Statistiek I

χ^2

John Nerbonne

CLCG, Rijks*universiteit* Groningen

http://www.let.rug.nl/nerbonne/teach/Statistiek-I/

Overview

Two Variables

We often wish to study not merely a single variable, but rather the relation between two variables — categorical or quantitative — as these are realized on a range of individuals.

no numeric var.	race and hair color, sex/gender and word choice, syndrome (e.g., Broca's) and symptom (e.g. pronunciation)

methods to visualize/analyse relations between categorical variables:

- cross tables
- side-by-side box diagrams
- side-by-side histograms
- χ^2 test of independence

Cross-Tables

CROSS-TABLES show frequencies of nominal variables by value—values of one var. horizontally against the values of another vertically. Here is ability class by sex (see first lecture on "descriptive statistics"):

%	SEX	male	female	
		1	2	Row Total
NL_CLASS				
0		3	3	6
beginner				15.0
1		12	6	18
advanced beginner				45.0
2		6	7	13
intermediate				32.5
3		2	1	3
advanced				7.5
Column Total		23 57.5	17 42.5	40 100.0

Cross-Tables

SPSS: `Statistics >> Descriptives >> Cross-Tabs`

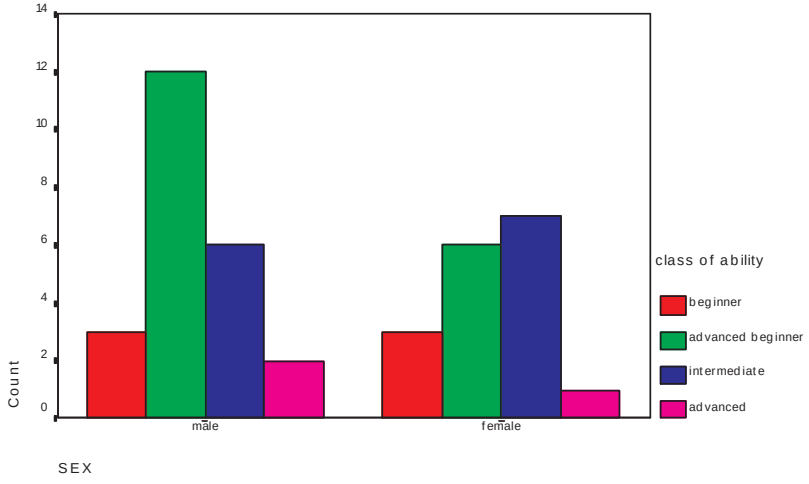
Each cell shows the number of cases with the combination of values. Upper left shows that 3 beginners are male, etc.

Presentation restructured and summed, but original data not lost.

We examine other ways of visualizing the data, focusing on the differences in the distribution depending on sex. We might also have looked at the distribution depending on area of origin, but we need to choose a perspective.

Side-by-Side Histograms

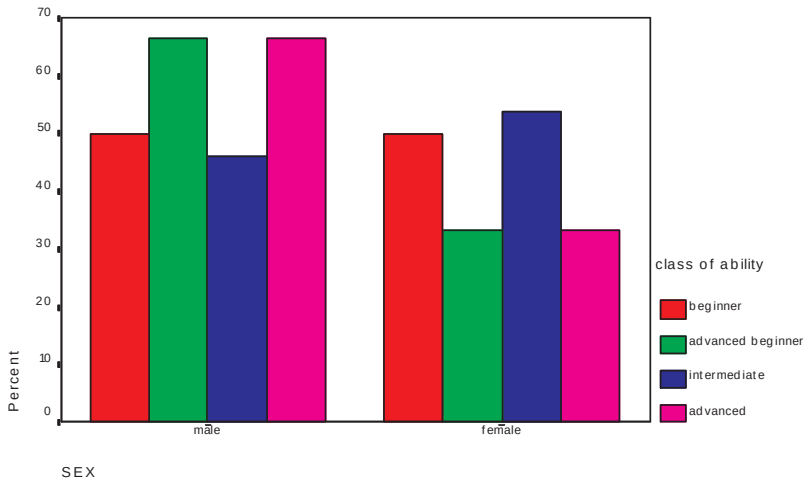
Visualize information in cross table using HISTOGRAMS, one for male, one female.



This retains all original information as well, in particular, frequencies.

Relative Histograms

Relative histograms, showing percentages **hide** some data (the frequencies).

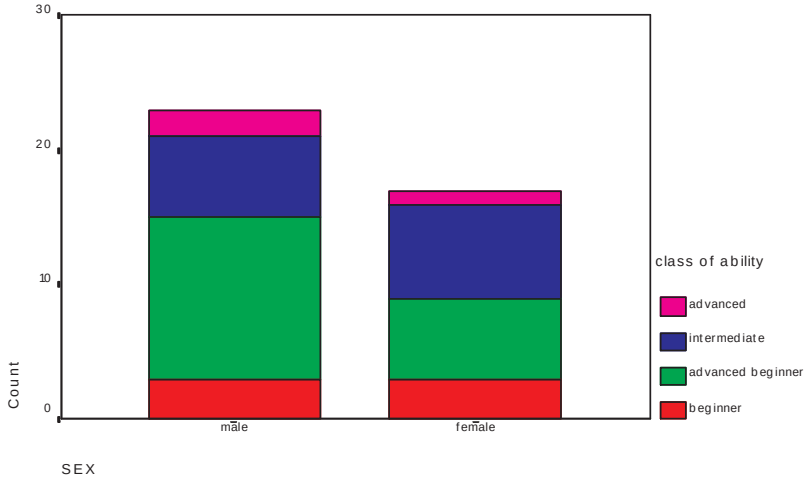


"Advanced" column shows that ca. 2/3 were male, 1/3 female—but hides frequencies. Most people don't see this directly.

Relative histograms may be appropriate when *rates* are significant, but use with caution (sparingly)!

Segmented Bar Charts

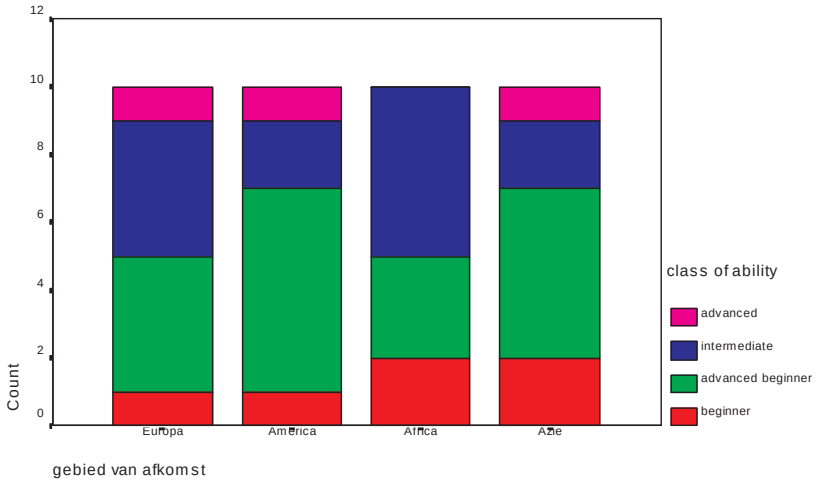
SEGMENTED BAR CHARTS show frequencies and ease visual comparison.



At a glance: *more* men than women involved, but more women in better classes.
Segmented bar charts **recommended** means of visualizing relation between nominal variables (side-by-side histograms also OK).

Area and Ability

To examine now the relation between area of origin and language proficiency level in the test, we may examine a segmented bar chart (and a cross-table, of course).



Quantitative vs. Categorical Variable

But recall the language proficiency was originally a quantitative variable, which we classified for convenience.

The classification of the quantitative scores hides the detail of the scores themselves.

There are other ways of analyzing the relation between one quantitative and one categorical variable, — e.g. *t*-tests.

χ^2 Background

Definition of CONDITIONAL PROBABILITY:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

Sometimes we're interested in the probability of a circumstance within a larger group, e.g.,

- the probability that people of strong religious conviction vote for the Christian Democrats $P(\text{CDA}|\text{Rel.})$
- the chance of dyslexia among boys $P(\text{dyslexia}|\text{sex=m})$
- the chance of internet shopping among the young $P(\text{www-shopper}|(\text{age} \leq 25))$

Statistical Independence

If the condition is irrelevant, then the variables are STATISTICALLY INDEPENDENT.

X, Y are statistically independent iff $P(X|Y) = P(X)$.

- The chance of rain on Tuesdays $P(\text{rain}|\text{day=Tues}) = P(\text{rain})$
- The chance of a web site being positively evaluated when frames are used. $P(\text{eval=+}|\text{frames-in-www-page})$
- The chance of aphasia in right-handed people $P(\text{aphasia=+}|\text{right-handed})$

We can check on statistical (independence) simply by counting, and checking whether proportions are the same.

Statistical Independence in Tables

The rows of a table show the effect of different conditions on the column variable.

Work	Verb	Noun	Other	Total
Text 1	32	28	40	100
Text 2	52	27	21	100
Totals	84	55	61	200

Where the row, column variables are **independent**, then the relative frequencies in the different rows should show about the same proportions (with respect to the totals).

x^2 test of independence test whether these proportions are the same.

Statistical Analysis of Nominal Data: χ^2

- suitable for nominal data
- compares frequencies in classes
- determine whether there is significant difference in observed freq, expected freq (in respective classes)
- frequent application: **test of independence** applied to cross-tables

$$\chi^2 = \sum_{\forall i} \frac{(o_i - e_i)^2}{e_i}$$

where o_i **observed** freq. in class i and e_i **expected** freq. in class i

Example Application of χ^2, I.

Wim Vuijk *Zicht op interne communicatie*, Ch.6

Vuijk compares two versions of a single text, *Aanpak Ziekteverzuim*, prepared for use in a municipal agency. One version *with* an introduction stating the purpose of the text, the second *without* that.

He then asked the question: What is your opinion on the text?

	'clear'	'unclear'	'other'
W.Intro			
Without			

χ^2 asks: does the variable 'with/without introduction' influence the variable 'clarity'?

Example Application of χ^2, II

Kos (thesis under **Bastiaanse**) replicated work by Blumstein, examining phonological errors in different sorts of aphasia.

	subst	addition	omission	transposition
paarden [pardə]	[tardə]	[spardə]	[ardə]	[pradə]

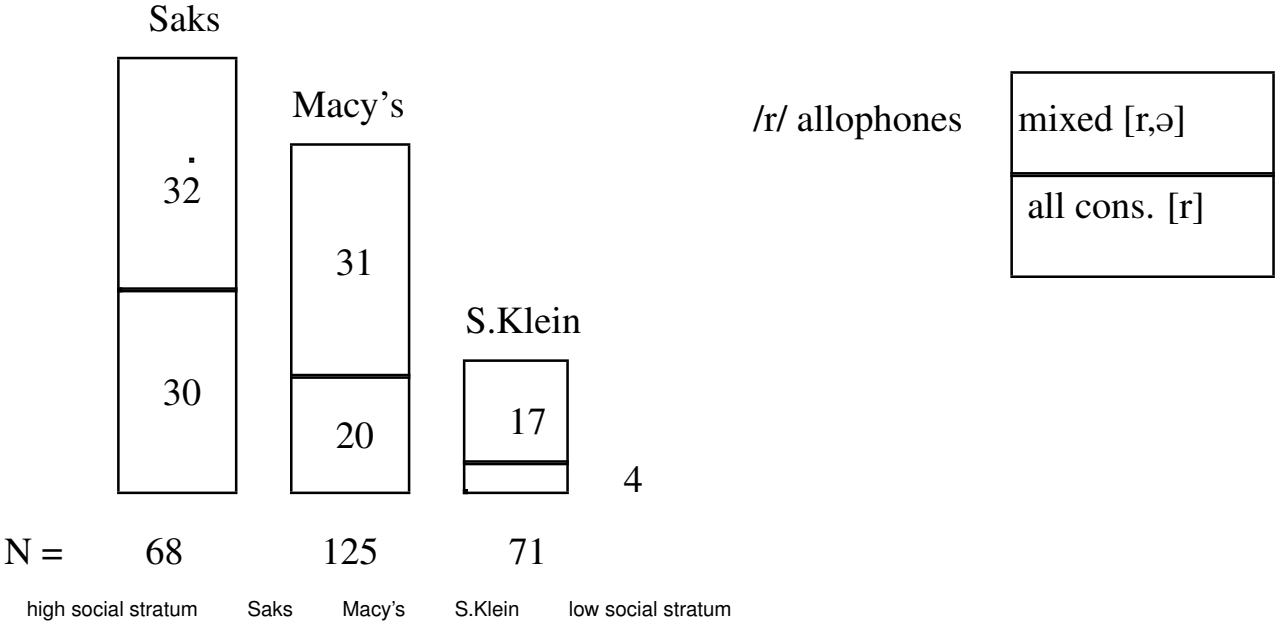
She then asked the question: Does the sort of aphasia influence the sort of phonological error?

	subst	addition	omission	transposition
conduction aphasia				
other aphasia				

χ^2 asks: does 'aphasia sort' influence the (frequency of the values of the variable) 'phonological error sort'?

Example Application of χ^2, III

William Labov examined variant pronunciations of syllable-final /r/ in American English ([r] vs [ə]). New York used to be like Boston, final /r/ is [ə], but it started changing in the 1950's and 1960's. Labov hypothesized a social basis for the change.



Example Application of χ^2, III

Labov asked: Does social status influence the pronunciation of final /r/?

Social Status	Pronunciation of /r/		
	cons. ([r])	vocalic ([ə])	mixed
high	30	6	32
medium	20	74	31
low	4	50	17

χ^2 asks: does 'social status' influence the (frequency of the values of the variable) 'pronunciation of final /r/'?

This led to new understanding of the social motivation for language change.

Example Application of χ^2, IV

Frequent "literary" application of χ^2: **authorship studies**.

Examples:

- Aristotle: *Rhetoric to Alexander* (Kenny)
- Hamilton, Madison and Jay: *Federalist Papers*
  cf. Welling's WWW site:
  `http://www.let.rug/~welling/usa/federalist.html`
  Who was "Publius"?
- W.F.Hermanns war diaries (Dorlein and Zwarts)
  editing/rewriting much later?

Literary Application of χ^2

We check check whether the distribution of a style variable is (proportionally) similar in two authors using χ^2.

1. **Preparation**
   1. identify quantifiable characteristic of style (literary)
   2. show that style element characterizes writer (or genre, or epoch, or writer at a particular period on a particular subject, etc.)
2. **Apply** χ^2
   1. identify null hypothesis (normally that style and text variables are independent)
   2. test likelihood of sample
   3. if sample unlikely (low p-value), then style and text are dependent.
3. **Combine (1) and (2)**: If H_0 is rejected, we conclude that texts differ in style. If style characterize authors (1), we conclude that texts have different authors.

Kenny on Aristotle

Procedure

1. identify characteristic of style (literary)
   syntax: which part of speech at end of sentence (last word): verb, noun or other
2. show that it identifies writer (or genre, or epoch, or subject *s*, etc.)
   unavailable!
3. identify null (authorship) hypothesis
   distribution (of part of speech at sentence-end) same in questionable work and similar work definitely by Aristotle
4. test likelihood of null hypothesis
   data on (i) *Rhetoric to Alexander* and (ii) *Rhetoric A*

Kenny's Sample

First 100 sentences of each text:

Work	Verb	Noun	Other	Total
Rhetoric A	32	28	40	100
Rhetoric to Alexander	52	27	21	100
Totals	84	55	61	200

This is a CROSS-TABLE or a 2×3 CONTINGENCY TABLE

Note generality: test whether distribution of variable values *sentence-ending-type* is affected by distribution of *author*.

If the relative frequencies are about the same, then the variables are statistically independent.

Calculating χ^2

Problem: what are expected frequencies?

Solution: use existing data (assume that both samples from same population), then relative frequencies (%'s) should be same.

	Observed				**Expected**		
Work	Verb	Noun	Other	Total	Verb	Noun	Other
Rhetoric A	32	28	40	100	42	27.5	30.5
Rhetoric to Alexander	52	27	21	100	42	27.5	30.5
Totals	84	55	61	200			

$$\text{expected freq.} = \frac{\text{row-total} \times \text{col.-total}}{\text{grand-total}}$$

Applying χ^2

$$\chi^2 = \sum_{\forall i} \frac{(o_i - e_i)^2}{e_i}$$

where

- o_i observed freq. in class i
- e_i expected freq. in class i

In this case i ranges over the cells of the table.

Calculating χ^2

$\chi^2 = \sum_{\forall i} \frac{(o_i - e_i)^2}{e_i}$

O	E	$(O-E)^2$	$(O-E)^2/E$
27	27.5	0.25	0.01
28	27.5	0.25	0.01
32	42	100	2.38
52	42	100	2.38
40	30.5	90.25	2.95
21	30.5	90.25	2.95
$\sum$			10.68

This is the χ^2 value.

Degrees of Freedom

In a table, degrees of freedom are **not** #cells−1

Work	Verb	Noun	Other	Total
Rhetoric A	x	y		100
Rhetoric to Alexander				100
Totals	84	55	61	200

Once x and y are known, other cells known.

For example, in Rhet-A, "other" must be $100 - (x + y)$.

dF $= 2$

For tables in general, dF $= (\#\text{rows} - 1) \times (\#\text{cols} - 1)$

Applying χ^2 tables

O_i	$(O_i - E_i)^2/E_i$
$\sum$	10.68

dF $= 2, \chi^2 = 10.68$

How likely is H_0 (refer to χ^2 tables)?

If $p \leq 0.005$, should we assume that the two works are by the same author?

SPSS (Remember the 'weight by frequency' trick!)

```
Input

1     1    32
1     2    28
1     3    40
2     1    52
2     2    27
2     3    21

Output

Chi-Square Tests
                   Value     df  Asymp. Sig. (2-sided)
Pearson Chi-Square  10,698a     2      ,005
N of Valid Cases 200

a 0 cells (0,0%) have expected count less than 5. The
minimum expected count is 27,50.
```

So the data prove the hypothesis to be significant. But how strong is it? What's the EFFECT SIZE?

Effect Size

Just as we attend to EFFECT SIZE in comparing means (size of difference in standard deviations), we wish to estimate effect size in χ^2 as well

Here we rely on SPSS and the Cramer's V to estimate effect size.

We can calculate that the variable interaction is strong to the degree $V = 0.23$. This is MODEST.

Roughly, Cramer's V is understood as the fraction of variance in the data that may be explained by the variable dependence. Complex tables are difficult to summarize in a single measures of association.

Remember: it is particularly important to attend to effect size with large samples (> 100), where significance may be (too) easy to attain, as in corpus linguistics or web experiments.

SPSS

Remember to set the option in "Crosstabs", "Statistics" that you want to see Cramer's V.

Symmetric Measures

	Value	Approx. Sig.
Cramer's V	,231	,005
N of Valid Cases	200	

So the influence has been proven significantly, but the effect size is modest.

SPSS: Examining Dependencies

Remember to ask for "standardized residuals" – these are z-scores, and can help to pinpoint where expectations are violated.

SText * Syntax Crosstabulation

		Syntax			Total
		Verb	Noun	Other	
Text					
Rhetoric A	Count	32	28	40	100
	Std. Res.	-1,5	,1	1,7	
Rhetoric to Alexander	Count	52	27	21	100
	Std. Res.	1,5	-,1	-1,7	
Total	Count	84	55	61	200

Here, the "other" category contributes the most to the imbalance, but it's still less than the $z > \pm 1.96$ that we use as a rule of thumb.

Alternative Measure of Effect Size

A simpler measure of effect size is often reported for tables with two rows, namely the DISSIMILARITY INDEX:

	Observed				Expected		
Work	Verb	Noun	Other	Total	Verb	Noun	Other
Rhetoric A	32	28	40	100	42	27.5	30.5
Rhetoric to Alexander	52	27	21	100	42	27.5	30.5
Totals	84	55	61	200			

N.B. if 10 elements moved in the first column, the counts would be the same, and similarly for 0.5 elements in the 2nd,and 9.5 in the 3rd. There are thus 20 elements in total (in a sample of 200) that make the rows dissimilar. The DISSIMILARITY INDEX is 20/200 = 0.10.

$$\frac{1}{2}\sum_{i=1}^{C}\left|\frac{c_{1i}}{\sum_j c_{1j}} - \frac{c_{2i}}{\sum_j c_{2j}}\right|$$

Odds Ratios

Gambler's probability: ODDS. If FC Groningen has a 20% chance of winning against Firenze, the odds are four to one against Groningen (80 vs. 20).

To compare one condition against another, we examine the ODDS RATIO.

Work	Verb	Noun
Rhetoric A	32	28
Rhetoric to Alexander	52	27
Totals	84	55

The odds seeing a verb at sentence end in *A* are 8-to-7 (1.14), and in *R to Alexander* nearly 2-to-1 (1.93). The odds ratio compares these.

The ODDS RATIO of seeing verbs in *R to Alexander* as opposed to *A* is $1.93/1.14 = 1.69$. So you're 1.69 times more likely to encounter verbs at the ends of sentences in *R to Alexander* than in *A*

Odds Ratios and Effect Size

Note that odds ratios are insensitive to table size. This is what makes them a good indication of effect size.

	Small		**Large**	
Work	Verb	Noun	Verb	Noun
Rhetoric A	32	28	320	280
Rhetoric to Alexander	52	27	520	270

The odds ratios are identical. Try it as an exercise!

(Log) Odds Ratios

Odds ratio ranges from 0 to ∞, where 1 indicates no effect.

Because this is asymmetric, most people prefer LOG ODDS RATIO, often referred to as "log odds", which is just the (natural) logarithm of the odds ratio. Since the log odds ratio was 1.69 for the two tables above the log odds is $\log_e 1.69 = 0.23$.

Remember to correct for the log in expressing the effect size! You're $e^{0.23}$ times more likely to see a sentence final verb in *R to Alexander*.

Note that the log ranges from $-\infty$ to $+\infty$, with 0 as the mid point. This is the symmetry that was missing from the odds ratio.

Conditions on χ^2

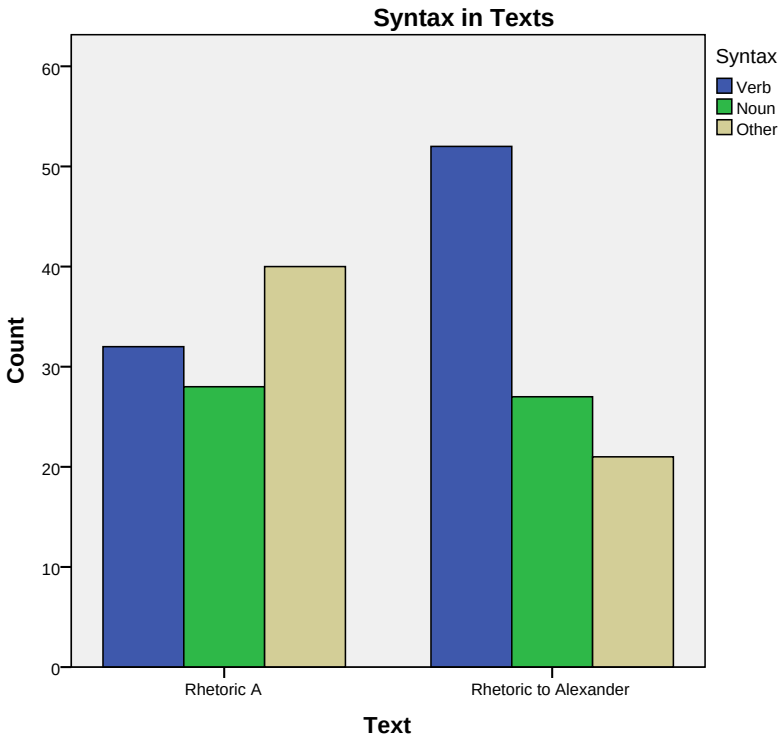
Test requirements

1. each o_i falls into exactly one class
2. outcomes for N independent observations
3. sample N large

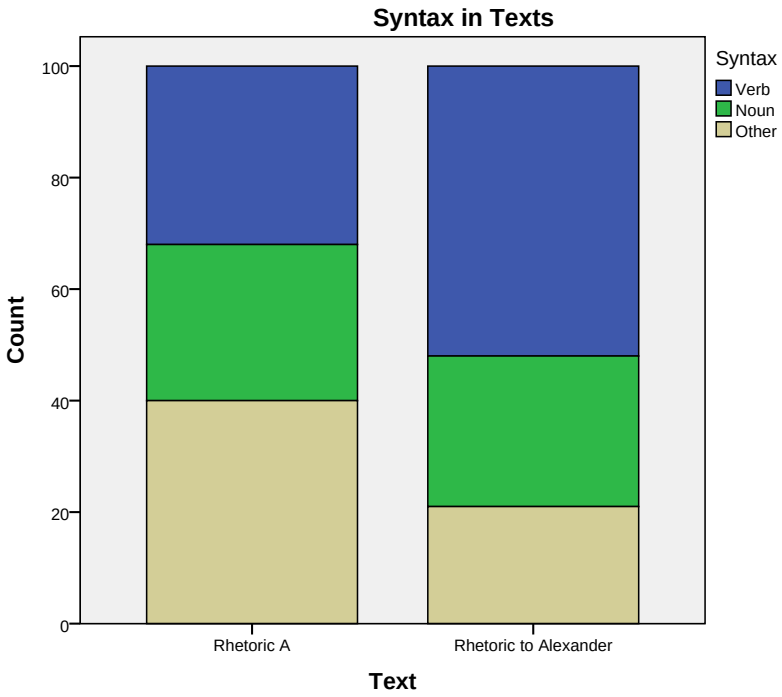
ad(1) no ambiguous classification, no differing classifications
ad(2) study specific. Tense use e.g. shouldn't study consecutive verbs, but every 100th
ad (3) mean expected frequency (of all cells) ≥ 5, no cell has an expected frequency of less than one

Visualizing the χ^2 Question via RELATIVE HISTOGRAM

χ^2 is asking whether freq. distributions are comparable.



Alternative Visualization: Segmented Bar Chart



Cases weighted by Frequency

Modifying χ^2

It can be useful to group var. values

Example: you've tested language attitudes among foreigners living in the Netherlands:

attitude: positive, neutral, negative
nationality: American, English, French, German, Maroccan, Turkish, ...

—maybe more revealing to group **nationality** into N-Am., Euro., and Other

—depends on research question!

Yates's Correction

When investigating 2×2 tables (4 cells), with small frequencies, the common χ^2 measure *overstates* the improbability of the sample

YATES'S CORRECTION must be applied in this case

$$\chi^2 = \sum_{\forall i} \frac{(|o_i - e_i| - 0.5)^2}{e_i}$$

where o_i, e_i observed, expected freq. in class i

If the correction is not applied, results may be reported as significant which are not.

It makes little difference when numbers are large (all cell frequencies above 30), and is not needed for tables larger that 2×2.

Consider also FISHER'S EXACT TEST when numbers are small.

Paired Data in χ^2

Analyze paired (binary) data in (a special variant of) χ^2, the McNemar Test

Example: Your work as a communication advisor and test the effectiveness of a promotional video promoting study at the University of Groningen.

You collect a sample of data from potential students before and after they see the video. Because the data is from the same subjects, it is paired data, and the samples are RELATED.

before/after	*RuG+*	*RuG−*	Total
RuG+	30	50	80
RuG−	10	32	42
Totals	40	82	

McNemar, Paired Data in χ^2

McNemar ignores the unchanging data, focuses on changes:

before/after	*RuG+*	*RuG−*	Total
RuG+	30	50	80
RuG−	10	32	42
Totals	40	82	

and compares (above) $(50-10)^2/(50+10)$ against the χ^2 distribution with one degree of freedom.

Use Yate's continuity correction for tables with under 30/cell.

See fourth SPSS exercise!

Avoid Overuse of χ^2

χ^2 lacks a general measure of effect size, such as number of standard deviations for differences in means. Cramer's V is OK for simple tables.

χ^2 used often in corpus linguistics, which nearly always results in significance due to sheer numbers (100's even 1,000's of examples).

Consider advanced alternatives if comparing large numbers, e.g. binomial test (later in this course) or (log) odds-ratios (advanced).

Alternatives to χ^2

χ^2 not very sensitive, so use numerical variables where appropriate

Example: Instead of testing:
This course was (check one)
very useful [] *useful []* *not useful []* *completely useless []*

Try so-called **Likert scales**
On a scale of 1-7, this course was
very useful (1) _____ *completely useless (7)*

Likert Scales

On a scale of 1-7, this course was
very useful (1) _____ *completely useless (7)*

- use at least 5 points
- allow a "center" (use 1 through odd number), unless you wish to force a positive or negative decision
- be consistent, keeping "positive" answers on one side
  —e.g., always low (as above)
- compare using interval-level tests (Mann-Whitney), perhaps even using quantitative tests (*t*-tests, ANOVA), if judgments normally distributed.

χ^2

χ^2 **Goodness of Fit**. Since χ^2 compares observations to expectations, we can apply it to check on how well observations fit a theoretical distribution.

Example: coin toss, results heads (*h*) or tails (*t*), series of 100 tosses recorded.

H_0 coin fair, dist. 50-50 and H_a coin unfair, dist. not 50-50

Critical region: let's reject H_0 only if 95% certain

dF: since there are two variable values (frequencies in respective classes), there is one dF

cf. χ^2 table, e.g., M&M Tabel G (p.642), $P_{df=1}(\chi^2 \leq 3.84) = 0.95$

If $\chi^2 > 3.84$, then we're 95% certain that coin is bad.

χ^2 Application to Coin Toss

χ^2 "Goodness of Fit" test: If $\chi^2 > 3.84$, then we're 95% certain that coin is bad.

$$\chi^2 = \sum_{\forall i} \frac{(|o_i - e_i|)^2}{e_i}$$

where o_i, e_i observed, expected freq. in class i. Expected is 50.

Calculating χ^2 Test of Independence

Since there are only two (symmetric) classes, we ask when is $2 \times (o - e)^2/e = 3.84$?

$$\begin{aligned} 2 \times (o - e)^2/e &\geq 3.84 \\ (o - 50)^2/50 &\geq 1.92 \\ (o - 50)^2 &\geq 96 \\ o^2 - 100o + 2500 &\geq 96 \\ o^2 - 100o &\geq -2404 \\ o &\leq 40 \end{aligned}$$

If we see ≤ 40 heads (or ≥ 60), then we're 95% certain that coin is dishonest.

Verifying Calculations (χ^2 Test of Indep.)

$$\begin{aligned} (36-50)^2/50 &\;?\; 3.84 \\ (14)^2/50 &\;?\; 3.84 \\ 192/50 &\;?\; 3.84 \\ 3.92 &\;>\; 3.84 \end{aligned}$$

If we use binomial distribution, then we obtain:

$$\text{SE} = \sqrt{p(1-p)100} = \sqrt{0.5(0.5)100} = \sqrt{(0.25)100} = \sqrt{25} = 5$$

The 95% CI is $50 \pm (2 \times SE) = (40, 60)$. The numeric test is similar.

But χ^2 still used to check more complicated apparatus, such as roulette wheels.

Puzzle about Relations

Here is data from two hospitals about patient survival after surgery. All of the patients who underwent surgery are included in the table. 'Survived' means that the patient was alive six weeks after the surgery.

	Hospital A	Hospital B
Deceased	63 (3%)	16 (2%)
Survived	2037	784
Total	2100	800

Which hospital would you choose for surgery?

Hidden Variables

With only the original information, Hospital B appears better.

But appearances can deceive. Consider:

	Good Condition		Poor Condition	
	A	B	A	B
Deceased	6 (1%)	8 (1.3%)	57 (3.8%)	8 (4%)
Survived	594	592	1433	192
Total	600	600	1500	200

Notate bene In *both* categories of patients, hospital A has better survival rates. But it attracts more difficult patients (perhaps because it's better?) So it's overall survival rate is worse.

Simpson's Paradox

Terminology: when we take all patients together, we AGGREGATE the values of the variable 'condition'.

Simpson's Paradox: The distribution of a nominal variable can be reversed under aggregation.

We blame this on HIDDEN VARIABLES—a variable which is important, but neglected in analysis. In this case 'condition of patients'.

Toward Causality

Terminology: STUDIES record variables as they are naturally found; EXPERIMENTS manipulate situations in order to record variables.

Very difficult to show cause-effect relationship without experiment

Hidden variables always possible

Example: Does smoking cause cancer?

Toward Causality

Example: Does smoking cause cancer?

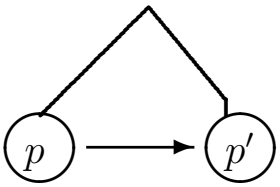
- smokers suffer lung cancer ten times more often than nonsmokers
- measure strength of smoking by cigarettes/day, and incidence of lung cancers in percentages (for all population groups) $r = 0.716, r^2 = 0.51$ amount of smoking

But the tobacco industry noted that there might be hidden variables:

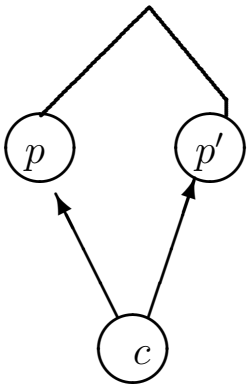
- genetic predisposition toward smoking **and** lung cancer?
  —later passive smoking linked to cancer, too
- "unhealthy liefstyle" contributes crucially?

Toward Causality

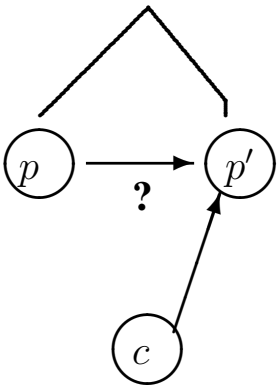
Very difficult to show cause-effect relationship without experiment



p causes p'



Common Dependence



Confounding c

Given highly correlating phenomena p, p', it could be that

- p causes p' (or *vice versa*);
- both are caused by another, hidden c (genetic predisposition); or
- there is a crucial CONFOUNDING factor ("unhealthy lifestyle")

university of groningen
Relations Statistical Independence Examples Calculations Variants, Discussion "Goodness of Fit"∗ Hidden Variables

Toward Causality

EXPERIMENT needed to systematically control potential causes and confounding factors.

"Future oriented" LONGITUDINAL STUDIES follow cases chosen for possible common dependence and/or possible.

Methodology course (Ensink, Stowe): design of experiments, data acquisition

. . . and for the smokers: animal experiments have proven the link between smoking and cancers (in animals)

Next Topics

Nonparametric Tests — Mann-Whitney, Wilcoxon