

SPSS Lab 8

Categorical Data – Chi-Square Tests

Chi-Square Goodness of Fit Test

Demo

In order to use parametric methods we need to assume that the data comes from a specific underlying probability distribution (Normal, binomial, Poisson, Chi-Square, etc) Now we will study a method to test for the *goodness of fit of a probability model*.

Frequency distribution of mean diastolic blood pressure for adults 30–69 years old in a community-wide screening program in East Boston, Massachusetts

Group (mm Hg)	Observed frequency	Expected frequency	Group	Observed frequency	Expected frequency
<50	57	77.9	≥80, <90	4604	4478.5
≥50, <60	330	547.1	≥90, <100	2119	2431.1
≥60, <70	2132	2126.7	≥100, <110	659	684.1
≥70, <80	4584	4283.3	≥110	251	107.2
			Total	14,736	14,736

We would like to assume the measurements above came from a normal distribution. The assumption can be tested by computing the expected frequencies if the data came from a normal distribution.

The expected frequencies within a group interval a to b are given by

$$14736 \times (\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma))$$

For the (≥50, <60) group we get

$$14736 \times (\Phi((60-80.68)/12) - \Phi((50-80.68)/12))$$

$$= 14736 \times (\Phi(-1.723) - \Phi(-2.557))$$

$$= 14736 \times (0.0424 - 0.0053) = 14736 \times 0.0371 = 547.1$$

We will run the Chi-Square Goodness of Fit using the book’s example on page 438.

- Create a new data file in SPSS.
- Type in the observed frequency in a variable called Observed.
- Create two new variables called RangeLeft for a and RangeRight for b and enter the ranges of each group. Enter a=0 for the first group where b=50. Enter a=110 and b=100000 for the last group.
- Use Transform/Compute Variable to create a new variable called “Expected” which will be $14736 * (\text{CDF.NORMAL}(\text{rangeRight}, 80.68, 12) - \text{PDF.NORMAL}(\text{rangeLeft}, 80.68, 12))$.
- Describe what values are in the Expected column
- Sort column “Observed” in ascending order
- Whenever you have grouped values in SPSS you need to weight the cases by going to Data/weight cases. In this case choose “observed” as the frequency variable.

- h. Now you will go to Analyze/Nonparametric Tests/Chi-Square... Select Observed as the test variable. Now unfortunately you will need to enter the “expected” values by hand one by one. That is the reason why you had to sort the values. Enter the expected values in the order they appear from top to bottom. Click Options/descriptives. Click Ok.
- i. You should obtain a chi-square statistic of 350.2. The significance is close to zero which means we reject H_0 . What does it mean to reject H_0 in the context of this test?

Chi-Square Test

Let’s solve Homework “Sports Medicine” problem 10.70, page 454-455 from the book. We will test if “tennis racket size” is related to episodes of the painful condition called “tennis elbow.”

- Download the data from the web site. Use the file Tennis1.sav to compute the chi-square test.
- To compare participants with 1 or more episodes with participants with no episodes of the condition, create a new field called “episodes” with 1 if 0 episodes and 2 if more than 1 episodes. Use the Transform\Visual Binning or Transform\Compute Variable Method.
- Filter out any current racket size that is not ($typ_curr = 1 = CONVENTIONAL_SIZE$) or ($typ_curr = 3 = OVER_SIZE$).
- Go to Analyze\Descriptive Statistics\Crosstabs...
- In Rows select “episodes” and columns = “typ_curr.” In Statistics select Chi-Square. In cells select “Observed” and “Expected” Click OK.
- The “Continuity Correction” is the Yates-Corrected Chi-Square Tests.
- Calculate the Chi-Square statistic by hand and compare to the SPSS value:
 - First calculate the expected value for each cell: $Expected(cell_{i,j}) = \frac{total(row_i) \times total(column_j)}{grand_total}$
 - Then plug the values into this formula to obtain the Yates corrected Chi-Square statistic: $\chi^2 = \sum (|O_{ij} - E_{ij}| - 0.5)^2 / E_{ij}$ for all cells i,j

Lab Assignment 8a

Test if “tennis string type” is related to episodes of the painful condition called “tennis elbow.” Do as before (except hand calculation of chi_square statistic) for variable *str_curr*. Describe your results.

More Chi-Square Tests

The feeding preference of goldfinches is measured by the number of seeds the birds eat on different days. Two types of seeds are studied, black oil and striped sunflower seeds.

The following table lists the results.

Type of Seed	Day				Total
	1	2	3	4	
Black Oil	19	14	9	45	87
Striped	5	10	6	39	60

What test can be used to assess if the feeding preference of goldfinches are the same on different days and why? Answer (Since Class takes place after lab): Use the RxC Chi-Square test because we have data with two possible outcomes Black Oil or Striped and a table that is larger than 2x2.

Lab Assignment 8b

- a. Implement the test from a and report a p-value
- b. You will need to enter the data into SPSS by hand. You will need to create 3 variables: Day, Type, and Count. In Type you enter 1 for Black Oil and 2 for Striped in Count you enter the count.
- c. You will need to weight the cases by count by going to Data/weight cases ...
- d. Go to Analyze/Descriptive Statistics/Crosstabs, add the variables Day and Type to row and column. Click statistics and select Chi-Square, Click cells and select observed and expected. Click OK
- e. Which of the tests is equivalent to the test you chose or to the test in the book? In order to find out you will need to calculate the statistic and compare the value with the 3 SPSS values:
 - a. First calculate the expected value for each cell: $\text{Expected}(\text{cell}_{i,j}) = \frac{\text{total}(\text{row}_i) \times \text{total}(\text{col}_j)}{\text{grand_total}}$
 - b. Then plug the values into this formula to obtain the Chi-Square statistics:
$$\chi^2 = \sum (\text{O}_{ij} - \text{E}_{ij})^2 / \text{E}_{ij}$$
 for all cells i,j
- f. What can you conclude from the results?

Lab Assignment 8c

McNemar's Test – Categorical Data

Run McNemar's Test for matched-pair data on the Valid_ch5.sav (lab 3) data to test the hypothesis that there is a difference in the Sat fat intake recorded with the FFQ method as compared to the DR method.

- a. State the hypothesis
- b. A t-test is more appropriate when using all the information in the data, but if you assume that the data only has two categories LOW\HIGH Sat FAT intake then the McNemar is appropriate here. Why?
- c. Create two columns one for FFQ , one for DR for LOW\HIGH Sat FAT intake, and chose $\text{sfat} > \text{mean}$ for high=1, and $\text{sfat} \leq \text{mean}$ for low=0.
- d. Run the Crosstabs analysis. Go to Analyze/Descriptive Statistics/Crosstabs, add the two new columns one each in row and column. Click statistics and select McNemar's test. Click cells and select observed and expected.
- e. What can you conclude from the McNemar's test
- f. Run the above test comparing another nutrient