统计学复习

司继春

[1]上海对外经贸大学

2025年2月

概览

协方差

对于随机向量$X = [X_1, X_2]'$，如果$\mathbb{E}\left(X_1^2\right) < \infty, \mathbb{E}\left(X_2^2\right) < \infty$，根据Cauchy-Schwarz 不等式，

$$\mathbb{E}\left|X_1 X_2\right| \leq \sqrt{\mathbb{E}\left|X_1\right|^2 \mathbb{E}\left|X_2\right|^2} < \infty$$

即$X_1 X_2$可积，我们可以定义两个随机变量的**协方差（covariance）**：

$$\begin{aligned}
\mathbb{C}\left(X_1, X_2\right) &= \mathbb{E}\left[\left(X_1 - \mathbb{E}\left(X_1\right)\right)\left(X_2 - \mathbb{E}\left(X_2\right)\right)\right] \\
&= \mathbb{E}\left[X_1 X_2 - \mathbb{E}\left(X_1\right) X_2 - X_2 \mathbb{E}\left(X_1\right) + \mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right)\right] \\
&= \mathbb{E}\left(X_1 X_2\right) - 2\mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right) + \mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right) \\
&= \mathbb{E}\left(X_1 X_2\right) - \mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right)
\end{aligned}$$

当$X_2 = X_1$时，

$$\mathbb{C}\left(X_1, X_1\right) = \mathbb{E}\left(X_1^2\right) - \left[\mathbb{E}\left(X_1\right)\right]^2 = \mathbb{V}\left(X_1\right)$$

协方差的性质

协方差性质

$$\mathbb{C}\left(aX_1+b,cX_2+d\right)=ac\mathbb{C}\left(X_1,X_2\right)$$

Proof.

根据定义，有

$$\begin{aligned}\mathbb{C}\left(aX_1+b,cX_2+d\right)&=\mathbb{E}\left[\left(aX_1+b\right)\left(cX_2+d\right)\right]-\mathbb{E}\left(aX_1+b\right)\mathbb{E}\left(cX_2+d\right)\\&=\mathbb{E}\left(acX_1X_2+adX_1+bcX_2+bd\right)\\&\quad-ac\mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right)-ad\mathbb{E}\left(X_1\right)-bc\mathbb{E}\left(X_2\right)-bd\\&=ac\left[\mathbb{E}\left(X_1X_2\right)-\mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right)\right]\\&=ac\mathbb{C}\left(X_1,X_2\right)\end{aligned}$$

□

相关系数

进而可以使用协方差定义**简单相关系数**（correlation coefficient）或称**皮尔森相关系数**（Pearson correlation coefficient）：

$$\rho_{X_1,X_2} = \frac{\mathbb{C}(X_1, X_2)}{\sqrt{\mathbb{V}(X_1)\,\mathbb{V}(X_2)}}$$

由于

$$\begin{aligned}
\mathbb{C}(X_1, X_2) &= \mathbb{E}\left[\left(X_1 - \mathbb{E}(X_1)\right)\left(X_2 - \mathbb{E}(X_2)\right)\right] \\
&\leq \mathbb{E}\left|\left(X_1 - \mathbb{E}(X_1)\right)\left(X_2 - \mathbb{E}(X_2)\right)\right| \\
&\leq \sqrt{\mathbb{E}\left|X_1 - \mathbb{E}(X_1)\right|^2 \mathbb{E}\left|X_2 - \mathbb{E}(X_2)\right|^2} \\
&= \sqrt{\mathbb{V}(X_1)\,\mathbb{V}(X_2)}
\end{aligned}$$

可知$-1 \leq \rho_{X_1,X_2} \leq 1$。

相关系数

- 如果$\rho_{X_1,X_2}=\pm 1$，那么$P\left(X_2=c_1X_1+c_2\right)=1, c_1\neq 0$，此时$X_1$和$X_2$之间存在完美的线性关系；
- 如果$\rho_{X_1,X_2}>0$，我们称随机变量X_1和X_2正相关，反之称为负相关；
- 如果$\rho_{X_1,X_2}=0$，我们称随机变量X_1和X_2不相关（uncorrelated），记为$X_1\perp X_2$。

相关系数

注意这里的相关系数实际上只度量了随机变量之间的线性相关性。相关系数等于0并不意味着两个随机变量没有非线性的相关性

简单相关系数与非线性相关

如果随机变量$Y = Z^2$，$Z \sim \mathcal{N}(0,1)$，那么

$$\begin{aligned}\mathbb{C}(Z,Y) &= \mathbb{E}ZY - \mathbb{E}Z\mathbb{E}Y \\ &= \mathbb{E}Z^3 \\ &= 0\end{aligned}$$

两者相关系数为0，然而显然两者存在着非线性的函数关系。

协方差

和的协方差

如果a, b为任意实数，Y和Z为一元随机变量，那么：

$$\begin{aligned}\mathbb{V}(aX_1 + bX_2) &= \mathbb{E}(aX_1 + bX_2)^2 - [a\mathbb{E}(X_1) + b\mathbb{E}(X_2)]^2 \\ &= \mathbb{E}\left(a^2X_1^2 + b^2X_2^2 + 2abX_1X_2\right) \\ &\quad - \left[a^2(\mathbb{E}(X_1))^2 + b^2(\mathbb{E}(X_2))^2 + 2ab\mathbb{E}(X_1)\mathbb{E}(X_2)\right] \\ &= a^2\mathbb{V}(X_1) + b^2\mathbb{V}(X_2) + 2ab\mathbb{C}(X_1, X_2)\end{aligned}$$

如果$X_1 \perp X_2$，那么

$$\mathbb{V}(aX_1 + bX_2) = a^2\mathbb{V}(X_1) + b^2\mathbb{V}(X_2)$$

协方差矩阵

如果对于一个随机向量：$X = [X_1, X_2, ..., X_n]'$，我们可以定义矩阵：

$$\begin{aligned}\mathbb{V}(X) &= [\mathbb{C}(X_i, X_j)] \\ &= \begin{bmatrix} \mathbb{V}(X_1) & \mathbb{C}(X_1, X_2) & \cdots & \mathbb{C}(X_1, X_n) \\ \mathbb{C}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \mathbb{C}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}(X_n, X_1) & \mathbb{C}(X_n, X_2) & \cdots & \mathbb{V}(X_n) \end{bmatrix}\end{aligned}$$

为**协方差矩阵**（covariance matrix）。易知协方差矩阵为实对称矩阵。

协方差矩阵的计算

- 根据协方差矩阵的定义，协方差矩阵可以如下计算：

$$\mathbb{V}(X) = \mathbb{E}\left([X - \mathbb{E}(X)]\,[X - \mathbb{E}(X)]'\right)$$

注意X为列向量，从而：

$$X - \mathbb{E}(X) = \begin{bmatrix} X_1 - \mathbb{E}(X_1) \\ \vdots \\ X_n - \mathbb{E}(X_n) \end{bmatrix}$$

从而：$\mathbb{E}\,[X - \mathbb{E}(X)]\,[X - \mathbb{E}(X)]' =$

$$\mathbb{E}\begin{bmatrix} (X_1 - \mathbb{E}(X_1))^2 & (X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2)) & \cdots & (X_1 - \mathbb{E}(X_1))(X_n - \mathbb{E}(X_n)) \\ (X_2 - \mathbb{E}(X_2))(X_1 - \mathbb{E}(X_1)) & (X_2 - \mathbb{E}(X_2))^2 & \cdots & (X_2 - \mathbb{E}(X_2))(X_n - \mathbb{E}(X_n)) \\ \vdots & \vdots & \ddots & \vdots \\ (X_n - \mathbb{E}(X_n))(X_1 - \mathbb{E}(X_1)) & (X_n - \mathbb{E}(X_n))(X_2 - \mathbb{E}(X_2)) & \cdots & (X_n - \mathbb{E}(X_n))^2 \end{bmatrix}$$

协方差矩阵的计算

根据定义，有

$$\begin{aligned}\mathbb{V}(X) &= \mathbb{E}\left[(X-\mathbb{E}X)(X-\mathbb{E}X)'\right] \\ &= \mathbb{E}\left[XX' - X\mathbb{E}\left(X'\right) - \mathbb{E}(X)X' + \mathbb{E}(X)\mathbb{E}\left(X'\right)\right] \\ &= \mathbb{E}\left(XX'\right) - \mathbb{E}(X)\mathbb{E}\left(X'\right)\end{aligned}$$

协方差矩阵的性质

- 根据协方差矩阵的定义，对于任意的n维向量c，我们有：

$$\begin{aligned} c'\mathbb{V}(X)c &= c'\left[\mathbb{E}(X-\mathbb{E}X)(X-\mathbb{E}X)'\right]c \\ &= \mathbb{E}\left[c'(X-\mathbb{E}X)(X-\mathbb{E}X)'c\right] \\ &= \mathbb{E}\left\{\left[c'(X-\mathbb{E}X)\right]\left[c'(X-\mathbb{E}X)\right]'\right\} \\ &= \mathbb{E}\left[\left(\left[c'(X-\mathbb{E}X)\right]\right)^2\right] \geq 0 \end{aligned}$$

因而协方差矩阵是一个半正定矩阵，我们记为$\mathbb{V}(X) \succeq 0$。

协方差矩阵的性质

- 当X的分量之间存在完美的线性关系时，即存在一个向量a使得$a'X = \sum_{i=1}^{n} a_i X_i = 0$以概率1成立（$P(a'X = 0) = 1$），从而自然有$\mathbb{E}(a'X) = 0$，那么

$$a'\mathbb{V}(X)\,a = \mathbb{E}\left[\left(\left[a'(X - \mathbb{E}X)\right]\right)^2\right] = 0$$

  此时等号成立。
- 否则，如果X的分量之间不存在完美的线性关系，那么$\mathbb{V}(X)$为正定矩阵，记为$\mathbb{V}(X) \succ 0$。

协方差矩阵的性质

线性变换的协方差矩阵

对于一个实数矩阵$A_{h\times n}$，以及h维向量$b=[b_1,...,b_h]'$，有

$$\mathbb{V}(AX+b)=A\mathbb{V}(X)A'$$

Proof.

令$A_{h\times n}=[a_1,a_2,...,a_h]$以及$b=[b_1,...,b_h]'$，有：

$$\begin{aligned}
\mathbb{V}(AX+b)&=\mathbb{E}\left[(AX+b-\mathbb{E}(AX+b))(AX+b-\mathbb{E}(AX+b))'\right]\\
&=\mathbb{E}\left[(AX-\mathbb{E}(AX))(AX-\mathbb{E}(AX))'\right]\\
&=\mathbb{E}\left[(AX-A\mathbb{E}(X))\left(X'A'-\mathbb{E}\left(X'\right)A'\right)\right]\\
&=\mathbb{E}\left[AXX'A'-AX\mathbb{E}\left(X'\right)A'-A\mathbb{E}(X)X'A'+A\mathbb{E}(X)\mathbb{E}\left(X'\right)A'\right]\\
&=A\left[\mathbb{E}\left(XX'\right)-\mathbb{E}(X)\mathbb{E}\left(X'\right)\right]A'\\
&=A\mathbb{V}(X)A'
\end{aligned}$$

随机向量的独立性

随机向量的独立性

随机向量$[X_1, ..., X_n]'$各分量相互独立的充要条件是其联合分布函数等于边缘分布乘积:

$$F(x_1, ...x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n) = \prod_{i=1}^{n} P(X_i \leq x_i) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

如果密度（质量）函数存在，那么:

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

随机向量的独立性

独立性

两个联合分布函数:

$$F_{U,V}(u,v) = \min\{u,v\}, [u,v]' \in [0,1]\times[0,1]$$
$$\tilde{F}_{U,V}(u,v) = u\cdot v, [u,v]' \in [0,1]\times[0,1]$$

其边缘分布都为均匀分布，即$F_U(u) = u, F_V(v) = v$，然而由于:

$$F_{U,V}(u,v) = \min\{u,v\} \neq F_U(u)\cdot F_V(v)$$
$$\tilde{F}_{U,V}(u,v) = u\cdot v = F_U(u)\cdot F_V(v)$$

因而联合分布服从$F_{U,V}(u,v)$的随机变量不是相互独立的，而服从$\tilde{F}_{U,V}(u,v)$的随机变量是相互独立的。

随机变量函数的独立性

随机变量函数的独立性

$[X_1, ..., X_n]'$为一系列相互独立的随机变量，$1 \leq n_1 \leq n_2 \leq \cdots \leq n_k = n$，对于函数$f_1, f_2, ...f_k$，随机向量

$$\left[f_1\left(X_1, ...X_{n_1}\right), f_2\left(X_{n_1+1}, ...X_{n_2}\right), \cdots, f_k\left(X_{n_{k-1}+1}, ...X_{n_k}\right)\right]'$$

的分量也是相互独立的。

独立与期望

独立随机变量乘积的期望

如果随机向量$X=[X_1,X_2]'$的分量X_1和X_2相互独立且可积，那么

$$\mathbb{E}(X_1X_2)=\mathbb{E}(X_1)\mathbb{E}(X_2)$$

独立随机变量和的矩母函数

如果随机向量$X=[X_1,...,X_n]'$的分量相互独立，常数向量$a=[a_1,...,a_n]'$，记

$$S_n=a'X=\sum_{i=1}^{n}a_iX_i$$

那么矩母函数满足

$$M_{S_n}(t)=\prod_{i=1}^{n}M_{X_i}(a_it)$$

独立与不相关

独立与不相关

如果X_1和X_2相互独立且方差存在，那么$\mathbb{C}(X_1, X_2) = 0$。

不相关与独立

然而反过来，不相关并不意味着独立。

- 比如，$Y = Z^2, Z \sim \mathcal{N}(0,1)$
- Y和Z之间不相关，但是$Y = Z^2$存在完美的函数关系，自然不会是独立的。
- 实际上，如果$Y \perp\!\!\!\perp Z$，那么Y和Z的任意函数$g(Y)$和$h(Z)$之间都应该是独立的，即$g(Y) \perp\!\!\!\perp h(Z)$从而$\mathbb{C}(g(Y), h(Z)) = 0$
- 即如果$Y \perp\!\!\!\perp Z$，那么Y和Z的任意函数之间都应该是不相关的。

条件期望

- 令$[Y, X']' \in \mathbb{R}^{n+1}$为一个随机向量，如何使用随机向量$X$预测随机变量$Y$?
- 所谓预测，就是找到一个X的函数$h(X)$，使得其与Y之间的差异最小。
- 在统计中，我们称这类问题为**回归**（**regression**）。
- 比较常见的做法是最小化**均方误差**（**mean squared error**）：

$$\min_{h \in L^2} \left\{ \mathbb{E}\left[(Y - h(X))^2 \right] \right\}$$

其中

$$L^2 = \left\{ h | h : \mathbb{R} \to \mathbb{R}, \mathbb{E}\left[(h(X))^2 \right] < \infty \right\}$$

条件期望

- 定义误差项$\epsilon = Y - h_0(X)$，对于随机变量X的任意函数$g(X)$，我们有：

$$\mathbb{E}[\epsilon \cdot g(X)] = 0$$

- 如果令$\tilde{g}(X) = 1$，那么我们有

$$\mathbb{E}[\epsilon \cdot \tilde{g}(X)] = \mathbb{E}(\epsilon) = \mathbb{E}[Y - h_0(X)] = 0$$

因而

- $\mathbb{E}[\epsilon \cdot g(X)] = 0$意味着

$$\mathbb{C}(\epsilon, g(X)) = \mathbb{E}[\epsilon \cdot g(X)] - \mathbb{E}(\epsilon)\mathbb{E}[g(X)] = 0$$

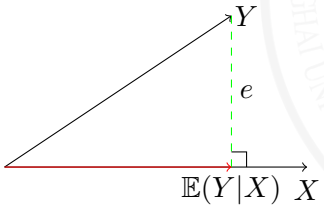
即$\epsilon \perp g(X)$。

条件期望

- 通过反证法证明，如果存在$g(X)$使得$\mathbb{E}[\epsilon \cdot g(X)] \neq 0$，那么我们令

$$h(X) = h_0(X) + \frac{\mathbb{E}[\epsilon g(X)]}{\mathbb{E}[g^2(X)]} g(X)$$

  根据这一构造，有：$\mathbb{E}\left[(Y - h(X))^2\right] < \mathbb{E}\left[(Y - h_0(X))^2\right]$，与$h_0(X)$最小化了均方误差矛盾。
- 我们称$h(X)$为Y在X上的**正交投影（orthogonal projection）**。



条件期望

- 我们知道，

$$\mathbb{E}(Y) = \arg\min_{c\in\mathbb{R}} \left\{ \mathbb{E}(Y-c)^2 \right\}$$

- 仿照上式，我们可以定义随机变量Y给定X的**条件期望**（conditional expectation）：

$$\mathbb{E}(Y|X) = h_0(X) = \arg\min_{h\in L^2} \left\{ \mathbb{E}\left[(Y - h(X))^2 \right] \right\}$$

因而随机变量Y给定X的条件期望$\mathbb{E}(Y|X)$是一个关于X的函数。

- $\mathbb{E}(\epsilon) = \mathbb{E}[Y - \mathbb{E}(Y|X)] = 0$，从而有**全期望定律**（law of total expectation）：

$$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y|X)]$$

注意到$\mathbb{E}(Y|X)$仅仅为X的函数，从而以上公式也可以写为

$$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y|X)] = \int_{\mathbb{R}} \mathbb{E}(Y|X)\, dF_X$$

条件期望与期望

- 期望可以看做是没有任何其他信息时的最优预测，即只用常数对Y进行预测，是条件期望的特例：

$$\mathbb{E}(Y|c) = \arg\min_{c\in\mathbb{R}} \left\{ \mathbb{E}\left[(Y-c)^2 \right] \right\}$$

- 这也就意味着：
  - 期望本身是对一个随机变量的最优预测
  - 具体的一个实现与期望之间的差异为误差项
  - 比如：
    - 如果全国所有人的平均体重为60公斤，那么随机从人群中选取一个人，对其体重的最优预测为60公斤
    - 单独每个人的体重与60公斤之间的差距为误差项

条件期望：离散情形

如果记$D \in \{0,1\}$代表性别，0代表女性、1代表男性，记Y为收入，那么根据以上全期望公式，有

$$\begin{aligned}\mathbb{E}\left[Y-h\left(D\right)\right]^2 &= p_0\mathbb{E}\left(\left[Y-h\left(D\right)\right]^2|D=0\right)+p_1\mathbb{E}\left(\left[Y-h\left(D\right)\right]^2|D=1\right)\\ &= p_0\mathbb{E}\left(\left[Y-h\left(0\right)\right]^2|D=0\right)+p_1\mathbb{E}\left(\left[Y-h\left(1\right)\right]^2|D=1\right)\end{aligned}$$

其中$p_d = P\left(D=d\right)$，从而最小化$\mathbb{E}\left[Y-h\left(D\right)\right]^2$等价于分别最小化$\mathbb{E}\left(\left[Y-h\left(0\right)\right]^2|D=0\right)$及$\mathbb{E}\left(\left[Y-h\left(1\right)\right]^2|D=1\right)$。

- 比如，全国所有男性的平均体重为70公斤，所有女性平均体重为50公斤
- 那么：

$$\begin{cases}\mathbb{E}\left(Y|D=1\right)=70\\ \mathbb{E}\left(Y|D=0\right)=50\end{cases}$$

即条件期望为分组期望。

条件期望：连续情形

或者，如果我们能看到一个变量X为连续型变量，那么

$$\mathbb{E}(Y|X) = \arg\min_{h \in L^2} \left\{ \mathbb{E}\left[(Y - h(X))^2 \right] \right\}$$

为一个未知的函数：

- 比如，如果我们现在可以观察到身高（X）
- 可以假想如果有无数个身高一样的人的平均体重，如：

$$\mathbb{E}(Y|X = 170)$$

即为条件期望。

条件期望的性质

条件期望的性质

对于任意的可测函数$g(X)$，条件期望有如下性质：

1. $\mathbb{E}[g(X)|X] = g(X)$；
2. $\mathbb{E}[(Y - \mathbb{E}(Y|X)) \cdot g(X)] = 0$；
3. $\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y|X)] = \int_{\mathbb{R}} \mathbb{E}(Y|X)\, dF_X$；
4. $\mathbb{E}[(g(X) \cdot Y)|X] = g(X) \cdot \mathbb{E}(Y|X)$；
5. $\mathbb{E}(aY_1 + bY_2|X) = a\mathbb{E}(Y_1|X) + b\mathbb{E}(Y_2|X)$。

条件期望的性质

银行到达人数

假设每天到达银行的人数服从泊松分布$N \sim \mathcal{P}(\lambda)$，而每个到达银行的人，办理外汇业务的概率为$p$。那么给定到达人数$N$，办理外汇业务的人数$M$服从二项分布，即$M|N \sim \mathcal{B}i(N,p), N \sim \mathcal{P}(\lambda)$。那么每天来银行办理外汇业务的人数的期望：

$$\mathbb{E}(M) = \mathbb{E}[\mathbb{E}(M|N)] = \mathbb{E}(Np) = p\mathbb{E}(N) = p\lambda$$

均值独立

- 注意到如果我们没有任何信息，因而只能用常数c去预测Y，那么

$$\mathbb{E}(Y|c) = c^* = \arg\min_{h \in L^2} \left\{ \mathbb{E}\left[(Y-c)^2 \right] \right\}$$

  - 即如果我们没有X，只能用常数预测Y，那么我们将得到Y的期望。
- 如果有其他随机变量X，但是$\mathbb{E}(Y|X) = \mathbb{E}(Y)$，那么$X$对$Y$的均值没有预测能力，退化成了一个常数而非$X$的函数，此时我们称$Y$对$X$是**均值独立**（**mean independence**）的。

均值独立

- 如果随机变量Y对X是均值独立的，那么：

$$\begin{aligned}\mathbb{C}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(\mathbb{E}(XY|X)) - \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(X\mathbb{E}(Y|X)) - \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(X\mathbb{E}(Y)) - \mathbb{E}(X)\mathbb{E}(Y) = 0\end{aligned}$$

  因而随机变量Y和X必然是不相关的。反之则不成立，不相关并不一定意味着均值独立。
- 实际上，可以证明，$\mathbb{C}(g(X),Y)=0$，即Y对X是均值独立的意味着Y与X的任意函数都不相关。
- 反过来不一定正确，即$\mathbb{C}(g(Y),X)=0$不一定成立
  - $Y=Z^2, Z\sim\mathcal{N}(0,1)$

条件方差

- 相应的，我们还可以定义**条件方差**

$$\mathbb{V}(Y|X) = \mathbb{E}\left[(Y - \mathbb{E}(Y|X))^2 |X\right]$$

- 根据条件期望的性质：

$$\begin{aligned}
\mathbb{V}(Y|X) &= \mathbb{E}\left[(Y - \mathbb{E}(Y|X))^2 |X\right] \\
&= \mathbb{E}\left\{\left[Y^2 + [\mathbb{E}(Y|X)]^2 - 2Y\mathbb{E}(Y|X)\right] |X\right\} \\
&= \mathbb{E}\left(Y^2|X\right) + \mathbb{E}\left\{[\mathbb{E}(Y|X)]^2 |X\right\} - 2\mathbb{E}\left[Y\mathbb{E}(Y|X)|X\right] \\
&= \mathbb{E}\left(Y^2|X\right) + [\mathbb{E}(Y|X)]^2 - 2\mathbb{E}(Y|X)\,\mathbb{E}[Y|X] \\
&= \mathbb{E}\left(Y^2|X\right) - [\mathbb{E}(Y|X)]^2
\end{aligned}$$

其中第4个等号由于$\mathbb{E}(Y|X)$也是X的函数

条件方差与方差

全方差定律（law of total variance）

$$\mathbb{V}(Y) = \mathbb{E}[\mathbb{V}(Y|X)] + \mathbb{V}[\mathbb{E}(Y|X)]$$

条件方差

银行到达人数的方差

在银行到达人数的例子中，可以计算每天办理外汇业务的人数的方差:

$$\mathbb{V}(M)=\mathbb{V}(\mathbb{E}(M|N))+\mathbb{E}(\mathbb{V}(M|N))$$

其中$\mathbb{E}(M|N)=Np$，因而

$$\mathbb{V}[\mathbb{E}(M|N)]=\mathbb{V}(Np)=p^2\mathbb{V}(N)=p^2\lambda$$

而$\mathbb{V}(M|N)=Np(1-p)$，从而

$$\mathbb{E}(\mathbb{V}(M|N))=\mathbb{E}(Np(1-p))=\lambda p(1-p)$$

从而

$$\mathbb{V}(M)=p^2\lambda+\lambda p-\lambda p^2=\lambda p$$

条件协方差

- 此外我们还可以定义**条件协方差**（conditional covariance）为

$$\mathbb{C}\left(Y_1, Y_2|X\right) = \mathbb{E}\left[\left(Y_1 - \mathbb{E}\left(Y_1|X\right)\right)\left(Y_2 - \mathbb{E}\left(Y_2|X\right)\right)|X\right]$$

- 同样根据条件期望的性质，有

$$\mathbb{C}\left(Y_1, Y_2|X\right) = \mathbb{E}\left(Y_1Y_2|X\right) - \mathbb{E}\left(Y_1|X\right)\mathbb{E}\left(Y_2|X\right)$$

以及**全协方差定律**（law of total covariance）：

$$\mathbb{C}\left(Y_1, Y_2\right) = \mathbb{E}\left[\mathbb{C}\left(Y_1, Y_2|X\right)\right] + \mathbb{C}\left[\mathbb{E}\left(Y_1|X\right), \mathbb{E}\left(Y_2|X\right)\right]$$

条件不相关

- 如果$\mathbb{C}\left(Y_1, Y_2|X\right) = 0$，可以称$Y_1$和$Y_2$给定$X$**条件不相关（conditionally uncorrelated）**，记为$Y_1 \perp Y_2|X$。
  - 条件不相关意味着在X相同的条件下，Y_1, Y_2之间是不相关的；
  - 然而即使$Y_1 \perp Y_2|X$，$\mathbb{C}\left[\mathbb{E}\left(Y_1|X\right), \mathbb{E}\left(Y_2|X\right)\right]$也可能不为0，从而条件不相关并不意味着不相关。

条件不相关

条件不相关与不相关

如果假设一个学习中同学的数学成绩$Y_1 = g(X) + Z_1$，物理成绩$Y_2 = f(X) + Z_2$，其中X为同学的努力程度，而$(Z_1, Z_2) \perp\!\!\!\perp X$为卷面成绩的随机干扰，且$Z_1 \perp Z_2$，那么

$$\begin{aligned}\mathbb{E}(Y_1 Y_2|X) &= \mathbb{E}[(g(X) + Z_1)(f(X) + Z_2)|X] \\ &= \mathbb{E}[g(X) f(X) + g(X) Z_2 + Z_1 f(X) + Z_1 Z_2|X] \\ &= g(X) f(X) + g(X)\mathbb{E}(Z_2) + f(X)\mathbb{E}(Z_1) + \mathbb{E}(Z_1 Z_2)\end{aligned}$$

同时

$$\begin{aligned}\mathbb{E}(Y_1|X)\mathbb{E}(Y_2|X) &= [g(X) + \mathbb{E}(Z_1|X)][f(X) + \mathbb{E}(Z_2|X)] \\ &= [g(X) + \mathbb{E}(Z_1)][f(X) + \mathbb{E}(Z_2)]\end{aligned}$$

条件不相关

条件不相关与不相关

从而

$$\begin{aligned}\mathbb{C}\left(Y_1, Y_2|X\right) &= \mathbb{E}\left(Y_1Y_2|X\right) - \mathbb{E}\left(Y_1|X\right)\mathbb{E}\left(Y_2|X\right)\\ &= \mathbb{E}\left(Z_1Z_2\right) - \mathbb{E}\left(Z_1\right)\mathbb{E}\left(Z_2\right)\\ &= 0\end{aligned}$$

注意

$$\begin{aligned}\mathbb{C}\left(Y_1, Y_2\right) &= \mathbb{E}\left[\mathbb{C}\left(Y_1, Y_2|X\right)\right] + \mathbb{C}\left[\mathbb{E}\left(Y_1|X\right), \mathbb{E}\left(Y_2|X\right)\right]\\ &= \mathbb{C}\left[g\left(X\right) + Z_1, f\left(X\right) + Z_2\right]\\ &= \mathbb{C}\left(g\left(X\right), f\left(X\right)\right)\end{aligned}$$

通常不为0。按照这个设定，数学和物理成绩的相关性完全是由共同的努力程度X导致的，然而如果给定一些同学的努力程度一样，其数学和物理成绩就不相关了。

条件协方差矩阵

条件协方差矩阵

令$Y=[Y_1,...,Y_K]'$为随机向量，条件协方差矩阵可以定义为

$$\begin{aligned}\mathbb{V}(Y|X) &= \mathbb{E}\left\{\mathbb{E}\left[Y-\mathbb{E}(Y|X)|X\right]\mathbb{E}\left[Y-\mathbb{E}(Y|X)|X\right]'\right\}\\ &= \mathbb{E}\left(YY'|X\right)-\mathbb{E}(Y|X)\left(Y'|X\right)\\ &= \begin{bmatrix} \mathbb{V}(Y_1|X) & \mathbb{C}(Y_1,Y_2|X) & \cdots & \mathbb{C}(Y_1,Y_n|X)\\ \mathbb{C}(Y_2,Y_1|X) & \mathbb{V}(Y_2|X) & \cdots & \mathbb{C}(Y_2,Y_n|X)\\ \vdots & \vdots & \ddots & \vdots\\ \mathbb{C}(Y_n,Y_1|X) & \mathbb{C}(Y_n,Y_2|X) & \cdots & \mathbb{V}(Y_n|X)\end{bmatrix}\end{aligned}$$

条件分布

- 如果对于随机向量$[Y, X']'$，我们记指示函数$\mathbb{1}_A(X) = \mathbb{1}\{X \in A\}$，这是一个随机变量$X$的函数，因而

$$\mathbb{E}(Y \cdot \mathbb{1}_A(X)) = \mathbb{E}[\mathbb{E}(Y\mathbb{1}_A(X)|X)] = \mathbb{E}[\mathbb{1}_A(X)\mathbb{E}(Y|X)]$$

- 若$X, Y \in \mathbb{Z}$是离散型随机变量，那么我们令$A = \{x\}$，有

$$\begin{aligned}\mathbb{E}(Y \cdot \mathbb{1}\{X = x\}) &= \mathbb{E}[\mathbb{1}\{X = x\}\mathbb{E}(Y|X)] \\ &= \mathbb{E}[\mathbb{1}\{X = x\}\mathbb{E}(Y|X = x)] \\ &= \mathbb{E}(Y|X = x) \cdot P(X = x)\end{aligned}$$

条件概率质量函数

- 从而

$$\begin{aligned}\mathbb{E}\left(Y|X=x\right)=h_0\left(x\right)&=\frac{\mathbb{E}\left(Y\cdot\mathbb{1}\left\{X=x\right\}\right)}{P\left(X=x\right)}\\&=\frac{\sum_{k=0}^{\infty}\left[k\cdot P\left(Y=k,X=x\right)\right]}{P\left(X=x\right)}\\&=\sum_{k=0}^{\infty}k\cdot\frac{P\left(Y=k,X=x\right)}{P\left(X=x\right)}\end{aligned}\tag{1}$$

条件概率质量函数

- 如果记

$$P(Y=k|X=x)=\frac{P(Y=k,X=x)}{P(X=x)}$$

可以验证$P(Y=k|X=x)\geq 0$且

$$\sum_{k=0}^{\infty}P(Y=k|X=x)=\sum_{k=0}^{\infty}\frac{P(Y=k,X=x)}{P(X=x)}=\frac{\sum_{k=0}^{\infty}P(Y=k,X=x)}{P(X=x)}=1$$

从而$P(Y=k|X=x)$是一个概率质量函数，我们称其为**条件概率质量函数**（conditional probability mass function）。

条件密度

- 对于连续型随机向量(X,Y)，可以证明

$$\mathbb{E}\left(Y|X=x\right)=h_0\left(x\right)=\frac{\int_{\mathbb{R}}yf\left(x,y\right)dy}{f_X\left(x\right)}=\int_{\mathbb{R}}y\frac{f\left(x,y\right)}{f_X\left(x\right)}dy \tag{2}$$

  - 由于$\mathbb{E}\left[Y-\mathbb{E}\left(Y|X\right)\right]=0$，从而

$$\int_{\mathbb{R}}\left[y-h_0\left(x\right)\right]f\left(x,y\right)dy=0$$

  - 固定x，那么以上条件意味着

$$\int_{\mathbb{R}}yf\left(x,y\right)dy=h_0\left(x\right)\int_{\mathbb{R}}f\left(x,y\right)dy=h_0\left(x\right)f_X\left(x\right)$$

条件密度

条件密度函数

- 如果记

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f(x,y)}{\int_{\mathbb{R}} f(x,y)\,dy}$$

- 可以验证$f_{Y|X}(y|x) > 0$且

$$\int_{\mathbb{R}} f_{Y|X}(y|x)\,dy = \int_{\mathbb{R}} \frac{f(x,y)}{\int_{\mathbb{R}} f(x,y)\,dy} dy = \frac{\int_{\mathbb{R}} f(x,y)\,dy}{\int_{\mathbb{R}} f(x,y)\,dy} = 1$$

从而$f_{Y|X}(y|x)$也是一个密度函数。

- 我们把$f_{Y|X}(y|x)$定义为**条件密度函数**（conditional denstiy function）。

条件密度函数

- 无论对于离散型随机变量还是连续型随机变量，或者他们的混合，不失一般性我们将条件概率质量函数和条件密度函数统称为条件密度函数$f_{Y|X}(y|x)$。
- 根据式(1)和式(2)，条件期望可以通过

$$\mathbb{E}(Y|X=x)=\int_{\mathbb{R}} y\cdot f_{Y|X}(y|x)\,dy$$

进行计算，即使用条件密度计算的期望即条件期望。

条件分布函数

条件分布函数

对于随机向量$[Y, X']'$，条件分布函数定义为

$$F_{Y|X}(y|x) = P(Y \leq y|X = x) = \mathbb{E}(\mathbb{1}\{Y \leq y\}|X = x)$$

- 如果给定$X = x$，Y的分布函数为$F_{Y|X}$，我们称Y给定X的条件分布为$F_{Y|X}$，记为$Y|X \sim F_{Y|X}$。
- 而对应于$F_{Y|X}$的密度（质量）函数就是条件密度（质量）函数。
- 比如，$M|N \sim \mathcal{B}i(N, p)$即条件分布：给定人数为$N$的情况下，$M$的分布为二项分布。
- 注意条件分布与无条件分布的区别，条件分布与无条件分布可以是完全不同的两个分布！

条件分布与无条件分布

银行人数的无条件分布

$M|N \sim \mathcal{B}i\,(N,p)$，$N \sim \mathcal{P}\,(\lambda)$，如果计算$M$的无条件分布，根据条件概率质量函数定义，有

$$\begin{aligned}
P\,(M=m) &= \sum_{n=m}^{\infty} P\,(M=m|N=n)\cdot P\,(N=n) \\
&= \sum_{n=m}^{\infty} \left(\begin{array}{c} n \\ m \end{array} \right) p^m\,(1-p)^{n-m}\,\frac{\lambda^n}{n!}e^{-\lambda} \\
&= \frac{p^m e^{-\lambda}}{m!} \sum_{n=m}^{\infty} \frac{n!}{(n-m)!}\,(1-p)^{n-m}\,\frac{\lambda^n}{n!} \\
&= \frac{p^m e^{-\lambda}}{m!} \sum_{h=0}^{\infty} (1-p)^h\,\frac{\lambda^{h+m}}{h!} \\
&= \frac{(\lambda p)^m\,e^{-\lambda}}{m!} \sum_{h=0}^{\infty} (1-p)^h\,\frac{\lambda^h}{h!}
\end{aligned}$$

条件分布与无条件分布

银行人数的无条件分布

根据$e^{\lambda x}$在$x=0$处的泰勒展开

$$e^{\lambda x}=\sum_{h=0}^{\infty}\frac{\lambda^h}{h!}x^h$$

将$x=1-p$带入，得到

$$e^{\lambda(1-p)}=\sum_{h=0}^{\infty}(1-p)^h\frac{\lambda^h}{h!}$$

带入得到

$$P(M=m)=\frac{(\lambda p)^m e^{-\lambda}}{m!}e^{\lambda(1-p)}=\frac{(\lambda p)^m}{m!}e^{-\lambda p}$$

从而无条件分布$M\sim\mathcal{P}(\lambda p)$。

独立与均值独立

- 如果随机变量X和Y是独立的，那么

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f_X(x) \cdot f_Y(y)}{f_X(x)} = f_Y(y)$$

即两个随机变量独立的充要条件是$f_{Y|X} = f_Y$。

- 在独立的条件下:

$$\mathbb{E}(Y|X) = \int_{\mathbb{R}} y \cdot f_{Y|X}(y|x)\, dy = \int_{\mathbb{R}} y \cdot f_Y(y)\, dy = \mathbb{E}(Y)$$

因而如果随机变量X和Y是独立的，那么其一定是均值独立的。

- 反之则不成立。比如$Y|X \sim \mathcal{N}(0, X^2)$，即使$\mathbb{E}(Y|X) = \mathbb{E}(Y) = 0$，但是$\mathbb{V}(Y|X) = X^2 \neq \mathbb{V}(Y)$，然而独立性一定要求$\mathbb{V}(Y|X) = \mathbb{V}(Y)$，从而均值独立推不出独立。

独立与均值独立

独立、均值独立、不相关的强弱关系

$$X \perp\!\!\!\perp Y \underset{\not\Leftarrow}{\Rightarrow} \mathbb{E}(Y|X) = \mathbb{E}(Y) \underset{\not\Leftarrow}{\Rightarrow} X \perp Y$$

条件密度函数

四面骰子

四面骰子的例子中，其条件密度可以如下计算：

$Z\backslash Y$	2	3	4	5	6	7	8
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0
4	0	0	0	0	0	0	$\frac{1}{16}$
$f_{Y\|Z}(y\|Z=1)$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{2}{7}$	$\frac{2}{7}$	0	0	
$f_{Y\|Z}(y\|Z=2)$	0	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	0	

条件密度函数

二元正态分布

对于联合正态

$$(X,Y)' \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$$

密度函数:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right\}$$

其中$-1 < \rho < 1$为X, Y的相关系数。

条件密度函数

二元正态分布

其边际密度函数为:

$$
\begin{aligned}
f_X(x) &= \int_{\mathbb{R}} f_{X,Y}(x,y)\,dy \\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}} \\
&\cdot \int_{\mathbb{R}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(1-\rho^2)(x-\mu_X)^2}{\sigma_X^2}+\left(\frac{y-\mu_Y}{\sigma_Y}-\frac{\rho(x-\mu_X)}{\sigma_X}\right)^2\right]\right\} dy \\
&= \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left\{-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\}
\end{aligned}
$$

或者$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$

条件密度函数

二元正态分布

其条件密度函数为:

$$
\begin{aligned}
f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\
&= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)}{\sigma_Y\sqrt{(1-\rho^2)}}\right)^2\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2}\left(\frac{y-\left[\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]}{\sigma_Y\sqrt{(1-\rho^2)}}\right)^2\right\}
\end{aligned}
$$

或者$Y|X \sim \mathcal{N}\left(\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), \sigma_Y^2(1-\rho^2)\right)$，也服从正态分布，进而:

- 条件期望$\mathbb{E}(Y|X=x) = \mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$
- 条件方差$\mathbb{V}(Y|X) = \sigma_Y^2(1-\rho^2)$。

贝叶斯公式

- 使用条件密度函数的定义，我们还可以得到随机变量的贝叶斯公式。
- 由于：$f(x,y)=f_X(x)\cdot f_{Y|X}(y|x)=f_Y(y)\cdot f_{X|Y}(x|y)$从而条件密度：

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y)\cdot f_Y(y)}{\int_{\mathbb{R}} f(x,y)\,dy} \\ &= \frac{f_{X|Y}(x|y)\cdot f_Y(y)}{\int_{\mathbb{R}} f_{X|Y}(x|y)\cdot f_Y(y)\,dy} \end{aligned}$$

以上方程即随机变量的**贝叶斯公式**，在贝叶斯统计中有大量的应用。

收敛的概念

收敛的定义

若$\{a_n, n = 1, 2, ...\}$为实数序列，如果对于任意的$\epsilon > 0$，存在$n_0 = n_0(\epsilon)$使得：

$$|a_n - a| < \epsilon, \forall n > n_0$$

那么我们称数列$\{a_n\}$ 的极限为a，或者$\{a_n\}$ 收敛到（converges to）a，记为

$$\lim_{n\to\infty} a_n = a$$

或者

$$a_n \to a \ as \ n \to \infty$$

收敛的概念

有界的定义

若$\{a_n, n = 1, 2, ...\}$为实数序列，如果存在常数$b < \infty$，使得

$$|a_n| < b$$

那么我们称数列$\{a_n\}$为有界的（bounded），否则称之为无界的（unbounded）。

小o符号

小o符号的定义

对于两个序列$\{a_n\},\{b_n\}$，如果随着$n \to \infty$，有：

$$\frac{a_n}{b_n} \to 0$$

那么我们记为$a_n = o(b_n)$。特别的，如果令$b_n = 1$，那么$a_n = o(1)$等价为$a_n \to 0$。

小o符号

小o实例

如果$a_n = \frac{1}{n^2}, b_n = \frac{1}{n}$，那么:

$$\frac{a_n}{b_n} = \frac{\frac{1}{n^2}}{\frac{1}{n}} = \frac{1}{n} \to 0$$

因而$\frac{1}{n^2} = o\left(\frac{1}{n}\right)$，即$\frac{1}{n^2}$以更快的速度收敛到0。 如果两个序列$a_n \to 0, b_n \to 0$, 且$a_n = o\left(b_n\right)$，那么我们称$a_n$为比$b_n$高阶的无穷小量。

小o符号

小o符号经常用来对一个复杂的式子进行化简，通过将无穷小量舍掉从而减少了计算量。比如：

小o的应用

假设有两个数列，$a_n = \frac{1}{n} + \frac{6}{n^2} - \frac{8}{n^3}$而另外一个序列：$b_n = \frac{1}{n}$如果定义$R_n = \frac{6}{n^2} - \frac{8}{n^3}$，显然$R_n = o\left(\frac{1}{n}\right)$，因而$a_n = b_n + o\left(\frac{1}{n}\right)$，即：

$$\frac{a_n}{b_n} = \frac{b_n + o\left(\frac{1}{n}\right)}{b_n} \to 1$$

因而尽管两个序列a_n和b_n并不相等，但是当$n \to \infty$时，两者误差趋向于0，因而我们可以舍去无穷小量R_n，使用更简单的序列b_n去逼近a_n。

小o符号的应用：泰勒展开

当$x \to a$时，$(x-a) = o(1)$，同时我们有$(x-a)^{k+1} = o\left((x-a)^k\right)$，即当$x \to a$时，$(x-a)$的高阶幂是低阶幂的无穷小量。对于一个单变量实值函数$f(x)$且k阶可微，那么有：

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + o\left(|x-a|^k\right)$$

因而对于一个难以计算的函数f，我们经常使用其前k阶泰勒多项式对其进行逼近。

泰勒展开

泰勒展开

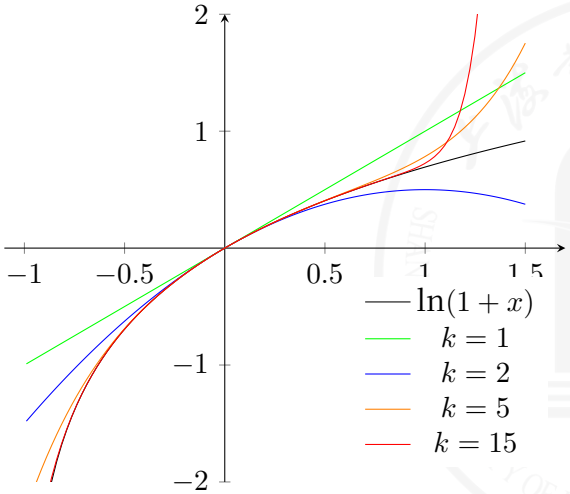
函数$f(x)=\ln(1+x)$在$x=0$处的泰勒展开为：

$$f(x)=\sum_{n=1}^{\infty}(-1)^{n+1}\frac{x^n}{n}=x-\frac{x^2}{2}+\frac{x^3}{3}-\cdots$$

因而当x充分靠近0时，我们可以使用前k阶泰勒展开对其进行逼近。特别的，如果令

$$k=1,\quad \ln(1+x)=x+o(x)\approx x$$

泰勒展开



多元函数的泰勒展开

更一般的，如果$f:\mathbb{R}^n\to\mathbb{R}$为多元实值函数，那么其泰勒级数为：

$$f(x)=f(a)+\frac{\partial f}{\partial x'}(a)(x-a)+\frac{1}{2!}(x-a)'\frac{\partial^2 f}{\partial x\partial x'}(a)(x-a)+o\left(\|x-a\|_2^2\right)$$

其中x和a为$n\times 1$向量。

多元函数的泰勒展开

多元函数泰勒展开示例

令$f(x)=e^{x_1}\ln(1+x_2)$，其中$x=(x_1,x_2)'$。那么：

$$\frac{\partial f}{\partial x}=\begin{bmatrix} e^{x_1}\ln(1+x_2) \\ \frac{e^{x_1}}{1+x_2} \end{bmatrix},\frac{\partial^2 f}{\partial x\partial x'}=\begin{bmatrix} e^{x_1}\ln(1+x_2) & \frac{e^{x_1}}{1+x_2} \\ \frac{e^{x_1}}{1+x_2} & -\frac{e^{x_1}}{(1+x_2)^2} \end{bmatrix}$$

因而其在$a=(0,0)'$处的二阶泰勒展开：

$$\begin{aligned} p_2(x)&=[0,1]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}+\frac{1}{2}[x_1,x_2]\begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &=x_2+\frac{1}{2}\left(2x_1x_2-x_2^2\right) \end{aligned}$$

小o符号

小o的性质

1. 若$a_n = o(b_n), b_n = o(c_n)$，那么$a_n = o(c_n)$
2. 对于任意的常数$c \neq 0$，及$a_n = o(b_n)$，有$ca_n = o(b_n)$
3. 对于任意的数列$c_n \neq 0$，及$a_n = o(b_n)$，有$c_n a_n = o(c_n b_n)$
4. 如果$d_n = o(b_n), e_n = o(c_n)$，那么$d_n e_n = o(b_n c_n)$
5. 如果$a_n, b_n > 0, c_n, d_n > 0$，$a_n = o(b_n), c_n = o(d_n)$，那么$a_n + c_n = o(b_n + d_n)$。

大O符号

大O符号的定义

对于两个序列$\{a_n\},\{b_n\}$，如果随着$n \to \infty$，$\left|\frac{a_n}{b_n}\right|$是有界的，即存在一个M使得:

$$\left|\frac{a_n}{b_n}\right| < M$$

那么我们记为$a_n = O(b_n)$。特别的，如果令$b_n = 1$，那么$a_n = O(1)$等价为a_n是有界的。

同阶的定义

对于两个序列$\{a_n\},\{b_n\}$，如果$a_n = O(b_n)$，且同时$b_n = O(a_n)$，那么我们称两个序列是同阶的，简记为$a_n \asymp b_n$。

大O符号

大O符号示例

对于序列$a_n = \frac{1}{n} + \frac{b}{n\sqrt{n}} + \frac{c}{n^2} + \frac{d}{n^2\sqrt{n}}$，同时定义$R_n = \frac{b}{n\sqrt{n}} + \frac{c}{n^2} + \frac{d}{n^2\sqrt{n}}$那么:

1. $a_n \sim \frac{1}{n}$
2. 若$b = 0$，$R_n = O\left(\frac{1}{n^2}\right)$
3. 若$b = 0$，$R_n \asymp \frac{1}{n^2}$
4. 若$b \neq 0$，$R_n \sim \frac{b}{n\sqrt{n}}$
5. 若$b = c = 0$，$R_n = o\left(\frac{1}{n^2}\right)$

大O符号

大O符号的性质

1. 若$a_n = O(b_n), b_n = O(c_n)$，那么$a_n = O(c_n)$
2. 对于任意的常数$c \neq 0$，及$a_n = O(b_n)$，有$ca_n = O(b_n)$
3. 对于任意的数列$c_n \neq 0$，及$a_n = O(b_n)$，有$c_n a_n = O(c_n b_n)$
4. 如果$d_n = O(b_n), e_n = O(c_n)$，那么$d_n e_n = O(b_n c_n)$
5. 如果$a_n = o(b_n)$，$c_n = O(b_n)$，那么$a_n c_n = o(b_n)$
6. 如果$a_n = o(b_n)$，$c_n = O(b_n)$，那么$a_n + c_n = O(b_n)$

依概率收敛

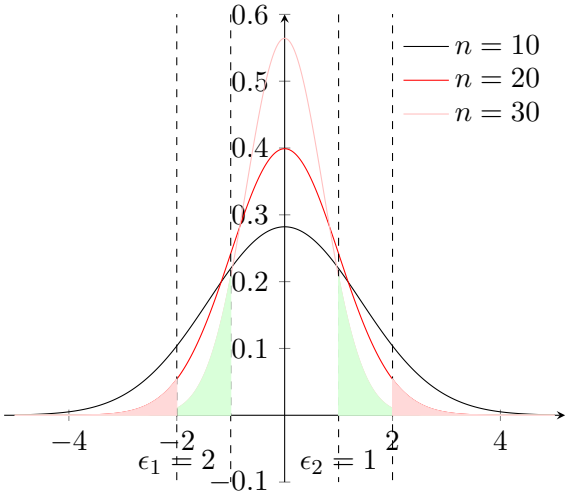
依概率收敛

如果对于任意的$\epsilon > 0$，概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的一系列随机变量序列$\{X_n\}$ 满足:

$$P\left(|X_n - X| > \epsilon\right) \to 0$$

那么我们称X_n依概率收敛于X，记为$X_n \xrightarrow{p} X$，或$\operatorname{plim} X_n = X$。

依概率收敛



o_p符号

o_p符号

$\{X_n\}$与$\{Y_n\}$为定义在概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的两个随机变量序列，如果

$$\frac{X_n}{Y_n} \xrightarrow{p} 0$$

那么我们记为$X_n = o_p(Y_n)$。特别的，当$Y_n = 1$时，即$X_n = o_p(1)$，等价于$X_n \xrightarrow{p} 0$。

依概率收敛

简单的大数定律

现在假设有一组独立的样本$x_i, i = 1, 2, ..., N$，其方差有界：$\mathbb{V}(x_i) < M$，其样本均值为$\bar{x}_N$。根据切比雪夫不等式，随着样本量$N \to \infty$，有：

$$\begin{aligned} P\left(\left|\bar{x}_N - \mathbb{E}(x_i)\right| > \epsilon\right) &\leq \frac{\mathbb{E}\left[\left(\bar{x}_N - \mathbb{E}(x_i)\right)^2\right]}{\epsilon^2} \\ &= \frac{\mathbb{V}(\bar{x}_N)}{\epsilon^2} = \frac{\frac{\mathbb{V}(x_i)}{N}}{\epsilon^2} \\ &< \frac{M}{N\epsilon^2} \to 0 \end{aligned}$$

从而$\bar{x}_N \xrightarrow{p} \mathbb{E}(x_i)$，或者$\bar{x}_N - \mathbb{E}(x_i) = o_p(1)$，或者$\bar{x}_N = \mathbb{E}(x_i) + o_p(1)$。

O_p符号

- 类似的，我们还可以定义大O_p符号。
- 与数列有界不同的是，随机变量的取值范围可能是$\mathbb{R}$，从而一定要求随机变量取值在一个有限的范围内是不可能的。
  - 比如，如果$X \sim \mathcal{N}(0,1)$，虽然X的分布范围看起来很小，但是$\text{supp}(X) = \mathbb{R}$意味着其“理论上”的取值范围仍然是$(-\infty, \infty)$
  - 所以根据$|X_n| < M$这样定义有界显然是不合适的。

O_p符号

- 对于一个随机变量X，如果对于任意的$\epsilon > 0$，都能找到一个C_ϵ使得$P(|X| > C_\epsilon) < \epsilon$，那么我们就可以认为随机变量$X$是有界的。
- 根据这一定义，我们可以将其推广到随机变量序列上：

O_p符号

$\{X_n\}$与$\{Y_n\}$为定义在概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的两个随机变量序列，如果对于任意的$\epsilon > 0$，存在一个$C_\epsilon$使得：

$$\sup_n P(|X_n| \geq C_\epsilon |Y_n|) < \epsilon$$

那么我们记$X_n = O_p(Y_n)$。特别的，当$Y_n = 1$时，我们称X_n**依概率有界（bounded in probability）**。

O_p符号

- O_p符号是对O符号的推广，当X_n退化为确定性序列时，O_p符号就变成了O符号。
- 相对于随机变量的有界，随机变量序列的依概率有界需要对于所有的$\{X_n\}$找到一个共同的C_ϵ:
  - 如果$X_n = O_p(1)$，那么意味着$|X_n|$不能太大，我们总是可以以很高的概率（$1-\epsilon$）保证对于任意的n，都有$|X_n| < C_\epsilon$成立即可。
- 如果$\{X_n\}$的期望趋向于正无穷，那么自然不是依概率有界的。

O_p符号

- 注意如果$\mathbb{E}\left(X_n^2\right) < M$，对于任意的$\epsilon > 0$，取$C_\epsilon = \sqrt{M/\epsilon + 1}$，那么：

$$P\left(|X_n| \geq C_\epsilon\right) \leq \frac{\mathbb{E}\left(X_n^2\right)}{C_\epsilon^2} = \frac{\mathbb{E}\left(X_n^2\right)}{M/\epsilon + 1} < \epsilon$$

  因而$X_n = O_p\left(1\right)$。
- 如果$\mathbb{E}\left|X_n\right| < M_0$，那么$\mathbb{E}\left(X_n^2\right) < M$意味着$\mathbb{V}\left(X_n\right) = \mathbb{E}\left(X_n^2\right) - \mathbb{E}\left(X_n^2\right) < M + M_0$从而方差有界
- 从而$\{X_n\}$的期望、方差同时有界意味着$X_n = O_p\left(1\right)$。
- 实际上，只需要$\mathbb{E}\left|X_n\right| < M$即可（why?）。

依概率有界

样本均值的阶数

现在假设有一组独立同分布的样本$x_i, i = 1, 2, ..., N$，其方差有界：$\mathbb{V}(x_i) < M$，其样本均值为$\bar{x}_N$。可以计算：

$$\mathbb{V}(\bar{x}_N) = \frac{\mathbb{V}(x_i)}{N}$$

因而$\bar{x}_N$的方差是有界的，$\mathbb{V}(\bar{x}_N) = O_p(1)$。此外：

$$\mathbb{V}\left(\sqrt{N}\bar{x}_N\right) = \mathbb{V}(x_i)$$

因而$\sqrt{N}\bar{x}_N = O_p(1)$。但是，如果我们考虑样本和：$S_N = \sum_{i=1}^{N} x_i$由于$\mathbb{V}(S_N) = N\mathbb{V}(x_i)$，该方差是无界的，因而不是$O_p(1)$。

O_p和o_p符号的性质

O_p和o_p的性质

如果$X_n = o_p(1), Y_n = o_p(1), Z_n = O_p(1), W_n = O_p(1)$，那么：

1. $X_n + Y_n = o_p(1)$
2. $X_n + Z_n = O_p(1)$
3. $Z_n + W_n = O_p(1)$
4. $X_n Y_n = o_p(1)$
5. $X_n Z_n = o_p(1)$
6. $Z_n W_n = O_p(1)$

均方收敛

均方收敛的定义

如果概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的一系列随机变量序列$\{X_n\}$随着$n \to \infty$满足：

$$\mathbb{E}\left(|X_n - X|^2\right) \to 0$$

那么我们称X_n均方收敛于X，记为$X_n \xrightarrow{L^2} X$。

均方收敛

均方收敛与依概率收敛

如果随机变量序列$X_n \xrightarrow{L^2} X$，那么$X_n \xrightarrow{p} X$。

Proof.

据切比雪夫不等式，对于任意的$\epsilon > 0$，有：

$$P\left(|X_n - X| > \epsilon\right) \leq \frac{\mathbb{E}\left(|X_n - X|^2\right)}{\epsilon^2} \to 0$$

□

依分布收敛

依分布收敛的定义

令F_n,F为分布函数，如果对于每一个$F(x)$连续的点x，有：

$$\lim_{n\to\infty} F_n(x) = F(x)$$

那么我们称$F_n(x)$弱收敛于$F(x)$，记为$F_n \xrightarrow{w} F$。

如果一系列随机变量$\{X_n\}$的分布函数$F_{X_n}(x) \xrightarrow{w} F_X$，我们称$X_n$依分布收敛于$X$，记为$X_n \xrightarrow{D} X$或者：$X_n \overset{a}{\sim} F$，其中$a$代表渐近的（asymptotically），即$X_n$渐近服从分布函数为$F$的分布。

依分布收敛

依分布收敛与O_p的关系

1. 如果$X_n \neq O_p(1)$，那么分布函数极限$\lim_{n\to\infty} F_n(x) = F(x)$不是一个分布函数。
2. 如果X_n依分布收敛，那么$X_n = O_p(1)$。

因而当我们讨论依分布收敛时，一定要保证我们讨论的$X_n = O_p(1)$。

大数定律

大数定律（Law of Large Numbers, LLN）讨论样本均值的极限，即在何种条件下，以下结论：

$$\frac{S_N - \mathbb{E}(S_N)}{N} = \bar{x} - \mathbb{E}(\bar{x}) \xrightarrow{p} 0$$

成立，其中：

$$S_N = \sum_{i=1}^{N} x_i$$

大数定律

实际上，$S_N - \mathbb{E}(S_N) = \sum_{i=1}^{N} [x_i - \mathbb{E}(x_i)]$，因而：

$$
\begin{aligned}
\mathbb{E}[S_N - \mathbb{E}(S_N)]^2 &= \mathbb{E}\left(\sum_{i=1}^{N} [x_i - \mathbb{E}(x_i)]\right)^2 \\
&= \mathbb{E}\left(\sum_{i=1}^{N} [x_i - \mathbb{E}(x_i)]^2\right) \\
&+ \mathbb{E}\left(2 \sum_{1 \le j < i \le N}^{N} [x_i - \mathbb{E}(x_i)][x_j - \mathbb{E}(x_j)]\right) \\
&= \sum_{i=1}^{N} \mathbb{V}(x_i) + 2 \sum_{1 \le j < i \le N}^{N} \mathbb{C}(x_i, x_j)
\end{aligned}
$$

大数定律

$$\mathbb{E}\left[S_N - \mathbb{E}\left(S_N\right)\right]^2 = \sum_{i=1}^{N} \mathbb{V}\left(x_i\right) + 2 \sum_{1 \le j < i \le N}^{N} \mathbb{C}\left(x_i, x_j\right)$$

如果我们假设$\mathbb{V}\left(x_i\right) < M$，那么：

1. $\sum_{i=1}^{N} \mathbb{V}\left(x_i\right) < NM = O\left(N\right)$
2. 而根据Cauthy-Schwartz不等式，$\mathbb{C}\left(x_i, x_j\right) \le \sqrt{\mathbb{V}\left(x_i\right)\mathbb{V}\left(x_j\right)} \le M$，而上式中有$\frac{N(N-1)}{2}$个协方差，所以

$$\sum_{1 \le j < i \le N}^{n} \mathbb{C}\left(x_i, x_j\right) = O\left(N^2\right)$$

大数定律

$$\mathbb{E}\left[S_N - \mathbb{E}\left(S_N\right)\right]^2 = O\left(N\right) + O\left(N^2\right)$$

从而:

$$\mathbb{E}\left[\frac{S_N - \mathbb{E}\left(S_N\right)}{N}\right]^2 = \frac{1}{N^2}O\left(N\right) + \frac{1}{N^2}O\left(N^2\right) = o\left(1\right) + O\left(1\right)$$

根据均方收敛定义，如果希望$\frac{S_N}{N} - \frac{\mathbb{E}(S_N)}{N} \xrightarrow{L^2} 0$，必须

$$\mathbb{E}\left[\frac{S_N - \mathbb{E}\left(S_N\right)}{N}\right]^2 = o\left(1\right)$$

大数定律

$$\mathbb{E}\left[\frac{S_N - \mathbb{E}(S_N)}{N}\right]^2 = o(1) + O(1)$$

其中的$O(1)$来源于$\frac{N(N-1)}{2}$个协方差，如果$\mathbb{C}(x_i, x_j) = 0$，那么自然有：

$$\mathbb{E}\left[\frac{S_N - \mathbb{E}(S_N)}{N}\right]^2 = o(1)$$

从而

$$\frac{S_N}{N} - \frac{\mathbb{E}(S_N)}{N} \xrightarrow{L^2} 0 \Rightarrow \frac{S_N}{N} - \frac{\mathbb{E}(S_N)}{N} \xrightarrow{p} 0$$

大数定律

大数定律1

如果概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的一个随机变量序列$\{x_i\}$两两不相关，且存在一个$M$使得对于所有的$i = 1, 2, ...$，都有$\mathbb{V}(x_i) < M$，那么：

$$\bar{x} - \mathbb{E}(\bar{x}) = \frac{S_N - \mathbb{E}(S_N)}{N} \xrightarrow{L^2} 0$$

从而

$$\bar{x} - \mathbb{E}(\bar{x}) = \frac{S_N - \mathbb{E}(S_N)}{N} \xrightarrow{p} 0$$

如果额外假设$\{x_i\}$是同分布的，那么$\mathbb{E}(\bar{x}) = \mathbb{E}(x_i) = \mu$，从而：

$$\bar{x} \xrightarrow{p} \mu$$

大数定律

以下的定理放松了二阶矩有限的假定以及独立的假定，保留了独立同分布的假定：

大数定律2

令$\{x_i\}$为概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的两两独立且同分布的随机变量序列，若

$$\mathbb{E}\,|x_i| < \infty$$

那么

$$S_N/N = \bar{x} \xrightarrow{p} \mu$$

其中$\mu = \mathbb{E}\,(x_i)$。

大数定律

以下的定理则同时放宽了同分布的假定以及二阶矩的假定。

大数定律3

令$\{x_i\}$为概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的相互独立的随机变量序列，如果存在一个常数$p \in [1, 2]$，随着$N \to \infty$，使得：

$$\frac{1}{N^p} \sum_{i=1}^{N} \mathbb{E} |x_i|^p \to 0$$

那么$S_N/N = \bar{x} \xrightarrow{p} \mu$，其中

$$\mu = \lim_{N \to \infty} \frac{\sum_{i=1}^{N} \mathbb{E}(x_i)}{N}$$

大数定律

大数定律的应用

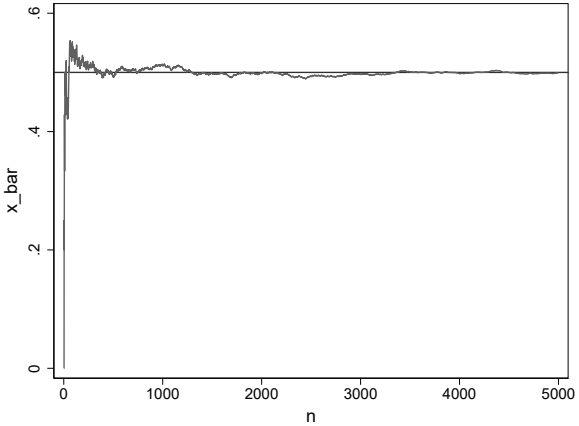
如果令$\{x_i\}$为一系列i.i.d的随机变量，且$x_i \sim \mathcal{B}er(p)$，那么$\mathbb{E}(x_i) = p, \mathbb{V}(x_i) = p(1-p) < \infty$，定义：

$$\hat{p} = \frac{\sum_{i=1}^{N} x_i}{N}$$

即成功的比例，那么根据上例，可以得到

$$\hat{p} \xrightarrow{p} p$$

大数定律



中心极限定理

中心极限定理（Central Limit Theorem, CLT）讨论样本均值的极限分布，即在大样本条件下，样本均值的分布情况，通常中心极限定理可以得到如下结论：

$$\sqrt{N}\left(\bar{x}-\mu\right)=\frac{S_N-N\mu}{\sqrt{N}}\overset{a}{\sim}\mathcal{N}\left(0,\mathbb{V}\left(x_i\right)\right)$$

其中：

$$S_N=\sum_{i=1}^{N}x_i$$

即样本均值的极限分布为正态分布。

中心极限定理

前面我们提到，如果希望讨论$\sqrt{N}(\bar{x}-\mu)$的极限分布，需要保证$\sqrt{N}(\bar{x}-\mu)=O_p(1)$，实际上：如果假设$\{x_i\}$之间两两不相关，那么：

$$\begin{aligned}\mathbb{V}(S_N) &= \mathbb{E}[S_N-\mathbb{E}(S_N)]^2 \\ &= \mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mathbb{E}(X_i))\right]^2 \\ &= \sum_{i=1}^{n}\mathbb{V}(X_i) \\ &= O(N)\end{aligned}$$

从而$\mathbb{V}\left(\sqrt{N}\frac{S_N}{N}\right)=O(1)$，或者其方差有界，因而$\sqrt{N}\frac{S_N}{N}=\sqrt{N}\bar{x}=O_p(1)$。

中心极限定理

中心极限定理（标量）

令$\{x_i\}$为概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上$i.i.d$的随机变量序列，且$\mathbb{E}(x_i) = \mu, \mathbb{V}(x_i) = \sigma^2$，那么:

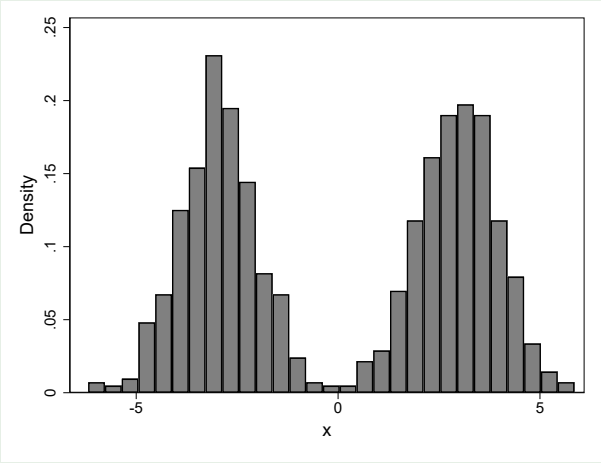
$$\sqrt{N}(\bar{x}_N - \mu) \overset{a}{\sim} \mathcal{N}(0, \sigma^2)$$

或:

$$\sqrt{N}\left(\frac{\bar{x}_N - \mu}{\sigma}\right) \overset{a}{\sim} \mathcal{N}(0, 1)$$

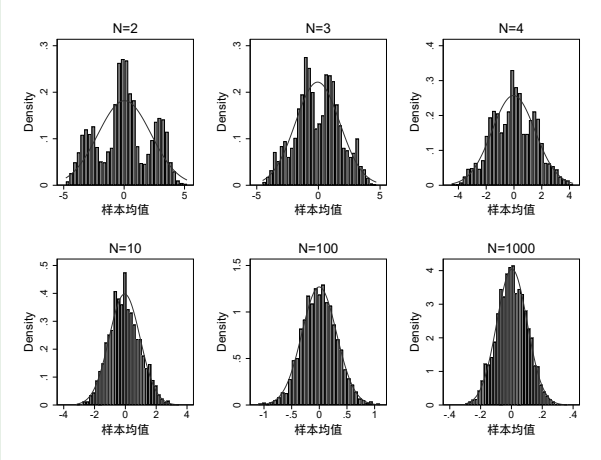
中心极限定理

x_i的分布



中心极限定理

$\bar{x}$的大样本分布



中心极限定理

中心极限定理示例

如果$\{x_i\}$为$i.i.d$的随机变量，且$x_i \sim \mathcal{B}er(p)$，令$\hat{p}_N$如前定义，那么：

$$\sqrt{N}(\hat{p}_N - p) \overset{a}{\sim} \mathcal{N}(0, p(1-p))$$

如果$\{x_i\}$为$i.i.d$的随机变量，且$x_i \sim \mathcal{N}(0,1)$，那么可知$\mathbb{E}(x_i^2) = 1, \mathbb{E}(x_i^4) = 3$，因而：

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - 1\right) \overset{a}{\sim} \mathcal{N}(0, 2)$$

中心极限定理

中心极限定理（向量）

令$\{x_i\}$为概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上$i.i.d$的随机向量序列，且$\mathbb{E}(x_i) = \mu, \mathbb{V}(x_i) = \Sigma$，那么：

$$\sqrt{N}(\bar{x}_N - \mu) \overset{a}{\sim} \mathcal{N}(0, \Sigma)$$

或：

$$\sqrt{N}\Sigma^{-\frac{1}{2}}(\bar{x}_N - \mu) \overset{a}{\sim} \mathcal{N}(0, I)$$

随机变量连续函数的收敛

连续函数的收敛

令$\{X_i\}$为k维随机向量，$g(x):\mathbb{R}^k\to\mathbb{R}^l$为连续函数，那么：

1. $X_n \xrightarrow{\text{a.s.}} X \Rightarrow g(X_n) \xrightarrow{\text{a.s.}} g(X)$
2. $X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X)$
3. $X_n \xrightarrow{D} X \Rightarrow g(X_n) \xrightarrow{D} g(X)$

随机变量连续函数的收敛

样本相关系数的极限

对于二维随机向量(x_i, y_i)，令(x_i, y_i)为i.i.d的样本，那么在可积性条件下，

$$\begin{cases} \frac{1}{N}\sum_{i=1}^{N} x_i \xrightarrow{p} \mathbb{E}(x_i) & \frac{1}{N}\sum_{i=1}^{N} y_i \xrightarrow{p} \mathbb{E}(y_i) \\ \frac{1}{N}\sum_{i=1}^{N} x_i y_i \xrightarrow{p} \mathbb{E}(x_i y_i) & \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \xrightarrow{p} \mathbb{E}(x_i^2) & \frac{1}{N}\sum_{i=1}^{N} y_i^2 \xrightarrow{p} \mathbb{E}(y_i^2) \end{cases}$$

从而

$$\frac{\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\sum_{i=1}^{N} x_i \frac{1}{N}\sum_{i=1}^{N} y_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N} y_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2}} \xrightarrow{p} \mathbb{C}\mathrm{orr}(x_i, y_i)$$

Slutsky定理

Slutsky定理

如果随机变量$X_n \xrightarrow{D} X$，$R_n = o_p(1)$，那么

$$X_n + R_n \xrightarrow{D} X$$

同时如果$Y_n \xrightarrow{p} a \neq 0$，那么

$$\frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{a}$$

如果$Y_n \xrightarrow{p} a$，那么

$$X_n Y_n \xrightarrow{D} aX$$

Slutsky定理

t统计量的大样本分布

之前曾讨论过，如果$x_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$ $i.i.d$，那么

$$\frac{\sqrt{N}\left(\bar{x}-\mu\right)}{\sqrt{\frac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}{N-1}}} \sim \mathrm{t}\left(N-1\right)$$

现在我们不假设x_i服从正态分布，而是假设其独立同分布且具有有限的二阶矩，那么我们有$\bar{x} \xrightarrow{p} \mathbb{E}\left(x_i\right)$, $\frac{1}{N}\sum_{i=1}^{n} x_i^2 \xrightarrow{p} \mathbb{E}\left(x_i^2\right)$

Slutsky定理

t统计量的大样本分布

因而:

$$\begin{aligned}\frac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}{N-1} &= \frac{\sum_{i=1}^{N}x_i^2}{N-1}-\frac{N}{N-1}\bar{x}^2 \\ &= \frac{\sum_{i=1}^{N}x_i^2-N\bar{x}^2}{N-1} \\ &\xrightarrow{p} \mathbb{E}\left(x_i^2\right)-\left[\mathbb{E}\left(x_i\right)\right]^2=\mathbb{V}\left(x_i\right)\end{aligned}$$

进而:

$$\frac{\sqrt{N}\left(\bar{x}-\mu\right)}{\sqrt{\frac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}{N-1}}} \xrightarrow{p} \frac{\sqrt{N}\left(\bar{x}-\mu\right)}{\sqrt{\mathbb{V}\left(x_i\right)}}=\sqrt{N}\left(\frac{\bar{x}-\mu}{\sqrt{\mathbb{V}\left(x_i\right)}}\right) \overset{a}{\sim} \mathcal{N}\left(0,1\right)$$

参数估计

- 在参数模型中，我们假设总体P属于某一个参数族$\{P_\theta, \theta \in \Theta\}$，从而推断总体等价于找到一个参数$\theta_0$，使得$P_{\theta_0} = P$。
- 我们一般把θ_0称为**真值（true value）**。
- 而由于总体是不可观测的，我们只能通过样本对总体进行推断，因而我们不可能得到θ_0的精确值，只能对其进行估计，即**参数估计（Estimation）**。

参数估计

- 参数估计包含两部分，即:
  - **点估计**（point estimation）：找到一个统计量$\hat{\theta}(\boldsymbol{x})$，对总体参数$\theta_0$进行推断，而统计量$\hat{\theta}$我们一般称为**估计量**（estimator）。
  - **区间估计**（interval estimation）：找到一组统计量$L(\boldsymbol{x}), U(\boldsymbol{x})$，使得由其组成的区间包含总体参数$\theta_0$的概率为已知的，即

$$P(L(\boldsymbol{x}) \leq \theta_0 \leq U(\boldsymbol{x})) = p$$

    其中统计量$L(\boldsymbol{x}), U(\boldsymbol{x})$被称为**区间估计量**（interval estimator）。
- 估计量即样本的一个函数，即用于估计参数的统计量，而**估计**（estimate）是对于某一个样本，估计量的实现。

评价估计量的标准

对于同一个参数，经常我们有不同的估计量，比如：

方差的估计

- 对于正态总体$x_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$ $i.i.d$，自然地，σ^2的一个估计量为：

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

而类似的，我们也可以使用：

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

作为σ^2的一个估计。

- 以上两个估计量的差别在于分母的不同，很显然，以上两个估计量具有不同的抽样分布。

均方误差

- 那么，在有很多统计量可供选择时，该如何评价这些统计量呢?
- 一个常用的标准是**均方误差**（**mean squared error**, MSE），即对于一个参数θ和它的估计量$\hat{\theta}$，其误差平方的期望$\mathbb{E}\left(\hat{\theta}-\theta_0\right)^2$为估计量$\hat{\theta}$的均方误差。
- 注意由于:

$$\begin{aligned}
\mathbb{E}\left(\hat{\theta}-\theta_0\right)^2 &= \mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}+\mathbb{E}\hat{\theta}-\theta_0\right)^2 \\
&= \mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right)^2+\mathbb{E}\left(\mathbb{E}\hat{\theta}-\theta_0\right)^2+2\mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right)\left(\mathbb{E}\hat{\theta}-\theta_0\right) \\
&= \mathbb{V}\left(\hat{\theta}\right)+\left(\mathbb{E}\hat{\theta}-\theta_0\right)^2+2\left(\mathbb{E}\hat{\theta}-\theta_0\right)\mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right) \\
&= \mathbb{V}\left(\hat{\theta}\right)+\left(\mathbb{E}\hat{\theta}-\theta_0\right)^2 \\
&= \mathbb{V}\left(\hat{\theta}\right)+\left[\text{Bias}\left(\hat{\theta}\right)\right]^2
\end{aligned}$$

- 其中定义**偏差**（**bias**）：$\text{Bias}\left(\hat{\theta}\right)=\mathbb{E}\left(\hat{\theta}-\theta_0\right)$

均方误差

- 从而$\mathrm{MSE}\left(\hat{\theta}\right) = \mathbb{V}\left(\hat{\theta}\right) + \left[\mathrm{Bias}\left(\hat{\theta}\right)\right]^2$，即均方误差等于估计量的方差与偏差平方的和。
  - 因而，降低均方误差有两种途径：降低估计的方差以及降低偏差。
- 此外，根据均方收敛的定义，只要

$$\mathbb{E}\left(\hat{\theta} - \theta_0\right)^2 = \mathbb{V}\left(\hat{\theta}\right) + \left[\mathrm{Bias}\left(\hat{\theta}\right)\right]^2 \to 0$$

那么$\hat{\theta} \xrightarrow{L^2} \theta_0$，从而$\hat{\theta} \xrightarrow{p} \theta_0$。尽管小样本情况下估计量的偏差不为0，但是我们希望当样本量趋向于无穷时，估计量收敛到真值，也是可以接受的。

评价估计量的三个标准

- 这就引出了评价点估计量的三条标准:
  - 无偏性（unbiasedness）
  - 有效性（efficiency）
  - 一致性（consistency）。

无偏性

- 无偏性要求估计量的偏差为0。
- 当估计量的偏差为0，即$\mathbb{E}\left(\hat{\theta}\right) = \theta_0$时，我们称估计量$\hat{\theta}$为**无偏的（unbiased）**。
- 无偏性意味着，尽管对于每个样本，$\hat{\theta}$对θ_0的估计不可能完全准确而是有误差的，但是估计量$\hat{\theta}$总是围绕在真值θ_0的周围，平均而言（其期望）误差为0。

无偏性

均值的偏差

如果$\mathbb{E}x_i = \mu_0$，那么样本均值的期望：

$$\mathbb{E}(\bar{x}) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \mu_0$$

因而$\bar{x}$是总体均值μ_0的无偏估计。

无偏性

方差的偏差

根据之前的计算，我们知道对于$x_i \sim \left(\mu_0, \sigma_0^2\right)\ i.i.d$:

$$\mathbb{E}\left(s^2\right) = \sigma_0^2$$

因而$\hat{\sigma}^2$的期望为:

$$\mathbb{E}\left(\hat{\sigma}^2\right) = \mathbb{E}\left(\frac{N-1}{N}s^2\right) = \frac{N-1}{N}\sigma_0^2$$

因而$\text{Bias}\left(s^2\right) = 0, \text{Bias}\left(\hat{\sigma}^2\right) = \frac{1}{N}\sigma_0^2$，只有$s^2$是$\sigma_0^2$的无偏估计量。

有效性

- 为了降低MSE，除了降低偏差以外，降低估计量的方差$\mathbb{V}\left(\hat{\theta}\right)$也是非常重要的手段。
- 一般而言，如果两个估计量$\hat{\theta}_1$和$\hat{\theta}_2$，如果

$$\mathbb{V}\left(\hat{\theta}_1\right) < \mathbb{V}\left(\hat{\theta}_2\right)$$

那么我们称$\hat{\theta}_1$相对于$\hat{\theta}_2$是**有效（efficient）**的。

有效性

样本方差的有效性

在样本方差的例子中，因为$\hat{\sigma}^2 = \frac{N-1}{N}s^2$，从而

$$\mathbb{V}\left(\hat{\sigma}^2\right) = \left(\frac{N-1}{N}\right)^2 \mathbb{V}\left(s^2\right) < \mathbb{V}\left(s^2\right)$$

因而$\hat{\sigma}^2$是相对于s^2更有效的估计量。

偏差-方差权衡

- 我们注意到，尽管s^2比$\hat{\sigma}^2$的偏差更小，但是s^2比$\hat{\sigma}^2$的方差更大。
- 在很多应用问题，比如非参数回归或者监督学习（supervised learning）中，我们都会碰到类似的**偏差-方差权衡**（**bias-variance tradeoff**），即很多时候，同时降低偏差和方差是不可能的。

一致性

- 很多时候，尽管在有限样本下，一个估计量的偏差不为零，但是如果样本量足够大时，估计量与真值之间的误差充分的小，那么在样本量比较大时，我们也可以得到一个足够好的估计量。
- 如果一个估计量$\hat{\theta}$依概率收敛到真值θ_0，即$\hat{\theta} \xrightarrow{p} \theta_0$，那么我们称估计量$\hat{\theta}$为一**致（consistent）**估计量。
- 如果一个估计量是不一致的，也就是说即便我们拥有无限多的样本，我们也不能获得真值θ_0的估计，因而一致性是对一个估计量的最低要求。

一致性

样本均值的一致性

在样本均值的例子中，如果对样本$\{x_i\}$做额外的假设（比如x_i独立同分布且可积），那么根据大数定律，有

$$\bar{x} \xrightarrow{p} \mathbb{E}(x_i) = \mu_0$$

因而样本均值是总体均值的一致估计量。

一致性

样本方差的一致性

在样本方差的例子中，由于

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \bar{x}^2$$

其中$\bar{x} \xrightarrow{p} \mu_0$，而

$$\frac{1}{N}\sum_{i=1}^{N}x_i^2 \xrightarrow{p} \mathbb{E}\left(x^2\right) = \mu_0^2 + \sigma_0^2$$

从而$\hat{\sigma}^2 \xrightarrow{p} \sigma_0^2$，即$\hat{\sigma}^2$是$\sigma_0^2$的一致估计量。而

$$s^2 = \frac{N}{N-1}\hat{\sigma}^2 \xrightarrow{p} \sigma_0^2$$

因而s^2也是σ_0^2的一致估计量。

无偏性与一致性

- 需要注意的是：
  - 无偏性关注的是估计量的期望
  - 而一致性则是当样本足够大时估计量的性质
- 两者并没有任何必然联系，无偏性和一致性并不是彼此的充分或者必要条件。

区间估计

- 在上一节中，无偏性、一致性都是使用单个估计量（如样本均值、样本方差）对未知参数进行估计，这种估计被称为点估计（point estimation）。
- 尽管我们可以使用点估计方法对参数值进行推断，然而我们知道，参数的点估计值$\hat{\theta}$与真值θ_0相等的概率一般为0，即$P\left(\hat{\theta}=\theta_0\right)=0$。
- 因而更进一步的，我们很多时候希望得到一个区间，使得这个区间能够以正的概率包含真值θ_0。这就诞生了**区间估计**（**interval estimation**）的概念。
- 区间估计即对于样本$\boldsymbol{x}=(x_1,...,x_N)$，通过一对统计量$L(\boldsymbol{x})$和$U(\boldsymbol{x})$，满足$L(\boldsymbol{x})\leq U(\boldsymbol{x})$，我们可以使用区间$[L(\boldsymbol{x}),U(\boldsymbol{x})]$对未知参数$\theta_0$进行推断。

区间估计

区间包含真值的概率

如果样本$x_i \sim \mathcal{N}(\mu_0, 1)\ i.i.d, i = 1, ..., N$，那么区间$[\bar{x} - 0.5, \bar{x} + 0.5]$包含真值$\mu_0$的概率为:

$$\begin{aligned} P(\mu_0 \in [\bar{x} - 0.5, \bar{x} + 0.5]) &= P(\mu_0 - 0.5 \leq \bar{x} \leq \mu_0 + 0.5) \\ &= P\left(-\frac{0.5}{\sqrt{\frac{1}{N}}} \leq \frac{\bar{x} - \mu_0}{\sqrt{\frac{1}{N}}} \leq \frac{0.5}{\sqrt{\frac{1}{N}}}\right) \end{aligned}$$

由于$\bar{x} \sim \mathcal{N}\left(\mu_0, \frac{1}{N}\right)$，因而

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{1}{N}}} \sim \mathcal{N}(0, 1)$$

区间估计

区间包含真值的概率

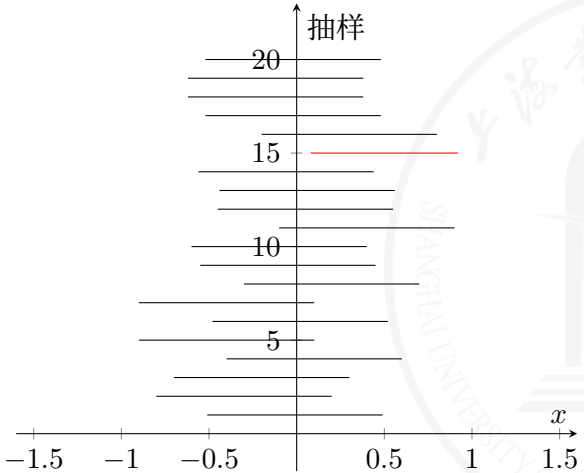
因而

$$P\left(\mu_0 \in [\bar{x}-0.5, \bar{x}+0.5]\right) = \Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right) - \Phi\left(-\frac{0.5}{\sqrt{\frac{1}{N}}}\right)$$

$$= 2\Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right) - 1$$

例如，当$N = 16$时，查表可得，$P\left(\mu_0 \in [\bar{x}-0.5, \bar{x}+0.5]\right) = 2\Phi(2) - 1 \approx 2 \times 0.9772 - 1 = 0.9544$，即区间$[\bar{x}-0.5, \bar{x}+0.5]$包含真值$\mu_0$的概率为$95.44\%$。

区间估计



置信区间

- 注意由于未知参数θ_0是一个未知的常数，而统计量$L(\boldsymbol{x})$和$U(\boldsymbol{x})$是随着抽样的变化而变化的，因此我们不能说「θ_0落入区间$[\bar{x}-0.5,\bar{x}+0.5]$的概率是多少」，而只能说「区间$[\bar{x}-0.5,\bar{x}+0.5]$包含$\theta_0$的概率是多少」。
- 我们把概率$P(\theta_0 \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$称为**覆盖概率（coverage probability）**。
- 注意由于总体参数θ_0未知，因而概率$P(\mu_0 \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$可能依赖于未知的参数$\theta_0$，因而我们通常将而覆盖概率的下界，即$\inf_\theta P_\theta(\theta_0 \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$称为**置信度（confidence coefficient）**或者**置信水平**，通常用$1-\alpha$表示。
- 在某一置信度下，区间$[L(\boldsymbol{x}), U(\boldsymbol{x})]$又被称为**置信区间（confidence interval）**。
  - 因而上例中，我们可以说在95.44%的置信水平下，置信区间为$[\bar{x}-0.5,\bar{x}+0.5]$。

基准统计量

- 此外还需要注意的是，在上例中，为了求得置信区间和覆盖概率，我们首先将统计量$\bar{x}$做了标准化处理，即使用$(\bar{x}-\mu_0)/\sqrt{\frac{1}{N}}$推算概率，而不是直接使用$\bar{x}$。
- 使用$(\bar{x}-\mu_0)/\sqrt{\frac{1}{N}}$的好处是，此统计量的抽样分布不依赖于任何未知参数，因而其分布不会随着未知参数的变化而变化，即服从一个“标准的”分布，这样一来，我们得到的覆盖概率不依赖于任何未知参数。
- $(\bar{x}-\mu_0)/\sqrt{\frac{1}{N}}$依赖未知参数$\mu_0$，严格意义上不是一个统计量，然而$\mu_0$是我们要构建区间估计的目标变量，我们在此允许$\mu_0$的存在。
- 一般的，我们把分布不依赖于未知参数的统计量成为**基准统计量（pivotal statistic）**。

基准统计量

样本均值与基准统计量

如果样本$x_i \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)\ i.i.d, i = 1, ..., N$，那么：

1. 统计量$\bar{x} \sim \mathcal{N}\left(\mu_0, \frac{\sigma_0^2}{N}\right)$，其分布依赖于两个未知参数；
2. 统计量$\bar{x} - \mu_0 \sim \mathcal{N}\left(0, \frac{\sigma_0^2}{N}\right)$，其分布仍然依赖于未知参数$\sigma_0^2$；
3. 统计量$\frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma_0^2}{N}}} \sim \mathcal{N}(0,1)$分布不依赖于任何未知参数，然而计算过程中$\sigma_0^2$未知；
4. 统计量$\frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{N}}} \sim \mathrm{t}(N-1)$分布不依赖于任何未知参数，且计算过程除需要构建置信区间的μ_0之外没有未知参数，因而是基准统计量。

基准统计量

样本方差与基准统计量

如果样本$x_i \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ $i.i.d, i = 1, ..., N$，那么

$$(N-1)\frac{s^2}{\sigma_0^2} \sim \chi^2(N-1)$$

其分布不依赖于任何未知参数，且计算过程除需要构建置信区间的σ_0^2之外没有未知参数，因而是基准统计量。

置信区间的构造

- 之前我们首先给出了区间，进而计算了该区间的置信度。
- 然而现实中，我们经常希望得到在一定置信水平下的置信区间，即一般的区间估计过程。
  - $1-\alpha$常取90%、95%、99%三个数值，其中95%最为常用。
- 由于基准统计量的分布不依赖任何未知参数，因而一般情况下，使用基准统计量可以很方便地构造置信区间。

置信区间的构造

正态分布均值的置信区间

如果样本$x_i \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)\ i.i.d, i = 1, ..., N$，为了得到$\mu_0$的95%的置信区间，我们首先找到基准统计量，要求在基准统计量中，只有μ_0是未知的，其他都是已知的（包括已知常数以及已知统计量）。在上例中，只有统计量

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} \sim \mathrm{t}\left(N - 1\right)$$

满足以上条件。

置信区间的构造

正态分布均值的置信区间

记$F_{\mathrm{t}(N-1)}(x)$为自由度为$N-1$的t分布的分布函数，令$\mathrm{t}_{\alpha/2}^{(N-1)} = F_{\mathrm{t}(N-1)}^{-1}\left(1-\frac{\alpha}{2}\right)$，我们有：

$$
\begin{aligned}
P\left(-\mathrm{t}_{\alpha/2}^{(N-1)} \le \frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{N}}} \le \mathrm{t}_{\alpha/2}^{(N-1)}\right) &= F_{\mathrm{t}(N-1)}\left(\mathrm{t}_{\alpha/2}^{(N-1)}\right) - F_{\mathrm{t}(N-1)}\left(-\mathrm{t}_{\alpha/2}^{(N-1)}\right) \\
&= 1 - 2F_{\mathrm{t}(N-1)}\left(\mathrm{t}_{\alpha/2}^{(N-1)}\right) \\
&= 1 - 2F_{\mathrm{t}(N-1)}\left(F_{\mathrm{t}(N-1)}^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1-\alpha
\end{aligned}
$$

因而我们可以得到：

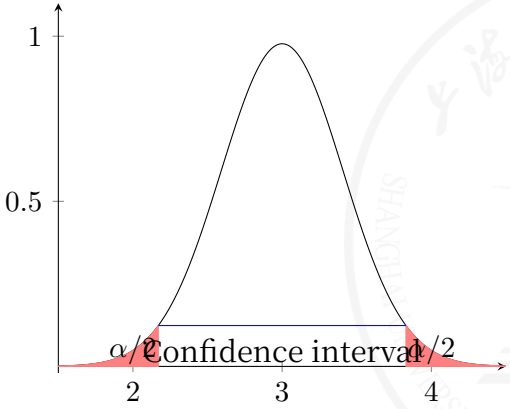
$$
P\left(\bar{x} - \mathrm{t}_{\alpha/2}^{(N-1)}\sqrt{\frac{s^2}{N}} \le \mu_0 \le \bar{x} + \mathrm{t}_{\alpha/2}^{(N-1)}\sqrt{\frac{s^2}{N}}\right) = 1-\alpha
$$

置信区间的构造

正态分布均值的置信区间

- 从而$\left[\bar{x}-\mathrm{t}_{\alpha/2}^{(N-1)}\sqrt{\frac{s^2}{N}},\bar{x}+\mathrm{t}_{\alpha/2}^{(N-1)}\sqrt{\frac{s^2}{N}}\right]$就是我们想要的置信区间。
- 例如，对于一个$N=30$的正态样本，$\bar{x}=3$，$s^2=5$，如果我们想要得到95%置信水平下的置信区间，查表得到$\mathrm{t}_{\alpha/2}^{29}=2.0452$，因而置信下界为$3-2.0452\times\sqrt{5/30}\approx 2.17$，置信上界为$3+2.0452\times\sqrt{5/30}\approx 3.83$。
- 如下图所示，其中红色区域为左右两个概率为$\alpha/2$的区域，中间的一块即为所要求的置信区间。

置信区间的构造



置信区间的构造步骤

总结上述两个置信区间的计算，一般而言我们得到置信区间的步骤如下：

1. 给定置信度$1-\alpha$；
2. 找到一个基准统计量，其中只有所要求的参数是未知的，其他都是已知的；
3. 找到这个基准统计量的分布函数$F(\cdot)$；
4. 查表或使用计算机计算$F^{-1}\left(\frac{\alpha}{2}\right)$以及$F^{-1}\left(1-\frac{\alpha}{2}\right)$；
5. 通过不等式变换得到置信区间。

置信区间的模拟

- 我们可以使用程序验证以上步骤得到的置信区间包含真值的概率刚好为置信度$1-\alpha$。
- 代码CI_small_sample.do给出了正态均值的区间估计中小样本均值的区间估计的模拟。
- 在以上程序中，我们首先定义了一个抽样过程，即从$\mathrm{N}(5,100)$的正态总体中进行抽样，接下来使用公式

$$\left[\bar{x}-\mathrm{t}_{\alpha/2}^{(N-1)}\sqrt{\frac{s^2}{N}},\bar{x}+\mathrm{t}_{\alpha/2}^{(N-1)}\sqrt{\frac{s^2}{N}}\right]$$

计算置信区间，并验证置信区间是否包含了真值（5）。重复抽样10000次，我们就得到了10000个置信区间，最终计算出置信区间包含真值的比率为94.93%，结果显示与与95%相差无几。

大样本下的置信区间

- 尽管上两例给出了正态总体的均值和方差的置信区间的计算方法，然而很多时候我们的总体并不是一定来自于正态总体，很多时候我们很难计算在非正态总体下样本均值的精确分布。
- 然而根据中心极限定理，独立同分布、二阶矩有限的条件下，有：

$$\sqrt{N}\left(\bar{x}-\mu_0\right) \overset{a}{\sim} \mathcal{N}\left(0, \mathbb{V}\left(x\right)\right)$$

- 根据上式，我们有：

$$\frac{\sqrt{N}\left(\bar{x}-\mu_0\right)}{\sqrt{\mathbb{V}\left(x_i\right)}} \overset{a}{\sim} \mathcal{N}\left(0,1\right)$$

然而以上公式中，$\mathbb{V}\left(x_i\right)$是未知的，因而不能直接用于假设检验。

大样本下的置信区间

- 根据Slutsky定理，我们可以使用样本方差s^2代替$\mathbb{V}(x_i)$，由于$s^2 \xrightarrow{p} \mathbb{V}(x_i)$，因而不改变分子上的渐近分布，即有：

$$\frac{\sqrt{N}(\bar{x}-\mu_0)}{\sqrt{s^2}} \overset{a}{\sim} \mathcal{N}(0,1)$$

因而我们可以使用上式进行区间估计。

- 使用类似的技巧，有：

$$P\left(\left|\frac{\sqrt{N}(\bar{x}-\mu_0)}{s}\right| \leq \mathrm{z}_{\alpha/2}\right) = 1-\alpha$$

其中$\mathrm{z}_{\alpha/2} = \Phi_t^{-1}\left(1-\frac{\alpha}{2}\right)$为标准正态分布的上$\alpha/2$分位数，从而置信区间为$\left[\bar{x}-\mathrm{z}_{\alpha/2}\sqrt{\frac{s^2}{N}}, \bar{x}+\mathrm{z}_{\alpha/2}\sqrt{\frac{s^2}{N}}\right]$。

大样本下的置信区间

- 注意$\sqrt{\frac{s^2}{N}}$实际上是$\bar{x}$的标准误$\sqrt{\frac{\sigma_0^2}{N}}$的估计：

$$\text{s.e.}\left(\bar{x}\right)=\sqrt{\frac{s^2}{N}}$$

从而置信区间也可以写为

$$\left[\bar{x}-\text{z}_{\alpha/2}\text{s.e.}\left(\bar{x}\right),\bar{x}+\text{z}_{\alpha/2}\text{s.e.}\left(\bar{x}\right)\right]$$

- 正态总体小样本可以用$\text{t}_{\alpha/2}^{(N-1)}$代替$\text{z}_{\alpha/2}$，即

$$\left[\bar{x}-\text{t}_{\alpha/2}^{(N-1)}\text{s.e.}\left(\bar{x}\right),\bar{x}+\text{t}_{\alpha/2}^{(N-1)}\text{s.e.}\left(\bar{x}\right)\right]$$

作业

1. 如果一个随机变量$X \sim \mathcal{N}(0,1)$，现如下定义随机变量Y：

$$Y = \begin{cases} X-2 & with\ prob\ 0.5 \\ X+2 & with\ prob\ 0.5 \end{cases}$$

   求$\mathbb{V}(Y)$。
2. 证明$\mathbb{V}(Y) = \mathbb{V}[\mathbb{E}(Y|X)] + \mathbb{E}[\mathbb{V}(Y|X)]$
3. 如果$\mathbb{E}(Y|X) = \mathbb{E}(Y)$，那么有没有可能$\mathbb{E}(X|Y) \neq \mathbb{E}(X)$？请举反例或者证明。
4. 如果$Y|(X,Z) \sim \mathcal{N}(X,Z)$，其中$X \sim \mathcal{N}(\mu,\sigma^2)$，$Z \sim \chi^2(K)$，求$\mathbb{E}(Y)$及$\mathbb{V}(Y)$。
5. 证明：
   1. $\mathbb{V}(g(X)+Y|X) = \mathbb{V}(Y|Y)$
   2. $\mathbb{V}(g(X)Y|X) = [g(X)]^2\mathbb{V}(Y|X)$

作业

1. 如果$x_i \sim \mathcal{N}(0,1)$ $i.i.d$，那么请写出：
   1. $\chi_n^2 = \sum_{i=1}^n x_i^2$的分布
   2. $\frac{1}{n}\chi_n^2$的渐进分布
   3. 根据以上的结论，能否认为χ^2分布的极限分布是正态分布（或者说$\chi_n^2 \overset{a}{\sim} \mathcal{N}(n, 2n)$是否成立？）？
2. 已知：如果$x \sim \mathcal{N}\left(\mu, \sigma^2\right)$，那么

$$\mathbb{E}\left[(x - \mathbb{E}(x))^4\right] = 3\sigma^4$$
$$\mathbb{E}\left[(x - \mathbb{E}(x))^6\right] = 15\sigma^6$$
$$\mathbb{E}\left[(x - \mathbb{E}(x))^8\right] = 105\sigma^8$$

对于样本偏度系数：

$$b_1 = \frac{N^2}{(N-1)(N-2)} \frac{\frac{1}{N}\sum_{i=1}^N (x_i - \bar{x})^3}{s^3}$$

请计算：

1. b_1的概率极限
2. $\sqrt{N} b_1$的极限分布