Math 214 – Introductory Statistics **Summer 2008**
6-10-08 Class Notes

Sections 3.4, 3.5
3.4: 9-15, 22-25a
3.5: 1-9

Measures of Position

Consider this question: "Suppose you scored 80 on your first sociology exam (which had a mean of 60 and a standard deviation of 5) and you scored 75 on the second exam (with a mean of 65 and a standard deviation of 2). On which exam did you do better?"

We often want to compare data values from two different sample or populations. To do this, we need to standardize the scores.

Definition: The ***z-score*** (or ***standard score***) is the number of standard deviations that a given data value is above (for a positive z-score) or below (for a negative z-score) the mean. It is found by using the formula $z = \frac{x - \overline{x}}{s}$ (for a sample) or $z = \frac{x - \mu}{\sigma}$ (for the population).

Example 1: IQ scores are normally distributed (i.e. bell-shaped) with a population mean of 100 and a standard deviation of 15. If you have an IQ score of 143, what is your z-score?

$$z = \frac{x - \mu}{\sigma} = \frac{143 - 100}{15} = 2.87$$

Note: Always round z-scores to two decimal places

As mentioned, z-scores are an effective way to compare data values from two different data sets.

Example 2: Recall the question I posed at the beginning of this section. Suppose you scored 80 on your first sociology exam (which had a mean of 60 and a standard deviation of 5) and you scored 75 on the second exam (with a mean of 65 and a standard deviation of 2). On which exam did you do better?

Your z-score on Exam 1 was $\frac{80-60}{5} = 4.00$ and your z-score on Exam 2 was $\frac{75-65}{2} = 5.00$. So even though you scored higher on Exam 1, relative to the sample data, you did better on Exam 2.

Recall that if a set of data is arranged in increasing order, the middle data value (or mean of the two middle data values) is the median. This is the value that divides the data set into two equal parts. We can extend this idea further: we can find the values that divide the data set into 4 equal parts. These values, denoted by Q_1, Q_2, and Q_3 are called the first, second, and third quartiles respectively. Note that Q_2 is the mean.

Similarly, the values that divide the data set into 10 equal parts are called deciles and are denoted by $D_1, D_2, \ldots, D_9$. The values that divide the data set into 100 equal parts are called percentiles and are denoted by $P_1, P_2, \ldots, P_{99}$. Note that $D_5 = P_{50} = Q_2$, $P_{25} = Q_1$, and $P_{75} = Q_3$.

To find the percentile for a *given data value*, we use the following formula:

$$\text{Percentile of } x = \frac{\#\text{of data values} < x}{n} \cdot 100$$

Example 3: Compute the percentile of 8 for the data set 1, 2, 4, 6, 8, 9, 10.

$$\text{Percentile of } 8 = \frac{\#\text{of data values} < 8}{7} \cdot 100 = \frac{4}{7} \cdot 100 = 57.14 = 57^{\text{th}} \text{ percentile.}$$

Definition: The k^{th} ***percentile*** is the value such that k percent of the data set is below that value.

How do we find the k^{th} ***percentile*** P_k?

(1) Arrange data in increasing order.

(2) Compute the L-value (location value) using the formula $L = \left(\frac{k}{100}\right) \cdot n$ (where n is the size of your sample).

(3) (a) If L is a whole number, then P_k is midpoint between the L^{th} data value and the next data value.

(b) If L is not a whole number, round it up, call this new number L, and then P_k is the L^{th} data value.

Example 4: Find each desired percentile for the data set 1.1, 1.7, 1.9, 2.1, 2.2, 2.5, 3.3, 6.2, 6.8, 20.3.

(a) 70^{th}

(b) 31^{st}

(a) $L = \left(\frac{70}{100}\right) \cdot 10 = 7$. The 7^{th} data value is 3.3 and the next one is 6.2. So $P_{70} = \frac{3.3 + 6.2}{2} = 4.8$.

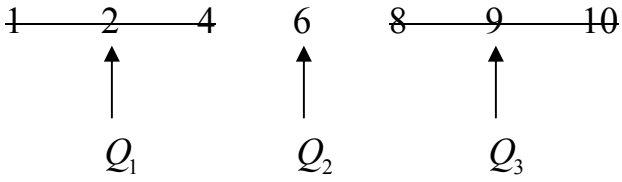
(b) $L = \left(\frac{31}{100}\right) \cdot 10 = 3.1$. This rounds up to 4, so $P_{31} = 2.1$ (the 4^{th} data value).

The Five-Number Summary

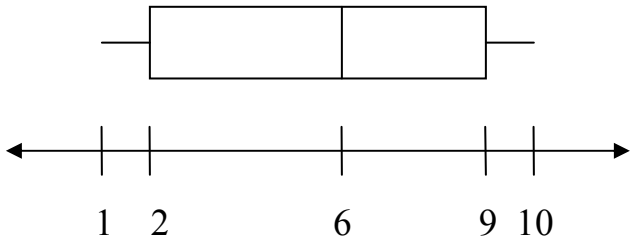
The five numbers $x_{\min}$, Q_1, Q_2, Q_3, and $x_{\max}$ constitute the ***5-number summary*** of a data set. This is often expressed using a ***boxplot***.

Example 5: Compute the 5-number summary and the boxplot for the data set 1, 2, 4, 6, 8, 9, 10.

Clearly, $x_{\min} = 1$, $Q_2 = 6$ (the median), and $x_{\max} = 10$. To find Q_1 and Q_3, we can use the formula given above (to find the 25^{th} percentile and the 75^{th} percentile) OR we can realize that just like Q_2 is the median of the data set, Q_1 and Q_3 are the medians of each half of the data set. So we can see that,



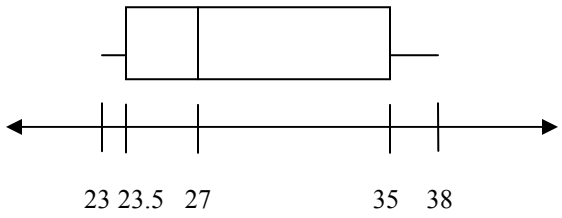
So our "five numbers" are: 1, 2, 6, 9, and 10. These are graphed using a boxplot as follows.



Visually you can see that the area between Q_1 and Q_3 makes up the middle half of the data. The size of this area is the ***interquartile range*** (and is simply $Q_3 - Q_1$).

Example 6: Find the 5-number summary, interquartile range, and boxplot for the data set 23, 23, 24, 26, 27, 31, 35, 35, 38.

$x_{\min} = 23$
$Q_1 = 23.5$
$Q_2 = 27$
$Q_3 = 35$
$x_{\max} = 38$



$IQR = 11.5$