Lesson Plan

- Discuss Quizzes/Answer Questions
- Statistical Graphics (Not Maps)
- Box Plots
- Linear Relationships
- Correlation

5.1 Statistical Graphics (Not Maps)

In the last lecture, we examined statistical graphics related to maps. We now begin with some other kinds of graphics.

William Playfair, the man who invented the histogram, made many plots for Parliament. The one shown here is a longitudinal plot, showing change over time.

Why story does the graph tell?

The second plot shows the age pyramid for the French population in 1968. The upper plot has unnecessary detail (chartjunk) which many graphics experts dislike.

- Why are there "notches" in the curve?
- Why are there fewer men in certain age groups?
- Where is the baby boom?

The third graph is a Pareto chart, based on responses students gave in their course evaluations a few years ago.

A Pareto chart is used in industry to identify patterns of problems.

The fourth graph is a histogram of the number of hours students reported working on an introductory statistics course. Why is there a spike at 9?

The fifth graph is a set of side-by-side boxplots that show how the number of hours that students spent on my course is related to their overall evaluation of the course. Why is there an irregularity for those giving the lowest rating?

The sixth graph is another set of side-by-side boxplots, showing draft rank against birth month for the 1970 lottery to go to Vietnam. What does this show?

The seventh graph is a scatterplot of the scores that my students got on the final against the number of hours of sleep that they had.

There were two sections for the class, denoted by open circles and dots. Is there any difference between the sections?

There is a weak trend to increasing scores with increasing hours of sleep. But the trend is not strong, and the correlation is probably only about .3 or so.

What are two possible reasons for this increasing trend?

5.2 Boxplots

The boxplot is a graphic that summarizes a dataset by plotting 5 numbers:

- the least
- the largest
- the median
- the 25th percentile
- the 75th percentile

The last two are the medians of the numbers less than the median and the numbers greater than the median, respectively (according to the convention for this class).

A boxplot looks like this:

Boxplots are often made for different groups and plotted side by side to highlight differences.

Suppose one made a side-by-side boxplot of the IQ scores for Duke students and the IQ scores for students at the University of Idaho. What do you think they would look like?

.

5.3 Lines

The algebraic equation for a line is

$$Y = mX + b$$

If you don't know this stuff, read the book carefully.

The use of coordinate axes to show functional relationships was invented by René Descartes (1596-1650). He was an artillery officer, and probably got the idea from pictures that showed the trajectories of cannonballs.

5.4 Correlation

Sir Francis Galton explored Africa, invented eugenics, studied whether ships that carried missionaries were less likely to be lost at sea, pioneered birth-and-death models and meteorology, and was Charles Darwin's cousin.

He also was the first to conceive of linear regression (although he did not have the mathematical skill to develop the formulae, and got a friend of his at Cambridge to do the derivations).

Correlation is a measure of the strength of the linear association between two continuous variables. The book discusses the relationship between the height of fathers and the height of sons.

Clearly, tall fathers tend to have tall sons, and short fathers tend to have short sons. If the father's height were a perfect predictor of the son's height, then all father-son pairs would lie on a straight line in a scatterplot.

Regression fits a line to the points in a scatterplot. The term comes from the father-son example. An exceptionally tall father tends to have sons that are shorter; an exceptionally short father tends to have sons that are taller. Thus the sons' height tend to "regress towards the mean".

The correlation tells you how much information one of the variables provides about the other. If the absolute value of the correlation is near 1, then knowing one tells the other almost perfectly (if the relationship is linear).

If the correlation is near zero, then knowing one variable gives essentially no information about the other.

All correlations are between -1 and 1, inclusive.

The square of the correlation is called the coefficient of determination. It is the proportion of the variation in Y that is explained by knowledge of X. If this is 0, then X provides no insight about Y (assuming linearity).

The assumption of linearity is important. Specifically, it asserts that:

$$Y_i = mX_i + b + \epsilon_i$$

where the Y_i and X_i are the observed values for the ith case (e.g., a father-son pair) and the ϵ_i is random error (due to genetics, measurement error, etc.).

If the assumption does not hold, then one can get strange and even misleading behavior.

To calculate the correlation coefficient, just take the average of the products of the z-transforms of the X values with the z-transforms of the Y values.

Example: Suppose your data are (1,0), (2,1), and (3,5). Then the mean of the X values is 2, the mean of the Y values is 2, and the sds are:

$$sd_X = \sqrt{\frac{1}{3}(1^2 + 2^2 + 3^2) - 2^2} = .8165$$

$$sd_Y = \sqrt{\frac{1}{3}(0^2 + 1^2 + 5^2) - 2^2} = 2.1602$$

The z-transformation subtracts the mean and divides by the standard deviation for each of the X and Y values, giving:

X	Z_X	Y	Z_Y
1	(1-2)/.8165 = -1.2245	0	(0-2)/2.1602 = -.9258
2	(2-2)/.8165 = 0	1	(1-2)/2.1602 = -.4629
3	(3-2)/.8165 = 1.2245	5	(5 - 2)/2.1602 = 1.3888

Then the correlation coefficient is

$$r = \frac{1}{3}[(-1.224)(-.9258) + (0)(-.4629) + (1.2245)(1.3888)] = .9446.$$

Do we really believe this?

How much of the variation in Y is explained by knowing X? This is $r^2 = (.9446)^2 = .89$.