Module 3: Random Variables and Normal Distributions

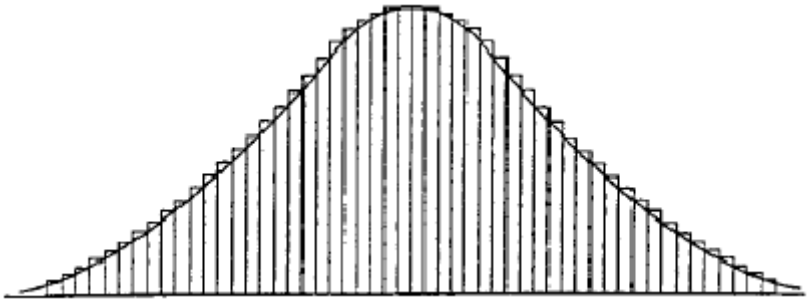
A **variable** is a measurement or characteristic on which individual observations are made. A **random variable** is a variable whose possible values are outcomes of a random phenomenon. A **domain** is a set of all possible values a variable can take.

Discrete random variables is a finite set or countably infinite sequence.

$p_x(x_i) = P(X = x_I)$ is called the **Probability Mass Function** (PMF).

- $0 <= p_x(x_i) <= 1$ as it is a probability,
- Sum of PMF for all values of X = 1.

Continuous random variable can lie on a numerical scale, such as all real numbers between (0, +infinity). If we mapped the data to a histogram, we would see the curve begin to smooth as the number of data points approaches infinity.



This density curve, $f_x(x)$, is called the **Probability Density Function** (PDF).

- $f_x(x) >= 0$ but can be greater than 1
- The integral of $f_x(x)$ over the domain of X is 1

The **Cumulative Distribution Function** (CDF) is defined as $F_x(X) = P(X <= x)$

- Non-decreasing
- The limit toward -infinity is 0, toward +infinity is 1
- For discrete random variables:

$$F_X(x) = \mathrm{P}(X \leq x) = \sum_{i:\, x_i \leq x} p_X(x_i)$$

- For continuous random variables:

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

$$\frac{dF_X(x)}{dx} = f_X(x)$$

The **Expected Value** of a random variable is an average of the possible values weighted by their probabilities. Also called mean and denoted by μ.

- For discrete random variables

$$E(X) = \sum_{i=1}^{K} x_i p_X(x_i)$$

- For continuous random variables

$$E(X) = \int_{-\infty}^{+\infty} x f_X(x)dx$$

A generalization for the expected values is the **r^{th} moment of a random variables**, $R(X^r)$.

- For discrete random variables:

$$E(X^r) = \sum_{i=1}^{K} x_i^r p_X(x_i)$$

- For continuous random variables:

$$E(X^r) = \int_{-\infty}^{+\infty} x^r f_X(x) dx$$

The first moment of a random variable is the expect value (mean). The r^{th} moment of a random variable about the mean, also called the r^{th} central moment, is defined as: $E[(X - \mu)^r]$

- The first central moment = 0
- The second central moment is the variance denoted as $\square\square^2$

The **variance** measures the spread around the mean of a random variable: $Var(X) = E[(X - \mu)^2]$

- Also equivelent to $Var(X) = E(X^2) - [E(X)]^2$
- The standard deviation is the square root of the variance

The **normal distribution:**

- is a continuous distribution
- can be expressed by a formula
- also called Gaussian distribution
- is a theoretical model for a population distribution that approximates the distribution of a number of measurement variables
- is appropriate for a number of measures, but not all. Only appropriate for some continuous measurements.
- is symmetric about the mean (i.e $P(X > \mu) = P(X < \mu) = .5$)
- is completely characterized mean and variance

The 68/95/99 rule:

- 68.25% of the data falls within 1 SD
- 95.45% of the data falls within 2 SD
- 99.74% of the data falls within 3SD

The standard normal random variable, referred to as Z, is in the scale of SD units from the mean.

Z has a $\mu=0$ and SD = 1 we can standardize any normal distribution with:

$$Z = \frac{X - \mu_X}{\sigma_X}$$

By converting to Z-scores we can easily compare the probability events in two different normal distribution.

The k^{th} percentile is defined as the score that holds the k percent of the scores below it. Ex. 90th percentile is the score that has 90% of the scores below it. We can compute percentiles with: X = μ + Z□□

The probability density function for normal curves:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \text{ for } -\infty < x < \infty$$

And in normal curves when mean is 0 and SD is 1 this can be simplified.

Relevant R Functions

qnorm(percentile, μ, □□) computes percentiles for normal variables

dnorm(x, μ, □□) will return the height of normal density function with a certain mean and SD at point x

pnorm(z, μ, □□) will return the cumulative distribution function of a normal distribution with certain mean and SD

Revision #4
Created 17 August 2022 14:03:18 by Elkip
Updated 17 August 2022 16:14:39 by Elkip