Math 316

Section 9.4 The Chi-Square-Goodness-of-Fit Test

In Sections 9.2 and 9.3 we looked at problems with one population (e.g. 3 to 5 year old children) whose variable (e.g. iron deficient?) had two possible values (e.g. either iron deficient or not). In this section there is still one variable (e.g. flower color) for the population but with multiple possible values (white, yellow, green) rather than just two. What we are measuring now is how the observed proportions of each possible value match the expected proportions. Let's work **HW 9.4.1**, **plus two variations**.

In **Table 9** we see that (as always) a larger test statistic gives us a smaller P value. If there are more than two possible values, then the test is inherently non-directional. For two possible values, you can have direction. See the **note** at the top of Table 9.

For the case of two (rather than more than two) possible values for our category, it turns out that χ_s^2 is actually t_s^2, that is, $\sqrt{\chi_s^2} = t_s$. The book doesn't discuss this, but I think it worth knowing, so we'll briefly look at this below. So χ_s^2 is simply a generalization of the test statistic t_s for more than one (or more) population/category and two or more possible values for each category variable. Put another way, t_s is the special case of χ_s^2 when there are just two categories (e.g. iron deficient or not). We'll see further generalizations of this in Chapter 10.

Reminders of several ideas:

- A binomial distribution is used when dealing with a variable that either has one value or another: made free throw or not, iron deficient or not, etc.
- For a binomial distribution, the standard deviation is $\sqrt{n\,p\,(1-p)}$. See Theorem 5.4.1 on book page 164. And if you like, you can see this by experimenting a bit with the **Binomial distribution calculator** online at the class homepage.
- For normally distributed data, the z-value is
$$z = \frac{Y-\mu}{\sigma} = \frac{Observed\ value - Expected\ value}{Standard\ deviation}$$
- A binomial distribution is approximately normal, and the larger the sample, the more normally distributed it is. So when dealing with proportions, we assume normal distribution.
- The area beyond a z-value is the likelihood of getting that z-value or larger. Similarly, the area beyond a t-value (i.e. t_s) is the likelihood of getting that t-value or larger.

Putting these ideas together, for a binomial distribution, with sample size n, the likelihood that of getting a particular Observed value (or larger) given the Expected value is the area corresponding to
$$z = \frac{Observed\ value - Expected\ value}{Standard\ deviation} = \frac{O-E}{\sqrt{n\,p\,(1-p)}}.$$
Let's work **Class Example 2**.

Let me remind you that one of the reasons that we usually find a t-value (e.g. t_s) and use Table 4 rather than a z-value and Table 3 is that we don't have perfect normal distribution in our data. The larger the sample, the more we have normal distribution, and you'll notice that for infinitely large samples we can assume perfect normal distribution. This is why the bottom row of Table 4 corresponds exactly with Table 3, but in reverse, as we've discussed in class.

We saw for Class Example 2 that $\chi_s^2 = t_s^2$, that is, $\sqrt{\chi_s^2} = t_s$. We can actually show this in general (reminder: this is just for the case when the category variable has *two* possible values). To make this a bit simpler to discuss, let's suppose this is still for the child iron deficiency example. Suppose for a sample of size n we have:

	Iron deficient	Not deficient	Total
Observed	a	$n-a$	n
Expected	b	$n-b$	n

We could compute

$$\chi_s^2 = \frac{(a-b)^2}{b} + \frac{((n-a)-(n-b))^2}{n-b} = \frac{(a-b)^2}{b} + \frac{(b-a)^2}{n-b} = (a-b)^2\left(\frac{1}{b} + \frac{1}{n-b}\right) = \frac{n\cdot(a-b)^2}{b\cdot(n-b)}.$$

Or we could compute

$$t_s = \frac{O-E}{\sqrt{n\,p\,(1-p)}} = \frac{a-b}{\sqrt{n\left(\frac{b}{n}\right)\left(1-\frac{b}{n}\right)}} = \frac{(a-b)}{\sqrt{\frac{b\cdot(n-b)}{n}}} = \frac{\sqrt{n}\cdot(a-b)}{\sqrt{b\cdot(n-b)}}.$$

(We can verify these two formulas in Class Example 2.) We see that $t_s^2 = \chi_s^2$. I'll reiterate that $t_s^2 = \chi_s^2$ only when there are two possible values for the variable (e.g. either *iron deficient* or *not*) rather than > 2 (e.g. three flower colors).

Here's one more way to see the connection between t_s and χ_s^2 values, this time using P values. Below are (1) the first row of Table 9, for which $df = 2-1 = 1$, and (2) parts of the final row of Table 4, for which $df = \infty$ (which means we have perfect normal distribution, and recall that we can treat a binomial distribution as normal).

Table	df	Area in both tails				
		P	$.20$	$.10$	$.05$	$.02$	etc.
9	1	χ_s^2	1.64	2.71	3.84	5.41	etc.
		$\sqrt{\chi_s^2}$	1.28	1.65	1.96	2.33	etc.
4	∞	t_s	1.28	1.65	1.96	2.33	etc.

Read through Section 9.5 on your own.