Statistics 31, Section 3, Midterm I, Solution
Tuesday, September 26, 2000

Name: ____

Pledge: I have neither given nor received aid on this examination.

Signature: ____

Instructions: Do not do any actual numerical calculations (e.g. answers in a form that you would type into an Excel field, with a working answer, are expected).
[points per problem]

1. Admissions policies at the Law School and Business School in a major university were compared for gender bias. Here is a breakdown of admissions during one period:

	Law	School	Business	School
	Admitted	Denied	Admitted	Denied
Male:	70	30	3	7
Female:	8	2	40	60

Note that in each school, the percent of females admitted (80% and 40%, respectively) is higher than the percentage of males (70% and 30% respectively). Yet, if the admissions across the schools are aggregated, the percent of females admitted (~44%) is paradoxically MUCH LOWER than the percent of males (~66%).

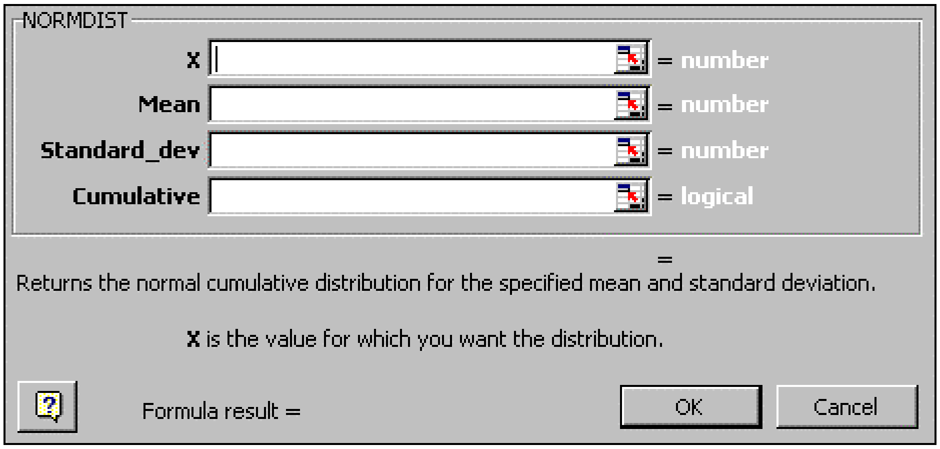
a. Which of the following is the name of this phenomenon? [2]
i. Confounding of Variables
ii. Simpson's Paradox
iii. The Law of Averages
iv. Extrapolation

b. What is the lurking variable in the aggregated scores?
[3]
The type of school, i.e. was it Business School or Law School?

c. Explain in 20 words or less why the aggregated percentages show women as being admitted less frequently.
[5]
Women applied to Business School much more often, but that school had a much higher overall rejection rate.

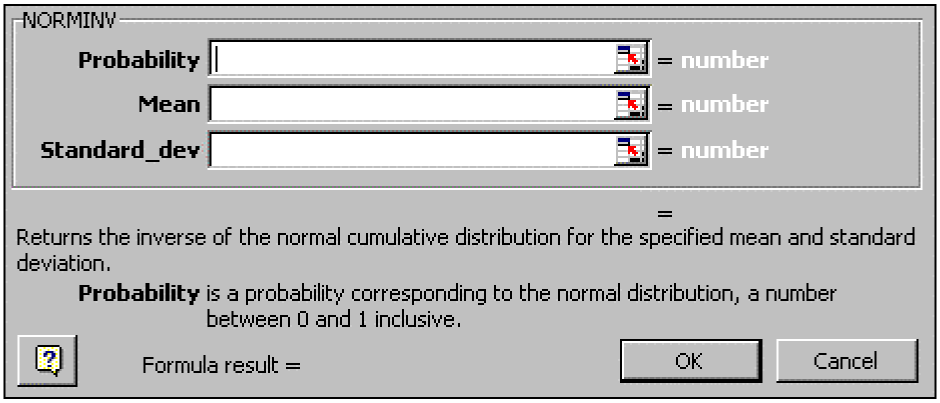
2. Lengths of pregnancies vary approximately according to a Normal distribution with mean 266 days and standard deviation 16 days.

(a) Choose an Excel menu below (only one!), and fill it out to find the pregnancy length at the third quartile of the population.



Returns the normal cumulative distribution for the specified mean and standard deviation.

X is the value for which you want the distribution.



Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation.

Probability is a probability corresponding to the normal distribution, a number between 0 and 1 inclusive.

[4] Use NORMINV, Prob. = 0.75, Mean = 266, S. D. = 16.

(b) Write an Excel command to calculate the percent of pregnancies between 250 and 290 days

[4]

=NORMDIST(290,266,16,TRUE)-NORMDIST(250,266,16,TRUE)

(c) Write an Excel command to calculate the percent of pregnancies that are within 1.5 standard deviations of the mean.

[4]

=NORMDIST(266+1.5*16,266,16,TRUE)-NORMDIST(266-1.5*16,266,16,TRUE)

(d) Write an Excel command to calculate how long the longest 10% of pregancies last.

[4]

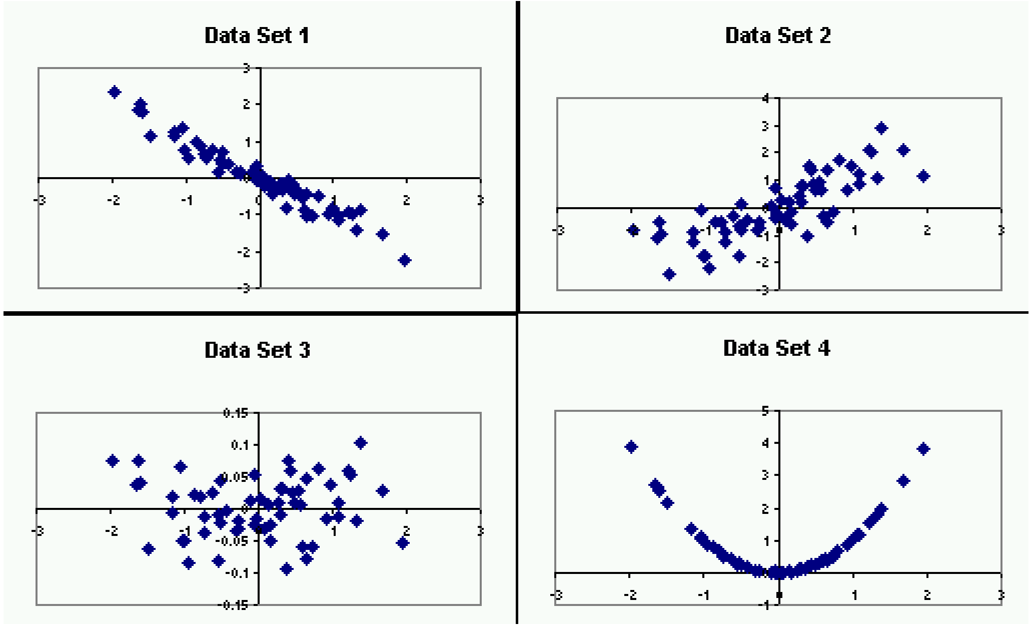
=NORMINV(1-0.1,266,16)

(e) Use the 68-95-99/7 rule to write Excel commands to calculate values between which the middle 95% of all pregnancies lie.

[4]

Mean + or – 2 sd's ie. =266+2*16 and =266+2*16

3. Here are scatterplots for 4 Toy Data Sets



Match the data sets to all statements which apply (matches can overlap, can be reused, and may not exist).

a.	Data Set 1	1,4	The variables are strongly associated with each other.
	Data Set 2	2	The variables are moderately associated with each other.
	Data Set 3	3	The variables are not associated with each other.
	Data Set 4		

[5]

b.	Data Set 1	1	The correlation is approximately	r = -0.95
	Data Set 2	3,4	The correlation is approximately	r = 0
	Data Set 3	2	The correlation is approximately	r = 0.8
	Data Set 4		The correlation is approximately	r = 0.95

[5]

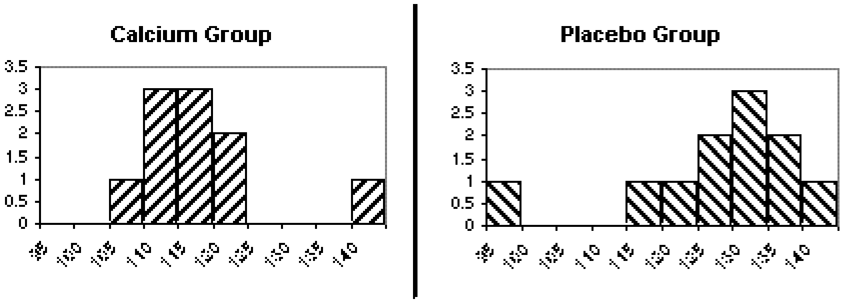
c.	Data Set 1	1234	There are no obvious outliers
	Data Set 2		There is one serious outliers
	Data Set 3		There are two probably outliers
	Data Set 4		There are many outliers

[5]

d.	Data Set 1	2	There is a positive linear relationship.
	Data Set 2	1	There is a negative linear relationship.
	Data Set 3	4	There is a curved relationship.
	Data Set 4	3	There is no apparent relationship.

[5]

4. In a medical experiment, one group of men was given calcium, and another group was given a placebo. After some time their blood pressures were recorded and are shown in these two histograms.



Circle one of True, False or N.E.I. ("Not Enough Information") for each of the following:

a. True [✓ False] N.E.I The Calcium Group had higher blood pressures overall.
[2]

b. [✓ True] False N.E.I The Placebo Group population shows more variability.
[2]

c. True False [✓ N.E.I] There is a mild positive correlation between Cal. and Pla. Groups
[2]

d. [✓ True] False N.E.I The Placebo Group has an outlier to the left.
[2]

e. True [✓ False] N.E.I When the outlier is ignored, the Placebo Dist'n is left skewed.
[2]

f. [✓ True] False N.E.I When the outlier is ignored, the Cal. Dist'n is fairly symmetric.
[2]

g. True [✓ False] N.E.I The Calcium Distribution has 3 modes.
[2]

h. True [✓ False] N.E.I The median is larger for the Calcium Group than for the Controls.
[2]

i. True [✓ False] N.E.I The Inter Quartile Range is larger for the Calcium Group.
[2]

j. True [✓ False] N.E.I The range is larger for the Calcium Group.
[2]

5. To understand erosion, water was released on a test bed, at different flow rates, and the amount of eroded soil was measured.

a. What is the response variable?

[5]

Amount of eroded soil

b. What is the explanatory variable?

[5]

Flow rate

c. If the x-data values are in Excel cells D4:D24, and the y data values are in the Excel cells E4:E24, write an Excel formula to calculate the y-intercept of the least squares regression line.

[5]

=INTERCEPT(E4:E24,D4:D24)

d. For data as in (c), write an Excel formula to calculate the slope of the least squares regression line.

[5]

=SLOPE(E4:E24,D4:D24)

e. If the y-intercept and slope from (c) are –3 and 2 (respectively), write an Excel formula to calculate the predicted y value corresponding to a new x value of 27.

[5]

Y = m * x + b   =2*27-3

f. If the x values range from 5 to 15, is the prediction in (e) likely to be reasonably accurate? Explain why or why not in 20 words or less.

[5]

No, x values is outside the range of the given ones, so have extrapolation.