

**STATISTICS 151 SECTION 2 MIDTERM 1  
OCTOBER 3 2006**

Your name (print!):

Identification number:

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.

All answers are to be written on this sheet and you are to hand the sheet in on completion. **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. If the space provided for working is insufficient, you may also use the back of the sheet.

Answer all three questions.

1. The following data represent the heights (in inches) of nine randomly selected men and twenty randomly selected women from the course.

MEN: 67, 74, 71, 66, 72, 74, 71, 67, 68

WOMEN: 65, 59, 65, 65, 67, 64, 65, 64, 54, 64, 63, 65, 62, 65, 71, 68, 65, 66, 70, 67

- (a) Calculate the mean and standard deviation of the men's heights (show all working — you are not asked to do this for the women's heights as well).

- (b) For both men and women, calculate the five-number summary.

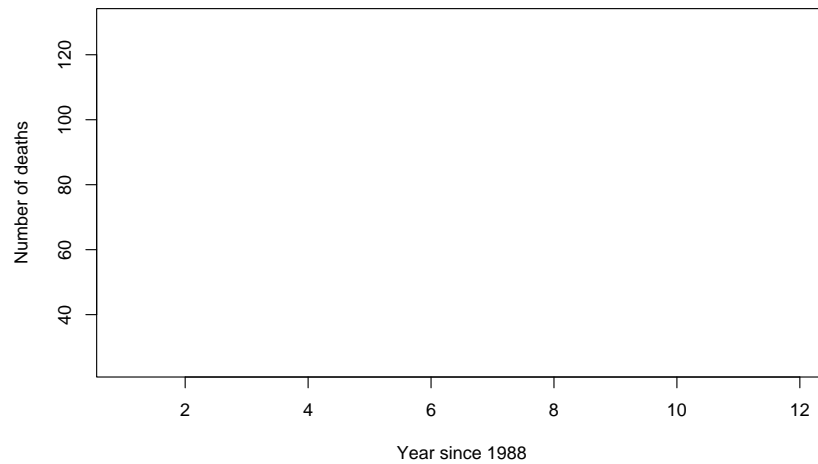
(c) Draw side-by-side boxplots for the men's and women's heights.

(d) Describe in words the shape of the two distributions.

2. The following table shows the number of deaths from tornadoes in the US for each year from 1989 (labelled year 1) through 2000 (year 12)

| Year ( $x$ ) | Deaths ( $y$ ) |
|--------------|----------------|
| 1            | 48             |
| 2            | 53             |
| 3            | 39             |
| 4            | 39             |
| 5            | 33             |
| 6            | 68             |
| 7            | 30             |
| 8            | 25             |
| 9            | 67             |
| 10           | 130            |
| 11           | 93             |
| 12           | 40             |

- (a) Plot the data as a scatterplot, using the following frame:



- (b) You are given:  $\bar{x} = 6.5$ ,  $s_x = 3.606$ ,  $\bar{y} = 55.417$ ,  $s_y = 30.420$ ,  $r = 0.409$ . Find the constants  $a$  and  $b$  in the regression line  $\hat{y} = a + bx$ .

- (c) Compute  $\hat{y}$  for  $x = 2$  and for  $x = 12$  and plot these points on the graph. Hence draw in the regression line.

- (d) Do you believe that this proves that deaths from tornadoes have been increasing? Explain why or why not.

3. At a late stage in the distribution of basketball season tickets, the president of the Carolina Athletic Association finds that she has an additional ten tickets to hand out. Because they have been underrepresented in previous lottery allocations, she decides to distribute these tickets by random lottery among residents of four campus dorms: Ehringhaus South (263 students), Horton (276 students), McIver (105 students) and Old East (67 students). Within these four dorms there are a total of 711 students — suppose each student is identified by a unique three-digit number from 001 through to 711.
- (a) The CAA president decides to select the sample using simple random sampling from all 711 students. Describe a scheme to do this using a random number table. Then, starting on row 4 of the table of random numbers in the book (Appendix A, Table E, page A6), carry out your procedure and say which ten students will be the ones who get the tickets.
- (b) Describe briefly how you would modify this procedure if the CAA decided to use instead (i) a stratified sample, or (ii) a cluster sample. There is no need to work out a specific sample for this part of the question.

**SKETCH SOLUTIONS {Max. Scores}**

1. (a) See the following table:

| $x$ | $x - \bar{x}$ | $(x - \bar{x})^2$ |
|-----|---------------|-------------------|
| 67  | -3            | 9                 |
| 74  | 4             | 16                |
| 71  | 1             | 1                 |
| 66  | -4            | 16                |
| 72  | 2             | 4                 |
| 74  | 4             | 16                |
| 71  | 1             | 1                 |
| 67  | -3            | 9                 |
| 68  | -2            | 4                 |
| 630 | 0             | 76                |

So  $\bar{x} = \frac{630}{9} = 70$  and  $s^2 = \frac{76}{8} = 9.5 = 3.082$ . **(6 points)**

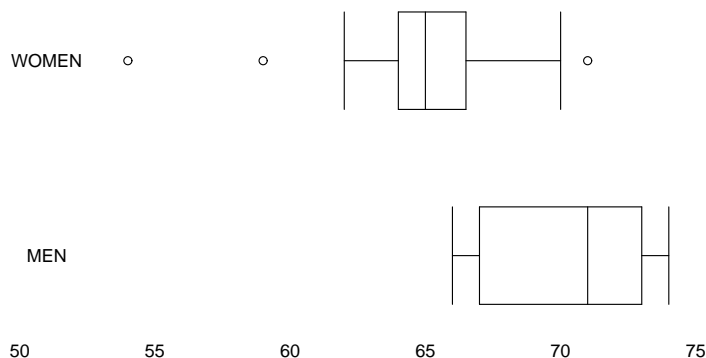
- (b) Arranged in order, the men's and women's heights are:

MEN: 66, 67, 67, 68, 71, 71, 72, 74, 74

WOMEN: 54, 59, 62, 63, 64, 64, 64, 65, 65, 65, 65, 65, 65, 66, 67, 67, 68, 70, 71

For men, 5NS is (66, 67, 71, 73, 74) (note that by convention, the median is not counted when calculating the quartiles, so that Q3 for instance is the median of (71, 72, 74, 74)). For women, 5NS is (54, 64, 65, 66.5, 71). The IQR is 6 for men and 2.5 for women. **(7 points)**

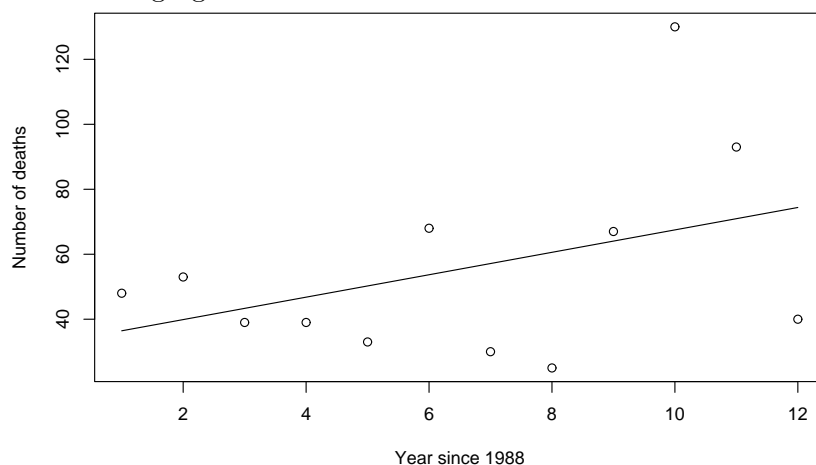
- (c) The outlier criteria are (for men)  $73 + 1.5 \times 6 = 82$  and  $67 - 1.5 \times 6 = 58$  and (for women)  $66.5 + 1.5 \times 2.5 = 70.25$  and  $64 - 1.5 \times 2.5 = 60.25$  so there are no outliers among men but 54, 59 and 71 are outliers among women. The box plots are as follows:



Many students forgot to look for outliers and consequently did not get full points. **(4 points)**

- (d) In general, the men are taller than the women, e.g. the median, Q1 and Q3 are all higher for men than for women (in fact Q1 for men is higher than Q3 for women, at least in this sample). Looking at the spread of the two distributions, the women have a smaller IQR than the men, indicating that the main part of the distribution of women's heights has less variability than for men's heights; however, there are also more outliers among the women. **(3 points)**

2. (a) See the following figure:



**(6 points)**

- (b)  $b = \frac{.409 \times 30.420}{3.606} = 3.4503$  and  $a = 55.417 - 3.4503 \times 6.5 = 32.99$ . **(6 points)**
- (c) For  $x = 2$ ,  $\hat{y} = 32.99 + 3.4503 \times 2 = 39.89$ . For  $x = 12$ ,  $\hat{y} = 32.99 + 3.4503 \times 12 = 74.39$ . The straight line is shown on the above plot. **(5 points)**
- (d) The regression does show an increasing trend over time. However, it looks very much as though the large values for years 10 and 11 are outliers, and the remaining points do not show a trend. (Another point you could make is that there is just a lot of variability, so that although there is an increasing straight line the data points do not follow this line too closely). **(3 points)**
3. (a) Number all the students 1–263 (ES), 264–539 (H), 540–644 (McI), 645–711 (OE). These are three-digit numbers, so we select numbers from the random number table in the range 001–711, ignoring anything outside that range. The result is 421, 679, 309, 306, 243, 616, 800 (drop), 785 (drop), 616 (drop because repetition), 376, 394, 405, 353. So the selected students are numbers:  
421, 679, 309, 306, 243, 616, 376, 394, 405, 353. **(12 points)**
- (b) For a stratified sample, we would decide first how many tickets to allocate to each residence, and then perform a SRS within that residence. For example, based on the relative populations of the residences, a possible allocation would be 4 for Ehringhaus South, 4 for Horton and 1 each for McIver and Old East. (Other possible ideas for a stratified sample: Instead of basing it on residence, could define strata based on class (FR,SO,JR,SR) or gender (M/F). Other classifications would be acceptable if clearly defined.) **(4 points)**

A cluster sample would first select one of the four dorms, and then randomly choose four students from that dorm. *Alternatively*, define 71 “clusters” of size 10 (e.g. allocate students by numerical order, or by dorm room or floor) then randomly select one cluster out of the 71 and give the tickets to all ten members of that cluster. Again, there are different ways to do it, but some students proposed a scheme that was really just another way to do stratified sampling, and that was not correct. **(4 points)**