Two Variables

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

We begin with case of nonnumeric variables.

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

methods to study nonnumeric variables:

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

A More Formal View

Terminology: we speak of CASES, e.g., Joe, Sam, . . . and VARIABLES, e.g. height (h) and native language (l). Then each variable has a VALUE for each case, h_j is Joe's height, and l_s is Sam's native language.

When we examine relations, we always examine the realization of two variables on each of a group of cases.

- height vs. weight on each of a group of Dutch adults
- effectiveness vs. a design feature of group of web sites, e.g. use of menus, use of frames, use of banners
- pronunciation correctness vs. syntactic category of a group of words
- phonetic vs. geographic distance on a group of pairs of Dutch towns

Tabular Presentation

Example: A test is given to students of Dutch from non-Dutch countries. Variables:

Variable	**Values**
area of origin	EUrope, AMerica, AFrica, ASia
test score	0-40
sex	Male, Female

Here is part of the results.

area	**score**	**sex**
EU	22	M
AM	21	F
AF	15	F
AZ	26	M
⋮	⋮	⋮

Three variables, where only score is numeric, & others nominal. Each row is a CASE.

Tables show *all* the data, which is nice, but can be confusing when they become too large.

Coding

It is often necessary to code information in a particular way for a particular software package.

In general, SPSS allows fewer manipulations and analyses for data coded in letters. Use numbers as a matter of course. This causes us to recode ‘area of origin’ and ‘sex’, since these were coded in letters.

area of origin	EUrope	AMerica	AFrica	ASia
	0	1	2	3
sex	Male	Female		
	1	2		

Notate bene: this is a weakness in SPSS. In general, it is good practice to use meaningful codings. But in SPSS, this will limit what you can do—use numbers!

Classifying

It is also sometimes useful to group numeric values into classes. We'll group score into 0-16 (beginner), 17-24 (advanced beginner), 25-32 (intermediate), and 33-40 (advanced).

area	score	sex	score class
0	22	1	1
1	21	2	1
2	15	2	0
3	26	1	2
⋮	⋮	⋮	⋮

Grouping numerical information into classes loses information. Care!

Reminder:

area of origin	EUrope	AMerica	AFrica	ASia
	0	1	2	3
sex	Male	Female		
	1	2		

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:

```
      %          SEX   male     female
                                          Row
                          1        2      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       23       17        40
            Total       57.5     42.5     100.0
```

Cross-Tables

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

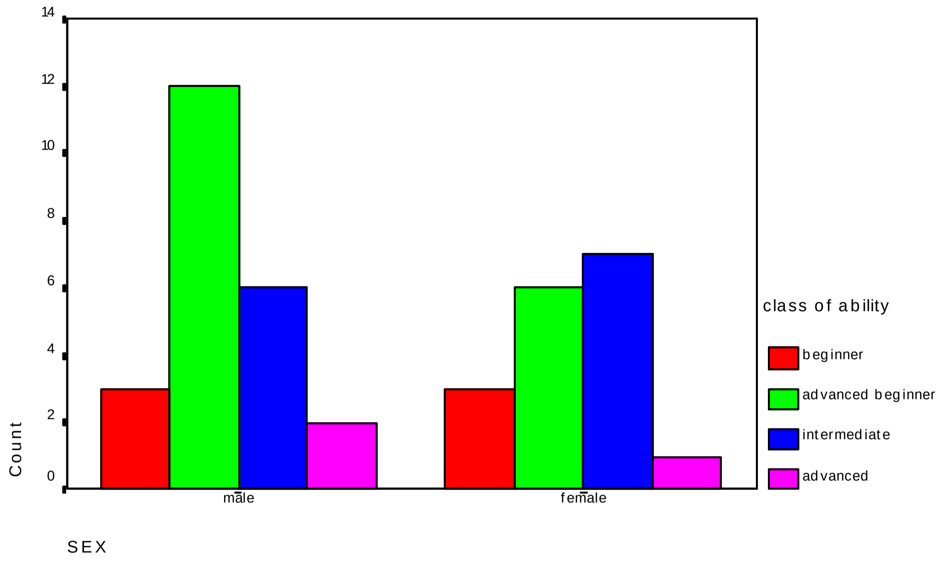
Each cell shows the number of cases with the combination of values. Upper left shows that 3 beginners are male, etc.

Presentation restructured and summed, but original data not lost.

Let's 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

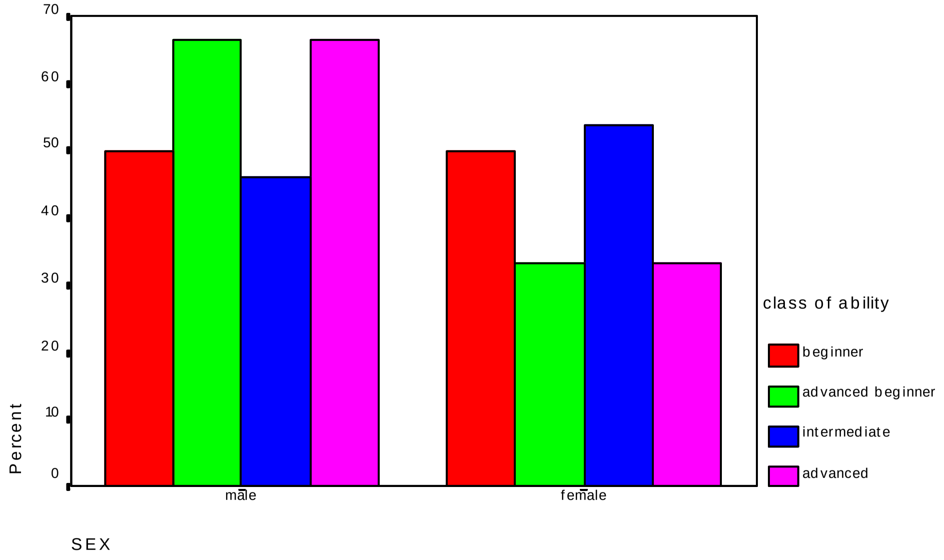
Visualize information in cross table using two HISTOGRAMS, one for male, one female.



This retains all original information as well. In particular, frequencies are shown (graphically).

Relative Histograms

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

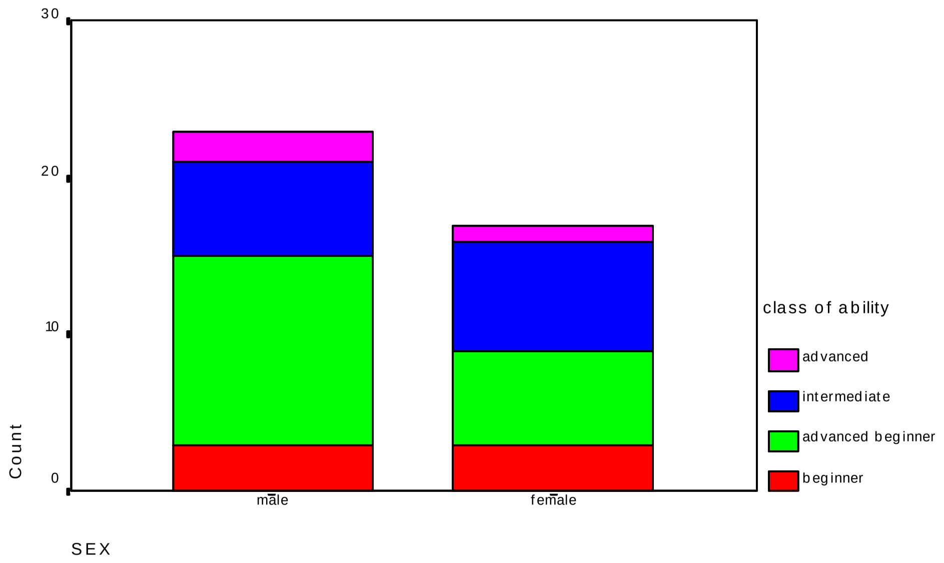


“Advanced” column shows that ca. $2/3$ were male, $1/3$ female—but hides frequencies.

Relative histograms appropriate when *rates* are significant.

Segmented Bar Charts

SEGMENTED BAR CHARTS show frequencies and ease visual comparison.

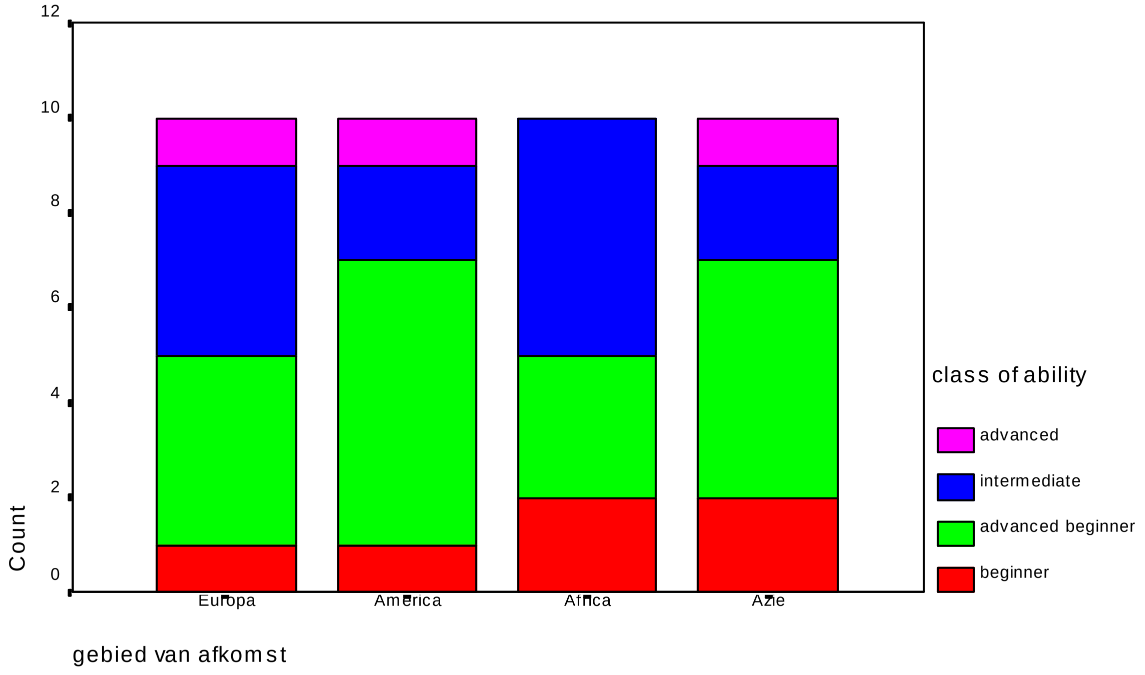


At a glance: *more* men than women were involved, but more women in better classes.

Segmented bar charts recommended means of visualizing relation between nominal variables.

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).



Numeric vs. Nonnumeric Variable

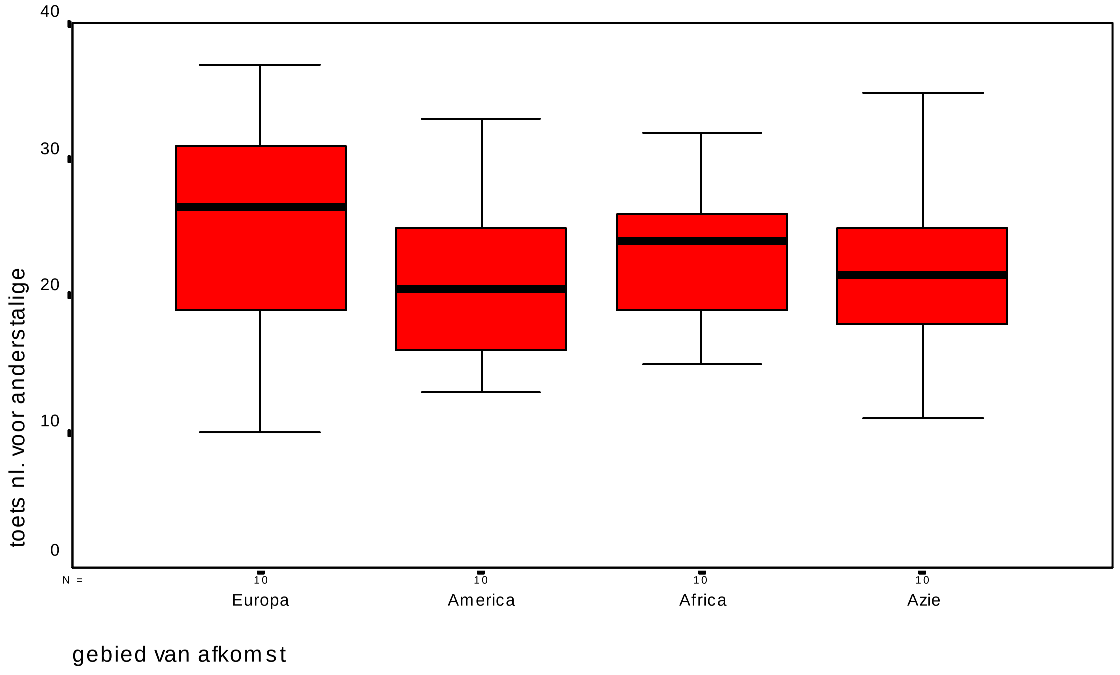
But recall the language proficiency was originally a numeric variable, which we classified for convenience.

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

There are other ways of analyzing the relation between one numeric and one nonnumeric variable.

Numeric/Nonnumeric Relation

Use multiple numeric displays (box-n-whisker diagrams) to compare cases of nominal variable.



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	16
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%)	Deceased	57 (3.8%)	8 (4%)
Survived	594	592	Survived	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)

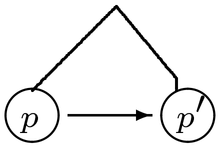
$r = 0.716, r^2 = 0.51$ amount of smoking

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

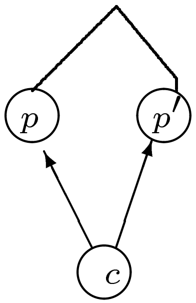
- genetic predisposition toward smoking **and** lung cancer?
  —later passive smoking linked to cancer, too
- "unhealthy liefstyle" contributes crucially?

Toward Causality

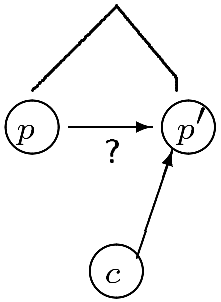
Very difficult to show cause-effect relationship without experiment



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”)

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)

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

χ^2 Test of Independence

Recall 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(CDA|Rel.)
- the chance of dyslexia among boys P(dyslexia|sex=m)
- the chance of internet shopping among the young P(www-shopper|(age ≤ 25))

Statistical Independence

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

$$X, Y \text{ are statistically independent iff } P(X|Y) = P(X).$$

- 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

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 frequencies in the different rows should show about the same proportions (including in the totals).

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

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]

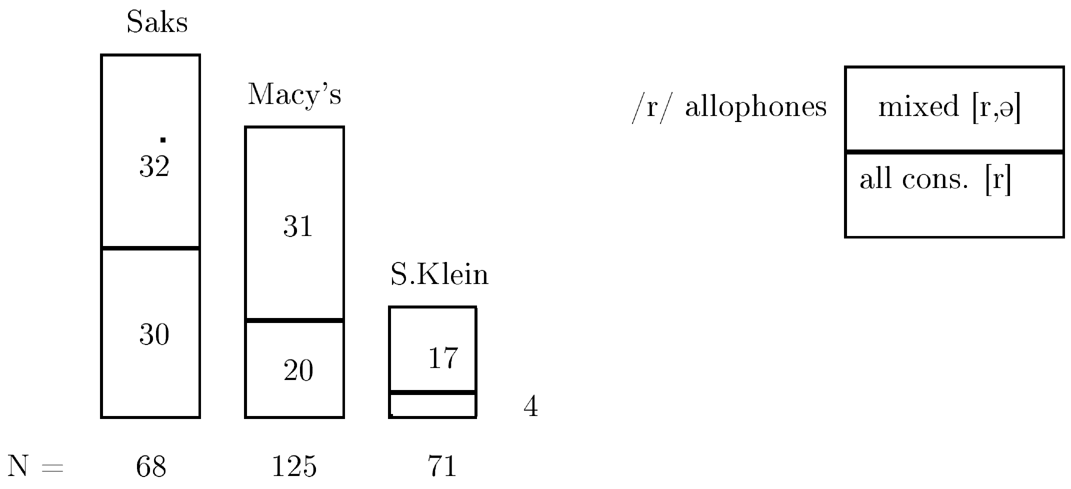
She then asked the question: Does the sort of aphasia influence the sort of phonological error?

	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.



high social stratum Saks Macy's S.Klein low social stratum

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

χ^2 Test of Independence

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

To put χ^2 to use to pose a question about authorship means that we obtain two texts and check on the distribution of a style variable.

1. **Preparation**
   (a) identify quantifiable characteristic of style (literary)
   (b) 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**
   (a) identify null hypothesis (normally hypothesis that style and text variables are independent)
   (b) test likelihood of sample
   (c) if sample unlikely (low p-value), then there is a dependence between style element and text
3. **Combine (1) and (2)**: In case H_0 is rejected, we conclude immediately that the texts differ in their style elements. If we are convinced by (1) that these style elements characterize authors, we conclude that the 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 writer at period p on a 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*.

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 $=$ (#rows $- 1) \times$ (#cols $- 1$)

Applying χ^2 tables

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

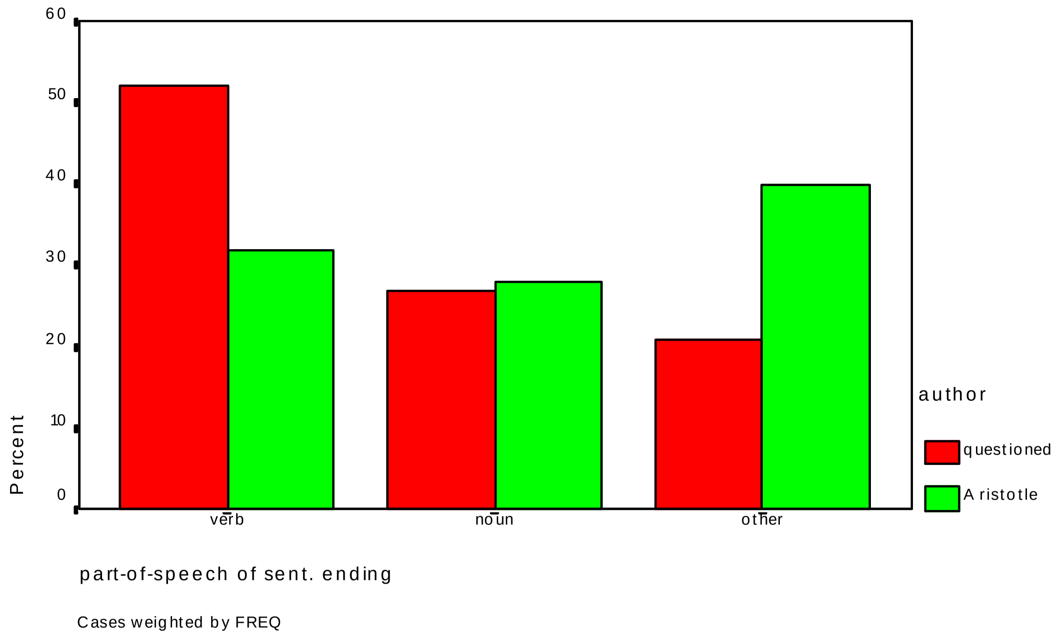
dF $= 2, \chi^2 = 10.68$

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?

Visualizing the χ^2 Question

χ^2 is asking whether freq. distributions are comparable. RELATIVE HISTOGRAM:



If distributions comparable, bars in each cluster should be about the same height. Better than segmented bar chart because of emphasis on relative 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, Canadian, English, French, German, Maroccan, Turkish, . . .

—maybe more revealing to group **nationality** into N-American, European, 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.

Avoid Overuse of 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)
- be consistent, keeping “positive” answers on one side
  —e.g., always low (as above)
- compare using t-test or ANOVA!

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), $P_{\text{df}=1}(\chi^2 \leq 3.84) = 0.95$

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.)

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

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.

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) each class has ≥ 5 elements (expected)