数理统计

——一本简明笔记

唐嘉琪

2025 年 6 月 8 日

编译时间 2025 年 6 月 8 日

纸张大小: A4

书本主页:www.cnblogs.com/TangJiaqiMath/p/18889902

唐嘉琪 | 上海立信会计金融学院

个人主页:www.cnblogs.com/TangJiaqiMath

前言

非常简明, 用于复习.
课程教材选取为

茆诗松, 《概率论与数理统计教程》, 第三版.

学习时参考用书

韦来生, 《数理统计》, 第二版.
李增沪, 张梅, 何辉, 《概率论 (上册)》,101 计划.

顺序以笔者学习顺序为主. 由于自用性质, 笔记用了不少笔者所习惯的符号体系.

唐嘉琪
2025 年 6 月 8 日
于上海

目录

第一章　统计量及其分布

1.1　总体与样本

Definition1.1.1 总体

一个统计问题所研究的对象的全体称为总体.

总体 (在数理统计学中) 可用一个随机变量及其分布来描述.

因此, 在统计意义下, **总体即指总体分布**. 有时也用 F 的密度 f 来描述.

Definition1.1.2 样本

设 $(x_1, \cdots, x_n)$ 是从总体中抽取的样本.

n 称为**样本容量**, 简称**样本量**, 样本中的个体称为**样本**.

样本 $(x_1, \cdots, x_n)$ 可能取值全体构成样本空间 (sample space).

Theorem1.1.3 样本的两重性

样本既可以看成具体的数, 又可以看成随机变量 (或随机向量).

在实施抽样后, 它是具体的数; 在实施抽样前, 它被看作随机变量.

我们说简单随机抽样满足:

- 样本具有**代表性**: 要求总体中每一个个体都有同等机会被选入样本. 数学的话是每一样本 x_i 与总体 X 有相同的分布;
- 样本要有**独立性**. 数学的话是 $x_1, \cdots, x_n$ 互相独立.

用简单随机抽样法得到的样本成为**简单随机样本**, 简称**样本**.

Definition1.1.4 简单随机样本

设有一总体 F, $(x_1, \cdots, x_n)$ 为从 F 中抽取出的容量为 n 的样本, 若

- $x_1, \cdots, x_n$ 互相独立;
- $x_1, \cdots, x_n$ 有相同分布 F.

则称 $(x_1, \cdots, x_n)$ 为简单随机样本.

Remark1.1.5 记法

若总体分布函数为 F, 从中抽取互相独立同分布的大小为 n 的样本 $(x_i)_{i=1}^n$, 常记为

$$x_1, \cdots, x_n \quad \text{i. i. d.} \sim F.$$

若 F 有密度 f, 也记作

$$x_1, \cdots, x_n \quad \text{i. i. d.} \sim f.$$

若总体用 r.v.X 表示. 其分布为 F. 样本 $(x_i)_{i=1}^n$ 可视作 r.v.X 的观察值, 记作

$$x_1, \cdots, x_n \quad \text{i. i. d.} \sim X.$$

Remark1.1.6

非特别说明, 本笔记中的样本均指简单随机样本.

Theorem1.1.7

总体 F, F 有密度 f, $(x_i)_{i=1}^n$ 为容量是 n 的简单随机样本, 则有联合密度如下

- $F(x_1, \cdots, x_n) = \prod_{i=1}^n F(x_i)$;
- $f(x_1, \cdots, x_n) = \prod_{i=1}^n f(x_i)$.

Remark1.1.8

上述定理中 x_i 可以是多维的. 灵活运用样本两重性是关键.

1.2 经验分布函数及其性质

Definition1.2.1 经验分布函数

设 $x_1,\cdots,x_n$ 为自总体 $F(x)$ 中抽取的 i.i.d. 样本, 将其按大小排列为 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$, 对任意实数 x, 称下列函数:

$$F_n(x)=\begin{cases}0, & x<x_{(1)},\\ k/n, & x_{(k)}\leqslant x<x_{(k+1)},\ k=1,2,\cdots,n-1,\\ 1, & x_{(n)}\leqslant x.\end{cases}$$

为经验分布函数 (empirical distribution function).

记 $v_n(x):=\sharp\{x_i\colon x_i\leqslant x\}$, 那么 $F_n(x)=v_n(x)/n$.

Definition1.2.2 示性函数

$$I_i(x)=\begin{cases}1, & x_i\leqslant x,\\ 0, & \text{其他}.\end{cases}$$

那么进一步有

$$F_n(x)=\frac{v_n(x)}{n}=\frac{1}{n}\sum_{i=1}^{n}I_i(x),\quad\forall x\in\mathbb{R}.$$

Propsition1.2.3

固定 x, 那么 $v_n(x)$ 与 $F_n(x)$ 都是样本 $(x_1,\cdots,x_n)$ 的函数, 从而都是随机变量, 由于 $\mathbb{P}\{I_i(x)=1\}=\mathbb{P}\{x_i\leqslant x\}=F(x)=:F$, 则有

$$I_i(x)\sim b(1,F)\Longrightarrow v_n(x)=\sum_{i=1}^{n}I_i(x)\sim b(n,F)$$

那么自然有期望和方差 $\mathbb{E}(v_n(x))=nF$, $\mathrm{Var}(v_n(x))=nF(1-F)$.
利用性质可以有

$$\mathbb{E}(F_n(x))=\mathbb{E}\left(\frac{v_n(x)}{n}\right)=\frac{\mathbb{E}(v_n(x))}{n}=F,$$

$$\mathrm{Var}(F_n(x))=\mathrm{Var}\left(\frac{v_n(x)}{n}\right)=\frac{\mathrm{Var}(v_n(x))}{n^2}=\frac{F(1-F)}{n}.$$

Recall1.2.4 中心极限定理 (一般形式)

随机变量序列 $\{\xi_n\}$. 用 S_n 表示 $\sum_{k=1}^n \xi_k$, 且对于每个 $n \geqslant 1$ 都有 $\mathrm{Var}(S_n) > 0$. 定义标准化部分和 S_n^* 如下:

$$S_n^* = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}} = \frac{1}{\sqrt{\mathrm{Var}(S_n)}} \sum_{k=1}^n (\xi_k - \mathbb{E}(\xi_k)).$$

称序列 $\{\xi_n\}$ 满足中心极限定理, 是指当 $n \to \infty$ 时 S_n^* 的分布函数弱收敛 (或依分布收敛) 到标准正态分布 $\mathcal{N}(0,1)$ 的分布函数, 即

$$\mathbb{P}(S_n^* \leqslant x) \xrightarrow{\mathrm{w}} \Phi(x). \text{ 下文约记 } S_n^* \overset{\mathrm{w}}{\sim} \mathcal{N}(0,1)$$

Propsition1.2.5

$$\frac{v_n(x) - nF}{\sqrt{nF(1-F)}} \overset{\mathrm{w}}{\sim} \mathcal{N}(0,1), \ (n \to \infty).$$

等价的是

$$\frac{\sqrt{n}(F_n(x) - F)}{\sqrt{F(1-F)}} \overset{\mathrm{w}}{\sim} \mathcal{N}(0,1), \ (n \to \infty).$$

Recall1.2.6 辛钦 (Khintchine) 大数定律

独立同分布的可积随机变量序列 $\{\xi_n\}$ 满足如下大数定律:

$$\bar{S}_n = \frac{1}{n} \sum_{k=1}^n \xi_k \xrightarrow{\mathrm{p}} \mathbb{E}(\xi_1)$$

Recall1.2.7 伯努利 (Bernoulli) 大数定律

假设 $0 < p < 1$. 令 S_n 是以 p 为成功概率的伯努利实验前 n 次实验中的成功次数. 则对于任意的 $\varepsilon > 0$ 有

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right) \leqslant \frac{1}{n^2\varepsilon^2} \mathrm{Var}(S_n) = \frac{p(1-p)}{n\varepsilon^2}.$$

故当 $n \to \infty$ 时有 $\bar{S}_n := S_n/n \xrightarrow{\mathrm{p}} p$

Propsition1.2.8

$\{I_i(x)\}$ 独立同分布, 故利用辛钦 (Khintchine) 大数定律直接有

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_i(x) \xrightarrow{\mathrm{p}} \mathbb{E}(I_1(x)) = F$$

类似的用伯努利 (Bernoulli) 大数定律也可以直接得到结论.

Theorem1.2.9 格里汶科定理 (Glivenko-Cantelli Theorem)

设 $F(x)$ 为 r. v. X 的分布函数, $x_1, \cdots, x_n$ 为取自总体 $F(x)$ 的 i. i. d. 样本, $F_n(x)$ 为其经验分布函数, 记 $D_n := \sup\limits_{-\infty<x<\infty} |F_n(x) - F(x)|$, 则有

$$\mathbb{P}\left(\lim_{n\to\infty} D_n = 0\right) = 1$$

1.3 样本均值与方差及其分布

Definition1.3.1

设 $x_1,\cdots,x_n$ 为取自某总体的样本, 若样本函数 $T=T(x_1,\cdots,x_n)$ 中不含有任何未知参数, 则称 T 为**统计量**, 统计量的分布成为**抽样分布**.

Definition1.3.2 样本均值 (sample mean)

设 $x_1,\cdots,x_n$ 为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.

在分组样本场合, 样本均值的近似公式为

$$\bar{x}=\sum_{i=1}^{k}x_if_i\Big/\sum_{i=1}^{k}f_i. \tag{1.3.1}$$

其中 k 为组数, f_i 为频数.

若把样本中的数据与样本均值之差成为偏差.

Propsition1.3.3 样本所有偏差之和为 0

$$\sum_{i=1}^{n}(x_i-\bar{x})=0$$

Propsition1.3.4 数据观测值与均值的偏差平方和最小

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\min_{c}\sum_{i=1}^{n}(x_i-c)^2$$

Theorem1.3.5 样本均值的抽样分布

$x_1,\cdots,x_n$ i. i. d. $\sim X$ 样本. $\bar{x}$ 为样本均值.

- $X\sim\mathcal{N}(\mu,\sigma^2)\Longrightarrow\bar{x}\sim\mathcal{N}(\mu,\sigma^2/n)$;
- $\exists\mathbb{E}(X)=\mu,\mathrm{Var}(X)=\sigma^2$, 当 n 较大时有渐进分布 (此处指近似分布): $\bar{x}\dot{\sim}\mathcal{N}(\mu,\sigma^2/n)$.

证明分别利用正态分布线性性质与中心极限定理给出, 即

$$\frac{n(\bar{x}-\mu)}{\sqrt{n}\sigma}\overset{\mathrm{w}}{\sim}\mathcal{N}(0,1)\Longrightarrow\bar{x}\dot{\sim}\mathcal{N}(\mu,\sigma^2/n).$$

Definition1.3.6 样本方差 (sample variance)

$$
\begin{aligned}
s_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2, && \text{称为样本方差;}\\
s_n &= \sqrt{s_n^2}, && \text{称为标准差;}\\
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2, && \text{称为无偏方差;}\\
s &= \sqrt{s^2}, && \text{称为修正样本标准差.}
\end{aligned}
$$

Remark1.3.7

今后方差指 s^2, 标准差指 s.

Propsition1.3.8 样本偏差平方和有三个常用表达式

$$
\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}=\sum_{i=1}^{n}x_i^2-n\bar{x}^2.
$$

分组样本场合, 样本方差的近似计算公式为

$$
s^2=\frac{1}{n-1}\sum_{i=1}^{k}f_i(x_i-\bar{x})^2=\frac{1}{n-1}\left(\sum_{i=1}^{k}f_ix_i^2-n\bar{x}^2\right),
$$

其中 x_i 为组中值, $\bar{x}$ 由(1.3.1)给出.

Theorem1.3.9

若总体 r. v. X 具有二阶矩, 即 $\mathbb{E}(X)=\mu$, $\mathrm{Var}(X)=\sigma^2<\infty$. $x_1,\cdots,x_n$ i. i. d. $\sim X$, 则

$$
\mathbb{E}(\bar{x})=\mu,\ \mathrm{Var}(\bar{x})=\sigma^2/n,\ \mathbb{E}(s^2)=\sigma^2
$$

1.4 样本矩, 样本偏度, 样本峰度

Definition1.4.1 样本矩

设 $x_1, \cdots, x_n$ 是 i. i. d. 样本.

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \qquad \text{称为样本 } k \text{ 阶原点矩;}$$

$$b_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k, \qquad \text{称为样本 } k \text{ 阶中心矩.}$$

样本的原点矩和中心矩统称为**样本矩** (sample moments).

Remark1.4.2

作为特例.

$$a_1 = \bar{x}, \quad b_2 = s_n^2 = (n-1)s^2/n.$$

Definition1.4.3 样本偏度

$x_1, \cdots, x_n$ i. i. d. $\sim F$, 称

$$\gamma_1 = \frac{b_3}{b_2^{3/2}} = \sqrt{n}\sum_{i=1}^{n} (x_i - \bar{x})^3 \Big/ \left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^{3/2}$$

为样本偏度 (sample skewness).

Remark1.4.4

$\gamma_1 = 0$ 说明是对称的, 类似于正态的去考虑;

$\gamma_1 > 0$ 说明数据中有几个较大的值, 反应总体分布正偏 (右偏);

$\gamma_1 < 0$ 反之.

Definition1.4.5 样本峰度

$x_1, \cdots, x_n$ i. i. d. $\sim F$, 称

$$\gamma_2 = \frac{b_4}{b_2^2} - 3 = n\sum_{i=1}^{n} (x_i - \bar{x})^4 \Big/ \left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^2 - 3$$

为样本峰度 (sample kurtosis).

Remark1.4.6

$\gamma_2 = 0$ 说明是标准的, 类似于正态的去考虑;

$\gamma_2 > 0$ 说明数据较集中 (在峰值附近分布密度曲线陡峭);

$\gamma_2 < 0$ 反之.

1.5 次序 (顺序) 统计量及其有关统计量

Definition1.5.1 次序统计量

$$x_1, \cdots, x_n \text{ i.i.d.} \sim \text{r.v. } X$$

$x_{(i)}$ 称为该样本的第 i 个**次序统计量**, 其取值是将样本观测值由小到大排列后得到的第 i 个观测值.

其中 $x_{(1)} = \min\{x_1, \cdots, x_n\}$ 称为该样本的**最小次序统计量**, 称 $x_{(n)} = \max\{x_1, \cdots .x_n\}$ 为该样本的**最大次序统计量**.

由于是 i.i.d. 取样. 我们可以有以下断言:

Propsition1.5.2

$$x_1, \cdots, x_n \text{ i.i.d.} \sim F$$

那么

$$\begin{aligned} F_{x_{(1)}}(z) =& 1 - \mathbb{P}\left(x_{(1)} > z\right) = 1 - \mathbb{P}\left(\bigcap_{k=1}^{n} x_k > z\right) \\ =& 1 - \prod_{k=1}^{n} \mathbb{P}\left(x_k > z\right) = 1 - \left[1 - F_X(z)\right]^n; \\ F_{x_{(n)}}(z) =& \mathbb{P}\left(x_{(n)} \leqslant z\right) = \mathbb{P}\left(\bigcap_{k=1}^{n} x_k \leqslant z\right) \\ =& \prod_{k=1}^{n} \mathbb{P}\left(x_k \leqslant z\right) = \left[F_X(z)\right]^n. \end{aligned}$$

Recall1.5.3 多项分布

多项分布是重要的多维离散分布, 它是二项分布的推广. 进行 n 次独立重复试验, 如果每次试验有 r 个互不相容的结果: $A_1, A_2, \cdots, A_r$ 之一发生, 且每次试验中 A_i 发生的概率为 $p_i = \mathbb{P}(A_i)$, $i = 1, 2, \cdots, r$, 且 $p_1 + p_2 + \cdots + p_r = 1$. 记 X_i 为 n 次独立重复试验中 A_i 出现的次数,$i = 1, 2, \cdots, r$, 则 $(X_1, X_2, \cdots, X_r)$ 取值 $(n_1, n_2, \cdots, n_r)$ 的概率, 即 A_1 出现 n_1 次, A_2 出现 n_2 次$\cdots\cdots A_r$ 出现 n_r 次的概率为

$$\mathbb{P}(X_1 = n_1, X_2 = n_2, \cdots, X_r = n_r) = \frac{n!}{n_1! n_2! \cdots n_r!} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$$

其中 $n = n_1 + n_2 + \cdots + n_r$.

这个联合分布称为 r 项分布, 又称多项分布, 记为 $M(n, p_1, p_2, \cdots, p_r)$. 这个概率是多项式 $(p_1 + p_2 + \cdots + p_r)^n$ 展开式中的一项, 故其和为 1. 当 $r = 2$ 时, 即为二项分布.

我们给出专属的记号:

$$\binom{n}{n_1\, n_2\, \cdots\, n_r} = \frac{n!}{n_1!n_2!\cdots n_r!}$$

Remark1.5.4

二项分布是一维随机变量的分布, 而在 r 项分布中, 因为 $p_1+p_2+\cdots+p_r=1$, 且 $n_1+n_2+\cdots+n_r=n$, 所以 r 项分布是 $r-1$ 维随机变量的分布.

Theorem1.5.5 单个次序统计量的分布

$$x_1,\cdots,x_n \overset{\text{i.i.d.}}{\sim} \mathrm{F}$$

p 为 F 的密度函数, 则第 k 个次序统计量 $x_{(k)}$ 的密度函数为

$$p_k(z) := p_{x_{(k)}}(z) = \binom{n}{k-1\ 1\ n-k} [F(z)]^{k-1} p(z) [1-F(z)]^{n-k}$$

证明. 证明是简单的, 基于下图.

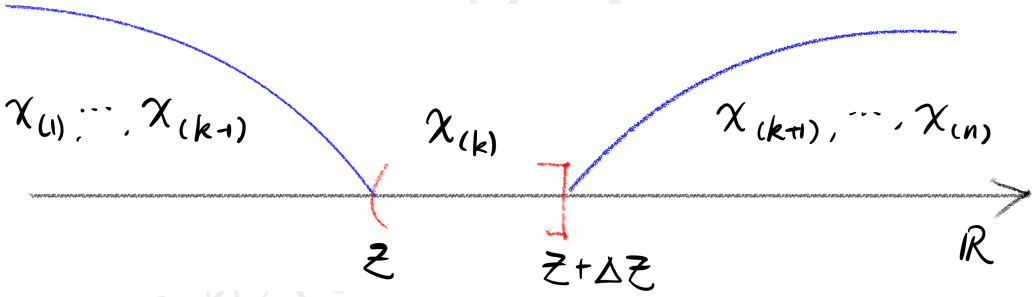

图 1.1: 单个次序统计量的分布

对任意的实数 z, 考虑次序统计量 $x_{(k)}$ 取值落在小区间 $(z, z+\Delta z]$ 内这一事件, 它等价于"样本容量为 n 的样本中有 1 个观测值落在区间 $(z, z+\Delta z]$ 之间, 而有 $k-1$ 个观测值小于等于 z, 有 $n-k$ 个观测值大于 $z+\Delta z$", 于是, 若以 $F_k(z)$ 记 $x_{(k)}$ 的分布函数, 则由多项分布可得

$$\begin{aligned}&\mathbb{P}(z \leqslant x_{(k)} \leqslant z+\Delta z) = F_k(z+\Delta z) - F_k(z)\\ \approx&\frac{n!}{(k-1)!(n-k)!}[F(z)]^{k-1}[F(z+\Delta z)-F(z)][1-F(z+\Delta z)]^{n-k}.\end{aligned}$$

两边同除以 Δz, 并令 $\Delta z \to 0$, 即有

$$\begin{aligned}p_k(z) =& \lim_{\Delta z\to 0}\frac{F_k(z+\Delta z)-F_k(z)}{\Delta z}\\ =&\frac{n!}{(k-1)!(n-k)!}[F(z)]^{k-1}p(z)[1-F(z)]^{n-k}.\end{aligned}$$

□

Theorem1.5.6 双个次序统计量的分布

次序统计量 $\left(x_{(i)}, x_{(j)}\right)$, $(i<j)$ 的联合分布密度函数为

$$\begin{aligned} p_{ij}(y,z) =& \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(y)]^{i-1}p(y)[F(z)-F(y)]^{j-i-1} \\ & \cdot p(z)[1-F(z)]^{n-j} \end{aligned}$$

证明. 证明与**定理1.5.5**的证明是类似的. □

Recall1.5.7 欧拉积分

回顾 Γ 函数与 B 函数:

$$\Gamma(s) = \int_0^{+\infty} x^{s-1}\mathrm{e}^{-x}\mathrm{d}x,\ s>0;$$
$$\mathrm{B}(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}\mathrm{d}x,\ p>0,\ q>0.$$

切有性质

- $\Gamma(s)$ 在定义域 $s>0$ 内连续且可导;
- $\Gamma(s+1)=s\Gamma(s)$, 当 s 取正整数 n, 有 $\Gamma(n+1)=n\Gamma(n)=n!\int_0^{+\infty}\mathrm{e}^{-x}\mathrm{d}x$;
- 延拓 $\Gamma(s)$ 至 $\mathbb{R}$, 方法是改写递推公式 $\Gamma(s)=\Gamma(s+1)/s$;
- $\mathrm{B}(p,q)$ 在定义域 $p>0$, $q>0$ 内连续;
- 对称: $\mathrm{B}(p,q)=\mathrm{B}(q,p)$;
- 递推公式:

$$\mathrm{B}(p,q) = \frac{q-1}{p+q-1}\mathrm{B}(p,q-1) \qquad (p>0,q>1);$$
$$\mathrm{B}(p,q) = \frac{p-1}{p+q-1}\mathrm{B}(p-1,q) \qquad (p>1,q>0);$$
$$\mathrm{B}(p,q) = \frac{(p-1)(q-1)}{(p+q-1)(p+q-2)}\mathrm{B}(p-1,q-1) \qquad (p>1,q>1).$$

- Γ 函数与 B 函数之间的关系

$$\mathrm{B}(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \quad (p>0,q>0)$$

Recall1.5.8 B 分布

$X \sim \mathrm{Be}(a,b),\ (a>0,b>0)$, 指密度函数为

$$\begin{aligned} p(x) =& \frac{1}{\mathrm{B}(a,b)} x^{a-1}(1-x)^{b-1},\ \ 0<x<1 \\ =& \begin{cases} \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, & 0<x<1; \\ 0, & \text{其他.} \end{cases} \end{aligned}$$

若 $X \sim \mathrm{B}(a,b)$, 则直接有

$$\mathbb{E}(X) = \frac{a}{a+b},\ \ \mathrm{Var}(x) = \frac{ab}{(a+b)^2(a+b+1)}$$

Recall1.5.9

总体 X 的分布函数 $F(x) = F_X(x)$ 是连续的, 则

$$Y = F(X) \sim \mathcal{U}(0,1)$$

Propsition1.5.10

$$x_1, \cdots, x_n \ \text{i. i. d.} \sim F \in \mathscr{C}(\mathbb{R})$$

以 $F\left(x_{(i)}\right), i = 1, \cdots, n$ 是来自总体为 $\mathcal{U}(0,1)$ 的次序统计量.

且 $x_{(k)} \sim \mathrm{Be}(k, n-k+1)$,

$$\mathbb{E}\left(x_{(k)}\right) = \frac{k}{n+1},\ \ \mathrm{Var}\left(x_{(k)}\right) = \frac{k(n-k+1)}{(n+1)^2(n+2)}.$$

证明. 证明是显然的, 由于 $F \in \mathscr{C}$ 是单调的, 自然有

$$F\left(x_{(1)}\right) \leqslant F\left(x_{(2)}\right) \leqslant \cdots \leqslant F\left(x_{(n)}\right)$$

第 k 个次序统计量 $x_{(k)}$ 的密度函数是

$$\begin{aligned} p_k(x) =& \begin{cases} \dfrac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}, & 0<x<1; \\ 0, & \text{其他.} \end{cases} \\ =& \begin{cases} \dfrac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} x^{k-1}(1-k)^{(n-k+1)-1}, & 0<x<1; \\ 0, & \text{其他.} \end{cases} \end{aligned}$$

□

Definition1.5.11 样本极差

对于 i. i. d. 样本 $x_1, \cdots, x_n$. 称统计量

$$R_n = x_{(n)} - x_{(1)}$$

为**样本极差** (sample range).

1.6 样本分位数与样本中位数

Recall1.6.1 分位数

设 $0<p<1$, 若 x_p 满足

$$\mathbb{P}(X \leqslant x_p) = F(x_p) = p$$

则称 x_p 为此分布的 p-分位数.

特别的, 称 $p=0.5$ 时的 p-分位数 $x_{0.5}$ 为**中位数**.

Definition1.6.2 总体分位数, 样本分位数

样本中位数 $m_{0.5}$ 是一个统计量, 定义为:

$$m_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n \text{ 为奇数}, \\ \dfrac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & n \text{ 为偶数}. \end{cases}$$

更一般的样本 p 分位数 m_p 定义为:

$$m_p = \begin{cases} x_{([np+1])}, & \text{若 } np \text{ 不是整数}, \\ \dfrac{1}{2}\left(x_{(np)} + x_{(np+1)}\right), & \text{若 } np \text{ 是整数}. \end{cases}$$

Theorem1.6.3 大样本下样本分为数的渐近分布

设总体密度函数为 $p(x)$, x_p 为其 p 分位数, $p(x)$ 在 x_p 处连续且 $p(x_p)>0$, 则当 $n \to \infty$ 时样本 p 分位数 m_p 的渐近分布为

$$m_p \dot{\sim} \mathcal{N}\left(x_p, \frac{p(1-p)}{n \cdot p^2(x_p)}\right).$$

特别, 对样本中位数, 当 $n \to \infty$ 时有渐近分布

$$m_{0.5} \dot{\sim} \mathcal{N}\left(x_{0.5}, \frac{1}{4n \cdot p^2(x_{0.5})}\right).$$

1.7 五数概括与箱线图

Definition1.7.1 五数概括

指五个数:

$$x_{\min} := x_{(1)}, \qquad Q_1 := m_{0.25}, \qquad m_{0.5}, \qquad Q_3 := x_{0.75}, \qquad x_{\max} := x_{(n)}$$

Definition1.7.2 箱线图

指通过以下步骤 (符合以下条件) 画出的图.

1. 箱体左右侧分别为 Q_1 与 Q_3, 在中位数位置上画上一条竖线.
2. 在箱子左右两侧分别引出一条水平线, 分别至最值位置.

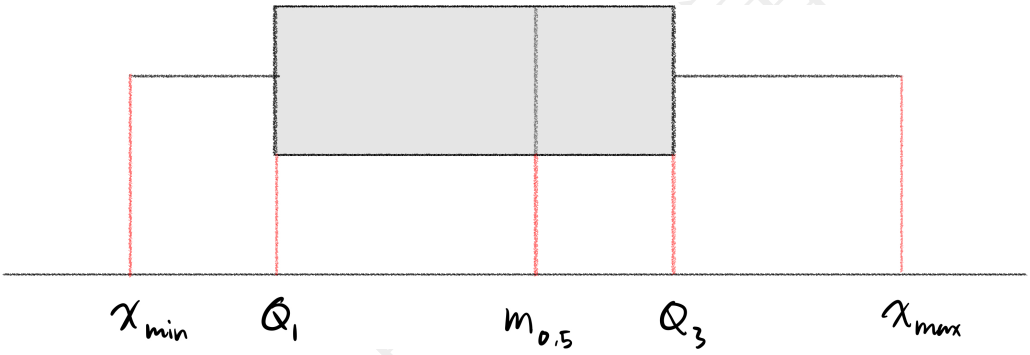

图 1.2: 箱线图

1.8 三大抽样分布

Recall1.8.1 Γ **分布** $\mathrm{Ga}(\alpha,\lambda)$

Γ 函数见回忆1.5.7.

若 r. v. X 的密度函数为

$$p_X(x)=\frac{x^{\alpha-1}}{\Gamma(\alpha)}\lambda^{\alpha}\mathrm{e}^{-\lambda x},\quad x\geqslant 0$$

则称 r. v. X 服从 Γ 分布, 记作 $X\sim\mathrm{Ga}(\alpha,\lambda)$, 其中 $\alpha>0$, $\lambda>0$.

Γ 分布的期望和方差分别为

$$\mathbb{E}(X)=\frac{\alpha}{\lambda},\qquad\qquad \mathrm{Var}(X)=\frac{\alpha}{\lambda^2}.$$

需要注意的是:

1. $\Gamma(1)=1$, $\Gamma(1/2)=\sqrt{\pi}$, $\Gamma(n+1)=n!$;
2. $\mathrm{Ga}(1,\lambda)=\mathrm{Exp}(\lambda)$, 先给出 $\mathrm{Ga}(n/2,1/2)=\chi^2(n)$;
3. $\mathbb{E}\left(\chi^2\right)=n$, $\mathrm{Var}\left(\chi^2\right)=2n$.

1.8.1 χ^2 分布

Definition1.8.2 χ^2 **分布**

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

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

的分布为**自由度**为 n 的 χ^2 **分布**, 记作 $\chi^2\sim\chi^2(n)$.

χ^2 分布的密度函数为:

$$p_{\chi^2}(x;n)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{x}{2}}, & x\geqslant 0\\ 0, & x<0.\end{cases}\tag{1.8.1}$$

Remark1.8.3

相关定理: 若 $X\perp\!\!\!\perp Y$(相互独立的), $h(x)$ 和 $g(y)$ 是实数域 $\mathbb{R}$ 上的连续函数, 那么随机变量 $h(X)\perp\!\!\!\perp g(Y)$.

由此可推出 $X_1^2,\cdots,X_n^2$ 相互独立.

Remark1.8.4

符号 “⫫” 表相互独立

Remark1.8.5

自由度是可以自由变化的观测值的个数.

提出疑问 “和分布的自由度就一定为 n 吗?”. 答案不一定, 因为如果 $X_1, X_2, \cdots, X_n$ 之间存在约束关系, 自由度就可能小于 n, 只有在它们相互独立且无约束时, 和分布的自由度才为 n.

Remark1.8.6

密度函数(1.8.1)的图像是一只取非负值的偏态分布.

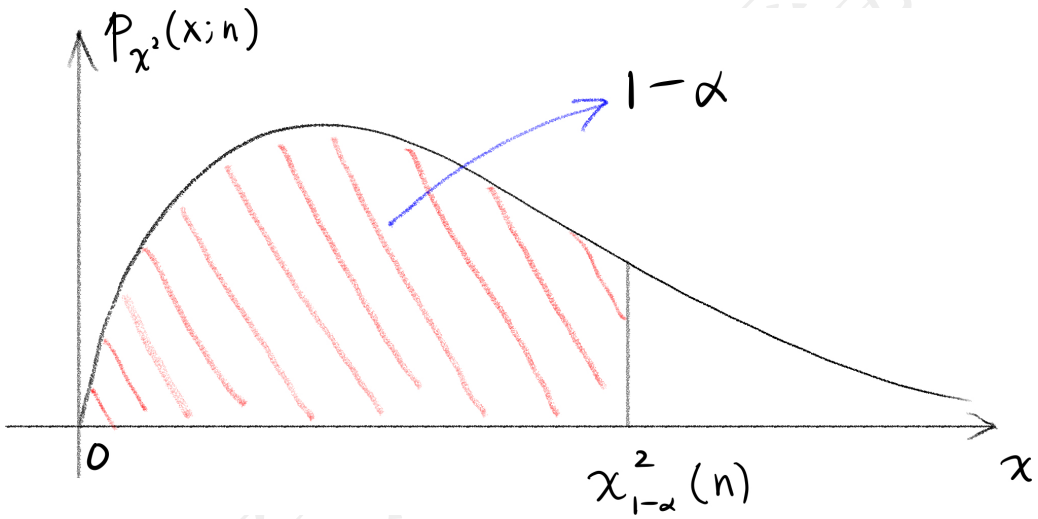

图 1.3: χ^2 分布的 $1-\alpha$ 分位数

Definition1.8.7 χ^2 分布的 $1-\alpha$ 分位数

当 r.v.$\chi^2 \sim \chi^2(n)$ 时, 对给定 $0<\alpha<1$, 称满足条件

$$\mathbb{P}\left\{\chi^2 \leqslant \chi^2_{1-\alpha}(n)\right\} = \int_0^{\chi^2_{1-\alpha}(n)} p_{\chi^2}(x;n)\mathrm{d}x = 1-\alpha$$

的点 $\chi^2_{1-\alpha}(n)$ 为 $\chi^2(n)$ 的 $1-\alpha$ 分位数.

Propsition1.8.8 χ^2 分布的性质

给出四个性质:

1. 设 $X_1, \cdots, X_n$ 互相独立, 切都服从正态分布 $\mathcal{N}(\mu, \sigma^2)$, 则

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

2. χ^2 分布的可加性: 设 $X_1 \sim \chi^2(n_1)$, $X_2 \sim \chi^2(n_2)$, 且 $X_1 \perp\!\!\!\perp X_2$, 那么

$$X_1 + X_2 \sim \chi^2(n_1 + n_2)$$

3. 若 r. v. $X \sim \chi^2(n)$, 则

$$\mathbb{E}(X) = n, \qquad \mathrm{Var}(X) = 2n.$$

4. 若 r. v. $X \sim \chi^2(n)$, 则对任意 x, 有

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X-n}{\sqrt{2n}} \leqslant x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}} \mathrm{d}t = \Phi(x)$$

即说明当 $n \to \infty$ 时, 有渐近分布

$$\frac{X-n}{\sqrt{2n}} \dot{\sim} \mathcal{N}(0,1), \qquad X \dot{\sim} \mathcal{N}(n, 2n).$$

证明. 现依次对命题1.8.8进行证明:

1. 显然

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

2. 正态分布可加性.

3. 由于 $X_i \sim \mathcal{N}(0,1)$, 故 $\mathbb{E}\left(X_i^2\right) = \mathrm{Var}(X_i) + \mathbb{E}^2(X_i) = 1$. 利用分布积分法与下式

$$\mathbb{E}\left(X_i^2\right) = \int_{-\infty}^{+\infty} \frac{x^2}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x, \qquad \mathbb{E}\left(X_i^4\right) = \int_{-\infty}^{+\infty} \frac{x^4}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x$$

显然有 $\mathbb{E}\left(X_i^4\right) = 3\mathbb{E}\left(X_i^2\right) = 3$. 于是

$$\mathrm{Var}\left(X_i^2\right) = \mathbb{E}\left(X_i^4\right) - \mathbb{E}^2\left(X_i^2\right) = 3 - 1 = 2$$

于是

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

$$\mathrm{Var}(X) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i^2\right) \overset{\text{独立}}{=} \sum_{i=1}^{n} \mathrm{Var}\left(X_i^2\right) = 2n.$$

4. 由中心极限定理得当 $n \to \infty$ 时, 有

$$\mathbb{P}\left\{\frac{X-n}{\sqrt{2n}} \leqslant x\right\} = \mathbb{P}\left\{\frac{\sum_{i=1}^{n} X_i^2 - \mathbb{E}\left(\sum_{i=1}^{n} X_i^2\right)}{\sqrt{\mathrm{Var}\left(\sum_{i=1}^{n} X_i^2\right)}} \leqslant x\right\} \xrightarrow{\mathrm{w}} \Phi(X).$$

从而

$$\frac{X-n}{\sqrt{2n}} \overset{\mathrm{w}}{\sim} \mathcal{N}(0,1), \qquad X \overset{\mathrm{w}}{\sim} \mathcal{N}(n, 2n).$$

□

Propsition1.8.9 简单样本方差的方差 (无偏)

设

$$x_1, x_2, \cdots, x_n \ \text{i. i. d.} \sim \mathcal{N}(\mu, \sigma^2)$$

则方差为

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

由于

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

则

$$\mathrm{Var}\left[\frac{(n-1)s^2}{\sigma^2}\right] = 2(n-1) \Longrightarrow \mathrm{Var}(s^2) = \frac{2\sigma^4}{n-1}$$

非中心 χ^2 分布

Definition1.8.10

设随机变量 $X_1, \cdots, X_n$ 互相独立, $X_i \sim \mathcal{N}(a_i, 1)$, $a_i (i = 1, \cdots, n)$ 不全为 0.

记 $Y = \sum_{i=1}^{n} X_i^2$, 则称 Y 的分布是自由度为 n 和非中心参数为 $\delta = \sqrt{\sum_{i=1}^{n} a_i^2}$ 的**非中心 χ^2 分布**, 记为 $Y \sim \chi^2(n;\delta)$ (有的书上记作 $Y \sim \chi^2_{n,\delta}$). 特别当 $\delta = 0$ 时称为中心的 χ^2 分布, 记前面所述的 χ^2 分布.

若 $Y \sim \chi^2_{n,\delta}$, 则其密度函数为

$$p_{\chi^2}(x;n,\delta) = \begin{cases} \mathrm{e}^{-\delta^2/2}\displaystyle\sum_{i=0}^{\infty}\frac{1}{i!}\left(\frac{\delta^2}{2}\right)^i \frac{x^{i+n/2-1}}{2^{i+n/2}\Gamma(n/2+i)}\mathrm{e}^{-x/2}, & x > 0, \\ 0, & x \leqslant 0 \end{cases}$$

$$= \begin{cases} \mathrm{e}^{-\delta^2/2}\displaystyle\sum_{i=0}^{\infty}\frac{(\delta^2/2)^i}{i!}p_{\chi^2}(x;2i+n), & x > 0, \\ 0, & x \leqslant 0. \end{cases}$$

参考韦来生《数理统计》第 35 页.

1.8.2 F 分布

Definition1.8.11 F 分布

设 $U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$, $U \perp\!\!\!\perp V$, 则称随机变量

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

服从自由度为 n_1 及 n_2 的 F 分布, n_1(resp. n_2) 称为分子 (第一)(resp. 分母 (第二)) 自由度.

若 $F \sim F(n_1, n_2)$, 则 F 有密度函数为

$$p_F(x; n_1, n_2) = \begin{cases} \dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)} \left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}} \cdot x^{\frac{n_1}{2}-1} \cdot \left(1+\dfrac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, & y > 0, \\ 0, & y \leqslant 0. \end{cases}$$

Propsition1.8.12 F 分布的性质

给出四个性质:

1. 若 $F \sim F(n_1, n_2)$, 则

$$\begin{aligned} \mathbb{E}(F) &= \frac{n_2}{n_2-2}, && \text{若} n_2 > 2; \\ \mathrm{Var}(F) &= \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}, && \text{若} n_2 > 4; \\ (\star\text{补充})\mathbb{E}(F^r) &= \left(\frac{n_2}{n_1}\right)^r \frac{\Gamma\left(\frac{n_1}{2}+r\right)\Gamma\left(\frac{n_2}{2}-r\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}, && 0 < 2r < n_2. \end{aligned}$$

2. 若 $F \sim F(n_1, n_2)$, 则

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

3. 若 $F \sim F(n_1, n_2)$, 则当 $n_2 > 4$ 时, 对任意 x, 有

$$\mathbb{P}\left\{\frac{F-\mathbb{E}(F)}{\sqrt{\mathrm{Var}(F)}}\right\} \xrightarrow{\mathrm{w}} \Phi(x), \quad n_1 \to +\infty.$$

4. 对于固定的 n_1, 有

$$F(n_1, n_2) \dot{\sim} \frac{1}{n_1}\chi^2(n_1), \quad n_2 \to +\infty.$$

证明. 对于命题1.8.12的性质4, 关键在于下式 (大数定律),

$$V/n_2 \xrightarrow{\mathrm{p}} 1, \quad n_2 \to +\infty.$$

□

Remark1.8.13

F 分布的密度函数图像有以下三个特点:

1. 一只取非负值的偏态 (右偏) 分布;
2. 固定 n_1, n_2 越大, 越渐近于 $\chi^2(n_1)/n_1$ 分布;
3. n_1 越大, 越渐近于正态分布.

Definition1.8.14 F 分布的分位数

对于给定的 $0<\alpha<1$, 称满足条件

$$\mathbb{P}\left\{F \leqslant F_{1-\alpha}\left(n_1, n_2\right)\right\}=\int_0^{F_{1-\alpha}\left(n_1, n_2\right)} p_F(x ; n_1, n_2) \mathrm{d} x=1-\alpha$$

的点 $F_{1-\alpha}\left(n_1, n_2\right)$ 是 $F\left(n_1, n_2\right)$ 分布的 $1-\alpha$ 分位数.

Propsition1.8.15

F 分布的分位数有以下等式,

$$F_\alpha(n_1, n_2)=\frac{1}{F_{1-\alpha}\left(n_1, n_2\right)} \tag{1.8.2}$$

证明. 现在来证明等式(1.8.15):

设 $F \sim F\left(n_1, n_2\right)$, 则 $1/F \sim F\left(n_2, n_1\right)$. 关键在于下式.

$$1-\alpha=\mathbb{P}\left\{F \leqslant F_{1-\alpha}\left(n_1, n_2\right)\right\}=\mathbb{P}\left\{\frac{1}{F} \geqslant \frac{1}{F_{1-\alpha}\left(n_1, n_2\right)}\right\}.$$

从而有

$$\alpha=1-\mathbb{P}\left\{\frac{1}{F} \geqslant \frac{1}{F_{1-\alpha}\left(n_1, n_2\right)}\right\}=\mathbb{P}\left\{\frac{1}{F} \leqslant \frac{1}{F_{1-\alpha}\left(n_1, n_2\right)}\right\}.$$

由于

$$\mathbb{P}\left\{\frac{1}{F} \leqslant F_\alpha\left(n_2, n_1\right)\right\}=\alpha.$$

从而得证. □

非中心 F 分布

Definition1.8.16

设随机变量 $X \sim \chi^2(n_1 ; \delta)$, $Y \sim \chi^2(n_2)$ 且 $X \perp\!\!\!\perp Y$, 则称

$$Z=\frac{X / n_1}{Y / n_2}$$

的分布是自由度为 n_1, n_2 和非中心参数为 δ 的**非中心 F 分布**, 记为 $Z \sim F(n_1, n_2 ; \delta)$. 当 $\delta=0$ 时, 称 Z 的分布为中心的 F 分布, 即前面定义的 $F(n_1, n_2)$.

若 $Z \sim F(n_1, n_2; \delta)$, 则 Z 的密度函数 $p_Z(x) := p_F(x; n_1, n_2, \delta)$ 为

$$p_Z(x) = \begin{cases} \dfrac{n_1^{\frac{n_2}{2}} n_2^{\frac{n_1}{2}}}{\Gamma\left(\frac{n}{2}\right)} \mathrm{e}^{\frac{-\delta^2}{2}} x^{\frac{n_1}{2}-1} \displaystyle\sum_{k=0}^{+\infty} \frac{\left(\frac{\delta^2 n_1 x}{2}\right)^k \Gamma\left(\frac{m+n}{2}+k\right)}{k!\,\Gamma\left(\frac{n_1}{2}+k\right)(mx+n)^{\frac{m+n}{2}+k}}, & x > 0 \\ 0, & \text{其他.} \end{cases}$$

参考韦来生《数理统计》第 41 页.

1.8.3 t 分布

Definition1.8.17

设 r. v. X_1 丄 X_2 , 且 $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \chi^2(n)$, 则称随机变量

$$t = \frac{X_1}{\sqrt{X_2/n}} \sim t(n).$$

服从自由度为 n 的 t 分布, 又称学生氏 (Student) 分布.

若 $t \sim t(n)$, 则 t 有密度函数为

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

Propsition1.8.18

t 分布的密度函数图像有以下性质:

1. 一只关于纵轴对称的图像;
2. 峰比正态低, 尾比正态厚;
3. 自由度 n 越大, 越渐近标准正态分布图像.

Recall1.8.19 Cauchy 分布

Cauchy 分布是一种连续型概率分布, 记为 $\mathcal{C}(\theta, \alpha)$, 其中 θ 为位置参数, 决定分布峰值位置, α 为尺度参数, 控制分布的分散程度.

其概率密度函数为

$$p_{\mathcal{C}}(x; \theta, \alpha) = \frac{1}{\pi\alpha\left(1 + \left(\frac{x-\theta}{\alpha}\right)^2\right)},$$

当 $\theta=0,\alpha=1$ 时, 得到标准 Cauchy 分布, 概率密度函数为

$$p_C(x;0,1)=\frac{1}{\pi(1+x^2)}$$

有如下性质

1. 期望和方差不存在.
2. 关于位置参数 θ 对称, 是左右对称的分布.
3. 两个独立的 Cauchy 分布变量相加, 结果仍服从 Cauchy 分布.
4. (重尾性) 相比于正态分布, Cauchy 分布的尾部下降至 0 的速度慢很多, 意味着极端值出现的概率相对较大.

Propsition1.8.20 t 分布的性质

t 分布有如下性质:

1. 自由度为 1 的 t 分布就是标准 Cauchy 分布, 他的均值不存在.
2. 当 $n>1$ 时, 若 $t\sim t(n)$, 则 $\mathbb{E}(t)=0$;
3. 当 $n>2$ 时, 若 $t\sim t(n)$, 则 $\exists\ \mathrm{Var}(t)=\dfrac{n}{n-2}$.
4. 当自由度较大, 可用标准正态近似, 数学一点就是

$$t(n)\dot{\sim}\mathcal{N}(0,1),\quad n\to+\infty.$$

Definition1.8.21 t 分布的分位数

当 r. v. $t\sim t(n)$ 时, 称满足

$$\mathbb{P}\{t\leqslant t_{1-\alpha}(n)\}=\int_{-\infty}^{t_{1-\alpha}(n)}p_t(x;n)\mathrm{d}x=1-\alpha$$

的 $t_{1-\alpha}(n)$ 是自由度为 n 的 t 分布的 $1-\alpha$ 分位数.

Propsition1.8.22

t 分布的分位数有如下性质:

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

Example1.8.23

设总体 X,Y i. i. d. $\sim\mathcal{N}(0,9)$, 而

$$X_1,\cdots,X_9\ \text{i. i. d.}\sim X,\qquad Y_1,\cdots,Y_9\ \text{i. i. d.}\sim Y.$$

那么

$$U=\frac{\sum_{i=1}^{9}X_i}{\sqrt{\sum_{j=1}^{9}Y_j^2}}\sim t(9).$$

证明是简单的, 注意到

$$\sum_{i=1}^{9}X_i\sim\mathcal{N}(0,81),\qquad\qquad \left(\frac{Y_j}{3}\right)^2\sim\chi^2(1).$$

那么自然有

$$\frac{1}{9}\sum_{i=1}^{9}X_i\sim\mathcal{N}(0,1),\qquad\qquad \sum_{j=1}^{9}\frac{Y_j^2}{9}=\frac{1}{9}\sum_{j=1}^{9}Y_j^2\sim\chi^2(9)$$

且

$$\left(\frac{1}{9}\sum_{i=1}^{9}X_i\right)\perp\!\!\!\perp\left(\frac{1}{9}\sum_{j=1}^{9}Y_j^2\right),$$

故

$$U=\frac{1}{9}\sum_{i=1}^{9}X_i\Bigg/\sqrt{\frac{1}{81}\sum_{j=1}^{9}Y_j^2}=\frac{\sum_{i=1}^{9}X_i}{\sqrt{\sum_{j=1}^{9}Y_j^2}}\sim t(9).$$

非中心 t 分布

Definition1.8.24

设随机变量 $X\sim\mathcal{N}(\delta,1)$, $Y\sim\chi^2(n)$, 且 $X\perp\!\!\!\perp Y$, 则称

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

的分布是自由度为 n 和非中心参数为 δ 的**非中心 t 分布**, 记为 $Z\sim t(n;\delta)$ (也有书上记作 $\sim t_{n,\delta}$). 特别当 $\delta=0$ 时的分布称为中心的 t 分布, 即前面所述的 $t(n)$ 分布.

非中心 t 分布的密度函数为

$$p_t(x;n,\delta)=\frac{n^{n/2}}{\sqrt{\pi}\cdot\Gamma(n/2)}\cdot\frac{\mathrm{e}^{-\delta^2/2}}{(n+x^2)^{\frac{n+1}{2}}}\sum_{i=0}^{+\infty}\Gamma\left(\frac{n+i+1}{2}\right)\frac{(\delta x)^i}{i!}\left(\frac{2}{n+x^2}\right)^{i/2},$$

$$-\infty<x<+\infty.$$

参考韦来生《数理统计》第 38 页.

1.9 三大抽样分布与正态总体间的关联

Theorem1.9.1 正态总体样本线性函数的分布

设

$$x_1,\cdots,x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu,\sigma^2).$$

令 $G=\sum_{i=1}^n a_ix_i$, 其中 $a_1,\cdots,a_n$ 为已知常数. 那么

$$G\sim\mathcal{N}(\mu\sum_{i=1}^n a_i,\sigma^2\sum_{i=1}^n a_i^2).$$

当 a_i 取到 $1/n$, $i=1,\cdots,n$. 那么归为定理1.3.5, 即

$$G=\bar{x}\sim\mathcal{N}\left(\mu,\frac{\sigma^2}{n}\right)\overset{\text{i.e.}}{\Longrightarrow}\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim\mathcal{N}(0,1).$$

Theorem1.9.2

设

$$x_1,\cdots,x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu,\sigma^2).$$

其样本均值与样本方差分别为

$$\bar{x}=\frac{1}{n}\sum_{i=1}^n x_i,\qquad s^2=\frac{1}{n-1}\sum_{i=1}^n(x_i-\bar{x})^2.$$

则有

1. $\bar{x}$ 丄 s^2;
2. $\bar{x}\sim\mathcal{N}(\mu,\sigma/n)$;
3. $(n-1)s^2/\sigma^2\sim\chi^2(n-1)$, 即 $ns_n^2/\sigma^2\sim\chi^2(n-1)$.

证明. 根据定理1.1.7, 有

$$\begin{aligned}p(x_1,\cdots,x_n)=&\prod_{i=1}^n\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=(2\pi\sigma^2)^{-\frac{n}{2}}\mathrm{e}^{-\sum\limits_{i=1}^n\frac{(x_i-\mu)}{2\sigma^2}}\\=&(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{\sum\limits_{i=1}^n x_i^2-2n\bar{x}\mu+n\mu^2}{2\sigma^2}\right\}\end{aligned}$$

记

$$\boldsymbol{X}=(x_1,\cdots,x_n)^\top$$

给出一个 n 维的正交阵

$$\boldsymbol{A}=\begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2\cdot 1}} & -\frac{1}{\sqrt{2\cdot 1}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{3\cdot 2}} & \frac{1}{\sqrt{3\cdot 2}} & -\frac{1}{\sqrt{3\cdot 2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & -\frac{n-1}{\sqrt{n(n-1)}} \end{pmatrix}$$

令

$$\boldsymbol{Y}=(y_1,\cdots,y_n)^\top=\boldsymbol{AX}$$

注意到该变换的 Jacobi 行列式 $\det(A)=1$. 且有

$$\bar{x}=\frac{1}{\sqrt{n}}y_1, \qquad \sum_{i=1}^n y_i^2=\boldsymbol{Y}^\top\boldsymbol{Y}=\boldsymbol{X}^\top\boldsymbol{A}^\top\boldsymbol{AX}=\sum_{i=1}^n x_i^2.$$

由 $p_{\boldsymbol{Y}}(\boldsymbol{Y})=p_{\boldsymbol{X}}(\boldsymbol{X})|\boldsymbol{A}|^{-1}=p_{\boldsymbol{X}}(\boldsymbol{X})$, 从而得到密度函数

$$\begin{aligned} p(y_1,\cdots,y_n)=&(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{\sum\limits_{i=1}^n y_i^2-2\sqrt{n}y_1\mu+n\mu^2}{2\sigma^2}\right\} \\ =&(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{\sum\limits_{i=1}^n y_i^2+(y_1-\sqrt{n}\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

从而 $\boldsymbol{Y}$ 的各个分量互相独立, 切都服从正态分布:

$$y_1\sim\mathcal{N}(\sqrt{n}\mu,\sigma^2), \qquad y_j\sim\mathcal{N}(0,\sigma^2),\ \forall j\neq 1.$$

从而结论2得证.

由于

$$(n-1)s^2=\sum_{i=1}^n(x_i-\bar{x})^2=\sum_{i=1}^n x_i^2-\left(\sqrt{n}\bar{x}\right)^2=\sum_{i=1}^n y_i^2-y_i^2=\sum_{i\neq 1}y_i^2$$

而

$$\bar{x}=\frac{1}{\sqrt{n}}y_1 \perp\!\!\!\perp \sum_{i\neq 1}y_i^2$$

从而结论1得证.

由于

$$y_2, \cdots, y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),$$

于是

$$\frac{y_i}{\sigma} \sim \mathcal{N}(0,1), \ \forall i \neq 1.$$

从而

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i \neq 1} \left(\frac{y_i}{\sigma}\right)^2 \sim \chi^2(n-1).$$

从而结论3得证. □

Corollary1.9.3

设

$$x_1, \cdots, x_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$$
$$y_1, \cdots, y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$$

且这两组样本互相独立, 则有

$$F = \frac{s_x^2/\sigma_1^2}{s_y^2/\sigma_2^2} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 \Big/ (m\sigma_1^2 - \sigma_1^2)}{\sum_{i=1}^{n} (y_i - \bar{y})^2 \Big/ (n\sigma_2^2 - \sigma_2^2)} \sim F(m-1, n-1).$$

其中

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i;$$
$$s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m} (x_i - \bar{x})^2, \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2.$$

特别, 若 $\sigma_1^2 = \sigma_2^2$, 则

$$F = \frac{s_x^2}{s_y^2} \sim F(m-1, n-1).$$

证明. 证明是基于定理1.9.2.

由于这两组样本互相独立, 那么 $s_x^2 \perp\!\!\!\perp s_y^2$, 由定理1.9.2有

$$\frac{(m-1)s_x^2}{\sigma_1^2} \sim \chi^2(m-1), \qquad \frac{(n-1)s_y^2}{\sigma_2^2} \sim \chi^2(n-1).$$

再利用定义1.8.11得证. □

Corollary1.9.4

设

$$x_1, \cdots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2).$$

$\bar{x}$ 与 s^2 分别为样本均值与样本方差, 则有

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

证明. 由定理1.9.2知, $\bar{x} \perp\!\!\!\perp s^2$, 且有

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1), \qquad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1).$$

由定义1.8.17立即有 (得证)

$$\frac{\dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)s^2}{\sigma^2} \Big/ (n-1)}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

类似的可以得到

$$\frac{\bar{x} - \mu}{s_n/\sqrt{n-1}} \sim t(n-1).$$

这事实上只是形式上的变形. □

Corollary1.9.5

再推论1.9.3的记号下, 设 $\sigma_1^2 = \sigma_2^2 = \sigma^2$, 并记

$$s_w^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2} = \frac{\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2}{m+n-2}$$

则

$$\frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{s_w\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}} \sim t(m+n-2)$$

证明. 由正态分布的可加性与

$$\bar{x} \sim \mathcal{N}(\mu_1, \sigma^2/m), \qquad \bar{y} \sim \mathcal{N}(\mu_2, \sigma^2/n), \qquad \bar{x} \perp\!\!\!\perp \bar{y}.$$

立刻得

$$\bar{x} - \bar{y} \sim \mathcal{N}(\mu_1 - \mu_2, \sigma^2/m + \sigma^2/n) \overset{\text{i.e.}}{\Longrightarrow} \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}} \sim \mathcal{N}(0,1)$$

利用定理1.9.2知

$$\frac{(m-1)s_x^2}{\sigma^2} \sim \chi^2(m-1), \qquad \frac{(n-1)s_y^2}{\sigma^2} \sim \chi^2(n-1).$$

且 $s_x^2 \perp\!\!\!\perp s_y^2$, 由 χ^2 分布得可加性得

$$\frac{(m-1)s_x^2+(n-1)s_y^2}{\sigma^2} = \frac{(m+n-2)s_w^2}{\sigma^2} \sim \chi^2(m+n-2).$$

由于 $(\bar{x}-\bar{y}) \perp\!\!\!\perp s_w^2$, 利用定义1.8.17立即有 (得证)

$$\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}} \Bigg/ \sqrt{\frac{(m+n-2)s_w^2}{\sigma^2} \Bigg/ (m+n-2)}$$
$$= \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{s_w\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(m+n-2).$$

□

我们将上述内容分类归位如下两部分:

1.9.1　单个正态总体的抽样分布

设 $x_1, \cdots, x_m$ i. i. d. $\sim \mathcal{N}(\mu, \sigma^2)$. 那么

1. $\dfrac{\bar{x}-\mu}{\sigma/\sqrt{m}} \sim \mathcal{N}(0,1), \quad \dfrac{\bar{x}-\mu}{s/\sqrt{m}} \sim t(m-1).$
2. $\dfrac{(m-1)s^2}{\sigma^2} \sim \chi^2(m-1), \quad \dfrac{ms_m^2}{\sigma^2} \sim \chi^2(m-1).$
3. $\displaystyle\sum_{i=1}^{m}\left(\frac{x_1-\mu}{\sigma}\right)^2 \sim \chi^2(m), \quad \sum_{i=1}^{m}\left(\frac{x_1-\bar{x}}{\sigma}\right)^2 \sim \chi^2(m-1).$

1.9.2　双个正态总体的抽样分布

设 $x_1, \cdots, x_m$ i. i. d. $\sim \mathcal{N}(\mu_1, \sigma_1^2)$, $y_1, \cdots, y_n$ i. i. d. $\sim \mathcal{N}(\mu_2, \sigma_2^2)$. 那么

1. $\dfrac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}} \sim \mathcal{N}(0,1).$
2. $\dfrac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{s_w\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(m+n-2), (\sigma_1^2=\sigma_2^2=\sigma^2).$

其中 $s_w^2 = \dfrac{(m-1)s_x^2+(n-1)s_y^2}{m+n-2} = \dfrac{\sum_{i=1}^{m}(x_i-\bar{x})^2+\sum_{i=1}^{n}(y_i-\bar{y})^2}{m+n-2}$

3. $F = \dfrac{s_x^2/\sigma_1^2}{s_y^2/\sigma_2^2} \sim F(m-1, n-1)$.

4. $F = \dfrac{s_x^2}{s_y^2} \sim F(m-1, n-1), (\sigma_1^2 = \sigma_2^2)$.

1.10 充分统计量

Definition1.10.1 1922,Fisher: 充分统计量

$$x_1,\cdots,x_n \overset{\text{i.i.d}}{\sim} F(x;\theta)$$

统计量 $T=T(x_1,\cdots,x_n)$ 称为 θ 的充分统计量, 如果在给定 T 的取值 t 后, 条件分布 $F(x_1,\cdots,x_n|T=t)$ 与参数无关.

Remark1.10.2 充分统计量的意义

如果知道了统计量 T 的观察值以后, 样本的条件分布与 θ 无关, 也就是样本的剩余部分不再包含关于 θ 的信息, 换言之, 在 T 中包含了关于 θ 的全部信息, 因此要做关于 θ 的统计推断, 只需用统计量 T 就足够.

要注意充分统计量不唯一.

Theorem1.10.3 因子分解定理

$$X_1,\cdots,X_n \overset{\text{i.i.d.}}{\sim} p(x;\theta)$$

则 $T=T(X_1,\cdots,X_n)$ 是充分统计量的充分必要条件是: 存在两个函数 $g(t;\theta)$ 和 $h(x_1,\cdots,x_n)$ 使得对任意的 θ 和任一组观测值 $x_1,\cdots,x_n$, 有

$$p(x_1,\cdots,x_n;\theta)=g(T(x_1,\cdots,x_n);\theta)h(x_1,\cdots,x_n)$$

其中 $g(t,\theta)$ 是通过统计量 T 的取值而依赖于样本的.

分布	参数	充分统计量
两点分布 $\mathrm{b}(1,p)$	p	$T=\overline{x}$
Poisson 分布 $\mathcal{P}(\lambda)$	λ	$T=\overline{x}$
几何分布 $\mathrm{Ge}(\theta)$	θ	$T=\overline{x}$
均匀分布 $\mathcal{U}(0,\theta)$	θ	$T=X_{(n)}$
均匀分布 $\mathcal{U}(\theta_1,\theta_2)$	(θ_1,θ_2)	$\left(X_{(1)},X_{(n)}\right)$
均匀分布 $\mathcal{U}(\theta,2\theta)$	θ	$\left(X_{(1)},X_{(n)}\right)$
正态分布 $\mathcal{N}(\mu,\sigma^2)$	(μ,σ^2)	$(\overline{x},s^2)$ 或 $(\overline{x},\sum_{i=1}^n x_i^2)$ 或 $(\sum_{i=1}^n x_i,\sum_{i=1}^n x_i^2)$
指数分布 $\mathrm{Exp}(\lambda)$	λ	$T=\overline{x}$
伽玛分布 $\mathrm{Ga}(\alpha,\lambda)$	(α,λ)	$(\prod_{i=1}^n x_i,\sum_{i=1}^n x_i)$

表 1.1: 常见分布的充分统计量

第二章　(参数) 点估计

参数估计 (parameterestimation) 问题常有两类: **点估计**和**区间估计**.

点估计就是用样本函数的一个具体数值去估计一个未知参数. 区间估计就是用样本函数的两个值构成的区间去估计未知参数的取值范围.

例如, 在某市居民年人均收入的调查中, 估计该市居民的年人均收入为 18250 元, 这是一个点估计, 若估计年人均收入在 16350 元到 19850 元之间, 这就是一个区间估计.

点估计与区间估计是互为补充的参数估计形式, 本章主要是讨论参数的点估计问题, 区间估计问题在第三章讨论.

Definition2.0.1 点估计 (point estimation)

$$X_1, \cdots, X_n \text{ i. i. d. } \sim \text{ r. v. } X.$$

$\hat{g}(X_1, \cdots, X_n)$ 是样本的函数, 用 $\hat{g}(X_1, \cdots, X_n)$ 作为 $g(\theta)$ 的估计, 称为**点估计**

2.1 矩估计

$$x_1, \cdots, x_n \overset{\text{i.d.d.}}{\sim} \text{r.v.} X.$$

对于 $k = 1, 2, \cdots$ 分别做如下约定记号:

$$\mu_k = \mathbb{E}(X^k), \qquad a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \qquad k \text{ 阶原点矩.}$$

$$v_k = \mathbb{E}[X - \mathbb{E}(X)]^k, \qquad b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k, \qquad k \text{ 阶中心矩.}$$

实际上, 若 $|\mathbb{E}(X)| = |\mu| < +\infty$, 基于大数定律, 则

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{\text{p}} \mathbb{E}(X) = \mu$$

$$\underbrace{a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k}_{\text{样本原点矩}} \xrightarrow{\text{p}} \underbrace{\mathbb{E}(X^k) = \mu_k}_{\text{总体原点矩}}, \quad (k = 1, 2, \cdots)$$

一句话来说

$$g(a_1, \cdots, a_k) \xrightarrow{\text{p}} g(\mu_1, \cdots, \mu_k), \quad (n \to \infty, g \in \mathscr{C}(\mathbb{R}^k)).$$

Definition2.1.1 矩估计法

用**样本 (原点) 矩**估计相应的**总体 (原点) 矩**, 又用样本 (原点) 矩的连续函数估计相应的总体 (原点) 矩的连续函数, 这种参数点估计法称为**矩估计法**, 或称**矩法 (moment method of estimation)**. 即

$$\mathbb{E}(X^K) = \mu_k \approx a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k,$$

$$g(\mu_1, \cdots, \mu_k) \approx g(a_1, \cdots, a_k).$$

矩法的是指是用经验分布函数去替换总体分布函数, 理论基础是 Glivenko-Cantelli 定理 (1.2.9).

矩法的优缺点: 优点: 直观简单, 不需知道总体分布形式, 对期望方差极其方便; 缺点: 若总体原点矩不存在则失效, 不具有唯一性, 没有充分利用分布.

2.2 最 (极) 大似然估计

Propsition2.2.1 最大似然原理

一次试验中发生了的时间的概率是所有可能的概率中最大的.

若 $X_1,\cdots,X_n$ 是取自总体 r. v. X(连续型) 的一个样本, 已知其概率密度为 $p(x;\theta)$, $\theta\in\Theta$ 为待估参数, 样本的联合概率密度为

$$p(x_1,\cdots,x_n;\theta)=\prod_{i=1}^{n}p(x_i;\theta)$$

设 $x_1^\star,\cdots,x_n^\star$ 为相应 $X_1,\cdots,X_n$ 的样本值, 则随机点 $(X_1,\cdots,X_n)$ 落在 $(x_1^\star,\cdots,x_n^\star)$ 的领域 (边长分别为 $\mathrm{d}x_1,\cdots,\mathrm{d}x_n$ 的 n 维立方体) 内的概率近似为:

$$\prod_{i=1}^{n}p(x_i;\theta)\mathrm{d}x_i$$

因此在得到样本值 $(x_1^\star,\cdots,x_n^\star)$ 的情况下, 自然应当选取使得 $\prod_i p(x_i^\star;\theta)\mathrm{d}x_i$ 达到最大的 θ 的值作为未知参数 θ 的估计值.

显然当未知参数 θ 等于这个值时, 出现给定的那个样本观测值的可能性最大.

注意到 $\prod_i \mathrm{d}x_i$ 不随 θ 而变, 故考虑

$$L(\theta)=p(x_1^\star,\cdots,x_n^\star;\theta)=\prod_{i=1}^{n}p(x_i^\star;\theta)$$

的最大值, 这里 $L(\theta)$ 称为**样本的似然函数**.

Definition2.2.2 最大似然估计原理

设 $X_1,\cdots,X_n$ 是取自总体 X 的一个样本, 样本的联合密度 (连续型) 或联合概率密度 (离散型) 为

$$p(x_1,\cdots,x_n;\theta)$$

当给定样本 $X_1,\cdots,X_n$ 时, 定义似然函数为:

$$L(\theta)=p(x_1^\star,\cdots,x_n^\star;\theta)$$

这里 $x_1^\star,\cdots,x_n^\star$ 是样本的观察值.

最大似然估计就是使样本的似然函数 $L(\theta)$ 达到最大值.

Definition2.2.3 似然函数

设总体的概率密度为 $p(x;\theta)$, Θ 是参数 θ 可能取值的参数空间, $x_1,\cdots,x_n$ 是样本, 将样本的联合概率函数看成 θ 的函数, 用 $L(\theta;x_1,\cdots,x_n)$ 表示, 简记为 $L(\theta)$.

$$L(\theta)=L(\theta;x_1,\cdots,x_n):=\prod_{i=1}^{n}p(x_i;\theta)$$

称为样本的**似然函数**.

Definition2.2.4 最 (极) 大似然估计值 (resp. 量)

最大似然估计法就是用使得 $L(\theta)$ 达到最大的 $\hat{\theta}$ 去估计 θ.

$$\hat{\theta} \in \left\{\hat{\theta}\colon L\left(\hat{\theta}\right) = \max_{\theta} L(\theta)\right\} \tag{2.2.1}$$

称 $\hat{\theta}$ 为 θ 的**最 (极) 大似然估计值 (MLE,Maximum Likelihood Estimate)**. 相应的统计量 $\theta(x_1, \cdots, x_n)$ 称为 θ 的**最 (极) 大似然估计量**.

Remark2.2.5

由于 $\ln(x)$ 是 x 的增函数, 故若 $\ln L(\theta)$ 关于 $\theta \in \mathbb{R}$ 可微, 则问题转化为

$$\frac{\mathrm{d} \ln L(\theta)}{\mathrm{d}\theta} = 0.$$

否则则需要用最大似然原则(2.2.1)来求.

Theorem2.2.6 最大似然估计的不变性

若果 $\hat{\theta}$ 是 θ 的最大似然估计, 则对任意函数 $g(\theta)$(具有单值反函数), 其最大似然估计为 $g\left(\hat{\theta}\right)$.

2.3 估计 (拟合) 评价

2.3.1 无偏性

Definition2.3.1

设 $\hat{\theta}=\hat{\theta}(x_1,\cdots,x_n)$ 是参数 θ 的估计量, θ 的参数空间为 Θ. 若对任意 $\theta\in\Theta$, 有

$$\mathbb{E}\left(\hat{\theta}\right)=\theta$$

则称 $\hat{\theta}$ 为无偏估计, 否则为有偏估计.

称 $b\left(\hat{\theta}\right)=\mathbb{E}\left(\hat{\theta}-\theta\right)$ 为 $\hat{\theta}$ 的偏差.

Example2.3.2

$\hat{\sigma^2}=s^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ 是 σ^2 的无偏估计量.

Example2.3.3

样本 k 阶原点矩 $A_k=\dfrac{1}{n}\sum\limits_i x_i^k$ 是总体 k 阶矩 $\mu_k=\mathbb{E}\left(X^k\right)$ 的无偏估计.

Definition2.3.4 渐近无偏估计

$$\lim_{n\to+\infty}\mathbb{E}\left(\hat{\theta}_n\right)=\theta$$

Example2.3.5

s_n^2 为 σ^2 的渐近无偏估计.

Propsition2.3.6

性质:

1. 无偏化修正技术: 如样本方差 $s_n^2=\frac{1}{n}\sum(X_i-\bar{X})^2$ 有偏, 修正后 $S^2=\frac{n}{n-1}s_n^2$ 无偏.
2. 非不变性: 若 $\hat{\theta}$ 无偏, $g(\hat{\theta})$ 不一定是 $g(\theta)$ 的无偏估计 (如 S 不是 σ 的无偏估计).

Propsition2.3.7

无偏性的缺点:

1. 可能不合理: 即可能出现负值估计正值的情况 (如 $x_1\sim\mathcal{P}(\lambda)$, 可以验证 $(-2)^{x_1}$ 是 $\mathrm{e}^{-3\lambda}$ 的无偏估计量).
2. 可能不可估: 即参数的无偏估计量可能不存在 (如 Cauchy 分布).

3. 可能不唯一: 若 $\hat{\theta}_1, \hat{\theta}_2$ 都是 θ 的无偏估计量, 那么对任意 $\lambda \in \mathbb{R}$, $\lambda\hat{\theta}_1 + (1-\lambda)\hat{\theta}_2$ 是参数 θ 的无偏估计.

2.3.2 有效性

Definition2.3.8

设 $\hat{\theta}_1, \hat{\theta}_2$ 是 θ 的两个无偏估计, 如果对任意的 $\theta \in \Theta$, 有

$$\operatorname{Var}\left(\hat{\theta}_1\right) \leqslant \operatorname{Var}\left(\hat{\theta}_2\right).$$

且至少有一个 $\theta \in \Theta$ 使得上述不等号严格成立, 则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效.

2.3.3 相合性 (一致性)

Definition2.3.9

设 $\theta \in \Theta$ 为未知参数, $\hat{\theta}_n = \hat{\theta}_n(x_1, \cdots, x_n)$ 是 θ 的一个估计量, n 是样本容量, 若对任何一个 $\varepsilon > 0$, 有

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geqslant \varepsilon\right) = 0$$

则称 $\hat{\theta}_n$ 为参数 θ 的想和估计.

上式可以理解为

$$\hat{\theta}_n \xrightarrow{\mathrm{w}} \theta, \quad n \to \infty.$$

Propsition2.3.10

矩估计, 最大似然估计都具有相合性.

Theorem2.3.11

设 $\hat{\theta}_n = \hat{\theta}_n(x_1, \cdots, x_n)$ 是 θ 的一个估计量, 若

$$\lim_{n\to\infty} \mathbb{E}\left(\hat{\theta}_n\right) = \theta, \quad \lim_{n\to\infty} \operatorname{Var}\left(\hat{\theta}_n\right) = 0.$$

则 $\hat{\theta}_n$ 是 θ 的相合估计

Theorem2.3.12

相合估计具有不变性, 即若 $\hat{\theta}_n$ 是 θ 的相合估计, 函数 $g(x)$ 在 $x = \theta$ 处连续, 则 $g(\hat{\theta}_n)$ 也是 $g(\theta)$ 的相合估计.

Propsition2.3.13

参数的相合估计不唯一.

Propsition2.3.14

矩估计一般具有相合性:

1. 样本均值是总体均值的相合估计.
2. 样本标准差 (resp. 方差) 是总体标准差 (resp. 方差) 的相合估计.
3. 样本变异系数是总体变异系数的相合估计.($\mathrm{CV}=100\%\sigma/\mu$, $\mathrm{SCV}=100\%s/\bar{x}$)

2.3.4 渐近正态性

Definition2.3.15

参数 θ 的相合估计 $\hat{\theta}_n$ 称为**渐近正态**的, 若存在趋于 0 的非负常数序列 $\sigma_n(\theta)$, 使得

$$\frac{\hat{\theta}_n-\theta}{\sigma_n(\theta)} \overset{\mathrm{w}}{\sim} \mathcal{N}(0,1) \iff \mathbb{P}\left\{\frac{\hat{\theta}_n-\theta}{\sigma_n(\theta)} \leqslant x\right\} \xrightarrow{\mathrm{w}} \Phi(x), \quad (n\to\infty).$$

也称 $\hat{\theta}_n$ 服从渐近正态分布 $\mathcal{AN}(\theta,\sigma_n^2(\theta))$, 记作 $\hat{\theta}_n \sim \mathcal{AN}(\theta,\sigma_n^2(\theta))$, $\sigma_n^2(\theta)$ 称为 $\hat{\theta}_n$ 的渐近方差.

Remark2.3.16

上述定义中没有要求 θ 为 $\hat{\theta}_n$ 的均值, 也没有要求 $\sigma_n^2(\theta)$ 为 θ 的方差. 但在渐近分布中起着类似的作用.

对于一个渐近正态估计 $\hat{\theta}_n$, 放样本容量 n 足够大, 可以用 $\mathcal{N}(\theta,\sigma_n^2)$ 作为 $\hat{\theta}_n$ 的近似分布, $\hat{\theta}_n$ 的渐近方差 σ_n^2 的大小标志着渐近正态估计 $\hat{\theta}_n$ 的优劣.

Propsition2.3.17

一些性质

1. 渐近正态估计一定是相合估计.
2. 最大似然估计具有渐近正态性.
3. 矩估计都是相合估计, 且在一般情况下是渐近正态估计.
4. 最大似然估计的渐近方差具有统一形式 $\sigma_n^2(\theta)=[nI(\theta)]^{-1}$, 其中 $I(\theta)$ 为 θ 的 Fisher 信息量 (会在后文中给出定义).

Theorem2.3.18

设总体 X 具有密度函数 $p(x;\theta)$, $\theta\in\Theta$, Θ 为非退化区间, 假定

1. 对任意的 x, 下述偏导数对所有 $\theta \in \Theta$ 都存在.

$$\frac{\partial^k \ln p(x;\theta)}{\partial \theta^k}, \quad k = 1, 2, 3.$$

2. 对任意的 $\theta \in \Theta$, 有

$$\left|\frac{\partial^k p(x;\theta)}{\partial \theta^k}\right| < F_k(x), F_k \in \mathscr{R}(\mathbb{R}), \quad k = 1, 2, 3.$$

3. 对任意的 $\theta \in \Theta$, 有

$$0 < I(\theta) = \int_{-\infty}^{\infty} \left(\frac{\partial \ln p(x;\theta)}{\partial \theta}\right)^2 p(x;\theta)\mathrm{d}x < \infty.$$

若 $x_1, \cdots, x_n$ i. i. d. $\sim X$, 则存在未知参数 θ 的 MLE $\hat{\theta}_n = \hat{\theta}_n(x_1, \cdots, x_n)$, 且 $\hat{\theta}_n$ 具有相合性和渐近正态性:

$$\hat{\theta}_n \dot{\sim} \mathcal{N}\left(\theta, \frac{1}{nI(\theta)}\right).$$

其中 $I(\theta)$ 称为 θ 的 Fisher 信息量.

2.4 均方误差

无偏估计量与参数真值的偏差由有效性来度量, 有偏估计量则由均方误差来度量.

Definition2.4.1 均方误差

设 $\hat{\theta}=\hat{\theta}(x_1,\cdots,x_n)$ 是总体参数 θ 的估计量, 则称

$$\mathrm{MSE}\left(\hat{\theta}\right)=\mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right],\quad \theta\in\Theta.$$

是总体参数 θ 的**均方误差 (Mean Squared Error)**.

均方误差是评价点估计的一般标准, 希望其越小越好.

Propsition2.4.2

$$\mathrm{MSE}\left(\hat{\theta}\right)=\mathrm{Var}\left(\hat{\theta}\right)+\left[\mathbb{E}\left(\hat{\theta}\right)-\theta\right]^2$$

若 $\hat{\theta}$ 是 θ 的无偏估计, 则

$$\mathrm{MSE}\left(\hat{\theta}\right)=\mathrm{Var}\left(\hat{\theta}\right).$$

Propsition2.4.3

在均方误差的汗一下有些有偏估计优于无偏估计.

Definition2.4.4 一致最小均方误差估计

设有一个估计类, 称 $\hat{\theta}^\star$ 是该估计类中的**一致最小均方误差估计**, 如果对该估计类中另外任意一个 θ 的估计 $\widetilde{\theta}$, 都有

$$\mathrm{MSE}\left(\hat{\theta}^\star\right)\leqslant\mathrm{MSE}\left(\widetilde{\theta}\right),\quad \forall\theta\in\Theta.$$

一致最小均方估计要在一个估计类里进行, 否则不存在这样的 $\hat{\theta}^\star$.

2.5 一致最小方差无偏估计

Definition2.5.1 一致最小方差无偏估计

对参数估计问题, 设 $\hat{\theta}$ 是 θ 的一个无偏估计, 若对另外任意一个 θ 的无偏估计 $\widetilde{\theta}$, 在参数空间 Θ 上都有

$$\mathrm{Var}_\theta\left(\hat{\theta}\right) \leqslant \mathrm{Var}_\theta\left(\widetilde{\theta}\right).$$

则称 $\hat{\theta}$ 是 θ 的**一致最小方差无偏估计 (Uniformly Minimum Variance Unbiased Estimation)**, 简记为 **UMVUE**.

如果 UMVUE 存在, 那么它一定是充分统计量的函数.

Theorem2.5.2

设 $x_1, \cdots, x_n$ i. i. d. $\sim X$, $\hat{\theta} = \hat{\theta}(x_1, \cdots, x_n)$ 是 θ 的任一无偏估计, 且

1. $\mathrm{Var}\left(\hat{\theta}\right) < +\infty$.
2. 满足 $\mathbb{E}[\varphi(x_1, \cdots, x_n)] = 0$ 的统计量 φ, 都有

$$\mathrm{Cov}_\theta\left(\hat{\theta}, \varphi\right) = 0, \quad \forall \theta \in \Theta.$$

那么 $\hat{\theta}$ 是 θ 的 UMVUE. 反之亦然.

Example2.5.3

设 $x_1, x_2, \cdots, x_n$ i. i. d. $\sim \mathrm{Exp}(1/\theta)$, 则 $T = x_1 + \cdots + x_n$ 是 θ 的充分统计量, 而 $\bar{x} = T/n$ 是 θ 的无偏估计. 现说明样本均值是参数 θ 的 UMVUE.

设 $\varphi = \varphi(x_1, x_2, \cdots, x_n)$ 是 0 的任一无偏估计, 即 $\mathbb{E}(\varphi) = 0$, 则

$$\int_{\mathbb{R}^n_{\geqslant 0}} \varphi(x_1, \cdots, x_n) \cdot \mathrm{e}^{-(x_1+\cdots+x_n)/\theta} \mathrm{d}\boldsymbol{x} = 0, \quad \mathrm{d}\boldsymbol{x} = \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

两端对 θ 求导得

$$\int_{\mathbb{R}^n_{\geqslant 0}} \frac{n\bar{x}}{\theta^2} \varphi(x_1, \cdots, x_n) \cdot \mathrm{e}^{-(x_1+\cdots+x_n)/\theta} \mathrm{d}\boldsymbol{x} = 0.$$

这说明 $E(\bar{x} \cdot \varphi) = 0$, 从而 $\mathrm{Cov}(\bar{x}, \varphi) = \mathbb{E}(\bar{x} \cdot \varphi) - \mathbb{E}(\bar{x}) \cdot E(\varphi) = 0$, 由定理2.5.2, 它是 θ 的 UMVUE. □

2.6 充分性原则 *

Theorem2.6.1 充分性原则

设 $T(X)$ 是一个充分统计量, 而 $\hat{g}(X)$ 是 $g(\theta)$ 的一个无偏估计, 则存在可以表示为 T 的函数的 $g(\theta)$ 的无偏估计 $h(t(X))$, 使

$$\mathrm{Var}_\theta(h(T(X))) \leqslant \mathrm{Var}_\theta\left(\hat{g}(X)\right).$$

等号当且仅当 $\hat{g}(X)$ 能表示为 $T(X)$ 的函数时才成立

证明.

$$h(T(X)) = \mathbb{E}\left[\hat{g}(X)|\, T(X)\right].$$

□

其说明了求 $g(\theta)$ 的 UMVUE, 只需在基于充分统计量 $T(X)$ 的函数的无偏估计类中进行即可. 若充分统计量和 UMVUE 都存在, 则 UMVUE 一定可以表示为充分统计量的函数

2.7 Fisher 信息量

2.7.1 Fisher 信息量

Definition2.7.1 Fisher 信息量

设总体的概率密度函数 $p(x;\theta)$, $\theta \in \Theta$ 满足:

1. 参数空间 Θ 是直线上的一个开区间.
2. 支撑集 $S = \{x\colon p(x;\theta) > 0\}$ 与 θ 无关.
3. 导数 $\dfrac{\partial p(x;\theta)}{\partial \theta}$ 对一切 $\theta \in \Theta$ 都存在.
4. 对 $p(x;\theta)$, 积分与微分运算可交换, 即

$$\frac{\partial}{\partial \theta}\int_{-\infty}^{\infty} p(x;\theta)\mathrm{d}x = \int_{-\infty}^{\infty}\frac{\partial}{\partial \theta}p(x;\theta)\mathrm{d}x.$$

5. 期望 $\mathbb{E}\left\{\left[\dfrac{\partial}{\partial \theta}\ln p(x;\theta)\right]^2\right\}$ 存在.

那么称

$$I(\theta) = \mathbb{E}\left\{\left[\frac{\partial}{\partial \theta}\ln p(x;\theta)\right]^2\right\} = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2}\ln p(x;\theta)\right]$$

为总体的 **Fisher 信息量**.

2.7.2 C-R 不等式

Theorem2.7.2 Cramer-Rao 不等式

设总体 X 分布满足定义2.7.1的条件, $x_1, \cdots, x_n$ $\text{i.i.d.} \sim X$, $T = T(x_1, \cdots, x_n)$ 为 $g(\theta)$ 的任一个无偏估计, 有

$$g'(\theta) = \frac{\partial g(\theta)}{\partial \theta}.$$

存在, 且对任意的 $\theta \in \Theta$, 对

$$g(\theta) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} T(x_1, \cdots, x_n)\prod_{i=1}^{n} p(x_i;\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

的微商可在积分号下进行, 即

$$\begin{aligned} g'(\theta) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} T(x_1, \cdots, x_n)\frac{\partial}{\partial \theta}\left[\prod_{i=1}^{n} p(x_i;\theta)\right]\mathrm{d}x_1\cdots\mathrm{d}x_n \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} T(x_1, \cdots, x_n)\left[\frac{\partial}{\partial \theta}\ln\prod_{i=1}^{n} p(x_i;\theta)\right]\prod_{i=1}^{n} p(x_i;\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n. \end{aligned}$$

对离散总体, 则将上述积分改为求和号后, 等式依然成立. 则有

$$\mathrm{Var}(T) \geqslant \frac{[g'(\theta)]^2}{nI(\theta)}. \tag{2.7.1}$$

称(2.7.1)为信息不等式, 或 Cramer-Rao 不等式 (简记为 C-R 不等式). 不等式的右端项称为参数函数 $g(\theta)$ 的无偏估计 $T(X)$ 的方差的 C-R 下界 (简记为 $g(\theta)$ 的 C-R 下界).

特别的, 对 θ 的无偏估计 $\hat{\theta}$, 有 $\mathrm{Var}\left(\hat{\theta}\right) \geqslant [nI(\theta)]^{-1}$.

如果等号成立, 即方差达到 C-R 下界, 那么称 $T = T(x_1, \cdots, x_n)$ 为 $g(\theta)$ 的**有效估计**. 有效估计一定是 UMVUE.

Propsition2.7.3

UMVUE 的性质:

1. UMVUE 可能不唯一.
2. UMVUE 可能不存在.
3. UMVUE 不一定是最优的估计, 例如在 MSE 下, 有偏估计可能比无偏更优.

第三章　(参数) 区间估计

设有一个参数分布族 $\mathscr{F}=\{f(x,\theta),\theta\in\Theta\}$, 其中 Θ 是参数空间. $\boldsymbol{X}=(x_1,\cdots,x_n)$ 为取自分布族中某总体 $f(x,\theta)$ 的样本, $g(\theta)$ 为定义在 Θ 上的一个已知函数, 要利用样本 $\boldsymbol{X}$ 对 $g(\theta)$ 的值作出估计, 就是参数估计问题. 参数估计有两类: 点估计和区间估计. 关于点估计的问题已在第二章中讨论过了. 在那里是用样本函数 $\hat{g}(\boldsymbol{X})$ 去估计 $g(\theta)$ 的, 称为点估计. 这种估计的缺点是: 单从 $\hat{g}(\boldsymbol{X})$ 所给出的估计值上, 无法看出它的精度有多大. 当然可以定义某种指标, 如估计的均方误差之类去刻画它的精度, 但也还是间接的. 更直接的方法是指出一个误差限 $d(\boldsymbol{X})$, 把估计写成 $\hat{g}(\boldsymbol{X})\pm d(\boldsymbol{X})$ 的形式. 在应用部门中常见到这种写法. 这实际上就是一种区间估计, 即估计 $g(\theta)$ 的取值在区间 $[\hat{g}(\boldsymbol{X})-d(\boldsymbol{X}),\hat{g}(\boldsymbol{X})+d(\boldsymbol{X})]$ 之内. 将其一般化, 给出区间估计的下列定义.

Definition3.0.1 区间估计

设有一个参数分布族 $\mathscr{F}=\{f(x;\theta),\theta\in\Theta\}$, $g(\theta)$ 是定义在参数空间 Θ 上的一个已知函数, $x_1,\cdots,x_n$ i. i. d. $\sim f(x;\theta)$. 令 $\hat{g}_1(x_1,\cdots,x_n)$ 和 $\hat{g}_2(x_1,\cdots,x_n)$ 为定义在样本空间 $\mathscr{X}$ 上, 取值在 Θ 上的两个统计量, 且

$$\hat{g}_1(x_1,\cdots,x_n)\leqslant\hat{g}_2(x_1,\cdots,x_n)$$

则称随机区间 $[\hat{g}_1(x_1,\cdots,x_n),\hat{g}_2(x_1,\cdots,x_n)]$ 为 $g(\theta)$ 的一个**区间估计 (interval estimation)**.

3.1 区间估计的基本概念

Definition3.1.1

设 θ 是总体的一个参数, 其参数空间为 Θ, $x_1,\cdots,x_n$ 是来自该总体的样本, 对给定的一个 α, $(0<\alpha<1)$, 若有统计量 $\hat{\theta}_L=\hat{\theta}_L(x_1,\cdots,x_n)$ 和 $\hat{\theta}_U=\hat{\theta}_U(x_1,\cdots,x_n)$, 若对于 $\forall\theta\in\Theta$, 有

$$\mathbb{P}_\theta\left(\hat{\theta}_L\leqslant\theta\leqslant\hat{\theta}_U\right)\geqslant 1-\alpha.$$

则称随机区间 $\left[\hat{\theta}_L,\hat{\theta}_U\right]$ 为 θ 的置信水平 (置信度) 为 $1-\alpha$ 的置信区间, 或简称 $\left[\hat{\theta}_L,\hat{\theta}_U\right]$ 为 θ 的 $1-\alpha$ 置信区间.

$\hat{\theta}_L$ 和 $\hat{\theta}_U$ 分别称为 θ 的 (双侧) 置信下限和置信上限.

Remark3.1.2

上述定义中理解为随机区间 $\left[\hat{\theta}_L,\hat{\theta}_U\right]$ 至少以 $1-\alpha$ 的概率包含这参数 θ 的真值, 而不能说参数 θ 以 $1-\alpha$ 的概率落入随机区间 $\left[\hat{\theta}_L,\hat{\theta}_U\right]$.

Definition3.1.3 同等置信区间

沿用定义3.1.1的记号, 若对给定的 α, 对任意的 $\theta\in\Theta$, 有

$$\mathbb{P}_\theta\left(\hat{\theta}_L\leqslant\theta\leqslant\hat{\theta}_U\right)=1-\alpha.$$

称 $\left[\hat{\theta}_L,\hat{\theta}_U\right]$ 为 θ 的 $1-\alpha$ **同等置信区间**.

Definition3.1.4 单侧置信区间

若对给定的 $\alpha,(0<\alpha<1)$ 和任意的 $\theta\in\Theta$, 有

$$\mathbb{P}_\theta\left(\hat{\theta}_L\leqslant\theta\right)\geqslant 1-\alpha.$$
$$(\text{resp. }\mathbb{P}_\theta\left(\hat{\theta}_U\geqslant\theta\right)\geqslant 1-\alpha.)$$

则称 $\hat{\theta}_L$ (resp.$\hat{\theta}_U$) 为 θ 的置信水平为 $1-\alpha$ 的 **(单侧) 置信下 (resp. 上) 限**. 若等号对一切 $\theta\in\Theta$ 都成立, 则称 $\hat{\theta}_L$ (resp.$\hat{\theta}_U$) 为 θ 的 $1-\alpha$ 同等置信下 (resp. 上) 限.

Definition3.1.5 枢轴量

样本函数 $G=G(x_1,\cdots,x_n;\theta)$ 称为**枢轴量**, 若满足:

1. G 包含待估参数 θ, 但除 θ 外 G 不含其他未知参数.
2. G 的分布已知, 但与未知参数 θ 无关.

Remark3.1.6

置信区间不唯一.

3.2 枢轴变量法——单个正态总体参数的置信区间

3.2.1 均值的置信区间——总体方差已知

设 $x_1,\cdots,x_n$ i. i. d. $\sim\mathcal{N}(\mu,\sigma^2)$, σ^2 已知, 现估计参数 μ 的 $1-\alpha$ 置信区间:
选取 μ 的点估计为 $\bar{x}$, 构造

$$G:=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim\mathcal{N}(0,1).$$

那么对给定的置信水平 $1-\alpha$ 有

$$\mathbb{P}\{|G|\leqslant y\}=\mathbb{P}\left\{\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right|\leqslant y\right\}=1-\alpha.$$

若记标准正态分布 α 分位数为 u_α, 则有

$$\mathbb{P}\left\{-u_{1-\frac{\alpha}{2}}\leqslant\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\leqslant u_{1-\frac{\alpha}{2}}\right\}=1-\alpha.$$

化简得

$$\mathbb{P}\left\{\bar{x}-u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\leqslant\mu\leqslant\bar{x}+u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\}=1-\alpha.$$

从而未知参数 μ 的一个 $1-\alpha$ 置信区间为

$$\left[\bar{x}-u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\bar{x}+u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

置信区间通常写成 $\bar{x}\pm u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$, 其置信区间长度为 $2\cdot u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$.

3.2.2 均值的置信区间——总体方差未知

设 $x_1,\cdots,x_n$ i. i. d. $\sim\mathcal{N}(\mu,\sigma^2)$. 用 $\bar{x}$ 对 μ 作为点估计, 并用 s 代替 σ, 故枢轴量选为

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}\sim t(n-1).$$

对与给定置信水平 $1-\alpha$, 有

$$\begin{aligned}&\mathbb{P}\left\{\left|\frac{\bar{x}-\mu}{s/\sqrt{n}}\right|\leqslant t_{1-\frac{\alpha}{2}}(n-1)\right\}=1-\alpha\\\iff&\mathbb{P}\left\{\bar{x}-\frac{s}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\leqslant\mu\leqslant\bar{x}+\frac{s}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\right\}=1-\alpha.\end{aligned}$$

故得到 μ 的 $1-\alpha$ 置信区间为

$$\left[\bar{x}-\frac{s}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1),\bar{x}+\frac{s}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\right]$$

3.2.3 方差的置信区间——总体均值已知

枢轴量:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \sim \chi^2(n).$$

σ^2 的 $1-\alpha$ 置信区间:

$$\left[\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)}, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}\right]$$

3.2.4 方差的置信区间——总体均值未知

枢轴量:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 \sim \chi^2(n-1).$$

σ^2 的 $1-\alpha$ 置信区间:

$$\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$$

3.2.5 方差的置信区间——利用 s^2 误差估计 σ^2

枢轴量:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1).$$

σ^2 的 $1-\alpha$ 置信区间:

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$$

3.2.6 (概率) 参数 p 的置信区间——总体为 $b(1,p)$

利用中心极限定理, 有

$$\bar{x} \sim \mathcal{AN}\left(p, \frac{p(1-p)}{n}\right).$$

枢轴量:

$$u = \frac{\bar{x}-p}{\sqrt{p(1-p)/n}} \dot{\sim} \mathcal{N}(0,1).$$

解得精准的 $1-\alpha$ 置信区间:

$$\left[\frac{1}{1+\frac{\lambda}{n}}\left(\bar{x}+\frac{\lambda}{2n}-\sqrt{\frac{\bar{x}(1-\bar{x})\lambda}{n}+\frac{\lambda^2}{4n^2}}\right),\frac{1}{1+\frac{\lambda}{n}}\left(\bar{x}+\frac{\lambda}{2n}+\sqrt{\frac{\bar{x}(1-\bar{x})\lambda}{n}+\frac{\lambda^2}{4n^2}}\right)\right]$$

其中

$$\lambda := u_{1-\frac{\alpha}{2}}^2.$$

由于在大样本下 $0<\lambda/n<<1$. 因此置信区间近似为

$$\left[\bar{x}-u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}},\bar{x}+u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right]$$

3.3 枢轴变量法——双个正态总体参数的置信区间

3.3.1 两个均值差的置信区间——两个总体方差已知

枢轴量:

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

$1-\alpha$ 置信区间:

$$(\bar{x}-\bar{y})\pm u_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}.$$

3.3.2 两个均值差的置信区间——两个总体方差未知

$\sigma_1^2=\sigma_2^2=\sigma^2$, σ^2 未知

枢轴量:

$$\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{s_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2),\quad s_w^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}.$$

$1-\alpha$ 置信区间:

$$(\bar{x}-\bar{y})\pm t_{1-\frac{\alpha}{2}}(n_1+n_2-2)s_w\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

$\sigma_2^2/\sigma_1^2=c$, c 已知

我们直接给出 $1-\alpha$ 置信区间:

$$(\bar{x}-\bar{y})\pm s_t\cdot t_{1-\frac{\alpha}{2}}(n_1+n_2-2)\sqrt{\frac{cn_1+n_2}{n_1n_2}},\quad s_t^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2/c}{n_1+n_2-2}.$$

σ_1^2,σ_2^2 未知, n_1,n_2 也不是很大

这是著名的 Behrens-Fisher 问题. 这里给出一种近似方法: 令 $s_0^2=s_1^2/s_2^2$. 枢轴量:

$$T=\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{s_0}\dot{\sim}t(l)$$

其中 l 由下式确定, 取最近整数:

$$l=\frac{s_0^4}{\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-2)}}.$$

则得到 $1-\alpha$ 近似置信区间:

$$(\bar{x}-\bar{y})\pm s_0t_{1-\frac{\alpha}{2}}(l).$$

3.3.3 两个方差比的置信区间——两个总体均值未知

枢轴量:

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

$1-\alpha$ 置信区间:

$$\left[\frac{1}{F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1)} \cdot \frac{s_1^2}{s_2^2}, \frac{1}{F_{\frac{\alpha}{2}}(n_1-1, n_2-1)} \cdot \frac{s_1^2}{s_2^2}\right]$$

3.3.4 两个方差比的置信区间——两个总体均值已知

枢轴量:

$$\frac{n_2 \sum_{i=1}^{n_1}(x_i-\mu_1)^2 \Big/ \sigma_1^2}{n_1 \sum_{i=1}^{n_2}(x_i-\mu_2)^2 \Big/ \sigma_2^2} \sim F(n_1, n_2).$$

$1-\alpha$ 置信区间:

$$\left[\frac{1}{F_{1-\frac{\alpha}{2}}(n_1, n_2)} \cdot \frac{n_2 \sum_{i=1}^{n_1}(x_i-\mu_1)^2}{n_1 \sum_{i=1}^{n_2}(x_i-\mu_2)^2}, \frac{1}{F_{\frac{\alpha}{2}}(n_1, n_2)} \cdot \frac{n_2 \sum_{i=1}^{n_1}(x_i-\mu_1)^2}{n_1 \sum_{i=1}^{n_2}(x_i-\mu_2)^2}\right]$$

3.4 单侧置信区间

定义已经在3.1.4给出. 下直接给出一些常见的单个 (resp. 双个) 正态总体下参数的单侧 $1-\alpha$ 置信区间.

3.4.1 单个正态总体

待估参数 μ, σ^2 已知

$$\overline{\mu}=\bar{x}+u_{1-\alpha}\frac{\sigma}{\sqrt{n}},\quad \underline{\mu}=\bar{x}-u_{1-\alpha}\frac{\sigma}{\sqrt{n}}.$$

待估参数 μ, σ^2 未知

$$\overline{\mu}=\bar{x}+t_{1-\alpha}(n-1)\frac{s}{\sqrt{n}},\quad \underline{\mu}=\bar{x}-t_{1-\alpha}(n-1)\frac{s}{\sqrt{n}}.$$

待估参数 σ^2, μ 未知

$$\overline{\sigma^2}=\frac{(n-1)s^2}{\chi^2_{\alpha}(n-1)},\quad \underline{\sigma^2}=\frac{(n-1)s^2}{\chi^2_{1-\alpha}(n-1)}.$$

3.4.2 双个正态总体

待估参数 $\mu_1-\mu_2$, σ_1^2,σ_2^2 已知

$$\overline{\mu_1-\mu_2}=(\bar{x}-\bar{y})+u_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\quad \underline{\mu_1-\mu_2}=(\bar{x}-\bar{y})-u_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}.$$

待估参数 $\mu_1-\mu_2$, $\sigma_1^2=\sigma_2^2=\sigma^2$ 未知

$$\overline{\mu_1-\mu_2}=(\bar{x}-\bar{y})+t_{1-\alpha}(n_1+n_2-2)s_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},$$
$$\underline{\mu_1-\mu_2}=(\bar{x}-\bar{y})-t_{1-\alpha}(n_1+n_2-2)s_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

待估参数 σ_1^2/σ_2^2, μ_1,μ_2 未知

$$\overline{\left(\frac{\sigma_1^2}{\sigma_2^2}\right)}=\frac{s_1^2/s_2^2}{F_{\alpha}(n_1-1,n_2-1)},\quad \underline{\left(\frac{\sigma_1^2}{\sigma_2^2}\right)}=\frac{s_1^2/s_2^2}{F_{1-\alpha}(n_1-1,n_2-1)}.$$

第四章　参数与非参数假设检验

参数估计和假设检验是统计推断的两个主要形式. 关于参数的点估计和区间估计的问题已在第二、三章中讨论. 本章讨论假设检验问题, 假设检验问题大致分为两类:

1. 参数假设检验: 即总体分布已知, 总体分布依赖于未知参数 (或参数向量)θ, 要检验的是关于未知参数的假设. 例如总体 $X \sim \mathcal{N}(a, \sigma^2)$, a 未知, 检验

$$H_0 \colon a = a_0 \leftrightarrow H_1 \colon a \neq a_0 \quad \text{或} \quad H_0 \colon a \leqslant a_0 \leftrightarrow H_1 \colon a > a_0.$$

2. 非参数假设检验: 总体分布形式未知, 此时就需要有一种与总体分布族的具体数学形式无关的统计方法, 称为非参数假设检验. 例如检验一批数据是否来自某个已知的总体, 这就属于这类问题.

4.1 假设检验的若干基本概念

4.1.1 检验问题的提法

为了说明假设检验问题的提法. 给出一个例子

Example4.1.1

某工厂生产的一大批产品, 要卖给商店. 按规定次品率 p 不得超过 0.01, 今在其中抽取 100 件, 经检验有 3 件次品, 问这批产品可否出厂?

关于这个问题, 有两个可能性:

$$甲: 0 < p \leqslant 0.01; \quad 乙: 0.01 < p < 1.$$

要通过这批产品中抽样来决定甲, 乙两种可能性中哪个成立.

这个问题通常以下述方式提出: 引入一个 “假设”

$$H_0: 0 < p \leqslant 0.01,$$

它称为**零假设 (null hypothesis)** 或**原假设**, 有时也简称假设. 另一个可能是

$$H_1: 0.01 < p < 1,$$

称为**对立假设**或**备择假设 (alternative hypothesis)**.

目的是要通过样本决定接受 H_0, 还是拒绝 H_0. 可以形象地把问题写成

$$H_0: 0 < p \leqslant 0.01 \leftrightarrow H_1: 0.01 < p < 1.$$

注意这个提法中将 H_0 放在中心位置, 它是检验的对象.H_0 **和** H_1 **的位置不可颠倒**.

Definition4.1.2

设有参数分布族 $\{f(x,\theta), \theta \in \Theta\}$. $X_1, \cdots, X_n$ 为从上述分布族中抽取的简单随机样本.

在参数假设检验问题中, 感兴趣的是 θ 是否属于 Θ 的某个非空真子集 Θ_0, 则命题 $H_0: \theta \in \Theta_0$ 称为**原假设**或**零假设**, 其确切含义是: 存在一个 $\theta_0 \in \Theta_0$ 使得 X 的分布为 $f(x,\theta_0)$.

记 $\Theta_1 = \Theta \setminus \Theta_0$, 则命题 $H_1: \theta \in \Theta_1$ 称为 H_0 的**对立假设**或**备择假设**, 则假设检验问题表述为

$$H_0: \theta \in \Theta_0 \leftrightarrow H_1: \theta \in \Theta_1. \tag{4.1.1}$$

在(4.1.1)中, 若 Θ_0 或 Θ_1 只包含参数空间 Θ 中的一个点, 则称为**简单假设 (simple hypothesis)**; 否则, 称为**复合假设 (composite hypothesis)**.

4.1.2 否定域, 检验函数和检验统计量

仍通过例子来说明这些概念.

Example4.1.3

设 $X=(X_1,\cdots,X_n)$ 为从正态总体 $\mathcal{N}(a,\sigma^2)$ 中抽取的随机样本, 其中 σ^2 已知. 考虑检验问题

$$H_0\colon a=a_0 \leftrightarrow H_1\colon a\neq a_0.$$

此处 a_0 为给定的常数.

这种检验的一种直观作法是: 先求 a 的一个估计量, 可以知道 $\bar{X}$ 是 a 的一个优良估计 (无偏, 有效, 一致). 若 $|\bar{X}-a_0|$ 较大, 就倾向于否定 H_0; 反之, 若 $|\bar{X}-a_0|$ 较小, 就认为抽样结果与 H_0 相接近, 因而倾向于接受 H_0.

具体的说, 要确定一个数 A, 由 $X_1,\cdots,X_n$ 算出样本均值 $\bar{X}$, 当 $|\bar{X}-a_0|>A$ 时就否定 H_0; 当 $|\bar{X}-a_0|\leqslant A$ 时就接受 H_0. 称

$$D=\{\boldsymbol{X}=(X_1,\cdots,X_n)\colon |\bar{X}-a_0|>A\}$$

为**否定域**或**拒绝域 (reject region)**, 即否定域是由样本空间 $\mathscr{X}$ 中一切使得 $|\bar{X}-a_0|>A$ 的那些样本 $\boldsymbol{X}=(X_1,\cdots,X_n)$ 构成.

有了否定域, 等价于将样本空间 $\mathscr{X}$ 分成不相交的两部分 D 和 $\overline{D}=\mathscr{X}\setminus D$. 一旦有了样本 $\boldsymbol{X}$, 当 $\boldsymbol{X}\in D$ 时, 就否定 H_0; 当 $\boldsymbol{X}\in\overline{D}$ 时, 就接受 H_0. 称 $\overline{D}$ 为**接受域 (acceptance region)**. 因此, 此问题中的检验可视为如下一种法则:

$$T\colon \begin{cases} \text{当 } |\bar{X}-a_0|>A \text{ 时}, & \text{拒绝 } H_0,\\ \text{当 } |\bar{X}-a_0|\leqslant A \text{ 时}, & \text{接受 } H_0. \end{cases}$$

为了便于数学处理, 应如检验函数 $\varphi(\boldsymbol{x})$ 的概念, $\varphi(\boldsymbol{x})$ 是与检验 T 一一对应的, 在本例中

$$\varphi(\boldsymbol{x})=\begin{cases} 1, & \text{当 } |\bar{x}-a_0|>A,\\ 0, & \text{当 } |\bar{x}-a_0|\leqslant A. \end{cases} \tag{4.1.2}$$

Definition4.1.4 检验函数

由(4.1.2)给出的**检验函数** $\varphi(\boldsymbol{x})$ 是定义在**样本空间** $\mathscr{X}$ 上, 取值于 $[0,1]$ 上的函数. 他表示当有了样本 $\boldsymbol{X}$ 后, 否定 H_0 的概率. 若只取 0, 1 两个值, 则称这种检验为**非随机化检验 (non-randomized test)**; 反之称为**随机化检验 (randomized test)**.

4.1.3 两类错误与势函数

统计推断是以样本为依据的, 由于样本的随机性, 不能保证统计推断方法的绝对正确, 而只能以一定的概率去保证这种推断的可靠性. 在假设检验问题中可能出现下列两种情形会犯错误:

1. 零假设 H_0 本来是对的, 由于样本的随机性, 观察值落入否定域 D, 错误地将 H_0 否定了, 称为弃真. 这时犯的错误称为**第一类错误 (type I error)**.

2. 零假设 H_0 本来不对, 由于样本的随机性, 观察值落入接受域 $\overline{D}$, 错误地将 H_0 接受了, 称为取伪. 这时犯的错误称为**第二类错误 (type II error)**.

应当注意, 在每一具体场合, 只会犯两类错误中的一个. 当检验确定后, 犯两类错误的概率也就确定了. 那么自然思考一个问题: 怎么取计算犯两类错误的概率?

Definition4.1.5 势函数

设 $\varphi(\boldsymbol{x})$ 是 $H_0: \theta \in \Theta_0 \leftrightarrow H_1: \theta \in \Theta_1$ 的一个检验函数, 则

$$g_\varphi(\theta) = \mathbb{P}_\theta\{\text{用检验 } \varphi \text{ 否定了 } H_0\} = \mathbb{E}_\theta[\varphi(\boldsymbol{X})], \quad \theta \in \Theta.$$

称为 φ 的**功效函数 (power function)**, 也称为**势函数**或**效函数**.

在样本容量 n 一定时, 若减少犯第一类错误的概率 α, 则犯第二类错误的概率 β 往往会增大. 若要使犯两类错误的概率都减少, 除非增加样本容量.

4.1.4 显著性检验

Definition4.1.6 显著性检验

支队犯第一类错误的概率加以控制, 而不考虑犯第二类错误的检验问题, 称为**显著性检验问题**.

显著性水平指犯第一类错误的概率.

Definition4.1.7 假设检验的三种形式

分为双侧检验和单侧检验, 左侧检验、右侧检验统称单侧检验.

双侧检验 $H_0: \theta = \theta_0 \leftrightarrow H_1: \theta \neq \theta_0$.

左侧检验 $H_0: \theta \geqslant \theta_0 \leftrightarrow H_1: \theta < \theta_0$.

右侧检验 $H_0: \theta \leqslant \theta_0 \leftrightarrow H_1: \theta > \theta_0$.

Propsition4.1.8 单侧检验的拒绝域

总体 $X \sim \mathcal{N}(\mu, \sigma^2)$, 其中 μ 为未知参数, σ^2 已知. $x_1, \cdots, x^n$ 为来自总体的简单样本, 给

定显著性水平 α, 则

$$\text{右侧检验的拒绝域为 } u=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \geqslant u_{1-\alpha},$$

$$\text{左侧检验的拒绝域为 } u=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \leqslant u_{\alpha}.$$

Definition4.1.9 检验的 p 值

在一个假设检验问题中, 利用样本观测值能够作出拒绝原假设的最小显著性水平称为**检验的 p 值**.(检验的 p 值是一个概率; 称为观察到的显著性水平 (observed level of significance); 原假设 H_0 能被拒绝的最小显著性水平.)

4.2 正态总体参数的假设检验

4.2.1 总体均值的检验

单个正态总体均值 μ 的检验:

1. $H_0: \mu = \mu_0 \leftrightarrow H_1: \mu \neq \mu_0$,
2. $H_0: \mu \leqslant \mu_0 \leftrightarrow H_1: \mu > \mu_0$,
3. $H_0: \mu \geqslant \mu_0 \leftrightarrow H_1: \mu < \mu_0$.

参数 σ^2 已知时的 u 检验

检验统计量:$u = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.
检验1的拒绝域: $|u| \geqslant u_{1-\frac{\alpha}{2}}$,
检验2的拒绝域: $u \geqslant u_{1-\alpha}$,
检验3的拒绝域: $u \leqslant -u_{1-\alpha} = u_\alpha$.

参数 σ^2 未知时的 t 检验, 小样本 $(n < 30)$

检验统计量:$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(n-1)$.
检验1的拒绝域: $|t| \geqslant t_{1-\frac{\alpha}{2}}(n-1)$,
检验2的拒绝域: $t \geqslant t_{1-\alpha}(n-1)$,
检验3的拒绝域: $t \leqslant -t_{1-\alpha}(n-1)$.

参数 σ^2 未知时的 t 检验, 大样本

检验统计量:$u = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim \mathcal{N}(0,1)$.
检验1的拒绝域: $|u| \geqslant u_{1-\frac{\alpha}{2}}$,
检验2的拒绝域: $u \geqslant u_{1-\alpha}$,
检验3的拒绝域: $u \leqslant -u_{1-\alpha} = u_\alpha$.

两个正态总体均值 μ_1, μ_2 的检验:

1. $H_0: \mu_1 = \mu_2 \leftrightarrow H_1: \mu_1 \neq \mu_2$,
2. $H_0: \mu_1 \leqslant \mu_2 \leftrightarrow H_1: \mu_1 > \mu_2$,
3. $H_0: \mu_1 \geqslant \mu_2 \leftrightarrow H_1: \mu_1 < \mu_2$.

参数 σ_1^2, σ_2^2 已知时两样本 u 检验

检验统计量: $u = \dfrac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{N}(0,1)$
检验1的拒绝域: $|u| \geqslant u_{1-\frac{\alpha}{2}}$,
检验2的拒绝域: $u \geqslant u_{1-\alpha}$,
检验3的拒绝域: $u \leqslant -u_{1-\alpha} = u_\alpha$.

参数 $\sigma_1^2 = \sigma_2^2$ 未知时的两样本的 t 检验, 小样本

检验统计量:

$$t = \frac{\bar{x} - \bar{y}}{s_w\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2), \quad s_w^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

检验1的拒绝域: $|t| \geqslant t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)$,
检验2的拒绝域: $t \geqslant t_{1-\alpha}(n_1 + n_2 - 2)$,
检验3的拒绝域: $t \leqslant -t_{1-\alpha}(n_1 + n_2 - 2)$.

参数 σ_1^2, σ_2^2 未知, 大样本 u 检验

检验统计量: $u = \dfrac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n_1 + s_y^2/n_2}} \sim \mathcal{N}(0,1)$, n_1, n_2 充分大.
检验1的拒绝域: $|u| \geqslant u_{1-\frac{\alpha}{2}}$,
检验2的拒绝域: $u \geqslant u_{1-\alpha}$,
检验3的拒绝域: $u \leqslant -u_{1-\alpha} = u_\alpha$.

参数 σ_1^2, σ_2^2 未知, 小样本近似 t 检验

检验统计量: $t = \dfrac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n_1 + s_y^2/n_2}} \sim t(l)$
其中 $s_0^2 = \dfrac{s_x^2}{n_1} + \dfrac{s_y^2}{n_2}$, l 为接近 $\dfrac{s_0^4}{s_x^4 n_1^{-2}/(n_1-1) + s_y^4 n_2^{-2}/(n_2-1)}$ 的整数.
检验1的拒绝域: $|t| \geqslant t_{1-\frac{\alpha}{2}}(l)$,
检验2的拒绝域: $t \geqslant t_{1-\alpha}(l)$,
检验3的拒绝域: $t \leqslant -t_{1-\alpha}(l)$.

4.2.2 总体方差的检验

单个正态总体方差 σ^2 的检验

1. $H_0: \sigma^2 = \sigma_0^2 \leftrightarrow H_1: \sigma^2 \neq \sigma_0^2$,
2. $H_0: \sigma^2 \leqslant \sigma_0^2 \leftrightarrow H_1: \sigma^2 > \sigma_0^2$,
3. $H_0: \sigma^2 \geqslant \sigma_0^2 \leftrightarrow H_1: \sigma^2 < \sigma_0^2$.

在总体均值 μ 未知, 总体方差 σ^2 的假设检验: χ^2 检验法.

检验统计量: $\chi^2=\dfrac{(n-1)s^2}{\sigma_0^2}\sim\chi^2(n-1)$.

检验1的拒绝域:$\chi^2\leqslant\chi^2_{\frac{\alpha}{2}}(n-1)$ 或 $\chi^2\geqslant\chi^2_{1-\frac{\alpha}{2}}(n-1)$,

检验2的拒绝域:$\chi^2\geqslant\chi^2_{1-\alpha}(n-1)$,

检验3的拒绝域:$\chi^2\leqslant\chi^2_{\alpha}(n-1)$.

两个正态总体方差的检验

1. $H_0\colon\sigma_1^2=\sigma_2^2\leftrightarrow H_1\colon\sigma_1^2\neq\sigma_2^2$,
2. $H_0\colon\sigma_1^2\leqslant\sigma_2^2\leftrightarrow H_1\colon\sigma_1^2>\sigma_2^2$,
3. $H_0\colon\sigma_1^2\geqslant\sigma_2^2\leftrightarrow H_1\colon\sigma_1^2<\sigma_2^2$.

在总体均值 μ_1,μ_2 未知, 总体方差的假设检验: F 检验法

检验统计量: $F=s_1^2/s_2^2\sim F(n_1-1,n_2-1)$

检验1的拒绝域:$F\leqslant F_{\frac{\alpha}{2}}(n_1-1,n_2-1)$ 或 $F\geqslant F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)$,

检验2的拒绝域:$F\geqslant F_{1-\alpha}(n_1-1,n_2-1)$,

检验3的拒绝域:$F\leqslant F_{\alpha}(n_1-1,n_2-1)$.

4.2.3 成对数据的检验

一般, 设有 n 对相互独立的观察结果:$\{(x_i,y_i)\}_{i=1}^n$, 令 $d_i=x_i-y_i$, 则 $d_1,\cdots,d_n$ 独立同分布, 在正态假定下,

$$d_1,\cdots,d_n\ \text{i.i.d.}\sim\mathcal{N}(\mu,\sigma_d^2)$$

其中 μ,σ_d^2 未知. 那么检验两组数据是否有显著差异转化为方差未知时正态总体的均值假设检验. 考虑下面三个检验问题:

1. $H_0\colon\mu=0\leftrightarrow H_1\colon\mu\neq0$,
2. $H_0\colon\mu\leqslant0\leftrightarrow H_1\colon\mu>0$,
3. $H_0\colon\mu\geqslant0\leftrightarrow H_1\colon\mu<0$.

检验统计量选为:$t=\dfrac{\bar{d}}{s_d/\sqrt{n}}\sim t(n-1)$.

4.2.4 假设检验与置信区间的关系

假设检验问题与置信区间是一一对应的.

参数估计是根据样本统计量估计总体参数的真值.

假设检验是根据样本统计量来检验对总体参数的先验假设是否成立.

4.2.5 关于假设检验拒绝域的说明

我们说下面这两种的拒绝域是一样的.

- $H_0: \mu = \mu_0 \leftrightarrow H_1: \mu > \mu_0$
- $H_0: \mu \leqslant \mu_0 \leftrightarrow H_1: \mu > \mu_0$

类似有其他类型上的等价.

4.3 其他分布参数的假设检验

4.3.1 指数分布参数的假设检验

设 $x_1, \cdots, x_n$ i. i. d. $\sim \mathrm{Exp}(1/\theta)$, 关于 θ 可以考虑如下三个检验问题:

1. $H_0: \theta = \theta_0 \leftrightarrow H_1: \theta \neq \theta_0$,
2. $H_0: \theta \leqslant \theta_0 \leftrightarrow H_1: \theta > \theta_0$,
3. $H_0: \theta \geqslant \theta_0 \leftrightarrow H_1: \theta < \theta_0$.

考虑 θ 的无偏估计 $\bar{x}$, 当 $\theta = \theta_0$ 时, 有

$$n\bar{x} = \sum_{i=1}^{n} x_i \sim \mathrm{Ga}(n, 1/\theta_0).$$

从而选取检验统计量:$\chi^2 = \dfrac{2n\bar{x}}{\theta_0} \sim \chi^2(2n)$.

检验1的拒绝域:$\chi^2 \leqslant \chi^2_{\frac{\alpha}{2}}(2n)$ 或 $\chi^2 \geq \chi^2_{1-\frac{\alpha}{2}}(2n)$.

检验2的拒绝域:$\chi^2 \geqslant \chi^2_{1-\alpha}(2n)$

检验3的拒绝域:$\chi^2 \leqslant \chi^2_{\alpha}(2n)$

4.3.2 比例 p 的检验

比例 p 可看做某事件发生的概率, 即可看做二点分布 $b(1,p)$ 中的参数. 作 n 次独立实验, 以 x 记该事件发生的次数, 则 $x \sim b(n,p)$. 我们可以根据 x 检验 p 的一些假设:

首先是 $H_0: p \leqslant p_0 \leftrightarrow H_1: p > p_0$:

直观上看拒绝域为 $W = \{x \geqslant c\}$, c 为一个非负整数. 对给定的 α, 要找到

$$\mathbb{P}(x \geqslant c; p_0) = \sum_{i=c}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} = \alpha$$

是困难的. 因此比较常见的是找一个 c_0, 使得

$$\sum_{i=c_0+1}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} < \alpha < \sum_{i=c_0}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}.$$

于是可取 $c = c_0 + 1$, 即满足不等式的最小整数, 此时相当于把显著性水平降低了.

然而在离散场合, 用 p 值作检验较为简便, 这时可以不用找 c_0, 而只需要更具观测值 $x = x_0$ 计算检验的 p 值, 即 $p = \mathbb{P}_{p_0}(x \geqslant x_0)$.

另外的情形是类似的

4.3.3 大样本检验

对二点分布 $b(1,p)$, 在大样本下利用中心极限定理有

$$\bar{p} \dot{\sim} \mathcal{N}(p, p(1-p)/n).$$

采用如下检验统计量

$$u = \frac{\sqrt{n}(\bar{x} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} \dot{\sim} \mathcal{N}(0,1), \quad \hat{p} = \hat{p}_{\text{MLE}}.$$

对于更一般的, 若

$$x_1, \cdots, x_n \ \text{i. i. d.} \sim F(x;\theta).$$

又设该总体均值为 θ, 方差为 $\sigma^2(\theta)$ 是 θ 的函数. 检验统计量采用:

$$u = \frac{\sqrt{n}(\bar{x} - \theta_0)}{\sqrt{\sigma^2(\hat{\theta})}} \dot{\sim} \mathcal{N}(0,1), \quad \hat{\theta} = \hat{\theta}_{\text{MLE}}.$$

对于检验问题

1. $H_0 \colon \theta = \theta_0 \leftrightarrow H_1 \colon \theta \neq \theta_0$,
2. $H_0 \colon \theta \leqslant \theta_0 \leftrightarrow H_1 \colon \theta > \theta_0$,
3. $H_0 \colon \theta \geqslant \theta_0 \leftrightarrow H_1 \colon \theta < \theta_0$.

拒绝域依次为

$$W_1 = \{|u| \geqslant u_{1-\frac{\alpha}{2}}\}, \quad W_2 = \{u \geqslant u_{1-\alpha}\}, \quad W_3 = \{u \leqslant u_\alpha\}.$$

关于两总体比例差 $p_1 - p_2$ 的大样本检验问题:

1. $H_0 \colon p_1 - p_2 = \delta \leftrightarrow H_1 \colon p_1 - p_2 \neq \delta$,
2. $H_0 \colon p_1 - p_2 \leqslant \delta \leftrightarrow H_1 \colon p_1 - p_2 > \delta$,
3. $H_0 \colon p_1 - p_2 \geqslant \delta \leftrightarrow H_1 \colon p_1 - p_2 < \delta$.

由中心极限定理有

$$\frac{(\bar{p}_1 - \bar{p}_2) - \delta}{\sqrt{\dfrac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \dfrac{\bar{p}_2(1-\bar{p}_2)}{n_2}}} \dot{\sim} \mathcal{N}(0,1).$$

在 H_0 中, 取 $\delta = 0$, 令

$$\hat{p} = \frac{1}{n_1 + n_2}\left(\sum x_i + \sum y_i\right) = \frac{1}{n_1 + n_2}(n_1\bar{p}_1 + n_2\bar{p}_2).$$

可以选取检验统计量:

$$u=\frac{(\bar{p}_1-\bar{p}_2)}{\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right)\hat{p}(1-\hat{p})}}\sim\mathcal{N}(0,1).$$

Poisson 分布的大样本检验: 设 $x_1,\cdots,x_n$ i. i. d. $\sim\mathcal{P}(\lambda)$, 有 $\mathbb{E}=\mathrm{Var}=\lambda$.

检验问题: $\lambda \leftrightsquigarrow \lambda_0$.

当样本比较大时, 由中心极限定理有:$\bar{x}\dot{\sim}\mathcal{N}(\lambda,\lambda/n)$.

选取检验统计量:

$$u=\frac{\sqrt{n}(\bar{x}-\lambda_0)}{\sqrt{\hat{\lambda}_{\mathrm{MLE}}}}=\frac{\sqrt{n}(\bar{x}-\lambda_0)}{\sqrt{\bar{x}}}=\frac{\sqrt{n}(\bar{x}-\lambda_0)}{\sqrt{\lambda_0}}\dot{\sim}\mathcal{N}(0,1).$$

4.4 似然比检验与分布拟合检验

4.4.1 似然比检验

Definition4.4.1 广义似然比

设 $x_1,\cdots,x_n$ i. i. d. $\sim p(x;\theta)$, 考虑检验问题: $H_0\colon \theta\in\Theta_0 \leftrightarrow H_1\colon \theta\in\Theta_1=\Theta\setminus\Theta_0$. 称统计量

$$\Lambda(x_1,\cdots,x_n)=\frac{\sup\limits_{\theta\in\Theta} p(x_1,\cdots,x_n;\theta)}{\sup\limits_{\theta\in\Theta_0} p(x_1,\cdots,x_n;\theta)} \tag{4.4.1}$$

为该假设的**似然比 (likelihood ratio)**, 也称为**广义似然比**.

(4.4.1)式可以写成如下形式:

$$\Lambda(x_1,\cdots,x_n)=\frac{p(x_1,\cdots,x_n;\hat{\theta})}{(x_1,\cdots,x_n;\hat{\theta}_0)},$$

其中 $\hat{\theta}$ 为 Θ 上 θ 的 MLE, $\hat{\theta}_0$ 为 Θ_0 上 θ 的 MLE.

Example4.4.2

设 $x_1,\cdots,x_n$ i. i. d. $\sim\mathcal{N}(\mu,\sigma^2)$, μ,σ^2 均未知, 试求检验问题 $H_0\colon \mu=\mu_0 \leftrightarrow H_1\colon \mu\neq\mu_0$ 的显著性水平为 α 的似然比检验.

$\theta=(\mu,\sigma^2)$, 从而

$$\Theta=\{(\mu,\sigma)\colon -\infty<\mu<\infty,\sigma^2>0\},\quad \Theta_0=\{(\mu_0,\sigma^2)\colon \sigma^2>0\}.$$

当 μ,σ^2 未知时, 易知在 Θ 上的 MLE 分别为

$$\hat{\mu}=\bar{x},\quad \widehat{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

当 $\mu=\mu_0$ 已知, 在 Θ_0 上 σ_0^2 的 MLE 为

$$\widehat{\sigma_0^2}\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2.$$

所以广义似然比取为

$$\begin{aligned}\Lambda(x_1,\cdots,x_n)=&\frac{\sup\limits_{\theta\in\Theta}p(x_1,\cdots,x_n;\theta)}{\sup\limits_{\theta\in\Theta_0}p(x_1,\cdots,x_n;\theta)}=\frac{p(x_1,\cdots,x_n;\theta)}{p(x_1,\cdots,x_n;\hat{\theta})}\\=&\frac{\left(2\pi\widehat{\sigma^2}\right)^{-\frac{n}{2}}\exp\left\{-\dfrac{1}{2\widehat{\sigma^2}}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2\right\}}{\left(2\pi\widehat{\sigma_0^2}\right)^{-\frac{n}{2}}\exp\left\{-\dfrac{1}{2\widehat{\sigma_0^2}}\sum\limits_{i=1}^{n}(x_i-\mu_0)^2\right\}}=\left(\frac{\widehat{\sigma_0^2}}{\widehat{\sigma^2}}\right)^{\frac{n}{2}}\\=&\left(\frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu_0)^2}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\right)^{\frac{n}{2}}=\left(1+\frac{n(\bar{x}-\mu_0)^2}{(n-1)s^2}\right)^{\frac{n}{2}}.\end{aligned}$$

若令 $t=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}\sim t(n-1)$, 则 $\Lambda(x_1,\cdots,x_n)=\left(1+\dfrac{t^2}{n-1}\right)^{\frac{n}{2}}$. 拒绝域为

$$W=\left\{|t|\geqslant t_{1-\frac{\alpha}{2}}(n-1)\right\}=\left\{\Lambda(x_1,\cdots,x_n)\geqslant\left(1+\frac{d^2}{n-1}\right)^{\frac{n}{2}}\right\},\quad d=t_{1-\frac{\alpha}{2}}(n-1).$$

若令 $F=\dfrac{(\bar{x}-\mu_0)^2}{s^2/n}\sim F(1,n-1)$, 则 $\Lambda(x_1,\cdots,x_n)=\left(1+\dfrac{F}{n-1}\right)^{\frac{n}{2}}$. 拒绝域为

$$W=\{F\geqslant F_{1-\alpha}(1,n-1)\}=\left\{\Lambda(x_1,\cdots,x_n)\geqslant\left(1+\frac{d_1}{n-1}\right)^{\frac{n}{2}}\right\},$$
$$d_1=F_{1-\alpha}(1,n-1).$$

这说明这时的 t 检验和 F 检验是等价的.

4.4.2　χ^2 拟合优度检验

分类数据的 χ^2 检验法

将随机试验可能的结果的全体 Ω 分为 m 个互不相容的事件,

$$A_1,\cdots,A_m\quad\text{s.t.}\left(\sum_{i=1}^{m}A_i=\Omega,A_iA_j=\varnothing,i\neq j,i,j=1,2,\cdots,m\right).$$

可以计算 $p_{i0}=\mathbb{P}(A_i),\quad i=1,2,\cdot,m.$

建立假设 H_0: $p_i=p_{i0},\forall i$, 其中 p_i 是已知量, 即存在一个已知分布 $\mathbb{P}(X=t_i)=p_i,\forall i$.

情形一: 诸 p_{i0} 均已知时, 理论分布完全已知. 此时若 H_0 成立, 则对每一类 A_i, 其频率 N_i/n 与概率 p_{i0} 应较接近. 即观测频数与理论频数 np_{i0} 应相差不大. 据此, 英国统计学家 K.Pearson 提出如下定理:

Theorem4.4.3 K.Pearson

若 n 充分大 ($\geqslant 50$), 则当 H_0 为真时 (不论 H_0 中的分布是什么分布), 检验统计量总是近似地服从自由度为 $m-1$ 的 χ^2 分布

$$\chi^2=\sum_{i=1}^{m}\frac{(N_i-np_{i0})^2}{np_{i0}}\dot{\sim}(m-1).$$

其中 N_i 是 n 个样本中属于 A_i 的样本观测值 (观测频数), np_{i0} 为期望频数 (理论频数).

称检验假设 H_0 的统计量为 Pearson 统计量

$$\chi^2=\sum_{i=1}^{m}\frac{(N_i-np_{i0})^2}{np_{i0}}=\sum_{i=1}^{m}\frac{N_i^2}{np_{i0}}-n.$$

拒绝域为: $W=\{\chi^2\geqslant\chi^2_{1-\alpha}(m-1)\}$.

于是在 H_0 为真的假设下

$$\chi^2=\sum_{i=1}^{m}\frac{(N_i-np_{i0})^2}{np_{i0}}\geqslant\chi^2_{1-\alpha}(m-1),$$

则在显著性水平 α 下拒绝 H_0, 否则就接受.

Example4.4.4

把一颗骰子重复抛掷 60 次, 结果为出现点数 $1,2,3,4,5,6$ 对应的频数为 $7,8,12,11,9,13$. 试检验这颗骰子的六个面是否均匀? (在显著性水平 $\alpha=0.05$ 下)

给出零假设 H_0: $\mathbb{P}\{X_i\}=1/6,\quad i=1,\cdots,6$ (即 H_0: 是均匀的.)

在 H_0 为真的前提下, $p_{i0}=1/6,\forall i.\ m=6,$

$$\chi^2=\sum_{i=1}^{m}\frac{(N_i-np_{i0})^2}{np_{i0}}=\frac{(7-10)^2}{10}+\frac{(8-10)^2}{10}+\frac{(12-10)^2}{10}$$
$$+\frac{(11-10)^2}{10}+\frac{(9-10)^2}{10}+\frac{(13-10)^2}{10}=2.8$$

然而 $\chi^2=2.8\leqslant\chi^2_{1-0.05}(5)=11.0705$, 故接受 H_0, 认为这颗骰子六个面是均匀的.

Example4.4.5

一家工厂分早、中、晚三班, 每班 8 小时, 近期发生了一些事故共 15 次, 记录早班 6 次, 中班 3 次, 晚班 6 次. 据此怀疑事故发生率与班次有关, 比方说, 中班事故率小些 (表面上的差异), 要用这些数据来检验一下.

根据题意需要检验假设 H_0: $\mathbb{P}\{X=i\}=1/3,\quad i=1,2,3.$ (即 H_0: 事故发生率与班次无关.)

$np_i=15/3=5$, $\chi^2=(1^2+2^2+1^2)/5=1.2$, 注意到 $\chi^2=1.2\leqslant\chi^2_{0.95}(2)=5.9915$. 从而接受 H_0, 即认为事故发生率与班次无关.

当总事故达到 90 而仍维持上述比例 6: 3: 6 时, $\chi^2 = 7.2 \geqslant \chi^2_{0.95}(2) = 5.9915$, 这时候认为事故发生率与班次有关. $(p = \mathbb{P}(\chi^2 \geqslant 7.2) = 0.03)$

这表明观察数 $n = 15$ 太小, 随机性的影响就大了. 当观察的总事故达到 90 而仍维持相同的比例, 则 p 值降到 0.05 以下, 因而有较充分的理由认为三个班次有差异. 在 15 这么小的观察数之下, 对米钱这个结果, 只宜解释为: 一方面数据能提供事故率与班次有关的支持, 一方面也认为表面上的差异究竟不宜完全忽视, 值得进一步观察.

情形二: 诸 p_{i0} 不完全已知时, 理论分布已知带有未知参数. 若诸 $p_{i0}, i = 1, \cdots, m$ 由 r $(r < m)$ 个未知参数 $\theta_1, \cdots, \theta_r$ 确定, 即 $p_{i0} = p_{i0}(\theta_1, \cdots, \theta_r), \quad i = 1, \cdots, m.$

首先给出 $\theta_1, \cdots, \theta_r$ 的 MLE$\hat{\theta}_1, \cdots, \hat{\theta}_r$, 然后给出诸 $p_{i0}, \quad i = 1, \cdots, m$ 的 MLE$\hat{p}_{i0} = p_{i0}(\hat{\theta}_1, \cdots, \hat{\theta}_r)$. Fisher 给出了如下定理

Theorem4.4.6 Fisher

在 H_0 为真时, 检验统计量

$$\chi^2 = \sum_{i=1}^{m} \frac{(N_i - n\hat{p}_{i0})^2}{n\hat{p}_{i0}} \dot{\sim} \chi^2(m - r - 1).$$

于是检验拒绝域为 $W = \{\chi^2 \geqslant \chi^2_{1-\alpha}(m - r - 1)\}$.

如上用来假设检验总体分布的检验方法称为 χ^2 拟合优度检验法.

Remark4.4.7

χ^2 拟合优度检验方法使用时, 必须注意 n 要足够大, 且 np_{i0} 不太小这两个条件. 一般要求样本容量 n 不小于 50, 以及各类观察值不小于 5, 每个 np_{i0} 都不小于 5, 而且 np_{i0} 最好都在 10 以上, 否则应当适当的合并区间.

Example4.4.8

教材 P346(文档 P368) 例 7.4.3.

一般分布的 χ^2 检验法 (总体为离散分布或连续分布)

假设检验的问题为 H_0: $F(x) = F(x_0)$, 其中 $F(x_0)$ 为一个已知分布.

仍取 $m - 1$ 个实数, 使得 $-\infty < a_1 < \cdots < a_{m-1} < +\infty$. 这将实数族分为 m 个区间

$$\mathbb{R} = \bigsqcup \{(-\infty, a_1], (a_1, a_2], \cdots, (a_{m-1}, +\infty)\}.$$

当观测值落入第 i 个区间内, 就把它看作属于第 i 类, 因此这 m 个区间就相当于 m 个类. 以 N_i 表示样本的观测值 $x_1, \cdots, x_n$ 落入区间 $(a_{i-1}, a_i]$ 内的个数.

记 $p_{10} = F_0(a_1)$, $p_{m0} = 1 - F_0(a_{m-1})$, $p_{i0} = F_0(a_i) - F_0(a_{i-1}), \quad i = 2, \cdots, m - 1.$

在 H_0 为真时, 观测值落入第 i 个区间的观测频数 N_i 与理论频数 np_{i0} 应该相差不大. 经过上述处理, 此问题转化为分类数据的检验问题. 选择 Pearson 统计量

$$\chi^2=\sum_{i=1}^{m}\frac{(N_i-np_{i0})^2}{np_{i0}}=\sum_{i=1}^{m}\frac{N_i^2}{np_{i0}}-n,$$

拒绝域为 $W=\{\chi^2\geqslant\chi^2_{1-\alpha}(m-1)\}$.

一般分布中含有未知参数的 χ^2 检验法

假设检验为 $H_0\colon F_0(x;\theta_1,\cdots,\theta_r)\leftrightarrow H_1\colon F(x)\neq F_0$, 其中 F_0 的形式已知, 参数 $\theta_1,\cdots,\theta_r$ 未知.

设 $(X_1,\cdots,X_n)$ $\text{i.i.d.}\sim X$, $(x_1,\cdots,x_n)$ 为其观测值, 首先用 MLE 得到参数的估计. 由此得到 $F_0(x;\hat\theta_1,\cdots,\hat\theta_r)$.

记 $\hat p_{10}=F_0(a_1;\hat\theta_1,\cdots,\hat\theta_r)$, $\hat p_{m0}=1-F(a_{m-1};\hat\theta_1,\cdots,\hat\theta_r)$, $\hat p_{i0}=F(a_i;\hat\theta_1,\cdots,\hat\theta_r)-F(a_{i-1};\hat\theta_1,\cdots,\hat\theta_r)$, $\quad i=2,\cdots,m-1$.

由此可以看到, 此问题又可以转化为分类数据的假设检验问题. 检验统计量为

$$\chi^2=\sum_{i=1}^{m}\frac{(N_i-n\hat p_{i0})^2}{n\hat p_{i0}}=\sum_{i=1}^{m}\frac{N_i^2}{n\hat p_{i0}}-n,$$

拒绝域为 $W=\{\chi^2\geqslant\chi^2_{1-\alpha}(m-r-1)\}$.

4.4.3 列联表独立性检验

提出假设 $H_0\colon p_{ij}=p_{i\cdot}p_{\cdot j}$. 检验统计量为

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-n\hat p_{ij})^2}{n\hat p_{ij}}=n\sum_i\sum_j\frac{(n_{ij}-n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}},$$

且在 H_0 真时, 有 $\chi^2\dot\sim\chi^2((r-1)(c-1))$. 其中 $\hat p_{i\cdot}=n_{i\cdot}/n$, $\hat p_{\cdot j}=n_{\cdot j}/n$. 拒绝域为 $W=\{\chi^2\geqslant\chi^2_{1-\alpha}((r-1)(c-1)\}$.

特别的, 对二乘二列联表, 拒绝域为

$$W=\left\{\chi^2=\frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot1}n_{\cdot2}}\geqslant\chi^2_{1-\alpha}(1)\right\}.$$

4.5 正态性检验

4.6 秩和检验

第五章　方差分析与回归分析

5.1　方差分析

Definition5.1.1 因子

影响试验指标的条件, 用 A, B, C 表示.

Definition5.1.2 水平

因子所处的不同状态, 因子 A 的不同水平用 $A_1, A_2, \cdots$ 表示.

Example5.1.3

在饲料养鸡增肥的研究中, 把饲料成为因子, 记作 A. 把三种不同的配方称为因子 A 的三个水平, 记作 A_1, A_2, A_3.

5.1.1　单因子试验的方差分析

Definition5.1.4

只考虑一个因子的试验称为单因子试验.

通常, 在单因子试验中, 记因子为 A, 设其有 r 个水平, 记为 $A_1, \cdots, A_r$, 在每一个水平下考察的指标可以看成一个总体, 现在有 r 个水平, 故有 r 个总体, 假定:

- 每一个总体均为正态总体, 记为 $\mathcal{N}(\mu_i, \sigma_i^2), \quad i = 1, \cdots, r.$
- 各总体的方差相同, 记为 $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 = \sigma^2.$
- 从每一个总体中抽取的样本都是互相独立的, 即所有的实验结果 y_{ij} 都相互独立.

对如下假设进行检验:

$$H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_r \leftrightarrow H_1 \colon \exists i, j \,\mathrm{s.t.}\, \mu_i \neq \mu_j.$$

若 H_0 成立, 称因子 A 不显著, 反之称其显著.

为了简单起见, 先假设各个水平下试验的重复数相同. 设从第 i 个水平下的总体获得 m 个实验结果, 用 y_{ij} 表示第 i 个总体第 j 次重复试验结果, 共得如下 $r \times m$ 个试验结果:

$$y_{ij}, \quad i = 1, \cdots, r, \quad j = 1, \cdots, m,$$

其中 r 为水平数, m 为重复数, i 为水平编号, j 为重复序号.

水平 A_i 下的实验结果 y_{ij} 与该水平下的指标均值 μ_i 总是有差距的, 记 $\varepsilon_{ij}=y_{ij}-\mu_i$, 称 ε_{ij} 为随机误差, 于是有

$$y_{ij}=\mu_i+\varepsilon_{ij}.$$

上式称为实验结果 y_{ij} 的数据结构式. 把三个假定用于数据结构式就可以有单因子方差分析的统计模型:

$$\begin{cases} y_{ij}=\mu_i+\varepsilon_{ij}, \quad i=1,\cdots,r, \quad j=1,\cdots,m, \\ \text{诸 } \varepsilon_{ij} \text{ 互相独立, 且都服从 } \mathcal{N}(0,\sigma^2). \end{cases}$$

Definition5.1.5

诸 μ_i 的平均

$$\mu=\frac{1}{r}\sum_{i=1}^{r}\mu_i$$

称为**总均值**, 或称**一般均值**. 第 i 个水平下的均值 μ_i 与总均值 μ 的差

$$a_i=\mu_i-\mu, \quad i=1,2,\cdots,r$$

称为因子 A 的第 i 个水平的**主效应**, 简称为 A_i 的**水平效应**.

容易看出

$$\sum_{i=1}^{r}a_i=0, \quad \mu_i=\mu+a_i.$$

这表明第 i 个总体的均值时由总均值与该水平的效应叠加而成的从而模型可以改写为

$$\begin{cases} y_{ij}=\mu+a_i+\varepsilon_{ij}, \quad i=1,\cdots,r, \quad j=1,\cdots,m, \\ \sum\limits_{i=1}^{r}a_i=0, \\ \text{诸 } \varepsilon_{ij} \text{ 互相独立, 且都服从 } \mathcal{N}(0,\sigma^2). \end{cases}$$

从而零假设改写为 $H_0\colon a_1,\cdots,a_r=0 \leftrightarrow H_1\colon \exists i \,\text{s.t.}\, a_i\neq 0$.

5.1.2 平方和分解

$$T_i=\sum_{j=1}^{m}y_{ij}, \quad \bar{y}_{i\cdot}=\frac{T_i}{m}, \quad i=1,\cdots,r,$$

$$T=\sum_{i=1}^{r}T_i, \quad \bar{y}=\frac{T}{r\cdot m}=\frac{T}{n}, \quad n=rm=\text{总试验次数}.$$

数据之间也是有差异的, 数据 y_{ij} 与总均值 $\bar{y}$ 间的偏差可用 $y_{ij}-\bar{y}$ 表示, 它可分解为两个偏差之和

$$y_{ij}-\bar{y}=(y_{ij}-\bar{y}_{i\cdot})+(\bar{y}_{i\cdot}-\bar{y}).$$

记

$$\bar{\varepsilon}_{i\cdot}=\frac{1}{m}\sum_{j=1}^{m}\varepsilon_{ij},\quad \bar{\varepsilon}=\frac{1}{r}\sum_{i=1}^{r}\bar{\varepsilon}_{i\cdot}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}\varepsilon_{ij}.$$

Definition5.1.6

称 $y_{ij}-\bar{y}_{i\cdot}=\varepsilon_{ij}-\bar{\varepsilon}_{i\cdot}$ 为组内偏差;

称 $\bar{y}_{i\cdot}-\bar{y}=a_i+\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}$ 为组间偏差;

称 $Q=\sum\limits_{i=1}^{k}(y_i-\bar{y})^2$ 为 k 个数据的偏差平方和. 自由度 $f_Q=k-1$;

Definition5.1.7

总偏差平方和为 $S_T=\sum\sum(y_{ij}-\bar{y})^2,\quad f_T=n-1=mc-1$;

组内偏差平方和, 也成误差偏差平方和为

$$S_e=\sum\sum(y_{ij}-\bar{y}_{i\cdot})^2,\quad f_e=n-r;$$

组间偏差平方和, 也称因子 A 的偏差平方和为

$$S_A=\sum\sum(\bar{y}_{i\cdot}-\bar{y})^2=m\sum_{i=1}^{r}(\bar{y}_{i\cdot}-\bar{y})^2,\quad f_A=r-1.$$

Theorem5.1.8 总平方和分解公式

$S_T=S_A+S_e,\quad f_T=f_A+f_e.$

5.1.3 检验方法

Definition5.1.9 均方

定义为 $MS=Q/f_Q$, 其意为平均每个自由度上有多少平方和, 它较好的度量了一组数据的离散程度.

组内方差: $MS_e=S_e/f_e$;

组间方差: $MS_A=S_A/f_A$.

用 $F=\dfrac{MS_A}{MS_e}=\dfrac{S_A/f_A}{S_e/f_e}\sim=F(r-1,n-r)$ 作为检验 H_0: $a_i=0,\forall i$ 的统计量.

Theorem5.1.10

在单因子方差分析模型及前述符号下, 则有

- $S_e/\sigma^2 \sim \chi^2(n-r)$, 从而 $\mathbb{E}(S_e) = (n-r)\sigma^2$.
- $\mathbb{E}(S_A) = (r-1)\sigma^2 + m\sum_{i=1}^{r} a_i^2$, 进一步, 若 H_0 成立, 则有 $S_A/\sigma^2 \simeq \chi^2(r-1)$.
- S_A 与 S_e 独立.

5.1.4 参数估计