

## Statistics 8 Practice Final Exam Solution

1. (a) Boxplots. The change in the centerline (median) across the boxplots will describe the variance between regions. (b) Two-sample  $t$ -test.  $H_0$  : Asian and African Americans have the same mean annual income.  $H_A$  : Their mean annual incomes differ. (c) One-sample  $z$ -test.  $H_0$  : The proportion of program participants who find work within three months of the program is less than or equal to 30%.  $H_A$  : The proportion is greater than 30%. (d) Matched pairs  $t$ -test.  $H_0$  : Diet does not affect the speed at which mice can complete the task.  $H_A$  : Diet does affect this speed. (e) ANOVA (one-way).  $H_0$  : The effect of diet is the same for all three breads.  $H_A$  : The effect of diet differs for different breads. Note that this is not a two-way ANOVA because of the pairing. The response is the difference in time required between diets, the variable in the ANOVA is breed. (f) Simple regression analysis, regress days missed on GPA.

2. (a) The study is an observational study. There is a treatment, but the researchers were not able to control who got what treatment. (b) The study was retrospective. The researchers looked for bacteria in patients that already had heart disease. A prospective study would have looked for bacteria and waited for heart disease to develop. (c) The treatment is the presence of bacteria. The response was whether or not the patient had heart disease. (d) No, it could be the case that people with heart disease are more likely to become infected with this type of bacteria. (e)  $\hat{p} = 72/90 = 0.8$ . A 90% confidence interval would be  $\hat{p} \pm 1.645\sqrt{\hat{p}(1-\hat{p})/90} = (0.731, 0.869)$

3. (a) Let  $\mu_D = \mu_{1990} - \mu_{1980}$  in the west. We would like to test  $H_0 : \mu_D = 0$  versus  $H_A : \mu_D > 0$ . The matched-pairs  $t$ -statistic is  $t = 0.047/(0.187/\sqrt{5}) = 0.56$  on 4 df. The p-value is greater than 0.25, and we conclude that there is no evidence of growth between 1980 and 1990. (b) The two-sample  $t$  95% confidence interval is given by  $(6.457 - 5.539) \pm 2.776\sqrt{\frac{(1.041)^2}{5} + \frac{(0.505)^2}{5}} = 0.92 \pm 1.40$ . The Confidence interval contains zero, so there is no evidence of a difference. Let  $\mu_D$  equal the difference between the average growth between 1980 and 1990 of cities in the west and that in the east. (c) We will test  $H_0 : \mu_D = 0$  versus  $\mu_D > 0$  with the two-sample  $t$ -test:  $t = (0.211 - 0.047)/\sqrt{\frac{(0.187)^2}{5} + \frac{(0.190)^2}{5}} = 1.37$  on 4 df. The p-value is between 0.15 and 0.10, so we accept the null hypothesis, there is no evidence of a difference in growth rates. (d) We were unable to find compelling evidence of growth in western cities in the eighties, a difference in the growth rate between eastern and western cities, or a difference between the size of cities in the west and in the east in 1990. A larger sample of cities might help us to see any differences that might exist.

4. (a) This is a survey. We are sampling from a population with the goal of learning about several population parameters. (b)  $X$  follows a binomial distribution with mean equal  $np = 100 \times .4 = 40$  and variance equal to  $np(1-p) = 100 \times .4 \times .6 = 24$ .  $\Pr(30 < X \leq 40)$  can be approximated using the normal distribution,  $\Pr(30 < X \leq 40) \approx \Pr[(30 - 40)/\sqrt{24} < Z \leq (40 - 40)/\sqrt{24}] = \Pr(-2.04 < Z \leq 0) = 47.93\%$ . (c)  $\hat{p} = 36/47 = 0.766$  and  $sd(\hat{p}) \approx \sqrt{\hat{p}(1-\hat{p})/n} = 0.062$ , Thus, a 97% confidence interval for  $p$  is given by  $0.766 \pm 2.17 \times 0.062 = 0.766 \pm 0.135 = (0.631, 0.901)$  (d) Let  $p$  be the proportion of Harvard undergraduate men who attended public school that wear boxers.  $H_0 : p = 0.5$  vs  $H_A : p \neq 0.5$ ,  $z = (0.528 - 0.50)/\sqrt{0.5^2/53} = 0.407$ . Thus, the two-sided p-value is 68%. We can not reject the null hypothesis that Harvard undergraduate men who attended public schools are equally likely to wear boxers or briefs.

5. (a)  $H_0$  : There is no difference between the mean coagulation time in the different diet groups.  $H_A$  : There is a difference. The  $F$ -test statistic is 13.57 with a p-value less than 0.0001. We can reject the null hypothesis, and conclude that there is evidence that diet affects mean coagulation time. (b) We can estimate the difference with the difference in means,  $\bar{x}_1 - \bar{x}_2 = -5$ . (c) The standard error of the difference in means is given by  $s_p \sqrt{\frac{1}{4} + \frac{1}{6}} = 1.61$  with 8 df. Thus the 95% CI is given by  $-5 \pm 2.306 * 1.61 = (-8.71, -1.29)$ . We know that 95% of such confidence intervals contain the actual difference, so we have reasonable evidence that diet 1 results in quicker blood coagulation than does diet 2.

If you did not catch the mistake in the problem, your answer would be:

6. (a) Plugging 11 into the regression equation, we get  $2898.467 + 11 * 170.4424 = 4773$ . (b)  $R^2 = 0.9289$ . (c) The conditional standard deviation is the Root MSE = 151.47. (d) Using the normal distribution,  $\Pr(Y > 4800) = \Pr[Z > (4800 - 4773)/151.47] = \Pr(Z > 0.18) = 42.68\%$ . (e) According to the regression line, minority enrollments increased by about 170 students per year. According to the  $t$ -test, we can reject the null hypothesis that enrollments are constant in favor of the alternative that they are increasing. (f) From looking at the residual plot (#2), there is a clear lack of linearity, so the above calculations can not be trusted. Plot (#1) indicates that the relationship looks rather linear, but the slope changes between 1987 and 1988.