第六章: 样本与统计量

6.1 总体与样本

6.2 统计量

6.3 正态总体的抽样分布

数理统计的定义

定义 0.1

数理统计是数学的一个分支, 它是研究如何有效地收集和使用带有随机性影响的数据的一门学科.

其核心为:

- 有效地收集数据.
- 有效地使用数据.
- 与概率论的联系与区别

数理统计的定义: 有效地收集数据

例 0.1 (抽样调查)

考察某地区 10000 户农户的经济状况, 从中挑选 100 户作抽样调查. 若该地区分为平原和山区两部分, 平原较富, 占该地区农户的 70%, 而占 30% 的山区农户较穷. 抽样方案规定在抽取的 100 户中, 从平原地区抽 70 户, 山区抽 30 户, 在各自范围内使用随机化方法抽取.

数理统计的定义: 有效地收集数据

例 0.2 (实验设计)

某化工产品的获得率与温度、压力和原配料有关. 为提高获得率, 通过试验寻找最佳生产条件. 试验因素和水平如下:

因素/水平	1	2	3	4
温度	800°C	1000°C	1200°C	1400°C
压力	10	20	30	40
配方	A	B	C	D

三个因素, 每个因素 4 个水平共要做 $4^3 = 64$ 次试验. 做那么多试验, 人力、物力和财力都不可能. 因此, 如何通过尽可能少的试验获得尽可能多的信息? 例如采用正交表安排试验就是一种有效的方法.

数理统计的定义: 有效地收集数据

在有效收集数据中, 一个重要的问题是数据须具有随机性.

- **调查数据**. 抽样调查中, 随机性体现在抽样的 100 户农户是从 10000 户中**按一定的方式"随机抽取"的**, 它具有一定的代表性.
- **实验数据**. 在实验设计中, 随机性体现在试验误差中. 化工产品的获得率除了受温度、压力和配方的影响, 还受到一些无法控制, 甚至仍未被人们认识的因素影响. **这些因素无法完全控制, 从而对试验结果产生随机性的影响, 带来不确定性.**

数理统计的定义: 有效地使用数据

例 0.3

某农村有 100 户农户, 要调查此农户是否脱贫. 脱贫的标准是每户年均收入超过 1 万元. 经调查此村 90 户农户年收入 5000 元, 10 户农户年收入 10 万元, 问此村是否脱贫?

解: (1) 用**样本平均值**计算该村农户年均收入如下:

$$\bar{x} = \frac{90 \times 0.5 + 10 \times 10}{100} = 1.45 \text{ 万元}.$$

按此方法得出结论: 该村农民已脱贫. 但 90% 的农户的年均收入只有 5000 元, 事实上并未脱贫.

数理统计的定义: 有效地使用数据

(2) 用**样本中位数**计算该村农户年均收入, 即 100 户的年收入分别记为

$$x_1, x_2, \cdots, x_{100},$$

将其按大小排列为 $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(100)}$. 样本中位数定义为排在最中间的两户的平均值, 即

$$\frac{x_{(50)} + x_{(51)}}{2} = 0.5\ (\text{万元}).$$

按此方法得出结论: 该村农民尚未脱贫.

不同的统计方法得出的结论不同. 有效地使用数据, 需针对不同问题选择合适的统计方法. □

数理统计的定义: 与概率论的联系与区别

例 0.4 (数理统计问题)

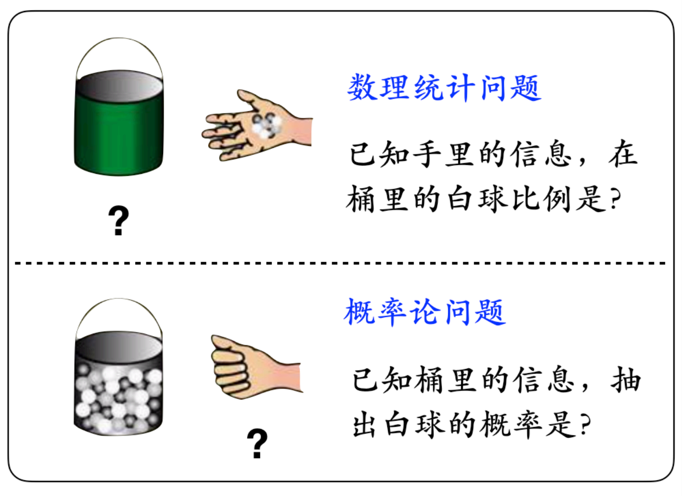
一大批产品共有 N 件, 其中废品 M 件, N 已知, 而 M 未知. 现在从中不放回地抽出 n 个检验其中废品的件数, 用以估计废品率 $p = M/N$ 和求出样本分布. 参数 p 未知, 用 n 个样本来估计参数, 这是典型的统计问题.

例 0.5 (概率论问题)

一大批产品共有 N 件, 其中废品率为 p, p 已知. 设随机变量 X 表示 n 件产品中的废品数, 求随机变量 X 的概率分布. 显然, X 服从超几何分布, $\mathcal{H}(n, N, N\theta)$. 参数 p 已知, 求随机变量 X 的概率分布, 这是典型的概率论问题.

数理统计的定义：与概率论的联系与区别

注意上面两个例子的差别，下图反映了二者的差别.



6.1 总体与样本

总体与样本

例 1.1

假定一批产品有 10000 件, 其中有正品也有次品. 为估计废品率, 往往从中抽取一部分, 如 100 件进行检查. 此时:

- 这批 10000 件产品称为总体 (*population*).
- 其中的每件产品称为个体 (*individual*).
- 从中抽取的 100 件产品称为样本 (*sample*).
- 样本中个体的数目称为样本容量 (*sample size*).
- 抽取样本的行为称为抽样 (*sampling*).

总体与样本

总体是由与所研究的问题有关的所有个体组成，而样本是从总体中抽取的一部分个体.

- 若总体中个体的数目有限, 则称为**有限总体** (finite population), 否则称为**无限总体**(infinite population).
- 总体可看成由所有个体上的某种数量指标构成的集合, 因此它是**数的集合**. 由于每个个体的出现都是随机的, 所以相应个体的数量指标也带有**随机性**.
- **从而此种数量可看成随机变量. 随机变量的分布是该数量指标在总体中的分布.**

例 1.2

以例 *1.1* 来说明, 假定 *1000* 只产品中废品数为 *100* 件, 其余为正品, 废品率未知设为 p. 定义随机变量 X:

$$X = \begin{cases} 1, & \text{废品}, \\ 0, & \text{正品}. \end{cases}$$

其概率分布为 $0-1$ 分布, 且 $P(X=1)=p$. 因为特定个体上的数量指标是随机变量 X 的观测值 (*observations*).

总体与样本

定义 1.1

一个统计问题所研究对象的的全体称为总体. 在数理统计中总体可用一个随机变量及其概率分布来描述.

定理 1.1

若总体分布函数为 F, 当有一个从该总体抽取的相互独立同分布 (*Independent Identically Distributed, i.i.d.*) 的大小为 n 的样本 $X_1, \cdots, X_n$, 则记为

$$X_1, \cdots, X_n \quad i.i.d. \sim F.$$

若 F 有密度 f, 可记为

$$X_1, \cdots, X_n \quad i.i.d. \sim f.$$

总体与样本

定义 1.2

样本 $\boldsymbol{X}=(X_1,\cdots,X_n)$ 所有可能取值构成样本空间 *(sample space)*, 记为 $\mathcal{X}$.

例 1.3

打靶试验, 每次打三发, 考察中靶的环数. 若 $\boldsymbol{X}=(5,1,9)$ 表示三次打靶分别中 *5* 环, *1* 环和 *9* 环. 此时样本空间为

$$\mathcal{X}=\{(x_1,x_2,x_3):x_i=0,1,2,\cdots,10,i=1,2,3\}.$$

这个样本空间中样本点数是有限的.

总体与样本: 样本的二重性

样本的二重性: 样本既可看成具体的数, 又可看成随机变量.

- 在实施抽样之后, 样本是**具体的数字**.
- 在实施抽样之前, 样本是**随机变量**.
- 因为在实施抽样之前, 无法预料抽样的结果, 只能预料可能取值的范围, 因此可看成随机变量 (向量).

例 1.4

样本 $\boldsymbol{X} = (X_1, X_2, X_3)$, $0 \le X_i \le 10$ $(i = 1, 2, 3)$ 为具体的数字. 但在进行下一次打靶之前, 并不能知道 X_i 的可能的取值, 出现的结果可看成随机向量 (X_1, X_2, X_3) 的具体观察值.

总体与样本: 简单随机样本

定义 1.3

设有一总体 F, $X_1, X_2, \cdots, X_n$ 为从 F 中抽取容量为 n 的样本, 若

(1) $X_1, X_2, \cdots, X_n$ 相互独立;

(2) $X_1, X_2, \cdots, X_n$ 相同分布, 即同有分布 F;

则称 $X_1, X_2, \cdots, X_n$ 为简单随机样本.

总体与样本: 简单随机样本

定理 1.2

设总体为 F, $X_1, X_2, \cdots, X_n$ 为此总体中抽取的简单随机样本, 则 $X_1, X_2, \cdots, X_n$ 的联合分布函数 $F(x_1, x_2, \cdots, x_n)$ 可表示为

$$F(x_1) \cdot F(x_2) \cdots F(x_n) = \prod_{i=1}^{n} F(x_i).$$

若 F 有密度 f, 则其联合密度函数 $f(x_1, \cdots, x_n)$ 可表示为

$$f(x_1) \cdot f(x_2) \cdots f(x_n) = \prod_{i=1}^{n} f(x_i).$$

例 1.5

某公司为制订营销策略, 需要研究一城市居民的收入情况. 假定该城市居民家庭年收入服从正态分布 $N(\mu,\sigma^2)$, 其概率密度函数为 $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty<x<\infty$. 现随机调查 n 户居民年收入, 记为 $X_1,X_2,\cdots,X_n$, 它们是从总体 $N(\mu,\sigma^2)$ 中抽取的样本, 它们是相互独立的, 且与总体 $N(\mu,\sigma^2)$ 有相同分布, 即 $X_i\sim N(\mu,\sigma^2)$, $i=1,2,\cdots,n$. 于是 $X_1,\cdots,X_n$ 的联合概率密度为

$$f(x_1,x_2,\cdots,x_n)=\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}.$$

这个函数概括了样本 $X_1,X_2,\cdots,X_n$ 中包含的总体 $N(\mu,\sigma^2)$ 的全部信息.

6.2 统计量

统计量: 定义

定义 2.1

由样本计算出的量称为统计量 (*statistic*). 或者说, 统计量是样本的函数.

统计量只与样本有关, 不能与未知参数有关.

- 例如, $X \sim \mathcal{N}(a, \sigma^2), X_1, \cdots, X_n$ 是从总体 X 中抽取的 i.i.d. 样本, a 和 σ^2 未知, 则

$$\sum_{i=1}^{n} X_i, \quad \sum_{i=1}^{n} X_i^2, \quad \sum_{i=1}^{n} (X_i - a), \quad \sum_{i=1}^{n} X_i^2/\sigma^2,$$

其中前两个是统计量, 后两个不是统计量.

统计量: 常用统计量

例 2.1 (样本均值)

$X_1, \cdots, X_n$ 是从某总体中抽取的样本, 则称

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

为样本均值 (*sample mean*). 它反映了总体均值的信息.

例 2.2 (样本方差)

称

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

为样本方差 (*sample variance*). 它反映了总体方差的信息. 而 S 称为样本标准差.

性质 2.1 (样本均值和方差的性质)

(1) $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$.

(2) 设非零实数 a 和 b 为常数, 作变换 $Y_i = aX_i + b$, $i = 1, 2, \cdots, n$, 则 $Y_1, \cdots, Y_n$ 的样本均值和方差分别是

$$\bar{Y} = a\bar{X} + b, \quad S_Y^2 = a^2 S_X^2.$$

(3) 对任意常数 c, 有

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

且等号只在 $c = \bar{X}$ 时成立.

统计量: 常用统计量

例 2.3 (样本原点矩和中心矩)

设 $X_1, \cdots, X_n$ 是从某总体中抽取的样本, 则称

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, 2, \cdots$$

为样本 k 阶原点矩. 特别 $k = 1$ 时, $A_1 = \bar{X}$, 即样本均值. 称

$$M_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k, \quad k = 2, 3, \cdots$$

为样本 k 阶中心矩. 特别 $k = 2$ 时, $M_2 = (n-1)S^2/n$.

统计量: 统计量抽样分布

根据样本的二重性, 样本是随机变量, 统计量是样本的函数, 故它也是随机变量, 其分布称为抽样分布.

- **精确抽样分布.** 当总体 X 的分布类型已知时, 若对任一自然数 n, 都能导出统计量 $T = T(X_1, \cdots, X_n)$ 的分布表达式, 这种分布称为 T 的精确抽样分布 (大多是在正态条件下得到).
- **大样本分布.** 当样本容量 $n \to \infty$ 时统计量的分布称为大样本分布. 只要样本容量足够大, 且大样本分布的形式比较简单, 就可用统计量的大样本分布作为精确分布的近似.

定理 2.1

假设 $X_1, X_2, \cdots, X_n$ 为来自均值为 μ、方差为 σ^2 的总体的一组样本, 则当 n 充分大时, 近似地有

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

证明: 因为 $X_1, X_2, \cdots, X_n$ 是来自均值为 μ、方差为 σ^2 的总体的样本, 是独立同分布的, 且 $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, $i = 1, 2, \cdots, n$. 根据中心极限定理, 对充分大的 n, 近似地有

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

统计量: 大样本分布

即对充分大的 n, 近似地有

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

等价地,

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

- 无论总体分布的具体形式如何, 只要它的均值为 μ, 方差为 σ^2.
- 则从这个总体抽取的样本均值 $\bar{X}$ **近似地服从均值为 μ, 方差为 σ^2 地正态分布.**

统计量: 大样本分布

例 2.4

某公司用机器向瓶子里灌装液体洗净剂, 规定每瓶装 μ 毫升. 但实际灌装量总有一定的波动, 假定罐装量的方差 σ^2 为 1. 如果每箱装 25 瓶这样的洗净剂, 试问这 25 瓶洗净剂的平均每罐装量与标定值 μ 相差不超过 0.3 毫升的概率是多少?

解: 记一箱中 25 瓶洗净剂灌装量为 $X_1, X_2, \cdots, X_{25}$, 它们是来自均值为 μ、方差为 1 的总体样本. 需计算

$$P\{|\bar{X} - \mu| \leq 0.3\}.$$

根据定理 2.1, 有

$$
\begin{aligned}
P\{|\bar{X}-\mu| \leq 0.3\} &= P\{-0.3 \leq \bar{X}-\mu \leq 0.3\} \\
&= P\left\{-\frac{0.3}{\sigma/\sqrt{n}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{0.3}{\sigma/\sqrt{n}}\right\} \\
&\approx \Phi\left(\frac{0.3}{\sigma/\sqrt{n}}\right) - \Phi\left(-\frac{0.3}{\sigma/\sqrt{n}}\right) \\
&= 2\Phi\left(\frac{0.3}{\sigma/\sqrt{n}}\right) - 1 \\
&= 2\Phi(1.5) - 1 = 0.8664.
\end{aligned}
$$

即对装 25 瓶的一箱, 平均每瓶罐装量与标定值之差不超过 0.3 毫升的概率近似为 86.64%. □

6.3 正态总体的抽样分布

定义 3.1

设 $X_1, X_2, \cdots, X_n$ $i.i.d. \sim \mathcal{N}(0,1)$, 则称

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

是自由度为 n 的 χ^2 变量, 其分布称为自由度为 n 的 χ^2 分布, 记为 $X \sim \chi_n^2$.

定理 3.1

设随机变量 X 是自由度 n 的 χ^2 随机变量, 其概率密度函数为

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

正态总体的抽样分布: χ^2 分布

若记伽马分布 $\Gamma(\alpha,\lambda)$ 为具有下列密度函数的概率分布

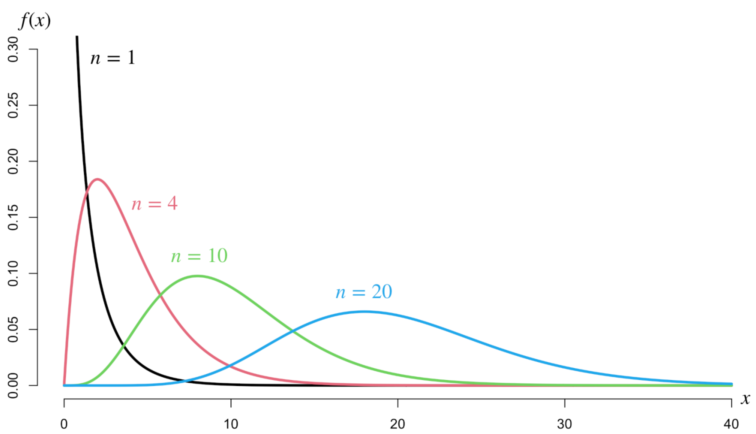
$$f(x;\alpha,\lambda)=\begin{cases}\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, & x>0.\\ 0, & x\le 0,\end{cases}$$

则自由度为 n 的 χ^2 分布与伽马分布的关系为

$$X=\sum_{i=1}^{n}X_i^2\sim\Gamma\left(\frac{n}{2},\frac{1}{2}\right).$$

即若随机变量 X 的概率密度为 $\Gamma(n/2,1/2)$, 则称 X 为服从自由度为 n 的 χ^2 分布.

正态总体的抽样分布: χ^2 分布



正态总体的抽样分布: χ^2 分布

性质 3.1 (可加性)

设 $Y_1 \sim \chi^2_m, Y_2 \sim \chi^2_n$, 且两者独立, 则 $Y_1 + Y_2 \sim \chi^2_{m+n}$.

性质 3.2

设 $X \sim \chi^2_n$. 则 $E(X) = n$, $Var(X) = 2n$.

证明: 因为 $X \sim \chi^2_n = \Gamma(n/2, 1/2)$, 利用伽马分布的期望与方差, 若 $X \sim \Gamma(\alpha, \lambda)$, 则

$$E(X) = \frac{\alpha}{\lambda} = \frac{\frac{n}{2}}{\frac{1}{2}} = n,$$

$$Var(X) = \frac{\alpha}{\lambda^2} = \frac{\frac{n}{2}}{\frac{1}{4}} = 2n.$$

$\square$

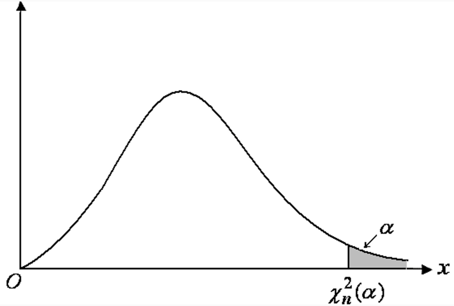
正态总体的抽样分布: χ^2 分布

定义 3.2

对给定的 α, $0 < \alpha < 1$, 称满足条件

$$P\{\chi_n^2 > \chi_n^2(\alpha)\} = \int_{\chi_n^2(\alpha)}^{\infty} f(x)dx = \alpha$$

的点 $\chi_n^2(\alpha)$ 为 χ_n^2 的**上分位点**.



正态总体的抽样分布: t 分布

定义 3.3

设随机变量 $X \sim \mathcal{N}(0,1), Y \sim \chi_n^2$, 且 X 和 Y 独立, 则称

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

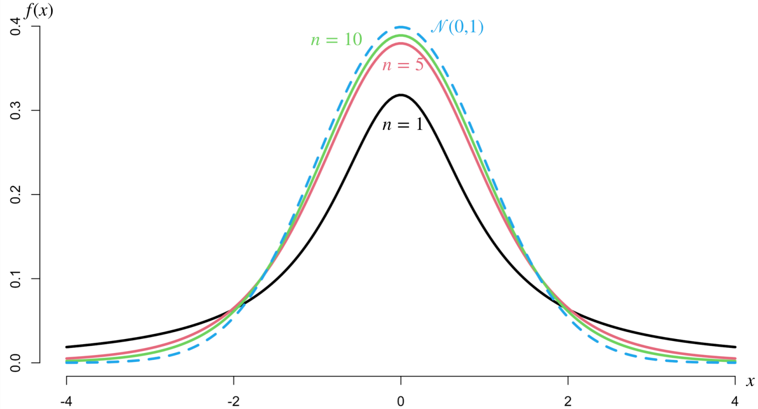
是自由度为 n 的 t 变量, 其分布称为自由度为 n 的 t 分布, 记为 $T \sim t_n$.

定理 3.2

设随机变量 $T \sim t_n$, 则其密度函数为

$$t_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < x < \infty.$$

正态总体的抽样分布: t 分布



正态总体的抽样分布: t 分布

性质 3.3

若随机变量 $T \sim t_n$, 则 $E(T^r)$ 只有当 $r < n$ 时存在, 且

$$E(T^r) = \begin{cases} n^{\frac{r}{2}} \dfrac{\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{n-r}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}, & r \text{ 为偶数}, \\ 0, & r \text{ 为奇数}. \end{cases}$$

特别地,

- 当 $n \geq 2$ 时, $E(T) = 0$.
- 当 $n \geq 3$ 时, $Var(T) = n/(n-2)$.

性质 3.4

当 $n \to \infty$ 时, t 变量的极限分布为 $N(0, 1)$.

正态总体的抽样分布: t 分布

性质 3.5

当 $n = 1$ 时, t 分布就是柯西 (*Cauchy*) 分布, 即

$$t_1(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < +\infty.$$

定义 3.4

设 $T \sim t_n$. 对给定的 α, $0 < \alpha < 1$, 称满足条件

$$P\{T > t_n(\alpha)\} = \int_{t_n(\alpha)}^{\infty} f(x)dx = \alpha$$

的点 $t_n(\alpha)$ 为 t_n 的**上分位点**.

正态总体的抽样分布: F 分布

定义 3.5

设随机变量 $X \sim \chi_m^2, Y \sim \chi_n^2$ 且 X 和 Y 独立, 则称

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

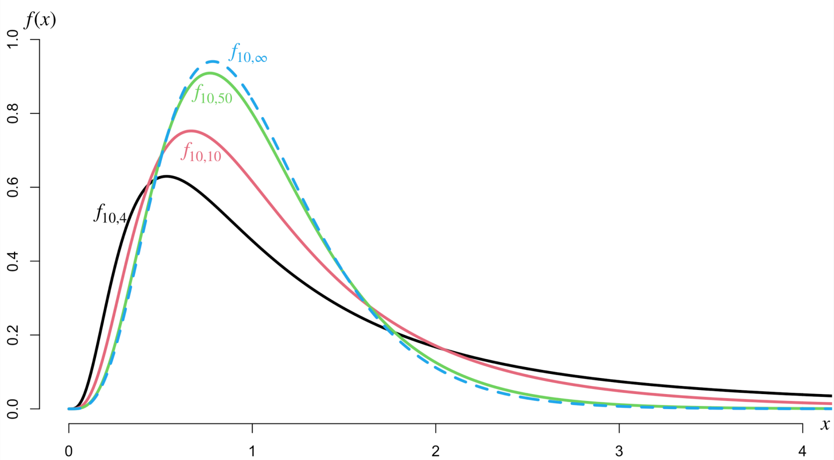
是自由度为 m 和 n 的 F 变量, 其分布称为自由度是 m 和 n 的 F 分布, 记为 $F \sim F_{m,n}$.

定理 3.3

若随机变量 $Z \sim F_{m,n}$, 则其密度函数为

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

正态总体的抽样分布: F 分布



性质 3.6

若 $Z \sim F_{m,n}$, 则对 $r > 0$ 有

$$E(Z^r) = \left(\frac{n}{m}\right)^r \frac{\Gamma\left(\frac{m}{2} + r\right)\Gamma\left(\frac{n}{2} - r\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)}, \quad 2r < n.$$

特别地,

$$E(Z) = \frac{n}{n-2}, \quad n > 2,$$
$$Var(Z) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \quad n > 4.$$

正态总体的抽样分布: F 分布

性质 3.7

若 $Z \sim F_{m,n}$, 则 $1/Z \sim F_{n,m}$.

性质 3.8

若 $T \sim t_n$, 则 $T^2 \sim F_{1,n}$.

定义 3.6

设 $F \sim F_{m,n}$. 对给定的 α, $0 < \alpha < 1$, 称满足条件

$$P\{F > F_{m,n}(\alpha)\} = \int_{F_{m,n}(\alpha)}^{\infty} f(x)dx = \alpha$$

的点 $F_{m,n}(\alpha)$ 为 $F_{m,n}$ 的**上分位点**.

正态总体的抽样分布: F 分布

性质 3.9

$F_{m,n}(1-\alpha) = 1/F_{n,m}(\alpha)$.

证明: 设 $X \sim F_{m,n}$, 则

$$\begin{aligned}
1-\alpha = P\{X > F_{m,n}(1-\alpha)\} &= P\left\{\frac{1}{X} < \frac{1}{F_{m,n}(1-\alpha)}\right\} \\
&= P\left\{Y < \frac{1}{F_{m,n}(1-\alpha)}\right\} = 1 - P\left\{Y \geq \frac{1}{F_{m,n}(1-\alpha)}\right\}
\end{aligned}$$

等价地,

$$P\left\{Y > \frac{1}{F_{m,n}(1-\alpha)}\right\} = \alpha \Rightarrow \frac{1}{F_{m,n}(1-\alpha)} = F_{n,m}(\alpha).$$

正态总体的抽样分布: 样本均值和样本方差的分布

定理 3.4

设 $X_1, \cdots, X_n$ *i.i.d.* $\sim \mathcal{N}(\mu, \sigma^2)$. 令

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

分别为样本均值和样本方差, 则有

(1) $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.

(2) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

(3) $\overline{X}$ 和 S^2 独立.

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

正态总体的抽样分布: 样本均值和样本方差的分布

证明: (2) 思路.

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

注意到, $(\sum_{i=1}^{n}(X_i-\mu)^2)/\sigma^2 \sim \chi_n^2$, $\sqrt{n}(\bar{X}-\mu)/\sigma \sim \chi_1^2$, **且 $\bar{X}$ 和 S^2 相互独立**, 由 χ^2 分布可加性, $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$.

(4) 由定理 3.4 可知

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

将其标准化得 $\sqrt{n}(\overline{X} - a)/\sigma \sim \mathcal{N}(0,1)$. 又因为

$$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1},$$

即 $S^2/\sigma^2 \sim \chi^2_{n-1}/(n-1)$ 且 $\overline{X}$ 和 S^2 相互独立, 按定义有

$$T = \frac{\sqrt{n}(\overline{X} - a)/\sigma}{\sqrt{S^2/\sigma^2}} = \frac{\sqrt{n}(\overline{X} - a)}{S} \sim t_{n-1}.$$

$\square$

正态总体的抽样分布: 样本均值和样本方差的分布

注 3.1

定理 *3.4* 有违背常识之处, 一般地, 若设 $h_1(X_1, \cdots, X_n)$ 和 $h_2(X_1, \cdots, X_n)$ 为 $X_1, \cdots, X_n$ 的函数, $X_1, \cdots, X_n$ 相互独立并不能保证 $h_1(\cdot)$ 和 $h_2(\cdot)$ 是相互独立的. 该定理却表明, 在正态条件下,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \equiv h_1(X_1, \cdots, X_n),$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \equiv h_2(X_1, \cdots, X_n).$$

均为 $X_1, \cdots, X_n$ 的函数, $\overline{X}$ 和 S^2 是相互独立的!

例 3.1

在设计导弹发射装置时, 重点之一是研究弹着点偏离目标中心的距离的方差. 对一类导弹发射装置, 弹着点偏离目标中心的距离服从正态分布 $N(\mu,\sigma^2)$. 这里 $\sigma^2 = 100$ 平方米. 现进行了 25 次发射试验, 用 S^2 记这 25 次试验中弹着点偏离目标中心的距离的样本方差. 试求 S^2 超过 50 平方米的概率.

解: 根据定理 3.4, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, 于是

$$
\begin{aligned}
P\{S^2 > 50\} &= P\left\{\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)50}{\sigma^2}\right\} \\
&= P\left\{\chi^2_{24} > \frac{24\times 50}{100}\right\} = P\{\chi^2_{24} > 12\} \\
&> 0.975.
\end{aligned}
$$