Contingency analysis: associations between categorical variables

Chapter 9

Contingency analysis

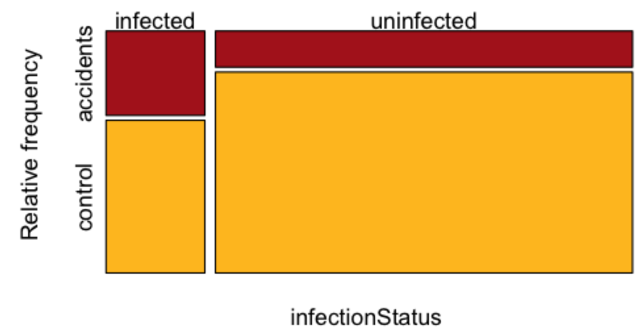
- Test the independence of two or more categorical variables
- We'll learn one kind: χ^2 contingency analysis

Toxoplasma and accidents

OBSERVED	Infected with *Toxoplasma*	Uninfected	Totals
Accident	21	38	59
No accident	38	211	249
Totals	59	249	308

Flegr, J. et al. 2002. Increased risk of traffic accidents in subjects with latent toxoplasmosis: a retrospective case-control study. BMC Infect. Dis. 2:11.

Mosaic plot



Hypotheses

H_0: Being infected with *Toxoplasma* does not affect chance of having a car accident.

H_A: Being infected with *Toxoplasma* does affect chance of having a car accident.

Calculating the expectations

With independence,

Pr[*Toxoplasma* AND accident] =

Pr[*Toxoplasma*] × Pr[accident]

Calculating the expectations

EXPECTED	Infected with Toxoplasma	Uninfected	Totals
Accident			59
No accident			249
Totals	59	249	308

Pr[Infection] = 59/308=0.1916

Calculating the expectations

EXPECTED	Infected with *Toxoplasma*	Uninfected	Totals
Accident			59
No accident			249
Totals	59	249	308

Pr[Infection] = 59/308=0.1916

Pr[No accident] = 249/308= 0.8084

Calculating the expectations

EXPECTED	Infected with *Toxoplasma*	Uninfected	Totals
Accident			59
No accident	47.7		249
Totals	59	249	308

Pr[Infection] = 59/308=0.1916

Pr[No accident] = 249/308= 0.8084

If H_0 is true,
Pr[Infection AND No accident] = (0.1916)(0.8084) = 0.1548

Expected = 0.1548 x 308 = 47.7

Calculating the expectations

EXPECTED	Infected with *Toxoplasma*	Uninfected	Totals
Accident	11.3	47.7	59
No accident	47.7	201.3	249
Totals	59	249	308

Calculating χ^2

$$\chi^2 = \sum_i \frac{(Oberved_i - Expected_i)^2}{Expected_i}$$

$$= \frac{(21-11.3)^2}{11.3} + \frac{(38-47.7)^2}{47.7} + \frac{(38-47.7)^2}{47.7} + \frac{(211-201.3)^2}{201.3}$$

$$= 12.7$$

Degrees of freedom

df= (# columns -1)(#rows -1)

For *Toxoplasma* example,
df = (2-1)(2-1) = 1

Conclusion

$\chi^2 = 12.7 \gg \chi^2_{1,\alpha=0.05} = 3.84$,

We can reject the null hypothesis of independence. Toxoplasma infection status did covary with having car accidents.

Conclusion, using R

```
chisq.test(toxoData$infectionStatus,
toxoData$driverType, correct = FALSE)
```

```
     Pearson's Chi-squared test

data:  toxoData$infectionStatus and
toxoData$driverType
X-squared = 12.733, df = 1, p-value =
0.0003593
```

Assumptions

This χ^2 test is just a special case of the χ^2 goodness-of-fit test, so the same rules apply.

You can't have any expectation less than 1, and no more than 20% < 5.

Fisher's exact test

For 2 x 2 contingency analysis

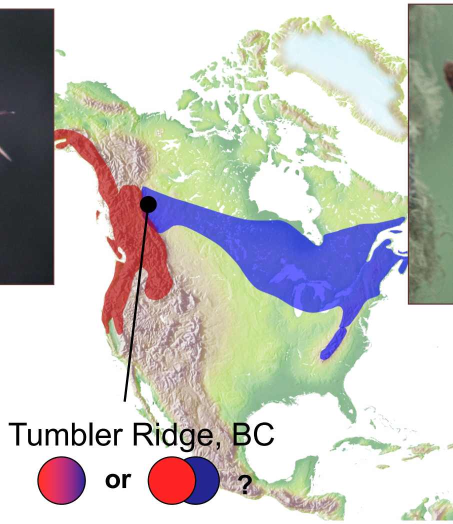
Does not make assumptions about the size of expectations

R (or other programs) will do it, but cumbersome to do by hand

Winter Wren (*Troglodytes troglodytes*)

Are western and eastern forms (currently considered subspecies) actually reproductively isolated, and therefore separate species?

T. (t.) pacificus



T. t. hiemalis

Photos by D. Irwin

Association of DNA and song: The winter wren contact zone

OBSERVED	Western song	Eastern song	Totals
Western mtDNA	12	0	12
Eastern mtDNA	0	4	4
Totals	12	4	16

Data from Toews & Irwin 2008, Molecular Ecology

Calculating the expectations

EXP.	Western song	Eastern song	Totals
Western mtDNA			12
Eastern mtDNA			4
Totals	12	4	16

A shortcut for calculating expectations (assuming H_0 is true):

Exp[row i, column j] =

$$\frac{(\textit{row i total})(\textit{column j total})}{\text{grand total}}$$

Exp[w mtDNA, w song] = 12*12/16 = 9

Comparing observed and expected

OBS.	Western song	Eastern song	Totals
Western mtDNA	12	0	12
Eastern mtDNA	0	4	4
Totals	12	4	16

EXP.	Western song	Eastern song	Totals
Western mtDNA	9	3	12
Eastern mtDNA	3	1	4
Totals	12	4	16

Too many of the expected are below 5, so we cannot use the χ^2 contingency test. Instead, we use a computer to do Fisher's exact test:

$P = 0.00055$, so we reject the H_0 of no association.

Fisher's exact test in R

```
fisher.test(wrenData$song,wrenData$mtDNA)
```

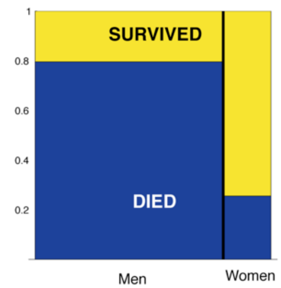
```
      Fisher's Exact Test for Count Data

data:  wrenData$song and wrenData$mtDNA
p-value = 0.0005495
alternative hypothesis: true odds ratio is not
equal to 1
95 percent confidence interval:
 4.616679       Inf
sample estimates:
odds ratio
       Inf
```

Odds

The probability of success divided by the probability of failure.

$$O = \frac{p}{1-p}$$



Odds of survival:

$$O_{men} = \frac{0.20}{1-0.20} = \frac{0.20}{0.80} = 0.25$$

Or "1 to 4"

$$O_{women} = \frac{0.74}{1-0.74} = \frac{0.74}{0.26} = 2.85$$

Or roughly "3 to 1"

Odds ratio

The odds of success in one group divided by the odds of success in another group.

Used often in medical research

$$OR = \frac{Odds\ in\ treatment\ group}{Odds\ in\ Control\ group} = \frac{\frac{Number\ of\ bad\ thing}{Number\ of\ good\ thing}}{\frac{Number\ of\ bad\ thing}{Number\ of\ good\ thing}}$$

OR<1 means treatment helps;
OR>1 means treatment makes things worse.

Odds ratio: ABO blood type and hospitalization for COVID-19

	A	other
Hospitalized	670	1105
Control	1188	2506

$$OR = \frac{\frac{670}{1188}}{\frac{1105}{2506}} = 1.28$$

Odds ratio: ABO blood type and hospitalization for COVID-19

```
ABOCOVIDData$Hospitalized = factor(ABOCOVIDData$Hospitalized,
levels = c("Hospital","Control"))
ABOCOVIDData$typeA = factor(ABOCOVIDData$typeA, levels =
c("A","notA"))

ABOtable = table(ABOCOVIDData$Hospitalized,
ABOCOVIDData$typeA)

oddsratio(ABOtable, method = "wald")
```

Odds ratio: ABO blood type and hospitalization for COVID-19

```
oddsratio(ABOtable, method = "wald")
```

```
             A    notA
Hospital     670  1105
Control      1188 2506

$data
               A  notA Total
Hospital     670  1105  1775
Control      1188 2506  3694
Total        1858 3611  5469

$measure
             odds ratio with 95% C.I.
             estimate    lower    upper
    Hospital 1.000000       NA       NA
     Control 1.279019 1.136407 1.439528
```