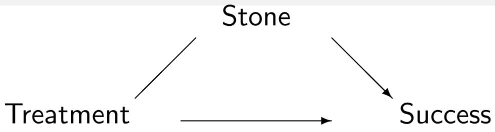
The Effect of Treatment for Large Stones

```
Large <- mytable[ , , 1]
oddsratio(Large, method = "wald")

$data
        Success1
Treatment  0   1 Total
    A     71 192   263
    B     25  55    80
   Total 96 247   343

$measure
         odds ratio with 95% C.I.
Treatment  estimate   lower    upper
        A 1.0000000      NA       NA
        B 0.8135417 0.47147 1.403801
```

Treatment A is better for large stones OR=0.81.

How can this happen?

Confounding



A confounder is:

- Associated with outcome:
  e.g., smaller kidney stones have higher rate of success.
- Associated with the treatment:
  e.g., doctors have chosen treatment A for difficult cases.
- Not a result of treatment, i.e. not an intermediate variable.
  Not a statistical property; cannot be seen from tables; common sense is required.

Smaller kidney stones have higher rate of success.

```
Stone_Succ <- margin.table(mytable, 3:2)
oddsratio(Stone_Succ, method = "wald")

$data
       Success1
Stone     0   1 Total
  Large  96 247   343
  Small  42 315   357
  Total 138 562   700

$measure
       odds ratio with 95% C.I.
Stone   estimate    lower    upper
  Large  1.00000       NA       NA
  Small  2.91498 1.955863 4.344429
```

Smaller kidney stones have higher rate of success.

```
Stone_Succ <- margin.table(mytable, 3:2)
oddsratio(Stone_Succ, method = "wald")

$data
       Success1
Stone     0   1 Total
  Large  96 247   343
  Small  42 315   357
  Total 138 562   700

$measure
       odds ratio with 95% C.I.
Stone   estimate    lower    upper
  Large  1.00000       NA       NA
  Small  2.91498 1.955863 4.344429
```

OR 2.9 (95% CI 1.96 to 4.34) for success with small stone compared to large.

Doctors have chosen treatment A for difficult cases.

```
Stone_Treat <- margin.table(mytable, c(3,1))
oddsratio(Stone_Treat, method = "wald")

$data
       Treatment
Stone     A   B Total
  Large 263  80   343
  Small  87 270   357
  Total 350 350   700

$measure
       odds ratio with 95% C.I.
Stone   estimate   lower    upper
  Large  1.00000      NA       NA
  Small 10.20259 7.20504 14.44721
```

Doctors have chosen treatment A for difficult cases.

```
Stone_Treat <- margin.table(mytable, c(3,1))
oddsratio(Stone_Treat, method = "wald")

$data
       Treatment
Stone      A   B Total
  Large 263  80   343
  Small  87 270   357
  Total 350 350   700

$measure
       odds ratio with 95% C.I.
Stone    estimate   lower    upper
  Large  1.00000      NA       NA
  Small 10.20259 7.20504 14.44721
```

Small stones have been treated with B.

Controlling for Confounding

- We could have randomized the treatment.
- We can keep the confounder constant.

Controlling for Confounding

- We could have randomized the treatment.
- We can keep the confounder constant.

Hold the confounder constant:

- Compare treatments within strata (small stones, and large).
- If the estimates are similar we calculate a combined estimate as a suitable average (No more on this today).
- Fit a logistic regression model (more about this in the afternoon).

Outline

RxC tables

Observed Expected	Caffeine Intake 0	 1-150	 151-300	 300+	 Total
Married	652 705.83	1537 1488.01	598 578.07	242 257.09	3029
Prev. Married	36 32.86	46 69.27	38 26.91	21 11.97	141
Single	218 167.31	327 352.72	106 137.03	67 60.94	718
Total	906	1910	742	330	3888

Chi-square test in RxC tables

- As for a 2x2 table.
- Hypothesis: Caffeine intake the same irrespective of marital status (independence in table).

$$\chi^2 = \sum \frac{(\text{Observed-Expected})^2}{\text{Expected}}$$

- Follows a χ^2 distribution with (r-1)(c-1) degrees of freedom.

Chi-square test in RxC tables, contd.

- Test for independence gives a p-value, but we are not finished yet.
- If the test is significant.
  - Describe the connections. The p-values does not show where the associations are found.
- If the test is not significant.
  - There might still be some associations.
- **In both cases describe the table with percentages and plots.**

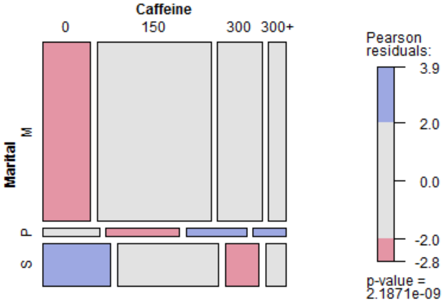
RxC tables

Observed Row %	Caffeine Intake 0	 1-150	 151-300	 300+	 Total
Married	652 21.53	1537 50.74	598 19.74	242 7.99	3029
Prev. Married	36 25.53	46 32.62	38 26.95	21 14.89	141
Single	218 30.36	327 45.54	106 14.76	67 9.33	718
Total	906	1910	742	330	3888

Pearsons $\chi^2(6) = 51.6556$, p < 0.0001

Mosaic Plot

```
library(vcd)
mosaic(mytable, shade = TRUE, legend = TRUE)
```



Outline

Exercises

- Exercise 1 Admission to Berkeley
- Exercise 2 Popular