Contingency tables

Theory

Sarah Kaspar

EMBL

Contingency tables

We count cases and divide them into categories:

	Disease	
treatment	yes	no
treated	*4*	*96*
untreated	*10*	*90*

<u>Proportions</u>:

$^{4}/_{100} = 0.04$

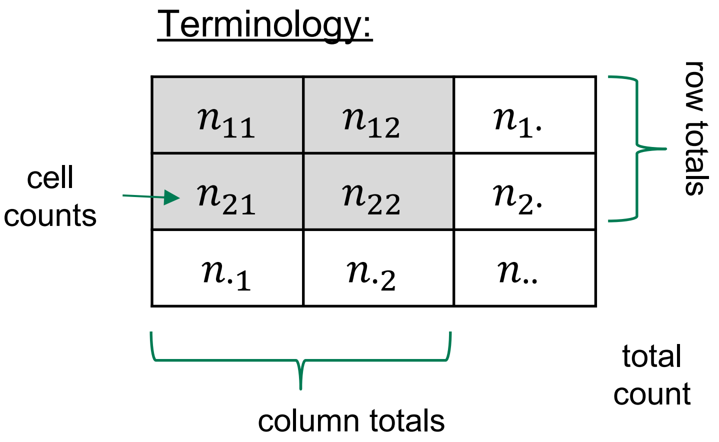
$^{10}/_{100} = 0.1$

<u>Question</u>:
Is there an association between disease and treatment?
Are the proportions of diseased persons different in the two treatment groups?

Contingency tables

We count cases and divide them into categories:

	Disease	
treatment	yes	no
treated	*4*	*96*
untreated	*10*	*90*



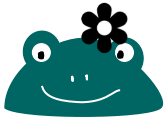
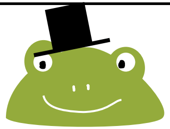
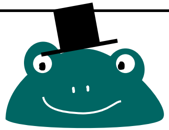
Terminology:

n_{11}	n_{12}	$n_{1.}$
n_{21}	n_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	$n_{..}$

Question:
Is there an association between disease and treatment?
Are the proportions of diseased persons different in the two treatment groups?

Contingency tables

... with frogs

	Colour	
sex	light	dark
female		
male		

To answer this question, we have to understand where the cell counts come from.

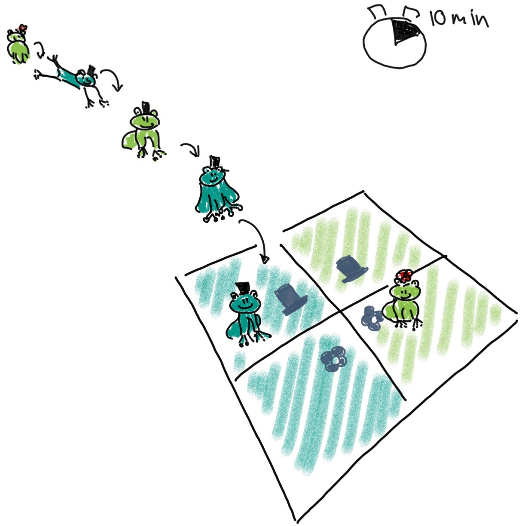
Question:
Is there an association between sex and colour?

Different study designs

Poisson sampling:

- Each category has its own Poisson rate:

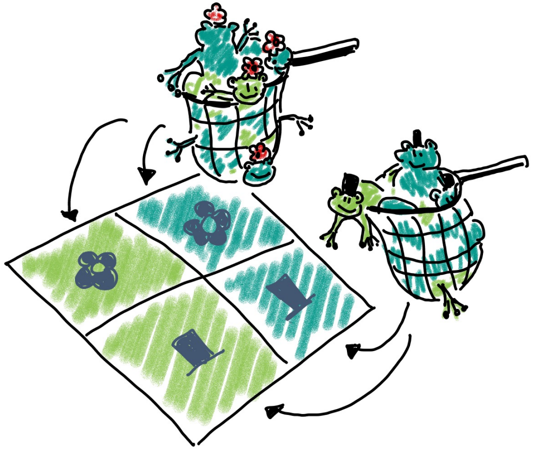
$$\lambda = \boldsymbol{n} * \boldsymbol{p}$$



Different study designs

Binomial sampling:

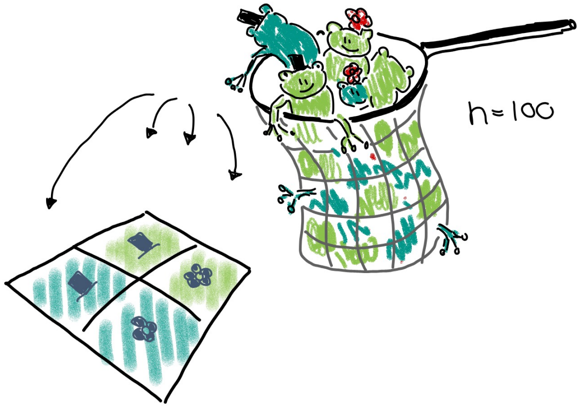
- Fixed number of n_f female frogs
- Fixed number of n_m male frogs
- For each sex, there is the probability of being light: p



Different study designs

Multinomial sampling:

- Fixed number of n frogs
- Each category has its own probability: p



Expected counts

The expected counts are the same for these study designs:

	Colour	
sex	light	dark
female	$n * p_{11}$	$n * p_{12}$
male	$n * p_{21}$	$n * p_{22}$

New question:
Are the probabilities dependent on each other?

Probability rules for independence

Independence:

$$P(A,B) = P(A) * P(B)$$

Example: Flip two coins

$$P(head, head) = P(head) * P(head) = {}^{1}/_{4}$$

The outcomes of the two coins are independent.

Association:

$$P(A,B) \neq P(A) * P(B)$$

Example: hair and eye colour

$$P(blond, blue) = P(blond) * \underline{P(blue|blond)}$$

conditional probability: blonds are more likely to have blue eyes than dark-haired

Hair and eye colour are associated.

Expected counts under H_0

Observed counts:

	disease	
treatment	yes	no
treated	$\boldsymbol{n_{11}}$	$\boldsymbol{n_{12}}$
untreated	$\boldsymbol{n_{21}}$	$\boldsymbol{n_{22}}$

$$P(treat) = \frac{n_{11} + n_{12}}{n_{..}}$$

$$P(disease) = \frac{n_{11} + n_{21}}{n_{..}}$$

marginal probabilities

Null hypothesis: disease and treatment are independent.

→ We expect that the product of the marginal probabilities is a good estimate for the cell count:

Expected counts:

$$\boldsymbol{\mu_{11} = P(disease) * P(treat) * n_{..}}$$

…

Expected counts

	Disease		
treatment	yes	no	total
treated	4	96	100
untreated	10	90	100
total	14	186	n = 200

$$P(treated) = \frac{100}{200} = 0.5$$

$$P(D) = \frac{14}{200} = 0.07$$

Expected counts
assuming independence:

$$E(n_{treated,disease})$$

$$= P(treated) * P(disease) * n_{..}$$

$$= 0.5 \ * 0.07 \ * 200$$

$$= 7$$

Chi-Square test

Statistic:

$$\chi^2 = \sum_{ij} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

O: observed count
E: expected count under H_0
i, j: row and column index
r : number of rows
c: number of columns

χ^2 quantifies the deviation from independence.

χ^2 follows a chi-squared distribution with $(r-1)*(c-1)$ degrees of freedom.

Question: How can we use this information to perform a test?

In our example

	Disease		
treatment	yes	no	total
treated	4	96	100
untreated	10	90	100
total	14	186	n = 200

Question: What else should we report?

Output from R:

```
chisq.test(array(c(4,96,10,90), dim=c(2,2)), correct=FALSE)

           Pearson's Chi-squared test

data:  array(c(4, 96, 10, 90), dim = c(2, 2))
X-squared = 2.765, df = 1, p-value = 0.09635
```

Quantifying association

Difference in proportions

$$D = P(\text{disease} \,|N) \;- P(\text{disease}|T)$$

The absolute difference between the proportions of disease cases in the two groups is **6%**.

Relative risk

$$RR = \frac{P(\text{disease} \mid T)}{P(\text{disease} \mid N)}$$

The proportion of disease cases in the treated group is **0.4** times the proportion of disease cases in the no treatment group**.**

Odds ratio

$$OR = \frac{P(\text{disease}|T)/(1 - P(\text{disease}|T))}{P(\text{diesese}|N)/(1 - P(\text{disease}|N))}$$

The odds for having the disease when treated is **0.38** times the odds when being untreated.

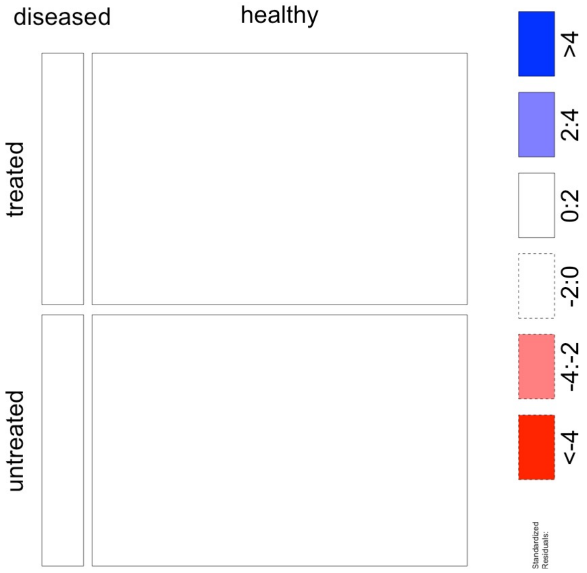
Log odds ratio

$$\log(OR)$$

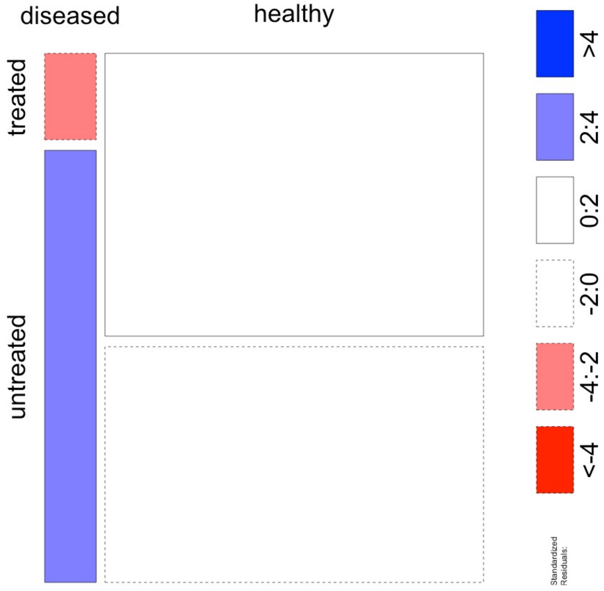
… a useful parameter for models (**-0.97**)
log(OR) > 0: disease more likely when untreated
log(OR)<0: disease more likely when treated

Visualizing association

Independence



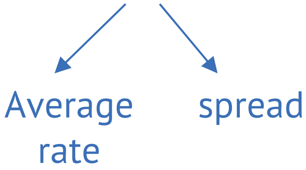
Association



Overdispersed data

Overdispersion: For each measurement, the rate lambda is slightly different (it’s drawn from a gamma-poisson distribution).

For each cell: $\lambda \sim GammaPois(\mu, \theta)$



Question: How can we estimate the spread from a contingency table?

Example: Expression counts

	Cell type	
treatment	control	cancer
treated	0	5
untreated	10	28

Research question: is there a cell-type specific response to the treatment?

Known sources of variation: individual cell-to-cell differences in expression.

Overdispersed data

→ We need more counts for each combination of variables.

→ Represent data in a data frame:

treatment	Cell type	count
untreated	control	0
untreated	control	28
untreated	control	5
untreated	cancer	37
untreated	cancer	20
…	…	…

Several measurements per combination

Gamma-poisson regression fits:

- Overdispersion
- Individual effects of cell type and treatment
- Interaction between cell type and treatment

Summary

- **Contingency table**:
  - Row and column = two different variables
  - Each cell is a count from 1 combination of the two variables (Poisson or binomial counts)

- We are interested in the association between the two variables:
  - A **chi-square test** gives a significance of the association
  - The effect size is a **measure or association**, e.g. relative risk

- **Limitations** of contingency tables:
  - Can only deal with one count per combination of variables
  - Does not allow to estimate overdispersion
  - Use regression in these cases

- Contingency tables can be **extended** to more dimensions.