Lecture 10

§4.1 Visualizing Variability with a Scatterplot and §4.2 Measuring Association with Correlation

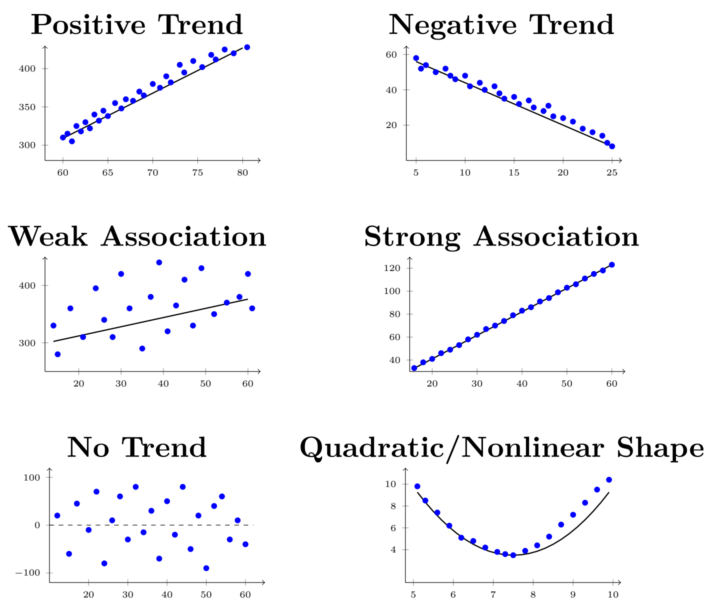
1. §4.1 Visualizing Variability with a Scatterplot

Definition.

A **scatterplot** displays the relationship between two numerical variables measured on the same individuals, where each point represents an individual observation.

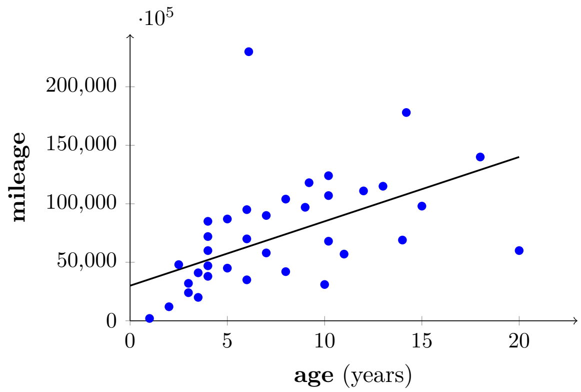
When describing and interpreting scatterplots, we always examine three primary characteristics:

1. **Trend (Direction):** The overall tendency of the scatterplot going from left to right.
   - *Positive:* As x increases, y generally increases.
   - *Negative:* As x increases, y generally decreases.
   - *No Trend:* Values of y show no systematic upward or downward change as x increases.
2. **Strength:** How closely the points follow the underlying curve or line. Strong associations show little vertical variation around the trend, whereas weak associations show considerable vertical dispersion.
3. **Shape (Form):** Whether the pattern is linear or nonlinear (e.g., curved or quadratic).



Example 1: Used Cars Age and Mileage

The scatterplot below shows the age (in years) and corresponding mileage for a sample of used cars:



Active Worksheet Analysis:

1. Identify the variables:
   - Explanatory Variable (x): ______________________
   - Response Variable (y): ______________________
2. Identify the **trend**: ______________________ (*positive* / *negative* / *none*)
3. Identify the **shape**: ______________________ (*linear* / *nonlinear*)
4. Identify the **strength**: ______________________ (*weak* / *moderate* / *strong*)
5. Are there any potential **outliers**? Identify at least one point: ______________________

2. §4.2 Measuring Strength of Association with Correlation

Definition.

The **correlation coefficient** (denoted r) is a single number that measures both the **strength** and **direction** of a linear association between two numerical variables.

Properties of the Correlation Coefficient (r)

$$\boxed{-1 \leq r \leq 1}$$

- $r > 0$: Indicates a **positive trend** (as x increases, y tends to increase).
- $r < 0$: Indicates a **negative trend** (as x increases, y tends to decrease).
- $r = +1$: Perfect, deterministic positive linear relationship.
- $r = -1$: Perfect, deterministic negative linear relationship.
- $r \approx 0$: No linear relationship.

Strict Requirement: The correlation coefficient only makes sense if the trend is **linear** and both variables are **numerical**!

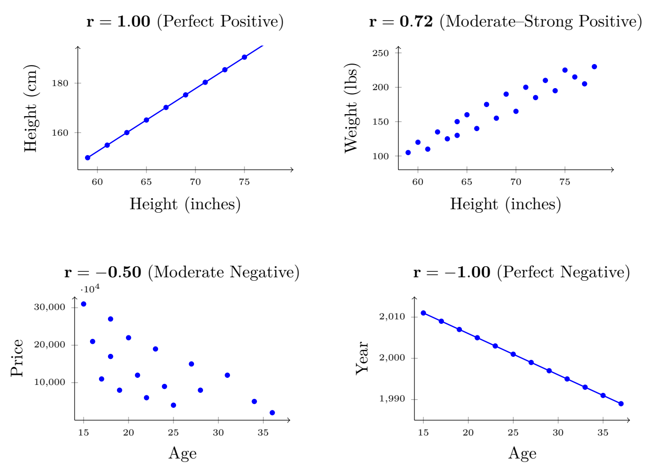
Correlation Does Not Imply Causation!

A correlation between two variables does **not**, by itself, establish that changes in one variable cause changes in the other.
A correlation may arise because:

- x causes y;
- y causes x;
- a third variable affects both x and y; or
- the observed association is due to chance.

Example: Ice cream sales and drowning deaths may be positively correlated. This does not mean that eating ice cream causes drowning. A third variable, **hot weather**, can increase both ice cream sales and swimming activity.
Correlation describes association; it does not establish causation.

2.1 Visual Guide: Interpreting Values of r



2.2 Formula for the Correlation Coefficient

Definition.

The sample correlation coefficient r between variables x and y is computed as:

$$\boxed{r = \frac{\sum z_x z_y}{n-1}}$$

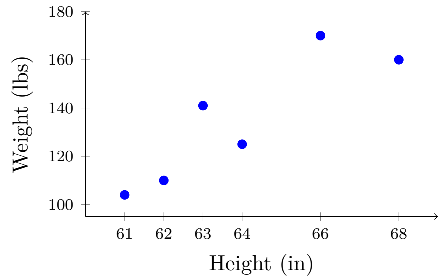
where z_x and z_y are the standardized z-scores for each entry:

$$z_x = \frac{x - \bar{x}}{s_x} \qquad \text{and} \qquad z_y = \frac{y - \bar{y}}{s_y}$$

Example 2: Computing Correlation by Hand

Below are the heights (in inches) and weights (in pounds) of six women:

Height (x)	61	62	63	64	66	68
Weight (y)	104	110	141	125	170	160



Compute the sample correlation coefficient r by hand.

Step 1: Compute Means and Standard Deviations

The sample size is $n = 6$.

For x (height):

$$\sum x = 61 + 62 + 63 + 64 + 66 + 68 = 384 \implies \bar{x} = \frac{384}{6} = 64$$

$$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{(-3)^2 + (-2)^2 + (-1)^2 + 0^2 + 2^2 + 4^2}{5}} = \sqrt{\frac{34}{5}} = \sqrt{6.8} \approx 2.6077$$

For y (weight):

$$\sum y = 104 + 110 + 141 + 125 + 170 + 160 = 810 \implies \bar{y} = \frac{810}{6} = 135$$

$$s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}} = \sqrt{\frac{(-31)^2 + (-25)^2 + 6^2 + (-10)^2 + 35^2 + 25^2}{5}} = \sqrt{\frac{3372}{5}} = \sqrt{674.4} \approx 25.9692$$

Step 2: Construct the Computation Table

x	y	$x-\bar{x}$	$y-\bar{y}$	$z_x = \frac{x-\bar{x}}{s_x}$	$z_y = \frac{y-\bar{y}}{s_y}$	$z_x \cdot z_y$
61	104	-3	-31	-1.1504	-1.1937	1.3732
62	110	-2	-25	-0.7670	-0.9627	0.7384
63	141	-1	6	-0.3835	0.2310	-0.0886
64	125	0	-10	0.0000	-0.3851	0.0000
66	170	2	35	0.7670	1.3477	1.0337
68	160	4	25	1.5340	0.9627	1.4768
					Sum of products $\sum z_x z_y =$	**4.5335**

Step 3: Calculate and Interpret r

$$r = \frac{\sum z_x z_y}{n-1} = \frac{4.5335}{6-1} = \frac{4.5335}{5} \approx \boxed{0.907}$$

Interpretation: Because $r \approx 0.907$ is positive and close to 1, there is a **strong, positive linear association** between the heights and weights of these women.

2.3 Properties of the Correlation Coefficient

Key Mathematical Properties:

1. **Symmetry:** Changing the order of the variables does not change r:

$$\text{corr}(x, y) = \text{corr}(y, x)$$

2. **Scale Invariance:** Adding a constant or multiplying by a positive constant does not affect r (e.g., converting height from inches to centimeters leaves r unchanged).
3. **Unitless:** r has no units attached to it because it is calculated from standardized z-scores.
4. **Linearity Dependency:** r only measures *linear* associations. A strong curved relationship may still yield $r \approx 0$.

StatCrunch Technology Guide

To obtain scatterplots and compute r using **StatCrunch**:

1. **For Scatterplots:** Select **Graph** → **Scatter Plot**. Choose your X-variable and Y-variable, then click **Compute!**.
2. **For Correlation r:** Select **Stat** → **Summary Stats** → **Correlation**. Select both columns, then click **Compute!**.