Introduction to Statistical Data Science

October 9, 2018

1 Lecture 1

1.1 Normal Distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-u)^2}{2\sigma^2})$$

$$F(x) \equiv P(X \leq x)$$

Central Limit Theorem (for normal distribution)

The more random variables we average over, the closer the resulting distribution will be to the Normal distribution

Parameter: u The mean is the location parameter.

Parameter: σ^2 The variance is the scale parameter

1.2 Uniform Distribution

$$X \sim U[0,1] \tag{1}$$

$$0 \leq x \leq 1 \tag{2}$$

$$p(x) = 1 \tag{3}$$

$$F(x) = P(X \leq X) = x \tag{4}$$

$$\tag{5}$$

Use uniform distribution to construct normal distribution

$$\mathbf{X} = \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \dots \\ X^{(n)} \end{bmatrix} \tag{6}$$

$$X^{(i)} \sim U[0,1]$$

$$Y = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}$$

subsampling $\mathbf{X}$ vector to construct Y. The distribution of $Y_j \sim ?, j = 1,2,3,\dots,p$ would close to normal distribution.

1.3 Poisson distribution

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$E(X) = V(X) = \lambda$$

1.4 empirical CDF

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x^{(i)} \leq x)$$

quantile is simply the inverse of the CDF: -The 0.9 quantile is the value of x such that $F(x) = 0.9$ i.e. $x = F^{-1}(0.9)$

2 Hypothesis

null hypothesis

In inferential statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured phenomena, or no association among groups.

p-value

The probability of obtaining results as or more extreme than that observed, assuming H_0 is true, is the p-value, under the assumption that the null hypothesis.

$$p \equiv P(X \leq 15; H_0)$$

$$p = \sum_{x=0}^{15} \binom{40}{x} 0.5^x (1-0.5)^{(40-x)}$$

The p-value is most certainly not the probability of H_0 being true

When used in practice with a threshold of 0.05 this is an informal method of reasoning and can be easily criticized

Statistical power

The power of a hypothesis test is the probability of avoiding a false negative

P-value distribution (CDF)

$$P(F(X) \leq z) = P(F^{-1}(F(x)) \leq F^{-1}(z)) = P(X \leq F^{-1}(z)) = F(F^{-1}(z)) = z$$

ł