Analysis of categorical data

S3

Michael Hauptmann
Netherlands Cancer Institute
Amsterdam, The Netherlands
m.hauptmann@nki.nl

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Categorical data

- One-way contingency table = frequency table

	Frequency (%)
Blue	14 (34)
Yellow	22 (54)
Green	5 (12)
Total	41 (100)

- Two-way contingency tables = cross tabs

  - 2-by-2

	Exposed		
Vital status	Yes	No	
Deceased	18	4	22
Nondeceased	18	365	383
	36	369	405

  - r-by-c

	Taste			
Age	Mild	Medium	Strong	Total
Young	9	14	3	26
Old	4	17	12	33
	13	31	15	59

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Which test to use when?

Groups	Outcome data scale: Binomial	Nominal	Ordinal (Continuous)
Single sample	Binomial	χ^2 goodness of fit	
Paired samples	McNemar Sign		Wilcoxon Signed Rank One-sample T
2 independent samples	Fisher's Exact Pearson's χ^2 Cochran-Mantel-Haenszel (stratified)	Fisher's Exact Pearson's χ^2	Wilcoxon Mann Whitney Two-sample T
K independent unordered samples	Fisher's Exact Pearson's χ^2	Fisher's Exact Pearson's χ^2	ANOVA Kruskal-Wallis
K independent ordered samples	Linear-by-linear		Linear-by-linear

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Anaplastic Large-Cell Lymphoma in Women With Breast Implants

Daphne de Jong, MD, PhD
Wies L. E. Vasmel, MD, PhD
Jan Paul de Boer, MD, PhD
Gideon Verhave, MD
Ellis Barbé, MD
Mariel K. Casparie, MD, PhD
Flora E. van Leeuwen, PhD

Context Recently, we identified 2 patients with anaplastic large T-cell lymphoma (ALCL) negative for tyrosine kinase anaplastic lymphoma kinase (ALK-negative) in the fibrous capsule of silicone breast prostheses, placed for cosmetic reasons. Similar cases have been reported in the literature. Although an increased risk of ALCL in patients with breast prostheses has been speculated, no studies have been conducted so far.

Objective To determine whether ALCL risk is associated with breast prostheses.

Design A search for all patients with lymphoma in the breast diagnosed in the Netherlands between 1990 and 2006 was performed through the population-based nationwide pathology database. Subsequently, we performed an individually matched

- De Jong et al., *JAMA* 2008
- Observation of 2 pts with ALCL in fibrous capsule of saline-filled silicone breast implant, some similar cases described in literature
- Question: Do breast implants cause ALCL?
- What type of study would you do to investigate this?

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Answer: Case-control study

- Identified all 429 pts with biopsy-proven primary NHL of the breast in 1990–2006 in NL from PALGA
- 11/389 female pts had ALCL
- 11 subjects with other lymphomas in the breast as controls
- Medical records obtained for all cases & controls
- Presence of breast implant asked by letter to treating physician

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Observed contingency table

Breast implant	Cases	Controls	
Yes	5	1	6
No	6	10	16
	11	11	22

What is the hypothesis (and alternative) of interest here?

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Observed contingency table

Breast implant	Cases	Controls	
Yes	5	1	6
No	6	10	16
	11	11	22

H_0: % pts w/ breast implants equal among cases & controls
H_1: proportions not equal

equivalently:

H_0: no association between breast implants & case status
H_1: association

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Proportion of implants under the null?

Breast implant	Cases	Controls	
Yes	5	1	6
No	6	10	16
	11	11	22

6/22=.27 (27%)

Confidence interval around a single proportion

$$\left[p - z_{\alpha/2}\sqrt{p(1-p)/n}, \quad p + z_{\alpha/2}\sqrt{p(1-p)/n}\right]$$

$$\Rightarrow \left[.27 - 1.96\sqrt{.27(1-.27)/22}, \quad .27 + 1.96\sqrt{.27(1-.27)/22}\right]$$

$$\Rightarrow [.08, \quad .46]$$

(for large n, and p not near zero or unity)

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

CI around single proportion in SPSS[1]

1. CI around single proportion
   Click Analyze – Descriptive Statistics – Frequencies and select variable `implant`. Click Transform – Compute Variable and fill in an arbitrary new variable name with value 1. Click Analyze – Descriptive Statistics – Ratio and select variable `implant` and the new variable. Click Statistics and request confidence intervals.

```
FREQUENCIES VARIABLES=implant
  /ORDER=ANALYSIS.
COMPUTE one=1.
EXECUTE.
RATIO STATISTICS implant WITH one
  /MISSING=EXCLUDE
  /PRINT=CIN(95) MEAN.
```

Dataset `alcl_small_casecontrol.sav`

implant

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	16	72.7	72.7	72.7
	1.00	6	27.3	27.3	100.0
	Total	22	100.0	100.0	

Ratio Statistics for implant / one

Mean		.273
95% Confidence Interval for Mean	Lower Bound	.071
	Upper Bound	.475

The confidence intervals are constructed by assuming a Normal distribution for the ratios.

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Back to hypothesis: Pearson chi-square test

Compare observed table w/ expected table under the null

Breast implant	Observed Cases	Observed Controls	Expected Cases	Expected Controls	
Yes	5	1	11*6/22=3	11*6/22=3	6
No	6	10	11*16/22=8	11*16/22=8	16
	11	11	11	11	22

$$\sum_{\text{cells}} \frac{(|O-E|-1/2)^2}{E} > \chi^2_{(\#\text{rows}-1)*(\#\text{columns}-1),.95}$$

where O & E are observed & expected frequencies, respectively, for each cell in 2×2 contingency table; 1/2 is continuity or Yates' correction (little difference unless $n < 40$ or Es very small)

Here:

$(|5-3|-.5)^2/3+(|1-3|-.5)^2/3+(|6-8|-.5)^2/8+(|10-8|-.5)^2/8 = 2.06$

at (2–1)*(2–1)=1 degrees of freedom

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Pearson chi-square in SPSS[2]

case * implant Crosstabulation

Count

		implant		
		.00	1.00	Total
case	.00	10	1	11
	1.00	6	5	11
Total		16	6	22

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for case (.00 / 1.00)	8.333	.776	89.470
For cohort implant = .00	1.667	.942	2.950
For cohort implant = 1.00	.200	.028	1.445
N of Valid Cases	22		

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.667[a]	1	.056		
Continuity Correction[b]	2.063	1	.151		
Likelihood Ratio	3.922	1	.048		
Fisher's Exact Test				.149	.074
Linear-by-Linear Association	3.500	1	.061		
N of Valid Cases	22				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 3.00.

b. Computed only for a 2x2 table

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Chi-square tests

- Pearson: based on O-E under independence
- Goodness of fit: same as Pearson, but can be used with E based on specific distribution other than independence
- Can all be extended to r-by-c tables w/ DF=(r–1)*(c–1)
- Chi-square tests are asymptotic tests (for large N)
  for small N: Fisher's exact test

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

An example of Fisher's exact test

Observed table		
7	2	9
5	6	11
12	8	20

Next stronger		
8	1	9
4	7	11
12	8	20

Strongest table		
9	0	9
3	8	11
12	8	20

- Obtain tables stronger than the observed table by reducing the cell with the lowest count by 1 in steps

- Compute the probability for each table $P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!}$

$$\left.\begin{aligned} P_{\text{observed}} &= \frac{9!11!12!8!}{20!7!2!5!6!} = .132 \\ P_{\text{stronger}} &= \frac{9!11!12!8!}{20!8!1!4!7!} = .024 \\ P_{\text{strongest}} &= \frac{9!11!12!8!}{20!9!0!3!8!} = .001 \end{aligned}\right\} P_{\text{total (one-tailed)}} = .157$$

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Interpretation of Fisher's exact test

- $p = .157$, i.e., there is a 15.7% chance under the null that, given the sample size and the margins, we would get a table as strong or stronger as the observed table by chance alone

- At $\alpha = .05$, distribution in observed table is not significantly different from independence

- Possible for r-by-c tables

- 2-tailed p is sum of probabilities of all tables with p equal or less than the observed p

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Is a 2-by-2 analysis sufficient? Confounding?

- A variable correlated with the variable of interest (e.g., exposure or treatment) and with the outcome is a potential confounder

- E.g., age and calendar year of diagnosis are potential confounders of the association between breast implants & ALCL

  1. Prevalence of breast implants is increasing with calendar year

  2. ALCL is more common at older age

  3. ALCL cases have fewer breast implants than random sample of non-ALCL cases: apparent protective effect, but due to confounding bias

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

ALCL study: matching to control confounding

- 11 controls matched to cases by age at DX (within 5 yrs) & yr of DX (within 2 yrs)

- More specifically: for each case, one woman is randomly chosen from all women with an age at DX within 5 yrs of that of the case & diagnosed within 2 yrs of the case

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

McNemar test for dependent samples

- 2 dependent samples
  - Before-after or other repeated measure of same units
  - Matched-pairs studies with similar units measured at different times

- Depends only on number of discordant pairs (diagonal)

	After	
Before	Exposed	Nonexposed
Exposed	a	b
Nonexposed	c	d

	Control	
Case	Exp.	Nonexp.
Exp.	a	b
Nonexp.	c	d

- χ^2=(|b-c|-1)^2/(b+c) at 1 DF (continuity-corrected)

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

- Discordant pairs split evenly $\rightarrow$ evidence that overall proportion about the same for both raters

- Discordant pairs skewed in one direction (e.g., more yes/no than no/yes) $\rightarrow$ evidence that overall proportion of yeses higher for one rater than the other

- Multivariate analogon to McNemar is conditional logistic regression

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

ALCL study: McNemar test for 1:1 matched design[3]

Data set `alcl_one_to_one_matched.sav`

	implant_cases	implant_controls
1	1	0
2	1	0
3	1	0
4	1	0
5	1	0
6	0	1
7	0	0
8	0	0
9	0	0
10	0	0
11	0	0

implant_controls & implant_cases

implant_controls	implant_cases 0	implant_cases 1
0	5	5
1	1	0

Test Statistics[b]

	implant_controls & implant_cases
N	11
Exact Sig. (2-tailed)	.219[a]
Exact Sig. (1-tailed)	.109
Point Probability	.094

a. Binomial distribution used.

b. McNemar Test

6 discordant pairs, under H_0: 3 in each direction, observed: 5 vs. 1

Binomial point probability: P(X=5)=.094

Two-sided p=2*[.094+P(X=6)]=.219 $\Rightarrow$ P(X=6)=.0155

Alternatively: $(|5-1|-1)^2/(5+1) = 9/6 = 1.5$ at χ^2_1

$\Rightarrow$ p=.22067

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Exact vs. asymptotic procedures

- Most standard statistical tests assume test statistic is asymptotically normally distributed (large N)
- May not be true for small studies
- Exact tests based on permutation or Monte Carlo simulation (resampling)
- Exact p-values smaller or larger than asymptotic p-values
- Included in most statistical software packages

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

What they really did: individually matched case-control study

- 35 controls from 389 pts with other lymphomas in the breast

- Matched by age at DX (within 5 yrs) and yr of DX (within 2 yrs)

Breast implant	Cases	Controls	
Yes	5	1	6
No	6	34	40
	11	35	46

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

- McNemar only for matched pairs (1:1)

- Standard 2-by-2 analysis[4]
  Crude OR=28.3, asymptotic 95% CI 2.8–287.1, p=.002

- Conditional logistic regression (asymptotic) for >1 controls per case via Cox regression[5]
  OR=18.2, 95% CI 2.1–156.8, p=.008 (reported)

- Matching is controlling for confounding at the design stage

- Confounding can also be controlled, however less efficiently, at the analysis stage: Cochran-Mantel-Haenszel test or adjustment in regression models

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Updated results 2018

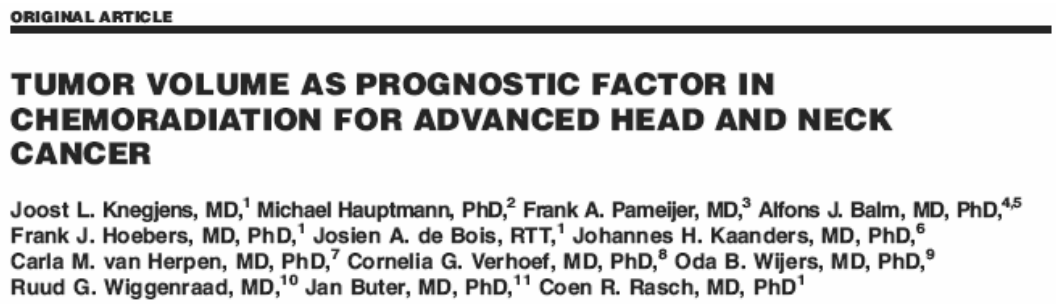
JAMA Oncology | Original Investigation

Breast Implants and the Risk of Anaplastic Large-Cell Lymphoma in the Breast

Mintsje de Boer, MD; Flora E. van Leeuwen, PhD; Michael Hauptmann, PhD; Lucy I. H. Overbeek, PhD; Jan Paul de Boer, MD, PhD; Nathalie J. Hijmering, MSc; Arthur Sernee, MSc; Caroline A. H. Klazen, MD, PhD; Marc B. I. Lobbes, MD, PhD; René R. W. J. van der Hulst, MD, PhD; Hinne A. Rakhorst, MD, PhD; Daphne de Jong, MD, PhD

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Tumor volume study



ORIGINAL ARTICLE

TUMOR VOLUME AS PROGNOSTIC FACTOR IN CHEMORADIATION FOR ADVANCED HEAD AND NECK CANCER

Joost L. Knegjens, MD,[1] Michael Hauptmann, PhD,[2] Frank A. Pameijer, MD,[3] Alfons J. Balm, MD, PhD,[4,5] Frank J. Hoebers, MD, PhD,[1] Josien A. de Bois, RTT,[1] Johannes H. Kaanders, MD, PhD,[6] Carla M. van Herpen, MD, PhD,[7] Cornelia G. Verhoef, MD, PhD,[8] Oda B. Wijers, MD, PhD,[9] Ruud G. Wiggenraad, MD,[10] Jan Buter, MD, PhD,[11] Coen R. Rasch, MD, PhD[1]

- 360 pts w/ advanced head & neck squamous cell carcinoma (oral cavity, oropharynx, hypopharynx)
- Diagnosed 1997–2006 in 6 hospitals
- Chemoradiation for functionally/anatomically unresectable disease
- Primary tumor volume measured by pretreatment MRI or CT
- 72% male, mean age 56 yrs (range, 25–85)
- Followed for local recurrence (median, 20 months)

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Data set `tumorvolume.sav`

- Includes several transformations of original variables based on code in `transformations_syntax.txt`

- Codebook `labels.doc` describes variable labels

- Categorizing continuous volume
  `volumeg4=1+(Tumor_volume>20)+(Tumor_volume>40)+(Tumor_volume>60)`
  SPSS: Click Transform – Compute Variable

volumeg4	Tumor_volume
1	$\leq$20
2	>20-$\leq$40
3	>40-$\leq$60
4	>60

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Cochran-Mantel-Haenszel test

- Test association in stratified 2-by-2 tables, i.e., is there association between volume & gender stratified by tumor site

- $\frac{\sum_i a_i * d_i / N_i}{\sum_i b_i * c_i / N_i} = \sum \frac{b_i * c_i / N_i}{\sum b_i * c_i / N_i} \times \frac{a_i * d_i}{b_i * c_i}$

- Stratified OR=weighted mean, weights approximate the inverse variance of individual ORs if near 1 (stratified Pearson)

- Usually preceded by Breslow-Day test of homogeneity of ORs across strata (but limited power unless sample size in all strata is large & heterogeneity is substantial)

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Association between volume & gender stratified by tumor site

- Overall odds ratio[6]

volumeg2 * Gender Crosstabulation

Count

		Gender		
		0	1	Total
volumeg2	1.00	195	84	279
	2.00	67	17	84
Total		262	101	363

- Pearson χ^2 p=.077, OR=.589

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

- Test of homogeneity, common OR & CMH test[7]

volumeg2 * Gender * Tumor_site Crosstabulation

Count

Tumor_site			Gender 0	Gender 1	Total
1	volumeg2	1.00	42	19	61
		2.00	14	6	20
	Total		56	25	81
2	volumeg2	1.00	124	51	175
		2.00	40	11	51
	Total		164	62	226
3	volumeg2	1.00	29	14	43
		2.00	13	0	13
	Total		42	14	56

Risk Estimate

Tumor_site		Value	95% Confidence Interval Lower	95% Confidence Interval Upper
1	Odds Ratio for volumeg2 (1.00 / 2.00)	.947	.316	2.844
	For cohort Gender = 0	.984	.705	1.372
	For cohort Gender = 1	1.038	.482	2.234
	N of Valid Cases	81		
2	Odds Ratio for volumeg2 (1.00 / 2.00)	.669	.318	1.405
	For cohort Gender = 0	.903	.760	1.073
	For cohort Gender = 1	1.351	.763	2.394
	N of Valid Cases	226		
3	For cohort Gender = 0	.674	.548	.830
	N of Valid Cases	56		

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	4.334	2	.115
Tarone's	4.334	2	.115

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	3.177	1	.075
Mantel-Haenszel	2.679	1	.102

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

- Common OR=.588 $\Rightarrow$ no confounding by tumor site

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Chi-square goodness of fit test with specified distribution

- Question: Is infusion side (single- vs. double-sided) distributed 50-50?

- Provide SPSS with expected proportions: 50-50[8]

Infusion_side

	Observed N	Expected N	Residual
1	170	180.0	-10.0
2	190	180.0	10.0
Total	360		

Test Statistics

	Infusion_side
Chi-Square	1.111[a]
df	1
Asymp. Sig.	.292

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 180.0.

- p=.292: data are consistent with 50-50 split

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

New question: is pre-treatment health status of NKI pts better vs. pts treated elsewhere?

- Variable `who` represents general health status of pts in 3 ordered categories (0, 1, 2)
- `nki`=1 for NKI pts, 0 otherwise
- Question rephrased: Does prevalence of NKI-pts change (increase/decrease) with WHO status?
- Is Pearson's χ^2 test appropriate?

WHO * nki Crosstabulation

			nki		Total
			.00	1.00	
WHO	0	Count	65	61	126
		% within WHO	51.6%	48.4%	100.0%
	1	Count	41	61	102
		% within WHO	40.2%	59.8%	100.0%
	2	Count	3	7	10
		% within WHO	30.0%	70.0%	100.0%
Total		Count	109	129	238
		% within WHO	45.8%	54.2%	100.0%

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Linear-by-linear association test[9]

- Scores 0-1-2: more power to detect ordered alternatives vs. Pearson
- Choice of scores: group mean or median (numeric variables), meaningful scores (ordinal variables)
- Null hypothesis: no trend, i.e., binomial proportion the same for all levels of explanatory variable
- Sign of standardized test statistic: increasing (decreasing) trend

WHO * nki Crosstabulation

			nki		Total
			.00	1.00	
WHO	0	Count	65	61	126
		% within WHO	51.6%	48.4%	100.0%
	1	Count	41	61	102
		% within WHO	40.2%	59.8%	100.0%
	2	Count	3	7	10
		% within WHO	30.0%	70.0%	100.0%
Total		Count	109	129	238
		% within WHO	45.8%	54.2%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	3.996[a]	2	.136	.141		
Likelihood Ratio	4.037	2	.133	.145		
Fisher's Exact Test	3.913			.137		
Linear-by-Linear Association	3.975[b]	1	.046	.056	.029	.012
N of Valid Cases	238					

a. 1 cells (16.7%) have expected count less than 5. The minimum expected count is 4.58.

b. The standardized statistic is 1.994.

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Association between tumor-volume & N-stage?

- N-stage as row scores & tumor volume categories as column scores[10]

volumeg4 * N_stage Crosstabulation

Count

		N_stage				Total
		0	1	2	3	
volumeg4	1.00	28	18	60	11	117
	2.00	31	15	71	9	126
	3.00	7	12	36	3	58
	4.00	7	4	42	8	61
Total		73	49	209	31	362

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16.181[a]	9	.063
Likelihood Ratio	16.897	9	.050
Linear-by-Linear Association	7.002	1	.008
N of Valid Cases	362		

a. 1 cells (6.3%) have expected count less than 5. The minimum expected count is 4.97.

- Trend test more powerful than Pearson χ^2: p=.008 vs. .063

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

- N-stage categories: degree of spread to regional lymph nodes
  N0: tumor cells absent from regional nodes
  N1: regional lymph node metastasis present
  N2: tumor spread to an extent between N1 & N3
  N3: tumor spread to more distant or numerous regional nodes

- Consider N3 much worse than N2 and use score of 10 instead of 3[11]

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16.181[a]	9	.063
Likelihood Ratio	16.897	9	.050
Linear-by-Linear Association	1.702	1	.192
N of Valid Cases	362		

a. 1 cells (6.3%) have expected count less than 5. The minimum expected count is 4.97.

- P=.192: tendency for more higher volume tumors with increasing N stage is not continuing for N3 as it did for N0-N2

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Which test to use when?

Groups	Outcome data scale: Binomial	Nominal	Ordinal (Continuous)
Single sample	Binomial	χ^2 goodness of fit	
Paired samples	McNemar Sign		Wilcoxon Signed Rank One-sample T
2 independent samples	Fisher's Exact Pearson's χ^2 Cochran-Mantel-Haenszel (stratified)	Fisher's Exact Pearson's χ^2	Wilcoxon Mann Whitney Two-sample T
K independent unordered samples	Fisher's Exact Pearson's χ^2	Fisher's Exact Pearson's χ^2	ANOVA Kruskal-Wallis
K independent ordered samples	Linear-by-linear		Linear-by-linear

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

SPSS code (syntax and clicking)

1. CI around single proportion
   Click Analyze – Descriptive Statistics – Frequencies and select variable `implant`. Click Transform – Compute Variable and fill in an arbitrary new variable name with value 1. Click Analyze – Descriptive Statistics – Ratio and select variable `implant` and the new variable. Click Statistics and request confidence intervals.

```
FREQUENCIES VARIABLES=implant
  /ORDER=ANALYSIS.
COMPUTE one=1.
EXECUTE.
RATIO STATISTICS implant WITH one
  /MISSING=EXCLUDE
  /PRINT=CIN(95) MEAN.
```


NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

2. Pearson chi-square
   Click Analyze – Descriptive Statistics – Crosstabs and select variables `case` and `implant`.

```
CROSSTABS
  /TABLES=case BY implant
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ RISK
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

3. McNemar test for 1:1 matched design
   Click Analyze – Nonparametric Tests – (Legacy Dialogs) – 2 Related Samples and select variables `implant_controls` and `implant_cases` and request McNemar's test. Click Exact and request calculation of an exact p-value.

```
NPAR TESTS
  /MCNEMAR=implant_controls WITH implant_cases (PAIRED)
  /MISSING ANALYSIS
  /METHOD=EXACT TIMER(5).
```


NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

4. Standard 2-by-2 analysis

   Click Analyze – Descriptive Statistics – Crosstabs and select the variables. Click Statistics and request Chi-square and Risk. Click Exact and request exact p-values.

```
CROSSTABS
  /TABLES=implant BY case
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ RISK
  /CELLS=COUNT
  /COUNT ROUND CELL
  /METHOD=MC CIN(99) SAMPLES(10000).
```


NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

5. Conditional logistic regression (asymptotic) for >1 controls per case via Cox regression

   Click Transform – Compute Variable and calculate a new variable `dv` with value "2-case". Click Analyze – Survival – Cox Regression and select variable `dv` as Time and `case` as Status. Define a value of 1 as an event. Select variable `implant` as a covariate and `setno` as a stratum.

```
COMPUTE dv=2-case.
EXECUTE.
COXREG dv
  /STATUS=case(1)
  /STRATA=setno
  /METHOD=ENTER implant
  /PRINT=CI(95)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```


NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

6. Overall odds ratio

   Click Analyze – Descriptive Statistics – Crosstabs and select the variables. Click Statistics and request Chi-square and Risk. Click Exact and request exact p-values.

```
CROSSTABS
  /TABLES=volumeg2 BY Gender
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ RISK
  /CELLS=COUNT
  /COUNT ROUND CELL
  /METHOD=EXACT TIMER(5).
```


NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

7. Test of homogeneity, common OR & CMH test
   As above. Select variable `Tumor_site` as a Layer. Click Statistics and request the Cochran-Mantel-Haenszel test.

```
CROSSTABS
  /TABLES=volumeg2 BY Gender BY Tumor_site
  /FORMAT=AVALUE TABLES
  /STATISTICS=RISK CMH(1)
  /CELLS=COUNT
  /COUNT ROUND CELL
  /METHOD=EXACT TIMER(5).
```

8. Provide SPSS with expected proportions: 50-50
   Click Analyze – Nonparametric Tests – Legacy Dialogs – Chi-square, select variable `Infusion_side` and add Expected Values.

```
NPAR TESTS
  /CHISQUARE=Infusion_side
  /EXPECTED=50 50
  /MISSING ANALYSIS.
```


NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

9. Linear-by-linear association test
   Click Analyze – Descriptive Statistics – Crosstabs, select variables and request Chi-square under Statistics.

```
CROSSTABS
  /TABLES=WHO BY nki
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT ROW
  /COUNT ROUND CELL
  /METHOD=EXACT TIMER(5).
```

10. Linear-by-linear test with N-stage as row scores & tumor volume categories as column scores
    Click Analyze – Descriptive Statistics – Crosstabs, select variables and request Chi-square under Statistics.

```
CROSSTABS
  /TABLES=volumeg4 BY N_stage
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT
  /COUNT ROUND CELL.
```


NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

11. Linear-by-linear test with score of 10 instead of 3 for N3
    Click Transform – Compute Variable and create variable `n_stage_new` as `n_stage*(n_stage<3)+10*(n_stage=3)`. Click Analyze – Descriptive Statistics – Crosstabs, select variables `volumeg4` and `n_stage_new` and request Chi-square under Statistics.

```
COMPUTE n_stage_new=n_stage*(n_stage<3)+10*(n_stage=3).
EXECUTE.
CROSSTABS
  /TABLES=volumeg4 BY n_stage_new
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT
  /COUNT ROUND CELL.
```


NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital