Lab 5: Chi-Square and Contingency Analysis

Lab Objectives:

- Conduct Chi-Square goodness-of-fit tests
- Enter frequency data and assign weighting
- Calculate a Poisson distribution
- Conduct a Chi-square contingency test and Fisher's Exact test for contingency tables
- Use SPSS to determine odds ratio and relative risk

Exercise 1: Chi-Square testing in SPSS

A Chi-Square test is used to analyze frequency data for categorical or discrete numerical variables. It compares observed frequencies to expected or predicted frequencies. Today we will conduct Chi-square goodness-of-fit tests to analyze frequency distributions of a single variable, and Chi-Square contingency tests to analyze associations between two or more categorical variables.

Note that Statistical Table A (χ^2 Distribution) can be found at the end of this file.

Days of the week

In class, we looked at a research study examining the number of births on each day of the week. We will examine this data set, but will consider a scenario where every day of the week has an equal probability (e.g. 1/7). This data is found in the file *Days of the week.sav*

- Open *Days of the week.sav*
- Select Analyze, Nonparametric tests, Legacy dialog, and **Chi-Square**.
- In the window that pops up, add Day to your test variable list.
- Note that under "Expected values" you can change whether all categories are equal, or whether the categories should occur with certain expected values. For this test we will consider all categories equal. Click OK.
- In the output you should see a table showing each day of the week, the observed N, the expected N and a value called "Residual". Below that is a table with the test statistics. It provides the Chi-Square value, the degrees of freedom, and the significance value ("asymp. sig.").

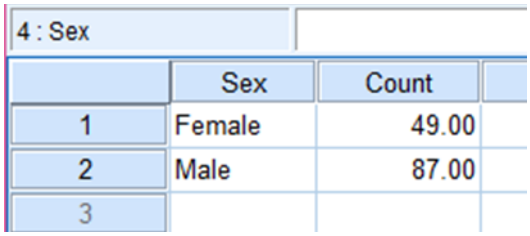
➢ Answer questions 1 and 2 on Sakai.

Birds in a storm

There are two ways that SPSS can work with frequency data. The first we have worked with extensively already: each row represents an individual, and each column represents a variable. The value of a categorical variable in a given cell describes which group the individual in that row belongs to.

Alternatively, in some cases you may have frequency data that is already organized into a table, or summarized as such. For example, you may be interested in the sex of birds collected at a particular site following a wind storm. You have recorded that 49 females and 87 males were caught. In this case it would be quicker to input this data into SPSS as a summary, rather than filling in 49 rows for females, and 87 rows for males.

- In SPSS, select File, New, Data.
- In Variable view, enter a variable for "Sex" and a variable for "Count". Both should be type numeric. Assign the appropriate measurements and values. For the sex variable, use a value of 1 for female, and a value of 2 for male.
- In Data view, in the Sex column enter 1 and 2 (for female and male)
- In the count column, enter 49 for the female count, and 87 for the male count. Check that your data looks like this (if you have data labels turned on):



- Now we need to indicate to SPSS that the values in column 2 are the frequency observations. Select Data, **Weight cases**. Select Weight cases by, then select "count" as your frequency variable. Click OK.

Recently we focused on conducting binomial tests. A binomial test compares the observed number of successes to that expected under the null hypothesis, and calculates an exact P-value. A binomial test can only be used when there are only two categories. The Chi-Square goodness-of-fit test also works when there are only two categories and can be a quick substitute for the binomial test when calculating statistics by hand. When calculated with a statistics program, the Chi-Square test is still an approximation based on the Chi-Square probability distribution with a particular number of degrees of freedom. It will provide a P-value, but this will not be as precise as the P-value calculated with the binomial test. Let's check the difference using the bird data.

- Select Analyze, Nonparametric tests, Legacy dialog, and **Binomial test**.
- Add Sex to the test variable, and test the null hypothesis that 50% of birds caught at this site should be females.
- Examine the output, and increase the number of decimal places for the P-value to four.
- Select Analyze, Nonparametric tests, Legacy dialog, and **Chi-Square**.
- Add Sex to the test variable, and test the null hypothesis that 50% of birds caught at this site should be females.

- Examine the output, and increase the number of decimal places for the P-value (Asymp sig.) to four.

➢ Answer question 3 on Sakai.

Not at all like me

Now let's try an example where the expected values are not equal across categories. The difference here is we have to specify the expected proportions or counts. Howell has a hypothesis that if you ask participants to sort one-sentence characteristics of themselves (such as "I eat too fast") into five piles ranging from "not at all like me", to "very much like me", the percentage of items placed in each pile will be approximately 10%, 20%, 40%, 20%, and 10% for the five piles. Let's test this hypothesis using data gathered from 50 sorted statements.

- Open *Not like me.sav*
- Choose Data, weight cases, and use frequency as the weighting variable.
- Select Analyze, Nonparametric tests, Legacy dialog, and **Chi-Square**.
- Select Category as the test variable. Under expected values we have to select "values" instead of "all categories equal", and then type in the expected number of cases that should fit each category. You do not necessarily need to calculate the exact expected numbers according to your data, as long as you enter the values in the correct relative proportions. For example, in this case you can enter 10, 20, 40, 20, and 10 even though you have only 50 measurements, and SPSS will convert these to expected values for a sample size of 50. You could also enter 0.1, 0.2, 0.4, 0.2 and 0.1. Or alternatively you could enter 5, 10, 20, 10, and 5. SPSS would use any of these inputs to calculate the same expected counts according to the total number of observations in your data. It is very important that you enter these values in the same order as your data.
- Enter the expected values by entering each value in order, then clicking Add. Once all values are entered click OK to run the Chi-Square test.
- Examine the output.

➢ Answer questions 4 and 5 on Sakai.

Exercise 2: Poisson Distribution

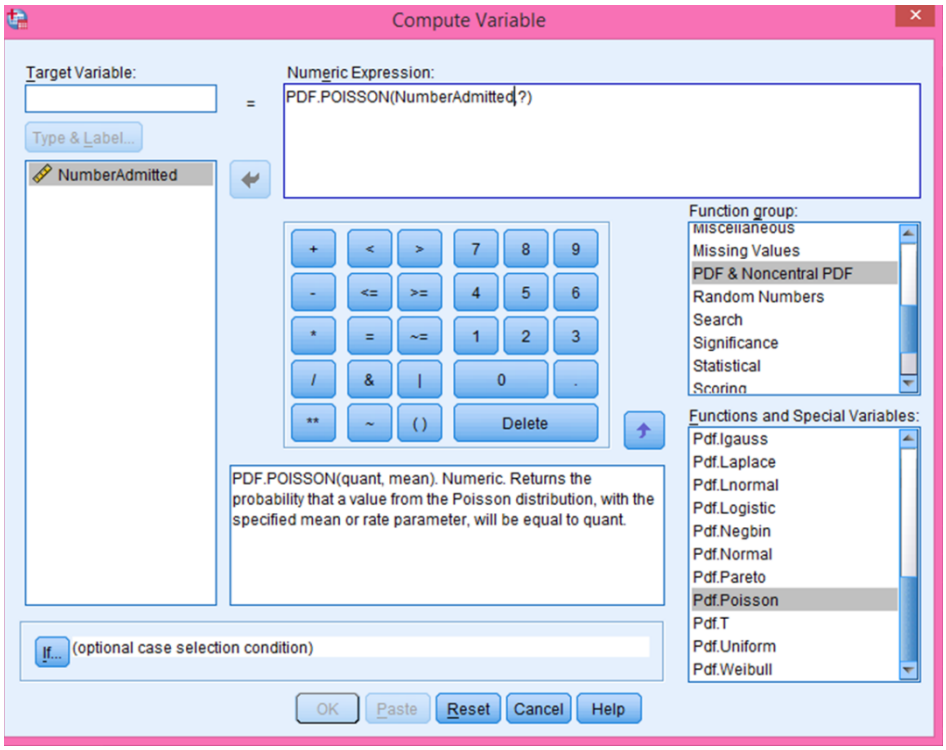
The Poisson distribution describes the frequency distribution of successes, when successes happen independently and with equal probability over time or space. Given a mean number of events observed in a given time or space, the Poisson distribution can be used to determine the probability of observing X number of events in that same unit of time or space. If successes occur "randomly" in time or space, then the distribution of values should follow the Poisson distribution.

Emergency

Let's check the Poisson distribution for the number of people admitted to the ER between 7 and 8 pm on Friday night, given an average admittance of 2.4 people. We will assume that admissions are

independent of one another, and are just as likely to land in one instant in time as another on a Friday night. Your boss at the hospital is adjusting staff schedules and is interested in the probability that 0 – 10 patients will be admitted during this time period.

- In SPSS, select File, New, Data.
- In Variable view, enter a variable for "NumberAdmitted", with a type of numeric, and a measurement of scale.
- Return to data view and enter values of 0 – 10.
- Select Transform, compute variable.
- In the box that pops up, under function group select "PDF & Noncentral PDF". Under functions and special variables, select **PDF.Poisson**. You should see a description pop up in the lower white box. Click the up arrow to move this formula to the upper white box (numeric expression).
- Note that the formula requires the quant (number of successes you are interested in), and the mean. Click on NumberAdmitted in the variable list and then the right arrow to move it to replace the first question mark. See the following image to make sure you're on the right track:



- Replace the second question mark with the mean value (2.4).
- In the target variable box, type Poisson. This will create a new variable in your sheet with the Poisson probability values.
- Click OK.

- Review the data in the data view. Note that these values indicate the probability of 0, 1, 2, 3, … 10 people admitted during this time period, given a mean number of admittances of 2.4, and each admittance being independent and occurring with equal probability.

➢ Answer question 6 on Sakai.

Truffles

Truffles are a great delicacy. A set of plots of equal size in an old-growth forest in Northern California was surveyed to count the number of truffles per plot. The resulting distribution is found in the file *Truffles.sav*. The data include a total of 288 plots, and have a mean value of 0.604 truffles per plot. You are interested in determining whether truffles are randomly located around the forest. If not, they may be either clumped or dispersed.

- Open the file *Truffles.sav*.
- Adjust the weighting according to the frequency.
- Calculate the probabilities of 0 – 4 truffles per plot using the Poisson distribution function.
- Increase the number of decimal places in the Poisson distribution to 4 so that you have a good view of the values.
- Calculate the expected frequencies for each number of truffles per plot (0 – 4) according to the expected probabilities of the Poisson distribution and the total number of truffles collected. Try using SPSS to perform this calculation directly. Hint: Use the transform / compute variable function to create a new variable called "Expected".
- Note that we cannot run a Chi-square test on this data, because one of our expected values is too low. Remember that in a Chi-square test, no expected value can be less than 1. To get around this, combine the data for 3 and 4 truffles per plot into a single row.
- Run a Chi-square test on the data. Enter each of the expected values you computed in SPSS with two decimal places.

➢ Answer question 7 on Sakai.

Exercise 3: Chi-Square for Contingency Analysis

The Chi-Square test can also be used to analyze whether two categorical variables are associated. That is, whether the two variables are independent, or whether the outcome of one variable depends on the other.

The gnarly worm gets the bird

Example 9.4 in your book examines the life cycle of a parasite that is transferred from snails, to fish, to birds. The researchers noticed that infected fish spend more time near the water surface, and wanted to examine whether this influenced their chances of being ingested by a bird. An outdoor

tank was stocked with uninfected, lightly infected, and heavily infected fish. The number of each that were eaten by birds was recorded. This data could be presented in a contingency table as follows:

	Uninfected	Lightly infected	Highly infected	Total
Eaten	1	10	37	48
Not eaten	49	35	9	93
Total	50	45	46	141

Remember that the response variable (eaten/not eaten in this case) is displayed in rows, while the explanatory variable is displayed in columns. This data is contained in the files *worms.sav*. We can use a Chi-Square test to determine whether parasite infection and being eaten are independent (H0) or not independent (HA).

- Open *worms.sav*
- Note that in this case we want to use Chi-Square to compare associations between two categorical variables. We therefore need to use a different option from the SPSS menu.
- Select analyze, descriptive statistics, **crosstabs**.
- To be consistent with the contingency table, add the fate variable to the rows and the infection variable to the columns.
- Click on Statistics, and select Chi-square.
- Click on cells, and select Observed and Expected under counts. Click OK.
- Check the box for "display clustered bar charts". Then click OK to run the analysis.
- Review the crosstabulation table. This is similar to a contingency table, but shows both the observed and expected counts.
- In the Chi-Square test table we are interested in the Pearson Chi-Square results. This row provides both the Chi-Square statistic and the P-value (Asymptotic significance).

➢ Answer question 8 on Sakai.

Small fry

A study by Miller et al examined the survival of rainbow trout fry (babies) in Lake Superior, comparing those that came from a government hatchery on the lake, and those that came from wild trout. The data is contained in fry.sav.

- Open this file and conduct a Chi-Square test through the crosstab function, as above. Add frysource to the columns and survival to the rows. Display the observed and expected counts.

- Note that in your output, since this is a 2 x 2 comparison, the results for Fisher's Exact Test are also displayed. The P-value for Fisher's exact test is an exact P-value, rather than a close approximation as obtained with Chi-Square. Fisher's Exact test must be used whenever the expected frequencies are too low for the Chi-Square test, but it is difficult to compute by hand.

➢ Answer question 9 on Sakai.

Exercise 4: Odds Ratio & Relative Risk

An odds ratio measures the magnitude of association between two categorical variables when each variable has only two categories. It is calculated from the odds of a focal outcome in one group, divided by the odds of the same outcome in a second group. Odds are calculated by the probability of the focal outcome (success) divided by the probability of the alternate outcome (failure). A similar measure used in medical studies is relative risk, which is equal to the probability of an undesired outcome in the treatment group, divided by the probability of the same undesired outcome in the control group.

Postnatal depression

Postnatal depression affects approximately 8 – 15% of new mothers. Patel et al (2005) examined whether the rates of postnatal depression differed between mothers who delivered vaginally (control), compared to mothers who delivered by C-section (treatment). The data is found in *delivery.sav*

- Open the file *delivery.sav*
- Select analyze, descriptive statistics, crosstabs.
- Note that for determining odds and relative risk in SPSS, the risk factor (delivery method) must be moved to the **rows** box, and the outcome we are interested in (depression) must be moved to the **columns** box.
- Select cells and check "row percentages".
- Select statistics and check "risk".
- Click OK.
- In the crosstabulation, you can see the observed counts for each combination of delivery & outcome. The percentages for depression and no depression are also shown for each type of delivery. This allows you to quickly see the percent of women with a C-section who experienced depression, compared to the percent of women with a vaginal birth who experienced depression.
- In the risk estimate box, the first row "Odds ratio for delivery (C-section/vaginal)" is the odds ratio. It shows the odds of developing depression with a C-section delivery compared to a vaginal delivery. If this number is greater than 1, then the odds of developing depression are higher with

a C-section delivery. If this value is less than 1, then the odds of developing depression are lower with a C-section delivery.

- The next two rows are risk ratios. The first is the relative risk for the outcome **depression**, with a C-section compared to vaginal birth. The second is the risk for the outcome no depression. SPSS shows both because it does not know which outcome we are interested in. In this case we are interested in the depression outcome (the undesired outcome).

- Return to crosstabs, and conduct a Chi-Square test on the data to see if there is a significant difference in these outcomes. This test will tell you if the two variables (delivery method and depression) are independent.

➢ Answer questions 10 and 11 on Sakai.

Exercise 5: Additional practice:

Heart failure

The file *heart failure.sav* contains data on the day of the week that patients were admitted to a hospital with heart failure. Analyze this data using a Chi-Square test to determine if patients are admitted in equal proportions on each day of the week.

Feline high rise syndrome

A more recent study of feline high-rise syndrome (FHRS) included data on the month in which each of 119 cats fell. The data are found in the file *Rainingcats.sav*. Do a hypothesis test to determine whether the probability of FHRS is the same every month.

MS and CCSVI

In 2012, a research group examined the association between MS and a vein condition known as chronic cerebrospinal venous insufficiency (CCSV). This data is found in the file *CCSVI.sav*. Conduct a Chi-Square test to determine whether there is a significant association between MS and CCSVI

Smoking Fingers

Researchers examined the presence of finger defects (fused fingers, extra fingers, or less than five fingers) in births of mothers who smoked (risk factor) during pregnancy, compared to control mothers who did not smoke during pregnancy. The data is found in the file *Smoking fingers.sav*. Use SPSS to calculate the Odds ratio and 95% confidence interval of the odds ratio. Use a Chi-Square test to determine if this difference is significant. Don't forget that the risk factor must go in the row, and the outcome in the column.

➢ Answer the remaining questions on Sakai.

Statistical Table A: χ^2 Distribution

	α									
df	0.999	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005	0.001
1	0.0000016	0.000039	0.00016	0.00098	0.00393	3.84	5.02	6.63	7.88	10.83
2	0.002	0.01	0.02	0.05	0.10	5.99	7.38	9.21	10.60	13.82
3	0.02	0.07	0.11	0.22	0.35	7.81	9.35	11.34	12.84	16.27
4	0.09	0.21	0.30	0.48	0.71	9.49	11.14	13.28	14.86	18.47
5	0.21	0.41	0.55	0.83	1.15	11.07	12.83	15.09	16.75	20.52
6	0.38	0.68	0.87	1.24	1.64	12.59	14.45	16.81	18.55	22.46
7	0.60	0.99	1.24	1.69	2.17	14.07	16.01	18.48	20.28	24.32
8	0.86	1.34	1.65	2.18	2.73	15.51	17.53	20.09	21.95	26.12
9	1.15	1.73	2.09	2.70	3.33	16.92	19.02	21.67	23.59	27.88
10	1.48	2.16	2.56	3.25	3.94	18.31	20.48	23.21	25.19	29.59
11	1.83	2.60	3.05	3.82	4.57	19.68	21.92	24.72	26.76	31.26
12	2.21	3.07	3.57	4.40	5.23	21.03	23.34	26.22	28.30	32.91
13	2.62	3.57	4.11	5.01	5.89	22.36	24.74	27.69	29.82	34.53
14	3.04	4.07	4.66	5.63	6.57	23.68	26.12	29.14	31.32	36.12
15	3.48	4.60	5.23	6.26	7.26	25.00	27.49	30.58	32.80	37.70
16	3.94	5.14	5.81	6.91	7.96	26.30	28.85	32.00	34.27	39.25
17	4.42	5.70	6.41	7.56	8.67	27.59	30.19	33.41	35.72	40.79
18	4.90	6.26	7.01	8.23	9.39	28.87	31.53	34.81	37.16	42.31
19	5.41	6.84	7.63	8.91	10.12	30.14	32.85	36.19	38.58	43.82
20	5.92	7.43	8.26	9.59	10.85	31.41	34.17	37.57	40.00	45.31
21	6.45	8.03	8.90	10.28	11.59	32.67	35.48	38.93	41.40	46.80
22	6.98	8.64	9.54	10.98	12.34	33.92	36.78	40.29	42.80	48.27
23	7.53	9.26	10.20	11.69	13.09	35.17	38.08	41.64	44.18	49.73