概率论

大数定律

对于随机变量序列$\{X_n\}$,对任意的$\epsilon>0$,有

$$\lim_{n\to\infty}P(|\frac{1}{n}\sum_{k=1}^{n}X_k-\frac{1}{n}\sum_{k=1}^{n}E(X_k)|\geq\epsilon)=0$$

切比雪夫大数定律

定理：对于独立同分布的随机变量$\{X_n\}$,$E(X_n)=\mu,D(X_n)=\sigma^2$,则$\{X_n\}$服从大数定律。

中心极限定理

X_i独立同分布

$$\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}\sim N(0,1)$$

统计量与抽样分布

统计量

统计量是随机变量，且不含任何未知参数

正态分布的一些性质

两个独立的正态分布，和也是正态分布。

正态分布的k阶原点矩
$X \sim N(0,1), E(X^k) = (k-1)!!, k$是偶数; $E(X^k) = 0, k$是奇数

正态总体

χ^2分布

$\chi_n^2 = \sum_{i=1}^{n} X_i^2$，$X_i$独立同分布，$X_i \sim N(0,1)$

性质:

1. $\chi_1^2 \sim \chi^2(n_1), \chi_2^2 \sim \chi^2(n_2)$且$\chi_1^2, \chi_2^2$相互独立，则有$\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$
2. $\chi^2 \sim \chi^2(n) \Rightarrow E(\chi^2) = n, D(\chi^2) = 2n$

t分布

$T = \frac{X}{\sqrt{Y/n}}, X \sim N(0,1), Y \sim \chi^2(n)$,且X和Y相互独立。

关于y轴对称。

F分布

$F = \frac{U/n_1}{V/n_2}, U \sim \chi^2(n_1), V \sim \chi^2(n_2)$且U和V相互独立。$F \sim F(n_1, n_2)$

性质:

1. $F \sim F(n_1, n_2) \Rightarrow \frac{1}{F} \sim F(n_2, n_1)$
2. $T \sim t(n) \Rightarrow T^2 \sim F(1, n)$

上分位点

$P(X > \lambda_\alpha) = \alpha$,$\lambda_\alpha$为X的$\alpha$分位点。

$u_{1-\alpha} = -u_\alpha$
$t_{1-\alpha}(n) = -t_\alpha(n)$
$F_{1-\alpha}(n_1, n_2) = \dfrac{1}{F_\alpha(n_2, n_1)}$

正态总体的样本均值与样本方差的分布

设$X_1, X_2, \ldots, X_n$是来自正态总体$N(\mu, \sigma^2)$的一个样本，则
$\bar{X} \sim N(\mu, \dfrac{\sigma^2}{n})$
$\dfrac{nS_n^2}{\sigma^2} \sim \chi^2(n-1) \quad or \quad \dfrac{(n-1)S_{n-1}^2}{\sigma^2} \sim \chi^2(n-1)$
$\bar{X}$与S^2相互独立

设$X_1, X_2, \ldots, X_n$是来自正态总体$N(\mu, \sigma^2)$的一个样本。则$T = \dfrac{(\bar{X} - \mu)}{S_{n-1}/\sqrt{n}} \sim t(n-1)$
$X_1, .., X_{n_1}$是来自正态总体$N(\mu_1, \sigma_1^2)$的一个样本,$Y_1, .., Y_{n_2}$是来自正态总体$N(\mu_2, \sigma_2^2)$的一个样本，且两样本相

记$S_1^2 = \dfrac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2$
记$S_2^2 = \dfrac{1}{n_2 - 2}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2$
$F = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 2)$

$\bar{X} - \bar{Y} \sim N(\mu_1, \mu_2, (\dfrac{1}{n_1} + \dfrac{1}{n_2})\sigma^2)$

参数估计

矩估计

以样本矩作为总体矩的估计从而得到参数的估计量

有几个参数就求几阶原点矩，然后得到方程组求解。

估计值在参数上面加一个$\hat{\lambda}$

注意：方差和期望之间的转换方式，以及样本方差S_{n-1}^2和S_n^2的不同，这里用的是后者。

无论总体X服从何种分布，总体均值$EX=\mu$, 总体方差$DX=\sigma^2$作为未知参数，其矩估计量一定是样本均值和样本方差，即

$$\hat{\mu}=\overline{X},\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=S_n^2$$

相关系数的矩估计：

$$\rho_{XY}=\frac{cov(X,Y)}{\sqrt{D(X)D(Y)}}=\frac{E((X-E(X))(Y-E(Y)))}{\sqrt{(E((X-EX)^2)E((Y-E(Y))^2)}}$$

然后用$\bar{X}$和S_n^2替换.

矩估计特殊情况

一阶不行时求二阶。

极大似然估计

选择出现样本情况概率最高的参数取值。

求出最大似然函数，对每个参数求偏导可得。

连续性随机变量，将概率密度相乘即可。

离散型随机变量将分布律相乘。

极大似然估计的不变性

设$\hat{\theta}$是θ的极大似然估计，$u(u(\theta))$是θ的函数，且有单值反函数：$\theta=\theta(u)$,则$\hat{u}=u(\hat{\theta})$是$u(\theta)$的极大似然估计。

$\hat{\theta}$是θ的极大似然估计，则$u(\hat{\theta})$是$u(\theta)$的极大似然估计

如果极大似然方程组无解，可以直接考虑极大似然函数，使其最大，求得其最大时参数的取值（例如均匀分布的极大似然估计）

估计量的评选标准

无偏性

$E(\hat{\theta})=\theta$

设总体X方差σ^2未知,σ^2的据估计量

$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{x})^2$是有偏的

$E(S_n^2)=\frac{n-1}{n}\sigma^2\neq\sigma^2$,所以$\hat{\sigma}^2=S_n^2$是有偏的。所以修正样本方差$\frac{n}{n-1}S_n^2=S_{n-1}^2$是无偏的。

有效性

$\hat{\theta_1},\hat{\theta_2}$是$\theta$的无偏估计量，方差小的较为有效。这里指无偏估计量的方差。若$D(\hat{\theta_1})\leq D(\hat{\theta_2})$,则称$\hat{\theta_1}$较$\hat{\theta_2}$有效(对于任意的n)。

一致性

$\hat{\theta_n}=\theta(x_1,\hat{\ldots},x_n),\lim\limits_{n\to\infty}\hat{\theta_n}\to\theta$

$\lim\limits_{n\to\infty}P(|\hat{\theta_n}-\theta|<\epsilon)=1$

样本k阶矩是总体k阶矩的一致性估计量（由大数定律证明）

$\frac{1}{n}\sum_{i=1}^{n}X_i^k\to\frac{1}{n}\sum_{i=1}^{n}E(X_i^k)=E(X^k)$

设θ_n是θ的无偏估计量，且$lim_{n\to\infty}D(\hat{\theta_n})=0$,则$\hat{\theta}$是$\theta$的一致估计量

矩法得到的估计量一般为一致估计量

区间估计

区间估计：根据样本给出未知参数的一个范围，并保证真参数以指定的较大概率属于这个范围。$P(\hat{\theta_1}<\theta<\hat{\theta_2})=1-\alpha$

基本方式是找一个分布（正态分布 or t or χ^2分布 or t分布 or F分布），这个分布中仅包含需要做区间估计得参数

置信区间与置信度

定义：设总体 含未知参数 θ；对于样本$X_1,...,X_n$找出统计量:
$\hat{\theta_i}=\theta_i(X_1,..,X_n),(i=1,2),\hat{\theta_1}<\hat{\theta_2}$
使得$P(\hat{\theta_1}<\theta<\hat{\theta_2})=1-\alpha$，$0<\alpha<1$

称区间$[\hat{\theta_1},\hat{\theta_2}]$为$\theta$的置信区间，$1-\alpha$为该区间的置信度。

正态总体，求均值的μ区间估计

已知方差，估计均值

已知方差σ^2,则$U=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$

$P(\lambda_1\le U\le\lambda_2)=1-\alpha$

代入U得：$[\bar{X}-u_{\alpha/2}\dfrac{\sigma}{\sqrt{n}},\bar{X}+u_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}]$

未知方差，估计均值

$T=\dfrac{\bar{X}-\mu}{S_{n-1}/\sqrt{n}}\sim t(n-1)$

$P(\lambda_1\le T\le\lambda_2)=1-\alpha$

$[\bar{X}-t_{\alpha/2}(n-1)\dfrac{S_n}{\sqrt{n-1}},\bar{X}+t_{\alpha/2}(n-1)\dfrac{S_n}{\sqrt{n-1}}]$

正态总体，求方差σ^2的区间估计

$\chi=\dfrac{nS_n^2}{\sigma^2}\sim\chi^2(n-1)$

使概率对称 $P(\chi^2<\lambda_1)=P(\chi^2>\lambda_2)=\dfrac{\alpha}{2}$

$\chi^2_{1-\frac{\alpha}{2}}\le\dfrac{nS_n^2}{\sigma^2}\le\chi^2_{\frac{\alpha}{2}}(n)$

$[\dfrac{nS_n^2}{\chi^2_{\alpha/2}(n-1)},\dfrac{nS_n^2}{\chi^2_{1-\alpha/2}(n-1)}]$

双正态总体情形

使用的是修正的样本方差S_{n-1}^2

求$\mu_1-\mu_2,\dfrac{\sigma_1^2}{\sigma_2^2}$的区间估计。

σ_1^2,σ_2^2已知，求μ_1,μ_2的置信区间

$\bar{X}\sim N(\mu_1,\dfrac{\sigma_1^2}{n_1}),\bar{Y}\sim N(\mu_2,\dfrac{\sigma_2^2}{n_2})$

$\bar{X}-\bar{Y}\sim N(\mu_1-\mu_2,\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2})$

化为标准正态分布后查表

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

$$[(\bar{X}-\bar{Y})-u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},(\bar{X}-\bar{Y})+u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}]$$

如果$\sigma_1.\sigma_2$位置，但是$\sigma_1=\sigma_2=\sigma,\sigma$未知，取$\sigma^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

方差比$\frac{\sigma_1^2}{\sigma_2^2}$的置信区间

$$F=\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$$

置信区间$(\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)},\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}(n_1-1,n_2-1)})$

单侧置信区间

在单侧置信区间中，都是分位点都是α

对$0<\alpha<1$,样本$X_1,..,X_n$,确定统计量$\hat{\theta}(X_1,...,X_n)$使$P(\theta>\hat{\theta_1})=1-\alpha$,则称$(\hat{\theta_1},+\infty)$是$\theta$的置信度$1-\alpha$的单侧置信区间，$\hat{\theta}$称为单侧置信下限。

类似有$P(\theta<\hat{\theta_2})=1-\alpha$，位单侧置信上限。

例如$X\sim N(\mu,\sigma^2)$，求μ的单侧置信下限，$T=\frac{\bar{X}-\mu}{S_{n-1}/\sqrt{n}}\sim t(n-1)$

分布：求上限从大于入手，求小于从小于入手

求单侧置信区间但未说明求上下限，根据具体问题判断，例如寿命问题求下限

非正态总体均值的区间估计（大样本法）

设$X_1,X_2,...,X_n$为来自均值为μ，方差为σ^2的总体的一组杨本，给定置信度$1-\alpha$，求均值μ的区间估计（注：非正态分布）

当n充分大时，根据中心极限定理有

$$\frac{\sum\limits_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}\to N(0,1)$$

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1)$$

若σ未知，可以用样本标准差S_{n-1}代替

$$U=\frac{\bar{X}-\mu}{S_{n-1}/\sqrt{n}} \sim N(0,1), (\text{近似})$$

主义使用的标准差，要给方差开方

假设检验

简单假设:$H_0: x=a, H_1: x \neq a$

复合假设:$x<a$

u检验法

一般根据拒绝的概率计算出拒绝域，检查样本是否在拒绝域之中。

第一步：统计假设

第二步：H_0成立时，考虑一个统计量U。（统计量及分布）

第三步：由$P(|U|>u_{\alpha/2})=\alpha$,得到拒绝域

第四步：根据样本得到U的观测值

第五步：得出结论

假设检验基本步骤

1. 根据问题提出原假设H_0和对立假设H_1
2. 构造一个合适的统计量（往往由参数估计而来），并在H_n成立的条件下推导出该统计量的分布
3. 给出小概率α，确定临界值和拒绝域W
4. 由样本算出统计量的观察值
5. 若观察值落在拒绝域W，则拒绝H_0,若在接受域，接受H_0

正态总体均值的假设检验

单个正态总体均值的假设检验

σ^2 已知(u检验法)

$$U=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1)$$

拒绝域为$W=\{|U|\geq u_{\alpha/2}\}$

单边检验

$H_0:\mu=\mu_0,H_1:\mu>\mu_0$

拒绝域$W=\{\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\geq u_\alpha\}$

$H_0:\mu=\mu_0,H_1:\mu<\mu_0$

拒绝域$W=\{\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\leq -u_\alpha\}$

σ^2 未知（t检验法）

$$T=\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\sim t(n-1)$$

拒绝域$W=\{|T|\geq t_{\alpha/2}(n-1)\}$

对于单边检验，判断大于号还是小于号后，使用的$t_\alpha(n-1)$

双正态总体的情形

σ_1,σ_2 已知

$$U=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

拒绝域(双边)$W=\{|U|\geq u_{\alpha/2}\}$

单边($H_1:\mu_1<\mu_2$时)$W=\{U\leq -u_\alpha\}$

单边($H_1:\mu_1>\mu_2$时)$W=\{U\geq u_\alpha\}$

σ_1,σ_2 未知但相等，$\sigma_1=\sigma_2=\sigma$

$$S_w=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\text{代替}\sigma$$

$$T=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

正太总体方差的假设检验

单正太总体

$H_0:\sigma^2=\sigma_0^2,H_1:\sigma\neq\sigma_0^2$

$$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1)$$

$W=\{\chi^2\leq\chi^2_{1-\alpha/2}(n-1)\}\cup\{\chi^2\geq\chi^2_{\alpha/2}(n-1)\}$

双正太总体（F检验法）

$H_0:\sigma_1^2=\sigma_2^2,H_1:\sigma_1^2\neq\sigma_2^2$

$$F=\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$$

在假设H_0成立的条件下，$F=\dfrac{S_1^2}{S_2^2}\sim F(n_1-1,n_2-1)$

拒绝域$W=\{F\leq F_{1-\alpha/2}(n_1-1,n_2-1)\}\cup\{F\geq F_{\alpha/2}(n_1-1,n_2-1)\}$

非正太总体均值的检验

一个总体均值的检验

假设X为任意总体，$EX=\mu,DX=\sigma^2$,$X_1,\ldots,X_n$是一组样本,
$\bar{X}$是样本均值,S^2是修正的样本方差，μ_0是已知参数，记$U=\dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$或$U=\dfrac{\bar{X}-\mu_0}{S.\sqrt{n}}$，当n充分大时，统计量U近似服从标准正态分布。

两个正态总体的检验

$X_1,\ldots,X_m,S_1^2,Y_1,Y_2,\ldots,Y_n,S_2^2$,修正样本方差

$$U=\frac{\bar{X}-\bar{Y}}{\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}}\text{或}U=\frac{\bar{X}-\bar{Y}}{\sqrt{\dfrac{S_1^2}{m}+\dfrac{S_2^2}{n}}}$$

拟合优度检验（分布拟合优度检验）(不考)

不知道总体的分布类型

$H_0:F(x)=F_0(x,\theta)$，F_0为某个已知的分布函数，$\theta=(\theta_1,\ldots,\theta_r)$为未知参数

利用事件的频率与概率之间的偏差构造检验统计量

皮尔逊统计量

$H_0: O(X=x_i), i=1,2,\dots,k$

(1)计算$X_1,\dots,X_n$中取x_i的实际频数$n_i=\{X_1,\dots,X_n$中取x_i的个数$\}$

(2)计算实际频数与理论频数的偏差平方和$\chi^2=\sum\limits_{i=1}^{k}\dfrac{(n_i-np_i)^2}{np_i}\sim\chi^2(k-1)$

(3)拒绝域为$W=\{\chi^2\geq\chi^2_\alpha(k-1)\}$

一般的假设检验问题

1. 将样本空间分为k个互不相交的事件$A_i,A_2,\dots,A_k$
2. 计算每个事件A_i上的理论频数，若参数θ未知，先算出θ的极大似然估计$\hat{\theta}$，计算理论上样本落在事件A_i中的概率$\hat{p_i}=P(X\in A_i|\theta=\hat{\theta}),i=1,2...,k$，最后得到每个事件的理论频数$n\hat{p_i}$
3. 计算$X_1,\dots,X_n$中取x_i的实际频数$n_i=\{X_1,\dots,X_n$中取x_i的个数$\}$
4. 计算实际频数与理论频数的偏差平方和$\chi^2=\sum\limits_{i=1}^{k}\dfrac{(n_i-n\hat{p_i})^2}{n\hat{p_i}}\sim\chi^2(k-1)$
5. 拒绝域为$W=\{\chi^2\geq\chi^2_\alpha(k-1)\}$

注意：通常要求$n\geq 50$,将样本空间划分为事件，要求每个事件的理论频数不应太小

期中之前的内容

基本概念

条件概率:$P(B|A)=\dfrac{P(AB)}{P(A)}\Leftrightarrow P(AB)=P(A)P(B|A)=P(B)P(A|B)$

全概率公式与贝叶斯公式:

全概率公式：A_i是Ω的一个划分，$P(B)=\sum\limits_{i=1}^{n}P(A_i)P(B|A_i)$

贝叶斯公式：A_i是Ω的一个划分，$P(A_j|B)=\dfrac{P(A_j)P(B|A_j)}{\sum\limits_{i=1}^{n}P(A_i)P(B|A_i)}$

分布函数

二项分布的峰值：当$(n+1)p$是整数时。$k_0=(n+1)p-1$或$(n+1)p$，当$(n+1)p$不是整数时，$k_0=[(n+1)p]$

若随机变量$X\sim B(n,p)$，则当n充分大，p充分小时，令$\lambda=np$，则有
$P\{X=k\}=C_n^k p^k(1-p)^{n-k}\approx\dfrac{\lambda^k}{k!}e^{-\lambda}$

离散型：几何分布:$X\sim g(p)$

连续型：均匀分布$X\sim U[a,b]$，$E(x)=\dfrac{a+b}{2},D(x)=\dfrac{(b-a)^2}{12}$

二项分布$X\sim B(n,p)$,$E(x)=np,D(x)=npq$

超几何分布$X\sim H(n,N,M)$

泊松分布:$p\{X=k\}=\dfrac{\lambda^k}{k!}e^{-\lambda},k=0,1,2,\ldots,\lambda>0$，记作$X\sim P(\lambda)$,$E(x)=\lambda,D(x)=\lambda$

指数分布（无记忆性）$X\sim E(\lambda)E(x)=\dfrac{1}{\lambda},D(x)=\dfrac{1}{\lambda^2}$

正态分布$p(x)=\dfrac{1}{\sqrt{2\pi}\sigma}e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}$,$X\sim N(\mu,\sigma^2)$.

若$X\sim N(\mu,\sigma^2),Y=\dfrac{X-\mu}{\sigma}\sim N(0,1)$，以及$3\sigma$原理

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

$F(x)=P(X\leq x)(-\infty<x<+\infty)$称为随机变量X的随机变量。

随机变量函数的分布：

对于连续型随机变量，其密度函数为$p(x),y=g(x)$是x的连续函数，$Y=g(x)$是连续性随机变量。求$Y=g(X)$的密度函数$p_Y(y)$

1. 分布函数法：先求$Y=g(X)$的分布函数，再求导。
2. 公式法。

随机向量的函数的分布：

同样求出对应的分布函数，然后求导，如
$Z=max\{X,Y\},F(Z<z)=P(X<z,y<Z)$，然后积分

随机变量的数字特征（期望，方差）

期望的性质：

期望的线性性质：不要求独立，$E(aX+bY)=aE(X)+bE(y)$

若X，Y相互独立，$E(XY)=E(X)E(Y)$

方差的性质：

$D(aX+b)=a^2D(X)$

$D(X+Y)=D(x)+D(Y)+E((X-E(X))(Y-E(Y)))$，X和Y独立时
$D(X+Y)=D(X)+D(Y)$

切比雪夫不等式$P(|X-EX|\geq\epsilon)\leq\dfrac{D(X)}{\epsilon^2}$

注意样本方差和总体方差的区别

协方差

$cov(X,Y)=E(XY)-E(X)E(Y)$

相关系数：$\rho=\dfrac{cov(X,Y)}{\sqrt{D(X)D(Y)}}$,$|\rho|=1\Leftrightarrow P(cX+aY=b)=1$,X和Y以概率1成线性关系。

X和Y不相关
$\Leftrightarrow\rho_{XY}=0\Leftrightarrow cov(X,Y)=0\Leftrightarrow E(XY)=E(X)E(Y)\Leftrightarrow D(X+Y)=D(X)+D(Y).$