INTELLIGENT SYSTEMS LABORATORY

Lesson 3
Descriptive Statistics

Mathematics and Statistics for Data Science

Tapanan Yeophantong
Vincent Mary School of Science and Technology
Assumption University

Content

- Point & variability estimations
- Percentiles, boxplots & histograms (Practical)

Summary Statistics

Let's start easy ...

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Sample Mean

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Sample Variance

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Sample Standard Deviation

Summary Statistics for Categorical Data

- Two most commonly used numerical summaries for categorical data are *frequencies* and *sample proportions*.
- The frequency for a given category is simply the number of sample items that fall into that category.
- The sample proportion is the frequency divided by the sample size.

Example 1

- A simple random sample of 5 men is chosen from a large population of men, and their heights are measured. The five heights (in inches) are 65.51, 72.30, 68.31, 67.05, and 70.68.
  - Find the sample mean.
  - Find the sample variance.
  - Find the sample standard deviation.

Example 1 - Solution

- A simple random sample of 5 men is chosen from a large population of men, and their heights are measured. The five heights (in inches) are 65.51, 72.30, 68.31, 67.05, and 70.68.

$$\overline{X} = \frac{1}{5}(65.51 + 72.30 + 68.31 + 67.05 + 70.68) = 68.77 \text{ in.}$$

$$s^2 = \frac{1}{4}[(65.51 - 68.77)^2 + (72.30 - 68.77)^2 + (68.31 - 68.77)^2 + (67.05 - 68.77)^2 + (70.68 - 68.77)^2] = 7.47665$$

$$s = \sqrt{7.47665} = 2.73$$

Example 2

- A process manufactures bearings for an engine. Bearings whose thicknesses are between 1.486 and 1.490 mm are classified as conforming. Bearings *thicker* than this are reground, and bearings *thinner* than this are scrapped. In a sample of 1,000 bearings, 910 were conforming (C), 53 were reground (R), and 37 were scrapped (S).
  - Find the frequencies of C, R, and S.
  - Find the sample proportions of C, R, and S.

Example 2 - Solution

- A process manufactures bearings for an engine. Bearings whose thicknesses are between 1.486 and 1.490 mm are classified as conforming. Bearings thicker than this are reground, and bearings thinner than this are scrapped. In a sample of 1,000 bearings, 910 were conforming (C), 53 were reground (R), and 37 were scrapped (S).

  - The frequencies of C, R, and S are 910, 53, and 37, respectively.
  - The sample proportions of C, R, and S are:
    - p(C) = 910/1000 = 0.910
    - p(R) = 53/1000 = 0.053
    - p(S) = 37/1000 = 0.037

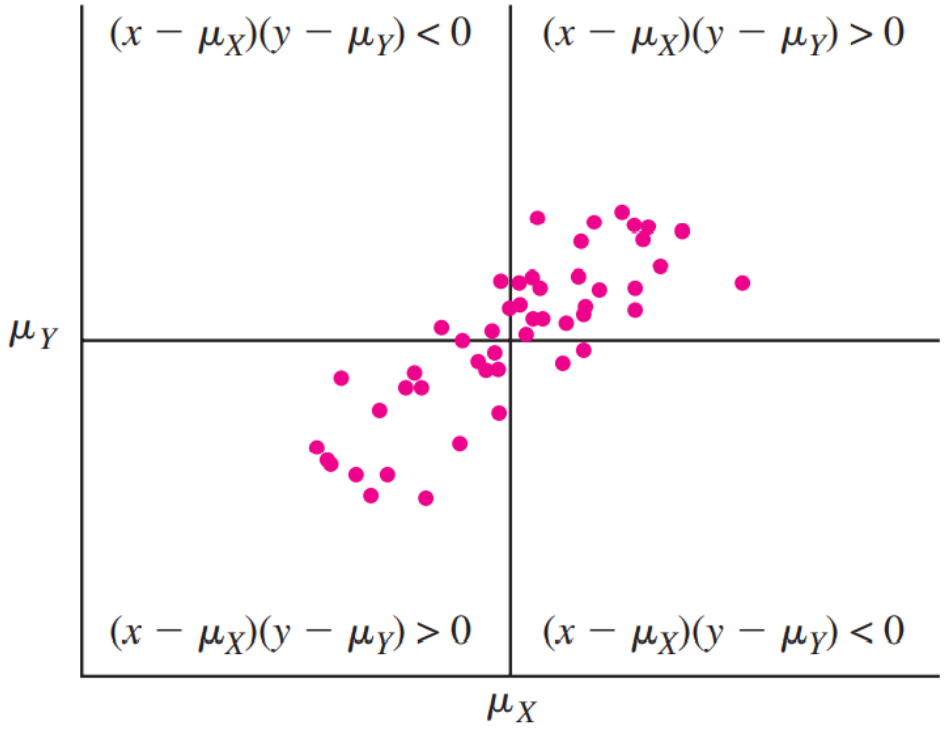
Covariance

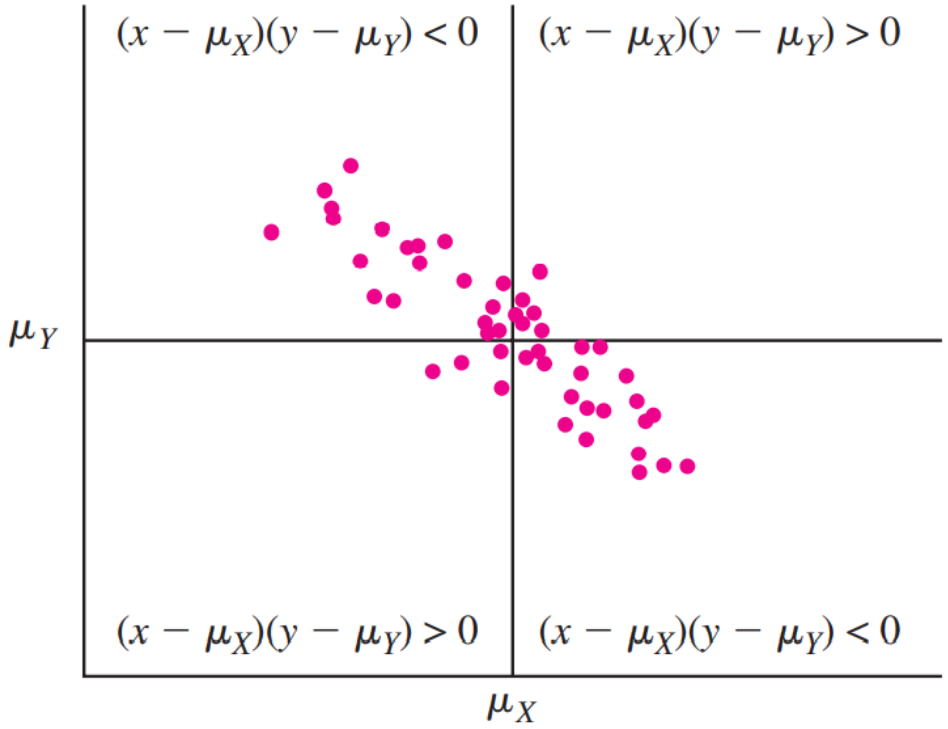
- When two random variables are not independent, it is useful to have a measure of the strength of the relationship between them.
- The covariance (*Cov(X,Y)*) is a measure of a linear relationship between the random variables.

$$\text{Cov}(X,Y) = \mu_{(X-\mu_X)(Y-\mu_Y)}$$

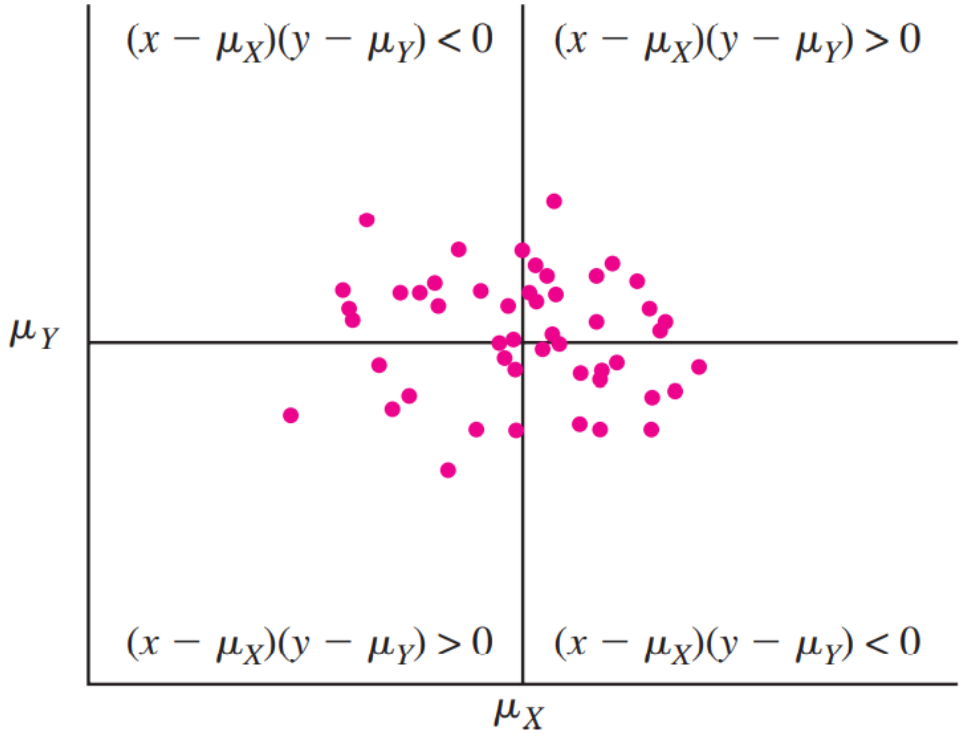
$$\text{Cov}(X,Y) = \mu_{XY} - \mu_X\mu_Y$$

Positive & Negative Covariances





Weak Covariances



Correlation

- Correlation is a measure of the strength of a linear relationship that is unitless, so that they can be compared.
- Let X and Y be jointly distributed random variables with standard deviations σX and σY, the correlation between X and Y (ρX,Y), is:

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$