Lecture 3

Continuous Probability Distributions

Continuous Probability Distributions

- Continuous distribution has an infinite number of values between any two values assumed by the continuous variable
- As the number of observations, n, approaches infinity, the width of the class intervals approaches zero, the graph of the frequencies (frequency polygon) approaches a smooth curve
- As with other probability distributions, the total area under the curve equals 1
- Probability of occurrence of values between any two points on the x-axis is equal to the total area bounded by the curve, the x-axis, and perpendicular lines erected at the two points on the x-axis
- Probability of any specific value of the random variable is 0

Area under a smooth curve

- Integration of the density function over the range a to b
  - Density function is a formula used to represent the distribution of a continuous random variable
- A nonnegative function f(x) is called a probability distribution or probability density function of the continuous random variable X if the total area bounded by its curve and the x-axis is equal 1 and if the subarea under the curve bounded by the curve, the x-axis, and perpendiculars erected at any two points a and b gives the probability that X is between the points a and b

$$\int_x f(x)\,dx = 1$$

Normal distribution

- Most important distribution in statistics
- Also called the Gaussian distribution
- Density given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2 / 2\sigma^2}$$

  - for $-\infty < x < \infty$
  - where μ is the mean and σ the standard deviation

Characteristics of Normal Distribution

- Symmetrical about mean, μ
- Mean, median, and mode are equal
- Total area under the curve above the x-axis is one square unit
- 1 standard deviation on both sides of the mean includes approximately 68% of the total area
  - 2 standard deviations includes approximately 95%
  - 3 standard deviations includes approximately 99%
- Normal distribution is completely determined by the parameters μ and σ
  - Different values of μ shift the distribution along the x-axis
  - Different values of σ determine degree of flatness or peakedness of the graph

Standard Normal Distribution

- Standard normal distribution is one with a $\mu = 0$ and $\sigma = 1$
  - Standard normal density given by:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

  - for $-\infty < x < \infty$
  - where $z = (x - \mu) / \sigma$

Applications of Normal Distribution

- Frequently, data are normally distributed
  - Essential for some statistical procedures
    - Due to the combination of many random factors/ influences having different effects on the outcome
  - If not, possible to transform to a more normal form
- Approximations for other distributions
- Because of the frequent occurrence of the normal distribution in nature, much statistical theory has been developed for it

Normal Approximation to the Binomial Distribution

- When n is moderate in size ($n \geq 25$) and p is not too extreme, then the normal distribution is a good approximation to the binomial distribution with $\mu = np$ and $\sigma^2 = npq$
- $Pr(a \leq X \leq b)$ is approximately the area under the N(np, npq) curve between a-1/2 and b+1/2

  - $B(x,n,p) \rightarrow N(\mu=np, \sigma^2=npq)$

Normal Approximation to the Poisson Distribution

- When μ is moderate in size ($\lambda \geq 10$), the Poisson distribution is cumbersome to use and the normal distribution is a good approximation with $\mu = \lambda$ and $\sigma^2 = \lambda$
- $Pr(X = x)$ is approximately the area under the $N(\mu,\sigma^2)$ curve between x-1/2 and x+1/2
  - $P(\lambda) \rightarrow N(\mu=\lambda, \sigma^2=\lambda)$

Linear Combination of Random Variables

- Frequently, we work with combinations of variables; ie., variables 'added' of combined together in some way
- Definition:
  - $L = c_1X_1 + c_2X_2 + \ldots + c_nX_n$
- Expected value (typically, the mean) of a linear combination is the sum of the expected values of the variables
  - $E(L) = c_1 E(X_1) + c_2 E(X_2) + \ldots$
  - $E(L) = \Sigma_i c_i E(X_i)$
- Variance of a linear combination of independent random variables is
  - $Var(L) = \Sigma_i c_i^2 var(X_i)$

Linear Combination of Random Variables

- Variables that are not independent have a covariance between each pair
  - $Cov(X,Y) = E[(X - \mu_x)(Y - \mu_Y)]$
- Covariance hard to interpret directly, so we calculate the correlation coefficient for descriptive purposes
  - $\rho = Corr(X,Y) = Cov(X,Y) / (\sigma_x \sigma_Y)$
- Correlation coefficient ranges from -1 to 1 and is without units
  - Frequently used to describe strength of relationship between variables

Linear Combination of Random Variables

- Variance of a linear combination takes into account the covariance
  - $Var(c_1X_1 + c_2X_2) = c_1^2 var(X_1) + c_2^2 var(X_2) + 2c_1c_2Cov(X_1,X_2)$
  - Typically, we calculate the variance-covariance matrix for these calculations
- Note that the mean of a linear combination is not affected by the covariance 'structure' of the combination, only the variance
  - Important for regression analysis and analysis of correlated analysis