

# Lecture 2: Correlation

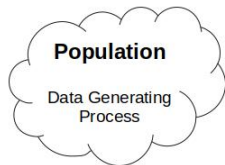
Module 1, part 1

Spring 2025

# Logistics

- Please take a look at the syllabus if you haven't already
- Population, data, and statistics
- Start Module 1 (3 lectures total)
- Correlation

# Sample data vs Population distribution



**Data**

A screenshot of a data table with multiple columns and rows of numerical data. The table is displayed in a software interface with a menu bar at the top.



**Statistic**



# Summarizing a data set

Suppose we observe  $n$  numbers,  $x_1, x_2, \dots, x_n$ . How might we summarize this set of number succinctly?

# Summarizing a data set

Suppose we observe  $n$  numbers,  $x_1, x_2, \dots, x_n$ . How might we summarize this set of number succinctly?

- Mean:

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

- Median: “middle value”
- Mode: most frequent value

## Alternative way

We can think about the mean through a different lens...

- Let  $\hat{b}_0$  be a “candidate”
- The residual for the  $i$ th observation is  $e_i = x_i - \hat{b}_0$

## Alternative way

We can think about the mean through a different lens...

- Let  $\hat{b}_0$  be a “candidate”
- The residual for the  $i$ th observation is  $e_i = x_i - \hat{b}_0$

Suppose we use the *residual sum of squares* to define how well a number “summarizes” a set:

$$RSS(\hat{b}_0) = \sum_i |x_i - \hat{b}_0|^2 = \sum_i |e_i|^2$$

How do we select the best  $b_0$ ?

## Alternative way

We can think about the mean through a different lens...

- Let  $\hat{b}_0$  be a “candidate”
- The residual for the  $i$ th observation is  $e_i = x_i - \hat{b}_0$

Suppose we use the *residual sum of squares* to define how well a number “summarizes” a set:

$$RSS(\hat{b}_0) = \sum_i |x_i - \hat{b}_0|^2 = \sum_i |e_i|^2$$

How do we select the best  $b_0$ ?

$$\frac{\partial RSS}{\partial \hat{b}_0} = -2 \sum_i^n (x_i - \hat{b}_0)$$



## Alternative way

We can think about the mean through a different lens...

- Let  $\hat{b}_0$  be a “candidate”
- The residual for the  $i$ th observation is  $e_i = x_i - \hat{b}_0$

Suppose we use the *residual sum of squares* to define how well a number “summarizes” a set:

$$RSS(\hat{b}_0) = \sum_i |x_i - \hat{b}_0|^2 = \sum_i |e_i|^2$$

How do we select the best  $b_0$ ?

$$\frac{\partial RSS}{\partial \hat{b}_0} = -2 \sum_i^n (x_i - \hat{b}_0)$$

If you need a refresher on notation:

<https://www.youtube.com/watch?v=bPvtv780h3k>

# Measure of centrality

The **mean** is the value  $\hat{b}_0$  which minimizes

$$RSS(\hat{b}_0) = \sum_i^n (x_i - \hat{b}_0)^2 = \sum_i |e_i|^2$$

We often also use  $\bar{y}$  to denote the mean of the  $x_1, x_2, \dots, x_n$ .

The **median** is a value  $\hat{b}_0$  which minimizes

$$\sum_i^n |x_i - \hat{b}_0| = \sum_i |e_i|$$

The **mode** is a value  $\hat{b}_0$  which minimizes

$$\sum_i^n |x_i - \hat{b}_0|^0 = \sum_i |e_i|^0,$$

with here the (unusual) convention  $0^0 = 0$ .

# Measuring spread of data

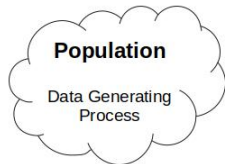
The **variance** of a data set is defined as:

$$\hat{\sigma}_X^2 = \text{var} = \frac{1}{n} \sum_i (x_i - \bar{x})^2 = \frac{RSS(\bar{x})}{n}$$

The **standard deviation** of a data set is defined as:

$$\text{sd} = \sqrt{\hat{\sigma}_X^2}$$

# Sample data vs Population distribution



**Data**

| id | name        | age | height | weight | gender | status |
|----|-------------|-----|--------|--------|--------|--------|
| 1  | John        | 25  | 180    | 75     | M      | 1      |
| 2  | Jane        | 22  | 165    | 60     | F      | 1      |
| 3  | Mike        | 30  | 190    | 85     | M      | 1      |
| 4  | Sarah       | 28  | 170    | 70     | F      | 1      |
| 5  | David       | 20  | 175    | 65     | M      | 1      |
| 6  | Emily       | 24  | 160    | 55     | F      | 1      |
| 7  | Chris       | 27  | 185    | 78     | M      | 1      |
| 8  | Alice       | 21  | 168    | 58     | F      | 1      |
| 9  | Bob         | 32  | 195    | 90     | M      | 1      |
| 10 | Anna        | 26  | 172    | 68     | F      | 1      |
| 11 | Tom         | 23  | 178    | 72     | M      | 1      |
| 12 | Lisa        | 29  | 162    | 62     | F      | 1      |
| 13 | Kevin       | 24  | 182    | 76     | M      | 1      |
| 14 | Nancy       | 27  | 166    | 64     | F      | 1      |
| 15 | Paul        | 31  | 192    | 88     | M      | 1      |
| 16 | Michelle    | 25  | 170    | 66     | F      | 1      |
| 17 | James       | 22  | 176    | 70     | M      | 1      |
| 18 | Stephanie   | 28  | 164    | 60     | F      | 1      |
| 19 | Benjamin    | 20  | 180    | 74     | M      | 1      |
| 20 | Karen       | 26  | 168    | 62     | F      | 1      |
| 21 | Robert      | 33  | 198    | 92     | M      | 1      |
| 22 | Julia       | 24  | 174    | 68     | F      | 1      |
| 23 | Michael     | 21  | 179    | 73     | M      | 1      |
| 24 | Elizabeth   | 29  | 161    | 59     | F      | 1      |
| 25 | William     | 27  | 184    | 77     | M      | 1      |
| 26 | Olivia      | 23  | 169    | 61     | F      | 1      |
| 27 | Alexander   | 30  | 191    | 89     | M      | 1      |
| 28 | Sophia      | 25  | 171    | 67     | F      | 1      |
| 29 | Matthew     | 22  | 177    | 71     | M      | 1      |
| 30 | Victoria    | 28  | 163    | 57     | F      | 1      |
| 31 | Christopher | 20  | 181    | 75     | M      | 1      |
| 32 | Christina   | 26  | 167    | 63     | F      | 1      |
| 33 | Andrew      | 34  | 200    | 95     | M      | 1      |
| 34 | Megan       | 24  | 173    | 69     | F      | 1      |
| 35 | Joshua      | 21  | 176    | 72     | M      | 1      |
| 36 | Amber       | 29  | 165    | 61     | F      | 1      |
| 37 | Daniel      | 27  | 183    | 79     | M      | 1      |
| 38 | Madison     | 23  | 169    | 62     | F      | 1      |
| 39 | Anthony     | 31  | 193    | 91     | M      | 1      |
| 40 | Emily       | 25  | 175    | 69     | F      | 1      |
| 41 | Benjamin    | 22  | 178    | 73     | M      | 1      |
| 42 | Grace       | 28  | 164    | 60     | F      | 1      |
| 43 | Robert      | 20  | 180    | 74     | M      | 1      |
| 44 | Chloe       | 26  | 168    | 62     | F      | 1      |
| 45 | William     | 33  | 198    | 92     | M      | 1      |
| 46 | Isabella    | 24  | 174    | 68     | F      | 1      |
| 47 | Michael     | 21  | 179    | 73     | M      | 1      |
| 48 | Elizabeth   | 29  | 161    | 59     | F      | 1      |
| 49 | William     | 27  | 184    | 77     | M      | 1      |
| 50 | Olivia      | 23  | 169    | 61     | F      | 1      |



**Statistic**



# Random variable notation

So far, we've discussed observing a sample of data, but now we will define some notation for random variables

# Random variable notation

So far, we've discussed observing a sample of data, but now we will define some notation for random variables

Let  $X_i$  denote a random variable (sometimes we will drop the subscript).

- Roughly speaking, random variables take a “process” and output a number
- $E(\cdot)$  will denote the “expectation” which roughly speaking means the average in the population or what we would get if we could take an infinite number of samples
- $E(X)$  denotes the (population) mean of  $X$ , also sometimes will use  $\mu_X$
- We will denote the (population) variance of  $X$  as

$$\sigma_X^2 = E[(X - \mu_X)^2]$$

# Random variable notation

So far, we've discussed observing a sample of data, but now we will define some notation for random variables

Let  $X_i$  denote a random variable (sometimes we will drop the subscript).

- Roughly speaking, random variables take a “process” and output a number
- $E(\cdot)$  will denote the “expectation” which roughly speaking means the average in the population or what we would get if we could take an infinite number of samples
- $E(X)$  denotes the (population) mean of  $X$ , also sometimes will use  $\mu_X$
- We will denote the (population) variance of  $X$  as

$$\sigma_X^2 = E[(X - \mu_X)^2]$$

We will generally use lower case letters to denote numbers

- Typically,  $x_i$  will denote the realization of random variable  $X_i$
- $\bar{x}$  denotes the mean of the observations  $x_1, x_2, \dots, x_n$
- $\hat{\sigma}_x^2$  denotes the variance of the observations

## Estimating the variance

Suppose we have some observations  $x_1, x_2, \dots, x_n$  which are sampled from a population with a true mean of  $\mu_X$  and true variance of  $\sigma_X^2$ . How would we estimate the true variance if it is unknown?

$$\sigma_X^2 = E[(X - \mu_X)^2]$$



## Estimating the variance

Suppose we have some observations  $x_1, x_2, \dots, x_n$  which are sampled from a population with a true mean of  $\mu_X$  and true variance of  $\sigma_X^2$ . How would we estimate the true variance if it is unknown?

$$\sigma_X^2 = E[(X - \mu_X)^2]$$

If we knew  $\mu_X$ , we could use

$$\hat{\sigma}_X^2 = \frac{1}{n} \sum_i^n (x_i - \mu_X)^2 = \frac{1}{n} \text{RSS}(\mu_X)$$

and

$$E(\hat{\sigma}_X^2) = \sigma_X^2$$

## Estimating the variance

Suppose we have some observations  $x_1, x_2, \dots, x_n$  which are sampled from a population with a true mean of  $\mu_X$  and true variance of  $\sigma_X^2$ . How would we estimate the true variance if it is unknown?

$$\sigma_X^2 = E[(X - \mu_X)^2]$$

If we knew  $\mu_X$ , we could use

$$\hat{\sigma}_X^2 = \frac{1}{n} \sum_i^n (x_i - \mu_X)^2 = \frac{1}{n} RSS(\mu_X)$$

and

$$E(\hat{\sigma}_X^2) = \sigma_X^2$$

When we don't know  $\mu_X$ , we can plug in  $\bar{x}$ , and use

$$s_X^2 = \frac{1}{n} \sum_i^n (x_i - \bar{x})^2 = \frac{1}{n} RSS(\bar{x})$$

# Estimating the variance

Unfortunately,  $\bar{x}$  minimizes RSS, so

$$\frac{1}{n}RSS(\bar{x}) \leq \frac{1}{n}RSS(\mu_x)$$

and

$$E(s_x^2) \leq \sigma_x^2$$

## Estimating the variance

Unfortunately,  $\bar{x}$  minimizes RSS, so

$$\frac{1}{n}RSS(\bar{x}) \leq \frac{1}{n}RSS(\mu_x)$$

and

$$E(s_x^2) \leq \sigma_x^2$$

Instead of dividing by  $n$ , we divide by  $n - 1$  and redefine

$$s_x^2 = \frac{1}{n-1} \sum_i^n (x_i - \bar{x})^2 = \frac{1}{n-1}RSS(\bar{x})$$

and we now have

$$E(s_x^2) = \sigma_x^2$$

# Group Discussion

- What is a scientific problem you are interested in?
- Describe the population process, the data you might gather, and the statistic you might be interested in

# Correlation

# Wine data

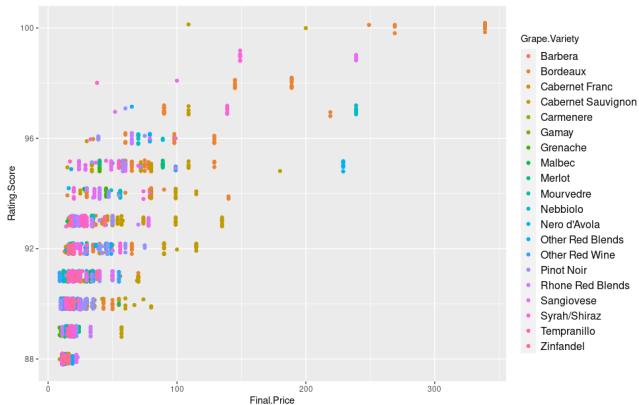


Figure: Wine Price vs Wine Rating from wine.com

# Correlation

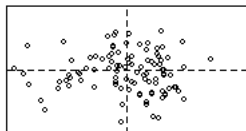
Correlation measures the linear dependence between two variables.

- For two variables,  $X$  and  $Y$ , correlation is denoted by  $r_{XY}$
- Correlation is between -1 and 1
- $r_{XY} = 0$  indicates no **linear** relationship
- $r_{XY} > 0$  indicates positive **linear** relationship
- $r_{XY} < 0$  indicates negative **linear** relationship
- $r_{XY} = \pm 1$  indicates perfect **linear** relationship

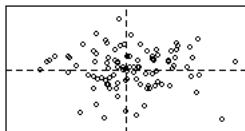


# Correlation

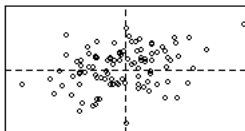
Cor = 0



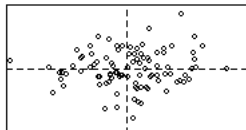
Cor = 0.1



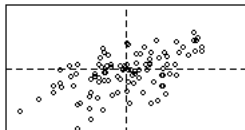
Cor = 0.2



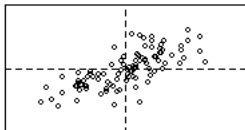
Cor = 0.3



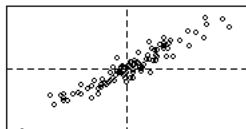
Cor = 0.5



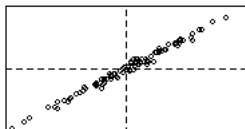
Cor = 0.8



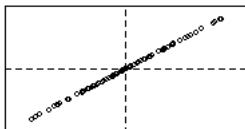
Cor = 0.95



Cor = 0.99



Cor = 1



# Correlation

For two variables,  $X$  and  $Y$ , the sample correlation is

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

where

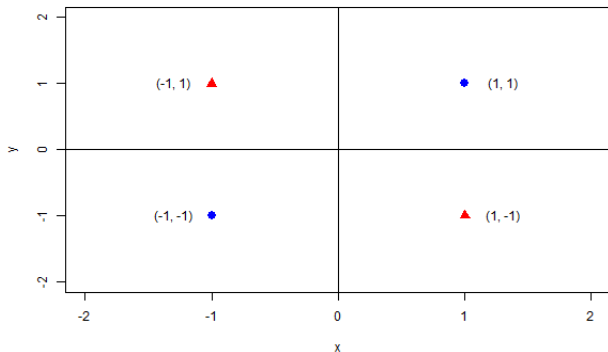
$$\text{Sample SD of } X = s_X = \sqrt{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2}$$

$$\text{Sample SD of } Y = s_Y = \sqrt{\frac{1}{n-1} \sum_i (y_i - \bar{y})^2}$$

$$\text{Sample Covariance} = s_{XY} = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

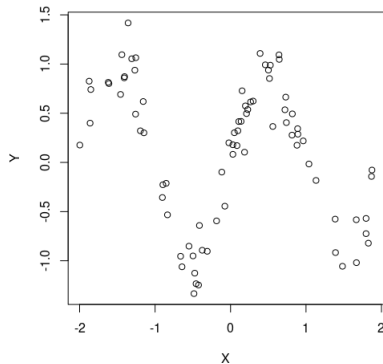
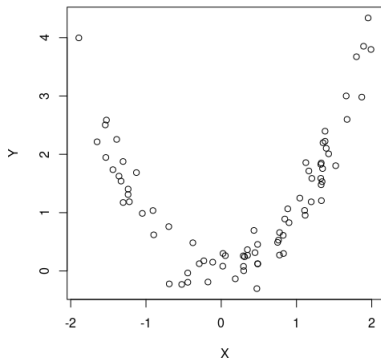
# Sample Covariance

$$s_{XY} = \frac{1}{n-1} \sum_i (x_i - \bar{x})(x_i - \bar{y})$$



# Non-linear association

Correlation only measure **linear** association



# Wrap-up

- Population: process of interest
- Data: measurements gathered
- Statistic: calculation based on data
- Describe linear relationship between two variables using correlation