

STAT 515 - Annotations to the Text

Brian Habing - University of South Carolina

Last Updated: December 31, 2003

In several places the textbook uses some shorthand notation that might seem confusing on first glance. There are also a few places where it presents things in a more complicated fashion than it needs to. The notes below are given by section number and will hopefully make reading the text a bit easier once in a while. There are also seven supplements available on the course web-page that contain additional material. Chapters and sections with supplements are noted below in the appropriate place.

Section 3.8: Notice that we can think of the permutations rule (page 156) and the combinations rule (page 158) as both being special cases of the partitions rule (page 157). A permutation is a partition where there are n groups of size 1 and one group of size $N-n$. This is the case where each of the first n selected are selected for different purposes and the order is important, but we don't care about the order of those not selected. A combination is a partition where one group is of size n and the other is of size $N-n$. This is the case where the first n are interchangeable with each other and the remaining $N-n$ are also interchangeable.

Section 4.4: $q=1-p$ is just shorthand. Don't let the q throw you off.

Section 5.4: We will always use method 4 in the box on page 237, the normal probability plot (also called the q-q plot). Exercise 5.45 on page 241 gives three sample q-q plots (with answers in the back of the text).

Histograms (method 1) are bad because they are easily manipulated by choosing different class intervals. Figuring out the percentages (method 2) is time consuming. Comparing the IQR to s (method 3) isn't built into most packages and doesn't give as much information as the q-q plot.

Section 5.5: The explanations in this section are a lot more complicated than they need to be. The entire key to the section can be seen by simply reading the material on page 244 and at the top of page 245.

First, notice in Figure 5.18 that they are looking for the probability that the binomial random variable x will be less than or equal to 10. In order to get all of the histogram bars for 10 and less they need to take everything from 10.5 and smaller. If you started at 10 you would miss half of the 10 bar. If you started at 11 you would include an extra half of the 10 bar. If they had asked

“< 10” we would have needed to start at 9.5, if they had asked “>10” we would have needed to start at 10.5, and if they had asked “≥10” we would have needed to start at 9.5. This going up or down by 0.5 is the *continuity correction*, and the safest way to see which way you need to go is to draw the picture.

Second, notice that they are using $\mu = \mu_x = np$ and $\sigma = \sigma_x = \sqrt{np(1-p)}$ because we are looking at the binomial distribution (box on the bottom of page 194).

Third, an easier rule for seeing if n is large enough is to say that n is large enough for the normal approximation to work if both $np \geq 5$ and $n(1-p) \geq 5$. Technically this condition is weaker than the one using $\mu \pm 3\sigma$, but it seems to work fairly well in practice and will match a rule we will use in Chapter 13. The accuracy of the normal approximation to the binomial, even for large n , is a topic of current research by statisticians, and we will see in section 7.3 that some tricks can be done to make it work even better in certain circumstances.

Applying this to example 5.12 we would do the following:

1. $np = 200(0.06) = 12 \geq 5$ and $n(1-p) = 200(1-0.06) = 200(0.94) = 188 \geq 5$ so the sample size is large enough for the normal approximation to the binomial to be reasonable. Notice that we've already found $\mu = np = 12$. Just using the formula for s gives us

$$\sigma = \sqrt{np(1-p)} = \sqrt{200(0.06)(1-0.06)} = \sqrt{11.28} \approx 3.359$$

2. If we want $P(x \geq 20)$ then we want to include the 20 bar, so we need to start at 19.5. (The uncolored in area in Figure 5.20... there is no reason to switch it around like the book does).

$$P(x \geq 20) = P(x \geq 19.5) = P\left(\frac{x - \mu}{\sigma} \geq \frac{19.5 - \mu}{\sigma}\right) = P\left(\frac{x - np}{\sqrt{np(1-p)}} \geq \frac{19.5 - 12}{3.359}\right) \approx P(z \geq 2.23)$$

This is now just a probability to look up on the normal table, and we get $0.5 - 0.4871 = 0.0129$.

Section 6.3: When reading the examples in this section it is important to note the blue box on the bottom of page 274 that says that $\mu_{\bar{x}} = \mu$ and that $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

See the supplement: Chapter 6 - More on Sampling Distributions: t, chi-square and F

Section 7.1: As in section 6.3, note that $\mu_{\bar{x}} = \mu$ and that $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

Also, Since we virtually never know σ , and since we can always use the t-table, you should probably never use $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ or $\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$. Instead you should always use $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ that is discussed in section 7.2 (unless for some bizarre reason you actually know σ).

Section 7.3: In the boxes on page 309 notice first that $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$ which we approximate by $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ if we aren't given p (and we never are for a confidence interval!) Second, notice that we can again use the $np \geq 5$ and $n(1-p) \geq 5$ rule discussed above instead of the $\hat{p} \pm 3\sigma_{\hat{p}}$ rule, we'll just have to use the \hat{p} instead of the p .

Finally, it is probably always best to use the Agresti and Coull formula in the box on the bottom of page 311 than it is to use the formulas in the box on the bottom of page 309. The correction basically says "pretend your sample was four larger and that two of those four were successes". The reason this works is fairly complicated to explain, but a computer simulation studies can be performed that show that in a large number of cases that the normal approximation to the binomial works better using this correction in the context of confidence intervals for a single percentage. We will not use it for a test of hypotheses or a confidence interval for two proportions.

Section 7.4: The entire key to this section can be found in the top equations in the blue boxes on page 316 and 318. The basic idea is to simply set the "plus or minus" portion of the confidence interval equal to the size you want it to be. The second equations in each box are gotten simply by solving the first equation for n .

See the supplement: Chapter 7 - Confidence Intervals for Variances

Section 8.1: Again, note that $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

Section 8.2: Just like with the confidence intervals, we will never have the σ , so we should always use $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ instead of $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.

Section 8.3: The note about p-values on page 344 can be important!

Section 8.5: In the boxes on page 309 notice first that $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$ and, unlike for confidence intervals, we actually have p . Second, notice that we can again use the $np \geq 5$ and $n(1-p) \geq 5$ rule discussed above instead of the $\hat{p} \pm 3\sigma_{\hat{p}}$ rule.

See the supplement: Section 8.6 - Power Curves

Section 9.1: Just like in sections 7.1 and 8.2 we will never use the large sample formulas with z and σ discussed on pages 380-385. Instead use the t formula on page 386 if the variances are equal. If the variances are not equal use the box on page 390.

Section 9.2: Again, do not use the large sample formula in the box on page 401.

Section 9.3: Note that $\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.

When we are making a confidence interval we know nothing about the values of p_1 and p_2 except what we have in the sample. In this case we just substitute in \hat{p}_1 and \hat{p}_2 . When we are making a test of the hypothesis that the two populations have equal percentages, however, it doesn't make sense to put in two different values (because we are assuming they are the same!). In this case we substitute in $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ for both p_1 and p_2 .

See the supplements:

Section 10.2 - The ANOVA Table

Section 11.3 - Checking the Regression Assumptions

Section 11.5 - The ANOVA Table for Regression

Section 13.3 - Chi-Square Test for Homogeneity