Testing Research and Statistical Hypotheses

Introduction

In the last lab we analyzed metric artifact attributes such as thickness or width/thickness ratio. Those were continuous variables, which as you learned can be described in terms of the central tendency, such as mean length, and the variability about that mean, as expressed by the variance and standard deviation. But many archaeological data come to us as discrete variables, which need to be treated very differently. For example, among an assemblage of artifacts including handaxes, cleavers and scrapers there is no "mean artifact type." In faunal assemblages there is no "mean animal species." For these kinds of data we need to treat the individual categories, whether they are artifact types, raw material types or categories of ceramic decorations. These are discrete variables, and to test hypotheses about them we need to use appropriate statistical procedures. We will give you the opportunity to do this by applying discrete statistical analysis to the testing of hypotheses about differences in lithic artifact samples.

Background

Statistical analysis is very common to archaeology and other sciences, because it is the way we qualify our conclusions, rather than just make claims that they are true. It is true that many scientific conclusions can be reached without statistics. Indeed many major contributions from science are observations.

Also, we need to understand when we need statistical support for our work and when we don't. For example, we do not need statistics to test the hypothesis that one group used pottery and another did not; we document this by the presence or absence of ceramics in their assemblages. But a hypothesis concerning the cultural relationships between prehistoric groups based on the similarities and differences in their pottery styles may very well require that we use statistical analysis, since this involves degrees of similarity, and a decision about what is "significantly" similar or different.

Statistics are most commonly applied when we are testing hypotheses about archaeological phenomena, such that it is the archaeological hypotheses that drive our work. We perform statistical analyses in order to verify that our observations on archaeological data are valid and not just impressions or erroneous conclusions based on small samples.

In most cases there are two kinds of hypotheses that are being tested. One, we'll call the **research hypothesis**, is based on the actual archaeological data, and our understanding of how those data are related. So if we pose the hypothesis that the cultural relationships between two groups are accurately reflected by their pottery styles, we need to be confident that pottery indeed can be used to measure cultural similarities. We can support this by analogy to modern potters, and from the archaeological record itself. Once we have defined our research hypothesis, we can decide what kind of **statistical hypothesis** is appropriate. In the following exercises we will illustrate the sequence of testing a research hypothesis with a statistical hypothesis that is appropriate for the situation. But in the end, it is the archaeological issue that we are concerned with.

There are many discrete statistical analytical methods. We will use the **chi-square** (X^2) **analysis**. The X^2 statistic is commonly used in testing archaeological hypotheses because we deal with so

many categorical variables as mentioned above. These can involve *single classifications* or *analysis of independence* in two-way (two categories compared at the same time) classification. In both, the X^2 statistic allows us to evaluate the observed frequency of certain categories against a theoretical frequency of those categories.

In some cases we are simply deciding if variations in our data are **random**, and therefore of little meaning. For example, I claim that in my site I found 46 handaxes and 54 choppers, and claim that the higher number of choppers is a significant difference, indicative of the cultural tradition that made those artifacts. You are not convinced, and ask me to test whether this difference could be due to chance. Let's use this as an example of hypothesis testing in archaeology, supported by use of the the X^2 statistic.

Testing for Randomness of Observations – One Variable

In this example, we are going to test the following research hypothesis:

> H_0: Any differences between handaxe and chopper frequencies in my assemblage are due to chance.

Statistically, we need to compare the observed frequencies of these artifact types against their theoretical frequencies that would be expected if their frequencies were indeed random. If two variables occur with random frequencies, such as tossing of coins, then their theoretical frequencies are 50-50. Once we have determined the theoretical frequencies, we can proceed with testing the hypothesis using the X^2 statistic that is calculated with the following equation:

$$X^2 = \sum \frac{(O_i - T_i)^2}{T_i}$$

where,

$\sum$	"the sum of"'
O	observed frequency (count)
T	theoretical frequency (count)
i	label for categories (1 for handaxes, 2 for choppers)

Let's consider this equation briefly, to make sure you understand what it's about. First, note that the X^2 statistic is going to be calculated by completing the operations shown on the right side of the equation. These can be summarized as follows:

1. For the first category (handaxes) subtract the theoretical frequency from that which we have observed.

2. Then square this difference to make all differences a positive value. We need to do this because some differences are negative and some are positive, but we're only interested in how large the difference is.

3. Next, divide the squared difference by the theoretical frequency. This is actually computing the ratio of the difference to the absolute value. If this ratio is a large number it will

make the X^2 larger, which will increase the chance that the difference is in fact a significant one.

4. Now perform steps 1-3 for the second variable (choppers).

5. Last, add the results for both variables to calculate the X^2. This will always be a positive number, and as it gets bigger, the chances that the difference between handaxe and chopper frequencies is significant (that is to say nonrandom) will be increased.After we have done this on our data, we will see how we determine whether or not my claim is supported.The first thing we need to do is determine the theoretical number of handaxes and choppers. We calculate this by dividing the sum of all cases by the number of categories. In this case, the sum of all artifacts is 100 and the number of categories (handaxes and choppers) is two. Thus, the theoretical frequency of handaxes and choppers is 100/2 =50.

	Observed #	Theoretical #	(O-T)	$(O-T)^2$	$(O-T)^2/T$
Handaxes	46	**50**			
Choppers	54	**50**			
Sum	100	**100**			

Next, we'll calculate the differences between observed and theoretical for each category and square those differences:

	Observed #	Theoretical #	(O-T)	$(O-T)^2$	$(O-T)^2/T$
Handaxes	46	50	**-4**	**16**	
Choppers	54	50	**4**	**16**	
Sum	100	100			

Next we divide each squared difference by the theoretical frequency for each category, which in our situation is 16/50 = 0.32 for each category:

	Observed #	Theoretical #	(O-T)	$(O-T)^2$	$(O-T)^2/T$
Handaxes	46	50	-4	16	**0.32**
Choppers	54	50	4	16	**0.32**
Sum	100	100			

Last, we simply add the last column in our table. This sum is the X^2 statistic.

	Observed #	Theoretical #	(O-T)	$(O-T)^2$	$(O-T)^2/T$
Handaxes	46	50	-4	16	0.32
Choppers	54	50	4	16	0.32
Sum	100	100			**0.64**

We now have to determine if the X^2 we calculated indicates that the difference in frequencies of handaxes and choppers is statistically significant, which is the way that we'll decide to accept or reject our hypothesis. So now we're going to compare our X^2 with the probability distribution of the X^2 statistic as shown in the **X^2 Table**.

The X^2 Table lists values of the X^2 statistic based on two considerations: **probability level** and **degrees of freedom**. What do these mean?

We determine the level of **probability** by deciding how often we are willing to make a mistake in our conclusion. If, for example, we are willing to be wrong 10% of the time, we will use the P_{90} column in the table. If we are willing to be wrong only 5% of the time, we'll select the P_{95} column.

Notice that the X^2 cut-off values for P_{90} and P_{95} increase in the first row from 2.71 to 3.84. Remember we said earlier that the larger the X^2 value, the greater the chance (probability) that the differences between observed and theoretical frequencies are significant. This should help you to see that the term "significance" is really a statement of how often we could be wrong in accepting or rejecting a hypothesis.

So hypothesis testing in a situation like this can yield an all-or-nothing result only if we define ahead of time how often we are willing to make a mistake. Alternatively, we could just see what the chances are for a mistake by looking at the X^2 probability we came up with.

The **degrees of freedom (df)** which make up the first column in the X^2 Table simply adjusts for the number of categories we are comparing. The degrees of freedom is one less than the number of categories (2-1) As you can see in the table under any probability column, that the cut-off X^2 values increase as the number of categories increase. We have two categories, so our degree of freedom is 1. If we had more categories, the X^2 cut-off would increase simply because we would add together more summed differences between observed and theoretical frequencies.

Making a Decision About the Hypothesis

We can now see the result of our X^2 analysis. The X^2 for our handaxe-chopper example is 0.64. First, we have to predetermine the probability level we want to use, let's say P_{95}, which would mean that we could make a mistake 5% of the time. By looking at the X^2 cut-off for the P_{95} probability and 1 df you can see that our X^2 is smaller than the expected X^2 value of 3.84.

If the X^2 value we calculate is *greater* than the selected X^2 probability, then the null

hypothesis is rejected. Thus, the likelihood is that the frequencies *are not due to chance* and the difference between the observed and theoretical frequencies of handaxes and choppers is *statistically significant*.

If the X^2 value we calculate is *less* than the selected X^2 probability then the null hypothesis cannot be rejected and the likelihood is that the frequencies *are due to chance* and the difference between the observed and theoretical frequencies of handaxes and choppers is *not statistically significant*.

So now you know that the frequencies of handaxes and choppers in my sample could easily be due to chance, and the difference in their frequencies is not statistically significant. Therefore we reject my hypothesis, and your suspicions are now statistically affirmed.

We can also use X^2 analysis to compare the frequencies of categories between two archaeological samples not to see if they are random, but to see if they are different or similar. This requires a slightly different approach, and one that you will use in your lab exercise to compare artifact samples you will analyze.

Testing for Randomness of Observations – Multiple Variables

In this next example, we are going to examine differences across two variables. The mechanics of conducting a X^2 analysis is similar to the previous example. The main difference is in the calculation of the theoretical values and the degrees of freedom. In this example, we will examine whether there are differences in the raw materials used to make handaxes and choppers.

Here are the frequencies of raw materials for handaxes and choppers from a site in East Africa:

	Basalt	Rhyolite	Total
Handaxes	30	24	54
Choppers	18	37	55
Total	48	61	109

Overall, it appears that rhyolite was somewhat preferred for making these tools than basalt, but were these raw materials used the same for the two tool categories or was there some preference by their makers? We can find out by testing the following null hypothesis:

H_o: Any differences between raw materials for handaxes and chopper are due to chance.

To calculate the theoretical values for our X^2 analysis, we use the following formula:

$$(R_i \times C_i)/T$$

where R_i is the row total (1 for handaxes and 2 for choppers), C_i is the column total (1 for basalt

and 2 for rhyolite) , and T is the total of all artifacts (109).

Here are the theoretical frequencies:

Basalt handaxes	(54x48)/109 = 23.8
Basalt choppers	(55x48)/109 = 24.2
Rhyolite handaxes	(54x61)/109 = 30.2
Rhyolite choppers	(55x61)/109 = 30.8

Calculating the X^2 is the same as in the previous example:

$$X^2 = \sum \frac{(O_i - T_i)^2}{T_i}$$

	Basalt	Rhyolite	Total
Handaxes	$(30-23.8)^2/23.8$=**1.62**	$(24-30.2)^2/30.2$=**1.27**	**2.89**
Choppers	$(18-24.2)^2/24.2$=**1.59**	$(37-30.8)^2/30.8$=**1.25**	**2.84**
Total			**X^2 =5.73**

The degrees of freedom for multiple variables is calculated slightly differently from in the previous example. Because we now have more than one column, we use the following formula:

$$df= (r-1)x(c-1)$$

where r is the number of rows and c is the number of columns. Thus in our example, there are two columns (basalt and rhyolite) and two rows (handaxes and choppers). So the degrees of freedom are calculated as follows: (2-1) x (2-1) = 1.

If we use the probability level of 95% with the degrees of freedom of 1, the selected probability cut-off (P_{95})in the X^2 Table 2 is 3.84. The X^2 value for our analysis (5.73) is greater than the expected value, thus the null hypothesis that raw material selection for handaxes and choppers is due to chance is rejected if we are willing to make an error in this decision no more than 5% of the time. On the other hand, if we wanted to risk an error in testing our hypothesis no more than 1% of the time (P_{99}) our X^2 is less than the cut-off of 6.63, and we would conclude that the differences in raw material selection are not significant. Thus, we get to decide the level of significance.

Percentiles of the X^2 Distribution

df	$P_{O.5}$	P_I	$P_{2.5}$	P_5	P_{IO}	P_{90}	P_{95}	$P_{97.5}$	P_{99}	$P_{99.5}$
1	0.000039	0.00016	0.00098	0.0039	0.158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.201	0.0506	0.1026	0.2107	4.61	5.99	7.38	9.21	10.60
3	0.0717	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.38	88.38	91.95
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64