

Lecture 2: Correlation

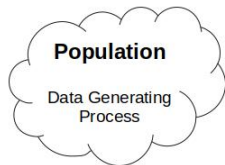
Module 1, part 1

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

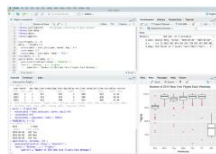


Data

A screenshot of a data table with columns labeled 'id', 'name', 'age', 'height', 'weight', 'gender', and 'status'. The table contains 100 rows of data.



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

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

Measuring spread of data

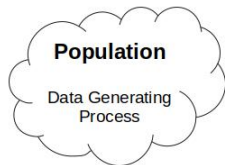
The **variance** of a data set is defined as:

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

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	parent1	parent2
101	John	15	170	65	M	101	101
102	Jane	14	160	55	F	101	101
103	Mike	16	180	75	M	102	102
104	Sarah	15	165	60	F	102	102
105	David	17	190	80	M	103	103
106	Emily	16	175	68	F	103	103
107	Chris	15	168	58	M	104	104
108	Alice	14	158	52	F	104	104
109	Bob	16	178	70	M	105	105
110	Anna	15	162	56	F	105	105
111	Tom	17	185	78	M	106	106
112	Lisa	16	172	66	F	106	106
113	Mark	15	167	59	M	107	107
114	Karen	14	155	50	F	107	107
115	James	16	179	71	M	108	108
116	Helen	15	164	57	F	108	108
117	Steven	17	188	79	M	109	109
118	Nancy	16	173	67	F	109	109
119	Paul	15	166	59	M	110	110
120	Rachel	14	156	51	F	110	110
121	Kevin	16	177	69	M	111	111
122	Michelle	15	161	54	F	111	111
123	Andrew	17	186	81	M	112	112
124	Stephanie	16	174	68	F	112	112
125	Timothy	15	169	60	M	113	113
126	Victoria	14	159	53	F	113	113
127	Jonathan	16	176	67	M	114	114
128	Christina	15	163	56	F	114	114
129	Benjamin	17	189	82	M	115	115
130	Samantha	16	175	69	F	115	115
131	Gregory	15	168	60	M	116	116
132	Kathleen	14	157	52	F	116	116
133	Anthony	16	179	72	M	117	117
134	Ashley	15	165	58	F	117	117
135	Christopher	17	187	80	M	118	118
136	Brittany	16	173	67	F	118	118
137	Nicholas	15	167	59	M	119	119
138	Heather	14	156	51	F	119	119
139	Timothy	16	178	70	M	120	120
140	Elizabeth	15	162	56	F	120	120
141	Matthew	17	188	81	M	121	121
142	Olivia	16	174	68	F	121	121
143	Christopher	15	169	60	M	122	122
144	Madison	14	158	53	F	122	122
145	Andrew	16	177	69	M	123	123
146	Chloe	15	164	57	F	123	123
147	Benjamin	17	189	82	M	124	124
148	Grace	16	175	69	F	124	124
149	Christopher	15	168	60	M	125	125
150	Isabella	14	157	52	F	125	125



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 μ_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 μ_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 μ_X and true variance of σ_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 μ_X and true variance of σ_X^2 . How would we estimate the true variance if it is unknown?

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

If we knew μ_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 μ_X and true variance of σ_X^2 . How would we estimate the true variance if it is unknown?

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

If we knew μ_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 μ_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

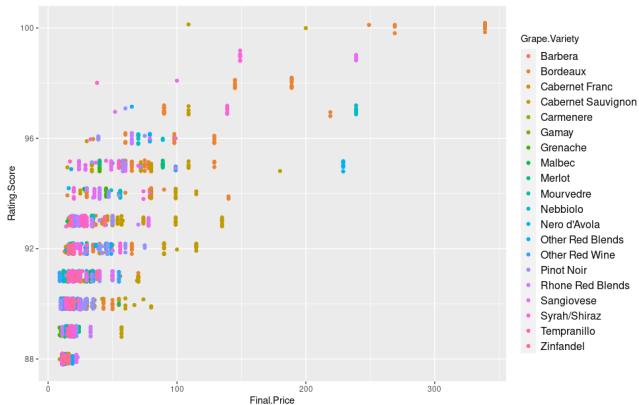


Figure: Wine Price vs Wine Rating from wine.com

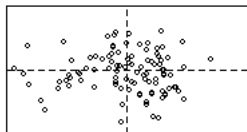
Correlation

Correlation measures the linear dependence between two variables.

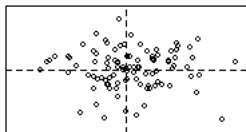
- For two variables, X and Y , correlation is denoted by r_{XY}
- Correlation is between -1 and 1
- $r_{XY} = 0$ indicates no **linear** relationship
- $r_{XY} > 0$ indicates positive **linear** relationship
- $r_{XY} < 0$ indicates negative **linear** relationship
- $r_{XY} = \pm 1$ indicates perfect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

Correlation

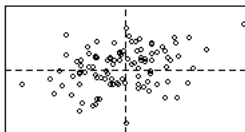
Cor = 0



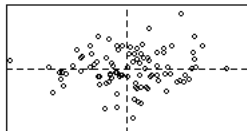
Cor = 0.1



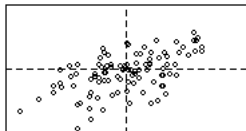
Cor = 0.2



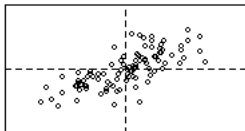
Cor = 0.3



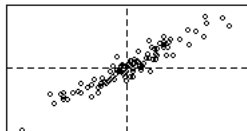
Cor = 0.5



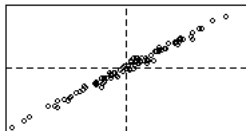
Cor = 0.8



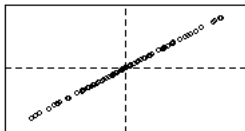
Cor = 0.95



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

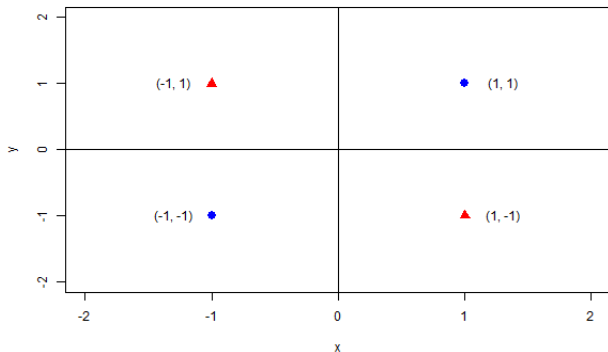
$$\text{Sample SD of } X = s_X = \sqrt{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2}$$

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

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

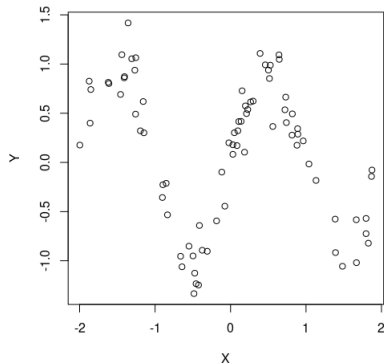
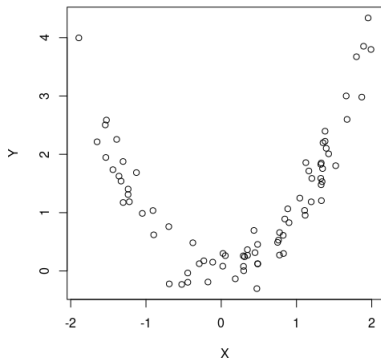
Sample Covariance

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



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