LAST NAME (Please Print): KEY

FIRST NAME (Please Print): _____

HONOR PLEDGE (Please Sign): _____

Statistics 111

Midterm 3

- This is a closed book exam.
- You may use your calculator and a single page of notes.
- The room is crowded. Please be careful to look only at your own exam. Try to sit one seat apart; the proctors may ask you to randomize your seating a bit.
- Report all numerical answers to at least two correct decimal places or (when appropriate) write them as a fraction.
- All question parts count for 1 point.

1. The Department Chair says that she wants the final to have about 20% A's, 30% B's, 30% C's, 10% D's and 10% F's. In his class of 80, Balderdash gives 22 A's, 28 B's, 25 C's, 5 D's and 0 F's. Did he follow the guidelines?

   What is the Department Chair's null hypothesis? (In words.)

   Balderdash follows the guidelines.

12.08 What is the value of the Chair's test statistic?

   This is a goodness-of-fit test.
   $ts = (22-16)^2/16 + (28-24)^2/24 + \ldots + (0-8)^2/8 = 12.08.$

χ^2_4 What table does the Chair use? (Include df if appropriate.)

   There are five categories, but you lose one degree of freedom.

$0.01 < P < 0.02$ What P-value does the Chair get? (Give a range if needed.)

   What conclusion does she reach? (In words. Use $\alpha = 0.05$)

   Balderdash does not follow her rules.

2. The chair wants to decide whether two instructors have the same grading curves. She observes:

	A	B	C	D	F
Balderdash	20	30	30	10	10
Crundel	50	80	50	10	10

   What is the chair's alternative hypothesis? (In words.)

   The curves are different; or, grades and teacher are dependent.

_____ What is the value of her test statistic?

The table of expected values is

	A	B	C	D	F
Balderdash	23.33	36.67	326.67	6.67	6.67
Crundel	46.67	73.33	53.33	13.33	13.33

Then calculate $\sum_{ij} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ to get 8.16.

χ^2_4 What table does she use? (Include df if appropriate.)

$0.05 < P < 0.1$ What P-value does she get? (Give a range if needed.)

What conclusion does she reach? (In words; use $\alpha = 0.05$)

Curves appear the same; insufficient evidence to conclude there is dependence.

3. You are doing a linear regression to predict average body temperature from the average daily amount of exercise one does. For a sample of 20 random people, you record their hours per day of exercise for a week, and their body temperatures at noon. You find that the average number of hours of exercise is 0.5 and the sample variance is 2. Your estimate of the intercept as 99.2 and the slope as -.08. And 30% of the variation in Y is explained by X. The standard deviation of the residuals is 0.6 and the standard error of $\hat{\beta}_1$ is 0.04.

99.04 Estimate temperature of someone who exercises for 2 hours per day.

$99.2 + (-0.08)2 = 99.04$.

99.46 What is upper bound of a two-sided 95% confidence interval on the average temperature of people who exercise 2 hours/day?

First, find SS_x. Since the sample variance of the X-values is $\frac{1}{n-1}SS_x$, then $SS_x = 38$. Then use the formula

$$L, U = \hat{\beta}_0 + \hat{\beta}_1 x \pm \text{rmse} * \sqrt{\frac{1}{n} + \frac{(x-\bar{X})^2}{SS_x}} t_{n-2,\alpha/2}$$

where $t_{18,0.025} = 2.101$, $x = 2$, and the rmse is 0.6.

100.37 What is the upper bound of a two-sided 95% confidence interval on the temperature of Ann, who exercises for 2 hours/day?

Just as before, except now the formula is

$$L, U = \hat{\beta}_0 + \hat{\beta}_1 x \pm \text{rmse} * \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{SS_x}} t_{n-2,\alpha/2}$$

-0.55 What is the value of the correlation coefficient?

It is the square root of 0.3, the percent of variation explained. It is the negative root since the slope is negative.

$0.05 < P < 0.1$ What is your significance probability for testing whether exercise is related to body temperature? Give a range if necessary.

The ts is $(pe - null)/se = -2$. Looking that up on the t_{18} table shows the probabilities lie between 0.05 and 0.1.

4. You want to place a confidence interval (CI) on the correlation between grades and income. For 100 people, you find that the sample correlation is 0.4 You resample 20 times from that sample (with replacement) and find the following correlations:

0.2, 0.3, 0.4, 0.5 0.25, 0.35, 0.45, 0.55, 0.23, 0.33, 0.43, 0.53, 0.21, 0.32, 0.42, 0.52, 0.24, 0.34, 0.44, 0.54

0.25 or 0.26 Set a 95% one-sided lower CI with the pivot bootstrap.

The formula is $2\hat{\theta} - \hat{\theta}^*_{0.95}$, where $\hat{\theta} = 0.4$ and $\hat{\theta}^*_{0.95}$ is the 95% percentile, or the number halfway between the two largest values.

0.37 What is the probability that Abelard's GPA and income were not included in the first bootstrap sample?

$(1 - \frac{1}{100})^{100}$. This is the probability that he does not get picked on the first draw, and the second, and so forth.

5. Freshmen at UNC work 12.2 hours a week for pay, on average, and the SD is 10.5 hours; at Duke, the average is 10.2 hours and the SD is 9.9 hours. You want to show that Duke students work less. Assume that these data are based on two independent simple random samples, each of size 100. (Subtract Duke from UNC.)

   What is the alternative hypothesis (in symbols)?

   $H_A : \mu_{NC} - \mu_D > 0$

   1.39 What is the value of your test statistic?

   $ts = \frac{12.2-10.2-0}{\sqrt{\frac{10.5^2}{100} + \frac{9.9^2}{100}}} = 1.386.$

   0.08 What is the P-value of the test statistic? (Give a range if needed.)

   Use the normal table.

   What is your conclusion? Consider P-values less than 0.05 as small.

   There is insufficient evidence to reject—no reason to think that the average Duke student works less.

6. A carnival barker guesses people's weight (WT) from estimates of their height (H) and width (W). Fill in the equation:

   $\ln WT = \beta_0 + \beta_1 \ln \text{ H } + \beta_2 \text{ ln W}$

7. Suppose you want to test whether a new diet pill leads to a five pound weight loss within one month. In fact, the average weight loss is six pounds, and the standard deviation in weight loss is eight pounds.

548 For a power of 0.9 with 0.05 Type I error rate, what sample size do you need?

$$\begin{aligned} 0.9 &= \mathbb{P}[ts > 1.645] \\ &= \mathbb{P}[\frac{\bar{X} - 5}{8/\sqrt{n}} > 1.645] \\ &= \mathbb{P}[\frac{\bar{X} - 6 + 6 - 5}{8/\sqrt{n}} > 1.645] \\ &= \mathbb{P}[Z > 1.645 - \frac{1}{8/\sqrt{n}}] \end{aligned}$$

so, from the normal table, $1.645 - \frac{1}{8/\sqrt{n}} = -1.28$. Solving for n gives 547.56, which must be rounded up to 548.

0.65 or 0.66 For $\alpha = 0.05$ and sample size of 100, what is your Type II error?

$$\begin{aligned} TypeII &= \mathbb{P}[ts < 1.645] \\ &= \mathbb{P}[\frac{\bar{X} - 5}{8/\sqrt{100}} < 1.645] \\ &= \mathbb{P}[\frac{\bar{X} - 6 + 6 - 5}{8/\sqrt{100}} < 1.645] \\ &= \mathbb{P}[Z < 1.645 - \frac{1}{8/\sqrt{100}}] \\ &= \mathbb{P}[Z < 0.395] \end{aligned}$$

so, from the normal table, this is 0.65 or 0.66.

8. The lifespan of light bulbs, in months, is exponential with parameter λ. Brand A has $\lambda = 1$, Brand B has $\lambda = 1/2$ and Brand C has $\lambda = 1/3$. Brand A has 30% of the market, Brand B has 20%, and Brand C has the rest.

0.81 Your light bulb burns out at 1 month. What is your belief about the probability that it is not Brand A?

Use Bayes. Find

$\pi(\text{Brand X}|\text{ one month }] =$

$$\mathbb{P}[\text{ 1 month } | \text{ Brand X }]\mathbb{P}[\text{ Brand X }] / \left\{ \sum_{A,B,C} \mathbb{P}[\text{ 1 month } | \text{ Brand X }]\mathbb{P}[\text{ Brand X }] \right\}$$

so the posterior probability of B is 0.206 and of C is 0.607.

2.42 How long do you think another bulb from the same package will last?
Brand A has mean lifetime 1, Brand B has mean lifetime 2, and Brand C has mean lifetime 3. So 0.187 * 1 + 0.206 * 2 + 0.607 * 3 = 2.42 months.

9. List all, and only, the true statements. C, D, G, H, J

   A. As points cluster more tightly around a line, the correlation increases.
   B. The percentile bootstrap is better than the pivot.
   C. Galton invented the idea of eugenics.
   D. In high dimensions, data tends to be too sparse.
   E. Making predictions outside the range of the explanatory variables in your training data is reasonable.
   F. If you make many tests, your true α decreases.
   G. With large samples, you can get results that are statistically significant but not scientifically significant.
   H. You should not plot your data before deciding upon your alternative hypothesis.
   I. The P-value is the probability of observing data that are as more more supportive of the null than the data observed, when the null is true.
   J. All models are wrong, but some models are useful.

10. What is multicollinearity? Here and below, be brief and clear.

Multicollinearity means that two or more of the explanatory variables are strongly correlated. Equivalently, they lie on an affine subspace of the space of explanatory variables.

11. Why is multicollinearity a problem?

It means there is large uncertainty when predicting Y for future values of $x_1, x_2, \ldots x_p$ that do not have the same strong correlation, or, equivalently, do not lie on the same affine subspace.

12. How do you calculate a moving average estimate for Y at a point x in nonparametric regression?

To find the moving average prediction of Y at x, one finds the shortest window centered at x that contains a prespecified number of x-values in the training sample, and then averages the y-values that correspond to the x-values in the window.

13. How does two-fold cross-validation work in nonparametric regression?

One divides the sample at random into two groups. Fit the nonparametric regression to the first group and use it to predict the other. Sum the squares of the error for each prediction. Then use the second group to fit the nonparametric regression, and use it to predict the first. Again, sum the squared errors. Finally, the average of all squared errors is your estimate of the squared error in using the nonparametric regression fitted from all the data.