

## Chapter 4: Summarizing Continuous Data

### Statistics Based on Ordered Values

Median  
Range  
Quartiles & Interquartile Range  
Symmetry & Skewness

### Statistics Based on Averages

Mean  
Standard Deviation  
Degrees of Freedom  
Mound-Shaped Distributions

### Linear Transformations & Standardization

Chap. 4 - page 1

### Statistics Based on Ordered Values

#### The Median (page 85)

1. Arrange the numbers in numerical order from smallest to largest
2. If the number of data points,  $n$ , is odd, the median is the middle value in the ordered list.
3. If  $n$  is even, the median is the average of the two middle data points.

The median is at position  $(n+1)/2$  if  $n$  is odd and halfway between the data at positions  $n/2$  and  $(n/2)+1$  when  $n$  is even.

Chap. 4 - page 3

## Statistics

A **statistic** is a number calculated from a sample of data.

It is typically used to describe a particular feature of the data—a **descriptive statistic**.

Our first goal is to find a statistic (or statistics) to describe the **middle** of a set of data.

Later we will describe the amount of variability in a set of data.

Chap. 4 - page 2

Example:

data: 4.5, 3.2, 5.4, 4.4, 3.9

In order: 3.2, 3.9, 4.4, 4.5, 5.4

Median is 4.4

Notice that actual numerical values are not used much after putting data into order.

If the 5.4 had been 54 the median would not change. It is still 4.4.

We say that **the median is resistant to outliers** or extreme data values.

The extremes do not influence the median.

Chap. 4 - page 4

## The Range (page 86)

The **range** of a set of numbers is the difference between the largest and smallest values.

The range measures the variability in the set of numbers.

The range is **very sensitive** to extreme data values.

Example:

$$5.4 - 3.2 = 2.2$$

but

$$54 - 3.2 = 50.8$$

To find the third quartile,  $Q_3$ :

1. Find the upper half of the data, that is, the ordered values above the median.
2. The third quartile is the median of the upper half.

Quartiles are resistant to extreme values in the data.

### Interquartile Range

= third quartile – first quartile

$$= Q_3 - Q_1$$

The **interquartile range is resistant to extreme values** in the data.

## Quartiles & Interquartile Range (page 87)

The **first quartile**,  $Q_1$ , is one-quarter up from the smallest data value.

The **third quartile**,  $Q_3$ , is one-quarter down from the largest data value.

To find the first quartile,  $Q_1$ :

1. Find the lower half of the data, that is, the ordered values below the median.
2. The first quartile is the median of the lower half.

## Symmetric and Skewed Distributions (page 89)

For a symmetric distribution the left and right sides are mirror images.

Data distributions can only be approximately symmetric.

If a distribution is stretched out longer on the high side than on the low side, we say that the distribution is **skewed** toward the high side (or right side).

If a distribution is stretched out longer on the low side than on the high side, we say that the distribution is **skewed** toward the low side (or left side).

Example: Edwin Moses' winning times  
(Histogram  
Dotplot  
Stem-and-Leaf)

$n = 122$  so median is at position  
 $123/2 = 61.5$  or the average of data at  
positions 61 and 62.

$$\text{Median} = (48.53 + 48.55)/2 = 48.54$$

Lower half is the ordered first 61 data values  
so  $Q_1$  is at position  $62/2 = 31$ .

$$Q_1 = 47.95$$

$Q_3$  is at position 31 down from the largest  
value (or at position  $122 - 31 + 1 = 92$  or at  
 $62 + 31 - 1 = 92$ ) so

$$Q_3 = 49.0$$

### The Standard Deviation $s$ (page 93)

- Compute the deviations from the mean  
 $y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y}$  (page 93)
- Square the deviations and sum them
- Divide the sum by  $n-1$
- Finally, take the positive square root to  
get  $s$

In symbols (page 94)

$$s = \sqrt{\frac{\sum (y_j - \bar{y})^2}{n - 1}}$$

## Statistics Based on Averages (page 91)

Individual observations denoted  $y_1, y_2, \dots, y_n$ .

### The Mean (page 92)

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{\sum y_j}{n}$$

The mean is the balance point in a  
distribution and is another measure of the  
center of a distribution.

The mean is sensitive to extreme  
observations or outliers.

**Outliers** are observations that are  
separated from the main body of the data.  
(page 93)

### Facts about $s$

- The larger  $s$  is more spread out the  
distribution
- the smaller  $s$  is the more compact the  
distribution
- The units for  $s$  are the same as the units  
for  $y$  (page 94)
- $s \geq 0$
- $s$  is zero only when all of the  
observations are identical

The standard deviation is sensitive to  
extreme observations or outliers.

Example: 2, 3, 1 with  $n = 3$ .

$$\bar{y} = (2 + 3 + 1)/3 = 6/3 = 2$$

Data	Deviation from mean	Deviation Squared
2	$2 - 2 = 0$	$0^2 = 0$
3	$3 - 2 = 1$	$1^2 = 1$
1	$1 - 2 = -1$	$(-1)^2 = 1$
Totals	6	2

so

$$s = \sqrt{\frac{2}{3-1}} = \sqrt{\frac{2}{2}} = \sqrt{1} = 1$$

## Mound-Shaped Distributions (page 95)

Mound-shaped distributions can be economically summarized by two statistics: the mean and the standard deviation.

For mound-shaped distributions we have the empirical rule (page 96)

- about **68%** of the observations lie within **one** standard deviation of the mean
- about **95%** of the observations lie within **two** standard deviations of the mean
- **nearly all** of the observations lie within **three** standard deviations of the mean

## Degrees of Freedom (page 95)

Only  $n-1$  of the deviations from the mean,  $y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y}$ , are “free bits” of data.

This is because these deviations “automatically” sum to zero.

$$\sum (y_j - \bar{y}) = 0$$

Given any  $n-1$  of the deviations from the mean, we could solve for the last deviation.

We say that the  $n$  deviations,  $y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y}$ , have  $n-1$  **degrees of freedom**.

Example: Edwin Moses data

$$n = 122, \bar{y} = 48.528 \text{ and } s = 0.732$$

Minitab output

Variable	N	Mean	Median	TrMean	StDev	SEMean
Time	122	48.528	48.540	48.522	0.732	0.066
Variable	Min	Max	Q1	Q3		
Time	47.020	50.190	47.948	49.000		

$$\bar{y} \pm 1s \text{ is } 48.528 \pm 0.732$$

or

47.796 up to 49.260

79 of the 122 observations are in this interval. This is 65%.

Similarly,  $\bar{y} \pm 2s$  is  $48.528 \pm 2(0.732)$

or  $48.528 \pm 1.464$

or 47.064 up to 49.992

118 or 97% of the 122 are within this interval.

Finally,  $\bar{y} \pm 3s$  is  $48.528 \pm 3(0.732)$  or

$48.528 \pm 2.196$  or 46.332 up to 50.724

All or 100% of the 122 times are within this interval.

The Minitab Macro NORMCHEK will do all of the work

Example:

```
MTB > exec 'normchek'
Executing from file: normchek.MTB
This macro uses C50 and K45-K50.

Enter the number of the column containing the data and the number
of standard deviations to check, in that order.

DATA> 1
DATA> 2
...working...
The percent within your selected standard deviations of the mean
is
Data Display
K46      96.7213
Compare this to the following percent for a normal distribution.
Data Display
K45      95.4500
```

## Linear Transformations & Standardization (page 98)

$$y_* = a + by$$

Example:

You do business in Canada. A group of accounts has a mean of \$20,000 and a standard deviation \$1000 in U.S. dollars.

One U.S. dollar is worth about 1.45 Canadian dollars. That is, the rate of exchange is 1.45 Canadian dollar per U.S. dollar (as of 2/2/2000).

If  $y$  is in U.S. dollars, then  $y_* = 1.45y$  gives the same amount in Canadian dollars.

What will the mean and standard deviation be if we measure our accounts in Canadian dollars?

## Facts:

$$\bar{y}_* = a + b\bar{y}$$

and

$$s_{y_*} = |b|s_y$$

Example:

$a = 0$  and  $b = 1.45$  so that, in Canadian dollars, the mean account size is  $1.45 \times 20,000 = \$29,000$  Canadian and the standard deviation of the account sizes is  $1.45 \times 1000 = \$1450$  Canadian.

Same sorts of results hold for medians, and quartiles and for ranges and interquartile ranges.

If a set of accounts has a median of \$100,000 Canadian dollars and an interquartile range of \$15,000 Canadian dollars, then the corresponding values in U.S. dollars are

$$(1/1.45) \times 100,000 = \$68,965.517 \text{ U.S.}$$

$$\text{and } (1/1.45) \times 15,000 = \$10,344.828 \text{ U.S.}$$

Examples:

Data	Deviation	(Dev) <sup>2</sup>	Standardized
4	0	0	0/2 = 0
6	2	4	2/2 = 1
2	-2	4	-2/2 = -1

$$\text{mean} = 12/3 = 4$$

$$s = \sqrt{\frac{8}{2}} = \sqrt{4} = 2$$

## Standardization (page 100)

$$z = \frac{y - \bar{y}}{s}$$

Facts:

$$\bar{z} = 0$$

and

$$s_z = 1$$

z measures the number of standard deviations that y is above (or below) the mean.

Example:

ID	Data	Standardized
1	2.2	-0.84607
2	3.1	0.38209
3	3.9	1.47379
4	2.1	-0.98253
5	2.8	-0.02729

Here

$$\text{mean} = 2.820$$

and

$$s = 0.73280$$

For mound-shaped distributions the empirical rule (page 96) can be rephrased in terms of standardized values as: (page 100)

- about 68% of the standardized observations lie within  $(-1, +1)$
- about 95% of the standardized observations lie within  $(-2, +2)$
- nearly all of the standardized observations lie within  $(-3, +3)$