

APPENDIX Statistics Formulas

Below are a few formulas on random variables that we are likely to encounter throughout the book.

DEFINITIONS

Let $X, Y, \text{ and } Z$ be random variables. Let $(x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_N, y_N, z_N)$ be N realization 3-tuples for these random variables.

Mean

- The mean or expected value of X is denoted by $E[X] = \mu_x$.
- The estimated value of the mean of a random variable is known as the average.
- The formula for the average is $x_{avg} = \frac{1}{N} \sum_{i=1}^N x_i$. [Unbiased Estimator]

Variance

- The variance of X is $var(X) = E[(x - \mu_x)^2] = \sigma_x^2$
- The estimated value of the square root of variance is the familiar standard deviation.
- Its value is calculated using the formula $x_{stddev} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - x_{avg})^2}$.

Covariance

- The covariance between X and Y is denoted as $cov(X, Y) = E[(x - \mu_x)(y - \mu_y)]$.
- An estimation of the covariance may be calculated using the formula $\left[\frac{1}{N} \sum_{i=1}^N (x_i - x_{avg})(y_i - y_{avg}) \right]$ [Unbiased Estimator for $cov(X, Y)$]

Correlation

- The correlation between X and Y is $corr(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X) var(Y)}} = \rho_{xy}$
- The formula for the estimate of correlation is given as $\frac{\left[\frac{1}{N} \sum_{i=1}^N (x_i - x_{avg})(y_i - y_{avg}) \right]}{\left(\frac{x_{stddev}}{y_{stddev}} \right)}$ [Unbiased Estimator]

Important Introduction
 If $Z = \alpha X + \beta Y$
 $\Rightarrow \sigma_z^2 = \alpha^2 \sigma_x^2 + \beta^2 \sigma_y^2 + 2\alpha\beta cov(X, Y)$
 $\Rightarrow \sigma_z^2 = \alpha^2 \sigma_x^2 + \beta^2 \sigma_y^2 + 2\alpha\beta (\rho_{xy} \sigma_x \sigma_y)$

- The correlation between any two random variables is always a value between +1 and -1.
- Every random variable is perfectly correlated with itself, that is, the correlation is +1.
- Two random variables are said to be uncorrelated when the correlation between them is 0. This happens when the random variables are statistically independent.

FORMULAS

If α, β are nonrandom numbers, then the following formulas hold:
 determines the variables, or just numbers

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y] \quad E[X] = E[E[X|I]]$$

$$var(\alpha X + \beta Y) = \alpha^2 var(X) + \beta^2 var(Y) + 2\alpha\beta cov(X, Y) \quad var[X|I] \leq var[X]$$

$$cov(\alpha X + \beta Y, Z) = \alpha cov(X, Z) + \beta cov(Y, Z) \quad E[X \cdot Y] = E[Y \cdot E[X|I]] + cov(X, Y)$$

$$cov(\alpha X, \beta Y) = \alpha\beta cov(X, Y) \quad E[X \cdot Y] = E[X \cdot E[Y|I]]$$

$$corr(\alpha X, \beta Y) = corr(X, Y) \text{ if } \alpha\beta > 0; = -corr(X, Y) \text{ if } \alpha\beta < 0$$

$$cov(X, X) = var[X]; \quad cov(X, Y) = cov(Y, X)$$

$$cov(X, Y) = +1 \Leftrightarrow Y = \alpha + \beta X \text{ where } \beta > 0$$

$$cov(X, Y) = -1 \Leftrightarrow Y = \alpha + \beta X \text{ where } \beta < 0$$

OLS Linear Regression

$Y = \alpha + \beta X + \epsilon, \quad \epsilon \sim N(0, \sigma_\epsilon^2)$

Read: Regress Y on X $\left\{ \begin{array}{l} cov(\epsilon, X) = 0 \\ cov(\epsilon_a, \epsilon_b) = 0 \text{ if } a \neq b \end{array} \right.$

$$\beta = \frac{cov(X, Y)}{\sigma_x^2} \text{ where } \sigma_x^2 = var[X]$$

$$\alpha = E[Y] - \beta E[X] \quad \left\{ \begin{array}{l} \sigma_y^2 = \beta^2 \sigma_x^2 + \sigma_\epsilon^2 \\ \Rightarrow R^2 + (\sigma_\epsilon^2 / \sigma_y^2) = 1 \end{array} \right.$$

$R^2 = \frac{\sigma_x^2 \beta^2}{\sigma_y^2}$ $\left\{ \begin{array}{l} 0 < R^2 < 1 \\ R^2 = 1 \Leftrightarrow \text{Exact Linear Relationship between } X, Y \text{ (i.e. } \epsilon = 0) \\ R^2 = 0 \Leftrightarrow X, Y \text{ are un-correlated} \end{array} \right.$

High $R^2 \rightarrow$ Good
 Low $R^2 \rightarrow$ Bad
 If X, Y are statistically independent $\Rightarrow R^2 = 0$
 eg: R^2 close to 0 \Rightarrow NO linear Relationship