NTNU
Norwegian University of
Science and Technology

Analysing categorical data
TMA4315 H2017

Mette Langaas

October 9, 2017

Analysis of 2×2 tables

Breast cancer and first childbirth

- Events occuring in the period between menarche and first childbirth is believed to partly be the cause of breast cancer.
- The hypothesis is that the risk of breast cancer increases as the length of this time interval increases.
- In a study breast cancer patients were identified among women in selected hospitals in different countries, and controls were chosen from women of comparable age (also in hospital). Comment: No direct matching.
- Date of first birth were collected and women were divided into two groups; $\leq$ 29 and $\geq$ 30 years at first childbirth. Comment: this threshold is “arbitrarily”.

Breast cancer and first childbirth (cont.)

	Age at first birth		
Status	≥ 30	≤ 29	Total
Case	683	2537	3220
Control	1498	8747	10245
Total	2181	11284	13465

Example (cont.)

	Age at first birth		
Status	≥ 30	≤ 29	Total
Case	683	2537	3220
Control	1498	8747	10245
Total	2181	11284	13465

— What if breast cancer and age at first childbirth are not associated?

— That is, what if the events D and E are independent?

- D = the woman has (had) breast cancer
- E = the age of the woman at first childbirth was ≥ 30

— D and E are independent if and only if $P(D \cap E) = P(D) \cdot P(E)$.

Example (cont.)

- The observed proportions

$$\widehat{P(D)} = \frac{3220}{13465} \text{ and } \widehat{P(E)} = \frac{2181}{13465}$$

- Under independence, $P(D \cap E) = P(D) \cdot P(E)$:

$$P\widehat{(D \cap E)} = \widehat{P(D)} \cdot \widehat{P(E)} = \frac{3220}{13465} \cdot \frac{2181}{13465}$$

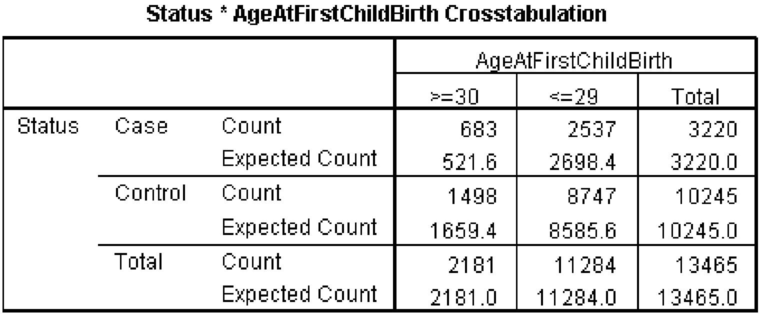
- and the expected number of events of $D \cap E$ is

$$E_{11} = N \cdot P\widehat{(D \cap E)} = 13465 \cdot \frac{3220}{13465} \cdot \frac{2181}{13465} = 521.6$$

- The observed number of events of $D \cap E$ is

$$O_{11} = 683$$

Observed and expected counts



Status * AgeAtFirstChildBirth Crosstabulation

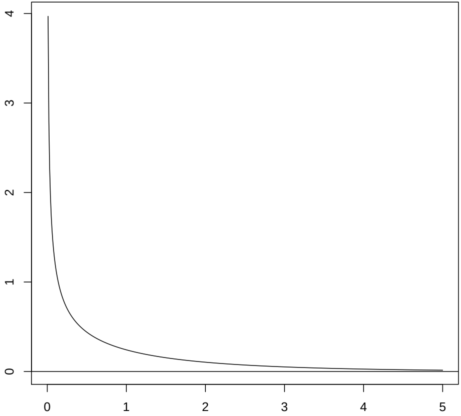
			AgeAtFirstChildBirth		
			>=30	<=29	Total
Status	Case	Count	683	2537	3220
		Expected Count	521.6	2698.4	3220.0
	Control	Count	1498	8747	10245
		Expected Count	1659.4	8585.6	10245.0
	Total	Count	2181	11284	13465
		Expected Count	2181.0	11284.0	13465.0

Pearson’s chi-square test statistic

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\\ &= \frac{(O_{11}-E_{11})^2}{E_{11}}+\frac{(O_{12}-E_{12})^2}{E_{12}}+\frac{(O_{21}-E_{21})^2}{E_{21}}+\frac{(O_{22}-E_{22})^2}{E_{22}}\end{aligned}$$

Under the null hypothesis of independence between columns and rows the Pearson test statistic is approximately χ^2-distributed with 1 degrees of freedom.

The χ_1^2 distribution



Analysing the data in R

Comparing two binomial probabilities

- Hypotheses:

$$H_0 : p_1 = p_2 \text{ vs. } H_1 : p_1 \neq p_2$$

- Estimators for p_1 and p_2:

$$\hat{p}_1 = \frac{X_1}{n_1} \text{ and } \hat{p}_2 = \frac{X_2}{n_2}$$

- When should we reject H_0?

Comparing two binomial probabilities

- When n_1 and n_2 are large, then the binomial distributions can be approximated by normal distributions.
- Under the null hypothesis

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\text{Var}(\hat{p}_1 - \hat{p}_2)}}$$

  is approximately normally distributed.
- The variance of the difference, under independence:

$$\begin{aligned} \text{Var}(\hat{p}_1 - \hat{p}_2) &= \text{Var}(\hat{p}_1) + (-1)^2\text{Var}(\hat{p}_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \\ &= (\frac{1}{n_1} + \frac{1}{n_2})p(1-p) \end{aligned}$$

Comparing two binomial probabilities

— By inserting common estimate for p for the variance

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

— we get an approximately normal test statistic Z

$$Z \approx \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\hat{p}(1-\hat{p})}}$$

— which can be used when $n_1\hat{p}_1(1-\hat{p}_1) \geq 5$ and $n_2\hat{p}_2(1-\hat{p}_2) \geq 5$.

The Z-test vs. Pearson’s chi-square test for 2×2 tables

- It can be shown that Z^2 (the square of the test statistic for comparing two binomial probabilities) is *identical* to the Pearson chi-square test statistic χ^2.
- This means that these two tests are equivalent and will always give exactly the same *p*-value.

Relation between the Pearson's chi-square test and logistic regression

Fit a logistic regression with "case or control" as the response. Then the Pearson's chi-square test statistic equals the score test statistics for testing if the coefficient for the "age at first childbirth" is significant.

This is beyond the scope of this course.

http://www.statsci.org/smyth/pubs/goodness.pdf

Test for Association for $R \times C$ contingency table

```
        H=0 H=1 H=2 H=3 H=4+  rowtots
A=0      27  29  10   8    2  76
A=1      59  53  14  12    4 142
A=2      28  32  14  12    4  90
A=3      19  14   7   4    1  45
A=4+      7   8  10   2    0  27
coltots 140 136  55  38   11 380
```

Expected cell count

- O_{ij} is observed count in cell (i,j) (row,column).
- If the table rows and columns are independent then the expected number of counts in cell (i,j) is

$$\begin{aligned} E_{ij} &= \frac{\text{row sum}}{\text{total sum}} \cdot \frac{\text{column sum}}{\text{total sum}} \cdot \text{total sum} \\ &= \frac{\text{row sum } \times \text{column sum}}{\text{total sum}} \end{aligned}$$

Pearson’s Chi-square test

$$
\begin{aligned}
\chi^2 &= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \\
&= \frac{(O_{11}-E_{11})^2}{E_{11}} + \frac{(O_{12}-E_{12})^2}{E_{12}} + \cdots + \frac{(O_{rc}-E_{rc})^2}{E_{rc}}
\end{aligned}
$$

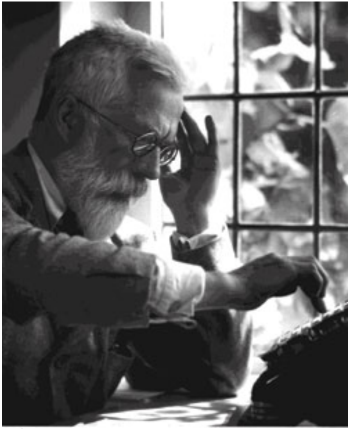
Under the null hypothesis of independence between columns and rows the Pearson test statistic is

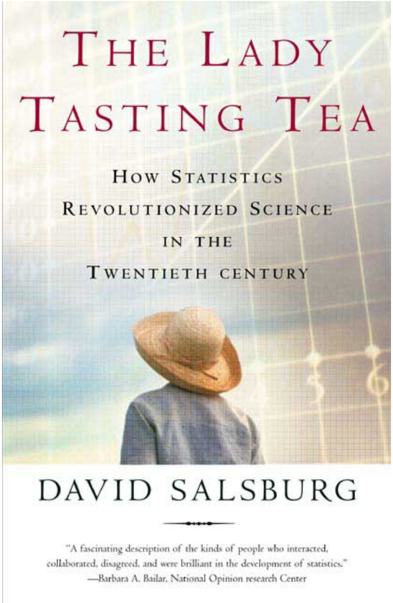
- approximately χ^2-distributed with $(r-1)\cdot(c-1)$ degrees of freedom
- given that
  - no more than 20% of the cells have $E_{ij} < 5$ and
  - no cell has expected count less than 1, $E_{ij} < 1$.

Analysing the data in R

Fisher’s exact test

- the solution to "what to do when the asymptotic requirements for using the Pearson chisquare test are not satisfied".





THE LADY TASTING TEA

HOW STATISTICS REVOLUTIONIZED SCIENCE IN THE TWENTIETH CENTURY

DAVID SALSBURG

"A fascinating description of the kinds of people who interacted, collaborated, disagreed, and were brilliant in the development of statistics."
—Barbara A. Bailar, National Opinion research Center

The lady tasting tea

Events:

- Lady says tea first.
- Truth is tea first.

And, the lady *knows* that 4 cups has milk first and 4 cups has tea first.

		Lady says tea first		
		Y	N	total
Truth is tea first	Y			4
	N			4
	total	4	4	8

The lady tasting tea

Fill in observed and expected values when we assume that the Lady identified 3 out of the 4 tea first infusions correctly, and 3 out of the 4 milk first infusions correctly.

		Lady says tea first		
		Y	N	total
Truth is tea first	Y			4
	N			4
	total	4	4	8

Can the Pearson chi-square contingency table method be used?

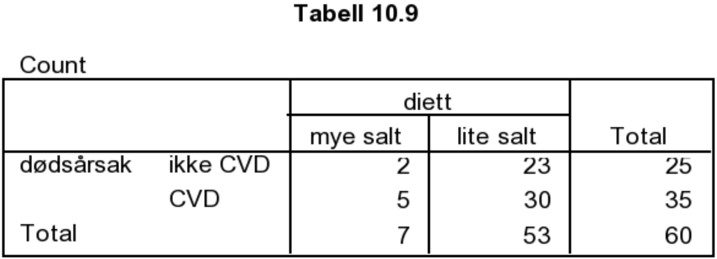
Fisher’s exact test

- In the tea-tasting example the margins are fixed.
- In general this is not the case.
- But, we may condition on the margins (i.e. assuming they are fixed).
- We will then end up with a conservative test (low power).

Quoting Rosner:
“For mathematical convenience, we shall assume that the margins (row sums and column sums) of this table are fixed”.

Cardiovascular Disease and Nutrition

- Investigate possible association between high salt intake and death from cardiovascular Disease (CVD).
- Retrospective study of males aged 50-54 in a specific country, who died within a chosen month.
- Include approximately the same number of cases and controls.



Tabell 10.9

Count

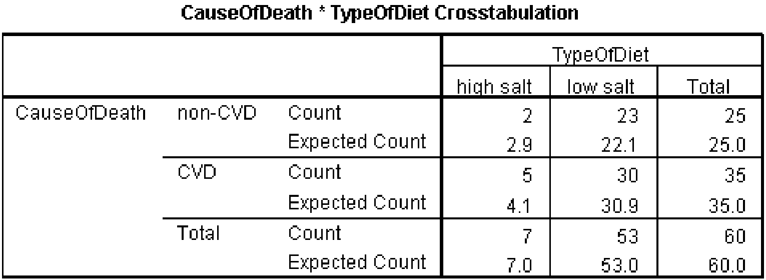
		diett		Total
		mye salt	lite salt	
dødsårsak	ikke CVD	2	23	25
	CVD	5	30	35
Total		7	53	60

CVD and salt, cont.

	Type of diet		
Cause of death	High salt	Low salt	Total
Non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

H_0 : P(Type of diet=high salt|Death by CVD)=
P(Type of diet=high salt|Death by non-CVD)?

Crosstabulation: observed and expected



CauseOfDeath * TypeOfDiet Crosstabulation

			TypeOfDiet		
			high salt	low salt	Total
CauseOfDeath	non-CVD	Count	2	23	25
		Expected Count	2.9	22.1	25.0
	CVD	Count	5	30	35
		Expected Count	4.1	30.9	35.0
	Total	Count	7	53	60
		Expected Count	7.0	53.0	60.0

More than one cell has expected count <5, thus it is not valid to use the Pearson chi-square test.

Fisher’s exact test

Quoting Rosner:

“For mathematical convenience, we shall assume that the margins (row sums and column sums) of this table are fixed”.

- If the cause of death and the diet are not associated, then we may think of this as if we can randomly distribute the diets among the cases and controls, but keeping the marginals fixed.
- Look at the $n = 60$ deaths, and draw randomly $(a + c) = 7$ of these, and assume that these are the ones with high salt intake.
- Then, let X be the number of those $(a + c)$ that did not die from CVD.
- X then follows the hypergeometric distribution.

Hypergeometric distribution

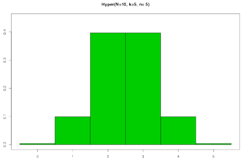
a	b	a+b
c	d	c+d
a+c	b+d	n=a+b+c+d

$$P(a) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)! \cdot (c+d)! \cdot (a+c)! \cdot (b+d)!}{n! \cdot a! \cdot b! \cdot c! \cdot d!}$$

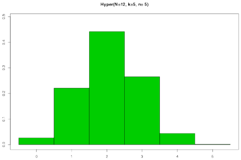
for all non-negative integers *a*, that gives non-negative integers in all cells of the table.
Remember: $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ and $0! = 1$.

Hypergeometric distribution

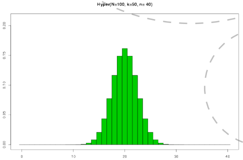
$N = 10, k = 5, n = 5$



$N = 12, k = 5, n = 5$



$N = 100, k = 50, n = 40$



Expected value and variance:

$$\mu = E(X) = \frac{nk}{N} \quad \text{og} \quad \sigma^2 = \text{Var}(X) = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N}(1 - \frac{k}{N})$$

Urn with balls

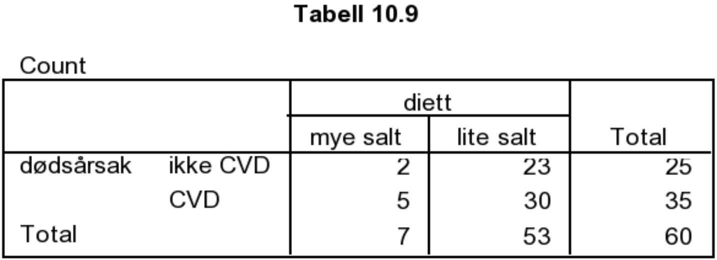
- Definition:
  - N=number of balls.
  - k=number of red balls.
- Procedure: do n times:
  - draw a ball at random
  - make a note of the colour
  - but the ball aside.
- Then the number of red balls follows a hypergeometric distribution.

Urn with balls

- Definition: $p = \frac{\text{number of red balls}}{\text{number of balls}}$
- Procedure: do n times:
  - draw a ball at random
  - make a note of the colour
  - put the ball back into the urn.
- Then the number of red balls follows a binomial distribution.

CVD and salt (cont.)

What is the probability of the observed table under the null hypothesis of no association between CVD and type of diet?



Tabell 10.9

Count

		diett		Total
		mye salt	lite salt	
dødsårsak	ikke CVD	2	23	25
	CVD	5	30	35
Total		7	53	60

$$P(a=2)=\frac{25!35!7!43!}{60!2!23!5!30!}=0.252$$

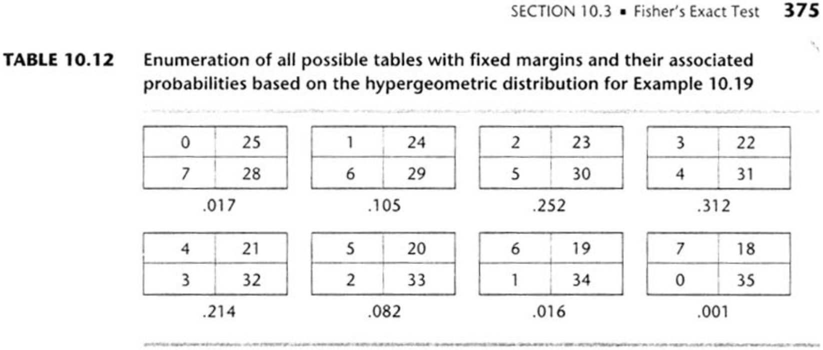
CVD and salt: *p*-value

- The *p*-value denotes the probability of what is observed or "something more extreme" given that the null hypothesis is true.
- We have calculated the probability of what we have observed,
- but what is more extreme?

Enumeration

- We need to *enumerate* all possible tables with the same margins (row and column) as the observed table.
- Then calculate the probability for each possible table.
- And, finally, sum the probability for the tables that are “more extreme” to get the *p*-value.

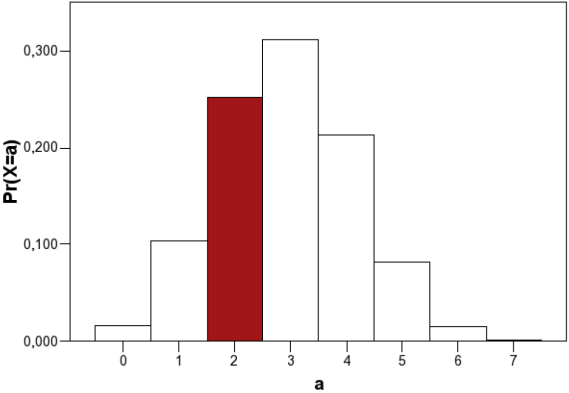
CVD and salt: enumeration



SECTION 10.3 ▪ Fisher's Exact Test 375

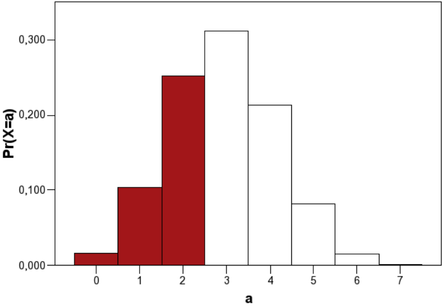
TABLE 10.12 Enumeration of all possible tables with fixed margins and their associated probabilities based on the hypergeometric distribution for Example 10.19

0	25	1	24	2	23	3	22
7	28	6	29	5	30	4	31
.017		.105		.252		.312	
4	21	5	20	6	19	7	18
3	32	2	33	1	34	0	35
.214		.082		.016		.001	



What we have observed; $a = 2$. What is more extreme?

Left-sided

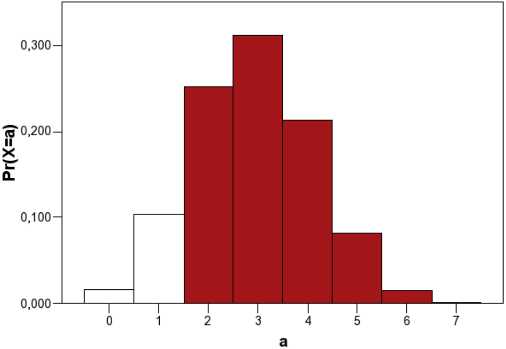


- p_1=P(Type of diet=high salt|non-CVD death)
- p_2=P(Type of diet=high salt|CVD death)

$$H_0 : p_1 = p_2 \text{ vs. } p_1 < p2$$

$$P_{left} = P(a \leq 2) = 0.375$$

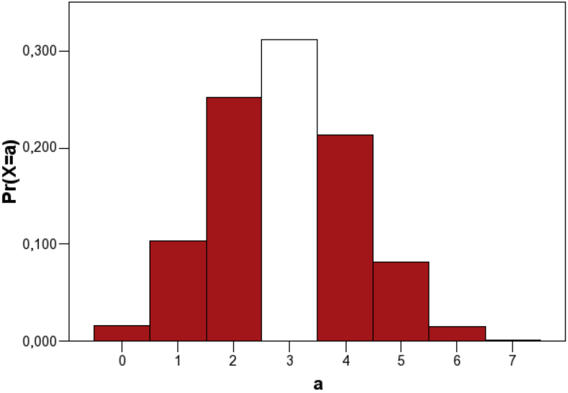
Right-sided



$$H_0 : p_1 = p_2 \text{ vs. } p_1 > p2$$

$$P_{right} = P(a \geq 2) = 0.878$$

Two-sided



$$\sum_{i:P(i)\leq P(2)} P(X=i) = P(X\leq 2) + P(X\geq 4) = 0.688$$

CVD and salt analyses in R

Lady tasting tea in R

Summing up

$R \times C$ tables: test for independence of row and column events.

- — Pearson's chi-square test with $(r-1) \times (c-1)$ degrees of freedom.
- — or Fisher's exact test.
- — If "test for homogeniety" - same methods can be used.
- — Other methods when data are matched or want to test for trend instead of independence.