Making statistic claims

Corpus Linguistics

Kron

Outline of the session

- Lecture
  - Raw and normalised frequency
  - Descriptive statistics (mean, mode, media, measure of dispersion)
  - Inferential statistics (chi squared, LL, Fisher’s Exact tests)
  - Collocation statistics

Quantitative analysis

- Corpus analysis is both qualitative and quantitative
- One of the advantages of corpora is that they can readily provide quantitative data which intuitions cannot provide reliably
- “The use of quantification in corpus linguistics typically goes well beyond simple counting” (McEnery and Wilson 2001: 81)
  - What can we do with those numbers and counts?

Raw frequency

- The arithmetic count of the number of linguistic feature (a word, a structure etc)
- The most direct quantitative data provided by a corpus
- Frequency itself does **NOT** tell you much in terms of the validity of a hypothesis
  - There are 250 instances of the *f**k* swearword in the spoken BNC, so what?
    - Does this mean that people swear frequently – or infrequently – when they speak?

Normalized frequency

- …in relation to what?
  - Corpus analysis is inherently comparative
- There are 250 instances of the swearword in the spoken BNC and 500 instances in the written BNC
  - Do people swear twice as often in writing as in speech?
    - Remember the written BNC is 9 times as large as the spoken BNC
- When comparing corpora of different sizes, we need to normalize the frequencies to a common base (e.g. per million tokens)
  - Normalised freq = raw freq / token number * common base
  - The swearword is 4 times as frequent in speech as in writing
    - Swearword in spoken BNC = 250 / 10 * 1 = 25 per million tokens
    - Swearword in written BNC = 500 / 90 * 1 = 6 per million tokens
  - …but is this difference statistically significant?

Normalized frequency

- The size of a sample may affect the level of statistical significance
- Tips for normalizing frequency data
  - The common base for normalization must be comparable to the sizes of the corpora
    - Normalizing the spoken vs. written BNC to a common base of 1000 tokens?
- Warning
  - Results obtained on an irrationally enlarged or reduced common base are distorted

Descriptive statistics

- Frequencies are a type of descriptive statistics
- Descriptive statistics are used to describe a dataset
- A group of ten students took a test and their scores are as follows
  - 4, 5, 6, 6, 7, 7, 7, 9, 9, 10
- How will you report the measure of *central tendency* of this group of test results using a single score?

The mean

- The mean is the arithmetic average
- The most common measure of central tendency
- Can be calculated by adding all of the scores together and then dividing the sum by the number of scores (i.e. 7)
  - 4+5+6+6+7+7+7+9+9+10=70/10=7
- While the mean is a useful measure, unless we also knows how dispersed (i.e. spread out) the scores in a dataset are, the mean can be an uncertain guide

The mode and the median

- The mode is the most common score in a set of scores
  - The mode in our testing example is 7, because this score occurs more frequently than any other score
    - 4, 5, 6, 6, 7, 7, 7, 9, 9, 10
- The median is the middle score of a set of scores ordered from the lowest to the highest
  - For an odd number of scores, the median is the central score in an ordered list
  - For an even number of scores, the median is the average of the two central scores
    - In the above example the median is 7 (i.e. (7+7)/2)

Measure of dispersion: range

- The range is a simple way to measure the dispersion of a set of data
  - The difference between the highest and lowest frequencies / scores
  - In our testing example the range is 6 (i.e. highest 10 – lowest 4)
- Only a poor measure of dispersion
  - An unusually high or low score in a dataset may make the range unreasonably large, thus giving a distorted picture of the dataset

Measure of dispersion: variance

- The variance measures the distance of each score in the dataset from the mean
  - In our test results, the variance of the *score 4* is 3 (i.e. 7–4); and the variance of the *score 9* is 2 (9–7)
- For the whole dataset, the sum of these differences is always zero
  - Some scores will be above the mean while some will be below the mean
- Meaningless to use variance to measure the dispersion of a whole dataset

Measure of dispersion: std dev

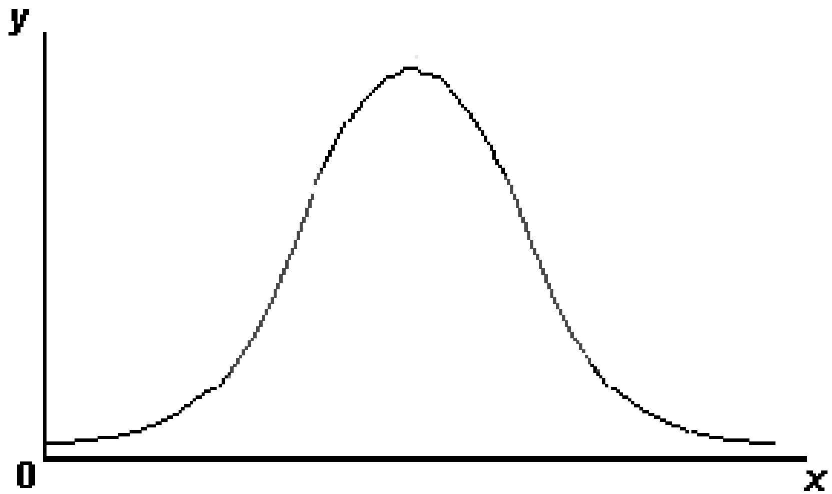
- Standard deviation is equal to the square root of the quantity of the sum of the deviation scores squared divided by the number of scores in a dataset

$$\sigma = \sqrt{\frac{\sum (F-\mu)^2}{N}}$$

  - *F* is a score in a dataset (i.e. any of the ten scores)
  - *μ* is the mean score (i.e. 7)
  - *N* is the number of scores under consideration (i.e. 10)
  - Std dev in our example of test results is 1.687

Measure of dispersion: std dev

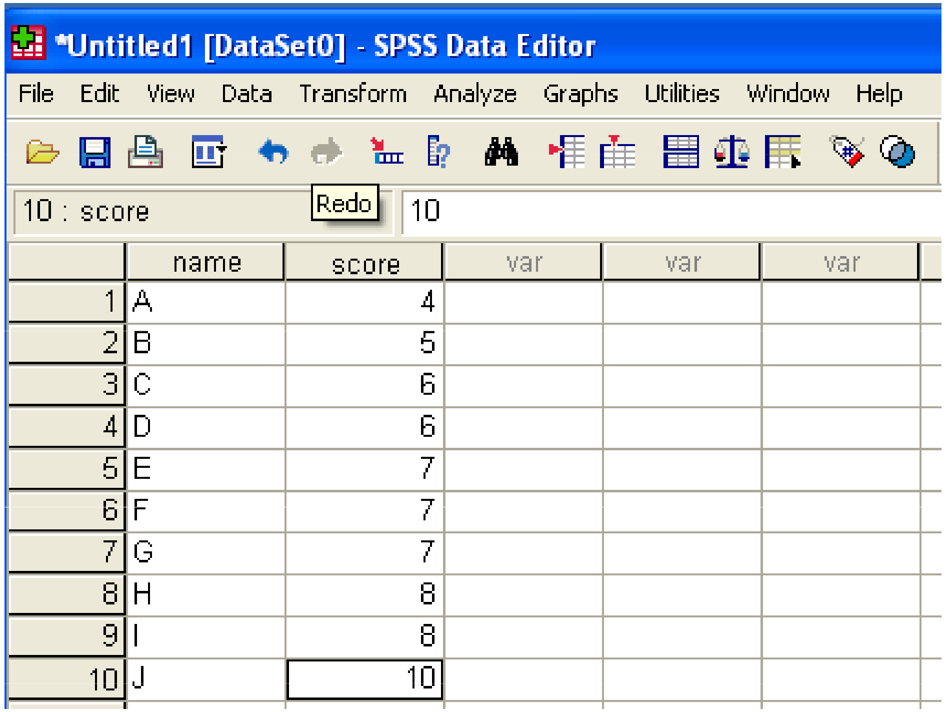
- For a normally distributed dataset (i.e. where most of the items are clustered towards the centre rather than the lower or higher end of the scale)
  - 68% of the scores lie within one standard deviation of the mean
  - 95% lie within two standard deviations of the mean
  - 99.7% lie within three standard deviations of the mean
- The standard deviation is the most reasonable measure of the dispersion of a dataset



©Copyright Robert Niles, http://www.robertniles.com/stats/stdev.shtml

Normal distribution
(bell-shaped curve)

Computing std dev with SPSS



SPSS Menu - Analyze – Descriptive statistics - Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
score	10	4	10	6.80	1.687
Valid N (listwise)	10				

Inferential statistics

- Descriptive statistics are useful in summarizing a dataset
- Inferential statistics are typically used to formulate or test a hypothesis
  - Using statistical measures to test whether or not any differences observed are statistically significant
- Tests of statistical significance
  - chi-square test
  - log-likelihood (LL) test
  - Fisher's Exact test
- Collocation statistics
  - Mutual information (MI)
  - *z* score

Statistical significance

- In testing a linguistic hypothesis, it would be nice to be 100% sure that the hypothesis can be accepted
- However, one can never be 100% sure in real life cases
  - There is always the possibility that the differences observed between two corpora have been due to chance
    - In our swearword example, it is 4 times as frequent in speech as in writing
    - We need to use a statistical test to help us to decide whether this difference is statistically significant
- The level of statistical significance = the level of our confidence in accepting a given hypothesis
  - The closer the likelihood is to 100%, the more confident we can be
    - One must be more than 95% confident that the observed differences have **not** arisen by chance

Commonly used statistical tests

- Chi square test
  - …compares the difference between **the observed values** (e.g. the actual frequencies extracted from corpora) and **the expected values** (e.g. the frequencies that one would expect if no factor other than chance was affecting the frequencies)
- Log likelihood test (LL)
  - Similar, but more reliable as LL does not assume that data is normally distributed
  - The preferred test for statistic significance

Commonly used statistical tests

- Interpreting results
  - The greater the difference (absolute value) between the observed values and the expected values, the less likely it is that the difference is due to chance; conversely, the closer the observed values are to the expected values, the more likely it is that the difference has arisen by chance
  - A probability value *p* close to 0 indicates that a difference is highly significant statistically; a value close to 1 indicates that a difference is almost certainly due to chance
  - **By convention, the general practice is that a hypothesis can be accepted only when the level of significance is less than 0.05 (i.e. $p<0.05$, or more than 95% confident)**

Online LL calculator

- http://ucrel.lancs.ac.uk/llwizard.html

	Corpus 1	Corpus 2
Frequency of word	250	500
Corpus size	10000000	90000000

Item	O1	%1	O2	%2		LL
Word	250	0.00	500	0.00	+	301.88

How to find the probability value p for an LL score of 301.88?

Contingency table

	right-handed	left-handed	TOTAL
male	43	9	52
female	44	4	48
TOTAL	87	13	100

degree of freedom (d.f.) = (No. of row -1) * (No. of column - 1)
= (2 - 1) * (2 – 1) =1 * 1 = 1

Critical values

d.f.	0.10	0.05	0.025	0.01	0.001
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588

The chi square test or LL test score must be greater than 3.84 (1 d.f.) for a difference to be statistically significant.
Oakes, M (1998) *Statistics for Corpus Linguistics*, EUP, p. 266

In the example of swearword in spoken/written BNC, LL 301.88 for 1 d.f.
More than 99.99% confident that the difference is statistically significant

Excel LL calculator by Xu

Log-likelihood Ratio Calculator

Step 1. Enter the **corpus sizes** in **A** and **B**.
Step 2. Enter the **frequency counts** in columns B and C.
* The white cells are data cells; the gray ones are result cells.

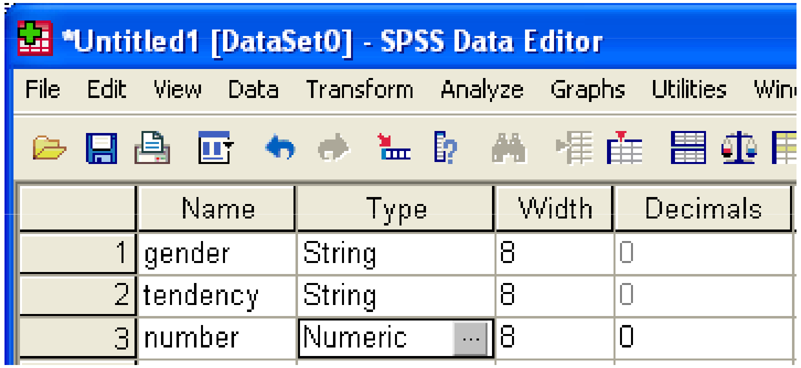
	A		B
Corpus Size 1	52191	Corpus Size 2	52877

Word	Freq. in Corpus 1	Freq. in Corpus 2	Log-likelihood	Sig.		
will	224	138	21.77	0.000	***	+
can	198	192	0.19	0.665		+
would	169	125	7.20	0.007	**	+
could	72	66	0.35	0.557		+
must	67	30	14.96	0.000	***	+
have to	132	41	51.56	0.000	***	+
should	130	55	32.29	0.000	***	+
may	51	35	3.21	0.073		+
might	67	8	53.82	0.000	***	+
ought to	10	3	4.07	0.044	*	+
shall	5	2	1.37	0.242		+

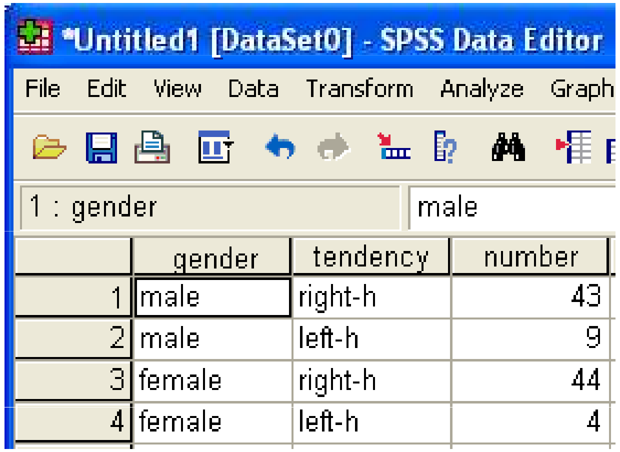
www.corpus4u.org/attachment.php?attachmentid=560&d=1240826440

SPSS: Left- vs. right-handed

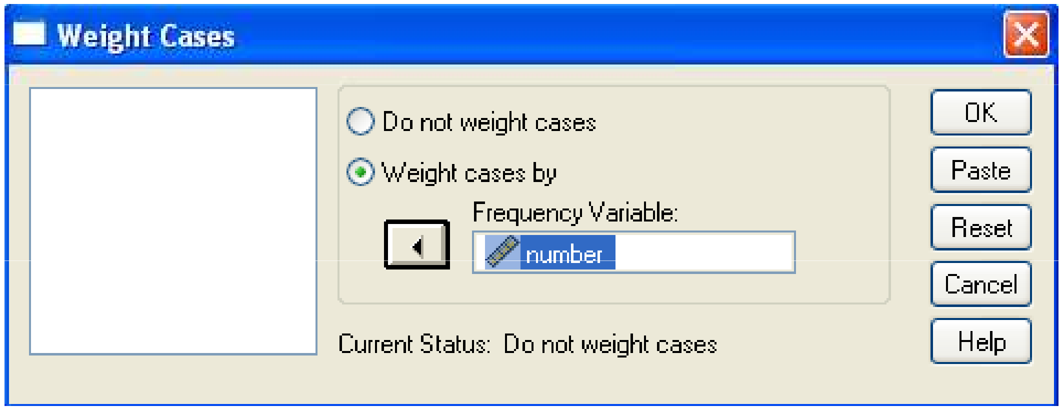
Define variables



Data view



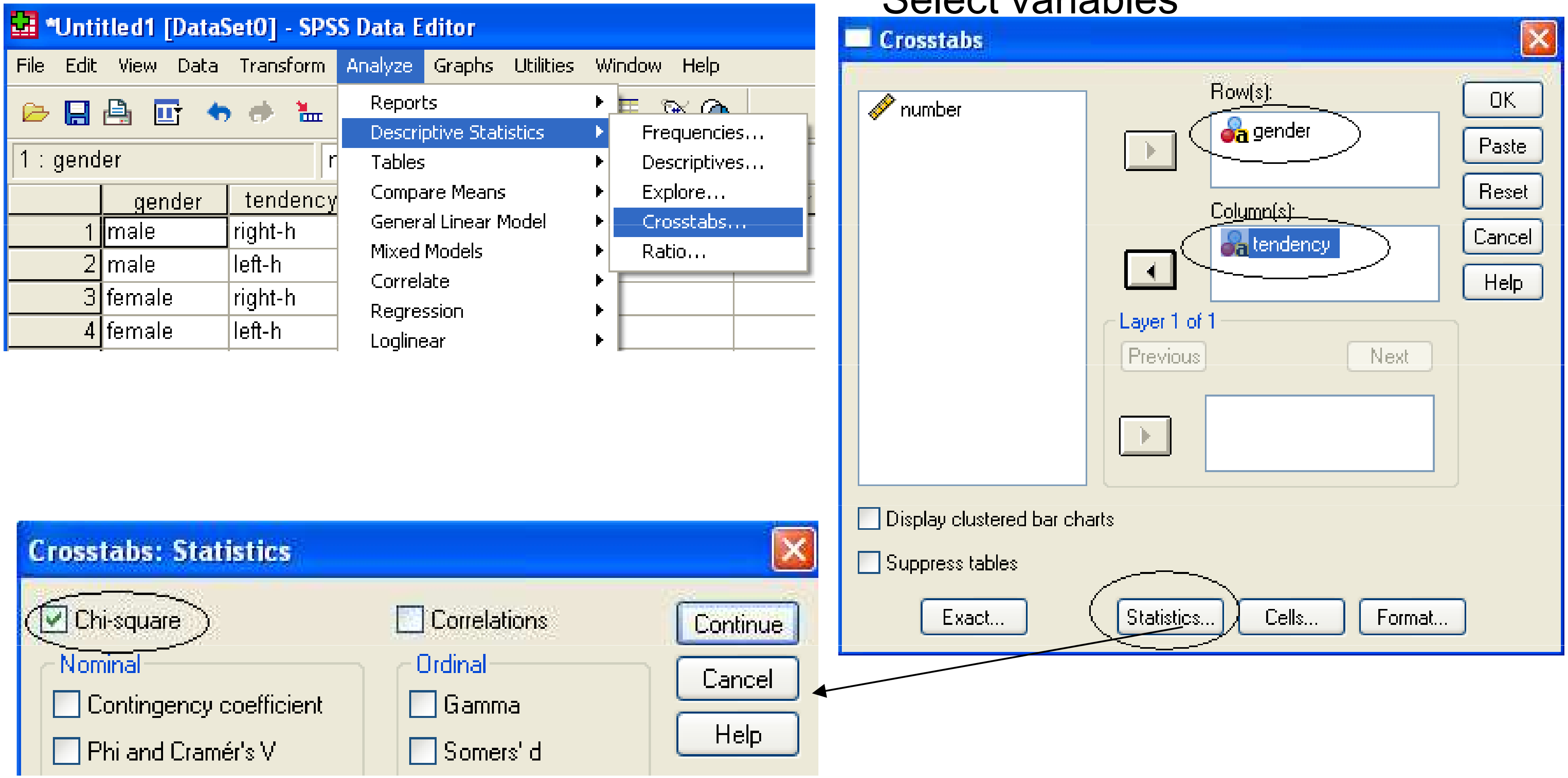
weight case (Data – Weight cases)



SPSS: Left- vs. right-handed

Cross-tab

Select variables



SPSS: Left- vs. right-handed

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.777[b]	1	.182		
Continuity Correction[a]	1.072	1	.300		
Likelihood Ratio	1.825	1	.177		
Fisher's Exact Test				.239	.150
N of Valid Cases	100				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.24.

Any cells with an expected value less than 5?

Critical value (X^2 / LL) for 1 d.f. at $p<0.05$ (95%): 3.84

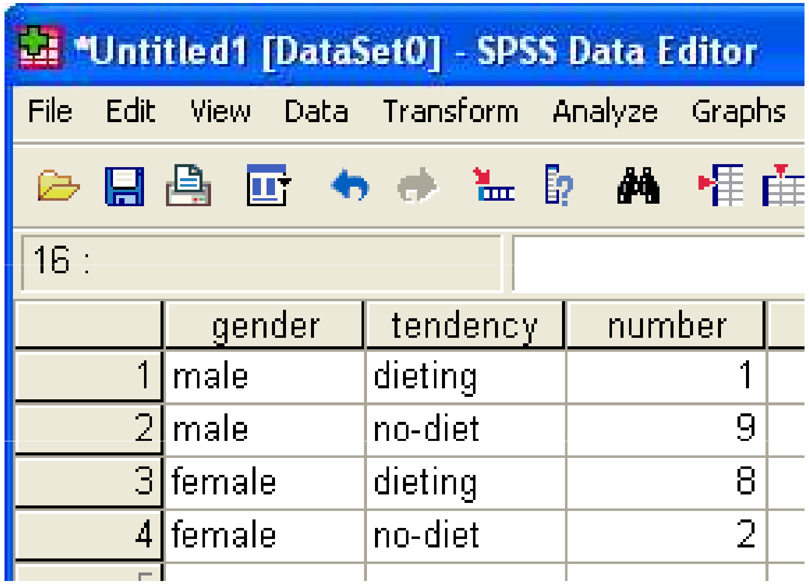
Is there a relationship between gender and left- or right-handedness?

Fisher’s Exact test

- The chi-square or log-likelihood test may not be reliable with **very low frequencies**
  - When a cell in a contingency table has an **expected value** less than 5, Fisher’s Exact test is more reliable
  - In this case, SPSS computes Fisher’s exact significance level automatically when the chi-square test is selected
    - SPSS Releases 15 and 16 have removed the Fisher’s Exact test module, which can be purchased separately

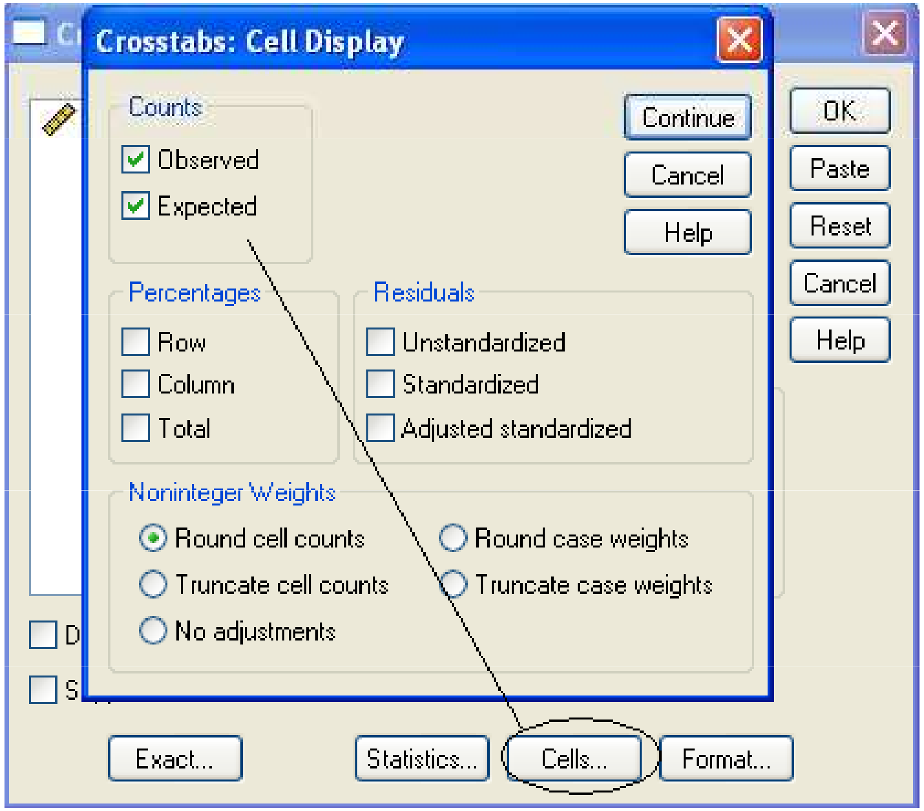
Fisher’s Exact test

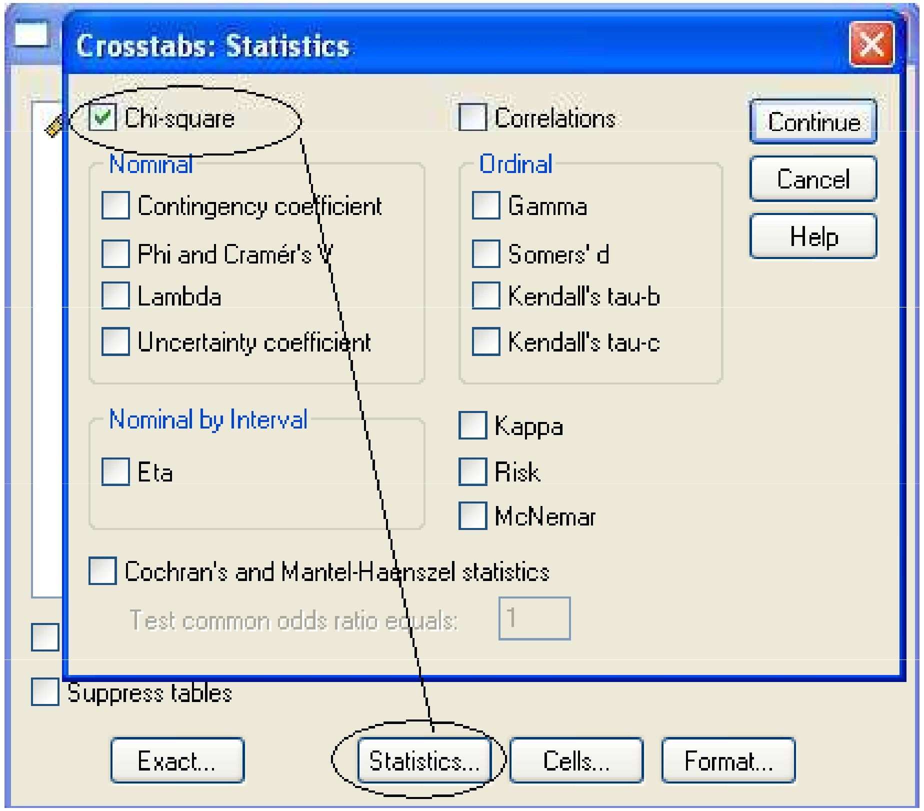
	men	women	TOTAL
dieting	1	8	9
no dieting	9	2	11
TOTAL	10	10	20



Don't forget to weight cases!

Fisher's Exact test





Fisher's Exact test

gender * tendency Crosstabulation

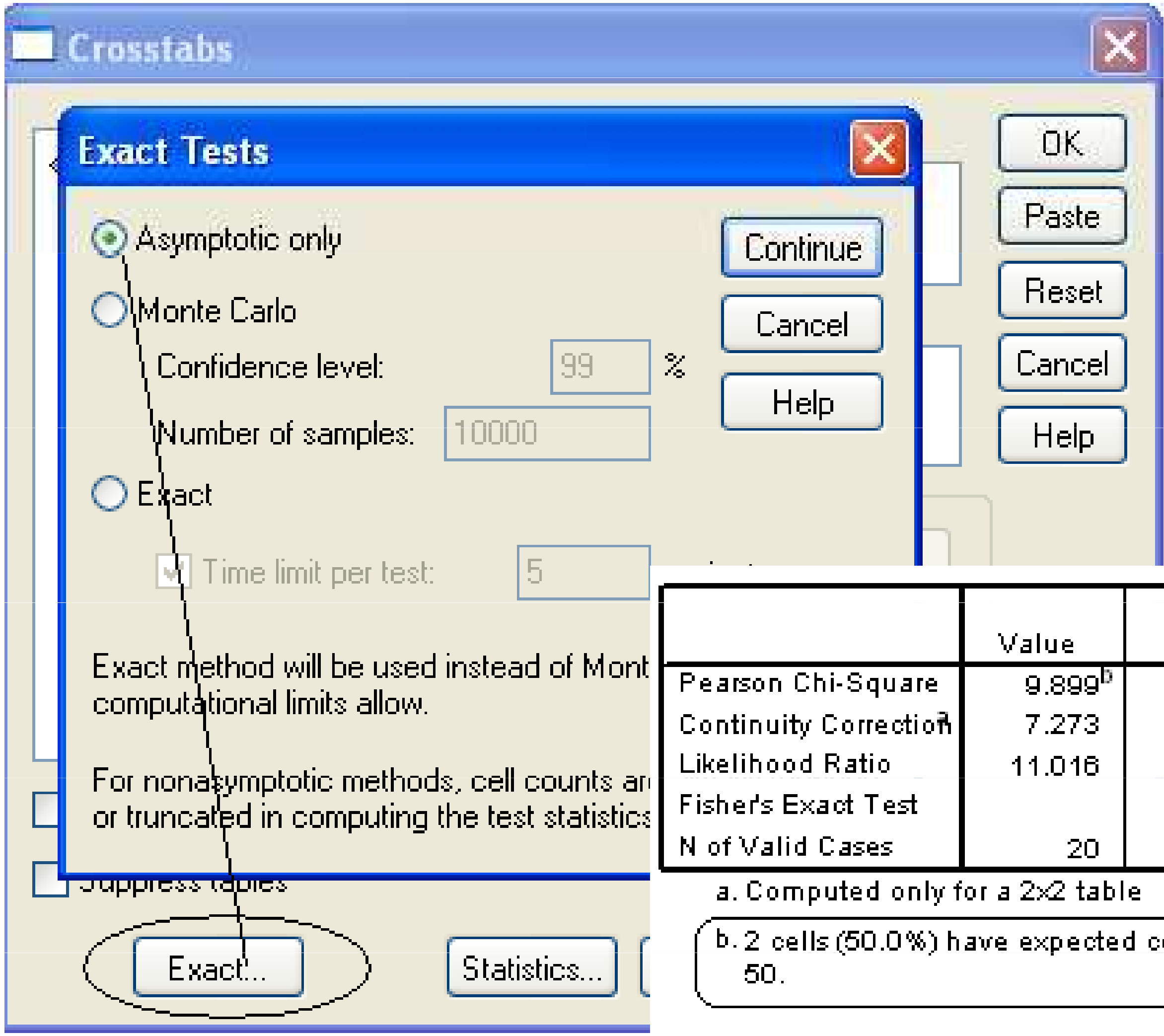
			tendency		Total
			dieting	no-diet	
gender	female	Count	8	2	10
		Expected Count	4.5	5.5	10.0
	male	Count	1	9	10
		Expected Count	4.5	5.5	10.0
Total		Count	9	11	20
		Expected Count	9.0	11.0	20.0

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	9.899[b]	1	.002		
Continuity Correction[a]	7.273	1	.007		
Likelihood Ratio	11.016	1	.001		
Fisher's Exact Test				.005	.003
N of Valid Cases	20				

a. Computed only for a 2x2 table

b. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 4.50.

Force an FE test



	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	9.899[b]	1	.002		
Continuity Correction[a]	7.273	1	.007		
Likelihood Ratio	11.016	1	.001		
Fisher's Exact Test				.005	.003
N of Valid Cases	20				

a. Computed only for a 2x2 table

b. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 4.50.

Practice

- Use both the UCREL/Xu’s LL calculator / SPSS to determine if the difference in the frequencies of passives in the CLEC and LOCNESS corpora is statistically significant
  - CLEC: 7,911 instances in 1,070,602 words
  - LOCNESS: 5,465 instances in 324,304 words

	Corpus 1	Corpus 2
Frequency of word	7911	5465
Corpus size	1070602	324304

Calculate LL   Clear form

Item	01	%1	02	%2	LL
Word	7911	0.74	5465	1.69	- 2039.12

Log-likelihood Ratio Calculator

Step 1. Enter the corpus sizes in A and B.
Step 2. Enter the frequency counts in columns B and C.
* The white cells are data cells; the gray ones are result cells.

	A		B
Corpus Size 1	1070602	Corpus Size 2	324304

Word	Freq. in Corpus 1	Freq. in Corpus 2	Log-likelihood	Sig.
passives	7911	5465	2039.12	0.000 *** -

*Untitled1 [DataSet0] - SPSS Data Editor

File Edit View Data Transform Analyze Graph

16 : corpus

	corpus	data	frequency
1	CLEC	passives	7911
2	CLEC	words	1070602
3	LOCNESS	passives	5465
4	LOCNESS	words	324304

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2290.475[b]	1	.000		
Continuity Correction[a]	2289.494	1	.000		
Likelihood Ratio	2017.086	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	1408282				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 3132.18.

Collocation statistics

- ***Collocation:*** the habitual or characteristic co-occurrence patterns of words
  - Can be identified using a statistical approach in CL, e.g.
    - Mutual Information (MI), *t* test, *z* score
  - Can be computed using tools like SPSS, Wordsmith, AntConc, Xaira
  - Only a brief introduction here
    - More discussions of collocation statistics to be followed

Mutual information

- Computed by dividing the observed frequency of the co-occurring word in the defined span for the search string (so-called ***node word***), e.g. a 4:4 window, by the expected frequency of the co-occurring word in that span and then taking the logarithm to the base 2 of the result

Mutual information

- A measure of collocational strength
- The higher the MI score, the stronger the link between two items
  - MI score of 3.0 or higher to be taken as evidence that two items are collocates
- The closer to 0 the MI score gets, the more likely it is that the two items co-occur by chance
- A negative MI score indicates that the two items tend to shun each other

The *t* test

- Computed by subtracting the expected frequency from the observed frequency and then dividing the result by the standard deviation
- A *t* score of 2 or higher is normally considered to be statistically significant
- The specific probability level can be looked up in a table of *t* distribution

The *z* score

- The *z* score is the number of standard deviations from the mean frequency
- The *z* test compares the observed frequency with the frequency expected if only chance is affecting the distribution
- A higher *z* score indicates a greater degree of collocability of an item with the node word