STATISTICS 220 **WINTER 2008** **MIDTERM EXAM**

Friday February 8, 2008; 8.30 a.m. - 9.20 a.m.

NAME _____ SECTION: (circle 1); AA AB AC AD AE AF

1.This is a closed-book, closed-notes exam, except that each student
may refer to one (two-sided) $8\frac{1}{2}$ by 11 page of notes (s)he has brought to the exam.
2.Use of a calculator is allowed.
However, in order to receive full credit, all computations must be shown.
3.Students must clearly explain each answer to receive full credit.
4. **A Normal table is on the last page of this exam**.
5.Students must follow a reasonable code of conduct.

Cheating or other dishonest practices will result in an examination grade of zero. Such practices include, but are not limited to (i) making use of books, papers, or memoranda other than those authorized, (ii) speaking or communicating with other students during the examination, (iii) purposely exposing written papers to the view of other students.

6.If you do not sign the Honor statement below, then your exam score will be zero.

A good strategy: Skim the entire exam.
Then work first the problems with which you feel most comfortable.
Answer as many of the questions as you have time for.
This is a long exam for 50 minutes; do not panic if you cannot complete it all.

Honor statement:
I have followed a reasonable code of conduct in conjunction with this exam, as outlined in point 5 above.

_____
Student signature

Please do not write in this space.

Question number	points	possible points
1		10
2		10
3		12
4		12
Total		44

1.(a) (4 points) The following are real data on the deaths from lung cancer for women in five age groups in a particular three-year period in Ohio State. Age-group 1 are the youngest; age-group 5 are the oldest. The total sizes of the population groups are given, and the death rates are deaths from lung cancer per 100,000 women. The final column gives the percentage of non-white women in each age group.

age group	white women		non-white women		% non-white women in population
	population	death rate	population	death rate	
1	9,775,375	1	1,351,616	2	12.1
2	1,528,328	35	180,917	54	10.6
3	1,555,481	82	148,513	106	8.7
4	1,136,655	118	98,228	119	8.0
5	878,424	93	61,589	106	6.6
overall		27		25	

We see that in every age group, the lung-cancer death rate is higher for non-white women than for white women in this population, but that overall (taking all five age groups together) the rate is higher for white women than for non-white women.

Explain briefly how this can happen.

EITHER
This is an example of Simpson's paradox. *(2 points)* **An overall association in one direction (white women having having higher risk), can be reversed (non-white women having higher risk) when we stratify by age.** *(2 points)*

OR
Age is a confounding factor, affecting both breast cancer death risk and the overall proportion of non-white/white women in the population. *(2 points)* **There is lower risk of death from breast cancer at the younger age groups, but the percentage of non-white women is higher in the lower age groups.** *(2 points,)*
So even though non-white women have higher risk at each age group, they tend to be in the low-risk age groups.

Part (b) of this questions is on the next page.

1.(b) In a given quarter, there are exactly 40,000 undergraduate students registered at UW. A simple random sample of 2,000 students are mailed a survey of student study habits. Responses are anonymous. A total of 800 completed survey responses were received by UW.

(i) (3 points) There are about the same numbers of juniors as freshmen registered, both just about 10,000. However, among the 800 responses there were 250 from freshmen and 150 from juniors. Explain how this might happen.

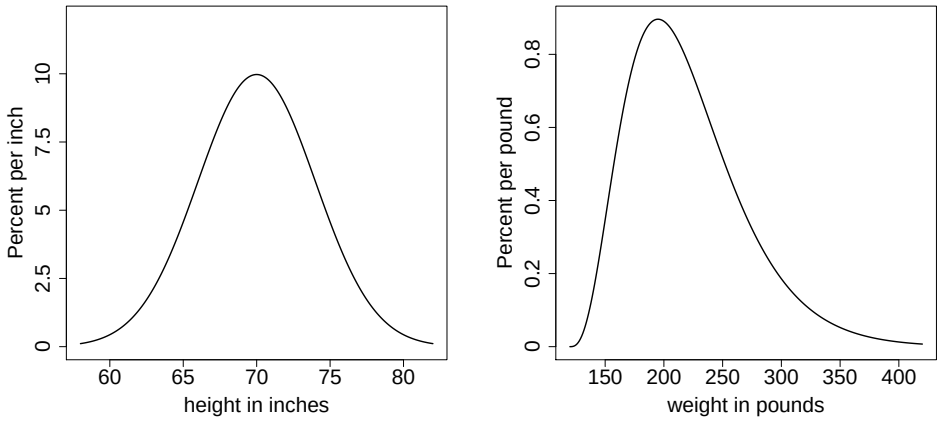
This is non-response bias. The Freshmen are being more reponsive to this UW survey. The Juniors are less likely to respond. *(3 points)*
The Juniors are probably fed up with all the surveys, and have learned to just ignore them.

There could be some randomness in the sampling, but this could not explain such a big difference. Comments only about the sampling: 1 point only.

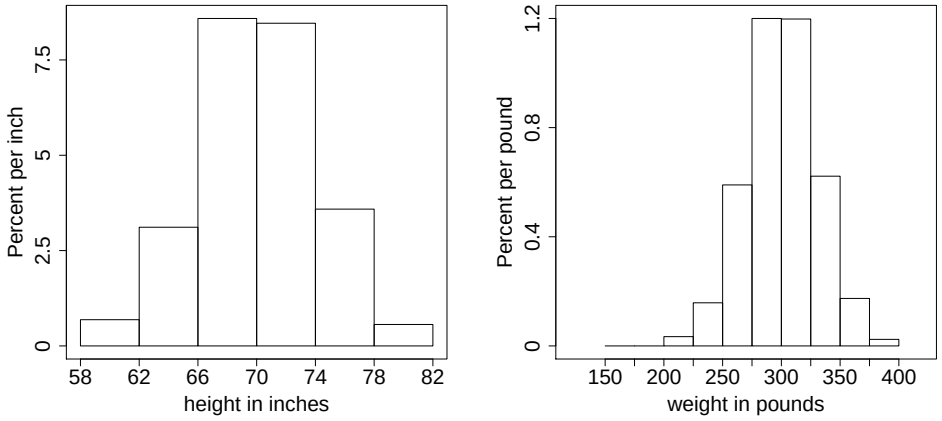
(ii) (3 points) By mistake, two slightly different versions of the survey were mailed out. In the first version, students were asked whether they considered they studied more than their peers, less than their peers, or about the same. Student responses were just about equal (about 33% each) for the three options. In the second version of the survey, the option "about the same" got omitted. For these surveys, many more students replied "more" (65%) than replied "less" (35%). Explain how this might happen.

This is response bias. How the question is answered depends how it is asked. Many are willing to say they study "about the same", but if that is not an option, almost all go for "more" rather than "less". *(3 points)*

2. The following two histograms show the distributions of height (on the left) and weight (right) of adult men in a large city population.



The following two histograms show the heights (left) and weights (right) of men who attended a particular hospital clinic in the city, over a 5-year period.



Answer the questions on the following page, by circling the answer you choose, or, in (c), writing in your answer.

NO explanations are required.

2 ctd. (1 point for each part) Answer the following questions, by circling the answer you choose, or (for (c)) writing in your answer.

NO explanations are required.

(a) Relative to the population, the sample of men who attended the clinic are mostly

much smaller in height, **quite typical in height**, much taller.

(b) Relative to the population, the sample of men who attended the clinic mostly

weigh much less, are quite typical in weight, **are much heavier**.

(c) The average height of men in the population is about **70** inches.

(d) The standard deviation (SD) of height of men in population is

2 inches, **4 inches**, 6 inches.

(e) The median weight of men in the population is about

180 lb **210 lb** 250 lb

(f) The average weight of men in the population is

larger than the median, equal to the median, smaller than the median.

(g) The percentage of men in the population who are over 74 inches tall is about

2.5%, **16%**, 45% 68%

(h) The percentage of men in the sample who are 62 inches to 66 inches in height is about

3%, 9%, **12%**, 25%

(i) The percentage of men in the population who weigh 300 lb to 350 lb is approximately

52%, 45%, 10%, **4%**.

(j) The percentage of men in the sample who weigh 300 lb to 350 lb is approximately

52%, **45%**, 10%, 4%

3. (**Show your work**: 3 points each part.) A company makes protective body-armor for police and security personnel. It is important that it be an excellent fit, to provide protection without impeding movement. The population of people who will use this body armor have an average back length of 35.5 inches with an SD of 2 inches, and an average shoulder width of 22 inches with an SD of 1 inch. Both back length and shoulder width have approximately a normal distribution.

The company makes a range of standard sizes, but for the larger and smaller personnel it is more cost-effective to fit them individually. The standard sizes fit people who have back lengths 32 inches to 40 inches, and shoulder widths from 20 inches to 24 inches.

(a) What percentage of people have shoulders either too broad or too narrow to fit the standard sizes of body-armor?

20 to 24 is mean (22) ± 2 SD *(1 point)*
so 95 % will fit *(1 point)*
so 5% will not fit *(1 point)*

(b) What percentage of people will have too large a back-length to fit the standard sizes of body-armor?

z-score = (40-35.5)/2 = 2.25 *(1 point)*
Middle area = 98% from table *(1 point)*
So 1% will have too large a back length to fit *(1 point)*

(c) Fred has shoulders width 21 inches, and is at the 5*th* percentile (5% point) of back-length distribution. Will he need to be fit individually, or will the standard sizes of body-armor fit him?

His shoulders fit. *(1 point)*
For 5 th percentile, need middle area 90%; z-score = -1.65 from Table *(1 point)*
so Fred's back = 35.5 - 1.65×2 = 32.2. Yes he just fits! *(1 point)*

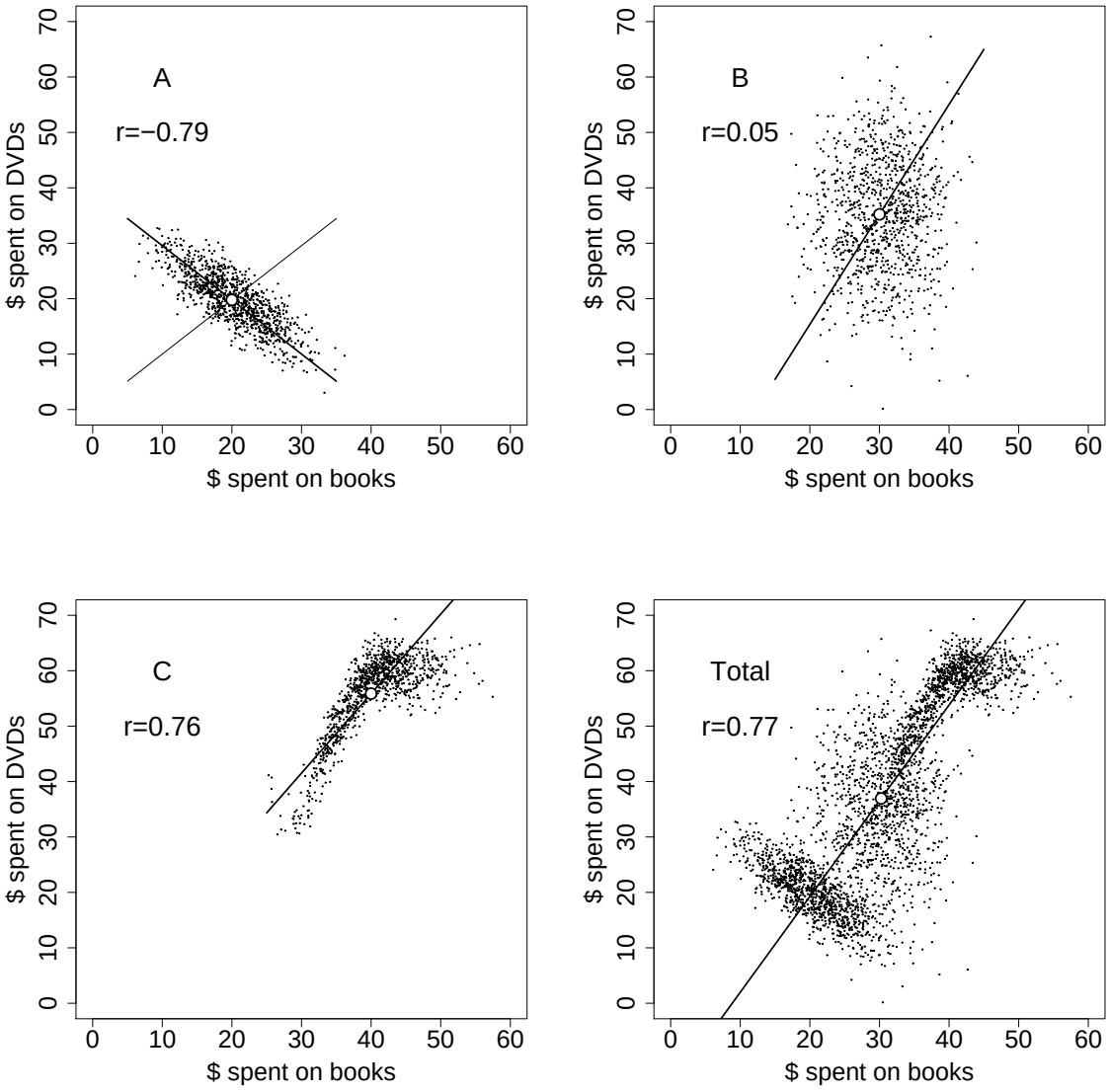
Or equally good to show lower limit 32" has z-score (32-35.5)/2=-1.75 ... Only 4% are too small, so at 5 th percentile he fits.

Or,compute both z-scores, and -1.75 is less than -1.65.

(d) The company says: if 5% of people have back lengths outside the range of our standard sizes, and 5% of people have shoulder widths outside the range of our standard sizes, then we are going to have to fit 10% of people individually for their body-armor. Do you agree? (Why? or why not?)

No, it will be somewhere between 5% and 10% *(1 point)*,
but likely some of the ones whose backs are too long/short will have shoulders too wide/narrow *(2 points)*
Can mention positive association between back length and shoulder width, but do not have to – it wd be true even if these were independent.

4. (2 points each part: see next page.) The BookCo company sells paperback books and DVDs in three regions of a country: A, B, and C. They use a "BookCo Member Bonus Card" system to keep record of sales to all their regular customers, and they compile the data on annual amounts spent on books and spent on DVDs by each customer, averaged over several years. The scatterplots are shown for each region separately, and for all three regions together. There are 1000 points in each of the first three plots. The point of averages is shown as a blank spot in the center of each scatter plot, and the solid line is the SD-line. (For region A, both positive and negative SD lines are shown.) The correlation coefficients for the points of each scatterplot is computed, and shown on the plot: for example "r= -0.79" on plot A means that the correlation coefficient computed for the points in plot A is -0.79.



Answer the questions on the next page, on the basis of the scatterplots shown (2 points each part). **For parts (a) and (b), no explanations are needed.**

4 ctd. (a) Approximately, what is the average annual $ amount spent on books by customers in each of the three separate regions?

A:	**20**	B:	**30**	C:	**40**

(b) By chance it happens that the standard deviations (SDs) of the annual $ spent on books are exactly the same in the three regions. The SDs of the annual $ spent on DVDs are approximately **5, 7,** and **10**. Which is which?

A:	**5**	B:	**10**	C:	**7**

(c) The President of BookCo calls his three Regional Managers for a meeting to discuss the data. He says: Overall there is a strong positive association between spending on books and spending on DVDs. There must be some mistake somewhere because in region A there is a strong negative association between spending on books and on DVDs, in region B there is almost no association, and only in C is there a positive association. Do you agree/disagree with all/part of his statement? Explain.

Yes, there is overall a strong positive association, but no mistake. *(1 point)* **The strong positive association comes from the 3 regions with low/middle/high average spending on both books and DVDs.** *(1 point)*
The relationship within each region is practically irrelevent.

(d) The Regional Manager of Region A says: In my region, people who buy more books tend to buy less DVDs. The data show they are making a choice to buy one or the other because they cannot afford both. Do you agree/disagree with all/part of her statement? Explain.

Yes, in her region there is a negative association. People who buy more books tend to buy less DVDs. (*1 point)* **No, there is no way the scatterplot can show this is because they cannot afford both.** (*1 point).*
Even though there is low spending on both books and DVDs, this may have nothing to do with lask of $; maybe they are buying MP3 players, or something quite else.

(e) The Regional Manager of Region B says: Even though on average my customers are between regions A and C in spending both on books and on DVDs, several of the largest and smallest amounts of DVD spending are from my region. How can this happen? Explain how this happens.

This is just because this region has the largest SD for DVD spending. (*1 point*). **The values are more spread out, so the extreme ones can be lower than region A, or higher than region C.** (*1 point*) *Note, with 1000 points, we expect a very few at* $\pm$ *3 SD from the average. There are NO true outliers.*

(f) The Regional manager of Region C says: The points in my Region are tightly clustered. Even if I plot the standardized values, my points still look a lot more clustered than the total set of points. But the correlation coefficients for the whole data set, and just for my region, are almost exactly the same. How can this happen? Explain how this happens.

Yes, his points are very tightly clustered, but about a curve, not about a line. (*1 point*) **The non-linear relationship is his region makes the value of the correlation coefficient not a good measure of association.** (*1 point*)
The same is somewhat true of the overall values; the combined plot is not very football-shaped. But it is much more non-linear for C.