第六章 样本及抽样分布

§3 抽样分布

统计推断的两个重要基础

收集数据——从总体$X \sim F(x)$抽取样本$X_1, X_2, \ldots, X_n$

加工整理数据——统计量

§3 抽样分布

三种重要的统计学分布

- χ^2分布主要是用于列联分析
- t分布主要是用于小样本分析
- F分布主要是用于方差分析

四个重要的抽样定理

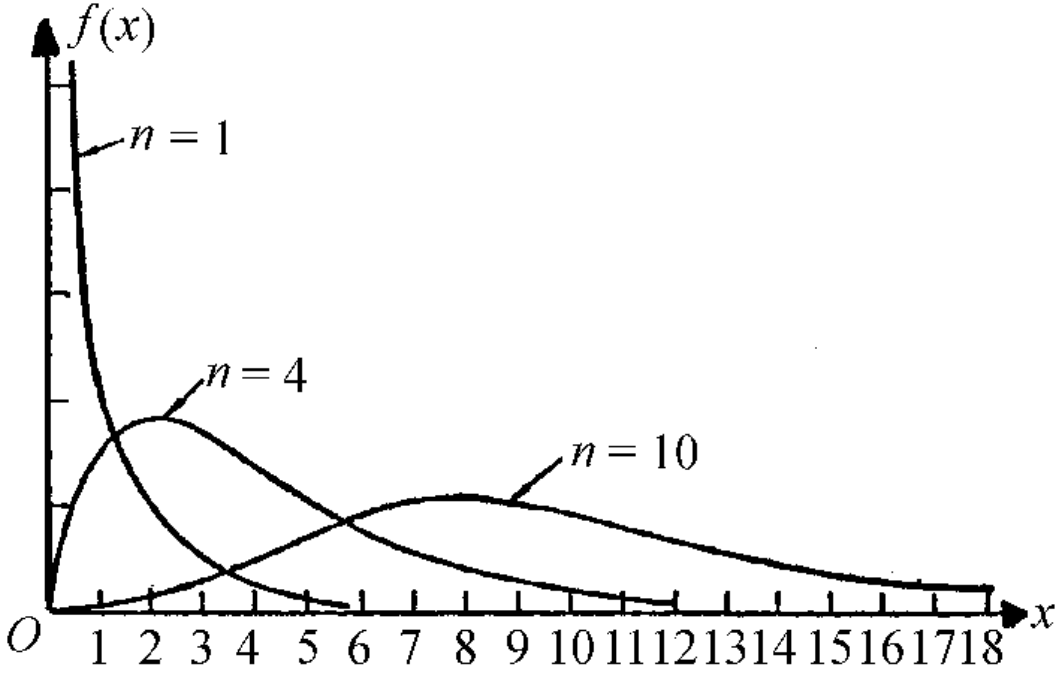
χ^2分布

设$X_1, X_2, \dots, X_n$ i.i.d $\sim N(0,1)$, 则称统计量

$$\chi^2 = \sum_{i=1}^{n} X_i^2$$

服从的分布为自由度为n的χ^2(卡方)分布，记为：

$\chi^2 \sim \chi^2(n)$



χ^2分布

这里自由度是指右端包含的独立变量的个数，$\chi^2(n)$的概率密度为

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

关于Γ函数的回顾(详见第二章内容)

$$\Gamma(\alpha)=\int_0^{+\infty}x^{\alpha-1}e^{-x}dx$$

$$\Gamma(\alpha+1)=\alpha\Gamma(\alpha)\ ,\ \Gamma(1)=1\ ,\ \Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$$

若α为正整数，记为n，则 $\Gamma(n)=(n-1)!$

χ^2分布的推导：

由第二章第5节例5有 $Y_i = X_i^2 \sim \Gamma\left(\frac{1}{2}, 2\right)$，其概率密度为

$$f(y,1) = \begin{cases} \dfrac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)} y^{-\frac{1}{2}} e^{-\frac{y}{2}}, & y > 0 \\ 0, & y \le 0 \end{cases}$$

Γ分布

$$\Gamma(k,\theta) = \begin{cases} \dfrac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-\frac{x}{\theta}}, & x > 0 \\ 0, & x \le 0 \end{cases}$$

下面证明相互独立的Γ分布的可加性：

设相互独立的$X_1 \sim \Gamma(\alpha,\theta), X_2 \sim \Gamma(\beta,\theta)$，$Z = X_1 + X_2$

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X_1}(x) f_{X_2}(z-x)\,dx$$

$$= \int_0^z \frac{1}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\theta}} \frac{1}{\theta^\beta \Gamma(\beta)} (z-x)^{\beta-1} e^{-\frac{z-x}{\theta}}\,dx$$

$$= \frac{e^{-\frac{z}{\theta}}}{\theta^{\alpha+\beta}\Gamma(\alpha)\Gamma(\beta)} \int_0^z x^{\alpha-1}(z-x)^{\beta-1}\,dx$$

$$= \frac{z^{\alpha+\beta-2} e^{-\frac{z}{\theta}}}{\theta^{\alpha+\beta}\Gamma(\alpha)\Gamma(\beta)} \int_0^z \left(\frac{x}{z}\right)^{\alpha-1} \left(1-\frac{x}{z}\right)^{\beta-1} dx$$

(代换$x = zt, dx = z\,dt$)

$$= \frac{\boldsymbol{z^{\alpha+\beta-1} e^{-\frac{z}{\theta}}}}{\theta^{\alpha+\beta}\Gamma(\alpha)\Gamma(\beta)} \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt \overset{\text{def}}{=} A\boldsymbol{z^{\alpha+\beta-1} e^{-\frac{z}{\theta}}}$$

注意：与z无关

于是，$A=\frac{1}{\theta^{\alpha+\beta}\Gamma(\alpha)\Gamma(\beta)}\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$

由概率性质有

$$\int_0^{\infty} Az^{\alpha+\beta-1}e^{-\frac{z}{\theta}}dz=1$$

$$A\theta^{\alpha+\beta}\int_0^{\infty}\left(\frac{z}{\theta}\right)^{\alpha+\beta-1}e^{-\frac{z}{\theta}}d\left(\frac{z}{\theta}\right)=1$$

$$A=\frac{1}{\theta^{\alpha+\beta}\int_0^{\infty}t^{\alpha+\beta-1}e^{-t}dt}=\frac{1}{\theta^{\alpha+\beta}\Gamma(\alpha+\beta)}$$

于是，当$z>0$时，

$$f_Z(z)=Az^{\alpha+\beta-1}e^{-\frac{z}{\theta}}=\frac{1}{\theta^{\alpha+\beta}\Gamma(\alpha+\beta)}z^{\alpha+\beta-1}e^{-\frac{z}{\theta}}$$

于是对于相互独立的$X_1 \sim \Gamma(\alpha,\theta), X_2 \sim \Gamma(\beta,\theta)$，若$Z = X_1 + X_2$，则有$Z \sim \Gamma(\alpha+\beta,\theta)$，即满足可加性。

于是对于Γ分布的特例χ^2同样满足可加性，即有

$$\chi^2 = \sum_{i=1}^{n} X_i^2 \sim \Gamma\left(\frac{n}{2}, 2\right) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, \qquad y > 0$$

χ^2分布的性质

- 自由度为n的χ^2分布的均值$\mu = n$

  可以直接从密度函数导出，或者，考虑$X_i \sim N(0,1)$，于是

$$E(X_i^2) = D(X_i) = 1$$

$$E(\chi^2) = E\left(\sum_{i=1}^{n} X_i^2\right) = \sum_{i=1}^{n} E(X_i^2) = n$$

- 自由度为n的χ^2分布的方差$\sigma^2 = 2n$

$$D(X_i^2) = E(X_i^4) - \left[E(X_i^2)\right]^2$$

$$= \int_{-\infty}^{+\infty} x^4 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx - 1 = 2$$

  于是，

$$D(\chi^2) = D\left(\sum_{i=1}^{n} X_i^2\right) = \sum_{i=1}^{n} D(X_i^2) = 2n$$

χ^2分布的可加性

$\chi_1^2 \sim \chi^2(m)$，$\chi_2^2 \sim \chi^2(n)$，且χ_1^2和χ_2^2相互独立，则

$$\chi_1^2 + \chi_2^2 \sim \chi^2(m+n)$$

推论：$X_1, X_2, \ldots, X_n$ i.i.d $\sim N(\mu, \sigma^2)$，随机变量

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

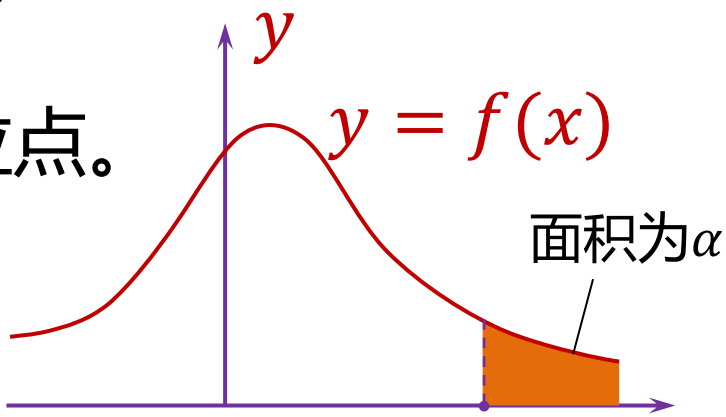
服从自由度为n的χ^2分布

上α分位点的定义

设$X \sim f(x)$，若$\forall 0 < \alpha < 1$，存在常数f_α满足

$$P\{X > f_\alpha\} = \int_{f_\alpha}^{+\infty} f(x)dx = \alpha$$

则称f_α为分布密度$f(x)$的上α分位点。



分位点的作用——

在统计推断时，需要知道给定概率下，对应随机变量的取值。

对χ^2分布而言，称满足条件

$$P\{\chi^2 > \chi_\alpha^2(n)\} = \alpha$$

的点$\chi_\alpha^2(n)$为$\chi^2(n)$分布的上α分位点

当n充分大时，$\chi^2(n) \approx \frac{1}{2}\left(z_\alpha + \sqrt{2n-1}\right)^2$，$z_\alpha$是标准正态分布的上$\alpha$分位点

χ^2分布表($P\{\chi^2 > \chi_\alpha^2(n)\} = \alpha$)

n \ α	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025
1	0.00004	0.00016	0.001	0.004	0.016	2.706	3.841	5.024
2	0.01	0.02	0.051	0.103	0.211	4.605	5.991	7.378
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143
5	0.412	0.554	0.831	1.145	1.61	9.236	11.07	12.833
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449
7	0.989	1.239	1.69	2.167	2.833	12.017	14.067	16.013
8	1.344	1.646	2.18	2.733	3.49	13.362	15.507	17.535
9	1.735	2.088	2.7	3.325	4.168	14.684	16.919	19.023
10	2.156	2.558	3.247	3.94	4.865	15.987	18.307	20.483
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.92
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337

t分布(Student(学生氏)分布)

定义：设$X\sim N(0,1)$，$Y\sim\chi^2(n)$，且X与Y相互独立，称随机变量

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

服从自由度为n的t分布，$t\sim t(n)$

t分布的概率密度函数为：

$$h(t)=\frac{\Gamma\left[\frac{n+1}{2}\right]}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}},-\infty<t<+\infty$$

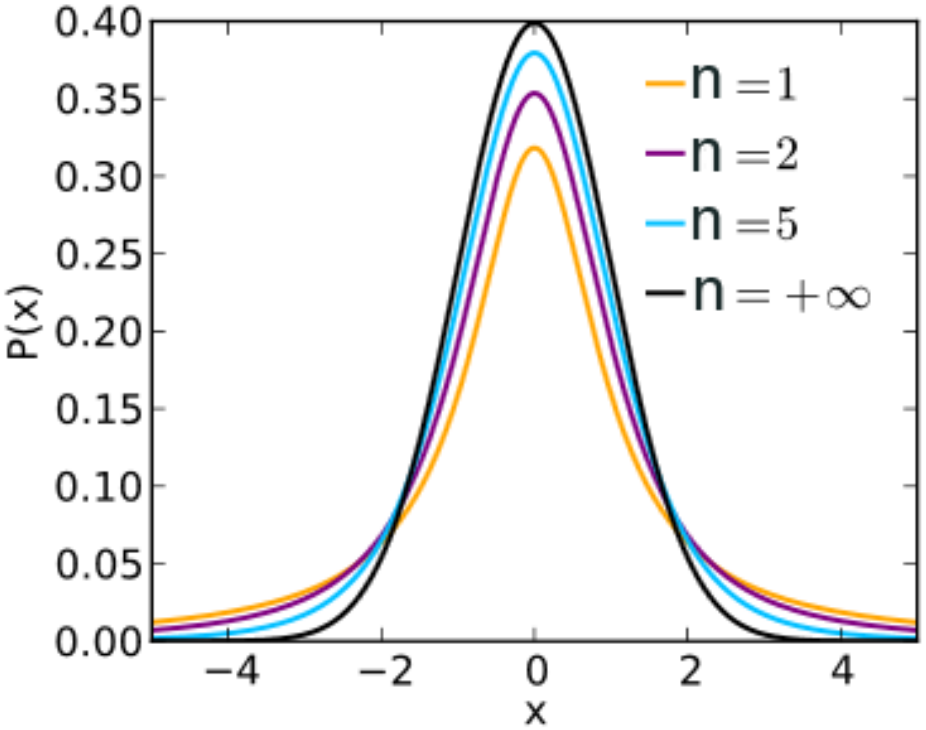
t分布特性

t分布关于x=0对称

- n=1时，E(t)不存在

$$h(t)=\frac{\Gamma(1)}{\sqrt{\pi}(1+t^2)\Gamma\left(\frac{1}{2}\right)}$$

- n≥2时，E(t) =0



n→∞时，t分布趋向于标准正态分布

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

n足够大(≥ 45)时， t分布可用标准正态分布近似

对t分布而言，称满足条件

$$P\{t > t_\alpha(n)\} = \alpha$$

的点$t_\alpha(n)$为t分布的上α分位点

由概率密度的对称性，$t_{1-\alpha}(n) = -t_\alpha(n)$

推论: 如果$X_1, X_2, \dots, X_n$ i.i.d, $\sim N(\mu, \sigma^2)$，则随机变量

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

为自由度为$n-1$的t分布。

t分布被大量用于总体均值的推断、样本比较等问题

t分布的来历

- t分布最先提出者——Helmert和Lüroth (1876)
- 1908年William Sealy Gosset用笔名“Student”在 Biometrika 上发表了题为“The probable error of a mean”的文章
- Gosset当时在Guinness啤酒厂工作，对小样本问题感兴趣——大麦样本可能只有极少数
- 命名的来历
  - Gosset的雇主不希望雇员用真名发表论文 vs. Guinness不希望对手掌握这一小样本方法
  - Ronald Fisher在其文章中称这一分布为“Student′s distribution”，并用t表示值，原因之一是Gosset和Student 的最后一个字母都是t

F分布（方差比分布）

定义：设$U\sim\chi^2(n_1)$，$V\sim\chi^2(n_2)$，且U与V相互独立，称随机变量

$$F=\frac{U/n_1}{V/n_2}$$

服从自由度为(n_1,n_2)的F分布，记为$F\sim F(n_1,n_2)$

注意到，

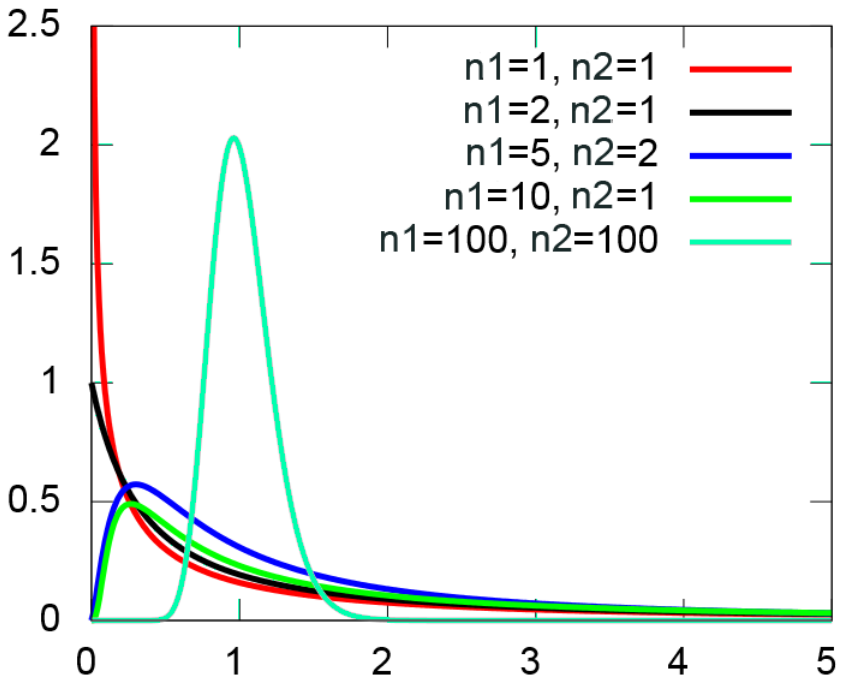
$$\frac{1}{F}=\frac{V/n_2}{U/n_1}\sim F(n_2,n_1)$$

若$F\sim F(n_1,n_2)$，则$1/F\sim F(n_2,n_1)$

F分布的密度函数

$$\psi(y)=\begin{cases}\dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)\left(\dfrac{n_1}{n_2}\right)^{\frac{n}{2}}y^{\frac{n}{2}-1}}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)\left(1+\dfrac{n_1}{n_2}y\right)^{\frac{n_1+n_2}{2}}}, & y>0\\ 0, & \text{其他}\end{cases}$$

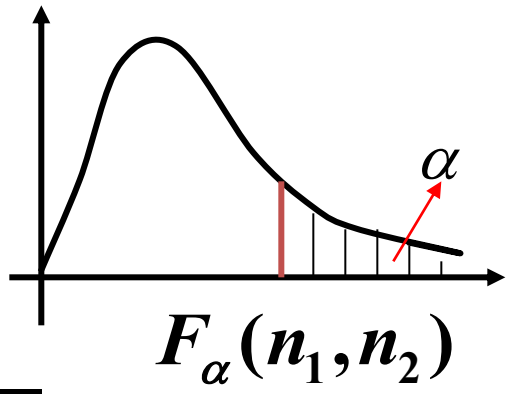
用于比较多种类型的样本方差问题，如：比较样本是否来源于不同总体



对F分布而言，称满足条件

$$P\{F > F_\alpha(n_1, n_2)\} = \alpha$$

的点$F_\alpha(n_1, n_2)$为F分布的上α分位点



F分布上α分位点的性质，

$$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_\alpha(n_2, n_1)}$$

证：由F分布方差比的特点

$$1-\alpha = P\{F > F_{1-\alpha}(n_1, \mathrm{n}_2)\} = P\left\{\frac{1}{F} < \frac{1}{F_{1-\alpha}(n_1, \mathrm{n}_2)}\right\}$$

$$= 1 - P\left\{\frac{1}{F} \geq \frac{1}{F_{1-\alpha}(n_1, \mathrm{n}_2)}\right\}$$

于是

$$P\left\{\frac{1}{F} \geq \frac{1}{F_{1-\alpha}(n_1, \mathrm{n}_2)}\right\} = \alpha$$

又因为$\frac{1}{F} \sim F(n_2, n_1)$，所以$F_\alpha(n_2, n_1) = \frac{1}{F_{1-\alpha}(n_1, n_2)}$

即，

$$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_\alpha(n_2, n_1)}$$

例：

F分布表中没有

F分布表中直接查得

$$F_{0.90}(12,9) = \frac{1}{F_{0.1}(9,12)} = \frac{1}{2.21} = 0.452$$

为何称之为F-分布？

- F-分布是为了纪念著名统计学家费歇耳(Ronald Aylmer Fisher 1890-1962)而命名
- 关于Fisher

  英国统计与遗传学家，现代统计科学的奠基人之一。在遗传学的研究中，引入并发展了统计学方法，其著作《研究工作者的统计方法》(1925)影响超过半世纪。著作《天择的遗传理论》将统计分析的方法带入进化论的研究。

正态总体的抽样定理

最重要的总体：$X \sim N(\mu, \sigma^2)$

问题：如何由样本获得对μ和σ^2的估计

方法：构造合适的统计量$g(X_1, X_2, \dots, X_n)$

问题:

1. 什么统计量是合适的？
2. $g(X_1, X_2, \dots, X_n)$服从什么分布？

抽样定理回答了上述问题

定理1. 设$X_1, X_2, \ldots, X_n$ 是来自总体$X \sim N(\mu, \sigma^2)$的样本，则

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

证明： 由$X_1, X_2, \ldots, X_n$独立同分布，以及正态分布的性质，于是有

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

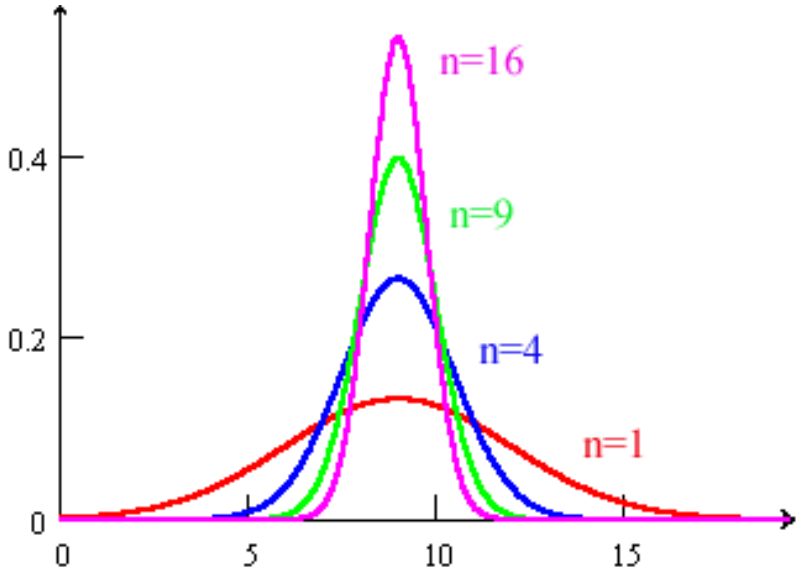
仍服从正态分布，且

$$E(\bar{X}) = \mu, \qquad D(\bar{X}) = \frac{\sigma^2}{n}$$

于是，

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

证明了$\bar{X}$逼近μ的合理性

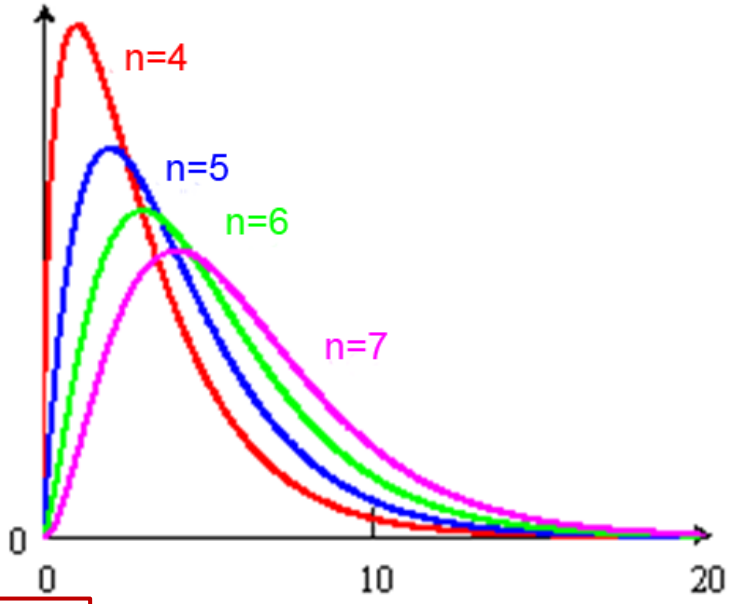


n取不同值时样本均值 $\bar{X}$的分布

定理2. 设$X_1, X_2, \ldots, X_n$是来自总体$X \sim N(\mu, \sigma^2)$的样本，$\bar{X}$和S^2分别为样本均值与样本方差，则有

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

(2) $\bar{X}$和S^2相互独立



n取不同值时$\frac{(n-1)S^2}{\sigma^2}$的分布

如果(1)成立，则

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1 \Rightarrow E[S^2] = \sigma^2$$

$$D\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1) \Rightarrow D[S^2] = \frac{2\sigma^4}{(n-1)}$$

证明了S^2逼近σ^2的合理性

证明：(1)

$$\frac{(n-1)S^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}=\sum_{i=1}^{n}\left[\frac{(X_i-\mu)-(\bar{X}-\mu)}{\sigma}\right]^2$$

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

于是，

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

注意到：$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi^2(n)$，而$\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2\sim\chi^2(1)$

于是由χ^2分布的性质，有$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$

定理3. 设$X_1, X_2, \dots, X_n$ 是来自总体$X \sim N(\mu, \sigma^2)$的样本，$\bar{X}$和S^2分别为样本的均值与样本方差，则有

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

证明：由定理1和定理2有：

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1), \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

于是由t分布定义有，

$$\frac{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} \sim t(n-1) \Rightarrow \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

定理4. 设$X_1, X_2, \dots, X_n$ 与$Y_1, Y_2, \dots, Y_n$分别是来自总体$X \sim N(\mu_1, \sigma_1^2)$和$Y \sim N(\mu_2, \sigma_2^2)$的样本，且两样本分别独立，两样本均值与样本方差分别为$\bar{X}, \bar{Y}, S_1^2, S_2^2$ ，则有

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

当$\sigma_1^2 = \sigma_2^2 = \sigma^2$时，

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

其中

$$S_w^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \qquad S_w = \sqrt{S_w^2}$$

证明：(1) 由定理2

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

由假设S_1^2,S_2^2相互独立，则由F分布的定义知

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

即，

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

(2) 易知

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

即有

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

又由给定条件知

$$\frac{(n_1-1)S_1^2}{\sigma^2}\sim\chi^2(n_1-1),\frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_2-1)$$

由χ^2分布的可加性有

$$V=\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2)$$

由U,V的独立性以及t分布的定义有

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

作业

概率论与数理统计

pp. 147-148, #4，#6，#7，#9