Probability and Statistics

SNU/NUKE/EHK

Properties of Probability

- To an event E_k, we assign a probability p_k, also denoted by $P(E_k)$, or "probability of the event E_k". The quantity p_k must satisfy the properties given in Fig. 1:

- $0 \leq p_k \leq 1$
- If E_k is certain to occur, $p_k = 1$.
- If E_k is certain not to occur, $p_k = 0$.
- If events E_i and E_j are mutually exclusive, then

$$P(E_i \text{ and } E_j) = 0; \quad \text{and} \quad P(E_i \text{ or } E_j) = p_i + p_j$$

- If events E_i, $(i=1, 2, \cdots, N)$ are mutually exclusive and exhaustive (one of the N events E_i is assured to occur), then

$$\sum_{i=1}^{N} p_i = 1$$

Fig. 1 properties of probabilities

Element of Probability Theory

- ✓ A random variable = a symbol that represents a quantity that can not be specified without the use of probability laws.
  - denoted by a capital letter X with the lower case x for a fixed value of the random variable.
  - a real value x_i that is assigned to an event E_i.
  - has a discrete distribution if X can take only a countable, finite or infinite, number of distinct values: [x_i, i=1, 2, . .].

$$\sum_i p(x_i) = 1.$$

Expectation Value: Discrete Random Variable

- An expectation value for the random variable X is defined by

$$E(x) = \bar{x} \equiv \sum_i p_i x_i \tag{1}$$

- Define a <u>unique</u>, real-valued function of x, g(x), which is also a random variable. The expected value of g(x) is

$$E[g(x)] = \bar{g}(x) = \sum_i p_i g(x_i) \tag{2}$$

- For a linear combination of random variables,

$$E[ag(x)+bh(x)] = aE[g(x)] + bE[h(x)] = a\bar{g}(x) + b\bar{h}(x) \tag{3}$$

Variance: Discrete Random Variable

- An variance for the random variable X is defined by

$$variance = var(x) = \sigma^2(x) \equiv \overline{(x-\overline{x})^2} = \sum_i p_i (x_i - \overline{x})^2 \qquad (4)$$

- The standard deviation of the random variable X is defined by

$$\sigma(x) \equiv [var(x)]^{1/2} \qquad (5)$$

- It is straightforward to show the following

$$\sigma^2(x) = \overline{x^2} - \overline{x}^2 \qquad (6)$$

$$\sigma^2(x) = \overline{x^2} - \overline{x}^2 \qquad (6)$$

$$\sigma^2(x) \equiv \overline{(x-\overline{x})^2} = \sum_i p_i \cdot (x_i - \overline{x})^2 = \sum_i p_i \cdot (x_i^2 - 2\overline{x}x_i + \overline{x}^2)$$

$$= \sum_i p_i \cdot (x_i^2 - 2\overline{x}x_i + \overline{x}^2) = \sum_i p_i \cdot x_i^2 - 2\overline{x}\sum_i p_i \cdot x_i + \overline{x}^2 \sum_i p_i$$

$$= \overline{x^2} - 2\overline{x}^2 + \overline{x}^2 = \overline{x^2} - \overline{x}^2$$

Linear/Non-linear Statistic

- The mean of linear combination of random variables equals the linear combination of the means because the mean is a linear statistic.

$$E\left[ag(x)+bh(x)\right]=aE\left[g(x)\right]+bE\left[h(x)\right]=a\overline{g}(x)+b\overline{h}(x) \qquad (3)$$

- The variance is not a linear statistic.

$$\sigma^2\left[ag(x)+bh(x)\right]=a^2\sigma^2\left[g(x)\right]+b^2\sigma^2\left[h(x)\right]+2ab\left[\overline{g(x)h(x)}-\overline{g}(x)\overline{h}(x)\right] \qquad (7)$$

$$E\left[ag(x)+bh(x)\right]=aE\left[g(x)\right]+bE\left[h(x)\right]=a\overline{g}(x)+b\overline{h}(x) \qquad (3)$$

$$E\left[ag(x)+bh(x)\right]=\frac{1}{N}\sum_{i=1}^{N}\left[ag(x_i)+bh(x_i)\right]$$

$$=\frac{1}{N}\left\{ag(x_1)+bh(x_1)+ag(x_2)+bh(x_2)+\ ...\ +ag(x_N)+bh(x_N)\right\}$$

$$=a\cdot\frac{1}{N}\left\{g(x_1)+g(x_2)+...+g(x_N)\right\}+b\cdot\frac{1}{N}\left\{h(x_1)+h(x_2)+...+h(x_N)\right\}$$

$$=a\cdot\overline{g}(x)+b\cdot\overline{h}(x)=aE\left[g(x)\right]+bE\left[h(x)\right]$$

$$\sigma^2\left[ag(x)+bh(x)\right] = a^2\sigma^2\left[g(x)\right] + b^2\sigma^2\left[h(x)\right] + 2ab\left[\overline{g(x)h(x)} - \bar{g}(x)\bar{h}(x)\right] \quad (7)$$

$$\sigma^2\left[ag(x)+bh(x)\right] = \overline{\left[ag(x)+bh(x)\right]^2} - \overline{\left[ag(x)+bh(x)\right]}^2$$

$$= \overline{a^2g^2(x)+2abg(x)h(x)+b^2h^2(x)} - \left[a\bar{g}(x)+a\bar{h}(x)\right]^2$$

$$= a^2\overline{g^2}(x)+2ab\overline{g(x)h(x)}+b^2\overline{h^2}(x) - \left[a^2\bar{g}^2(x)+2ab\bar{g}(x)\bar{h}(x)+b^2\bar{h}^2(x)\right]$$

$$= a^2\left[\overline{g^2}(x)-\bar{g}^2(x)\right] + b^2\left[\overline{h^2}(x)-\bar{h}^2(x)\right] + 2ab\left[\overline{g(x)h(x)}-\bar{g}(x)\bar{h}(x)\right]$$

Product of Random Variables

- The average value of the product of two random variables is

$$E(xy) = \sum_{i,j} p_{ij} x_i y_i \quad (8)$$

- If x and y are independent,

$$p_{ij} = f_i g_j \quad (9)$$

- Then $E(xy) = \sum_{i,j} p_{ij} x_i y_i = \sum_{i,j} f_i g_j x_i y_i = \sum_i f_i x_i \sum_j g_j y_j = E(x)E(y)$ (10)

- Eq. (7) is reduced to Eq. (11) if g(x) and h(x) are independent.

$$\sigma^2 \left[ag(x) + bh(x) \right] = a^2 \sigma^2 \left[g(x) \right] + b^2 \sigma^2 \left[h(x) \right] \quad (11)$$

Covariance and Correlation Coefficient

- The covariance is defined by

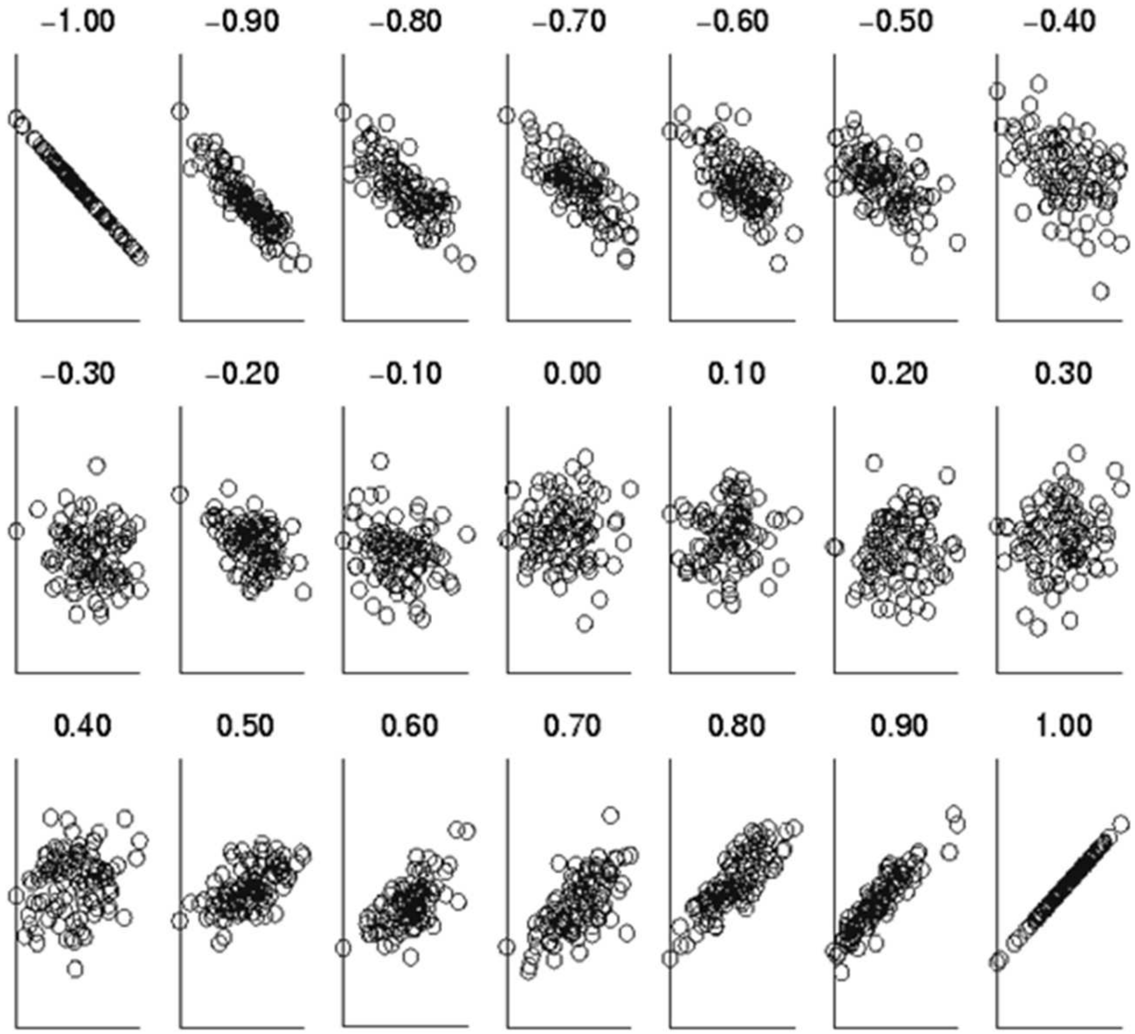
$$\text{covariance} = \text{cov}(x,y) = \overline{xy} - \bar{x}\cdot\bar{y} \qquad (12)$$

  - If x and y are independent, cov(x,y) = 0.
    - (x and y are independent: sufficient cond.; cov(x,y) = 0: necessary cond.)
  - It is possible to have cov(x,y) = 0 even if x and y are not independent.
  - The covariance can be negative.

- Correlation coefficient is a convenient measure of the degree to which two random variables are correlated (or anti-correlated).

$$\text{correlation coefficient} = \rho(x,y) \equiv \text{cov}(x,y) \Big/ \left[\sigma^2(x)\cdot\sigma^2(y)\right]^{1/2} \qquad (13)$$

where $-1 \le \rho(x,y) \le 1$.



$\rho = 1$ when $y = ax + b$
$\rho = -1$ when $y = -ax + b$
$(a>0)$

ρ = *correlation coefficient*

Continuous Random Variable

✓ A random variable <u>has a continuous distribution</u> if X can take any value between the limits x_1 and x_2 : $[x_1 \leq X \leq x_2]$.

$$Prob\ (x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)dx$$

where $p(x)$ is the probability density of $X = x$ and

$$\int_{-\infty}^{\infty} p(x)dx = 1.$$

- The probability of getting exactly a specific value is zero because there are an infinite number of possible values to choose from.

$$p_i \equiv p(x_i) \sim 0$$

Expected Value and Variance for Continuous pdf's

- For a continuous pdf, $f(x)$,

$$E(x) = \mu = \overline{x} \equiv \int_{-\infty}^{\infty} f(x')x'dx'$$

$$var(x) = \sigma^2(x) \equiv \int_{-\infty}^{\infty} f(x')(x' - \mu)^2 dx'$$

- For a real-valued function $g(x)$,

$$E[g(x)] = \overline{g}(x) \equiv \int_{-\infty}^{\infty} f(x')g(x')dx'$$

$$var[g(x)] = \sigma^2[g(x)] \equiv \int_{-\infty}^{\infty} f(x')[g(x') - \overline{g}(x)]^2 dx'$$

- Keep in mind that $\overline{x}$ and $\overline{g}(x)$ are the true means.

Discrete vs. Continuous pdf's

- For a discrete pdf of random variable x,

$$E_d(x) = \sum_{i=1}^{N} p_i x_i \quad \text{and} \quad var_d(x) = \sum_{i=1}^{N} p_i (x_i - \mu)^2 .$$

- Take the limit $N \to \infty$, then

$$\lim_{N\to\infty} E_d(x) = \lim_{N\to\infty} \sum_{i=1}^{N} p_i x_i = \lim_{N\to\infty} \sum_{i=1}^{N} x_i \frac{p_i}{\Delta x_i} \Delta x_i$$

$$= \lim_{N\to\infty} \sum_{i=1}^{N} x_i f_i \Delta x_i = \int_{-\infty}^{\infty} f_i(x')x' dx' \equiv E(x)$$

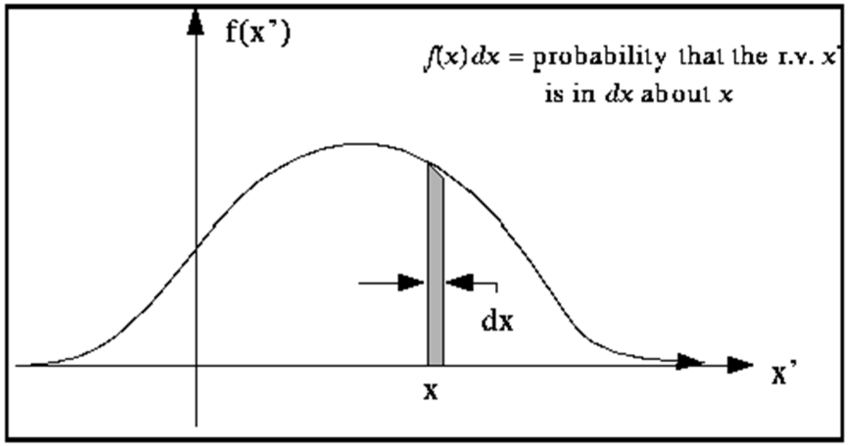
$$\lim_{N\to\infty} var_d(x) = \lim_{N\to\infty} \sum_{i=1}^{N} p_i (x_i - \mu)^2 = \lim_{N\to\infty} \sum_{i=1}^{N} (x_i - \mu)^2 \frac{p_i}{\Delta x_i} \Delta x_i$$

$$= \lim_{N\to\infty} \sum_{i=1}^{N} (x_i - \mu)^2 f_i \Delta x_i = \int_{-\infty}^{\infty} f_i(x')(x_i - \mu)^2 dx' \equiv var(x)$$

Probability Density Function

- Let $f(x)$ be a probability density function (pdf).
- $f(x)dx$ = the probability that the random variable X is in the interval $(x, x+dx)$:

$$f(x)dx = P(x \leq X \leq x+dx) .$$



- Since $f(x)dx$ is unitless, $f(x)$ has a unit of inverse random variable.

Probability Density Function (cont.)

- The probability of finding the random variable somewhere in the finite interval [a, b] is then

$$P(a \leq x \leq b) = \int_a^b f(x')dx' \quad ,$$

which is the area under the curve $f(x)$ from $x=a$ and $x=b$.

- Since $f(x)$ is a probability density, it must be positive for all values of the random variable x, or

$$f(x) \geq 0, \qquad -\infty < x < \infty$$

- Furthermore, the probability of finding the random variable somewhere on the real axis must be unity.

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

Cumulative Distribution Function

- The cumulative distribution function (cdf) F(x) gives the probability that the random variable is less than or equal to x:

$$cdf = F(x) \equiv P(x' \leq x) = \int_{-\infty}^{x} f(x')dx'$$

- Since F(x) is the <u>indefinite</u> integral of f(x),

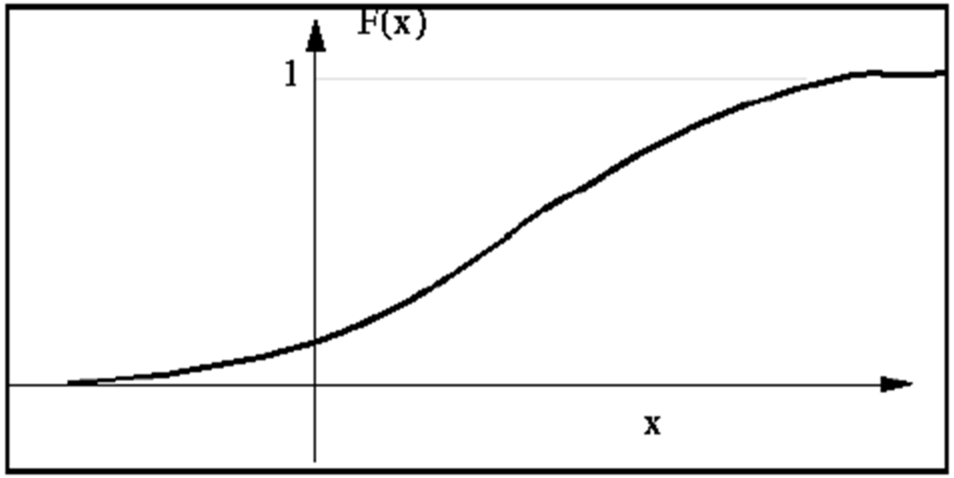
$$f(x) = F'(x).$$

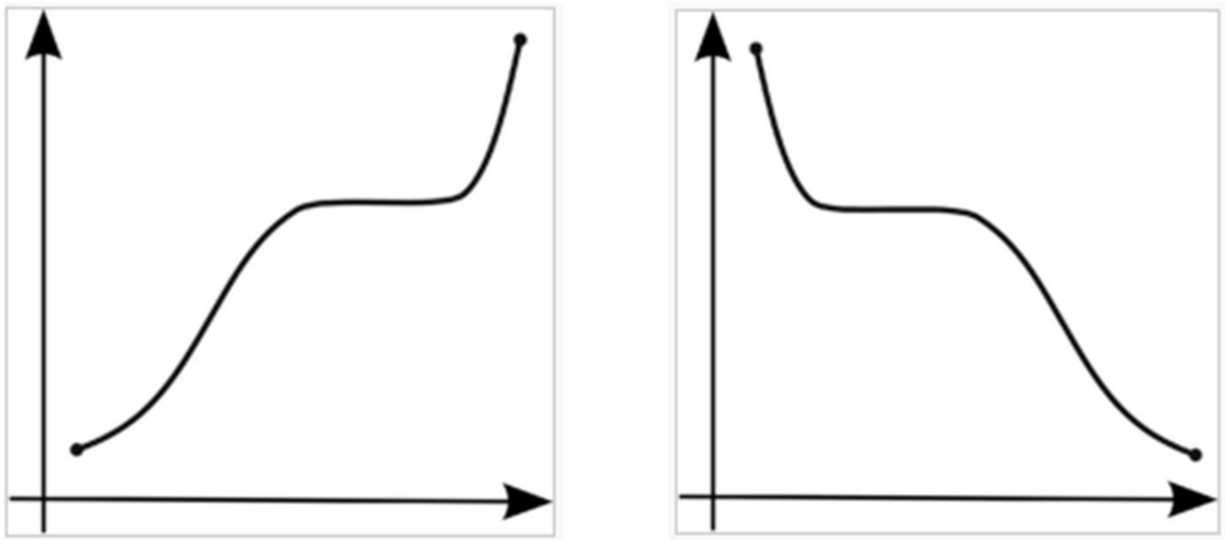
  – <u>definite</u> integral of f(x) in [a, b]

$$\int_a^b f(x)dx = F(b) - F(a) \quad vs. \quad \frac{dF(x)}{dx} = f(x) \;\leftrightarrow\; F(x) = \int f(x)dx + C$$

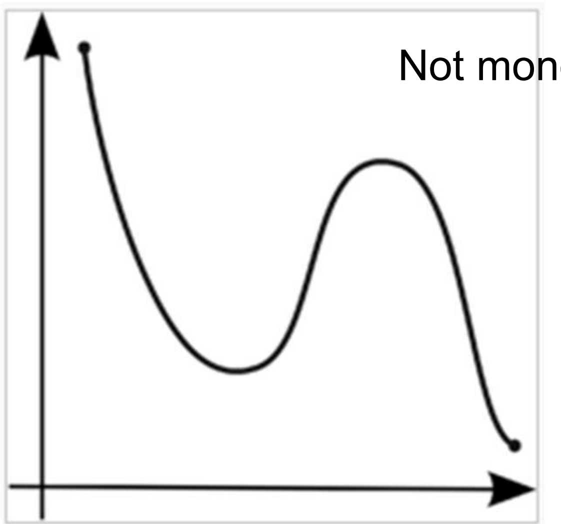
Cumulative Distribution Function (cont.)

- Since $f(x) \geq 0$ and the integral of $f(x)$ is unity, $F(x)$ obeys the following conditions:
  - $F(x)$ is monotonously increasing.
  - $F(-\infty) = 0$
  - $F(+\infty) = 1$





monotonically increasing (left) and decreasing (right)



Not monotonic

✓ *A <u>strictly</u> <u>monotonic</u> function $F(x)$:*
with $f(x) \neq 0$.