# Categorical Data Analysis I: Associations with nominal and ordinal data

Contents

## 1. Nominal/nominal association

A randomized clinical trial was conducted to estimate incidence of HPV and assess the effectiveness of the HPV 16 vaccine. 414 subjects aged 15-25 were assigned to receive the vaccine, while a control group of 385 did not receive the vaccine. The table below indicates the number in each group that acquired HPV infection during the study period.

| Group | Infection | |
|---|---|---|
| | No | Yes |
| Control | 366 | 19 |
| Vaccine | 413 | 1 |

Question 1: What is the incidence of HPV in each group?
Question 2: Is the incidence of HPV lower in the vaccine group?

### 1.1. Estimating a population proportion based on a single sample.

Binomial experiment:

- Series of identical, independent "trials" (Observe subject throughout the study period)
- Each trial results in one of two possible outcomes (Acquires HPV or does not)
- Count the number of "successes" (number that acquire HPV)
- Interest is in the proportion of successes (proportion that acquire HPV)

95% Confidence interval for population proportion

Basic form of the interval: sample estimate +/- margin of error

Wald interval ("textbook" interval)

Sample estimate: $\hat{p} = \frac{\#\ successes}{n}$; margin of error: $1.96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Works "OK" for

large samples
population proportion not close to 0 or 1
suffers from bias and undercoverage otherwise
bias: systematically lower or higher than population proportion
undercoverage: Actual confidence level less
than 95% (intervals tend to be too narrow)

Agresti-Coull interval (new and improved “textbook” interval)

Helps to “fix” problems with the Wald interval—add 2 successes and 2 failures

Sample estimate: $\tilde{p} = \frac{\#\ successes + 2}{n+4}$; margin of error: $1.96 * \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$

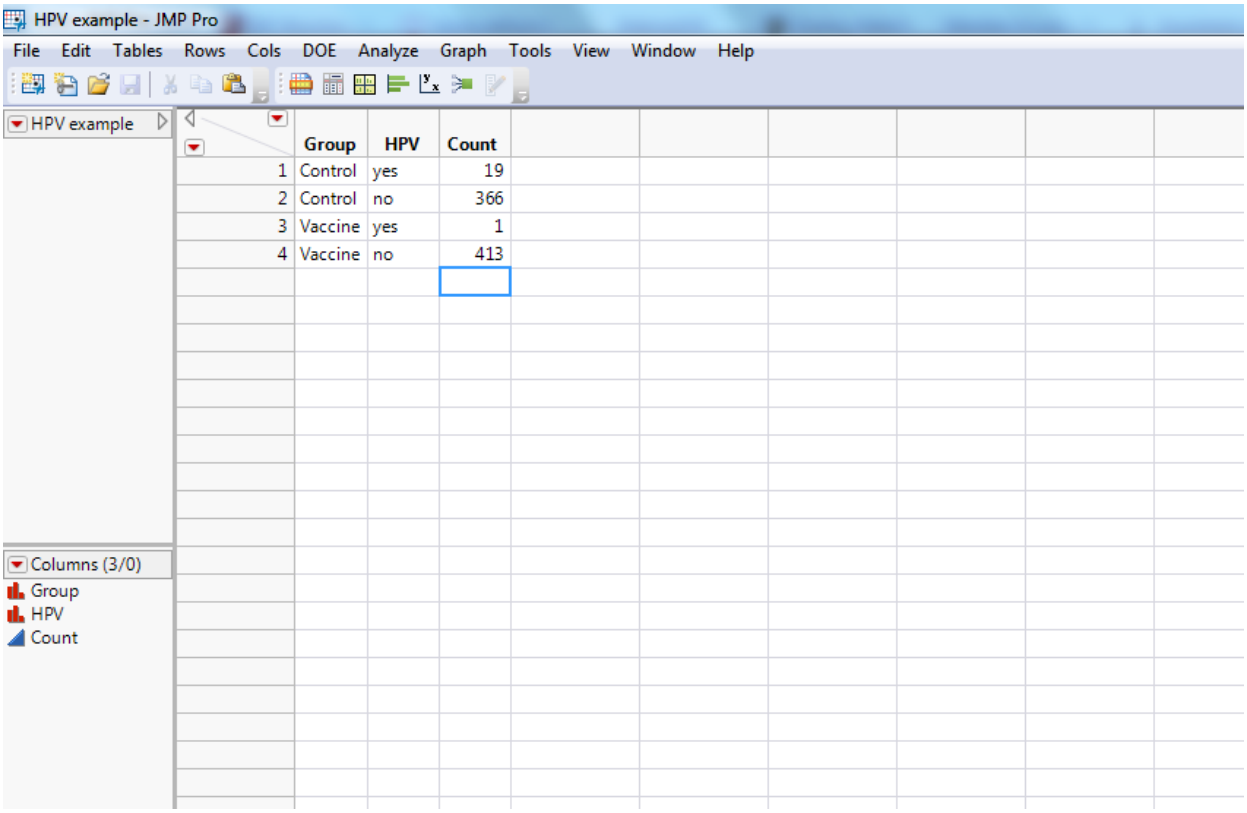
Works better for smaller samples, population proportions close to 0 or 1

Score interval (“Ideal” interval, but more complicated-doesn’t appear in most textbooks)

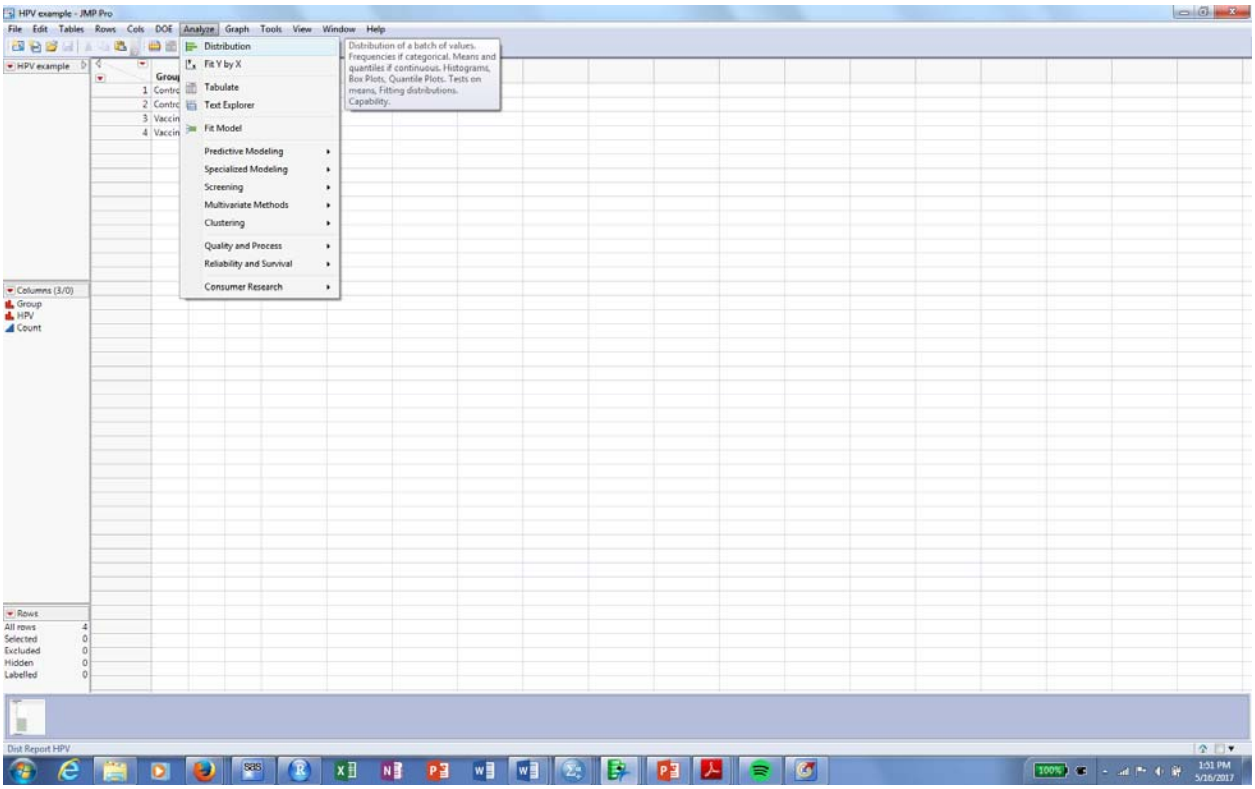
HPV example

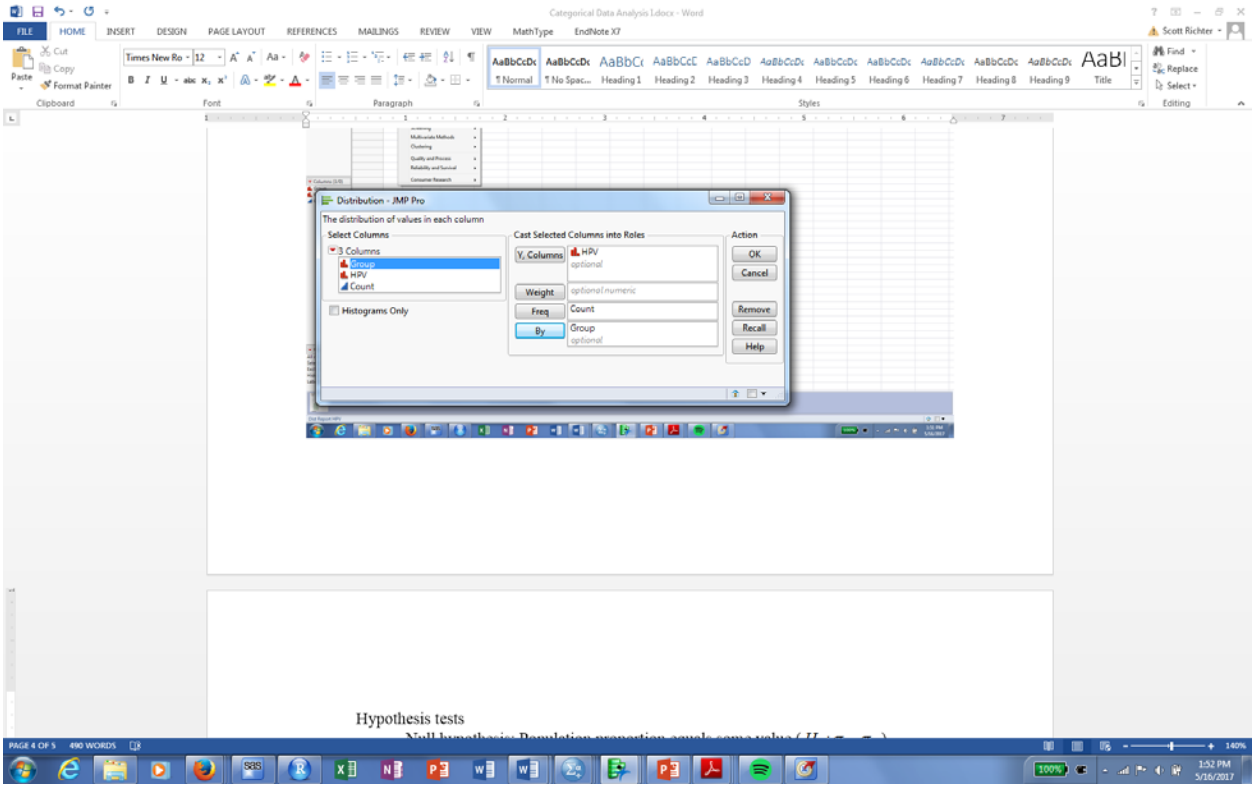
| | | 95% confidence interval | |
|---|---|---|---|
| Group | Method | Lower limit | Upper limit |
| Control | Wald | 0.0292 | 0.0695 |
| | Agresti-Coull | 0.0315 | 0.0764 |
| | Score | 0.0318 | 0.0757 |
| | | | |
| Vaccine | Wald | -0.0023 | 0.00714 |
| | Agresti-Coull | -0.0009 | 0.01526 |
| | Score | 0.0004 | 0.01355 |

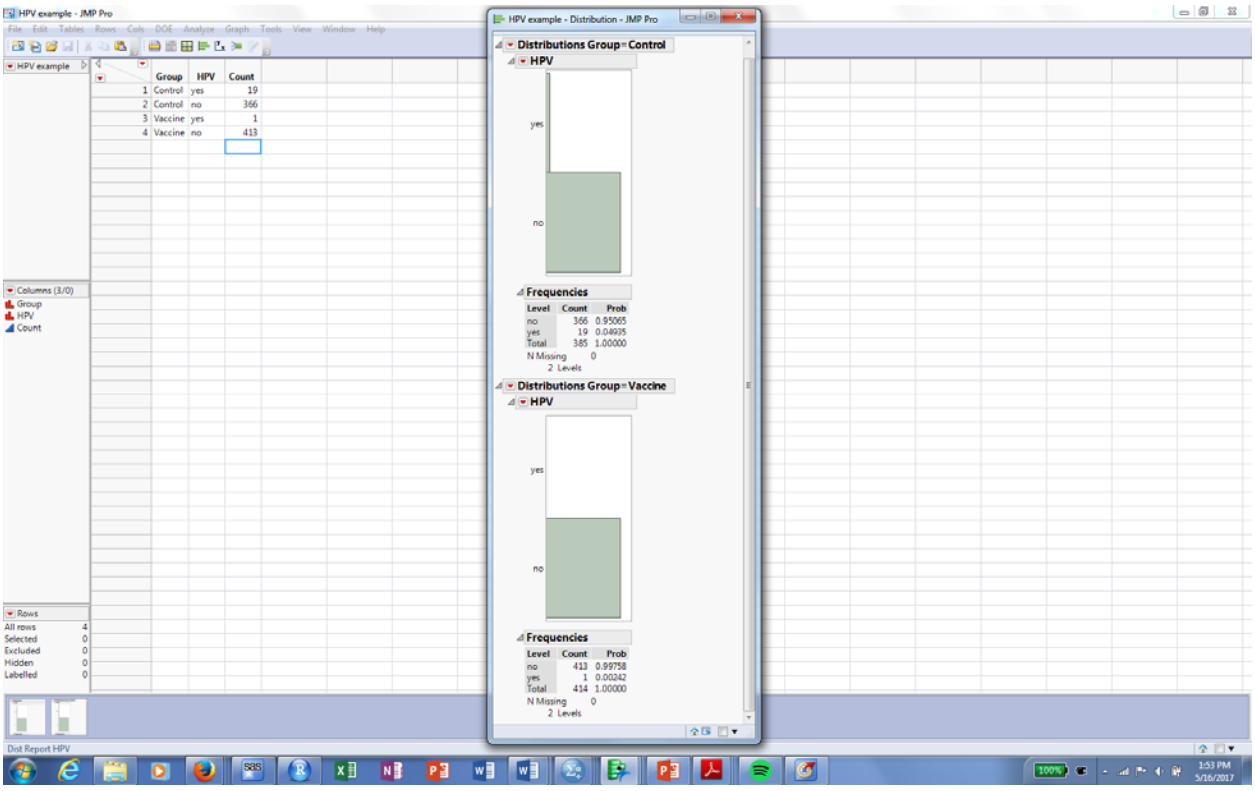
JMP

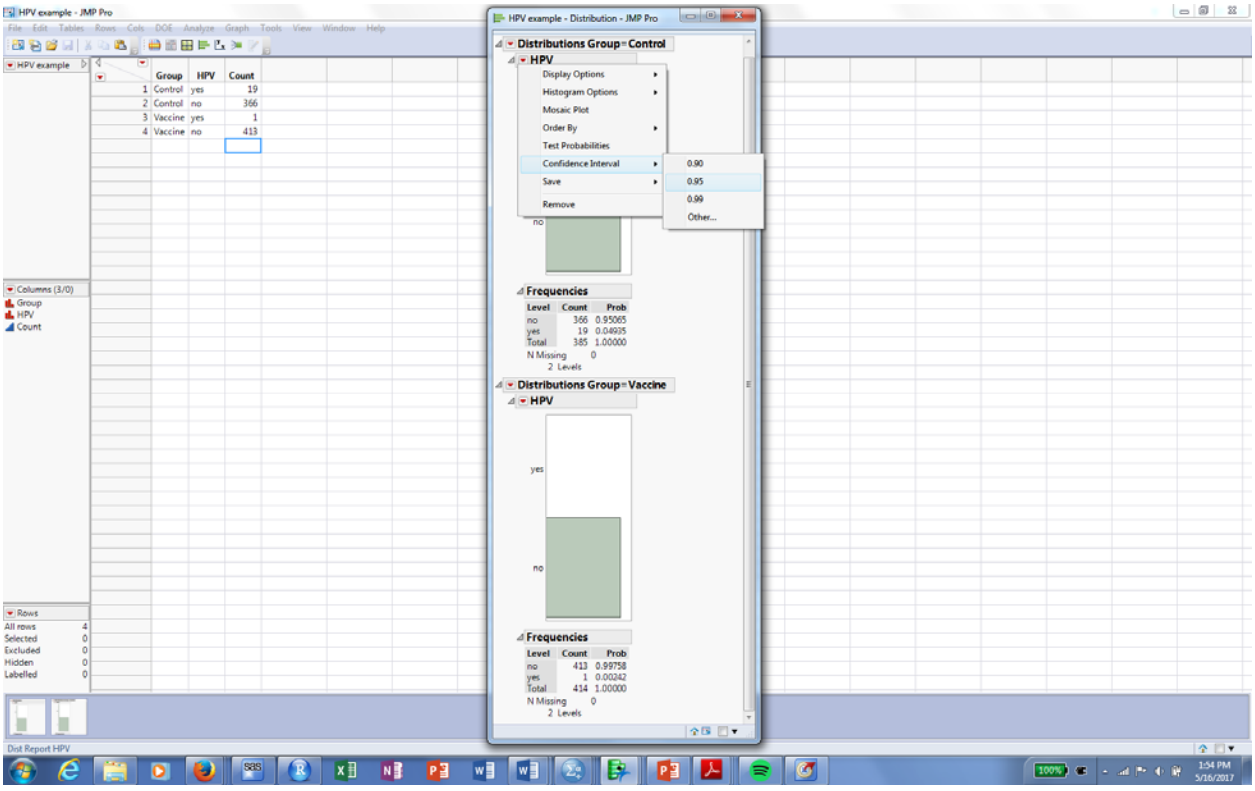


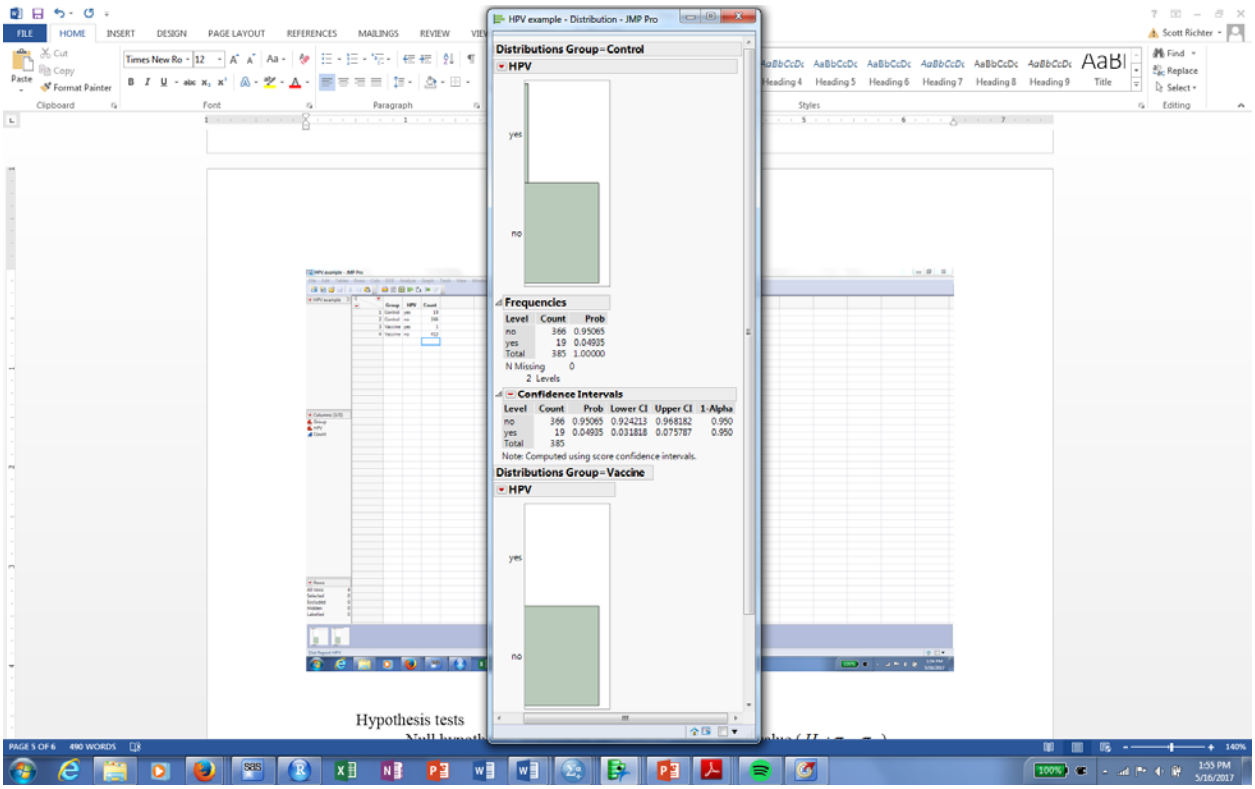

| | Group | HPV | Count |
|---|---|---|---|
| 1 | Control | yes | 19 |
| 2 | Control | no | 366 |
| 3 | Vaccine | yes | 1 |
| 4 | Vaccine | no | 413 |


HPV example - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
Distribution
Fit Y by X
Tabulate
Text Explorer
Fit Model
Predictive Modeling
Specialized Modeling
Screening
Multivariate Methods
Clustering
Quality and Process
Reliability and Survival
Consumer Research
Distribution of a batch of values. Frequencies if categorical. Means and quantiles if continuous. Histograms, Box Plots, Quantile Plots. Tests on means. Fitting distributions. Capability.
Columns (3/0)
Group
HPV
Count
Rows
All rows 4
Selected 0
Excluded 0
Hidden 0
Labelled 0
Dist Report HPV


Distribution - JMP Pro
The distribution of values in each column
Select Columns
3 Columns
Group
HPV
Count
Histograms Only
Cast Selected Columns into Roles
Y, Columns
HPV
optional
Weight
optional numeric
Freq
Count
By
Group
optional
Action
OK
Cancel
Remove
Recall
Help
Hypothesis tests


HPV example - JMP Pro
HPV example - Distribution - JMP Pro
Distributions Group=Control
HPV
yes
no
Frequencies
Level Count Prob
no 366 0.95065
yes 19 0.04935
Total 385 1.00000
N Missing 0
2 Levels
Distributions Group=Vaccine
HPV
yes
no
Frequencies
Level Count Prob
no 413 0.99758
yes 1 0.00242
Total 414 1.00000
N Missing 0
2 Levels
Group HPV Count
1 Control yes 19
2 Control no 366
3 Vaccine yes 1
4 Vaccine no 413
Columns (3/0)
Group
HPV
Count
Rows
All rows 4
Selected 0
Excluded 0
Hidden 0
Labelled 0
Dist Report HPV


HPV example - JMP Pro
HPV example - Distribution - JMP Pro
Distributions Group=Control
HPV
Display Options
Histogram Options
Mosaic Plot
Order By
Test Probabilities
Confidence Interval
Save
Remove
0.90
0.95
0.99
Other...
yes
no
Frequencies
Level Count Prob
no 366 0.95065
yes 19 0.04935
Total 385 1.00000
N Missing 0
2 Levels
Distributions Group=Vaccine
HPV
yes
no
Frequencies
Level Count Prob
no 413 0.99758
yes 1 0.00242
Total 414 1.00000
N Missing 0
2 Levels
Group HPV Count
1 Control yes 19
2 Control no 366
3 Vaccine yes 1
4 Vaccine no 413
Columns (3/0)
Group
HPV
Count
Rows
All rows 4
Selected 0
Excluded 0
Hidden 0
Labelled 0
Dist Report HPV




R

Control group:

```
#Wald
19/385-1.96*sqrt(19/385*366/414/414)
```

```
## [1] 0.02922997
```

```
19/385+1.96*sqrt(19/385*366/414/414)
```

```
## [1] 0.06947133
```

```
#AC
21/389-1.96*sqrt(21/389*368/389/389)
```

```
## [1] 0.03152688
```

```
21/389+1.96*sqrt(21/389*368/389/389)
```

```
## [1] 0.07644227
```

```
#Score
prop.test(19,385,correct=F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  19 out of 385, null probability 0.5
## X-squared = 312.75, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.03181847 0.07578697
## sample estimates:
##          p
## 0.04935065
```

Vaccine group

```
#Vaccine
#Wald
1/414-1.96*sqrt(1/414*412/414/414)
```

```
## [1] -0.002307391
```

```
1/414+1.96*sqrt(1/414*412/414/414)
```

```
## [1] 0.007138309
```

```
#AC
3/418-1.96*sqrt(3/418*414/418/418)
```

```
## [1] -0.0009055918
```

```
3/418+1.96*sqrt(3/418*414/418/418)
```

```
## [1] 0.01525966
```

```
#Score
prop.test(1,414,correct=F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  1 out of 414, null probability 0.5
## X-squared = 410.01, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.0004265151 0.0135535692
## sample estimates:
##           p
## 0.002415459
```

SAS

```
data gibbs;
input Group$ HPV$ count @@;
datalines;
Control Yes 19  Control No 366
Vaccine Yes 1   Vaccine No 413
;

proc freq data=gibbs;
weight count;
tables HPV /
   binomial (level='Yes' CL=all) /*Request confidence
                        intervals for proportion 'Yes'*/;
by Group;
run;
```

Group=Control

| HPV | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| No | 366 | 95.06 | 366 | 95.06 |
| Yes | 19 | 4.94 | 385 | 100.00 |

| Binomial Proportion | |
|---|---|
| HPV = Yes | |
| Proportion | 0.0494 |
| ASE | 0.0110 |

| Type | 95% Confidence Limits | |
|---|---|---|
| Wald | 0.0277 | 0.0710 |
| Wilson | 0.0318 | 0.0758 |
| Agresti-Coull | 0.0314 | 0.0762 |

Group=Vaccine

| HPV | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| No | 413 | 99.76 | 413 | 99.76 |
| Yes | 1 | 0.24 | 414 | 100.00 |

| Binomial Proportion | |
|---|---|
| HPV = Yes | |
| Proportion | 0.0024 |
| ASE | 0.0024 |

| Type | 95% Confidence Limits | |
|---|---|---|
| Wald | 0.0000 | 0.0071 |
| Wilson | 0.0004 | 0.0136 |
| Agresti-Coull | 0.0000 | 0.0150 |

**1.2 Comparing two proportions—independent samples**

Proportion difference—interpretation depends on incidence rates

Risk ratio (relative risk)—may not be valid for retrospective studies

Odds ratio—most obscure for practitioners

HPV example

| Comparison | Estimate | Interpretation |
|---|---|---|
| Control--Vaccine | $\frac{19}{385}-\frac{1}{414}=0.04935-0.00242=0.047$ | Incidence of HPV higher in Control group by 4.7% |
| $\frac{Incidence(HPV)[Control]}{Incidence(HPV)[Vaccine]}$ | $\frac{19}{385}\Big/\frac{1}{414}=\frac{0.04935}{0.00242}=20.4$ | Incidence of HPV in Control group 20.4 times higher |
| $\frac{Odds(HPV)[Control]}{Odds(HPV)[Vaccine]}$ | $\frac{19}{366}\Big/\frac{1}{413}=\frac{0.04935}{0.00242}=21.4$ | Odds of HPV in Control group 21.4 times higher |

1.2.1 Confidence intervals

Proportion difference

Wald interval

Sample estimate: $\hat{p}_1 - \hat{p}_2$; margin of error: $1.96 * \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Similar issues as in the one-sample case

Agresti-Caffo interval

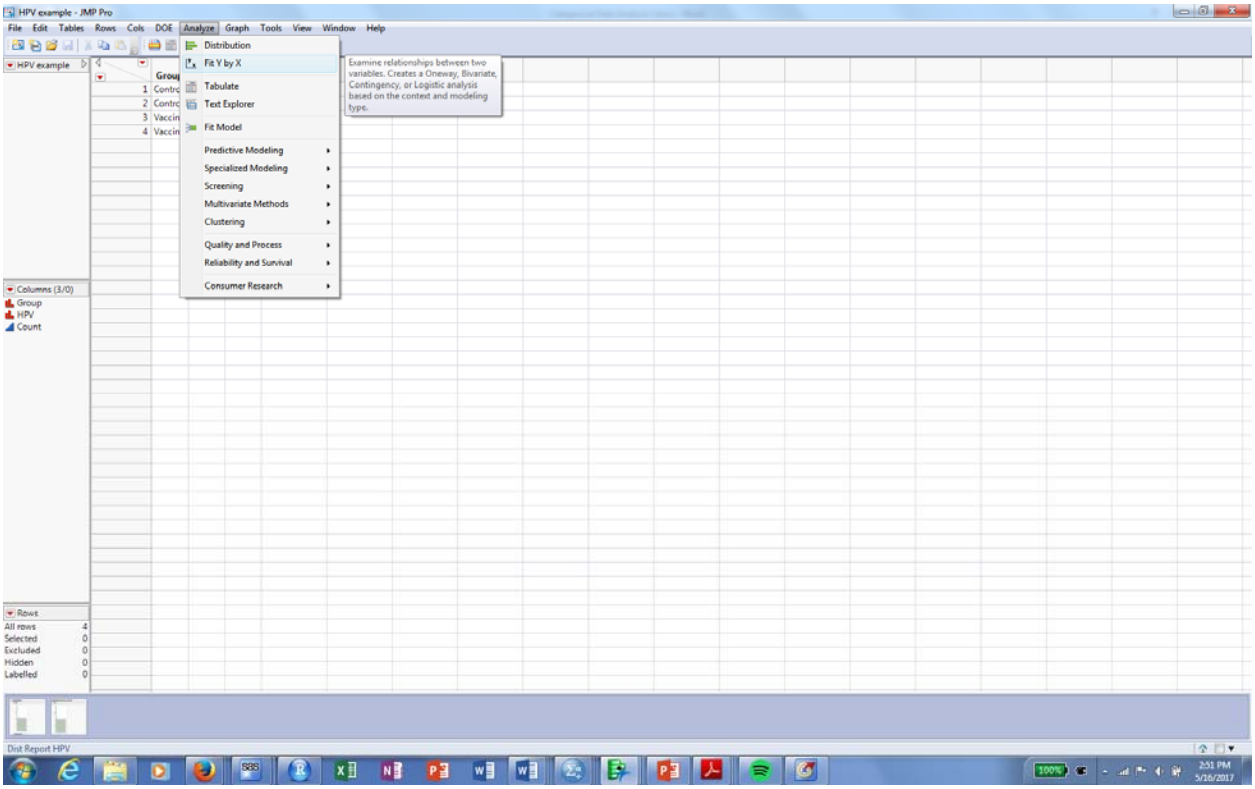
Add 1 success and 1 failure to each group

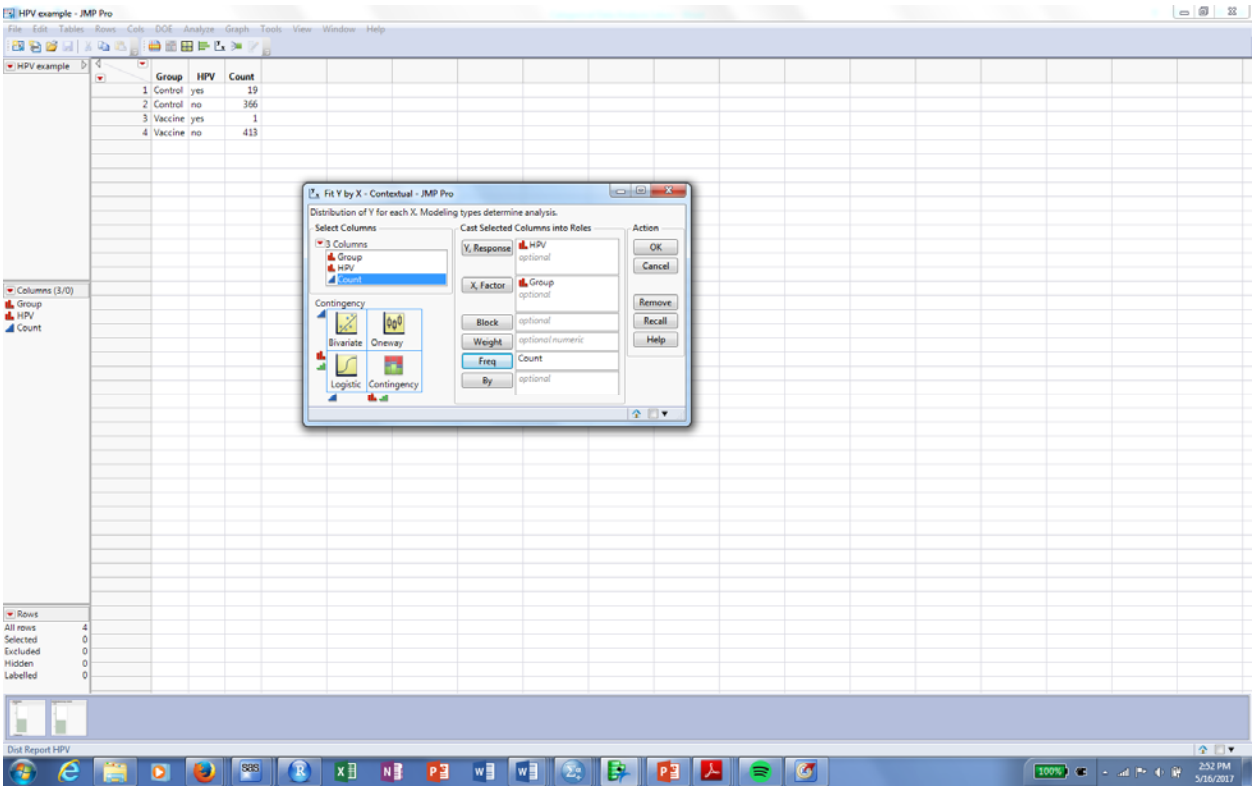
Use $\tilde{p}_i = \frac{\#\ successes + 1}{n+2}$ instead of $\hat{p}_i$

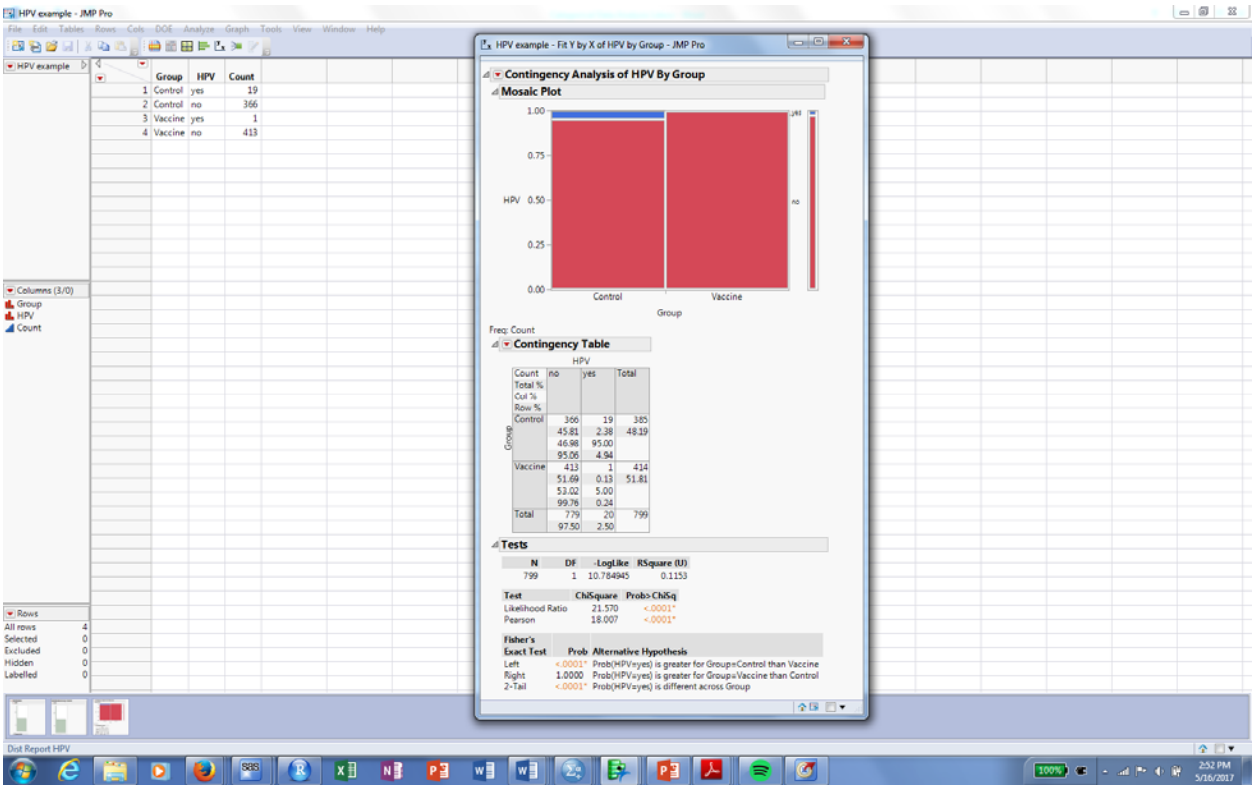
Risk ratio (relative risk) and odds ratio

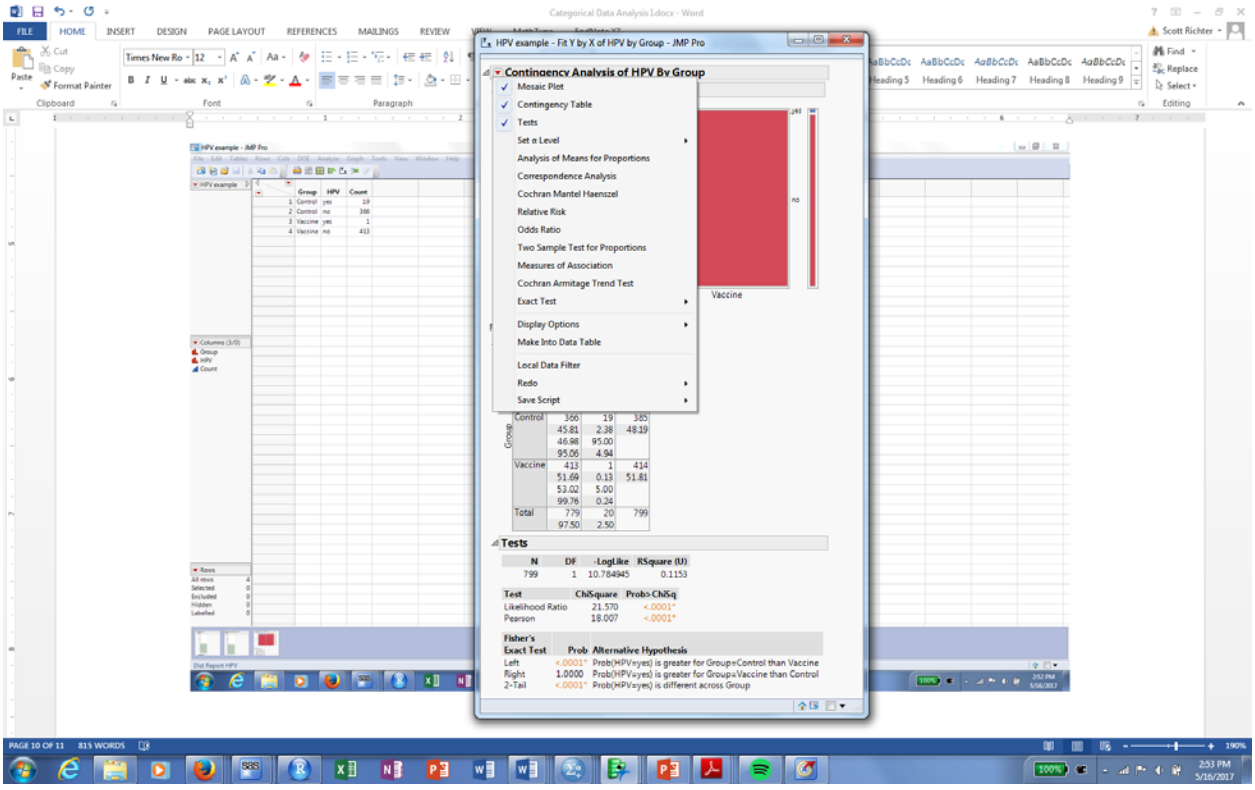
Inference usually based on $\ln(ratio)$ and using Wald interval

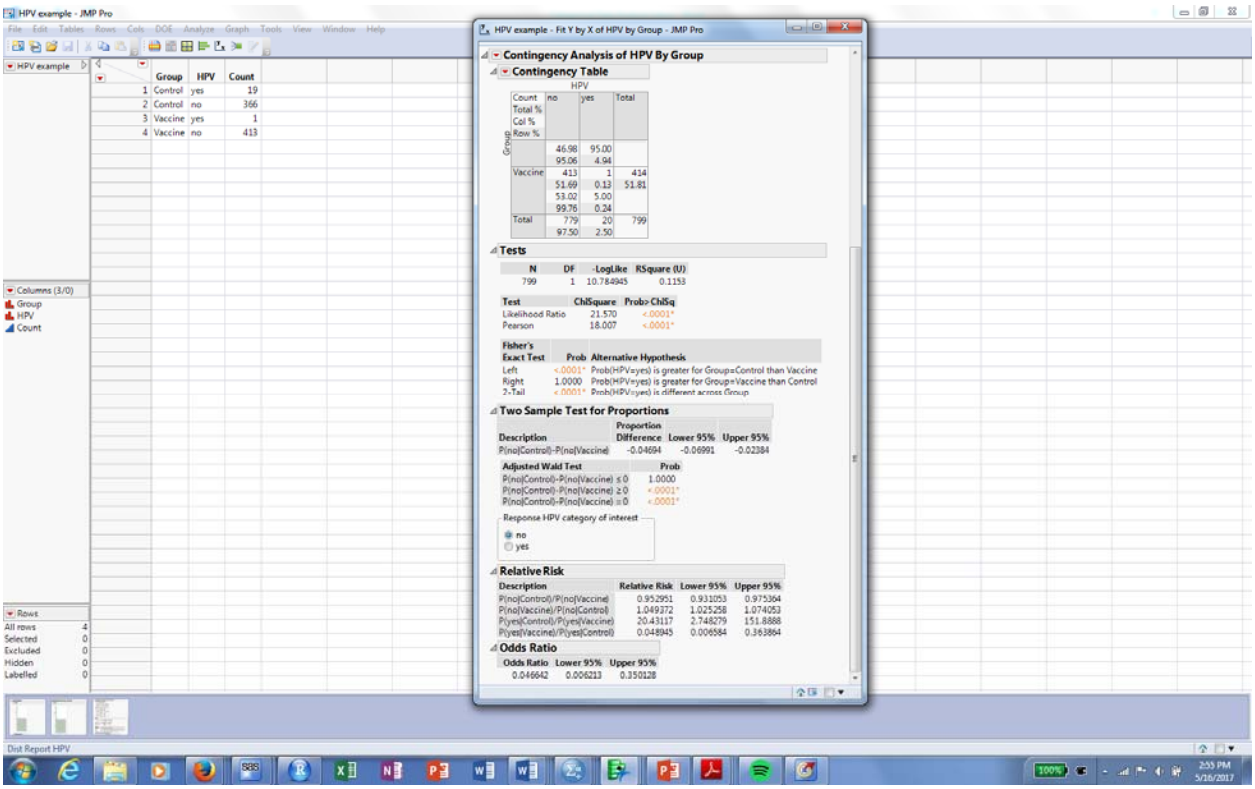
JMP


HPV example - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
Distribution
Fit Y by X
Tabulate
Text Explorer
Fit Model
Predictive Modeling
Specialized Modeling
Screening
Multivariate Methods
Clustering
Quality and Process
Reliability and Survival
Consumer Research
Examine relationships between two variables. Creates a Oneway, Bivariate, Contingency, or Logistic analysis based on the context and modeling type.
Columns (3/0)
Group
HPV
Count
Rows
All rows 4
Selected 0
Excluded 0
Hidden 0
Labelled 0
Dist Report HPV


HPV example - JMP Pro
Group HPV Count
1 Control yes 19
2 Control no 366
3 Vaccine yes 1
4 Vaccine no 413
Fit Y by X - Contextual - JMP Pro
Distribution of Y for each X. Modeling types determine analysis.
Select Columns
Cast Selected Columns into Roles
Action
3 Columns
Group
HPV
Count
Y, Response
X, Factor
Block
Weight
Freq
By
OK
Cancel
Remove
Recall
Help
Contingency
Bivariate
Oneway
Logistic
Contingency
Columns (3/0)
Rows
All rows 4
Selected 0
Excluded 0
Hidden 0
Labelled 0
Dist Report HPV


HPV example - JMP Pro
HPV example - Fit Y by X of HPV by Group - JMP Pro
Contingency Analysis of HPV By Group
Mosaic Plot
1.00
0.75
HPV 0.50
0.25
0.00
Control
Vaccine
Group
yes
no
Freq: Count
Contingency Table
HPV
Count
Total %
Col %
Row %
no
yes
Total
Control
366
45.81
46.98
95.06
19
2.38
95.00
4.94
385
48.19
Vaccine
413
51.69
53.02
99.76
1
0.13
5.00
0.24
414
51.81
Total
779
97.50
20
2.50
799
Tests
N
DF
-LogLike
RSquare (U)
799
1
10.784945
0.1153
Test
ChiSquare
Prob>ChiSq
Likelihood Ratio
21.570
<.0001*
Pearson
18.007
<.0001*
Fisher's Exact Test
Prob
Alternative Hypothesis
Left
<.0001*
Prob(HPV=yes) is greater for Group=Control than Vaccine
Right
1.0000
Prob(HPV=yes) is greater for Group=Vaccine than Control
2-Tail
<.0001*
Prob(HPV=yes) is different across Group
Columns (3/0)
Group
HPV
Count
Rows
All rows
Selected
Excluded
Hidden
Labelled
Dist Report HPV


Categorical Data Analysis 1.docx - Word
HPV example - Fit Y by X of HPV by Group - JMP Pro
Contingency Analysis of HPV By Group
Mosaic Plot
Contingency Table
Tests
Set α Level
Analysis of Means for Proportions
Correspondence Analysis
Cochran Mantel Haenszel
Relative Risk
Odds Ratio
Two Sample Test for Proportions
Measures of Association
Cochran Armitage Trend Test
Exact Test
Display Options
Make Into Data Table
Local Data Filter
Redo
Save Script
Vaccine
Tests
Fisher's Exact Test
Prob
Alternative Hypothesis
Left
<.0001*
Prob(HPV=yes) is greater for Group=Control than Vaccine
Right
1.0000
Prob(HPV=yes) is greater for Group=Vaccine than Control
2-Tail
<.0001*
Prob(HPV=yes) is different across Group
PAGE 10 OF 11
815 WORDS




SAS

```
proc freq data=gibbs;
weight count;
tables Group*HPV /
  riskdiff (Column=2 CL=Wald CL=AC)/*Estimate difference
                        between proportions*/
  relrisk /*Estimate relative risk and odds ratio*/;
run;
```

Column 2 Risk Estimates

| | Risk | ASE | (Asymptotic) 95% Confidence Limits | | (Exact) 95% Confidence Limits | |
|---|---|---|---|---|---|---|
| Row 1 | 0.0494 | 0.0110 | 0.0277 | 0.0710 | 0.0300 | 0.0760 |
| Row 2 | 0.0024 | 0.0024 | 0.0000 | 0.0071 | 0.0001 | 0.0134 |
| Total | 0.0250 | 0.0055 | 0.0142 | 0.0359 | 0.0154 | 0.0384 |
| Difference | 0.0469 | 0.0113 | 0.0248 | 0.0691 | | |

Difference is (Row 1 - Row 2)

| Confidence Limits for the Proportion (Risk) Difference | | |
|---|---|---|
| Column 2 (HPV = Yes) | | |
| Proportion Difference = 0.0469 | | |
| Type | 95% Confidence Limits | |
| Agresti-Caffo | 0.0238 | 0.0699 |
| Wald | 0.0248 | 0.0691 |

| Estimates of the Relative Risk (Row1/Row2) | | | |
|---|---|---|---|
| Type of Study | Value | 95% Confidence Limits | |
| Case-Control (Odds Ratio) | 0.0466 | 0.0062 | 0.3501 |
| Cohort (Col1 Risk) | 0.9530 | 0.9311 | 0.9754 |
| Cohort (Col2 Risk) | 20.4312 | 2.7483 | 151.8888 |

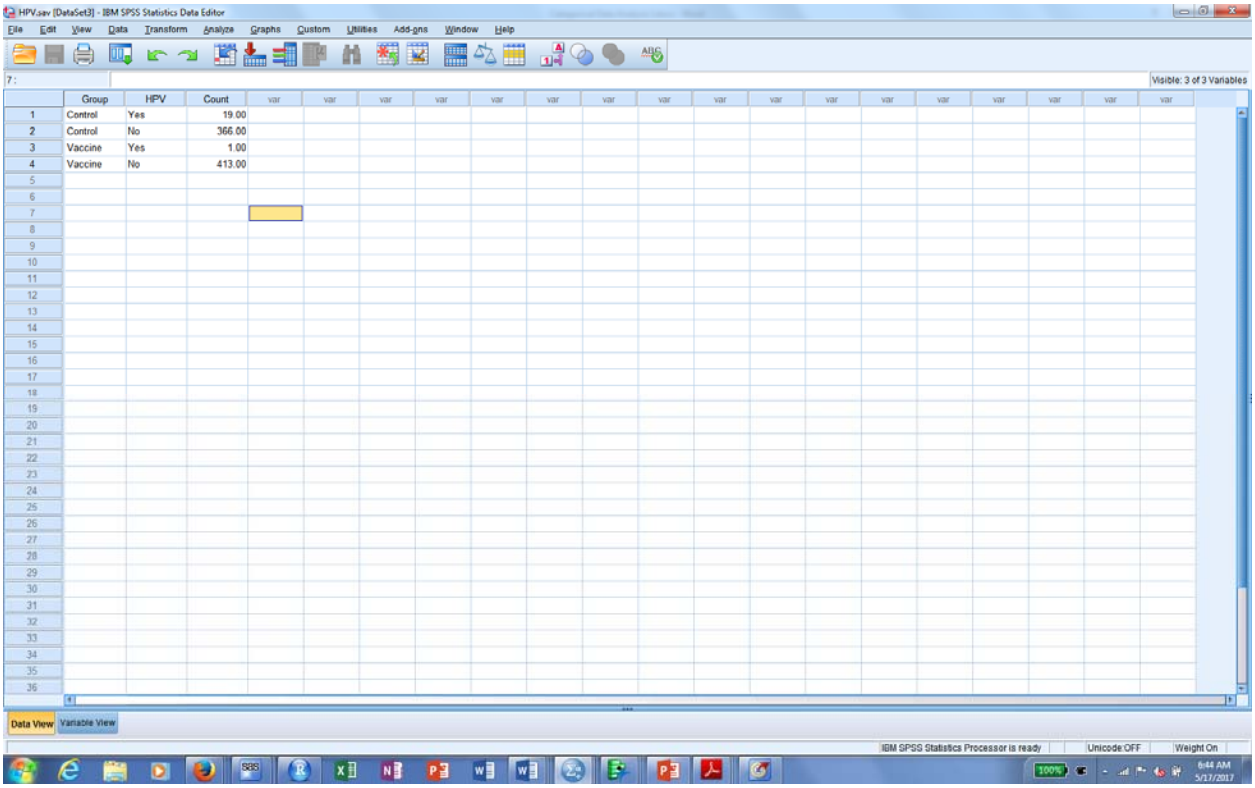
R

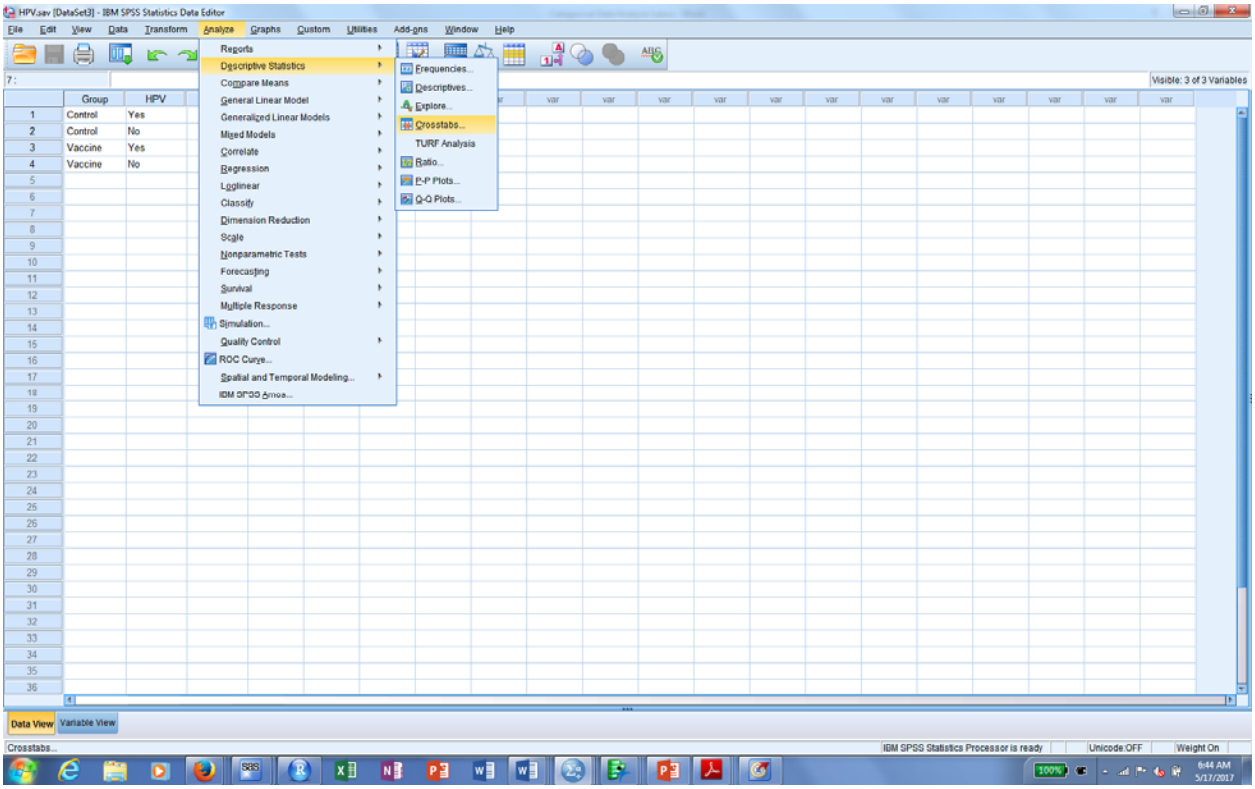
Confidence interval for proportion difference

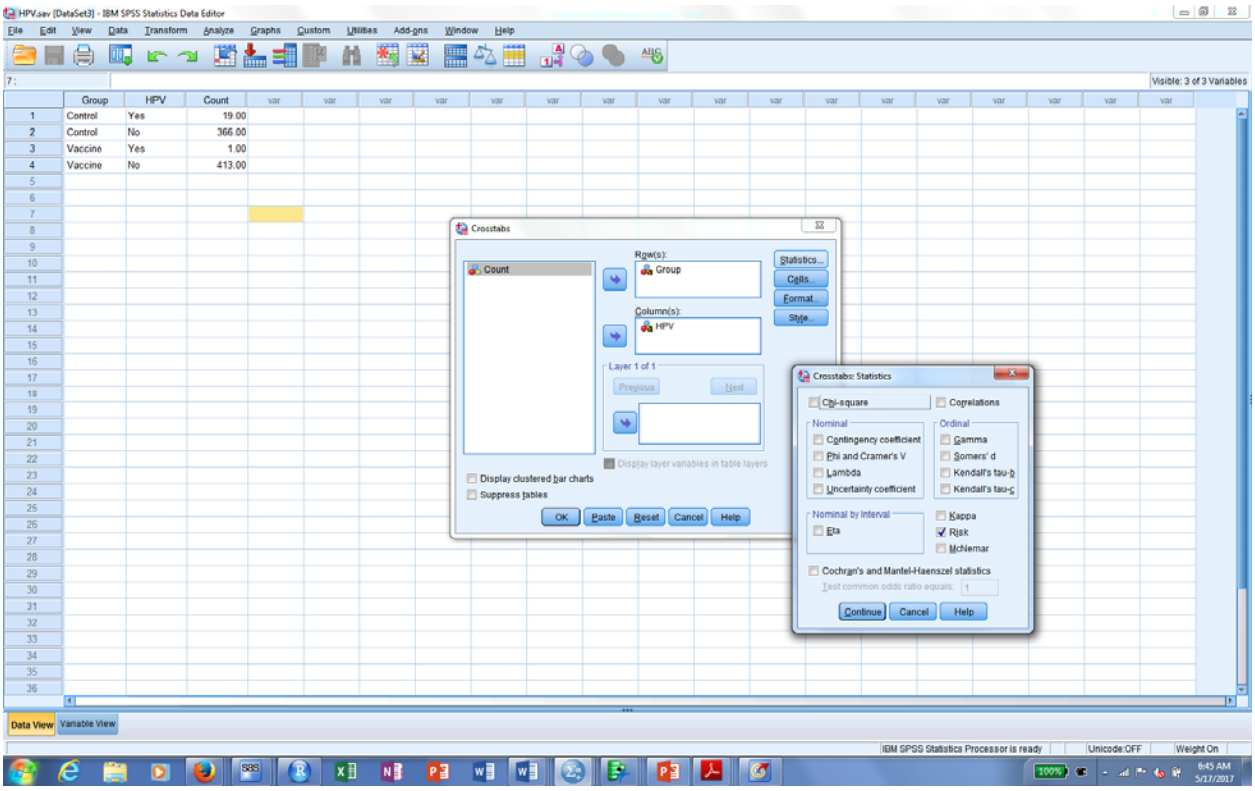
```
prop.test(x=c(19,1),n=c(385,414),correct=F)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(19, 1) out of c(385, 414)
## X-squared = 18.007, df = 1, p-value = 2.201e-05
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.02478865 0.06908174
## sample estimates:
##      prop 1      prop 2
## 0.049350649 0.002415459
```

SPSS


HPV.sav [DataSet3] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Custom Utilities Add-ons Window Help
Visible: 3 of 3 Variables
Group HPV Count
1 Control Yes 19.00
2 Control No 366.00
3 Vaccine Yes 1.00
4 Vaccine No 413.00
Data View Variable View
IBM SPSS Statistics Processor is ready
Unicode:OFF
Weight On


HPV.sav [DataSet3] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Custom Utilities Add-ons Window Help
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Simulation...
Quality Control
ROC Curve...
Spatial and Temporal Modeling...
Frequencies...
Descriptives...
Explore...
Crosstabs...
TURF Analysis
Ratio...
P-P Plots...
Q-Q Plots...
Visible: 3 of 3 Variables
Group HPV
1 Control Yes
2 Control No
3 Vaccine Yes
4 Vaccine No
Data View Variable View
Crosstabs...
IBM SPSS Statistics Processor is ready
Unicode:OFF
Weight On




Risk Estimate

| | Value | 95% Confidence Interval | |
|---|---|---|---|
| | | Lower | Upper |
| Odds Ratio for Group (Control / Vaccine) | .047 | .006 | .350 |
| For cohort HPV = No | .953 | .931 | .975 |
| For cohort HPV = Yes | 20.431 | 2.748 | 151.889 |
| N of Valid Cases | 799 | | |

1.2.2 Hypothesis tests

HPV example. Suppose the research hypothesis is that the vaccine reduces the incidence rate. Then we wish to test one of three sets of equivalent hypotheses:

1. $H_0 : \pi_V = \pi_C$ vs. $H_A : \pi_V < \pi_C$,
2. $H_0 : \frac{\pi_V}{\pi_C} = 1$ vs. $H_A : \frac{\pi_V}{\pi_C} < 1$, or
3. $H_0 : \frac{Odds(HPV)_V}{Odds(HPV)_C} = 1$ vs. $H_A : \frac{Odds(HPV)_V}{Odds(HPV)_C} < 1$

Test statistic 1-- $Z = \frac{\hat{\pi}_V - \hat{\pi}_C}{SE\left(\hat{\pi}_V - \hat{\pi}_C\right)}$, $SE\left(\hat{\pi}_V - \hat{\pi}_C\right) = \sqrt{\frac{\pi(1-\pi)}{n_1} + \frac{\pi(1-\pi)}{n_2}}$. $\pi$ is the common true incidence rate under the null hypothesis and is estimated by computing the combined sample incidence rate over both groups, $\hat{\pi} = \frac{total\ number\ of\ HPV\ cases}{n_1 + n_2} = \frac{19+1}{385+414} = 0.025$. Then the test statistic value is $Z = \frac{\frac{1}{414} - \frac{19}{385}}{\sqrt{\frac{0.025(0.975)}{414} + \frac{0.025(0.975)}{385}}} = -4.243$, with corresponding p-value less than 0.0001.

R is the only software that produced a test statistic (X-squared = $Z^2$ ) and p-value, although JMP also showed the p-value. However, as we will see, the p-value can be calculated by all software using a chi-squared test.

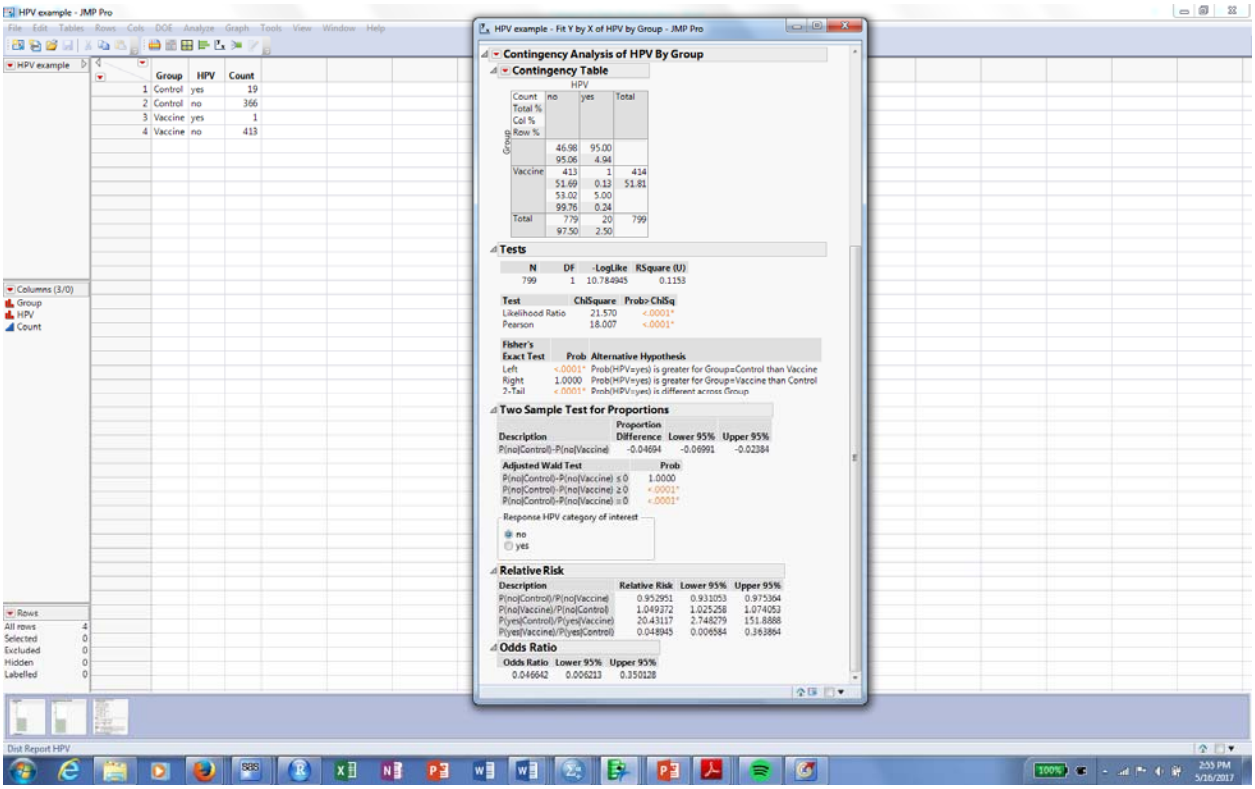
R

p-value for proportion difference

```
prop.test(x=c(19,1),n=c(385,414),correct=F, alternative="greater")

##
##  2-sample test for equality of proportions without continuity
##  correction
##
```

```
## data:  c(19, 1) out of c(385, 414)
## X-squared = 18.007, df = 1, p-value = 1.101e-05
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.02834922 1.00000000
## sample estimates:
##      prop 1      prop 2
## 0.049350649 0.002415459
```

JMP



## 1.3. Chi-squared test

Generalizes the Z-test to

1. 2 or more groups,
2. outcomes with 2 or more categories

### *1.3.1. 2x2 table*

Compares the observed table with what would be expected if the probabilities were the same:

Observed table:

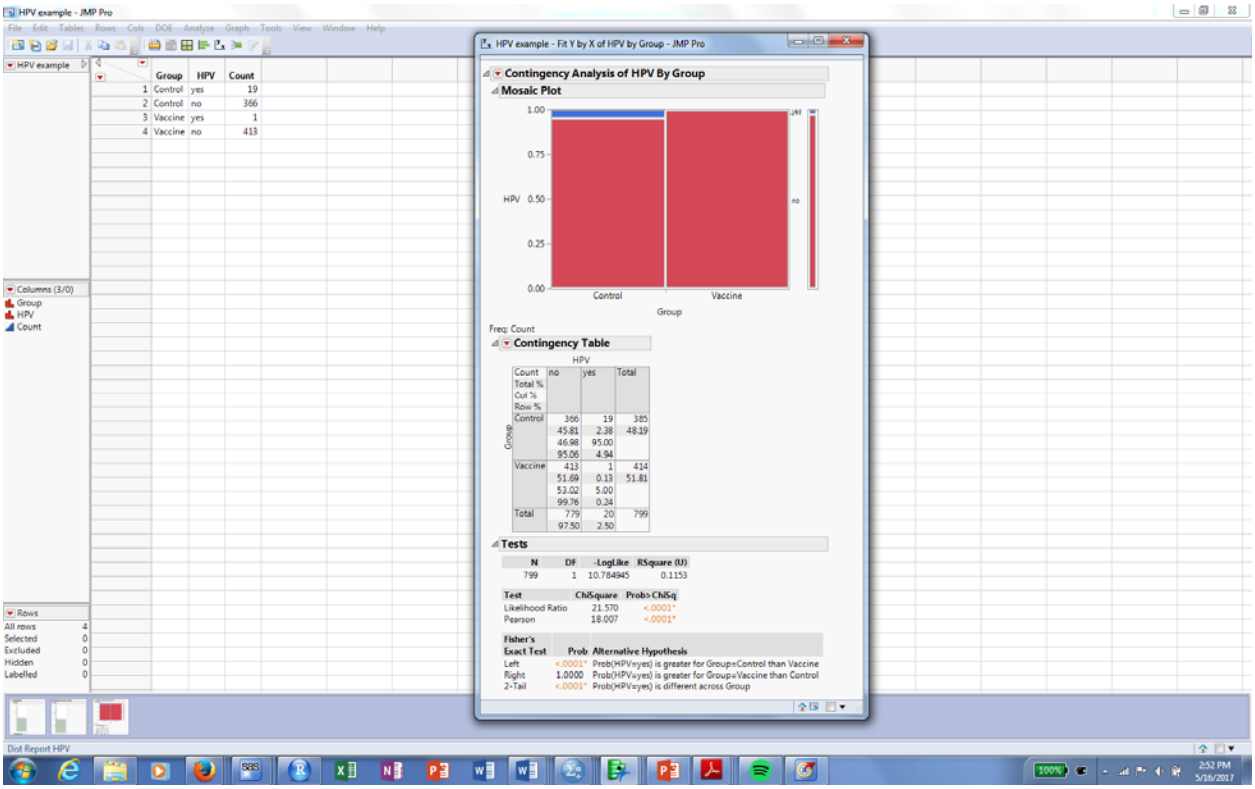
| Group | Infection | | |
|---|---|---|---|
| | No | Yes | Total |
| Control | 366 | 19 | 385 |
| Vaccine | 413 | 1 | 414 |
| Total | 779 | 20 | 799 |

Expected table:

| Group | Infection | |
|---|---|---|
| | No | Yes |
| Control | 385*779/799 = 375.36 | 385*20/799 = 9.64 |
| Vaccine | 414*779/799 = 403.64 | 414*20/799 = 10.36 |

(Pearson) chi-squared test statistic is the sum across all cells in the table, of $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$. For the HPV example, the value of the test statistic is $X^2 = 18.007$ (this was the value given by the R output above). The p-value is usually based on the chi-squared distribution. All software packages will compute this statistic and corresponding p-value.

JMP

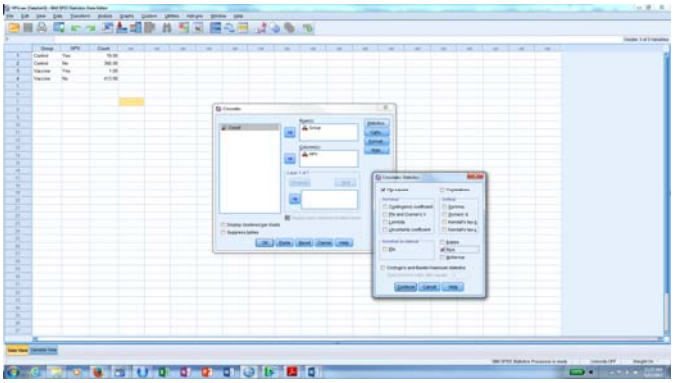



SAS

```
proc freq data=gibbs;
weight count;
tables Group*HPV /
  chisq /*chi-squared test*/;
run;
```

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 18.0068 | <.0001 |
| Likelihood Ratio Chi-Square | 1 | 21.5699 | <.0001 |
| Continuity Adj. Chi-Square | 1 | 16.1350 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 17.9843 | <.0001 |
| Phi Coefficient | | -0.1501 | |
| Contingency Coefficient | | 0.1485 | |
| Cramer's V | | -0.1501 | |

SPSS



Chi-Square Tests

| | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 18.007[a] | 1 | .000 | | |
| Continuity Correction[b] | 16.135 | 1 | .000 | | |
| Likelihood Ratio | 21.570 | 1 | .000 | | |
| Fisher's Exact Test | | | | .000 | .000 |
| N of Valid Cases | 799 | | | | |

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 9.64.
b. Computed only for a 2x2 table

*1.3.2. More than 2 rows/columns*

Example: Alsunni et. al (2014) studied the relationship between patient misconceptions about diabetes with several sociodemographic variables. One such variable was age group, and they obtained the following data:

| Age Group | Misconception score | | | |
|---|---|---|---|---|
| | Low | Moderate | High | Total |
| <20 | 16 | 14 | 2 | 32 |
| 21-40 | 32 | 28 | 2 | 62 |
| 41-60 | 56 | 24 | 1 | 81 |
| >60 | 11 | 11 | 3 | 25 |
| Total | 115 | 77 | 8 | 200 |

2 Types of tests—

1. "homogeneity"—"ANOVA-type" hypothesis, where one variable represents a factor and the other a response,
2. "independence"—"correlation-type" hypothesis, where a single sample is measured on two variables

Computation is exactly the same, however.

Misconception score example.

1. Hypotheses: $H_0$ : Misconception score is not associated with age
   $H_A$ : Misconception score is associated with age
2. Test statistic: $X^2 = 12.228$ ; p-value (based on chi-squared distribution with 6 df) = 0.057.

SAS

```
data alsunni_age;
input age score count @@;
datalines;
1 1 16 1 2 14 1 3 2
2 1 32 2 2 28 2 3 2
3 1 56 3 2 24 3 3 1
4 1 11 4 2 11 4 3 3
;

proc freq data=alsunni_age;
weight count;
tables treat*resp / chisq;
run;
```

Statistics for Table of age by score

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 6 | 12.2285 | 0.0571 |
| Likelihood Ratio Chi-Square | 6 | 11.4164 | 0.0763 |
| Mantel-Haenszel Chi-Square | 1 | 0.3005 | 0.5836 |
| Phi Coefficient | | 0.2473 | |

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Contingency Coefficient | | 0.2400 | |
| Cramer's V | | 0.1748 | |
| WARNING: 33% of the cells have expected counts less than 5. (Asymptotic) Chi-Square may not be a valid test. | | | |

P-value will be approximately correct if sample size is large, or more precisely if expected cell frequencies are not too small.

1. Cochran (1952): "if any expected frequency is less than 1 or if more than 20% are less than 5, the approximation may be poor"
2. Conover (1999): "if any expected frequency is less than 0.5 or if most are less than 1, the approximation may be poor".

Alternatives?

1. Combine columns/rows

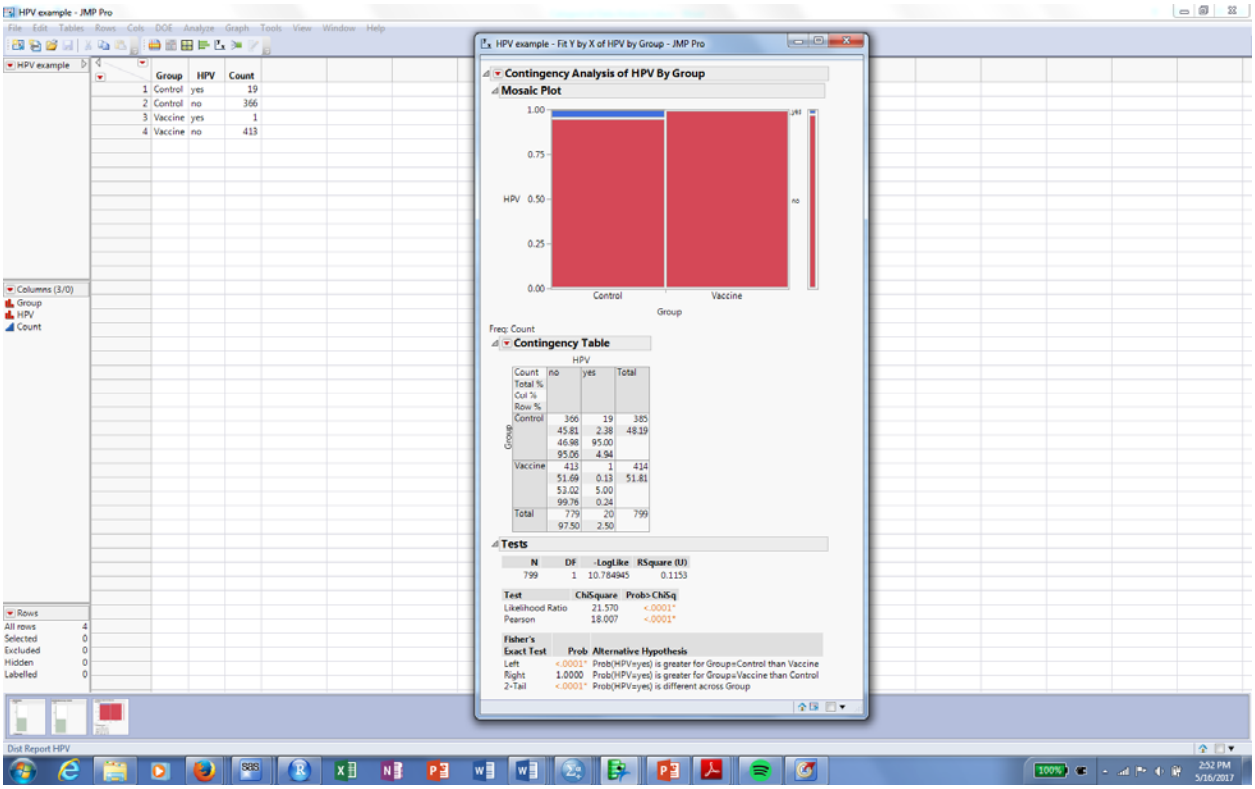
Misconception example. Combine Moderate and High categories.

| Age Group | Misconception score | | |
|---|---|---|---|
| | Low | Moderate/High | Total |
| <20 | 16 | 16 | 32 |
| 21-40 | 32 | 30 | 62 |
| 41-60 | 56 | 25 | 81 |
| >60 | 11 | 14 | 25 |
| Total | 115 | 77 | 200 |

Changes interpretation

2. Exact test
   a. $2 \times 2$ table--Fisher's Exact test (usually output by default)
   b. $R \times C$ table—Permutation test

HPV example--JMP



Misconception example--SAS

```
data alsunni_age;
input age score count @@;
datalines;
1 1 16 1 2 14 1 3 2
2 1 32 2 2 28 2 3 2
3 1 56 3 2 24 3 3 1
4 1 11 4 2 11 4 3 3
;

proc freq data=alsunni_age;
weight count;
exact chisq;
tables age*score / chisq;
run;
```

| Pearson Chi-Square Test | |
|---|---|
| Chi-Square | 12.2285 |
| DF | 6 |
| Asymptotic Pr > ChiSq | 0.0571 |
| Exact Pr >= ChiSq | 0.0547 |

Notice that even though software printed a warning, the approximate p-value is very close to the exact p-value.

### 1.3. Measures of association

In the previous section a larger chi-squared statistic implied a stronger association, provided the degrees of freedom remains the same. In the Alsunni et. al (2014) example, the chi-squared statistic, with 6 df, was $X^2 = 12.23$, which corresponded to an exact p-value of 0.055. However, for a 3x3 table with 4 df, a chi-squared value of $X^2 = 12.23$ would correspond to a p-value of 0.016. Thus, it is clear that $X^2$ cannot easily be used as a measure of the degree of association across tables of different sizes. However, several measures have been proposed to do this.

*Phi coefficient*
For 2x2 tables, phi ranges between -1 and 1 and thus can measure "direction" of the association. For the 2x2 table

| | |
|---|---|
| $a$ | $b$ |
| $c$ | $d$ |

$$\varphi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

A positive value suggests higher proportions of responses on the diagonal (cells a and d), while a negative value suggests higher proportion on the off-diagonal. Perfect positive association occurs when b and c are both 0, while perfect negative when a and d are both 0

*Cramer's contingency coefficient*
Cramer's coefficient is defined as

$$C = \sqrt{\frac{X^2}{n(q-1)}},$$

where $q$ is the smaller of the number of rows and the number of columns. The value $n(q-1)$ is the maximum possible value of $X^2$ for a given set of fixed row and column totals.

HPV example

| Group | Infection | | |
|---|---|---|---|
| | No | Yes | Total |
| Control | 366 | 19 | 385 |
| Vaccine | 413 | 1 | 414 |
| Total | 779 | 20 | 799 |

$$\varphi = \frac{366*1 - 19*413}{\sqrt{(385)(414)(779)(20)}} = -0.15$$

The negative coefficient results from the fact that a higher proportion of control patients had infections while a higher proportion in the vaccine group did not.

$$C = \sqrt{\frac{18.0068}{799(1)}} = 0.15$$

SAS

```
proc freq data=gibbs;
weight count;
tables Group*HPV /
  chisq /*chi-squared test*/;
run;
```

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 18.0068 | <.0001 |
| Likelihood Ratio Chi-Square | 1 | 21.5699 | <.0001 |
| Continuity Adj. Chi-Square | 1 | 16.1350 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 17.9843 | <.0001 |
| Phi Coefficient | | -0.1501 | |
| Contingency Coefficient | | 0.1485 | |
| Cramer's V | | -0.1501 | |

Alsunni et. al (2014) example.

Statistics for Table of age by score

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 6 | 12.2285 | 0.0571 |
| Likelihood Ratio Chi-Square | 6 | 11.4164 | 0.0763 |
| Mantel-Haenszel Chi-Square | 1 | 0.3005 | 0.5836 |
| Phi Coefficient | | 0.2473 | |
| Contingency Coefficient | | 0.2400 | |
| Cramer's V | | 0.1748 | |
| WARNING: 33% of the cells have expected counts less than 5. (Asymptotic) Chi-Square may not be a valid test. | | | |

## 2. Nominal/ordinal association

### 2.1 Comparing groups on an ordinal variable—independent samples

Rank tests for comparing groups can be used

Wilcoxon rank-sum/Mann-Whitney test (2 groups)

Kruskal-Wallis test (3 or more groups)

Misconception score example.

Hypotheses: $H_0$ : Misconception score is not associated with age

$H_A$ : Misconception score is associated with age

Since there are 4 age groups, Kruskal-Wallis test is performed:

Test statistic: $KW = 9.0896$ ; p-value = 0.0271 (exact)/ 0.0281 (based on chi-squared distribution with 3 df).

Stronger evidence of association than chi-squared test (p-value = 0.0571)

SAS

```
proc npar1way data=alsunni_age
        wilcoxon /*request WRS/MW/KW test*/;
class age;
var score;
freq count;
exact wilcoxon /*Calculate exact p-value*/;
run;
```

Wilcoxon Scores (Rank Sums) for Variable score
Classified by Variable age

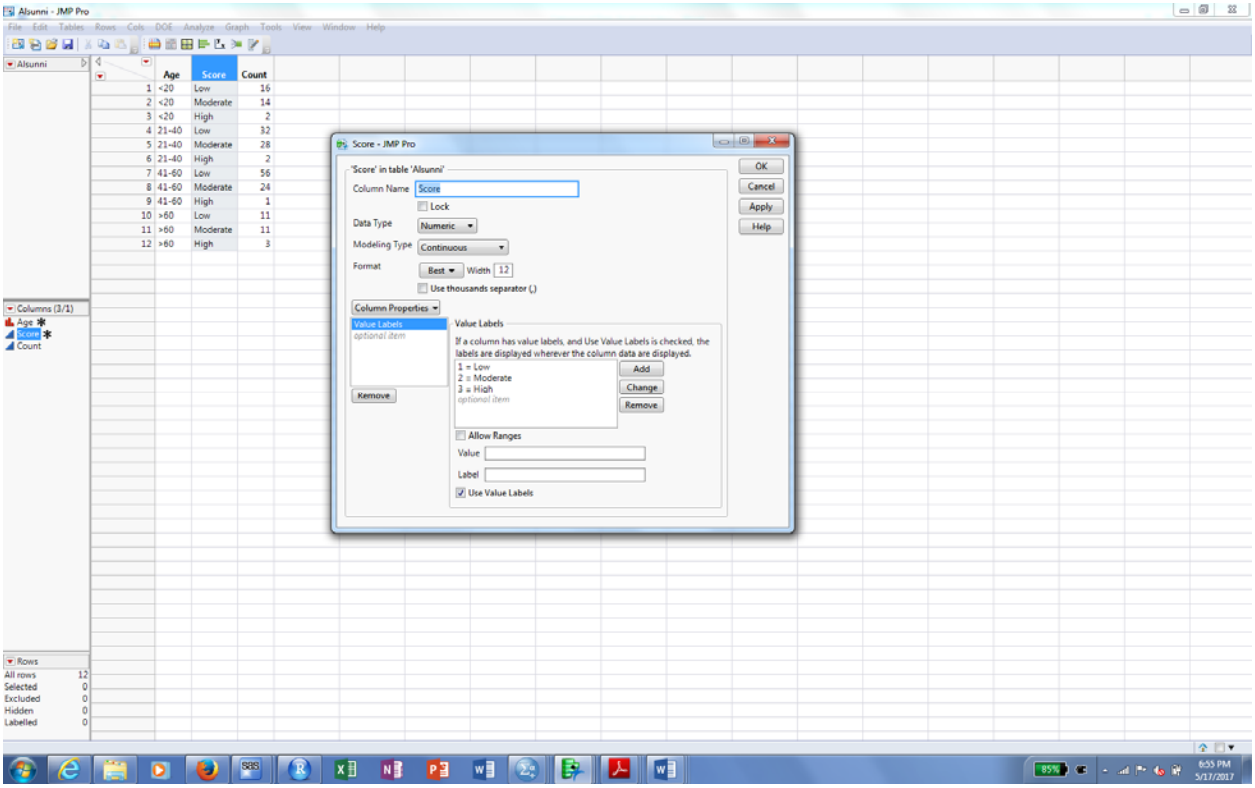
| age | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
|---|---|---|---|---|---|
| 1 | 32 | 3477.00 | 3216.00 | 260.357883 | 108.656250 |
| 2 | 62 | 6561.00 | 6231.00 | 328.455463 | 105.822581 |
| 3 | 81 | 7140.50 | 8140.50 | 348.623847 | 88.154321 |
| 4 | 25 | 2921.50 | 2512.50 | 234.871396 | 116.860000 |

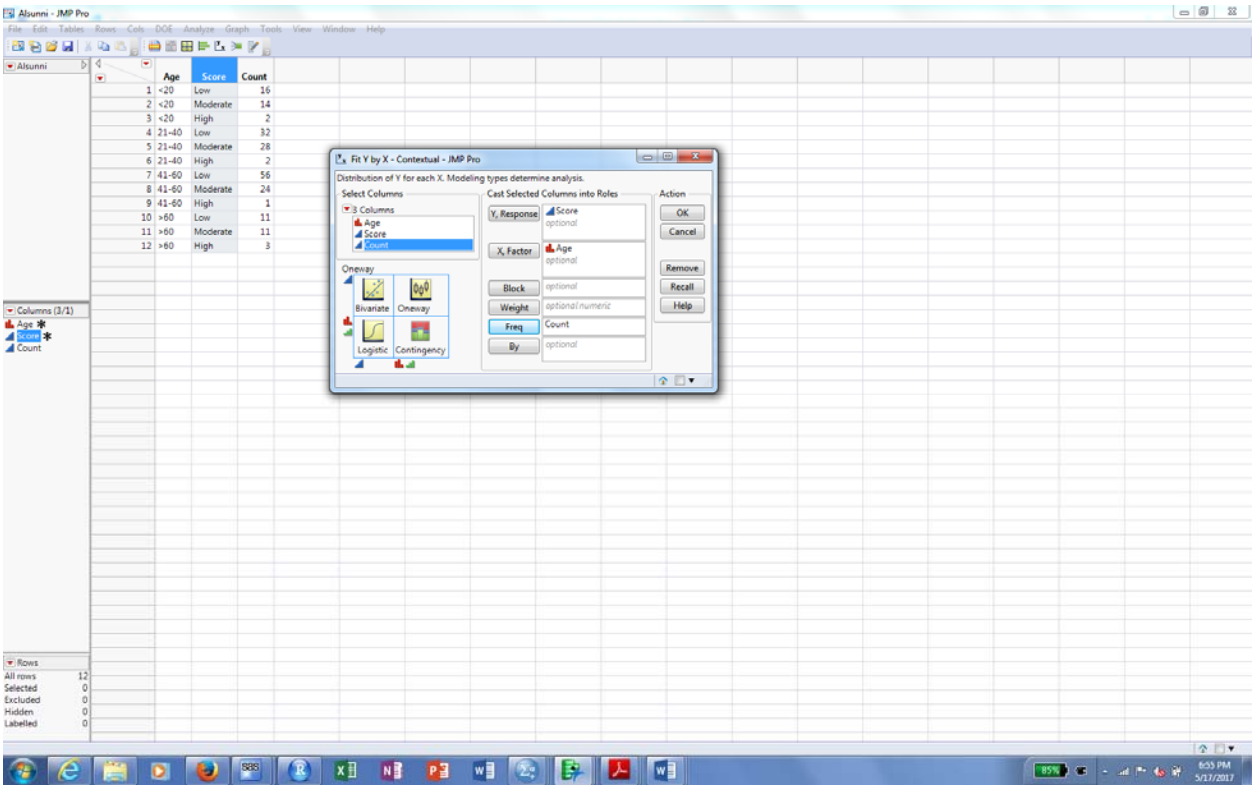
Average scores were used for ties.

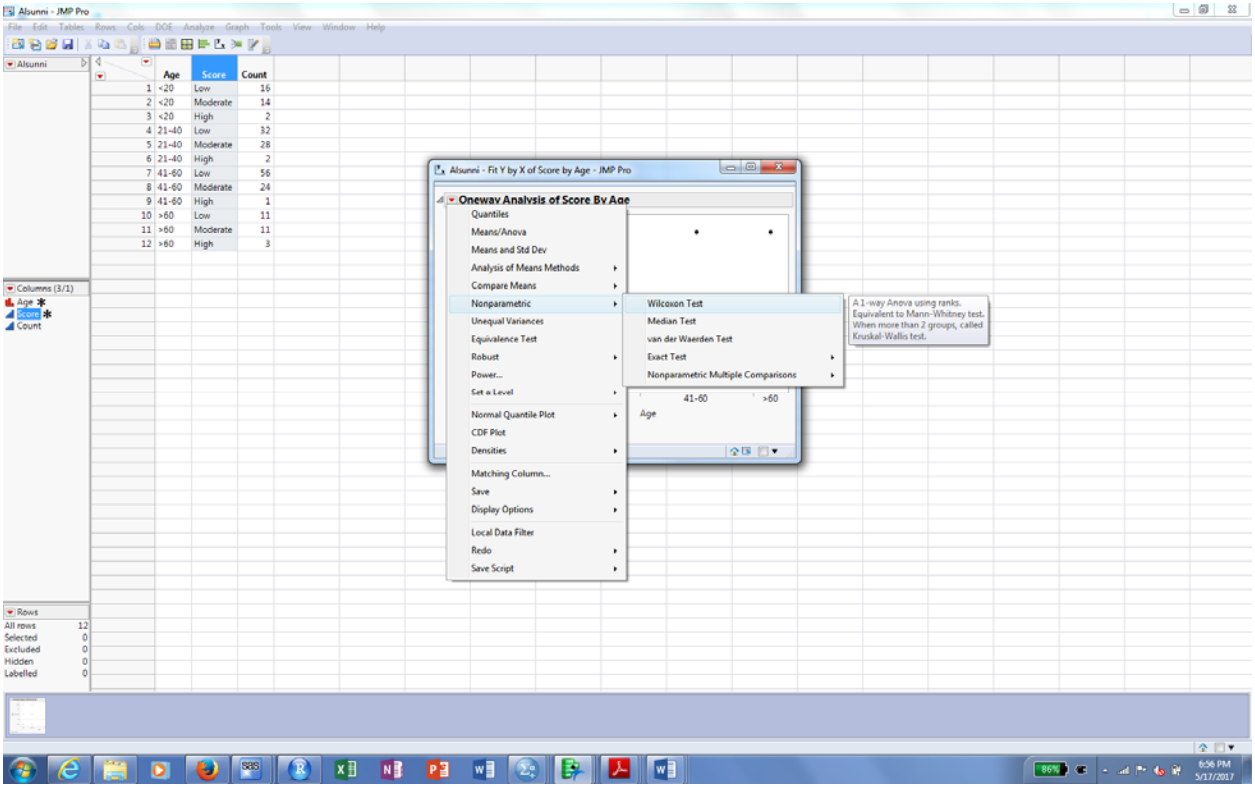
| | |
|---|---|
| Kruskal-Wallis Test | |
| Chi-Square | 9.0896 |
| DF | 3 |
| Asymptotic Pr > Chi-Square | 0.0281 |
| Exact Pr >= Chi-Square | 0.0271 |

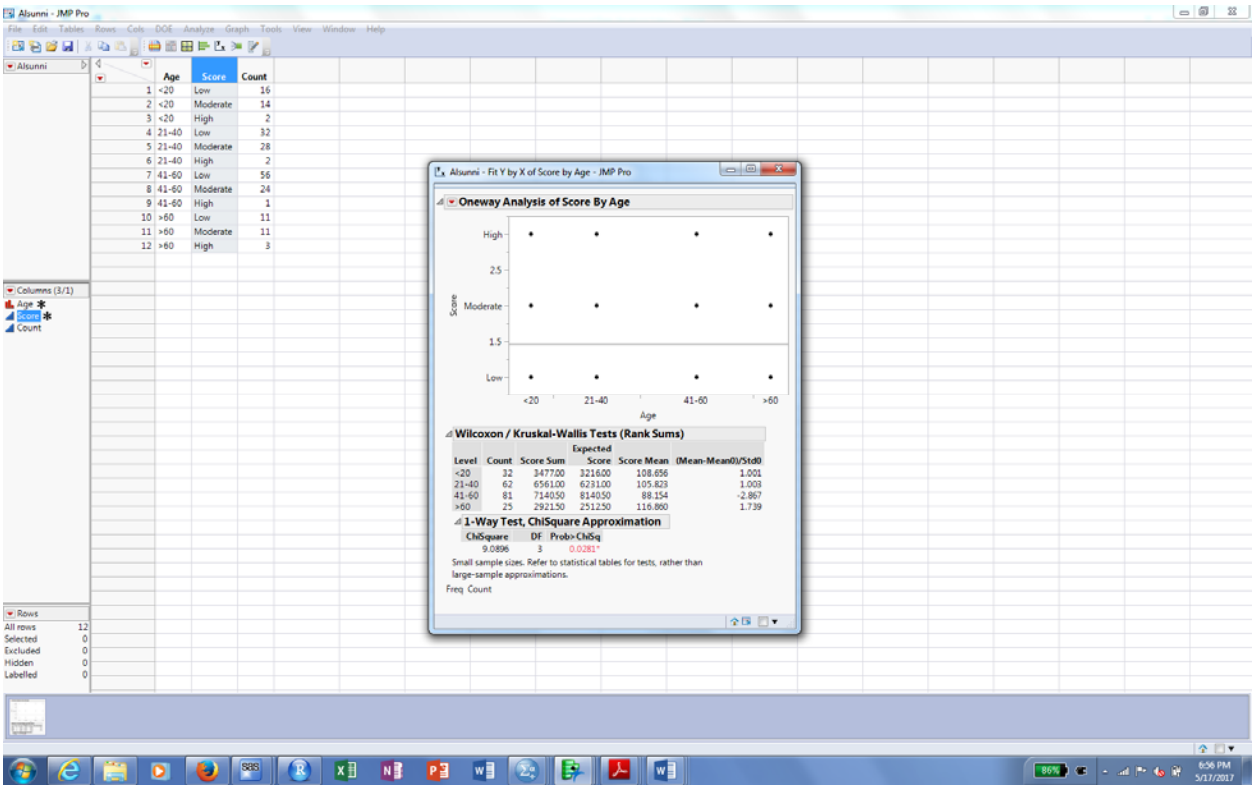
JMP

Note: Response (Y) variable must by identified as continuous, Explanatory (X) as Nominal.




Alsunni - JMP Pro
Fit Y by X - Contextual - JMP Pro
Distribution of Y for each X. Modeling types determine analysis.
Select Columns
Cast Selected Columns into Roles
Action
Y, Response
X, Factor
Block
Weight
Freq
By
OK
Cancel
Remove
Recall
Help
Bivariate
Oneway
Logistic
Contingency


Alsunni - JMP Pro
Alsunni - Fit Y by X of Score by Age - JMP Pro
Onewav Analysis of Score By Age
Quantiles
Means/Anova
Means and Std Dev
Analysis of Means Methods
Compare Means
Nonparametric
Unequal Variances
Equivalence Test
Robust
Power...
Set a Level
Normal Quantile Plot
CDF Plot
Densities
Matching Column...
Save
Display Options
Local Data Filter
Redo
Save Script
Wilcoxon Test
Median Test
van der Waerden Test
Exact Test
Nonparametric Multiple Comparisons
A 1-way Anova using ranks.
Equivalent to Mann-Whitney test.
When more than 2 groups, called
Kruskal-Wallis test.
41-60
>60
Age




The WRS/MW/KW tests are usually thought of in the same way as T/ANOVA tests as for testing for group differences, rather than testing for association. However, the distinction only affects interpretation of test results. However, as with the $X^2$ statistic, it is difficult to use these tests statistics to compare degree of association between different data sets.

## 3. Ordinal-ordinal association

Several rank-based methods

1. Spearman correlation—Pearson correlation on rank scores
2. Kendall's tau—measure of "concordance
   (Called the Jonckheere Terpstra test if testing for group differences)

Both are measures of either increasing or decreasing (monotonic) association, range between -1 and 1, and yield similar p-values.

Misconception example.

Spearman and Kendall coefficients are -0.060 and -0.056, respectively, with large sample p-values 0.399 and 0.387, respectively. Thus, there is not evidence of monotonic association between age and misconception score. That is, there is not statistical evidence that misconception score tends to increase or decrease with age.

SAS

```
proc corr data=alsunni_age spearman kendall;
var age score;
freq count;
run;

proc freq data=alsunni_age;
weight count;
exact measures jt;
tables age*score / measures jt;
run;
```

Spearman Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

| | age | score |
|---|---|---|
| age | 1.00000 | -0.05993<br>0.3992 |
| score | -0.05993<br>0.3992 | 1.00000 |

Kendall Tau b Correlation Coefficients, N = 200
Prob > |tau| under H0: Tau=0

| | age | score |
|---|---|---|
| age | 1.00000 | -0.05577<br>0.3871 |
| score | -0.05577<br>0.3871 | 1.00000 |

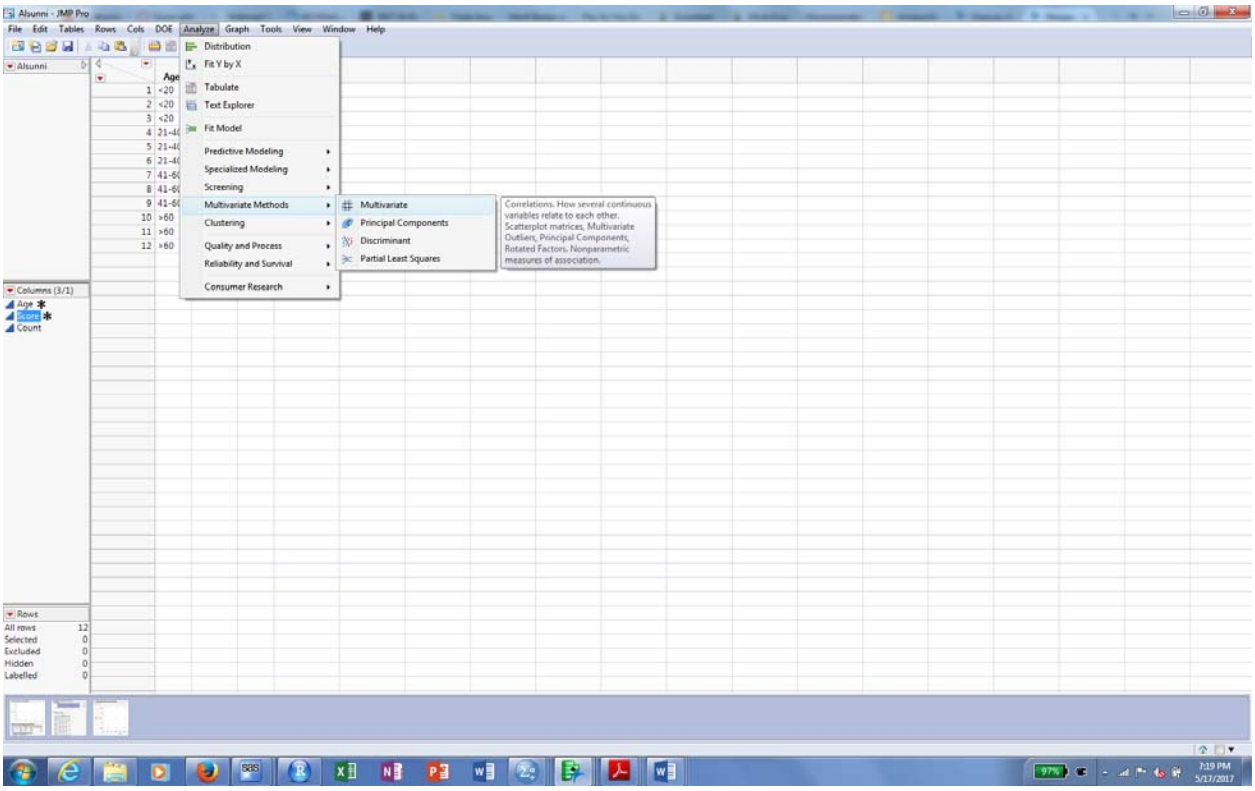
| Spearman Correlation Coefficient | |
|---|---|
| Correlation (r) | -0.0599 |
| ASE | 0.0749 |

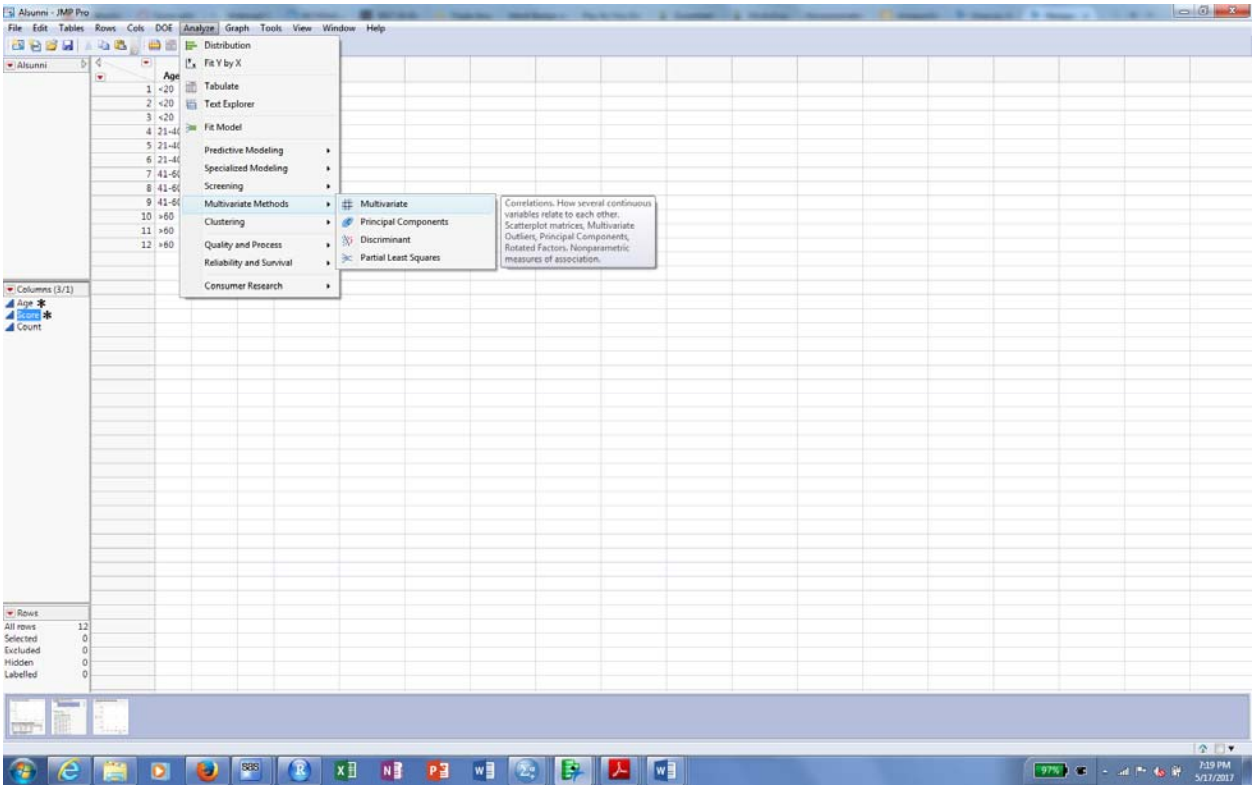
| Spearman Correlation Coefficient | |
|---|---|
| 95% Lower Conf Limit | -0.2067 |
| 95% Upper Conf Limit | 0.0868 |

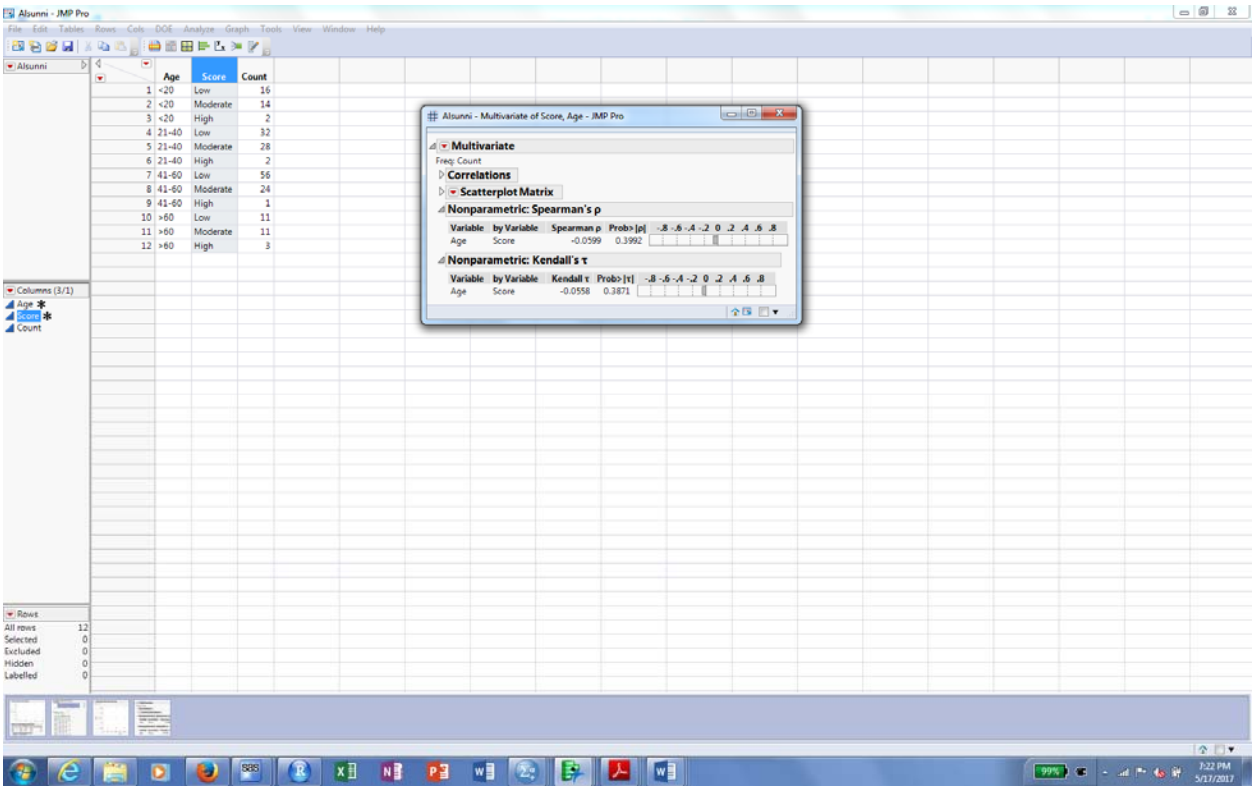
| Test of H0: Correlation = 0 | |
|---|---|
| ASE under H0 | 0.0748 |
| Z | -0.8012 |
| One-sided Pr < Z | 0.2115 |
| Two-sided Pr > \|Z\| | 0.4230 |
| | |
| Exact Test | |
| One-sided Pr <= r | 0.1994 |
| Two-sided Pr >= \|r\| | 0.3988 |

| Jonckheere-Terpstra Test | |
|---|---|
| Statistic (JT) | 6650.5000 |
| Z | -0.8649 |
| | |
| Asymptotic Test | |
| One-sided Pr < Z | 0.1936 |
| Two-sided Pr > \|Z\| | 0.3871 |
| | |
| Exact Test | |
| One-sided Pr <= JT | 0.1941 |
| Two-sided Pr >= \|JT - Mean\| | 0.3883 |

JMP
Both variables need to be recognized as continuous.






We found good statistical evidence of an association using the Kruskal-Wallis test, and moderate evidence using the chi-squared test, but virtually no evidence using rank correlations.

In general,

- if one variable is ordinal and the other nominal, the WRS/MW/KW test will have more power to detect an association than the chi-squared test
- if both variables are ordinal,
    - the WRS/MW/KW tests will have more power to detect an association than the chi-squared test
    - the Spearman/Kendall/JT tests will have more power than the chi-squared test to detect an increasing or decreasing association, but may have less power otherwise.

Hypothetical example

| Age Group | Misconception score | | | |
|---|---|---|---|---|
| | Low | Moderate | High | Total |
| <20 | 2(6%) | 14(44%) | 16(50%) | 32 |
| 21-40 | 6(10%) | 24(39%) | 32(52%) | 62 |
| 41-60 | 50(62%) | 25(31%) | 6(7%) | 81 |
| >60 | 20(80%) | 4(16%) | 1(4%) | 25 |
| Total | 78 | 67 | 55 | 200 |

Now the Spearman and Kendall coefficients are -0.607 and -0.534, respectively, with p-values less than 0.0001.

Spearman Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

| | age | score |
|---|---|---|
| age | 1.00000 | -0.60660 |
| | | <.0001 |
| score | -0.60660 | 1.00000 |
| | <.0001 | |

Kendall Tau b Correlation Coefficients, N = 200
Prob > |tau| under H0: Tau=0

| | age | score |
|---|---|---|
| age | 1.00000 | -0.53241 |
| | | <.0001 |
| score | -0.53241 | 1.00000 |
| | <.0001 | |

## 4. Comparing proportions—dependent samples

Mc Nemar's Test

Example. Participants are asked their preferred candidate before and after a debate. Each subject gives a response before and after:

| *Subject* | *Before* | *After* | |
|---|---|---|---|
| 1 | *A* | *A* | |
| 2 | *A* | *A* | |
| 3 | *A* | *A* | |
| 4 | *A* | *B* | ↙ |
| 5 | *A* | *B* | ↙ |
| 6 | *A* | *B* | ↙ |
| 7 | *A* | *B* | ↙ |
| 8 | *A* | *B* | ↙ |
| 9 | *A* | *B* | ↙ |
| 10 | *A* | *B* | ↙ |
| 11 | *A* | *B* | ↙ |
| 12 | *A* | *B* | ↙ |
| 13 | *B* | *A* | ↙ |
| 14 | *B* | *A* | ↙ |
| 15 | *B* | *B* | |
| 16 | *B* | *B* | |
| 17 | *B* | *B* | |
| 18 | *B* | *B* | |
| 19 | *B* | *B* | |
| 20 | *B* | *B* | |

Observed Table

| | $A$ | $B$ |
|---|---|---|
| $A$ | $X_{AA}$ | $X_{AB}$ |
| $B$ | $X_{BA}$ | $X_{BB}$ |

Population Table

| | |
|---|---|
| $P_{AA}$ | $P_{AB}$ |
| $P_{BA}$ | $P_{BB}$ |

$P_{A\cdot} = P_{AA} + P_{AB}$

$P(A$ is first response$)$

$P_{\cdot A} = P_{AA} + P_{BA}$

$P(A$ is second response$)$

If there is no effect of the debate, then $A$ is equally likely to be chosen before and after, i.e. $P_{\cdot A} = P_{A\cdot}$.

$H_0 : P_{\cdot A} = P_{A\cdot}$ or $P_{AA} + P_{AB} = P_{AA} + P_{BA}$ or $P_{AB} = P_{BA}$

Test Statistic: $T = X_{AB} = \#$ switched from $A$ to $B$. We can consider just people who switched (The rest are "ties"). Then under $H_0$ the switches to $B$ are just as likely as to $A$. So, we can calculate a one-sided p-value as $P(X \geq X_{AB} \mid n = X_{AB} + X_{BA}, p = .5)$.

Example

$9 + 2 = 11$ people switched, and of those $X_{AB} = 9$ switched to $B$.

$P(X \geq 9 \mid n = 11, p = .25) = .027 + .005 + .000 = .032$

Here the alternative is that more likely to switch to $B$, or $H_a : P_{\cdot A} < P_{A\cdot}$

SAS

```
data ta5_8_1;
input before $ after $ count @@;
datalines;
A A 3 A B 9
B A 2 B B 6
;

proc freq data=ta5_8_1;
weight count;
exact mcnem;                    /* Requests McNemar test, exact p-value */
tables before*after;
run;
```

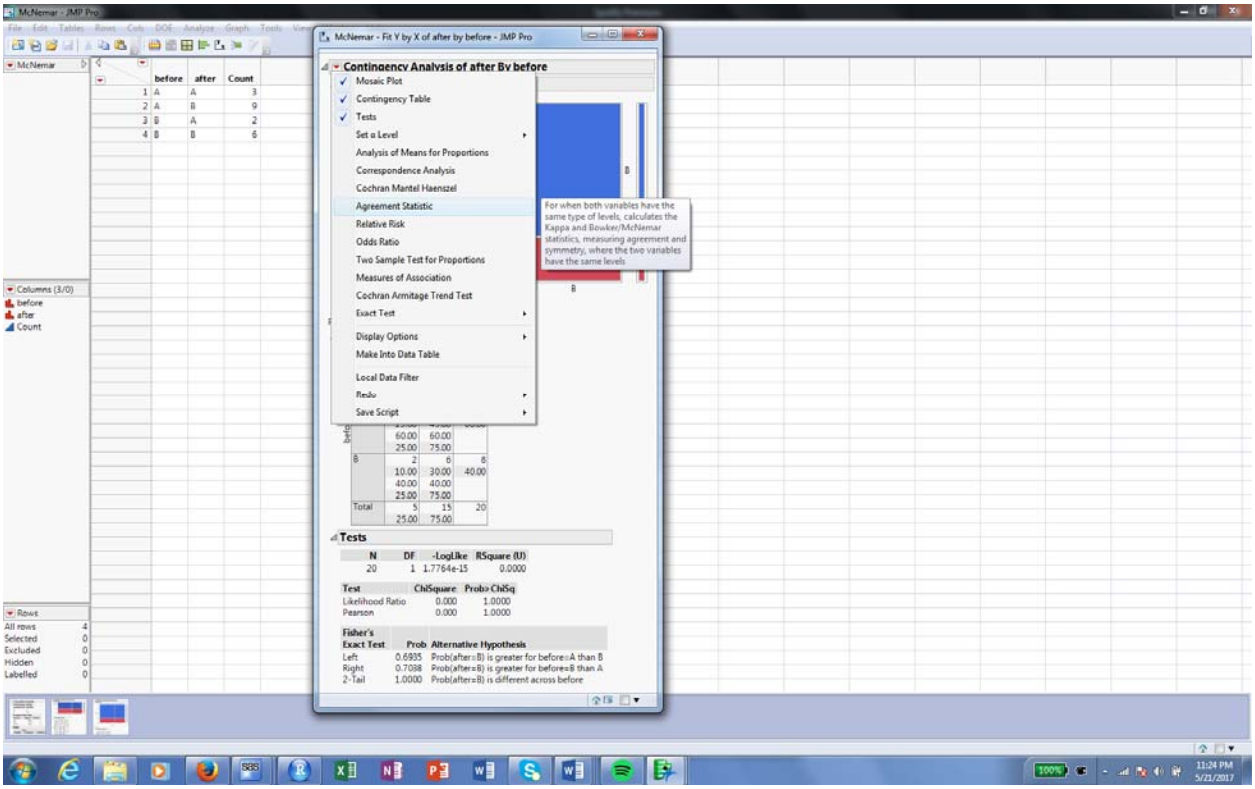
Statistics for Table of before by after

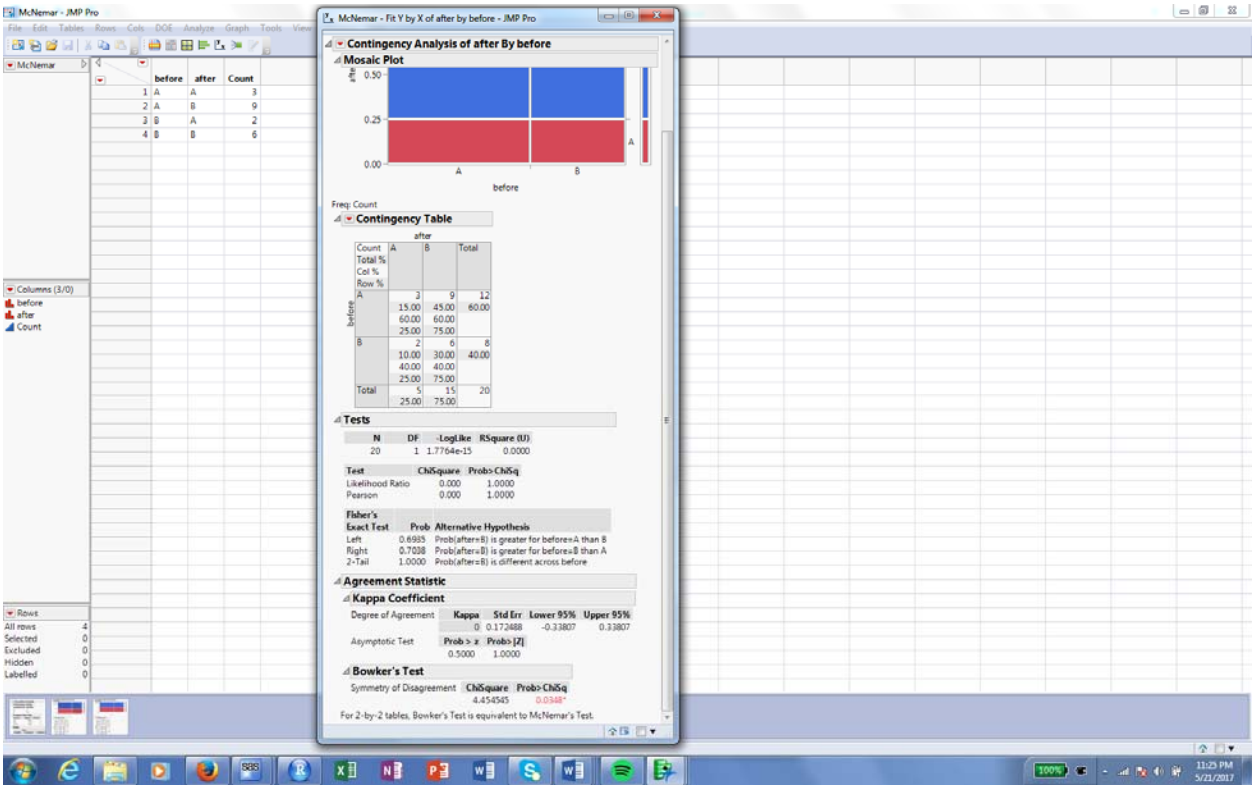
| McNemar's Test | |
|---|---|
| Statistic (S) | 4.4545 |
| DF | 1 |
| Asymptotic Pr > S | 0.0348 |
| Exact Pr >= S | 0.0654 |

R

```
table <- matrix(
c(3, 9,
2, 6),
nrow = 2, byrow = TRUE,
dimnames = list(
"First" = c("A", "B"),
"Second" = c("A", "B")
)
)
library(coin)

## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:epitools':
##
##     ratetable

mh_test(as.table(table), distribution = "exact")

##
##  Exact Marginal Homogeneity Test
##
## data:  response by
##   conditions (First, Second)
##   stratified by block
## chi-squared = 4.4545, p-value = 0.06543
```

JMP


McNemar - JMP Pro
McNemar - Fit Y by X of after by before - JMP Pro
Contingency Analysis of after By before
Mosaic Plot
Contingency Table
Tests
Set α Level
Analysis of Means for Proportions
Correspondence Analysis
Cochran Mantel Haenszel
Agreement Statistic
Relative Risk
Odds Ratio
Two Sample Test for Proportions
Measures of Association
Cochran Armitage Trend Test
Exact Test
Display Options
Make Into Data Table
Local Data Filter
Redo
Save Script
For when both variables have the same type of levels, calculates the Kappa and Bowker/McNemar statistics, measuring agreement and symmetry, where the two variables have the same levels
Tests
Likelihood Ratio
Pearson
Fisher's Exact Test
Left
Right
2-Tail


McNemar - JMP Pro
McNemar - Fit Y by X of after by before - JMP Pro
Contingency Analysis of after By before
Mosaic Plot
Freq: Count
Contingency Table
Tests
Likelihood Ratio
Pearson
Fisher's Exact Test
Left
Right
2-Tail
Agreement Statistic
Kappa Coefficient
Degree of Agreement
Asymptotic Test
Bowker's Test
Symmetry of Disagreement
For 2-by-2 tables, Bowker's Test is equivalent to McNemar's Test.