

Chapter 10

Nonparametric Statistics

All of the procedures covered so far are parametric statistics. Parametric statistics are very powerful, meaning that even small differences can be readily detected. The reason for this is that parametric statistics assume that the sample(s) being analyzed came from populations that are normally distributed. As discussed in chapter 4, normally distributed populations can be characterized by two parameters: the mean (μ) and standard deviation (σ). Parametric statistics take advantage of this, and are very sensitive to differences in location (mean) and dispersion (standard deviation).

However, not all samples come from normally distributed populations so parametric statistics are not appropriate for all data sets. For such samples there are two possibilities. The data can be transformed from an arithmetic scale to another, which will sometimes normalize the data (Figure 10.1; we will talk about data transformation later). Alternatively, a different category of statistics, non-parametric, can be used.

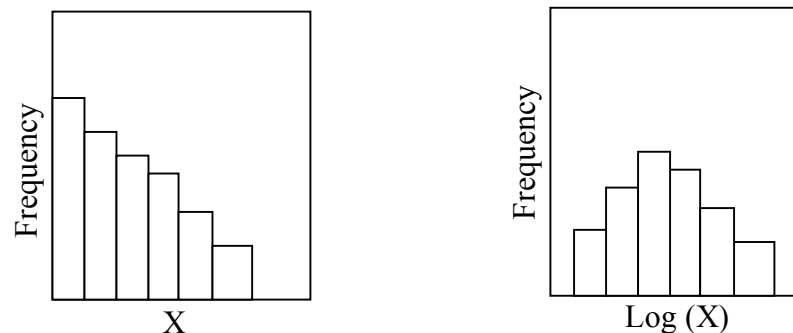


Figure 10.1. An example of non-normally distributed data (left panel) that has been transformed by taking the log of each measurement (right panel). Transforming the data is valid because it only changes the scale in which the data are expressed.

Nonparametric statistics can be used to analyze data when either the shape of the distribution is not known or not normal. Unlike parametric statistics, nonparametric procedures do not test for differences in specific parameters (i.e. means or standard deviations), but only whether the data have the same distribution. The trade-off is that while non-parametric procedures can be used for all types of data, they are less powerful than their parametric counterparts. Because of this the risk of type II errors, not recognizing actual differences, is high.

So, how can you tell if your data are appropriate for parametric statistics? There are specific procedures that test for normality, but in most cases examining the frequency distribution will give you a good idea of whether the data are normally distributed.

The first step in creating a frequency distribution is to divide the data into different classes and then count the number, or frequency, of occurrences of values within each class. The frequency of observations is then plotted as a bar graph versus the class. An example of how a frequency distribution is created is illustrated using the data in Table 10.1, which is

plotted in Figure 10.2. As you can see the data has a shape that approximates a bell shaped curve. These data could, therefore, be analyzed using parametric statistics. It is unlikely that you will actually observe that your data form a classic bell shape due to the small sample sizes that are usually used for student research projects. However, parametric statistics are powerful enough that they can handle slight deviations from normal distributions so as long as your data aren't skewed (as is the case with the original data in Figure 10.1) you are probably okay with parametric statistics.

Table 10.1. Tree heights (m) measured within a hectare plot on the island of Tiki Tiki.

6	3	4	5	7	6	4	6	5	7	5	6	6	8	6
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

<u>Height Class</u>	<u>Frequency</u>
1	0
2	0
3	1
4	2
5	3
6	6
7	3
8	1
9	0

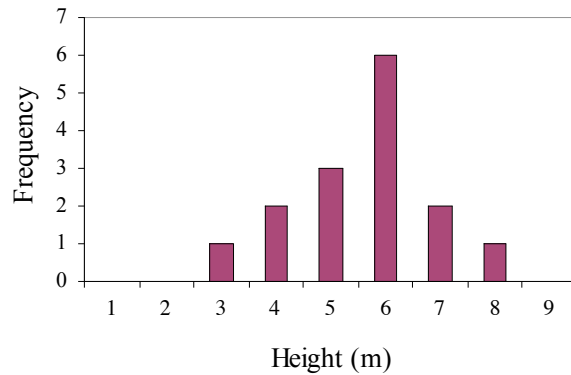


Figure 10.2. Frequency distribution for tree heights (m) measured within a plot on the island of Tiki Tiki (Data from Table 10.1).

As previously mentioned, if data are not normally distributed then parametric statistics are not appropriate. Sometimes changing the scale using mathematical transformations as is shown in Figure 10.1 can normalize the data. Methods of data transformations will be covered in chapter 13. As a general rule, if transforming the data results in a normal distribution then using parametric statistical procedures on the transformed data is the best course of action. However, not all data can be normalized in which case nonparametric statistics are the only option for analysis.

There are many different nonparametric techniques. For every parametric technique there is usually a nonparametric equivalent. Most students have used nonparametric statistics without realizing it. Chi-square analysis is a commonly used nonparametric procedure that is easy to use and familiar to most students. Thus, chi-square (which is categorized as a goodness of fit test) will be the main nonparametric procedure we will utilize in this class. We will also learn how to do heterogeneity chi-square and contingency tables.

Chi-square analysis

Chi-square analysis is especially suited for nominal scale data where the variable under study is classified by some quality it possesses rather than by a numerical measurement. In such cases the variable is called an attribute and counts are made of the number of items (e.g. animals) that exhibit certain attributes. Examples familiar to most students are those from genetics where phenotypes are the attributes analyzed. These types of data are referred to as discrete or discontinuous data since they can take on only certain values (usually integers). They are obviously different from most measurement data where a datum can be of any value on a continuum between two extremes (continuous data).

Chi-square analyses are different than others covered in this course in that they do not involve consideration of the parameters of the population from which samples have been drawn. Thus, they fall into the category of non-parametric statistics.

In a Chi-square analysis one compares values that are **observed** to those that would be **expected** based on the null hypothesis. Chi-square analyses are often referred to as "goodness of fit" tests; that is, how good do the observed data fit the expected. Calculated chi-square can be as small as zero in cases where the fit is perfect or very large. They can never be negative (because the difference between the observed and expected numbers is squared). The chi-square statistic is always calculated using actual frequencies observed. It is not valid (illegal, immoral and dishonest!) to use relative frequencies or percentages.

Chi-square (χ^2) is calculated using the following formula

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad [11.1]$$

where O is the observed frequency (number of counts) in a particular class or category, E is the expected frequency in that class if the null hypothesis is true, and the Σ indicates that summation is performed over all of the classes or groups.

Example Problem

A frequent problem encountered in genetics is to infer whether a population conforms to a certain theoretical distribution. For example, assume that coat color in a little known species of mice (*Fattus rotundus*) appears to be inherited as a single gene with incomplete dominance and that the phenotypes and genotypes are: brown hair, BB; black hair, Bb; white hair, bb. Crosses between mating pairs of mice with black hair would be expected to produce offspring in the ratio of 0.25 brown to 0.50 black to 0.25 white. In observing the results of such crosses the geneticist notes that the results were 65 brown, 130 black and 45 white. The question the geneticist must answer is whether the observed frequencies deviate significantly from the hypothesized frequencies expected.

I. STATISTICAL STEPS

A) Concise statement of Ho and Ha

Ho: offspring follow 1:2:1 ratio of coat color

Ha: offspring do not follow 1:2:1 ratio

B) Choice of appropriate statistical test.

Since data are counts or discrete, then the only appropriate test is Chi Square.

C) Computation of descriptive and test statistics.

	<u>brown</u>	<u>black</u>	<u>white</u>	<u>total</u>
observed	65	130	45	240
expected	$=(240)(0.25)=60$	$=(240)(0.5)=120$	$=(240)(0.25)=60$	240
difference	5	10	15	--
difference ²	25	100	225	--
$\chi^2 =$	0.417	+ 0.833	+ 3.75	= 5.00

D) Determination of the probability of the test statistic.

Assume you use a computer to construct a population of a billion mice with coat colors of 1 brown: 2 black: 1 white. You randomly sample 240 mice coat colors and compute the chi-square values a great number of times. Then you compute the relative frequency of each observed chi-square. (The relative frequency is equal to the number of observations of a specific occurrence divided by the total number of all possible occurrences.) What value of chi-square would have the highest relative frequency? The lowest?

This sort of simulation results in Figure 11.1 where the relative frequency is plotted on the Y axis with the associated calculated chi-square value being plotted on the X axis. This figure represents the relationship only in the case of two degrees of freedom (*df*), which is equal to the number of classes minus one in this sort of a problem ($=3$ coat colors - 1). The relationship between the relative frequency and the chi-square changes with the degrees of freedom so there are a multitude of these sorts of curves. When $df=2$, as shown to the right, a chi-square value of 0 would have the highest relative frequency, and as the chi-square values increase the relative frequencies would decrease.

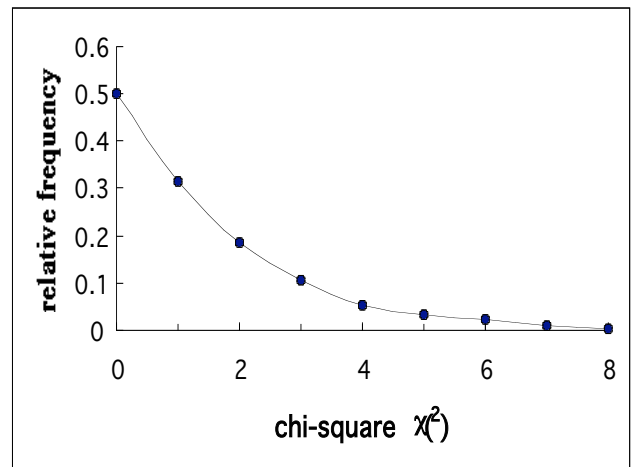


Figure 11.1. Relative frequency of χ^2 's when $df=2$.

Another way of thinking about this relationship is that one would expect to get most of the time very close to 60 brown, 120 black, and 60 white mice when randomly sampling 240 mice from a populations of mice that are known to have the coat colors in that ratio. If one calculates the chi-square for this sort of sample it will generally come out fairly close to zero. On the other hand, there is a chance that all 240 mice sampled could be, say, white.

Such an event is not impossible; however, it is highly unlikely, would be observed very infrequently, and would result in a huge chi-square (720!).

Statisticians have packaged the chi-square values at various probabilities and degrees of freedom into tables similar to those used for the other test statistics. Table 5 is such a table and is packaged in its entirety in Appendix B. It is partially reproduced below (Table 11.1). The left-hand column is the degree of freedom for each row (df; typically df = classes -1). In the body of the table are chi-square values. Across the top are the probabilities of obtaining those chi-square or higher values. Given df = 2 one can see in the table that the probability of getting a chi-square of 5.991 or greater is 0.05. In other words, one would expect to get from such a population a chi-square of 5.991 or greater only 5% of the time.

Table 2.1. Cumulative chi-square distribution for dfs from 1 to 3. The body of the table contains chi-square values at different degrees of freedom (df, left column) at decreasing cumulative probabilities (top row) between the tabulated value of chi-square and infinity.

<u>df</u>	<u>0.3</u>	<u>0.25</u>	<u>0.2</u>	<u>0.15</u>	<u>0.1</u>	<u>0.05</u>	<u>0.02</u>	<u>0.01</u>	<u>0.001</u>
1	1.074	1.323	1.642	2.072	2.706	3.841	5.412	6.635	10.827
2	2.408	2.773	3.219	3.794	4.605	5.991	7.824	9.210	13.815
3	3.665	4.108	4.642	5.317	6.251	7.815	9.837	11.345	16.266

The probability associated with the chi-square value of 5.0, as obtained in the sample problem, can now be determined. In the sample problem the number of classes is three so df=2. Examination of Table 11.1 show that a chi-square value of 5 falls between the tabular chi-square values of 4.605 and 5.991. Obtaining a chi-square of 4.605 or greater has a probability of 0.1. This statement can be written in shorter notation as $P(\chi^2_{[2]} \geq 4.61) = 0.1$. Similarly $(\chi^2_{[2]} \geq 5.99) = 0.05$. Thus, the P of getting a chi-square of 5.0 or greater is above 0.05 but below 0.1 or in notational form, $0.10 > P(\chi^2_{[2]} > 5.00) > 0.05$.

E) Decision of whether differences are "significant" or "not significant."

Accept Ho and reject Ha.

II. BIOLOGICAL INTERPRETATION

The observed biological and statistical inference is that the distribution of coat colors in *Fattus rotundus* does not significantly differ from a theoretical distribution of 1:2:1 (0.25 brown: 0.50 black: 0.25 white). It can be further inferred that coat color in this rodent is inherited as a single gene with two alleles and incomplete dominance.

Yates Correction

When there are only two classes or groups being compared with Chi-square then a correction must be applied in computing the chi-square value. That correction, the **Yates Correction**, is to subtract 0.5 from the absolute difference (ignore the sign) between the

observed and expected before squaring that difference. Thus, equation 11.1 would be rewritten as

$$\frac{(O - E - 0.5)^2}{E} \quad [11.2]$$

Calculating Chi-square Using EXCEL (TEMPLATE AVAILABLE ON MY WEBSITE)

An EXCEL template can be used to quickly generate the chi-square test statistic and its associated probability value from the observed and expected frequencies of the attributes involved. An example of such a template is provided in Figure 11.2, which shows a chi-square analysis for Problem 11.1. This problem is typical of chi-square problems involving more than two classes in that its first solution (Figure 2.3) indicates only that there is a difference among the classes—it does not indicate where the differences lie. To find this out, the investigator has to use common sense along with statistics and examine the data in further detail. First, let's examine the template in greater detail.

The top two rows of the chi-square analysis from column B to I include a brief descriptor of the data and the actual observed data for each group. These data, the descriptor and observed numbers, have to be put into the template. The only other data that has to be added are the expected relative frequencies (*expected rf* in template).

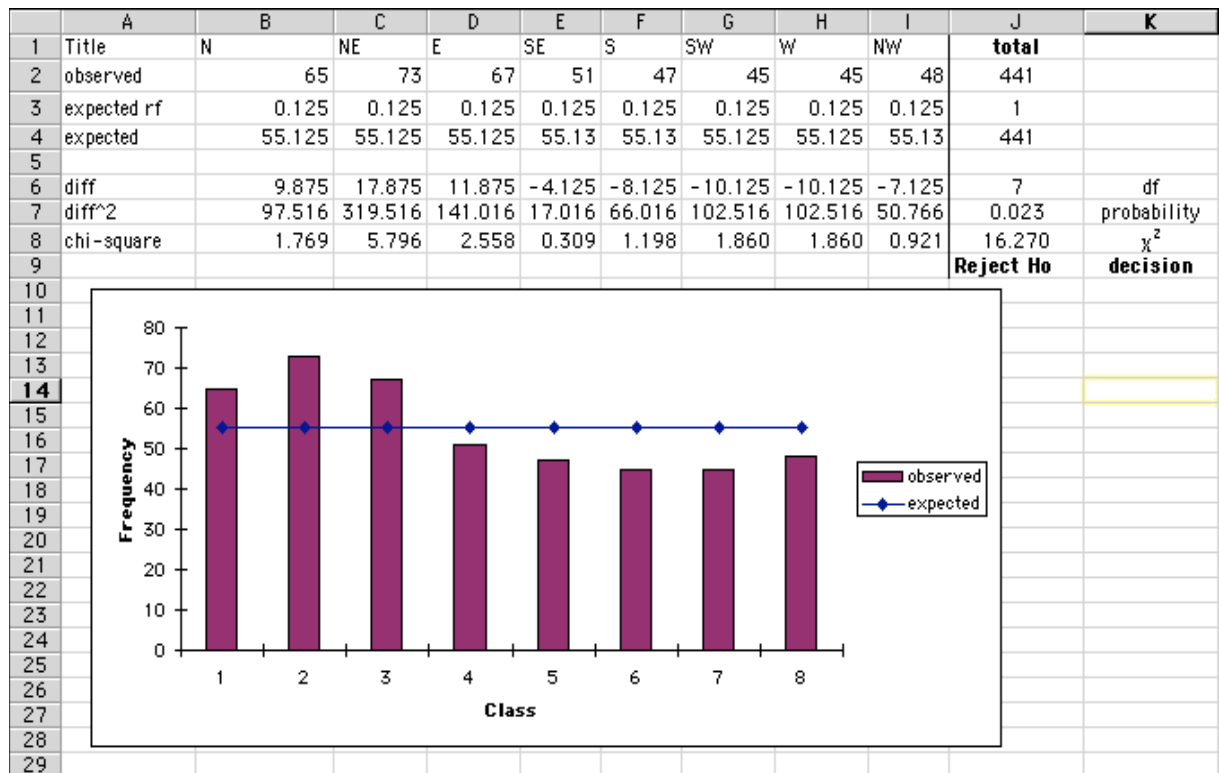


Figure 11.2. Chi-square analysis of number of bird nests observed at eight different compass directions. Data from Problem 11.1.

All the other values in the analysis are generated from the underlying formulas in the template. The expected values are simply the expected relative frequencies times the total number of observations. A good check to ensure that the expected relative frequencies have been entered properly is that the sum of the relative frequencies for all groups *must* equal 1 (shown in cell J3). In row 6 the difference between the observed and expected values is calculated and the resulting number is squared in row 7. Row 8 of the template is the individual chi-square value for each group, which is then summed in cell J8. This is the final test statistic, with the associated *df* and probability located in the cells above.

Example Problem (Problem 11.1)

The first chi-square calculation for this particular problem has already been examined in Figure 11.2. Now it will be put into the statistical procedural framework.

I. STATISTICAL STEPS

A. Statement of H_0 and H_a

H_0 : the observed frequencies of nests are **not** different that an equal distribution among the eight compass directions; that is, 1/8 of the nests should be found at each compass direction.

H_a : the observed frequencies of nests are different from an equal distribution among the eight compass directions.

B. Statistical test

Because the H_0 states an expected (hypothetical) frequency, use chi-square.

C. Computation of chi-square

$\chi^2_{[df=7]} = 16.270$ as shown in Figure 11.2

D. Determination of the P of the test statistic

$P(\chi^2_{[df=7]} = 16.270) < 0.023$, as shown in Figure 11.2

E. Statistical Inference

The differences between the observed frequencies and expected value of 1/8 at each compass direction is significant because its probability is lower than 0.05. Therefore, the H_0 is rejected and the H_a accepted.

One can and should try to determine where the major difference between observed and expected occurs. Inspection of the data indicates that the first three compass directions (N, NE, E) have observed values above expected while the remaining five directions are all lower. It thus appears that the nests fall into two distinct groups with one group have a larger number of nests than the other. Before the two groups can be compared, however, one has to

first demonstrate that the various members thought to be in that group are indeed not different from another. This leads to two new sets of hypotheses.

F. 2nd Statement of H₀

H₀₂: the observed frequencies of nests in the N, NE, and E are **not** different than an equal distribution among the three compass directions; that is, 1/3 of the nests should be found at each compass direction.

H_{a2}: the observed frequencies of nests in the N, NE, and E are different than an equal distribution among the three compass directions.

G. 2nd Computation of chi square

$\chi^2_{[2]} = 0.507$, as shown to right

Title	N	NE	E	total	
observed	65	73	67	205	
expected rf	0.3333	0.333	0.333333	1	
expected	68.333	68.33	68.33333	205	
diff	-3.333	4.667	-1.33333	2	df
diff^2	11.111	21.78	1.777778	0.776	probability
chi-square	0.1626	0.319	0.026016	0.507	χ^2
				Accept Ho	decision

H. 2nd Determination of the P of the test statistic

$P(\text{test } \chi^2_{[2]} = 0.507) > 0.05$

I. 2nd Statistical Inference

Difference between observed and expected ratio of 1:1:1 is not significant; therefore, accept H₀ and reject H_{a2}. Thus, the three compass direction groups tested can be legitimately lumped together into one group.

J. 3rd Statement of H₀

H₀₃: the observed frequencies of nests in the SE, S, SW, W, NW are **not** different than an equal distribution among the five compass directions; that is, 1/5 of the nests should be found at each compass direction.

H_{a3}: the observed frequencies of nests in the SE, S, SW, W, NW are different than an equal distribution among the five compass directions.

K. 3rd Computation of chi square

$\chi^2_{[4]} = 0.525$ as shown to right

Title	SE	S	SW	W	NW	total	
observed	51	47	45	45	48	236	
expected rf	0.2	0.2	0.2	0.2	0.2	1	
expected	47.2	47.2	47.2	47.2	47.2	236	
diff	3.8	-0.2	-2.2	-2.2	0.8	4	df
diff^2	14.44	0.04	4.84	4.84	0.64	0.971	probability
chi-square	0.3059	8E-04	0.102542	0.102542	0.013559	0.525	χ^2
						Accept Ho	decision

L. 3rd Determination of the P of the test statistic

$P(\text{test } \chi^2_{[4]} = 0.525) > 0.05$

M. 3rd Statistical Inference

Difference between observed and expected ratio of 1:1:1:1:1 (=1/5) in the five compass directions of SE, S, SW, W, NW are not significant; therefore, accept H_{03} and reject H_{a3} . Thus, the five compass direction groups tested can be legitimately lumped together into one group.

N. 4th (& last!) Statement of H_0

H_{04} : The observed distribution of the nests are **not** different than 3/8 of the total observed frequencies of nests in the compass directions of N, NE, and E and 5/8 of the total in the compass directions of SE, S, SW, W, NW.

H_{a4} : the observed frequencies of nests are different than the 3/8 and 5/8 expected of the two groups

O. 4th Computation of chi square

$\chi^2_{[1]} = 14.810$, as shown to right. Note that Yates correction was applied.

Yates applied				
	N, NE, E	Other		
observed	205	236	441	
expected rf	0.375	0.625	1	
expected	165.375	275.625	441	
diff-0.5	39.125	39.125	1	df
diff^2	1530.766	1530.766	0.000	probabili
chi-square	9.25633	5.554	14.810	χ^2
			Reject H_0	decision

P. 4th Determination of the P of the test statistic

$P(\text{test } \chi^2_{[1]} = 14.810) < 0.001$

Q. Final Statistical Inference

Nests are in significantly higher numbers than expected in the compass of N, NE, and E while they are in significantly lower numbers in the other five compass directions.

II. BIOLOGICAL INTERPRETATION

Results

The number of bird nests varied from 45 – 73 for various compass directions (Fig. 1). Nests were not randomly placed (Table 1; All Nests). The number of bird nest found in the N, NE, and E direction were not significantly different among themselves, but were significantly greater than the number found in the other directions (Table 1). These nests averaged slightly above 15% of the total nests observed. Nest in the other five compass directions also exhibited no significant differences among one another (Table 1), but each averaged only 11% of the total nests.

Table 1. Results of Chi-square analyses of nests observed in bushes at eight compass directions testing all nests and groupings of nests. Values shown are the Chi-square (χ^2), degrees of freedom (df) and P value.

Nests	χ^2	df	P
All	16.27	7	0.023
N, NE, E	0.057	2	0.78
SE, S, SW, W, NW	0.525	4	0.97
N - E vs SE - NW	14.81	1	< 0.001*

* Significant difference

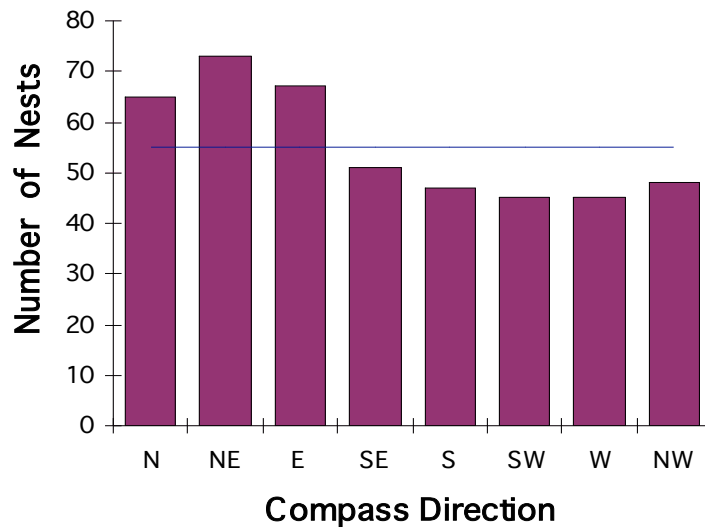


Figure 1. The number of bird nests observed in bushes at eight compass directions. The line is the expected number of nests that would be found at each compass heading if there were no preference in nest location.

Heterogeneity Chi-Square (Example Problem 11.6)

Heterogeneity chi-square is an elaboration of chi-square that is used when you are testing expected frequencies across more than one group (different sites, localities, etc). For example consider problem 11.6, you want to know if sex ratios differ, but you also have four localities to test for sex ratio differences. The use of heterogeneity chi-square determine whether samples can be pooled across localities, which, if statistically allowable, increases the sample size tested, thereby making the test and its result more reliable. You first compute χ^2 for each locality separately and add up the computed χ^2 values, this is called the total χ^2 . Then you compute χ^2 for a pooled sample of each sex irrespective of locality. The heterogeneity chi-square, the test statistic, is simply the different between the total χ^2 and the pooled χ^2 . The probability of the heterogeneity chi-square is then determined. If is not significant, then it is appropriate to pool all the data and simply test for sex ratio. If significant, you need to test various locality combinations.

I. STATISTICAL STEPS

A. Statement of H_0 and H_a

For each sample (locality) and the pooled sample (all localities summed together):

Ho: sex ratios are **not** different than 1:1

Ha: sex ratios are different than 1: 1

Because the data were collected from four different localities, the investigator must first determine if the data are homogenous (not heterogeneous). This requires testing the following hypotheses.

Ho: The samples do not differ from one another

Ha: The samples do differ from one another

B. Statistical test

Because the first Ho states an expected (hypothetical) ratio, use chi-square.

C. Computation of heterogeneity chi- square

First, one has to compute χ^2 for each locality using the stated Ho as shown to the right. The χ^2 and *dfs* are summed to obtain a total χ^2 and *df*. Then, the observations are summed to obtain a pooled set of data; localities are ignored. The heterogeneity χ^2 and *df* are simply computed as the difference between the total and pooled χ^2 and *df*. The heterogeneity χ^2 is:

$$\chi^2_{[3]} = 0.454$$

	Male	Female	χ^2	df	Prob
Locality A	44	54	1.020	1	
Locality B	31	40	1.141	1	
Locality C	12	18	1.200	1	
Locality D	15	16	0.032	1	
Total			3.394	4	
Pooled	102	128	2.939	1	
heterogeneity			0.454	3	0.929
Apply Yates Correction					
Pooled	102	128	2.717	1	0.099

D. Determination of the P of the test statistic

$$P(\chi^2_{[3]} = 0.454) > 0.05$$

E. Statistical Inference

The probability of the heterogeneity χ^2 is much greater than 0.05, which indicates that the four samples come from the same population. Thus, the pooled data set can be legitimately used. Because there are only two classes, the Yates correction has to be applied; it is not applied in computing the heterogeneity χ^2 . Based on the pooled and Yates corrected $\chi^2_{[2]} = 2.717$ and its associated $P > 0.05$, it would be concluded that the sex ratio follows the expected 1 male:1 female ratio.

Note that if it is concluded that the sample came from two or more populations then one has to separate the data appropriately. For example, if the data for locality A were 62 males and 36 females instead of the 44 and 54, then it would be concluded that the data were heterogeneous. Localities B, C, and D would then be shown to be from one population with a different sex ratio than locality A.

Problem Set - Chi-square

11.1) The positioning of nests in shrubs was noted for a species of birds. Determine if there was a directional preference in placing the nests (f_i = frequency of observation i).

Nest Position (i):	N	NE	E	SE	S	SW	W	NW
f_i :	65	73	67	51	47	45	45	48

11.2) Each of 126 individuals of a certain mammal species was placed in an enclosure containing equal amounts of each of six different foods. The frequency with which the animals chose each of the foods is given below. Determine if there is a preference among the food items, and, if so, what is the most preferred food.

Food Items (i):	A	B	C	D	E	F
f_i :	13	26	31	14	28	14

11.3) A sample of hibernating bats consisted of 44 males and 54 females. Determine if the hibernating population consists of equal numbers of males and females.

11.4) Genetic theory predicts three red-winged flies for each blue-winged fly among the offspring from a certain cross. If 76 red-winged flies and 22 blue-winged flies result from such a cross, is there a significant indication that the underlying theory is inaccurate?

11.5) A geneticist observed that the homozygous recessive of a particular di-hybrid cross did not appear as fit as the other offspring phenotypes. The geneticist also suspected that the number of homozygous recessives was reduced from the expected ratios and obtained the following data from the offspring of such a cross. Analyze and interpret.

Phenotype:	yellow	yellow	green	green
	smooth	wrinkled	smooth	wrinkled
Genotype:	Y_ S_	Y_ ss	yy S_	yy ss
Observed frequency:	152	39	53	6

11.6) In attempting to determine whether there is a 1:1 sex ratio among hibernating bats, samples were taken from four different locations. Determine whether the four samples may justifiably be pooled and if there is a 1:1 sex ratio.

Location:	<u>A</u>	<u>B</u>	<u>C</u>	<u>D</u>
Males:	44	31	12	15
Females:	54	40	18	16

Problems will be assigned in lab.