Categorical Data Analysis

CHAPTER 7

Categorical Data Analysis

Independent (Explanatory) Variable is Categorical (Nominal or Ordinal)

Dependent (Response) Variable is Categorical (Nominal or Ordinal)

Special Cases:

- 2x2 (Each variable has 2 levels)
- Nominal/Nominal
- Nominal/Ordinal
- Ordinal/Ordinal

Contingency Tables

Tables representing all combinations of levels of explanatory and response variables

Numbers in table represent **Counts** of the number of cases in each cell

Row and column totals are called **Marginal counts**

Example – EMT Assessment of Kids

Explanatory Variable – Child Age (Infant, Toddler, Pre-school, School-age, Adolescent)

Response Variable – EMT Assessment (Accurate, Inaccurate)

Assessment			
Age	Acc	Inac	**Tot**
Inf	168	73	**241**
Tod	230	73	**303**
Pre	254	53	**307**
Sch	379	58	**437**
Ado	652	124	**776**
Tot	**1683**	**381**	**2064**

Source: Foltin, et al (2002)

2x2 Tables

Each variable has 2 levels

- Explanatory Variable – Groups (Typically based on demographics, exposure, or Trt)
- Response Variable – Outcome (Typically presence or absence of a characteristic)

Measures of association

- Relative Risk (Prospective Studies)
- Odds Ratio (Prospective or Retrospective)
- Absolute Risk (Prospective Studies)

2x2 Tables - Notation

	Outcome Present	**Outcome Absent**	**Group Total**
Group 1	n_{11}	n_{12}	$n_{1.}$
Group 2	n_{21}	n_{22}	$n_{2.}$
Outcome Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Relative Risk

Ratio of the probability that the outcome characteristic is present for one group, relative to the other

Sample proportions with characteristic from groups 1 and 2:

$$\hat{\pi}_1 = \frac{n_{11}}{n_{1.}} \qquad \hat{\pi}_2 = \frac{n_{21}}{n_{2.}}$$

Relative Risk

Estimated Relative Risk:

$$RR = \hat{\pi}_1 / \hat{\pi}_2$$

95% Confidence Interval for Population Relative Risk:

$$\left(RR(e^{-1.96\sqrt{v}}) , RR(e^{1.96\sqrt{v}}) \right)$$

$$e = 2.71828 \quad v = \frac{(1-\hat{\pi}_1)}{n_{11}} + \frac{(1-\hat{\pi}_2)}{n_{21}}$$

Relative Risk

Interpretation

- Conclude that the probability that the outcome is present is higher (in the population) for group 1 if the entire interval is above 1
- Conclude that the probability that the outcome is present is lower (in the population) for group 1 if the entire interval is below 1
- Do not conclude that the probability of the outcome differs for the two groups if the interval contains 1

Example - Coccidioidomycosis and TNFα-antagonists

• Research Question: Risk of developing Coccidioidmycosis associated with arthritis therapy?

• Groups: Patients receiving tumor necrosis factor α (TNFα) versus Patients not receiving TNFα (all patients arthritic)

	COC	No COC	Total
TNFα	7	240	247
Other	4	734	738
Total	11	974	985

Source: Bergstrom, et al (2004)

Example - Coccidioidomycosis and TNFα-antagonists

- Group 1: Patients on TNFα
- Group 2: Patients not on TNFα

$$\hat{\pi}_1 = \frac{7}{247} = .0283 \qquad \hat{\pi}_2 = \frac{4}{738} = .0054$$

$$RR = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{.0283}{.0054} = 5.24 \qquad v = \frac{1-.0283}{7} + \frac{1-.0054}{4} = .3874$$

$$95\%CI: \quad (5.24e^{-1.96\sqrt{.3874}}, 5.24e^{1.96\sqrt{.3874}}) \equiv (1.55, 17.76)$$

Entire CI above 1 $\Rightarrow$ Conclude higher risk if on TNFα

Odds Ratio

Odds of an event is the probability it occurs divided by the probability it does not occur

Odds ratio is the odds of the event for group 1 divided by the odds of the event for group 2

Sample odds of the outcome for each group:

$$odds_1 = \frac{n_{11} / n_{1.}}{n_{12} / n_{1.}} = \frac{n_{11}}{n_{12}}$$

$$odds_2 = \frac{n_{21}}{n_{22}}$$

Odds Ratio

- Estimated Odds Ratio:

$$OR = \frac{odds_1}{odds_2} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

95% Confidence Interval for Population Odds Ratio

$$\left(OR(e^{-1.96\sqrt{v}}),\ OR(e^{1.96\sqrt{v}})\right)$$

$$e = 2.71828 \qquad v = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

Odds Ratio

Interpretation

- Conclude that the probability that the outcome is present is higher (in the population) for group 1 if the entire interval is above 1
- Conclude that the probability that the outcome is present is lower (in the population) for group 1 if the entire interval is below 1
- Do not conclude that the probability of the outcome differs for the two groups if the interval contains 1

Example - NSAIDs and GBM

Case-Control Study (Retrospective)

- Cases: 137 Self-Reporting Patients with Glioblastoma Multiforme (GBM)
- Controls: 401 Population-Based Individuals matched to cases wrt demographic factors

	GBM Present	GBM Absent	Total
NSAID User	32	138	170
NSAID Non-User	105	263	368
Total	137	401	538

Source: Sivak-Sears, et al (2004)

Example - NSAIDs and GBM

$$OR = \frac{32(263)}{138(105)} = \frac{8416}{14490} = 0.58$$

$$v = \frac{1}{32} + \frac{1}{138} + \frac{1}{105} + \frac{1}{263} = 0.0518$$

$$95\%\ CI: \quad (0.58e^{-1.96\sqrt{0.0518}}, 0.58e^{1.96\sqrt{0.0518}}) \equiv (0.37, 0.91)$$

Interval is entirely below 1, NSAID use appears to be lower among cases than controls

Absolute Risk

Difference Between Proportions of outcomes with an outcome characteristic for 2 groups

Sample proportions with characteristic from groups 1 and 2:

$$\hat{\pi}_1 = \frac{n_{11}}{n_{1.}} \qquad \hat{\pi}_2 = \frac{n_{21}}{n_{2.}}$$

Absolute Risk

Estimated Absolute Risk:

$$AR = \hat{\pi}_1 - \hat{\pi}_2$$

95% Confidence Interval for Population Absolute Risk

$$AR \pm 1.96\sqrt{\frac{\hat{\pi}_1\left(1-\hat{\pi}_1\right)}{n_{1.}} + \frac{\hat{\pi}_2\left(1-\hat{\pi}_2\right)}{n_{2.}}}$$

Absolute Risk

Interpretation

- Conclude that the probability that the outcome is present is higher (in the population) for group 1 if the entire interval is positive
- Conclude that the probability that the outcome is present is lower (in the population) for group 1 if the entire interval is negative
- Do not conclude that the probability of the outcome differs for the two groups if the interval contains 0

Example - Coccidioidomycosis and TNFα-antagonists

- Group 1: Patients on TNFα
- Group 2: Patients not on TNFα

$$\hat{\pi}_1 = \frac{7}{247} = .0283 \quad \hat{\pi}_2 = \frac{4}{738} = .0054$$

$$AR = \hat{\pi}_1 - \hat{\pi}_2 = .0283 - .0054 = .0229$$

$$95\%CI: \quad .0229 \pm 1.96\sqrt{\frac{.0283(.9717)}{247} + \frac{.0054(.9946)}{738}}$$

$$\equiv .0229 \pm .0213 \equiv (0.0016\ ,\ 0.0242)$$

Interval is entirely positive, TNFα is associated with higher risk

Fisher's Exact Test

Method of testing for association for 2x2 tables when one or both of the group sample sizes is small

Measures (conditional on the group sizes and number of cases with and without the characteristic) the chances we would see differences of this magnitude or larger in the sample proportions, if there were no differences in the populations

Example – Echinacea Purpurea for Colds

Healthy adults randomized to receive EP ($n_{1.}$=24) or placebo ($n_{2.}$=22, two were dropped)

Among EP subjects, 14 of 24 developed cold after exposure to RV-39 (58%)

Among Placebo subjects, 18 of 22 developed cold after exposure to RV-39 (82%)

Out of a total of 46 subjects, 32 developed cold

Out of a total of 46 subjects, 24 received EP

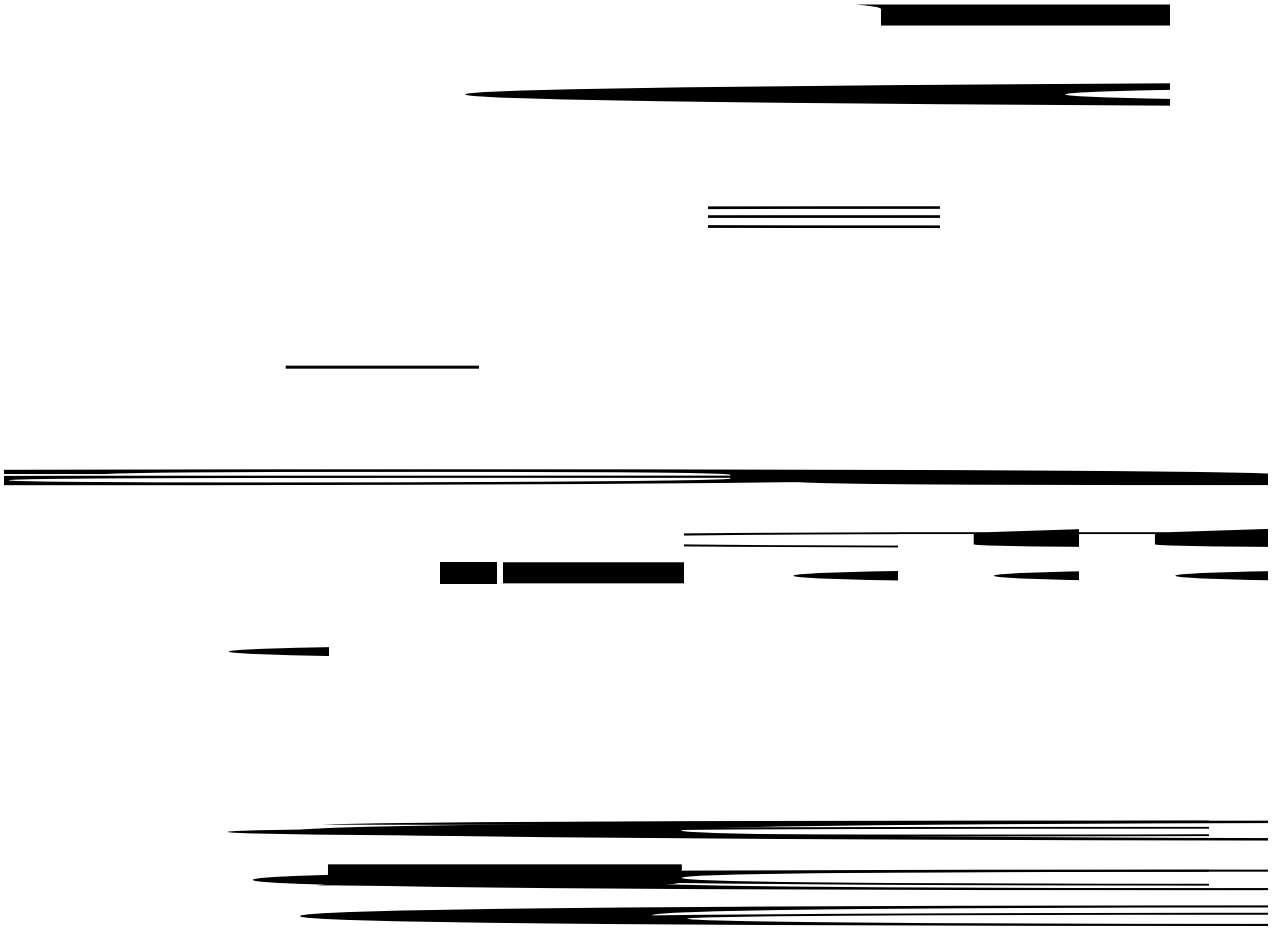
Source: Sperber, et al (2004)

Example – Echinacea Purpurea for Colds

Conditional on 32 people developing colds and 24 receiving EP, the following table gives the outcomes that would have been as strong or stronger evidence that EP reduced risk of developing cold (1-sided test). *P*-value from SPSS is .079.

EP/Cold	Plac/Cold
14	18
13	19
12	20
11	21
10	22

Example - SPSS Output



McNemar's Test for Paired Samples

Common subjects being observed under 2 conditions (2 treatments, before/after, 2 diagnostic tests) in a crossover setting

Two possible outcomes (Presence/Absence of Characteristic) on each measurement

Four possibilities for each subjects wrt outcome:

- Present in both conditions
- Absent in both conditions
- Present in Condition 1, Absent in Condition 2
- Absent in Condition 1, Present in Condition 2

McNemar's Test for Paired Samples

Condition 1\2	Present	Absent
Present	n_{11}	n_{12}
Absent	n_{21}	n_{22}

McNemar's Test for Paired Samples

H_0: Probability the outcome is Present is same for the 2 conditions

H_A: Probabilities differ for the 2 conditions (Can also be conducted as 1-sided test)

$$T.S.: z_{obs} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

$$R.R.: | z_{obs} | \geq z_{\alpha/2} \qquad (1.96 \; if \; \alpha = 0.05)$$

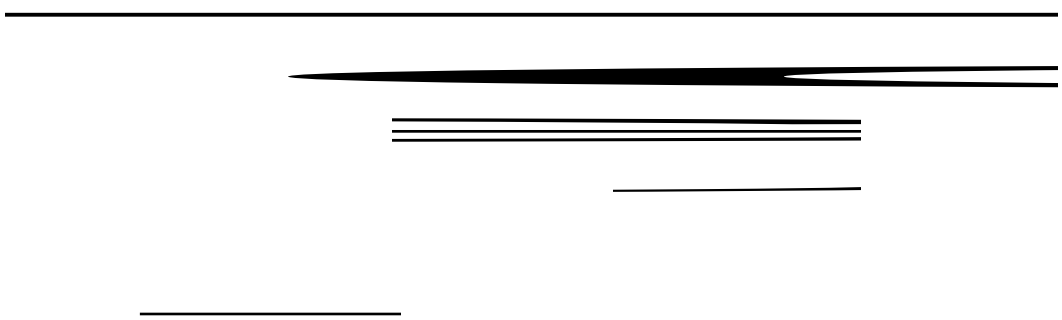
$$P - val = 2P(Z \geq | z_{obs} |)$$

Example - Reporting of Silicone Breast Implant Leakage in Revision Surgery

Subjects - 165 women having revision surgery involving silicone gel breast implants

Conditions (Each being observed on all women)

- Self Report of Presence/Absence of Rupture/Leak
- Surgical Record of Presence/Absence of Rupture/Leak



Source: Brown and Pennello (2002)

Example - Reporting of Silicone Breast Implant Leakage in Revision Surgery

H_0: Tendency to report ruptures/leaks is the same for self reports and surgical records

H_A: Tendencies differ

$$T.S.: z_{obs} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{28 - 5}{\sqrt{28 + 5}} = 4.00$$

$$R.R.: | z_{obs} | \geq 1.96$$

$$P - val = 2P(Z \geq | z_{obs} |) \approx 0$$

Pearson's Chi-Square Test

Can be used for nominal or ordinal explanatory and response variables

Variables can have any number of distinct levels

Tests whether the distribution of the response variable is the same for each level of the explanatory variable (H_0: No association between the variables

r = # of levels of explanatory variable

c = # of levels of response variable

Pearson's Chi-Square Test

Intuition behind test statistic

- Obtain marginal distribution of outcomes for the response variable
- Apply this common distribution to all levels of the explanatory variable, by multiplying each proportion by the corresponding sample size
- Measure the difference between actual cell counts and the expected cell counts in the previous step

Pearson's Chi-Square Test

Notation to obtain test statistic

- Rows represent explanatory variable (*r* levels)
- Cols represent response variable (*c* levels)

	1	**2**	**…**	**c**	**Total**
1	n_{11}	n_{12}	…	n_{1c}	$\boldsymbol{n_{1.}}$
2	n_{21}	n_{22}	…	n_{2c}	$\boldsymbol{n_{2.}}$
…	…	…	…	…	**…**
r	n_{r1}	n_{r2}	…	n_{rc}	$\boldsymbol{n_{r.}}$
Total	$\boldsymbol{n_{.1}}$	$\boldsymbol{n_{.2}}$	**…**	$\boldsymbol{n_{.c}}$	$\boldsymbol{n_{..}}$

Pearson's Chi-Square Test

Marginal distribution of response and expected cell counts under hypothesis of no association:

$$\hat{\pi}_1 = \frac{n_{.1}}{n_{..}} \quad \cdots \quad \hat{\pi}_c = \frac{n_{.c}}{n_{..}}$$

$$E(n_{ij}) = n_{i.}\hat{\pi}_j = \frac{n_{i.}n_{.j}}{n_{..}}$$

Pearson's Chi-Square Test

H_0: No association between variables

H_A: Variables are associated

- $T.S.: X^2 = \sum_i \sum_j \frac{(n_{ij} - E(n_{ij}))^2}{E(n_{ij})}$
- $R.R.: X^2 \geq \chi^2_{\alpha,(r-1)(c-1)}$
- $P - val = P(\chi^2 \geq X^2)$

Example – EMT Assessment of Kids

Observed

Assessment			
Age	Acc	Inac	**Tot**
Inf	168	73	**241**
Tod	230	73	**303**
Pre	254	53	**307**
Sch	379	58	**437**
Ado	652	124	**776**
Tot	**1683**	**381**	**2064**

Expected

Assessment			
Age	Acc	Inac	**Tot**
Inf	197	44	**241**
Tod	247	56	**303**
Pre	250	57	**307**
Sch	356	81	**437**
Ado	633	143	**776**
Tot	**1683**	**381**	**2064**

Example – EMT Assessment of Kids

Note that each expected count is the row total times the column total, divided by the overall total. For the first cell in the table:

$$E(n_{11}) = \frac{n_{1.} n_{.1}}{n_{..}} = \frac{241(1683)}{2064} = 197$$

- The contribution to the test statistic for this cell is

$$\frac{(168-197)^2}{197} = 4.27$$

Example – EMT Assessment of Kids

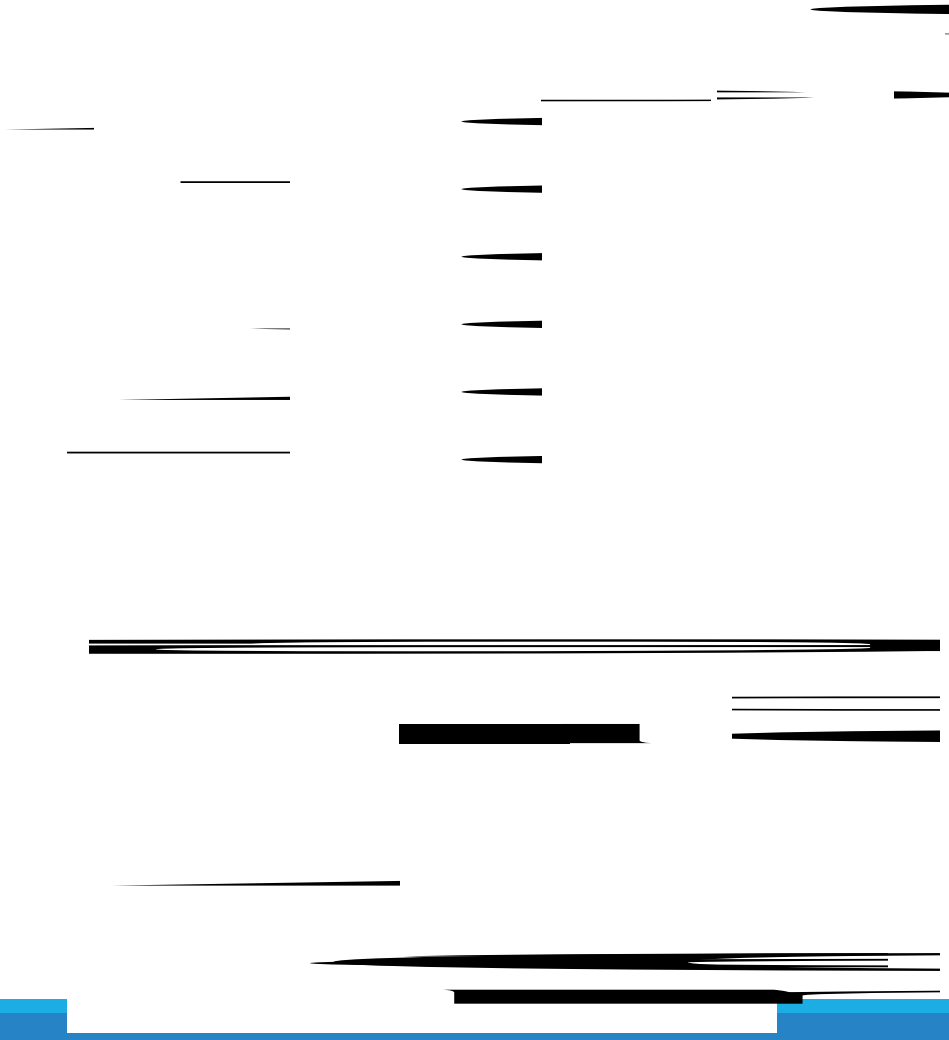
H_0: No association between variables

H_A: Variables are associated

- $T.S.: X^2 = \frac{(168-197)^2}{197} + \cdots + \frac{(124-143)^2}{143} = 40.1$
- $R.R.: X^2 \geq \chi^2_{.05,(5-1)(2-1)} = \chi^2_{.05,4} = 9.488$

Reject H_0, conclude that the accuracy of assessments differs among age groups

Example SPSS Output



Ordinal Explanatory and Response Variables

Pearson's Chi-square test can be used to test associations among ordinal variables, but more powerful methods exist

When theories exist that the association is directional (positive or negative), measures exist to describe and test for these specific alternatives from independence:

- Gamma
- Kendall's τ_b

Concordant and Discordant Pairs

Concordant Pairs - Pairs of individuals where one individual scores "higher" on both ordered variables than the other individual

Discordant Pairs - Pairs of individuals where one individual scores "higher" on one ordered variable and the other individual scores "higher" on the other

C = # Concordant Pairs D = # Discordant Pairs

- Under Positive association, expect $C > D$
- Under Negative association, expect $C < D$
- Under No association, expect $C \approx D$

Example - Alcohol Use and Sick Days

Alcohol Risk (Without Risk, Hardly any Risk, Some to Considerable Risk)

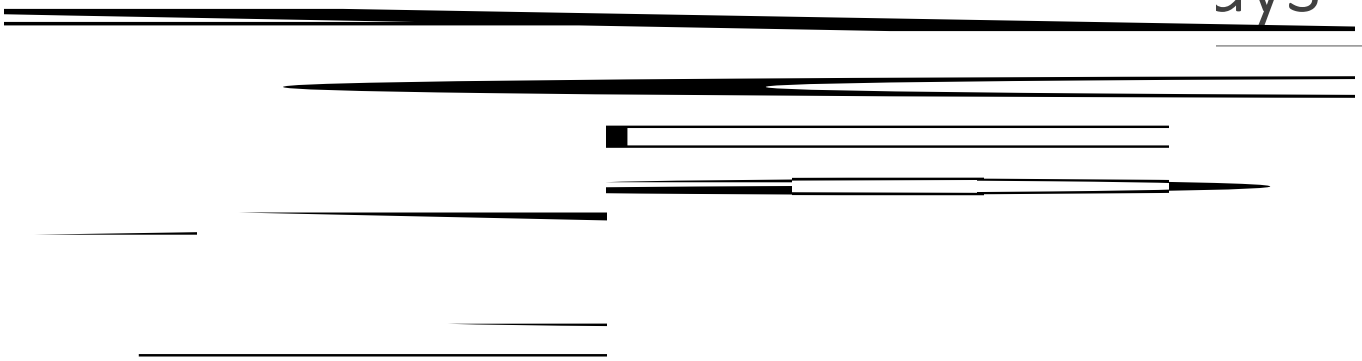
Sick Days (0, 1-6, $\geq$7)

Concordant Pairs - Pairs of respondents where one scores higher on both alcohol risk and sick days than the other

Discordant Pairs - Pairs of respondents where one scores higher on alcohol risk and the other scores higher on sick days

Source: Hermansson, et al (2003)

Example – Alcohol Use and Sick Days



• Concordant Pairs: Each individual in a given cell is concordant with each individual in cells “Southeast” of theirs

•Discordant Pairs: Each individual in a given cell is discordant with each individual in cells “Southwest” of theirs

Example – Alcohol Use and Sick Days

$$C = 347(63+56+25+34)+113(56+34)+154(25+34)+63(34) = 83164$$

$$D = 145(154+63+52+25)+113(154+52)+56(52+25)+63(52) = 73496$$

Measures of Association

- Goodman and Kruskal's Gamma:

$$\hat{\gamma} = \frac{C - D}{C + D} \qquad -1 \leq \hat{\gamma} \leq +1$$

- Kendall's τ_b:

$$\hat{\tau}_b = \frac{C - D}{\sqrt{(n^2 - \sum n_{i.}^{\ 2})(n^2 - \sum n_{.j}^{\ 2})}}$$

When there's no association between the ordinal variables, the population based values of these measures are 0. Statistical software packages provide these tests.

Example - Alcohol Use and Sick Days

$$\hat{\gamma} = \frac{C - D}{C + D} = \frac{83164 - 73496}{83164 + 73496} = 0.0617$$