Extra Problem Solutions

Dr. Erin K. Melcon

Name: ______________________________

True/False

For each of the following questions indicate true or false, then **explain** your answer. You may use examples to illustrate your answer.

(I) For two events A and B (where $Pr\{B\} > 0$ and $Pr\{A\} > 0$), $Pr\{A|B\}$ must be greater than or equal to $Pr\{A\}$.

- Note: For this problem you may acutally answer TRUE or FALSE. It depends on the data. If you answered TRUE, you would have to argue that yes they can be larger than or equal (address both situations). For example, they are equal when the events are independent. Suppose B = women, A = glasses. It is absolutely possible that out of all women, the probability of wearing glasses is higher than the general probability.
  If you answer FALSE, you have to explain how it could be less, which the above explanation also works for. It is absolutely possible that out of all women, the probability of wearing glasses is lower than the general probability.

(II) If a 95% confidence interval for β_1 is: (5.6, 9.4), it is appropriate to interpret this interval as "There is a 95% chance that the true slope β_1 is in the interval 5.6 to 9.6".

- FALSE. A confidence interval means that out of many many random samples which result in many many confidence intervals, 95% of the many many confidence intervals will contain the true slope. For any particular confidence interval, the true slope is either in the interval, or not.

(III) One of the assumptions for a binomial random variable is that the probability of success remains constant (the same value) for all n trials.

- TRUE. Because the trials are independent, we then can use p^j as the probability that j trials have successes, which uses the property $Pr\{A \text{ and } B\} = Pr\{A\}Pr\{B\}$

(IV) If we wanted to minimize the probability of a Type II error, we should pick the largest value of α possible.

- TRUE. As α increases, the probability of a Type II error decreases. So we would choose the largest value of α to have the smallest value of Type II error.

(V) The larger $(o_i - e_i)$ is for a χ^2 Goodness of fit test, the larger the p-value for the two-sided test will be.

- FALSE. A larger value of $(o_i - e_i)$ means that the chi-squared test statistic will be larger, which would result in a smaller p-value.

(VI) If the sample correlation is zero, that means the estimated slope for the regression line will also be zero.

- TRUE. This can be seen through the equation $b_1 = r\frac{s_y}{s_x}$. When you plug in $r = 0$, you get b_1

Full Detail

Work out the following problems. **Show your work.**

1. A study to determine the effectiveness of a drug for arthritis was conducted, with 200 arthritic patients randomly put in a placebo group or a drug group, which has 201 patients. They were later asked if their arthritis improved:

	Improved	No Improvement
Drug Group	117	83
Placebo Group	72	129

(a) Find a 95% confidence interval for the difference in the probability of improvement for the drug vs placebo group.

- **Solution**: $\tilde{p}_1 = \frac{y_1+1}{n_1+2} = \frac{117+1}{200+2} = 0.5842$ and $\tilde{p}_2 = \frac{y_2+1}{n_2+2} = \frac{72+1}{201+2} = 0.3596$
  Then, the confidence interval is: $(\tilde{p}_1 - \tilde{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$ or
  $(0.5842 - 0.3596) \pm (1.96)\sqrt{\frac{0.5842(1-0.5842)}{200+2`} + \frac{0.3596(1-0.3596)}{201+2`}}$ or $(0.2246) \pm (1.96)0.0483$ or $(0.1299, 0.3193)$.

(b) Interpret your interval in terms of the problem.

- **Solution**: We are 95% confident that the true proportion of subjects in the drug group who show improvement is higher than those in the placebo group by between 0.1299 and 0.3193.

(c) Does your interval from (a) suggest that improvement is independent of group? Explain.

- **Solution**: No, it does not. Since the interval does not contain zero, this means there is significant evidence that there is a difference in the conditional probabilities. I.e. $Pr\{Improved|Drug\} \neq Pr\{Improved|Placebo\}$. If they were independent, we know that $Pr\{Improved|Drug\} = Pr\{Improved|Placebo\} = Pr\{Improved\}$

(d) List all the necessary assumptions for a confidence interval for a difference in two proportions.

- **Solution**: We must assume:
  i. A random sample was taken from each group.
  ii. Both groups are independent.
  iii. There are at least 5 subjects with and without the trait in each group.

2. The following data summarizes the incidence of Coronary Heart Disease (CHD) for people who smoked regularly and for people who did not smoke regularly:

	CHD	No CHD
Smoked	84	87
Did not Smoke	296	491

For the hypothesis $H_0 : Pr\{CHD|Smoke\} = Pr\{CHD|Smoke^C\}$, the p-value was: 0.00528.

(a) Find the estimated probability of having CHD if you smoke, and the estimated probability of having CHD if you do not smoke.

- **Solution**: $\hat{Pr}\{CHD|Smoke\} = \frac{84}{171} = 0.4912$, $\hat{Pr}\{CHD|Smoke^C\} = \frac{296}{787} = 0.3761$

(b) If the goal is to test if smoking status is independent of CHD, state the alternative hypothesis and find the p-value.

- **Solution**: $H_A : Pr\{CHD|Smoke\} \neq Pr\{CHD|Smoke^C\}$ vs. $H_0 : Pr\{CHD|Smoke\} = Pr\{CHD|Smoke^C\}$ The p-value was given as 0.00528.

(c) What value of α (out of the three 0.01, 0.05, and 0.10) should we pick if we want the probability of a type I error to be as small as possible? Explain your answer.

- **Solution**: Since α is the probability of a type I error, we would want the smallest α, which is 0.01. We would make it more difficult to reject the null, and thus less likely to reject the null when it was true (since we are less likely to reject the null period).

(d) Estimate the probability that someone has CHD.

- **Solution**: The estimated probability would then be: $\hat{Pr}\{CHD\} = \frac{380}{958} = 0.3966597$

3. 60 guinea pigs were measured, and the amount of vitamin C (in mg) and the length of their adult teeth (in mm) was measured. We believe that vitamin C may increase the length of their teeth. Summary statistics are:

	Vitamin C (mg)	Tooth Length (mm)
sample mean	1.17	18.81
sample std. dev	0.63	7.65

In addition, $r = 0.803$, and $SS(Resid) = 1227.905$.

(a) Find the estimated regression line.

- **Solution**: The estimated slope is: $\hat{\beta}_1 = r\frac{s_y}{s_x} = 0.803\frac{7.65}{0.63} = 9.7507$.
  The estimated intercept is: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 18.81 - (1.17)(9.7507) = 7.4017$
  Thus, the line is: $\hat{y} = 7.4017 + 9.7507X$

(b) Find the 95% confidence interval for the slope, and interpret it in terms of the problem.

- **Solution**: The 95% confidence interval is: $b_1 \pm t_{\alpha/2}\frac{s_e}{s_x\sqrt{n-1}}$ or $9.7507 \pm (2.009)\frac{4.6012}{0.63\sqrt{59}}$ or (7.8405, 11.6609).
  We are 95% confident that When the dosage of vitamin C increases by 1 mg, we expect the length of the adult tooth to increase by between 7.8405mm and 11.6609 mm on average.

(c) If a guinea pig that received 2 mg of vitamin C has a tooth length of 29.4 mm, find the error based on your regression line. Did the line over estimate or under estimate the actual value?

- **Solution**: Our prediction is : $\hat{y} = 7.4017 + 9.7507(2) = 26.9031$.
  Thus, our error is: $e_i = y_i - \hat{y}_i = 29.4 - 26.9031 = 2.4969$.
  Thus, we underestimated the actual value.

(d) List two of the assumptions for linear regression.

- **Solution**: The assumptions are:
  i. A linear relationship is appropriate.
  ii. A random sample of pairs was taken.
  iii. The errors have mean 0
  iv. The errors have constant variance.
  v. The errors are normally distributed.

(e) Find and interpret s_e.

- **Solution:** $s_e = \sqrt{\frac{SS(resid)}{n-2}} = 4.6012$. A typical error when using our regression line to predict tooth length is 4.6012 mm.

(f) Find and interpret r^2.

- **Solution:** $r^2 = (r)^2 = (0.803)^2 = 0.6448$. The reduction in error when using linear regression between dose of vitamin C and tooth length (instead of the sample mean) is 0.6448

4. A biology professor believes the proportion of males taking their class is 60%, and the proportion of females taking their class is 40%. For a class of size 489, she finds that 223 are female, and 266 are male.

   (a) State the appropriate null and alternative for the professors belief.

      - **Solution**: H_0: $Pr\{male\} = 0.60$, $Pr\{female\} = 0.40$ (using one or the other is fine)
        $H_A : Pr\{male\} \neq 0.60$, $Pr\{female\} \neq 0.40$

   (b) Calculate the test-statistic for the hypothesis in (a).

      - **Solution**: $\chi^2_S = \sum_{i=1}^{K} \frac{(o_i - e_i)^2}{e_i} = \frac{(223-195.6)^2}{195.6} + \frac{(266-293.4)^2}{293.4} = 3.838 + 2.559 = 6.397$

   (c) Calculate the p-value for the test in (a).

      - **Solution**: At d.f $= k - 1 =$ 2 -1 $= 1$, we find $Pr\{\chi^2 > 5.41\} = 0.01$, and $Pr\{\chi^2 > 6.63\} = 0.02$
        Thus, the range for our p-value is : $0.01 <$ p-value < 0.02

   (d) State your conclusion in terms of the problem, using $\alpha = 0.01$.

      - **Solution**: Since our p-value is larger than α, we fail to reject the null hypothesis and conclude we support the biology professors claim that 60% of her students are male, and 40% are female.

   (e) Were there more or less females taking the class than the professor expected? Explain.

      - **Solution**: Since the expected number was 195.6 and the observed was 223, there were more than expected.