

# Lecture 2: Correlation

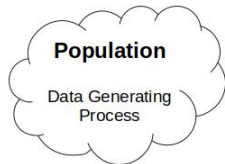
Module 1, part 1

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

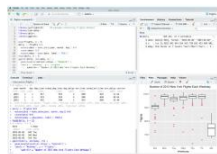


**Data**

id	name	age	sex	height	weight	hair
204	John	25	M	183	80.5	Black
304	John	25	M	183	80.5	Black
404	John	25	M	183	80.5	Black
504	John	25	M	183	80.5	Black
604	John	25	M	183	80.5	Black
704	John	25	M	183	80.5	Black
804	John	25	M	183	80.5	Black
904	John	25	M	183	80.5	Black
1004	John	25	M	183	80.5	Black
1104	John	25	M	183	80.5	Black
1204	John	25	M	183	80.5	Black
1304	John	25	M	183	80.5	Black
1404	John	25	M	183	80.5	Black
1504	John	25	M	183	80.5	Black
1604	John	25	M	183	80.5	Black
1704	John	25	M	183	80.5	Black
1804	John	25	M	183	80.5	Black
1904	John	25	M	183	80.5	Black
2004	John	25	M	183	80.5	Black
2104	John	25	M	183	80.5	Black
2204	John	25	M	183	80.5	Black
2304	John	25	M	183	80.5	Black
2404	John	25	M	183	80.5	Black
2504	John	25	M	183	80.5	Black
2604	John	25	M	183	80.5	Black
2704	John	25	M	183	80.5	Black
2804	John	25	M	183	80.5	Black
2904	John	25	M	183	80.5	Black
3004	John	25	M	183	80.5	Black
3104	John	25	M	183	80.5	Black
3204	John	25	M	183	80.5	Black
3304	John	25	M	183	80.5	Black
3404	John	25	M	183	80.5	Black
3504	John	25	M	183	80.5	Black
3604	John	25	M	183	80.5	Black
3704	John	25	M	183	80.5	Black
3804	John	25	M	183	80.5	Black
3904	John	25	M	183	80.5	Black
4004	John	25	M	183	80.5	Black
4104	John	25	M	183	80.5	Black
4204	John	25	M	183	80.5	Black
4304	John	25	M	183	80.5	Black
4404	John	25	M	183	80.5	Black
4504	John	25	M	183	80.5	Black
4604	John	25	M	183	80.5	Black
4704	John	25	M	183	80.5	Black
4804	John	25	M	183	80.5	Black
4904	John	25	M	183	80.5	Black
5004	John	25	M	183	80.5	Black
5104	John	25	M	183	80.5	Black
5204	John	25	M	183	80.5	Black
5304	John	25	M	183	80.5	Black
5404	John	25	M	183	80.5	Black
5504	John	25	M	183	80.5	Black
5604	John	25	M	183	80.5	Black
5704	John	25	M	183	80.5	Black
5804	John	25	M	183	80.5	Black
5904	John	25	M	183	80.5	Black
6004	John	25	M	183	80.5	Black
6104	John	25	M	183	80.5	Black
6204	John	25	M	183	80.5	Black
6304	John	25	M	183	80.5	Black
6404	John	25	M	183	80.5	Black
6504	John	25	M	183	80.5	Black
6604	John	25	M	183	80.5	Black
6704	John	25	M	183	80.5	Black
6804	John	25	M	183	80.5	Black
6904	John	25	M	183	80.5	Black
7004	John	25	M	183	80.5	Black
7104	John	25	M	183	80.5	Black
7204	John	25	M	183	80.5	Black
7304	John	25	M	183	80.5	Black
7404	John	25	M	183	80.5	Black
7504	John	25	M	183	80.5	Black
7604	John	25	M	183	80.5	Black
7704	John	25	M	183	80.5	Black
7804	John	25	M	183	80.5	Black
7904	John	25	M	183	80.5	Black
8004	John	25	M	183	80.5	Black
8104	John	25	M	183	80.5	Black
8204	John	25	M	183	80.5	Black
8304	John	25	M	183	80.5	Black
8404	John	25	M	183	80.5	Black
8504	John	25	M	183	80.5	Black
8604	John	25	M	183	80.5	Black
8704	John	25	M	183	80.5	Black
8804	John	25	M	183	80.5	Black
8904	John	25	M	183	80.5	Black
9004	John	25	M	183	80.5	Black
9104	John	25	M	183	80.5	Black
9204	John	25	M	183	80.5	Black
9304	John	25	M	183	80.5	Black
9404	John	25	M	183	80.5	Black
9504	John	25	M	183	80.5	Black
9604	John	25	M	183	80.5	Black
9704	John	25	M	183	80.5	Black
9804	John	25	M	183	80.5	Black
9904	John	25	M	183	80.5	Black
10004	John	25	M	183	80.5	Black



**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

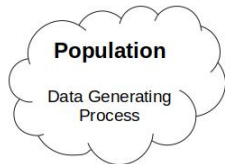
The **variance** of a data set is defined as:

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

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

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

# Sample data vs Population distribution

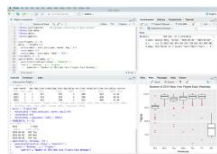


**Data**

id	name	age	sex	height	weight	hair
204	John	25	M	183	80.5	Black
304	John	25	M	183	80.5	Black
404	John	25	M	183	80.5	Black
504	John	25	M	183	80.5	Black
604	John	25	M	183	80.5	Black
704	John	25	M	183	80.5	Black
804	John	25	M	183	80.5	Black
904	John	25	M	183	80.5	Black
1004	John	25	M	183	80.5	Black
1104	John	25	M	183	80.5	Black
1204	John	25	M	183	80.5	Black
1304	John	25	M	183	80.5	Black
1404	John	25	M	183	80.5	Black
1504	John	25	M	183	80.5	Black
1604	John	25	M	183	80.5	Black
1704	John	25	M	183	80.5	Black
1804	John	25	M	183	80.5	Black
1904	John	25	M	183	80.5	Black
2004	John	25	M	183	80.5	Black
2104	John	25	M	183	80.5	Black
2204	John	25	M	183	80.5	Black
2304	John	25	M	183	80.5	Black
2404	John	25	M	183	80.5	Black
2504	John	25	M	183	80.5	Black
2604	John	25	M	183	80.5	Black
2704	John	25	M	183	80.5	Black
2804	John	25	M	183	80.5	Black
2904	John	25	M	183	80.5	Black
3004	John	25	M	183	80.5	Black
3104	John	25	M	183	80.5	Black
3204	John	25	M	183	80.5	Black
3304	John	25	M	183	80.5	Black
3404	John	25	M	183	80.5	Black
3504	John	25	M	183	80.5	Black
3604	John	25	M	183	80.5	Black
3704	John	25	M	183	80.5	Black
3804	John	25	M	183	80.5	Black
3904	John	25	M	183	80.5	Black
4004	John	25	M	183	80.5	Black
4104	John	25	M	183	80.5	Black
4204	John	25	M	183	80.5	Black
4304	John	25	M	183	80.5	Black
4404	John	25	M	183	80.5	Black
4504	John	25	M	183	80.5	Black
4604	John	25	M	183	80.5	Black
4704	John	25	M	183	80.5	Black
4804	John	25	M	183	80.5	Black
4904	John	25	M	183	80.5	Black
5004	John	25	M	183	80.5	Black
5104	John	25	M	183	80.5	Black
5204	John	25	M	183	80.5	Black
5304	John	25	M	183	80.5	Black
5404	John	25	M	183	80.5	Black
5504	John	25	M	183	80.5	Black
5604	John	25	M	183	80.5	Black
5704	John	25	M	183	80.5	Black
5804	John	25	M	183	80.5	Black
5904	John	25	M	183	80.5	Black
6004	John	25	M	183	80.5	Black
6104	John	25	M	183	80.5	Black
6204	John	25	M	183	80.5	Black
6304	John	25	M	183	80.5	Black
6404	John	25	M	183	80.5	Black
6504	John	25	M	183	80.5	Black
6604	John	25	M	183	80.5	Black
6704	John	25	M	183	80.5	Black
6804	John	25	M	183	80.5	Black
6904	John	25	M	183	80.5	Black
7004	John	25	M	183	80.5	Black
7104	John	25	M	183	80.5	Black
7204	John	25	M	183	80.5	Black
7304	John	25	M	183	80.5	Black
7404	John	25	M	183	80.5	Black
7504	John	25	M	183	80.5	Black
7604	John	25	M	183	80.5	Black
7704	John	25	M	183	80.5	Black
7804	John	25	M	183	80.5	Black
7904	John	25	M	183	80.5	Black
8004	John	25	M	183	80.5	Black
8104	John	25	M	183	80.5	Black
8204	John	25	M	183	80.5	Black
8304	John	25	M	183	80.5	Black
8404	John	25	M	183	80.5	Black
8504	John	25	M	183	80.5	Black
8604	John	25	M	183	80.5	Black
8704	John	25	M	183	80.5	Black
8804	John	25	M	183	80.5	Black
8904	John	25	M	183	80.5	Black
9004	John	25	M	183	80.5	Black
9104	John	25	M	183	80.5	Black
9204	John	25	M	183	80.5	Black
9304	John	25	M	183	80.5	Black
9404	John	25	M	183	80.5	Black
9504	John	25	M	183	80.5	Black
9604	John	25	M	183	80.5	Black
9704	John	25	M	183	80.5	Black
9804	John	25	M	183	80.5	Black
9904	John	25	M	183	80.5	Black
10004	John	25	M	183	80.5	Black



**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

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

and

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

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

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

## 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

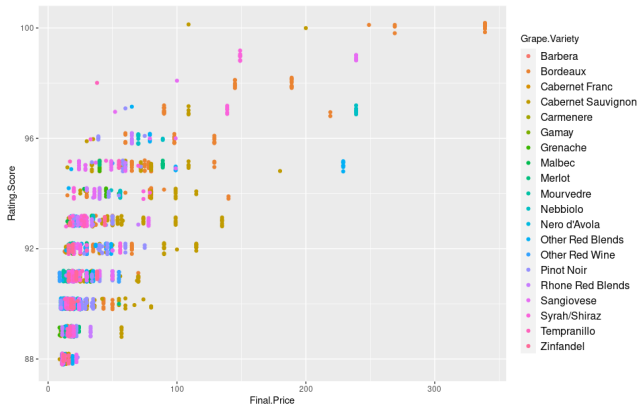


Figure: Wine Price vs Wine Rating from wine.com

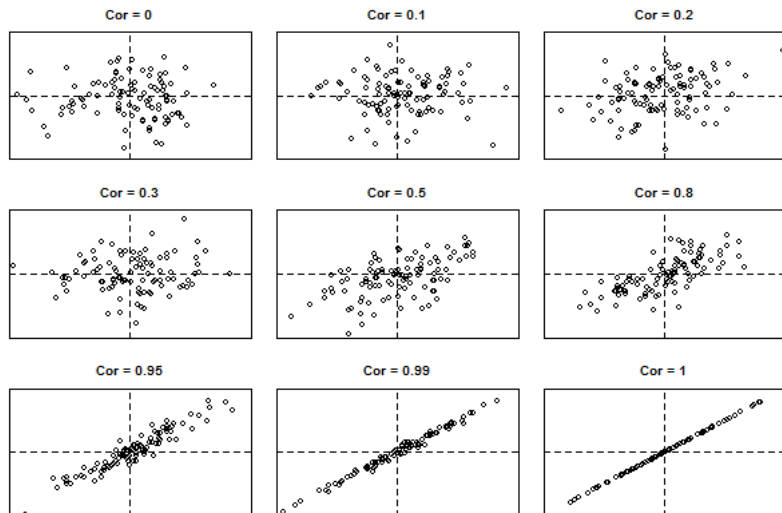
# Correlation

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



# Correlation

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

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

where

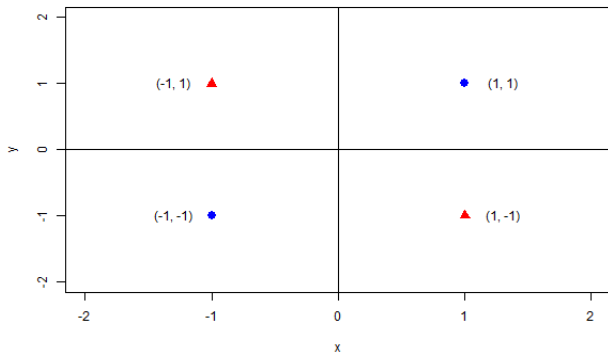
$$\text{Sample SD of } X = s_X = \sqrt{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2}$$

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

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

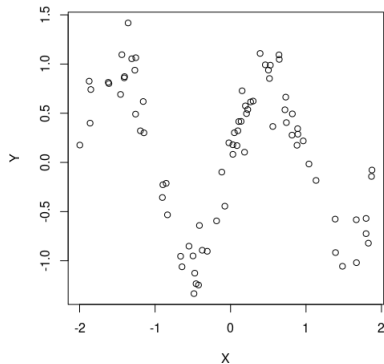
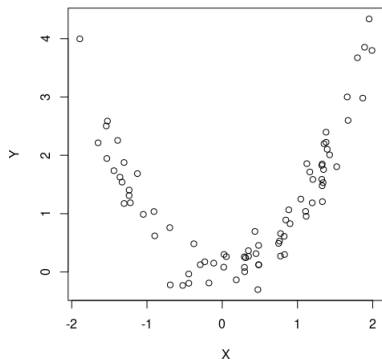
# Sample Covariance

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



# 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