

Counts and Proportions

Anton Rask Lundborg
arl@math.ku.dk

Copenhagen Causality Lab
Department of Mathematical Sciences

November 29, 2023

Plan for lecture

- Summary of Day 1.
 - Recap of t -tests.
 - Solution to Exercise 1.2 and 1.5.
- Non-parametric tests:
 - Wilcoxon / Mann-Whitney, Kruskal-Wallis.
- Discussion of papers.
 - Stern & Smith: “Sifting the evidence — What’s wrong with significance tests?”
 - Gelman & Carlin: “Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors”
- Analysis of 2-way tables.
 - More assumptions \implies more power.
 - Interpretations and associated tests.

Summary of Day 1

- Statistics answers four important questions:
 - ① Is there an effect? (falsification of null hypotheses via p -values)
 - ② Where is the effect? (p -values from post hoc analyses)
 - ③ What is the effect? (confidence and prediction intervals)
 - ④ Can the conclusions be trusted? (model validation)
- We do model-based frequentist statistics: Interpretation of p -values and confidence intervals via the meta-experiment.
- Tidy data: Datasets consists of **variables** (columns) and **observations** (rows).
- t -tests, data transformation, and validation of normality assumption.

Data example 4 from Day 1: Phosphorus in lakes

Two independent samples (not necessarily of the same length)

```
> lakes
# A tibble: 627 x 2
  location      phosphorus
  <chr>         <dbl>
1 East-Denmark  255
2 East-Denmark  102.
3 East-Denmark  166.
4 East-Denmark   42.5
5 East-Denmark  102.
6 East-Denmark   60.6
7 East-Denmark   89.8
8 East-Denmark  182.
9 East-Denmark  243.
10 East-Denmark   30.9
# ... with 617 more rows
```

R code (here log-transformation needed for normality of data)

```
t.test(log(phosphor) ~ location, data = lakes)
```

Solution to Exercise 1.2

- Example 1: Growth of rats, $N = 12$.

Variable	Type	Range	Usage
antibiotic	Nominal	0, 40	fixed effect
vitamin	Nominal	0, 5	fixed effect
growth	Continuous	[1.00 ; 1.56]	response

- Example 2 (for paired analysis): Tenderness of meat, $N = 24$.

Variable	Type	Range	Usage
pH.group	Nominal	low, high	fixed effect
Tunnel	Continuous	[3.11 ; 8.78]	response
Fast	Continuous	[3.33 ; 8.44]	response

- Example 2 (for mixed model): Tenderness of meat, $N = 48$.

Variable	Type	Range	Usage
Pig	Nominal	24 levels	random effect
pH.group	Nominal	low, high	fixed effect
method	Nominal	tunnel, fast	fixed effect
tenderness	Continuous	[3.11 ; 8.44]	response

Solution to Exercise 1.2 (continued)

- Example 3: Weight loss (raw binary data), $N = 160$.

Variable	Type	Range	Usage
diet	Nominal	2 levels	fixed effect
week	Ordinal	$1 < \dots < 8$	fixed effect
person	Nominal	20 levels	random effect
weight.loss	Binary	no, yes	response

- Example 3: Weight loss (for binomial analysis), which is also available from the table, $N = 20$.

Variable	Type	Range	Usage
diet	Nominal	2 levels	fixed effect
weeks.with.weight.loss	Count	$0 < \dots < 8$	response

Solution to Exercise 1.2 (continued)

- Example 4 (with univariate end-point): Stress and metabolism, $N = 8 \cdot 96 = 768$.

Variable	Type	Range	Usage
number.of.rats	Continuous	5,6	weight
group	Nominal	8 levels	random effect
sex	Nominal	male, female	fixed effect
stabling	Nominal	no, yes	fixed effect
food.additive	Nominal	no, yes	fixed effect
gene	Nominal	96 levels	fixed effect
expression	Continuous	[-5.7060 ; 13.2240]	response

Solution to Exercise 1.5 (slide 1 of 4)

- Open data frame hypertension from the R datafile hypertension.RData to see the variables:

change1 = change of blood pressure over study period 1
change2 = change of blood pressure over study period 2
average = (change1+change2)/2
diff = change1 - change2
E_diff_N = changeE - changeN

- Four tests:

Two-sample test: E_diff_N in E/N-group vs. N/E-group
Two-sample test: average in E/N-group vs. N/E-group
Two-sample test: diff in E/N-group vs. N/E-group
One-sample test: E_diff_N against 0

- Question: What do the four hypotheses mean?

Solution of Exercise 1.5 (slide 2 of 4)

- Time for some algebra!

e_1 = effect of drug E in period 1

e_2 = effect of drug E in period 2

n_1 = effect of drug N in period 1

n_2 = effect of drug N in period 2

- E/N patients experience effects (e_1, n_2)

- N/E patients experience effects (n_1, e_2)

- Two first null hypotheses stipulate the following:

Test 1: $e_1 - n_2 = e_2 - n_1$ (No spill-over: $e_1 + n_1 = e_2 + n_2$)

Test 2: $e_1 + n_2 = e_2 + n_1$ (No interaction: $e_1 - n_1 = e_2 - n_2$)

- If these hypotheses are not rejected, then $e_1 = e_2 = e$ and $n_1 = n_2 = n$

- The two last null hypotheses stipulate:

Test 3: $e - n = n - e$ (No difference: $e = n$)

Test 4: $e - n = 0$ (No difference: $e = n$)

Solution to Exercise 1.5 (slide 3 of 4): R code

t-test 1: example of a two-sample test

```
ggplot(hypertension, aes(sample = E_diff_N)) + geom_qq() + facet_grid(. ~ order)
t.test(E_diff_N ~ order, data = hypertension)
```

t-test 4: example of a one-sample test

```
ggplot(hypertension, aes(sample = E_diff_N)) + geom_qq()
t.test(hypertension$E_diff_N)
```

Solution to Exercise 1.5 (4/4): Conclusion from analysis

Test number: null hypothesis	Statistical test	Assumptions	p-value
1: No spill-over	Welch t -test	Normality ok?	0.2011
2: No interaction	Welch t -test	Normality questionable	0.5136
3: No drug difference	Welch t -test	Normality questionable	0.0932
4: No drug difference	t -test	Normality ok!	0.1108

- Neither spill-over nor interaction.
- However, drug effect is also non-significant.
- Here, the two-sample test is more powerful than the one-sample test (0.0932 vs. 0.1108). But the one-sample test can be easier to interpret and is thus often preferred (e.g. confidence interval for drug difference).

All results are contingent on the (sometimes dubious) normality assumption. **Is this avoidable?**

Checkpoint

- Questions?
- After the break we discuss **non-parametric** tests that may be used instead of t -tests and 1-way ANOVA when the normality assumption is not satisfied:
 - Wilcoxon's signed rank test (one sample).
 - Wilcoxon's rank sum test (two independent samples). Also known as the Mann-Whitney U test.
 - Kruskal-Wallis test (1-way design with more than 2 groups).

Time for a break!

Nonparametric hypothesis testing

Data example: Density of nerve cells – Motivation for non-parametric tests

Density of nerve cells measured at two sites of the intestine, midregion and mesenteric region, for $n = 9$ horses.

horse	mid	mes	diff
1	50.6	38.0	12.6
2	39.2	18.6	20.6
3	35.2	23.2	12.0
4	17.0	19.0	-2.0
5	11.2	6.6	4.6
6	14.2	16.4	-2.2
7	24.2	14.4	9.8
8	37.4	37.6	-0.2
9	35.2	24.4	10.8

Densities of nerve cells are significantly different at the two sites:

$$p = 2 \cdot \mathbb{P}\left(t_{df=8} > \frac{7.33 - 0}{2.60}\right) = 0.0222$$

This conclusion, however, **relies on the assumption** that diff is normally distributed.

But what if this assumption is wrong?

Usually, we check if $\log\left(\frac{\text{mid}}{\text{mes}}\right)$ is normal, or use a **non-parametric** test.

Non-parametric methods – Tests that do not assume normality

Pro and Cons:

- + Often much fewer (or no) assumptions to check.
- No model, no estimates, no confidence intervals \implies harder to interpret.
- Only available for some situations.

Today:

- One sample: Sign test (appealing, but very weak and never used), Wilcoxon signed rank test (preferable).
- Two samples: Wilcoxon rank sum test (Mann-Whitney).
- 1-way ANOVA: Kruskal-Wallis test.

Sign test for the density of nerve cells example

Recall that the paired two-sample t -test gave $p = 0.0222$.

horse	mid	mes	diff	sign
1	50.6	38.0	12.6	+
2	39.2	18.6	20.6	+
3	35.2	23.2	12.0	+
4	17.0	19.0	-2.0	-
5	11.2	6.6	4.6	+
6	14.2	16.4	-2.2	-
7	24.2	14.4	9.8	+
8	37.4	37.6	-0.2	-
9	35.2	24.4	10.8	+

- Test statistic $r = \text{number of positive signs} = 6$
- If H_0 (+/- are exchangeable) is true, then $r \sim \text{bin}(n, 0.5)$:

$$p = 2 \cdot \mathbb{P}(\text{bin}(9, 0.5) \geq 6) = 0.5078$$

- Absolutely no evidence against H_0 .

Wilcoxon's signed rank test for Data example 2

Recall that the paired two-sample t -test gave $p = 0.0222$.

horse	mid	mes	diff	abs_diff	sign	rank
1	50.6	38.0	12.6	12.6	+	8
2	39.2	18.6	20.6	20.6	+	9
3	35.2	23.2	12.0	12.0	+	7
4	17.0	19.0	-2.0	2.0	-	2
5	11.2	6.6	4.6	4.6	+	4
6	14.2	16.4	-2.2	2.2	-	3
7	24.2	14.4	9.8	9.8	+	5
8	37.4	37.6	-0.2	0.2	-	1
9	35.2	24.4	10.8	10.8	+	6

- Test statistic $S_+ = \text{sum of ranks of positive diff.'s}$
 $= 8 + 9 + 7 + 4 + 5 + 6 = 39$ (out of total=45)
- If H_0 (distribution of diffs is symmetric) is true, then

$$p = 2 \cdot \mathbb{P}(S_+ \geq 39) = 0.05469$$

- Almost significant at 5%. Some evidence of larger values at "mid".

R code available in script: `cell_density.R`

Try to execute the script on your laptop and read off the results from the previous slides

```
# Insert data into two vectors
mid <- c(50.6, 39.2, 35.2, 17.0, 11.2, 14.2, 24.2, 37.4, 35.2)
mes <- c(38.0, 18.6, 23.2, 19.0, 6.6, 16.4, 14.4, 37.6, 24.2)

# Check normality and perform t-test
qqnorm(mid - mes)
shapiro.test(mid - mes)
t.test(mid, mes, paired = TRUE)

# Perform sign test
binom.test(sum(mid > mes), length(mid))

# Perform Wilcoxon rank sum test
wilcox.test(mid, mes, paired = TRUE)
```

Non-parametric two-sample tests – two groups of observations

- Null hypothesis: Both groups have the same continuous distribution.
- The **Wilcoxon rank sum test**, also known as **the Mann-Whitney U test**, is based on the ranks R_j (among all observations) via the test statistic

$$S = \sum_{\text{observations } j \text{ from the first group}} R_j$$

- To compute p -values we need the distribution of S under the null hypothesis. On slide 20 we discuss how this may be done.
- There exist many other non-parametric tests. Doing several tests and choosing the test with the minimal p -value is bad research practice (amounts to cheating)!
 - Quiz: Can you explain why?

p -values: Exact / Simulated exact / Approximate

- In some situations, e.g. two-sample t -test with equal variances, the distribution of the test statistics is known mathematically. In these situations we may compute the exact p -value.
- Simulated p -values:
 - 1 Simulate 10000 datasets, say, assuming the null hypothesis.
 - 2 Compute the test statistic for each simulated dataset.
 - 3 Is the observed test statistic extreme among the 10000 simulated test statistics?
 - 4 Leads to estimate and confidence interval for the exact p -value.

On Day 1 we tried this for the milk yield example.

- Use an approximation of the test statistic distribution, e.g.

$$p = 2 \cdot \mathbb{P}(Z \geq |z_{\text{obs}}|), \quad Z = \frac{S - \text{mean}(S)}{\sqrt{\text{var}(S)}} \underbrace{\sim \mathcal{N}(0, 1)}_{\text{approximately}}$$

but several other approximations (Satterthwaite, Kenward-Roger, ...) appear in various situations.

Non-parametric equivalent of 1-way ANOVA – More than 2 groups

- Null hypothesis: All groups have the same continuous distribution.
- Test statistic constructed via the ranks of the observations.
- Approximate p -value via χ^2 -distribution.
- The equivalent (allowing or more than 2 groups) of the **Wilcoxon** test is also known as **Kruskal-Wallis** test.

Remarks:

- ① A falsification of the null hypothesis only tell us that the groups do not have the same continuous distribution.
- ② In principle none of the non-parametric methods discussed today tolerate **ties**, e.g. that two observations are identical. However, there exists remedies for this.

Nonparametric tests in Exercise 1.5: R code

Test 1: Example of a two-sample Wilcoxon (Mann-Whitney) test

```
# Wilcoxon Rank Sum test  
wilcox.test(E_diff_N~order,data=hypertension)
```

Test 4: Example of a one-sample Wilcoxon test

```
# Wilcoxon Signed Rank test  
wilcox.test(hypertension$E_diff_N)
```

Standard practice: only use Wilcoxon tests if the t -tests are not valid.

Test comparisons in Exercise 1.5

Test number: null hypothesis	Statistical test	Assumptions	p-value
1: No spill-over	Welch t -test Wilcoxon	Normality ok? None	0.2011 0.1067
2: No interaction	Welch t -test Wilcoxon	Normality questionable None	0.5136 0.71
3: No drug difference	Welch t -test Wilcoxon	Normality questionable None	0.0932 0.09022
4: No drug difference	t -test Wilcoxon	Normality ok! None	0.1108 0.1327

- Broadly similar results with both tests.
- The pragmatic statistician will consider this as evidence that we are doing something reasonable!

Checkpoint

- Questions?
- After the break we discuss the two papers. These papers address problems that may arise in experiments with low statistical power:
 - Is there an effect? (Sterne & Smith)
 - What is the effect? (Gelman & Carlin)
- We then begin the last main topic for today: the analysis of contingency tables.

Time for a break!

Hypothesis testing in practice

Discussion of Gelman & Carlin 2024 and Sterne & Smith 2001

Type S (sign) and Type M (magnitude) errors

Gelman & Carlin, Persp. Psych. Science, 1–11, 2014.

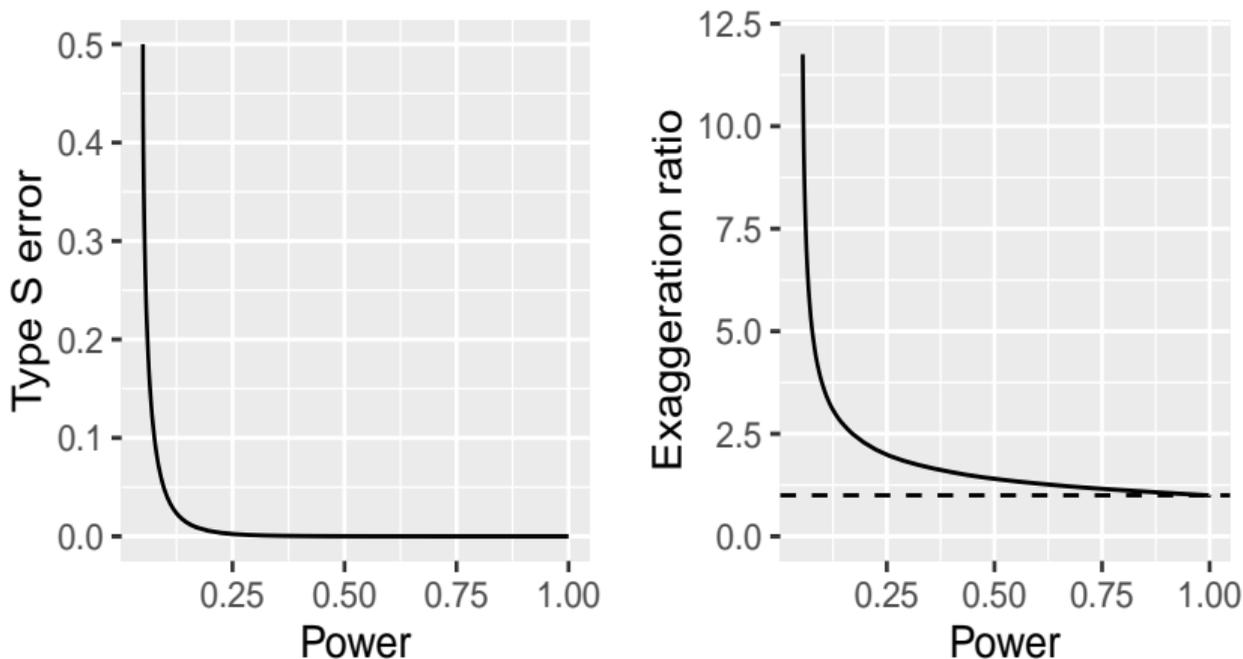
- Sterne & Smith discuss risk of **false positives** (Is there an effect?), i.e. the proportion of true null hypothesis among rejected tests.

		Null hypothesis:	
		True	False
Test:	Accept	Correct	Type II error
	Reject	False positive	Possibly Type S/M error

- Gelman & Carlin discuss risk of **wrong conclusions** (What is the effect?) given correctly rejected (and false) null hypothesis!
 - **Type S error** = the estimate (from a rejected false null hypothesis) has the wrong sign.
 - **Type M error** = the estimate (from a rejected false null hypothesis) has a too large magnitude.
- Risk of **false positives** and of **Type S** and **Type M** errors is only problematic when the **power** of the test is low.

Risk of Type S and Type M error

Figure 2, Gelman & Carlin, 2014: Effect size from 0 to 100, SE=1, $\alpha = 0.05$, $df=\infty$



Remark: Exaggeration ratio approaches ∞ as effects size approaches 0.

What's wrong with significance tests?

Table 2, Sterne & Smith, BMJ, 226–231, 2001

	Null hypothesis:	
	True	False
Test: Do not reject (accept)	Correct	Type II error
Reject	Type I error (size)	Correct (power)

- Specificity = $\mathbb{P}_{H_0}(\text{accept}) = 1 - \text{size}$
- Sensitivity = $\mathbb{P}_{H_1}(\text{reject}) = \text{power}$
- The p -value is **not** the probability that the hypothesis is true (this is not random!). As a “counterexample” consider the following 1000 tests ($\frac{45}{95} > 0.05$):

Result of experiment	Null hypothesis true (no association)	Null hypothesis false (association!)	Total
Accept H_0	855	50	905
Reject H_0	45	50	95
Total	900	100	1000

Percentage of significant results that are false positives

Table 3, Sterne & Smith, BMJ, 226–231, 2001

Ideas correct	Power	Size		
		$\alpha = 5\%$	$\alpha = 1\%$	$\alpha = 0.1\%$
80%	20%	5.9	1.2	0.10
	50%	2.4	0.5	0.05
	80%	1.5	0.3	0.03
50%	20%	20.0	4.8	0.50
	50%	9.1	2.0	0.20
	80%	5.9	1.2	0.10
10%	20%	69.2	31.0	4.30
	50%	47.4	15.3	1.80
	80%	36.0	10.1	1.10
1%	20%	96.1	83.2	33.10
	50%	90.8	66.4	16.50
	80%	86.1	55.3	11.00

Analysis of contingency tables

Statistics and tests for 2-way contingency tables

- Overview:

Number of categories		See slide
Variable 1	Variable 2	
2	2	34–39
2+ (nominal)	2+ (nominal)	42
2+ (nominal)	2+ (ordinal)	43
2+ (ordinal)	2+ (ordinal)	44–45
2 (paired)	2 (paired)	48–52
2+ (paired)	2+ (paired)	53–54

- We will focus on the statistical principles of validity and power.
- We will discuss data representations and R functions.
- Multi-way tables are analyzed via graphical models which we do not cover here.

How to describe these methods in publications?

- The methods discussed today are so simple/standard that no description should be included in the **Methods** section.
 - In particular, these models do not require validation (although significance testing in these models is approximate!).
 - The simulated χ^2 -test (which we will discuss later) is not yet standard, so you perhaps need to make a comment if you use that test.

- In the **Results** section you should present tables with both raw observations and model estimates. This should be summarized with a statement like (cf. chicken activity example):

Spearman's rank correlation shows strong evidence of treatment effect on gait score ($\hat{r}_s = -0.1968$, $p = 0.00066$).

Table-of-Variables: Overview of data examples

- Data example 1: Avadex and cancer

Variable	Type	Range	Usage
Avadex	nominal	no, yes	fixed effect
Tumor	nominal	no, yes	response

- Data example 2: Activity of chicken

Variable	Type	Range	Usage
Treatment	ordinal	$A < B < C < D$	fixed effect
Gait.score	ordinal	$0 < 1 < 2 < 3.5$	response

- Data example 3: Marijuana and sleeping problems

Variable	Type	Range	Usage
Marijuana	nominal	no, yes	response
Control	nominal	no, yes	response

Data example 1: Fungicide Avadex in mice – comparing 2 proportions

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

$n_1 = 16$ mice got Avadex in their food
 $n_2 = 79$ mice got no Avadex (controls)

Note that there is a clear interpretation of **treatment** and **response** for the two variables.

Presumably, most of you have learned (in your basic statistics course) about the three tests listed below. What are the differences, and which test should we use for the example above?

- χ^2 -test for independence (between Tumor and Avadex).
- χ^2 -test for homogeneity (of Tumor risks across the Avadex groups).
- Fishers exact test.

Chi-squared test (2×2 tables)

Null hypothesis: Independence or homogeneity of proportions.

	Var. 2		Total
Var. 1	a	b	a+b
	c	d	c+d
Total	a+c	b+d	n

Rule of thumb: Valid when

80% of cells: expected ≥ 5

all cells: expected ≥ 1

For 2×2 tables: all expected ≥ 5 .

- **Chi-squared test statistic:**

$$\chi^2 = \sum_{\text{cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

- **Yates' continuity correction (classically only for 2×2 tables):**

$$\chi_{\text{Yates}}^2 = \sum_{\text{cells}} \frac{(|\text{observed} - \text{expected}| - \frac{1}{2})^2}{\text{expected}}$$

Independence test using R

If you do not want the (recommended) continuity correction use option:
`correct=FALSE`

```
> chisq.test(matrix(c(4, 5, 12, 74), 2, 2))
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: matrix(c(4, 5, 12, 74), 2, 2)
X-squared = 3.4503, df = 1, p-value = 0.06324
```

Warning message:

```
In chisq.test(matrix(c(4, 5, 12, 74), 2, 2)) :
  Chi-squared approximation may be incorrect
```

Homogeneity test in 2×2 -table using R

If you do not want the (recommended) continuity correction use option:
`correct=FALSE`

```
> prop.test(c(4, 5), n = c(16, 79))
```

2-sample test for equality of proportions with continuity correction

```
data:  c(4, 5) out of c(16, 79)
```

```
X-squared = 3.4503, df = 1, p-value = 0.06324
```

```
alternative hypothesis: two.sided
```

```
95 percent confidence interval:
```

```
-0.06973073  0.44314845
```

```
sample estimates:
```

```
prop 1    prop 2
```

```
0.25000000 0.06329114
```

```
Warning: In prop.test(c(4, 5), n = c(16, 79)) :
```

```
Chi-squared approximation may be incorrect
```

What if the Chi-squared test is invalid?

Classically, Fisher's Exact Test is recommended

```
fisher.test(matrix(c(4, 5, 12, 74), 2, 2))
```

- However, Fisher's test has a slightly different **interpretation** of the null hypothesis!
- Alternative approach, which is always valid: keep the X^2 -statistic and **simulate** the p -value. This gives different tests for the hypotheses of **independence of variables** and **homogeneity of proportions**:
 - Testing for independence between tumors and avadex requires conditioning on the **total sum**.
 - Comparing proportions of tumors requires conditioning on the **row marginals**.
 - Quiz: What is correct in the present example?

Implemented in LabApplStat-package

```
chisq.test.simulate(matrix(c(4, 5, 12, 74), 2, 2),"row")
```

R function for power computations: `power.chisq.test.simulate()`

Avadex: Presentation of results – see `avadex.R`

		Tumor		Total
		+	-	
Avadex	+	4 (25%)	12 (75%)	16 (100%)
	-	5 (6%)	74 (94%)	79 (100%)
Total		9 (9%)	86 (91%)	95 (100%)

Chi-squared: $\chi^2 = 5.4083$, $p = 0.0200$ (*, validity questionable)
Yates: $\chi^2_{\text{Yates}} = 3.4503$, $p = 0.0632$ (NS, validity questionable)
Fisher: $p = 0.0411$ (*, “wrong” null hypothesis)
Simulated Chi-squared: $\chi^2 = 5.4083$, $p = 0.0224$ (*, valid)

- Only the simulated Chi-squared is needed, and we conclude that Avadex increases risk of tumors ($p = 0.02$).
- Estimated tumor probabilities:

Avadex	$\mathbb{P}(\text{Tumor} = +)$	95% confidence interval
+	0.250	(0.083 ; 0.526)
-	0.063	(0.024 ; 0.148)

Checkpoint

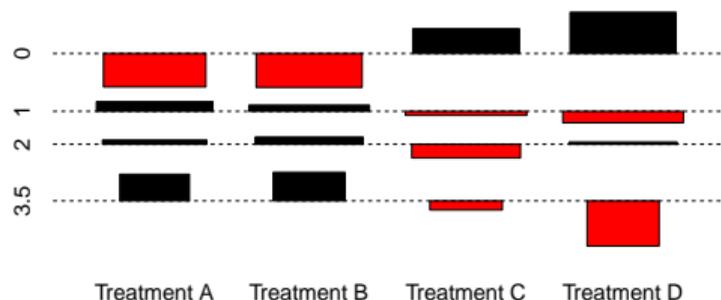
- Questions?
- After the break we continue discussing tests for **2-way contingency tables**.
- In particular, we will discuss how using **ordinal** instead of **nominal** structure \implies more power!

Time for a break!

Data example 2: Activity of chicken

Treatment	GAIT-score: 0	1	2	3.5	Total
A (ad libitum feeding)	12	26	20	12	70
B (fasting 8 H/day, 1 week)	13	27	22	13	75
C (fasting 8 H/day, 2 weeks)	25	25	18	8	76
D (fasting 8 H/day, 3 weeks)	28	23	21	3	75
Total	78	101	81	36	296

Graphical display using R: `assocplot()`



Unordered categories ($r \times k$ table)

Null hypothesis: No association between row and column variables

- **Chi-squared test statistic:**

$$\chi^2 = \sum_{\text{cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}, \quad \text{expected} = \frac{\text{row total} \cdot \text{column total}}{\text{grand total}}$$

Under the null hypothesis, χ^2 approximately (for large sample sizes) follows a χ^2 -distribution with $df = (r - 1) \cdot (k - 1)$.

Chicken example: $\chi_{\text{obs}}^2 = 17.68$, $p\text{-value} = P(\chi_{df=9}^2 \geq \chi_{\text{obs}}^2) = 0.04$ (*)

Treatment	GAIT-score				Total
	0	1	2	3.5	
A	12	26	20	12	70
B	13	27	22	13	75
C	25	25	18	8	76
D	28	23	21	3	75
Total	78	101	81	36	296

Can we do better than this, i.e.

- Get more power?
- Get more precise statement of the effect?

One ordered category ($r \times k$ table)

Compare rank sums of treatment groups using **Kruskal-Wallis** (an exact test so also works with small sample sizes!)

- **Chicken example:** GAIT-score is an ordinal variable.
- Ranks for GAIT-score with resolved ties:

A:	$\underbrace{39.5, \dots, 39.5}_{12}$	$\underbrace{129, \dots, 129}_{26}$	$\underbrace{220, \dots, 220}_{20}$	$\underbrace{278.5, \dots, 278.5}_{12}$
B:	$\underbrace{39.5, \dots, 39.5}_{13}$	$\underbrace{129, \dots, 129}_{27}$	$\underbrace{220, \dots, 220}_{22}$	$\underbrace{278.5, \dots, 278.5}_{11}$
C:	$\underbrace{39.5, \dots, 39.5}_{25}$	$\underbrace{129, \dots, 129}_{25}$	$\underbrace{220, \dots, 220}_{18}$	$\underbrace{278.5, \dots, 278.5}_{8}$
D:	$\underbrace{39.5, \dots, 39.5}_{28}$	$\underbrace{129, \dots, 129}_{23}$	$\underbrace{220, \dots, 220}_{21}$	$\underbrace{278.5, \dots, 278.5}_{3}$

- Test statistic = 13.0272.
- p -value = 0.00457 (**)

Two ordered categories ($r \times k$ table)

- Arguably, treatment is also an ordinal variable:
(A: ad libitum feeding, B: fasting 8 H/day, 1 week, C: 2 weeks, D: 3 weeks)
- Separate ranks for both variables (also resolving ties):

35.5:	$\underbrace{39.5, \dots, 39.5}_{12}$	$\underbrace{129, \dots, 129}_{26}$	$\underbrace{220, \dots, 220}_{20}$	$\underbrace{278.5, \dots, 278.5}_{12}$
108:	$\underbrace{39.5, \dots, 39.5}_{13}$	$\underbrace{129, \dots, 129}_{27}$	$\underbrace{220, \dots, 220}_{22}$	$\underbrace{278.5, \dots, 278.5}_{11}$
183.5:	$\underbrace{39.5, \dots, 39.5}_{25}$	$\underbrace{129, \dots, 129}_{25}$	$\underbrace{220, \dots, 220}_{18}$	$\underbrace{278.5, \dots, 278.5}_{8}$
259:	$\underbrace{39.5, \dots, 39.5}_{28}$	$\underbrace{129, \dots, 129}_{23}$	$\underbrace{220, \dots, 220}_{21}$	$\underbrace{278.5, \dots, 278.5}_{3}$

Two ordered categories ($r \times k$ table)

- Spearman's rank correlation r_s :

$$\hat{r}_s = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Pearson correlation of the ranks x and y

- Null hypothesis: $r_s = 0$. Alternative hypothesis: $r_s \neq 0$.
- Test statistic and p -value done using either an exact algorithm or an approximate t -test.
 - In R the exact algorithm is used if there are **no ties** and if the sample size n is less than 1290.
 - For applications, it is of minor importance which tests are used, so we simply ignore the issue. This remains true when Spearman's rank correlation is used in papers!
- **Chicken example:** $\hat{r}_s = -0.1969$, $p = 0.0006596$ (***)

Activity of Chicken: Comparison of hypothesis tests

Statistical power gained when using that both treatment and response are ordinal!

Test	X^2	Effective DF	p-value
Spearman rank correlation	NA	1	0.00066 (***)
Kruskal-Wallis	13.0272	3	0.00458 (**)
Chi-squared	17.6187	9	0.03909 (*)

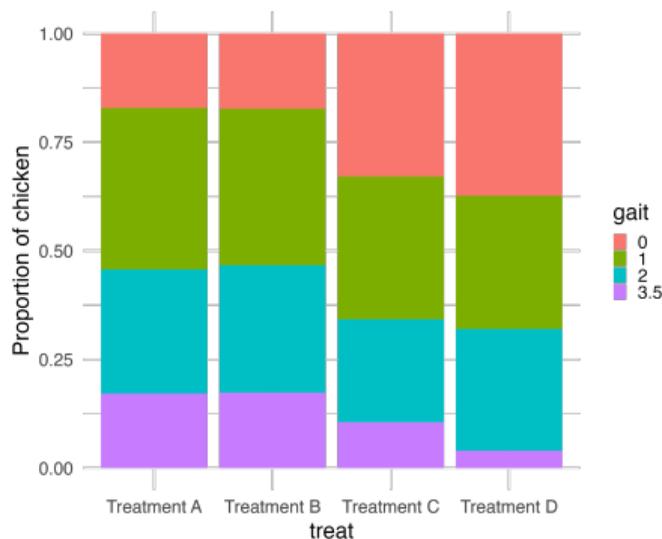
- Performing the tests in R can be seen in the R script `chicken.R`.
- **Kruskal-Wallis** and **Spearman's rank correlation** require that data is given with individual observations, i.e. in a data frame with 2 variables and 296 observations.
 - One way of transforming **table data** to **individual data** is shown in the R script `chicken.R`.

Activity of Chicken: Presentation of results

Spearman's rank correlation shows strong evidence of treatment effect on gait score ($\hat{r}_s = -0.1968$, $p = 0.00066$).

The estimated proportions are given in the following table:

Count, Row %	GAIT-score				Total
	0	1	2	3.5	
A	12 17.14	26 37.14	20 28.57	12 17.14	70 100
B	13 17.33	27 36.00	22 29.33	13 17.33	75 100
C	25 32.89	25 32.89	18 23.68	8 10.53	76 100
D	28 37.33	23 30.67	21 28.00	3 4.00	75 100
Total	78 26.35	101 34.12	81 27.36	36 12.16	296 100



- Confidence intervals on the proportions might be given.
- Quantification of the association might be improved.
- Analysis might be refined by pairwise comparisons of the treatment groups (plot suggests: A=B, C=D).

Agreement between two binary responses

Student	Correct answer	
	Pre-test	Post-test
1	No	No
2	No	Yes
3	No	No
4	No	Yes
5	No	Yes
6	Yes	Yes
7	No	No
...		
...		
30	No	Yes

Correct answer to question		Post-test		Total
		Yes	No	
Pre-test	Yes	2	0	2
	No	21	7	28
Total		23	7	30

- Did the teaching improve the students' performance, or is there **agreement** between performance before and after the teaching?
 - Are the responses the same at the two time points? (See slides 49–50)
 - Is there a direction of the disagreement? (See slides 51–52)
- Similar issues arise when **two raters categorize the same probes**, either on nominal or ordinal scale. Possibly with more than 2 categories.

Measuring agreement: Kappa coefficient

A straightforward way to measure agreement: count observations on the diagonal

		Variable 2		Total
		+	-	
Variable 1	+	a	b	a+b
	-	c	d	c+d
Total		a+c	b+d	n

- Observed fraction on the diagonal:

$$p_{\text{obs}} = \frac{a + d}{n}$$

- Expected fraction on the diagonal (given the marginals):

$$p_{\text{exp}} = \frac{a + b}{n} \cdot \frac{a + c}{n} + \frac{c + d}{n} \cdot \frac{b + d}{n}$$

- Kappa coefficient** (R: `kappa2()` from `irr`-package)

$$\kappa = \frac{p_{\text{obs}} - p_{\text{exp}}}{1 - p_{\text{exp}}}$$

Measuring agreement: Kappa coefficient (II)

- Maximal agreement if $\kappa = 1$. Positive agreement if $\kappa > 0$.
- How large κ should be to claim strong agreement is contextual. However, here is a guideline:

Value of κ	Strength of agreement
< 0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very good

- Making hypothesis tests on κ makes less sense (as the only a priori interpretable values are 0 and 1). But we may of course make confidence intervals.
- If there are more than two levels we may also use the **weighted kappa coefficient** which incorporates the distance to the diagonal.

Direction of disagreement – McNemar's test for paired binary observations

		Variable 2		Total
		+	-	
Variable 1	+	a	b	a+b
	-	c	d	c+d
Total		a+c	b+d	n

- **Null hypothesis:** known as equality of **marginal proportions**

$$\mathbb{P}(\text{Variable 1} = +) = \mathbb{P}(\text{Variable 2} = +) \quad \implies \quad \mathbb{P}(-, +) = \mathbb{P}(+, -)$$

- **Estimates (under the model):**

$$\hat{\mathbb{P}}(+, -) = \frac{b}{n}, \quad \hat{\mathbb{P}}(-, +) = \frac{c}{n}$$

- **Test statistic and continuity correction:**

$$z = \frac{b - c}{\sqrt{b + c}}, \quad z_c = \frac{|b - c| - 1}{\sqrt{b + c}}$$

Squared test statistic evaluated in a chi-squared distribution with 1 degree of freedom.

Data example 3: Marijuana and sleeping problems – Matched case-control

Null hypothesis: $p_{\text{marijuana}} = p_{\text{controls}}$

- **Design:** 32 cases (marijuana users)
32 controls (cases matched: age, sex, job, ...)
- **Observation:**

Sleeping problems		Marijuana group		Total
		Yes	No	
Controls	Yes	4	9	13 (41 %)
	No	3	16	19
Total		7 (22 %)	25	32 (100 %)

- **Estimates:** $\hat{p}_{\text{marijuana}} = \frac{7}{32} = 0.22$, $\hat{p}_{\text{controls}} = \frac{13}{32} = 0.41$
- **McNemar's test with continuity correction:**

$$z_c = \frac{|9 - 3| - 1}{\sqrt{9 + 3}} = \frac{5}{\sqrt{12}}$$

- **p-value:** $p = 2 \cdot \mathbb{P}(Z^2 > z_c^2 = \frac{25}{12}) = 0.1489$ (NS)

Stuart-Maxwell test for marginal homogeneity – paired ≥ 2 categories

Eye-testing (unaided distance vision performance) of $N = 7477$ female employees in Royal Ordnance factories between 1943 and 1946.

Right eye	Left eye				Total
	1st grade	2nd grade	3rd grade	4th Grade	
1st grade	1520	266	124	66	1976
2nd grade	234	1512	432	78	2256
3rd grade	117	362	1772	205	2456
4th Grade	36	82	179	492	789
Total	1907	2222	2507	841	7477

- Null hypothesis: green distribution = blue distribution.
- Here this means that left and right eyes perform equally well.
- R analysis: `stuart.maxwell.mh()` from `irr`-package.

Bowker's test for asymmetry – paired ≥ 2 categories

Right eye	Left eye				Total
	1st grade	2nd grade	3rd grade	4th Grade	
1st grade	1520	266	124	66	1976
2nd grade	234	1512	432	78	2256
3rd grade	117	362	1772	205	2456
4th Grade	36	82	179	492	789
Total	1907	2222	2507	841	7477

- Null hypothesis: **green counts** and **blue counts** mirror each other across the diagonal.
- Here this means that there is complete “symmetry” between joint performance of left and right eyes. Note that this hypothesis is more restrictive than marginal homogeneity.
- R analysis: `mcnemar.test()`

Summary of lecture

- Non-parametric tests can be used when normality assumptions are not satisfied.
 - Common non-parametric tests available for designs corresponding to t -tests and 1-way ANOVA.
- Risk of wrong conclusions when the statistical **power** is low.
 - Gelman & Carlin even recommend retrospective power analysis. This is, however, somewhat controversial.
- Analysis of 2-way contingency tables.
 - Association vs. marginal homogeneity (what is the relevant null hypothesis?)
 - Simulation of p -value available in the LabApp1Stat-package in cases, where the approximation by the χ^2 -distribution is invalid.
 - Nominal vs. ordinal variables (power to be gained!)