Probabilistic Model
概率模型

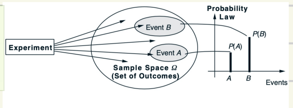


Figure 1.2: The main ingredients of a probabilistic model.

1. 是什么：对一个underlying process，i.e. $\boxed{\text{experiment}}$ 的不确定性的量化描述 (quantitive description)

   只考虑一个实验，3次抛硬币被视为一个 a single experiment，而不是三个实验

2. 包括2个元素：

   (1) Sample Space/样本空间 Ω

   Def. The set of all possible outcomes (w) of an experiment.

   (2) Probability Law/概率律

   - Def. 假设我们已经确定了与实验相关的样本空间.

     Intuitively，这个definition规定了任意结果/outcome或任意事件/Event (a subset of the sample space；a collection of outcomes) 的 "likelihood". More precisely，概率率为每一个事件A赋予一个 $\boxed{\text{数值}P(A)}$，i.e. the probability of A.

     → 满足的公理 := 概率公理

     $P(A)=\frac{2}{3}$ 可粗略理解为在大量重复实验中，事件A会在大约 $\frac{2}{3}$ 的实验中出现.

   - Discrete Probability Law

     If the sample space consists of a finite number of possible outcomes, then the probability law is specified by the probabilities of the events that consist of a single element. In particular, the probability of any event $\{s_1, s_2, \ldots, s_n\}$ is the sum of the probabilities of its elements:

     $$\mathbf{P}(\{s_1, s_2, \ldots, s_n\}) = \mathbf{P}(\{s_1\}) + \mathbf{P}(\{s_2\}) + \cdots + \mathbf{P}(\{s_n\}).$$

     概率律由 $\boxed{\text{单个元素组成的事件}}$ 的概率指定.

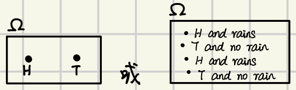
     Special Case: All single-element events have the same probability: $P(A) = \frac{\#\text{elements of } A}{n}$

3. 与 random variable 的联系.

   在许多概率模型中，outcomes 具有数值性质/numerical nature. 比如outcomes correspond to stock prices.
   在其它实验中，outcome并非数值，但它们可能与某些 numerical values of interest 相关联. 比如实验是 select students from a given population，we may wish to consider their GPA. 处理这些数值时，可以通过随机变量的概念来为它们分配概率.

- 符合以下3个特点的试验叫随机试验
  - 可重复：可在相同条件下重复进行
  - 多样性：每次试验的可能结果不止一个，且能事先明确试验的所有结果 has a well-defined set of possible outcomes infinitely repeated
  - 随机性：一次试验前不确定哪个结果会出现 has more than one possible outcome, and deterministic if it has only one

- 样本空间Ω须满足的条件：
  - mutually exclusive/互斥：与扔骰子相关的样本空间不能既包含 "1 or 3" 又包含 "1 or 4" 作为可能的outcome. 否则当掷出1时，实验 outcome 不唯一
  - collectively exhaustive/完全穷尽：无论实验发生什么情况，我们总能获得包含在样本空间内的 outcome.
  - at the right granularity：根据研究的目的确定适当的细节水平. 如对于 "Coin Flip" 随机试验可定义



- Probability Axioms / 柯尔莫果洛夫公理 / Kolmogorov Axioms

  Let $\mathbb{A} = \mathrm{Pot}(\Omega)$ (Event Space/die Menge aller Ereignisse über Ω/所有可能事件的集合).
  - Nonnegativity/非负性 $0 \le P(A) \le 1 \quad \forall A \in \mathbb{A}$
  - Additivity/可加性
    - If A and B are two disjoint events: $P(A\cup B) = P(A) + P(B)$
    - If Ω has infinite number of elements and $A_1, A_2, \cdots$ is the sequence of disjoint events: $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$
  - Normalization/归一性 $P(\Omega) = 1$

e.g. The experiment of rolling a pair of 4-sided dice.

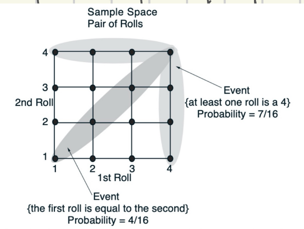


Figure 1.4: Various events in the experiment of rolling a pair of 4-sided dice, and their probabilities, calculated according to the discrete uniform law.

Example 1.5. Romeo and Juliet have a date at a given time, and each will arrive at the meeting place with a delay between 0 and 1 hour, with all pairs of delays being equally likely. The first to arrive will wait for 15 minutes and will leave if the other has not yet arrived. What is the probability that they will meet?

Let us use as sample space the square $\Omega = [0, 1] \times [0, 1]$, whose elements are the possible pairs of delays for the two of them. Our interpretation of "equally likely" pairs of delays is to let the probability of a subset of Ω be equal to its area. This probability law satisfies the three probability axioms. The event that Romeo and Juliet will meet is the shaded region in Fig. 1.5, and its probability is calculated to be 7/16.

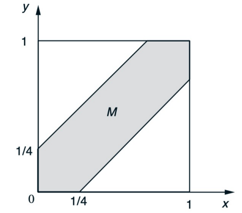


Figure 1.5: The event M that Romeo and Juliet will arrive within 15 minutes of each other (cf. Example 1.5) is

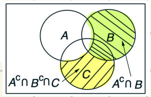
$$M = \left\{(x, y) \;\middle|\; |x - y| \le 1/4,\ 0 \le x \le 1,\ 0 \le y \le 1\right\},$$

and is shaded in the figure. The area of M is 1 minus the area of the two unshaded triangles, or $1 - (3/4)\cdot(3/4) = 7/16$. Thus, the probability of meeting is 7/16.

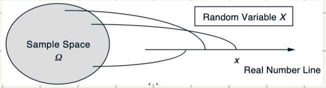
- Some properties of Probability Laws

Consider a probability law, and let A, B, and C be events.

(a) If $A \subset B$, then $\mathbf{P}(A) \le \mathbf{P}(B)$.
(b) $\mathbf{P}(A\cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A\cap B)$.
(c) $\mathbf{P}(A\cup B) \le \mathbf{P}(A) + \mathbf{P}(B)$.
(d) $\mathbf{P}(A\cup B\cup C) = \mathbf{P}(A) + \mathbf{P}(A^c\cap B) + \mathbf{P}(A^c\cap B^c\cap C)$.



Discrete Random Variables
离散随机变量



1. Def. $X: \Omega \to \mathbb{R}$

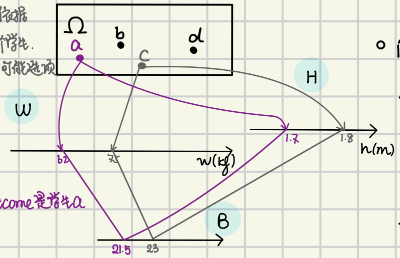
A random variable X is a real-valued function from the outcome of the experiment to the real numbers. X is called discrete if its range (the set of values that it can take) is finite or at most countably infinite.

- 定义理解:

Our probabilistic experiment is to pick a student at random $\boxed{\text{according to some probability law}}$. 意味着选择每个学生的概率可能不同，这些概率值是根据某种特定的概率分布来设置的

$\Omega = \{a, b, c, d\}$, and then record their weights and heights. 身高体重值本身不是样本空间的一部分，只是学生属性值的一个描述，是选择某个学生后所记录的一个附加信息

样本空间的定义是依据实验设计 — 随机选择一个学生. a, b, c, d 是实验的全部可能选项



假定这次实验的outcome是学生a

- 用W来表示"Weight"这个抽象概念，一旦我们确定哪位学生被选中，就能确定W具体的取值.

从这个意义上说，W is a function that assigns a numerical value (62) to each possible outcome (a) of the experiment.

  - 我们还可以根据已有的随机变量 W 和 H 创建新的随机变量 / A function of a random variable defines another random variable.

  例如身体质量指数 Body Mass Index $B = \frac{W}{H^2}$ 也是定义在样本空间上的一个函数 (一旦确定outcome, BMI也随之确定)

随机过程的结果是 Ω 中的元素 w，但是人们往往关注的不是结果本身 $w \in \Omega$，不是对其产生过程的精确描述 / die exakte Beschreibung des Zustandekommens，而是函数值 $X(w)$. 从这个意义上说，随机变量代表了对研究对象 / Untersuchungsgegenstand 的关注.

- e.g.

I toss a coin five times. This is a random experiment and the sample space can be written as

$$S = \{TTTTT, TTTTH, \ldots, HHHHH\}.$$

Note that here the sample space S has $2^5 = 32$ elements. Suppose that in this experiment, we are interested in the number of heads. We can define a random variable X whose value is the number of observed heads. The value of X will be one of $0, 1, 2, 3, 4$ or 5 depending on the outcome of the random experiment.

$X(TTTTT) = 0 \qquad X(THTHT) = 2$

2. Framework to describe discrete random variable.

(1) PMF / Probability Mass Function / 概率质量函数

- Def.

描述了离散随机变量 X 所有可能取值的概率.

PMF of X, denoted p_X is the "probability law" or "probability distribution" of a discrete random variable X

对于X每一个可能的数值 x, $p_X(x)$ 给出了事件 $\{X = x\}$ 的概率. 即 X 取到该值 x 的概率. $p_X(x) = P(\{X = x\}) = P(X = x)$

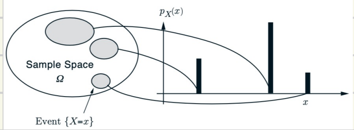
- PMF Properties

$\sum_x p_X(x) = 1$ 事件 $\{X = x\}$ 互不相交 (Ω 须满足的条件), x 遍历 X 所有可能的数值. 根据 Additivity 和 Normalization 公理可推导得出

- 计算

对于 X 的每个可能取值 x, collect 所有可以使事件 $\{X = x\}$ 发生的 possible outcomes;

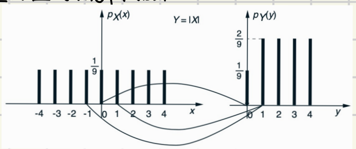
将这些结果的概率相加以获得 $p_X(x)$



- PMF for Functions of Random Variables $Y = g(X)$
  - Y 也是一个随机变量, since it provides a numerical value for each possible outcome.
  - $p_Y(y) = \sum_{\{x \mid g(x) = y\}} p_X(x)$

从已知的 X 的 PMF 推导出 Y 的 PMF: 为计算 Y 在某一特定值 y 处的概率 $p_Y(y)$, 需找到所有使得 $g(x) = y$ 的 x 值, 并将这些 x 值的概率相加.

e.g.



Let $Y = aX + b$ $\quad p_Y(y) = p_Y(y = ax + b) = p_X\left(\frac{y - b}{a}\right)$

(2) Mean and Variance

Expectation / Mean / Expected Value $\mathbb{E}[X]$

- Motivation

多次旋转一个fortune wheel，每次旋转出现代表金钱奖励的数字m_i $(i\in[n])$的概率是p_i.

What is the amount of money that you "expect" to get "per spin"? Average in large number of independent repetitions of the experiment.

— 假设共旋转K次，k_i是outcome为m_i的次数.

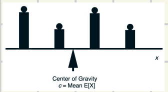
total amount received $= m_1k_1 + m_2k_2 + \cdots + m_nk_n$

total amount received per spin $= \dfrac{m_1k_1 + m_2k_2 + \cdots + m_nk_n}{n}$

$= m_1\cdot\frac{k_1}{n} + m_2\cdot\frac{k_2}{n} + \cdots + m_n\cdot\frac{k_n}{n}$

$= m_1\cdot p_1 + m_2\cdot p_2 + \cdots + m_n\cdot p_n$ $\quad (p_i := \text{relative frequency})$

- Interpretation



假设有一个bar，在每个点x处放置一个weight $p_X(x)$，且$p_X(x)>0$.

则在重心C处，左侧权重的力矩之和等于右侧权重的力矩之和 $\sum_x (x-c)\,p_X(x)=0$, or $c=\sum_x p_X(x)$

The center of gravity is the point where you would place your finger to balance the diagram so that it doesn't tip in either direction. $\mathbb{E}[X]$反映了X概率分布的中心位置.

- Properties

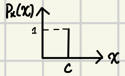
If $X\geq 0$, then $\mathbb{E}[X]\geq 0$ $\qquad X\geq 0$ i.e. $\forall w: X(w)\geq 0$

If $a\leq X\leq b$, then $a\leq \mathbb{E}[X]\leq b$ $\qquad a\leq X\leq b$ i.e. $\forall w: a\leq X(w)\leq b$

Sketch proof: $\mathbb{E}[X]=\sum_x x\cdot p_X(x) \geq \sum_x a\cdot p_X(x) = a\cdot\sum_x p_X(x) = a\cdot 1 = a$

If c is a constant, $\mathbb{E}(c)=c$

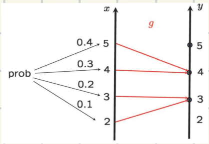
We can think of constant as being a random variable of a very special type.



In this case, there will be only one term in the formula.

$\mathbb{E}(c) = c\cdot \underbrace{P(c)}_{1} = c$

计算 $\mathbb{E}[X]=\sum_x x\cdot p_X(x)$ $\quad$ X可能取值的概率加权平均



- Expected Value Rule for Functions of Random Variables $\mathbb{E}[g(X)]=\sum_x g(x)\cdot p_X(x)$

方法I. Averaging over y $\qquad \mathbb{E}[Y]=\sum_y y\cdot p_Y(y)$

range over the values of Y one at a time. $\quad \mathbb{E}[Y] = 3\cdot(0.1+0.2) + 4\cdot(0.3+0.4)$

方法II. Averaging over x $\qquad \mathbb{E}[Y]=\mathbb{E}[g(X)]=\sum_x g(x)\cdot p_X(x)$

range over the values of X one at a time, and take into account their individual contributions.

$\mathbb{E}[Y] = \boxed{2\cdot 0.1} + 3\cdot 0.2 + 4\cdot 0.3 + 5\cdot 0.3$

10% of the time, X is going to be 2. and when that happens, Y takes on the value of 3.

Proof: $\sum_x g(x)\cdot p_X(x)$ $\quad$ fix a particular value of y, and add over all those x's that correspond to that particular y.

$= \sum_y \sum_{x:g(x)=y} g(x)\cdot p_X(x)$

$= \sum_y \sum_{x:g(x)=y} y\cdot p_X(x)$

$= \sum_y y \boxed{\sum_{x:g(x)=y} p_X(x)}$ $\quad$ 将所有映射到同一y值的x的概率相加 其实就等同于这个y的概率.

$= \sum_y y\cdot p_Y(y)$

$= \mathbb{E}[Y]$

应用: $\mathbb{E}[X^2] \overset{g(x)=x^2}{=\!=\!=} \sum_x x^2\cdot p_X(x)$

- Linearity of Expectation $\mathbb{E}[aX+b] = a\cdot\mathbb{E}[X]+b$

In general, $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ 当g是线性函数时，等式成立.

Variance $\mathrm{Var}(X)$

- The Variance provides a measure of dispersion of X around its mean

- Standard deviation of X $\quad \sigma_X = \sqrt{\mathrm{Var}(X)}$

- 计算

$\mathrm{var}(X) = \mathbb{E}[(X-\mu)^2] \underset{\mathbb{E}[g(X)]=\sum_x g(x)\cdot p_X(x)}{\overset{g(x)=(x-\mu)^2}{=\!=\!=}} \sum_x (x-\mu)^2\cdot p_X(x)$

$\mathrm{var}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2-2\mu X+\mu^2] = \mathbb{E}[X^2]-\mathbb{E}[2\mu X]+\mathbb{E}[\mu^2] = \mathbb{E}[X^2]-2\mu\mathbb{E}[X]+\mu^2 = \mathbb{E}[X^2]-(\mathbb{E}[X])^2$

- Linearity of Variance $\quad \mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$

Let $Y = X+b \qquad \mu=\mathbb{E}[X]$

$\mathbb{E}[Y] = \mathbb{E}[X+b] = \mu+b$

$\mathrm{var}(Y) = \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \mathbb{E}[(X+b-(\mu+b))^2] = \mathbb{E}[(X-\mu)^2] = \mathrm{var}(X)$

Let $Y = aX$

$\mathbb{E}[Y] = \mathbb{E}[aX] = a\mu$

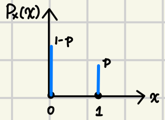
$\mathrm{var}(Y) = \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \mathbb{E}[(aX-a\mu)^2] = \mathbb{E}[a^2(X-\mu)^2] = a^2\cdot\mathbb{E}[(X-\mu)^2] = a^2\cdot\mathrm{Var}(X)$

伯努利分布 / 0-1分布 / Bernoulli Distribution

Parameter $p \in [0,1]$

- Def.

$$X = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$$



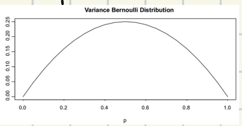
- 期望 $\mathbb{E}[X] = 1\cdot p + 0\cdot(1-p) = p$

  If X is the indicator of an event A, $X = I_A$

  $\mathbb{E}[I_A] = P(A)$

- 方差 $Var(X) = \sum_x (x-\mathbb{E}[X])^2 \cdot P_X(x)$

$$= (1-p)^2\cdot p + (0-p)^2\cdot(1-p)$$
$$= p - p^2$$
$$= p(1-p)$$



/ $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

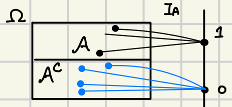
$$= \mathbb{E}[X] - (\mathbb{E}[X])^2 \quad (1^2 = 1,\ 0^2 = 0)$$
$$= p - p^2$$
$$= p(1-p)$$

Interpretation: 方差是对随机变量的uncertainty的度量，即对randomness的度量. 只有coin的随机性最大，即正反两面出现的概率都是$\frac{1}{2}$时$p = 1-p = \frac{1}{2}$，coin才公平.

Bernoulli random variable 的方差的确验证了这种直觉；

另一方面，在极端情况下，若$p=0$，即coin总是朝上(或朝下)，我们没有任何随机性. 相应地，方差为0.

- Application
  - Models a trial that results in 2 alternative outcomes: success / failure, Heads / Tail, etc.
  - Connect events and random variable, i.e., the indicator random variable I_A of an event A



$I_A = 1$ iff A occurs

$P_{I_A}(1) = P(I_A = 1) = P(A)$

$\Rightarrow$ I_A is a Bernoulli random variable, with parameter p equals to the probability of the event of interest.

I_A allows us to translate a manipulation of events to a manipulation of random variables.

离散型均匀分布 / discrete uniform distribution

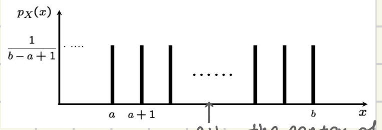
Parameters: a, b $\quad a \le b$

- Setting

  Experiment: Pick one of $a, a+1, \cdots, b$ at random, all equally likely

  Sample Space: $\{a, a+1, \cdots, b\}$ $\quad b-a+1$ possible values

  Random Variable: X: $X(\omega) = \omega$



$\frac{a+b}{2}$ the center of the symmetry

- 期望 $\mathbb{E}[X] = a\cdot\frac{1}{b-a+1} + \cdots + b\cdot\frac{1}{b-a+1}$

$$= \frac{(a+b)\cdot(b-a+1)}{2}\cdot\frac{1}{b-a+1} = \frac{a+b}{2}$$

- 方差 $var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

$$= \frac{1}{b-a+1}(a^2 + \cdots + b^2) - \left(\frac{b+a}{2}\right)^2$$
$$= \frac{1}{12}(b-a)(b-a+2)$$

- Application

  Model of: Complete ignorance, no reason to believe that one value is more likely than the other.

二项分布 / Binomial Distribution

Parameters: positive integer n, $p \in [0,1]$

- Setting

  Experiment: n independent tosses of a coin, $P(head) = p$

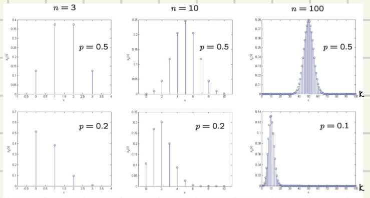
  Describe the # of successes in n independent Bernoulli trials.

  Sample Space: Set of sequences of H and T, of length n.

  Random Variable: X, # Heads observed

- Def.

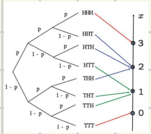
$$P_X(k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}, \quad k = 0, 1, \cdots, n$$



$p = \frac{1}{2}$

PMF is symmetric around $\frac{n}{2}$

e.g.



$P(X=2) = P(HHT) + P(HTH) + P(THH)$

$$= 3\cdot p^2\cdot(1-p)$$
$$= \binom{3}{2} p^2 \cdot (1-p)$$

- Calculation Preperation

  Let X be the indicator variable

$$X_i = \begin{cases} 1 & \text{if } i\text{-th trial is a success} \\ 0 & \text{otherwise} \end{cases} \qquad \mathbb{E}[X_i] = p$$

  # successes in n dependent trials $X = X_1 + \cdots + X_n$

  - 每次试验都是独立的且具有相同的成功概率p.

    因此每个X_i都有相同的分布，X_i is a Bernoulli random variable with parameter p

  - X是n次独立试验中成功的总次数，因此X是各独立随机变量X_i的和.

    X is a binomial random variable with parameters n and p

- 期望 $\mathbb{E}[X] = np$
- 方差 $var(X) = Var(X_1) + \cdots + Var(X_n)$

$$= n\cdot Var(X_1)$$
$$= n\cdot p(1-p)$$

几何分布 / Geometric Distribution

Parameter p: $0 < p \le 1$

- Setting

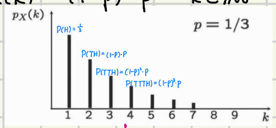
  Experiment: infinite many independent tosses of a coin $P(head) = p$. Describe the # of trials until a success.

  Sample Space: Set of infinite sequences of H and T.

  Random Variable: X, # tosses until the first Heads.

- Def

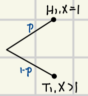
$$P_X(k) = (1-p)^k \cdot p \qquad k \in \mathbb{N}_0 \qquad p \in (0,1]$$



- 期望 $\mathbb{E}[X] = \frac{1}{p}$
  - 法I: Formula plug in

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k\, P_X(k) = \sum_{k=1}^{\infty} k\cdot(1-p)^{k-1}\cdot p$$

  - 法II: 随机变量X的期望值包括两部分: (期望的线性性质)



首先我们无论如何会有一次投掷，剩余X-1次投掷

$$\mathbb{E}[X] = \underbrace{P(X=1)}_{p}\cdot\underbrace{\mathbb{E}[X \mid X=1]}_{1^{\circ}\text{ try is a success } = 1} + \underbrace{P(X>1)}_{1-p}\cdot\underbrace{\mathbb{E}[X \mid X>1]}_{1+\mathbb{E}[X]}$$

(全期望值定理)

We've wasted one try, and we are back where we started.

$$\Rightarrow \mathbb{E}[X] = p + (1-p)\cdot(1+\mathbb{E}[X])$$
$$\Rightarrow \mathbb{E}[X] = \frac{1}{p}$$

Interpretation: 当P较小时，意味着观察到head的概率较小也就是说我们要等待较长的时间才能第一次看到正面. 随机变量X的值也会变得较大.

- 方差 $Var(X) = \frac{1-p}{p^2}$

  几何分布具有无记忆性，即使已经进行了K次失败的试验(X>K)，剩余的试验次数X-K仍服从与原几何分布相同的分布.

$$\mathbb{E}[X^2] = \underbrace{P(X=1)}_{p}\cdot\underbrace{\mathbb{E}[X^2 \mid X=1]}_{1} + \underbrace{P(X>1)}_{1-p}\cdot\mathbb{E}[X^2 \mid X>1]$$

$$\mathbb{E}[X^2 \mid X>1] = \mathbb{E}[(X+1)^2] = \mathbb{E}[X^2 + 2X + 1]$$

$$= \frac{2}{p^2} - \frac{1}{p}$$

$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1-p}{p^2}$$

(3) Joint PMF / 联合概率质量函数

- Motivation

  当我们需要同时处理多个与同一实验相关联的随机变量时，如随机变量 X 和 Y，单独的 p_X 和 p_Y 无法提供 X 和 Y 相互作用的信息.

- Def.

  The joint PMF of X and Y is defined by $p_{X,Y}(x,y) = P(X=x, Y=y)$

  PMF

  the function depends on 2 arguments, x and y

  Subscripts: the random variables we are dealing with

  表示随机变量 X 取值 x 的同时随机变量 Y 取值 y 的概率.

- Properties

  $\sum_x \sum_y p_{X,Y}(x,y) = 1$ 所有可能的 (x,y) pair 的概率总和为 1. 因为联合PMF覆盖了随机变量 X 和 Y 在其整个取值范围内的组合.

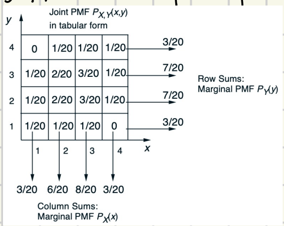
  $p_X(x) = P(X=x)$

  $= \sum_y P(X=x, Y=y)$

  $= \sum_y p_{X,Y}(x,y)$

  the marginal distribution of X

  事件 $\{X=x\}$ 的概率是 disjoint events $\{X=x, Y=y\}$ 在 y 遍历 Y 的所有值的并集

  ∘ The marginal probability is the probability of a single event occuring, independent of other events.

  $p_Y(y) = \sum_x p_{X,Y}(x,y)$

- PMF for Functions of Multiple Random Variables $Z = g(X,Y)$

  $p_Z(z) = P(Z=z) = P(g(x,y)=z) = \sum_{\{(x,y) \mid g(x,y)=z\}} p_{X,Y}(x,y)$

- Expected Value Rule for Functions of Multiple Random Variables $E[g(X,Y)] = \sum_x \sum_y g(x,y) \cdot p_{X,Y}(x,y)$

  Interpretation:

  函数 $g(X,Y)$ 的输出值依赖于特定的 (x,y) 值对. 每一对 (x,y) 发生的概率由它们的联合 PMF $p_{X,Y}(x,y)$ 给出.

  每一对 (x,y) 对函数期望值的贡献是 $g(x,y) \cdot p_{X,Y}(x,y)$

  最后累加所有的 (x,y) 对.

- $E[aX+bY+c] = aE[X] + bE[Y] + c$ is always true, no matter X, Y, Z are independent or not.

(4) Conditional PMF / 条件概率质量函数

- Motivation

  条件概率让我们能够"divide-and-conquer"，将复杂的probability models分解为更简单的submodels，然后逐个分析。
  比如给定an experiment, a corresponding sample space and a probability law. 假设我们知道某个特定事件B已经发生，那么我们可以只关注那些在事件B发生的前提下可能发生的事件，只关注样本空间中与事件B相交的部分，缩小了样本空间。

  $P(A|B)$ captures the partial information that event B provides about event A.

- Conditional Probability
  - Def. $P(A|B) = \frac{P(A\cap B)}{P(B)}$ $\quad P(B) > 0$
  - Conditional Probabilities form a legitimate Probability Law

Non-negativity	Additivity	Normalization
✓	For any two disjoint events A_1, A_2: $P(A_1 \cup A_2 \mid B) = \frac{P((A_1\cup A_2)\cap B)}{P(B)} = \frac{P((A_1\cap B)\cup(A_2\cap B))}{P(B)} = P(A_1|B) + P(A_2|B)$	$\boxed{P(\Omega\mid B)} = \frac{P(\Omega\cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$ 所有概率测度都集中在事件B的范围内，相当于将整个概率空间缩小到了B. i.e. 条件概率定义了在new universe B内的概率律

  - Calculation Rules

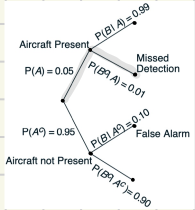
  I. Multiplication Rule $\quad P(\bigcap_{i=1}^{n} A_i) = P(A_1)\cdot P(A_2|A_1)\cdot P(A_3|A_1\cap A_2)\cdots P(A_n \mid \bigcap_{i=1}^{n-1} A_i)$
    - Setting: Experiments that have sequential character
    - 乘法规则反映了每个事件在前一个事件已经发生的基础上发生的依赖性。每一步的成功都依赖于前一步的成功。这种类型的问题通常用树状图表示。
    - e.g.

Example 1.9. Radar detection. If an aircraft is present in a certain area, a radar correctly registers its presence with probability 0.99. If it is not present, the radar falsely registers an aircraft presence with probability 0.10. We assume that an aircraft is present with probability 0.05. What is the probability of false alarm (a false indication of aircraft presence), and the probability of missed detection (nothing registers, even though an aircraft is present)?

A sequential representation of the sample space is appropriate here, as shown in Fig. 1.8. Let A and B be the events

$$A = \{\text{an aircraft is present}\},$$

$$B = \{\text{the radar registers an aircraft presence}\},$$

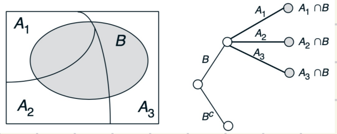


Leaf := event of interest

Branch := conditional probability

$P(\text{missed detection}) = P(A\cap B^c)$
$= P(A)\cdot P(B^c|A)$
$= 0.05 \cdot 0.01 = 0.0005$

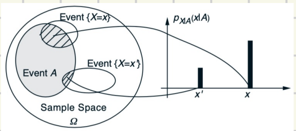
  II. Total Probability Theorem / 全概率定理 $\quad P(B) = P(A_1\cap B) + \cdots + P(A_n \cap B)$
  $= P(A_1)\cdot P(B|A_1) + \cdots + P(A_n)\cdot P(B|A_n)$
    - Setting: 事件的发生由多个互斥的途径决定
    - Let $A_1, \cdots, A_n$ be partition of Ω, i.e. each possible outcome is included in one and only one of the events $A_1, \cdots, A_n$



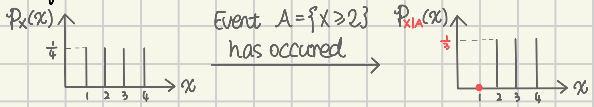
The probability that B occurs is a weighted average of its conditional probability under each scenario/event A_i, where each scenario is weighted according to its unconditional probability.

- Conditional PMF

Conditioning a random variable X on an Event A	Conditioning a random variable X on another
$p_{X\mid A}(x) = P(X=x\mid A) = \frac{P(\{X=x\}\cap A)}{P(A)}$	$p_{X\mid Y}(x\mid y) = P(X=x \mid Y=y) = \frac{P(X=x,\ Y=y)}{P(Y=y)} = \frac{p_{X,Y}(x,y)}{p_Y(y)}$

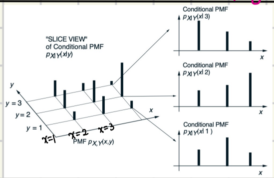


- Interpretation: 不同的x值对应的事件$\{X=x\}\cap A$之间互斥，且这些事件的并集是A，即$P(A) = \sum_x P(\{X=x\}\cap A)$
  结合上面公式可得 $\sum_x p_{X|A}(x) = 1$
  e.g.



$E[X] = \frac{a+b}{2} = 2.5$ $\qquad$ $E[X|A] = \frac{a+b}{2} = 3$

$Var(X) = \frac{(b-a)(b-a+2)}{12} = \frac{5}{4}$ $\qquad$ $var(X|A) = \frac{1}{3}(4-3)^2 + \frac{1}{3}(3-3)^2 + \frac{1}{3}(2-3)^2 = \frac{2}{3}$



- Interpretation: 当固定y并考察$p_{X|Y}(x|y)$时，实际上是在考察一个以x为变量的函数，i.e., 给定$Y=y$的条件下，X取每一个可能的值的概率。因此，我们实际上有一个由所有可能的Y值定义的PMF family，每一个不同的y值都会定义一个不同的$p_{X|Y}(x|y)$且$\sum_x p_{X|Y}(x|y) = 1$，无论Y取何值，X的条件分布都是合法的概率分布。
- e.g.

y \ x	1	2	3	4
4	1/20	2/20	2/20	
3	2/20	4/20	1/20	2/20
2		1/20	3/20	1/20
1		1/20		

P(Y=2)

$\boxed{P(Y=2) = \frac{5}{20}}$

$p_{X|Y}(1|2) = \frac{0}{P(Y=2)}$ $\qquad$ $p_{X|Y}(2|2) = \frac{\frac{1}{20}}{P(Y=2)} = \frac{1}{5}$

$p_{X|Y}(3|2) = \frac{3}{5}$ $\qquad$ $p_{X|Y}(4|2) = \frac{1}{5}$

X, Y are independent? —NO. $p_X(1) = \frac{3}{20} \neq p_{X|Y}(1|2) = 0$

What if we condition on $\boxed{X\le 2 \text{ and } Y \ge 3}$? — Yes.

y \ x	1/3	2/3
1/3	1/9	2/9
2/3	2/9	4/9

- Calculation Rules

  I. Multiplication Rule $\quad p_{X,Y,Z}(x,y,z) = p_X(x)\cdot p_{Y|X}(y|x)\cdot p_{Z|X,Y}(z|x,y)$

  II. Conditional Expectation / 条件期望
    - 条件期望与普通期望的概念在本质上是相同的，计算方式都是加权随机变量的所有可能值。
      但条件期望不是在原有的概率框架内进行计算，而是在一个通过条件改变了概率分布的new universe中进行。

      $E[X|A] = \sum_x x\cdot p_{X|A}(x|A)$ $\qquad$ $E[X] = \sum_x x\cdot p_X(x)$ $\qquad$ $E[g(X)] = \sum_x g(x)\cdot p_X(x)$

      $E[X|Y=y] = \sum_x x\cdot p_{X|Y}(x|Y=y)$

      $E[g(X)|Y=y] = \sum_x g(x)\cdot p_{X|Y}(x|Y=y)$

      $E[g(X)|A] = \sum_x g(x)\cdot p_{X|A}(x|A)$

  III. Total Probability Theorem / 全概率定理 $\quad$ Let $B = \{X=x\}$ $\quad p_X(x) = P(A_1)\cdot p_{X|A_1}(x) + \cdots + P(A_n)\cdot p_{X|A_n}(x)$

  IV. Total Expectation Theorem / 全期望值定理

  The total probability theorem is true for all x's: $\underbrace{\sum_x x\cdot p_X(x)}_{E[X]} = P(A_1)\cdot \underbrace{\sum_x x\cdot p_{X|A_1}(x)}_{E[X|A_1]} + \cdots + P(A_n)\cdot \underbrace{\sum_x x\cdot p_{X|A_n}(x)}_{E[X|A_n]}$

  $E[X] = \sum_{i=1}^{n} P(A_i)\cdot E[X|A_i]$ $\qquad$ Let $A = \{Y=y\}$

  $= \sum_y p_Y(y)\cdot E[X|Y=y]$ $\qquad$ 全期望值定理将一个r.v. X的期望分解为

  $= E[E[X|Y]]$ $\qquad$ X在给定另一个r.v. Y的不同条件下的期望的加权平均。

(5) Independence

① Independence of two Events A, B $\quad P(A\cap B) = P(A)\cdot P(B)$

- $P(A|B)$ captures the partial information that event B provides about event A.
  When the occurence of B provides no information and does not alter the probability that A has occured, i.e., $P(A|B) = P(A)$, we say that A is independent of B.
- By definition $P(A|B) = \frac{P(A\cap B)}{P(B)}$, in case of independent A and B, $P(A\cap B) = P(A)\cdot P(B)$ 我们采用此公式作为独立性的定义，因为即使 $P(B)=0$，公式依然成立，而 $P(A|B)$ is undefined.
- independent ≠ disjoint

  独立事件：一个事件 A 的发生不影响另一个事件 B 的发生概率．独立事件可以联合发生，其联合概率是各自概率的乘积．
  Independence is not easily visualized in terms of the sample space.

  不相交事件：两个事件不能同时发生．$P(A\cap B) = 0$.
  如果两个不相交的事件 A, B 都有正概率，即 $P(A)>0$, $P(B)>0$，那么它们绝对不可能是独立的．

② Conditional Independence of two Events A, B given an event C

- $P(A\cap B|C) = P(A|C)\cdot P(B|C)$ $\quad (*)$

  条件独立性表明，在已知事件 C 发生的情况下，事件 A 和 B 的发生相互独立．

  $P(A|C) = P(A|B\cap C)$

  推导：
  $$P(A\cap B|C) = \frac{P(A\cap B\cap C)}{P(C)} \qquad \text{(Def. Conditional Probability)}$$
  $$= \frac{P(C)\cdot P(B|C)\cdot P(A|B\cap C)}{P(C)} \qquad \text{(Multiplication Rule)}$$
  $$= P(B|C)\cdot P(A|B\cap C)$$
  $$P(A|C)\cdot P(B|C) = P(B|C)\cdot P(A|B\cap C) \qquad (*)$$
  $$P(A|C) = P(A|B\cap C)$$

  在事件 C 发生的前提下，事件 B 的发生对事件 A 的发生概率没有额外的影响．
  从信息论的角度来看，All the informations that B provides regarding the occurence of A are completely encapsulated in C. 某些看似重要的条件实际上不提供额外信息．

- unconditionally independent ⇏ conditionally independent

  2 independent fair coin tosses
  $H_1 = \{$1st toss is a head$\}$
  $H_2 = \{$2nd toss is a head$\}$
  $D = \{$the two tosses have different results$\}$

  ◦ H_1 and H_2 are unconditionally independent
  ◦ But NOT conditionally independent:
  $P(H_1\cap H_2|D) = 0$
  $P(H_1|D) = P(H_2|D) = \frac{1}{2}$
  $P(H_1\cap H_2|D) \neq P(H_1|D)\cdot P(H_2|D)$
  若已知抛出了一正一反，那么 H_1 发生时必有 H_2 不发生

- conditionally independent ⇏ unconditionally independent

  2 Coins ● $P(\text{Head}) = 0.99$ ● $P(\text{Head}) = 0.01$
  Choose one of the two at random
  Proceed with 2 independent coin tosses
  $H_i = \{$i-th toss is a head$\}$
  $B = \{$the blue coin was selected$\}$, $P(B) = \frac{1}{2}$

  ◦ H_1 and H_2 are conditionally independent:
  选定硬币后，两次抛掷都是独立的：
  $P(H_1\cap H_2|B) = P(H_1|B)\cdot P(H_2|B) = 0.99\cdot 0.99$
  ◦ But NOT unconditionally independent:
  若没有给定硬币信息，只知道 H_1 发生，我们自然会倾向于认为选择的是蓝色硬币．这将直接提高对 H_2 发生的预期．
  $P(H_1) = P(B)\cdot P(H_1|B) + P(B^c)\cdot P(H_1|B^c) = \frac{1}{2}\cdot 0.99 + \frac{1}{2}\cdot 0.01 = \frac{1}{2}$ 同样 $P(H_2) = \frac{1}{2}$
  $P(H_1\cap H_2) = P(B)\cdot P(H_1\cap H_2|B) + P(B^c)\cdot P(H_1\cap H_2|B^c) = \frac{1}{2}\cdot 0.99^2 + \frac{1}{2}\cdot 0.01^2 = 0.4901$

③ Independence of a Random Variable from an Event $\quad P(X=x \text{ and } A) = P(X=x)\cdot P(A) = p_X(x)\cdot P(A)$, for all x

- Interpretation I
  $p_{X|A}(x) = p_X(x)$, for all x $\qquad$ 无论 A 是否发生，X 的概率分布保持不变
  Using the definition of the conditional PMF $p_{X|A}(x) = \frac{P(\{X=x\}\cap A)}{P(A)}$
- Interpretation II.
  $P(A) = P(A|X=x)$, for all x $\qquad$ 对事件 A 的概率判断不受随机变量 X 取值的影响．

④ Independence of Random Variables $\quad p_{X,Y}(x,y) = p_X(x)\cdot p_Y(y)$, for all x, y

- Interpretation: $p_{X|Y}(x|y) = p_X(x)$, for all y with $p_Y(y)>0$, and all x
- If X and Y are independent variables, then:

  $\mathbb{E}[XY] = \mathbb{E}[X]\cdot\mathbb{E}[Y]$

  Proof: Expected Value Rule for Functions of Multiple Random Variables
  $\mathbb{E}[g(X,Y)] = \sum_x\sum_y g(x,y)\cdot p_{X,Y}(x,y)$ $\quad$ Let $g(X,Y) = XY$
  $\mathbb{E}[XY] = \sum_x\sum_y x\cdot y\cdot p_X(x)\cdot p_Y(y)$ $\quad$ By independence
  $= \underbrace{\left(\sum_x x\cdot p_X(x)\right)}_{\mathbb{E}[X]}\underbrace{\left(\sum_y y\cdot p_Y(y)\right)}_{\mathbb{E}[Y]}$

  $\mathbb{E}[g(X)\cdot h(Y)] = \mathbb{E}[g(X)]\cdot\mathbb{E}[h(Y)]$

  $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

  Proof: Let $\mathbb{E}[X] := a$ $\quad \mathbb{E}[Y] := b$ $\quad \mathbb{E}[X+Y] = a+b$
  $$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[((X+Y)-(a+b))^2] \\ &= \mathbb{E}[((X-a)+(Y-b))^2] \\ &= \mathbb{E}[(X-a)^2 + 2\cdot(X-a)\cdot(Y-b) + (Y-b)^2] \\ &= \mathbb{E}[(X-a)^2] + 2\cdot\mathbb{E}[(X-a)\cdot(Y-b)] + \mathbb{E}[(Y-b)^2] \\ &= \mathrm{Var}(X) + 2\cdot\mathbb{E}[X-a]\cdot\mathbb{E}[Y-b] + \mathrm{Var}(Y) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y)\end{aligned}$$

  Covariance: $\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = 0$
  Proof: $= \mathbb{E}[X-\mathbb{E}[X]]\cdot\mathbb{E}[Y-\mathbb{E}[Y]] = 0$

- In general: $\mathrm{Var}(X+Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$
  e.g. $X = Y$ $\qquad \mathrm{Var}(X+Y) = \mathrm{Var}(2X) = 4\,\mathrm{Var}(X)$
  $\quad\;\; X = -Y$ $\qquad \mathrm{Var}(X+Y) = \mathrm{Var}(0) = 0$

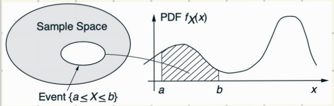
  In general $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

⑤ Conditional Independence of two Random Variables X, Y given an event A

$p_{X,Y|A}(x,y) = p_{X|A}(x)\cdot p_{Y|A}(y)$, for all x, y

Again, equivalent to $\underbrace{p_{X|Y,A}(x|y)}_{\{Y=y\}\text{和事件}A\text{同时发生的条件下，}X\text{取值为}x\text{的概率}} = p_{X|A}(x)$ for all x and y s.t. $p_{Y|A}(y) > 0$

Continuous Random Variables
连续随机变量



1. Def.

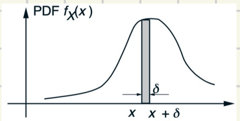
A random variable X is continuous if its possible values compromise either a single interval or a union of disjoint intervals on the number line.

A random variable X is continuous if its probability law can be described by a PDF.

2. Framework to describe continuous random variable

(1) PDF / Probability Density Function / 概率密度函数

- Intuition



We can think of the bars in the PMF as point masses with positive weight, each sitting on top of a specific numerical value. To calculate the probability that the random variable falls within a specific interval, we add up all the masses that sit on top of the numerical values within that interval.

In the case of PDF, there's still a total of one unit of probability mass assigned to the possible values of the random variable. However, this mass is spread across the entire real line, rather than being concentrated at specific points. The distribution of this mass is not uniform: some parts of the real line contain more mass per unit length than others. 我们可以想象有1磅雪落在了实数线上. PDF告诉我们在特定点上积累的雪的高度, 然后我们通过计算曲线下面的面积来得到在某个区间上积累的整体雪量的重量.

- PDF Properties

① $P(a \le X \le b) = \int_a^b f_X(x)\,dx \quad [a,b] \subseteq \Omega$ PDF用于确定小区间的概率

② $f_X(x) \ge 0$, for every x

③ $\int_{-\infty}^{+\infty} f_X(x) = P(-\infty < X < +\infty) = 1$ The entire area under the graph of the PDF must be equal to 1.

④ $P(X=a) = 0$ 在连续的范围内选取一个点的概率无穷小.

- For an interval $[x, x+\Delta]$ with very small length δ, we have: 由此可以看出, PDF $f_X(x)$ 的单位是

$\boxed{P([x, x+\delta])} = \int_x^{x+\delta} f_X(t)\,dt \boxed{\approx f_X(x)\cdot\delta} \Rightarrow f_X(x) \approx \frac{P([x, x+\delta])}{\delta}$ "Probability mass per unit length"

连续随机变量X取值在某个非常小的区间内的概率 ≈ PDF在该点的取值 × 小区间的长度. 随着这个区间长度趋向于0,

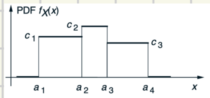
$P([x, x+\delta])$ 趋于0 $\Rightarrow P(X=a)=0$

PDF在某一点x处的值 是连续随机变量X在点x附近取值概率密度的度量, 而不是概率本身.

- 或代入公式理解: $P(a \le x \le a) = \int_a^a f(x)\,dx = 0$

⑤ $P(a<X<b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b)$

⑥ PDFs don't have to be continuous functions.



A piecewise constant PDF involving three intervals.

- 一旦我们得到了PDF, 可以不再关注其样本空间的构成. PDF为我们提供了计算和推导连续随机变量行为的所有必要信息. 之后我们默认 $\Omega = \mathbb{R}$. 若 $\Omega \subsetneq \mathbb{R}$, 我们可以通过将不属于Ω的x值在R上的概率密度定义为0来将模型扩展到整个R.

(2) CDF / Cumulative Density Function / (累积)分布函数

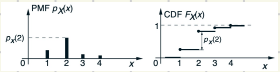
- Motivation

一些性质在离散随机变量和连续随机变量之间是共通的. CDF提供了一个统一的框架来表示随机变量的分布情况: $\{X \le x\}$ is always an event and therefore has a well-defined probability.

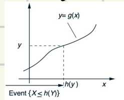
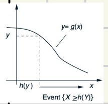
CDF是一个单调函数, 它给出随机变量取 ≤ 某个具体值x的概率.

Loosely speaking, the CDF $F_X(x)$ "accumulates" probability "up to" the value x. $F_X(x) = P(X \le x)$

- Def. and Properties

Discrete Random Variable X	Continuous Random Variable X
$F_X(x) = \sum_{k \le x} p_X(k)$	$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$
 $F_X(x)$ has a piecewise constant and staircase-like form. $p_X(k) = P(X \le k) - P(X \le k-1) = F_X(k) - F_X(k-1)$, $\forall k \in \mathbb{Z}$ 每当离散随机变量X的PMF在某个值k上有正值 $p_X(k)$, $F_X(k)$在k处会有一个跳跃, 其大小等于$p_X(k)$	• $F_X(x)$ has a continuously varying form. • The 2-step procedure to find a general PDF $Y = g(X)$ of a continuous random variable: ① Find the CDF of Y: $F_Y(y) = P(Y \le y)$ ② Differentiate: $f_Y(y) = \frac{dF_Y(y)}{dy}$
$F_X(x)$ is monotonically non-decreasing: $x \le y \Rightarrow F_X(x) \le F_X(y)$ $x \to \infty, F_X(x) \to 1 \qquad x \to -\infty, F_X(x) \to 0$	

- PDF for Functions of Random Variables

$Y = aX+b$ $f_Y(y) = \frac{1}{\lvert a\rvert} f_X\left(\frac{y-b}{a}\right)$ s.t. PDF integrates to 1		$Y = g(X)$, g is strictly monotonic and differentiable $f_Y(y) = f_X(h(y)) \left\lvert \frac{dh(y)}{dy} \right\rvert$	
$(a>0)$ $F_Y(y) = P(Y \le y)$ $= P(aX+b \le y)$ $= P(X \le \frac{y-b}{a})$ $= F_X(\frac{y-b}{a})$ $f_Y(y) = f_X(\frac{y-b}{a}) \cdot \frac{1}{a}$	$(a<0)$ $F_Y(y) = P(Y \le y)$ $= P(aX+b \le y)$ $= P(X \ge \frac{y-b}{a})$ $= 1 - P(X \le \frac{y-b}{a})$ $= 1 - F_X(\frac{y-b}{a})$ $f_Y(y) = -f_X(\frac{y-b}{a}) \cdot \frac{1}{a}$	• An important fact is that a monotonic function can be "inverted" in the sense that there is some function h, called the inverse of g, such that for all $x \in I$, we have $y = g(x)$ if and only if $x = h(y)$. g monotonically increasing $F_Y(y) = P(g(X) \le y)$ $= P(X \le h(y))$ $= F_X(h(y))$  $f_Y(y) = f_X(h(y)) \cdot \frac{dh(y)}{dy}$	g monotonically decreasing  $F_Y(y) = P(g(X) \le y)$ $= P(X \ge h(y))$ $= 1 - F_X(h(y))$ $f_Y(y) = -f_X(h(y)) \cdot \frac{dh(y)}{dy}$

(3) Joint PDF / 联合概率质量函数

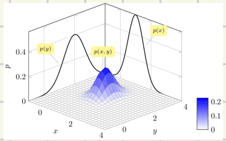
- Def

  Two continuous random variables associated with a common experiment are jointly continuous (联合连续) if they can be described by a joint PDF $f_{X,Y}$ that satisfies:

  $$\begin{cases} f_{X,Y}(x,y) \geq 0 \\ P((X,Y)\in B) = \iint_{(x,y)\in B} f_{X,Y}(x,y)\,dx\,dy \end{cases}$$ for every subset B of the two-dimensional plane.

- Interpretation

  我们不直接定义联合PDF为某个具体的概率值，而是通过定义联合PDF的功能来间接定义其概念，i.e., 联合PDF被用于计算概率，本身不表示概率.



  具体来说，联合PDF描述两个连续随机变量 X, Y 落在任意给定的二维区域 B 内的概率

  $f_{X,Y}(x,y)$ 在点 (x,y) 的值可以想象为点 (x,y) 上方的高度，这个高度代表了随机变量 (X,Y) 落在此点的概率密度.

  若在点 (x,y) 附近有一个非常小的矩形区域 $\Delta x \times \Delta y$，那么 X 和 Y 同时落在这个小矩形内的概率约为

  $$f_{X,Y}(x,y)\cdot\Delta x\cdot\Delta y$$

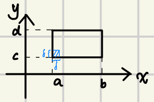
  将这种几何理解扩展到二维区域 B 时，计算 $P((X,Y)\in B)$ 的过程本质上是计算由 $f_{X,Y}(x,y)$ 形成的曲面与区域 B 所围成的体积.

  这个体积代表了在区域 B 上对 $f_{X,Y}(x,y)$ 和 $dx\cdot dy$ 的乘积进行整体积分.

- Properties

  ① $P(a\leq X\leq b,\ c\leq Y\leq d) = \int_c^d\int_a^b f_{X,Y}(x,y)\,dx\,dy$

  B is a rectangular of the form $B=[a,b]\times[c,d]$



  ② $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx\,dy = 1$

  ③ $\text{area}(B)=0 \Rightarrow P((X,Y)\in B)=0$ e.g. B是一条曲线

  如果矩形的边长都是某个很小的数值 δ:

  $$P(a\leq X\leq a+\delta,\ c\leq Y\leq c+\delta) = \int_c^{c+\delta}\int_a^{a+\delta} f_{X,Y}(x,y)\,dx\,dy \approx f_{X,Y}(a,c)\cdot\delta^2$$

  $f_{X,Y}(a,c) \approx \dfrac{P(a\leq X\leq a+\delta,\ c\leq Y\leq c+\delta)}{\delta^2}$ 由此可以看出，$f_{X,Y}(a,c)$ 的单位是 "Probability per unit area in the vinicity of (a,c)"

  ④ X, Y are continuous $\not\Rightarrow$ X and Y are jointly continuous.

  e.g. $X=Y$

  可以确定实验的结果将会落在 $x=y$ 这条线上，i.e. 所有的概率都集中在这条线上，而这条线的面积为0.

  我们在面积为0的集合上得到了正的概率，看似与性质③相矛盾，但这实际上意味着 X 和 Y 并不是联合连续的

  Joint continuity requires more than individual continuity of the random variables involved; it requires a genuine two-dimensional spread of probability.

- From the joint to the marginals

  $f_X(x) = \int f_{X,Y}(x,y)\,dy$ $\qquad f_Y(y) = \int f_{X,Y}(x,y)\,dx$

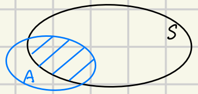
  e.g.

  (i)

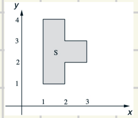
  $$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{\text{area of } S} & \text{if } (x,y)\in S, \\ 0 & \text{otherwise.} \end{cases}$$

  For any set $A\subset S$, the probability that the experimental value of (X,Y) lies in A is

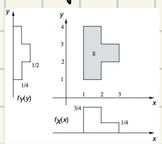
  $$\mathbf{P}\big((X,Y)\in A\big) = \iint_{(x,y)\in A} f_{X,Y}(x,y)\,dx\,dy = \frac{1}{\text{area of } S}\iint_{(x,y)\in A\cap S} dx\,dy = \frac{\text{area of } A\cap S}{\text{area of } S}.$$



  (ii) We are told that the joint PDF of the random variables X and Y is a constant c on the set S shown in Fig. and is zero outside. Find the value of C and the marginal PDFs of X and Y.



  The area of the set S is equal to 4 and, therefore, $f_{X,Y}(x,y)=c=1/4$, for $(x,y)\in S$. To find the marginal PDF $f_X(x)$ for some particular x, we integrate (with respect to y) the joint PDF over the vertical line corresponding to that x. The resulting PDF is shown in the figure. We can compute f_Y similarly.



- Expected Value Rule for Functions of Multiple Random Variables $E[g(X,Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)\cdot f_{X,Y}(x,y)\,dx\,dy$
- $E[aX+bY+c] = aE[X]+bE[Y]+c$

(4) Joint CDF / 联合分布函数

- Def

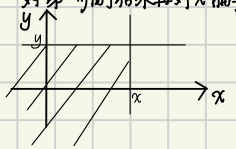
If X and Y are two continuous random variables associated with a common experiment, we define their joint CDF by:

$F_{X,Y}(x,y) = P(X \le x, Y \le y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(s,t)\,ds\,dt$ PDF → CDF

- CDF → PDF

$F_{X,Y}(x,y)$对y偏导：$\frac{\partial}{\partial y} F_{X,Y}(x,y) = \int_{-\infty}^{x} f_{X,Y}(s,y)\,ds$ $X \le x$且$Y = y$条件下的 marginal PDF

对第一步的结果再对x偏导：$\frac{\partial}{\partial x}\left(\int_{-\infty}^{x} f_{X,Y}(s,y)\,ds\right) = f_{X,Y}(x,y)$ X和Y在点(x,y)处的 joint PDF



$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}(x,y)$

(5) Mean and Variance

$E[X] = \int_{-\infty}^{+\infty} x f_X(x)\,dx$

离散随机变量X：$E[X] = \sum_x x \cdot p_X(x)$ (An integral is just a limiting form of a sum)

Can be interpreted as: Average in large number of independent repetitions of the experiment.

& "Center of gravity" of the probability

$Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x)\,dx$

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

- Linearity of Variance $Var(aX+b) = a^2 Var(X)$

(6) Conditional PDF / 条件概率密度函数

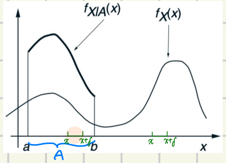
- Conditioning a random variable X on an Event A

$P(X\in B|A)=\int_B f_{X|A}(x)\,dx$, $P(A)>0$. for any subset B of the real line

- Intuition

	PMF	PDF	
Ordinary Probability Definition Calculation	$p_X(x)=P(X=x)$ $P(X\in B)=\sum_{x\in B}p_X(x)$	$f_X(x)\,\delta\approx P(x\le X\le x+\delta)$ $P(X\in B)=\int_B f_X(x)\,dx$	(**)
Conditional Probability Definition Calculation	$p_{X\mid A}(x)=P(X=x\mid A)$ $P(X\in B\mid A)=\sum_{x\in B}p_{X\mid A}(x)$	$f_{X\mid A}(x)\cdot\delta\approx P(x\le X\le x+\delta\mid A)$ $P(X\in B\mid A)=\int_B f_{X\mid A}(x)\,dx$	(*)

- Special Case: Conditional PDF of X, given that $X\in A$ A is a subset of the real line.



$$P(x\le X\le x+\delta\mid X\in A)\approx f_{X|X\in A}(x)\cdot\cancel{\delta}\quad(*)$$
$$=\frac{P(x\le X\le x+\delta,\ X\in A)}{P(X\in A)}$$
$$=\frac{P(x\le X\le x+\delta)}{P(X\in A)}\approx\frac{f_X(x)\cdot\cancel{\delta}}{P(X\in A)}\quad(**)$$

$$\Rightarrow\quad f_{X|X\in A}(x)=\begin{cases}\dfrac{f_X(x)}{P(X\in A)} & \text{if } x\in A\\ 0 & \text{otherwise}\end{cases}$$

$f_{X|A}(x)$ has exactly the same shape as the unconditional one, except that it's scaled by the constant factor $\frac{1}{P(X\in A)}$

- Calculation Rules

I. Total Probability Theorem / 全概率定理 $f_X(x)=\sum_{i=1}^{n}P(A_i)\cdot f_{X|A_i}(x)$

验证: $P(X\le x)=P(A_1)\cdot P(X\le x|A_1)+\cdots+P(A_n)\cdot P(X\le x|A_n)$

$\Rightarrow\ F_X(x)=P(A_1)\cdot F_{X|A_1}(x)+\cdots+P(A_n)\,F_{X|A_n}(x)$ Take the derivatives of both sides w.r.t. x

$\Rightarrow\ f_X(x)=P(A_1)\cdot f_{X|A_1}(x)+\cdots+P(A_n)\cdot f_{X|A_n}(x)$

II. Total Expectation Theorem / 全期望值定理 $\mathbb{E}[X]=\sum_{i=1}^{n}P(A_i)\cdot\mathbb{E}[X|A_i]$

验证: Multiply both sides of $f_X(x)=\sum_{i=1}^{n}P(A_i)\cdot f_{X|A_i}(x)$ by x and then integrate from $-\infty$ to $+\infty$.

Application: mixed distributions

Conditioning a random variable X on another $f_{X|Y}(x|y)=\dfrac{f_{X,Y}(x,y)}{f_Y(y)}$ if $f_Y(y)>0$

- 推导 $P(X\in A|Y=y)=\int_A f_{X|Y}(x|y)\,dx$

$f_{X|A}(x)\cdot\delta\approx\boxed{P(x\le X\le x+\delta\mid A)}$ where $P(A)>0$ (Conditional Probability Definition)

在事件A发生的条件下, 小区间 $[x,x+\delta]$ 的条件概率

我们不使用 $A=\{Y=y\}$, 因PDF用于确定小区间的概率, $P(Y=y)=0$

所以定义 $A=\{Y\approx y\}$, i.e., $y\le Y\le y+\varepsilon$

$\boxed{P(x\le X\le x+\delta\mid y\le Y\le y+\varepsilon)}$ (利用条件概率的定义 $P(A|B)=\frac{P(A\cap B)}{P(B)}$, $P(B)>0$)

$=\dfrac{P(x\le X\le x+\delta,\ y\le Y\le y+\varepsilon)}{P(y\le Y\le y+\varepsilon)}$ (x,y) 附近有一个非常小的矩形区域 $\delta\times\varepsilon$. X和Y同时落在这个小矩形内的概率约为 $f_{X,Y}(x,y)\cdot\delta\cdot\varepsilon$

$=\dfrac{f_{X,Y}(x,y)\cdot\delta\cdot\cancel{\varepsilon}}{f_Y(y)\cdot\cancel{\varepsilon}}$ $\left(f_{X|Y}(x|y)=\dfrac{f_{X,Y}(x,y)}{f_Y(y)}\right)$

$\boxed{=f_{X|Y}(x|y)\cdot\delta}$ 表示Y落在小区间 $[y,y+\varepsilon]$ 时, X落在小区间 $[x,x+\delta]$ 内的概率.

由于 $f_{X|Y}(x|y)\cdot\delta$ 不依赖于 ε, 可以考虑当 ε 趋近于0的极限情况, 得到 $P(x\le X\le x+\delta\mid Y=y)\approx f_{X|Y}(x|y)\cdot\delta$

More generally: $P(X\in A|Y=y)=\int_A f_{X|Y}(x|y)\,dx$

我们之前无法定义在给定0概率事件 $\{Y=y\}$ 下的条件概率, 但现在可以通过公式 $P(X\in A|Y=y)=\int_A f_{X|Y}(x|y)\,dx$ 处理这种情况.

In addition, it allows us to view the conditional PDF $f_{X|Y}(x|y)$ (as a function of x) as a description of the probability law of X, given that the event $\{Y=y\}$ has occured.

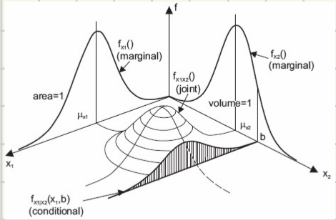
- Property

$$\int_{-\infty}^{+\infty}f_{X|Y}(x,y)\,dx=\frac{\int_{-\infty}^{+\infty}f_{X,Y}(x,Y)\,dx}{f_Y(y)}=1$$

当我们将y视为a fixed number, 并将 $f_{X|Y}(x|y)$ 视为单变量x的函数: 条件PDF $f_{X|Y}(x|y)$ 和联合PDF $f_{X,Y}(x,y)$ 具有相同的形状, 因为normalizing factor $f_Y(y)$ 不依赖于 x.

- Calculation Rules

I. Multiplication Rule $f_{X,Y}(x,y)=f_Y(y)\cdot f_{X|Y}(x|y)=f_X(x)\cdot f_{Y|X}(y|x)$



II. Conditional Expectation / 条件期望

$\mathbb{E}[X|A]=\int_{-\infty}^{+\infty}x\cdot f_{X|A}(x)\,dx$

$\mathbb{E}[X|Y=y]=\int_{-\infty}^{+\infty}x\cdot f_{X|Y}(x|y)\,dx$

$\mathbb{E}[g(X)|Y=y]=\int_{-\infty}^{+\infty}g(x)\cdot f_{X|Y}(x|y)\,dx$

$\mathbb{E}[g(X)|A]=\int_{-\infty}^{+\infty}g(x)\cdot f_{X|A}(x)\,dx$

III. Total Probability Theorem / 全概率定理 $f_X(x)=\int_{-\infty}^{+\infty}\underbrace{\boxed{f_Y(y)\cdot f_{X|Y}(x|y)}}_{=f_{X,Y}(x,y)}dy$

IV. Total Expectation Theorem / 全期望值定理 $\mathbb{E}[X]=\int_{-\infty}^{+\infty}f_Y(y)\cdot\mathbb{E}[X|Y=y]\,dy$

验证: $\int_{-\infty}^{+\infty}f_Y(y)\cdot\mathbb{E}[X|Y=y]\,dy$ 代入 $\mathbb{E}[X|Y=y]=\int_{-\infty}^{+\infty}x\cdot f_{X|Y}(x|y)\,dx$

$=\int_{-\infty}^{+\infty}f_Y(y)\cdot\int_{-\infty}^{+\infty}x\cdot f_{X|Y}(x|y)\,dx\,dy$

$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f_Y(y)\,x\cdot f_{X|Y}(x|y)\,dx\,dy$

$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f_Y(y)\,x\cdot f_{X|Y}(x|y)\,dy\,dx$

$=\int_{-\infty}^{+\infty}x\boxed{\int_{-\infty}^{+\infty}f_Y(y)\cdot f_{X|Y}(x|y)\,dy}\,dx$ (全概率定理)

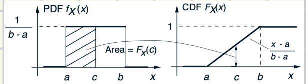
$=\int_{-\infty}^{+\infty}x\,f_X(x)\,dx\,dx$

$=\mathbb{E}[X]$

Continuous Uniform Random Variable

Parameters: a, b $\quad a \le b$

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \le x \le b \\ 0, & \text{otherwise} \end{cases}$$



$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \le x \le b \\ 0 & \text{if } x > b \end{cases}$$

- $P(X \in I) = \int_{[a,b]\cap I} \frac{1}{b-a}\,dx = \frac{1}{b-a}\int_{[a,b]\cap I} dx$

  $= \frac{\text{length of } [a,b]\cap I}{\text{length of } I}$

- 期望 $\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x)\,dx = \int_a^b x \cdot \frac{1}{b-a}\,dx = \frac{a+b}{2}$

- 方差 $var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{(b-a)^2}{12}$

  $\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f_X(x)\,dx = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx = \left(\frac{1}{b-a}\right)\left(\frac{b^3}{3} - \frac{a^3}{3}\right)$

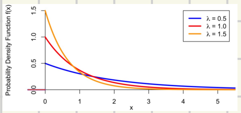
- 标准差 $\sigma = \frac{b-a}{\sqrt{12}}$

  $\frac{b-a}{\sqrt{12}}$ is proportional to the width of the uniform distribution. This aligns with the intuition that the standard deviation captures the "width" or dispersion of a particular distribution.

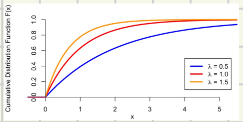
Exponential Random Variable

Parameter: $\lambda > 0$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \ge 0 \\ 0, & \text{otherwise} \end{cases}$$



$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \ge 0 \\ 0, & \text{otherwise} \end{cases}$$



$$\begin{aligned} P(X \ge a) &= \int_a^\infty \lambda e^{-\lambda x}\,dx \\ &= \lambda \int_a^\infty e^{-\lambda x}\,dx \\ &= \lambda \cdot \left(\frac{1}{-\lambda}\right) e^{-\lambda x}\Big|_a^\infty \\ &= -e^{-\lambda \cdot \infty} - (-e^{-\lambda a}) \\ &= e^{-\lambda a} \end{aligned}$$

(recall: $\int e^{ax}dx = \frac{1}{a}e^{ax}$
let $a = -\lambda$)

Observation: Let $a = 0$, we obtain the integral of the PDF over the entire range of X.

In this case, $P(X \ge 0) = 1$. Thus, we have verified that the integral of PDF is 1, as it should be.

- 期望 $\mathbb{E}[X] = \int_0^\infty x \cdot \lambda e^{-\lambda x}\,dx = \frac{1}{\lambda}$

- 方差 $var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{1}{\lambda^2}$

  $\mathbb{E}[X^2] = \int_0^\infty x^2 \cdot \lambda e^{-\lambda x}\,dx = \frac{2}{\lambda^2}$

Observation:

当λ较小时，PDF下降得较慢，这意味着X取较大值的概率增加，因此X的均值会变高，较大值的分布跨越范围更广，所以方差变大.

Application

Model the amount of time until a piece of equipment breaks down, until a light bulb burns out, or until an accident occurs.

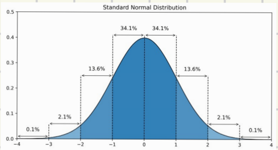
Standard Normal Random Variable

Parameters: $N(0,1)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

recall: $\int_{-\infty}^{+\infty} e^{-x^2/2}\,dx = \sqrt{2\pi}$

$F_X(x)$: standard normal table



- 期望 $\mathbb{E}[X] = 0$

  x^2 is symmetric around 0

  $\Rightarrow$ PDF $f_X(x)$ is symmetric around 0.

- 方差 $var(X) = 1$

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx \\ &= (2\pi)^{-1/2}\int_{-\infty}^{\infty} x^2 \exp\left(-\tfrac{1}{2}x^2\right)dx \\ &= (2\pi)^{-1/2}\left\{\int_{-\infty}^{0} x\left(x\exp\left(-\tfrac{1}{2}x^2\right)\right)dx + \int_0^\infty x\left(x\exp\left(-\tfrac{1}{2}x^2\right)\right)dx\right\} \\ &= (2\pi)^{-1/2}\left\{\left[-x\exp\left(-\tfrac{1}{2}x^2\right)\right]_{-\infty}^{0} + \int_{-\infty}^{0}\exp\left(-\tfrac{1}{2}x^2\right)dx + \left[-x\exp\left(-\tfrac{1}{2}x^2\right)\right]_0^\infty \right. \\ &\quad \left. + \int_0^\infty \exp\left(-\tfrac{1}{2}x^2\right)dx\right\} \quad \text{(integrating by parts)} \\ &= (2\pi)^{-1/2}\left\{(0-0) + (0-0) + \int_{-\infty}^{0}\exp\left(-\tfrac{1}{2}x^2\right)dx + \int_0^\infty \exp\left(-\tfrac{1}{2}x^2\right)dx\right\} \\ &= (2\pi)^{-1/2}\int_{-\infty}^{\infty}\exp\left(-\tfrac{1}{2}x^2\right)dx \\ &= \int_{-\infty}^{\infty} f_X(x)\,dx = 1 \quad \text{(the integral of a pdf over its support is equal to 1)} \end{aligned}$$

$E[X]^2 = 0^2 = 0$

$Var[X] = E[X^2] - E[X]^2 = 1 - 0 = 1$

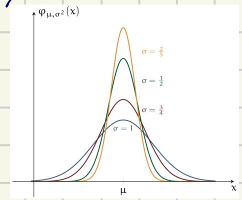
Application

- It models well the additive effect of many independent factors in a variety of contexts. It's important in the theory of probability, e.g. Central Limit Theorem.

Normal Random Variable

Parameters: $N(\mu, \sigma^2)$ $\quad \mu \in \mathbb{R}, \sigma > 0$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



$F_X(x)$: standard normal table

$$\begin{aligned} P(X \le x) &= P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) \\ &= P\left(Y \le \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

Y衡量X与均值μ之间的偏差. 除以标准差后, Y以标准差为单位来量化这个偏差. 若Y=3, 则意味着X与均值之间有3个标准差的距离.

Y是一个标准正态随机变量:

$\mathbb{E}[Y] = 0 \quad (\because X - \mu)$

$Var(Y) = \frac{1}{\sigma^2}\cdot Var(X) = 1$

- 期望 $\mathbb{E}[X] = \mu$

  $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is symmetric around μ

- 方差 $var(X) = \sigma^2$

$$\text{var}(X) = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty}(x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2}\,dx.$$

Using the change of variables $y = (x-\mu)/\sigma$ and integration by parts, we have

$$\begin{aligned} \text{var}(X) &= \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y^2 e^{-y^2/2}\,dy \\ &= \frac{\sigma^2}{\sqrt{2\pi}}\left(-ye^{-y^2/2}\right)\Big|_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\,dy \\ &= \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\,dy \\ &= \sigma^2. \end{aligned}$$

The last equality above is obtained by using the fact

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\,dy = 1,$$

which is just the normalization property of the normal PDF for the case where $\mu = 0$ and $\sigma = 1$.

- Normality is preserved by linear transformations

  Let $X \sim N(\mu, \sigma^2)$ $\quad Y = aX + b$

  $\mathbb{E}[Y] = a \cdot \mathbb{E}[X] + b = a\mu + b$

  $Var(Y) = a^2 \cdot Var(X) = a^2\sigma^2$

  或利用之前得出的结论 $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) = \frac{1}{|a|}\left(\frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\left(\frac{y-b}{a}-\mu\right)^2/2\sigma^2}\right)$

  $= \frac{1}{\sqrt{2\pi}\sigma \cdot |a|}\left(e^{-(y-b-a\mu)^2/2a^2\sigma^2}\right)$

  $Y \sim N(a\mu + b, a^2\sigma^2)$ $\quad = \frac{1}{\sqrt{2\pi}\sigma \cdot |a|}\left(e^{-(y-(b+a\mu))^2/2(a\sigma)^2}\right)$

  若$a=0$, 则$Y=b$, 即Y退化为一个常数, 但我们仍可使用正态分布的表示方法: $Y \sim (b, 0)$

Application

- In signal processing and communication engineering, model noise and unpredictable distortions of signals