BT1101 Introduction to Business Analytics

Hands-on Tutorial 04

Probability & Statistics Concepts

Discrete p.m.f Continuous p.d.f

Probability : ① Random Variable ② Distribution — mean, variance, parameters

↓

Statistics — Descriptive Stats. : ① descriptive statistics ② graphing

Mathematical Stats : ① parameter estimation ② hypothesis testing e.g. shapiro.test

* Random Variable X :

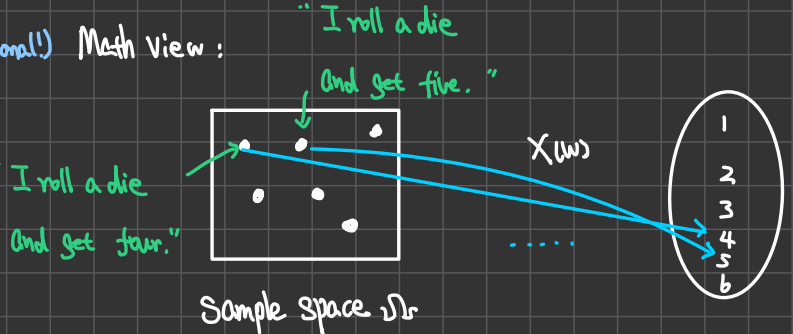
Intuition : X has some possible values. unknown before doing the experiment.

Each possible value has a chance (probability)

e.g. Rolling a die. Let X denote the number appears.

X has possible values : 1, 2, 3, 4, 5, 6

(optional!) Math view :



Random Variable is a function $X: \Omega \to R$ (or N)

Sample space : the set of all possible outcomes

Discrete Random variable & Continuous Random Variable

e.g Die, Coin, books I have··· e.g. height, Waiting time.

Probability Distribution: ① How to describe? ② How to calculate $P(X \in A)$?

Discrete R.V : Die Number

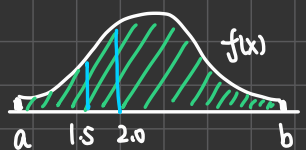
1	2	3	4	5	6
$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{2}{7}$

we define the probability

p.m.f : $P(X = w_i) = p_i$, $\sum_{\text{all } i} p_i = 1$ assign!

Continuous R.V : Height. Hard to assign the probability to each value.

p.d.f :



$\int_a^b f(x)\,dx = 1$. $f(x) \geq 0$

$P(X \in [1.5, 2.0]) = \int_{1.5}^{2.0} f(x)\,dx$

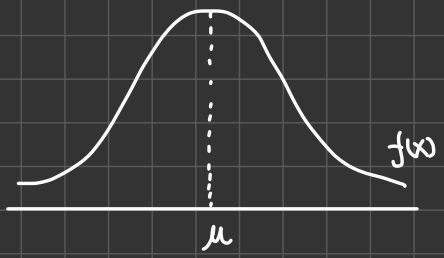
Rmk: Using p.m.f or p.d.f to describe the probability distribution of a random variable.

* important!

Special Distributions: { Binomial, Poisson, Geometric .... / Normal, Gamma ..... }

Fixed the "framework" of the pmf/pdf but leave some parameters flexible!

e.g. Normal Distribution $X \sim N(\mu, \sigma^2)$ $\quad f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$



Question: Assume the heights of all Singaporean Men are normally distributed.

Want to explore the distribution!

⇒ The nature of this question is to find μ and σ^2.

Mathematical Statistics:

Samples —inference→ population X

before observed: $(X_1, X_2, \cdots, X_n)$ ← n random variables

after observed: $(x_1, x_2, \cdots, x_n)$ ← n fixed constants

known: it has a normal distribution

unknown: parameters μ, σ^2

$N(\mu, \sigma^2)$

Height of a person in the world.

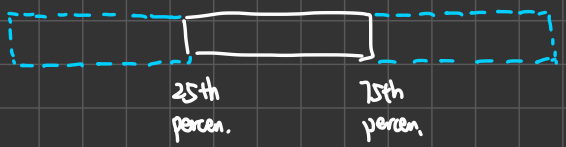
Descriptive Statistics:

NOTE: Sample Variance: $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ ($\bar{x}$: sample mean)

population variance $= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$

Outlier Analysis:

Method 1 IQR:



Method 2 Z-score:

Chebyshev's Thm: For any dataset, no matter what distribution, we can calculate the mean value and s.d. σ.

Then, the interval $[\text{mean} - z\cdot\sigma, \text{mean} + z\cdot\sigma]$ will include $(1-\frac{1}{z^2})\times 100\%$ data of this dataset.

e.g. we set $z=3$. Then $[\text{mean}-3\sigma, \text{mean}+3\sigma]$ will include around 89% data of this dataset.

(the rest 11% data we can regard as outliers)

Empirical Rule: For a dataset, normal distribution. Then:

$[\text{mean}-2\sigma, \text{mean}+2\sigma]$ will include around 95.44%

$[\text{mean}-3\sigma, \text{mean}+3\sigma]$ will include around 99.72%

Z-score of the samples of a random variable:

$$\boxed{\text{Z-score Amount}} = \frac{\text{Amount} - E[\text{Amount}]}{\sigma(\text{Amount})}$$

Sample mean ≈ 0

Sample std ≈ 1

Rmk: After transforming the sample data into z-score, we can easily identify who is in the interval

$[\text{mean} - 3\sigma, \text{mean} + 3\sigma]$

$= [0 - 3\times 1, 0 + 3\times 1] = [-3, 3]$