

Lecture 05. Probability Distributions

- A **distribution** is a mathematical function used to describe (model, modeling) a set of observed data.
- A **random variable** is a function defined on a set of events, and takes real value.

Example 1. For tossing a dice one time, let $A=\{\text{odd}\}$, and $B=\{\text{even}\}$. And if you get an odd at one tossing, you win 1 dollar; otherwise you loss 1 dollar. It is of course a fair game if the dice is 'fair'! That means, if we denote X as a random variable representing the money that you win or loss, then $X(A)=1$, and $X(B)=-1$. [Note that, in this case, $B=A^c$, which implies $A \cup B = \Omega$.] The function X thus defined is called a *discrete* random variable; it only takes values 1 and -1 . Moreover, The 'probability distribution' of X is: $P(X=1)=0.5$ and $P(X=-1)=0.5$, because we have just claimed that the dice is fair.

Example 2. Let $F(\cdot)$ denote a population with size 10^5 . It comprises people who are older than 50 years of age. If we denote w to be a sample point representing an sampled individual, and $X(w)$ is her/his measured systolic blood pressure. $X(\cdot)$ is suitably to be viewed as a continuous random variable, and the probability distribution of X is a continuous function, say, $f(x)$, which satisfies the following properties: (1) $f(x) \geq 0$, for $-\infty < x < \infty$. (2) If we define the 'support' S of $f(\cdot)$ as $S=\{x:f(x)>0, x \text{ is in } R\}$, then $\int_S f(x)dx=1$, and $F(x)=\int^x f(u)du$. As stated in a previous lecture, $f(x)$ is called the *probability density function* (pdf), and $F(x)$ the *cumulative distribution function* (cdf) with $dF(x)/dx=f(x)$.

The Bernoulli distribution Ber(p)

Formulae

Tossing a coin, $\Pr(\text{Head})=p$, $\Pr(\text{Tail})=1-p$

Random Variable X

DEF. :

If 'Head', then "+1"; if 'Tail', then "+0". We have

$$P(X=1)=p; P(X=0)=1-p; 0 < p < 1; \text{ or,}$$
$$P(X=x) = p^x(1-p)^{1-x}; x=0,1, 0 < p < 1. (**)$$

Alternatively, if 'Head', then "+1" (or "a" ; if 'Tail', then "-1" (or "b"); the distribution will be written as:

$$P(X=x)=p^{(1+x)/2}(1-p)^{(1-x)/2}; x=-1,1, 0 < p < 1.$$

Without loss of mathematical generality, we usually use formula (**), because it is easy to transform the problem into any "a" and "b".

In order to check that it is a probb. distribution:

$$\sum p^x(1-p)^{1-x}=(1-p)+p=1$$

亦即，所有的可能加起來必須等於 1

Example 3

1. 第一胎生男生的機率= p ；生女生的機率= $1-p$
2. 生第一胎的年齡 >35 歲之婦女，其終生得乳癌之機率= p_1 ，不會得乳癌之機率= $1-p_1$
3. 生第一胎的 <20 歲之婦女，其終生得乳癌之機率= p_2 ，不會得乳癌之機率= $1-p_2$
4. 男性肺腺癌患者接受化學治療一個療程後存活時間超過3年的機率= s ； <3 年的機率= $1-s$ 。

Comment:

In most of the cases, the probability of a concerned event differs if the definition of that event varies with different populations, time, space, or other characteristics. For example, the probability of becoming as an incidence case of breast cancer depends on the 'age' of a female individual, as well as on her lifestyle, family history, genetic susceptibility, etc. Further, a male lung-cancer patient's 3-year survival also depends on his personal characteristics; including past history of cigarette smoking or relevant exposure, occupation, genetic factors, cancer's stage, and many others.

So, the problem is complicated, which usually necessitate a complicated analysis tool and statistical 'model' to explore it.

Expectation and Variance (期望值與變異數) of a

Bernoulli random variable: $X \sim \text{Ber}(p)$

$$\begin{aligned} EX &= 1 \cdot P(X=1) + 0 \cdot P(X=0) \\ &= 1 \cdot p + 0 \cdot (1-p) = p \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= E[(X-EX)^2] = E(X^2 - 2 \cdot X \cdot EX + [EX]^2) \\ &= E(X^2) - 2 \cdot EX \cdot EX + (EX)^2 \\ &= E(X^2) - (EX)^2, \end{aligned}$$

$$\text{but } E(X^2) = 1^2 \cdot P(X=1) + 0^2 \cdot P(X=0) = p$$

$$\therefore \text{Var}(X) = p - p^2 = p(1-p)$$

The Binomial Distribution

Notation: Bin(n,p)

Question

If there are n independent Bernoulli trials

$$X_1, =1 (\text{probb.}=p); =0 (\text{probb.}=1-p)$$

$$X_2, =1 (\text{probb.}=p); =0 (\text{probb.}=1-p)$$

....

$$X_n, =1 (\text{probb.}=p); =0 (\text{probb.}=1-p)$$

$$+) \quad \Sigma X_i = \text{????}$$

ANS:

$$P(\Sigma X_i=0)=(1-p)(1-p)(1-p)\dots(1-p)$$

$$P(\Sigma X_i=1)= p(1-p)(1-p)\dots(1-p)+$$

$$(1-p)p(1-p)\dots(1-p)+$$

$$(1-p)(1-p)p(1-p)\dots(1-p)+$$

$$\dots\dots\dots+(1-p)(1-p)\dots(1-p)p$$

$$= np^1(1-p)^{n-1}$$

$$= nC_1 p^1(1-p)^{n-1}$$

..., it can be derived similarly for other values of ΣX_i

Formulae

$$P(\Sigma X_i=x) = nC_x p^x(1-p)^{n-x}, x=0,1,2,\dots,n; 0 < p < 1$$

In order to check that it is a probb. distribution:

$$\Sigma_x nC_x p^x(1-p)^{n-x} = (p+(1-p))^n, \text{ 二項式展開}$$

=1, 所有的可能加起來必須等於 1

Figures

(See Fig. 7.2~7.4 for various n and p configurations.)

Tables

(See Table A.1.)

Example 4 (homework !)

- 1 父親血型為 A(AO)型，母親血型為 B(BO)型，則 5 個兄弟姊妹中有 3 個 AB 型之機率
2. 一種新開發出來經過第二階段(phase II)人體試驗的藥，若其產生副作用的機率為 p (p=10%=0.1)，則當有 10 個人參加此試驗時，會有超過(≥)3 人以上產生副作用的機率=?
3. 同上題，當有 100 個人參加此試驗時，會有超過(≥)20 人以上產生副作用的機率=?

Remark. For the 3rd problem, it is of course difficult and even not possible to calculate the probability of a binomial variable, when n is really large. In this regard, an approximation (or called 'normal approximation') which borrow strength of the well-known central limit theorem (CLT) will be employed to solve the problem. See the next lecture for further details.

Expectation and Variance of a binomial variate

Let $Y = \sum X_i \sim \text{Bin}(n, p)$,

$$\begin{aligned} EY &= E(\sum X_i) = E(X_1) + E(X_2) + \dots + E(X_n) \\ &= p + p + \dots + p = np. \end{aligned}$$

(See what follows for a more dedicated derivation)

$$\begin{aligned} Y &\sim \text{Bin}(n, p) \\ \implies \sum_{y=0}^n C_y^n p^y (1-p)^{n-y} &= 1 \quad (*1) \\ \text{or } P(Y=0) + P(Y=1) + P(Y=2) + \dots + P(Y=n) &= 1 \quad (*2) \\ \implies E(Y) &= \sum_{y=0}^n y \cdot P(Y=y), \text{ by definition,} \\ &= \sum_{y=0}^n y \cdot \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y} \\ &= 0 + \sum_{y=1}^n y \cdot \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y} \\ &= \sum_{y=1}^n \frac{n!}{(y-1)!(n-y)!} p^y (1-p)^{n-y} \text{ because } \frac{y}{y!} = \frac{1}{(y-1)!} \\ &= \sum_{x=0}^{n-1} \frac{(n-1)! \cdot n}{x!(n-x-1)!} p^x (1-p)^{n-x-1} \cdot p, \text{ by setting } y-1 = x \\ &= np \sum_{x=0}^m \frac{m!}{x!(m-x)!} p^x (1-p)^{m-x}, \text{ by setting } n-1 = m \\ &= np \cdot 1, \text{ by } (*1). \end{aligned}$$

$$\begin{aligned} \text{Var}(Y) &= \text{Var}(\sum X_i) = \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n), \\ &\text{if } X_1, X_2, \dots, X_n \text{ are mutually independent} \\ &= p(1-p) + p(1-p) + \dots + p(1-p) = np(1-p). \end{aligned}$$

(More detailed derivation is omitted here, but will leave as a home work.)

The Poisson Distribution

Assumptions

(See Sec.7.3)

Example 5

1. 一個小時內辦公室電話通數
2. 一分鐘內高速公路上通過某點之車輛數
3. 一個月內台中市死亡車禍之數目
4. 一年內全台灣死亡於肺癌的人口數
- ...

Figure

(See Fig. 7.5 & Fig. 7.6)

Table (See Table A.2)

Formulae

$X \sim \text{Poisson}(\lambda) ; \text{Poi}(\lambda)$

$$P(X=x) = e^{-\lambda} \lambda^x / x!, \quad x=0,1,2,3,\dots; \lambda > 0$$

In order to check that it is a probb. distribution:

$$\sum_x e^{-\lambda} \lambda^x / x! = 1, \text{ 所有的可能加起來必須等於 } 1$$

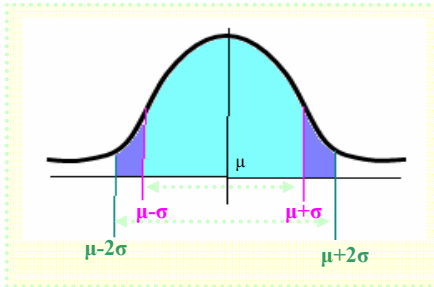
Expectation and Variance

$$E(X) = \lambda \quad \text{and} \quad \text{Var}(X) = \lambda$$

Normal (Gaussian) Distribution

常態分布；高斯分布

Figure



Table

Formulae

$$f(x) = [1/\sigma\sqrt{2\pi}] \exp[-(x-\mu)^2/2\sigma^2]$$

In order to check that $f(x)$ is a p.d.f. (probability density function), note that $\int f(x)dx=1$.

Expectation and Variance

$$E(X) = \int xf(x)dx = \mu$$

$$\text{Var}(X) = E[(X-EX)^2]$$

$$= E[(X-\mu)^2]$$

$$= E(X^2) - (EX)^2$$

$$= E(X^2) - \mu^2$$

$$= \int x^2 f(x) dx - \mu^2$$

$$= (\mu^2 + \sigma^2) - \mu^2$$

$$= \sigma^2$$

Characteristics

1. Bell-shaped (鐘形分布)
2. Symmetric about the mean μ (對稱性)
3. mean=median=mode
4. If $X \sim \text{Normal}(\mu, \sigma^2)$, or simply $N(\mu, \sigma^2)$,

$$P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$$

Standardization:

$$Z = (X - \mu) / \sigma$$

$$f(z) = [1/\sqrt{2\pi}] \exp[-z^2/2]$$

(See Figures 7.8~7.11)

$$P(-1 < Z < 1) \approx 0.68 \text{ (exactly, } = 0.6826)$$

$$P(-2 < Z < 2) \approx 0.95 \text{ (exactly, } = 0.9544)$$

Moreover,

$$P(-1.96 < Z < 1.96) = 0.95$$

$$P(-1.645 < Z < 1.645) = 0.90$$