

4. ANALYSING FREQUENCY TABLES

Categorical (nominal) data are usually summarized in frequency tables. Continuous numerical data may also be grouped into intervals and the frequency of observations in each interval may also be summarized in a frequency table (or in a histogram; see earlier lab on “Exploring and Describing Data”). In this lab we will explore two kinds of frequency tables and the ideas they may be used to test.

One-way Frequency Table

The first type of frequency table lists the number of observations in different categories of a single list. An example is the following table evaluating how good humans are at choosing random numbers. The data are from the early years of a US State Lottery, in which players would buy a ticket and choose any number they wanted between 000 and 999. Winnings would be divided between all holders of the winning number, which was chosen randomly. The following data are based on a random sample of 100 players of the Lottery (these are not the winning numbers, but rather they are the numbers selected by players). Listed are the frequencies of numbers chosen that have 0 to 9 as the first digit:

First digit of chosen number	Frequency (f_i)
0	4
1	16
2	14
3	15
4	13
5	8
6	9
7	7
8	8
9	6
Total	100

These frequencies may be compared to those predicted by different hypotheses. For example, when players pick a number between 000 and 999, are some first digits more popular than others? It looks like numbers beginning with “0” are unpopular, and those beginning with 1 through 4 are excessively popular. A **goodness of fit test** is an appropriate method for testing these data against the null hypothesis that there is no preference for different digits in the population.

Two-way (Contingency) Table

The second kind of frequency table is the two-way table, or **contingency table**. Here, every observation is cross-classified by two category variables instead of just one. The usual goal is to test whether true (population) relative numbers of individuals falling into the different classes for one variable is the same regardless of individual values for the second variable. An example given below lists the number of survivors and non-survivors in two classes of mountaineers descending

from the peak of Mount Everest between 1978 and 1999: those using supplemental oxygen, and those descending without supplemental oxygen. (Most deaths on Mount Everest occur during the descent, not the ascent.)

Survival	Used supplemental oxygen	Did not use supplemental oxygen
Survived descent	1045	88
Did not survive descent	32	8
Total	1077	96

(data from Huey and Eguskitza 2000, JAMA 284: 181)

In this case we are interested in knowing whether the relative numbers of survivors and nonsurvivors depends on whether or not supplemental oxygen was used. This is not an experimental study, so we are unable to test whether a difference in survival between classes is **caused** by oxygen use, but at least we can decide whether supplemental oxygen and survival are **associated**. The null hypothesis is once again the skeptical point of view: survival and oxygen use are **not associated** with one another (i.e., survival and oxygen use are **independent**). A test of differing survival frequencies between the two categories of mountaineers is carried out using a **contingency test**.

Hypothesis Testing

Forming and testing hypotheses is one of the most basic endeavors in statistical analysis of biological data. With your notes and the course textbook, review your knowledge of the following concepts:

- **null hypothesis** (H_0) and **alternate hypothesis** (H_a).
- **Type I errors** and **Type II errors**
- significance level
- degrees of freedom

Test Statistics for Goodness of Fit and Contingency Tests

The chi-squared statistic, χ^2 , is a measure of discrepancy between **observed** and **expected** frequencies, where expected frequencies are those expected under the null hypothesis.

$$\chi^2 = \sum \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$$

A second measure of discrepancy is the G -statistic (the log likelihood ratio):

$$G = 2 \sum f_i \ln \frac{f_i}{\hat{f}_i}$$

Under the null hypothesis both statistics have a distribution that conforms approximately to the theoretical chi-squared distribution. The degrees of freedom will usually be $k - 1$, where k is the number of classes of the category variable, except in special situations to be dealt with later in the course.

Analogous statistics contrast observed and expected frequencies in contingency tables. Here, however, the expected frequencies are based on the null hypothesis that relative frequencies are the same in each set. The expected frequency for row i and column j in the contingency is obtained as

$$\hat{f}_{ij} = \frac{R_i C_j}{N}$$

Where R_i and C_j are the row and column totals, respectively, and N is the grand total number of independent observations.

With your notes and the course textbook, review your knowledge of the following concepts:

- independence
- rules of thumb for low expected frequencies in chi-square tests
- Yates correction for continuity [JMP IN does not employ this correction]
- Fisher's exact test

Using the program

In the case of one-way tables, only a single categorical variable is required (e.g., “First digit of chosen number”). Two categorical variables are needed for a two-way (contingency) table (e.g., “Use of supplemental oxygen” and “Survival”). Make sure that after entering the data, the category variable(s) have the **nominal** attribute (this can be reset in the columns section of the left frame, or by selecting **Column Info** in the **Cols** pull-down menu). The observed frequencies may be entered directly to a new column (call it “observed frequency” or “number of observations”).

To produce a **bar graph** of frequencies from a one-way table, use the **Distribution** menu option and select the categorical variable as your Y column in the pop-up window. In the same window you also need to select the observed frequency column as your “Freq” variable. To carry out a goodness of fit test, click the red “▼” symbol next to the categorical variable name above the bar graph and select **Test Probabilities**. This action will open a new display box below the frequency table in the **Distribution** output window. Here you will need to enter the expected frequencies for your test. Click on each row and enter either the expected *frequency* or the expected *proportion* for that row (it doesn't matter which, as long as you are consistent; the goodness of fit test will be carried out using the expected frequencies in either case). Unfortunately, JMP IN doesn't display the expected frequencies it uses to calculate the test statistic, so these will be lacking if you have simply entered the expected proportions. In this case you will be unable to ensure that the expected frequencies are large enough to fulfill the assumptions of the χ^2 goodness of fit test. To calculate expected frequencies you will need to use your own calculator, or better yet the JMP IN calculator.

To produce a **mosaic plot** for a two-way (contingency) table, use the **Fit Y by X** menu option. In the pop-up window, select one of the categorical variables as your Y column and the other as your X column. Once again, select the observed frequency column as your “Freq” variable. A two-way table will also appear beneath the mosaic plot, giving the observed frequencies (the program will also display the expected frequencies but you need to select this option by pressing the red “▼” symbol next to the **Contingency Table** title). Unfortunately, JMP IN does not include the Yates correction for continuity when the *G*-test and chi-square (Pearson) tests are carried out on 2x2 tables (you will need to include the Yates correction with 2x2 tables on your assignments and written exams). However, it does include the Fisher exact test, which you can use to validate the results of the chi-square and *G* tests.

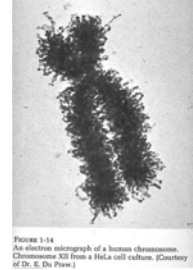
One-way and two-way frequency tables can be constructed from raw data on individual subjects using the **Tables -> Summary** option in the pull-down menu or by selecting **Summary** in the **Tables** tab on the JMP Starter. In the pop-up menu choose one (in one-way tables) or two (for two-way tables) categorical variables and click the **Group** button. Then click the **Statistics** button in the same window and select **N**. When you click “OK” a new data table will appear that tallies the frequency of observations corresponding to each category or combination of categories.

Problems



1. Enter the Lottery data given above and generate the corresponding bar graph.
 - a) Examine the bar graph. Do the frequencies appear to vary greatly between classes?
 - b) Carry out a statistical test of the hypothesis that players favor some first digits over others when choosing a number between 000 and 999. In your work, present **all** steps (i.e., state hypotheses, give the *P*-value, the significance level for the test, and state your conclusion). Since the computer provides the *P*-value directly, there is no need to provide the critical value from the tables in Zar.
 - c) Compare the results for the chi-squared (Pearson) to those for the *G* test (Likelihood ratio). Why are they different?
 - d) Compare the results from your visual appraisal of the data to the goodness of fit tests. Which approach provides qualitative information and which one provides quantitative information? What level of uncertainty is associated with those quantitative probabilities?
 - e) What are the degrees of freedom for these tests? Why do we lose a degree of freedom?
 - f) Why would it be necessary to alter the analysis if expected values are small?

2. A physical gene map of the human genome was published in 1998 that contained the estimated locations of 30075 human genes. The table below lists the estimated number of genes on each chromosome. The second column lists the fraction of the total human genome made up by each chromosome. For example, the X chromosome constitutes a little more than 5% of the total genome size. These data are in the data file **genemap98.jmp** on the shared directory.



Chromosome	Proportion of total genome	Observed number of genes
1	0.0834	3114
2	0.0809	2257
3	0.0679	2015
4	0.0644	1478
5	0.0615	1529
6	0.0580	1893
7	0.0542	1594
8	0.0492	1206
9	0.0460	1248
10	0.0457	1371
11	0.0457	1755
12	0.0453	1585
13	0.0311	703
14	0.0295	1047
15	0.0282	1029
16	0.0311	849
17	0.0292	1263
18	0.0269	523
19	0.0212	1114
20	0.0228	758
21	0.0123	305
22	0.0136	565
X	0.0520	874

Data from Deloukas et al (1998). *A physical map of 30,000 human genes*. Science 282:744–746 (see <http://www.ncbi.nlm.nih.gov/genemap98/page.cgi?F=GeneDistrib.html>).

- a) Display the estimated numbers of genes on different chromosomes using a bar graph and mosaic plot. Describe the differences between chromosomes. Which chromosomes have the most genes? Which have the fewest?
- b) We would not expect each chromosome to have the same number of genes because chromosomes differ in size. Use the variable “Proportion of total genome” to calculate the expected number of genes on each chromosome, taking into account chromosome size differences. (Create a new column to receive the expected frequencies, and use the JMP IN

calculator to compute them. The total number of genes is 30075. Call the new variable “Expected No. genes”). Generate a bar graph for the expected frequencies and place it beside the bar graph for the observed numbers of genes. Do larger chromosomes tend to have more genes?

- c) Generate a new column and compute the following quantity for each chromosome using the JMP IN calculator:

$$\frac{(\text{Observed No. genes}) - (\text{Expected No. genes})}{\sqrt{(\text{Expected No. genes})}}$$

This quantity (sometimes called a z-score) measured the difference between observed and expected frequencies scaled by the square root ($\sqrt{}$) of the expected frequencies. On this scale, which chromosomes have a dramatic deficiency of genes for their size? Which chromosomes have the most dramatic excesses?

- d) Use the observed and expected frequencies to test the null hypothesis that gene number is determined purely by chromosome size. To have the program do this for you automatically you will need to enter the expected frequencies for each chromosome, one at a time. Alternatively, you could use the JMP IN calculator to compute the chi-square statistic directly. This is easily done by squaring the quantities calculated in (c) and summing them up (**Col Sum** is an option in the **Statistical** functions provided in the **Functions** panel of the JMP IN calculator window). You can also use the calculator to provide you with the P -value for the calculated χ^2 statistic using **Probability -> ChiSquareDistribution**, or you can look up the appropriate critical value in Zar.

Note: The substantially lower than expected gene density on the X chromosome might result from expression bias. Gene expression from the X chromosome is reduced because in females the second copy of the X chromosome is inactivated, and in males a second X is lacking (males are XY). Gene with reduced expression are more difficult to detect by the method the researchers used to find them.

3. Enter the Mount Everest mountaineer survival and supplemental oxygen data from the above table into a JMP IN data table. The most useful way to do this is to create a new data table with 4 rows and three columns. Call the first column “Survival” and the second column “Oxygen use”. Enter the four combinations of these two variables into the four rows. Finally, put the observed frequencies into the third column.



- a) Inspect the mosaic plot for these data. Describe the pattern in words. Are the relative frequencies of individuals surviving similar or different in the two oxygen groups?
- b) Test whether survival of mountaineers descending from Mount Everest is significantly associated with use of supplemental oxygen. Show all steps in your work (a good habit, as always).
- c) Why does the P -value for the Fisher’s exact test differ from that of the Pearson χ^2 and the G tests?

- d) Repeat the calculation of the Pearson χ^2 by hand. Did you obtain the same number as JMP IN? Why?
- e) Do the expected frequencies satisfy the assumptions of the chi-square test? What strategy do you recommend?
- f) The authors who compiled the Everest data also presented results from the teams of mountaineers descending Everest (climbers tend not to go alone). These data are given below. Which data set is the most appropriate to test an association between survival and supplemental oxygen? Why?

Survival	Used supplemental oxygen	Did not use supplemental oxygen
All team members survived	85	24
At least one team member died	8	4
Total	93	28

- g) Should we conclude from the test in (f) that supplemental oxygen has no effect on survival?
- h) The same authors also compiled similar data for K2, a nearby summit in the Himalayas. Analyse these data in the same way as for Mount Everest. Are the results the same as those in (f) ?

Survival	Used supplemental oxygen	Did not use supplemental oxygen
All team members survived	12	24
At least one team member died	0	12
Total	12	36

4. Open the data file **student_data.jmp** from the shared directory. This file records the data taken from Biology 300 students on the first day of class, January 2001. The variables are:

- height, Student height in cm
- hand, Student handedness (left or right; “both” was classified as left)
- parent.first, Parent listed first by student when giving their heights (mom or dad)
- mom.height, Student’s mother’s height, in cm
- dad.height, Student’s father’s height, in cm
- mom.hand, Whether mother is left or right-handed
- sex, Whether student is male or female



- a) Use **Distribution** to test whether male and female students occur with equal frequency in the Bio 300 class. Note that in the pop-up window you will not need to specify a column for the **Freq** button because you are working now with the raw data instead of the frequency table.

- b) Use an appropriate method to test whether there is a statistical association between handedness of student (left or right-handed) and that of his/her mother.
- c) Use **Tables** to generate a two-way (contingency) table for handedness of student and mother. This method shows how JMP IN may be used to construct frequency tables from raw data.
- d) Some students listed their dad first when giving their height, whereas some student listed their mother first. Does this depend on the sex of the student?