Exploratory Data Analysis: Two Variables

FPP 7-9

Exploratory data analysis: two variables

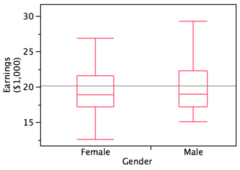
- 2 qualitative/categorical variables
  - Contingency tables (we will cover these later in the semester)
- 1 qualitative/categorical, 1 quantitative variables
  - Side-by-side box plots
- 2 quantitative variables
  - Scatter plots, correlations, regressions

Box plots

- A box plot is a graph of five numbers
  - minimum,
  - Maximum
  - Median
  - 1st quartile
  - 3rd quartile
- We know how to compute three of the numbers (min,max,median)
  - To compute the 1st quartile find the median of the 50% of observations that are smaller than the median
  - To compute the 3rd quartile find the median of the 50% of observatins that are bigger than the median

Side-by-side box plots

- Box plots are very useful when comparing distributions of a quantitative variable for levels of some qualitative variable

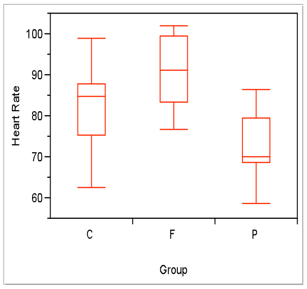


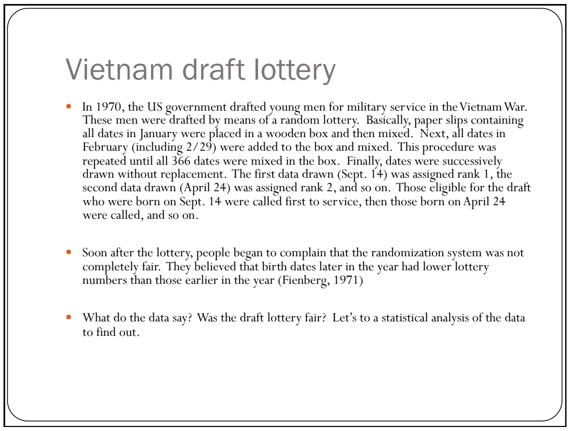
Pets and stress

- Are there any differences in stress levels when doing tasks with your pet, a good friend, or alone?
- Allen et al. (1988) asked 45 people to count backwards by 13s and 17s.
- People were randomly assigned to one of the three groups: pet, friend, alone.
- Response is subject's average heart rate during task

Pets and stress

- It looks like the task is most stressful around friends and least stressful around pets



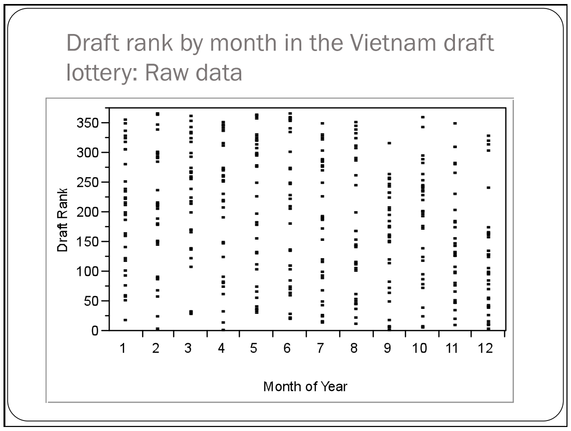


Vietnam draft lottery

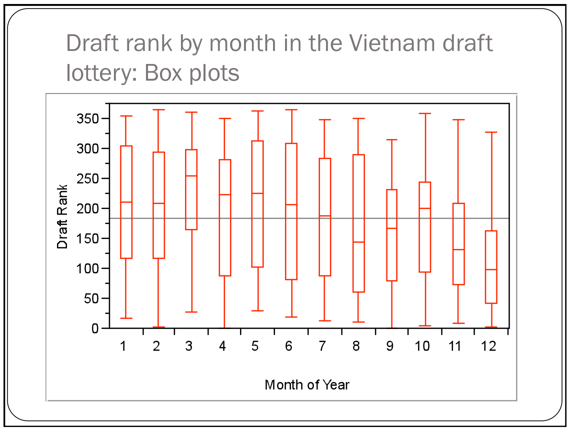
- In 1970, the US government drafted young men for military service in the Vietnam War. These men were drafted by means of a random lottery. Basically, paper slips containing all dates in January were placed in a wooden box and then mixed. Next, all dates in February (including 2/29) were added to the box and mixed. This procedure was repeated until all 366 dates were mixed in the box. Finally, dates were successively drawn without replacement. The first data drawn (Sept. 14) was assigned rank 1, the second data drawn (April 24) was assigned rank 2, and so on. Those eligible for the draft who were born on Sept. 14 were called first to service, then those born on April 24 were called, and so on.

- Soon after the lottery, people began to complain that the randomization system was not completely fair. They believed that birth dates later in the year had lower lottery numbers than those earlier in the year (Fienberg, 1971)

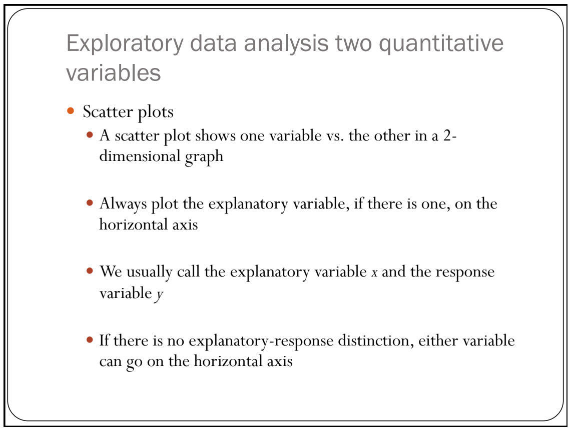
- What do the data say? Was the draft lottery fair? Let's to a statistical analysis of the data to find out.



Draft rank by month in the Vietnam draft lottery: Raw data

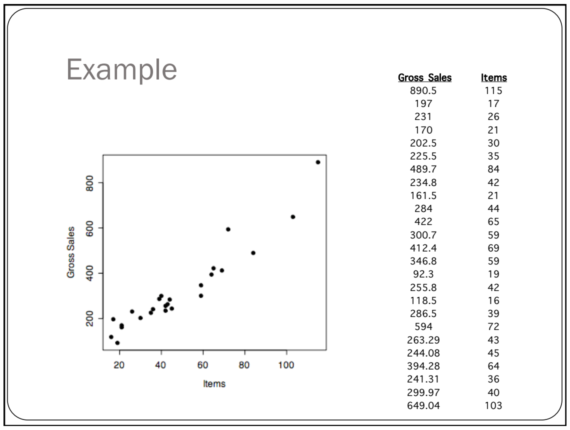


Draft rank by month in the Vietnam draft lottery: Box plots



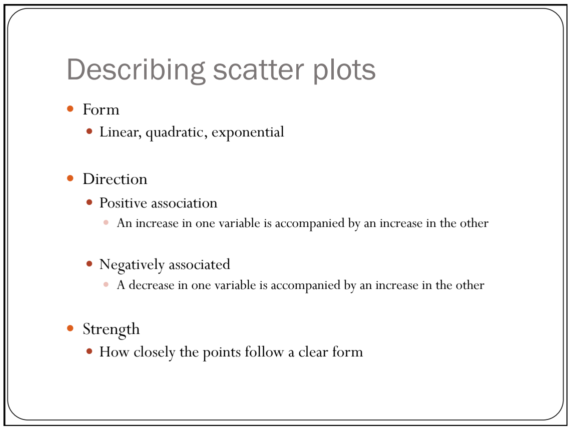
Exploratory data analysis two quantitative variables

- Scatter plots
  - A scatter plot shows one variable vs. the other in a 2-dimensional graph
  - Always plot the explanatory variable, if there is one, on the horizontal axis
  - We usually call the explanatory variable *x* and the response variable *y*
  - If there is no explanatory-response distinction, either variable can go on the horizontal axis



Example

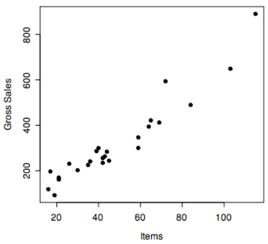
Gross Sales	Items
890.5	115
197	17
231	26
170	21
202.5	30
225.5	35
489.7	84
234.8	42
161.5	21
284	44
422	65
300.7	59
412.4	69
346.8	59
92.3	19
255.8	42
118.5	16
286.5	39
594	72
263.29	43
244.08	45
394.28	64
241.31	36
299.97	40
649.04	103



Describing scatter plots

- Form
  - Linear, quadratic, exponential
- Direction
  - Positive association
    - An increase in one variable is accompanied by an increase in the other
  - Negatively associated
    - A decrease in one variable is accompanied by an increase in the other
- Strength
  - How closely the points follow a clear form

Describing scatter plots



- Form:
  - Linear
- Direction
  - Positive
- Strength
  - Moderately strong?

Correlation coefficient

- We need something more than an arbitrary ocular guess at how strong an association is between two variables.
- We need a value that can summarize the strength of a relationship
  - That doesn't change with when units change
  - That makes no distinction between the response and explanatory variables

Correlation coefficient

- Definition: Correlation coefficient is a quantity used to measure the direction and strength of the **linear** relationship between two **quantitative** variables.
- We will denote this value as *r*

Computing correlation coefficient

- Let x, y be any two quantitative variables for n individuals

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where $\bar{x}$ and $\bar{y}$ are means and s_x *and* s_y are standard deviations of the variables x and y respectively

Correlation coefficient

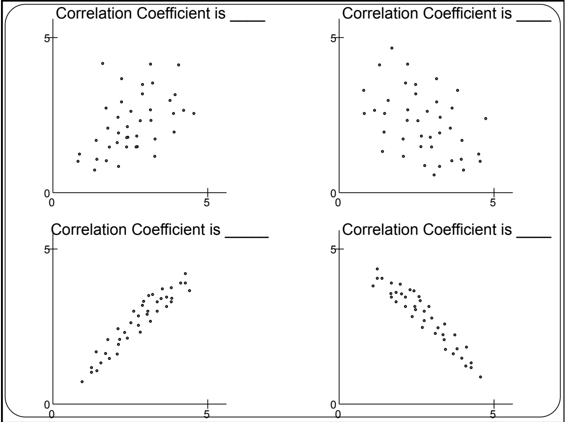
- Remember $\frac{x_i - \bar{x}}{s_x}$ and $\frac{y_i - \bar{y}}{s_y}$ are standardized values of variable x and y respectively
- The correlation r is an average of the products of the standardized values of the two variables x and y for the n observations

Properties of *r*

- Makes no distinction between explanatory and response variables
- Both variables must be quantitative
  - No ordering with qualitative variables
- Is invariant to change of units
- Is between -1 and 1
- Is affected by outliers
- Measures strength of association for only linear relationships!

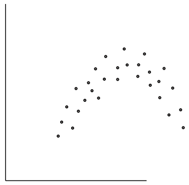
True or False

- Let X be GNP for the U.S. in dollars and Y be GNP for Mexico, in pesos. Changing Y to U.S. dollars changes the value of the correlation.



Correlation coefficient

- Correlation is not an appropriate measure of association for non-linear relationships



- What would r be for this scatter plot

Correlation coefficient

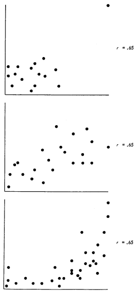


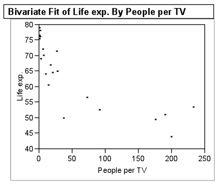
FIGURE 3-18 Three scatterplots with the same correlation

Correlation coefficient

- CORRELATION IS NOT CAUSATION
- A substantial correlation between two variables might indicate the influence of other variables on both
- Or, lack of substantial correlation might mask the effect of the other variables

Correlation coefficient

- CORRELATION IS NOT CAUSATION



- Plot of life expectancy of population and number of people per TV for 22 countries (1991 data)

Correlation coefficient

- CORRELATION IS NOT CAUSATION

- A study showed that there was a strong correlation between the number of firefighters at a fire and the property damage that the fire causes.
  - We should send less fire fighters to fight fires right??

  - Example of a lurking variable what might it be?

Interpreting correlations

- A newspaper article contains a quote from a psychologist, who says, “The evidence indicates the correlation between the research productivity and teaching rating of faculty members is close to zero.” The paper reports this as “The professor said that good researchers tend to be poor teachers, and vice versa.”

  Did the newspaper get it right?

Correlation coefficient

- What’s wrong with each of these statements?

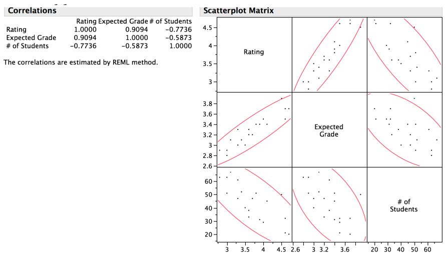
  - There exists a high correlation between the gender of American workers and their income.

  - The correlation between amount of sunlight and plant growth was $r = 0.35$ centimeters.

  - There is a correlation of $r = 1.78$ between speed of reading and years of practice

Examining many correlations simultaneously

- The correlation matrix displays correlations for all pairs of



Correlations

	Rating	Expected Grade	# of Students
Rating	1.0000	0.9094	-0.7736
Expected Grade	0.9094	1.0000	-0.5873
# of Students	-0.7736	-0.5873	1.0000

The correlations are estimated by REML method.

Scatterplot Matrix