扫地才子的学习笔记

——数理统计

扫地才子

2023 年 3 月 15 日

前言

数理统计是我认为很麻烦的一门课.

数不清的公式, 数不清的数据, 乱七八糟的定义.

但却很有用.

本文参考陈家鼎数理统计学讲义茆诗松概率论与数理统计

为初等统计知识, 没有涉及到测度理论

有些真的是很粗略的直观理解, 还待更进一步的深入理解. 但这些粗略的理解是很有必要的, 因为我上课的时候, 老师说的每个字我都知道是什么意思, 但就是理解不了这是在干什么, 为什么要这么干

这也是学数学的通病吧, 一开始就站在比较高的高度取看待一个问题, 反而会让人摸不着头脑

至此结束我的复习笔记的撰写.

附上孙七七的主页.home.ustc.edu.cn/~tysun/

但是孙七七的笔记里并没有数理统计

高

2023 年 3 月 15 日

目录

第一章　总体与样本

1.1　总体与样本的概念

我擅长以例子来理解数学概念，所以样本与总体，我同样以例子作为起始。

假设我要统计一个地区的平均身高。我们不妨设这个地区的人的身高满足正态分布。即 $N(\mu, \sigma^2)$

事实上，我们一般情况下对于参数 μ, σ 是不知道的，这正是我们统计意义。

在这个例子中，总体为这个地区（为了方便我记该地区为 A）所有人的身高，是“所有人的身高”，而并非“所有人”，我想这很容易理解。

在这个例子中，我们可以考虑将 A 地区的所有人的身高都测量一遍，但那无疑是愚蠢的，费时费力。

我们常用的手法是抽取一定的样本。为了方便我设 A 地区共有 100 人。

其中我抽取 10 个个体的身高，$X_1, X_2, \cdots, X_{10}$，构成一个随机向量 $X = (X_1, X_2, \cdots, X_{10})$。该随机向量我们称之为样本。

事实上，我们有时候也会将 X_i 称为样本，注意不要引起歧义，在本文中，我一般称之为“个体”。

这里将是第一个难点。

X_i 是随机变量！

X_i 是随机变量！

X_i 是随机变量！

重要的话说三遍，我知道这很难理解，但事实就是如此。

样本具有随机性与确定性。

- 随机性是指，我们抽取样本时，是随机抽取的。
- 确定性是指，我们抽取的样本是会以数据的形式呈现。

也就是说，样本是随机变量，又是数据。在此，我认为理解样本时，暂且不要将其理解为数据。

样本 $X_1, X_2, \cdots, X_{10}$ 代表的仅仅是随机抽选的 10 个个体的身高，而并非是“165cm，170cm，$\cdots$，155cm”这些具体的数值。只是代表“身高”这个概念，而并不是具体的数值。

我们将这些具体的数值以另一种形式去呈现,即样本的观测值 $x_1, x_2, \cdots, x_{10}$。

如果你对分析中的集合论部分足够了解，应该很容易构造一个二者的映射。

将样本与观测值区分开是有好处的，尽管很多时候，他们可以合二为一。

比如上述例子中的身高，就可以合二为一，但我还是倡导分开理解。

再比如测量一批产品的次品率，这时候 X_i 表示“好产品”或“坏产品”，而对应的观测值 x_i 可以为 1 或 0。1 代表好产品，2 代表坏产品也就是次品。

也就是说，X_i 是随机的，而其观测值 x_i 是确定的。

也就是对同一个 X_i 进行多次观测，得到的 x_i 是不一样的。这话其实说的并不合理，因为 X_i 未必是同一个，而是随机抽取的，得到的 x_i 自然是不同的。

说了这么多，就是为了说明，样本是随机变量，为什么要着重强调这一点，因为既然是随机变量，就会满足一定的分布，在后面我们会详细的说明这一点。

1.2　抽样

抽样是十分重要的，不当的抽样会让我们对整体信息的把控出现偏差，有时候这些偏差是致命的。

这让我想起了盲人摸象的故事。

第一个盲人摸了大象的身体，认为这是一堵墙。

第二个盲人摸了大象的耳朵，认为这是一把扇子。

这是因为他们的抽样次数太少了，只掌握了部分信息，才会出现如此大的错误。

在此，不仅仅是抽样次数，如何抽样也是一个问题。

试想水油分层的小实验，如果我事先并不知道这是水油的混合物，我在上层抽取的很多样本，于是我得到结论，这杯液体是油。这显然也是荒谬的。

一般我们抽取的叫做简单随机样本，其具有以下特点。

- 样本要简单，换言之个体之间相互独立，这会大大简化我们对样本的处理。
- 样本要和总体具有同样的随机性，换言之，同分布。

1.3　统计量

统计量：完全由样本所决定的量，不由参数决定。当然这其中有一些我们比较常用的统计量

1.3.1　样本均值

设 $X_1, X_2, \cdots, X_n$ 为样本，则样本均值为：

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

我们知道 X_i 是随机变量，因此，$\bar{X}$ 也是随机变量，尽管一般情况下，它的观测值只有一个。

我们通常在简单随机样本的情况下，用样本均值估计总体分布的均值，也就是期望。

1.3.2　样本方差

设 $X_1, X_2, \cdots, X_n$ 为样本，则样本方差为：

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

我们通常在简单随机样本的情况下，用样本方差估计总体分布的方差。

同样的, 样本方差也是个随机变量, 尽管通常情况下表现为一个确定的数.

样本方差与方差的关系

$Var(\bar{x}) = \frac{\sigma^2}{n}$

证明:

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

$E(S^2) = \sigma^2$

证明:

$$
\begin{aligned}
E(S^2) &= E(\frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)) \\
&= \frac{1}{n-1}(\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2))
\end{aligned}
$$

根据方差的性质我们有以下式子

$$
\begin{aligned}
\sigma^2 &= E(X^2) - \mu^2 \\
E(X^2) &= \sigma^2 + \mu^2 \\
E(\bar{X}^2) = Var(\bar{X}) + [E(\bar{X})]^2 &= \frac{\sigma^2}{n} + \mu^2
\end{aligned}
$$

代入得到:

$$
\begin{aligned}
E(S^2) &= \frac{1}{n-1}(\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2)) \\
&= \frac{1}{n-1}(\sum_{i=1}^{n} (\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)) \\
&= \sigma^2
\end{aligned}
$$

对此, 我们有大数定律:

$$
\lim_{n\to\infty} P(|\bar{X} - \mu| < \epsilon) = 1
$$

假设我们取 n 个样本, 计算出每个样本的样本方差 $S_1^2, S_2^2, \cdots, S_n^2$, 求出算数平均值:

$$\bar{S^2} = \frac{S_1^1 + S_2^2 + \cdots + S_n^2}{n}$$

代入得到:

$$\lim_{n\to\infty} P(|\bar{S^2} - \sigma^2| < \epsilon) = 1$$

我们可以拿这个平均值逼近总体的方差.

自由度

这里 $n-1$ 称为自由度.

自由度有两种理解方式:

- 一共有 n 个数值 $X_1, X_2, \cdots, X_n$ 应该有 n 个自由度 (因为每个样本可以自由变化, 不受其他样本的影响牵连), 但已经有一个自由度用于估计 $\bar{X}$, 所以还剩 $n-1$ 个自由度. 也就是说, 因为 $\bar{X}$ 的存在, 我们只需要知道 $n-1$ 个 X_i, 就可以知道所有的 X_i
- 若以 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 代入 $\sum_{i=1}^{n}(X_i - \bar{X})^2$ 中, 而将其整理成二次型 $\sum_{i,j=1}^{n} a_{ij}X_iX_j (a_{ij} = a_{ji})$, 不难验证: 方阵 $A = (a_{ij})$ 的秩为 $n-1$, 自由度就定义为这个秩.

当然, 为什么样本方差要除以这个自由度, 在之后将予以阐述.

1.3.3　次序统计量

在此引入次序统计量, 何为次序统计量?

设 $X_1, X_2, \cdots, X_n$ 为样本.

之后得到相应的观测值 $x_1, x_2, \cdots, x_n$

将观测值从小到大排列 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$

再找出对应的样本 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$

称 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 为次序统计量, 有时候我们也称 $X_{(i)}$ 为次序统计量

次序统计量也是随机变量, 因为样本是随机变量.

这个其实理解起来可能有些困难, 因为在次序统计量中涉及到了观测值, 这可能让人对样本造成误解, 认为样本就是数据.

但其实可以这么想, 样本是随机抽取的, 每次抽取样本, 对其进行观测, 都会有最小值, 第二小值,....., 最大值.

每次抽取样本, 观测的最小值都是不一样的, 但会呈现一定的分布, 也就是会围绕均值出现.

这就代表这个最小值, 第二小值,..... 最大值都是随机变量.

也就是次序统计量是随机变量.

1.3.4　由次序统计量衍生出的统计量

中位数 (第 p 百分位数)

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

中位数也被称为第 50 百分位数.

极值

样本的最大值和最小值, 即 X_n 和 X_1

极差

最大值与最小值之差, 即 $X_n - X_1$

经验分布函数

为什么引入经验分布函数, 因为有些时候我们可以通过样本分布来估计总体分布.

而经验分布函数是依概率收敛于总体分布

定义: 设 $X_1, X_2, \cdots, X_n$ 为样本. 对应的次序统计量为 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$, 当给定次序统计量的观测值 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ 时, 对任意实数 x 时, 对任意实数 x, 称下列函数

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ \frac{k}{n} & x_{(k)} \geq x \geq x_{(k+1)} \\ 1 & x_{(n)} \geq x \end{cases}$$

当然, 其实这并不是特别重要, 因为很多时候我们并不需要估计总体的分布, 而是估计总体分布的某些参数.

Clivenkotheorem

设总体 X 的分布函数为 $F(x)$, 经验分布函数为 $F_n(x)$, 对任意实数 x, 记

$$D_n = sup\{|F_n(x) - F(x)|, -\infty < x < +\infty\}$$

则 $P\{\lim_{n\to\infty} D_n = 0\} = 1$

当然, 上述的所有前提, 都是样本足够大.

第二章　抽样分布

我们已经知道了, 在数量足够大的情况下, 总体的均值和方差是可以用样本均值和样本方差替代的.

然而很多时候, 数量 n 都不会足够的大, 这时候我们用样本均值和方差预测是十分危险的, 运气不好的话会和总体的均值和方差有很大的偏差.

但所幸的是, 样本均值和样本方差都是随机变量, 既然是随机变量, 就会满足一定的分布.

2.1　初论样本均值和方差的分布

设 $X_1, X_2, \cdots, X_n$ 为取自正态总体 $N(\mu, \Sigma^2)$ 的简单随机样本. 对应的样本均值为 $\bar{X}$, 则样本均值的随机分布为:

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

或

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

证明:

由正态分布性质得到:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

又根据正态分布的性质:$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

上述是在总体是正态分布的情况下得到的.

但事实上, 如果总体不是正态分布, 只要 n 足够的大, 那么样本均值同样符合正态分布

设 $X_1, X_2, \cdots, X_n$ 为取自总体的简单随机样本. 对应的样本均值为 $\bar{X}$, 其中 $E(X)=\mu, Var(X)=\sigma^2$ 当 n 足够大时, 则样本均值的随机分布为:

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

证明:

假设

$$Y = \frac{X_1 + X_2 + \cdots + X_n}{\sigma\sqrt{n}}$$

根据中心极限定理:

$$\lim_{n\to\infty} = F_r(y) = \lim_{n\to\infty} P(Y \geq y) = \Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{t^2}{2}} dt$$

也就是说, 在 n 足够大的前提条件下

$Y \sim N(0,1) \quad \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

我们看似已经知道了样本均值的分布, 但实际上这是没有意义的, 因为 $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, 要想知道样本均值的分布, 我们还是要知道总体的均值和方差.

但正是因为我们得不到样本均值的均值和方差, 我们才不得不去寻求样本的分布.

但我们并非一无所获, 因为很多时候我们是可以知道总体方差的, 所以我们可以去用样本均值来估计总体均值, 并且由于我们得到了样本均值的分布, 这个估计会更加准确.

但并不是每次都会那么幸运的知道总体分布, 这就又陷入了一个困境.

幸运的是这个问题同样是可以被解决的.

2.2　三大分布

为了解决上述问题, 我们要引入三大分布.

2.2.1　卡方分布

随机变量 $X_1, X_2, \cdots, X_n$ 相互独立, 都服从 $N(0,1)$ 则随机变量:

$$X^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

的概率密度函数为:

$P(x) = \frac{1}{\Gamma(\frac{n}{2})2^{\frac{2}{n}}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \quad (x > 0)$

其中伽马函数为:$\Gamma(x) = \int_0^\infty \frac{t^{x-1}}{e^t} dt$

称 X^2 服从自由度为 n 的卡方分布, 记作:$X^2 \sim X^2(n)$. 其期望和方差分别为:

$$E(X^2) = n, \quad Var(X^2) = 2n$$

证明:

$(1)E(X^2) = n$

$$E(X_i^2) = D(X_i) + EX_i^2 = 1 + 0 = 1$$
$$E(X^2) = E(X_1^2 + X_2^2 + \cdots + X_n^2) = n$$

$(2)Var(X^2) = 2n$

$$Var(X_i^2) = E(X_i^4) - E(X_i^2)^2 = 3 - 1 = 2$$
$$Var(X^2) = Var(\sum_{i=1}^{n} X_i^2) = 2n$$

2.2.2　t 分布

设 $X \sim N(0,1), Y \sim X^2(n)$

X 和 Y 相互独立, 则随机变量:

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

的概率密度函数为:

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

称 T 服从自由度为 n 的 t 分布, 记作:$T \sim t(n)$, 其期望和方差为:

$$E(T) = 0 \quad (n > 1), Var(T) = \frac{n}{n-2} \quad (n > 2)$$

特别的当 $n = 1$ 时, 这个分布被称为柯西分布, 是个极其特殊的分布, 没有 k 阶距.

证明:

2.2.3　F 分布

设 $X \sim X^2(n) \quad Y \sim X^2(m)$

X 和 Y 相互独立, 则随机变量:

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

的概率密度函数为:

$$P(x) = \frac{\Gamma(\frac{m+n}{2})(\frac{m}{n})^{\frac{m}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} x^{\frac{m}{2}-1}(1+\frac{m}{n}x)^{-\frac{m+n}{2}}$$

称 F 服从自由度为 (n,m) 的 F 分布, 记作:

$$F \sim F(n,m)$$

, 期望和方差为:

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

证明:

2.3　再论样本均值与方差的分布

与上次不同的是, 这次我们将先从样本方差入手, 并且我们更加关注的是正态总体

2.3.1　样本方差的分布

设 $X_1, X_2, \cdots, X_n$ 为取自正态总体 $N(\mu, \Sigma^2)$ 的简单随机样本. 对应的样本均值为 $\bar{X}$, 则样本方差为:

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

我们有 $\frac{(n-1)S^2}{\sigma^2} \sim X^2(n-1)$ 并且 $\bar{X}$ 与 S^2 独立

证明:

这无疑是个很好的结果, 我们可以用这个结果来限定 $S^2 : \sigma^2$, 将之限制在自由度为 n 的卡方分布当中了.

这样就可以确定我们的样本方差与总体方差的逼近程度了.

再来看样本均值

2.3.2　样本均值的分布

当我们知道总体方差的时候, 我们可以用 $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

当我们不知道总体方差的时候呢? 我们依旧有方法来描述样本均值与总体均值的逼近程度

设 $X_1, X_2, \cdots, X_n$ 为取自正态总体 $N(\mu, \Sigma^2)$ 的简单随机样本, 则有:

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

证明:

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

然后:

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

根据 t 分布的定义我们有:

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

根据这个定理, 我们就不需要去知道总体方差 σ^2, 只需要算出样本方差 S^2 就可以知道 $\bar{X} - \mu$ 被限制在自由度为 $n-1$ 的 t 分布范围内了.

2.3.3　关于 F 分布

最后, 为什么要引入 F 分布

分别取自两个正态总体 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的两个简单随机样本:

$$X_1, X_2, \cdots, X_n \quad Y_1, Y_2, \cdots, Y_m$$

这两个样本相互独立, 其样本均值为:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \bar{Y} = \frac{1}{m}\sum_{i=1}^{n} Y_i$$

样本方差为:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2$$

则有:

$\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} = \frac{S_X^2/S_Y^2}{\sigma_1^2/\sigma_2^2} \sim F(n-1, m-1)$

证明:

根据之前的方差分布:

$$\frac{(n-1)S_X^2}{\sigma_1^2} \sim X^2(n-1), \frac{(m-1)S_Y^2}{\sigma_2^2} \sim X^2(m-1)$$

根据条件可知 S_X^2 和 S_Y^2 独立, 所以根据 F 分布的定义有:

$$\frac{(n-1)S_X^2}{(n-1)\sigma_1^2} / \frac{(m-1)S_Y^2}{(m-1)\sigma_2^2} \sim F(n-1, m-1)$$

整理之后有:

$\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} = \frac{S_X^2/S_Y^2}{\sigma_1^2/\sigma_2^2} \sim F(n-1, m-1)$

第三章　参数估计

很多时候我们知道总体的分布类型, 但其中含有未知的参数，我们需要估计这些参数，因此这就需要估计一下。

下面来看两种估计方式

研究某地小麦亩产量问题，已知小麦亩产量 $X \sim N(\mu, \sigma^2)$,μ, σ^2 未知, 假设有 1000 亩地, 分别记录每亩地的小麦产量, 若随机抽查 50 亩地的产量数据, 问小麦的亩产量是多少?

回答 1: 小麦每亩产量 600 斤

回答 2: 小麦每亩产量以 95% 的概率在 550-650 之间.

回答 1 就是我们所谓的" 点估计", 而回答 2 就是我们所谓的" 区间估计"

本章主要介绍这两类估计

3.1　点估计

定义: 设 $X_1, X_2, \cdots, X_n$ 为取自某总体的样本, 若构造某统计量 $\hat{\theta}(x_1, x_2, \cdots, x_n)$ 作为总体分布中未知参数 θ 的近似, 可以称此统计量为估计量, 并且称此估计量 $\hat{\theta}$ 为 θ 的点估计.

点估计的方法有很多, 但在此只介绍两种主要方法: 矩估计和极大似然估计 (MLE)(也叫最大似然估计法, 网上有观点认为这两种方法有区别, 在本文中对两种叫法不作区分).

3.1.1 矩估计

$X \sim F(X;\theta), \theta = (\theta_1, \cdots, \theta_k)$ 为未知参数向量, 若 X 的 m 阶原点矩,$\alpha_m = E(X^m) = \int_{-\infty}^{+\infty} x^m dF(X;\theta)$ 存在, $m = 1, 2, \cdots, k$, 且 α_m 是 $\theta = (\theta_1, \cdots, \theta_k)$ 的函数, 记作 $\alpha_m(\theta_1, \cdots, \theta_k)$

$$\begin{cases} \alpha_1(\theta_1, \cdots, \theta_k) = A_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \alpha_2(\theta_1, \cdots, \theta_k) = A_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ \cdots \\ \alpha_m(\theta_1, \cdots, \theta_k) = A_m = \frac{1}{n}\sum_{i=1}^{n} X_i^m \end{cases}$$

解上述方程组, 解出 $\hat{\theta}_i = \hat{\theta}_i(X_1, X_2, \cdots, X_n), i = 1, 2, \cdots, k$

$\hat{\theta}_i$ 就是 θ_i 的矩估计量;

例题 1: 设总体 X 的密度函数 $f(x) = \begin{cases} \alpha x^{\alpha-1}, & 0 < x < 1 \\ 0, & \text{其他} \end{cases}$, 其中 $\alpha > 0$ 为未知参数,$X_1, X_2, \cdots, X_n$ 是样本, 求 α 的矩估计.

解答:

总体的一阶原点矩为:

$E(X) = \int_{-\infty}^{+\infty} x f(x) dx = \int_0^1 x\alpha x^{\alpha-1} dx = \frac{\alpha}{\alpha+1} = \bar{X}$

求出 α 即可

例题 2: 设总体 X 的均值 $E(X) = \mu$, 方差 $DX = \sigma^2$,$X_1, \cdots, X_n$ 为取自总体的样本, 求 μ 和 σ^2 的矩估计.

我们知道二阶原点矩 $EX^2 = DX + (EX)^2 = \sigma^2 + \mu^2$

令

$$\begin{cases} \mu = \bar{X} \\ \sigma^2 + \mu^2 = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^2 \end{cases}$$

解方程组得到:

$$\begin{cases} \mu = \bar{X} \\ \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \end{cases}$$

看起来流程还是很简单的, 但学数学的人不懂原理的话用起来就会很不安心.

当然, 这个原理也很直观, 我们遇到 n 阶原点矩, 很自然的就会想到辛钦大数定律.

设 $X_1, X_2, \cdots, X_n$ 取自某总体的简单随机样本, 设随机变量的 k 阶距与样本的 k 阶距分别为: $E(X^k) = \mu_k, A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k (k = 1, 1, \cdots)$

则有

$$\lim_{n\to\infty} P(|A_k - \mu_k| < \epsilon) = 1$$

也就是说 A_k 依概率收敛于 μ_k

3.1.2 矩估计习题

1. 设 $X_1, X_2, \cdots, X_n$ 是来自下列分布的样本:

$$f(x;\theta) = \begin{cases} \frac{\Gamma(\theta+1)}{\Gamma(\theta)\Gamma(1)} x^{\theta-1}, & 0 \geq x \geq 1 \\ 0, & otherwise \end{cases}$$

的总体的样本, 其中 $\theta \in (0, +\infty)$ 试用矩法估计 θ

分析: 思路还是很简单的, 样本给了, 分布给了, 一个参数, 不出意外求总体一阶矩就可以得到估计了.

解答:

$E(X) = \int_0^1 x \cdot \theta x^{\theta-1} dx = \frac{\theta}{\theta+1}$ 这是总体一阶矩

$A(X) = \bar{X}$ 这是样本一阶矩

二者相等得出估计结果

$\hat{\theta} = \frac{\bar{X}}{1-\bar{X}}$

2. 设 $X_1, X_2, \cdots, X_n$ 是来自分布密度为

$f(x;c,\theta)=\frac{1}{2\theta}I_{[c-\theta,c+\theta]}(x)$

的总体分布的样本,$-\infty<c<+\infty,\theta>0$, 试用矩法估计 c 和 θ

分析: 比上一题难一点, 当然, 也就一点, 两个参数, 不出意外需要求导二阶矩, 有一点需要提示的是,$I_{[c-\theta,c+\theta]}(x)$ 为特征函数, 也就是说在指定区间 $[c-\theta,c+\theta]$ 取 1, 其他区间为 0.

解答:

$$E(X)=\int_{c-\theta}^{c+\theta}x\cdot\frac{1}{2\theta}dx=c$$

$A(X)=\bar{X}$

$E(X^2)=\int_{c-\theta}^{c+\theta}x^2\cdot\frac{1}{2\theta}dx=c^2+\frac{1}{3}\theta^2$

$A(K^2)=c^2+s^2$

对应相等, 解得估计值:

$$\begin{cases}\hat{c}=\bar{X}\\ \hat{\theta}=\sqrt{\frac{3}{n}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2}\end{cases}$$

3.1.3　极大似然估计法

有一个不透明的盒子, 里面有黑色与白色两种类型的小球, 随机的抽取一个, 是黑球.

这个时候会认为黑球要多于白球.

当然, 一次试验的说服力显然不够, 如果再抽取一个, 依旧是黑球.

再抽取一个, 依旧是黑球.

这个时候, 随着试验次数的增加, 我们会越来越坚信, 黑球要多于白球.

这就是最大似然思想.

设总体 X 为离散型, 其分布律为 $p(X=x)=p(x;\theta)$,θ 为未知参数,Θ 为 θ 的取值范围

设 $(x_1,x_2,\cdots,x_n)$ 的概率为

$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \prod_{i=1}^{n} P(X = x_i) = \prod_{i=1}^{n} P(x_i; \theta)$

定义上式为似然函数, 记作 $L(\theta) = L = (x_1, x_2, \cdots, x_n; \theta)$

既然在一次试验中, 样本 $(x_1, x_2, \cdots, x_n)$ 出现, 说明试验条件 θ 有利于该结果的出现, 参数 θ 应选取使该概率最大的参数值, 称为最大似然估计法.

最大似然估计法就是在参数 θ 的可能的取值范围内, 找出使似然函数(样本结果出现的概率)$L(\theta)$ 最大的参数值 $\hat{\theta}$, $\hat{\theta}$ 即为 θ 的最大似然估计值, 相应的统计量 $\hat{\theta}(X_1, X_2, \cdots, X_n)$ 称为最大似然估计量.

也就是说, 求参数的最大似然估计, 就是求似然函数的最大值点的问题.

当然, 因为很多时候 $L(\theta)$ 的单调性并不好研究, 但是由于 $L(\theta)$ 与 $lnL(\theta)$ 的单调性是一致的.

于是我们更多时候用 $lnL(\theta)$

当然, 连续型的时候类似, 将在例题中体现.

例题 1: 已知 $X \sim B(1, p), (x_1, x_2, \cdots, x_n)$ 为一个样本, 设求 p 的最大似然估计量

X 的分布律为 $P(X = x) = p^x(1-p)^{1-x}, x = 0, 1$

构造似然函数 $L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

取对数 $lnL(p) = (\sum_{i=1}^{n})lnp + (n - \sum_{i=1}^{n} x_i)ln(1-p)$

然后求导,$\frac{d}{d\theta}lnL(p) = 0$ 得到最大值点

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

最后得到的最大似然估计量为 $\hat{p} = \bar{x}$

例题 2: 设总体 $X \sim N(\mu, \sigma^2), \mu, \sigma^2$ 均未知,$(x_1, x_2, \cdots, x_n)$ 为 X 的样本值, 求 μ, σ^2 的最大似然估计

X 的概率密度为:$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

构造似然函数:

$$L(\mu, sigma^2) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

取对数以后变成:

$lnL(\mu,\sigma^2) = -\frac{n}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$

分别对 μ,σ 求偏导令偏导等于 0, 得到

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

例题 3: 设已知总体 X 在 $[a,b]$ 上服从均匀分布,a,b 未知,$(x_1,x_2,\cdots,x_n)$ 为 X 的一组样本. 如何求 a,b 的最大似然估计量

X 的概率密度函数为 $f(x) = \begin{cases} \frac{1}{b-a}, & a \geq x \geq b \\ 0 & otherwise \end{cases}$

构造似然函数 $L(a,b) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{b-a} = (\frac{1}{b-a})^n$

取对数, 求偏导

但是这道题的偏导为零的点确是不存在的

$$\begin{cases} \frac{\partial}{\partial a} lnL(a,b) = \frac{n}{b-a} = 0 \\ \frac{\partial}{\partial b} lnL(a,b) = -\frac{n}{b-a} = 0 \end{cases}$$

在分析学中, 我们知道偏导不存在未必会没有最值, 偏导存在也未必会有最值

而我们的目标是要找 $(\frac{1}{b-a})^n$ 的最大值

也就是找 $b-a$ 的最小值, 但是 $b-a$ 却又不能尽可能的小, 因为有如下式子成立

$$\begin{cases} a \geq min\{x_i\} \\ b \leq max\{x_i\} \end{cases}$$

为了让 $b-a$ 最小, 只需要

$$\begin{cases} a = min\{x_i\} \\ b = max\{x_i\} \end{cases}$$

3.1.4　最大似然估计法习题

1. 设 X 服从几何分布

$$P(X=k)=p(1-p)^{k-1}, k=1,2,\cdots$$

$X_1, X_2, \cdots, X_n$ 是 X 的简单随机样本, 试找出 p 的最大似然估计

分析: 没什么可分析的, 按步骤算就好, 此题的流程是标准流程

解答:

构造似然函数 $L(p)=\prod_{i=1}^{n} p(1-p)^{k_i}=(\frac{p}{1-p})^n(1-p)^{\sum_{i=1}^{n} k_i}$

取对数得到 $lnL(p)=n[lnp+(\bar{p}-1)ln(1-p)]$

对其求导 $\frac{d}{dp}lnL(p)=n[\frac{1}{p}-(\bar{k}-1)\frac{1}{1-p}]=0$

得到极值点 $\hat{p}=\frac{1}{\bar{k}}$ 其中 $\bar{k}=\sum_{i=1}^{n} k_i$

2. 设 X 的分布密度是

$$f(x)=\frac{1}{2\sigma}e^{-\frac{1}{\sigma}|x|} \quad (\sigma>0)$$

$X_1, X_2, \cdots, X_n$ 是 X 的简单随机样本, 试求 σ 的最大似然估计

分析: 标准流程

构造似然函数 $L(p)=\prod_{i=1}^{n} \frac{1}{2\sigma}e^{-\frac{1}{\sigma}|x_i|}=(\frac{1}{2\sigma})^n e^{\frac{1}{\sigma}\sum_{i=1}^{n}|x_i|}$

取对数得到 $lnL(p)=-nln2\sigma-\frac{1}{\sigma}\sum_{i=1}^{n}|x_i|$

对其求导 $\frac{d}{dp}lnL(p)=-(\frac{n}{\sigma}-\frac{1}{\sigma^2}\sum_{i=1}^{n}|x_i|)=0$

得到极值点 $\hat{\sigma}=\frac{\sum_{i=1}^{n}|x_i|}{n}$

3. 设 $X_1, X_2, \cdots, X_n$ 是来自下列两参数指数分布的样本:

$$f(x;\theta_1,\theta_2)=\begin{cases}\frac{1}{\theta_2}e^{-\frac{1}{\theta_2}(x-\theta_1)}, & x\le\theta_1\\ 0, & x<\theta_1\end{cases}$$

其中 $\theta_1\in(-\infty,+\infty), \theta_2\in(0,+\infty)$, 试求出 θ_1 和 θ_2 的最大似然估计

分析: 依旧是标准流程, 难一点, 真就一点

解答:

构造似然函数 $L(p)=\prod_{i=1}^{n}\frac{1}{\theta_2}e^{-\frac{1}{\theta_2}(x_i-\theta_1)}=(\frac{1}{\theta})^n e^{-\frac{1}{\theta_2}(\sum_{i=1}^{n}-n\theta_1)}$

上述这个式子的前提是 $x_1,x_2,\cdots,x_n\geq\theta_1$ 也就是说 $\theta_1\leq min\{x_1,x_2,\cdots,x_n\}$

我们只关注这一部分, 因为其他部分的概率为 0, 没有研究的必要

取对数 $lnL(\theta_1,\theta_2)=-nln\theta_2-\frac{\sum_{i=1}^{n}x_i}{\theta_2}+n\frac{\theta_1}{\theta_2}$

分别求偏导

$\frac{\partial}{\partial\theta_1}lnL(\theta_1,\theta_2)=\frac{n}{\theta_2}>0$

$\frac{\partial}{\partial\theta_2}lnL(\theta_1,\theta_2)=-\frac{n}{\theta_2}+\frac{\sum_{i=1}^{n}x_i}{\theta_2^2}=0$

解得 $\hat{\theta_2}=\bar{X}$

在此情况下, 要想使对数似然函数取得最大, 只需要 θ_1 取最大, 而幸运的是, 这个恰好有范围, 只需要取 $\hat{\theta}=min\{x_1,x_2,\cdots,x_n\}$

4. 设随机变量 X 以均等机会按 $N\sim(0,1)$ 分布取值和按 $N\sim(\mu,\sigma)$ 分布取值 (μ,σ^2 未知,$-\infty<\mu<+\infty,\sigma^2>0$) 这时 X 的分布密度为这两个分布的密度的平均, 即

$f(x;\mu,\sigma^2)=\frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}+\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}r^{-\frac{(x-\mu)^2}{2\sigma^2}}$

设 $X_1,X_2,\cdots,X_n$ 为此混合分布的简单随机样本, 试证明 μ 和 σ^2 不存在最大似然估计.

分析: 这题还算比较难, 这是两个分布的混合分布, 单独拿出每一个正态分布,μ 和 σ^2 都存在最大似然估计, 但题目要证明的是混合分布不存在, 这就说明, 一个正态分布必然会受另一个分布的影响, 导致不存在最大似然估计. 其实我们可以简化这道题, 题目取 n 个样本, 这样构造似然函数会涉及到连乘, 会给我们的求解造成很大的困扰. 不妨取 $n=1$ 这样分析起来会比较容易, 但证明的时候还需要一般化的证明, 我们只需要证明似然函数无解即可. 但这个似然函数为二元函数找这个无界点还是很困难的.

求解:

构造似然函数:

$$L(\mu,\sigma^2)=(\frac{1}{2})^n\frac{1}{\sqrt{2\pi}}e^{\sum\limits_{i=1}^{n}-\frac{x_i^2}{2}}\prod_{i=1}^{n}[1+exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}+\frac{x_i^2}{2}-ln\sigma^2\}]$$
$$=(\frac{1}{2})^n\frac{1}{\sqrt{2\pi}}e^{\sum\limits_{i=1}^{n}-\frac{x_i^2}{2}}[1+exp\{-\frac{(x_1-\mu)^2}{2\sigma^2}+\frac{x_1^2}{2}-ln\sigma^2\}]$$
$$\cdot[1+exp\{-\frac{(x_2-\mu)^2}{2\sigma^2}+\frac{x_2^2}{2}-ln\sigma^2\}]\cdot\ldots\cdot[1+exp\{-\frac{(x_n-\mu)^2}{2\sigma^2}+\frac{x_n^2}{2}-ln\sigma^2\}]$$

为了方便, 我们考察其中一项

$f_i(\mu,\sigma^2)=1+exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}+\frac{x_i^2}{2}-ln\sigma^2\}$

$\lim_{(\mu,\sigma^2)\to(x_i,0)}f(\mu,\sigma^2)=+\infty$

不失一般性, 令 $(\mu,\sigma^2)\to(x_1,0)$ 此时 $lim_{(\mu,\sigma^2)\to(x_1,0)}f(\mu,\sigma^2)=+\infty$

而对于 $i=2,3,\cdots,n$

则有 $f_i(\mu,\sigma^2)=1+exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}+\frac{x_i^2}{2}-ln\sigma^2\}>1$

则 $\lim_{(\mu,\sigma^2)\to(x_1,0)}L(\mu,\sigma^2)=+\infty$

无界

故无最大似然估计.

3.2　点估计的优良性

我们知道, 点估计可以有很多个, 但如何判断哪个点估计是最优的, 可以从以下三方面入手.

3.2.1　相合性

相合性也被称为一致性. 它的直观上的理解就是, 随着样本量 n 的增加, 估计量会越来越接近真值.

这个特性实际上是很有必要的, 因为在实际中, 我们能控制的只有样本量, 我们希望样本量越大, 我们的估计越接近真值.

相合性用数学语言表达为:

如果对 $\forall \epsilon > 0, \hat{\theta_n}$ 依概率收敛到参数 θ

$$\lim_{N \to +\infty} P(|\theta_n - \theta| > \epsilon) = 0$$

则称 $\hat{\theta}_n$ 是 θ 的相合估计量或一致估计量.

特别的, 如果将依概率收敛改为几乎处处收敛, 那么则称为强相合估计.

但其实, 要想证明相合性是一件十分困难的事情, 在这里我们不给予证明的给出两种方法

1.$\hat{\theta_n}$ 是 θ 的无偏估计, 如果 $\lim_{n \to \infty} D(\hat{\theta}) = 0$, 则 $\hat{\theta_n}$ 必为 θ 的相互估计.(无偏性的概念将在后面阐述)

2. 如果 $\hat{\theta_n}$ 是 θ 的相合估计,$g(\theta)$ 连续, 则 $g(\hat{\theta_n})$ 为 $g(\theta)$ 的相合估计.

3.2.2　无偏性

我们希望我们的点估计不会偏离真值太远, 而是在真值周围浮动. 这就需要这个点估计量具有无偏性, 即:

$$E(\hat{\theta}) = \theta$$

在这里的时候也许会有疑问, 其实我们并不知道真值是多少, 那么要如何去计算 $E(\hat{\theta}) = \theta$

其实并不需要我们知道真值是多少, 因为每一个样本都是与总体同分布的, 我们可以根据期望的性质来判断上式是否成立 $E(\hat{\theta}) = \theta$

如果用一个比喻的话, 无偏性就是一把带有准星的枪, 无论怎样都可以打到靶心附近

这也就是为什么样本方差的分母是 $n-1$ 了.

3.2.3　有效性

对于两个估计量, 如何判断其更有效.

这里引入均方误差的概念

如果有两估计量 $\hat{\theta_1}$ 和 $\hat{\theta_2}$, 前者的均方误差更小, 则称前者更为有效.

我们很希望找到找到这样一个估计值, 它同时满足相合性, 无偏性, 以及有着较小的均方误差.

但事实上总是事与愿违, 找到这样的一个统计量并不容易.

3.2.4　完全充分统计量

之前在统计量的章节说过一些常用的统计量, 那么为什么我们要用比如样本均值, 样本方差这些统计量来描述样本呢?

事实上, 样本均值是充分统计量.

什么是充分统计量, 粗略的讲, 充分统计量就是能充分描述总体的量, 也就是说, 有了充分统计量, 丢掉样本也没什么大关系.

在知乎上找到了一个有趣的例子.

假设你辛辛苦苦收集的 500 个数据全都写在了一张纸上，这些数据是给你写论文用的，非常重要。突然有一天你的狗把你这张写满数据的纸吃掉了，这个时候假如你的数据满足正态分布，且你已经提前把这些数据的均值和方差记录在另外一张纸上了，那你的狗也没坏了什么大事——因为这两个充分统计量包含了这 500 个数据的所有有用信息。

当然这只是粗略的理解.

在我们做估计的时候, 比如正态分布我们要估计均值和方差, 我们会用样本均值和样本方差来估计总体的均值和方差.

当然, 估计参数的信息是包含在样本当中, 但样本当中确实有很多无用的信息.

充分统计量的意义在于, 如果给定了样本统计量的值, 那么分布就与参

数无关了

还是在知乎上找到的一个理解, 通俗却不失准确

得到一个统计量，那么我们就是想用统计量来对未知参数进行分析，对未知参数估计的数值总是来自于统计量的一个取值，那么，在给定统计量的一个取值之后，我们就用那个数值当做参数的值去刻画总体的分布了，但是在统计量取定那个值以后，再次去观察样本的联合分布时，也就是最能够反应所得到样本信息的分布时，如果发现联合分布里面还有未知参数的话，那说明现在手中这个参数估计值没啥用，还不能把样本联合分布刻画出来，所以，当给定了参数估计值后，样本的联合分布里面没有未知参数，这个估计值才算是足够充分的刻画了样本。

现在再来看这个没头没脑的定理, 我想会简单很多.

定理 3.2.1. 因子分解定理 $T(X_1, X_2, \cdots, X_n)$ 是 θ 的充分统计量的充要条件是:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = g(T(x_1, x_2, \cdots, x_n);\theta)h(x_1, x_2, \cdots, x_n)$$

其中 $f(x;\theta)$ 是总体的密度函数,g 是 $T(X_1, X_2, \cdots, X_n)$ 的函数依赖于 $\theta, h(x_1, x_2. \cdots, x_n) \geq 0$ 不依赖于 θ

事实上, 充分统计量的值是由参数所决定的, 也就是说, 给出充分统计量的值, 参数也就随之确定了, 所以充分统计量描述了参数的信息.

注 3.2.1.　1. 充分统计量不是唯一的

2. 维数越少的充分统计量越有价值

3. 并非所有分布都存在降维的充分统计量

4. 若 T 为 θ 的充分统计量, 函数 $u(T)$ 具有单值反函数, 则 $u(T)$ 也是 θ 的充分统计量, 也就是说, 一一映射之后还是充分统计量

在点估计的优良性章节, 我们其实首先要保证无偏性, 再去保证有效性.

也就是说, 我们想找到无偏估计中最有效的那个, 被称为最小方差无偏估计 (MVUE)

定理 3.2.2. 设 $T(X_1, X_2, \cdots, X_n)$ 是 θ 的充分统计量,$S(X_1, X_2, \cdots, X_n)$ 是 $g(\theta)$ 的无偏估计量, 且对 $\forall\theta \in \Theta, D(S) < \infty$, 令 $T^* = E[S|T]$, 则对 $\forall\theta \in \Theta$ 有

$E(T^*) = g(\theta), D(T^*) \leq D(S)$

且除 $P(S = T^*) = 1$ 外, 上述不等式中严格不等号成立.

定理写的太难看了, 其实的意思就是, 要找最小方差无偏估计, 只需要构造充分统计量, 之后构造充分统计量的一一映射, 使其无偏即可.

定义 3.2.3. 完全统计量 称统计量 T 为完全统计量, 如果对任意定义在 T 的值域上的实函数 h, 只要 $E(h(T)) = 0$, 就有 $h(T) = 0$ 几乎处处成立

定理 3.2.4. *Cramer-Rao* 不等式 设总体 X 为连续型随机变量, 密度函数为 $f(x;\theta), \theta$ 为未知参数,$\theta \in \Theta, (X_1, X_2, \cdots, X_n$ 为来自该总体的一个样本, $T(X_1, X_2, \cdots, X_n)$ 为 $g(\theta)$ 的无偏估计量. 如果满足下列正则条件:

1. $I(\theta) = E[\frac{\partial ln f(X;\theta)}{\partial\theta}]^2 > 0$, 其中 $E = \{x : f(x;\theta) \neq 0\}$ 与 θ 无关
2. $\frac{f(x;\theta)}{\partial\theta}$ 存在, 并且 (积分与求导可交换次序) $\int_{-\infty}^{+\infty} \frac{\partial f(X;\theta)}{\partial\theta}dx = \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f(x;\theta)dx =$ $0 \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} \frac{\partial}{\partial\theta}L(x_1, x_2, \cdots, x_n;\theta)dx_1\cdots dx_n = \frac{\partial}{\partial\theta}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} L(x_1, x_2, \cdots, x_n;\theta)dx_1\cdots$ 0
3. $g'(\theta)$ 存在, 且 $g'(\theta) = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} T(x_1, x_2, \cdots, x_n)\frac{\partial}{\partial\theta}L(x_1, x_2, \cdots, x_n;\theta)dx_1\cdots dx_n$

则有 $D(T) \geq \frac{|g'(\theta)|^2}{nI(\theta)}$, 特别当 $g(\theta) = \theta$ 时,$D(T) \geq \frac{1}{nI(\theta)}$

3.3　区间估计的概念

点估计实际上只给出了真实值大概是多少, 但事实上, 我们并不知道点估计的值距离真实值到底有多远, 因此, 区间估计应运而生.

以样本均值为例, 样本不同得到的样本均值 $\bar{X}$ 也就不一样, 样本均值 $\bar{X}$ 到真实值 μ 的距离也不相同, 那么怎么才能保证取的 $\bar{X}$ 周围的区间能涵盖真值呢?

当然, 取一个特别大的区间当然是可以的, 但那就没有价值了.

于是, 我们要求区间宽度要要有足够的覆盖率, 这个覆盖率一般都会取 95%, 也就是说, 计算出 100 次样本均值, 平均下来要有 95 次可以覆盖到真值, 并且在保证覆盖率的情况下, 区间宽度要尽可能的短.

这样构造的区间就是 μ 的置信区间, 覆盖率为 95% 的置信区间也可以称为置信水平为 95% 的置信区间.

如果总体 $X \sim N(\mu, \sigma^2)$, 样本 $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

也就是说 $\bar{X}$ 的变动是一个以真值为中心的正态分布, 因此可以取以中间 95% 对应的区间作为置信区间.

为了方便还是稍微做一下变换 $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

那么求这个区间就变成了

$P(-z_{0.975} < Z < z_{0.975}) = 95\%$

这里的分位点采用陈家鼎数理统计学讲义中的定义方式.

$[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{0.975}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{0.975}]$

定义 3.3.1. *置信区间 设总体 X 的分布函数含有一个未知参数 $\theta \in \Theta$(Θ 为 θ 可能的取值范围), 对于给定值 $\alpha, 0 < \alpha < 1$, 若由来自 X 的样本 $X_1, X_2, \cdots, X_n$ 确定的两个统计量 $\underline{\theta} = \underline{\theta}(X_1, X_2, \cdots, X_n)$ 和 $\overline{\theta} = \overline{\theta}(X_1, X_2, \cdots, X_n)$, 对于任意 $\theta \in \Theta$ 满足:*

$$P(\underline{\theta} \leq \theta \leq \overline{\theta}) \geq 1 - \alpha$$

则称随机区间 $(\underline{\theta}, \overline{\theta})$ 是 θ 的置信水平为 $1-\alpha$ 的置信区间,$\underline{\theta}, \overline{\theta}$ 分别称为置信下限和置信上限,$1-\alpha$ 称为置信水平

3.4　一元正态总体下的区间估计

3.4.1　均值的置信区间

如果 σ 已知

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

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

$P(\underline{\mu} < Z < \overline{\mu}) = 1-\alpha$

如果 σ 未知

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

$P(\underline{\mu} < T < \overline{\mu}) = 1-\alpha$

3.4.2　方差的置信区间

μ 已知的情况

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

μ 未知的情况下, 有:

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

3.5　二元正态总体下的区间估计

$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$

当 σ_1, σ_2 已知:

$\mu_1 - \mu_2$ 的置信区间取枢轴量 $Z = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

因为 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, 所以 $\bar{X}-\bar{Y} \sim N(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})$

当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知:

$\mu_1 - \mu_2$ 的置信区间取枢轴量 $T = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

因为:

$\bar{X}-\bar{Y}\sim N(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})$
$\frac{(n_1-1)S_X^2}{\sigma^2}\sim\chi^2(n_1-1),\frac{(n_2-1)S_Y^2}{\sigma^2}\sim\chi^2(n_2-1)$
所以 $\frac{1}{\sigma^2}[(n_1-1)S_x+(n_2-1)S_y]\sim\chi^2(n_1+n_2-2)$ 由 t 分布的定义

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2}}/\sqrt{\frac{(n_1-1)S_X^2+(n_2-1)S_Y^2}{(n_1+n_2-2)\sigma^2}}\sim t(n_1+n_2-2)$$

$\frac{\sigma_1^2}{\sigma_2^2}$ 的置信区间选取枢轴量 $F=\frac{S_Y^2/\sigma_2^2}{S_X^2/\sigma_1^2}\sim F(n_2-1,n_1-1)$
$\frac{(n_1-1)S_X^2}{\sigma^2}\sim\chi^2(n_1-1),\frac{(n_2-1)S_Y^2}{\sigma^2}\sim\chi^2(n_2-1)$ 所以 $\frac{(n_2-1)S_Y^2}{(n_2-1)\sigma^2}/\frac{(n_1-1)S_X^2}{(n_1-1)\sigma_1^2}\sim F(n_2-1,n_1-1)$

第四章　假设检验

4.1　假设检验的概念

从 Fisher 的女士品茶故事开始, 一位女士认为自己可以分别出先放茶叶再放奶 (茶奶) 和先放奶再放茶叶 (奶茶) 的口感区别 Fisher 对此做出假设

H_0 : 不具备正确分辨的能力, H_1 : 具备正确分辨的能力

H_0 习惯性称为零假设或原假设,H_1 被称为备择假设

我们想把这两段文字转化成数学语言, 因此将其转化为概率问题, 将女士品茶猜对的杯数转换成二项分布

$H_0 : X \sim b(10, p), (p \leq 0.5), H_1 : X \sim b(10, p), p > 0.5$

Fisher 认为, 如果猜对 9 杯以上, 那么就说明原假设不成立. 这里的 9 杯以上被称为单边拒绝域.

之所以划定 9 杯以上作为拒绝域, 是因为它们的概率和小于 0.05

$P(X \geq 9) = 0.0011 \leq 0.05$

选择 0.05 是因为 Fisher 觉得 0.05 足够的小, 在此将 0.05 称作显著性水平

如果我们改变一下假设

H_0 : 不具备分辨的能力, H_1 : 具备分辨的能力

在此, 只需要分辨出来就好了, 并不需要正确, 因此猜错的足够多或者足够少时, 都可以认为具有分辨能力

因此假设可以改成

$H_0 : p = 0.5, H_1 : p \neq 0.5$

在 H_0 成立的前提条件下,$P(X \leq 1) + P(X \geq 9) = 0.022 \leq 0.05$

以上被称为双边拒绝域

事实上,H_0, H_1 的地位不是平等的, 假设检验更倾向于保护原假设.

打个比方 H_0 就像是某个案件中的嫌疑人.

H_0 : 无罪, H_1 : 有罪

由于总体和样本都是具有随机性的, 所以有可能原假设是对的, 但我们却拒绝了它, 或者原假设是假的, 我们却接受了它.

	H_0 为真	H_1 为真
拒绝 H_0	第一类错误 (弃真)	正确决定
拒绝 H_0	正确决定	第二类错误 (纳伪)

冤枉一个好人要比放走一个坏人更无法容忍.

因此我们可以看出第一类错误更加严重, 因此实际中我们会更加注意降低第一类错误的概率, 因为这两个错误发生的概率不能同时降低

4.2　一元正态总体下的假设检验

4.2.1　期望的假设检验 σ 已知

如果总体 $X \sim N(\mu, \sigma^2)$, 容量为 n 的样本均值为 $\bar{X}$, 假设如下:

$$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$$

在 H_0 成立的前提下, 有:

$$\bar{X} \sim N(\mu_0, \frac{\sigma^2}{n})$$

还是为了方便稍作变换 $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$

这个拒绝域很容易得到

当然如果是单边检验, 以右边检验为例:

如果总体 $X \sim N(\mu, \sigma^2)$, 容量为 n 的样本均值为 $\bar{X}$, 假设如下:

$$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$$

如果 σ 已知, 在 H_0 成立的前提下, 有:

$$\bar{X} \sim N(\mu_0, \frac{\sigma^2}{n})$$

还是为了方便稍作变换 $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

这个拒绝域很容易得到

$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$

4.2.2　期望的假设检验 σ 未知

如果总体 $X \sim N(\mu, \sigma^2)$, 容量为 n 的样本均值为 $\bar{X}$, 假设如下:

$$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$$

在 H_0 成立的前提下, 有:

$T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t(n-1)$

这个拒绝域也很容易得到

当然, 如果是大样本, 那么样本方差可直接替代总体方差

$Z = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim N(0,1)$

4.2.3　方差的假设检验 (μ 已知)

如果总体 $X \sim N(\mu, \sigma^2)$, 容量为 n 的样本, 假设如下:

$$H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 \neq \sigma_0^2$$

在 H_0 成立的前提下, 有:

$$\chi^2 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2(n)$$

这个拒绝域很容易得到

4.2.4　方差的假设检验 (μ 未知)

如果总体 $X \sim N(\mu, \sigma^2)$, 容量为 n 的样本, 假设如下:

$$H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 \neq \sigma_0^2$$

在 H_0 成立的前提下, 有:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

这个拒绝域很容易得到

4.3　两个正态总体下的假设检验

4.3.1　期望的假设检验 σ_1, σ_2 已知

$H_0 : \mu_1 = \mu_2 \Leftrightarrow H_1 : \mu_1 \neq \mu_2$

检验统计量:

$Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

4.3.2　期望的假设检验 $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$H_0 : \mu_1 = \mu_2 \Leftrightarrow H_1 : \mu_1 \neq \mu_2$

检验统计量:

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

其中 $S_w = \frac{(n_1-1)S_X^2+(n_2-1)S_Y^2}{n_1+n_2-2}$

4.3.3　期望的假设检验 σ_1^2, σ_2^2 未知,$\sigma_1 \neq \sigma_2$, 大样本

$H_0 : \mu_1 = \mu_2 \Leftrightarrow H_1 : \mu_1 \neq \mu_2$

检验统计量

$Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

4.3.4　期望的假设检验 σ_1^2, σ_2^2 未知,$\sigma_1 \neq \sigma_2$, 小样本

$T = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}} \sim t(l)$

$l = \frac{(\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2})^2}{\frac{S_X^4}{n_1^2(n_1-1)}+\frac{S_Y^4}{n_2^2(n_2-1)}}$

4.3.5　方差的假设检验

$H_0 : \sigma_1^2 = \sigma_2^2 \Leftrightarrow H_1 : \sigma_1^2 \neq \sigma_2^2$

检验统计量:

$F = \frac{S_X^2}{S_Y^2} \sim F(n_1 - 1, n_2 - 1)$

4.4　成对数据的比较

设 X 总体和 Y 总体服从正态分布, 两总体样本数据匹配或者配对.

观测序号	样本 1	样本 2	差值
1	x_1	y_1	$z_1 = x_1 - y_1$
2	x_2	y_2	$z_2 = x_2 - y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i	$z_i = x_i - y_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	$z_n = x_n - y_n$

$H_0 : \mu_1 - \mu_2 = d \Leftrightarrow H_1 : \mu_1 - \mu_2 \neq d$

检验统计量:

$T = \frac{\bar{Z}-d}{S_Z/\sqrt{n}} \sim t(n-1)$

其中 $S_Z = \sqrt{\frac{\sum_{i=1}^{n}(Z_i-\bar{Z})^2}{n-1}}$

4.5　似然比检验

定义 4.5.1. *功效函数 W 是拒绝域,$\rho_W(\theta) = P((X_1, X_2, \cdots, X_n) \in W|\theta)$*

定义 4.5.2. *一致最大功效检验 (UMP) 如果 W 是水平为 α 的否定域, 且对一切水平不超过 α 的否定域 $\tilde{W}$ 有 $\rho_W(\theta) \geq \rho_{\tilde{W}}(\theta)$ (对一切 $\theta \in \Theta_1$), 则称 W 为一致最大功效检验.*

定义 4.5.3. *无偏检验 如果 W 是水平为 α 的否定域, 若对一切 $\theta \in \Theta_1$, 有 $\rho_W(\theta) \geq \alpha$, 则称 W 是无偏检验*

定义 4.5.4. *一致最大功效无偏检验 UMPU 如果 W 是无偏否定域中, 水平为 α 的一致最大功效否定域, 则称 W 是一致最大功效无偏检验.*

设 X 为连续型随机变量, 密度函数为 $f(x;\theta), \theta \in \Theta = \{\theta_1, \theta_2\}, x = (x_1, x_2, \cdots, x_n)$ 为样本值. 考虑检验问题 $H_0 : \theta = \theta_1 \Leftrightarrow H_1 : \theta = \theta_2$

定理 4.5.5. *N-P 引理 给定检验水平 α, 令*

$$W_0 = \{x : \lambda(x) \triangleq \frac{L(\theta_2; x)}{L(\theta_1, x)} > \lambda_0\}$$

为水平为 α 的否定域, 则它是唯一水平为 α 的 UMP 否定域且无偏.

即对任意的 W 满足 $\rho_{W_0}(\theta_2) \geq \rho_W(\theta_2)$

定理 4.5.6. *设 X 具有单参数指数型分布, 给定检验问题*

$H_0 : \theta \leq \theta_1 \Leftrightarrow H_1 : \theta > \theta_1$

令 $W_0 = \{(X_1, X_2, \cdots, X_n) : T > C\}$

对 $\alpha \in (0, 1)$, 若存在 C 满足

$$P(T > C|\theta_1) = \alpha$$

则 W_0 是检验水平为 α 的 UMP 否定域

很抱歉, 我一点也看不懂.

4.6　广义似然比检验

对于假设 $H_0 : \theta \in \Theta_0 \Leftrightarrow H_1 : \theta \in \Theta_1$, 如果 H_0 成立, 由极大似然原理, 最可能有

$$\sup_{\theta\in\Theta_0}\{L(x;\theta)\} > \sup_{\theta\in\Theta_1}\{L(x;\theta)\}$$

其中 $x = (x_1, x_2, \cdots, x_n)$ 记

$$\lambda(x) = \frac{\sup\limits_{\theta\in\Theta}\{L(x;\theta)\}}{\sup\limits_{\theta\in\Theta_0}\{L(x;\theta)\}} \triangleq \frac{L(\Theta)}{L(\Theta_0)}$$

则当 H_0 成立时,$\lambda(x)$ 应该较小, 否则, 就不能认为 H_0 成立, 而应认为 H_1 成立故 H_0 的拒绝域应为

$$W_0 = \{x : \lambda(x) > \lambda_0\}$$

其中 λ_0 依赖于犯第一类错误的概率 α, 当 α 给定后,λ_0 由下式确定

$$\alpha = P(\lambda(x) > \lambda_0|\Theta_0)$$

这种检验法, 称为广义似然比检验法.

4.7　拟合优度检验

回忆高中生物, 孟德尔得出 9331 的比例.

但实际试验中几乎不可能是达到完美的 9331 的比例

我们对该比例进行的假设检验就是拟合优度检验.

离散卡方拟合优度检验

$$\sum_{\text{所有类}} \frac{(\text{实际频数-理论频数})^2}{\text{理论频数}} \to \chi^2(\text{类的个数}-1)$$

建立假设:

$H_0 : P(A_i) = p_i$

其中 p_i 分别为 $\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}$

类别	p_i	np_i	n_i	$\frac{(n_i-np_i)^2}{np_i}$
1	$\frac{9}{16}$	312.75	315	0.016
2	$\frac{3}{16}$	104.25	108	0.135
3	$\frac{3}{16}$	104.25	101	0.101
4	$\frac{1}{16}$	34.75	32	0.218
合计	1	556	556	0.47

$\chi^2 = \sum_{i=1}^{r} \frac{(n_i-np_i)^2}{np_i} = 0.47$

在接受域, 因此可以接受原假设

以上为无参数的情况, 若有参数, 则首先需要对参数进行最大似然估计

卡方的自由度变为 $k-1-r$ 即在无参数的情况下减去一个参数数量

连续卡方拟合优度检验

连续情况下则需要对其进行分类

之后与离散型一致

4.8　列联表

自己看吧, 电脑没电了, 敲不动了.

4.9　其他非参数检验

累瘫, 敲不动了

第五章　一元线性回归

5.1　相关系数

皮尔逊相关系数:

$$r = \frac{\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n}}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}\sqrt{\frac{\sum(y_i-\bar{y})^2}{n}}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} = \frac{l_{xy}}{\sqrt{l_{xx}}\sqrt{l_{yy}}}$$

斯皮尔曼相关系数:

$r = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$

5.2　一元线性回归模型

$Y = \beta_0 + \beta_1 X + e, e \sim N(0, \sigma^2)$

用最小二乘思想得到未知参数 β_1, β_0 的估计值,

$\hat{\beta}_1 = \frac{l_{XY}}{l_{XX}}$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

可证明回归直线穿过点 $(\bar{X}, \bar{Y})$

5.3　估计的性质

1. $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 皆服从正态分布

2. $E(\hat{\beta}_1) = \beta_1, D(\hat{\beta}_1) = \frac{\sigma^2}{l_{XX}}$

3. $E(\hat{\beta}_0) = \beta_0, D(\hat{\beta}_0) = (\frac{1}{n} + \frac{\bar{X}^2}{l_{XX}})\sigma^2$

4. $E(SSE) = (n-2)\sigma^2$, 即 $\hat{\sigma^2} = \frac{SSE}{n-2}$ 是 σ^2 的无偏估计

5.4　显著性检验

离差的分解:$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

两边平方后有:

$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

离差平方和 = 残差平方和 + 回归平方和

SST = SSE+SSR

如果 SSR 的数值比较大, 则 SSE 的数值比较小, 回归效果比较好

如果 SSR 的数值比较小, 则 SSE 的数值比较大, 回归效果比较差

在回归方程中, 我们应当判断 X,Y 之间确实有一个线性关系, 这才有实际意义.

$H_0 : \beta_1 = 0$(回归不显著) $\Leftrightarrow$ $H_0 : \beta_1 \neq 0$ (回归显著)

5.4.1　F 检验法

因为 $\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$;$H_0$ 为真时,$\frac{SSE}{\sigma^2} \sim \chi^2(1)$; 并且二者相互独立, 所以

$$F = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$$

F 表明了 Y 与 X 的线性关系与残差之比

F 很大时, 表示 SSR 值比较大, 应拒绝原假设, 认为有很好的线性关系

5.4.2　判别系数 r 检验法

容易验证 $SSR = \hat{\beta}_1 L_{XY} = \frac{l_{XY}}{l_{XX}}, r^2 = \frac{l_{XY}^2}{l_{XX} l_{XY}}$
$r = \sqrt{\frac{l_{XY}^2}{l_{XX} l_{YY}}} = \sqrt{\frac{SSR}{SST}}$
$|r|$ 较大, 则线性回归显著, 回归效果比较好.

5.4.3　t 检验法

因为 $\frac{SSE}{\sigma^2} \sim \chi^2(n-2); \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{l_{XX}})$

$$T = \hat{\beta}_1 \sqrt{\frac{l_{XX}}{SSE/(n-2)}} \sim t(n-2)$$

5.5　点估计与区间估计

5.5.1　点估计

当 $x = x_0$ 时, 用 $\hat{y_0} = \beta_0 + \beta_1 x_0$ 预测 Y_0 的观测值 y_0 称为点估计, 可以证明点估计无偏

5.5.2　区间估计

假定与 x_0 相对应的实际观测值为 y_0, 则

$$y_0 = \beta_0 + \beta_1 x_0 + e_0$$

其中 $e_0 \sim N(0, \sigma^2)$
对给定置信水平 $1-\alpha$, 求一个 δ, 使得 $P(|y_0 - \hat{y_0}| < \delta) = 1 - \alpha$
很容易知道 $y - y_0$ 服从正态分布, 且

$E(y_0 - \hat{y_0}) = 0$

$D(y_0 - \hat{y_0}) = \sigma^2[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{l_{XX}}]$

所以 $y_0 - \hat{y_0} \sim N(0, \sigma^2[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{l_{XX}}])$

并且 $\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$

因此

$$\frac{(y_0 - \hat{y_0})}{\sqrt{\frac{SSE}{(n-2)}[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{l_{XX}}]}} \sim t(n-2)$$

因此 δ 的选择为

$$\delta = t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{SSE}{(n-2)}[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{l_{XX}}]}$$