DeweyBurdPubEm Resource

From: Burrows, Ronald
Sent: Wednesday, August 25, 2010 6:04 PM
To: Burrows, Ronald
Subject: Intro to Statistical Methods R.L. Ott and M. Longnecker.pdf
Attachments: Intro to Statistical Methods R.L. Ott and M. Longnecker.pdf

Hearing Identifier: Powertech_Uranium_Dewey_Burdock_LA_Public
Email Number: 124

Mail Envelope Properties (44CD2E65B0FF0E499CB32BC30CF781F0108E914016)

Subject: Intro to Statistical Methods R.L. Ott and M. Longnecker.pdf
Sent Date: 8/25/2010 6:04:05 PM
Received Date: 8/25/2010 6:04:56 PM
From: Burrows, Ronald

Created By: Ronald.Burrows@nrc.gov

Recipients:
"Burrows, Ronald" <Ronald.Burrows@nrc.gov>
Tracking Status: None

Post Office: HQCLSTR01.nrc.gov

Files	**Size**	**Date & Time**
MESSAGE	4	8/25/2010 6:04:56 PM
Intro to Statistical Methods R.L. Ott and M. Longnecker.pdf		378187

Options
Priority: Standard
Return Notification: No
Reply Requested: No
Sensitivity: Normal
Expiration Date:
Recipients Received:

GEORGE MASON UNIVERSITY
UNIVERSITY LIBRARIES

An Introduction to Statistical Methods and Data Analysis

Fifth Edition

R. Lyman Ott

Michael Longnecker
Texas A&M University

DUXBURY
THOMSON LEARNING

Australia • Canada • Mexico • Singapore • Spain • United Kingdom • United States

Thus, the 90th percentile is

$$P_{90} = 4.65 + \frac{1}{7}[.9(100) - 84] = 4.74$$

This means that 90% of the measurements lie below this value and 10% lie above it.

interquartile range

The second measure of variability, the **interquartile range,** is now defined. A slightly different definition of the interquartile range is given along with the box plot (Section 3.5).

DEFINITION 3.6 The **interquartile range (IQR)** of a set of measurements is defined to be the difference between the upper and lower quartiles; that is,

$$\text{IQR} = \text{75th percentile} - \text{25th percentile}$$

The interquartile range, although more sensitive to data pileup about the midpoint than the range, is still not sufficient for our purposes. In fact, the IQR can be very misleading when the data set is highly concentrated about the median. For example, suppose we have a sample consisting of 10 data values:

20, 50, 50, 50, 50, 50, 50, 50, 50, 80

The mean, median, lower quartile, and upper quartile would all equal 50. Thus, IQR equals $50 - 50 = 0$. This is very misleading because a measure of variability equal to 0 should indicate that the data consist of *n identical* values, which is not the case in our example. The IQR ignores the extremes in the data set completely. In fact, the IQR only measures the distance needed to cover the middle 50% of the data values and, hence, totally ignores the spread in the lower and upper 25% of the data. In summary, the IQR does not provide a lot of useful information about the variability of a single set of measurements, but can be quite useful when comparing the variabilities of two or more data sets. This is especially true when the data sets have some skewness. The IQR will be discussed further as part of the boxplot (Section 3.6).

In most data sets, we would typically need a minimum of five summary values to provide a minimal description of the data set: smallest value, $y_{(1)}$, lower quartile, $Q(.25)$, median, upper quartile, $Q(.75)$, and the largest value, $y_{(n)}$. When the data set has a unimodal, bell-shaped, and symmetric relative frequency histogram, just the sample mean and a measure of variability, the sample variance can represent the data set. We will now develop the sample variance.

We seek now a sensitive measure of variability, not only for comparing the variabilities of two sets of measurements but also for interpreting the variability of a single set of measurements. To do this, we work with the **deviation** $y - \bar{y}$ of a measurement y from the mean $\bar{y}$ of the set of measurements.

deviation

To illustrate, suppose we have five sample measurements $y_1 = 68$, $y_2 = 67$, $y_3 = 66$, $y_4 = 63$, and $y_5 = 61$, which represent the percentages of registered voters in five cities who exercised their right to vote at least once during the past year. These measurements are shown in the dot diagram of Figure 3.23. Each measurement is located by a dot above the horizontal axis of the diagram. We use the

Also, because the median is closer to the lower quartile than the upper quartile and because the upper whisker is a little longer than the lower whisker, the distribution is slightly nonsymmetrical. To see that this conclusion is true, construct a frequency histogram for these data (or refer to your results in Exercise 3.8).

The skeletal boxplot can be expanded to include more information about extreme values in the tails of the distribution. To do so, we need the following additional quantities:

lower inner fence: $Q_1 - 1.5(\text{IQR})$

upper inner fence: $Q_3 + 1.5(\text{IQR})$

lower outer fence: $Q_1 - 3(\text{IQR})$

upper outer fence: $Q_3 + 3(\text{IQR})$

Any score beyond an inner fence on either side is called a *mild outlier,* and a score beyond an outer fence on either side is called an *extreme outlier.*

EXAMPLE 3.14

Compute the inner and outer fences for the data of Example 3.13. Identify any mild and extreme outliers.

Solution For these data, we found the lower and upper quartiles to be 464 and 719, respectively; IQR = 719 − 464 = 255. Then

lower inner fence = 464 − 1.5(255) = 81.5

upper inner fence = 719 + 1.5(255) = 1,101.5

lower outer fence = 464 − 3(255) = −301

upper outer fence = 719 + 3(255) = 1,484

Also, from the stem-and-leaf plot we see that the lower and upper adjacent values are 189 and 1,020. Because the upper and lower fences are 1,101.5 and 81.5, respectively, there are no observations beyond the inner fences. Hence, there are no mild or extreme outliers.

We now have all the quantities necessary for constructing a boxplot.

Steps in Constructing a Boxplot

1. As with a skeletal boxplot, mark off a box from the lower quartile to the upper quartile.
2. Draw a solid line across the box to locate the median.
3. Mark the location of the upper and lower adjacent values with an x.
4. Draw a dashed line between each quartile and its adjacent value.
5. Mark each extreme outlier with the symbol o.

EXAMPLE 3.15

Construct a boxplot for the data of Example 3.13.

Solution The boxplot is shown in Figure 3.27.

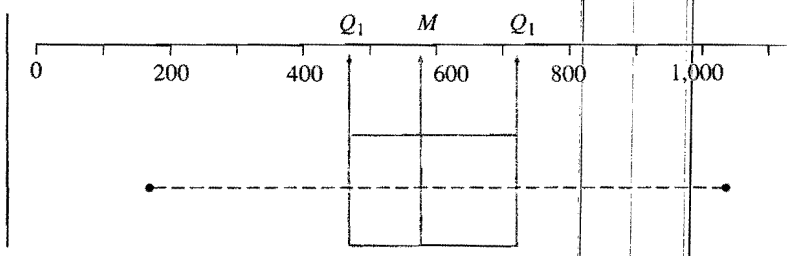


FIGURE 3.27 The boxplot for the data of Example 3.13

What information can be drawn from a boxplot? First, the center of the distribution of scores is indicated by the median line in the boxplot. Second, a measure of the variability of the scores is given by the interquartile range, the length of the box. Recall that the box is constructed between the lower and upper quartiles so it contains the middle 50% of the scores in the distribution, with 25% on either side of the median line inside the box. Third, by examining the relative position of the median line, we can gauge the symmetry of the middle 50% of the scores. For example, if the median line is closer to the lower quartile than the upper, there is a greater concentration of scores on the lower side of the median within the box than on the upper side; a symmetric distribution of scores would have the median line located in the center of the box. Fourth, additional information about skewness is obtained from the lengths of the whiskers; the longer one whisker is relative to the other one, the more skewness there is in the tail with the longer whisker. Fifth, a general assessment can be made about the presence of outliers by examining the number of scores classified as mild outliers and the number classified as extreme outliers.

Boxplots provide a powerful graphical technique for comparing samples from several different treatments or populations. We will illustrate these concepts using the following example. Several new filtration systems have been proposed for use in small city water systems. The three systems under consideration have very similar initial and operating costs, and will be compared on the basis of the amount of impurities that remain in the water after passing through the system. After careful assessment, it is determined that monitoring 20 days of operation will provide sufficient information to determine any significant difference among the three systems. Water samples are collected on a hourly basis. The amount of impurities, in ppm, remaining in the water after the water passes through the filter is recorded. The average daily values for the three systems are plotted using a side-by-side boxplot, as presented in Figure 3.28.

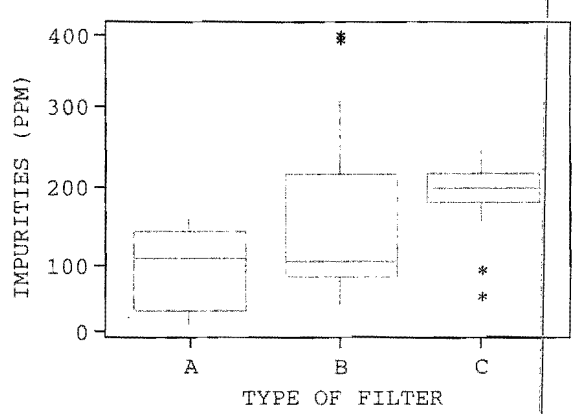


FIGURE 3.28 Removing impurities using three filter types

An examination of the boxplots in Figure 3.28 reveals the shapes of the relative frequency histograms for the three types of filters based on their boxplots. Filter A has a unimodal, symmetric distribution, filter B is unimodal, skewed to the right, and filter C is unimodal, skewed to the left. Filters A and B have nearly equal medians. However, filter B is much more variable than both filters A and C. Filter C has a larger median than both filters A and B but smaller variability than A with the exception of the two very small values obtained using filter C. The extreme values obtained by filters C and B, identified by *, would be examined to make sure that they are valid measurements. These measurements could be either recording errors or operational errors. They must be carefully checked because they have such a large influence on the summary statistics. Filter A would produce a more consistent filtration than filter B. Filter A generally filters the water more thoroughly than filter C. We will introduce statistical techniques in Chapter 8 that will provide us with ways to differentiate among the three filter types.

EXERCISES

Basic Techniques

3.48 Find the median and the lower and upper quartiles for the following measurements: 5, 9, 3, 6, 5, 7, 9, 2, 8, 10, 4, 3, 9.

3.49 Repeat Exercise 3.48 for the following measurements: 29, 22, 26, 20, 19, 16, 24, 11, 22, 23, 20, 29, 17, 18, 15.

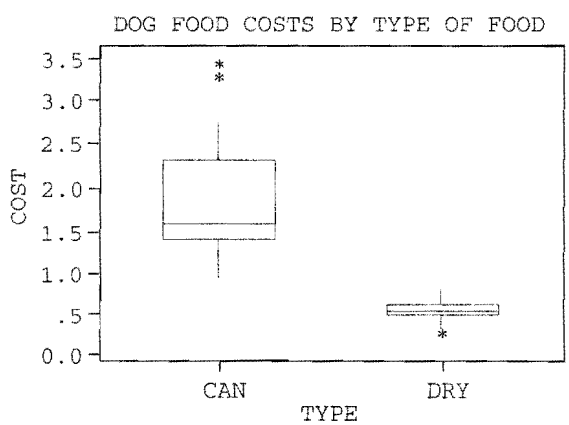
Applications

Med. **3.50** The number of persons who volunteered to give a pint of blood at a central donor center was recorded for each of 20 successive Fridays. The data are shown here:

320	370	386	334	325	315	334	301	270	310
274	308	315	368	332	260	295	356	333	250

a. Construct a stem-and-leaf plot.
b. Construct a boxplot and describe the shape of the distribution of the number of persons donating blood.

Bus. **3.51** *Consumer Reports* in its May 1998 issue provides cost per daily feeding for 28 brands of dry dog food and 23 brands of canned dog food. Using the Minitab computer program, the side-by-side boxplot for these data follow.



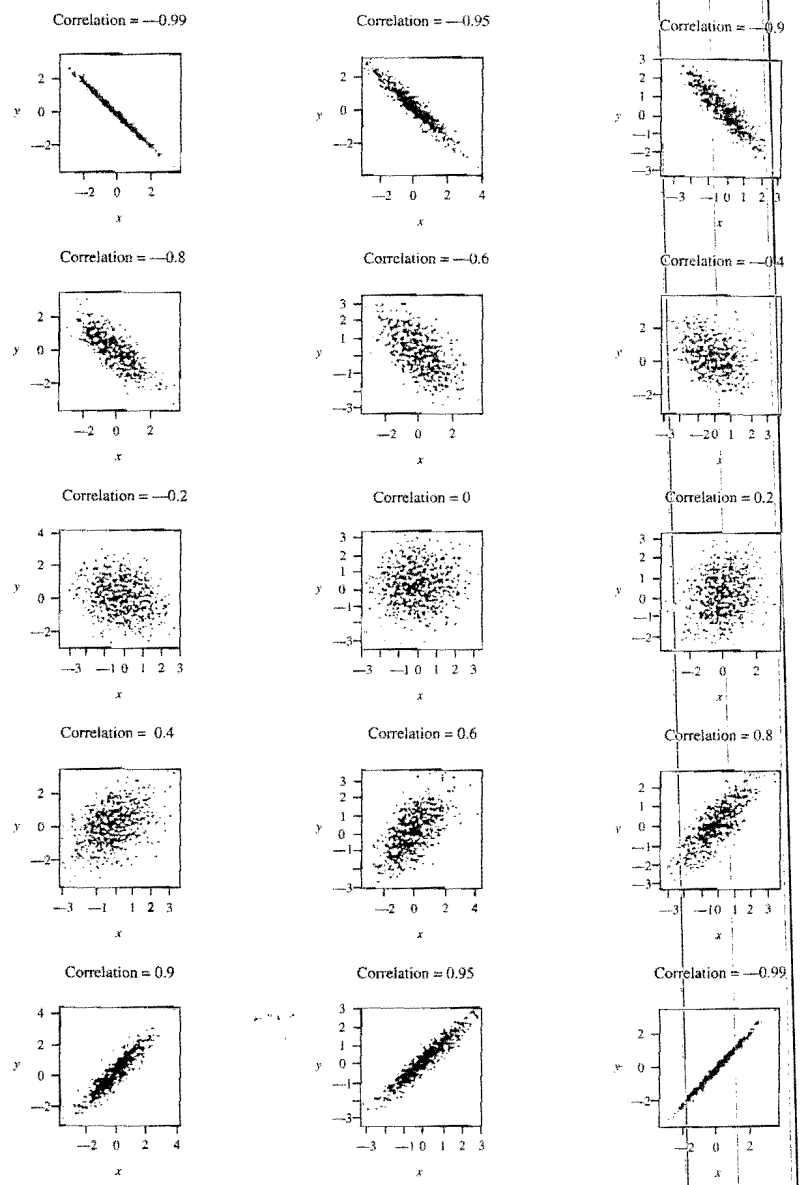


FIGURE 11.20
Samples of size 1,000 from the bivariate normal distribution

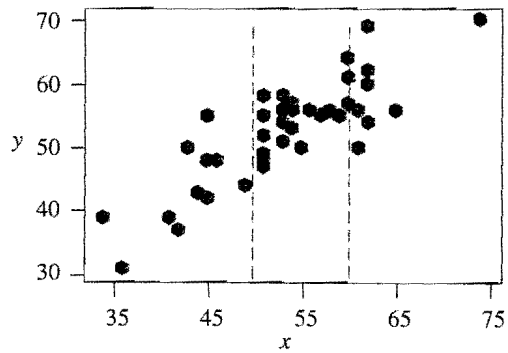


FIGURE 11.21
Effect of limited x range on sample correlation coefficient