STAT 3332 Statistics for Life Sciences

Fall 2007

Comments on Solutions to Recommended Text Problems

3.70–3.77

Note: The "test" is whether the SCN level exceeds 100 μg/mL or not. Each subject is tested.

For **3.70**, this test is explored in a subpopulation, the self-reported light smokers ($\leq$ 14 cigarettes per week). For this subpopulation, the sensitivity is $P(A \mid B)$, where A = "test positive" and B = "smoke $\leq$ 14 cigarettes per week". Using Table 3.9 to find $P(A \mid B)$, you first need to write $B = B_1 \cup B_2$, where B_1 = "smoke 1-4 cigarettes per week" and B_2 = "smoke 5-14 cigarettes per week". Then write

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A \cap B_1) + P(A \cap B_2)}{P(B_1) + P(B_2)}. \end{aligned}$$

Now use

$$P(A \cap B_1) = P(A \mid B_1) \times P(B_1) = 0.043 \times \frac{70}{1332}$$

and similar steps to complete the computation of $P(A \mid B)$.

For **3.74, 3.75**, use Bayes Rule (p. 60), with $P(A \mid B)$, $P(B)$, $P(A \mid \overline{B})$, and $P(\overline{B}) = 1 - P(B)$, where A = "SCN $\geq$ 100" and B = "smokes $\geq$ 1 cigarettes per week.

Get $P(\overline{B})$ from Table 3.9:

$$P(\overline{B}) = \frac{1163}{1332}.$$

Also from that table,

$$P(A \mid \overline{B}) = 0.033.$$

Now to get $P(A \mid B)$, use

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

You have the denominator and now need the numerator. For this use the law of total probability:

$$P(A \cap B) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) + P(A \cap B_4) + P(A \cap B_5),$$

where B_1 = "smokes 1-4 per week", B_2 = "smokes 5-14 per week", etc. Now from Table 3.9

$$P(A \cap B_1) = P(B_1) \times P(A \mid B_1) = \frac{70}{1332} \times 0.043.$$

and you can get the other $P(A \cap B_i)$ similarly.

With all this, you can now complete the computation of the inputs for use of Bayes Rule to get PV$^+$ and PV$^-$.

4.82-4.84 First use Appendix Table 2 to find x_0 such that

$$P(\text{Poisson}(2) \leq x_0) \geq 0.95.$$

Next repeat this with Poisson(4) instead of Poisson(2).

Finally, use

$$\begin{aligned} P(4 \text{ admissions}) &= P(\text{normal day}) \times P(4 \mid \text{normal day}) + P(\text{high day}) \times P(4 \mid \text{high day}) \\ &= P(\text{normal day}) \times P(\text{Poisson}(2) = 4) + P(\text{high day}) \times P(\text{Poisson}(4) = 4). \end{aligned}$$

5.31-5.34 First, assume FEV is $N(4.0, (0.5)^2)$ for nonsmoking men and $N(3.5, (0.6)^2)$ for smoking men.

For **5.31**, calculate $P(\text{FEV} < 2.5)$, using the FEV distribution for smoking men and the standardization approach to reduce to a computation with $N(0, 1)$.

For **5.32**, follow the same approach with the FEV distribution for nonsmoking men.

For **5.33**, you first need to figure out the FEV distribution for a 75-year-old man whose FEV at age 45 is 4.0. For $n = 30$ years, the *decline* is random: normal with mean $0.03 \times 30 = 0.9$ and standard deviation $0.02 \times 30 = 0.6$. Then *subtract* from mean 4.0 to get $N(3.1, (0.6)^2)$ as the resulting distribution. Now proceed as in the previous problems.

For **5.34**, proceed similarly to **5.33**.

5.78-5.80 For **5.78**, calculate $P(X \geq 2)$, where X is the mean decline for placebo: $N(1.8, (4.3)^2)$. (The mean change is -1.8, or equivalently the mean decline is $+1.8$.)

For **5.79**, proceed along similar lines with appropriate alterations. I.e., calculate $P(X \geq 2)$ for X now $N(-3.5, (4.2)^2)$.

For **5.80**, follow the hint. Thus (result of **5.79**) $= 0.10 \times$ (the result of **5.78**) $+\ 0.90 \times$ (the desired result). Now solve for the "desired result".

6.1-6.2 The idea here is to make a rule, using random numbers, to assign either treatment A or B to 20 patients. For example, you can use the odd random numbers for assignment to A and even ones for assignment to B. The number assigned to A in 20 trials is binomial$(20, 0.5)$. Now proceed with the problem.

6.5-6.6 Recall that the "standard error of the mean" is just other terminology for the standard deviation of the mean, which is $\sigma/\sqrt{n}$. Now use the σ and n values in Table 6.8.

6.24-6.25

For **6.24**, you use the three columns for "different media" in Table 6.10, one column for each type of control strain. For *each* column, you have $n = 9$ observations. Calculate their mean $\overline{X}$ as your point estimate. Also, calculate the sample standard deviation s, and then compute a 95% C.I. using $\overline{X}$, s, and the t_8 distribution. You thus have 3 point estimates and 3 C.I.s, one of each for each control strain.

For **6.25**, proceed similarly but with the other three columns of data in Table 6.10.

8.180-8.183

For **8.180-8.181**, we test the hypothesis $H_0 : \mu = 0$ versus the two-sided alternative $H_1 : \mu \neq 0$, where μ is the underlying mean change in BMI among heavy smokers 6 years after quitting. Here the appropriate procedure, based on the normality assumption, is to take differences across the last two columns of Table 8.41 and then use the paired t-test (i.e., the one-sample t-test) on the differences.

For **8.182-8.183**, we test the hypothesis $H_0 : \mu_1 - \mu_2 = 0$ versus the two-sided alternative $H_1 : \mu_1 - \mu_2 \neq 0$, where μ_1 is the underlying mean change in BMI among heavy smokers 6 years after quitting and μ_2 is the underlying mean change in BMI over 6 years among never-smokers. Here the appropriate procedure, based on the normality assumption, is to carry out the two-sample t-test using the first two columns of Table 8.41. Do this for each of the cases

(a) equal variances assumed,
(b) equal variances *not* assumed.

8.180-8.183 revisited using nonparametric procedures

For **8.180-8.181**, we test the hypothesis $H_0 : \nu = 0$ versus the two-sided alternative $H_1 : \nu \neq 0$, where ν is the underlying *median* change in BMI among heavy smokers 6 years after quitting. Here an appropriate procedure not requiring the normality assumption is again to take differences across the last two columns of Table 8.41, but now, instead of the one-sample t-test on the differences, try

(a) the *sign test*,
(b) the *Wilcoxon signed rank test*.

For **8.182-8.183**, we test the hypothesis $H_0 : \mu_1 - \mu_2 = 0$ versus the two-sided alternative $H_1 : \nu_1 - \nu_2 \neq 0$, where ν_1 is the underlying median change in BMI among heavy smokers 6 years after quitting and ν_2 is the underlying median change in BMI over 6 years among never-smokers. Here an appropriate procedure not requiring the normality assumption is to carry out the *Wilcoxon rank-sum test* using the first two columns of Table 8.41.

If this H_0 is *accepted*, you can also test for *scale difference* using the *Siegel-Tukey test* (covered in class). As an exercise, carry out this test whether or not the location H_0 is accepted.

10.19

Test for equality of two binomial proportions, using the normal approximation approach since these sample sizes are large enough.

10.68-10.69

This is a 3×2 contingency table. Test for independence of family income level and HIV-status using an appropriate chi-square procedure.

10.81

We wish to test whether the distribution of serum retinol is *normal* in shape. For the Vitamin A group in Year 0 in this study, Table 10.33 gives sample mean and sample standard deviation for $n = 73$ subjects. Therefore, for the purpose of testing goodness of fit to normality using the count data in Table 10.34, the relevant distribution for computing expected counts is $N(1.89, (0.36)^2)$. Using this distribution, compute the probabilities $\pi_1 = P(N(1.89, (0.36)^2) \le 1.40)$, $\pi_2 = P(1.40 \le N(1.89, (0.36)^2) \le 1.75)$, etc. Then get expected frequencies via $E_1 = 73 \times \pi_1$, $E_2 = 73 \times \pi_2$, etc. Finally, compute the chi-square statistic

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i},$$

and use the null hypothesis distribution χ^2 with degrees of freedom $g - 1 - 2 = 5 - 3 = 2$ (based on $g = 5$ categories and $k = 2$ estimated parameters) to get a p-value.

11.1 Straightforward.

11.3 The purpose is to measure the goodness of the regression fit, as explained in §11.6, using the square of the sample correlation coefficient between the X and Y values. As per Definition 11.15, p. 492, the sample correlation is

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}},$$

with L_{xy}, L_{xx}, and L_{yy} as illustrated in Example 11.8, p. 471, for example. Calculate r and then take r^2 as your measure.

The interpretation of R^2 (equivalently r^2 in the simple linear regression model with just one dependent variable) is found on p. 478 (bottom). Here it is defined as the ratio of the "regression sum of squares" to the "total sum of squares",

$$R^2 = \frac{L_{xy}^2 / L_{xx}}{L_{yy}},$$

which is the same as r^2 as given above. (In the *multiple* regression case, where more than one dependent variable is involved, R^2 is again a ratio of regression SS to total SS, but the regression SS is defined a bit differently. Also, it can be defined as the sample correlation between the Y-values and the *fitted* values $\widehat{Y}$.)

11.6 Follow your nose.

11.7 Straightforward.

11.30 Straightforward.

11.31-11.35 Enjoy.