Lecture 2: Correlation

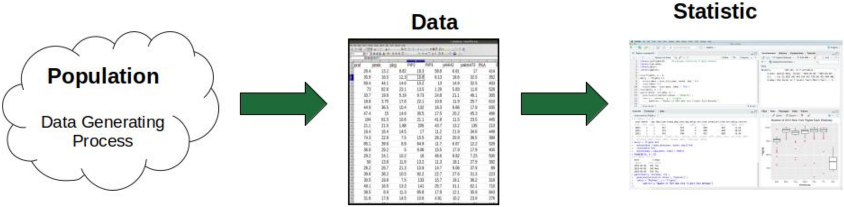
Module 1, part 1

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



Summarizing a data set

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

Summarizing a data set

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

- Mean:

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

- 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 ith 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 ith 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 ith 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 $x_1, x_2, \ldots x_n$.
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

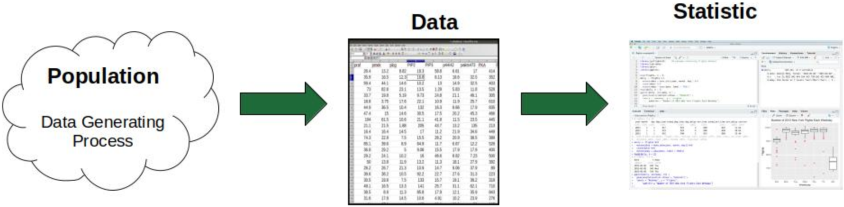
The **variance** of a data set is defined as:

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

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

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

Sample data vs Population distribution



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

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

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

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

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 $x_1, x_2, \ldots, x_n$
- $\hat{\sigma}_x^2$ denotes the variance of the observations

Estimating the variance

Suppose we have some observations $x_1, x_2, \ldots, x_n$ 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?

$$\sigma_x^2 = E\left[(X - \mu_X)^2\right]$$

Estimating the variance

Suppose we have some observations $x_1, x_2, \ldots, x_n$ 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?

$$\sigma_x^2 = E\left[(X - \mu_X)^2\right]$$

If we knew μ_X, we could use

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

and

$$E(\hat{\sigma}_x^2) = \sigma_x^2$$

Estimating the variance

Suppose we have some observations $x_1, x_2, \ldots, x_n$ 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?

$$\sigma_x^2 = E\left[(X - \mu_X)^2\right]$$

If we knew μ_X, we could use

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

and

$$E(\hat{\sigma}_x^2) = \sigma_x^2$$

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

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

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

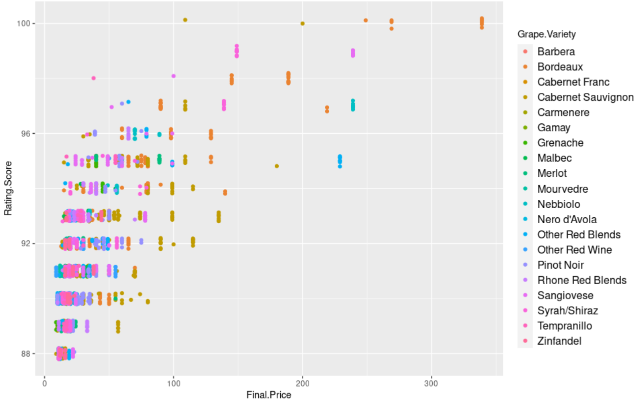


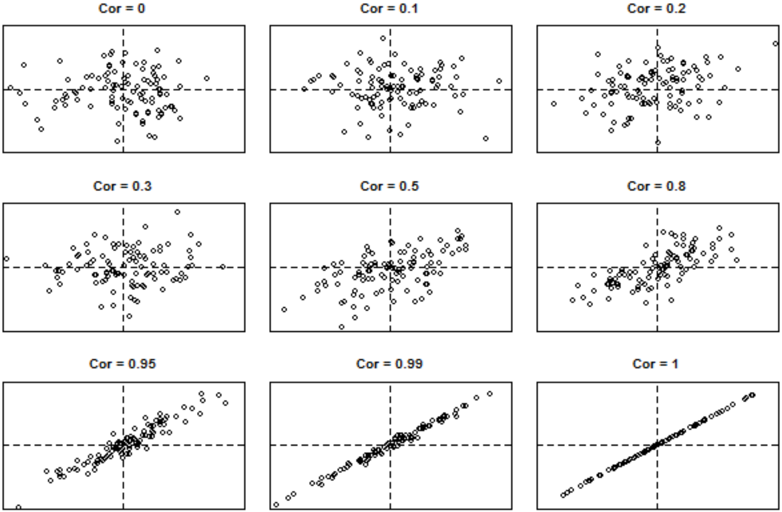
Figure: Wine Price vs Wine Rating from wine.com

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

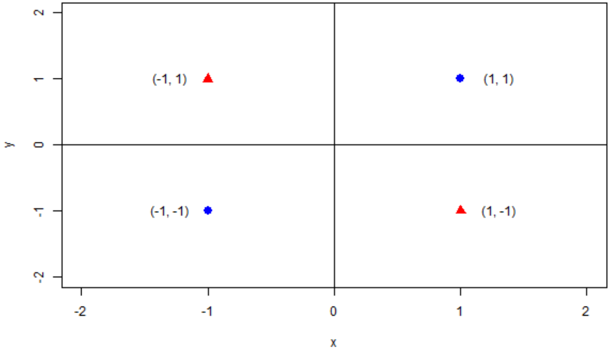
$$\text{Sample SD of X} = \text{s}_X = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$$

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

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

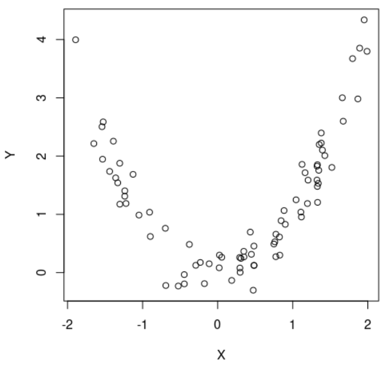
Sample Covariance

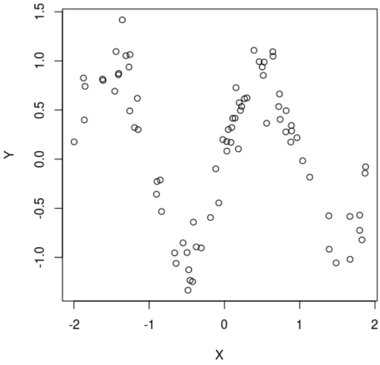
$$s_{XY} = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$



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