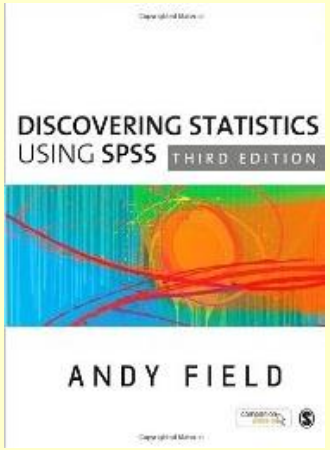
Analysis of Categorical Data

	A	a
A	AA Red	Aa Pink
a	Aa Pink	aa White

http://www.pelagicos.net/classes_biometry_fa16.htm

Reading - Field: Chapter 18

AIMS



- Analysis of Categorical Variables
- Contingency Tables
- Chi-Square Test, Likelihood Ratio, Exact Test
- Odds Ratio

Variable Types

- Sometimes we have data consisting of the frequency of cases scored by categories
- Examples:
  - Number of votes for different politicians
  - Number of goals scored by each player

Testing Hypotheses

- These categorical data can be used to test patterns and hypotheses in two ways:

- Example 1: Comparing several observations
  Numbers of students who pass or fail a final exam in different subject areas

- Example 2: Comparing pattern from the experimental treatment, against a control
  Number of treated patients or waiting list controls with different diagnosis following a given treatment

Variable Types in SPSS

- Numeric vs Categorical
  (e.g., 7, 0, 120) (e.g., 'Color', 'Sex')

- Analyzing categorical variables
  - Mean of categorical variables meaningless
    e.g., (blue + red) / 2 = ?

  - Because the numeric values you attach to different categories are arbitrary

Categorical Data Analysis – Example

Instead, analyze frequencies of categorical data:

- Example: Can animals be trained to dance with different rewards ?
  - Participants: 200 cats
  - Training
    - Animal trained using either food or affection, not both)
  - Dance
    - Animal either learnt to dance or it did not.
  - Outcome:
    - Number of animals (frequency) that could dance or could not dance in both reward condition.
  - We tabulate these frequencies in a **contingency table**.

Contingency Table – Example

Tabulate frequencies of categorical data using a contingency table:

Training vs Outcome:

		Training		
		Food as Reward	*Affection as Reward*	*Total*
Could They Dance?	Yes	28	48	76
	No	10	114	124
	Total	38	162	200

What is the null hypothesis?

No Pattern or Association:
Training and Outcome are independent.

Contingency Table – Example

		Training		
		Food as Reward	Affection as Reward	Total
Could They Dance?	Yes	28	48	76
	No	10	114	124
	Total	38	162	200

Pearson Chi-Square Test:

- Determine whether there is a relationship between two categorical variables

- Compare frequencies observed in certain categories to frequencies you expect in those categories by chance

Pearson Chi-Square Test – Example

- The Formula:

$$\chi^2 = \sum \frac{\left(Observed_{ij} - Model_{ij}\right)^2}{Model_{ij}}$$

  - i represents rows in table and j represents columns
  - observed data are frequencies in contingency table
  - model data are expected frequencies (from Ho)

- The 'Model' is based on the 'expected frequencies'

$$\text{Model}_{ij} = E_{ij} = \frac{\text{Row Total}_i \times \text{Column Total}_j}{n}$$

  - Calculated for each cell in the contingency table
  - n is total number of observations (in this case 200)

Pearson Chi-Square Test – Example

Observed Values:

		Training		
		Food as Reward	Affection as Reward	Total
Could They Dance?	Yes	28	48	76
	No	10	114	124
	Total	38	162	200

Expected Values:

$$\text{Model}_{\text{Food, Yes}} = \frac{\text{RT}_{\text{Yes}} \times \text{CT}_{\text{Food}}}{n} = \frac{76 \times 38}{200} = 14.44$$

$$\text{Model}_{\text{Food, No}} = \frac{\text{RT}_{\text{No}} \times \text{CT}_{\text{Food}}}{n} = \frac{124 \times 38}{200} = 23.56$$

$$\text{Model}_{\text{Affection, Yes}} = \frac{\text{RT}_{\text{Yes}} \times \text{CT}_{\text{Affection}}}{n} = \frac{76 \times 162}{200} = 61.56$$

$$\text{Model}_{\text{Affection, No}} = \frac{\text{RT}_{\text{No}} \times \text{CT}_{\text{Affection}}}{n} = \frac{124 \times 162}{200} = 100.44$$

Pearson Chi-Square Test – Example

Observed Values:

		Training		
		Food as Reward	Affection as Reward	Total
Could They Dance?	Yes	28	48	76
	No	10	114	124
	Total	38	162	200

Expected Values:

		Training		
		Food as Reward	Affection as Reward	Total
Could They Dance?	Yes	14.44	61.56	76
	No	23.56	100.44	124
	Total	38	162	200

Pearson Chi-Square Test – Example

- Test Statistic: Summed squared observed deviations, divided by model predictions:
- Sum of four terms (one for each cell in the table):

(28 – 14.44) ^ 2 divided by 14.44
(48 – 61.56) ^ 2 divided by 61.56
(10 – 23.56) ^ 2 divided by 23.56
(114 – 100.44) ^ 2 divided by 100.44

- Test Statistic
  - Checked against a chi-square distribution with $(r - 1)(c - 1) = 1$ degrees of freedom
  - If significant, then there is a significant association between the categorical variables in the population
  - The test distribution is approximate, so for small samples, use the **Fisher's exact test**

Pearson Chi-Square Test – Example

- The Chi-square test statistic gives an 'overall' result
- We can using investigate further using standardized residuals. Yet, we need to consider two important considerations about these standardized residuals:
  - Standardized residuals have direct relationship with the test statistic (provide standardized difference of observed / expected frequencies)
  - These standardized values are z-scores (e.g. values outside of ±1.96 significant at $p < 0.05$)

- Effect Size
  - **Odds ratio** can be used as an effect size measure

The Odds Ratio – Example

The **odds ratio** is a measure of effect size, describing the strength of association or non-independence between two binary data values

NOTE: Odds ratio works only for two outcomes

It is used as a descriptive statistic in contingency tests and logistic regression (binary outcome)

The odds ratio treats the two variables being compared symmetrically

Pearson Chi-Square Test - Example

- The Odds Ratio: Compares frequencies of each outcome, given a shared condition. NOTE: This is not a probability

$$\text{Odds}_{\text{dancing after food}} = \frac{\text{Number that had food and danced}}{\text{Number that had food but didn't dance}} = \frac{28}{10} = 2.8$$

$$\text{Odds}_{\text{dancing after affection}} = \frac{\text{Number that had affection and danced}}{\text{Number that had affection but didn't dance}} = \frac{48}{114} = 0.421$$

Quantifies effect size of treatment:
- 2.8 more likely to dance after food
- 0.421 times more likely to dance after affection

$$\text{Odds Ratio} = \frac{\text{Odds}_{\text{dancing after food}}}{\text{Odds}_{\text{dancing after affection}}} = \frac{2.8}{0.421} = 6.65$$

Larger than 1

Pearson Chi-Square Test – Assumptions

- The chi-square test has two important assumptions:
  - Independence:
    - Each person, item or entity contributes to only one cell of the contingency table
  - The expected frequencies should be greater than 5
    - In larger contingency tables up to 20% of expected frequencies can fall below 5, with a slight loss of statistical power
    - Even in larger contingency tables, no expected frequencies should fall below 1
    - In this situation: combine cells (treatments) for chi-square OR use Fisher's exact test (default)

Pearson Chi-Square Test - SPSS

Training	Dance	Frequency
0	0	28
0	1	10
1	0	48
1	1	114

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
Training	Numeric	8	0	Type of Training	{0, Food as Reward}...	None	14	Right	Nominal
Dance	Numeric	8	0	Did they dance?	{0, Yes}...	None	8	Right	Nominal
Frequency	Numeric	8	0	Frequency	None	None	8	Right	Scale

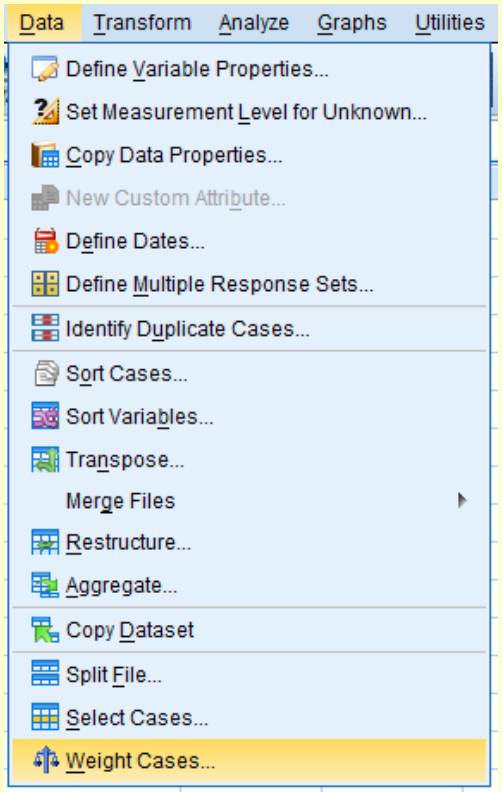
- Training (0 or 1)
- Dance (0 or 1)

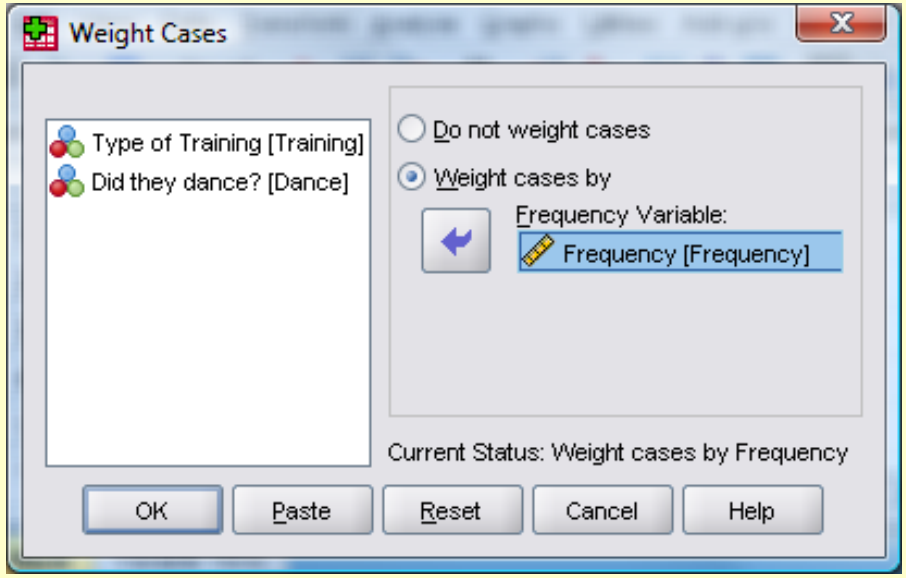
Each cells reports the frequency (count of cats)

Pearson Chi-Square Test - Step 1

Weighing Cases: Use the observed frequency data to calculate the expected values

Go to Data > Weight Cases

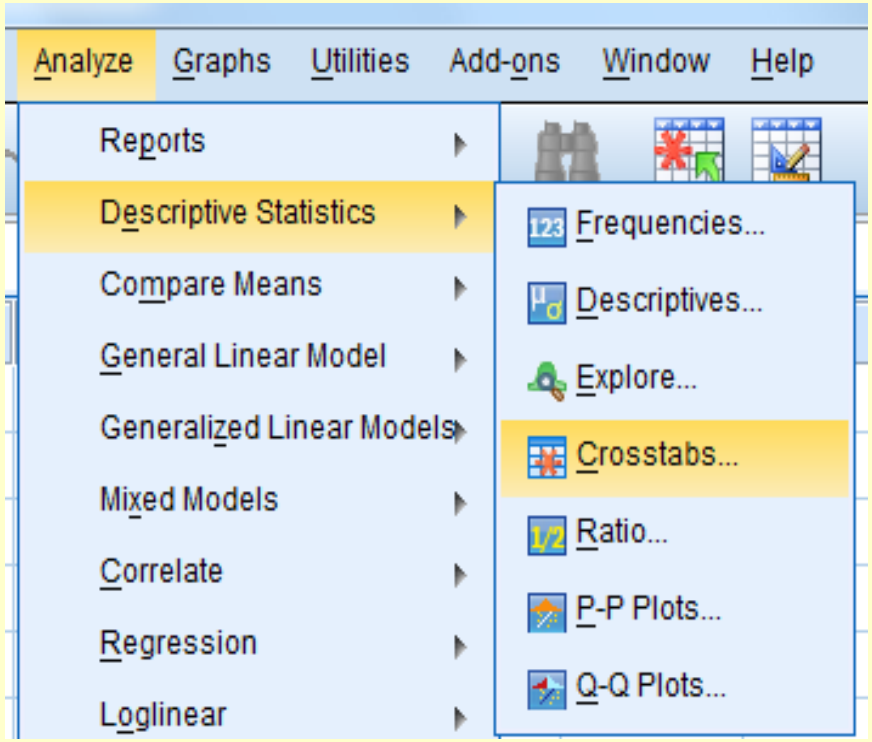




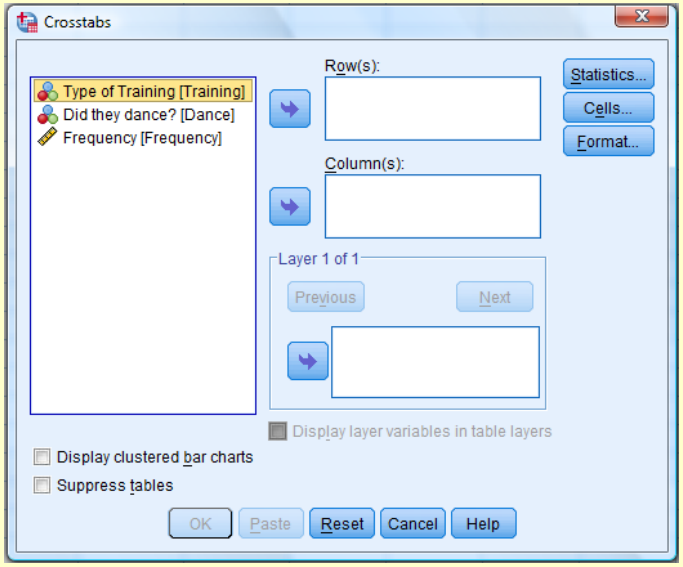

```
GET
  FILE='F:\Biometry\HW#8\CatsWeight.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
WEIGHT BY Frequency.
```

Pearson Chi-Square Test - Step 2

Go to Analyze > Descriptive Statistics > Crosstabs

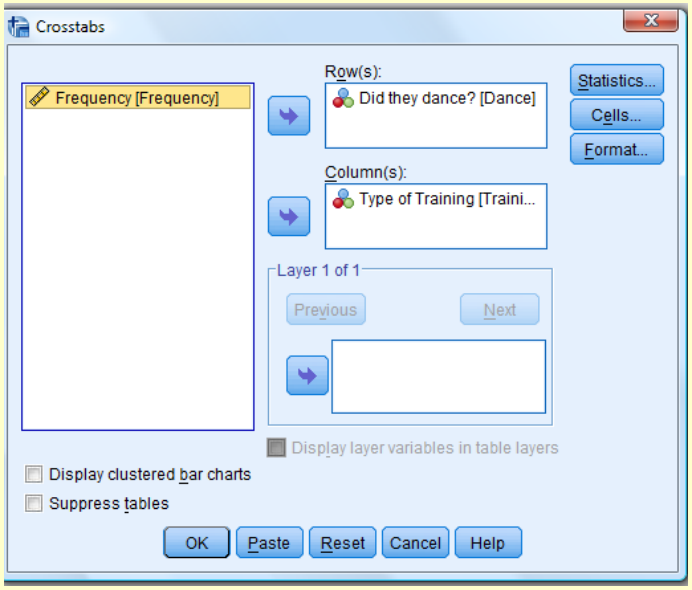


Pearson Chi-Square Test - Step 3



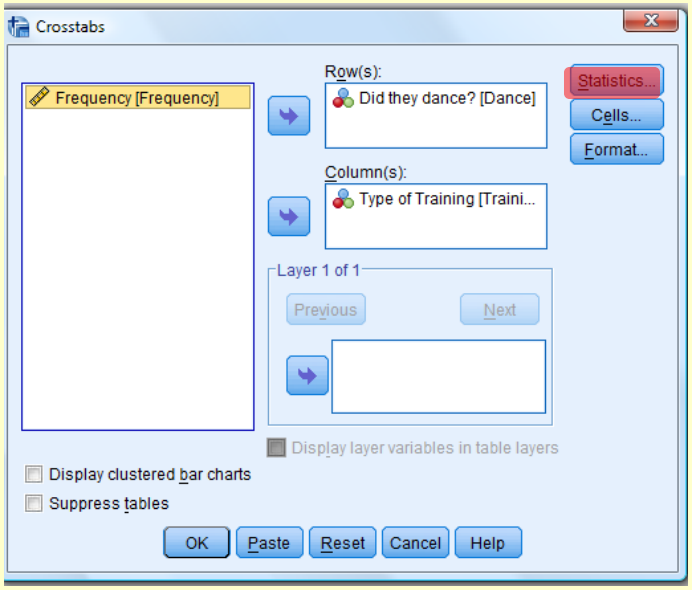
Select Row(s) / Column(s):

- Training (0 or 1)
- Dance (0 or 1)

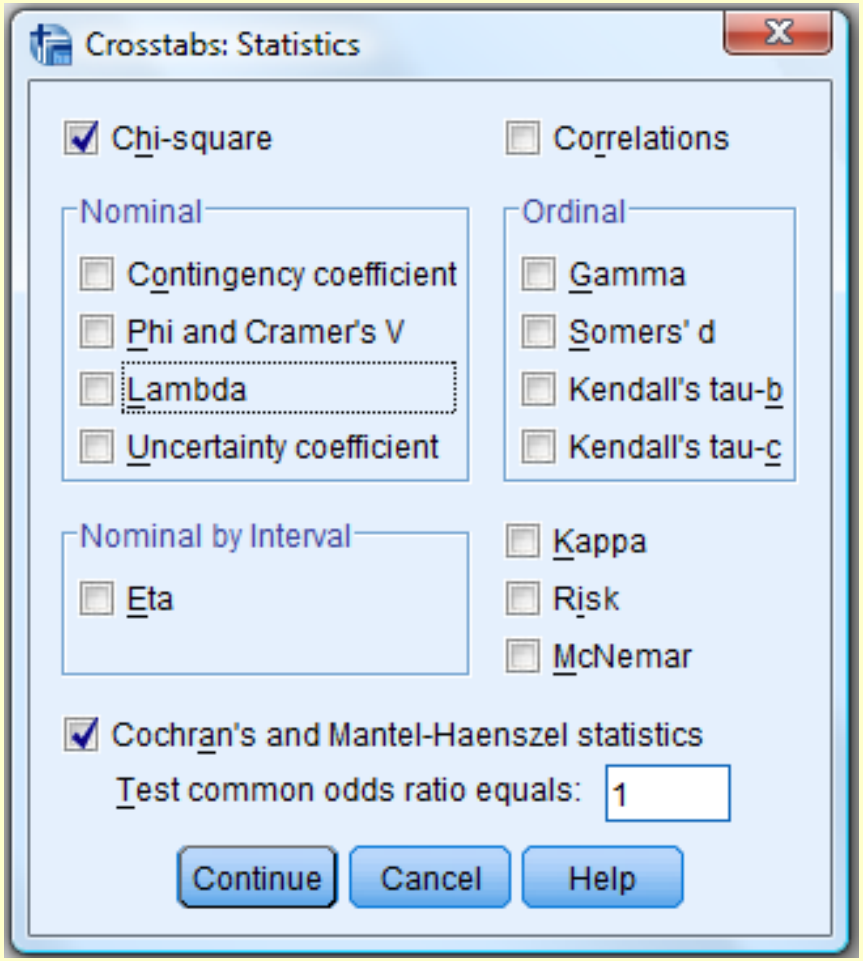


NOTE: It does not matter which variable used to define the rows or the columns

Pearson Chi-Square Test - Step 4



Select Statistics :
(in addition to default tests)



Chi-square Test

Odds Ratio:
Compares observed estimate to expected (null) value of 1 (same change in probability)

Pearson Chi-Square Test – Step 4

Default Statistics - I :

Likelihood Ratio

$$L\chi^2 = 2\sum Observed_{ij} \ln\left(\frac{Observed_{ij}}{Model_{ij}}\right)$$

- Based on maximum-likelihood theory.
  - Create model for which the probability of obtaining the observed set of data is maximized
  - This model is compared to the probability of obtaining those data under the null hypothesis
  - i and j are the rows and columns of the contingency table and ln is the natural logarithm
- Test Statistic
  - Has chi-square distribution with (r – 1)(c – 1) df
  - Preferred to chi-square when samples are small

Pearson Chi-Square Test – Step 4

Default Statistics - II :

Fisher Exact Test

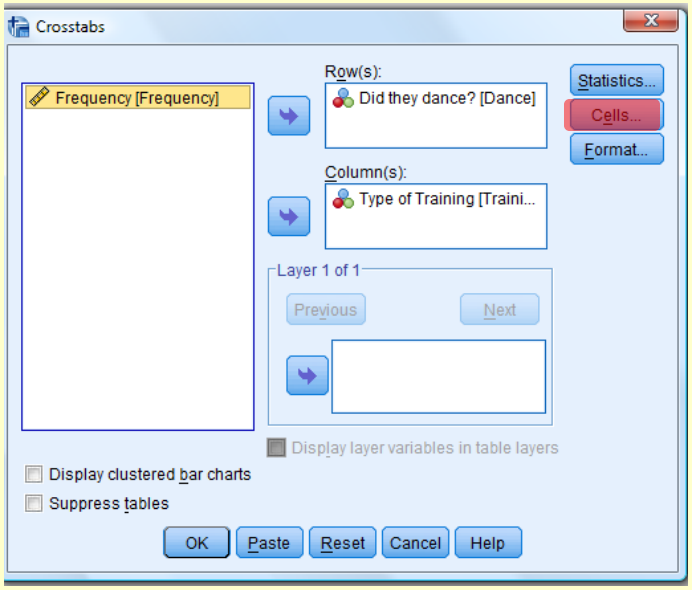
- Given a contingency table, test calculates exact probability of specific outcome, as follows:

a	b	$a + b$
c	d	$c + d$
$a + c$	$b + d$	$a + b + c + d\ (=n)$

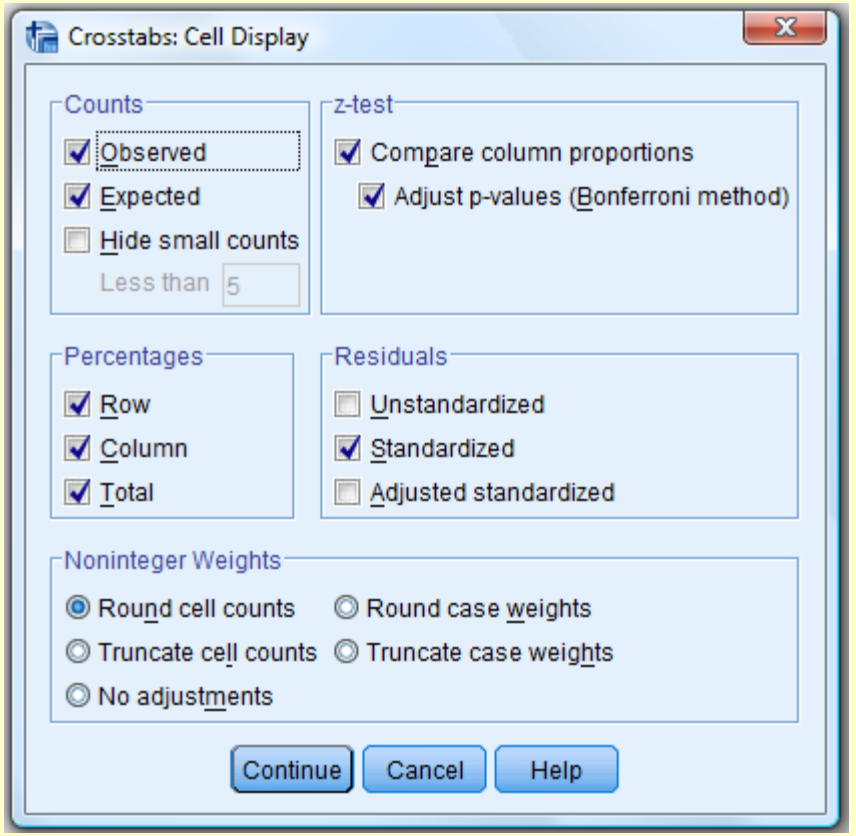
$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

- Test Statistic
  - Has chi-square distribution with (r – 1)(c – 1) df
  - Preferred to chi-square when samples are small
  - Conservative: Some argue that it rejects Ho at a higher alpha level, p > 0.05 (for small sample sizes)

Pearson Chi-Square Test - Step 5



Select Output:



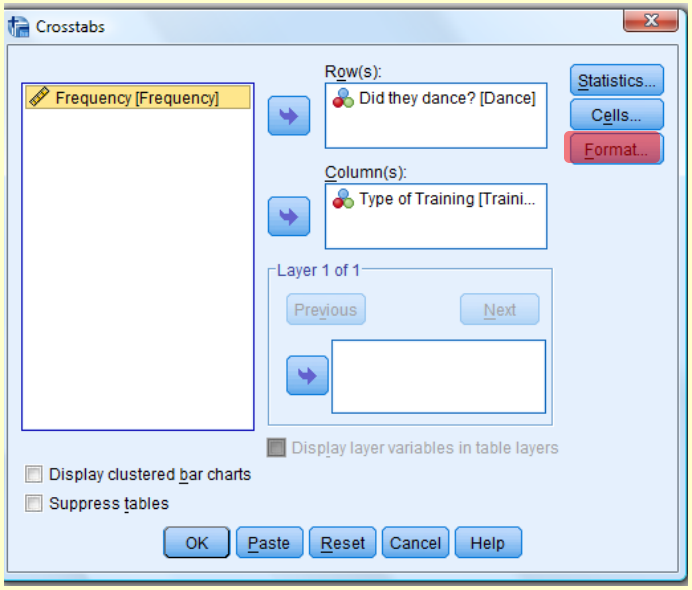
Observed / Expected Counts

Show Percentages

Show Residuals (and Z tests)

Round cell counts (weights)

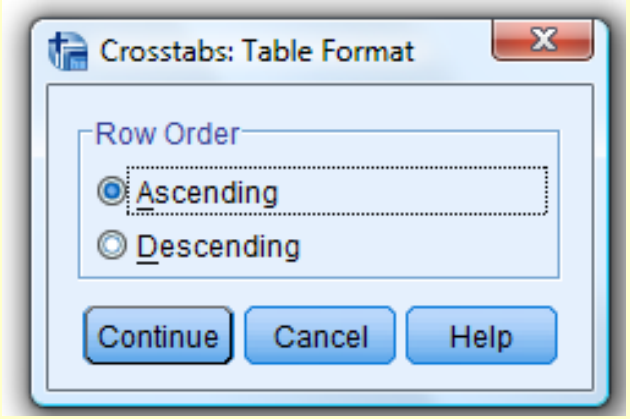
Pearson Chi-Square Test - Step 5



Select Format:

Arrange Results in order:

Ascending / Descending



Pearson Chi-Square Test - Step 6

Summary of Cases / Percentages

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Did they dance? * Type of Training	200	100.0%	0	.0%	200	100.0%

Cross-tabulation Summary:

Observed / Expected

Count / Proportions

Deviations (residuals)

Did they dance? * Type of Training Crosstabulation

			Type of Training		
			Food as Reward	Affection as Reward	Total
Did they dance?	Yes	Count	28[a]	48[b]	76
		Expected Count	14.4	61.6	76.0
		% within Did they dance?	36.8%	63.2%	100.0%
		% within Type of Training	73.7%	29.6%	38.0%
		% of Total	14.0%	24.0%	38.0%
		Std. Residual	3.6	-1.7	
	No	Count	10[a]	114[b]	124
		Expected Count	23.6	100.4	124.0
		% within Did they dance?	8.1%	91.9%	100.0%
		% within Type of Training	26.3%	70.4%	62.0%
		% of Total	5.0%	57.0%	62.0%
		Std. Residual	-2.8	1.4	
Total		Count	38	162	200
		Expected Count	38.0	162.0	200.0
		% within Did they dance?	19.0%	81.0%	100.0%
		% within Type of Training	100.0%	100.0%	100.0%
		% of Total	19.0%	81.0%	100.0%

Each subscript letter denotes a subset of Type of Training categories whose column proportions do not differ significantly from each other at the .05 level.

Pearson Chi-Square Test - Step 6

		Training		
		Food as Reward	Affection as Reward	Total
Could They Dance?	Yes	28	48	76
	No	10	114	124
	Total	38	162	200

Standardized Residuals:

+: excess of observed counts

-: deficit of observed counts

Did they dance? * Type of Training Crosstabulation

			Type of Training		
			Food as Reward	Affection as Reward	Total
Did they dance?	Yes	Count	28[a]	48[b]	76
		Expected Count	14.4	61.6	76.0
		% within Did they dance?	36.8%	63.2%	100.0%
		% within Type of Training	73.7%	29.6%	38.0%
		% of Total	14.0%	24.0%	38.0%
		Std. Residual	3.6	-1.7	
	No	Count	10[a]	114[b]	124
		Expected Count	23.6	100.4	124.0
		% within Did they dance?	8.1%	91.9%	100.0%
		% within Type of Training	26.3%	70.4%	62.0%
		% of Total	5.0%	57.0%	62.0%
		Std. Residual	-2.8	1.4	
Total		Count	38	162	200
		Expected Count	38.0	162.0	200.0
		% within Did they dance?	19.0%	81.0%	100.0%
		% within Type of Training	100.0%	100.0%	100.0%
		% of Total	19.0%	81.0%	100.0%

Pearson Chi-Square Test - Step 6

Subscripts:

Do the two columns differ significantly?

"food reward" differs from "affection reward"

Did they dance? * Type of Training Crosstabulation

			Type of Training		
			Food as Reward	Affection as Reward	Total
Did they dance?	Yes	Count	28_a	48_b	76
		Expected Count	14.4	61.6	76.0
		% within Did they dance?	36.8%	63.2%	100.0%
		% within Type of Training	73.7%	29.6%	38.0%
		% of Total	14.0%	24.0%	38.0%
		Std. Residual	3.6	-1.7	
	No	Count	10_a	114_b	124
		Expected Count	23.6	100.4	124.0
		% within Did they dance?	8.1%	91.9%	100.0%
		% within Type of Training	26.3%	70.4%	62.0%
		% of Total	5.0%	57.0%	62.0%
		Std. Residual	-2.8	1.4	
Total		Count	38	162	200
		Expected Count	38.0	162.0	200.0
		% within Did they dance?	19.0%	81.0%	100.0%
		% within Type of Training	100.0%	100.0%	100.0%
		% of Total	19.0%	81.0%	100.0%

Each subscript letter denotes a subset of Type of Training categories whose column proportions do not differ significantly from each other at the .05 level.

Pearson Chi-Square Test - Step 6

Chi-Square / Likelihood Ratio Results:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	25.356[a]	1	.000		
Continuity Correction[b]	23.520	1	.000		
Likelihood Ratio	24.932	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	25.229	1	.000	(Use 2-tailed)	
N of Valid Cases	200				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 14.44.
b. Computed only for a 2x2 table

Yates' Continuity Correction: Decreases test statistic (if expected values are too low) NOTE: 20% < 5, 0% < 1

Likelihood Ratio: Also significant

NOTE: Fisher Exact Test - More Conservative

Pearson Chi-Square Test - Step 6

Interpretation: Significant Chi-Square ($p < 0.001$)

$Df = (r-1) * (c-1) = 1*1 = 1$

We have more:
"Dance & Food Reward" than expected

We have fewer:
"No-Dance & Food-Reward" than expected

Did they dance? * Type of Training Crosstabulation

			Type of Training		
			Food as Reward	Affection as Reward	Total
Did they dance?	Yes	Count	28_a	48_b	76
		Expected Count	14.4	61.6	76.0
		% within Did they dance?	36.8%	63.2%	100.0%
		% within Type of Training	73.7%	29.6%	38.0%
		% of Total	14.0%	24.0%	38.0%
		Std. Residual	3.6	-1.7	
	No	Count	10_a	114_b	124
		Expected Count	23.6	100.4	124.0
		% within Did they dance?	8.1%	91.9%	100.0%
		% within Type of Training	26.3%	70.4%	62.0%
		% of Total	5.0%	57.0%	62.0%
		Std. Residual	-2.8	1.4	
Total		Count	38	162	200
		Expected Count	38.0	162.0	200.0
		% within Did they dance?	19.0%	81.0%	100.0%
		% within Type of Training	100.0%	100.0%	100.0%
		% of Total	19.0%	81.0%	100.0%

Pearson Chi-Square Test - Step 7

Interpretation: Test of Conditional Independence

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	25.356	1	.000
Mantel-Haenszel	23.403	1	.000

The two variables not independently distributed in our observations

Two different test statistics: Recommend use Cochran (1 df)

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.

Pearson Chi-Square Test - Step 7

Interpretation: Tests of Odds Ratio Estimates

$$\text{Odds Ratio} = \frac{\text{Odds}_{\text{dancing after food}}}{\text{Odds}_{\text{dancing after affection}}}$$

$$= \frac{2.8}{0.421}$$

$$= 6.65$$

Larger than 1

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			6.650
ln(Estimate)			1.895
Std. Error of ln(Estimate)			.407
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	2.997
		Upper Bound	14.754
	ln(Common Odds Ratio)	Lower Bound	1.098
		Upper Bound	2.692

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

Dancing after food is much more likely than dancing after affection

95% C.I. does not overlap 1.00

Pearson Chi-Square Test – Step 8

Describing the Results:

There was a significant association between the type of training and whether or not cats would dance

χ^2 (1 df) = 25.36, $p < 0.001$

This finding illustrates the fact that, based on the odds ratio, the likelihood of cats dancing is 6.65 times higher if they were trained with food than with affection

Contingency Tests – Summary

- We approach the analysis of categorical data in much the same way we approach the analysis of other data:
  - we fit a model, we calculate the deviations between our model and the observed data, and we use those deviations to evaluate the model we have fitted
- Three methods for Analyzing categorical variables
  - Pearson's chi-square test
  - Likelihood ratio test
  - Fisher Exact test
- Yates' Continuity Correction: Assume that expected frequencies are large (80% > 5, 100% > 1)

Contingency Tests – Recommendations

- Use all available tests for Contingency Tests
  - If they agree: GREAT
  - If they disagree: use more conservative method (Fisher Exact Test)

- Continuity Correction (Yates' Correction)
- Use it if expectation assumption violated

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

where:

O_i = an observed frequency

E_i = an expected (theoretical) frequency, asserted by the null hypothesis

N = number of distinct events

- Quantify Effect Size
  - The odds ratio is a useful measure of the size of the effect for categorical data