Statistics 1040, Sections 002, 003 & 004, Midterm 1 (200 Points)

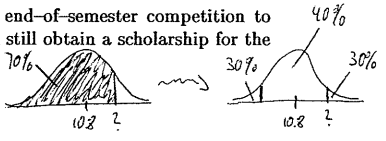
February 15, 2002

Your Name: _____

Question 1: Normal Distribution **(40 Points)**

A college that has an excellent track–and–field athletics program runs short on scholarships and cannot further support all of its 100m track athletes. The athletics director wants to make a decision which athletes to support in the future based on their athletic capabilities. Based on the athletes performance over the last few years, it is known that the distribution of running times approximately follows the normal curve with an average of 10.8 sec and a standard deviation of 0.2 sec. Answer the following 2 questions:

1. Which time does an athlete have to run in the end–of–semester competition to belong to the fastest 70% of the runners that will still obtain a scholarship for the next year? **(20 Points)**



Question asks for the 70th percentile:

From table: area between −0.50 to 0.50 is 38.29%
area between −0.55 to 0.55 is 41.77%

Since we are looking for a percentile (70th) that is larger than 50%, we have to work with 0.50 (or 0.55).

Original time: $10.8\,\text{sec} + 0.50 \cdot 0.2\,\text{sec} = \underline{10.90\,\text{sec}}$
$10.8\,\text{sec} + 0.55 \cdot 0.2\,\text{sec} = \underline{10.91\,\text{sec}}$

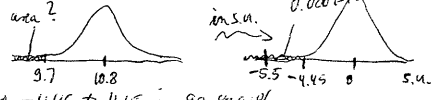
A time of about 10.90 sec (up to 10.91 sec) would be needed to belong to the fastest 70% of the runners.

- −5 incorrect area used (e.g. 70% and not 40%)
- −5 incorrect side
- −5 incorrect table value
- −5 incorrect final time

2. What are the chances that a randomly selected athlete from this college will set a new world record of 9.7 sec or better in the end–of–semester run? **(20 Points)**

17 correct value, no explanation

$\frac{9.7\,\text{sec} - 10.8\,\text{sec}}{0.2\,\text{sec}} = -5.5$ s.u.



10 correct s.u.

10 correct explanation/interpretation

From the table we get that the area between −4.45 to 4.45 is 99.9991%.
area outside is 0.0009%
area below −4.45 is $\frac{1}{2} \cdot 0.0009\% = 0.00045\%$

Note that −5.5 is even smaller than −4.45. This means the area is also smaller than 0.00045%.
So, for practical purposes, the chance is basically 0%, i.e., no athlete from this college has a realistic chance to set a new world record of 9.7 sec or better.

From: Stat 1040, Spring 2000, Final Test, Friday May 5, 2000.

Question 2: Controlled Experiment/Observational Study **(40 Points)**

In a recent study on SIDS (Sudden Infant Death Syndrome), one hospital collected data on 128 babies who died from SIDS in the last 12 months. They took a random sample of 500 babies (of similar ages) who did not die from SIDS (the "controls"), and they compared the two groups with respect to several variables of interest (e.g. whether the child slept on his or her stomach, birthweight, time of year, whether the mother smoked, whether she breast–fed, socio–economic status, etc.).

1. Is this a controlled experiment or an observational study? Explain. **(10 Points)**

7 correct keyword
3 correct explanation
1 some explanation

Worklook:

It is an observational study - there was no intervention.

2. One physician noticed that 63% of the SIDS babies had mothers who smoked during pregnancy, whereas only 26% of the control babies had mothers who smoked during pregnancy. Another physician claimed that low birthweight could be a "confounding factor". Explain what it means for low birthweight to be a "confounding factor". Be specific. **(15 Points)**

common for explanation
15/10/5

Worklook:

Perhaps smoking causes low birthweight and it is the low birthweight rather than smoking itself that is leading to higher rates of SIDS.

3. If you had access to the data, what would you do to "control for" birthweight? **(15 Points)**

common for explanation
15/10/5

Worklook:

Study babies with similar birthweights separately. eg. break up the comparison into groups of, say, babies 6–6.5 lb, 6.5–7 lb, 7–7.5 lb, etc.

Note: - independent (explanatory) variables are:
- sleeping on stomach (yes/no)
- birthweight
- time of year
- mother smokes (yes/no)
- mother breast-fed (yes/no)
- socio-economic status, etc.

- dependent (response) variable is: 2 • child dies of SIDS (yes/no)

In a controlled experiment, one might put half of the children sleep on stomach and the other on their back, breast-feed half of the children, etc. – dying of SIDS always would be the response!

Question 3: Guessing the Correlation Coefficient **(40 Points)**

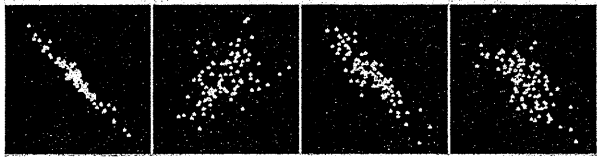
The correlation coefficients for the data points displayed in these four scatterplots are 4 out of the following 12 values:

~~−1.03~~, −0.97, −0.88, −0.69, −0.46, ~~−0.05~~, ~~0.05~~, 0.46, 0.69, 0.88, 0.97, ~~1.03~~

For each plot below, indicate the corresponding correlation coefficient r:

−0.97 0.46 −0.88 −0.69

Plot A Plot B Plot C Plot D



for each plot: 5 for 1 level off, i.e.,
A: −0.88
B: 0.05, 0.69
C: −0.97, −0.69
D: −0.88, −0.46
3 for correct value, but incorrect sign (+/−)
2 for correct direction (+/−)
0 for −1.03 or 1.03

(10) Correlation for Plot A: −0.97

(10) Correlation for Plot B: 0.46

(10) Correlation for Plot C: −0.88

(10) Correlation for Plot D: −0.69

Explanation (not required for your answer):
−1.03 and 1.03 are absolutely impossible as values for the correlation coefficient r; only values between −1 to 1 are possible
−0.05 and 0.05 relate to almost no correlation; it would be very difficult to imagine an increasing or decreasing line in such cases; we don't have this situation here
Four magnitudes remain: ±0.97, ±0.88, ±0.69, ±0.46; sort plots from strongest to weakest association:
Plots A, C & D show a negative association, with A strongest (−0.97), then C (−0.88), and D weakest (−0.69).
Plot B shows a positive association which is weaker (0.46) than the association in D (−0.69).

From: Freedman, Pisani, Purves, Chapter 10, Review Exercise 4, page 176-177

Question 4: Regression **(40 Points)**

$avg_x = 12$ $SD_x = 3$ $r = 0.50$
$avg_y = 12$ $SD_y = 3$

In one study, the correlation between the educational level of husbands and wives in a certain town was about 0.50; both averaged 12 years of schooling completed, with an SD of 3 years.

1. Predict the educational level of a woman whose husband has completed 18 years of schooling. **(15 Points)**

$$s.u._x = \frac{x - avg_x}{SD_x} = \frac{18-12}{3} = 2$$
$$s.u._y = r \cdot s.u._x = 0.50 \cdot 2 = 1$$
$$y = avg_y + s.u._y \cdot SD_y = 12 + 1 \cdot 3 = 15 \text{ years}$$

15 for correct result
5 for each correct step
1 for incorrect result (and no work shown)

2. Predict the educational level of a man whose wife has completed 15 years of schooling. **(15 Points)**

as above

$$\text{"}s.u._x\text{"} = \frac{15-12}{3} = 1$$
$$\text{"}s.u._y\text{"} = 0.50 \cdot 1 = 0.50$$
$$\text{"}y\text{"} = 12 + 0.50 \cdot 3 = 13.5 \text{ years}$$

3. Apparently, well–educated men marry women who are less well–educated than themselves. But the women marry men with even less education. How is this possible? **(10 Points)**

Nothing unexpected - this is just the regression effect!

10 for correct keyword
5 for reasonable explanation (without keyword)
1 for some explanation

Workbook answer:

4. (a) 15 years.
   (b) 13.5 years.
   (c) This is just the regression effect.

From: Stat 1040, Spring 2000, Midterm 1, February 17, 2000.
similar question: Stat 1040, Spring 1999, Final Test, Monday May 3, 1999.

Question 5: Representative Sample **(40 Points)**

In 1998 a researcher took a large representative sample of women in the U.S. She found that the older these women were, the less their daily average meat consumption. True or false and explain: "The data show that as women grow older, their daily average meat consumption drops."

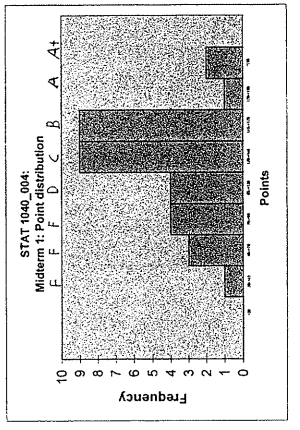
Worksheet:

False - this is a cross-sectional study. To find out what happens as people age, you need to watch them age. An alternative explanation could be that women who are young today eat more meat than those who were young in the past.
Not: "association is not causation" was not enough - even association is questionable here! (i.e. the point is, we don't even know that age & meat consumption are associated, and we are not trying to establish a cause).

20 for correct diagnosis (cross-sectional study)
10 for stating "false" (circling, in bold, etc.)
10/5/1 common for explanation

Sum Meit	
Mean	127.1
Standard Error	7.1
Median	132
Mode	132
Standard Deviation	40.7
Sample Variance	1658.2
Kurtosis	-0.5
Skewness	-0.3
Range	152
Minimum	44
Maximum	196
Sum	4195
Count	33

Points	Frequency
<20	0
20-<45	1
45-<70	3
70-<95	4
95-<120	4
120-<145	9
145-<170	9
170-<195	1
≥195	2



Sum Meit	
Mean	128.4
Standard Error	5.9
Median	130
Mode	118
Standard Deviation	35.6
Sample Variance	1270.6
Kurtosis	-0.5
Skewness	-0.2
Range	145
Minimum	43
Maximum	188
Sum	4749
Count	37

Points	Frequency
<20	0
20-<45	1
45-<70	0
70-<95	6
95-<120	10
120-<145	9
145-<170	4
170-<195	7
≥195	0