Recitation assignment due at the end of recitation 4 - 6 - 10.

Readings for the week are: for Monday Chapter 7; for Tuesday Chapter 9.

Textbook exercises for recitation Tuesday are:

Any plot of (x, y) scores is called a scatter plot. Dependencies observed between two scores or more scores are central to statistical method.

Much of the statistical thinking brought to bear on scatter plots has its origins in two early grand successes: Gauss (Astronomy, predicting future positions of heavenly bodies from a few observations of past positions) and Galton (Heredity, a scatter plot whose correlation actually established beyond doubt a measure of the force of heredity in the weights of seeds). Both were startling achievements at the time and even now. With hindsight, knowing all the mathematics, even knowing how he cleverly chose his subjects and found ways around the computational difficulties by using medians instead of means, it would take years for me duplicate. Galton's project, involving as it did the cooperation of his friends to grow seeds in many places, count and sort data and various ingenious simplifications remains a work of genius.

In both of the above cases the data was by nature "noisy." Measurements of astronomical positions were subject to various inaccuracies and so too natural variability was seen in the yields of seeds and growing conditions. A key idea which surfaced was that practical, even profound advances in scientific tools could be found in the emerging ideas of probability and statistics. Before that, noisy data was distrusted as bad science.

Such early successes brought needed insights to science and beyond but have been curiously and needlessly rendered as dogma by some.

In Chapter 7, page 182, item 169 is a case in point. It is said that you are to "Assign to the y-axis the response variable that you hope to predict or explain. Assign to the x-axis the explanatory or predictor variable that accounts for, explains, predicts, or is otherwise responsible or the y-variable." That is the usual convention, it is true. However, the better view is that (x, y) may have some relationship between them which you wish to exploit. Correlation r (page 172) is a measure of straight line fit of y on x but it is the same measure when the roles of x, y are interchanged (i.e. r[x, y] = r[y, x]).

We will see in Chapter 8 that if (x, y) are **jointly normal distributed** then there is a natural line for y on x and a second natural line for x on y. These two lines are the same only if $|r| = 1$ (perfect fit, all (x, y) points fall perfectly on a line). If your purpose is to utilize one variable to predict or explain the other then yes, the convention is that y is the one to be explained and the line for y on x is used. The slope and intercept of this line are given as item(s) 196 on page 212.

1. Do 7-1 (pg. 186). If you are unable to identify one of the two variables as **necessarily** y just say so.

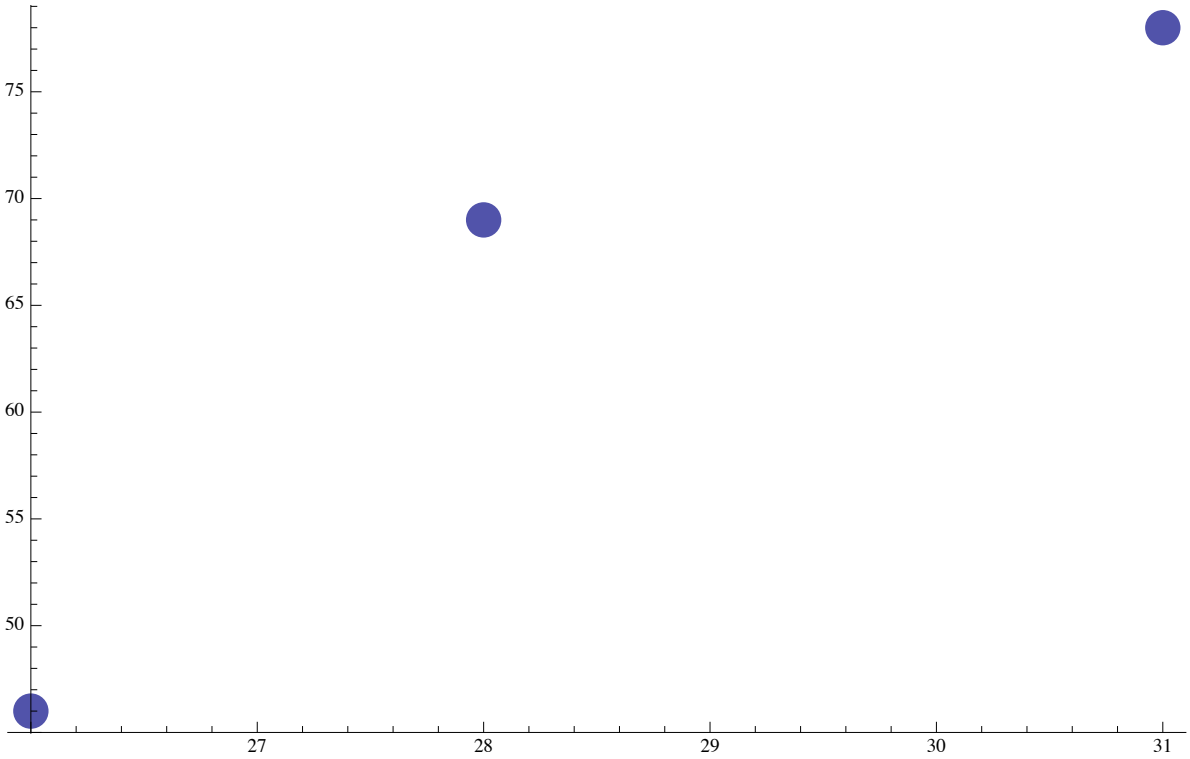
2. Do 7-2. As above.

3. For the (x, y) data

x	y	x deviations = $(x - \overline{x})$	y deviations = $(y - \overline{y})$	product
31	78			
26	46			
28	69			

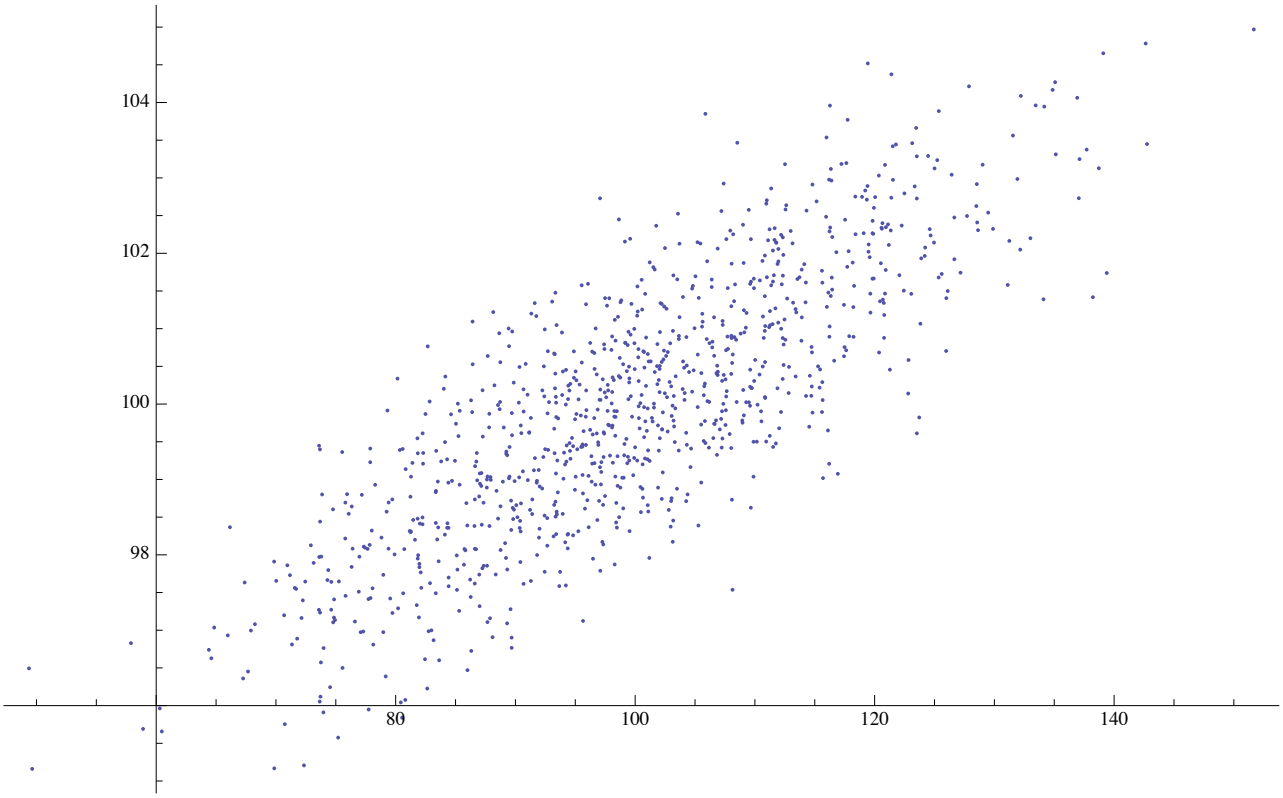
a. Fill out the above table to complete the table find the correlation r by hand as is shown on page 173.

b. In the plot of the data (a) we have not taken the origin of the plot to be (0, 0) since that would place the interesting part of the plot well away from the origin necessitating a fine scale (to get both in view) that would compress the data into a little blob.



Overlay on the scatter plot the line passing through the point of averages ($\overline{x}$, $\overline{y}$) and having slope equal to (r s_y / s_x). This line is known as the **least squares line of y on x** and is taken up in the following chapters. Does the line seem, to your eye, to be a reasonable fit to the three points?

4. Here is a plot of data having a **jointly normal** (i.e. 2-dimensional normal) distribution for (x, y). Note the roughly elliptical form of the plot.



a. For a jointly normal plot the regression line is just the line passing through the mid-height of the points at each x. You can easily draw it by hand. Just go through the middle of the plot. Label this line "regression of y on x." Your line should pass through the point of averages which is approximately (100, 100) for this particular data. If you were to calculate the slope as in #3 you would get this very line you can fit by eye for jointly normal (x, y) data.

b. Now look at things with the roles of x, y interchanged. Turn the picture (just above) so the y-axis is horizontal and draw another line, the regression line of x on y, which passes through the middle x for each y. Galton saw all of this in his remarkably nearly **jointly normal** data for x = parental seed weight and y = filial seed weight. So too for his data on x = father's adult height and y = son's adult height. The perfect expression of all this in planting conducted by his friends literally amazed him, led to the definition of correlation, defined a measure of the parental contribution to filial seed weight, tied in with continental research in astronomy, and helped propel statistics and the various scientific disciplines into an increasingly close relationship.

5. Do 7-4 . Be creative here. Play devil's advocate against what the authors may be urging.

6. Do 7-5.

7. Do 7-6.

8. Do 7-8.

9. Do 7-11.

10. Do 7-12.

11. r[x, y]= 08. What are

r[y, x] =

r[-x, y] =

r[-x, -y]]

r[2 x, 4 y] =

r[3x - 6, 4 y + 11] =

12. On page 173 we find "don't apply correlation to categorical data masquerading as quantitive." If I followed that advice I would not be able to narrow the z-based CI for μ_y (based on independent samples (x_i, y_i), i ⩽ n from

$$\overline{y} \pm z\, s_y / \sqrt{n}$$

the (except in rare conditions the narrower) regression-based

$$\boxed{\overline{y} + (\mu_x - \overline{x})\, r[x, y]\, s_y / s_x} \pm z \boxed{\sqrt{1 - r[x, y]^2}}\, s_y / \sqrt{n}$$

when the population mean μ_x is known, even for the case in which score x is 1 for males and 0 for females (and I know the population mean = rate of males in the population). To the extent that the sample correlation r[x, y] is near to 1 the regression-based method is both adjusting the estimator of μ_y from $\overline{y}$ to

regression-based estimator $\boxed{\overline{y} + (\mu_x - \overline{x})\, r[x, y]\, s_y / s_x}$

to which applies the (generally narrower than $z\, s_y / \sqrt{n}$) CI half width

$$z \boxed{\sqrt{1 - r[x, y]^2}\ s_y / \sqrt{n}}.$$

Determine these two CI for data on bar patrons having

x = 1 if male, 0 if female
y = age
n = 100 (sex and age obtained for each of a random sample of 100)

$\bar{x}$ = sample mean = 0.54 (given)
μ_x= 0.62 (given)
(suppose **we know** that 62% of patrons are male in the population)
s_x = sample sd of x-scores = $\sqrt{.54 \times .46} \sim 0.55$
(better to use $\sigma_x = \sqrt{.62 \times .38}$ throughout but don't do that now)

$\bar{y}$ = 26.7
μ_y= unknown and to be estimated
s_y = 3.9
r[x, y] = sample correlation = 0.58

a. Regular estimator of population mean age (ignores x-scores altogether) is
$\bar{y}$ =

b. Regular 95% z-based CI for population mean age (ignores x-scores altogether) is
$\bar{y} \pm z\, s_y / \sqrt{n}$ =

c. Regression-based estimator of population mean age is
$\boxed{\bar{y} + (\mu_x - \bar{x})\, r[x, y]\, s_y / s_x}$ =

d. Regression-based 95% CI for population mean age is
$\boxed{\bar{y} + (\mu_x - \bar{x})\, r[x, y]\, s_y / s_x} \pm z \boxed{\sqrt{1 - r[x, y]^2}}\, s_y / \sqrt{n}$ =

e. Experiment to find an n* for which the regular CI (b) would give the same CI half width as (d), i.e. find n* for which

$1 / \sqrt{n^*} = \boxed{\sqrt{1 - 0.58^2}} / \sqrt{100}$

If the cost to obtain each sample is \$100 dollars, how much money has been saved using the regression-based approach?