

3.1 Measures of Central Tendency

The word “average: is very ambiguous and can actually refer to the mean, median, mode or midrange.

Notation: n = sample size
 N = population size

A **statistic** is a characteristic or measure obtained by using a data value from a sample.

A **parameter** is a characteristic or measure obtained by using all the data values for a specific population.

A. The **mean** (commonly called the average) of a data set is defined to be the sum of the data divided by the number of data items. Your text says to “round your means to one more decimal place than occurs in the raw data.” We will always take everything out to 4 decimal places as a rule, so ignore what your book says.

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} \text{ or } \bar{x} = \frac{\sum x}{n}$$

Rounding Rule for the Mean: The mean should be rounded to one more decimal place than occurs in the raw data.

B. The **mode** of a data set is the value that occurs most frequently. A data set can be uni-modal, bi-modal, multi-modal, or have no mode at all. If more than one number shows up as the mode, we list each as part of our answer. If no value shows up the most, we say that there is **no mode**.

C. The **median** of a data set is the "middle" value when the data are listed in numerical order. If n is odd, the median is the middle data value. If n is even, the median is the mean (average) of the two middle data values.

D. The **midrange** of a data set is found by calculating the mean of the maximum and minimum values of the data set:

$$\text{midrange} = \left(\frac{\text{lowest value} + \text{highest value}}{2} \right)$$

example:

DATA: 10 12 10 13 12 8 12 25 15 14 13 7

List the data in order first:

median:

mode:

midrange:

mean:

example: Here the data is **grouped in classes**:

weekly salary	frequency	
\$200	6	mode =
\$300	2	
\$350	2	median =
\$700	1	
\$840	1	mean =
\$950	1	

Sometimes you don’t have the raw data itself, but only the classes. Find the mean of this distribution: (Hint: Use the class midpoint from each class.)

Intake (mg)		x	f
under 200			11
200-under 400			85
400-under 600			90
600-under 800			115
800-under 1000			135
1000-under 1200			37
1200-under 1400			22

For the distribution above, what is the median, mode and midrange?

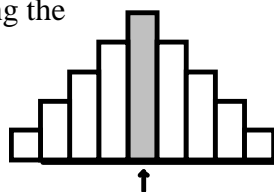
Sometimes one must find the mean of a data set in which not all values are equally represented. Find the weighted mean of a variable X by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\bar{X} = \frac{w_1X_1 + w_2X_2 + \dots + w_nX_n}{w_1 + w_2 + \dots + w_n} = \frac{\sum wX}{\sum w}$$

where w_1, w_2, \dots, w_n are the weights and X_1, X_2, \dots, X_n are the values.

Which measure of central tendency should you use?

1. The mean is very sensitive to large or small data values; the median is not.
2. The mode is not always near the center.
3. The perfect case is a bell curve. It has perfect symmetry. (mean = mode = median)
4. Ordinal data are data about order or rank. Most statisticians recommend using the median for indicating the center of an ordinal data set.



Properties and Uses of Central Tendency

The Mean

1. Once computes the mean by using all the values of the data.
2. The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.
3. The mean is used in computing other statistics, such as the variance.
4. The mean for a data set is unique and not necessarily one of the data values.
5. The mean cannot be computed for an open-ended frequency distribution.
6. the mean is affected by extremely high or low values, called outliers, and may not be the appropriate average to use in these situations.

The Median

1. The median is used when ne must find the center or middle value of a data set.
2. The median is used when one must determine whether the data values fall into the upper half or lower half of the distribution.
3. The median is used for all open-ended distributions.
4. The median is affected less than the mean by extremely high or extremely low values.

The Mode

1. The mode is used when the most typical case is desired.
2. The mode is the easiest average to compute.
3. The mode can be used when the data are nominal, such as preferences, gender, or political affiliation.
4. The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

The Midrange

1. The midrange is easy to compute.
2. The midrange gives the midpoint.
3. The midrange is affected by extremely high or low values in a data set.

Relative positions of the mean and median for (a) right-skewed, (b) symmetric, and (c) left-skewed distributions

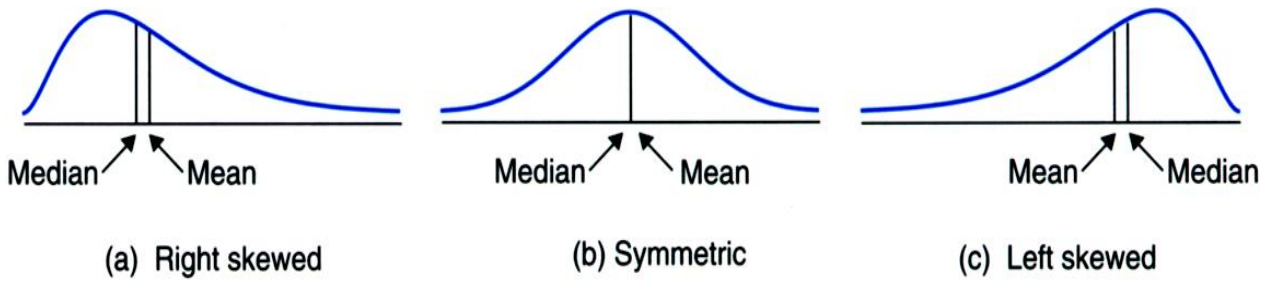


Table 3-1 Summary of Measures of Central Tendency

Measure	Definition	Symbol(s)
Mean	Sum of values, divided by total number of values	μ, \bar{X}
Median	Middle point in data set that has been ordered	MD
Mode	Most frequent data value	None
Midrange	Lowest value plus highest value, divided by 2	MR

3.2 Measures of Variation

1. **Range**- measures the "spread" of the data. The Range= (highest value – lowest value)
2. **Standard Deviation**- measures the variation in a data set by determining *how far the data values are from the mean, on the average.*
3. The **variance** is the square of the standard deviation. It is the average of the squares of the distance each value is from the mean.

The standard deviation is a measure of variation- the more variation there is in a data set, the larger its standard deviation.

Here are the formulas for the population standard deviation (left) and sample standard deviation (right):

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X - \mu)^2}{N}} \quad s = \sqrt{s^2} = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

The shortcut formula for computing s is

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

Example: Find the sample variance and sample standard deviation for the data set: 1, 2, 6, 7

a. Use $s = \sqrt{s^2} = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$

b. Use $s = \sqrt{\frac{n(\sum X^2) - (\sum X)^2}{n(n-1)}}$

The Range Rule of Thumb: A rough estimate of

the standard deviation is $s \approx \frac{\text{range}}{4}$

example:

DATA: **64 66 66 68 69 70 72 73**

\bar{x} = _____ s = _____

σ = _____

Turn back to page 16 in your lecture notes and find the sample and population standard deviation for this grouped data:

Intake (mg)	x	f
under 200		11
200-under 400		85
400-under 600		90
600-under 800		115
800-under 1000		135
1000-under 1200		37
1200-under 1400		22

In general, we use the following notation:

	size	mean	std. Dev.
sample	n	\bar{x}	s
population	N	μ	σ

Uses of the Variance and Standard Deviation

1. Variances and standard deviations can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed. This information is useful in comparing two (or more) data sets to determine which is more (most) variable.
2. The measure of variance and standard deviation determine the consistency of a variable. For example, in the manufacture of fittings, such as nuts and bolts, the variation in the diameters must be small, or the parts will not fit together.

3. The variance and standard deviation are used to determine the number of data values that fall within a specified interval in a distribution.
4. The variance and standard deviation are used quite often in inferential statistics.

When finding the mean of a data set, we can either consider the mean to be a sample mean \bar{x} or a population mean μ , **depending on how the data is being interpreted.**

Suppose a data set consists of the heights of an 11-man basketball team:

78 80 78 77 80 76 76 81 75 79 80 (inches)

If we were interested in this team only, we would call the mean a population mean and write $\mu = 78.2$ inches with $N = 11$.

If this team were to be considered to be a sample of all NBA teams, we would call the mean a sample mean and write $\bar{x} = 78.2$ inches with $n = 11$.

example: 1988-89 Phoenix Suns- Frequency Distribution of heights:

height (inches)	74	75	76	77	78	79	80	81	82	83
frequency	2	2	1	0	2	2	1	2	3	1

\bar{x} = _____ s = _____

example: Consider the following data sets:

DATA SET 1 **30 20 16 24 22 19 23 13 18 9 18 28**

DATA SET 2 **14 9 56 32 13 8 26 3 9 16 31 23**

- a) Which data set appears to have more deviation?
- b) Compute \bar{x} and s for each data set:

\bar{x}_1 = _____ \bar{x}_2 = _____ s_1 = _____ s_2 = _____

c) Draw dot plots for each data set:

DATA SET 1 30 20 16 24 22 19 23 13 18 9 18 28

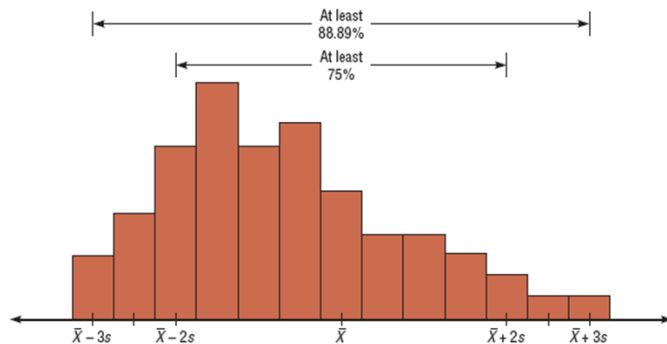
DATA SET 2 14 9 56 32 13 8 26 3 9 16 31 23

d) There seems to be more variance in data set 2. (The numbers are further apart.) Hence, the standard deviation for data set 2 is larger.

Chebyshev’s Theorem: The proportion of values from a data set that will fall within k standard deviations of the mean will be at least $1 - \frac{1}{k^2}$, where k is a number greater than 1 (k is not necessarily an integer.) Chebyshev’s Theorem can also be used to find the minimum percentage of data values that will fall between any two given values.

K	$\frac{1}{k^2}$	$1 - \frac{1}{k^2}$	At least _____ % of the data values will fall within k standard deviations to either side of the mean

Chebyshev's Theorem



The Empirical (Normal) Rule: Chebyshev’s theorem applies to ANY distribution regardless of its shape. However, when a distribution is bell-shaped, or what we call normal, the following statements are true:

1. Approximately 68% of the data values will fall within 1 standard deviation of the mean.
2. Approximately 95% of the data values will fall within 2 standard deviations of the mean.
3. Approximately 99.7% of the data values will fall within 3 standard deviations of the mean.

KEY FACT: In any data set, almost all of the data will lie within 3 standard deviations to either side of the mean. We can write this as an interval: $\bar{x} \pm 3s$

Pg. 137 #8 The increase (in cents) in cigarette taxes for 17 states in a 6-month period are

60, 20, 40, 40, 45, 12, 34, 51, 30, 70, 42, 31, 69, 32, 8, 18, 50

Use the range rule of thumb to estimate the standard deviation. Compare the estimate to the actual standard deviation.

Pg. 140. #40 The average farm in the United States in 2004 contained 443 acres. The standard deviation is 42 acres. Using Chebyshev’s theorem, find the minimum percentage of data values that will fall in the range of 338 – 548 acres.

Page 140, #34 In a distribution of 160 values with a mean of 72, at least 120 fall within the interval 67-77. Approximately what percentage of values should fall in the interval 62-82? Use Chebyshev’s Theorem.

Pg. 140 #42 The average full-time faculty member in a post-secondary degree-granting institution works an average of 53 hours per week.

a. If we assume the standard deviation is 2.8 hours, what percentage of faculty members work more than 58.6 hours a week?

b. If we assume a bell-shaped distribution, what percentage of faculty members work more than 58.6 hours a week?

3.3 Measures of Position

A **z-score** or **standard score** for a data value x is *the number of standard deviations x is away from the mean.*

For samples, the formula is $z = \frac{x - \bar{x}}{s}$ and for

populations the formula is $z = \frac{x - \mu}{\sigma}$.

If an x -value is below the mean, its corresponding z -score is negative. The z -score helps explain where a data value *is with respect to the mean and the rest of the sample.*

example: Consider a data set with $\bar{x}=80$ and $s=4$.

a) Find the z -score when $x=70$.

b) Interpret its meaning in words.

example: A student scores 60 on a mathematics test that has a mean of 54 and a standard deviation of 3, and she scores 80 on a history test with a mean of 75 and a standard deviation of 2. On which test did she perform better?

Quartiles **Q1**, **Q2**, **Q3** separate data into four parts, when the data is listed in order.

example: DATA: **11 13 14 17 18 19 21 28**
 13 13 14 17 18 21 25 17

List the data in order:

Find $Q1=$ _____ $Q2=$ _____ $Q3=$ _____

When the number of data values is not divisible by 4, first find the median. This is $Q2$. Then find the median of all values below $Q2$ and above $Q2$. These medians will be $Q1$ and $Q3$, respectively.

On the TI-83, the quartiles are given to you automatically when you enter the data in a list and use the 1-Var Stats command.

For the data above, calculate the trimean and the interquartile range.

The **Trimean** = $0.3 Q1 + 0.4 Q2 + 0.3 Q3$

The **Interquartile Range, IQR** = $Q3 - Q1$

The IQR measures the “middle 50%” of the data.

Outliers are observation that fall well outside the overall pattern of the data. An outlier requires special attention: It may be the result of a measurement or recording error, an observation from a different population, or an unusual extreme observation. Note that an extreme observation need not be an outlier; it may instead be an indication of skewness.

An **outlier** is defined to be any value that is more than 1.5 IQRs below $Q1$ or more than 1.5 IQRs above $Q3$.

Procedure for Identifying Outliers

1. Arrange the data in order and find Q_1 and Q_3 .
2. Find the interquartile range, IQR.
3. Multiply the IQR by 1.5.
4. Subtract the value obtained in step 3 from Q_1 and add the value to Q_3 .
5. Check the data set for any data value that is smaller than $Q_1 - 1.5(IQR)$ or larger than $Q_3 + 1.5(IQR)$

On the TI graphing calculator, you can also create a modified box plot in order to identify outliers.

Percentiles and **deciles** are defined in a similar manner; to find the deciles D1 through D9, for example, you would split the data up into *ten* evenly spaced parts.

Pg. 155 #30 Check the data set for outliers.

- a. 16, 18, 22, 19, 3, 21, 17, 20
- b. 24, 32, 54, 31, 16, 18, 19, 14, 17, 20
- c. 321, 343, 350, 327, 200
- d. 88, 72, 97, 84, 86, 85, 100
- e. 145, 119, 122, 118, 125, 116
- f. 14, 16, 27, 18, 13, 19, 36, 15, 20

3.4 Exploring Data Analysis

The **Five-Number Summary** of a data set consists of the five values:

{ min value, Q_1 , Q_2 , Q_3 , max value }

A **boxplot** is a graph of a data set that depicts the five-number summary in a visual way. It is also useful in helping you compare data sets. It is also sometimes referred to as a box-and-whisker-plot.

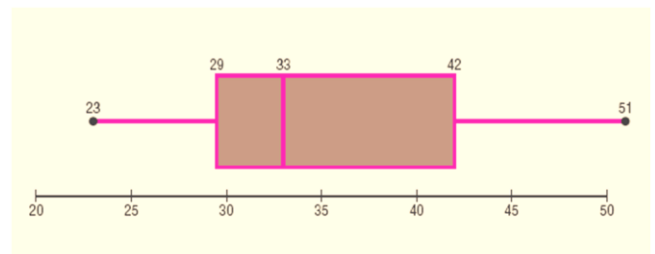
Information Obtained from a Box Plot

1. a. If the median is near the center of the box, the distribution is approximately symmetric.
b. If the median falls to the left of the center of the box, the distribution is positively skewed.
c. If the median falls to the right of the center, the distribution is negatively skewed.
2. a. If the lines are about the same length, the distribution is approximately symmetric.
b. If the right line is larger than the left line, the distribution is positively skewed.
c. If the left line is larger than the right line, the distribution is negatively skewed.

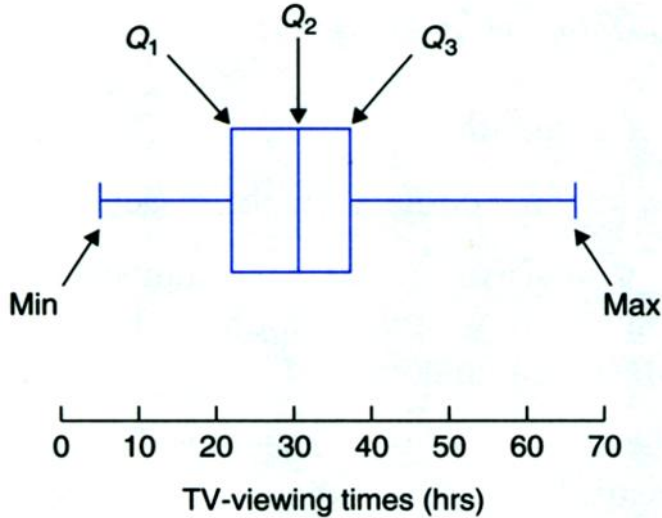
Example: Find the five-number summary for the following data set:

33, 38, 43, 30, 29, 40, 51, 27, 42, 23, 31

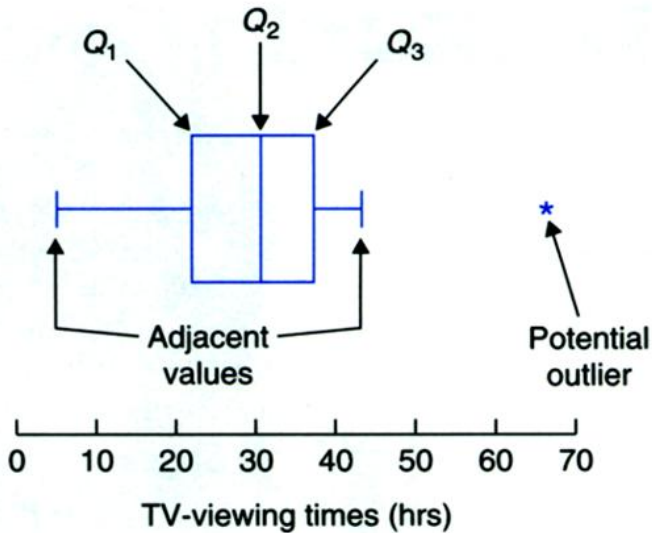
Boxplot for the data above:



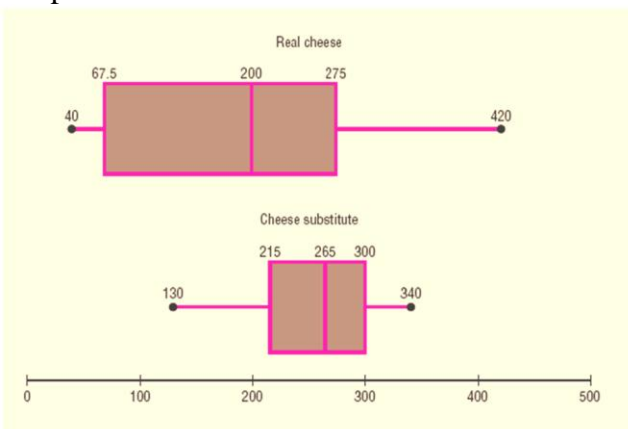
(a) Boxplot for TV-viewing times



(b) **Modified** boxplot for TV-viewing times



Sometimes you can use multiple boxplots to compare distributions:



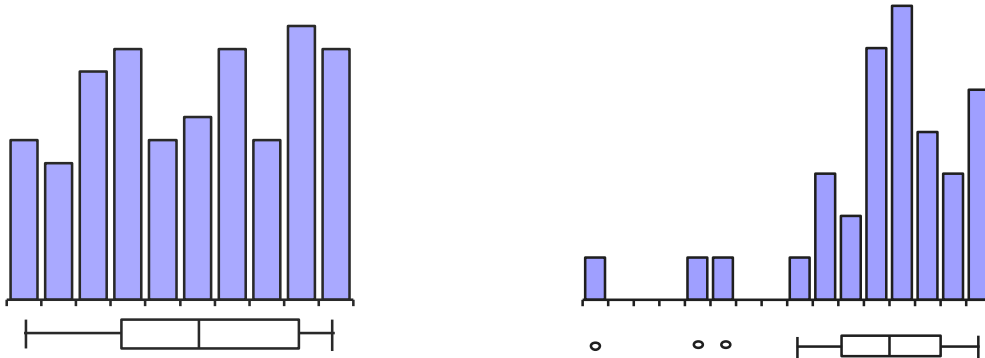
Traditional versus Exploratory Data Analysis

Traditional	Exploratory Data Analysis
Frequency distribution	Stem and leaf plot
Histogram	Boxplot
Mean	Median
Standard deviation	Interquartile Range

An important point to remember is that summary statistics (such as medians and IQRs) used in explanatory data analysis are said to be resistant statistics. A **resistant statistic** is relatively less affected by outliers than a nonresistant statistic. (The mean and standard deviation are examples of nonresistant statistics.)

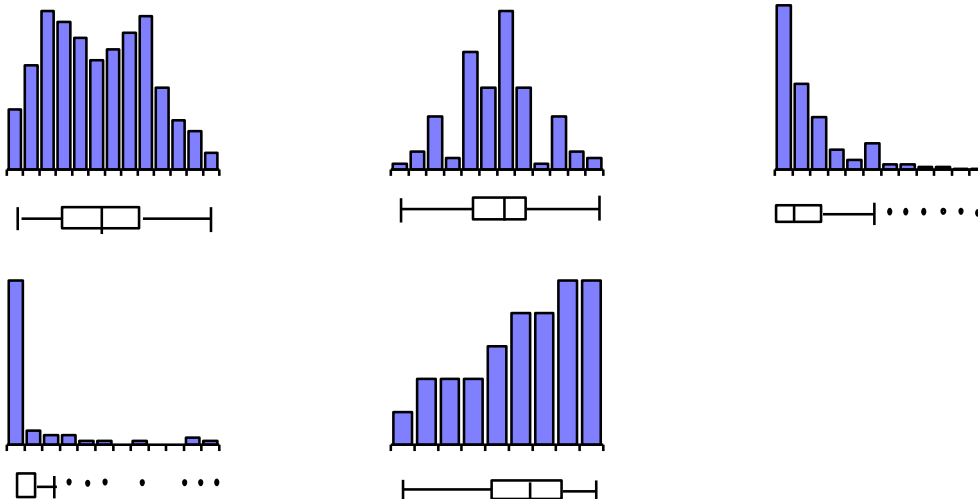
Matching Graphs

1. Consider the following two variables: A. age at death of a sample of 34 people
 B. the last digit of a social security number of each of 40 people
- Match these variables to their graphs:

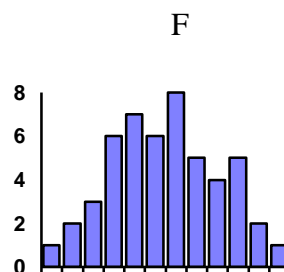
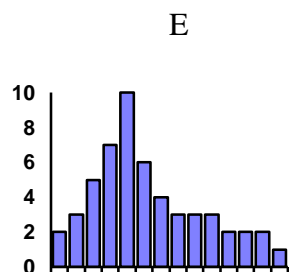
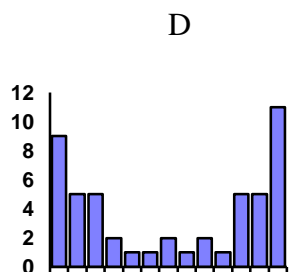
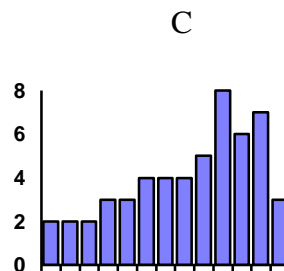
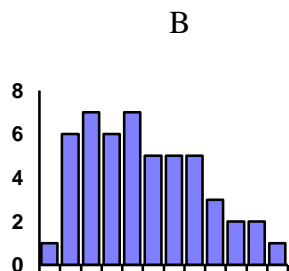
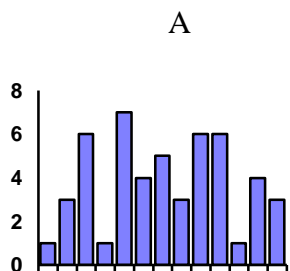


We know that there are relatively few deaths among young people; the death rate rises with age. Thus we would expect the histogram of ages of death to be skewed to the left. On the other hand, the social security data should have a distribution that is close to uniform.

2. Consider the following list of variables and match them to the appropriate graphs:
- A. scores on a fairly easy examination
 - B. number of menstrual cycles required to achieve pregnancy for a sample of women who attempted to get pregnant. Note that the data were self-reported from memory.
 - C. heights of a group of college students
 - D. numbers of medals won by medal-winning countries in the 1992 Winter Olympics
 - E. SAT scores for a group of college students



2. Match the following histograms to their summary statistics in the table below.



Variable	Mean	Median	Standard Deviation
1	50	50	10
2	50	50	15
3	53	50	10
4	53	50	20
5	47	50	10
6	50	50	5

2. Match the following histograms to their respective boxplots.

