Statistics 1

Sayan Das (dassayan0013@gmail.com)

December 14, 2023

Probability
Random variables, Probability distribution, Cumulative Probability distribution, Probability mass function, Probability density function, Expectations, Mean, Variance, Moment about a point, Raw moments, Central moments, Skewness, Kurtosis, Probability distribution function of two variables.

Statistics
Bivariate data, Scatter Diagram, Two-way frequency distribution, Marginal frequency distribution, Conditional frequency distribution, Covariance, Simple Correlation, Correlation coefficient, Cauchy Schwartz inequality, Properties of correlation coefficient, Regression analysis, Least square analysis, Regression lines and their properties, Rank data, Rank Correlation, Spearman's Rank Correlation - it's derivation and properties, Spearman's Rank Correlation with perfect agreement, Spearman's Rank Correlation with perfect disagreement.

Practical
Calculation of correlation coefficient, Regression and Calculation of Spearman's Rank Correlation. [50]

Contents

§0 Preliminaries

Definition 0.1

Consider n values $\{x_i\}_{i=1}^n$ of the variable x. Then we define the following:

1. **Mean**, or **expectation**: $\bar{x} = \mu = \mathbb{E}[x] := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$.
2. **Variance** : $\mathrm{Var}(x) := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2 = \mathbb{E}[(x-\bar{x})^2]$.
3. **Standard deviation**: $\mathrm{S}_x := \sqrt{\mathrm{Var}(x)} = \sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2}$.
4. **Mean deviation about mean**: $\mathrm{MD}_{\bar{x}} := \mathbb{E}[\ |x-\bar{x}|\ ] = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}|x_i - \bar{x}|$.

§1 Probability

§1.1 Probability distributions and densities

Definition 1.1 (Random variable)

If S is a sample space with a probability measure and X is a real-valued function defined over S, then X is called a **random variable**. If the sample space is finite or countably infinite then X is a discrete random variable, whilst for continuous sample spaces we have continuous random variables.

§1.1.1 Univariate case

Definition 1.2

If X is a discrete random variable, then the function

$$f(x) = P(X = x)\ \forall\ \text{x within the range of X}$$

is called the **probability mass function** or probability distribution of X.
An association such as,

$$\begin{array}{rccccc} X: & x_1 & x_2 & \dots & x_i & \dots \\ P(X=x): & p_1 & p_2 & \dots & p_i & \dots \end{array}$$

is called a **discrete probability distribution** of the random variable X.

Theorem 1.3

If X is a discrete random variable, then the function $f(x)$ can serve as the probability distribution of X iff,

1. $f(x) \geq 0$ for all values in the function's domain,

2. $\sum_x f(x) = 1$, where the summation extends over all values within the function's domain.

Definition 1.4

If X is a discrete random variable, then the function

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \quad \text{for} \quad -\infty < x < \infty$$

is called the **cumulative probability distribution function** or distribution function of X.
($f(t)$ being the probability distribution of X at t.)

Theorem 1.5

If X is a discrete random variable, then the following hold for the cumulative probability distribution $F(X)$,

1. $F(-\infty) = 0$ and $F(\infty) = 1$,
2. $\forall a, b \in \mathbb{R},\ a < b \implies F(a) \leq F(b)$.

Theorem 1.6

If X is a discrete random variable with its range consisting of the values

$$x_1 < x_2 < x_3 < ... < x_n$$

then $f(x_1) = F(x_1)$ and

$$f(x_i) = F(x_i) - F(x_{i-1}) \quad \text{for } i \geq 2.$$

Definition 1.7

If X is a continuous random variable and a function $f(x)$ is defined for all $x \in \mathbb{R}$, then it is called a **probability density function** of X iff

$$P(a \leq X \leq b) = \int_a^b f(x)dx, \quad \text{for any real constants } \ a \ \text{ and } \ b, \ a \leq b.$$

Remark. We have $P(X = c) = 0$ for any real constant c.

Corollary 1.8

If X is a continuous random variable and a and b are real constants with $a \leq b$, then

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b).$$

Theorem 1.9

If X is a continuous random variable, then a function $f(x)$ can serve as a probability density function of X iff,

1. $f(x) \geq 0 \quad \text{for} \quad -\infty < x < \infty$,
2. $\int_{-\infty}^{\infty} f(x) = 1$.

Definition 1.10

If X is a continuous random variable and the value of its probability density at t is $f(t)$, then the function given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt, \quad \text{for} \quad -\infty < x < \infty$$

is called the **cumulative probability distribution function** or distribution function of X.

Theorem 1.11

If $f(x)$ and $F(x)$ are the values of the probability density and distribution function of X at x respectively, then

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a),$$

$$\forall \text{ real constants } a \text{ and } b, \ a \leq b, \quad \text{and}$$

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$$

provided the derivative exists.

§1.1.2 Bivariate case

Definition 1.12

If X and Y are discrete random variables, then the function

$$f(x, y) = P(X = x, Y = y)$$

$$\forall \text{ pair } (x, \ y) \text{ within the range of X and Y}$$

is called the **joint probability mass function** or joint probability distribution of X and Y.

Theorem 1.13

A bivariate function $f(x, y)$ can serve as the joint probability distribution of a pair of discrete random variables X and Y iff,

1. $f(x,y) \geq 0$ for all pairs $(x,\ y)$ within the function's domain,
2. $\sum\limits_{x}\sum\limits_{y} f(x,y) = 1$, where the double summation extends over all possible pairs $(x,\ y)$ within the function's domain.

Definition 1.14

If X and Y are discrete random variables, then the function

$$F(x,y) = P(X \leq x, Y \leq y) = \sum_{s \leq x}\sum_{t \leq y} f(s,t),$$

$$\text{for} \quad -\infty < x < \infty,\ -\infty < x < \infty$$

is called the **joint cumulative distribution function** or joint distribution function of X and Y.

Definition 1.15

A bivariate function $f(x,y)$ defined over the xy-plane is called a **joint probability density function** of the continuous random variables X and Y iff,

$$P[(X,Y) \in A] = \iint_A f(x,y)dxdy$$

for any region A in the xy-plane.

Theorem 1.16

A bivariate function $f(x,y)$ can serve as the joint probability density function of the continuous random variables X and Y iff,

1. $f(x,y) \geq 0 \quad \text{for} \quad -\infty < x < \infty, -\infty < y < \infty,$
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1.$

Definition 1.17

If X and Y are continuous random variables, then the function

$$F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(s,t)dsdt,$$

$$\text{for} \quad -\infty < x < \infty,\ -\infty < x < \infty$$

is called the **joint cumulative distribution function** or joint distribution function of X and Y.

Remark. We have

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$$

whenever the partial derivatives exist.

§2 Statistics

§2.1 Standard deviation

Theorem 2.1

Standard deviation cannot be less than the mean deviation about mean.

Proof. Consider n pairs of values $x_i \ \forall i = 1, \ ..., \ n$ of the variable x. By definition,

$$\mathrm{S}_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

is the standard deviation of x and

$$\mathrm{MD}_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

is the mean deviation about mean of x. We define

$$a_i = |x_i - \bar{x}| \ \forall i = 1, \ ..., \ nb_i = 1 \ \forall i = 1, \ ..., \ n$$

then by the Cauchy-Schwarz inequality, we have

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \geq \left(\sum_{i=1}^{n} a_i b_i\right)^2$$

$$\iff \left(\sum_{i=1}^{n} |x_i - \bar{x}|^2\right)\left(\sum_{i=1}^{n} 1^2\right) \geq \left(\sum_{i=1}^{n} |x_i - \bar{x}| \cdot 1\right)^2$$

$$\iff \left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) n \geq \left(\sum_{i=1}^{n}|x_i - \bar{x}|\right)^2$$

$$\iff \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) \geq \frac{1}{n^2}\left(\sum_{i=1}^{n}|x_i - \bar{x}|\right)^2$$

$$\iff \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \geq \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

$$\iff \mathrm{S}_x \geq \mathrm{MD}_{\bar{x}}.$$

□

§2.2 Bivariate data

Data that is collected simultaneously for two variables is bivariate. Consider a class of students with their heights and weights collected simultaneously for the variables, say, X_a and Y_a respectively, such that,

X_a	Y_a
x_1	y_1
x_2	y_2

Consider another bivariate data, where the marks of the students in maths and physics are collected simultaneously for the variables, say, X and Y respectively.
In bivariate analysis there are two fundamental types of problems,

1. correlation analysis, and
2. regression analysis.

§2.2.1 Correlation analysis

In correlation analysis we want to study the nature of existence of association (if any) between the variables. Considering the second piece of bivariate data for the class, our aim is to make an association, or **correlation** between the marks in maths and marks in physics, i.e. a relation between X and Y.

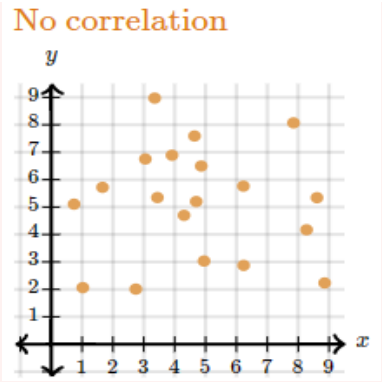
§2.2.2 Regression analysis

In regression analysis we want to find express one variable (dependent) as a function of another variable (independent), so that the value of the dependent variable can be predicted whenever the independent variable's value is known. Of course, this is only possible once an association has been established between the variables. Thus regression analysis is only valid when there is a correlation between the variables.

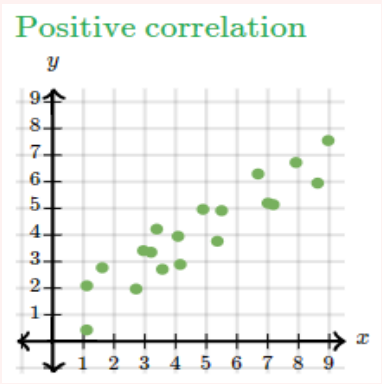
§2.2.3 Scatter or dot diagram

Scatter diagram is a graphical representation of bivariate data. Suppose we have n pairs of values $\{(x_i, y_i)\}_{i=1}^n$. If we plot the points (x_i, y_i) in $\mathbb{R}^2$ (or the 2-dimensional xy-plane), the diagram so obtained is called a scatter or dot diagram.
From a scatter diagram we may study the nature of existence or association between the two variables.

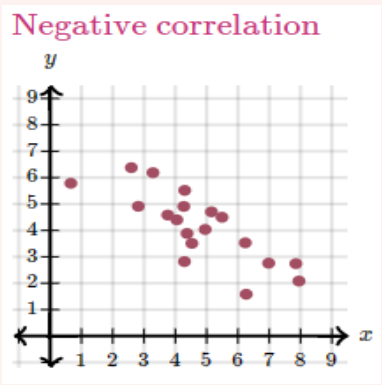
Example 2.2 1. When there is no clear relation between x_i and y_i,



2. When there is a linear relation between x_i and y_i with positive slope,



3. When there is a linear relation between x_i and y_i with negative slope,



§2.3 Covariance

Definition 2.3

Consider n pairs of values $\{(x_i, y_i)\}_{i=1}^n$ of the variables x and y. Then

$$\mathrm{Cov}(x, y) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

is the **covariance of x and y**.

§2.4 Correlation coefficient

§2.5 Regression

§2.6 Spearman's Rank Correlation