Chapter 9 Class Notes – One-Sample Categorical Data

§9.1 Return to Dichotomous Data and Proportions

Our ultimate goal here is to set a confidence interval (CI) for p, a Binomial population proportion, deferred to §9.2. Recall for Y~Binomial(n,p), the mean and variance are $\mu = np$ and $\sigma^2 = np(1-p)$ respectively. It follows that a good estimator of p might be the Wald estimate, $\hat{p} = \frac{y}{n}$, which has mean $\mu_{\hat{p}} = p$ and standard error $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. We'll find, however, that this Wald estimator can produce unreliable CI's, so we instead will use the Wilson-adjusted sample proportion:

$$\tilde{p} = \frac{y+2}{n+4}$$

Ex.9.1.1/9.12 – soft drinks and bacteria: y = 5 successes out of n = 30, so $\hat{p} = \frac{5}{30} = 0.167$ and $\tilde{p} = \frac{7}{34} = 0.206$. Note that $\tilde{p}$ is *always* closer to ½ than is $\hat{p}$.

Ex.9.1.3 illustrates the sampling distribution of $\tilde{P}$; we can contrast this with the sampling distribution of $\hat{P}$: let Y be Binomial with n = 2 and p = 0.17. Then here are the sampling distributions:

Y	$\hat{P}$	$\tilde{P}$	Probability
0	0	⅓	0.6889
1	½	½	0.2822
2	1	⅔	0.0289

Additional examples are on pp.338-340.

§9.2 Confidence Interval for P

In §6.3, we used $\bar{y} \pm t_{\alpha/2} SE_{\bar{Y}}$ as a CI for μ. We also saw that the normal approximation can be used for binomial situations when n is large, so the 95% CIs here will use 1.96 in place of $t_{0.025}$. It is tempting to use $\hat{p} \pm 1.96 SE_{\hat{p}}$, but as noted above, there are often problems associated with this Wald method, so we use $\tilde{p}$ instead.

Also, note that $SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$; 95% CIs for p are thus:

$$\tilde{p} \pm 1.96 SE_{\tilde{p}}$$

To illustrate (p.346, ex.9.2.5), y = 69 adverse reactions (AR) to a vaccine were recorded in n = 339 studied infants. Thus, $\tilde{p} = \frac{71}{343} = 0.2070$ and $SE_{\tilde{p}} = \sqrt{\frac{0.2070(0.7930)}{343}} = 0.0219$, and the 95% CI for p is 0.2070 ± 0.0429 = (0.1641, 0.2499). In conclusion, we are 95% confident that the proportion of all infants receiving this vaccine injection experiencing adverse reaction is between 16.4% and 25.0%.

For *planning a study to estimate p* using a CI (p.345), we have two approaches – one where we have an educated guess of p (called $\tilde{p}^*$) and one where we do not and want to use a conservative approach. Suppose that we desire that $SE_{\tilde{p}} \leq c$. Thus, we choose n so that

$$\sqrt{\frac{\tilde{p}^*(1-\tilde{p}^*)}{n+4}} \leq c \rightarrow \text{ thus, } \quad n \geq \frac{\tilde{p}^*(1-\tilde{p}^*)}{c^2} - 4$$

For the left-handedness example (p.345), $\tilde{p}^* = \frac{40+2}{400+4} = 0.1040$. If c = 0.01, then $n \geq \frac{0.1040(1-0.1040)}{0.01^2}$ - 4 = 931.8 - 4 = 927.8, so n = 928.

If we don't have an educated guess, to be conservative we use the above formula but with $\tilde{p}^* = ½$ – so we choose $n \geq \frac{1}{4c^2} - 4$

For the above left-Handedness example, we would now have
If we had no educated guess for p, we would choose a sample size of $n \geq \frac{1}{4(0.01)^2} - 4 = 2496$.

Note that we could have provided formulas here using the margin of error: $ME_{\tilde{p}} = z_{\alpha/2} SE_{\tilde{p}}$ but instead simply used $SE_{\tilde{p}}$.

§9.3 (p.347): What about other CIs – other than 95%?
Here, we use:

$$\tilde{p} = \frac{y + \frac{1}{2}z_{\alpha/2}^2}{n + z_{\alpha/2}^2} \qquad \text{and} \qquad SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + z_{\alpha/2}^2}}$$

So, continuing the above infant/AR example for a 90% CI,
$\tilde{p} = \frac{69 + \frac{1}{2}1.645^2}{339 + 1.645^2} = 0.2059$ and $SE_{\tilde{p}} = \sqrt{\frac{0.2059(0.7941)}{339 + 1.645^2}} = 0.0219$;
the 90% CI for p is 0.2059 ± 0.0360 = (0.1699, 0.2419). We are 90% confident that the proportion of adverse reaction in all infants receiving this vaccine injection is between 17.0% and 24.2%.

In §9.4, we discuss Goodness of Fit (GOF) tests. Ex. 9.4.1 on p.357 gives data related to progeny of crosses of W and Y summer squash:

Color	W	Y	G	Total
# of progeny:	155	40	10	205

Using these 205 progeny, do these data support the 12:3:1 ratio that comes from the genetic model/theory? Here the relevant TS is:

$$\chi_s^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

We use the genetic theory (H_0) to calculate the respective e_i – here they are $\frac{12}{16} \times 205 = 153.75$, $\frac{3}{16} \times 205 = 38.4375$ and $\frac{1}{16} \times 205 = 12.8125$, so the calculated TS is

$$\chi_s^2 = \frac{(155 - 153.75)^2}{153.75} + \frac{(40 - 38.4375)^2}{38.4375} + \frac{(10 - 12.8125)^2}{12.8125}$$

$$= 0.1008^2 + 0.2520^2 + (-0.7857)^2 = 0.6910$$

To find the associated p-value, we use the χ^2 table (Table 9 on p.627) with df = # of categories – 1. Here, df = 3-1 = 2. Using Table 9, we get p-value > 20% since 0.69 = $\chi_s^2 < \chi_{0.20}^2 = 3.22$. We conclude that *there is no evidence in these data to contradict the genetic theory (i.e., the data are consistent with the model)*. Note here that the test is not based on relative frequencies, but on frequencies!

<u>Note on Sample Size</u>: Suppose that we found exactly the same sample proportions as above ($\hat{p}_1 = 0.756, \hat{p}_2 = 0.195, \hat{p}_3 = 0.0.049$), but with ten times the sample size – i.e., with observed frequencies:

Color	W	Y	G	Total
# of progeny:	1550	400	100	2050

Then, the TS is $\chi_s^2 = 10 \times 0.69 = 6.90$, 0.02 < p-value < 0.05, so we would then reject H_0 (i.e., reject the genetic theory).

We could state this H_0 as H_0: "12:3:1 ratio applies here", or better yet

$$H_0: p_1 = Pr(W) = \frac{12}{16}, p_2 = Pr(Y) = \frac{3}{16}, \text{ and } p_3 = Pr(G) = \frac{1}{16}$$

This is called a ***compound hypothesis*** since it contains several parameters and more than one equal sign. Pay particular attention to the statement of H_A here! One way to state H_A would be H_A: at least one of the p's specified in H_0 is incorrect, that is,

$$H_A: p_1 \neq \frac{12}{16} \text{ and/or } p_2 \neq \frac{3}{16} \text{ and/or } p_3 \neq \frac{1}{16}$$

We don't run into this challenge with dichotomous variables (i.e., when there are just two categories) – as we see next.

Ex.9.S.18 on p.362, let p be the probability that a randomly selected man correctly identifies his partner. Here, the null hypothesis is that the man is guessing versus the alternative that he does indeed have some identifying ability. Thus, we have:

H_0: p = ⅓ (men are guessing) versus H_A: p > ⅓

The data are: Correct: 16 (12) Wrong: 20 (24)
And the TS, df, p-value are:

$$\chi^2_s = \frac{(16-12)^2}{12} + \frac{(20-24)^2}{24} = 2.0; \text{ df} = 2 - 1 = 1; 0.05 < p < 0.10$$

It's important to note here that since this is a one-tailed test, we half the areas to get the p-value. We can get the exact p-value here from the Normal table by finding $Pr(Z > \sqrt{2.0} = 1.41) = 0.0793 > 0.05$.
There is insufficient evidence (0.05 < P < 0.10) to conclude that the men have some ability to detect their partners by touching the backs of their hands; the data are consistent with guessing.