Discrete probability distribution

分布列

总概率=1

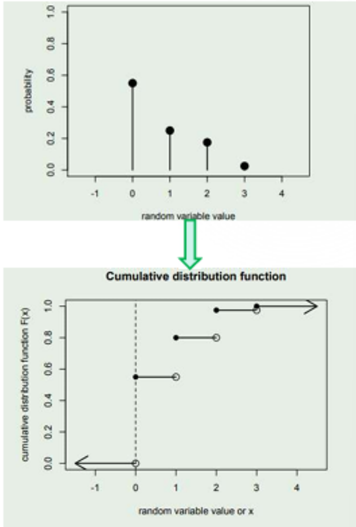
- **期望**
  - $$\mu = E(X) = \sum x \cdot P(X = x)$$
  - 加和/积分
- **方差**
  - $$\sigma^2 = Var(X) = E\left[\left(x - E(X)\right)^2\right] = \sum_x \left(x - E(X)\right)^2 P(X = x)$$
    $$\sigma = SD(X) = \sqrt{\mathrm{Var}(X)}$$
- **cumulative distribution function**概率分布函数（aka累积分布函数 分布函数 CDF上图）
  - $$F(x) = \mathbb{P}(X \leq x), \quad x \in R,$$
  - CDF是非减函数，0≤f≤1
  - 反之 1-P(X≤x)=P(X > x) 生存函数
  - probability mass function PMF下图的加和
  step function阶梯状

2.3 Properties of CDF – discrete r.v.

- For discrete r.v., CDF is an non-decreasing **step function** with left-closed and right-open intervals.

$$P(X = x_i) = F(x_i) - \lim_{x \uparrow x_i} F(x_i)$$

x	$P(X=x)$	$P(X \le x) = F(x)$ (Cumulative Probabilities...)	
0	0.550	0.550	$P(X \le 0) = F(0)$
1	0.250	0.800	$P(X \le 1) = F(1)$
2	0.175	0.975	$P(X \le 2) = F(2)$
3	0.025	1.000	$P(X \le 3) = F(3)$



- $0 \le F(x) \le 1$
- If $x \le y$, then $F(x) \le F(y)$

清华大学统计学研究中心

两段之间的差值：取到该值的概率？ 右连续

Binomial Distribution

只有两种可能结果
每一个样本完成一次伯努利试验，总体符合二项分布

Bernoulli trial：是在同样的条件下重复地、相互独立地进行的一种随机试验，其特点是该随机试验只有两种可能结果：发生或者不发生。

每个人只做一次也可以是重复

$$X \sim \text{Binom}(n, p)$$

$$P(X = k|n, p) = f(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

P(x=1)=p
p(x=0)=1-p
所以总的就可以写成指数形式↑

x1、x2、x3…分布都满足伯努利
p(x1=1)=p(x2=1)=....=p
k=Σxi=0~n最后出现1的个数（n个独立的伯努利试验的加和的含义）
Y=(Σxi=)k时该分布表现为二项分布)~B(n,p)

$\binom{n}{k}$

这是啥

图像

右偏/左偏/对称

2.4 Properties of a Binomial RV

22

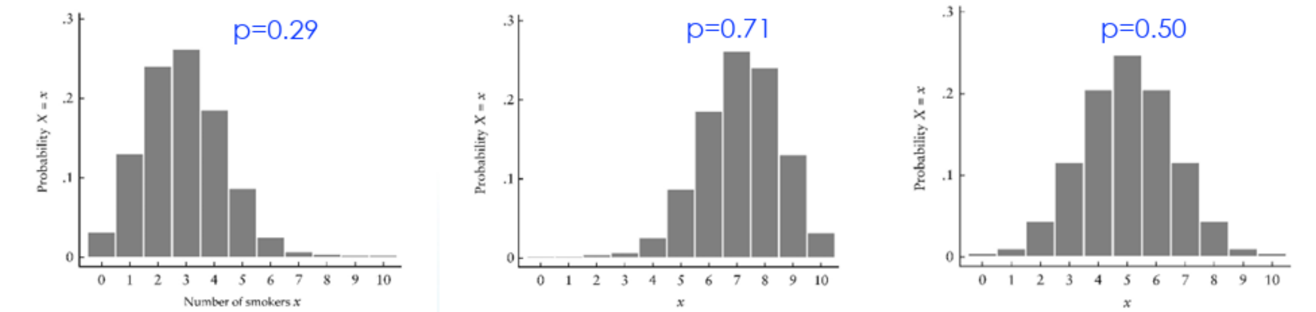


FIGURE 7.2 Probability distribution of a binomial random variable for which $n = 10$ and $p = 0.29$

FIGURE 7.3 Probability distribution of a binomial random variable for which $n = 10$ and $p = 0.71$

FIGURE 7.4 Probability distribution of a binomial random variable for which $n = 10$ and $p = 0.50$

Let $X \sim \text{Binom}(n, p)$ then $X = \sum_{i=1}^{n} Y_i$ where $Y_1, \cdots, Y_n \sim \text{Bern}(p)$.

$$E(X) = E\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} E(Y_i) = \sum_{i=1}^{n} p = np$$

$$Var(X) = Var\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} Var(Y_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

清华大学统计学研究中心

期望：平均/长期来看最可能出现的值

用右式（加和的方差=方差的加和）算方差的**前提是yi互相独立**否则存在协方差

Multinomial Distribution多项分布

对应的试验中每一个个体的试验结果多于两个（推广） 每一个结果对应的人数的分布为多项分布

独立

我最喜爱的食堂

人数	x_1	x_2	x_3	x_4
入选率	p_1	p_2	p_3	p_4

p=x/n 每一个人选择的比重相同不存在加权ry（ *呈现* 概率相同）

最终的概率表达（假设六种分类 $\sum_{i=1}^{6} p_i = 1$）：

$$P(x_1, x2, \cdots, x_6) = \frac{n!}{x_1!x_2!\cdots x_6!} p_1^{x_1} p_2^{x_2} \cdots p_6^{x_6}$$

Joint probability共同出现该结果的概率?

Poisson Distribution泊松分布

时间

在一个特定时间内，某件事情会在任意时刻随即发生，且每一次发生都是独立的

当我们把时间段分称非常非常小的时间片构成时，在每个**时间片**内，该事件可能发生，可能不发生

特定时间段被分称 n 个**时间片**，当时间片很小时，时间段内发生事件的概率 p 成比例减小

但该事件在指定时间段内发生的频度相同， $n * p = \mu$ 为常数。

营业时间 T 内有 k 个顾客到达超市的概率为： $P = \lim_{n\to\infty} C_n^k\, p^k(1-p)^{n-k}$ 且 $p = \frac{\mu}{n}$

$$P = \lim_{n\to\infty} C_n^k \left(\frac{\mu}{n}\right)^k \left(1-\frac{\mu}{n}\right)^{n-k} = \frac{\mu^k}{k!}e^{-\mu} = P(X=k)$$

通常在泊松分布里， μ 被写成 λ, 某特定时间内发生的频数。

时间片很小：每个时间片内 **最多** 只能有一个事件发生or不发生 可能有时间片概率=0

二项分布的极限分布

k没有上限

μ为常数（平均）

- describes occurrences or objects which are distributed randomly in space or time
- often used to describe distribution of th enumber of occurrences of a rare event
- underlying assumptions similar to those for binomial distribution

- useful when there're counts with no denominator

2.6 Common Discrete R.V. with Poisson Distribution

28

The probability of x occurrence of an event in an interval is:

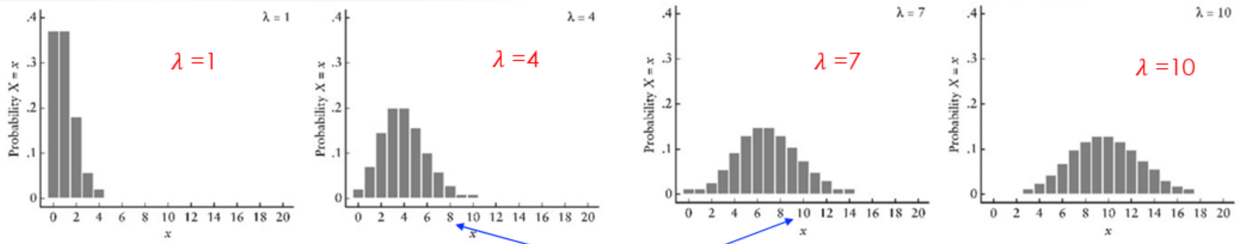
$$P(X=x)=\frac{e^{-\lambda}\cdot\lambda^{x}}{x!}, x=0,1,2,\ldots$$

怎么求区间的概率呢?

where λ = the expected number of occurrences in the interval

e = a constant (≈ 2.718)

$\boxed{\text{For the Poisson distribution: mean = variance} = \lambda}$



清华大学统计学研究中心

mean=variance=λ

成比例放缩eg λ（半天）=5 求整天p（x=20）

怎么求区间的概率呢?

2.6 Common Discrete R.V. with Poisson Distribution

Examples:
- Spatial distribution of stars, weeds, bacteria, flying-bomb strikes
- Emergency room or hospital admissions
- Deaths due to a rare disease
- More ….

Assumptions:
- The occurrences of a random event in an interval of time are **independent**
- In theory, an infinite number of occurrences of the event are possible (though perhaps rare) within the interval
- In any extremely small portion of the interval, the probability of more than one occurrence of the event is approximately zero

平均值→【某数量】概率

Continuous Random Variable

取值可能性 *无穷多* 不可能取到特定值

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

interval

$$P(a \leq X \leq b)$$

density probability density function(pdf), denoted by $f_X(x)$ (区分F(x) F(x)在此处也适用)

- Every continuous random variable X has a probability density function (pdf), denoted by $f_X(x)$.
- PDF is a function such that

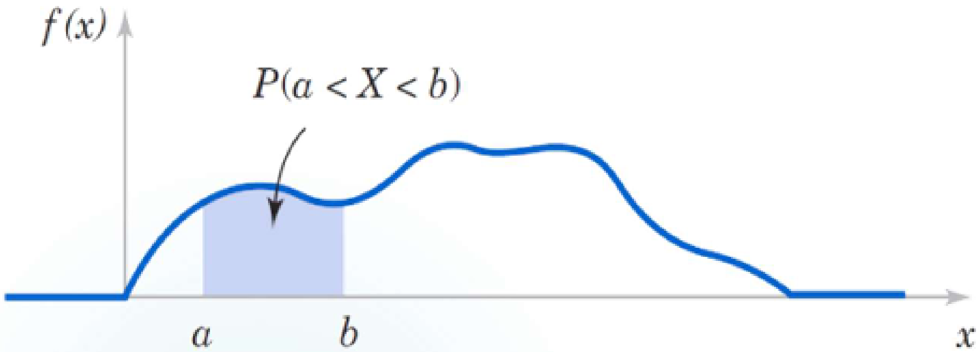
a $f_X(x) \geq 0$ for any $x \in \mathbb{R}$

b $\int_{-\infty}^{\infty} f_X(x)dx = 1$

c $P(a \leq X \leq b) = \int_a^b f_X(x)dx$, which represents the area under $f_X(x)$ from a to b for any $b > a$.

d If x_0 is a specific value, then $P(X = x_0) = 0$. We assign 0 to area under a point.

概率=面积



Let X_0 be a specific value of interest, the **cumulative distribution function (CDF)** is defined via

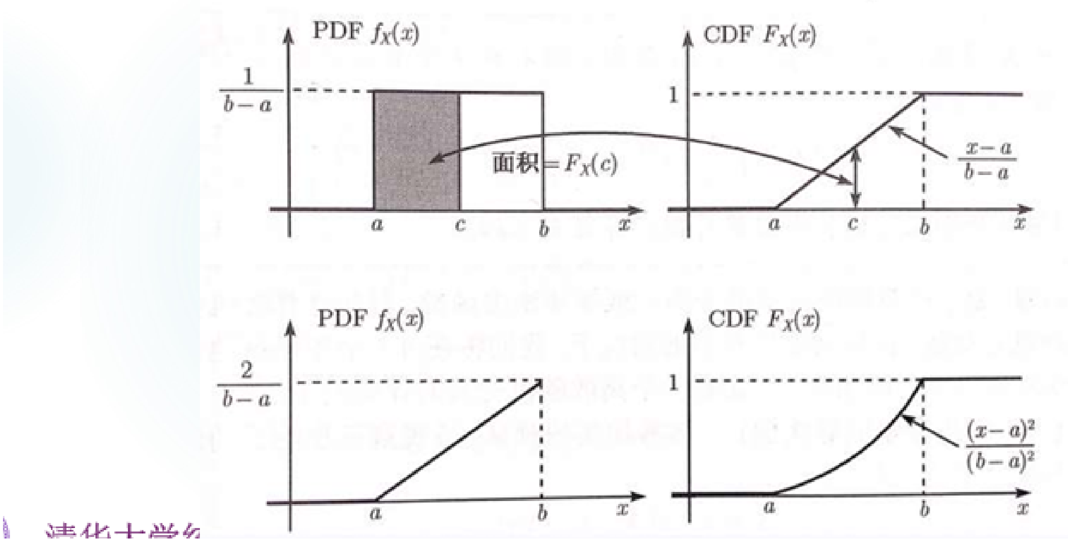
$$F_X(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} f_X(x)dx.$$

等号随意

If X is a **continuous r.v.** with probability density function (pdf) f(x), then the CDF of X

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)\,dt, \quad x \in R$$

is a continuous function, and f(x) = F'(x).



可通过CDF图像区分离散型和连续型（连续曲线） 一直增加到1为止，PDF就不一定了
☆f(x)=F'(x)

续

协方差
cov(X,Y)=E((X-E(x))(Y-E(Y)))=E(XY)-E(X)E(Y)
Var(X±Y)=Var(X)+Var(Y)±2Cov(X,Y)
与是否独立有关
两个变量相减时离散性同样会扩大

Cumulative Distribution Function

确定CDF→求导得pdf（一一对应）
CDF

- 非减
- 右连续
- F(-∞)=0，F(+∞)=1 端点
- $$P(a < X \leq b) = \int_a^b f(t)\,dt = F(b) - F(a)$$
  区间
  (-∞~b - -∞~a)

Mean of a Continuous RV

加和→积分

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

μx是啥

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

- The variance of X, denoted as Var(X) or σ^2 is

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

- Standard deviation

$$\sigma = \sqrt{\sigma^2}.$$

- The computational formula for variance is the same as the discrete case

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.$$

Summary

3.7 Comparison：Continuous and Discrete R.V.

Continuous Random Variable	Discrete Random Variable
X can take on all possible values in an interval of real numbers. e.g. $X \in [0,1]$	X can take on only distinct 'discrete' values in a set. e.g. $X \in \{0,1,2,3,\ldots,\infty\}$
Probability density function, $f(x)$	Probability mass function, $f(x)$
Cumulative distribution function, $F(x) = P(X \le x) = \int_{\infty}^{x} f(u)du$	Cumulative distribution function, $F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i)$
$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$	$\mu = E(X) = \sum_x x f(x)$
$\sigma^2 = V(X) = E(X-\mu)^2$	$\sigma^2 = V(X) = E(X-\mu)^2$
$= \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$	$= \sum_x (x-\mu)^2 f(x)$
$= E(X^2) - [E(X)]^2$	$= E(X^2) - [E(X)]^2$
$= \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$	$= \sum_x x^2 f(x) - \mu^2$

清华大学统计学研究中心

Expected Value

E(px+q-y)=pE(x)+q-E(y)
E(XY)=E(X)E(Y)(independent)
$E(XY)=\iint xyf_{xy}(xy)$ 联合积分（总之就是用定义求）

Variance

Var(c)=0 if c is constant
Var(px+q-y)=p^2Var(x)-Var(y)(independent)注意常数没了 整体平移

Normal distribution

连续型

正态分布/高斯分布/常态分布

单峰、对称、钟形 取值-∞~+∞
Mean=Median=Mode
$N(\mu,\sigma^2)$
位置参数μ=mean location parameter || 幅度参数σ^2=variance scale parameter
☆☆☆记

- The normal probability distribution is:

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\cdot e^{-(x-\mu)^2/2\sigma^2}, -\infty < x < +\infty$$

- $\pi \approx 3.14$ and $e \approx 2.72$

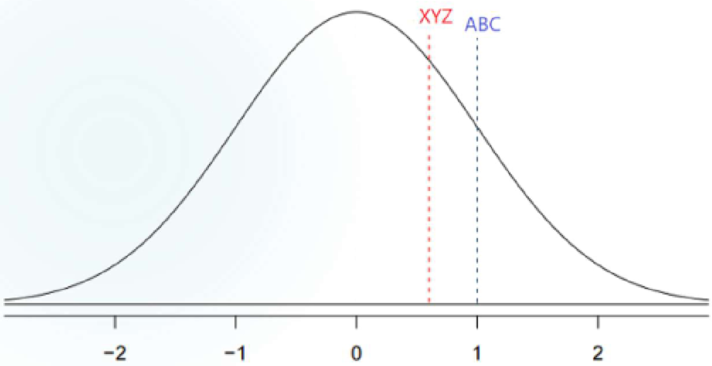
68-95-99.7 Rule

about 68% falls within 1 SD of the mean
about 95% falls within 2 SD of the mean *绝大多数* 但如果特别集中的话也不一定能作为评判标准
about 99.7% falls within 3 SD of the mean

标准化 相对位置

Since we cannot just compare these two raw scores, we instead compare how many standard deviations beyond the mean each observation is.

- For the candidate from ABC, score is (180-150)/30 = 1 SD above the mean.
- For the candidate from XYZ, score is (24-21)/5 = 0.6 SD above the mean.



$$Z = \frac{observation - mean}{SD}$$

- Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate the percentiles
- **|Z|>2 unusual**
-
  - $Z \sim N(0,1)$ 标准正态分布
  - If $X \sim N(\mu, \sigma)$, then $\frac{X-\mu}{\sigma} \sim N(0,1)$
    任意正态分布

percentile

below a given data point
quantile

```
qnorm() # 百分位数

qnorm(0.5)
> 0 (标准正态分布)
```

CDF

3.9.1 CDF for Standard Normal Distribution

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)\,dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt.$$

Since $\varphi(t)$ is symmetric, $\Phi(x) + \Phi(-x) = 1$, or $\Phi(-x) = 1 - \Phi(x), x \in R$.

If X ~ N(μ, σ^2),

$$\begin{aligned}\mathbb{P}(X \le a) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(a-\mu)/\sigma} \mathrm{e}^{-y^2/2}\,dy \\ &= \Phi\left(\frac{a-\mu}{\sigma}\right).\end{aligned}$$

Φ(X)：标准正态分布(pdf)对应的CDF
Calculating Probability

```
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
```

Caculating Quantile

```
qnorm(p)
```

---

mutually exclusive
independence

express the distribution
poisson distribution：n较大p较小 保证np=μ constant

Joint probabilities, CLT and sampling distribution