Probability and Statistics – P_1

离散概率空间(discrete probability space) (S, M, P) 由三部分组成

样本空间(sample space)，指一个包含所有可能样本的有限或可数无限集合，记为 S

事件集合(set of event)，通常是样本空间的幂集 $P(S)$，记为 $M = P(S)$

概率函数(probability function)，从事件集合 M 到 $[0,1]$ 内的实数的函数 $P: M \to [0,1]$

事件集合 M 可知 空集 $\emptyset$ 和 样本空间 S 本身都是事件集合 M 中的元素

对于集合 A_1, A_2，有 $A_1 \in M$ 且 $A_2 \in M$.

$A_1 \cup A_2 \in M$，并且类似地有 $A_1 \cap A_2 \in M$

$\overline{A_1} \in M$，注意这里定义 $\overline{A_1} = \{x \in S \mid x \notin A_1\}$

于是类似地有 $A_1 - A_2$，$A_1 \oplus A_2 \in M$

由此扩展为 如果集合 $\{A_1, A_2, \cdots, A_n, \cdots\} \subseteq M$，则有 $\bigcup_{n=1}^{\infty} A_n \in M$

closed under finite，指从一组初始的集合开始，经过有限次数的交集和并集生成属于这个集合的新元素

即 如果 $\{A_1, A_2, \cdots, A_n\} \subseteq M$，其中 $n \in \mathbb{Z}^+$，则 $\bigcup_{i=1}^{n} A_i \in M$，

则称集合 M 是 closed under finite unions.

类似地可定义 closed under finite complementation / intersection

closed under countable，与 closed under finite 的差别在于是否允许初始集合是可数无限的

即如定义 对于非负整数 n 的集合 $A_n = \{1, 2, \cdots, n\}$，于是有 $A_0 = \emptyset$

定义集合 $\Omega = \{A_i \mid 0 \le i \le n, n \in N\}$，

则对于有限多个 n_k，有 构成的子集合 $\{n_1, n_2, \cdots, n_k\}$

$A_{n_1} \cup A_{n_2} \cup \cdots \cup A_{n_k} = A_{\max\{n_1, n_2, \cdots, n_k\}} \in \Omega$

于是有集合 Ω 是 closed under finite unions 的

但是如果对于可数无限多个 n_k 的子集合 $\{n_1, n_2, \cdots, n_k, \cdots\}$

$\bigcup_{n \in \{n_1, \cdots, n_k, \cdots\}} A_n = \{1, 2, \cdots, n, \cdots\} \notin \Omega$.

所以集合 Ω 不是 closed under countable union 的

概率函数 P，函数 P 是从事件集合 M 到 $[0,1]$ 内的实数的函数，即 $P: M \to [0,1]$

对于空集 $\emptyset$，$P(\emptyset) = 0$

对于样本空间 S，$P(S) = 1$

如果事件 $A, B \in M$，且 A, B 是不相交事件，则有 $P(A) + P(B) = P(A \cup B)$

而根据容斥原理，$P(A \cup B) = P(A) + P(B) - P(A \cap B)$，可知 $P(A \cap B) = 0$

扩展到事件序列 $\{E_1, E_2, \cdots, E_n, \cdots\} \subseteq M$，如果 $E_1, E_2, \cdots$ 是两两不相交的(pairwise disjoint)

则有 $P(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} P(E_n)$

Probability and Statistics – P2

密度函数 f (density)，对于给定的离散概率空间 (S, M, P)，有与之对应的概率密度函数 f

已知对于任意样本 $x \in S$，有 $\{x\} \in M$，即 $P(\{x\})$ 有定义

定义从样本空间 S 到 $[0,1]$ 上的实数的概率密度函数 $f: S \to [0,1]$

为 $f(x) = P(\{x\})$，即 x 上的概率密度等于仅当 x 发生的事件 $\{x\}$ 的概率

独立性 (independence)，对于给定的离散概率空间 (S, M, P)

对于事件 $A, B \in M$，称事件 A 和 B 是独立的

当且仅当 $P(A \cap B) = P(A) \cdot P(B)$，记为 $A \perp B$ 或 $A \perp\!\!\!\perp B$，

对于一个有限的事件集合 $\{A_1, A_2, \cdots, A_n\}$，$n \in \mathbb{Z}^+$，

如果对于任意正整数 $1 \le i < j \le n$，有 $P(A_i \cap A_j) = P(A_i) P(A_j)$

则称事件集合是两两独立的 (pairwise independent)

对于一个有限的事件集合 $\{A_1, A_2, \cdots, A_n\}$，$n \in \mathbb{Z}^+$

如果对于该集合的任意子集，即对于任意 $\{B_1, \cdots, B_k\} \subseteq \{A_1, \cdots, A_n\}$，其中 $0 \le k \le n$，

有 $P(\bigcap_{i=1}^{k} B_i) = \prod_{i=1}^{k} P(B_i)$ 成立，

则称事件集合是相互独立的 (mutually independent)

注意两两独立与相互独立之间的区别

如掷一个标准四面骰，即 $C(d4)$，则有 $S = \{1, 2, 3, 4\}$，$M = P(S)$，$\forall x \in S$，$f(x) = \frac{1}{4}$

则取事件 $A = \{1, 2\}$，$B = \{2, 3\}$，$C = \{1, 3\}$.

可知，$P(A) = P(B) = P(C) = \frac{1}{2}$，$P(\{1\}) = P(\{2\}) = P(\{3\}) = \frac{1}{4}$

则 $P(A \cap B) = P(A) P(B)$，$P(B \cap C) = P(B) P(C)$，$P(A \cap C) = P(A) P(C)$

即，事件 A, B, C 是两两独立的.

但是注意有 $P(A \cap B \cap C) = P(\emptyset) = 0 \ne \frac{1}{8} = P(A) P(B) P(C)$

所以事件 A, B, C 不是相互独立的

离散随机变量 (discrete random variable)，对于离散概率空间 (S, M, P)，且有事件集合 $M = P(S)$，

则定义从样本空间 S 到实数集合 R 的函数 $X: S \to R$，称为离散随机变量

从而产生一个新的离散概率空间 (S', M', P')，其中 S' 为 X 的值域 $S' = \{X(s) \mid s \in S\}$

事件集合 $M' = P(S')$，概率函数 P' 由概率密度函数决定

即 $f'(x) = P(\{s \in S \mid X(s) = x\})$.

如连续掷一枚硬币两次，则样本空间为 $S = \{00, 01, 10, 11\}$.

取函数 $X(s)$ 为样本 s 中正面 1 的次数.

假定硬币是均匀的，则 $f'(1) = P(\{s \in S \mid X(s) = 1\}) = P(\{01, 10\}) = \frac{1}{2}$

Probability and Statistics - P3

伯努利试验(Bernoulli Trial)，对于离散概率空间(S, M, P)，其中$M = P(S)$，

有样本空间S有且仅有两个元素(exactly two possible outcomes)，如{success, failure}

成功率(probability of success) $p \in [0,1]$，

且有$f(\text{success}) = p$，$f(\text{failure}) = 1-p$，

则称为成功率为p的伯努利试验(Bernoulli Trial with probability of success p)

多重伯努利试验(multiple Bernoulli trials)，用于描述重复n次的独立伯努利试验的结果，其中单次成功率为p

model the experiment of n repetitions of independent Bernoulli trials with probability of success p

即有离散概率空间(S, M, P)，其中$M = P(S)$

样本空间S为所有长度为n，由success, failure组成的序列的集合，

set of all sequences of n total success's and failure's

即 $S = \{(a_1, a_2, \cdots, a_n) \mid a_i \in \{\text{success}, \text{failure}\}, i = 1, 2, \cdots, n\}$

或者可表示为 $S = \{\text{success}, \text{failure}\}^n$，其中$n \in \mathbb{Z}^+$

密度$f((a_1, a_2, \cdots, a_n)) = p^k(1-p)^{n-k}$，其中$k$为$(a_1, a_2, \cdots, a_n)$中success的个数，$0 \le k \le n$

或者有 $f((a_1, \cdots, a_n)) = \prod_{i=1}^{n} g(a_i)$，其中$g(\text{success}) = p$，$g(\text{failure}) = 1-p$

二项式分布(binomial distribution)，描述一个成功率为p的n重伯努利试验中成功次数的分布

即有离散随机变量$X: S \to N$，其中$X((a_1, a_2, \cdots, a_n))$等于$(a_1, \cdots, a_n)$中success的个数

于是对于离散概率空间(S', M', P')，其中$M' = P(S')$

令集合$S = \{\text{success}, \text{failure}\}^n$，其中$n \in \mathbb{Z}^+$

则X的值域(range) $X(S) = \{0, 1, 2, \cdots, n\}$，即样空间$S' = X(S)$

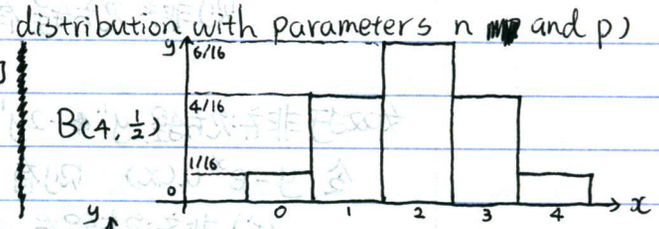
密度函数 $f'(k) = \binom{n}{k} p^k (1-p)^{n-k}$，其中$0 \le k \le n$，或$k \in S'$

或者有 $f'(k) = \sum_{\{s \in S \mid X(s) = k\}} f(s)$，又$f(s) = p^k(1-p)^{n-k}$，$|\{s \in S \mid X(s) = k\}| = \binom{n}{k}$

于是有$f'(k) = \binom{n}{k} p^k (1-p)^{n-k}$

称为参数为n, p的二项式分布(binomial distribution with parameters n and p)

记为$X \sim B(n, p)$，其中$n \in N$，$p \in [0,1]$



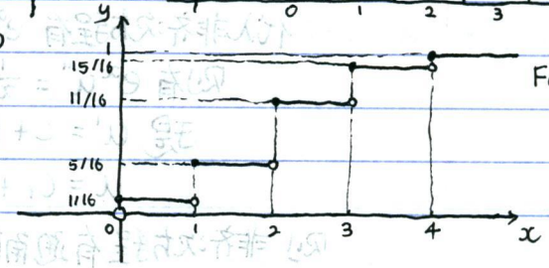
累积分布函数(cumulative distribution function, CDF)

对于离散概率空间(S, M, P)，其中$M = P(S)$

如果有样本空间$S \subseteq \mathbb{R}$

则定义累积分布函数$F: \mathbb{R} \to [0,1]$为

$F(t) = P(\{x \in S \mid x \le t\})$



Probability and Statistics - P4

连续概率空间(continuous probability space)，与离散概率空间类似，由$(S, \mathcal{M}, P)$三部分组成

样本空间S是一个可测集(measurable set)，且有$S \subseteq \mathbb{R}$，包含所有可能结果

事件集合$\mathcal{M}$是包含$\mathbb{R}$的子集的集合，且其元素均为可测集

概率函数P，定义从事件集合$\mathcal{M}$到$[0,1]$上的实数的函数，即$P: \mathcal{M} \to [0,1]$

事件集合$\mathcal{M}$　连续概率空间的事件集合有与离散概率空间相似的性质

空集$\emptyset$和样本空间S都是事件集合$\mathcal{M}$的元素

如果集合A_1, A_2是事件集合$\mathcal{M}$的元素，

则有$A_1 \cup A_2 \in \mathcal{M}$，类似地有$A_1 \cap A_2 \in \mathcal{M}$

$\overline{A_1} \in \mathcal{M}$，这是定义$\overline{A_1} = \{x \in S \mid x \notin A_1\}$

并且可以扩展为，如果$\{A_1, A_2, \cdots, A_n, \cdots\} \subseteq \mathcal{M}$，则$\bigcup_{i=1}^{\infty} A_i \in \mathcal{M}$，$\bigcap_{i=1}^{\infty} A_i \in \mathcal{M}$，$\forall_{i \in \mathbb{Z}^+} \overline{A_i} \in \mathcal{M}$

则称事件集合$\mathcal{M}$是closed under countable union (complementation, intersection)

实际上，事件集合包含样本空间S中的所有区间(interval)，

特别地有，$\mathcal{M}$的元素全部是可测集($\mathcal{M}$ is all Lebesgue measurable sets)

概率函数P　对于一个连续概率空间$(S, \mathcal{M}, P)$，概率函数P是定义从事件集合$\mathcal{M}$到$[0,1]$上实数的函数

存在一个可测的函数f，在样本空间S上是可积的

则对于任意$A \in \mathcal{M}$，有$P(A) = \int_A f(x)dx$，

即可以一般地称概率函数P在事件A上的值，等于函数$f(x)$在事件A的区间上的积分

称函数f为概率密度函数(probability density function, pdf)

且有对于$\forall x \in \mathbb{R}$ $f(x) \geq 0$，如果$x \notin S$，$f(x)$可能无定义

则对于$x \in \{x \in \mathbb{R} \mid x \notin S\}$，定义$f(x) = 0$

有$\int_{\mathbb{R}} f(x)dx = \int_S f(x)dx + \int_{\overline{S}} f(x)dx = 1 + 0 = 1$

即$P(S) = \int_S f(x)dx = 1$

$P(\emptyset) = \int_\emptyset f(x)dx = 0$

对于任意$x \in S$，$P(\{x\}) = \int_x^x f(x)dx = 0$

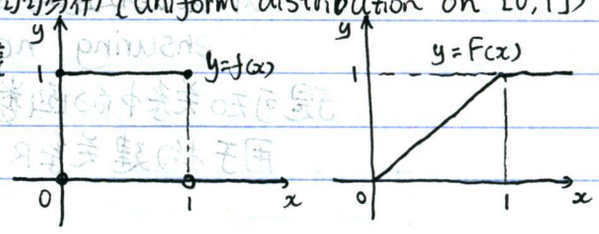
直观上即为，对于一个有无限多个可能样本的概率空间

取到任一单个样本的概率为0

连续均匀分布(uniform distribution)，对于一个0-1的连续均匀分布(uniform distribution on $[0,1]$)

有样本空间$S = [0,1]$，事件集合$\mathcal{M}$包含S的可测子集

概率密度函数$f(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$



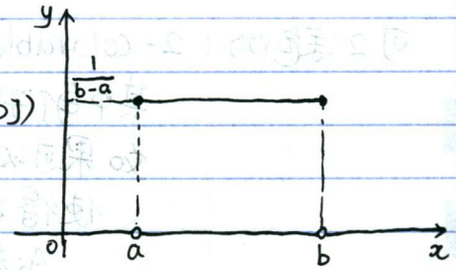
Probability and Statistics - P5

一般均匀分布 (general uniform distribution). 对于实数 $a<b$,

有 $[a,b]$ 上的均匀分布 (general uniform distribution on $[a,b]$)

样本空间 S 为 $[a,b]$, 事件集合 M 包含 S 的可测子集

概率密度函数 $f(x)=\begin{cases}\frac{1}{b-a}, & x\in[a,b]\\ 0, & \text{otherwise}\end{cases}$



正态分布 (normal distribution). 又称高斯分布 (Gaussian distribution)

样本空间 S 为 $\mathbb{R}$, 事件集合 M 包含 S 的可测子集

对于给定的实数 μ, σ, 则定义概率密度函数

$f(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, 其中 $x\in\mathbb{R}$

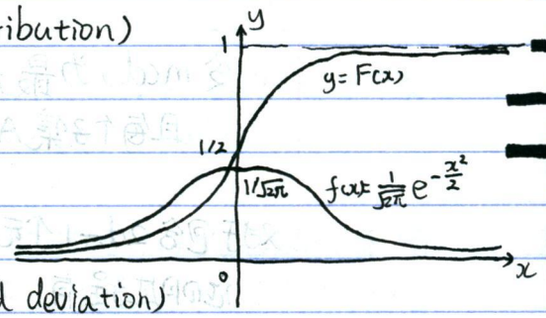
其中 μ 为均值 (mean), 或称期望 (expectation)

σ^2 为方差 (variance), σ 为标准差 (standard deviation)

随机变量 X 服从参数为 (μ,σ^2) 的正态分布 记为 $X\sim N(\mu,\sigma^2)$

当 $\mu=0$, $\sigma^2=1$, 称 X 服从标准正态分布 (standard normal distribution)

有 $f(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, $x\in\mathbb{R}$, 记为 $X\sim N(0,1)$



柏松分布 (Poisson distribution), 适用于描述单位时间内独立随机事件发生次数的离散概率空间

样本空间 S 为 $\mathbb{N}$, 事件集合 M 为 S 的幂集 $P(S)$

对于给定的参数 λ, λ 是一个非负的实数, 即 $\lambda\in\mathbb{R}^{\geq 0}$

则定义概率密度函数 $f(k)=P(\{k\})=e^{-\lambda}\frac{\lambda^k}{k!}$

其中 λ 通常表示单位时间内独立随机事件发生次数的均值

通常记为 $X\sim P(\lambda)$, 称随机变量 X 服从参数为 λ 的柏松分布

柏松小数定理 (Poisson limit theorem)

对于服从参数 n 和 p 的二项式分布 $X\sim B(n,p)$,

当 n 足够的大、而 p 足够的小, 使得 $\lambda=np$ 相对适中时,

$X\sim B(n,p)$ 会逼近于服从参数 $\lambda=np$ 的柏松分布 $Y\sim P(np)$

对于 $X=k$ $(k=0,1,2,\cdots)$, 有 $P(\{X=k\})=\binom{n}{k}p^k(1-p)^{n-k}$

假设当 $n\to+\infty$ 时, $p\to 0$, 使得 $\lambda=np$ 有 $\lim_{n\to+\infty}np=\lambda$ 为一个常数

$$\lim_{n\to+\infty}P(\{X=k\})=\lim_{n\to+\infty}\binom{n}{k}p^k(1-p)^{n-k}, \quad \text{取 } p=\lambda/n$$

$$=\lim_{n\to+\infty}\frac{n!}{k!(n-k)!}\cdot\left(\frac{\lambda}{n}\right)^k\cdot\left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$=\lim_{n\to+\infty}\frac{n!}{n^k(n-k)!}\cdot\left(\frac{\lambda^k}{k!}\right)\cdot\left(1-\frac{\lambda}{n}\right)^n\cdot\left(1-\frac{\lambda}{n}\right)^{-k}$$

又当 $n\to+\infty$ 时 $\frac{-\lambda}{n}\to 0$

$$=\left(\frac{\lambda^k}{k!}\right)\cdot\left[\lim_{n\to+\infty}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\cdot\lim_{n\to+\infty}\left(1+\frac{-\lambda}{n}\right)^{\frac{n}{-\lambda}\cdot(-\lambda)}\cdot\lim_{n\to+\infty}\left(1-\frac{\lambda}{n}\right)^{-k}\right]$$

$$=\left(\frac{\lambda^k}{k!}\right)\cdot\left[1\cdot e^{-\lambda}\cdot 1\right]$$

$$=e^{-\lambda}\frac{\lambda^k}{k!}=Q(\{Y=k\})$$

$X\sim P(\text{mean})$ 来近似地表示

即通常对于大量独立随机事件以小概率发生时, 如果观察到的均值稳定, 可以用柏松分布

Probability and Statistics – P6

极大似然估计(maximum likelihood estimation)，用于对类条件概率 $P(x|\theta)$ 进行估计

通常而言，对类条件概率的估计存在巨大困难

在于 1. 实际中可能没有充足的样本数据

2. 概率密度函数包含随机变量分布的全部信息

3. 特征向量 x 的维度过大

极大似然估计是将概率密度估计问题转换为参数估计问题的参数估计方法

即利用已知的样本结果，反推最大概率导致已知结果的未知参数值

前提　进行极大似然估计需要两个前提

首先是选择正确的概率密度函数、或者说正确描述随机变量分布的模型

如果选择的模型不准确，尽管有充足的样本数据

或者无法得到正确的参数估计，或者得到参数但仍无法描述随机变量分布

其次是样本必须满足服从独立同分布的随机变量

独立同分布 (independent, identically distributed, iid)

指一组随机变量样本中的每个变量都服从相同的概率分布，且相互独立

则有 $P(x_1, x_2, \cdots, x_n) = P(x_1)P(x_2)\cdots P(x_n)$

对于独立同分布的样本集 $D = \{x_1, x_2, \cdots, x_n\}$，估计参数 θ 的值

似然函数(likelihood function)，定义 $l(\theta)$ 为联合概率密度函数 $P(D|\theta)$ 相对于参数 θ 的函数

即 $l(\theta) = P(D|\theta) = P(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$

则求使该样本组的概率最大的参数值 $\hat{\theta}$

即 $\hat{\theta} = \arg\max_\theta l(\theta) = \arg\max_\theta \prod_{i=1}^{n} P(x_i|\theta)$

有时可定义对数似然函数 $H(\theta) = \ln l(\theta) = \sum_{i=1}^{n} \ln P(x_i|\theta)$

当似然函数连续且可微时，$\hat{\theta}$ 为使导数为0的点

即 $\frac{dl(\theta)}{d\theta} = 0$ 或 $\frac{dH(\theta)}{d\theta} = \frac{d\ln l(\theta)}{d\theta} = \frac{dl(\theta)}{d\theta} \cdot \frac{1}{l(\theta)} = 0$，即 $\frac{dH(\theta)}{d\theta} = \frac{dl(\theta)}{d\theta} = 0$

二项式分布的极大似然估计　如果有 n 次伯努利试验结果为 k 次成功与 n-k 次失败，则估计成功率 P

假设 $X \sim B(n, p)$，且样本 $\{x_1, x_2, \cdots, x_n\}$ 服从独立同分布

且 $x_1, x_2, \cdots, x_n$ 中有 k 次成功和 n-k 次失败，假设 $1<k<n$

则似然函数 $l(\theta) = \prod_{i=1}^{n} P(x_i|\theta) = \theta^k(1-\theta)^{n-k}$

则有 $l'(\theta) = k\theta^{k-1}(1-\theta)^{n-k} - \theta^k(n-k)(1-\theta)^{n-k-1} = 0$

$[k(1-\theta) - (n-k)\theta]\theta^{k-1}(1-\theta)^{n-k-1} = 0$

$k - n\theta = 0$，即 $\theta = \frac{k}{n}$

可知对成功率的极大似然估计为 k/n，

与直觉上的成功率为成功次数除以试验次数相同

Probability and Statistics - P7

正态分布的极大似然估计

对于n次服从正态分布的随机变量的样本 $\{x_1, x_2, \cdots, x_n\}$

估计参数 μ 和 σ^2 使得随机变量服从 $N(\mu, \sigma^2)$

假设 $\{x_1, x_2, \cdots, x_n\} \sim N(\mu, \sigma^2)$ 且 $x_1, \cdots, x_n$ 服从独立同分布

则可知 $f(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

于是有似然函数 $L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$

则取对数似然函数 $M(\mu, \sigma^2) = \ln L(\mu, \sigma^2)$

于是 $M(\mu, \sigma^2) = -\frac{n}{2} \ln 2\pi\sigma^2 - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$

$= -\frac{n}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2$

分别对参数 μ 和 σ^2 取 $M(\mu, \sigma^2)$ 的导数，并取导数为0

于是
$$\begin{cases} \frac{dM(\mu,\sigma^2)}{d\mu} = \frac{d - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \\ \frac{dM(\mu,\sigma^2)}{d\sigma^2} = [d - \frac{n}{2}\ln\sigma^2 + d - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2] / d\sigma^2 \\ \quad = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \end{cases}$$

则有
$$\begin{cases} \mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \\ \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = D(x) \end{cases}$$

于是可知对 μ 和 σ^2 的参数估计为 $\{x_1, \cdots, x_n\}$ 的期望与方差

多参数估计

如果极大似然估计中需要估计多个参数，即 $\hat{\theta} = [\theta_1, \theta_2, \cdots, \theta_n]^T$

则有梯度算子 $\nabla_\theta = [\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \cdots, \frac{\partial}{\partial\theta_n}]^T$

如果似然函数满足连续可导的条件

则有 $\nabla_\theta H(\hat{\theta}) = \nabla_\theta \ln L(\hat{\theta}) = \nabla_\theta \ln \prod_{i=1}^{N} P(x_i|\hat{\theta})$

$= \sum_{i=1}^{N} \nabla_\theta \ln P(x_i|\hat{\theta}) = 0$

对于概率密度函数不连续而无法求导的情形

如样本服从 $[a,b]$ 上的均匀分布，即 $f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases}$

则似然函数 $L(a,b) = \begin{cases} 0, & \text{othermise} \\ (\frac{1}{b-a})^n, & a \le x_i \le b, \ i = 1, 2, \cdots, n \end{cases}$

由于 $L(a,b)$ 是不连续，于是无法用求导的方式求解

则考虑似然函数 $L(a,b)$ 的定义

可知当 $b-a$ 尽可能小时，$L(a,b) = (\frac{1}{b-a})^n$ 尽可能大

如果 $b < \max(x_1, \cdots, x_n)$ 或 $a > \min(x_1, \cdots, x_n)$

则 $L(a,b) = 0$；

所以 $b \ge \max(x_1, \cdots, x_n)$ 且 $a \le \min(x_1, \cdots, x_n)$

与直观上服从均匀分布的样本最大值与最小值即分布边界相符

Probability and Statistics - P8

最小均方误差(minimum squared-error, MSE)

对于给定的n个序偶 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$

求出y的关于x的线性函数 $y=mx+b$,

使得平方误差 $[y_i-(mx_i+b)]^2$ 的和尽可能小

可以将这个问题视为一个最大似然估计的问题

即取 $L(m,b)=\sum_{i=0}^{n}[y_i-(mx_i+b)]^2$,并求使 $L(m,b)$ 取得最大值的参数 m,b

分别为m,b求偏导数,并令偏导数为0

$$\begin{cases}\frac{\partial L(m,b)}{\partial m}=\partial\sum_{i=0}^{n}[y_i-(mx_i+b)]^2/\partial m=0\\ \frac{\partial L(m,b)}{\partial b}=\partial\sum_{i=0}^{n}[y_i-(mx_i+b)]^2/\partial b=0\end{cases}$$

于是有 $\begin{cases}\sum_{i=0}^{n}-2[y_i-(mx_i+b)]\cdot x_i=0\\ \sum_{i=0}^{n}-2[y_i-(mx_i+b)]=0\end{cases}$

则 $\sum_{i=0}^{n}mx_i+b=\sum_{i=0}^{n}y_i$

$b=\frac{1}{n}[\sum_{i=0}^{n}y_i-m\sum_{i=0}^{n}x_i]=\bar{y}-m\bar{x}$

则 $\sum_{i=0}^{n}-2[y_i-(mx_i+\bar{y}-m\bar{x})]x_i=0$

$\sum_{i=0}^{n}[y_i-\bar{y}]x_i-m\sum_{i=0}^{n}(x_i-\bar{x})x_i=0$

$\sum_{i=0}^{n}x_iy_i-n\bar{x}\bar{y}=m(\sum_{i=0}^{n}x_i^2-n\bar{x}^2)$

$m=(\frac{1}{n}\sum_{i=0}^{n}x_iy_i-\bar{x}\bar{y})/(\frac{1}{n}\sum_{i=0}^{n}x_i^2-\bar{x}^2)$

于是可知参数 $\begin{cases}m=(\frac{1}{n}\sum_{i=0}^{n}x_iy_i-\bar{x}\bar{y})/(\frac{1}{n}\sum_{i=0}^{n}x_i^2-\bar{x}^2)\\ b=\bar{y}-m\bar{x}\end{cases}$

为使平方误差和 $\sum_{i=0}^{n}[y_i-(mx_i+b)]^2$ 最小的参数

无偏估计 (unbiased estimate),指用样本统计量来估计总体参数时的一种无偏判断

即如果被估计参数为θ,估计量为 $\hat{\theta}$,则估计量的数学期望 $E[\hat{\theta}]=\theta$

称估计量 $\hat{\theta}$ 为被估计参数θ的无偏估计,即具有无偏性(unbiasedness)

对于一个参数为实数θ的统计模型(statistical model)

记 $P_\theta(x)=P(x|\theta)$ 为基于概率的概率分布(probability distribution for observed data)

则统计量 $\hat{\theta}$ 为参数θ的估计量(estimator),基于观测到的数据x

则定义 $\hat{\theta}$ 与θ之间的偏差(bias)为

$$Bias_\theta[\hat{\theta}]=E_{x|\theta}[\hat{\theta}]-\theta=E_{x|\theta}[\hat{\theta}-\theta]$$

其中 $E_{x|\theta}$ 为分布 $P(x|\theta)$ 的期望值 the conditional distribution)

另外等式二成立是由于θ在条件分布 $P(x|\theta)$ 是可测的(mesurable with respect to

当偏差对于参数θ的所有值均为0时,称为无偏的(unbiased)

bias = 0 for all values of parameter θ

Probability and Statistics - P9

独立性 (independence)，又称统计独立(statistically independent)或随机独立(stochastically independent)

对于事件(event)而言，指一个事件的发生不影响另一个事件发生的概率

occurence of one doesn't affect the propability of occurence of the other

对于随机变量(random variable)，指一个变量的实现不影响另一个变量的概率分布

对于事件A，B：用 $A \perp B$ 或 $A \perp\!\!\!\perp B$ 表示A与B是相互独立的。

A和B是独立的当且仅当A与B的联合概率等于A与B概率之积

independent if and only if joint probability equals the product of probability

即 $P(A \cap B) = P(A)P(B)$

或者说，事件A发生的概率与给定事件B发生的条件概率相等。

即 $P(A) = P(A|B)$，与直观上事件A不受事件B影响相符

两者的关系为，$P(A|B) = P(A \cap B)/P(B)$

则如果A与B独立当且仅当 $P(A \cap B) = P(A)P(B)$

则同时有A与B独立当且仅当 $P(A|B) = P(A \cap B)/P(B) = P(A)$

对于随机变量X，Y，累积概率分布 $F_X(x) = P(\{X \le x\})$，$F_Y(y) = P(\{Y \le y\})$

X与Y是独立的当且仅当对于任意 x, y，$\{X \le x\}$ 与 $\{Y \le y\}$ 是独立事件

即联合累积概率分布 $F_{X,Y}(x,y) = F_X(x)F_Y(y)$] 对任意 x, y

等价地有联合概率密度 $f_{X,Y}(x,y) = f_X(x)f_Y(y)$

推导过程有，
$$
\begin{aligned}
F_{X,Y}(x,y) &= \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{y}\int_{-\infty}^{x} f_X(x)f_Y(y)\,dx\,dy \\
&= \int_{-\infty}^{y} f_Y(y)\,dy \int_{-\infty}^{x} f_X(x)\,dx = F_X(x)F_Y(x) \\
f_{X,Y}(x,y) &= \partial^2 F_{X,Y}(x,y)/\partial x\partial y = \partial^2 F_X(x)F_Y(y)/\partial x\partial y \\
&= f_X(x)\cdot \partial F_Y(y)/\partial y = f_X(x)\cdot f_Y(y)
\end{aligned}
$$

推广到多于两个事件或随机变量的情形

两两独立 (pairwise independent)，指对于一组事件或随机变量，其中每一对都是独立的

即对于事件 $A_1, A_2, \cdots, A_n$ 两两独立当且仅当 $\forall_{1 \le i < j \le n}\ P(A_i \cap A_j) = P(A_i)P(A_j)$

对于随机变量，$X_1, X_2, \cdots, X_n$，两两独立当且仅当 $\forall_{1 \le i < j \le n}\ F_{X_i}(x)F_{X_j}(y) = F_{X_i,X_j}(x,y)$

相互独立 (mutually independent)，指对于一组事件或随机变量，其中任一子集都是独立的

即对于事件 $A_1, A_2, \cdots, A_n$，相互独立当且仅当 $\forall_{\{B_1,\cdots,B_k\} \subseteq \{A_1,\cdots,A_n\}}\ P(\bigcap_{i=1}^{k} B_i) = \prod_{i=1}^{k} P(B_i)$

对于随机变量 $X_1, \cdots, X_n$，相互独立当且仅当 $F_{X_1,\cdots,X_n}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$

注意 相互独立蕴含了两两独立，但反之则不行

Probability and Statistics - P10

总体 (population)，指试验全部可能的观察值所构成的集合

其中每一个可能的观察值称为个体(individual)

总体中所包含的个体数量称为总体的容量(size of population)

分为有限总体和无限总体

注意当可能的观察值足够多从而无法列举的情况下，有限总体可视为无限总体

简单随机样本(simple random sample, SRS)，又简称样本(sample)

指对于 $X_1, X_2, \cdots, X_n$ 是服从分布函数F的独立同分布的随机变量

则称 $X_1, \cdots, X_n$ 为从分布函数F(或总体X)得到的容量为n的样本

其观察值(observed value) $x_1, x_2, \cdots, x_n$ 称为样本值(sample value)

又称X的n个独立的观测值(independent observations)

记 随机向量 $(X_1, X_2, \cdots, X_n)$ 为样本，对应样本值为 $(x_1, x_2, \cdots, x_n)$

通常认为相应于相同样本的样本值 $(x_1, x_2, \cdots, x_n)$ 和 $(y_1, y_2, \cdots, y_n)$ 是不同的

由于 $(X_1, \cdots, X_n)$ 为服从分布函数F的独立同分布随机变量，设概率密度函数为f

则 $(X_1, X_2, \cdots, X_n)$ 的分布函数为

$$F^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i)$$

而 $(X_1, X_2, \cdots, X_n)$ 的概率密度函数为

$$f^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i)$$

统计量 如果 $g(X_1, X_2, \cdots, X_n)$ 为 $X_1, \cdots, X_n$ 的函数且没有未知参数，则称为统计量

由于g是随机变量 $X_1, X_2, \cdots, X_n$ 的函数，所以统计量g本身也是随机变量

对于 $(X_1, X_2, \cdots, X_n)$ 的观测值 $(x_1, x_2, \cdots, x_n)$

称 $g(x_1, x_2, \cdots, x_n)$ 为 $g(X_1, X_2, \cdots, X_n)$ 的观测值

无偏估计(unbiased estimate)，指用样本统计量来估计总体参数时的一种无偏推断

即估计量的数学期望等于被估计参数的真实值

即对于总体中的参数 θ，$\hat{\theta}(X_1, X_2, \cdots, X_n)$ 为基于样本 $X_1, \cdots, X_n$ 的参数 θ 的估计量

称偏差 $b(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$.

当 $b(\hat{\theta}) = 0$，即 $E(\hat{\theta}) = \theta$ 时，称 $\hat{\theta}(X_1, \cdots, X_n)$ 为参数 θ 的无偏估计

直观意义为估计量 $\hat{\theta}$ 相对于真实值 θ 没有系统误差(systematic error)

而仅有随机误差(random error)

样本均值(sample mean)，记为 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，即字面意义的样本平均值

注意有 $E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n} E(\sum_{i=0}^{n} x_i) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n} \cdot n E(x_1) = E(x)$

于是可知样本均值 $\bar{X}$ 为总体均值 $E(x)$ 的无偏估计

Probability and Statistics -P11

期望 (expectation)，对于服从分布 $P\{X=x_k\}=p_k$，$k=1,2,\cdots$ 的离散随机变量 X

记 $E(X)=\sum_{k=1}^{\infty}x_kp_k$ 为随机变量 X 的期望

对于服从概率密度函数 $f(x)$，$x\in R$ 的连续随机变量 X

记 $E(X)=\int_{-\infty}^{+\infty}xf(x)dx$ 为随机变量 X 的期望

注意：对于离散随机变量，要求级数和 $\sum_{k=1}^{\infty}x_kp_k$ 绝对收敛

对于连续随机变量，要求反常积分 $\int_{-\infty}^{\infty}xf(x)dx$ 绝对收敛

定理 如果随机变量 Y 为随机变量 X 的函数，即有 $Y=g(X)$，且 $g(x)$ 为连续函数

如果 $\sum_{k=1}^{\infty}g(x_k)p_k$ 绝对收敛

则有 $E(Y)=E[g(X)]=\sum_{k=1}^{\infty}g(x_k)p_k$

如果 $\int_{-\infty}^{\infty}g(x)f(x)dx$ 绝对收敛

则有 $E(Y)=E[g(X)]=\int_{-\infty}^{\infty}g(x)f(x)dx$

性质 对常数 C，有 $E(C)=C$

证明过程有，$E(C)=\sum_{k=1}^{\infty}Cp_k$ / $\int_{-\infty}^{\infty}Cf(x)dx$

$=C\sum_{k=1}^{\infty}p_k$ / $C\int_{-\infty}^{\infty}f(x)dx=C$

对于常数 C 和随机变量 X，有 $E(CX)=CE(X)$

证明过程有，$E(CX)=\sum_{k=1}^{\infty}Cx_kp_k$ / $\int_{-\infty}^{\infty}Cxf(x)dx$

$=C\sum_{k=1}^{\infty}x_kp_k$ / $C\int_{-\infty}^{\infty}xf(x)dx=CE(X)$

对于随机变量 X,Y，有 $E(X+Y)=E(X)+E(Y)$

证明连续随机变量部分有，$E(X+Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)dxdy$

$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)dxdy+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)dxdy$

$=\int_{-\infty}^{\infty}xdx\int_{-\infty}^{\infty}f(x,y)dy+\int_{-\infty}^{\infty}ydy\int_{-\infty}^{\infty}f(x,y)dx$

$=\int_{-\infty}^{\infty}xf_X(x)dx+\int_{-\infty}^{\infty}yf_Y(y)dy=E(X)+E(Y)$

对于相互独立的随机变量 X,Y，有 $E(XY)=E(X)E(Y)$

证明连续随机变量部分有，$E(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf(x,y)dxdy$

又随机变量 X,Y 相互独立，即 $f(x,y)=f_X(x)f_Y(y)$

则 $E(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf_X(x)f_Y(y)dxdy$

$=[\int_{-\infty}^{\infty}xf_X(x)dx][\int_{-\infty}^{\infty}yf_Y(y)dy]$

$=E(X)E(Y)$