统计量与抽样分布

基本概念

总体：研究的对象的全体，是随机变量，有一定的分布

样本：从总体中随机地抽取一些个体，抽取的过程叫做**抽样。**

- 样本观测值：在对总体进行 n 次观测后可以得到一组**值** $(x_1, x_2, \ldots, x_n)$ ，称为样本的观测值
- 作为随机变量的样本：在不同的观测中样本的值会受到随机因素的影响而发生变化，此时的样本是一个**随机变量**，记为 $(X_1, X_2, \ldots, X_n)$

具体计算时将样本看作值，但是在讨论一般问题时将样本当作随机变量看待

一般研究随机抽样得到的样本时要求其满足

- 代表性：即样本能够代表总体，样本的每个分量 X_i 与总体 X 有相同的分布
- 独立性：样本的各个分量为相互独立的随机变量

满足上述要求的称为**简单随机样本**

如果总体的分布函数为 $F(x)$ ，则样本 $(X_1, X_2, \ldots, X_n)$ 的分布函数为

$$F(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F(x_i)$$

同样的，对于连续型随机变量，若总体的概率密度为 $p(x)$ ，则样本的概率密度为

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$$

统计量与抽样分布

统计量：设 $(X_1, X_2, \ldots, X_n)$ 为总体 X 的一个样本，$T(x_1, x_2, \ldots, x_n)$ 为**不含任何未知参数**的函数，则称 $T(x_1, x_2, \ldots, x_n)$ 为一个统计量

统计量不含任何未知参数，因此有样本即可计算出统计量

类似样本，在一次具体的观察中统计量是一个具体的数值，但是在讨论一般问题时统计量可以看作随机变量，统计量的分布称为**抽样分布。**

有一些常用的统计量如下

样本均值

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

样本方差

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

样本方差这样定义是为了**无偏性**

样本标准差

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

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

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

正态总体

在正态总体下，一些统计量具有精确的分布

$\mathcal{X}^2$ 分布

设 $X_1,X_2,\ldots,X_n$ 为独立同分布的随机变量，且服从 $N(0,1)$ ，则定义

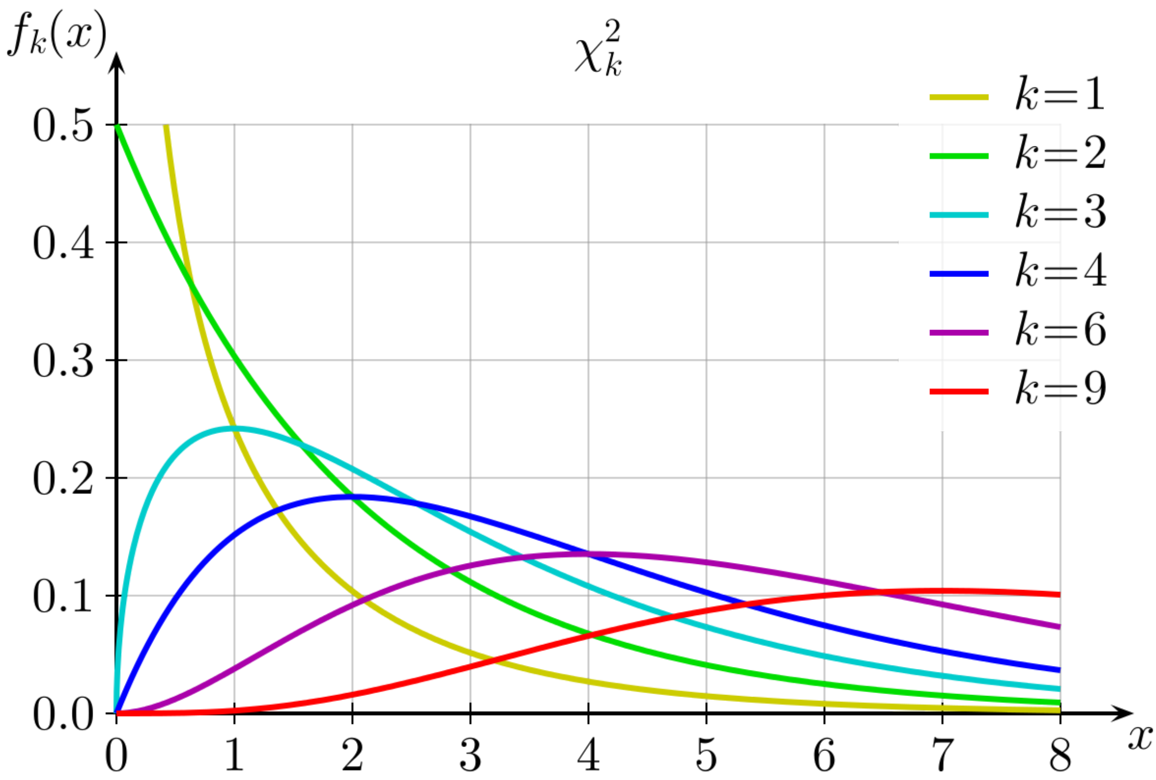
$$\mathcal{X}_n^2=\sum_{i=1}^{n}X_i^2$$

为服从自由度为 n 的 $\mathcal{X}^2$ 分布，记为 $\mathcal{X}_n^2\sim\mathcal{X}^2(n)$

其密度函数为

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

图像为



$\mathcal{X}^2$ 分布有性质

可加性：若 $\mathcal{X}_1^2 \sim \mathcal{X}^2(n_1), \mathcal{X}_2^2 \sim \mathcal{X}^2(n_2)$，且相互独立，则 $\mathcal{X}_1^2 + \mathcal{X}_2^2 \sim \mathcal{X}^2(n_1 + n_2)$

若 $\mathcal{X}^2 \sim \mathcal{X}^2(n)$，则 $E[\mathcal{X}^2] = n$， $D[\mathcal{X}^2] = 2n$

t 分布

若 $X \sim N(0,1), Y \sim \mathcal{X}^2(n)$，且 X, Y 相互独立，则定义

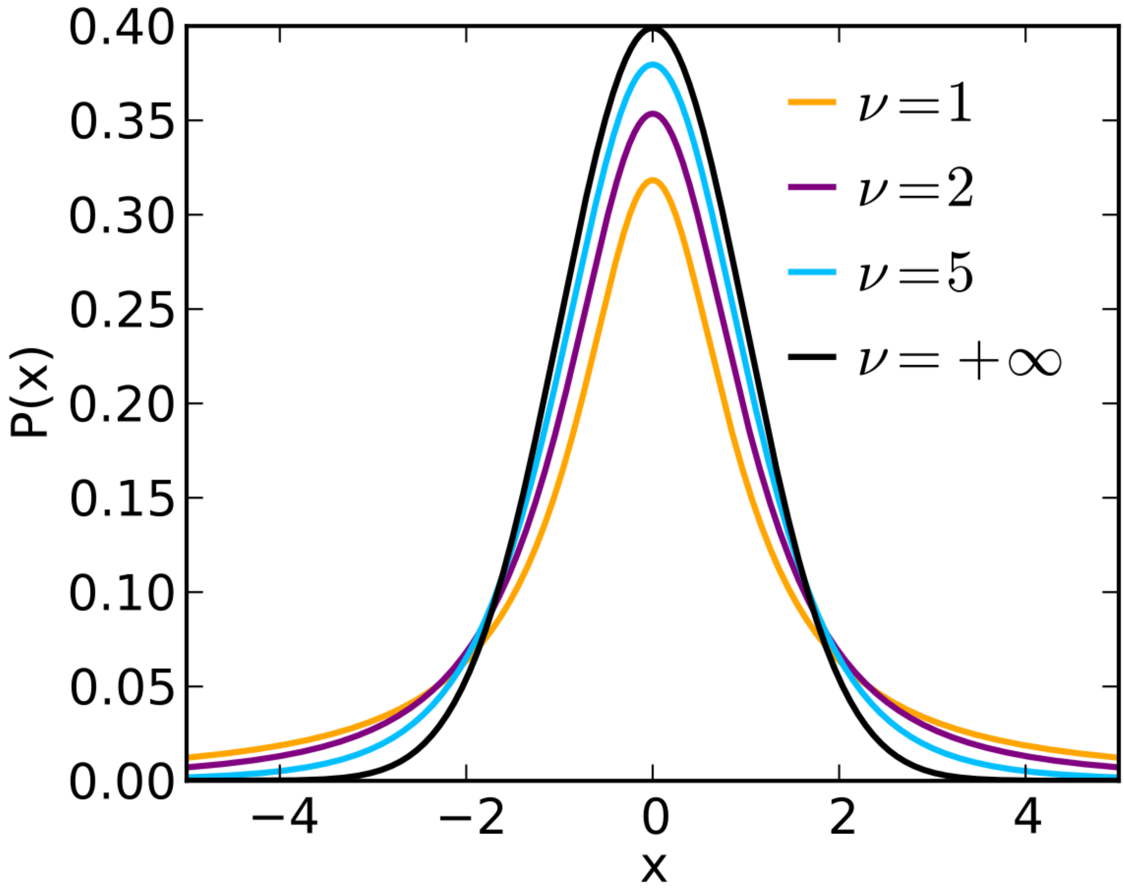
$$T = \frac{X}{\sqrt{Y/n}}$$

为服从自由度为 n 的 t 分布，记为 $T \sim t(n)$

其密度函数为

$$p(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

t 分布的密度函数为偶函数，图形为



当 n 较大时，$p(x)$ 收敛于标准正态分布 $\varphi(x)$ ，即 t 分布近似 $N(0,1)$

F 分布

设 $U\sim\mathcal{X}^2(n_1),V\sim\mathcal{X}^2(n_2)$ ，且 U,V 相互独立，则定义

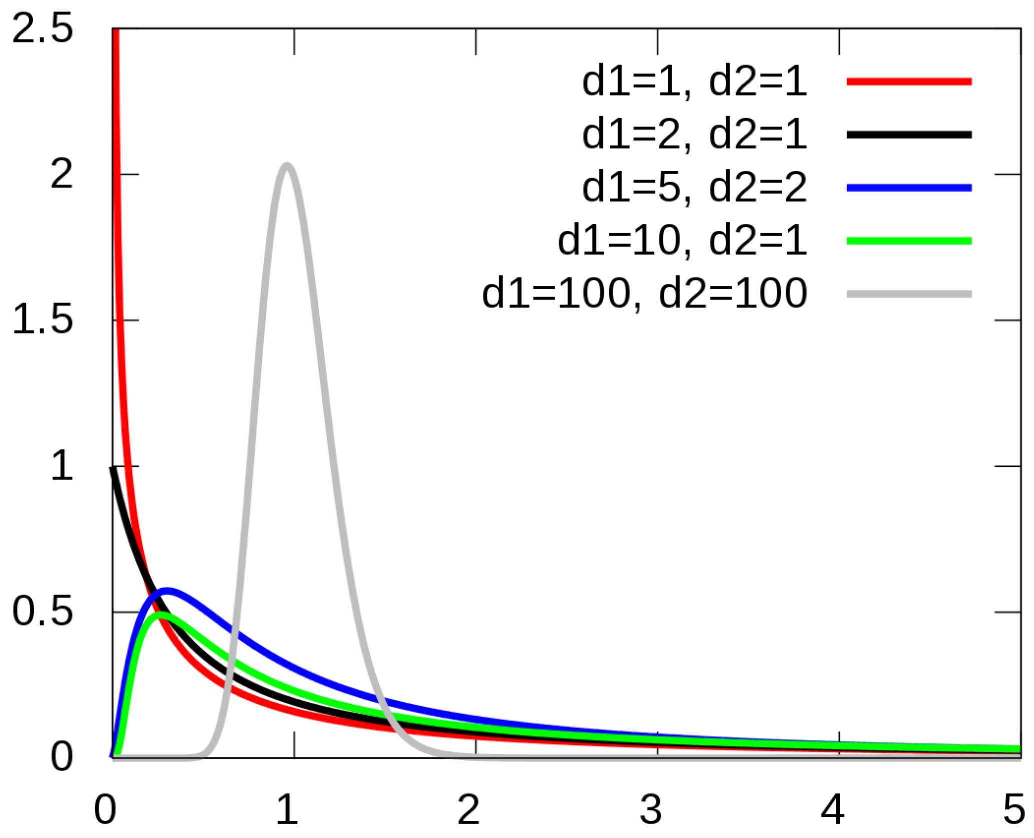
$$F=\frac{U/n_1}{V/n_2}$$

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

其密度函数为

$$p(x)=\begin{cases}\frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(n_1/2)\Gamma(n_2/2)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}} & x>0\\ 0 & x\leqslant 0\end{cases}$$

图形为



F 分布有性质

如果 $F \sim F(n_1, n_2)$ ，则 $\frac{1}{F} \sim F(n_2, n_1)$

若 $T \sim t(n)$ ，则 $T^2 \sim F(1, n)$

上 α 分位点

设 X 是一个随机变量，则对于给定实数 $\alpha(0 < \alpha < 1)$ ，若有实数 λ_α 满足

$$P(X > \lambda_\alpha) = \alpha$$

则称 λ_α 为 X 的上 α 分位点

记 X 服从 $N(0,1), \mathcal{X}^2(n), t(n), F(n_1, n_2)$ 时的上 α 分位点分别为 $u_\alpha, \mathcal{X}^2_\alpha(n), t_\alpha(n), F_\alpha(n_1, n_2)$ ，则有如下的性质

根据对称性可得

$$u_{1-\alpha} = -u_\alpha$$
$$t_{1-\alpha}(n) = -t_\alpha(n)$$

且对于 F 分布有

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

对于 $\mathcal{X}^2$ 分布，在 n 充分大时，有

$$\mathcal{X}_\alpha^2(n) \approx \frac{1}{2}(u_\alpha + \sqrt{2n-1})^2$$

抽样分布定理

设总体 X 的均值为 μ ，方差为 σ^2 ，$(X_1, X_2, \dots, X_n)$ 是来自总体的一组样本，则有

$$\begin{aligned} E[\overline{X}] &= \mu \\ D[\overline{X}] &= \frac{\sigma^2}{n} \\ E[S^2] &= \sigma^2 \end{aligned}$$

如果总体 X 是正态总体，即 $X \sim N(0,1)$ ，则还有

$$\begin{aligned} \overline{X} &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \\ \frac{(n-1)S^2}{\sigma^2} &\sim \mathcal{X}^2(n-1) \\ \frac{\overline{X}-\mu}{S/\sqrt{n}} &\sim t(n-1) \end{aligned}$$

且满足 $\overline{X}$ 与 S^2 独立

如果考虑两个正态总体 $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ 以及来自其的两组独立的样本 $(X_1, X_2, \dots, X_{n_1}), (Y_1, Y_2, \dots, Y_{n_2})$ ，则

$$\begin{aligned} F = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} &\sim F(n_1-1, n_2-1) \\ T = \sqrt{\frac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{(n_1-1)S_1^2+(n_2-1)S_2^2}} &\sim t(n_1+n_2-2) \end{aligned}$$

其中第二条成立需要 $\sigma_1^2 = \sigma_2^2$