概率论与数理统计

何志坚

华南理工大学数学学院

hezhijian@scut.edu.cn

第六章：数理统计的基本概念

数理统计是一门什么样的学科?

它使用概率论和其它数学方法，研究怎样收集（通过试验和观察）带有随机误差的数据，并在设定的模型（称为统计模型）之下，对这种数据进行分析（称为统计分析），以对所研究的问题作出推断（称为统计推断）。由于所收集的统计数据（资料）只能反映事物的局部特征，数理统计的任务就在于从统计资料所反映的局部特征以概率论作为理论基础去推断事物的整体特征。

- 数据收集
- 模型假定
- 数据分析/推断

本质：由局部（有限样本）推断整体（总体）

案例

- 数据收集：用n部手机进行测试，记录通话时间$X_1, \ldots, X_n$
- 模型假定：假设通话时间X服从正态分布$N(\mu, \sigma^2)$
- 数据分析：通过观测数据$x_1, \ldots, x_n$估计μ以及σ^2

目录

总体与样本

总体：研究对象X的概率分布，$X \sim F_X(\cdot, \theta)$，其中$\theta \in \Theta$为参数，$\Theta$为参数空间。

样本：从总体中随机抽取n个个体$X_1, X_2, \ldots, X_n$，称为总体的样本，n称为样本容量，简称样本量。样本具有二重性：

- 抽取之前无法预知它们的数值，因此$(X_1, \ldots, X_n)$为n维随机向量。
- 抽取后样本为具体的数，用小写字母$(x_1, \ldots, x_n)$表示，称为样本观测值。

简单随机样本：X_i相互独立，与X同分布。本课程只讨论简单随机样本。以后简称样本。

样本的分布

定义： 设总体X具有分布函数$F(x)$，则样本$(X_1, \ldots, X_n)$的联合分布函数为

$$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} F(x_i).$$

- 若X为离散型随机变量，分布列为$p_i = P(X = x_i)$, 则样本的联合分布列为

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p_i.$$

- 如X为连续型随机变量，密度函数为$p(x)$, 则样本的联合密度函数为

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i).$$

正态样本的分布

例：如果$X \sim N(\mu, \sigma^2)$, 其样本$(X_1, \ldots, X_n)$的联合密度函数为：

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

统计量

样本是总体的反映，但样本所含信息不能直接用于解决我们所要研究的问题，而需要把样本所含的信息进行数学上的加工使其浓缩起来，从而解决我们的问题。为此，数理统计学往往构造一个合适的依赖于样本的函数，我们称之为统计量。

定义：如果$(X_1, \ldots, X_n)$为来自总体的样本，若样本函数

$$T = T(X_1, \ldots, X_n)$$

中不含有任何未知参数，则称T为统计量。统计量的分布称为抽样分布。

例

例： 设$X_1, \ldots, X_n$为来自$X \sim N(\mu, \sigma^2)$的样本, 若μ已知，σ未知，判断以下样本函数是否为统计量

- $$T_1 = \frac{\sqrt{n}(\sum_{i=1}^{n} X_i - \mu)}{\sigma}$$
- $$T_2 = \min(X_1, \ldots, X_n)$$

常见的统计量

设$X_1, \dots, X_n$为来自X的样本。

- 样本均值:

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

- 样本方差:

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

- 修正样本方差:

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

- 样本标准差:

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

常见的统计量

设$X_1, \dots, X_n$为来自X的样本。

- 样本k阶原点矩：

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

- 样本k阶中心矩：

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

- 顺序统计量：

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$$

  其中$X_{(1)} = \min\{X_1, \dots, X_n\}$，$X_{(n)} = \max\{X_1, \dots, X_n\}$，$X_{(k)}$为$X_1, \dots, X_n$的递增排序的第$k$位。$X_{(n)} - X_{(1)}$称为样本极差。

- 样本中位数：

$$\tilde{X} = \begin{cases} X_{(\frac{n+1}{2})}, & n\text{为奇数} \\ (X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)})/2, & n\text{为偶数} \end{cases}$$

目录

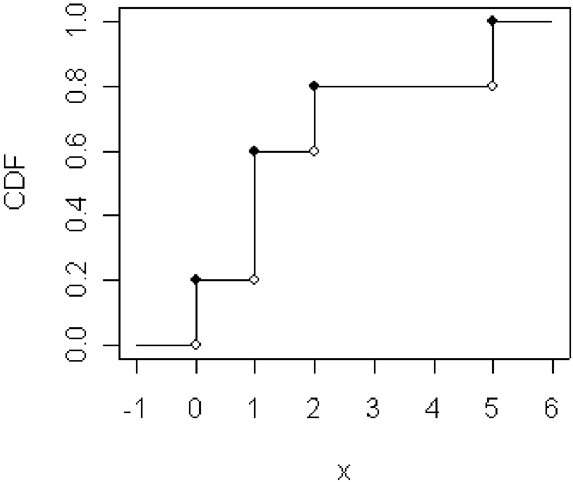
经验分布函数

目标： 通过样本的观测值构造一种函数来近似总体的分布函数

定义： 设总体X的样本$(X_1,\dots,X_n)$的一次观测值$(x_1,\dots,x_n)$, 并将它们由小到大排列$x_{(1)}\le x_{(2)}\le\cdots\le x_{(n)}$, 经验分布函数(或称样本分布函数)定义为

$$F_n^X(x)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{x_i\le x\}=\begin{cases}0, & x<x_{(1)}\\ 1/n, & x_{(1)}\le x<x_{(2)}\\ 2/n, & x_{(2)}\le x<x_{(3)}\\ & \vdots\\ k/n, & x_{(k)}\le x<x_{(k+1)}\\ & \vdots\\ 1, & x>x_{(n)}\end{cases}.$$

经验分布函数示意图



经验分布函数的性质

固定的x和n，$F_n^X(x)$表示事件$\{X \le x\}$的频率，由强大数定律知，

$$F_n^X(x) \stackrel{\text{a.s}}{\to} P(X \le x) = F_X(x),$$

即

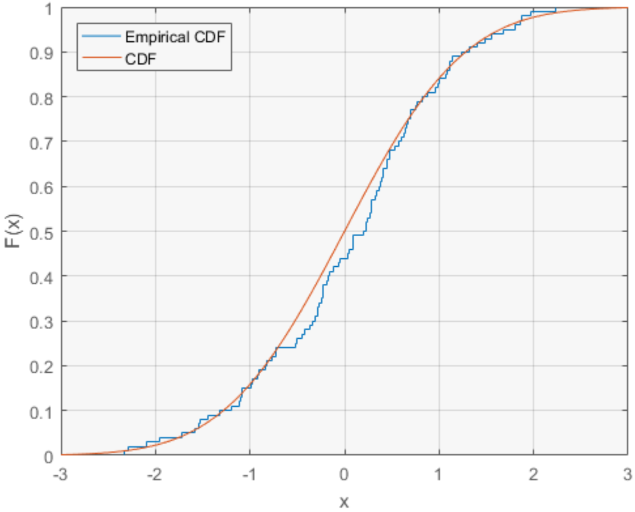
$$P\left(\lim_{n\to\infty} F_n^X(x) = F_X(x)\right) = 1.$$

格里汶科定理给出更强的结果（几乎处处一致收敛）：

$$P\left(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left|F_n^X(x) - F_X(x)\right| = 0\right) = 1.$$

注： 由此可见，当n相当大时，经验分布函数$F_n^X(x)$是母体分布函数$F_X(x)$的一个良好近似。数理统计学中一切都以样本为依据，其理由就在于此。

经验分布函数示意图



目录

抽样分布

统计量的分布称为抽样分布。样本均值和样本方差的数字特征:

定理: 设$X_1,\ldots,X_n$为来自X的样本，$\mathbb{E}[X]=\mu$, $\mathrm{Var}[X]=\sigma^2$, 则

- 样本均值的期望和方差分别为

$$\mathbb{E}[\bar{X}]=\mu,\ \ \mathrm{Var}[\bar{X}]=\frac{\sigma^2}{n}$$

- 样本方差和修正样本方差的期望分别为

$$\mathbb{E}[S_n^2]=\frac{n-1}{n}\sigma^2,\ \ \mathbb{E}[S_n^{*2}]=\sigma^2.$$

证:

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

$$\mathbb{E}[S_n^2]=\mathbb{E}[X^2]-\mathbb{E}[\bar{X}^2]=\sigma^2-\frac{\sigma^2}{n}=\frac{n-1}{n}\sigma^2$$

三种重要的概率分布

1. χ^2分布
2. t分布
3. F分布

χ^2分布

定义：设$X_i \overset{\text{iid}}{\sim} N(0,1), i=1,\ldots,n$，则称随机变量

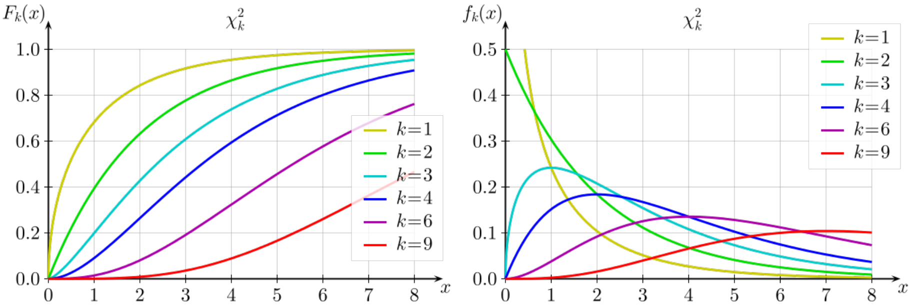
$$X = X_1^2 + \cdots + X_n^2$$

服从自由度为n的卡方分布，记为$X \sim \chi^2(n)$(或者写成χ_n^2). 其密度函数为

$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(n/2)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \leq 0, \end{cases}$$

其中$\Gamma(s) = \int_0^\infty x^{s-1} e^{-x} \, \mathrm{d}x (s > 0)$. 容易发现$\Gamma(n) = (n-1)!$.

χ^2分布的分布函数与密度函数曲线



χ^2分布的性质

定理： 如果$X \sim \chi^2(n)$, 则

$$\mathbb{E}[X] = n, \ \mathrm{Var}[X] = 2n.$$

如果$X \sim \chi^2(n)$, $Y \sim \chi^2(m)$且它们独立，则

$$X + Y \sim \chi^2(n+m)(\text{可加性}).$$

注： 英国统计学家费歇（R.A.Fisher）曾证明，当$n \to \infty$时，

$$\frac{X-n}{\sqrt{2n}} \xrightarrow{\mathrm{d}} N(0,1).$$

这就是中心极限定理的应用。

学生t分布

定义：设$X \sim N(0,1)$, $Y \sim \chi^2(n)$, 且它们独立，则称随机变量

$$T = \frac{X}{\sqrt{Y/n}}$$

服从自由度为n的学生t分布（简称t分布），记为$T \sim t(n)$. 其密度函数为

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

休息一会儿

1954 年该厂开始出版《吉尼斯世界纪录大全》.

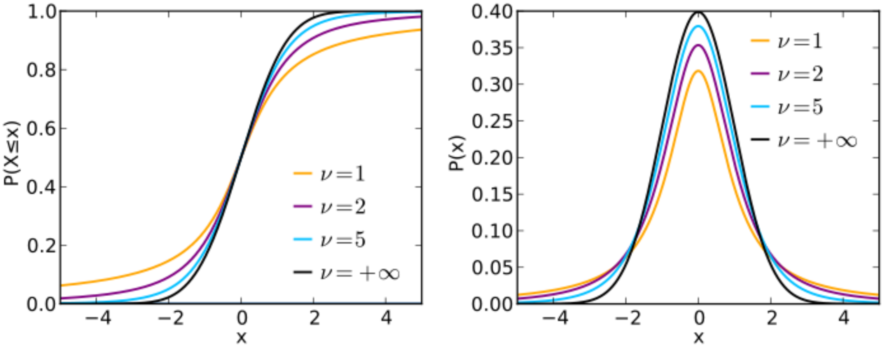
小故事: t 检验、啤酒、“学生”与 威廉·戈瑟特

1899 年, 由于爱尔兰都柏林的吉尼斯啤酒厂热衷于聘用剑桥、牛津的优秀毕业生, 学化学的牛津毕业生威廉·戈瑟特 (William Gosset, 1876—1937) 到该厂就职, 希望将他的生物化学知识用于啤酒生产过程. 为降低啤酒质量监控的成本, 戈瑟特发明了 t 检验法, 1908 年在 *Biometrika* 发表. 为防止泄漏商业机密, 戈瑟特发表文章时用了笔名“学生”, 于是该方法被称为“学生氏 t 检验” (Student's t-test).

吉尼斯啤酒厂是一家很有远见的企业, 为保持技术人员的高水准, 该厂像高校一样给予技术人员“学术假”, 1906—1907 年戈瑟特得以到“统计学之父”卡尔·皮尔逊 (Karl Pearson, 1857—1936) 教授在伦敦大学学院 (University College London, 简称 UCL) 的实验室访问学习. 因此, 很难说 t 检验法是戈瑟特在啤酒厂还是在 UCL 访学期间提出的, 但“学生”与戈瑟特之间的联系是被 UCL 的统计学家们发现的, 尤其因为皮尔逊教授恰是 *Biometrika* 的主编.

推荐阅读《女士品茶》

t分布的分布函数与密度函数曲线



注： 密度函数是偶函数

t分布的性质

设$X \sim t(n)$.

- 当$n=1$时，t分布成为柯西分布.

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

  此时，方差和期望都不存在。

- 当$n\geq 2$时, $\mathbb{E}[X]=0$. 对于k阶矩，$\mathbb{E}[X^k]$存在当且仅当$k<n$.
- 当$n\geq 3$时，t分布的方差才存在，即

$$\mathrm{Var}[X]=\frac{n}{n-2}.$$

- $$\lim_{n\to\infty} f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

  当$n>45$时,可以认为t分布与$N(0,1)$接近。

例

例： 设X, Y这两个总体相互独立，$X \sim N(0, 16)$，$Y \sim N(0, 9)$，$X_1, \ldots, X_9$为X的样本，$Y_1, \ldots, Y_{16}$为Y的样本，求统计量

$$Z = \frac{X_1 + \cdots + X_9}{\sqrt{Y_1^2 + \cdots + Y_{16}^2}}$$

所服从的分布。

证： $X_1 + \cdots + X_9 \sim N(0, 12^2)$，所以$(X_1 + \cdots + X_9)/12 \sim N(0, 1)$. 又因为，$(Y_1^2 + \cdots + Y_{16}^2)/9 \sim \chi^2(16)$. 所以，

$$Z = \frac{(X_1 + \cdots + X_9)/12}{\sqrt{\frac{(Y_1^2 + \cdots + Y_{16}^2)/9}{16}}} \sim t(16)$$

F分布

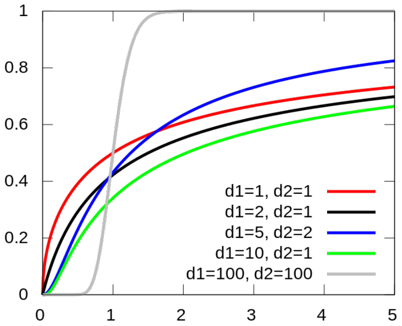
定义：设$X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, 且X, Y相互独立，则随机变量

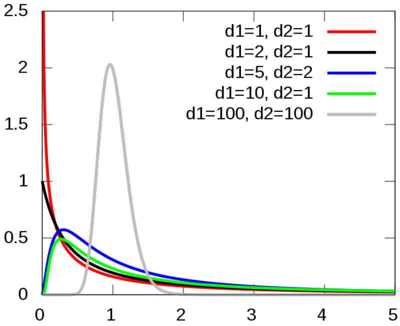
$$Z = \frac{X/m}{Y/n}$$

称为服从第一自由度为m、第二自由度为n的F分布，记$Z \sim F(m, n)$. 其密度函数为

$$f(x) = \begin{cases} \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} x^{\frac{m}{2}-1}(1 + mx/n)^{-(m+n)/2}, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

F分布的分布函数与密度函数曲线





F分布的性质

- 如果$Z \sim F(m, n)$, 则$1/Z \sim F(n, m)$.
- 如果$T \sim t(n)$, 则$T^2 \sim F(1, n)$.

分位数

定义：对于任意$\alpha \in (0,1)$, α分位数x_α满足$F(x_\alpha) = \alpha$.

- 标准正态分布分位数记为u_α, 满足$P(|Z| \le u_{1-\alpha/2}) = 1-\alpha$
- t分布分位数记为$t_\alpha(n)$, 满足$P(|T| \le t_{1-\alpha/2}(n)) = 1-\alpha$
- χ^2分布分位数记为$\chi^2_\alpha(n)$
- F分布分位数记为$F_\alpha(m,n)$

注1：在分位点表中对于标准正态分布、t分布和F分布只能查到$\alpha > 1/2$的分位数，需利用以下对称性间接查$\alpha < 1/2$的分位数：

$$u_\alpha = -u_{1-\alpha},\ t_\alpha(n) = -t_{1-\alpha}(n),\ F_\alpha(m,n) = \frac{1}{F_{1-\alpha}(n,m)}$$

注2：对于$t(n)$分布，由于当$n \to \infty$时，其极限分布为$N(0,1)$, 所以自由度n比较大时，$t_\alpha(n) \approx u_\alpha$.

注3：若$X \sim \chi^2(n)$分布，由于当$n \to \infty$时，$(X-n)/\sqrt{2n} \xrightarrow{d} N(0,1)$, 所以自由度$n$比较大时，$\chi^2_\alpha(n) \approx u_\alpha\sqrt{2n} + n$.

分位数

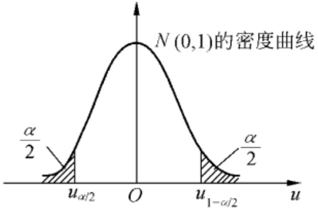


图 6.3.4. $N(0,1)$ 分布分位点示意图.

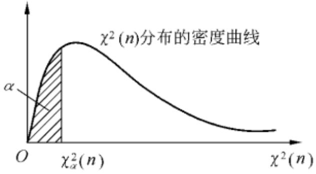


图 6.3.5. $\chi^2(n)$ 分布分位点示意图.

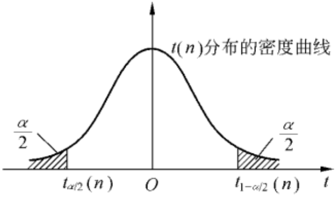


图 6.3.6. $t(n)$ 分布分位点示意图.

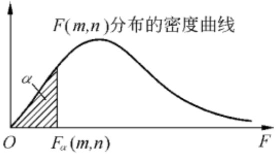


图 6.3.7. $F(m,n)$ 分布分位点示意图.

例

例：查表求以下分位数：

- $\chi^2_{0.99}(10)$，$\chi^2_{0.05}(10)$和$\chi^2_{0.95}(60)$
- $t_{0.95}(10)$，$t_{0.1}(10)$和$t_{0.9}(50)$
- $F_{0.99}(5,4)$，$F_{0.05}(3,7)$

解：

$$\chi^2_{0.99}(10) = 23.209,\ \chi^2_{0.05}(20) = 10.851,\ \chi^2_{0.95}(60) = u_{0.95}\sqrt{2\times 60} + 60 = 78.018$$

$$t_{0.95}(10) = 1.8125,\ t_{0.1}(10) = -t_{0.9}(10) = -1.3722\ t_{0.9}(50) = u_{0.9} = 1.28$$

$$F_{0.99}(5,4) = 5.5,\ F_{0.05}(3,7) = 1/F_{0.95}(7,3) = 1/8.89 = 0.112$$

单个正态总体的抽样分布

设$(X_1,\dots,X_n)$为来自总体$X\sim N(\mu,\sigma^2)$的样本，则有

$$\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2}=\sum_{i=1}^n\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi^2(n)$$

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

注意到

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

$$\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2}=\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\sigma^2}+\frac{(\bar{X}-\mu)^2}{\sigma^2/n}$$

单个正态总体的抽样分布

定理：设$(X_1,\ldots,X_n)$为来自总体$X\sim N(\mu,\sigma^2)$的样本，则

- 样本均值$\bar{X}$与样本方差S_n^2相互独立，
- $\bar{X}\sim N(\mu,\sigma^2/n)$，
- $$\frac{nS_n^2}{\sigma^2}=\frac{\sum_{i=1}^n(X_i-\bar{X})^2}{\sigma^2}\sim\chi^2(n-1).$$

证：由正态分布的可加性得$\bar{X}\sim N(\mu,\sigma^2/n)$. 下证独立性，令$Y_i=X_i-\bar{X}$. 则有$\sum_{i=1}^n Y_i=0$. 所以，$S_n^2$为$Y_1,\ldots,Y_{n-1}$的函数。由于$(\bar{X},Y_1,\ldots,Y_{n-1})$服从$n$维正态分布，所以只需验证$\mathrm{Cov}(\bar{X},Y_i)=\mathrm{Cov}(\bar{X},X_i-\bar{X})=0$即可。

当$n=2$时，

$$\frac{nS_n^2}{\sigma^2}=\left(\frac{X_1-X_2}{\sqrt{2}\sigma}\right)^2\sim\chi^2(1)$$

假设$n=k$时结论成立，当$n=k+1$时，

$$\frac{(k+1)S_{k+1}^2}{\sigma^2}=\frac{kS_k^2}{\sigma^2}+\frac{k(X_{k+1}-\bar{X}_k)^2}{(k+1)\sigma^2}=\chi^2(k-1)+\chi^2(1)=\chi^2(k)$$

推论一：标准正态总体

推论：设$(X_1, \ldots, X_n)$为来自总体$X \sim N(0,1)$的样本，则

- 样本均值$\bar{X}$与样本方差S_n^2相互独立，
- $$\bar{X} \sim N(0, \frac{1}{n}),$$
- $$nS_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

推论二

推论：设$(X_1,\dots,X_n)$为来自总体$X\sim N(\mu,\sigma^2)$的样本，则

$$\frac{\sqrt{n-1}(\bar{X}-\mu)}{S_n}=\frac{\sqrt{n}(\bar{X}-\mu)}{S_n^*}\sim t(n-1)$$

证明:令

$$Y=\frac{\bar{X}-\mu}{\sqrt{\mathrm{Var}[\bar{X}]}}=\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\sim N(0,1)$$

因为，$\bar{X}$与S_n^2独立，所以Y与$Z=\frac{nS_n^2}{\sigma^2}$独立。由于$Z\sim\chi^2(n-1)$, 所以

$$\frac{Y}{\sqrt{Z/(n-1)}}=\frac{\sqrt{n-1}(\bar{X}-\mu)}{S_n}\sim t(n-1)$$

例

例： 设总体$X \sim N(\mu, 4)$, 有样本$X_1, \ldots, X_n$, 当样本容量n至少多大时，才使得$P(\left|\bar{X} - \mu\right| \leq 0.01) \geq 0.95$?

解： $\bar{X} \sim N(\mu, 4/n)$. 所以$Z = (\bar{X} - \mu)/(2/\sqrt{n}) \sim N(0, 1)$.

$$P(\left|\bar{X} - \mu\right| \leq 0.01) = P(|Z| \leq 0.05\sqrt{n}) \geq 0.95.$$

所以，$0.05\sqrt{n} \geq u_{0.975}$，$n \geq 400u_{0.975}^2 = 400 \times 1.96^2 = 1536.6 \approx 1537$.

两个正态总体下的抽样分布

定理：设总体$X \sim N(\mu_1, \sigma_1^2)$, $(X_1, \ldots, X_m)$为其样本，样本均值为$\bar{X}$, 样本方差为S_{1m}^2; 另有与X独立的总体$Y \sim N(\mu_2, \sigma_2^2)$, $(Y_1, \ldots, Y_n)$为其样本，样本均值为$\bar{Y}$, 样本方差为S_{2n}^2. 则

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

$$\frac{mS_{1m}^2/\sigma_1^2/(m-1)}{nS_{2n}^2/\sigma_2^2/(n-1)} = \frac{S_{1m}^{*2}\sigma_2^2}{S_{2n}^{*2}\sigma_1^2} \sim F(m-1, n-1).$$

证明：利用单个正态整体下抽样分布的结论：$\bar{X} \sim N(\mu_1, \sigma_1^2/m)$, $\bar{Y} \sim N(\mu_2, \sigma_2^2/n)$, 且$\bar{X}, \bar{Y}$独立，所以，$\bar{X}-\bar{Y} \sim N(\mu_1-\mu_2, \sigma_1^2/m+\sigma_2^2/n)$. 又$mS_{1m}^2/\sigma_1^2 \sim \chi^2(m-1)$, $nS_{2n}^2/\sigma_2^2 \sim \chi^2(n-1)$, 且两者独立，根据$F$分布的定义可得第二个结论。

两个正态总体（相同方差）下的抽样分布

定理： 设总体$X \sim N(\mu_1, \sigma^2)$, $(X_1, \ldots, X_m)$为其样本，样本均值为$\bar{X}$, 样本方差为S_{1m}^2; 另有与X独立的总体$Y \sim N(\mu_2, \sigma^2)$, $(Y_1, \ldots, Y_n)$为其样本，样本均值为$\bar{Y}$, 样本方差为S_{2n}^2. 则

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w\sqrt{1/m + 1/n}} \sim t(m + n - 2),$$

其中$S_w = \sqrt{(mS_{1m}^2 + nS_{2n}^2)/(m + n - 2)}$.

证明： 令

$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2/m + \sigma^2/n}} \sim N(0, 1)$$

因为$mS_{1m}^2/\sigma^2 \sim \chi^2(m - 1)$, $nS_{2n}^2/\sigma^2 \sim \chi^2(n - 1)$ 且它们独立，所以

$$V = mS_{1m}^2/\sigma^2 + nS_{2n}^2/\sigma^2 \sim \chi^2(n + m - 2).$$

又U, V独立，所以$U/\sqrt{V/(m + n - 2)} \sim t(m + n - 2)$.

例1：$\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$的值已知

例： 设总体$X \sim N(6,1)$，$Y \sim N(5,1)$有$m = n = 10$两个独立样本，求$\bar{X} - \bar{Y}$小于1.3的概率。

解： 已知

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} = \frac{(\bar{X} - \bar{Y}) - 1}{\sqrt{1/10 + 1/10}} \sim N(0,1)$$

所以，

$$P(\bar{X} - \bar{Y} < 1.3) = P(Z < (1.3 - 1)/\sqrt{1/5}) = P(Z < 0.67) = \Phi(0.67) = 0.7486.$$

例2：μ_1, μ_2的值已知，且$\sigma_1^2 = \sigma_2^2$

例：设总体$X \sim N(6, \sigma^2)$，$Y \sim N(5, \sigma^2)$有$m = n = 10$两个独立样本，$S_{1m}^{*2} = 0.9130$，$S_{2n}^{*2} = 0.9816$，求$\bar{X} - \bar{Y}$小于1.3的概率。

解： 已知

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{1/m + 1/n}} \sim t(m + n - 2) = t(18),$$

其中$S_w = \sqrt{((m-1)S_{1m}^{*2} + (n-1)S_{2n}^{*2})/(m+n-2)} = 0.9733$. 所以，

$$P(\bar{X} - \bar{Y} < 1.3) = P(Z < (1.3 - 1)/(\sqrt{1/5} \times 0.9733)) = P(Z < 0.6884) = 0.74.$$

例3：σ_1^2, σ_2^2的值已知，但μ_1, μ_2的值未知

例：设总体$X \sim N(\mu, 3)$，$Y \sim N(\mu, 5)$有$m = 10$，$n = 15$两个独立样本，求两个修正样本方差之比S_{1m}^{*2}/S_{2n}^{*2}大于1.272的概率。

解： 已知

$$Z = \frac{S_{1m}^{*2}\sigma_2^2}{S_{2n}^{*2}\sigma_1^2} \sim F(m-1, n-1) = F(9, 14)$$

所以，

$$P(S_{1m}^{*2}/S_{2n}^{*2} > 1.272) = P(Z > 1.272 \times 5/3) = P(Z > 2.12) = 0.1.$$

顺序统计量分布

定理： 若$X_1,\dots,X_n$独立同分布，分布函数和密度函数分别为$F_X(x), f_X(x)$.
则$X_{(1)}=\min(X_1,\dots,X_n)$的分布函数和密度函数分别

$$\begin{cases} F_{X_{(1)}}(x) = 1-(1-F_X(x))^n \\ f_{X_{(1)}}(x) = n(1-F_X(x))^{n-1}f_X(x). \end{cases}$$

$X_{(n)}=\max(X_1,\dots,X_n)$的分布函数和密度函数分别

$$\begin{cases} F_{X_{(n)}}(x) = F_X(x)^n \\ f_{X_{(n)}}(x) = nF_X(x)^{n-1}f_X(x). \end{cases}$$

更一般地，

$$f_{X_{(k)}}(x) = \frac{n!}{(n-k)!(k-1)!}F_X(x)^{k-1}(1-F_x(x))^{n-k}f_X(x), k=1,\dots,n.$$

作业

第六章习题：2, 3, 4, 8, 11, 13, 17, 18

6月6日（周三）交！