Statistics

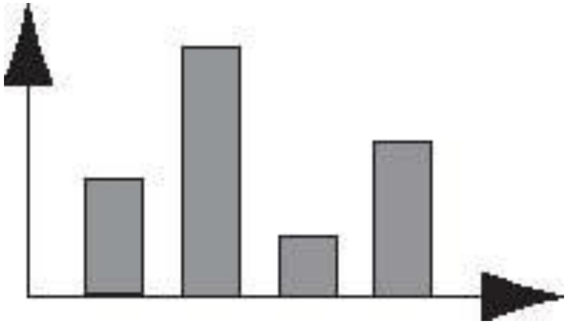
Chapter 2: Descriptive Statistics

Where we’ve been

- Descriptive and Inferential Statistics
- Randomness and Variability
- Experimental unit, variable
- uni/bi/multi-variate data
- Population, census, sample
- Measure of reliability
- Qualitative, Quantitative: Discrete, Continuous
- Sources: Published, Observational Study, Designed Experiment
- Errors: Selection, Response, Nonresponse, Measurement

Where We're Going

- Describe Data by Using Graphs
- Describe Data by Using Numerical Measures
  - Summation Notation
  - Central Tendencies
  - Variability
  - The Standard Deviation
  - Relative Standing
  - Outliers
  - Graphing Bivariate Relationships
  - Distorting the Truth



2.1: Describing Qualitative Data

- Qualitative Data are nonnumerical
- Summarized in two ways:
  - Class Frequency
  - Class Relative Frequency

2.1: Describing Qualitative Data

- Class Frequency
  - A class is one of the categories into which qualitative data can be classified
  - Class frequency is the number of observations in the data set that fall into a particular class

2.1: Describing Qualitative Data

Example: Adult Aphasia

Subject	Type of Aphasia	Subject	Type of Aphasia
1	Broca's	12	Broca's
2	Anomic	13	Anomic
3	Anomic	14	Broca's
4	Conduction	15	Anomic
5	Broca's	16	Anomic
6	Conduction	17	Anomic
7	Conduction	18	Conduction
8	Anomic	19	Broca's
9	Conduction	20	Anomic
10	Anomic	21	Conduction
11	Conduction	22	Anomic

2.1: Describing Qualitative Data

Example: Adult Aphasia

Type of Aphasia	Frequency
Anomic	10
Broca’s	5
Conduction	7
Total	22

2.1: Describing Qualitative Data

- Class Relative Frequency
  - Class frequency divided by the total number of observations in the data set
- Class Percentage
  - Class relative frequency multiplied by 100

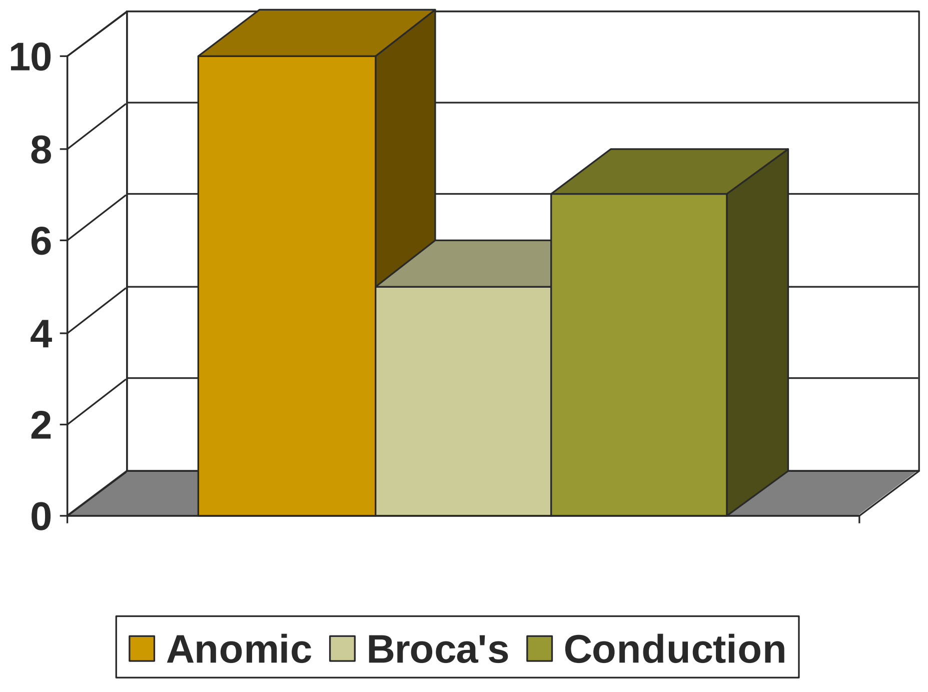
2.1: Describing Qualitative Data

Example: Adult Aphasia

Type of Aphasia	Relative Frequency	Class Percentage
Anomic	10/22 = .455	45.5%
Broca's	5/22 = .227	22.7%
Conduction	7/22 = .318	31.8%
Total	22/22 = 1.00	100%

2.1: Describing Qualitative Data

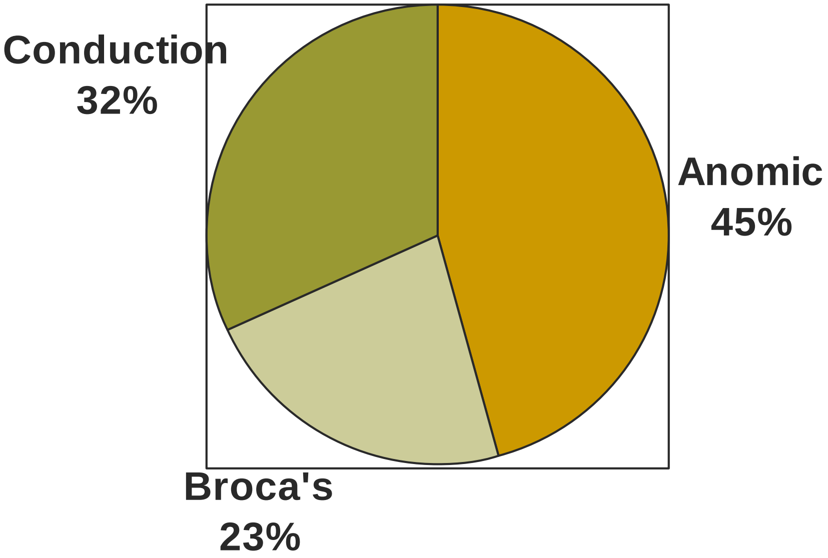
Example: Adult Aphasia



Bar Graph: The categories (classes) of the qualitative variable are represented by bars, where the height of each bar is either the class frequency, class relative frequency or class percentage.

2.1: Describing Qualitative Data

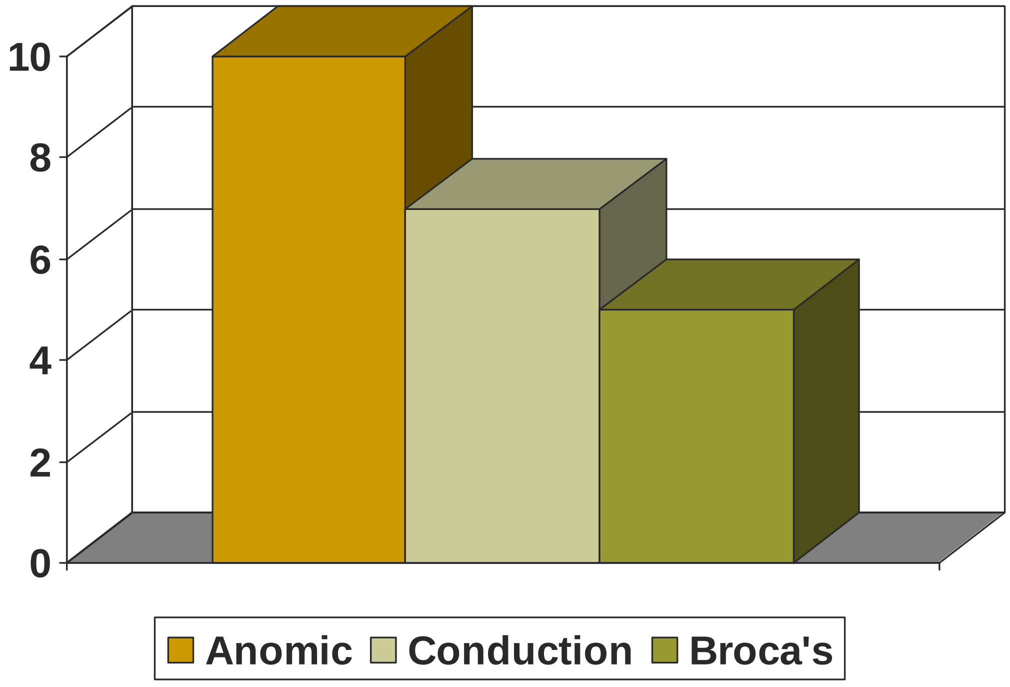
Example: Adult Aphasia



Pie Chart: The categories (classes) of the qualitative variable are represented by slices of a pie. The size of each slice is proportional to the class relative frequency.

2.1: Describing Qualitative Data

Example: Adult Aphasia



Pareto Diagram: A bar graph with the categories (classes) of the qualitative variable (i.e., the bars) arranged in height in descending order from left to right.

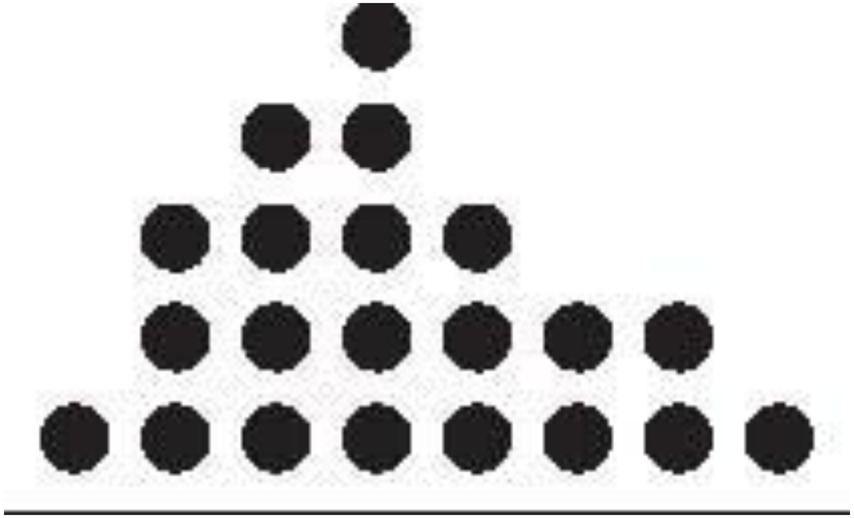
2.2: Graphical Methods for Describing Quantitative Data

- **Quantitative Data** are recorded on a meaningful numerical scale
  - Income
  - Sales
  - Population

2.2: Graphical Methods for Describing Quantitative Data

- Dot plots
- Stem-and-leaf diagrams
- Histograms

2.2: Graphical Methods for Describing Quantitative Data

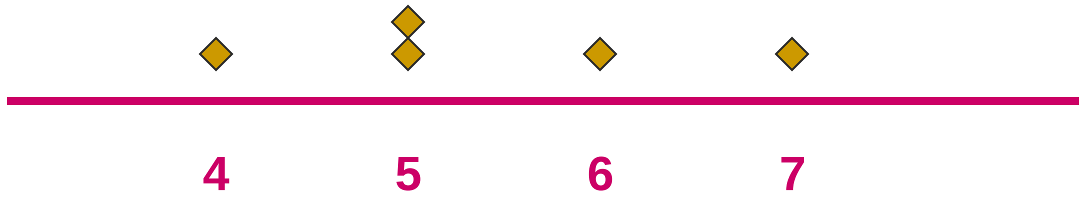


- **Dot plots** display a dot for each observation along a horizontal number line
  - Duplicate values are piled on top of each other
  - The dots reflect the shape of the distribution

2.2: Graphical Methods for Describing Quantitative Data

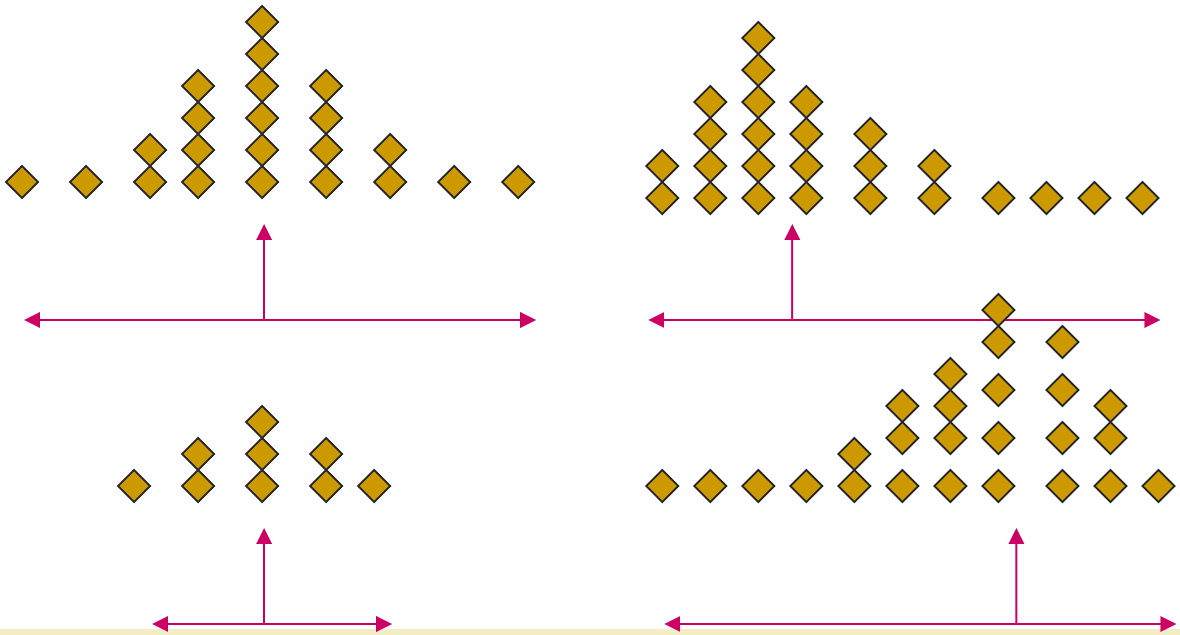
Dotplots

- The simplest graph for quantitative data
- Plots the measurements as points on a horizontal axis, stacking the points that duplicate existing points.
- **Example:** The set 4, 5, 5, 7, 6



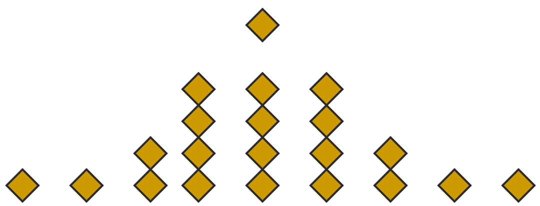
2.2: Graphical Methods for Describing Quantitative Data

Interpreting Graphs: Location and Spread

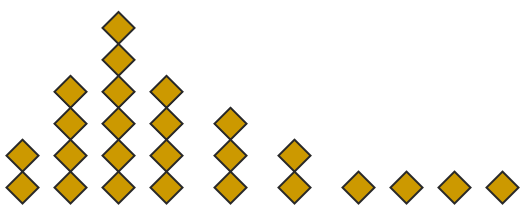


- Where is the data centered on the horizontal axis, and how does it spread out from the center?

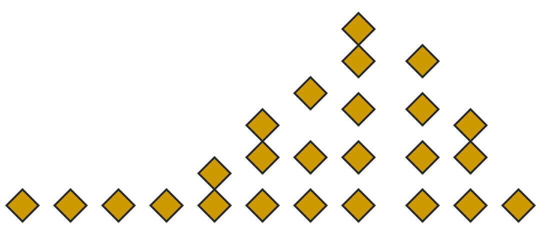
2.2: Graphical Methods for Describing Quantitative Data



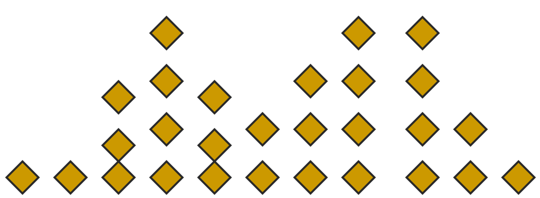
Mound shaped and symmetric (mirror images)



Skewed right: a few unusually large measurements



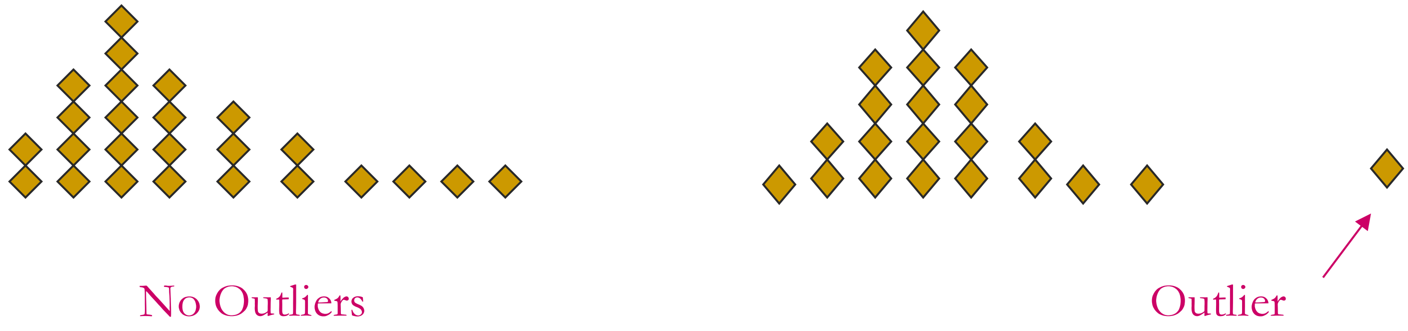
Skewed left: a few unusually small measurements



Bimodal: two local peaks

2.2: Graphical Methods for Describing Quantitative Data

Interpreting Graphs: Outliers

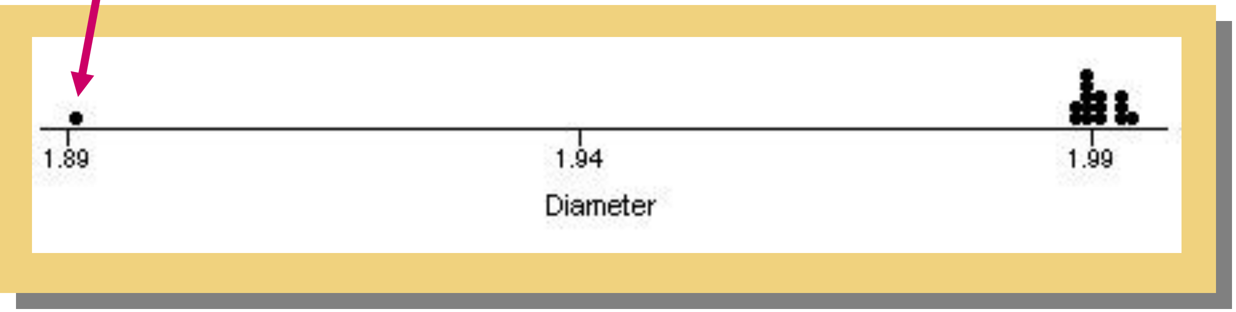


- Are there any strange or unusual measurements that stand out in the data set?

2.2: Graphical Methods for Describing Quantitative Data

- **Example:** A quality control process measures the diameter of a gear being made by a machine (cm). The technician records 15 diameters, but inadvertently makes a typing mistake on the second entry.

1.991	1.891	1.991	1.988	1.993	1.989	1.990	1.988
1.988	1.993	1.991	1.989	1.989	1.993	1.990	1.994



2.2: Graphical Methods for Describing Quantitative Data

- Dot Plots
  - Dots on a horizontal scale represent the values
  - Good for small data sets
- Stem-and-Leaf Displays
  - Divides values into “stems” and “leafs.”
  - Good for small data sets

2.2: Graphical Methods for Describing Quantitative Data

1	3
2	2489
3	126678
4	37
5	2

- A **Stem-and-Leaf Display** shows the number of observations that share a common value (the stem) and the precise value of each observation (the leaf)

Example: 13, 22, 24, 28, 29, 31, 32, 36, 36, 37, 38, 43, 47, 52.

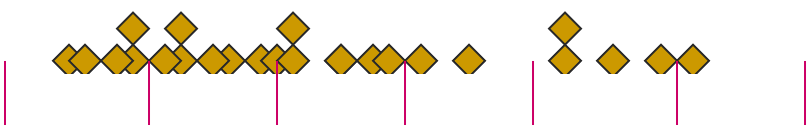
Stem-and-Leaf Display of these observations is shown above.

2.2: Graphical Methods for Describing Quantitative Data

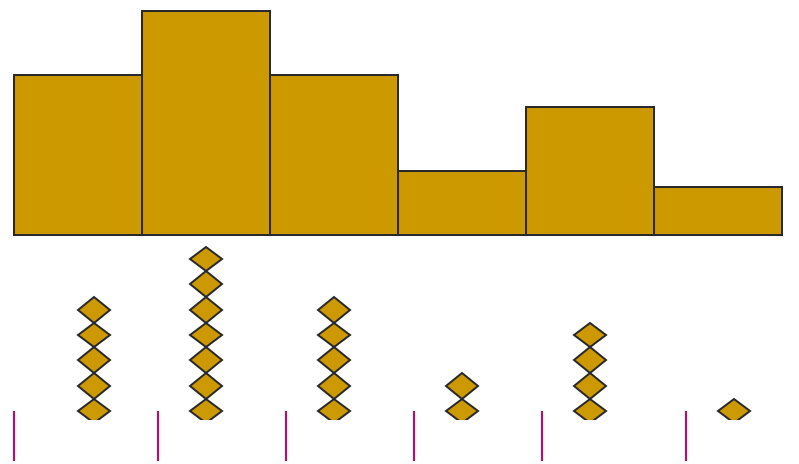
- Dot Plots and Stem-and-Leaf Displays are cumbersome for larger data sets
- Histograms
  - Frequencies or relative frequencies are shown for each class interval
  - Useful for larger data sets, but the precise values of observations are not shown

2.2: Graphical Methods for Describing Quantitative Data

- A **relative frequency histogram** for a quantitative data set is a bar graph in which the height of the bar shows “how often” (measured as a proportion or relative frequency) measurements fall in a particular class or subinterval.



Create intervals



Stack and draw bars

2.2: Graphical Methods for Describing Quantitative Data

- Divide the range of the data into **5-12 subintervals** of equal length.
- Calculate the **approximate width** of the subinterval as Range/number of subintervals.
- Round the approximate width up to a convenient value.
- Use the method of **left inclusion** including the left endpoint, but not the right in your tally.

2.2: Graphical Methods for Describing Quantitative Data

- Create a **statistical table** including the subintervals, their frequencies and relative frequencies.
- Draw the **relative frequency histogram** plotting the subintervals on the horizontal axis and the relative frequencies on the vertical axis.

2.2: Graphical Methods for Describing Quantitative Data

- The height of the bar represents
  - The **proportion** of measurements falling in that class or subinterval.
  - The **probability** that a single measurement, drawn at random from the set, will belong to that class or subinterval.

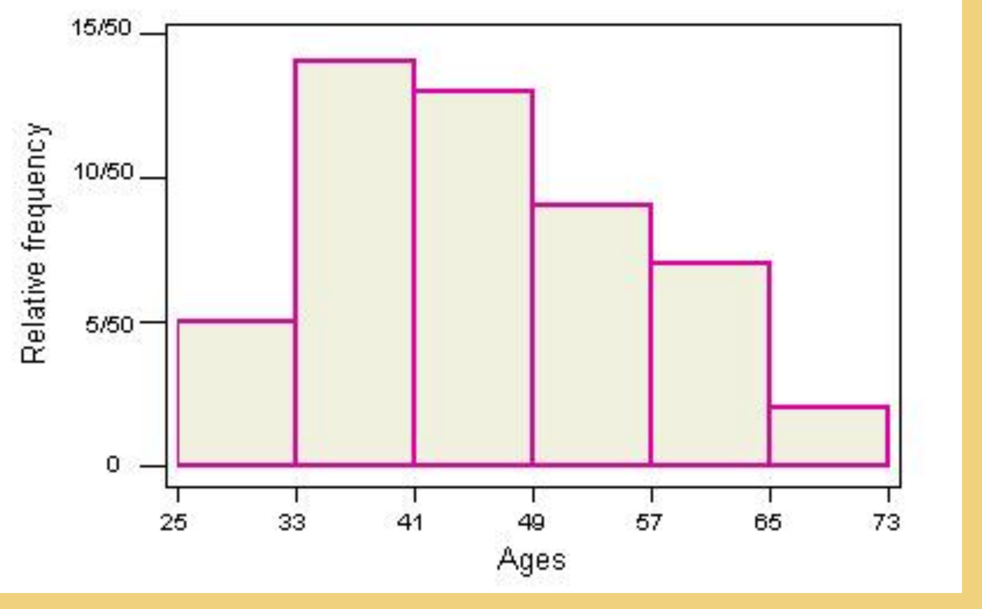
2.2: Graphical Methods for Describing Quantitative Data

The ages of 50 professors at a university.

34	48	**70**	63	52	52	35	50	37	43	53	43	52	44
42	31	36	48	43	**26**	58	62	49	34	48	53	39	45
34	59	34	66	40	59	36	41	35	36	62	34	38	28
43	50	30	43	32	44	58	53						

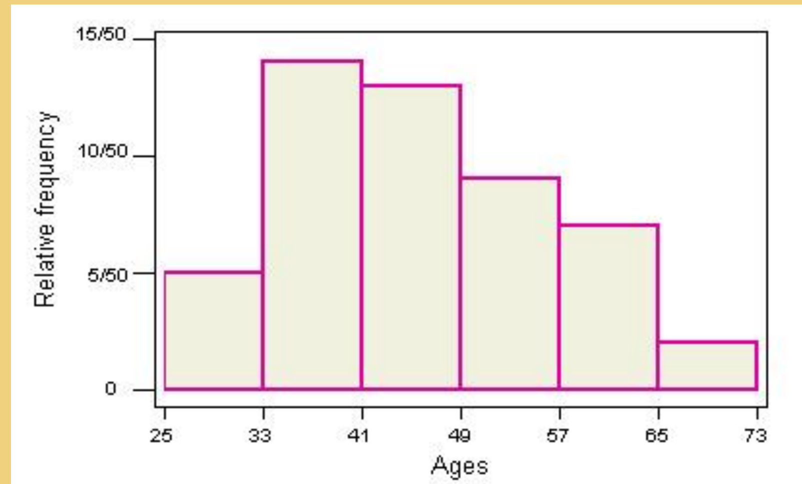
- We choose to use **6** intervals.
- Minimum class width **= (70 – 26)/6 = 7.33**
- Convenient class width **= 8**
- Use **6** classes of length **8**, starting at **25.**

Age	Tally	Frequency	Relative Frequency	Percent
25 to < 33	~~1111~~	5	5/50 = .10	10%
33 to < 41	~~1111~~ ~~1111~~ 1111	14	14/50 = .28	28%
41 to < 49	~~1111~~ ~~1111~~ 111	13	13/50 = .26	26%
49 to < 57	~~1111~~ 1111	9	9/50 = .18	18%
57 to < 65	~~1111~~ 11	7	7/50 = .14	14%
65 to < 73	11	2	2/50 = .04	4%



2.2: Graphical Methods for Describing Quantitative Data

Describing the Distribution



Shape? Skewed right

Outliers? No.

What proportion of professors are younger than 41?

(14 + 5)/50 = 19/50 = 0.38

What is the probability that a randomly selected professor is 49 or older?

(9 + 7 + 2)/50 = 18/50 = 0.36

2.2: Graphical Methods for Describing Quantitative Data

- How many classes?
  - <25 observations: 5-6 classes
  - 25-50 observations 7-14 classes
  - >50 observations 15-20 classes

2.3: Summation Notation

- Individual observations in a data set are denoted

$$x_1, x_2, x_3, x_4, \ldots x_n .$$

2.3: Summation Notation

- We use a summation symbol often:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$$

- This tells us to add all the values of variable x from the first (x_1) to the last (x_n).

- If $x_1 = 1$, $x_2 = 2$, $x_3 = 3$ and $x_4 = 4$,

$$\sum_{i=1}^{n} x_i = 1 + 2 + 3 + 4 = 10$$

2.3: Summation Notation

- Sometimes we will have to square the values before we add them:

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2$$

- Other times we will add them and then square the sum:

$$\left(\sum_{i=1}^{n} x_i \right)^2 = \left(x_1 + x_2 + x_3 + \ldots + x_n \right)^2$$

2.4: Numerical Measures of Central Tendency

- Summarizing data sets numerically
  - Are there certain values that seem more typical for the data?
  - How typical are they?

2.4: Numerical Measures of Central Tendency

- **Central tendency** is the value or values around which the data tend to cluster
- **Variability** shows how strongly the data cluster around that (those) value(s)

2.4: Numerical Measures of Central Tendency

- The **mean** of a set of quantitative data is the sum of the observed values divided by the number of values

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

2.4: Numerical Measures of Central Tendency

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- The mean of a *sample* is typically denoted by x-bar, but the *population mean* is denoted by the Greek symbol *μ*.

2.4: Numerical Measures of Central Tendency

- If $x_1 = 1$, $x_2 = 2$, $x_3 = 3$ and $x_4 = 4$,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = (1 + 2 + 3 + 4)/4 = 10/4 = 2.5$$

2.4: Numerical Measures of Central Tendency

- Exercise: Show that the mean minimizes the sum of squared deviations for a set of values.
- That is, given $(x_1, x_2, \ldots x_n)$, the value that minimizes $\sum_{i=1}^{n}(x_i - a)^2$ is $\mathrm{a} = \bar{x}$.

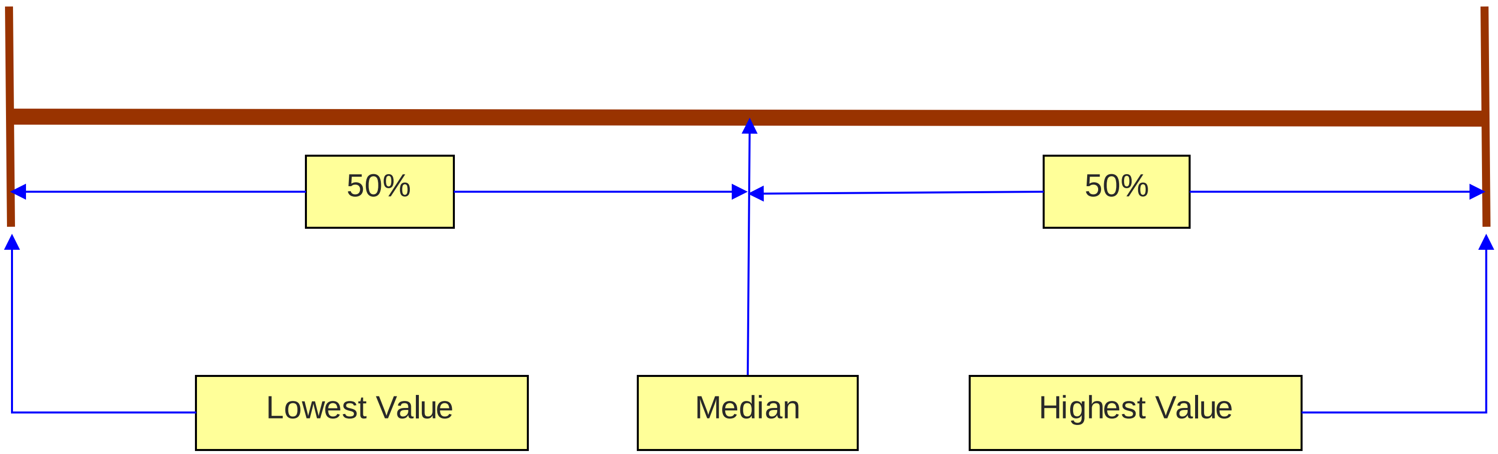
2.4: Numerical Measures of Central Tendency

- Exercise: Show that the mean minimizes the sum of squared deviations for a set of values.
- That is, given $(x_1, x_2, \dots x_n)$, the value that minimizes $\sum_{i=1}^{n}(x_i - a)^2$ is $a = \bar{x}$.
  1. Differentiate
  2. Add and subtract $\bar{x}$

2.4: Numerical Measures of Central Tendency

- The **median** of a set of quantitative data is the value which is located in the middle of the data, arranged from lowest to highest values (or vice versa), with 50% of the observations above and 50% below.

2.4: Numerical Measures of Central Tendency



2.4: Numerical Measures of Central Tendency

- Finding the Median, *M*:
  - Arrange the *n* measurements from smallest to largest
    - If *n* is odd, *M* is the middle number
    - If *n* is even, *M* is the average of the middle two numbers

2.4: Numerical Measures of Central Tendency

Examples

The set: 2, 4, 9, 8, 6, 5, 3 $n = 7$

Sort: 2, 3, 4, 5, 6, 8, 9

Position: $.5(n + 1) = .5(7 + 1) = 4^{th}$

Median = 4^{th} largest measurement =8

The set: 2, 4, 9, 8, 6, 5 $n = 6$

Sort: 2, 4, 5, 6, 8, 9

Position: $.5(n + 1) = .5(6 + 1) = 3.5^{th}$

Median = (5 + 6)/2 = 5.5 — average of the 3^{rd} and 4^{th} measurements

2.4: Numerical Measures of Central Tendency

The number of litres of milk purchased by 25 households:

0 0 1 1 1 1 1 2 2 2 2 2 2 2 2 2
3 3 3 3 3 4 4 4 5

- **Mean?**

$$\boxed{\bar{x} = \frac{\sum x_i}{n} = \frac{55}{25} = 2.2}$$

- **Median?**

$$\boxed{m = 2}$$

2.4: Numerical Measures of Central Tendency

- **The mean is more easily affected by extremely large or small values than the median.**
- In the previous example, if the consumption of the household buying 5 litres changes to 10 litres, the median remains same, but mean changes to 60/25=2.4

2.4: Numerical Measures of Central Tendency

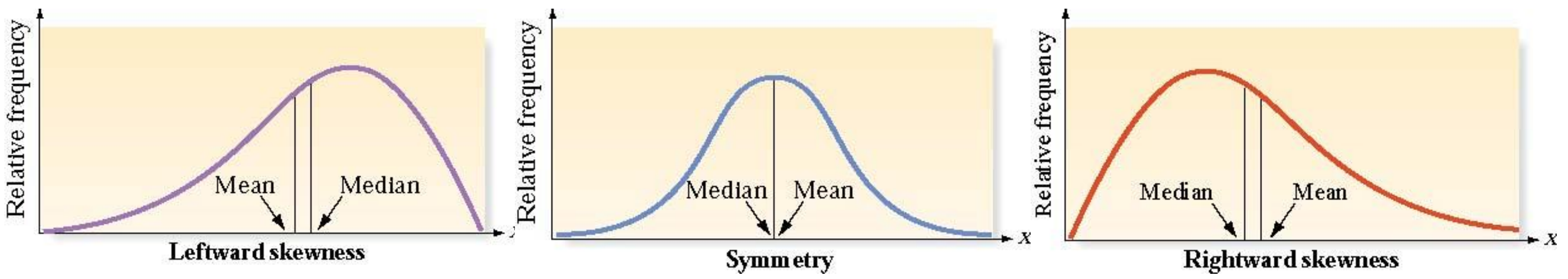
- The **mode** is the most frequently observed value.
- The **modal class** is the class with the highest relative frequency.
- For grouped data, the mode is the midpoint of the modal class.
- For the data on age of professors, the modal class is 33-40, hence the mode is 36.5.

2.4: Numerical Measures of Central Tendency

- Perfectly symmetric data set:
  - Mean = Median = Mode
- Extremely high value in the data set:
  - Mean > Median > Mode (Rightward skewness)
- Extremely low value in the data set:
  - Mean < Median < Mode (Leftward skewness)

2.4: Numerical Measures of Central Tendency

- A data set is **skewed** if one tail of the distribution has more extreme observations that the other tail.



2.5: Numerical Measures of Variability

- The mean, median and mode give us an idea of the central tendency, or where the “middle” of the data is.
- Variability gives us an idea of how spread out the data are around that middle. We shall discuss
  - Range
  - variance,
  - standard deviation
  - interquartile range.

2.5: Numerical Measures of Variability

- The **range** is equal to the largest measurement minus the smallest measurement.
  - Easy to compute, but not very informative
  - Considers only two observations (the smallest and largest)

2.5: Numerical Measures of Variability

- The **sample variance**, s^2, for a sample of n measurements is equal to the sum of the squared distances from the mean, divided by $(n - 1)$.

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

2.5: Numerical Measures of Variability

- The **sample standard deviation**, *s*, for a sample of *n* measurements is equal to the square root of the sample variance.

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

2.5: Numerical Measures of Variability

- Say a small data set consists of the measurements 1, 2 and 3.
  - μ = 2

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \left[(3-2)^2 + (2-2)^2 + (1-2)^2\right] / (3-1)$$

$$s^2 = \left(1^2 + 0^2 + 1^2\right) / 2 = 2/2 = 1$$

$$s = \sqrt{s^2} = \sqrt{1} = 1$$

2.5: Numerical Measures of Variability

- Greek letters are used for populations and Roman letters for samples

  s^2 = sample variance

  s = sample standard deviation

  σ^2 = population variance

  σ = population standard deviation

2.5: Numerical Measures of Variability

- The value of s is **ALWAYS** positive.
- The larger the value of s^2 or s, the larger the variability of the data set.
- **Why divide by n –1?**
  - The sample standard deviation s is often used to estimate the population standard deviation s. Dividing by n –1 gives us a better estimate of s.

2.5: Numerical Measures of Variability

- The **lower quartile (Q_1)** is the value of *x* which is larger than 25% and less than 75% of the ordered measurements.
- The **upper quartile (Q_3)** is the value of *x* which is larger than 75% and less than 25% of the ordered measurements.
- The range of the “middle 50%” of the measurements is the **interquartile range,**

$$\text{IQR} = \mathbf{Q_3 - Q_1}$$

2.5: Numerical Measures of Variability

- The **lower and upper quartiles (Q_1 and Q_3),** can be calculated as follows:
- The **position of Q_1** is **$.25(n + 1)$**
- The **position of Q_3** is **$.75(n + 1)$**

once the measurements have been ordered. If the positions are not integers, find the quartiles by interpolation.

2.5: Numerical Measures of Variability

The prices (in Rs 100) of 18 brands of walking shoes:

40 60 65 65 65 68 68 70 70

70 70 70 70 74 75 75 90 95

Position of Q_1 = .25(18 + 1) = 4.75

Position of Q_3 = .75(18 + 1) = 14.25

Q_1 is 3/4 of the way between the 4th and 5th ordered measurements, or

$Q_1 = 65 + .75(65 - 65) = 65.$

2.5: Numerical Measures of Variability

The prices (in Rs 100) of 18 brands of walking shoes:

40 60 65 65 65 68 68 70 70

70 70 70 70 74 75 75 90 95

Position of $Q_1 = .25(18 + 1) = 4.75$

Position of $Q_3 = .75(18 + 1) = 14.25$

Q_3 is 1/4 of the way between the 14th and 15th ordered measurements, or

$$Q_3 = 74 + .25(75 - 74) = 74.25$$

and IQR = $Q_3 – Q_1 = 74.25 - 65 = 9.25$

2.6: Interpreting the Standard Deviation

- Chebyshev's Rule
- The Empirical Rule

Both tell us something about where the data will be relative to the mean.

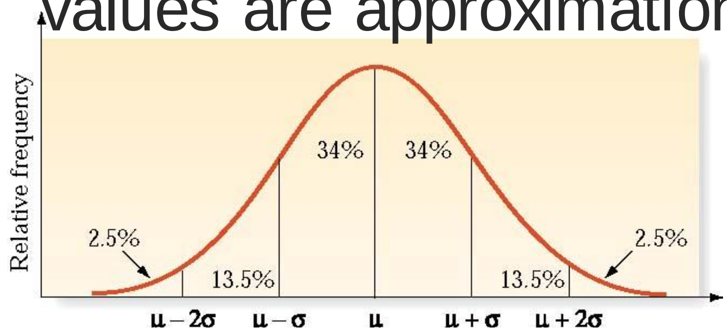
2.6: Interpreting the Standard Deviation

- **Chebyshev's Rule**
- Valid for *any* data set
- For any number k >1, at least $(1-1/k^2)$% of the observations will lie within k standard deviations of the mean

k	k^2	$1/k^2$	$(1-1/k^2)$%
2	4	.25	75%
3	9	.11	89%
4	16	.0625	93.75%

2.6: Interpreting the Standard Deviation

- The Empirical Rule
  - Useful for mound-shaped, symmetrical distributions
  - If not perfectly mounded and symmetrical, the values are approximations



- For a perfectly symmetrical and mound-shaped distribution,
  - ~68% will be within the range $(\bar{x}-s, \bar{x}+s)$
  - ~95% will be within the range $(\bar{x}-2s, \bar{x}+2s)$
  - ~99.7% will be within the range $(\bar{x}-3s, \bar{x}+3s)$

2.6: Interpreting the Standard Deviation

- Hummingbirds beat their wings in flight an average of 55 times per second.
- Assume the standard deviation is 10, and that the distribution is symmetrical and mounded.
  - Approximately what percentage of hummingbirds beat their wings between 45 and 65 times per second?
  - Between 55 and 65?
  - Less than 45?

2.6: Interpreting the Standard Deviation

- Hummingbirds beat their wings in flight an average of 55 times per second.
- Assume the standard deviation is 10, and that the distribution is symmetrical and mounded.
  - Approximately what percentage of hummingbirds beat their wings between 45 and 65 times per second?
  - Between 55 and 65?
  - Less than 45?

Since 45 and 65 are exactly one standard deviation below and above the mean, the empirical rule says that about 68% of the hummingbirds will be in this range.

2.6: Interpreting the Standard Deviation

- Hummingbirds beat their wings in flight an average of 55 times per second.
- Assume the standard deviation is 10, and that the distribution is symmetrical and mounded.
  - Approximately what percentage of hummingbirds beat their wings between 45 and 65 times per second?
  - Between 55 and 65?
  - Less than 45?

This range of numbers is from the mean to one standard deviation above it, or one-half of the range in the previous question. So, about one-half of 68%, or 34%, of the hummingbirds will be in this range.

2.6: Interpreting the Standard Deviation

- Hummingbirds beat their wings in flight an average of 55 times per second.
- Assume the standard deviation is 10, and that the distribution is symmetrical and mounded.

$$z = \frac{x - \overline{x}}{s}$$

An individual hummingbird is measured with 75 beats per second. What is this bird's z-score?

$$z = \frac{75 - 55}{10} = 2.0$$

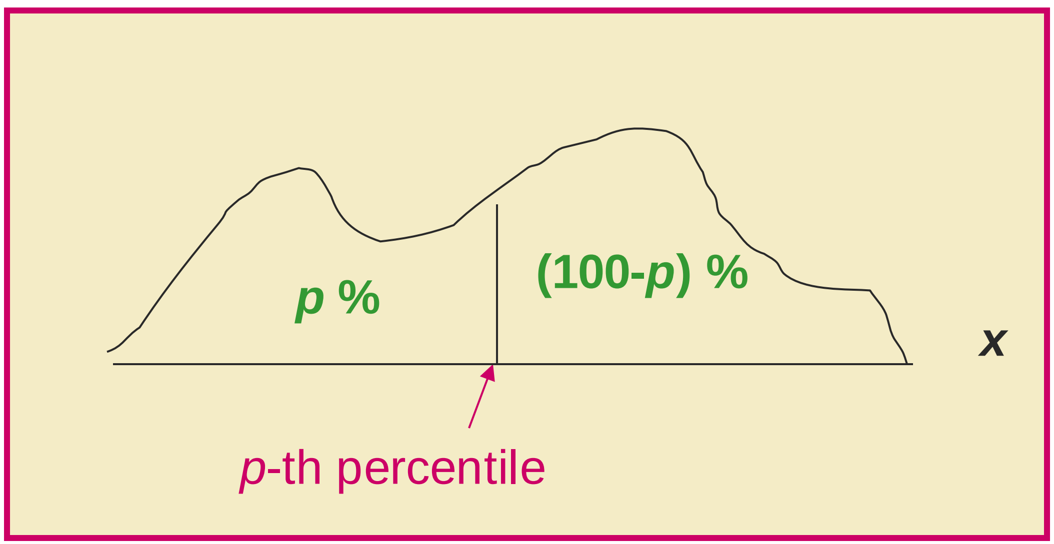
2.6: Interpreting the Standard Deviation

- Hummingbirds beat their wings in flight an average of 55 times per second.
- Assume the standard deviation is 10, and that the distribution is symmetrical and mounded.
  - Approximately what percentage of hummingbirds beat their wings between 45 and 65 times per second?
  - Between 55 and 65?
  - Less than 45?

Half of the entire data set lies above the mean, and ~34% lie between 45 and 55 (between one standard deviation below the mean and the mean), so ~84% (~34% + 50%) are above 45, which means ~16% are below 45.

2.7: Numerical Measures of Relative Standing

- **Percentiles**: for any (large) set of *n* measurements (arranged in ascending or descending order), the *p*th *percentile* is a number such that *p*% of the measurements fall below that number and (100 – *p*)% fall above it.



2.7: Numerical Measures of Relative Standing

- Finding percentiles is similar to finding the median – the median is the 50th percentile.
  - If you are in the 50th percentile for the GRE, half of the test-takers scored better and half scored worse than you.
  - If you are in the 75th percentile, you scored better than three-quarters of the test-takers.
  - If you are in the 90th percentile, only 10% of all the test-takers scored better than you.

2.7: Numerical Measures of Relative Standing

- The *z-score* tells us how many standard deviations above or below the mean a particular measurement is.

- Sample z-score

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

- Population z-score

$$z = \frac{x - \mu}{\sigma}$$

2.7: Numerical Measures of Relative Standing

- Z scores are related to the empirical rule:

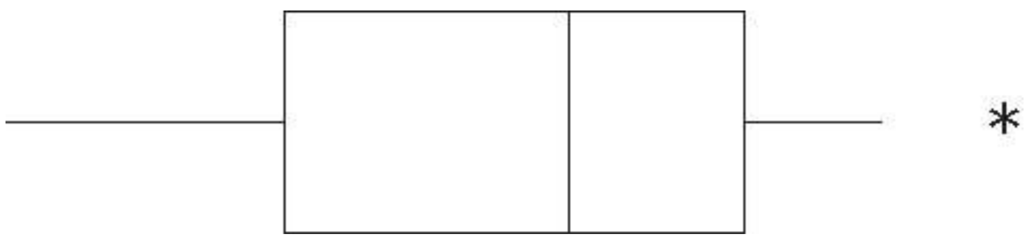
  For a perfectly symmetrical and mound-shaped distribution,
  - ~68 % will have z-scores between -1 and 1
  - ~95 % will have z-scores between -2 and 2
  - ~99.7% will have z-scores between -3 and 3

2.8: Methods for Determining Outliers

- An **outlier** is a measurement that is unusually large or small relative to the other values.
- Three possible causes:
  - *Observation, recording or data entry error*
  - *Item is from a different population*
  - *A rare, chance event*

2.8: Methods for Determining Outliers

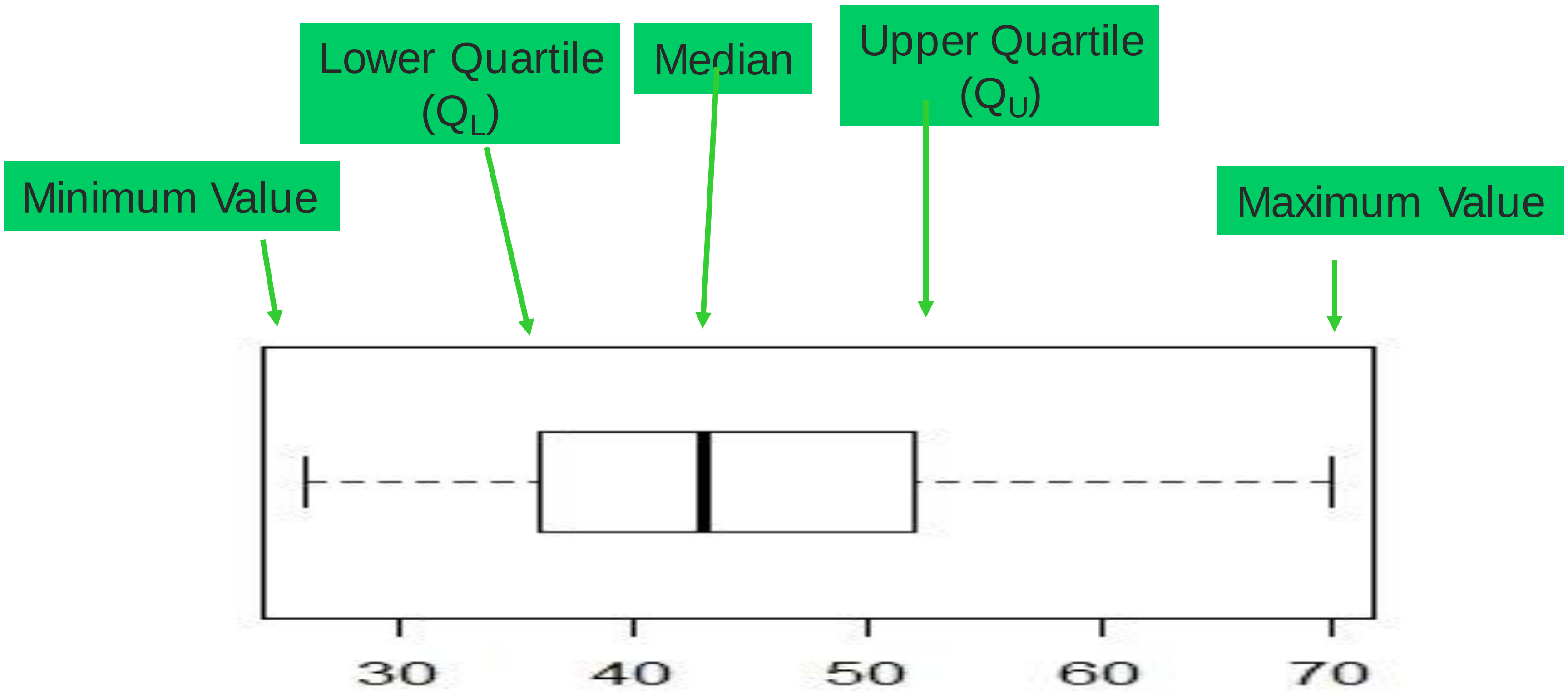
- The **box plot** is a graph representing information about certain percentiles for a data set and can be used to identify outliers



2.8: Methods for Determining Outliers

Boxplot: Data on age of professors

No value is outside the whiskers (1.5 times the IQR)



2.8: Methods for Determining Outliers

- Outliers and z-scores
  - The chance that a z-score is between -3 and +3 is over 99%.
  - Any measurement with $|z| > 3$ is considered an outlier.

2.8: Methods for Determining Outliers

#Observations	n = 50
Mean	44.90
Sample Variance	115.07
Sample Standard Deviation	10.73
Minimum	26
Maximum	70

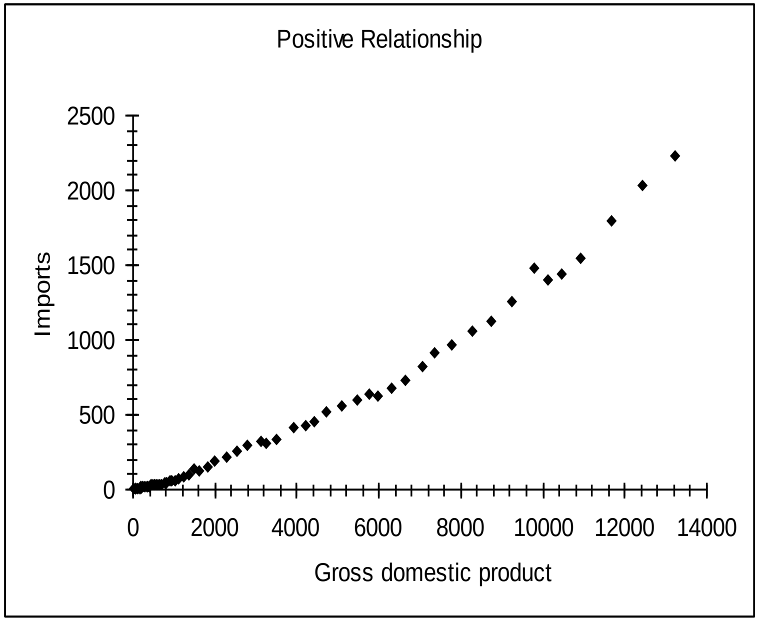
Here are the descriptive statistics

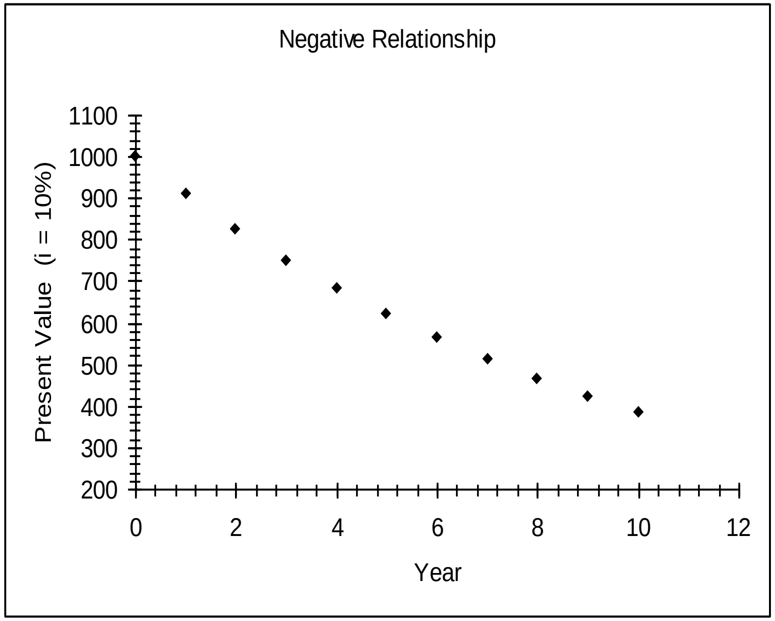
The z score corresponding to age 70 is (70-44.9)/10.73=2.34

So it is not an outlier.

2.9: Graphing Bivariate Relationships

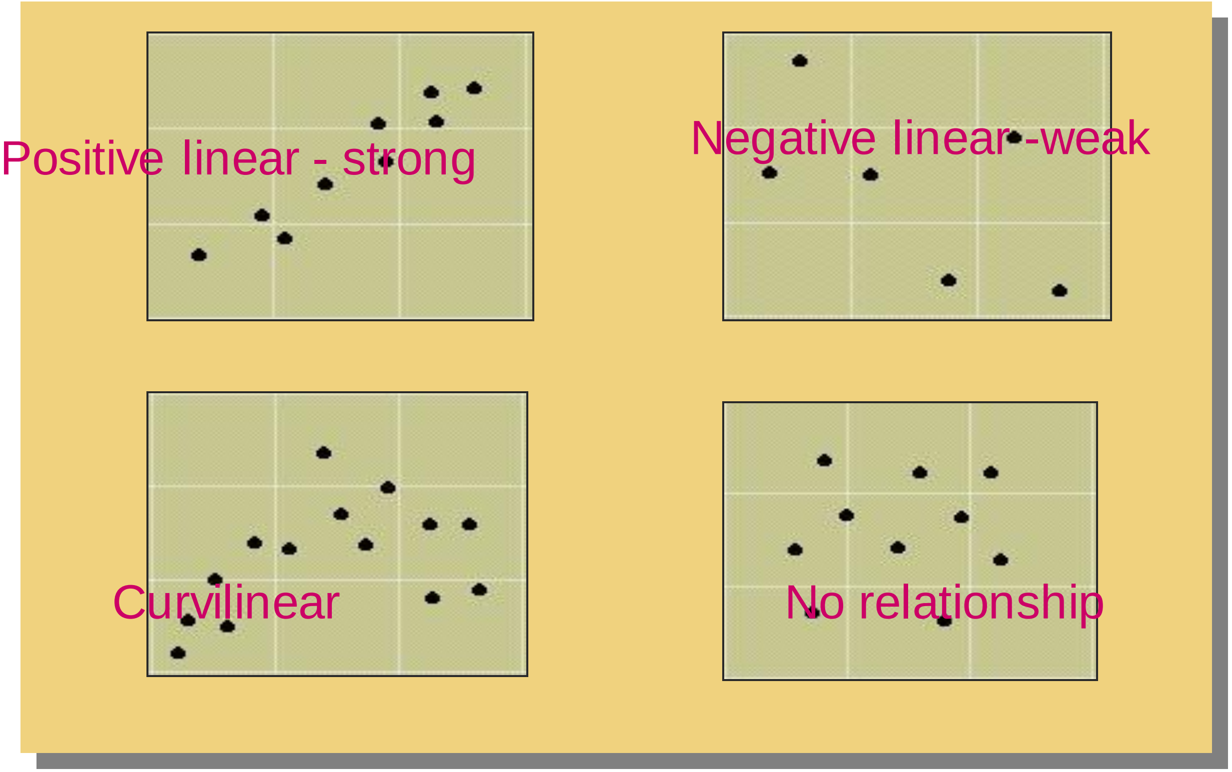
- **Scattergram** (or **scatterplot**) shows the relationship between two quantitative variables





2.9: Graphing Bivariate Relationships

If there is no linear relationship between the variables, the scatterplot may look like a cloud, a horizontal line or a more complex curve.



2.10: Distorting the Truth with Deceptive Statistics

- Distortions
  - Stretching the axis (and the truth)
  - Is average average?
    - Mean, median or mode?
  - Is average relevant?
    - What about the spread?