Chi-Square Analyses

Chi-square analysis is one of the more complicated analyses in the AP Statistics curriculum. The complication is a result of several features: 1) the plurality of different analyses (Goodness-of-Fit, Independence, Homogeneity), 2) the more extensive computations required, 3) a computational formula that makes little "sense" to students, and 4) the focus on distributions rather than means. My experience suggests that the Chi-square computation is one that students have seen before (usually as Goodness-of-Fit in Biology or Genetics class). In this session, we will try to bring some clarity and understanding to the various Chi-square analyses.

Goodness of Fit

Often, the first introduction to Chi-square is the Goodness-of-Fit model. Consider the following problem:

> The chocolate candy Plain M&M's have six different colors in each bag: brown, red, yellow, blue, orange, and green. According to the M&M's website
> http://www.m-ms.com/factory/history/faq1.html
> the intended distribution of the colors is 30% brown, 20% red, 20% yellow, 10% blue, 10% orange, and 10% green. If several bags are opened, and we consider these to be a random sample of M&M's made, how well does this distribution describe the observed M&M's. Suppose we have 270 M&M's from 5 bags, and we have:

Color	Brown	Red	Yellow	Blue	Orange	Green
Observed Count	76	47	53	30	35	29
Expected Count	81	54	54	27	27	27

Are these counts consistent with the stated proportions? Are they what we expect to see, or are they surprising? We can perform a hypothesis test. The null hypothesis is that the proportions are as stated. The alternative hypothesis is that at least one proportion is different from those stated.

We compute $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$, which in this case is

$$\chi^2 = \frac{(76-81)^2}{81} + \frac{(47-54)^2}{54} + \frac{(53-54)^2}{54} + \frac{(30-27)^2}{27} + \frac{(35-27)^2}{27} + \frac{(29-27)^2}{27} = 4.086$$

Does this value of χ^2 raise questions about the validity of the null hypothesis?

If the null hypothesis is true, what values of $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$ would likely be computed and where does this value of 4.086 fit in that spectrum? Is it a typical value or an atypical value? We can create a simulation that will help answer this question, and see how our simulation compares to the theoretical result.

teague@ncssm.edu

Simulation

The following program will simulate this situation and allow us to see how unusual or how typical is our value of 4.086.

```
Sim (NumMMs) := | for  i ∈ 0 .. NumMMs
                |   color_i ← | x ← rnd(1)
                |             | 1  if  x ≥ 0   ∧  x < 0.3
                |             | 2  if  x ≥ 0.3 ∧  x < 0.5
                |             | 3  if  x ≥ 0.5 ∧  x < 0.7
                |             | 4  if  x ≥ 0.7 ∧  x < 0.8
                |             | 5  if  x ≥ 0.8 ∧  x < 0.9
                |             | 6  if  x ≥ 0.9 ∧  x ≤ 1.0
                | for  j ∈ 0 .. 6
                |   intervals_j ← j + 1
                | tally ← hist(intervals, color)
                | proportions ← (0.3, 0.2, 0.2, 0.1, 0.1, 0.1)^T
                | for  k ∈ 0 .. 5
                |   expected_k ← NumMMs · proportions_k
                | Σ_{k=0}^{5} (tally_k − expected_k)^2 / expected_k
```

If we run this simulation 10,000 times, we can plot a bar graph of the distribution of Chi-square values under the null hypothesis (see Figure 1). The sample distribution shown below with the dotted curve illustrating the sampling distribution of Chi-square with 5 degrees of freedom stretched by a factor of 10,000. (The appendix contains a TI-83 program that performs the same simulation, only much more slowly.)

From the bar graph in Figure 1, we see that the value of 4.086 is quite typical of the Chi-square values generated when the null hypothesis is true. We also see that the distribution of our 10,000 trials matches the continuous distribution fairly well, although it seems that we have a few too many very small values of Chi-square.

We can also see from Figure 1 that we shouldn't consider a value of Chi-square to be unusually large unless it is larger than 10 or 12. If we look at the continuous distribution, we find that 5% of the scores are expected to be larger than 11.07 when the null hypothesis is true. This is the critical value of Chi-square with 5 degrees of freedom for a significance level of 0.05.

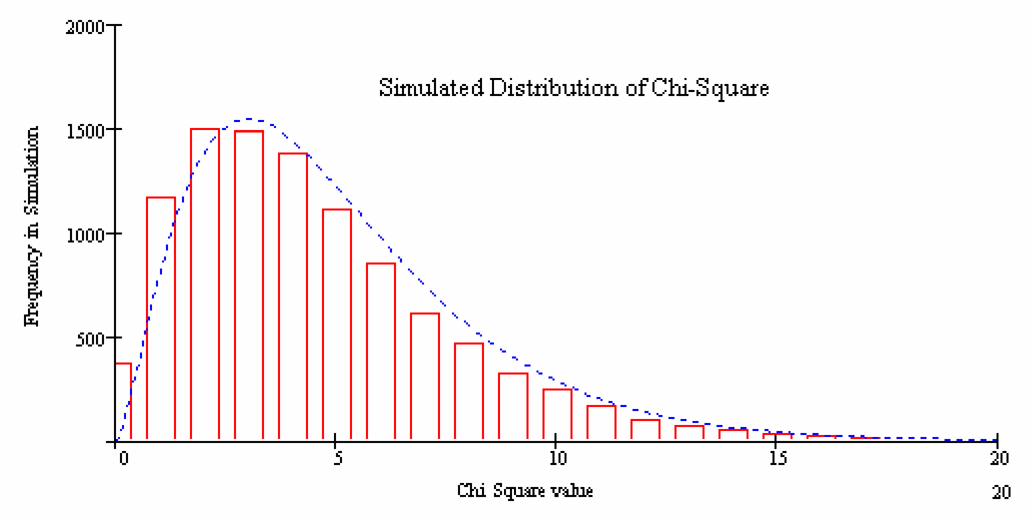

Figure 1: Bar graph of simulation Chi-square values and Theoretical Distribution

Expected Cell Counts > 5

The simulation also allows us to investigate the conditions for using the Chi-square distribution to approximate the theoretical (discrete) distribution of counts under the null hypothesis. Most texts require all cell counts to have an expected value of at least 5. What happens if this is not true?

Suppose the number in the sample is only 10. In this case, the largest expected cell count is only 3. The critical value of Chi-square for 5 degrees of freedom at the 0.05 level of significance is 11.07. We should expect to find approximately 50 of the 1000 runs of the simulation to have a Chi-square value greater than 11.07. However, if the number in the sample is 10, we find many more simulations with Chi-square scores larger than 11.07. We reject the null hypothesis much more frequently than we should. We ran the simulation of 1000 trials ten times each with $N = 10, 20, 30, 40, 50, 60\ 70$, and 100. As N increases, more of the expected counts are 5 or greater. Once N = 50, all expected counts are at least 5, so we would expect to find close to 5% of the 1000 trials with a Chi-square value greater than 11.07. Table 1 gives the number of trials over the critical value for each of the 10 repetitions.

	N = 10	N = 20	N = 30	N = 40	N = 50	N = 60	N = 70	N = 100
1	79	66	57	58	56	50	49	53
2	75	58	64	48	48	57	68	57
3	79	87	57	70	51	40	51	47
4	74	57	58	57	54	53	48	53
5	72	47	61	46	50	61	54	62
6	82	61	61	41	47	44	47	53
7	65	55	56	55	50	61	55	47
8	79	60	55	49	63	45	52	46
9	70	53	51	59	52	49	51	50
10	81	53	73	68	46	50	48	42
Mean	75.6	59.7	59.3	55.1	51.7	51.0	52.3	51.0

Table 1: The Number of Chi-square Scores above 11.07

teague@ncssm.edu

Notice that as the number of cells whose expected value exceeds 5 increases, the mean number of Chi-square values larger than 11.07 decreases to around 50 per 1000 trials. Once the number of M&M's sampled exceeds 50, all of the means remain essentially constant at around 50, as expected.

Understanding the Formula $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$

Just what is being measured in the computation of $\sum_i \frac{(O_i - E_i)^2}{E_i}$? How can we make sense out of this expression? We are interested in comparing our observed counts with the expected counts, so $(O_i - E_i)$ certainly makes sense. However, there are three problems with just summing up the terms $(O_i - E_i)$.

1) The first problem is one of scale or size. Suppose $O - E = 4$? Is this a large error or a small error? If $E = 5$ and $O = 9$, and error of 4 is quite large. However, if $E = 5000$ and $O = 5004$, an error of 4 is quite small. The absolute difference in O and E is often not the most important measure. We have a standard way to deal with this. We use the relative error. That is, we consider $\sum_i \frac{O_i - E_i}{E_i}$.

2) But adding up all of the relative errors leads to a second problem, the problem of sign. Some of the errors are positive and some are negative, so the total sum of the relative errors will be zero. There is also a standard way to handle this difficulty. We square the terms prior to adding them. So we are interested in $\sum_i \left(\frac{O_i - E_i}{E_i}\right)^2$.

3) But this also produces a different problem of scale. Suppose the squared relative error is 1/25. Is this a large error or a small error? If the expected count is 5, then we have a small error (1 out of 5) and so our squared relative error of 1/25 should not count too much against the null hypothesis. If the expected count is 5000, we are off by quite a lot (1000 out of 5000), so our squared relative error of 1/25 should count a lot against the null hypothesis. We also have a standard way to handle this difficulty. We use a weighted sum. We weight the squared relative error by the expected count size, so we end up with

$$\sum_i \left(\frac{O_i - E_i}{E_i}\right)^2 \cdot E_i = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

which is our standard computational formula.

Tests of Independence and Homogeneity

There are two forms of the Chi-square test that often get confused, the Chi-square test of independence and the Chi-square test of homogeneity. The test of independence attempts to determine whether two characteristics associated with subjects in a single population are independent. The test of homogeneity attempts to determine whether several populations are similar (homogeneous) with respect to some variable.

teague@ncssm.edu

We can use Problem #2 from the 1999 AP exam as an example to compare these two tests. In this problem, a random sample of 200 hikers was taken. The hikers were asked if they would walk uphill, downhill, or remain where there were if they became lost in the woods. They hikers were classified as either a Novice or an Experienced hiker. The question asks if there is an association between the responses "Uphill", "Downhill", and "Remain", and the classification as Novice or Experienced.

Independence

As stated, this is a Chi-square test of independence since a single sample from the population of hikers was taken, and the individuals sorted into the cells classified as Novice/Experienced and Uphill/Downhill/Remain. The null hypothesis is that there is no association between the two variables. As in all hypothesis tests, we use the null hypothesis to compute the expected values. The table of counts looked like this:

	Uphill	*Downhill*	*Remain*
Novice	20	50	50
Experienced	10	30	40

Table 2: Counts for Test of Independence

If the null hypothesis is true, how would the hikers be sorted into the 6 cells? We need to determine the expected counts in each cell. To determine the expected counts, consider the following three questions.

1. If a hiker from this sample were selected at random, what is the probability that they would be a Novice? Answer: p(N) = 120/200. Also note that p(E) = 80/200.

2. If a hiker from this sample were selected at random, what is the probability that they would go Uphill? Answer: p(U) = 30/200. Also note that p(D) = 80/200 and p(R) = 90/200.

3. Based on these two probabilities, if the level of hiking experience and the direction a hiker would travel if lost are independent, what is the probability that a hiker selected at random would be a Novice who would head Uphill?
Answer: p(N and U) = p(N)*p(U) = $\left(\frac{120}{200}\right)\left(\frac{30}{200}\right) = \frac{36}{400} = \frac{9}{100}$.

The assumption of independence allows us to compute the expected number to be found in this cell of our table. Since there are 200 hikers altogether, if the row and column variables are independent, we would expect to see $\left(\frac{9}{100}\right)\cdot 200 = 18$ hikers in this cell. We actually have 20 hikers in this cell. We repeat this calculation for each cell in the table and ask the question: Are these observed counts consistent with the expected counts computed under the assumption of independence? Notice that this computation is equivalent to $\frac{120\cdot 30}{200}$, the $\frac{(\text{row count})(\text{column count})}{\text{total}}$, which is given by most texts as the computational formula.

Observations	*Uphill*	*Downhill*	*Remain*
Novice	20	50	50
Experienced	10	30	40

Expectations	*Uphill*	*Downhill*	*Remain*
Novice	$200\left(\frac{120}{200}\right)\left(\frac{30}{200}\right)=18$	$200\left(\frac{120}{200}\right)\left(\frac{80}{200}\right)=48$	$200\left(\frac{120}{200}\right)\left(\frac{90}{200}\right)=54$
Experienced	$200\left(\frac{80}{200}\right)\left(\frac{30}{200}\right)=12$	$200\left(\frac{80}{200}\right)\left(\frac{80}{200}\right)=32$	$200\left(\frac{80}{200}\right)\left(\frac{90}{200}\right)=36$

Table3: Computing Expected Counts Under the Assumption of Independence

Degrees of Freedom

There were 200 hikers in the sample, 120 classified as being Novices and 80 classified as Experienced. We also know that 30 of the hikers answered they would move Uphill, 80 that they would move Downhill, and 90 that they would Remain where they were. We describe this information as "knowing the marginal totals". When computing the expected values, once we know that there are 18 Novices expected to answer "go Uphill", we know without computing that there must be 12 Experienced hikers expected to answer "go Uphill" (since there were 30 going uphill altogether). Also, once we have determined that there are 48 Novices expected to answer "go Downhill", we know there must be 32 Experienced hikers expected to give that answer (since there were 80 in total who answered Downhill). Moreover, the expected number of Novices answering "Remain" must be 54, since there were 120 Novices and 18 and 48 have already been allocated. In a similar manner, we know the expected number of Experienced hikers who answer "Remain" must be 36, since there were 80 in total and 44 have been allocated. In short, knowing only the expected values in two cells allows us to complete the table of expected counts. We say there are "two degrees of freedom" and our Chi-square statistics will have 2 degrees of freedom. This is classically stated as $df=(\text{row}-1)(\text{column}-1)$.

Historical Note (as told by Chris Olsen)**:** The Chi-square statistic was invented by Karl Pearson about 1900. Pearson knew what the Chi-square distribution looks like, but he was unsure about the degrees of freedom. About 15 years later, Fisher got involved. He and Pearson were unable to agree on the degrees of freedom for the two-by-two table, and they could not settle the issue mathematically. Pearson believed there was 1 degree of freedom and Fisher 3 degrees of freedom. They had no nice way to do simulations, which would be the modern approach, so Fisher looked at lots of data in two-by-two tables where the variables were thought to be independent. For each table he calculated the Chi-square statistic. Recall that the expected value for the Chi-square statistic is the degrees of freedom. After collecting many Chi-square values, Fisher averaged all the values and got a result he described as "embarrassingly close to 1." This confirmed that there is one degree of freedom for a two-by-two table. Some years later this result was proved mathematically.

Homogeneity

Now, suppose instead that a random sample of 120 Novice hikers was taken from a population of Novice hikers and a random sample of 80 Experience hikers was taken from a population of Experienced hikers. Each member in the two samples was asked about the direction they would travel when lost. This is a test of homogeneity since the are two populations (Novice and Experience) being classified on one variable (direction when lost). Suppose the results were the same as those above and the table of counts is shown below.

	Uphill	Downhill	Remain
Novice	20	50	50
Experienced	10	30	40

Table 2: Counts in Example Problem

The null hypothesis for this test of homogeneity is that the proportions of hikers falling into the three direction categories are the same for Novice and Experienced hikers. We use this null hypothesis to compute the expected values. Since this is a different null hypothesis from the test of independence, we would expect our computations to differ, and they do.

To find the expected value of Novice-Uphill, we note that there were 200 hikers altogether, and 30 of them would travel uphill. So the proportion of hikers who would travel uphill is 30/200. The null hypothesis is that this proportion of Novice hikers and this proportion of Experience hikers would travel uphill. It is the same for both. There are 120 Novice hikers, so the expected number of hikers in the Novice-Uphill cell is $\left(\frac{30}{200}\right)\cdot 120 = 18$ and the expected number of hikers in the Experienced-Uphill cell is $\left(\frac{30}{200}\right)\cdot 80 = 12$. Continuing in this manner, we complete the expected counts in the table.

Observations	*Uphill*	*Downhill*	*Remain*
Novice	20	50	50
Experienced	10	30	40

Expectations	*Uphill*	*Downhill*	*Remain*
Novice	$\left(\frac{30}{200}\right)\cdot 120 = 18$	$\left(\frac{80}{200}\right)\cdot 120 = 48$	$\left(\frac{90}{200}\right)\cdot 120 = 54$
Experienced	$\left(\frac{30}{200}\right)\cdot 80 = 12$	$\left(\frac{80}{200}\right)\cdot 80 = 32$	$\left(\frac{90}{200}\right)\cdot 80 = 36$

Table 4: Computing Expected Counts Under the Assumption of Homogeneity

Notice two things:

1) the computations for determining the expected number in each cell for tests of independence and for homogeneity are different, since they are derived from the different null hypotheses.

2) the results of these computations are the same. Since the results of the computations are the same, the distinction between tests of independence and tests of homogeneity is often considered academic in an introductory course like AP Statistics.

Equivalence of Chi-square Tests and *Z*-Tests for Proportions

Suppose we want to perform a test of proportions. Do we use a z-test or a Chi-square test? Suppose the national population proportion of overweight high school students is 58%. You take a random sample of 100 student records and determine that 64% are overweight. Does this provide evidence that your student population differs with regard to the percentage overweight from the national percentage?

We can do a one-proportion *z*-test.

$H_0 : p = 0.58$

$H_a : p \neq 0.58$

$$z = \frac{0.64 - 0.58}{\sqrt{\frac{(0.58)(0.42)}{100}}} = \frac{0.06}{0.04936} = 1.21566$$

so $P = 2(0.112) = 0.224$.

We can do a Chi-square test.

$H_0 : p = 0.58$

$H_a : p \neq 0.58$

$$\chi^2 = \frac{(64-58)^2}{58} + \frac{(36-42)^2}{42} = 1.47783$$

with 1 degree of freedom, so $P = 0.224$.

There is no evidence that your school population proportion differs from the national proportion. Notice that the *p*-values are the same and that the value of χ^2 is the square of the value of *z*. The two tests are identical.

We can show that *z*-tests for proportions are algebraically equivalent to Chi-square tests. For a single proportion where the hypotheses are

$$H_0: p = p_0 \text{ and } H_a: p \neq p_0.$$

To perform a Chi-square test, we would have observed and expected values as shown in the table below:

	Yes	No	Total
Observed	y	$n-y$	n
Expected	$n\,p_0$	$n(1-p_0)$	n

Since $\hat{p} = y/n$, $y = n\hat{p}$, this table is equivalent to

	Yes	No	Total
Observed	$n\,\hat{p}$	$n - n\,\hat{p}$	n
Expected	$n\,p_0$	$n(1-p_0)$	n

The Chi-square statistic with one degree of freedom is

$$\chi^2 = \sum_{i=1}^{2} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$= \frac{(y-np_0)^2}{np_0} + \frac{[(n-y)-n(1-p_0)]^2}{n(1-p_0)} = \frac{(n\hat{p}-np_0)^2}{np_0} + \frac{[(n-n\hat{p})-n(1-p_0)]^2}{n(1-p_0)}$$

$$= \frac{n^2(\hat{p}-p_0)^2}{np_0} + \frac{n^2[(1-\hat{p})-(1-p_0)]^2}{n(1-p_0)} = \frac{n(\hat{p}-p_0)^2}{p_0} + \frac{n(-\hat{p}+p_0)^2}{(1-p_0)}$$

teague@ncssm.edu

$$= n(\hat{p} - p_0)^2\left[\frac{1}{p_0} + \frac{1}{1-p_0}\right] = n\left(\hat{p} - p_0\right)^2\left[\frac{1}{p_0(1-p_0)}\right] = \frac{(\hat{p}-p_0)^2}{p_0(1-p_0)/n}$$

This last expression we recognize as Z^2 with $Z^2 = \left[\dfrac{\hat{p}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}\right]^2$

Note that large values of n are required for $\sum_{i=1}^{2}\dfrac{(\text{observed}-\text{expected})^2}{\text{expected}}$ to be approximately χ^2 distributed and for $\dfrac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$ to be approximately Z. We already know that Z^2 is χ^2 with 1 degree of freedom. It should be no surprise that the critical value for χ^2 with 1 degree of freedom with $\alpha = 0.05$ is $1.96^2 = 3.84$.

To see the derivation for 2-proportions see page 69-70 in *Hypothesis Testing* under *Theory of Inference* at http://courses.ncssm.edu/math/Stat_Inst/Notes.htm. For other concepts related to Chi-square, see *Categorical Data Analysis* under *Categorical Data Analysis and Surveys* at the same website.

Another look at cell counts greater than 5 (as described by Floyd Bullard)

Recognizing the equivalence between χ^2 and two-proportion *z*-tests gives us another way to consider the conditions on cell counts for χ^2. The cell count greater than 5 is the same condition as $np > 5$ for proportions. One way to think about the *np* condition is to consider confidence intervals for proportions. If, when we compute a 95% confidence interval for proportions, we obtain an interval (−0.1, 0.3) or (0.85, 1.05), we know that our interval is based on too few data, since the endpoints exceed the limits of 0 and 1. To have confidence in our confidence interval, the endpoints of the interval must be between 0 and 1. How large must *n* be to achieve this?

We need $p - 2\sqrt{\dfrac{p(1-p)}{n}} > 0$ and $p + 2\sqrt{\dfrac{p(1-p)}{n}} < 1$ where p is the population parameter (approximated by $\hat{p}$).

If $p - 2\sqrt{\dfrac{p(1-p)}{n}} > 0$, then $p > 2\sqrt{\dfrac{p(1-p)}{n}}$ or $p^2 > \dfrac{4p(1-p)}{n}$. Simplifying, we find that $np > 4(1-p)$. This lower bound is only an issue when *p* is small and $1-p$ is close to 1. So if $np > 4$, we will get an interval whose lower bound is greater than 0. But we don't know *p*, we only know $\hat{p}$. By requiring $n\hat{p} > 5$, we give ourselves some wiggle-room and confidence that, unless our $\hat{p}$ is far from *p*, our confidence interval will be OK. Solving $p + 2\sqrt{\dfrac{p(1-p)}{n}} < 1$ in a similar manner and adding in the wiggle-room for $\hat{p}$ will generate $n(1-\hat{p}) > 5$. If we want 99% confidence intervals we need $np > 9$ or $n\hat{p} > 10$.

teague@ncssm.edu

Appendix: TI-83 Program for Chi-Square Simulation

```
PROGRAM:MMS
ClrHome
Disp "NUM. MMS?"
Input M
Disp "NUM. SIMS?"
Input S
ClrHome
ClrList ʟSAMP, ʟCOLOR, ʟTALLY, ʟEXPTD, ʟCHISQ
For(I,1,S)
Output(1,1,"    ")
Output(1,1,S+1-I)
For(C,1,M)
prgmGENCOLOR
B→ʟSAMP(C)
End
seq(X,X,1,6)→ʟCOLOR
For(X,1,6)
0→D
For(Y,1,M)
If ʟSAMP(Y)=X
D+1→D
End
D→ʟTALLY(X)
End
{.3,.2,.2,.1,.1,.1}*M→ʟEXPTD
sum((ʟTALLY-ʟEXPTD)²/ʟEXPTD)→ʟCHISQ(I)
End
```

```
PROGRAM:GENCOLOR
0→B
rand→A
If A≤.3
1→B
If A>.3 and A≤.5
2→B
If A>.5 and A≤.7
3→B
If A>.7 and A≤.8
4→B
If A>.8 and A≤.9
5→B
If A>.9 and A≤1
6→B
```


teague@ncssm.edu