

STATISTICS 151 SECTION 1 MIDTERM 1

FEBRUARY 17 2009

Your name (print!):

Your PID:

Five-digit code number (**OPTIONAL**):

(This should be personal to you. Do not use your PID and do not disclose it to any other person, but do keep a record of it. This is for display of results on the course webpage. If you would prefer that your results not be displayed, leave blank.)

Honor pledge: On my honor, I have neither given nor received unauthorized aid in this exam.

Sign here:

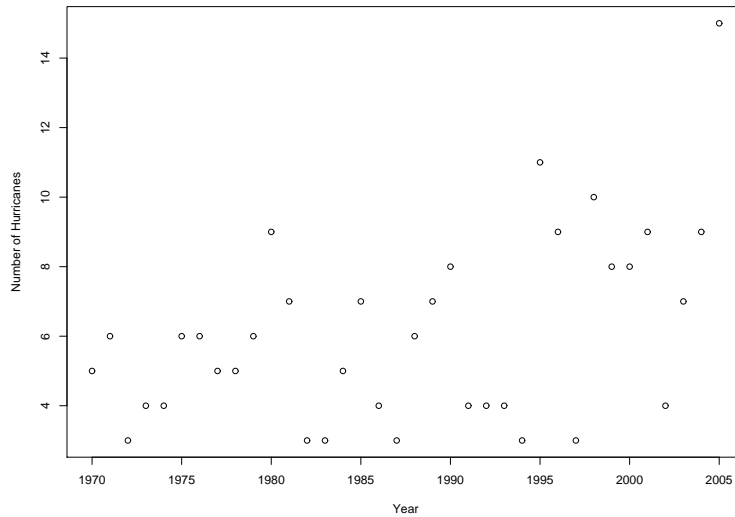
This is an open book exam. Course text, personal notes and calculator are permitted. You have 75 minutes to complete the test. Personal computers are not allowed. If you have any queries about the meaning of the questions, ask the instructor for assistance.

Answers are to be written in a blue book except for the plot in question 1. **HAND THIS SHEET IN** as it will be treated as part of your answer.

SHOW ALL WORKING — even correct answers will not get full credit if it's not clear how they were obtained. Incorrect answers will gain substantial credit if the method of working is substantially correct.

Answer all questions. The score for each part is indicated at the end of the question (100 points total).

1. The total numbers of Atlantic hurricanes, per year, for the period 1970–2005 (in year order) are as follows:
5, 6, 3, 4, 4, 6, 6, 5, 5, 6, 9, 7, 3, 3, 5, 7, 4, 3, 6, 7, 8, 4, 4, 4, 3, 11, 9, 3, 10, 8, 8, 9, 4, 7, 9, 15.
 - (a) Draw a stem and leaf plot of the data. Be sure to choose a scale for the stem that reasonably shows the overall shape of the distribution. How would you describe that shape? [**16 points.**]
 - (b) Calculate the five-number summary and the IQR. [**10 points.**]
 - (c) Are there any outliers in the data? State clearly what criteria you use to determine this. [**6 points.**]
 - (d) If x represents the year and y represents the number of hurricanes, then the following calculations may be assumed: $\bar{x} = 1987.5$, $\bar{y} = 6.11111$, $s_x = 10.53565$, $s_y = 2.72321$, $r = 0.46307$. Calculate the constants a and b for the linear regression formula that could be used to predict y from x . [**10 points.**]
 - (e) Draw the regression line on the plot on the other side of this sheet. [**8 points.**]
 - (f) How many hurricanes do you think there will be in the year 2030? Do you think this is a realistic answer (say why or why not)? [**5 points.**]
 - (g) On the basis of this analysis, do you think there has been an increasing trend in the number of Atlantic hurricanes per year (say why or why not)? [**5 points.**]



2. In an experiment to study whether the style of music played in a restaurant influences the amount of money people spend in the restaurant, the proprietor decides to test the following options: classical music (C), jazz (J) or no music (N). Each night (Monday through Saturday) for two weeks, he will randomly choose one of these three options, so that each option is chosen exactly four times during the experiment. At the end of the two weeks, he will calculate the mean amount of money spent per customer under each of the three options.
- Why might it be important to randomize in this experiment (instead of, for example, just repeating the three options in a fixed order, say CJNCJN...)? Give at least two reasons. [9 points.]
 - Based on the random numbers given below (starting at the top left), determine an exact scheme for doing the randomization. Remember that you need to randomize in a way that ensure that each of the three music options will be chosen exactly four times during the two-week trial. Your answer should state exactly which option will be used on which day. Be sure to describe your method in sufficient clarity that I can see exactly how you got your answer. [16 points.]
 - Suppose, in addition to studying the effect of music, the proprietor decides to vary the color of the paper on which the menu is printed (either blue or white). What is the name given to this type of experiment? [6 points.]
 - Without doing the actual calculations*, describe how you would vary the randomization scheme described in (b) to handle this more elaborate experiment. [9 points.]

91880519094707169160478560853363927804301508165985133873725298800418115855123396
42143113420621333766057657390734184308690609270794059183102492093728864224520878
24098059977038659329334551475784953923859392498491974943440935648061506064985002
84561652562318887249916001179344686272272875095508130109889567578068853556275163
90528098182916376784282571045056022950352288386944275455140426140120965386163430
65063619049141617074999602964016091081880796050781550476792046549512122623394659
05777702382682464721273211204580618452262279534934455969988457401031528232051058
78012109093257913217332364895473903069946485484513827664470179191705165232350452
36364566011438342462768543783500661565064175016656476419213681457513911035013519
07710572076883679285119111908673063632303734186597766765139226566642443558545385
79719959756034549273847519036762020126148702237108671581306950557615807967027729
58508579151018886602622118017748067048364506224201748033600087073910357751305036
29799758242912214013717310276071447193454427299885527913770261187354079856480197
30663206725589633628684139737742905618414978326484173794596908626930871773651585
34025299939472172429243214073902211965575839095526165211842824018258951430204195
61705361667173126505093693227033576546551339955321652388615375709956747843575812

SOLUTIONS

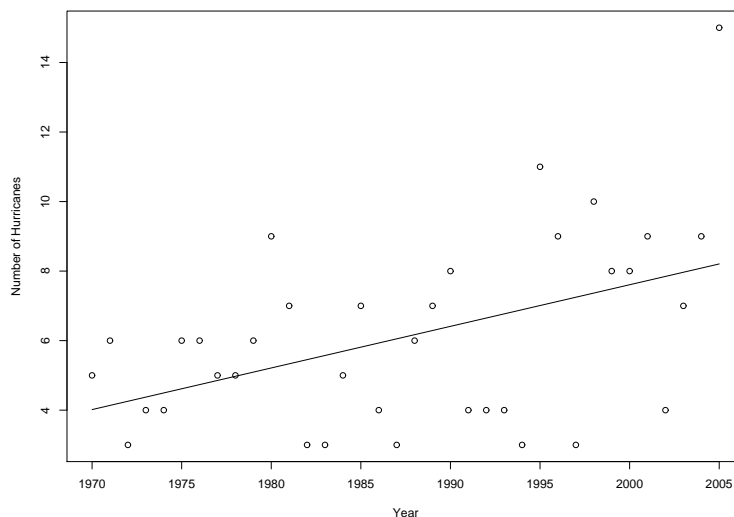
1. (a) Various ways to do it but best to split each set of 10 counts into 5 groups of 2, leading to the following stem and leaf plot:

```

0 |
0 | 333333
0 | 44444445555
0 | 666667777
0 | 8889999
1 | 01
1 |
1 | 5
    
```

the shape is largely symmetrical except for the one apparent outlier (15).

- (b) 3, 4, 6, 8, 15; IQR is 4.
 (c) Based on the 1.5IQR rule the limits are $4 - 6 = -2$ and $8 + 6 = 14$. Therefore, 15 is an outlier.
 (d) $b = \frac{rs_y}{s_x} = \frac{0.46307 \times 2.72321}{10.53565} = 0.11969$ and $a = \bar{y} - b\bar{x} = 6.11111 - 0.11969 \times 1987.5 = -231.773$.
 (e) Use the formula in (d) to calculate $\hat{y} = a + bx$ for $x = 1970, 2005$: the results are $\hat{y} = 4.02$ and 8.21. Now draw the line on the graph to connect the points (1970, 4.02) and (2005, 8.21).



- (f) Use the formula in (d) to calculate $\hat{y} = a + bx$ for $x = 2030$: answer is $\hat{y} = 11.2$. This represents a substantial extrapolation beyond the data and therefore should probably not be trusted.
 (g) The slope of the regression is positive which indicates a positive trend, but this seems strongly influenced by the anomalous 2005 value. On the other hand, you could delete this point and there still seems (mild) evidence of an upward trend.

2. (a) (i) There probably is a day of week effect (e.g. Saturday is usually the biggest day in a restaurant) and if we followed a fixed order, the same music style would be playing on two Saturdays. (ii) If there was advance information about which music style would play on which day, that could influence customers and restaurant staff in other ways.
- (b) Here there really isn't a fixed answer but it's important to state clearly how you would do it. *For example:* assign the digits 1,2,3 to C, 4,5,6 to J, 7,8,9 to N and leave out all 0s. This translates to NCNNJCNNJNNJCNCJNNJJNJCCJ... but we don't want more than four of any one so the final order will be NCNNJCNJCJCJ. In other words, the first week will be N (Monday), C (Tuesday), N (Wednesday) through to C (Saturday); the second week will be NJCJCJ in that order.
- (c) It's an example of a multi-factor experiment.
- (d) The simplest way is to assign 1 digit to each combination, e.g. 1=classical, blue menu, 2=classical, white menu etc. through to 6=no music, white menu. Then go through the random number table deleting 0,7,8,9 until you have two of each combination. Other solutions will be accepted if they lead to a correct randomization.