Biostatistics

林 建 甫
C.F. Jeff Lin, MD. PhD.

台 北 大 學 統 計 系 助 理 教 授
台 北 榮 民 總 醫 院 生 物 統 計 顧 問
美 國 密 西 根 大 學 生 物 統 計 博 士

Random Variable and Probability Distribution Function

C.F. Jeff Lin, MD. PhD.
林建甫
台北大學統計系助理教授

Random Variables

- Sample space is often too large to deal with directly
- Recall that flipping a coin 100 times
- Record 1 for head and 0 for tail
- Sample space: 2^{100}
- If we don't need the detailed actual pattern of 0's and 1's, but only the number of 0's and 1's, we are able to reduce the sample space from size 2^n to size (100+1) as {0, 1, 2, …, 100}
- Abstractions lead to the notion of a **random variable**

Random Variables

- A **random variable** is a **function** that assigns a **real number** to each outcome in sample space of a random experiment
- **A function represented by a symbol X(·) or X**
- **Not an observed value of a variable**
- **Domain: sample space** of some experiment
- **Range:** a subset of the **real numbers**.

Random Variables

- a random variable X takes a series real number
  - X= 0, 1, 2, …, x, … ,100
  - X denotes the number of heads of tossing 100 coins
  - Y = 65, 49, 73, … , y, …,
  - Y denotes the body weight in kg
- Each occurrence of a random variable, X, has an associated probability
  - $P_X(X=0)$, $P_X(X=1)$, $P_X(X=2)$, …, $P_X(X=x)$, …
  - $f_Y(Y=65)$, $f_Y(Y=49)$, $f_Y(Y=73)$, …, f_Y

Random Variables

- A **capital letter** is typically used as an **abstract symbol** for a random variable as X, Y, Z …
  - X could represent the total number of "head"
  - Y could represent the body weight in kg
- After an experiment is conducted, the measured value (**actual numerical value**) of the random variable is denoted by a **lowercase** as x, y, z … and are called as the **realization** of the random variable or the **observed value**
  - x = 2
  - y = 65

Random Variables

In Symbols	In Words
X = x	an individual's body weight **equals** a specific value x
P(X=x)	the **probability** of an individual's body weight is a specific value x
X > x	an individual's body weight is **greater** than a specific value x
P (X > x)	the **probability** of an individual's body weight is **greater** than a specific value x
P (a < X < b)	the **probability** of an individual's body weight is **greater** than a specific value **a** and **less** than a specific value **b**

Random Variables

- Toss 3 fair coins, let X be number of Head appearing, then X is a random variable with possible values (0,1,2,3)

s	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
X(s)=X	3	2	2	2	1	1	1	0

- With probability

x	0	1	2	3
P(X=x)	1/8	3/8	3/8	1/8

Random Variables

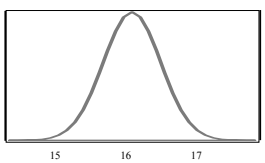
- A **discrete** random variable is a random variable with a finite (or countably infinite) range.
- A **continuous** random variable is a random variable with an interval (either finite or infinite) of real numbers for its range.

Random Variables

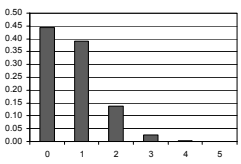
- Examples of **discrete** random variables:
  number of scratches on a surface, proportion of defective parts among 1000 tested, number of transmitted bits received in error.
- Examples of **continuous** random variables:
  electrical current, length, pressure, temperature, time, voltage, weight

Probability Distributions

- Since values of a random variable change from experiment to experiment, we have a distribution of possible outcomes.



Fill on 16 oz bottle of Pepsi



# of defects in a random sample of 5

Distribution of a Random Variable

- The **cumulative distribution function** or **cdf** of a random variable X, denoted by $F_X(x)$ is defined by
  $F_X(x) = P_X(X \leq x)$, for all x.
- We can treat "distribution" as "probability".
  - cdf: a **function**
  - cdf tells how the values of the r.v. are **distributed**
  - cdf is a cumulative distribution function since it gives the distribution of values in **cumulative form**

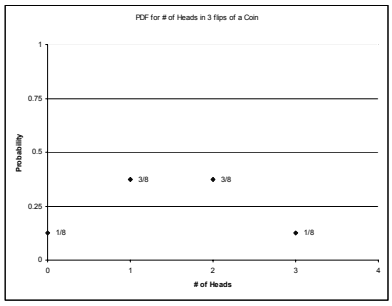
Cumulative Distribution Function

- Toss 3 fair coins, let X be number of Head appearing, then X is a random variable with possible values (0,1,2,3) with cdf

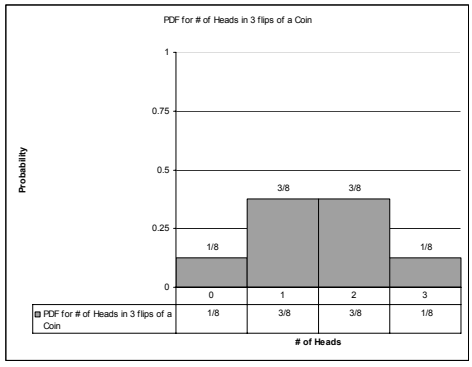
x	0	1	2	3
P (X=x)	1/8	3/8	3/8	1/8
x	$0 \le x < 1$	$1 \le x < 2$	$2 \le x < 3$	$3 \le x < \infty$
F ($X \le x$)	1/8	1/2	7/8	1

Probability Density Functions (pdf)

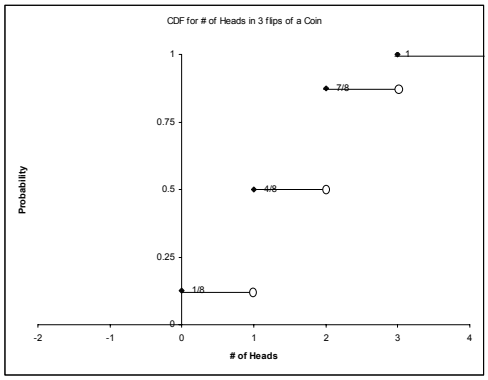
- We can graph pdf's usefully.
- For instance we can graph the pdf for flipping a coin three times using a "discrete density graph" or a histogram.
- We can also display them tabularly as in the table below the histogram.



Probability Density Functions (pdf)



Cumulative Distribution Functions



Discrete Random Variables and Probability Distributions

Probability Distribution of a Discrete Random Variable

The *probability distribution* or *probability mass function (pmf)* of a discrete rv is defined for every number x by $p(x) = P(\text{all } s \in S : X(s) = x)$.

pmf: Proposition

For any two numbers a and b with $a \le b$,

$$P(a \le X \le b) = F(b) - F(a-)$$

"$a-$" represents the largest possible X value that is strictly less than a.

Note: For integers

$$P(a \le X \le b) = F(b) - F(a-1)$$

Cumulative Distribution Function of a Discrete Random Variable

The cumulative distribution function (cdf) $F(x)$ of a discrete rv variable X with pmf $p(x)$ is defined for every number by

$$F(x) = P(X \le x) = \sum_{y:y \le x} p(y)$$

For any number x, $F(x)$ is the probability that the observed value of X will be at most x.

Probability Mass Function a Discrete Random Variable

- Random Variable, X, has possible variables, $\{x_1, x_2, x_3, \ldots x_n\}$
  - $P(X=x_i) = f(x_i)$
  - $f(x_i) \ge 0$
  - $\sum f(x_i) = 1$
- $f(x_i)$ is a probability mass function (**pmf**)
- For example:
  - $P(X=0\} = 0.04$
  - $P(X=1\} = 0.32$
  - $P(X=2\} = 0.64$

Example: Probability Distribution for the Random Variable *X*

A probability distribution for a random variable X:

x	−8	−3	−1	0	1	4	6
$P(X=x)$	0.13	0.15	0.17	0.20	0.15	0.11	0.09

Find

a. $P(X \le 0)$ 0.65

b. $P(-3 \le X \le 1)$ 0.67

Example

- Suppose we do an experiment that consists of tossing a coin until a head appears.
- Let p = probability of a head on any given toss
- Define a random variable X = number of tosses required to get a head. Then, for any X=1, 2, …
- P(X=0)=1-p
- P(X=1)=p
- P(X=2)=P(ap)=(1-p)×(p)
- P(X=3)=P(aap)=$(1-p)^2$×(p)
- P(X=x)=$(1-p)^{x-1}$× (p)
- Geometric Distribution with pmf of f(x)=$(1-p)^{x-1}$× (p)

Discrete Density Function

- Discrete Random Variable (Equivalence):
  - Probability mass function (**pmf**)
  - Discrete probability function
  - Discrete frequency function

  (consider integer valued random variable)

$$p_k = P(X = k)$$

- cdf: $F(x) = \sum_{k=0}^{\lfloor x \rfloor} p_k$
- pmf: $p_k = F(k) - F(k-1)$

Continuous Random Variables and Probability Distributions

Continuous Random Variables

- If there is a nonnegative function f(x) defined over the whole line such that

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f(x)dx$$

for any x_1, x_2 satisfying $x_1 \le x_2$, then X is a continuous random variable and *f(x)* is called its **density function**

Probability Density Function (pdf)

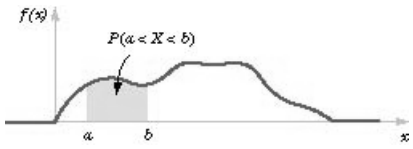
- For a continuous random variable X, a probability density function is a function such that :

(1) $f(x) \ge 0$

(2) $\int_{-\infty}^{\infty} f(x)dx = 1$

(3) $p(a \le X \le b) = \int_a^b f(x)\,dx$ = area under curve

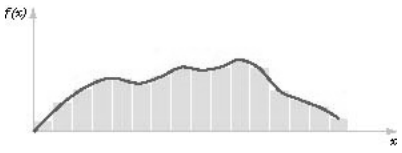
of f(x) from a to b for any a and b



Probability determined from the area under f(x)

Probability Density Function

- **f(x) is zero for x values** that cannot occur and it is assumed to be zero wherever it is not specifically defined.
- **Histogram:** An approximation to f(x). For each interval of the histogram, the area of the bar equals the relative frequency (proportion) of the measurements of the interval. This is an estimate of the probability that a measurement falls in the interval.
- The area under f(x) over any interval equals the true probability that a measurement falls in the interval.



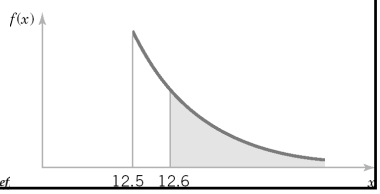
Histogram approximates a probability density function

Probability Density Function

- By appropriate choice of the shape of f(x), we can represent the probabilities associated with any continuous random variable X.
- **The shape of f(x)** determines how the probability that X assumes a value in [a,b] compares to the probability of any other interval of equal or different length.
- **Since p(X = x) = 0, to get p(X = x), we integrate f(x) over a small interval around X=x.**

Probability Density Function

- If X is a continuous random variable, for any x_1 and x_2,
- $P(x_1 \le X \le x_2)$
  $= p(x_1 < X \le x_2)$
  $= p(x_1 \le X < x_2)$
  $= p(x_1 < X < x_2)$.



Cumulative Distribution Functions

- The cumulative distribution function of a continuous random variable X is

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(u)du$$

For $-\infty < x < \infty$

- **F(x) is a continuous function (compared with F(x) for a discrete random variable that is not continuous).**
- A continuous random variable may be defined as one that has a continuous cumulative distribution function.

Cumulative Distribution Functions

- The cdf F of a continuous random variable has the same definition as that for a discrete random variable. That is,

$$F(x) = P(X \le x)$$

- In practice this means that F is essentially a particular antiderivative of the pdf since

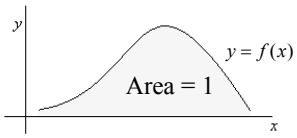
$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$$

- Thus at the points where f is continuous F'(x)=f(x).

Probability Density Function

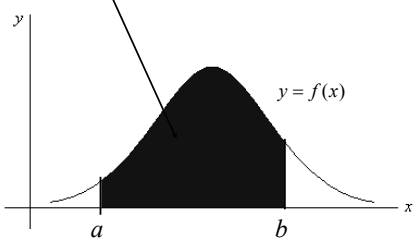
For $f(x)$ to be a pdf

1. $f(x) > 0$ for all values of x.
2. The area of the region between the graph of f and the x – axis is equal to 1.



Probability Density Function

$P(a \le X \le b)$ is given by the area of the shaded region.



Find Probability from PDF

PDF is used to find $P(a \le X \le b)$

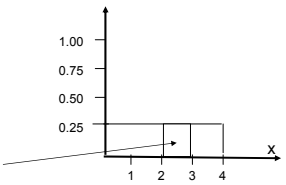
$$P(a \le x \le b) = \int_a^b f(x)dx$$

For example

$$f(x) = 0.25$$

$$P(2 \le x \le 3) = \int_2^3 0.25dx = 0.25x|_2^3 = 0.25$$

$$P(x = 3) = \int_3^3 0.25dx = 0.25x|_3^3 = 0$$



$$P(a \le x \le b) = P(a < x \le b) = P(a \le x < b) = P(a < x < b)$$

$$P(X = x) = 0$$

Cumulative Distribution Functions

- Knowing the cdf of a random variable greatly facilitates computation of probabilities involving that random variable since, by the Fundamental Theorem of Calculus,

$$P(a \le X \le b) = F(b) - F(a)$$

Cumulative Distribution Function (CDF)

- pdf *f(x)=0.25, for 0<x<4*
  =0, otherwise

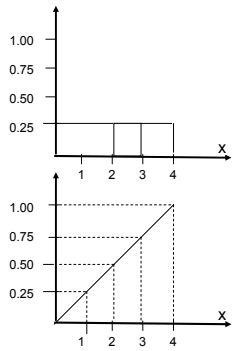
Cumulative density function, *F(x)*

$$F(x) = \int_{-\infty}^{x} f(u)du, \ -\infty < x < \infty$$

$$F(x) = \int_{-\infty}^{x} 0.25du = 0.25u\Big|_0^x = 0.25x, \quad 0 < x < 4$$

Uniform distribution:

$$f(x) = \frac{1}{b-a}; \ a < x < b$$



Revision

(1) f(x) ≥ 0

(2) $\int_{-\infty}^{\infty} f(x)dx = 1$

(3) p(a ≤ X ≤ b) = area under the curve

Expected Values and Variance of Discrete Random Variables

The Expected Value (Mean) of *X*

Let X be a discrete rv with set of possible values D and pmf $p(x)$. The ***expected value*** or ***mean value*** of X, denoted $E(X)$ or μ_X, is

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

$$\mu = E(X) = \sum_x xf(x)$$

Ex. Use the data below to find out the expected number of the number of credit cards that a student will possess.

x = # credit cards

x	$P(x=X)$
0	0.08
1	0.28
2	0.38
3	0.16
4	0.06
5	0.03
6	0.01

$$E(X) = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n$$
$$= 0(.08) + 1(.28) + 2(.38) + 3(.16) + 4(.06) + 5(.03) + 6(.01)$$
$$= 1.97$$

About 2 credit cards

Sampling Variation

28 51 54 26 38 41 41 37 31 50 33 42 34 26 34

Random sampling: 3 subjects per sample

Sample i	sa01	sa02	sa03	sa04	sa05
x_{i1}	50	26	41	50	34
x_{i2}	51	28	37	33	26
x_{i3}	54	31	31	32	34
Sample Mean	51.66	28.33	36.33	38.33	31.33
Ssample Variance	4.33	6.33	25.33	102.33	21.33

Population Mean = Expected Value of Population

- **Sample means vary** from sample to sample.
- Sample mean values vary because different samples are made up of different observations, called **sampling variation**.
- The **expected value E (X) = μ of a population** is not subject to sampling variation and depends entirely on the components of the probability distribution.

(Population) Variance of Discrete Random Variable

Let X have pmf $p(x)$, and expected value μ
Then the ***variance*** of X, denoted $V(X)$ (or σ_X^2 or σ^2), is

$$V(X) = \sum_D (x-\mu)^2 \cdot p(x) = E[(X-\mu)^2]$$

The ***standard deviation*** (SD) of X is

$$\sigma_X = \sqrt{\sigma_X^2}$$

Ex. The quiz scores for a particular student are given below:

22, 25, 20, 18, 12, 20, 24, 20, 20, 25, 24, 25, 18

Find the variance and standard deviation.

Value	12	18	20	22	24	25
Frequency	1	2	4	1	2	3
Probability	.08	.15	.31	.08	.15	.23

$\mu = 21$

$$V(X) = p_1(x_1-\mu)^2 + p_2(x_2-\mu)^2 + \ldots + p_n(x_n-\mu)^2$$

$$\sigma = \sqrt{V(X)}$$

$$V(X) = .08(12-21)^2 + .15(18-21)^2 + .31(20-21)^2$$
$$+.08(22-21)^2 + .15(24-21)^2 + .23(25-21)^2$$

$$V(X) = 13.25$$

$$\sigma = \sqrt{V(X)} \quad = \sqrt{13.25} \approx 3.64$$

Expected Values and Variance

of

Continuous Random Variable

(Population) Mean

= Expected Value

The *expected* or *mean value* of a continuous rv X with pdf $f(x)$ is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

Variance and Standard Deviation

The *variance* of continuous rv X with pdf $f(x)$ and mean μ is

$$\sigma_X^2 = V(x) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx$$
$$= E[(X-\mu)^2]$$

The *standard deviation* is $\sigma_X = \sqrt{V(x)}$.

Example: Continuous Distribution

- A continuous random variable X with probability density function

  f(x) = 1 / (b-a), $a \le x \le b$

  is a continuous uniform random variable.

$$\mu = E(X) = \frac{b+a}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12}$$

Rules of the Expected Value and Variance

Rules of the Expected Value

$$E(aX+b) = a \cdot E(X) + b$$

This leads to the following:

1. For any constant *a*, $E(aX) = a \cdot E(X)$.
2. For any constant *b*, $E(X+b) = E(X) + b$.

Rules of Variance

$$V(aX+b)=\sigma_{aX+b}^2=a^2\cdot\sigma_X^2$$

and $\sigma_{aX+b}=|a|\cdot\sigma_X$

This leads to the following:

1. $\sigma_{aX}^2=a^2\cdot\sigma_X^2,\ \sigma_{aX}=|a|\cdot\sigma_X$
2. $\sigma_{X+b}^2=\sigma_X^2$

The Expected Value of a Function

If the rv X has the set of possible values D and pmf $p(x)$, then the *expected value* of any function $h(x)$, denoted $E[h(X)]$ or $\mu_{h(X)}$, is

$$E[h(X)]=\sum_D h(x)\cdot p(x)$$

Expected Value of $h(X)$

If X is a continuous rv with pdf $f(x)$ and $h(x)$ is any function of X, then

$$E\left[h(x)\right]=\mu_{h(X)}=\int_{-\infty}^{\infty}h(x)\cdot f(x)dx$$

Thanks !