

Data Files for the homework problems can be downloaded from the website http://www.duxbury.com/cgi-brookscole/course_products_bc.pl?fid=M67&discipline_number=17

Chapter 3 Data Description

- Our objective is to extract general impressions from a set of data. We will discuss more formal methods (tests and confidence intervals) after reviewing probability
- The main topics in this Chapter are: graphical methods, univariate summary statistics, and multivariate summaries

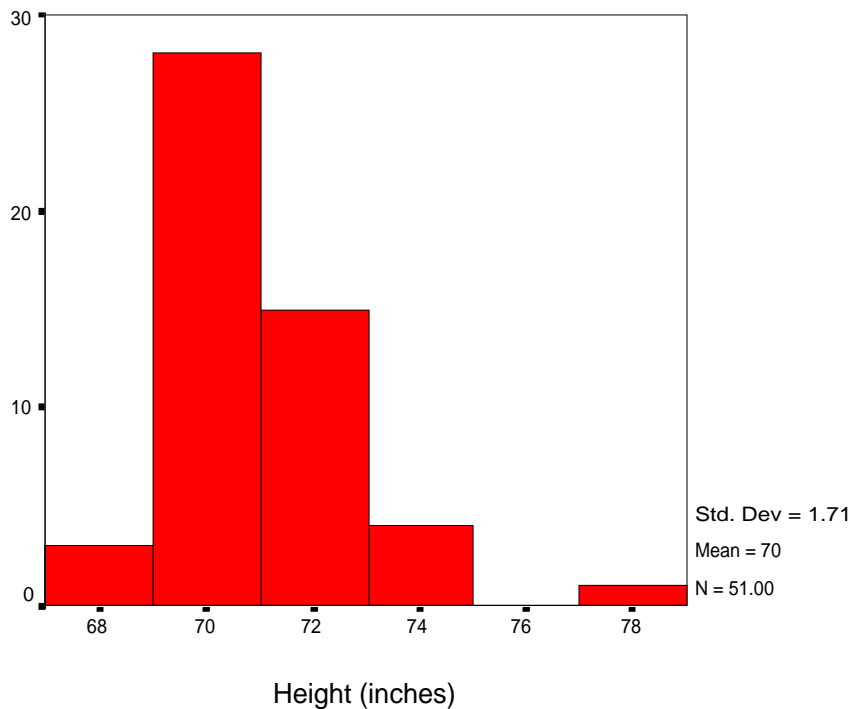
Definitions

- A *variable* is an attribute or characteristic of interest that is measured on the population units. For example, if the population units are humans, variables of interest might be: 1. age, 2. religious affiliation, 3. gender.
- A *quantitative variable* is a variable that is measured on a numerically ordered scale. E.g., age, weight, height
- A *qualitative variable* is a variable that is not measured on a numerically ordered scale. E.g., religious affiliation, gender

3.3 Describing a Single Variable by Graphical Methods

- Histograms are used to show the distribution of a variable. The *distribution* refers, in general sense, to the pattern of variation of the values.
- To illustrate, a sample of 51 adult males was collected and their heights measured to the nearest inch. Below is a frequency histogram of the heights

Figure. Frequency histogram for a sample of 51 adult male heights (in inches)



- The frequency of a value is how often the value occurs in the data set. For example, values of 67 and 68 occurred 3 times, so the height of the leftmost bar is 3
- Histograms are constructed by forming categories (if none exist), and counting the number of observations in each category. SPSS has formed the following categories

$$C_1 = [67, 68.99]$$

$$C_2 = [69, 70.99]$$

⋮

$$C_6 = [77, 78.99].$$

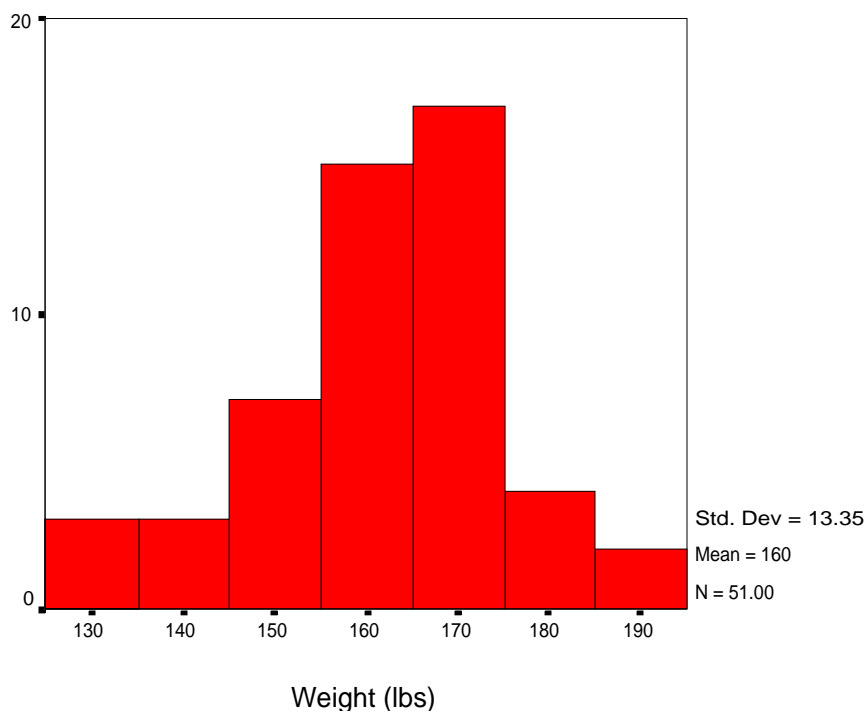
- A frequency table is sometimes used to present frequency, and relative frequency of values. The *relative frequency* of a value is the frequency divided by the number of observations in the set. Specifically, the *relative frequency* of the i th value is

$$f_i = \frac{n_i}{n},$$

where n_i is the *frequency* of the i th value, and n is the total number of observations. The sum of the relative frequencies is 1. For the heights example, $f_1 = 3/51$ because $n = 51$

- Ott and Longnecker (p. 47) provide guidelines for constructing histograms by hand. I encourage you to use software to produce them
- If a variable is categorical, then there is usually one category per value (e.g., males and females). If a variable is quantitative, then we usually use 5 to 20 categories (more categories for more values)
- When the variable is qualitative, the term *bar chart* is often used instead of histogram

Figure. Frequency histogram of the weights (lbs.) from the 51 adult males



Shapes of Histograms (refer to Ott and Longnecker, p. 52)

- The mode of a data set is the value with the largest (relative) frequency. The mode of the height data is 70". There are 3 modes of the weight data (160, 165, and 170), though there is no way to know this from the histogram
- A histogram with a single, dominant peak (mode) is called *unimodal*. If there are two prominent peaks, the histogram is *bimodal*.
- A distribution that looks approximately the same on both sides is called *symmetric*. The normal distribution is symmetric. The weight distribution is symmetric.
- A distribution with a longer right tail is called right-skewed; a distribution with a longer left tail is called left-skewed
- The bars of a *uniform distribution* are all approximately the same height.

Stem and Leaf Plots (see OL, p. 56)

- These are useful hand-drawn summary plots

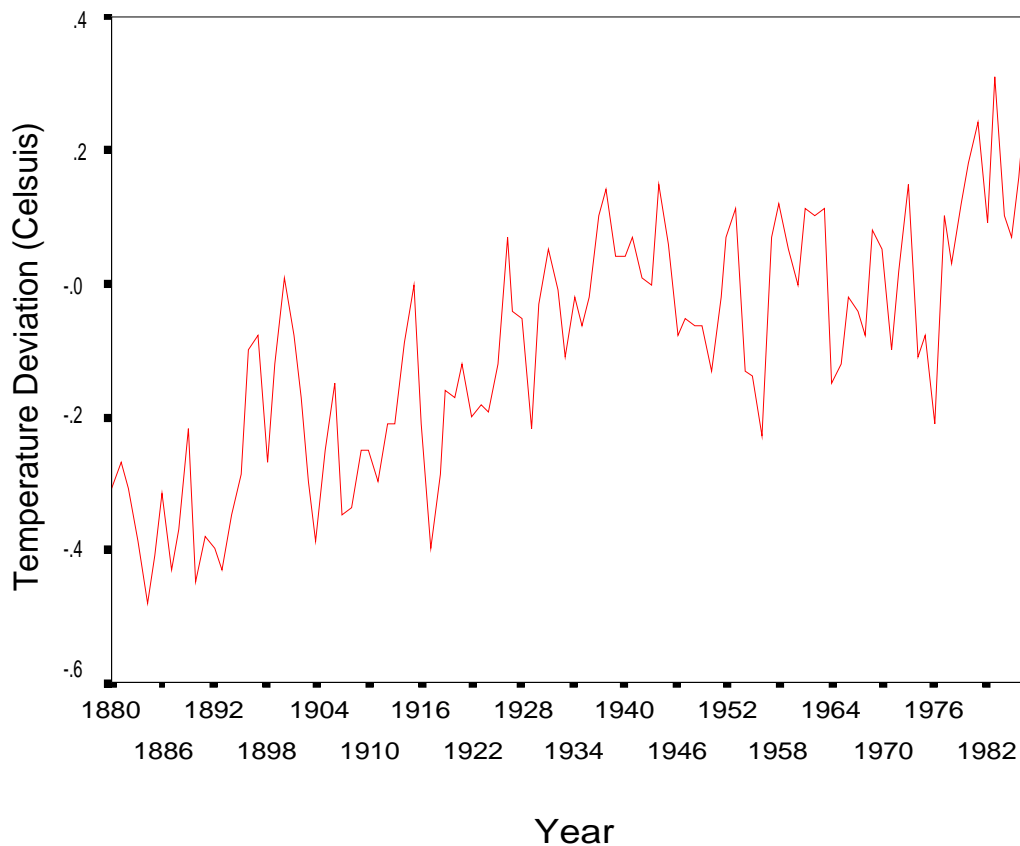
Construction:

- the rightmost digit of a number is a leaf
 - all other numbers are the stem (usually)
1. sort the data, from smallest to largest
 2. write down the stems
 3. write each leaf on the corresponding stem; use the same amount of space for each digit; use no other characters (no commas)

Time Series Plots

- These are useful if a variable is measured over time, and time is a variable that is suspected of being associated with the values.
- For example, atmospheric CO₂ has increased over the past 120 years. We say that there is an *association* (or a relationship) between time and CO₂. A plot of CO₂ measurements against time will provide some insight into the pattern of change.
- Example: From Jones, P.D. 1988, "Hemispheric surface air temperature variations—recent trends plus an update to 1987," *J. of Climatology*, **1**, 654-60. Data are annual average Northern Hemisphere temperatures (Celsius).

Figure. Deviation from average temperature, northern Hemisphere.



3.3 Measures of Central Tendency

- Before further discussion of graphical descriptors of data, we need to discuss several quantitative descriptors. We are interested in two types of measures - those of *central tendency*, and those of *variability*.
- The term *central tendency* refers to the center of a distribution. The center is a mathematically ambiguous term. There are three mathematically precise terms: mode, median, and mean

Definitions

- The *mode* is the most commonly occurring value in a data set. There may be more than one mode if there are more than one value with the same maximum number of observations. For the weight guesses, there were 3 modes 160, 165 and 170 (each occurring 7 times in the data set)
- Both qualitative and quantitative variables have modes. The mode of gender was *male*, because there were 25 observations with the value *female*, and 26 observations with the value *male*
- The *median* is the middle value of the data when the values are arranged in order. Only quantitative variables have medians because only quantitative variables can be unambiguously ordered.
- The median is also called the second quartile, so I will use the symbol Q_2 to denote the median

Finding the Median of a collection of n Values

- Order the data from smallest to largest
- If n is odd, then the $(n + 1)/2$ largest value is the median. E.g., suppose that the data are $D = (3, 3, 4, 6, 9)$. Then, $n = 5$,

$$(n + 1)/2 = (5 + 1)/2 = 3,$$

and the median is third largest value, namely, $Q_2 = 4$

- If n is even, then there is no middle value (what is the middle value if $D = (4, 5)$?) For convenience, we will define the median to be the average of the $n/2$ and $n/2 + 1$ largest values. E.g., suppose that the data are $D = (3, 3, 7, 9)$. Then, $n = 4$,

$$n/2 = 2,$$

and the median is the average of the second and third values. Hence,

$$Q_2 = (3 + 7)/2 = 5.$$

Quartiles and Percentiles

- The p th percentile is value x such that $p\%$ of the data is less than or equal to x , and $100 - p$ values are larger
- For example, 95% of all values are less than the 95th percentile; 5% of values are larger than the 5th percentile
- The 25th percentile is called the first quartile (denoted by Q_1), the 50th percentile is called the second quartile (or the median, and denoted by Q_2) and the 75th percentile is called the third quartile (denoted by Q_3)
- There is no universal method of computing percentiles (or quartiles)
- One simple method of finding Q_1 and Q_3 is to find the median, split the data into a set consisting of all numbers **strictly** less than the median, and another consisting of all numbers **strictly** greater than the median
- Q_1 is the median of the first subset and Q_3 is the median of the second subset

Using SPSS to Compute Quartiles

- Open a data file
- Follow the following sequence of drop-down menus:

Analyze

Descriptive Statistics

Frequencies

- Choose one or more variables
- Open the Statistics window and select the statistics of interest. Hit OK

The Mean

- The *mean* of a collection of n values is the average of the values
- Notation: we generically denote a set of n values by x_1, x_2, \dots, x_n , and the sample average of these values by \bar{x} . The sum of these values is

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n,$$

and the mean of these values is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Ott and Longnecker also use the notation

$$\frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_i x_i$$

to denote the sum

- For example, if $D = (3, 3, 4, 6, 9)$, then $x_1 = 3, x_2 = 3, x_3 = 4, x_4 = 6$, and $x_5 = 9$, and

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{3 + 3 + \dots + 9}{5} = \frac{25}{5} = 5.$$

- We are often interested in *population* means. For example, we might like to know the mean length of time it takes a chimpanzee to learn the American sign language word for *listen*. Conceptually, we imagine a population of learning times associated with the collection of all young adult captive chimpanzees

- The Greek letter μ ("mu") is used as a generic symbol for a population mean.

- For example, using the chimp data set, an estimate of μ is $\bar{x} = (12 + 10 + 2 + 15)/4 = 10.25$

- If we compute a mean from a sample of data, then the mean is called a *sample mean*.

- Ott and Longnecker discuss computing means and medians from grouped data (p. 72 and 74). We will not study that material

The Mean Versus the Median

- The weakness of the sample mean is that observations do not have the same importance in determining the sample.
- Suppose that we have 2 data sets: $D_1 = (1, 2, 3)$ and $D_2 = (1, 2, 30)$. The corresponding sample means are 2 and 11. The difference between the two means is very large, namely, $|3 - 11| = 8$, and entirely attributable to the third values. The value 30 has a great deal of *leverage* towards determining \bar{x} . The value 30 is called an outlier
- An *outlier* is an extreme value; specifically, a value that is distant from the center of the data compared to majority of the data
- The median is *resistant* to outliers. For example, the medians for the data sets $D_1 = (1, 2, 3)$ and $D_2 = (1, 2, 30)$ are both 2
- The strength of the sample mean is that it is (usually) a more precise estimate of μ than the median

Homework Assignment for Friday, Sept. 15 (Subject to modification)

p. 95. - 3.44, 3.47
p. 100 - 3.50, 3.51

Trimmed Means

- The *trimmed mean* is a resistant version of the sample mean. It resists the effect of outliers
- A 5% trimmed mean is computed by deleting the largest 5% of the values and the smallest 5% of the values from the data set, and computing the sample average from all the others.
- 10% trimmed means are sometimes used, but should be viewed with skepticism because so much (20%) of the data is deleted
- Deleting observations decreases the accuracy of the estimate of μ , and may bias the estimate of μ

Skewness and Measures of Central Tendency

- Differences among the mode, median, mean, and trimmed mean as measures of central tendency are related to the degree of *skewness* in the distribution of values
- The mode is largely unaffected by skewness because the mode is a most common value
- The median is resistant to skewness because a long thin tail will have relatively few observations in it. Unless there are many observations in the tail, the tail does not really affect the median
- The trimmed mean is fairly resistant to skewness; the degree of resistance depends on the amount of trimming and the thickness of the tail
- The mean is most sensitive to skewness
- Ott and Longnecker (p. 76, Fig. 3.17) has a helpful figure showing these relationships

Physical Interpretations of Measures of Central Tendency

Suppose that the histogram of a data set is thin sheet of metal of uniform thickness and weight

- The mode is the point on the x -axis corresponding to the highest point on the histogram
- The median is the point on the x -axis that divides the histogram into two regions of equal surface area
- The mean is the point on the x -axis at which a fulcrum would perfectly balance the histogram. The center of mass lies directly above the mean

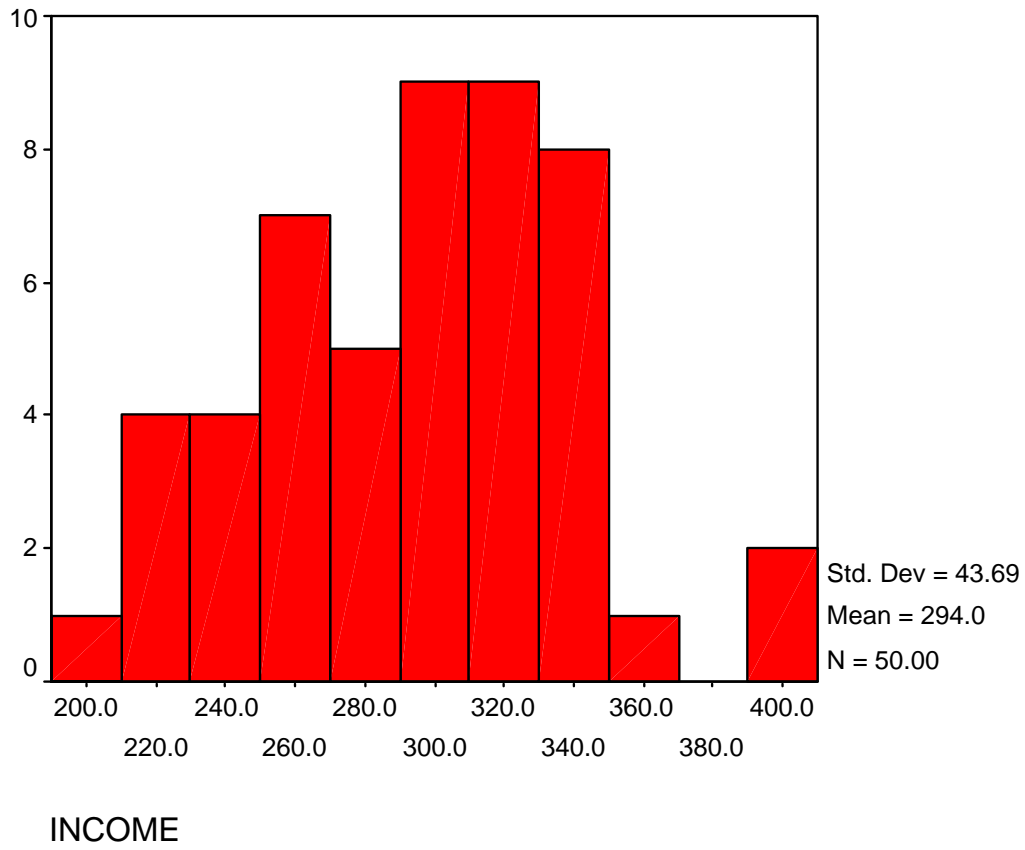
A Informal Introduction to the Boxplot

- A boxplot is an alternative to the histogram for showing the distribution of the sample data
- The lower and upper boundaries of the box are the first and third quartiles (Q_1 and Q_3)
- The median is identified by drawing line (necessarily within the box)
- The fences identify the upper and lower extent of the data, and outliers are shown as individual points
- We return to the SAT data to compare these methods

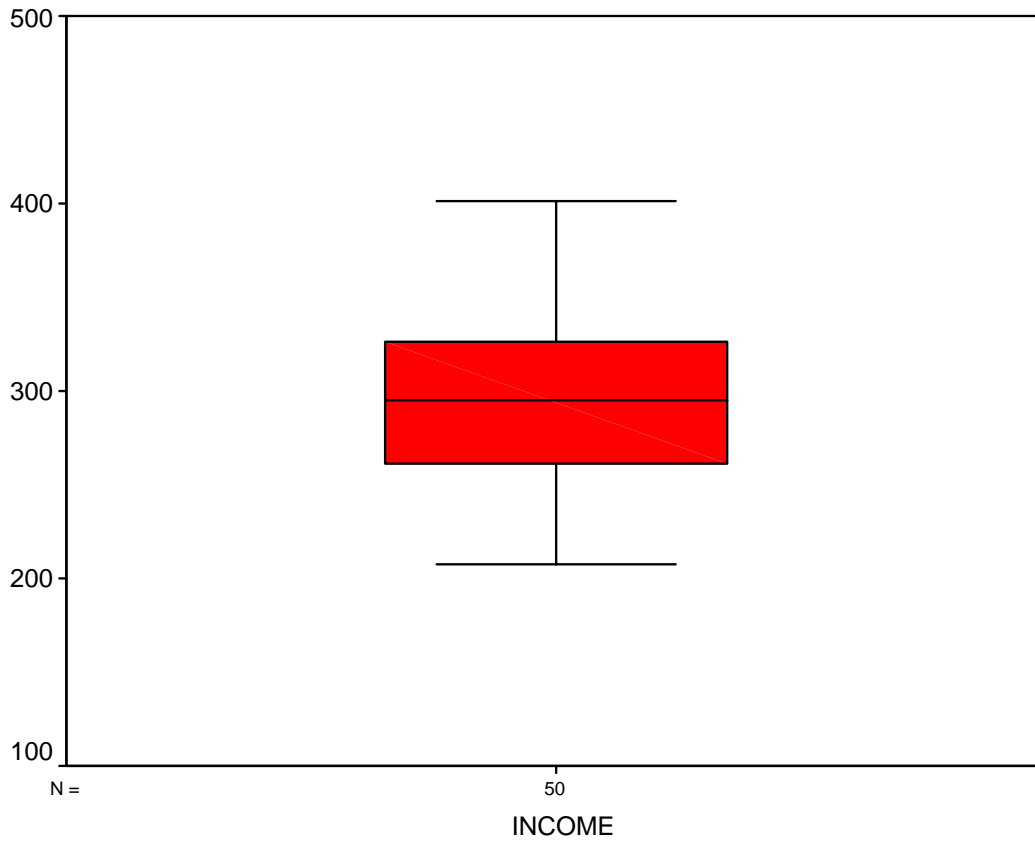
Case Study - State Average SAT Scores

- Citations: B. Powell and L.C. Steelman, 1984. "Variations in state SAT performance: Meaningful or Misleading?" *Harvard Educational Review* **54**(4), 389-412, and Ramsey, F.L. and Schafer, D.W., 1997, *The Statistical Sleuth*, Duxbury Press, p. 327.
- The variables are
 1. takers - (test-takers) percentage of total eligible students that took the exam
 2. income - median family income of the test-takers (hundreds of dollars)
 3. years - average number of years that the takers took formal courses in social sciences, humanities, and natural sciences
 4. public - percentage of takers that attended public secondary school
 5. expend - total state expenditure on secondary schools (hundreds of dollars per student)
 6. rank - median percentile ranking of the test takers among the students in their classes

- Histogram for the income variable

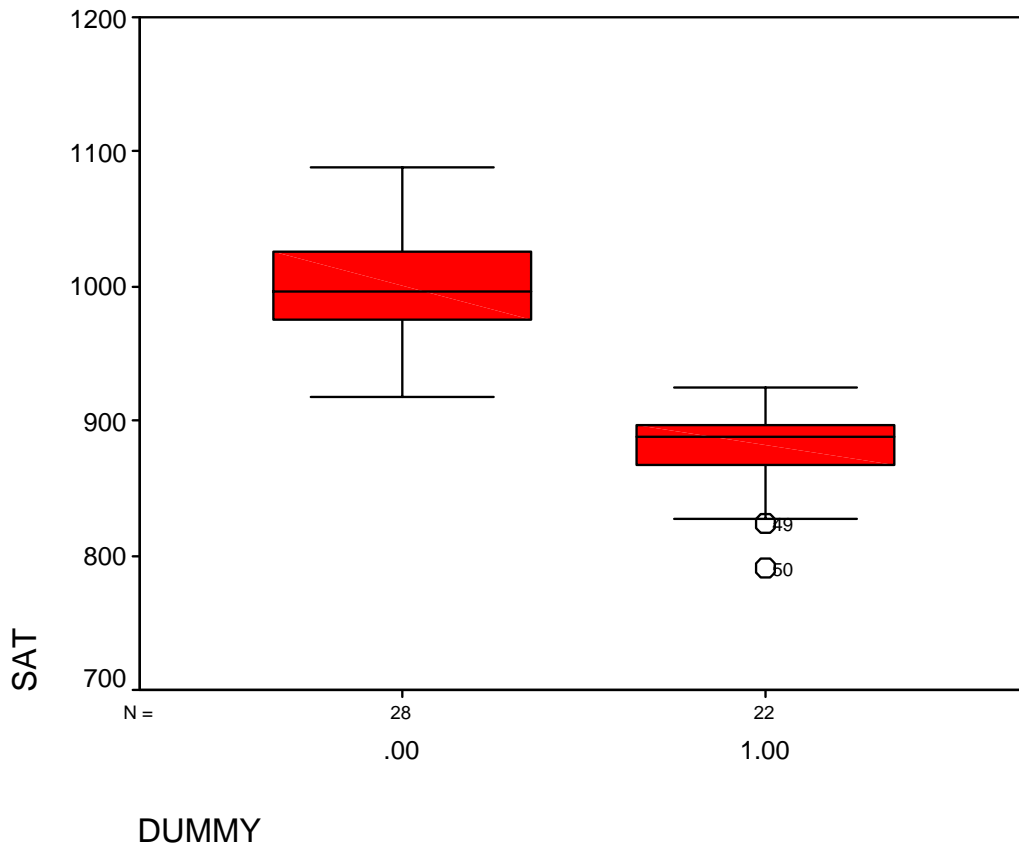


- Boxplot for the income variable



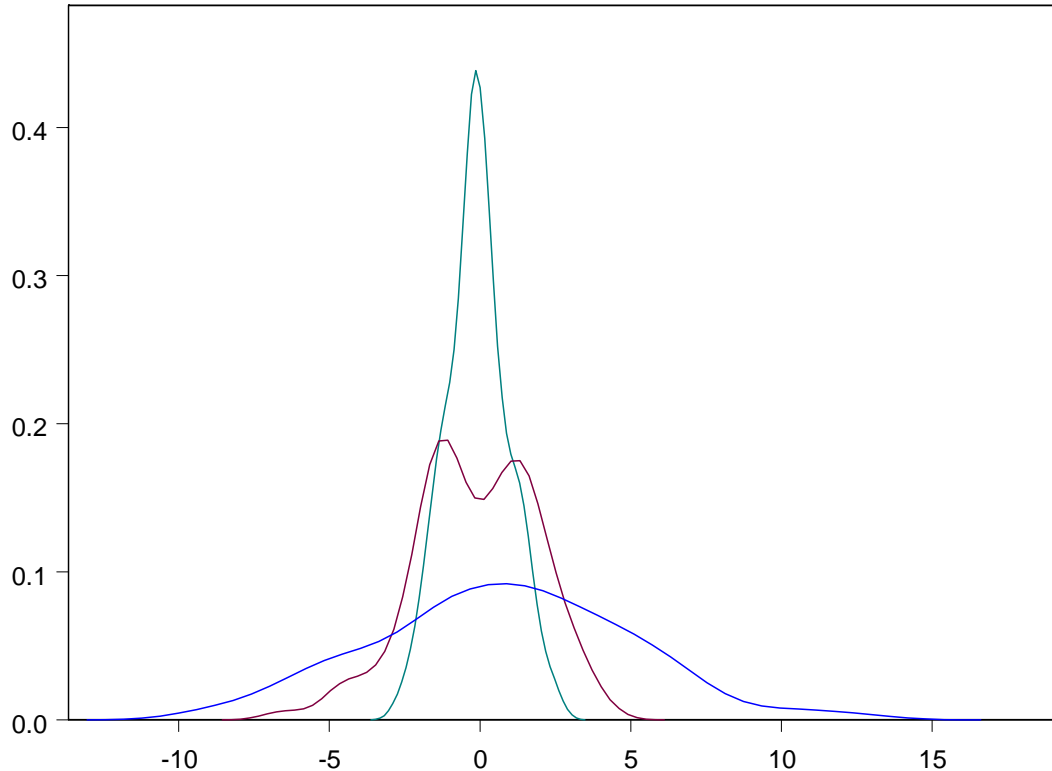
- Both figures effectively display the distribution of the data. The boxplot allows us to make more quantitative statements (e.g., the median income is about \$29,500)

- *Side-by-side* boxplots show separate boxplots for two or more subsets of the data. For example, I defined the variable *dummy* to be 0 if takers was less than 25, and 1 if takers was at least 25. Then, I produced a boxplot for those states with $dummy = 0$, and an another for those states with $dummy = 1$:



3.5 Measures of Variability

- Figure. Three smoothed histograms computed from three data sets of 100 observations each. In what sense are they different?



- We will limit the discussion of variability to quantitative variables
- There are three main measures of variability: 1. range, 2. percentiles and the IQR, and 3. standard deviation and variance

1. The *range* is the difference between the sample maximum and minimum values. It is a simple, but not very useful measure of variation because 2 values completely determine the range. Hence, it is extremely sensitive to outliers.

- For the **green distribution**, the range is $r = 2.48 - -2.61 = 5.09$; for the **blue distribution**, the range is $r = 12.4 - -8.86 = 21.00$

2. The *percentiles* of a distribution are much better. We define the percentile by stating:

$p\%$ of the values are less than or equal to the p th percentile

For example, the 25th percentile (also called Q_1), and the 75th percentile

(also called Q_3) are $Q_1 = -.74$ and $Q_3 = 0.46$ for the green distribution, and $Q_1 = -1.42$ and $Q_3 = 1.40$ for the red distribution

- The *interquartile range* (IQR) is the difference between Q_3 and Q_1 :

$$\text{IQR} = Q_3 - Q_1$$

For example, the IQR for the green distribution is

$$\text{IQR} = .46 - (-.74) = 1.20;$$

for the red distribution, the IQR is

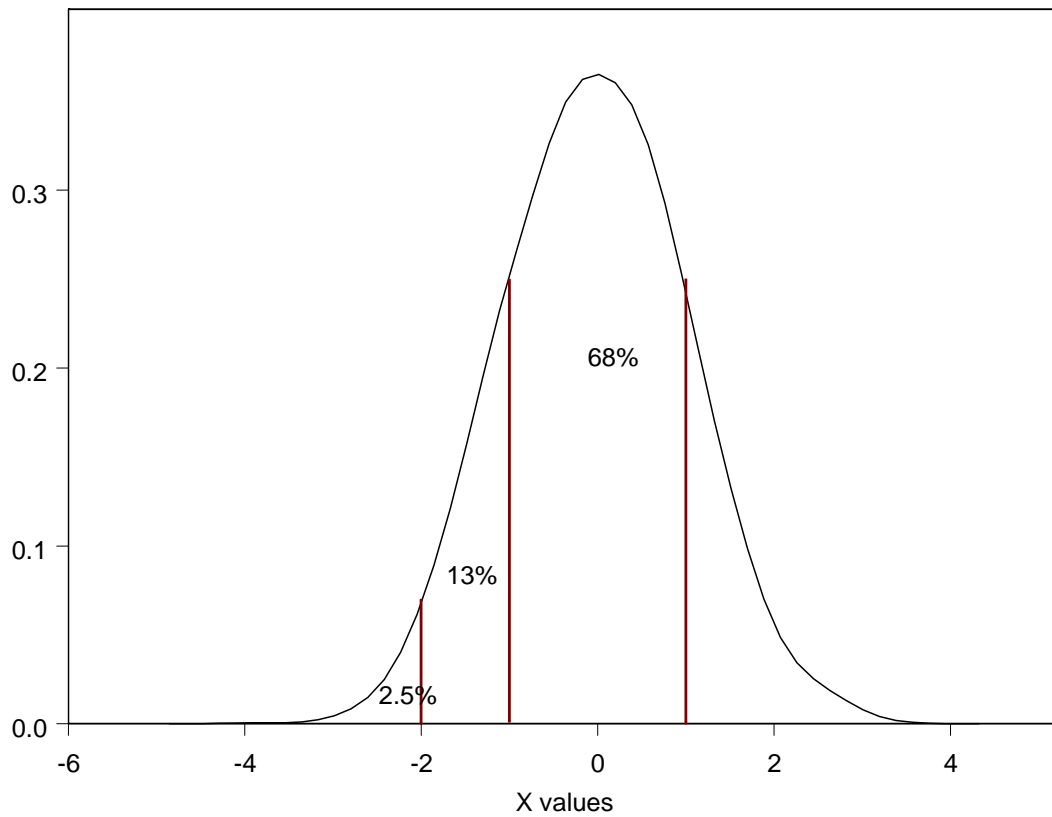
$$\text{IQR} = 1.40 - (-1.42) = 2.82$$

- The IQR is the distance about the median needed to contain the middle 50% of the data

The Variance and Standard Deviation

- The standard deviation is similar to the IQR. For the normal distribution, one standard deviation is the distance about μ that contains the middle two-thirds of the distribution

Figure. The *standard normal* distribution showing the proportion of values



- Approximately 68% of the area under the standard normal curve is between -1 and 1 . Said another way, approximately 68% of the area under the curve is within 1 standard deviation of the mean $\mu = 0$
- If a distribution is normal, and the standard deviation is 2, then approximately 68% of the area under the curve is between -2 and 2
- We can calculate the standard deviation for other distributions, but this rule usually will not be very accurate. For example, if a distribution is uniform, the approximately 57% of the data is within 1 standard deviation of the mean

A Mathematical View of the Standard Deviation

- A good way to measure variability is with respect to a point, and the center of the distribution is the best reference point
- A good definition of the variability of a set of data is *the average distance of the observations to the mean*
- The distance of the i th observation y_i to the sample mean \bar{y} is

$$d_i = |y_i - \bar{y}|$$

- The average of these n distances is a measure of variability:

$$v_2 = \frac{1}{n} \sum_{i=1}^n d_i = \frac{1}{n} \sum_{i=1}^n |y_i - \bar{y}|$$

- This is called the city-block distance between the data $\{y_1, \dots, y_n\}$ and \bar{y}
- Alternatively, we could use the Euclidean distance from the data y_1, \dots, y_n to \bar{y} :

$$v_3 = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$$

- v_3 is *almost* the standard deviation
- The *sample standard deviation* is

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

- s is measured in the same units as the original data
- s is the most commonly used estimator of the *population standard deviation*, denoted by σ

- The sample variance is sometimes used instead of s to describe variability. It is denoted by s^2 , and can be computed by squaring s . However, when calculations are done by hand, s^2 is computed first, then s . That is, first we compute

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2;$$

then we compute

$$s = \sqrt{s^2}$$

- s^2 is measured in square units. For example, for the weight data, $s^2 = 176.9 \text{ lbs}^2$

The Coefficient of Variation

- Occasionally, the coefficient of variation is a useful measure of variation. If the CV is computed from a sample, then it is defined to be

$$CV = \frac{s}{|\bar{y}|}$$

- The idea is to measure variation while taking into account the size of the values. Size is measured by $|\bar{y}|$.
- The CV may be badly misleading. If all measurements are between 1000 and 1100, then it is convenient to code the data using only the first two digits (e.g., 1023 is replaced by 23). The coefficient of variations for the actual and coded data will be grossly different because the sample means will differ by 1000, but the coded and actual data have the same standard deviation
- Ott and Longnecker discuss the coefficient of variation on page 93, and provide an example. We will not use the coefficient of variation.

The Empirical Rule

Suppose that a population of values has a normal distribution. Then,

- Approximately 68% of all values lie within one standard deviation of the population mean μ
- Approximately 95% of all values lie within two standard deviations of the population mean μ
- Approximately 99% of all values lie within three standard deviations of the population mean μ
- The *empirical rule* uses these properties of the normal distribution. It can be stated as follows:
 - *If* we have a large sample with a histogram that looks approximately normal in shape, and its sample mean and standard deviation are \bar{y} and s , then

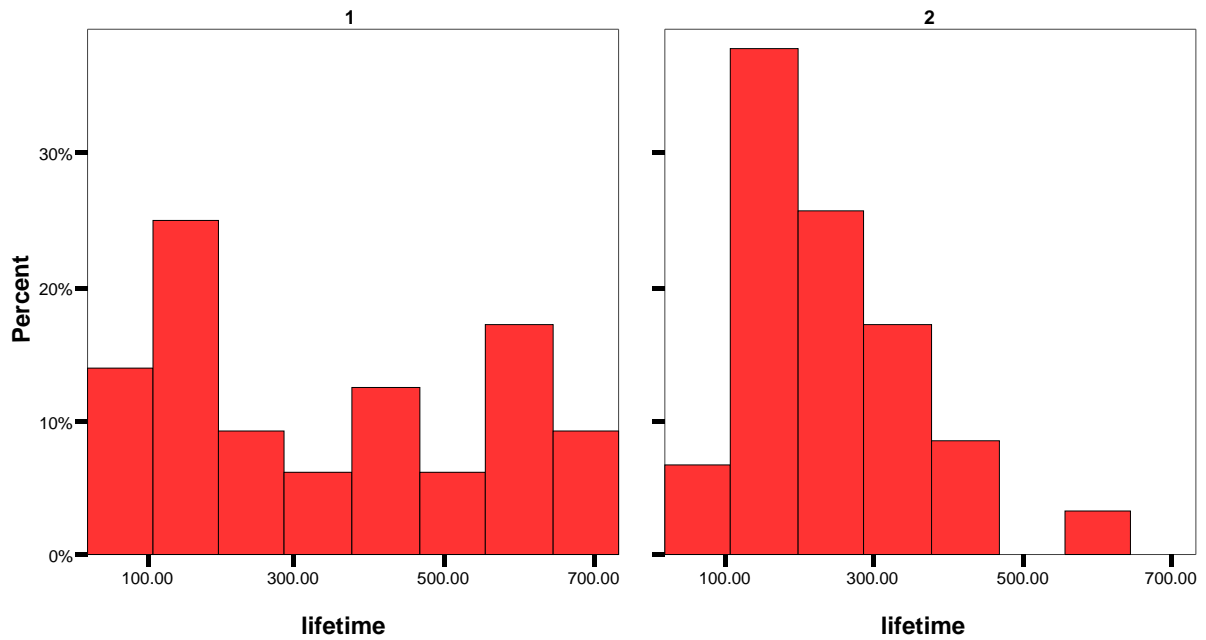
1. the interval $[\bar{y} - s, \bar{y} + s]$ contains approximately 68% of the data values
2. the interval $[\bar{y} - 2s, \bar{y} + 2s]$ contains approximately 95% of the data values
3. the interval $[\bar{y} - 3s, \bar{y} + 3s]$ contains approximately 99% of the data values

Example The sample mean and standard deviation of the state average SAT scores are $\bar{y} = 947.9$ and $s = 70.8$, and the distribution of values is approximately normal in shape. Therefore,

1. The empirical rule states that approximately 68% of the 50 states (34 states) have averages between $947.9 - 70.8 = 877.1$ and $947.9 + 70.8 = 1018.7$.
 - The actual number of states in this interval is 32.
2. The empirical rule states that approximately 95% of the 50 states (47.5 states) have averages between $947.9 - 2 \times 70.8 = 806.3$ and $947.9 + 2 \times 70.8 = 1089.5$.
 - The actual number is 49.

Case Study From Doksum, K. 1974. "Empirical probability plots and statistical inference for nonlinear models in the two-sample case," *Annals of Statistics*, 2, 267-77.

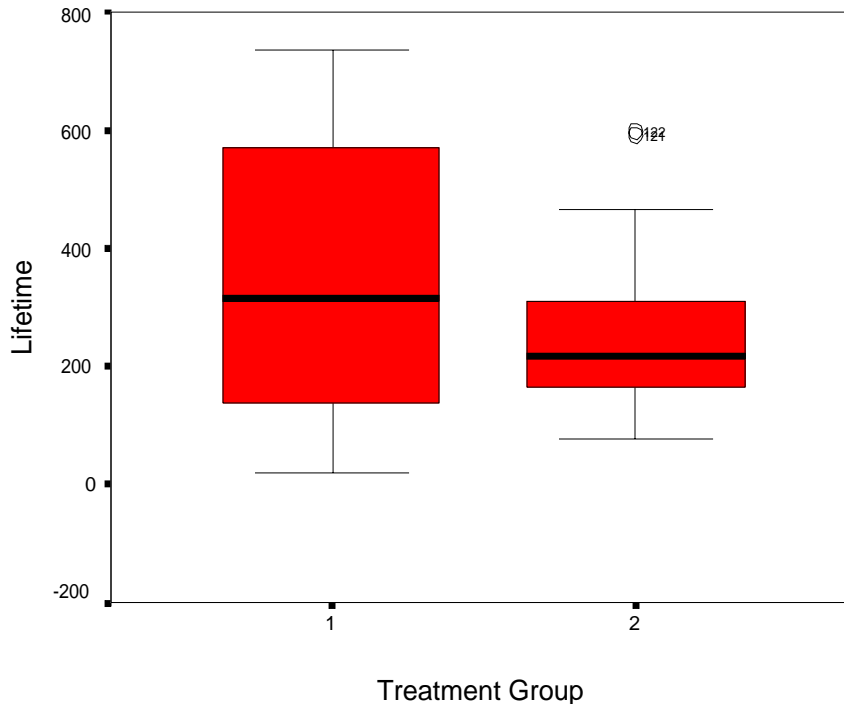
- The data in the figure below are survival times (days) of guinea pigs that were randomly assigned to either a control group or a treatment group that received a dose of turbercle bacilli.
- The left histogram is constructed the control observations ($n = 64$); the right histogram is constructed from the bacilli observations ($n = 58$).



- Is there a meaningful difference in lifetimes between treatment groups?

Box Plots represent the middle 50% of the data by a box and identify important features of the upper and lower 25% of the data.

- Another view of the guinea pig data:



- There is more variability in the **control** group
- There is some visual evidence that lifetimes tend to be longer for the control group, but the visual evidence is not convincing

- **Details on the Box Plot**

1. The lower fence (or whisker) is located at the smallest data value that is larger than $Q_1 - 1.5 \times \text{IQR}$.
2. The upper fence (or whisker) is located at the largest data value that is less than $Q_3 + 1.5 \times \text{IQR}$

- For example, for Treatment group 1, $\text{IQR} = 447$ and

$$Q_3 = \frac{569 + 576}{2} = 572.5.$$

Then,

$$Q_3 + 1.5 \times \text{IQR} = 572.5 + 1.5 \times 447 = 1243.$$

The largest observation that is less than 1243 is 735. So the fence is drawn at 735.

3. Observations that are larger than the upper fence are denoted by a symbol, and sometimes the record (or row) number; observations that are smaller than the lower fence are similarly noted

4. A *mild outlier* is an observation that is between $Q_1 - 3 \times \text{IQR}$ and $Q_1 - 1.5 \times \text{IQR}$ or between $Q_3 + 1.5 \times \text{IQR}$ and $Q_3 + 3 \times \text{IQR}$

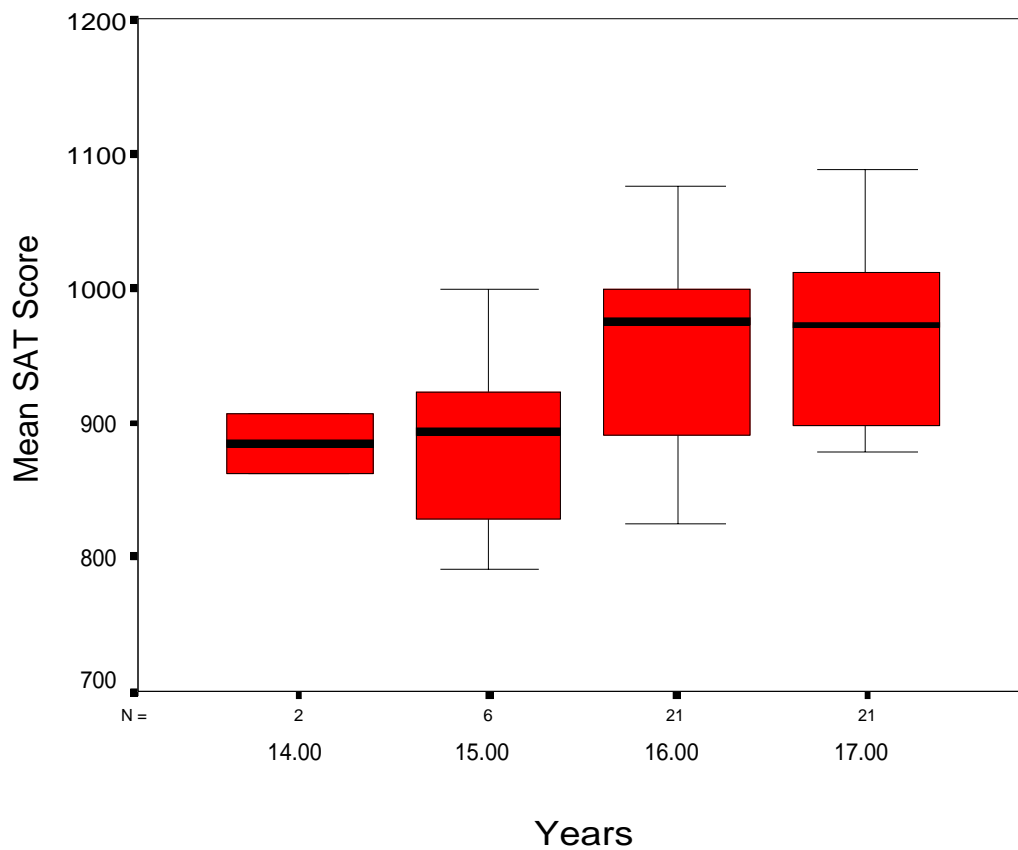
5. A *severe outlier* is an observation that is smaller than $Q_1 - 3 \times \text{IQR}$ or larger than $Q_3 + 3 \times \text{IQR}$

6. Ott and Longnecker have a detailed discussion on boxplot construction in 3.6 (p. 96-100)

Examples

1. The SAT dataset. To what extent is there an association between the state average number of years that the takers took formal courses in social sciences, humanities, and natural sciences and the state average SAT score?

- One way to address this question is to round the state average number of years down to the nearest integer and construct side-by-side boxplots.

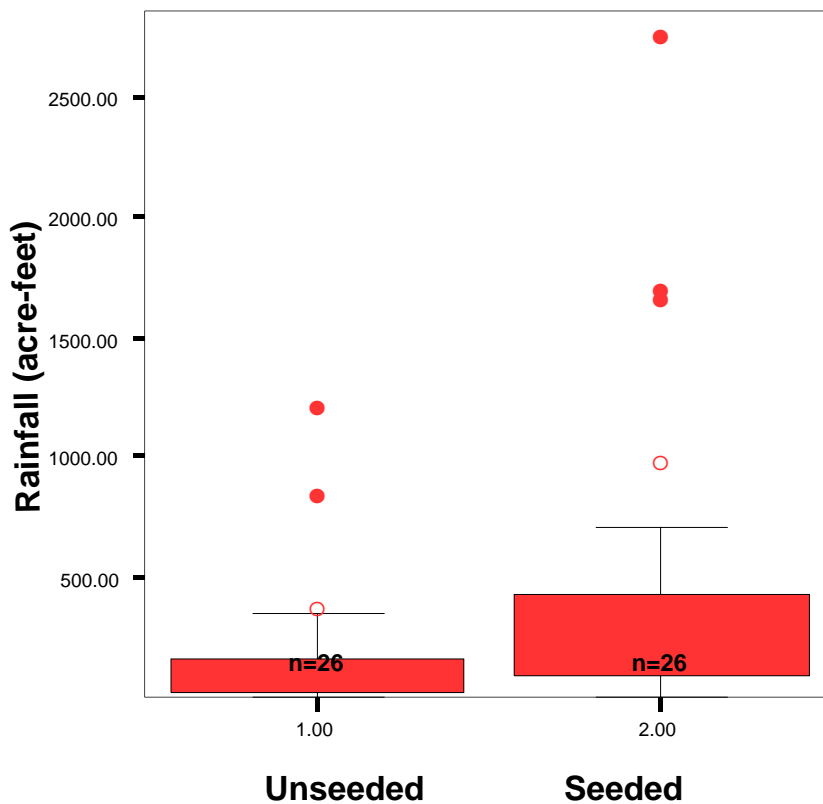


- The first box plot is based on 2 observations, so it may as well be ignored
- The first and second sample distributions (collectively, years 14 to 15.99) appears to be shifted downwards compared to those for years 16 and greater
- In conclusion, these data present some evidence of a difference in state averages associated with years of formal coursework.

- We might have a higher level on confidence in our conclusions if other factors are affecting the relationship (e.g., percent takers) have been taken into account and controlled

Case Study Simpson, J. Olsen, A., and Eden, J. 1975. "A Bayesian analysis of a multiplicative treatment effect in weather modification," *Technometrics*, **17**, 161-166, and Ramsey, F.L. and Schafer, D.W., 1997, *The Statistical Sleuth*, Duxbury Press, p. 54.

- On each of 52 days deemed suitable for cloud seeding, a random mechanism was used to determine whether to seed a target cloud or reserve the day as a control. An airplane flew through the cloud regardless, and the amount of rainfall falling from the cloud was measured by radar. The data are summarized in the following boxplot.

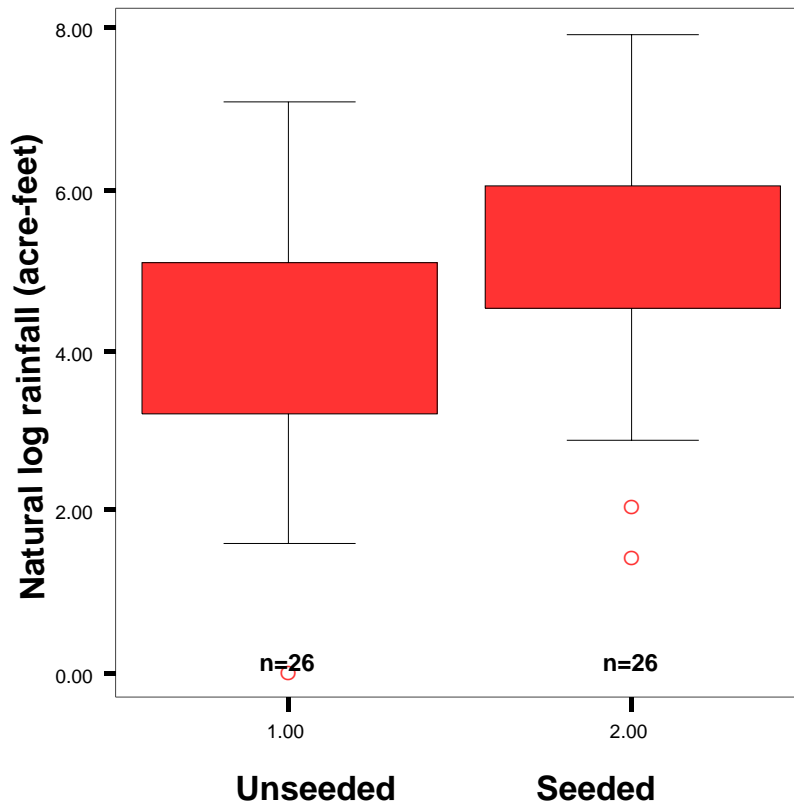


- What distributional feature is common in both sets of data?

- Skewness sometimes can be removed by a nonlinear transformation. For example, rainfall amounts can be transformed via the natural logarithm. If r_i is the rainfall on the i th day, then,

$$l_i = \ln(r_i) = \log_e(r_i), i = 1, \dots, n$$

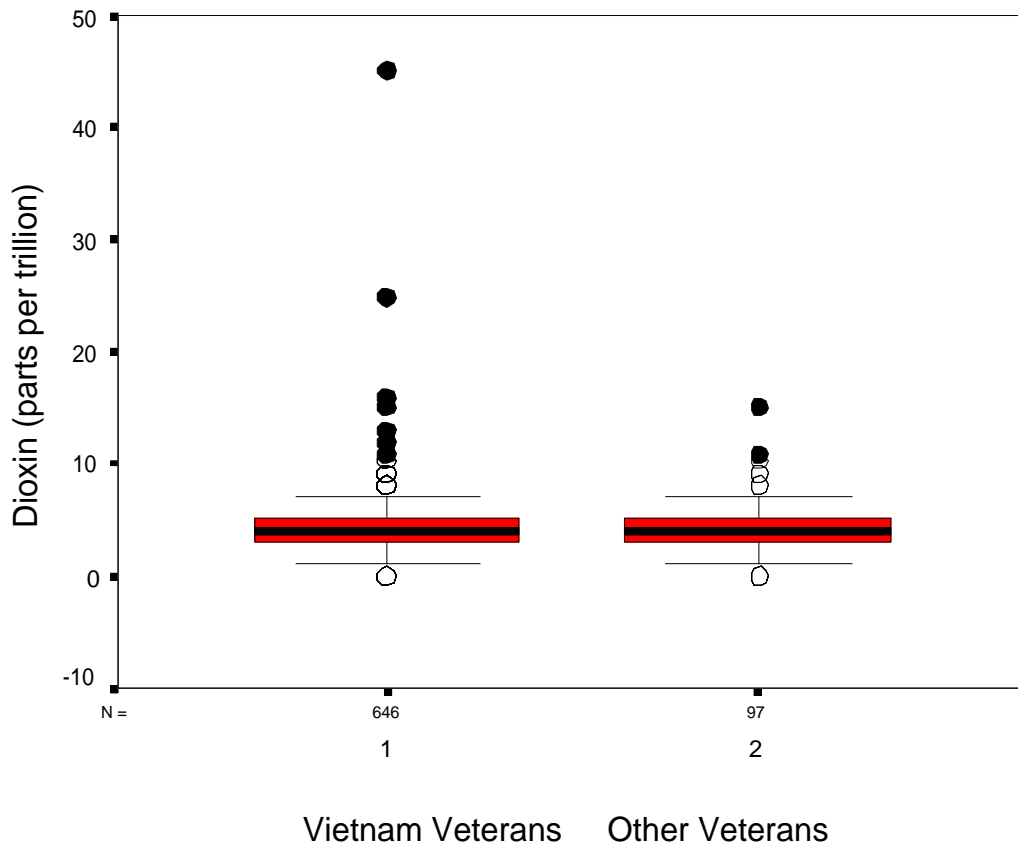
are the log-transformed data. These data are summarized below



- The log-rainfall medians are located roughly in the centers of the boxes. For the unseeded group, the median is 3.79 log-acre-feet; for the seeded group, the median is 5.40 log-acre-feet
- Converting back from natural logs to the original scale yields 44.2 and 221.6 acre-feet (same as the medians on the original scale). Seeding results in an estimated increase of $221.6/44.2 = 5$ times as much rainfall as not seeding. A formal test shows that there is strong evidence that seeding increases the amount of rainfall. Causation is attributable to seeding because this was a controlled experimentation that used randomization

Case Study Centers for Disease Control Veterans Health Studies. 1988. "Serum 2,3,7,8 tetrachlorodibenzo-*p*-dioxin levels in U.S. Army Vietnam-era veterans," *Journal of the American Medical Association*, **260**, 1249-54, and Ramsey, F.L. and Schafer, D.W., 1997, *The Statistical Sleuth*, Duxbury Press, p. 57.

- There is concern that exposure to Agent Orange (a defoliant) during the Vietnam war has affected the health of Vietnam veterans. One particularly dangerous component of Agent Orange is TCDD dioxin, known to be associated with certain cancers. TCDD dioxin has been detected in the blood 20 or more years after exposure. Researchers from the Centers for Disease Control compared 1987 blood levels of 646 Vietnam veterans and 97 veterans who served during the Vietnam era, but not in Vietnam.
- What kind of study is this? Is it possible to conclude that exposure to Agent Orange is responsible for increased dioxin levels?
- The observed levels of TCDD dioxin are displayed below



- Describe and compare the distributions.

3.7 Summarizing Data From More Than One Variable

- The general objective is to assess the pattern of variation and the association between the variables.
- There are three cases to consider:
 1. All variables are qualitative
 2. All variables are quantitative
 3. Some are qualitative and some are quantitative
- For simplicity, we will limit the discussion to comparing 2 variables

1. Qualitative vs. Qualitative

Example Anderson, T.W., Reid, D.B., and Beaton, G.H. 1972. "Vitamin C and the Common Cold," *Canadian Medical Assoc. Journal.* **107**, 503-508.

- 818 volunteers were randomly divided into two treatment groups
- The *vitamin C* group received a supply of pills containing 1000 mg of Vitamin C (enough to last the entire cold season)
- The *placebo* group received a similar supply of pills containing 0 mg of Vitamin C
- None of the subjects knew whether their pills contained Vitamin C. This is done to eliminate any possible psychological effects (the *placebo* effect)
- At end of season, each subject was interviewed by a doctor who determined whether the subject had suffered a cold. The doctor did not know which group the subject was assigned to. This was done to eliminate any possible biases related to the doctor's assessment
- The experiment is called *double-blind* because neither patient nor doctor knew which individuals were assigned to the Vitamin C group.
- There are two variables: measured on each subject, and both variables have two possible values
 1. Treatment group (Vit. C or placebo)
 2. Response (colds or no colds)

- The main question of interest is whether there is an association between treatments and response. That is, were there a greater percentage of colds in the placebo group, and a lesser percentage in the Vitamin C group?

Contingency Tables

- The rows of the table correspond to the categories (or levels) of one variable, and the columns correspond to the categories of the other variable
- The tabled values are the number of observations falling into each combination of categories, or *cells*.
- A contingency table of the counts is

	Outcome		Total
	Cold	No Cold	
Placebo	335	76	411
Vitamin C	302	105	407
Total	637	181	818

- A more informative table shows the percentages individuals in the two groups that got colds, i.e.,

among the placebo group, $100 \times 335/411 = 81\%$ got colds,

among the Vitamin C group, $100 \times 302/407 = 74\%$ got colds

among all subjects, $100 \times 637/818 = 78\%$ got colds

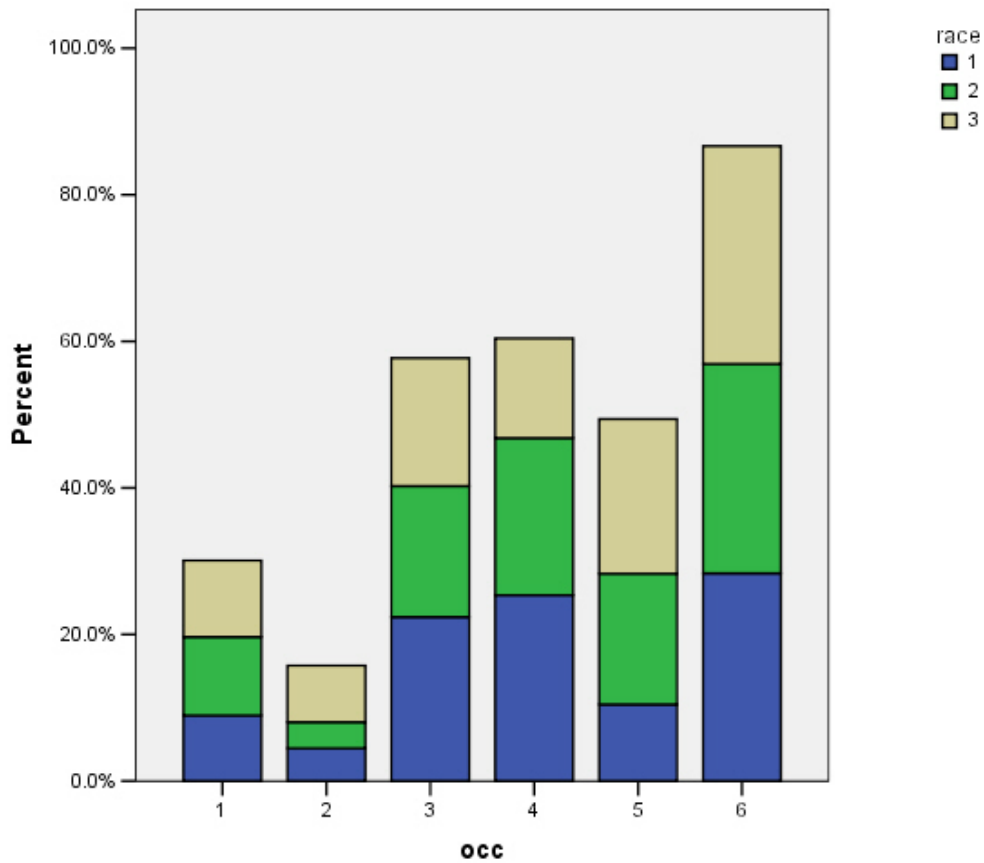
- A contingency table of the row percentages is

	Outcome		Total
	Cold	No Cold	
Placebo	81	19	100 ($n = 411$)
Vitamin C	74	26	100 ($n = 407$)

- There is relatively little association between the incidence of colds and consumption of Vitamin C because the differences in row percentages are not very large

Wages from the 1985 Current Population Survey

- Are there differences in the distribution of occupations among the races? To address this question in SPSS, Use Analyze, Descriptive Statistics, Crosstabs (cross-tabulation).
- Set the row variable to be race, and the column variable to be occupation. Then, change the options under *Cells* to get row percentages
- A bar chart of these data can be produced as well:



Case Study: The Donner Party (from Gayson, D.K., 1990, "Donner Party deaths: A demographic assessment," *Journal of Anthropological Research*, **46**, 223-42, and Ramsey, F.L. and Schafer, D.W., 1997, *The Statistical Sleuth*, Duxbury Press, p. 565.

- In 1846, the Donner and Reed families left Illinois for California by covered wagon (87 people, 20 wagons). They attempted a new and untried crossing of the region between Ft. Bridger, Wyoming and the Sacramento Valley. After numerous problems and delays in Utah, they reached the Eastern Sierra Nevada in late October. They were stranded near Lake Tahoe by a series of snowstorms that left as much as 8 feet of snow by some accounts. By the time they were rescued in April of the following year, 40 members had died. Some (or perhaps all) of those that survived did so by resorting to cannibalism
- The researchers attempted to address questions such as whether females are better able to withstand harsh conditions than men, and whether the odds of survival varied with age. Grayson was able to reconstruct records on survival, age and gender for 45 individuals.

Table 1. Survival by gender.

		Gender		Total	
		Females	Males		
Survival	No	Count	5	20	25
		% within gender	33.3	66.7	55.6
	Yes	Count	10	10	20
		% within gender	66.7	33.3	44.4
Total		Count	15	30	45

Table 2. Survival by age group, and gender within age group. The age groups are 31 years and younger, and older than 31.

Age Group	Survival		Gender		Total
			Females	Males	
≤ 31	No	Count	1	14	15
		% within gender	12.5	66.7	51.7
	Yes	Count	7	7	14
		% within gender	87.5	33.3	48.3
Total		Count	8	21	29

Age Group	Survival		Gender		Total
			Females	Males	
> 31	No	Count	4	6	10
		% within gender	57.1	66.7	62.5
	Yes	Count	3	3	6
		% within gender	42.9	33.3	37.5
Total		Count	7	9	16

- These data indicate that males had a greater mortality rate which was apparently not related to age. Female mortality rates were greater for the older women than the young.
- In SPSS, the *column* variable is sex, the *row* variable is survival, and the *layer* variable is ageclass. *Column* percentages were computed

2. Qualitative vs. Quantitative

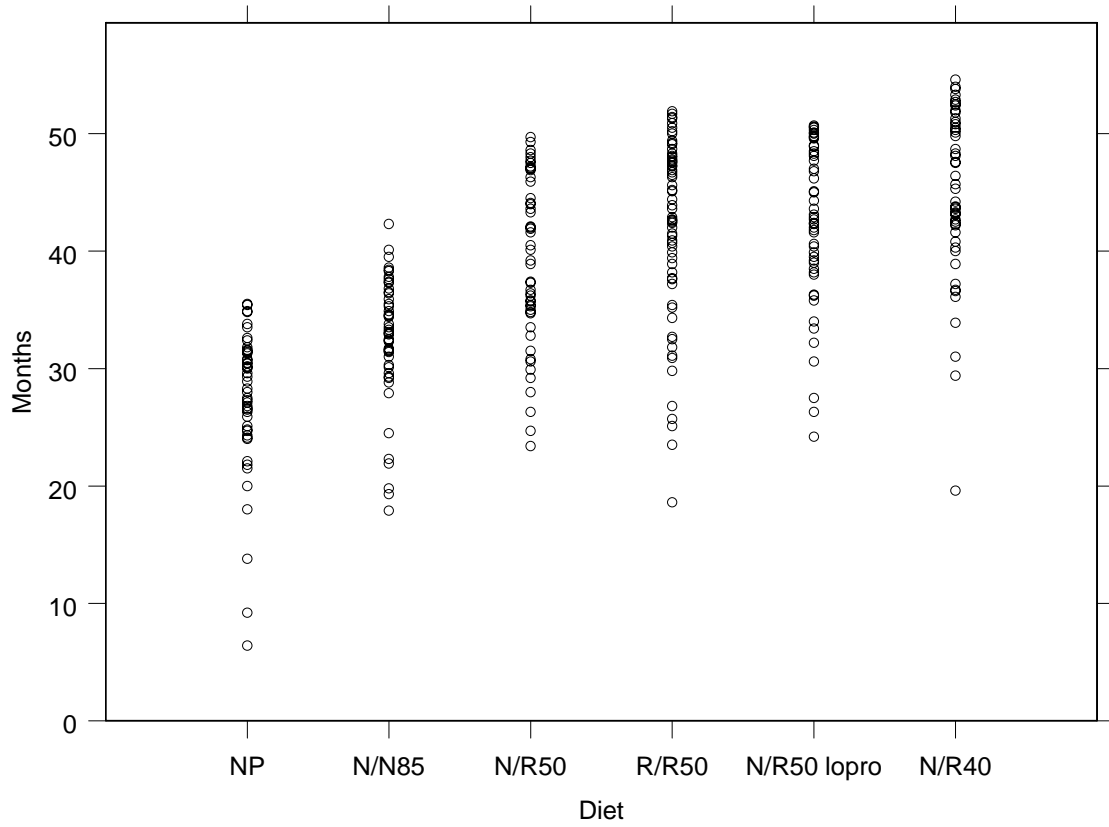
For two variables, a convenient summary is a plot with the qualitative variable plotted on the x-axis and the quantitative variable plotted on the y-axis.

Example Weindruch, R. et al. (1986). "The retardation of aging in mice by dietary restriction: longevity, cancer, immunity, and lifetime energy intake," *Journal of Nutrition*, **116**, 641-54, and Ramsey, F.L. and Schafer, D.W., 1997, *The Statistical Sleuth*, Duxbury Press, p. 109.

- A question of interest was whether restricting caloric intake can increase life expectancy. These researchers addressed this question by randomly assigning mice to six diets that varied with respect to calories and protein, and the time at which calorie reductions were imposed.

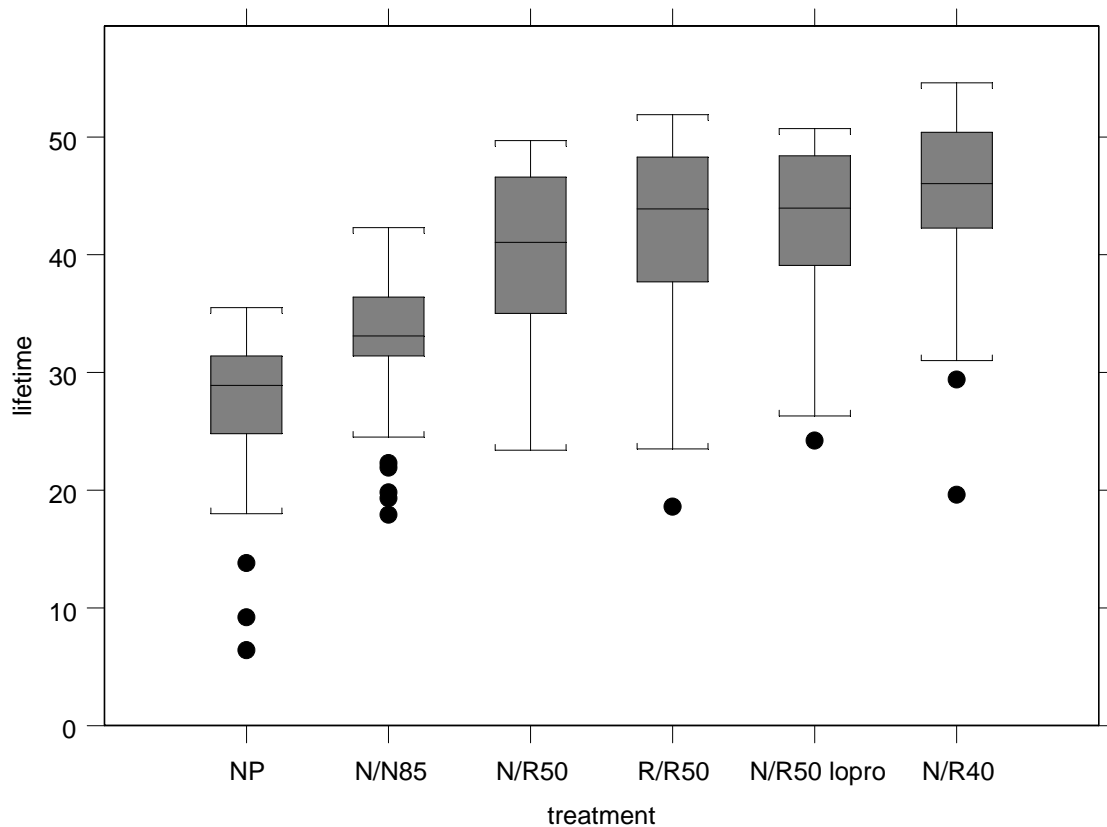
- The diets were
 1. NP: nonpurified and unlimited food
 2. N/N85: normal before weaning; normal food but limited to 85 Kcal/week after weaning
 3. N/R50: normal before weaning; normal but reduced to 50 Kcal/week after weaning
 4. R/R50: reduced 50 Kcal/week before weaning; normal food but reduced to 50 Kcal/week after weaning
 5. N/R50 lopro: normal before weaning; normal food but reduced to 50 Kcal/week after weaning and protein reduced with advancing age
 6. N/R40: normal before weaning; normal but reduced to 40 Kcal/week after weaning
- There are two variables of interest: lifetime (quantitative), and diet (qualitative with 6 levels)
- A plot of lifetime against diet is useful to display the association between diet and lifetime.

Figure. Lifetimes of female mice fed 6 different diets.



- Note that the diets have been ordered to best show the differences among diets with respect to lifetime
- A set of side-by-side box plots is very informative

Figure. Boxplots of lifetimes of female mice fed 6 different diets.

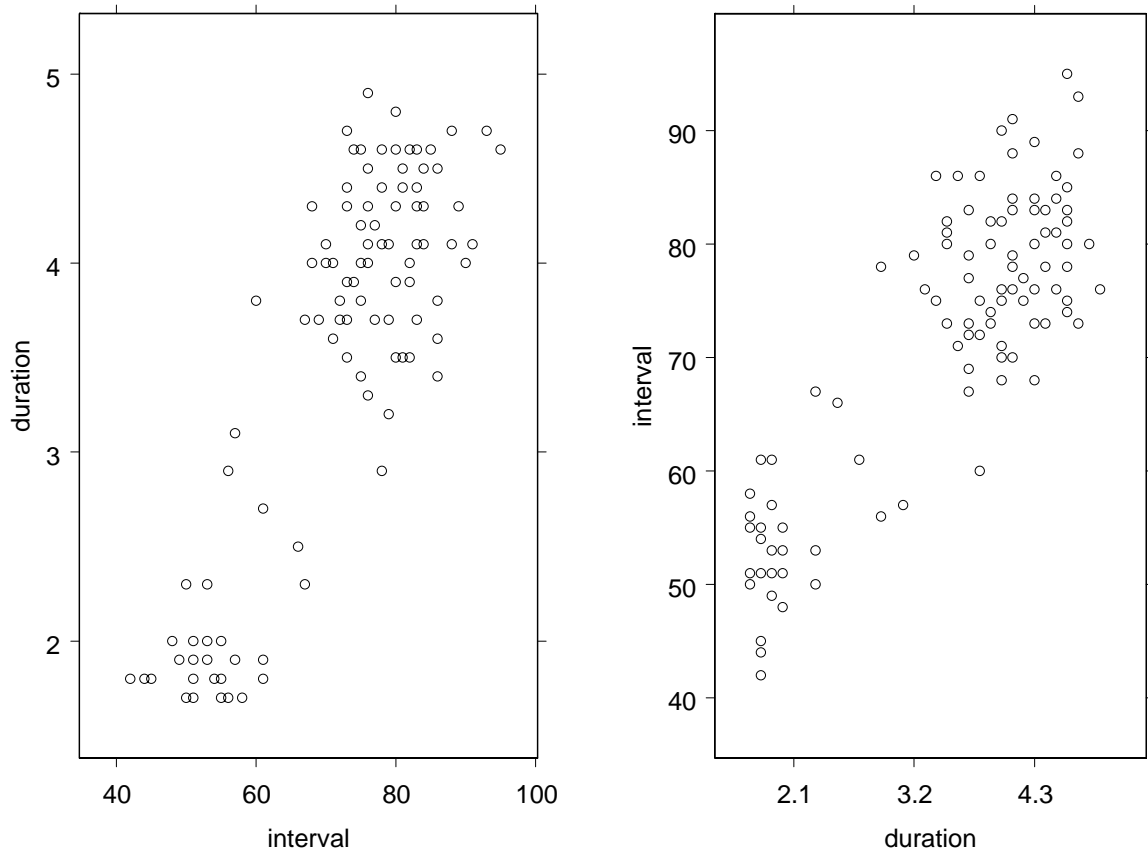


3. Quantitative versus Quantitative

- Scatterplots are commonly used to display the relationship between two quantitative variables.
- Two variables, or attributes, are measured on the sample (or population) units.
- The data thought of as a set of pairs, e.g., (Var 1, Var 2). One variable is assigned to the x-axis, and the second to the y-axis. Then, the pairs are plotted in the Cartesian plane.
- In talking about scatterplots, the convention is to name the *y*-axis variable first, then the *x*-axis variable

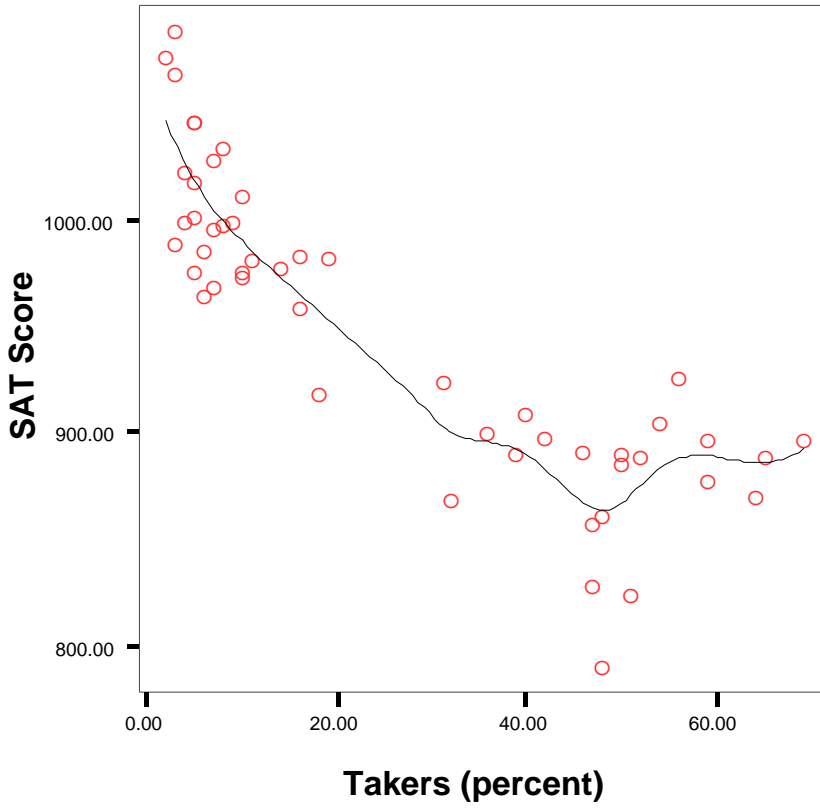
Example A popular recreation in Yellowstone NP is to monitor the time between, and the length of, Old Faithful eruptions. Some of these data are plotted below.

Figure. Time between, and length of, Old Faithful eruptions from Aug.1 to Aug. 8, 1978. The left panel is a plot of duration against interval, and the right panel is a plot of interval against duration.



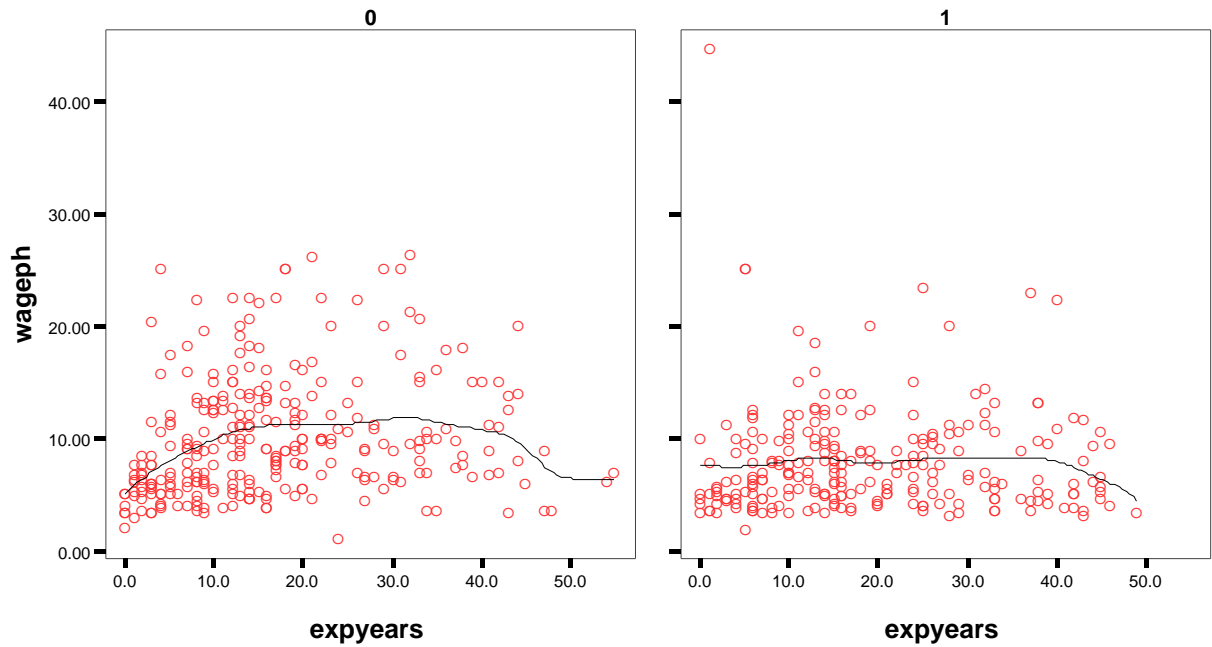
- Sometimes, it is helpful to impose a *smooth* on a scatterplot. The smooth provides a fairly flexible summarization of the relationship and trends in the data.
- For example, the smooth below helps to show that the SAT scores vary in an informative way with takers (the percent of all eligible students that took the SAT) only over the range of takers from about 0 to 40%.

Figure. State mean SAT score plotted against takers.



- Visual display of the relationship between 3 or variables is difficult; visual display of the relationship between more than 3 is even more difficult

- A plot of wage per hour versus experience, by sex is easier to interpret with a smooth. This can be obtained in SPSS through the Graphs, Interactive, Scatterplot sequence of menus:



- What can be said about experience as a predictor of wages?

Review Problems

p. 15, 1.1, 1.5

p. 26, 2.9

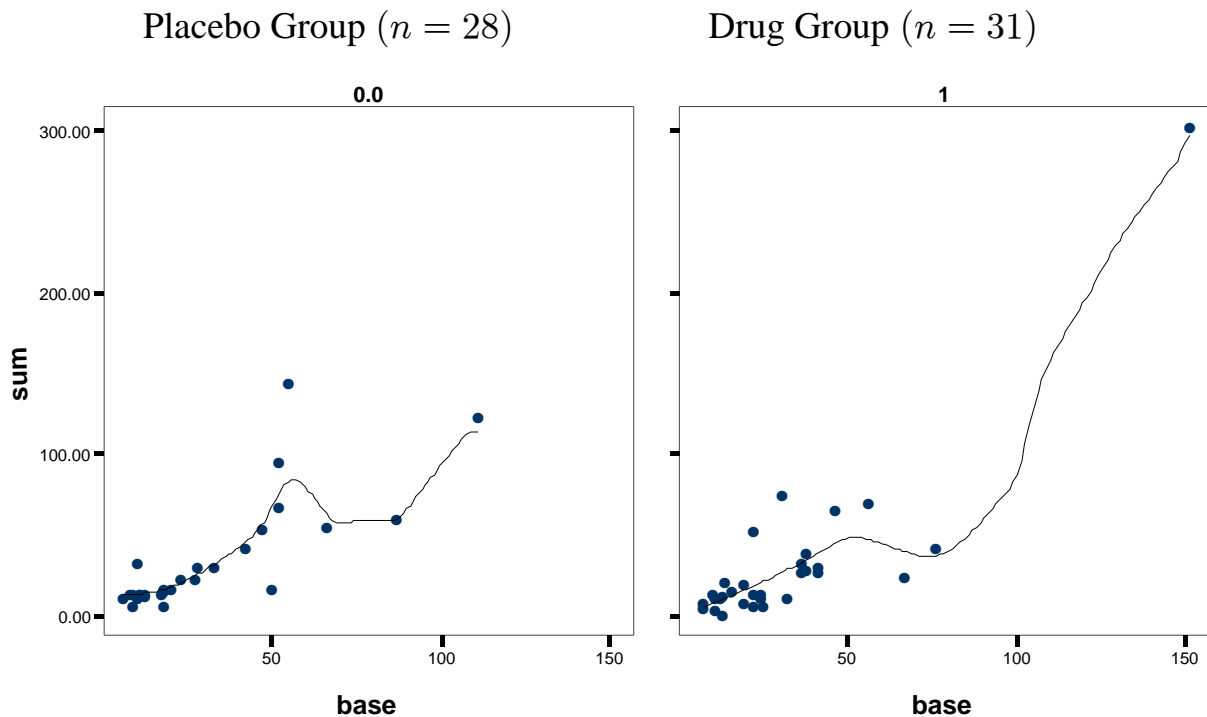
p. 115, 3.79, 3.80

Case Study Thall, P.F. and Vail, S.C. 1990. *Some covariance models for longitudinal count data with overdispersion*. *Biometrics*, **46**, 657-671. These are data from a study of the effectiveness of a drug for suppressing seizures in epileptics. 59 subjects were randomly assigned to treatment and placebo groups, and each subject reported the numbers of seizures experiences in each of 4 consecutive two-week periods. Also recorded was age and baseline seizure count. The baseline seizure count was the average number of seizures per two weeks, estimated by the patient and doctor.

- The total number of seizures was computed (called *sum* in Figure 1), and used for analysis.

a) Based on Figure 1, is there a association between the baseline count and the total number of seizures? Why or why not?

Figure 1. Scatterplots of total (sum) against baseline seizure count, by treatment group.



2) Based on the summary statistics and boxplots below,

a) Describe the two sample distributions

b) Is there evidence of differences between treatment groups with respect to the total number of seizures? Why or why not?

c) What variables might explain the differences between groups? What might be done to the data to help clarify the situation?

Figure 2. Boxplots of total, by treatment group.

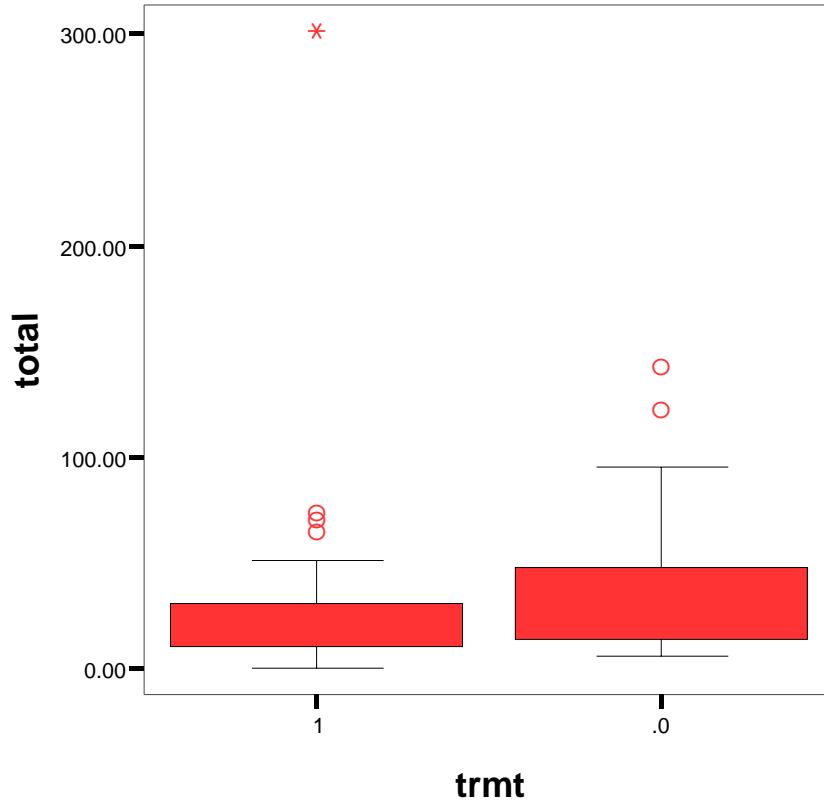


Table 1. Summary Statistics for total, by treatment group

	Placebo	Drug
Min:	6.00000	0.00000
1st Qu.:	13.75000	10.00000
Mean:	34.39286	31.83871
Median:	16.00000	15.00000
3rd Qu.:	44.75000	30.50000
Max:	143.00000	302.00000
Total N:	28.00000	31.00000
Std Dev.:	35.13291	53.88141