《概率论与数理统计》概念与常用分布

2021 秋季学期　授课教师：刘杰　整理者：徐小航

1. **条件概率**：$P(A|B) \triangleq P(AB)/P(B)$
乘法定理：$P(AB) = P(B)P(A|B)$
全概率公式：$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i)$
Bayes 公式：

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}$$

2. 多维变量：
n维离散型随机变量的概率函数：$p(j_1, \dots, j_n) = P(X_1 = a_{1j_1}, \dots, X_n = a_{nj_n})$
n维随机变量的(联合)分布函数：$F(x_1, x_2, \dots, x_n) = P(X_1 \le x_1, \dots, X_n \le x_n)$
边缘分布：对已知分布F的n维随机变量$(X_1, \dots, X_n)$，其任意一个子集的分布称为F的一个边缘分布。
边缘密度函数：

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y \ , f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x$$

条件概率密度：$f_{X|Y}(x|y) \triangleq P(X \le x | y \le Y \le y + \varepsilon) \equiv f(x,y)/f_X(x)$，记作$X|y \sim f_{X|Y}(x|y)$，反之同理。对高维情形，$h(\boldsymbol{y}|\boldsymbol{x}) = f(\boldsymbol{x},\boldsymbol{y})/g(\boldsymbol{x})$。

3. **独立事件**：$P(A_1A_2 \dots A_n) = P(A_1)P(A_2) \dots P(A_n)$
独立与不相容是两个不同概念。
随机变量独立性：若$f_{X|Y}(x|y) = f_1(x), f_{Y|X}(y|x) = f_2(y)$，则$f(x,y) = f_1(x)f_2(y)$，称$X, Y$相互独立。可推广到多变量。此时，联合分布函数也是各边缘分布的乘积。
离散型随机变量独立性：$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \dots P(X_n = X_n)$
*独立时，联合密度函数是各条件概率密度函数乘积，联合分布函数是各边缘分布函数乘积。
事件独立与变量独立关系：对事件组$\{A_n\}$与变量组$\{X_n = I_{A_n}\}$，A_n独立$\Leftrightarrow X_n$独立。

4. 随机变量的函数的概率分布：（已知$\{X_n\}$分布求$Y = g(X_1, \dots, X_n)$的分布）
离散随机变量的情形：$P(Y = y_j) = P(g(X) = y_j) = \sum_{x_i: g(x_i) = y\_j} P(X = x_i)$
离散卷积公式：相互独立的非负整值随机变量ξ, η各有分布律$\{a_k\}, \{b_k\}$，则$\xi + \eta$的分布律为：

$$P(\xi + \eta = n) = \sum_{k=0}^{n} a_k b_{n-k}$$

密度变换公式：连续随机变量X有密度函数$f(x), x \in (a,b)$，$y = g(x)$在$x \in (a,b)$上严格单调连续，存在唯一的反函数$x = h(y)$且$h'(y)$存在并连续，则$Y = g(x)$是连续性随机变量且：

$$p(y) = f(h(y)) \cdot |h'(y)| \ , y \in (g(a), g(b))$$

求解的一般方法：$P(Y \le y) = P(g(X) \le y) = P\big(X \le g^{-1}(y)\big)$

2 维情形：对有联合密度函数$p(x_1, x_2)$的 2 维连续随机向量(ξ_1, ξ_2)，设$\zeta_j = f_j(\xi_1, \xi_2), j = 1,2$，若$(\zeta_1, \zeta_2)$与$(\xi_1, \xi_2)$一一对应，有逆映射$\xi_j = h_j(\zeta_1, \zeta_2), j = 1,2$，且每个$h_j$都有一阶偏导，则$(\zeta_1, \zeta_2)$为连续随机向量，联合概率密度为：

$$q(y_1, y_2) = \begin{cases} p\big(h_1(y_1, y_2), h_2(y_1, y_2)\big)|J| & , (y_1, y_2) \in \mathbb{D} \\ 0 & , (y_1, y_2) \notin \mathbb{D} \end{cases} , J = \begin{vmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_2}{\partial y_1} \\ \frac{\partial h_1}{\partial y_2} & \frac{\partial h_2}{\partial y_2} \end{vmatrix}$$

卷积公式：X, Y的联合概率密度为$f(x, y)$，则$X + Y$的概率密度$p(z)$为：

$$p(z) = \int_{-\infty}^{+\infty} f(x, z - x)\,\mathrm{d}\,x = \int_{-\infty}^{+\infty} f(z - y, y)\,\mathrm{d}\,y$$

特别当X, Y独立时，有卷积公式：

$$p(z) = \int_{-\infty}^{+\infty} f_1(x) f_2(z - x)\,\mathrm{d}\,x = f_1 * f_2(z)$$

随机变量之商的概率密度：对联合密度为$f(x, y)$的连续随机变量X, Y，其商为连续随机变量，密度函数为：

$$p_{\frac{X}{Y}}(x) = \int_{-\infty}^{+\infty} |t| f(xt, t)\,\mathrm{d}\,t \ , p_{\frac{Y}{X}}(x) = \int_{-\infty}^{+\infty} |u| f(u, xu)\,\mathrm{d}\,u$$

5. 随机变量的数字特征

数学期望：

$$EX \triangleq \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}\,x \quad , \int_{-\infty}^{+\infty} |x| f(x)\,\mathrm{d}\,x = \infty \Leftrightarrow X\text{的数学期望不存在}$$

数学期望具有线性性。

对独立变量，$E(X_1 X_2 \dots X_n) = EX_1 EX_2 \dots EX_n$。

若$E\big(g(x)\big)$存在，则：

$$E(g(x)) = \begin{cases} \sum_i g(a_i) p_i \\ \int_{-\infty}^{+\infty} g(x) f(x)\,\mathrm{d}\,x \end{cases}$$

条件期望：

$$E(Y|X = x) = \triangleq \int_{-\infty}^{+\infty} y f(y|x)\,\mathrm{d}\,y \ , \sum_i a_i p_i$$

条件期望满足数学期望的性质，包括线性相加、线性相乘、变量函数的期望。

全期望公式：$EX = E\big(E(X|Y)\big)$

求解期望的第二种方法：先求解$h(x) = E(Y|X = x)$，再求解$Eh(X)$，根据全期望公式$Eh(X) = EY$。

全期望公式的推广：$Eg(X) = E\big(E(g(X)|Y)\big)$。

方差：

$$Var(X) \triangleq E(X = EX)^2 \triangleq \sigma^2$$

标准差：$\sigma \triangleq \sqrt{VarX}$。

$Var(X) = EX^2 - (EX)^2$。$0 \le Var(X) \le EX^2$，当且仅当$P(X \ne EX) = 0$时$Var(X) = 0$。

$Var(cX) = c^2 Var(X)$。

对独立的X, Y，$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$。

对$\forall c \in \mathbb{R}$，$Var(X) \leq E(x-c)^2$，等号只在$c = EX$时成立。

标准化随机变量：

$$X^* \triangleq \frac{X - EX}{\sqrt{Var(X)}}$$

易知，$EX^* = 0$，$Var(X^*) = 1$。标准化随机变量的引入可以消除计量单位带来的影响。

矩：对$c \in \mathbb{R}, r \in \mathbb{Z}^+$，称$E((X-c)^r)$为$X$关于$c$的$r$阶矩。

r阶原点矩：$c = 0$，$a_r \triangleq EX^r$。

r阶中心矩：$c = EX$，$\mu_r = E(X - EX)^r$。

1 阶原点矩就是期望，二阶中心矩就是方差。

协方差：

$$Cov(X,Y) = E(X - EX)(Y - EY)$$

协方差的性质：

$Cov(X,Y) = Cov(Y,X)$，$Cov(X,X) = Var(X)$;

$Cov(X,Y) = EXY - EXEY$，若X,Y独立则$Cove(X,Y) = 0$;

$Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$;

协方差具有多重线性相加性。

相关系数：

$$\rho_{X,Y} \triangleq \frac{Cov(X,Y)}{\sqrt{VarX}\sqrt{VarY}}$$

当$\rho_{X,Y} = 0$，称X,Y不相关。$\rho_{X,Y} = Cov(X^*, Y^*)$，因此相关系数可以视为标准尺度下的协方差。

X,Y独立$\Longrightarrow \rho_{X,Y} = 0$。独立与不相关是不同的，但独立一定不相关。

$|\rho_{X,Y}| \leq 1$，等号成立当且仅当X,Y间有严格线性关系，1 对应正相关，-1 对应负相关。

平均绝对差：$E|X - EX|$。

矩母函数：$Ee^{tX}, t \in \mathbb{R}$。

特征函数：$Ee^{itX}, t \in \mathbb{R}$。

6. 记：

$$(\forall \varepsilon > 0)\left(\lim_{n\to\infty} P(|\xi_n - \xi| \geq \varepsilon) = 0\right)$$

为$\xi_n \xrightarrow{p} \xi$，称为随机变量序列$\{\xi_n\}$依概率收敛到随机变量$\xi$。

大数定律：设$\{X_n\}$是独立同分布的随机变量列，有公共期望μ，则：

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{p} \mu$$

Chebyshev 不等式：

$$P(|X - EX| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}, \forall \varepsilon > 0$$

中心极限定理：对独立同分布的$\{X_n\}$，有：

$$\bar{X}^* \xrightarrow{p} N(0,1)$$

棣莫弗-拉普拉斯定理：0-1 分布的中心极限定理，中心极限定理的最早形式。根据该定理，可以用正态分布估算高n的二项分布。

7. 常用统计量:

样本均值:

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

样本方差:

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

S称为样本标准差。

样本矩:

样本k阶原点矩:

$$a_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$$

样本k阶中心矩:

$$m_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$$

次序统计量：将从总体F中抽取的样本F按大小排列为$X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)}$，则称$\left(X_{(1)},\dots,X_{(n)}\right)$及其任一部分为次序统计量。

样本中位数:

$$m_{\frac{1}{2}}=\begin{cases}X_{\left(\frac{n+1}{2}\right)} & \text{当}n\text{为奇数}\\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right) & \text{当}n\text{为偶数}\end{cases}$$

极值：极小值$X_{(1)}$与极大值$X_{(n)}$。

8. 点估计：设总体X的分布函数形式已知，但有一或多个参数未知，例如参数θ未知，根据样本$X_1,\dots,X_n$来估计参数θ，就是要构造适当的统计量$\hat{\theta}=\hat{\theta}(X_1,\dots,X_n)$，则这个构造出来的统计量$\hat{\theta}$称为$\theta$的**估计量**，这种估计方式叫做**点估计**。

矩估计：用样本矩估计总体矩。

1° 假设X的分布函数中有n个未知参数$\{\theta_n\}$，根据X的分布函数形式选取n个矩$\alpha_1,\dots,\alpha_n=f_1,\dots,f_n(\theta_1,\dots,\theta_n)$，再计算这$n$个矩对应的样本矩$a_1,\dots,a_n$。矩的选取原则是：能用低阶矩就不用高阶矩。

2° 根据大数定律，样本矩依概率收敛于对应矩，因此有方程组:

$$\begin{cases}\alpha_1=f_1(\theta_1,\dots,\theta_n)\\ \vdots\\ \alpha_n=f_n(\theta_1,\dots,\theta_n)\end{cases}\Longrightarrow\begin{cases}a_1=f_1\left(\widehat{\theta_1},\dots,\widehat{\theta_n}\right)\\ \vdots\\ a_n=f_n\left(\widehat{\theta_1},\dots,\widehat{\theta_n}\right)\end{cases}$$

3° 解该方程组，得到$\theta_1,\dots,\theta_n$的一个估计:

$$\begin{cases}\widehat{\theta_1}=g_1(a_1,\dots,a_n)\\ \vdots\\ \widehat{\theta_n}=g_n(a_1,\dots,a_n)\end{cases}$$

矩估计的优点是简单易行，不需要事先知道总体分布，缺点是当总体类型已知时，没有充分利用分布提供的信息。一般情况下，矩估计量不具有唯一性。

极大似然估计：当参数θ固定时，分布密度函数$f(x;\theta)$可看作得到样本观察值x的可能性。若把θ当作可变的自变量，就能得到在不同θ下能得到样本观察值为x的可能性$L(x;\theta)$。我们观察到的x是已知的，因此要寻找在哪个θ下$L(x;\theta)$最大，这个θ就是**极大似然估计值**。

简单样本、连续分布下的极大似然估计流程：

1° 若有n个简单样本$\{X_n\}$，$L(x;\theta)=\prod_{i=1}^{n} f(x_i;\theta)$。

2° 定义对数似然函数$l(\theta)=\log L(\theta)$，解方程得到l或L的驻点：

$$\frac{\mathrm{d}\,l}{\mathrm{d}\,\theta}=0 \text{ 或 } \frac{\mathrm{d}\,L}{\mathrm{d}\,\theta}=0$$

通常来说，解前者更为方便。当有多个参数$\{\theta_n\}$时，则分别求l对各个参数偏导的驻点。

3° 判断解得的驻点$\hat{\theta}$是否为极大值点。如果是，则$\hat{\theta}$是参数θ的极大似然估计值，$\hat{\theta}(x)=\hat{\theta}(X)$是极大似然估计量。

9. 点估计的优良准则：

无偏性：若$\hat{g}(X_1,\dots,X_n)$是待估参数函数$g(\theta)$的一个估计量，有$E\hat{g}(X_1,\dots,X_n)=g(\theta)$，则称$\hat{g}(X_1,\dots,X_n)$是$g(\theta)$的无偏估计量。无偏性是估计量的最基本要求，其意义是没有系统误差。有偏估计量可以修正为无偏估计量。

有效性：对$g(\theta)$的两个估计量$\hat{g}_1(X_1,\dots,X_n),\hat{g}_2(X_1,\dots,X_n)$，若：

$$Var\big(\hat{g}_1(X_1,\dots,X_n)\big)\le Var\big(\hat{g}_2(X_1,\dots,X_n)\big)$$

则称$\hat{g}_1$比$\hat{g}_2$有效。

相合性：设总体分布依赖于参数$\{\theta_k\}$，$g(\theta_1,\dots,\theta_k)$是待估参数函数。若对样本$\{X_n\}$，$g(\theta_1,\dots,\theta_k)$有估计量$T(X_1,\dots,X_n)$，且对$\forall\varepsilon>0$与$\theta_1,\dots,\theta_k$的一切可能值有：

$$\lim_{n\to\infty} P_{\theta_1,\dots,\theta_k}(|T(X_1,\dots,X_n)-g(\theta_1,\dots,\theta_k)|\ge\varepsilon)=0$$

则称$T(X_1,\dots,X_n)$为$g(\theta_1,\dots,\theta_k)$的相合估计量。相合性是估计量的最基本要求，意义是可以根据样本的多少，控制估计的精度。

渐近正态性：形式复杂的统计量，当样本n很大时，分布都渐近于正态分布。

10. 区间估计：若$P_\theta\big(\underline{\theta}\le\theta\le\bar{\theta}\big)=1-\alpha$，则称$[\underline{\theta},\bar{\theta}]$是$\theta$的置信水平为$1-\alpha$的**置信区间**。一般要先寻找$\theta$的一个估计（多数是基于其充分统计量构造的），然后基于此估计量构造参数θ的置信区间。

枢轴变量法：设待估参数为$g(\theta)$：

1° 找一个与待估参数$g(\theta)$有关的统计量T，一般是其一个良好的点估计，多数通过极大似然方法构造。

2° 设法找出T与$g(\theta)$的某一函数$S\big(T,g(\theta)\big)$的分布，其分布F要与参数θ无关，S即枢轴变量。

3° 对任何常数$a<b$，不等式$a<S\big(T,g(\theta)\big)<b$要能写成等价形式$A\le g(\theta)\le B$，其中$A,B$只与$T,a,b$有关，与参数无关。

4° 取分布F的上$\alpha/2$分位数与上$1-\alpha/2$分位数$\omega_{1-\alpha/2}$，有$F\big(\omega_{\alpha/2}\big)-F\big(\omega_{1-\alpha/2}\big)=1-\alpha$，因此$P\big(\omega_{1-\alpha/2}\le S\big(T,g(\theta)\big)\le\omega_{\alpha/2}\big)=1-\alpha$。这就是我们需要的置信区间。

大样本法：通过中心极限定理建立枢轴变量。在样本足够大时，得到估计$P\big(-u_{\alpha/2}\le\bar{X}^*\le u_{\alpha/2}\big)\approx1-\alpha$。由此可以解得$\bar{X}$的置信区间，但这种估计只有在样本数较大时才相去不远。如果假定方差已知，可以得到 Wald 置信区间：

$$\hat{p} \pm u_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

确定样本大小：以区间长度为精度准则下，置信区间越窄越好，而只有在样本数足够大，才能事先把置信区间压缩的足够窄。根据置信区间的构造方法，我们可以逆向构造出达到该置信区间需要的样本量，从而确定样本大小。

11. 假设检验问题的概念：假设检验问题就是研究如何根据抽样后获得的样本来检验抽样前所作出的假设。设有假设检验问题$H_0:\theta \in \Theta_0 \leftrightarrow H_1:\theta \in \Theta_1$，其中$H_0$为**零假设/原假设**，$H_1$为**对立假设/备择假设**，构造一个适当的**检验统计量**$T = T(X_1,\dots,X_n)$，其中$X_1,\dots,X_n$为从总体中抽得的一个样本，根据对立假设的形状构造一个检验的**拒绝域**$W = \{T(X_1,\dots,X_n) \in A\}$，以$\{T(X_1,\dots,X_n) > \tau\}$为例，$\tau$称为**临界值**。如果零假设成立但拒绝了零假设，则称犯了**第Ⅰ类错误**；如果对立假设成立但接受了零假设，则称犯了**第Ⅱ类错误**。如果对$\forall\theta \in \Theta_0$，犯第Ⅱ类错误的概率≤某个正的常数$\alpha$，则称$\alpha$为**显著性水平**。显著性水平取最小的那个。称$\beta(\theta) = P_\theta(\text{Refuse } H_0)$为检验的**功效函数**，当$\theta \in \Theta_0$，$\beta(\theta) \leqq \alpha$，而当$\theta \in \Theta_1$，我们希望$\beta(\theta)$越大越好，这样犯第Ⅱ类错误的概率越小。功效函数的评价检验优劣的一个标准。

假设检验提法的原则：

① 将受保护对象置为零假设。

② 如果要证明某个命题，则将相反命题置为零假设。

统一地来说，就是不要让受保护对象去自证一个命题。尽可能避免犯第Ⅰ类错误。

12. 重要参数检验：

一样本正态总体：

检验对象	检验统计量	分布	拒绝域
μ（σ^2已知）	$U = \sqrt{n}\dfrac{\bar{X}-\mu_0}{\sigma}$	$N(0,1)$	$\begin{cases}\lvert U\rvert > u_{\alpha/2}\\ U > u_\alpha\\ U < -u_\alpha\end{cases}$
μ（σ^2未知）	$T = \sqrt{n}\dfrac{\bar{X}-\mu_0}{S}$	t_{n-1}	$\begin{cases}\lvert T\rvert > t_{n-1}(\alpha/2)\\ T > t_{n-1}(\alpha)\\ T < -t_{n-1}(\alpha)\end{cases}$
σ^2（μ已知）	$\chi^2 = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$	χ_n^2	$\begin{cases}\chi^2 > \chi_n^2(\alpha/2)\text{或}\chi^2 < \chi_n^2(1-\alpha/2)\\ \chi^2 > \chi_n^2(\alpha)\\ \chi^2 < \chi_n^2(1-\alpha)\end{cases}$
σ^2（μ未知）	$\chi^2 = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2$	χ_{n-1}^2	$\begin{cases}\chi^2 > \chi_{n-1}^2(\alpha/2)\text{或}\chi^2 < \chi_{n-1}^2(1-\alpha/2)\\ \chi^2 > \chi_{n-1}^2(\alpha)\\ \chi^2 < \chi_{n-1}^2(1-\alpha)\end{cases}$

以上表格中，拒绝域一栏的第一、二、三式分别是对立假设取$\theta \neq \theta_0, \theta > \theta_0, \theta < \theta_0$的情况。

二样本正态总体：取$Z = X - Y$：

检验对象	检验统计量	分布	拒绝域
均值 方差已知	$U = \dfrac{\bar{X}-\bar{Y}}{\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}}$	$N(0,1)$	$\begin{cases}\lvert U\rvert > u_{\alpha/2}\\ U > u_\alpha\\ U < -u_\alpha\end{cases}$
均值 方差未知	$T = \dfrac{\bar{X}-\bar{Y}}{S_w\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}}$	t_{m+n-2}	$\begin{cases}\lvert T\rvert > t_{m+n-2}(\alpha/2)\\ T > t_{m+n-2}(\alpha)\\ T < -t_{m+n-2}(\alpha)\end{cases}$

方差 均值已知	$F=\frac{\sum_{i=1}^{m}(X_i-\mu_1)^2/m}{\sum_{i=1}^{n}(X_i-\mu_2)^2/n}$	$F_{m,n}$	$\begin{cases}F>F_{m,n}(\alpha/2)\text{或}\chi^2<F_{n,m}(\alpha/2)^{-1}\\ \chi^2>F_{m,n}(\alpha)\\ \chi^2<F_{n,m}(\alpha)^{-1}\end{cases}$
方差 均值未知	$F=\frac{S_1^2}{S_2^2}$	$F_{m-1,n-1}$	$\begin{cases}F>F_{m-1,n-1}(\alpha/2)\text{或}F<F_{n-1,m-1}(\alpha/2)^{-1}\\ F>F_{m-1,n-1}(\alpha)\\ F<F_{n-1,m-1}(\alpha)^{-1}\end{cases}$

以上表格中，拒绝域栏的第一、二、三式分别是对立假设取$\theta_1\neq\theta_2,\theta_1>\theta_2,\theta_1<\theta_2$的情况。

0-1 分布中p的假设检验：p的极大似然估计为$\bar{X}$，故检验统计量为：

$$T=\sqrt{n}\frac{\bar{X}-p_0}{\sqrt{p_0(1-p_0)}}$$

该统计量服从分布$N(0,1)$。p_0是检验域的临界值。

成对数据：两样本正态分布要求样本独立同分布，而成对数据不要求。成对数据只要求$Z_n(X_n,Y_n)$独立同分布，从而将成对数据的假设检验变为单样本假设检验问题。

13. 拟合优度检验：假设检验基本上是假定总体服从正态分布的条件下做的，但这个假设本身需要检验。一般地，检验H_0: X服从某种分布，可以采用χ^2拟合优度检验。

离散总体不含未知参数的情形：

1° 设总体X服从一个离散分布，该离散分布规定了总体落在类别$a_1,\dots,a_k$的理论频率分别为$p_1,\dots,p_k$，现抽取一个样本量为n的样本，落在$\{a_k\}$的观测数分别为$\{n_k\}$，检验理论频率是否正确即：

$$H_0: P(X\in a_1)=p_1,\dots,P(X\in a_k)=p_k$$

这类问题只提零假设，不提对立假设，检验方法称为拟合优度检验。

2° 零假设成立时，n_i/n依概率收敛于p_i，检验统计量取为：

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}=\sum\frac{(O-E)^2}{E}$$

其中O为观测频数，E为期望频数。χ^2的极限分布是χ^2_{k-1}。

3° 取显著性水平为α，则零假设的接受域为$\chi^2\leq\chi^2_{k-1}(\alpha)$。

离散总体含若干未知参数的情形：

1° 用适当的估计$\widehat{p_i}$，如极大似然估计，代替p_i，得到统计量：

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-n\widehat{p_i})^2}{n\widehat{p_i}}$$

χ^2的极限分布是χ^2_{k-1-r}，其中r为估计的独立参数个数。

2° 取显著性水平为α，则零假设的接受域为$\chi^2\leq\chi^2_{k-1}(\alpha)$。

列联表的独立性检验：如果一个事物有两个属性，第一个属性有a个水平，第二个属性有b个属性，则该列联表为$a\times b$表。常见问题是考察两个属性是否独立，即零假设是H_0: A,B独立。

1° 若样本量为n，第(i,j)格的频数为n_{ij}，记$p_{ij}=P(\text{属性}A,B\text{分别处于水平}i,j)$，$u_i=P(\text{属性}A\text{有水平}i)$，$v_i=P(\text{属性}B\text{有水平}j)$，则零假设是$p_{ij}=u_iv_j$。将$u_i,v_j$看作参数，则总独立参数有$a+b-2$个，极大似然估计为：

$$\widehat{u_i}=\frac{n_{i\cdot}}{n}\ ,\widehat{v_j}=\frac{n_{\cdot j}}{n}$$

其中$n_{i\cdot}=\sum_{j=1}^{b}n_{ij}\ ,n_{\cdot j}=\sum_{i=1}^{a}n_{ij}$。

2° 取检验统计量：

$$\chi^2=\sum_{i=1}^{a}\sum_{j=1}^{b}\left(n_{ij}-n\widehat{p_{ij}}\right)=\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\left(n_{ij}-n_{i\cdot}n_{\cdot j}/n\right)^2}{n_{i\cdot}n_{\cdot j}/n}$$

χ^2的极限分布是$\chi^2_{(a-1)(b-1)}$。对四格表，服从χ^2_1。

3° 取显著性水平为α，则零假设的接受域为$\chi^2\le\chi^2_{(a-1)(b-1)}(\alpha)$。

列联表的齐一性检验：检验某个属性A的各个水平对应的另一个属性B分布全部相同，所采用的检验方法与独立性检验完全相同。

连续总体情形的拟合优度检验：设在总体X中取样本$(X_1,\dots,X_n)$，记X的分布函数为$F(x)$，需要检验的那种分布含有r个总体参数$\theta_1,\dots,\theta_r$，我们要在显著性水平α下检验$H_0:F(x)=F_0(x;\theta_1,\dots,\theta_r)$，其中$F_0(x;\theta_1,\dots,\theta_r)$是需要检验的那种分布的分布函数。上述假设可以通过适当的离散化总体分布，采用拟合优度法检验。

1° 把实数轴分为k个子区间，$(a_{j-1},a_j],j=1,\dots,k$，其中$a_0$可以取$-\infty$，$a_k$可以取$+\infty$。这样，我们就构造了一个离散总体，记：

$$p_j=P_{H_0}\left(a_{j-1}<X\le a_j\right)=F_0\left(a_j;\theta_1,\dots,\theta_r\right)-F_0\left(a_{j-1};\theta_1,\dots,\theta_r\right)$$

如果H_0成立，p_j应该与数据落在$(a_{j-1},a_j]$的频率$f_j=n_j/n$接近，其中n_j表示相应的频数。

2° 取检验统计量：

$$\chi^2=\sum_{j=1}^{k}\frac{\left(n_j-np_j\right)^2}{np_j}\quad p_j\text{的取值不含未知参数}$$

$$\chi^2=\sum_{j=1}^{k}\frac{\left(n_j-n\widehat{p_j}\right)^2}{n\widehat{p_j}}\quad p_j\text{的取值含未知参数}$$

3° 拒绝域为$\chi^2>\chi^2_{k-r-1}(\alpha)$，$r$为独立参数量。

用χ^2进行拟合优度检验时，一般要求$n\ge50,n\widehat{p_j}\ge5,j=1,\dots,k$，否则最好将一些组合并。

《概率论与数理统计 B》常用分布

1. 0-1 分布/伯努利分布$\mathrm{Bern}(p)$:

$$\begin{cases} P(X=1)=p \\ P(X=0)=1-p \end{cases}$$

$$EX=p\ ,VarX=p(1-p)$$

2. 二项分布$B(n,p)$:

$$P(x=k)=\binom{n}{k}p^k(1-p)^{n-k}\ \ ,k=0,1,2,\dots,n$$

物理意义：把伯努利实验重复n次，其中事件A发生k次的概率。
服从条件：各次试验条件稳定；各次实验相互独立。
再生性：$X\sim B(n,p),Y\sim B(m,p)$，$X,Y$独立$\Longrightarrow X+Y\sim B(m+n,p)$。
$X_i\sim B(1,p)\Longrightarrow\sum_{i=1}^{n}X_i\sim B(n,p)$。

$$EX=np\ ,VarX=np(1-p)$$

3. 泊松分布$P(\lambda)$:

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}\ \ ,k\in\mathbb{N},\lambda>0$$

物理意义：$B(n,p)$在$p\to 0,np\to\lambda$时的极限。因此泊松分布是n很大的二项分布的近似，尤其在n未知的时候。
再生性：$X\sim P(\lambda),Y\sim P(\mu)$，$X,Y$独立$\Longrightarrow X+Y\sim P(\lambda+\mu)$。

$$EX=\lambda\ ,VarX=\lambda$$

4. 离散均匀分布$U(n)$:

$$P(X=a_k)=\frac{1}{n}\ \ ,k=1,\dots,n$$

离散均匀分布是经典概型的抽象。

5. 一维正态分布$N(\mu,\sigma^2)$:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\ \ ,x,\mu,\sigma\in\mathbb{R}$$

标准正态分布$N(0,1)$的分布函数表示为$\Phi(x)$，密度函数表示为$\phi(x)$。
σ又称正态分布的形状参数。

$$F(x)=\Phi\left(\frac{x-\mu}{\sigma}\right)$$

再生性：$X\sim N(\mu_1,\sigma_1^2),Y\sim N(\mu_2,\sigma_2^2)$，$X,Y$相互独立$\Longrightarrow X+Y\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$。

$$EX=\mu\ ,VarX=\sigma^2$$

中心极限定理：对于独立同分布的$\{X_n\}$，有：

$$\frac{1}{\sqrt{n}\sigma}(X_1+\cdots+X_n-n\mu)\xrightarrow{n\to+\infty}N(0,1)$$

6. 指数分布$Exp(\lambda)$:

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, x>0\\ 0, x\le 0\end{cases},\lambda>0$$

$$F(x)=\begin{cases}1-e^{-\lambda x}, x>0\\ 0, x\le 0\end{cases}$$

指数分布用于寿命近似，描述无老化时候的寿命分布。
指数函数具有无记忆性，即对$\forall s,t>0$有$P(X>s+t|X>s)=P(X>t)$。指数函数是唯一具有无记忆性的连续分布。

$$EX=\frac{1}{\lambda},VarX=\frac{1}{\lambda^2}$$

7. 连续均匀分布$U(a,b)$:

$$f(x)=\begin{cases}\frac{1}{b-a}, a\le x\le b\\ 0, \text{其它}\end{cases}$$

连续均匀分布可用于描述因四舍五入产生的误差。

$$EX=\frac{a+b}{2},VarX=\frac{(b-a)^2}{12}$$

8. 二维连续均匀分布$U([a,b]\times[c,d])$:

$$f(x_1,x_2)=\begin{cases}\frac{1}{(b-a)(c-d)}, a\le x_1\le b, c\le x_2\le d\\ 0, \text{其它}\end{cases}$$

X,Y相互独立。如果(X,Y)服从单位圆上的均匀分布，则不独立。

9. 二维正态分布$N(a,b,\sigma_1^2,\sigma_2^2,\rho)$:

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-a)^2}{\sigma_1^2}-\frac{2\rho(x-a)(y-b)}{\sigma_1\sigma_2}+\frac{(y-b)^2}{\sigma_2^2}\right)\right)$$

其中$a,b\in\mathbb{R},\sigma_1,\sigma_2\in\mathbb{R}^+,\rho\in[-1,1]$。$X,Y$相互独立的充要条件是$\rho=0$。
协方差$\rho_{X,Y}=\rho$。X的边缘分布为$N(a,\sigma_1^2)$，Y的边缘分布为$N(b,\sigma_2^2)$。

10. 自由度为n的卡方分布χ_n^2:

$$f_n(x)=\begin{cases}\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, x>0\\ 0, x\le 0\end{cases}$$

对$X_1,\dots,X_n$ i. i. d. $\sim N(0,1)$:

$$X=\sum_{i=1}^{n}X_i^2\sim\chi_n^2$$

$n=1,2$时曲线单调下降；$n\ge 3$时曲线有单峰；n越大曲线越对称。

$$EX=n,VarX=2n$$

再生性：$Z_1\sim\chi_n^2,Z_2\sim\chi_m^2$，$Z_1,Z_2$独立$\Longrightarrow Z_1+Z_2\sim\chi_{n+m}^2$。

11. 自由度为n的t分布t_n:

$$f_n(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},x\in\mathbb{R}$$

若$X_1\sim N(0,1),X_2\sim\chi_n^2$，$X_1,X_2$独立，则：

$$Y=\frac{X_1}{\sqrt{X_2/n}}\sim t_n$$

$$\lim_{n\to+\infty}t_n(x)=\phi(x)$$

$$n\geq 2\Longrightarrow EX=0\ ,n\geq 3\Longrightarrow VarX=\frac{n}{n-2}$$

其概率密度函数形状与标准正态分布相似，但尾部更粗。

12. 自由度为m,n的F分布$F_{m,n}$：

若$X_1\sim\chi_m^2,X_2\sim\chi_n^2$，$X_1,X_2$独立，则：

$$Y=\frac{X_1/m}{X_2/n}\sim F_{m,n}$$

该分布不对称！给定m，n越大偏态越严重。

$Z\sim F_{m,n}\Longrightarrow 1/Z\sim F_{n,m}$；$T\sim t_n\Longrightarrow T^2\sim F_{1.n}$；$F_{m,n}(1-\alpha)=1/F_{n,m}(\alpha)$。性质 3 用于根据$F_{n,m}(\alpha)$求$F_{m,n}(1-\alpha)$。

14. 对$\{X_n\}$ i. i. d. $\sim N(\mu,\sigma^2)$：

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\ ,S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

有：

$$\bar{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$$

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi_{n-1}^2$$

$\bar{X},S^2$独立

$$T=\frac{\sqrt{n}(\bar{X}-\mu)}{S}\sim t_{n-1}$$

若$\{X_m\}$ i. i. d. $\sim N(\mu_1,\sigma_1^2)$，$\{Y_n\}$ i. i. d. $\sim N(\mu_2,\sigma_2^2)$，且$\{X_m\},\{Y_n\}$独立，则：

$$\sigma_1^2=\sigma_2^2=\sigma^2\Longrightarrow\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_\omega}\sqrt{\frac{mn}{m+n}}\sim t_{n+m-2}\ ,S_\omega^2=\frac{1}{m+n-2}\sum_{\substack{i=1\\A=X,Y}}^{m,n}(A_i-A)^2$$

$$\frac{S_1^2}{S_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2}\sim F_{m-1,n-1}$$

*统计三大分布的分布函数表达式无需背诵。

20 级 少转地空 徐小航

2022.2.5