Continuous Random Variables

General Information

- A function $f\colon \mathbb{R} \to \mathbb{R}$ is a *probability mass function* (pdf) of a continuous random variable X iff f is nonnegative and $\int_{-\infty}^{\infty} f(x)\,dx = 1$.
- For any probability mass function f, we have $\mathrm{P}(a \le X \le b) = \int_a^b f(x)\,dx$. Whether the inequality is strict or nonstrict does not affect the above identity.
- A *mode* of X is any value m such that $f(m)$ is maximum.
- A *cumulative distribution function* (cdf) $F\colon \mathbb{R} \to [0,1]$ of a random variable X is defined by
$$F(x) := P(X \le x) = \int_{-\infty}^{x} f(x)\,dx.$$
- When writing out the cdf as a piecewise function, we explicitly write out the range of values for each case. We reserve the use of "otherwise" for pdf's.
- Any cdf is continuous and nondecreasing.
- Let X be a continuous random variable with cdf F. To find the pdf g of any $y(X)$, we first find its cdf, then differentiate. We achieve this by reverse engineering $y(X) \le y$ to find an inequality that relates X with y. E.g. $e^X \le y$ iff $X \le \ln(y)$.
- A *median* of X is any value m such that $\mathrm{P}(X \le m) = F(m) = 1/2$.
- Mean/Expectation:
$$\mu = \mathrm{E}(X) := \int_{-\infty}^{\infty} x f(x)\,dx \qquad \text{and} \qquad \mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$
- Important property:
$$\mathrm{E}(ag(X) \pm bh(x)) = a\,\mathrm{E}(g(X)) \pm \mathrm{E}(h(X)).$$
- Variance:
$$\mathrm{Var}(X) := \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.$$
- Important property:
$$\mathrm{Var}(aX \pm b) = a^2\,\mathrm{Var}(X).$$
- A continuous random variable X has a *uniform distribution* over the interval $[a,b]$ iff its pdf f is such that
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b, \\ 0 & \text{otherwise.} \end{cases}$$

Special Continuous Random Variables

Definition 2.1

A continuous random variable X has a *normal distribution* with mean μ and standard deviation σ, denoted by $X \sim \mathrm{N}(\mu, \sigma^2)$, iff its pdf f is such that

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

General Information

1. A normal distribution is symmetrical about the line $x = \mu$. That is

   $$\mathrm{P}(X \leq \mu - \delta) = \mathrm{P}(X \geq \mu + \delta)$$

   for each $\delta > 0$. Note that the mean, median, and mode coincide with μ.

2. Properties of the normal distribution. Let X and Y be independent, such that $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Y \sim \mathrm{N}(m, s^2)$. Then, for any $n \in \mathbb{N}$ and x, $y \in \mathbb{R}$,

   (a) $nX \sim \mathrm{N}(n\mu, n^2\sigma^2)$,

   (b) $X_1 + X_2 + \cdots + X_n \sim \mathrm{N}(n\mu, n\sigma^2)$,

   (c) $aX \pm bY \sim \mathrm{N}(a\mu \pm bm, a^2\sigma^2 + b^2s^2)$.

3. A variable $Z \sim \mathrm{N}(0, 1)$ is said to follow the *standard* normal distribution.

   Note: Z is reserved for this purpose.

4. Let $X \in \mathrm{N}(\mu, \sigma^2)$. Then, $\frac{X-\mu}{\sigma}$ follows the standard normal distribution.

Correlation and Linear Regression

Note

A good scatter diagram should follow the guidelines below.

- The relative position of each point on the scatter diagram should be clearly shown.
- The range of values for the set of data should be clearly shown by marking out the extreme x and y values on the corresponding axis.
- The axes should be labeled clearly with the variables.

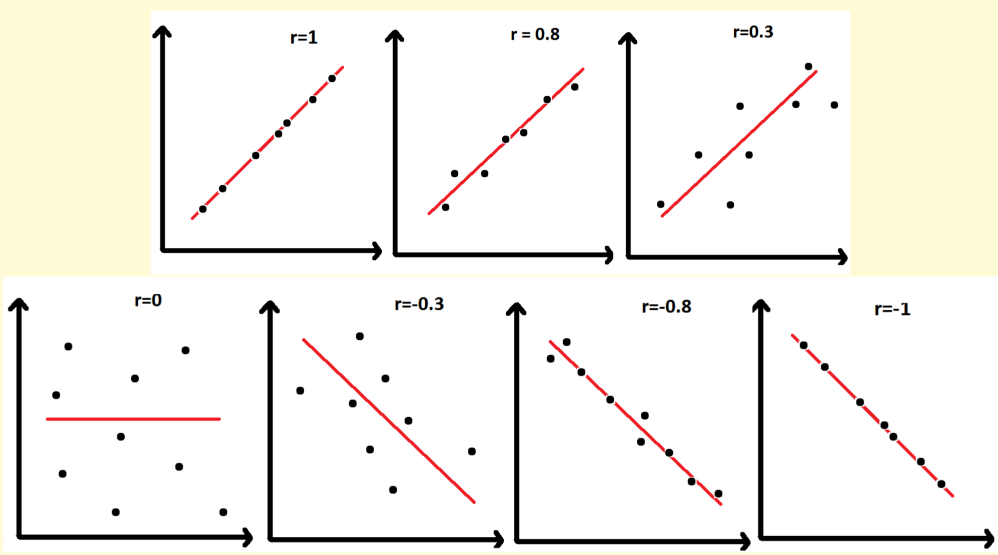
General Information

- The Product Moment Correlation Coefficient is a measure of the linear correlation between two variables. It is defined by

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \dfrac{(\sum x)^2}{n}\right]\left[\sum y^2 - \dfrac{(\sum y)^2}{n}\right]}},$$

  which takes on a value from 0 to 1.

- When $r = 0$, there is no linear relationship. But, a nonlinear relationship may be present. Additionally, the regression lines are perpendicular.
- The closer the value of r is to 1 (or -1), the stronger the positive (or negative) linear correlation. Furthermore, the regression lines coincide.



- The regression line of y on x minimises the sum of squares deviation (error) in the y-direction. (i.e. we are assuming x is the independent variable whose values are known exactly.) It is given by

$$y = \bar{y} + b(x - \bar{x}), \qquad \text{where} \qquad b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}.$$

- The point $(\bar{x}, \bar{y})$ always lies on both the regression lines of y on x, and x on y.
- Say we are given the value of one variable, and asked to approximate the the value of the other variable. Then, we should always use the line of the *dependent* variable on the *independent*.
- Estimations should not be taken for data outside the range of the sample provided, even if the value of r is close to 1.