Lecture 5

Displaying and Summarizing Data

Thought Question 1:

If you were to read the results of a study showing that daily use of a certain exercise machine resulted in an average 10-pound weight loss, what more would you want to know about the numbers in addition to the average?

(*Hint:* Do you think everyone who used the machine lost 10 pounds?)

Thought Question 2:

Suppose you are **comparing two job offers**, and one of your considerations is the **cost of living in each area**. You get the local newspapers and record the price of 50 advertised apartments for each community.

What summary measures of the rent values for each community would you need in order to make a useful comparison?

Would lowest rent in list be enough info?

Thought Question 3:

A real estate website reported that the ***median*** price of single family homes sold in the past 9 months in the local area was $136,900 and the ***average*** price was $161,447.

How do you think these values are computed? **Which do you think is more useful** to someone considering the purchase of a home, the median or the average?

What to do when you have the data

- We saw in the previous chapters how to collect data. We will spend the rest of this course looking at how to analyse the data that we have collected.

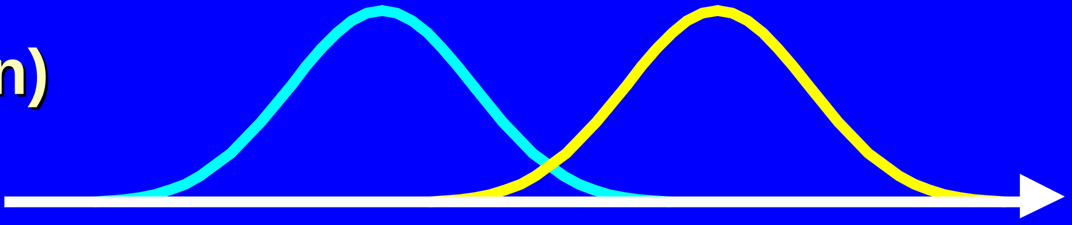
7.1 Turning Data Into Information

Four kinds of useful information about a set of data:

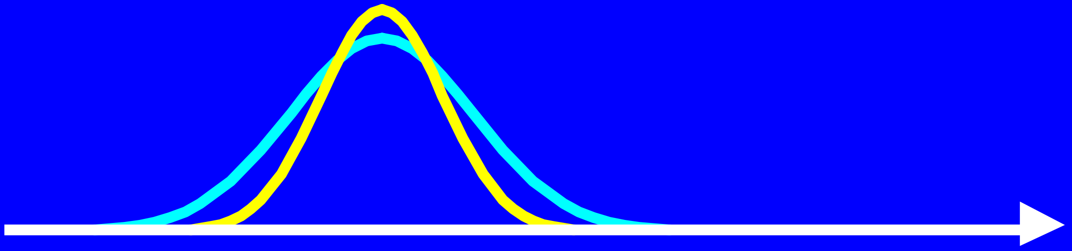
1. Center
2. Unusual values (outliers)
3. Variability
4. Shape

Numerical Data Properties

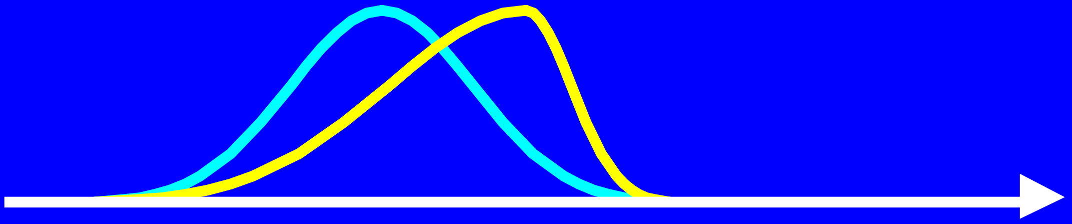
Center (Location)



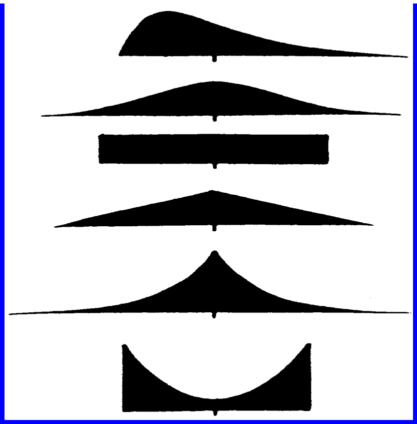
Variation (Dispersion)



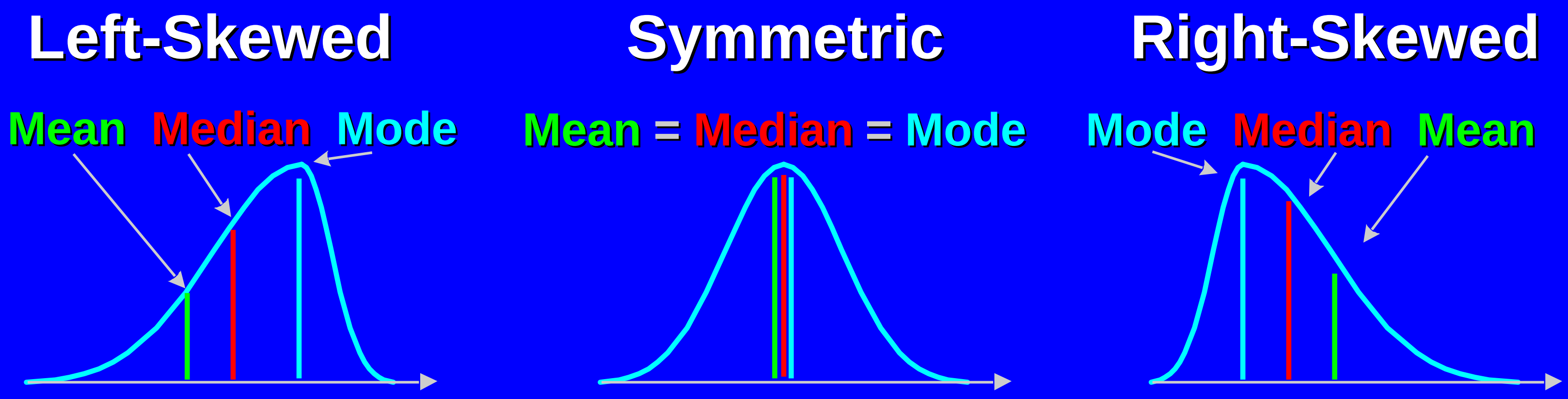
Shape



Shape



- 1. Describes How Data Are Distributed
- 2. Measures of Shape
  - – Skew = Symmetry



Left-Skewed

Mean Median Mode

Symmetric

Mean = Median = Mode

Right-Skewed

Mode Median Mean

Definitions

- **Mean:** The Mean of a quantitative dataset is the sum of the observations in the dataset divided by the number of observations in the dataset.
- **Median:** The Median **(m)** of a quantitative dataset is the middle number when the observations are arranged in ascending order.
- **Mode:** The Mode of a datset is the observation that occurs most frequently in the dataset.

Procedure to calculate the median: *M*

1. Arrange all observations in order of size, from smallest to largest.

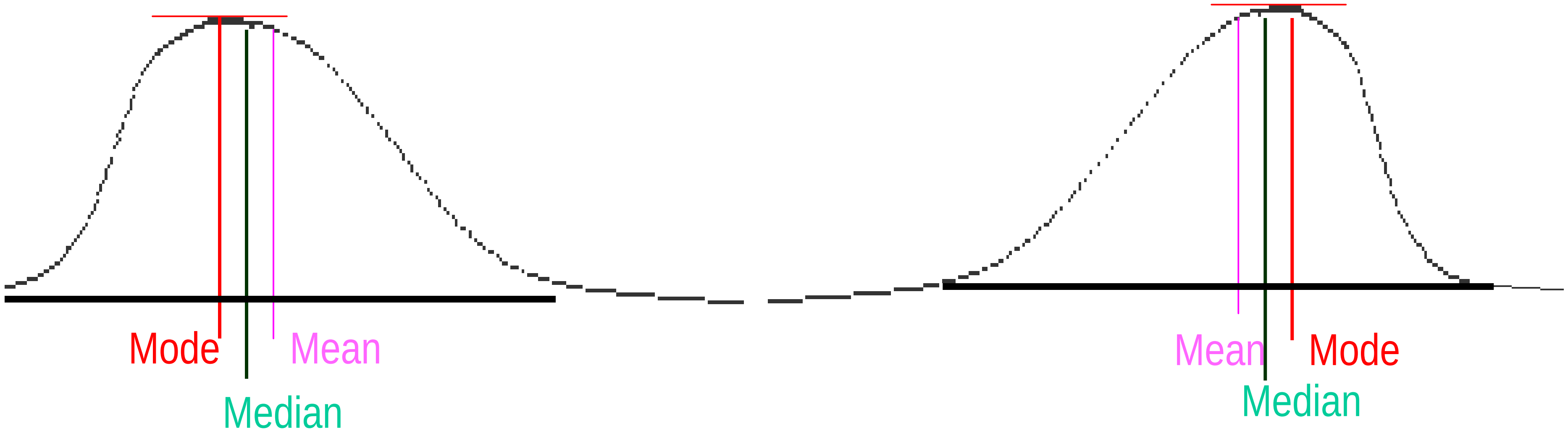
2. If the number of observations, *n*, is odd, the median *M* is the center observation in the ordered list.

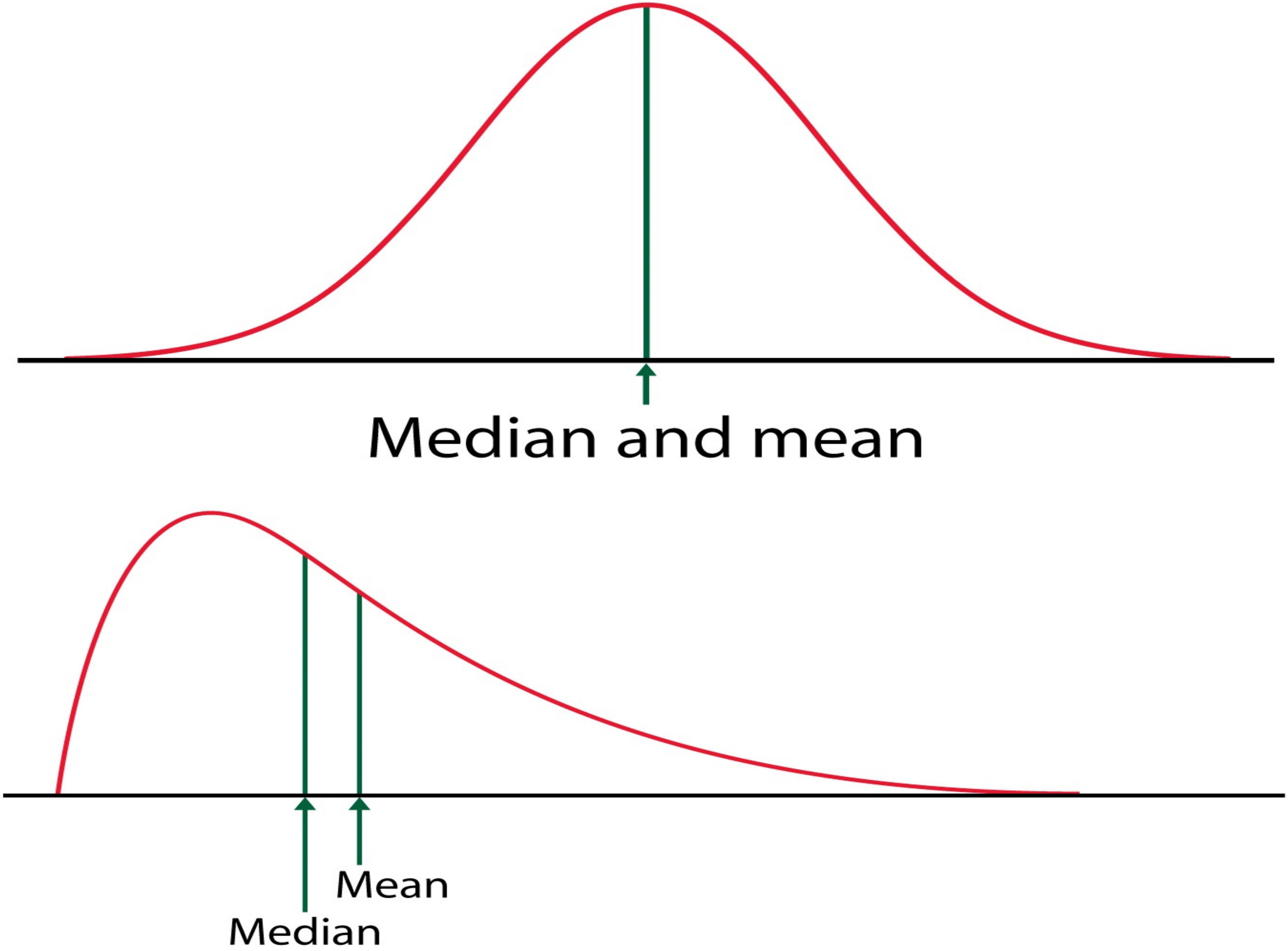
3. If *n* is even, then *M* is the mean of the two center observations in the ordered list.

Note: $(n+1)/2$ is the location of the median, not the median itself.

Mean, Median and Mode

- If the distribution is exactly symmetric, the mean, the median and the mode are exactly the same.
- If the distribution is skewed, the three measures differ.




Median and mean
Mean
Median

Mean vs. Median

- Mean:
  - easy to calculate
  - easy to work with algebraically
  - highly affected by outliers
  - Not a resistant measure
- Median:
  - can be time consuming to calculate
  - more resistant to a few extreme observations (sometimes outliers)
  - robust

The Mean, Median, and Mode

Ordered Listing of 28 Exam Scores

32, 55, 60, 61, 62, 64, 64, 68, 73, 75, 75, 76, 78, 78, 79, 79, 80, 80, 82, 83, 84, 85, 88, 90, 92, 93, 95, 98

- **Mean (numerical average): 76.04**
- **Median: 78.5 (halfway between 78 and 79)**
- **Mode (most common value): no single mode exists, many occur twice.**

Ordered Listing of 28 Exam Scores

32, 55, 60, 61, 62, 64, 64, 68, 73, 75, 75, 76, 78, 78, 79, 79, 80, 80, 82, 83, 84, 85, 88, 90, 92, 93, 95, 98

Outliers:

Outliers = values far removed from rest of data. Median of 78.5 higher than mean of 76.04 because one very low score (32) pulled down mean.

Variability:

How spread out are the values? A score of 80 compared to mean of 76 has different meaning if scores ranged from 72 to 80 versus 32 to 98.

Ordered Listing of 28 Exam Scores

32, 55, 60, 61, 62, 64, 64, 68, 73, 75, 75, 76, 78, 78, 79, 79, 80, 80, 82, 83, 84, 85, 88, 90, 92, 93, 95, 98

Minimum, Maximum and Range:

Range = max – min = 98 – 32 = 66 points. Other variability measures include interquartile range and standard deviation.

Shape:

Are most values clumped in middle with values tailing off at each end? Are there two distinct groupings? Pictures of data will provide this info.

7.2 Picturing Data: Stemplots and Histograms

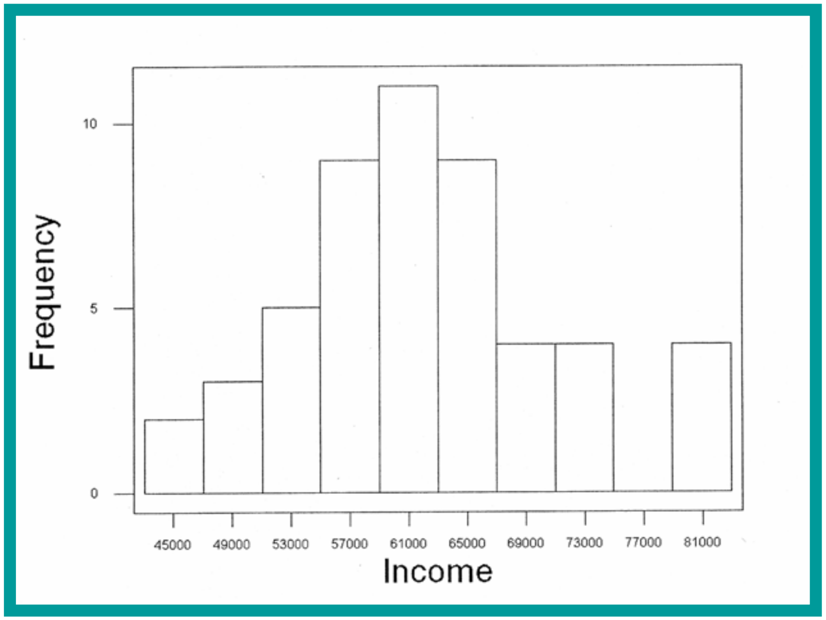
Stemplot: quick and easy way to order numbers and get picture of shape.

Stemplot for Exam Scores

3|2
4|
5|5
6|024418
7|56598398
8|5430820
9|53208

Example: 3|2 = 32

Histogram: better for larger data sets, also provides picture of shape.



Creating a Stemplot

Step 1: Create the Stems

Divide range of data into equal units to be used on **stem**. Have 6 – 15 stem values, representing equally spaced intervals.

Step 1: Creating the stem

3|
4|
5|
6|
7|
8|
9|

Example: each of the 7 stems represents a range of 10 points in test scores

Creating a Stemplot

Step 2: Attach the Leaves

Attach a **leaf** to represent each data point. Next digit in number used as leaf; drop remaining digits.

Example: Exam Scores

75, 95, 60, 93, …

First 4 scores attached.

Step 2: Attaching leaves

```
3|
4|
5|
6|0
7|5
8|
9|53
```

Optional Step: order leaves on each branch.

Further Details for Creating Stemplots

Splitting Stems:

Reusing digits two or five times.

Stemplot A:

```
5|4
5|789
6|023344
6|55567789
7|00124
7|58
```

Stemplot B:

```
5|4
5|7
5|89
6|0
6|233
6|44555
6|677
6|89
7|001
7|2
7|45
7|
7|8
```

Two times:
1st stem = leaves 0 to 4
2nd stem = leaves 5 to 9

Five times:
1st stem = leaves 0 and 1
2nd stem =leaves 2 and 3, etc.

Example 1: Stemplot of Median Income for Families of Four

Median incomes range from $46,596 (New Mexico) to $82,879 (Maryland).

Stems: 4 to 8, reusing two times with leaves truncated to $1,000s. Note leaves have been ordered.

Example:
$46,596 would be truncated to 46,000 and shown as 4|6

Stemplot of Median Incomes:

```
4|66789
5|11344
5|56666688899999
6|011112334
6|556666789
7|01223
7|
8|0022
```

Example: 4|6 = $46,xxx

Source: Federal Registry, April 15, 2003

Obtaining Info from the Stemplot

Determine shape, identify outliers, locate center.

Pulse Rates:

5|4
5|789
6|023344
6|55567789
7|00124
7|58

Bell-shape
Centered mid 60’s
no outliers

Exam Scores

3|2
4|
5|5
6|024418
7|56598398
8|5430820
9|53208

Outlier of 32.
Apart from 55,
rest uniform from
the 60’s to 90’s.

Median Incomes:

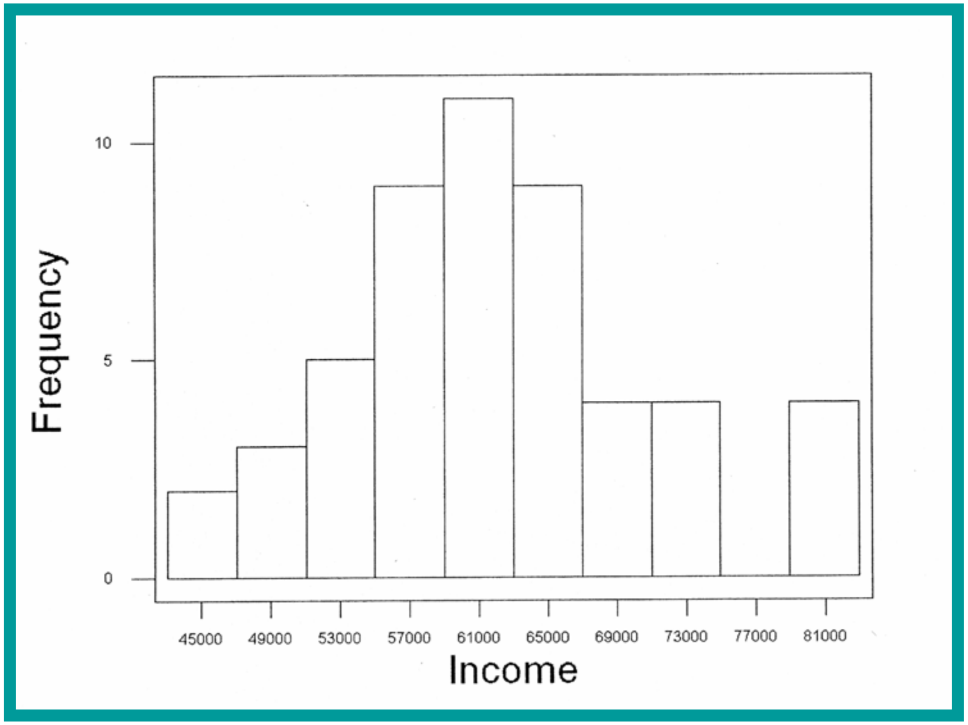
4|66789
5|11344
5|566666688899999
6|011112334
6|556666789
7|01223
7|
8|0022

Wide range with 4
unusually high values.
Rest bell-shape around
high $50,000s.

Creating a Histogram

- Divide range of data into intervals.
- Count how many values fall into each interval.
- Draw bar over each interval with height = count (or proportion).

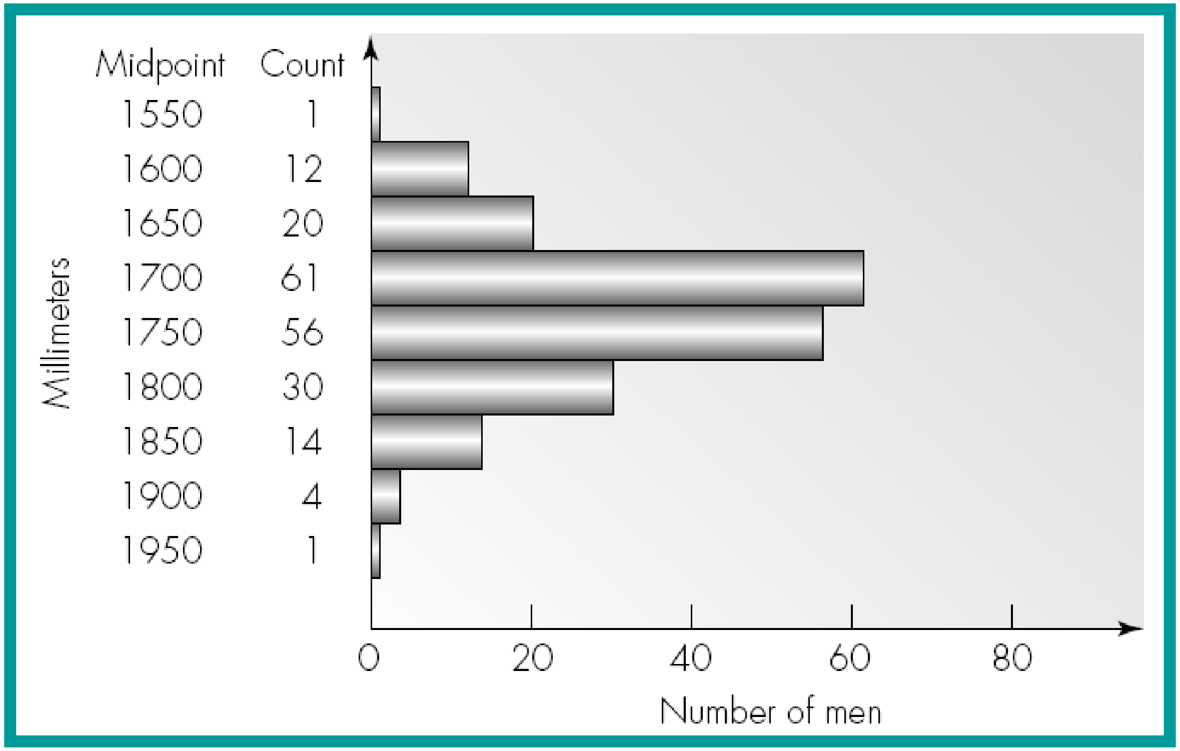
Histogram of Median Family Income Data



- **Histogram for Discrete Numerical Data**
- **1.** Draw a horizontal X-axis and on it mark the possible values taken by the observations
- **2.** Draw a vertical Y-axis marked with either relative frequencies or frequencies
- **3.** Above each possible value on the X-axis draw a rectangle centred on the value with width 1 and height equal to the relative frequency or frequency of that value.

Example 2: Heights of British Males

Heights of 199 randomly selected British men, in millimeters. Bell-shaped, centered in the mid-1700s mm with no outliers.

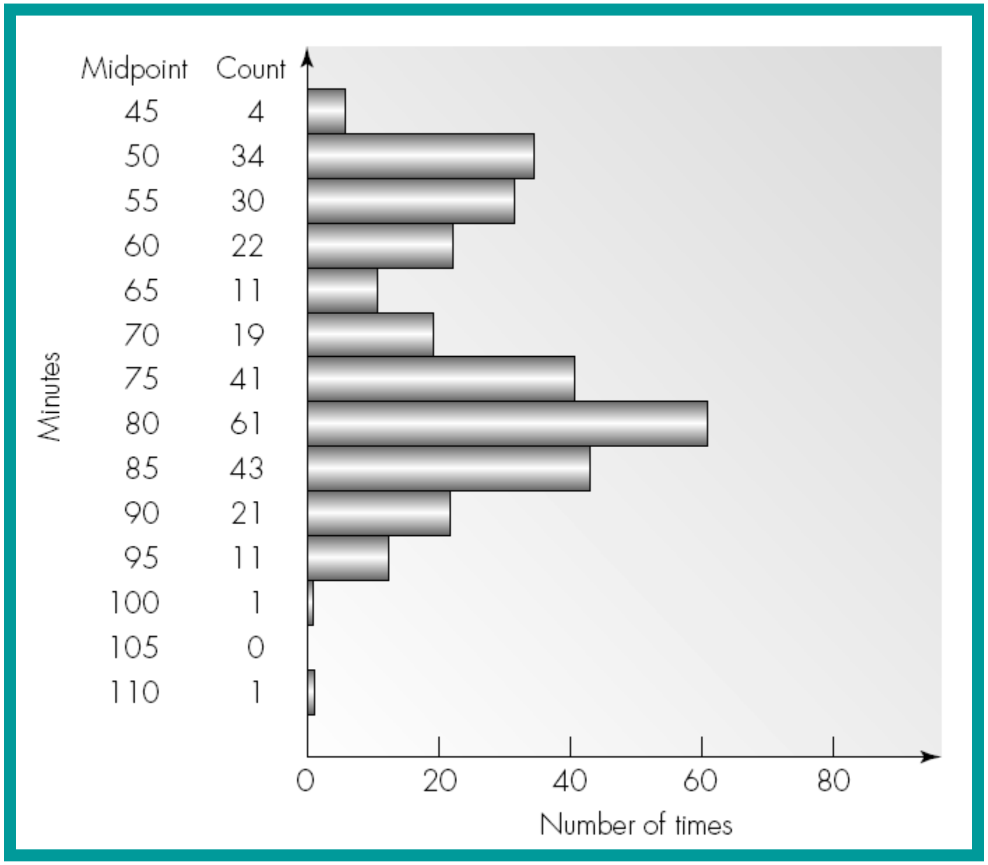


Midpoint	Count
1550	1
1600	12
1650	20
1700	61
1750	56
1800	30
1850	14
1900	4
1950	1

Source: Marsh, 1988, p. 315; data reproduced in Hand et al., 1994, pp. 179-183

Example 3: The Old Faithful Geyser

Times between eruptions of the
Old Faithful geyser.
Two clusters,
one around 50 min.,
other around 80 min.



Midpoint	Count
45	4
50	34
55	30
60	22
65	11
70	19
75	41
80	61
85	43
90	21
95	11
100	1
105	0
110	1

Source: Hand et al., 1994

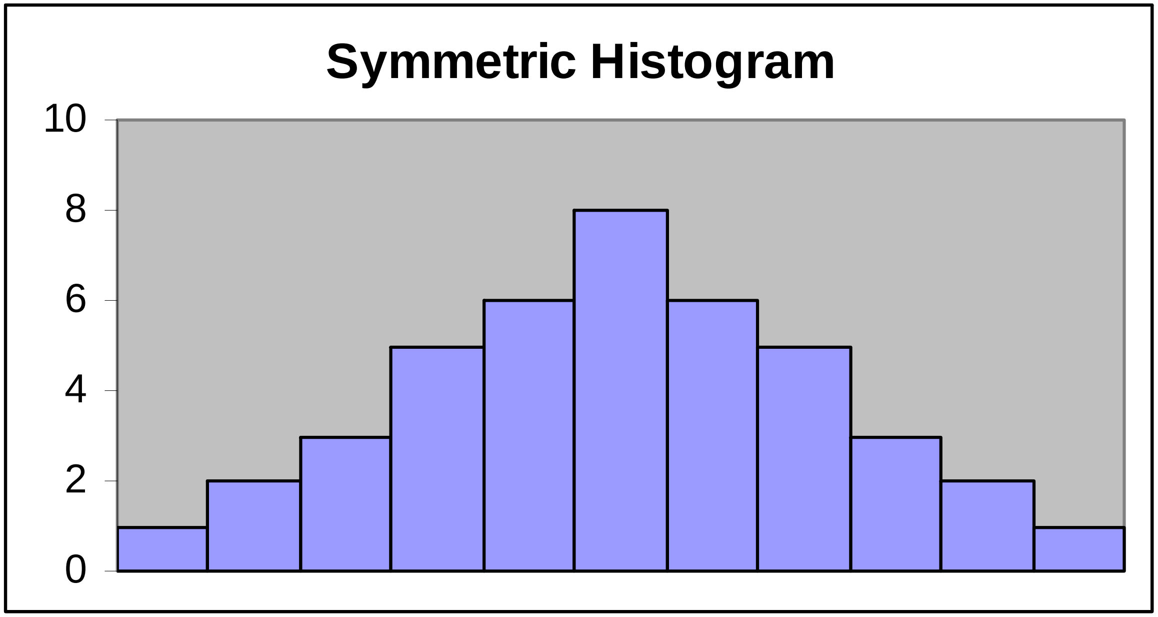
Defining a Common Language about Shape

- **Symmetric:** if draw line through center, picture on one side would be mirror image of picture on other side. *Example*: bell-shaped data set.
- **Unimodal:** single prominent peak
- **Bimodal:** two prominent peaks
- **Skewed to the Right:** *higher* values more spread out than lower values
- **Skewed to the Left:** *lower* values more spread out and higher ones tend to be clumped

Defining a Common Language About Shape

When investigators speak about the "shape" of the data, they are referring to the shape of the histogram resulting from the data.

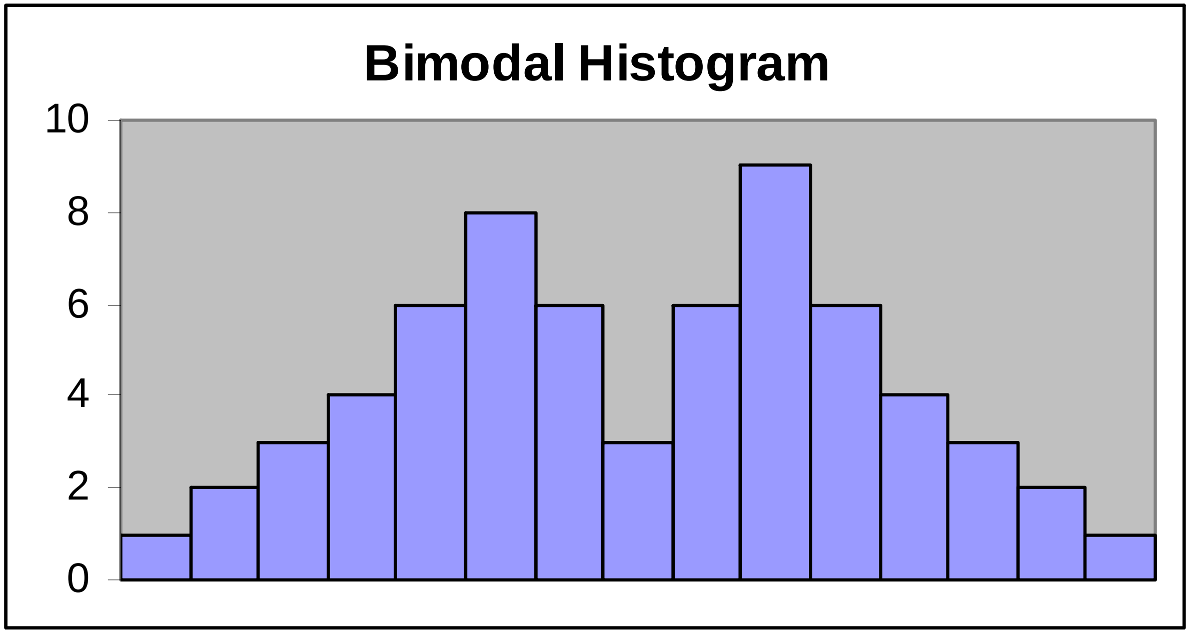
Symmetric Data Sets: A data set for which the histogram is (approximately) symmetric.



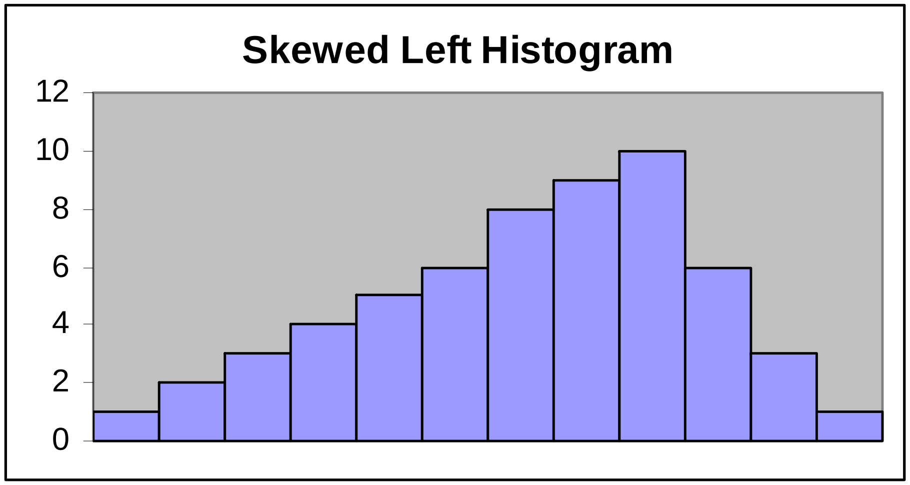
Unimodal or Bimodal

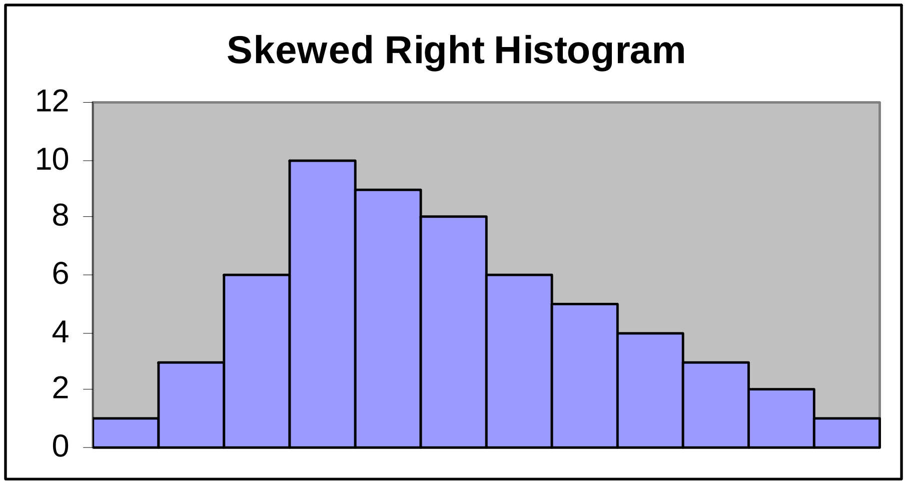
A data set is referred to as Unimodal if there is a single prominent peak in the histogram. An example is the Symmetric histogram.

A data set is referred to as Bimodal is there are two prominent peaks in the histogram.



A Skewed Data Set is one that is basically unimodal but is substantially off from being bell-shaped.

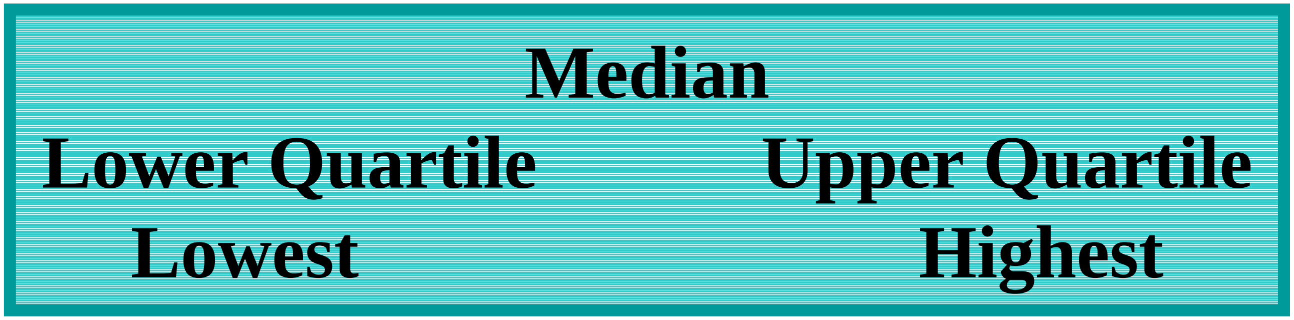




NOTE: The direction of the skew is in the direction of the long tail.

7.3 Five Useful Numbers: A Summary

The five-number summary display



Median

Lower Quartile Upper Quartile

Lowest Highest

- **Lowest** = Minimum
- **Highest** = Maximum
- **Median** = number such that half of the values are at or above it and half are at or below it (middle value or average of two middle numbers in ordered list).
- **Quartiles** = medians of the two halves.

Five-Number Summary for Income

***n* = 51 observations**

- **Lowest:** $46,xxx => $46,596
- **Highest:** $82,xxx => $82,879
- **Median:** (51+1)/2 => 26th value $61,xxx => $61,036
- **Quartiles:** Lower quartile = median of lower 25 values => 13th value, $56,xxx => $56,067; Upper quartile = median of upper 25 values => 13th value, $66,xxx => $66,507

Median Incomes:

4|66789
5|11344
5|566666688899999
6|011112334
6|556666789
7|01223
7|
8|0022

Five-number summary for family income

$61,036	
$56,067	**$66,507**
$46,596	**$82,879**

Provides center and spread. Can compare gaps between extremes and quartiles, gaps between quartiles and median.

Example: Give the 5 number summary for the following list.
3, 5, 2, 1, 9, 7, 8, 2, 4, 5, 8, 3, 0

FIRST ORDER THE DATA!!!!

0 1 2 2 3 3 4 5 5 7 8 8 9

Min=0

Max= 9

Median = 4

Q1 = 2

Q3 = 7

Five Number Summary:

Min =0

Q1 = 2

Median = 4

Q3 = 7

Max = 9

Percentiles (Quantiles)

- Percentiles (Quantiles) are derived from the ordered data values.

- The *p*th percentile (also called the *p*% quantile) is the value such that *p* percent of the observations fall at or below it.

- The median = the 50th percentile.

Quartiles

- The sample quartiles are the values that divide the sorted sample into quarters, just as the median divides it into half.

- The most commonly used quantiles are
  - The median M = 50th percentile
  - The first quartile Q1 = 25th percentile
  - The third quartile Q3 = 75th percentile

Calculations of Quartiles

- The first quartile Q1 is the median of the observations who are less than the overall median.

- The third quartile Q3 is the median of the observations who are greater than the overall median.

7.4 Boxplots

Visual picture of the five-number summary

Example 5: How much do statistics students sleep?

190 statistics students asked how many hours they slept the night before (a Tuesday night).

Five-number summary for number of hours of sleep

	7	
6		**8**
3		**16**

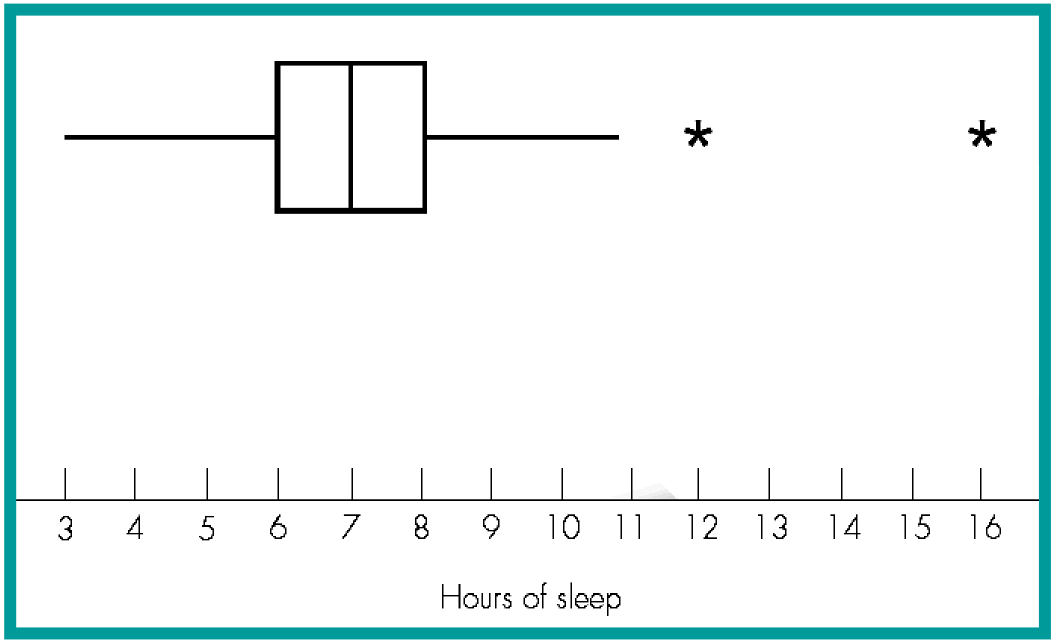
Two students reported 16 hours; the max for the remaining 188 students was 12 hours.

Creating a Boxplot

1. Draw horizontal (or vertical) line, label it with values from lowest to highest in data.
2. Draw rectangle (box) with ends at quartiles.
3. Draw line in box at value of median.
4. Compute IQR = distance between quartiles.
5. Compute 1.5(IQR); outlier is any value more than this distance from closest quartile.
6. Draw line (whisker) from each end of box extending to farthest data value that is not an outlier. (If no outlier, then to min and max.)
7. Draw asterisks to indicate the outliers.

Creating a Boxplot for Sleep Hours

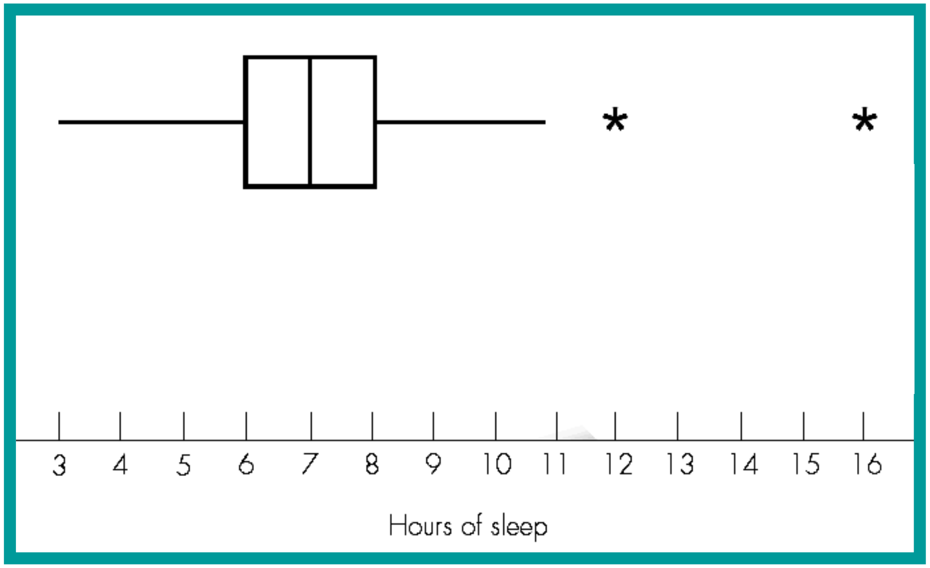
1. Draw horizontal line and label it from 3 to 16.
2. Draw rectangle (box) with ends at 6 and 8.
3. Draw line in box at median of 7.
4. Compute IQR = 8 – 6 = 2.
5. Compute 1.5(IQR) = 1.5(2) = 3; outlier is any value below 6 – 3 = 3, or above 8 + 3 = 11.
6. Draw line from each end of box extending down to 3 but up to 11.
7. Draw asterisks at outliers of 12 and 16 hours.



Interpreting Boxplots

- Divide the data into fourths.
- Easily identify outliers.
- Useful for comparing two or more groups.

Outlier: any value more than 1.5(IQR) beyond closest quartile.



¼ of students slept between 3 and 6 hours, ¼ slept between 6 and 7 hours, ¼ slept between 7 and 8 hours, and final ¼ slept between 8 and 16 hours

Example 6: Who Are Those Crazy Drivers?

What's the fastest you have ever driven a car? ____ mph.

Males (87 Students)

110

95 **120**

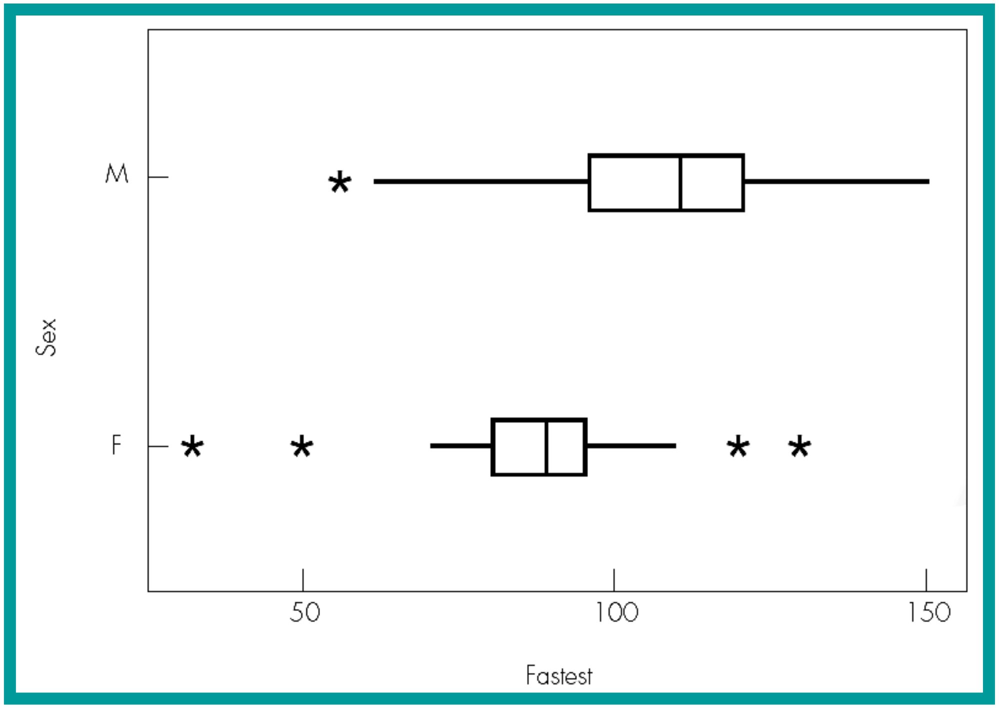
55 **150**

Females (102 Students)

89

80 **95**

30 **130**



- About 75% of men have driven 95 mph or faster, but only about 25% of women have done so.
- Except for few outliers (120 and 130), all women's max speeds are close to or below the median speed for men.

7.5 Traditional Measures:

Mean, Variance, and Standard Deviation

- **Mean:** represents center
- **Standard Deviation:** represents spread or variability in the values;
- **Variance = (standard deviation)^2**

Mean and standard deviation most useful for *symmetric* sets of data with *no outliers*.

The Mean and When to Use It

Mean most useful for symmetric data sets with no outliers.

Examples:

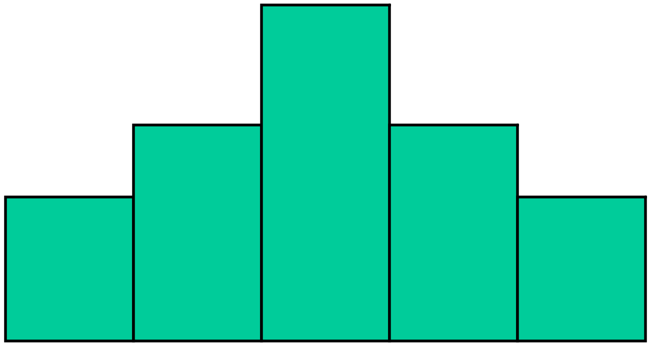
- Student taking four classes. Class sizes are 20, 25, 35, and 200. What is the typical class size? Median is 30. Mean is 280/4 = 70 (distorted by the one large size of 200 students).
- Incomes or prices of things often skewed to the right with some large outliers. Mean is generally distorted and is larger than the median.
- Distribution of British male heights was roughly symmetric. Mean height is 1732.5 mm and median height is 1725 mm.

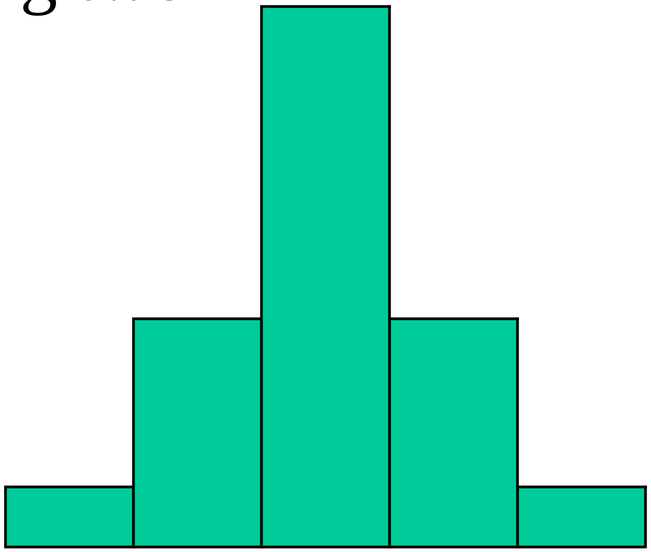
Variability: Why do we need "Spread"?

- Knowing the center of a distribution alone is not a good enough description of the data.
  - Two basketball players with the same shooting percentage may be very different in terms of consistency.
  - Two companies may have the same average salary, but very different distributions.
- We need to know the spread, or the variability of the values.

Numerical Measures of Variability

- When we want to describe a dataset providing a measure of the centre of that dataset is only part of the story. Consider the following two distributions:





- Both of these distributions are symmetric and
- meanA = meanB, modeA=modeB and medianA=medianB. However these two distributions are obviously different, the data in A is quite spread out compared to the data in B.
- This spread is technically called variability and we will now examine how best to measure it.

The Standard Deviation and Variance

Consider two sets of numbers, both with mean of 100.

Numbers	Mean	Standard Deviation
100, 100, 100, 100, 100	100	0
90, 90, 100, 110, 110	100	10

- **First** set of numbers has **no spread** or variability at all.
- **Second** set has some spread to it; **on average, the numbers are about 10 points away from the mean**.

The standard deviation is roughly the average distance of the observed values from their mean.

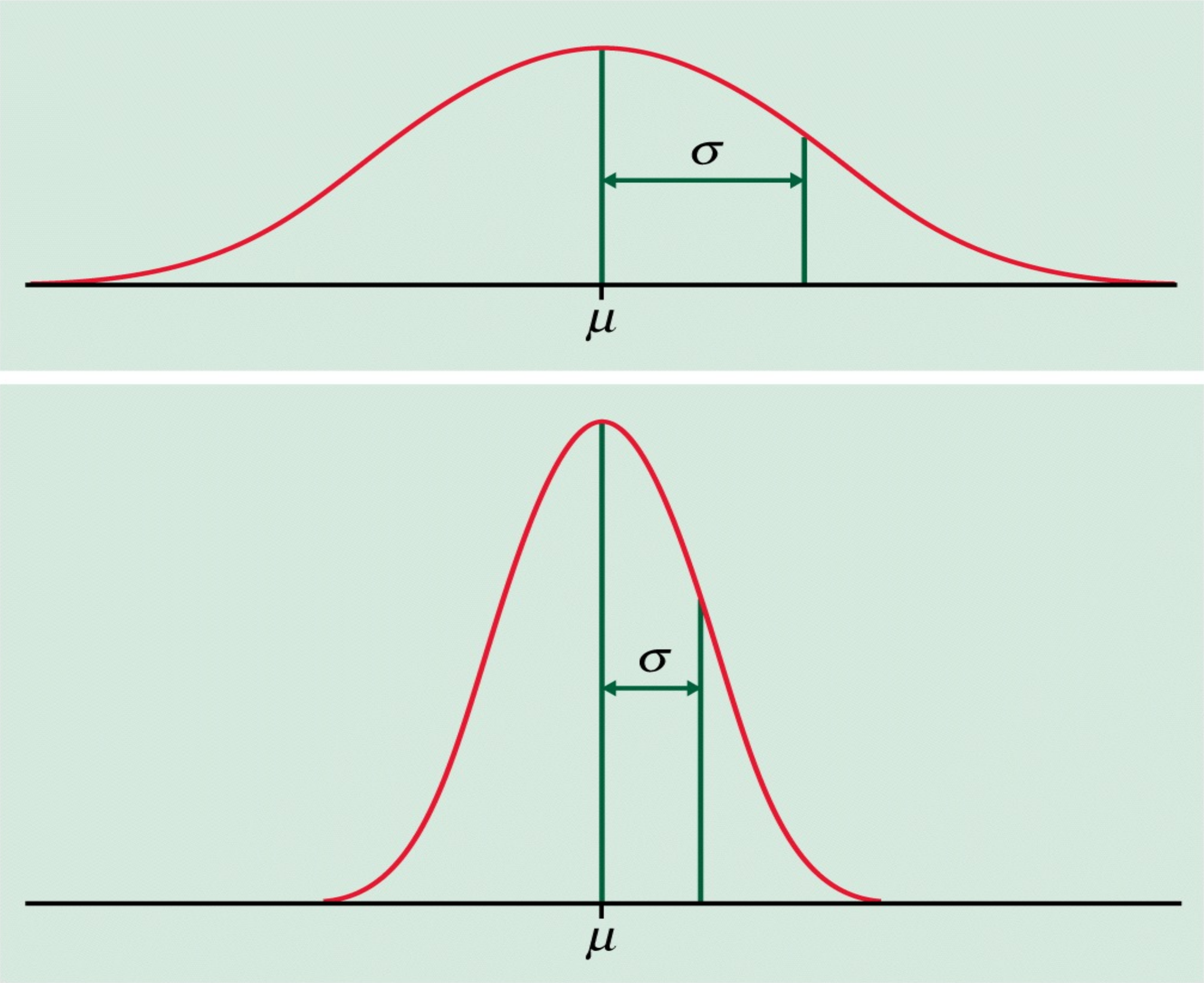
Computing the Standard Deviation

1. Find the mean.
2. Find the deviation of each value from the mean.
   Deviation = value – mean.
3. Square the deviations.
4. Sum the squared deviations.
5. Divide the sum by (the number of values) – 1, resulting in the variance.
6. Take the square root of the variance.
   The result is the standard deviation.

Computing the Standard Deviation

Try it for the set of values: 90, 90, 100, 110, 110.

1. The mean is 100.
2. The deviations are -10, -10, 0, 10, 10.
3. The squared deviations are 100, 100, 0, 100, 100.
4. The sum of the squared deviations is 400.
5. The variance = 400/(5 – 1) = 400/4 = 100.
6. The standard deviation is the square root of 100, or 10.


σ
μ
σ
μ

7.6 Caution: Being Average Isn’t Normal

Common mistake to confuse “average” with “normal”.

Example 7: How much hotter than normal is normal?

> “October came in like a dragon Monday, hitting 101 degrees in Sacramento by late afternoon. That temperature tied the record high for Oct. 1 set in 1980 – and was 17 degrees *higher than normal for the date*. (Korber, 2001, italics added.)”

Article had thermometer showing “normal high” for the day was 84 degrees. High temperature for Oct. 1st is quite variable, from 70s to 90s. While 101 was a record high, it was not “17 degrees higher than normal” if “normal” includes the range of possibilities likely to occur on that date.

Case Study 7.1: *Detecting Exam Cheating with a Histogram*

Details:

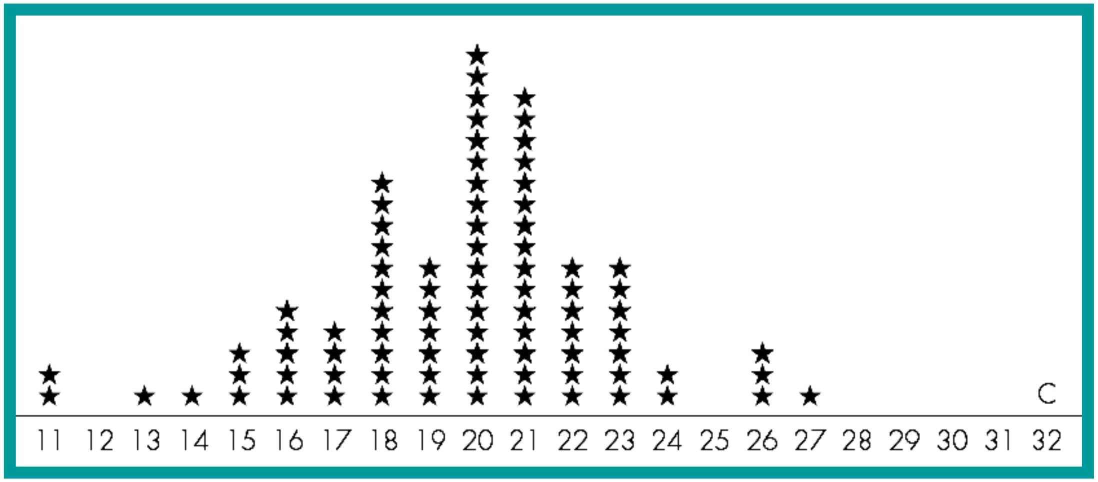
- Summer of 1984, class of 88 students taking 40-question multiple-choice exam.
- Student C accused of copying answers from Student A.
- Of 16 questions *missed* by both A and C, both made same wrong guess on 13 of them.
- Prosecution argued match that close by chance alone very unlikely; Student C found guilty.
- Case challenged. Prosecution unreasonably assumed any of four wrong answers on a missed question equally likely to be chosen.

Source: Boland and Proschan, Summer 1991, pp. 10-14.

Case Study 7.1: *Detecting Exam Cheating with a Histogram*

Second Trial:

For each student (except A), counted how many of his or her 40 answers matched the answers on A’s paper. Histogram shows Student C as obvious outlier. Quite unusual for C to match A’s answers so well without some explanation other than chance.



Defense argued based on histogram, A could have been copying from C. Guilty verdict overturned. However, Student C was seen looking at Student A’s paper – jury forgot to account for that.

For Those Who Like Formulas

The Data

n = number of observations

x_i = the *ith* observation, $i = 1,2, \ldots, n$

The Mean

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

The Variance

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The Computational Formula for the Variance

$$s^2 = \frac{1}{(n-1)}\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)$$

The Standard Deviation

Use either formula to find s^2; then simply take the square root to get the standard deviation s.