Faculty of Life Sciences

Analysis of count tables

Claus Ekstrøm and Ib Skovgaard
E-mail: ims@life.ku.dk

Program

- Examples of count tables
- Goodness-of-fit test
  - The χ^2-distribution
- 2×2 tables
- $r \times k$ tables
- Fisher's exact test
- One-sided hypothesis in a 2×2 table
- Odds and odds-ratio

Example: Traps and coyotes

	Injury		
Type of trap	I	II	III
Northwood	9	7	4
Soft-Catch	11	3	6

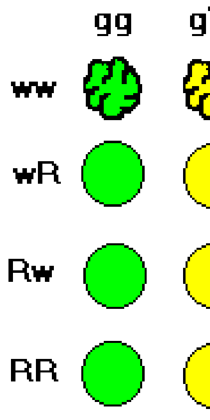
- I: No or little leg injury
- II: Injury to the skin
- III: Leg broken

Do the two types of traps give different injuries?

Example: Mendel's pea plants

Class	Number
Round, yellow	315
Round, green	108
Wrinkled, yellow	101
Wrinkled, green	32

According to Mendel's theory the frequencies should be in proportions $9:3:3:1$.



Are the data in agreement with Mendel's theory?

Goodness-of-fit test

Test of the hypothesis that the probabilities of the categories agree with specified values.

Test statistic:

$$X^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

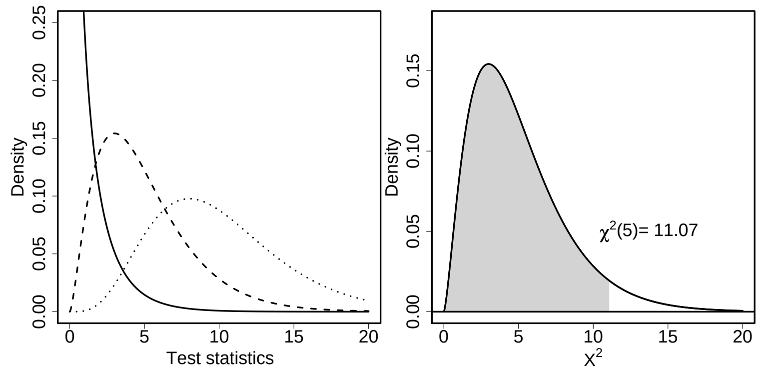
where the sum is over all categories.

- Small values of X^2: data and hypothesis agree well.
- Large values of X^2: data are in disagreement with the hypothesis.

It may be shown that if the hypothesis is true then X^2 follows (approximately) a so-called $\chi^2(r)$-distribution with r degrees of freedom, where

$$r = \text{degrees of freedom} = \text{number of categories - 1}$$

The $\chi^2(r)$-distribution



Example: Mendel's pea plants (cont.)

Mendel's hypothesis

$$H_0 : p_{ry} = \frac{9}{16}, \quad p_{rg} = \frac{3}{16}, \quad p_{wy} = \frac{3}{16}, \quad p_{wg} = \frac{1}{16}.$$

$$X^2 = \ldots = 0.470$$

This value is to be compared with a $\chi^2(3)$-distribution. The P-value is

$$1 - P(\chi^2_3 \geq 0.470) = 1 - 0.075 = 0.925,$$

which is far from significant. The data are in excellent agreement with Mendel's hypothesis.

Rule of thumb

To calculate the P-value the χ^2-distribution may be used if the expected number of observations for each category is at least 5.

2×2 count tables

With two binary classification criteria the results may be summarized in a 2 by 2 table:

	Response 1	Response 2	Total
Pop 1	y_{11}	y_{12}	$n_1 = y_{11} + y_{12}$
Pop 2	y_{21}	y_{22}	$n_2 = y_{21} + y_{22}$
Total	$c_1 = y_{11} + y_{21}$	$c_2 = y_{12} + y_{22}$	$n = n_1 + n_2 = c_1 + c_2$

Estimates

$$\hat{p}_{11} = \frac{y_{11}}{n_1}, \qquad \hat{p}_{21} = \frac{y_{21}}{n_2}.$$

Question: Is $p_{11} = p_{21}$?

Example: Avadex and mice

	Tumor	No tumor	Total
Avadex	4	12	16
Control	5	74	79
Total	9	86	95

Does the pesticide Avadex increase the occurrence of tumors?

Hypothesis $p_t = p_c$. Estimates are

$$\hat{p}_t = \frac{4}{16} = 0.25 \quad \hat{p}_c = \frac{5}{79} = 0.0633.$$

and a 95% confidence interval for $p_t - p_c$ is

$$(-0.0321, 0.4056)$$

which contains 0.

Test in 2×2 count tables

Expected counts

The expected counts in cell (i,j) in a two-way count table is

$$\text{Expected count} = \frac{(\text{Row total}) \cdot (\text{Column total})}{\text{Grand total}}$$

Test statistic: as with goodness-of-fit test we use the "chi-squared"statistic

$$X^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

and the test statistic follows a $\chi^2(1)$-distribution! [Notice that df$=1$]

Example: Avadex and mice (cont.)

	Tumor	No tumor	Total
Avadex			16
Control		71.52	79
Total	9	86	95

The test statistic is

$$X^2 = \frac{(4-1.52)^2}{1.52} + \frac{(12-14.48)^2}{14.48} + \frac{(5-7.48)^2}{7.48} + \frac{(74-71.52)^2}{71.52}$$
$$= 5.4083$$

Using the χ^2 approximation (with df$=1$) we get the P-value

$$1 - P(\chi^2_1 \geq 5.4083) = 0.020,$$

seemingly in disagreement with the confidence interval. Conclusion?

Two situations — one method

The chi-squared test is used for two situations:

1. Test for homogeneity: Two (or more) independent populations, in which all units are categorized according to one criterion.
2. Test for independence: A single sample in which each unit is classified according to two criteria (rows and columns).

For both situations we may use the chi-squared test as described. It is not always easy to distinguish the two situations.

$r \times k$ count tables

The method applies to an $r \times k$-table as well as a 2×2 table.

	Column 1	Column 2	$\cdots$	Column k	Total
Row 1	y_{11}	y_{12}	$\cdots$	y_{1k}	n_1
Row 2	y_{21}	y_{22}	$\cdots$	y_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Row r	y_{r1}	y_{r2}	$\cdots$	y_{rk}	n_r
Total	c_1	c_2		c_k	n

Test statistic

$$X^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i},$$

where

$$E_{ij} = \frac{(\text{Row total}) \cdot (\text{Column total})}{\text{Grand total}}$$

which is (approximately) χ^2-distributed with

$$\text{df} = (r-1) \cdot (k-1).$$

Example: 1111 cats

1111 cats received at an animal shelter

Other animals	Behavioral problems	
	yes	no
Yes	53	502
No	115	410
No information	17	14

$$X^2 = \frac{(53 - 92.42)^2}{92.42} + \cdots + \frac{(14 - 25.84)^2}{25.84} = 63.1804$$

Two situations — still the same method

The chi-squared test applies to two situations:

1. Test for homogeneity: r independent populations, in which all units are categorized according to one criterion.
2. Test for independence: A single sample in which each unit is classified according to two criteria (rows and columns).

- For each of three populations, behavioral status is determined on a sample.
- For a single population, behavioral status *and* presence of other animals in the home are determined.

One-sided hypotheses

For 2×2-tables the alternative hypothesis may occasionally be one-sided, that is,

$$H_0 : p_{11} = p_{21} \quad \text{vs.} \quad H_A : p_{11} > p_{21}$$

meaning that we regard it as unthinkable that $p_{11} < p_{21}$, whether the hypothesis is true or not.
Test method of for one-sided test:

1. Check whether the estimates are in the direction of alternative hypothesis.
2. If no then there is no indication of disagreement with the null-hypothesis and the p-value is 1.
3. If yes then the p-value is half of the p-value, you would get by a two sided test using the χ^2 distribution.

One-sided hypothesis: diabetes

	Diabetes	No diabetes	Total
Neutered group	26	24	50
Non-neutered group	12	38	50
Total	38	62	100

Null-hypothesis

$$H_0 : p_t = p_c \quad \text{mod} \quad H_A : p_t > p_c.$$

Test statistic:

$$X^2 = 8.3192.$$

What is the conclusion?

UNIVERSITY OF COPENHAGEN FACULTY OF LIFE SCIENCES

Example: Avadex and mice

	Tumor	No tumor	Total
Avadex	4	12	16
Control	5	74	79
Total	9	86	95

Does the pesticide Avadex increase the occurrence of tumors?

Expected values under the hypothesis $H_0 : p_t = p_c$:

	Tumor	No tumor	Total
Avadex	1.52	14.48	16
Control	7.48	71.52	79
Total	9	86	95

UNIVERSITY OF COPENHAGEN FACULTY OF LIFE SCIENCES

Fishert's exact test

Exact test in 2×2-tables. No approximations needed.

Advanced: The idea is that the row- and column-totals tells nothing about the relation. Hence we consider those as given in advance and compute the conditional probability of the test statistic given the marginal totals.

What to do? Compute the p-value as the sum of probabilities of tables that are equal to or more extreme than the observed.

UNIVERSITY OF COPENHAGEN FACULTY OF LIFE SCIENCES

Other tables with same marginal totals

	+	-	Total
Avadex	4	12	16
Control	5	74	79
Total	9	86	95

	+	-	Total
Avadex	2	14	16
Control	7	72	79
Total	9	86	95

	+	-	Total
Avadex	3	13	16
Control	6	73	79
Total	9	86	95

	+	-	Total
Avadex	1	15	16
Control	8	71	79
Total	9	86	95

The probability of a table

$$\Pr(Y=3) = \frac{\binom{16}{3}\binom{79}{6}}{\binom{95}{9}} = 0.1325$$

$$\Pr(Y=2) = \frac{\binom{16}{2}\binom{79}{7}}{\binom{95}{9}} = 0.2960$$

$$\Pr(Y=1) = \frac{\binom{16}{1}\binom{79}{8}}{\binom{95}{9}} = 0.3553$$

$$\Pr(Y=0) = \frac{\binom{16}{0}\binom{79}{9}}{\binom{95}{9}} = 0.1752$$

$$\Pr(Y \leq 4) = 0.0349 + 0.1325 + 0.2960 + 0.3553 + 0.1752 = 0.994$$

$$\Pr(Y \geq 4) = 1 - 0.994 + 0.0349 = 0.041$$

UNIVERSITY OF COPENHAGEN FACULTY OF LIFE SCIENCES

Odds

Odds: a different way to express a probability.

Odds 1:1 for the event A corresponds to $P(A) = 1/2$.

Odds 3:1 for the event A corresponds to $P(A) = 3/4$.

Odds 1:7 for the event A corresponds to $P(A) = 1/8$.

Odds for an event A are defined as

$$\text{Odds}(A) = \frac{P(A)}{1 - P(A)}$$

Notice that while the probability is between 0 and 1,

$$0 \leq \text{Odds}(A) \leq \infty, \qquad 0 \leq P(A) \leq 1$$

UNIVERSITY OF COPENHAGEN FACULTY OF LIFE SCIENCES

Comparison of risks

Three common ways to compare probabilities, for example probability of a disease for two age groups:

- Difference in probability: $p_1 - p_2$
- Relative risk: p_1/p_2
- Odds ratio: $\text{Odds}_1(A)/\text{Odds}_2(A) = \frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2}$.

The hypotheses that the two probabilities are the same ($p_1 = p_2$) may be written

$$\text{probability difference} = 0, \quad \text{RR} = 1, \quad \text{OR} = 1,$$

in terms of probabilities, relative risk and odds ratio, respectively, but the magnitude of the deviation depends on which of the three representations is used.

UNIVERSITY OF COPENHAGEN FACULTY OF LIFE SCIENCES

Comparison of two risks

Example: Risk of finding pesticide residues in strawberries

- before wash: $p_1 = 0.04$
- after wash: $p_2 = 0.01$

Probability difference $= 0.03$, RR $= 0.25$, OR $= 0.242$.

Compare the cases

$(p_1, p_2) =$	(0.04. 0.01)	(0.40, 0.10)	(0.99, 0.96)
Difference, $p_1 - p_2$	0.03	0.30	0.03
RR	4.00	4.00	1.03
OR	4.13	6.00	4.13

Odds-ratio

Odds-ratio is ratio between the odds in two groups:

$$\mathrm{OR} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1 \cdot (1-p_2)}{p_2 \cdot (1-p_1)}$$

The value 1 means that the two odds are the same. Hence, a test for $H_0 : \mathrm{OR} = 1$ is a test of equality of odds (and hence probabilities in the two groups!

Example: Avadex and mice (cont.)

	Tumor	No tumor	Total
Avadex	4	12	16
Control	5	74	79
Total	9	86	95

The estimate of OR is

$$\mathrm{OR} = \frac{4 \cdot 74}{5 \cdot 12} = 4.93, \quad \ln(\mathrm{OR}) = 1.596$$

The standard error of the *logarithm* of OR is

$$\mathrm{SE}(\ln(\mathrm{OR})) = \sqrt{\frac{1}{4} + \frac{1}{12} + \frac{1}{5} + \frac{1}{74}} = 0.74$$

The 95%-confidence interval for ln(OR) then becomes $1.60 \pm 1.96 \cdot 0.74 = (0.15, 3.05)$, which is translated back to OR by the exponential function:

$$\exp(0.15) < \mathrm{OR} < \exp(3.05).$$

Case-control investigation

		Infected	Not infected	Sum
Have eaten the product	yes	10	4	14
	no	6	41	47
	sum	16	45	61

After a number of cases of *Salmonella Manhattan* 16 persons with *Salmonella* infection were ask whether they had eaten sliced smoked fillet from a certain slaughterhouse. A control group was asked the same question.

Case-control and odds-ratio

Row- and column probabilities are not the same ($\frac{10}{14} \neq \frac{10}{16}$), but odds-ratio is the same if rows and columns are switched.

	Infected	Not infected
yes	10	4
no	6	41

$$\mathrm{OR} = \frac{10 \cdot 41}{6 \cdot 4}$$

Thus, the OR for infection among the “yes"group vs. the “no"group is identical to the OR for “yes"among infected vs. non-infected. OR= 17.1 seems to indicate that the product has to do with the risk of infection.

Lecture summary: main points

- Analysis of count tables
- Goodness-of-fit test
- Analysis of two-way count tables
  - 2×2 tables
  - $r \times k$ tables
  - Fishers exact test for 2×2 tables
  - Test of hypotheses with one-sided alternatives
  - Odds and odds-ratio