Lecture 3:
Summarizing and Displaying Numerical Data
STAT 310, Spring 2023

Measures of Central Tendency

Let $x_1, x_2, \cdots, x_n$ be observations of a sample of size n. The **sample mean** is given by

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

Example: The heights of 5 people: 63, 64, 66, 72, 62.

$$\bar{x} = \frac{63 + 64 + 66 + 72 + 62}{5} = 65.4$$

Measures of Central Tendency

The **sample median** of a set of observations is the middle value when values are ordered from smallest to largest.

Example: (n odd) Find the median of $63, 64, 66, 72, 62$.

First, order the data: $62, 63, 64, 66, 72$
median $= 64$

Example: (n even) Find the median of $63, 64, 66, 72, 62, 77$.

First, order the data: $62, 63, 64, 66, 72, 77$
median $= (64 + 66)/2 = 65$

Measures of Central Tendency

An **outlier** is an observation that appears extreme relative to the rest of the data. The median is resistant to outliers, while the mean is affected by outliers.

Example: How do the mean and median compare for the sample: $62, 63, 64, 66, 72, 1000$?

```
> x <- c(62, 63, 64, 66, 72, 1000)
> mean(x)
[1] 221.1667
> median(x)
[1] 65
```

Solution: The sample mean is much larger than the median, since it is affected by the outlier. The median is a better measure of central tendency in this example.

Quartiles

- The **first quartile**, denoted by Q_1, is the value such that 25% of the data falls below, i.e., the 25^{th} percentile.
- The **third quartile**, denoted by Q_3, is the value such that 75% of the data falls below, i.e., the 75^{th} percentile.
- Note that the second quartile, Q_2, is the median.

A method for finding Q_1 and Q_3 by hand:

1. Order the data from smallest to largest
2. Divide the data into two sets using the median
3. Q_1 is the median of the first half, and Q_3 is the median of the second half

Quartiles

Example: Find Q_1 and Q_3 for the following sample of height measurements of $n = 10$ people:

$68, 76, 66, 63, 70, 66, 71, 71, 64, 71$

Solution:
First, order the data: $63, 64, 66, 66, 68, 70, 71, 71, 71, 76$

median $= (68 + 70)/2 = 69$
$Q_1 = 66$
$Q_3 = 71$

Useful R commands:

```
> x <- c(68, 76, 66, 63, 70, 66, 71, 71, 64, 71)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   63.0    66.0    69.0    68.6    71.0    76.0
> mean(x)
[1] 68.6
> median(x)
[1] 69
> min(x)
[1] 63
> max(x)
[1] 76
> sort(x)
 [1] 63 64 66 66 68 70 71 71 71 76
```

Measures of Variation

- **Range** = Max - Min
- **Interquartile range**: IQR = Q_3 - Q_1
- Let $x_1, x_2, \cdots, x_n$ be a sample of n observations. The **sample variance** is defined as

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

and the **sample standard deviation** is defined as

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

Measures of Variation

- The sample variance can be thought of as the average of the squared deviations between the observations x_i and the sample mean $\bar{x}$. It measures how concentrated values are around the sample mean.
- The standard deviation is in the same units as the data (e.g., if the data are in ft, then s is in ft and s^2 is in ft^2).
- s^2, s, and the range are affected by outliers, while the IQR is resistant to outliers.

Measures of Variation

Example: Calculate the variance and standard deviation of the following sample of $n = 5$ observations: $2, 5, 10, 15, 18$

$$\bar{x} = \frac{2 + 5 + 10 + 15 + 18}{5} = \frac{50}{5} = 10$$

$$\begin{aligned} s^2 &= \frac{1}{5-1}[(2-10)^2 + (5-10)^2 + (10-10)^2 + (15-10)^2 + (18-10)^2] \\ &= \frac{1}{4}(8^2 + 5^2 + 0^2 + 5^2 + 8^2) \\ &= \frac{178}{4} = 44.5 \\ s &= \sqrt{44.5} = \boxed{6.67} \end{aligned}$$

Useful R commands:

```
> x <- c(68, 76, 66, 63, 70, 66, 71, 71, 64, 71)
> var(x)
[1] 15.6
> sd(x)
[1] 3.949684
> max(x) - min(x) # range
[1] 13
> IQR(x)
[1] 5
```

Example: Without doing any calculations, which of the following data sets do you think has the largest standard deviation? Which has the smallest standard deviation? Use R to verify.

Set 1: $100, 99, 98, 50, 2, 1, 0$
Set 2: $53, 52, 51, 50, 49, 48, 47$
Set 3: $51, 51, 51, 50, 49, 49, 49$

Solution: Set 1 has the largest standard deviation since the values are most spread out around the mean. Set 3 has the smallest standard deviation since the values are most concentrated around the mean. Note that $\bar{x} = 50$ for all three sets.

To verify in R:

```
> x1 = c(100, 99, 98, 50, 2, 1, 0); sd(x1)
> x2 = c(53, 52, 51, 50, 49, 48, 47); sd(x2)
> x3 = c(51, 51, 51, 50, 49, 49, 49); sd(x3)
```

Box Plot

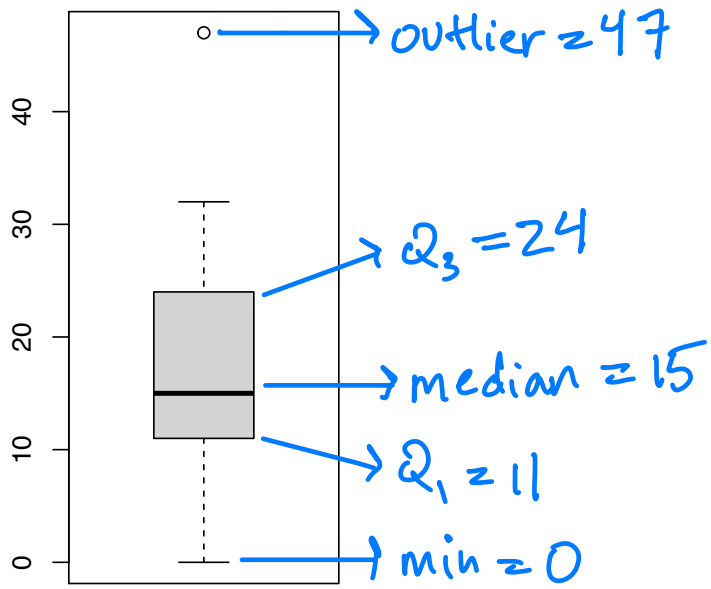
A box plot is a useful way to display the distribution of data and identify outliers.

Upper Fence $= Q_3 + 1.5(IQR)$
Lower Fence $= Q_1 - 1.5(IQR)$

Values outside the fences are potential outliers.

Example: Box Plot

```
> x <- c(0, 18, 15, 32, 5, 22, 47, 15, 26, 13, 9)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   11.00   15.00   18.36   24.00   47.00
> boxplot(x)
```



$$IQR = Q_3 - Q_1$$
$$= 24 - 11 = 13$$

$$UF = Q_3 + 1.5(IQR)$$
$$= 24 + 1.5(13) = 43.5$$

$$LF = Q_1 - 1.5(IQR)$$
$$= 11 - 1.5(13) = -8.5$$

Histogram

- A histogram is a useful way to visualize the distribution of a numerical variable.
- To construct a histogram, the range of the data is divided into bins of equal width. Then the number of observations falling in each bin are counted. The counts are plotted as rectangles over each bin.
- Histograms are especially convenient for understanding the shape of the data distribution.

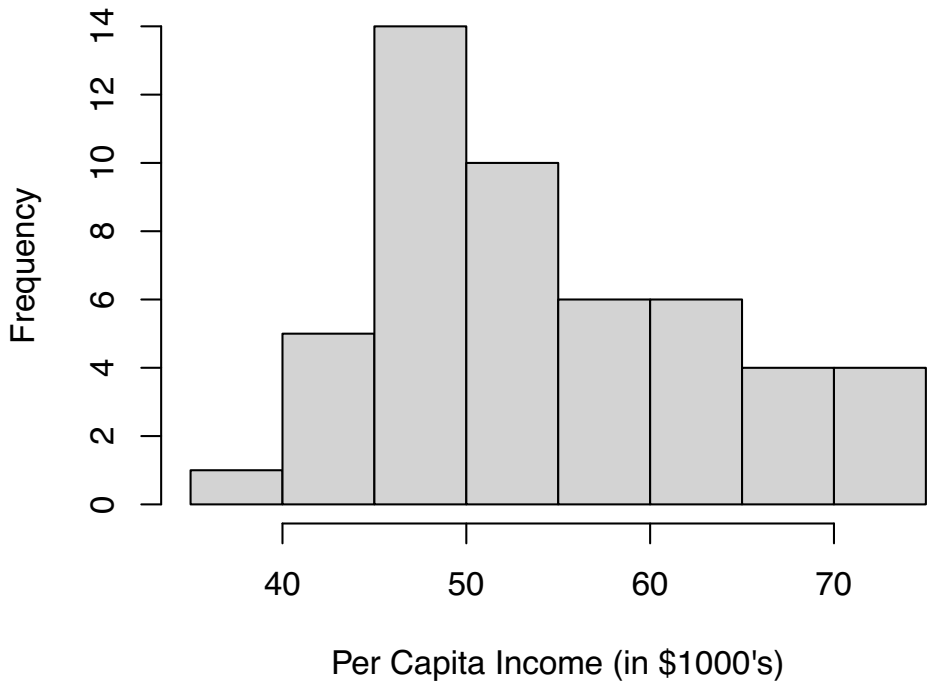
Example: Histogram

Data set from 2008 with the per capita income (in thousands of dollars) and percent of the population with a college education for each of the 50 states.

	State	Income	Pct College
1	Alabama	43.62	23.50
2	Alaska	72.52	28.00
3	Arizona	50.26	27.50
4	Arkansas	41.37	21.10
5	California	61.82	31.40
6	Colorado	60.63	38.10
⋮	⋮	⋮	⋮
49	Wisconsin	53.36	27.80
50	Wyoming	58.84	25.70

Example: Histogram

Bin	(35, 40]	(40, 45]	(45, 50]	(50, 55]	(55, 60]	(60, 65]	(65, 70]	(70, 75]
Count	1	5	14	10	6	6	4	4

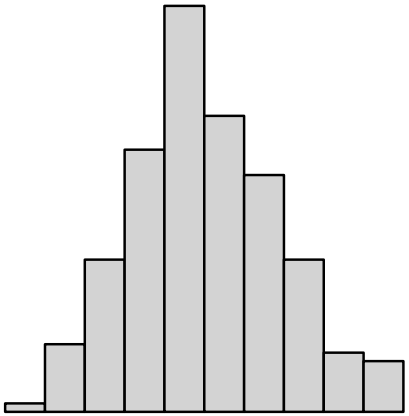


Based on the histogram, how many states have a per capita income between \$40,000-\$50,000?

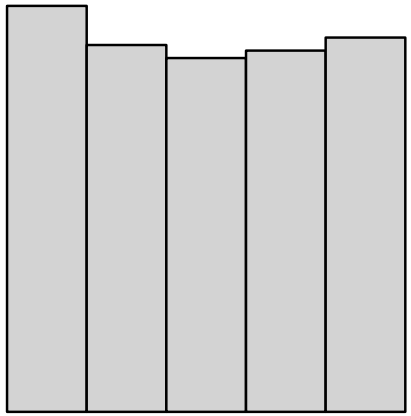
Solution: 19 states have a per capita income between \$40,000-\$50,000.

Describing the Shape of a Distribution

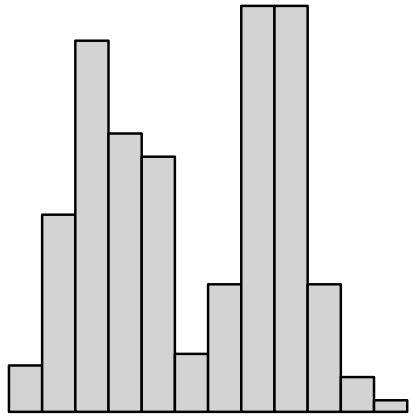
Symmetric, Normal



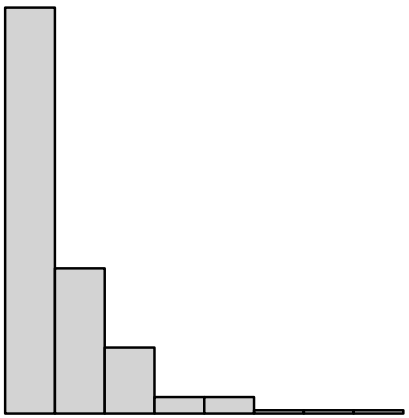
Symmetric, Uniform



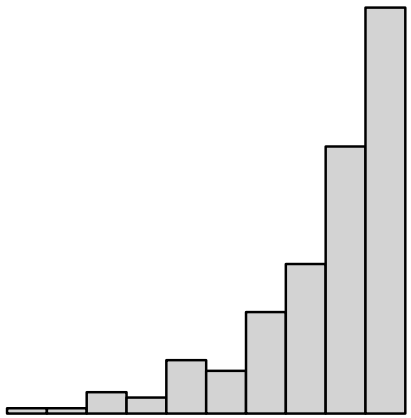
Bimodal



Right Skewed



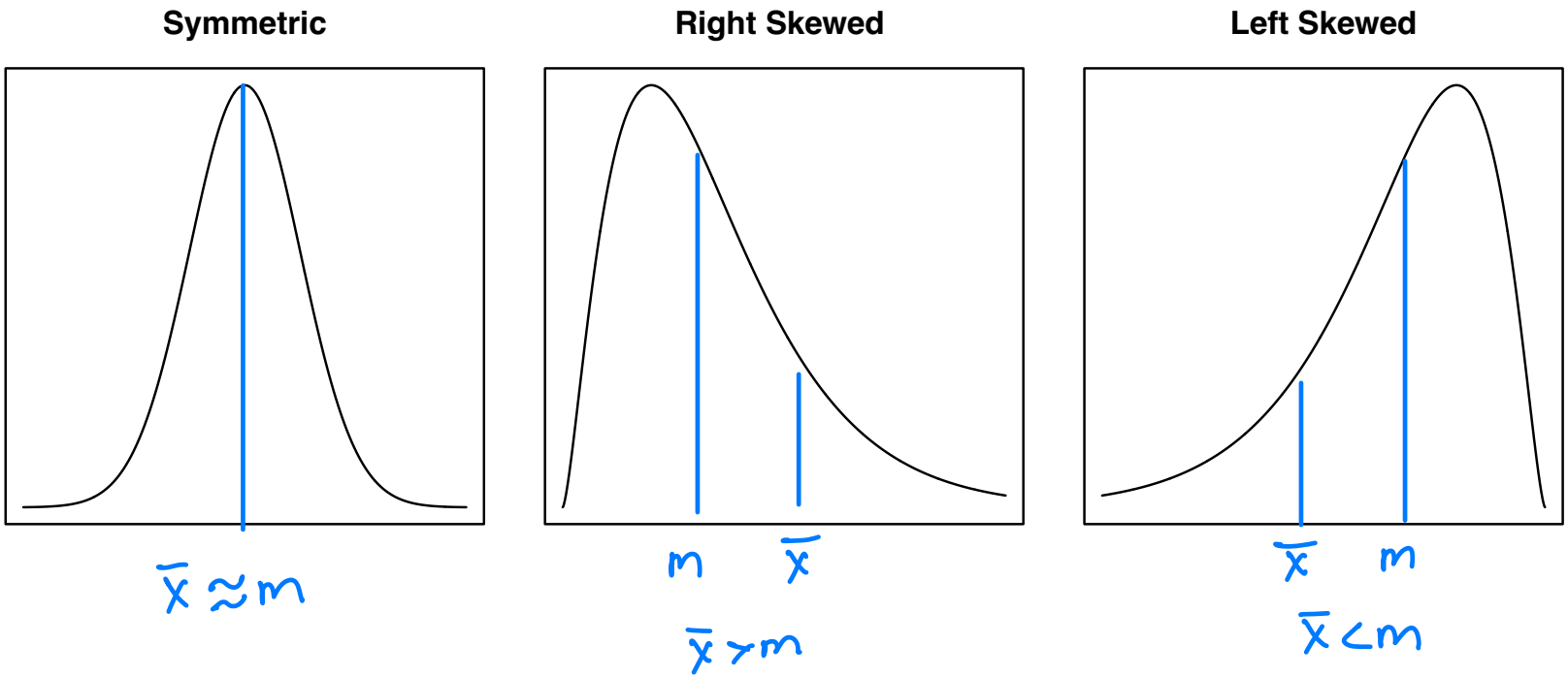
Left Skewed



$\bar{x} \Rightarrow$ mean

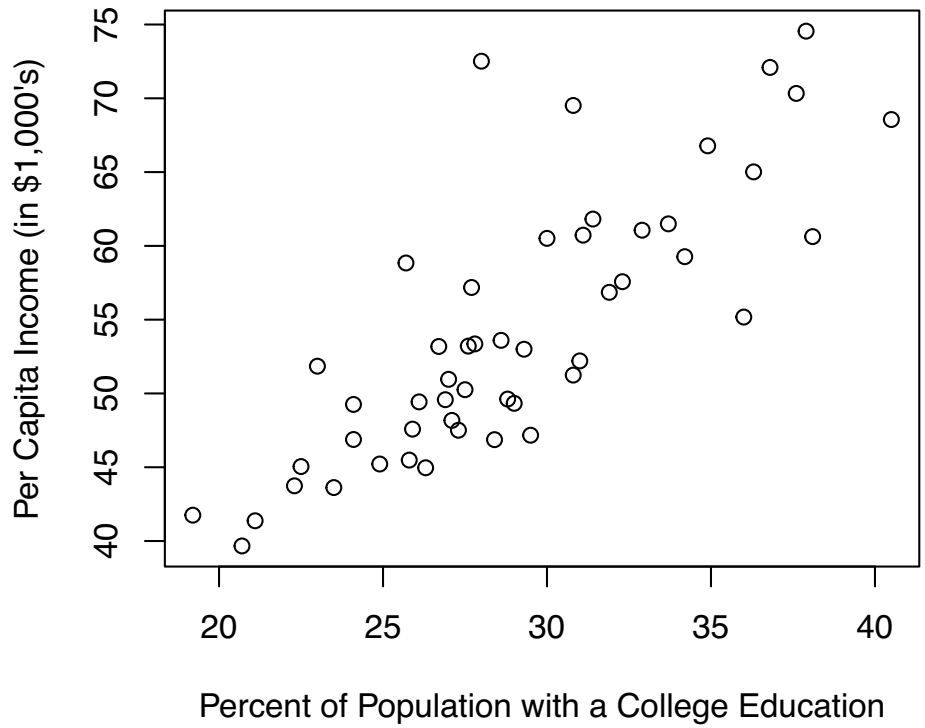
$m \Rightarrow$ median

- For symmetric distributions, the mean and median are approximately equal.
- For right skewed distributions, the mean is greater than the median.
- For left skewed distributions, the mean is less than the median.



Scatter Plot

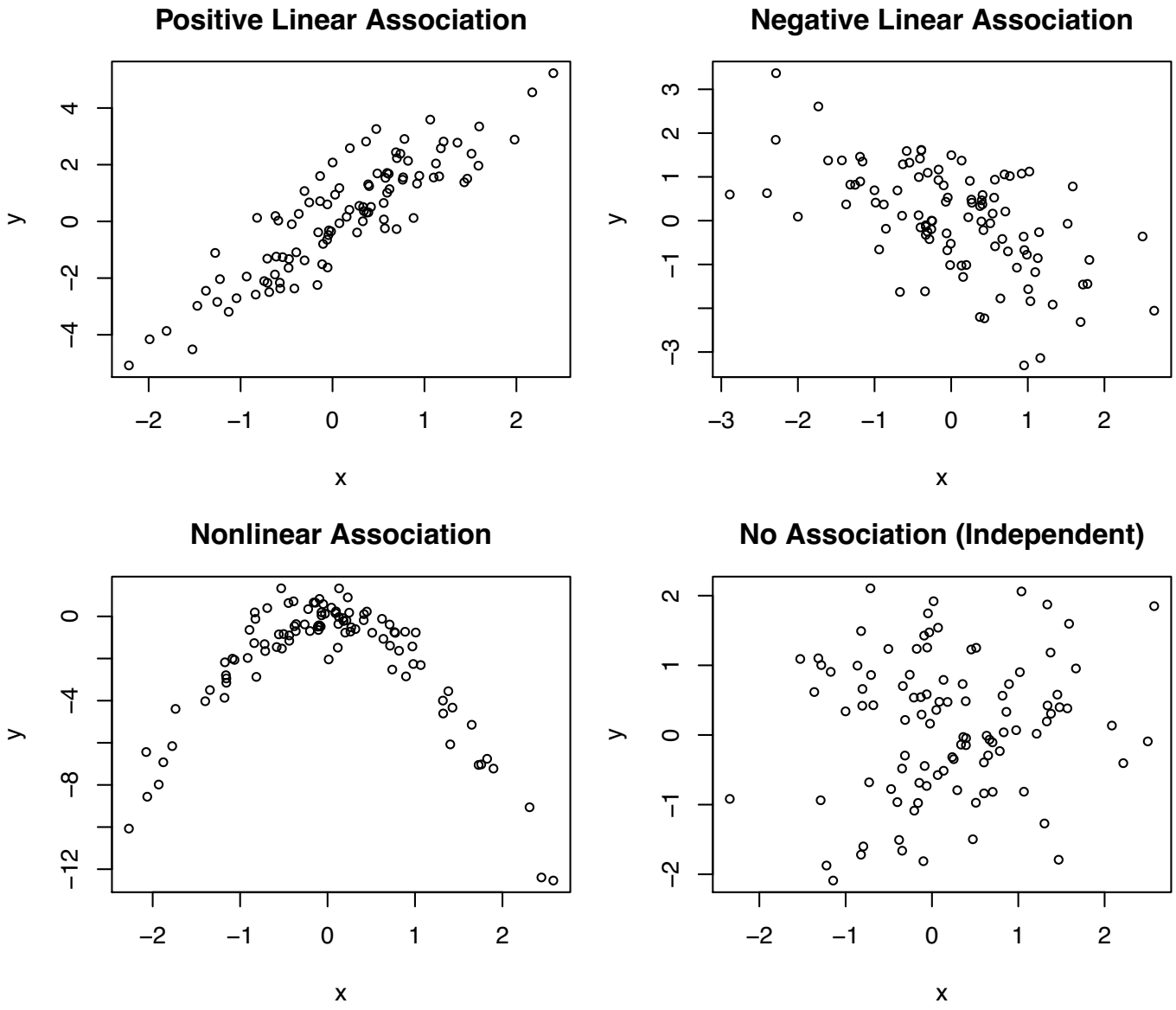
Scatter plots are useful for visualizing the relationship between two numerical variables. For example, the scatter plot below shows the relationship between per capita income and percent college graduates. There are 50 points since each point represents a state in the U.S.



There is a positive, linear relationship in the scatter plot. That is, states with a higher percentage of college graduates tend to have a higher per capita income.

Describing Relationships in Scatter Plots

- Two numerical variables are said to be **associated** if the scatter plot shows a discernible pattern or trend.
- An association is **positive** if y increases as x increases.
- An association is **negative** if y decreases as x increases.
- An association is **linear** if the scatter plot between x and y has a linear trend; otherwise, the association is called **nonlinear**.


Positive Linear Association
Negative Linear Association
Nonlinear Association
No Association (Independent)
x
y