第十一章 数理记录初步

- 第一节 基本概念
- 第二节 参数的点估计
- 第三节 参数的区间估计
- 第四节 参数的假设检查
- 第五节 一元线性回归

从宿舍到教室需要花费多少时间？

相信大家在心里均有一种大概的“数”了

问题

你是怎么得到这个“数”的？

这就是一种经典的记录思维过程。

数据 → 归纳 → 成果

数理记录就是一种归纳推断的过程。

数理记录是以概率论为基础，有关试验数据的搜集、整顿、分析、推断的一门科学与艺术。

问题 什么是实验数据？

科学试验，或对某事物、现象进行观测获得的数据称为试验数据。

特点 数据受随机因素的影响。

——可以通过某种概率分布老描述

问题 实验数据的处理过程？

《数理记录》就是围绕着四个过程来进行研究的.

数据 **搜集、整顿、分析、推断**

第一节 基本概念

1.总体、样本、记录量

2.几种常用记录量的分布

1 总体、样本

引例1

某工厂为检测出厂的100 000只灯泡的寿命，随机抽取了1 000只灯泡进行检测.

引例2

为了解某城市职工的年收入情况，随机抽取一少部分职工进行调查统计.

引例3

某电器公司开发一种使用新型灯丝的灯泡，为了了解新型灯丝灯泡的使用寿命，可抽取200只新型灯丝灯泡，测试其使用寿命.

上面这些例子均有一种共同的特点，就是为了研究某个对象的兴致，只研究对象包括的一部分元素，而不是研究对象包括的所有元素，通过这部分元素的研究，推断对象全体的性质，这就引出了总体、个体、和样本的概念.

将试验的全部可能的观察值称为**总体**(也称为**母体**)，每一个可能的观察值称为**个体**．总体中所包含的个体的个数称为总体的**容量**．容量为有限的总体称为**有限总体**，容量为无限的总体称为**无限总体**．

2 样本

在数理统计中，人们一般通过从总体中抽取一部分个体，根据获得的数据来对总体分布进行推断，从总体中抽出的这一部分个体组成的集合称为**样本**(也称子样)，样本中样品的个数称为**样本容量**(也称样本量).

从总体中抽取样本时，一般满足两个规定：

要求每个个体都有相同机会被选入样本，这便意味着每一样本与总体有相同的分布.

1.代表性

要求样本中每个样品取什么值不受其它样品取值的影响，这意味着 相互独立.

2.独立性

满足上述两条的样本称为简朴随机样本，获得简朴随机样本的抽样措施称为简朴随机抽样.在此后，假如不作特殊申明，所说的样本将理解为简朴随机样本.

3 统计量

定义1

在对样本进行观测时，每个个体的取值成果都是一种随机变量．假如样本包括n个个体，则这n个个体的指标可视为n个变量，常用 $(X_1, X_2, X_3, \mathrm{L}, X_n)$ 来表达．样本观测的成果就是这些随机变量的取值，称为样本值，常用 $(x_1, x_2, x_3, \mathrm{L}, x_n)$ 来表达．

设 $(x_1, x_2, x_3 \cdots x_n)$ 是总体 X 的样本，$\theta(x_1, x_2, x_3 \cdots x_n)$ 是样本的函数，如果其中不包括总体的任何未知的参数，那么称 $\theta(x_1, x_2, x_3 \cdots x_n)$ 为一个统计量。

在引例1中，我们希望知道全体灯泡的平均寿命，一个简单的方法就是用样本 $(X_1, X_2, X_3, \mathrm{L}, X_{1000})$ 的平均寿命 $\frac{X_1 + X_2 + X_3 + \mathrm{L} + X_{1000}}{1000}$ 去估计总体的平均寿命．在此过程中，称 $\frac{X_1 + X_2 + X_3 + \mathrm{L} + X_{1000}}{1000}$ 为统计量.

设总体 $X \sim N(\mu, \sigma^2)$，其中 μ 已知，σ^2 未知，$X_1, X_2, \cdots, X_n$ 是 X 的一个样本，则：

$\sum_{i=1}^{n}(X_i - \mu)^2$ 是记录量

$\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i^2$ 不是记录量

设总体$X \sim N(\mu, \sigma^2)$，其中μ未知，σ^2已知，$X_1, X_2, \cdots, X_n$是X的一个样本，则：

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

不是记录量

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

是记录量

几种常用的记录量

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一种样本，$x_1, x_2, \cdots, x_n$是这个样本的一组观测值.

样本均值：

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

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

几种常用的记录量

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一种样本，$x_1, x_2, \cdots, x_n$是这个样本的一组观测值.

样本方差：

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

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

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

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

几种常用的记录量

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一种样本，$x_1, x_2, \cdots, x_n$是这个样本的一组观测值.

样本原则差：

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

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

1 υ分布

设总体服从正态分布，即 $X \sim N(\mu, \sigma^2)$，$(X_1, X_2, X_3, \mathrm{L}, X_n)$ 是X的一个样本，则称

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

为U 统计量其中，μ为总体均值，σ^2 为总体方差.

在记录学中，常用到原则正态分布的上α分位点这个概念，简介如下：

设 $X \sim N(0,1)$ ，对给定的 α $(0<\alpha<1)$ ，称满足条件

$$P(X > U_\alpha) = \alpha \qquad (7)$$

或

$$P(X \geq U_\alpha) = 1-\alpha \qquad (8)$$

的点Uα为原则正态分布的上α分位点或上侧临界值，简称上α点.

$$P(|X|>U_{\frac{\alpha}{2}})=\alpha$$

的点 $U_{\frac{\alpha}{2}}$ 为标准正态分布的双侧 分位点或双侧临界值，简称双α点，其几何意义如图2所示.

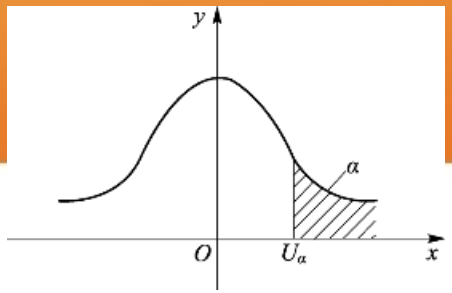


图1

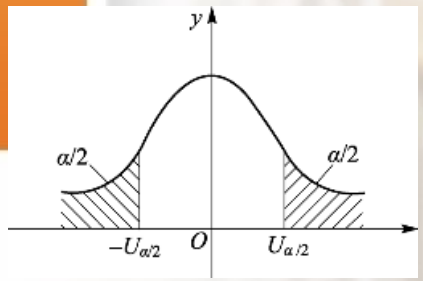


图2

在统计中，U_α可直接根据式（8）查书后附录一（正态分布表）求得；$U_{\frac{\alpha}{2}}$ 可由 $P(X>U_{\frac{\alpha}{2}})=\frac{\alpha}{2}$ 查表求得.

例1

已知某单位职工的月奖金服从正态分布，总体均值为200，总体原则差为40，从该总体中抽取一种容量为20的简朴随机样本，求这同样本的均值介于190~210的概率.

解： 因为 $X \sim N(200, 40^2), n = 20$ ，所以

$$\mu = 200, \frac{\sigma^2}{n} = \frac{40^2}{20} = 80$$

故 $\bar{X} \sim N(200, 80)$ ，

$$P(190 < \bar{x} < 210) = P\left(\frac{190-200}{\sqrt{80}} < \frac{\bar{x}-200}{\sqrt{80}} < \frac{210-200}{\sqrt{80}}\right)$$

$$= 2\Phi(1.118) - 1 = 0.868\ 6 \times 2 - 1$$

$$= 0.737\ 2$$

所以，样本均值介于190~210的概率是0.737 2.

2 χ^2分布

设 $X \sim N(\mu, \sigma^2)$ ，$(X_1, X_2, X_3, \mathrm{L}, X_n)$ 是X的一个样本，

则称 $\dfrac{(n-1)s^2}{\sigma^2}$ 为 χ^2 统计量，且 $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$: $\chi^2(n-1)$

类似于标准正态分布，对于给定的 α $(0<\alpha<1)$，满足条件

$$P\{[\chi^2 < \chi^2_{1-\frac{\alpha}{2}}(n-1)] \cup [\chi^2 > \chi^2_{\frac{\alpha}{2}}(n-1)]\} = \alpha$$

的点 $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ ， $\chi^2_{\frac{\alpha}{2}}(n-1)$ 为 χ^2分布的双侧α分位点或双侧临界值，自由度n-1.

密度函数的图形

2 t分布

设 $X \sim N(\mu, \sigma^2)$，$(X_1, X_2, X_3, \mathrm{L}, X_n)$ 是X的一个样本，

则称 $\dfrac{\bar{X}-\mu}{\dfrac{s}{\sqrt{n}}}$ 为分布其中，且 $t=\dfrac{\bar{X}-\mu}{\dfrac{s}{\sqrt{n}}} \sim t(n-1)$

类似于标准正态分布，对于给定的 α $(0<\alpha<1)$，满足条件

$$P\{|t|>t_{\frac{\alpha}{2}}(n-1)\}=\alpha$$

的点$t_{\frac{\alpha}{2}}(n-1)$为t分布的双侧$\alpha$分位点或双侧临界值，自由度n-1.

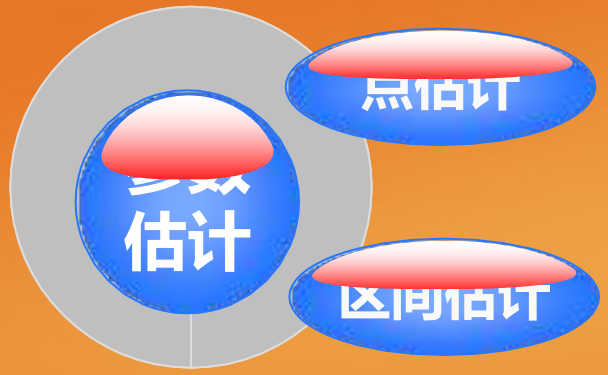
第二节 参数的点估计

1.点估计的措施

2.估计量的评比原则

引例

工厂生产一批铆钉，铆钉头的直径 ξ 是一个随机变量，现在要问这批铆钉头部的平均直径是多少？根据经验知道，ξ 服从正态分布 $N(\mu,\sigma^2)$ ，但参数 μ 和 σ^2 未知，而铆钉头部的平均直径就是参数 $E(\xi)=\mu$ ，因此需设法估计 μ 的值．通常我们从中抽取若干铆钉进行直径的测定，以这些测定量的平均值作为整批铆钉头部直径的平均值的近似值



点估计是以样本的某个函数值来估计总体的未知参数；

区间估计则是用一个区间来估计总体未知参数所在的范围，即把未知参数值估计在某两个界限之间.

估计中常用的方法是：

用一个样本的统计量 $\hat{\theta}$ 估计总体的参数 θ，并称它为**估计量**，其具体值称为**估计值**.

用一种数值来估计某个参数，称为参数的点估计.

例1 既有一批支援灾区的衣裤，共500箱，每箱内放的衣裤数量差不多，估计这批衣裤有多少件.

解： 为估计衣裤总数，随机抽查其中30箱，清点的数量如下：

101，104，98，111，103，97，110，99，99，100

103，97，104，102，96，102，98，101，96，105

105，98，102，101，107，97，104，96，103，94

样本的平均数是 $\bar{x}=\frac{\sum x}{30}=\frac{3033}{30}=101.1$. 以此为总体平均数的估计值，也就是说，每箱平均有衣裤101.1件，500箱共计50 550件衣裤，也可以说，这批支援灾区的衣裤大约有5万件.

例2 估算例1的原则差.

解：

$$S=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=4.147$$

因此，总体原则差的估计值为4.147.

1 数字特征法

用样本的数字特征来估计相应总体的数字特征的方法称为数字特征法．在实际问题中常需要对总体的数学期望 $E(\xi)$ 和方差 $D(\xi)$ 进行点估计。

设 $(x_1, x_2, \cdots, x_n)$ 是来自总体 $\xi \sim N(\mu, \sigma^2)$ 的一种样本，即：总体均值 μ 的估计量 $\hat{\mu}$ 就可以选择样本均值 $\bar{x}$ ，同样样本方差 S^2 也可以作为总体方差 σ^2 的估计量 $\hat{\sigma}^2$ 。即：

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

例2 某厂生产一批铆钉，目前检查铆钉头部的直径，从产品中抽取12只，测得直径（单位：mm）分别为：13.30，13.38，13.40，13.32，13.43，13.48
13.51，13.31，13.34，13.47，13.44，13.50

设铆钉头部直径总体 $\xi \sim N(\mu, \sigma^2)$ ，其中 μ 和 σ^2 未知，用数字特性法估计 μ 和 σ^2 。

解： μ和 σ^2的估计量分别为

$$\xi \sim N(\mu, \sigma^2) \hat{\mu} = \bar{x} = \frac{1}{12}(13.30 + 13.38 + 13.40 + 13.32 + 13.43 + 13.48 + 13.51 + 13.31 + 13.34 + 13.47 + 13.44 + 13.50) = 13.41.$$

$$\hat{\sigma}^2 = s^2 = \frac{1}{12}[(13.30-13.41)^2 + (13.38-13.41)^2 + (13.40-13.41)^2 + (13.32-13.41)^2 + (13.43-13.41)^2 + (13.48-13.41)^2 + (13.51-13.41)^2 + (13.31-13.41)^2 + (13.34-13.41)^2 + (13.47-13.41)^2 + (13.44-13.41)^2 + (13.50-13.41)^2] \approx 0.0053$$

2 顺序统计量法

估计总体参数除数字特征法之外，还有顺序统计量法.

将样本的一组观察值$(x_1, x_2, \square, x_n)$，按大小顺序排列为

$x_1^* \le x_2^* \le \square \le x_n^*$

x_n^*，取最大值

x_1^* 与最小值之差为 R

R，则称为样本的**极差**，取居中的一个（n为偶数，则取居中两数的平均值）为 $\tilde{x}$

$\tilde{x}$，则称为样本的**中位数**，记作

$$\tilde{x} = \begin{cases} x_{k+1}^* & (n = 2k+1) \\ \dfrac{1}{2}(x_k^* + x_{k+1}^*) & (n = 2k) \end{cases}$$

统计量$\tilde{x}$　　R　和　　称为顺序统计量，构成顺序统计量的方法称为顺序统计量法.

对于正态总体，用

$\tilde{x}$　　μ　　R　　σ

来估计　　，用　　来估计

$\tilde{x}$　　$\hat{\mu}$　　R　　$\hat{\sigma}$

是较适宜的，这时，

与　　，　　和　　有以下关系：

$$\hat{\mu} = \tilde{x}$$

$$\hat{\sigma} = \frac{1}{d_n} R \qquad (1)$$

其中

$$\frac{1}{d_n} \approx \frac{1}{n\sqrt{n - \frac{1}{2}}} \qquad (2)$$

n

（2≤

≤10）

例3 设某种灯泡寿命总体服从$N(\mu,\sigma^2)$，其中μ,σ^2未知，今随机取得6只灯泡，测得寿命（单位：h）为

1400，1502，1453，1367，1650，1660

用顺序统计量法估计μ和σ^2的值.

解： 按顺序排列为:

1367<1400<1453<1502<1650<1660，所以

$$\hat{\mu}=\tilde{x}=\frac{1}{2}(1453+1502)=1477.5 \quad h$$

$$R=1660-1367=293$$

$$\frac{1}{d_n}\approx\frac{1}{6}\sqrt{6-\frac{1}{2}}=\frac{1}{6}\sqrt{\frac{11}{2}}\approx 0.3908$$

$$\hat{\sigma}=0.3908\times 293=114.52(h)$$

例3 设某种灯泡寿命总体服从$N(\mu,\sigma^2)$，其中μ,σ^2未知，今随机取得6只灯泡，测得寿命（单位：h）为

1400，1502，1453，1387，1650，1660

用顺序统计量法估计μ和σ^2的值.

解： 在例 2 中，同样可以用数字特征法来估计μ和σ^2.

$$\hat{\mu}=\bar{x}=1505.3(h)$$

$$\hat{\sigma}=113.89(h)$$

这样，对同一正态总体的均值 μ

σ^2

估计量是随机变量，对不一样的样本观测值它有不一样的估计值，这些估计值在未知参数 的真值附近波动．我们但愿估计值的数学期望等于未知参数 的真值，并且但愿的方差越小越好．下面给出估计量的两个评比原则．

1 无偏性

定义 1 设 $\hat{\theta}(X_1, X_2, \cdots, X_n)$ 是未知参数 θ 的一个估计量，若

$$E\left[\hat{\theta}(X_1, X_2, \cdots, X_n)\right] = \theta,$$

对所有 $\theta \in \Theta$ 成立，则称 $\hat{\theta}(X_1, X_2, \cdots, X_n)$ 为 θ 的无偏估量，否则称 $\hat{\theta}(X_1, X_2, \cdots, X_n)$ 为 θ 的有偏估计量．

2 有效性

定义 2 设 $\hat{\theta}_1$, $\hat{\theta}_2$ 均为未知参数 θ 的无偏估计量，若

$D(\hat{\theta}_1) < D(\hat{\theta}_2)$

则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效.

区间估计的具体做法是：

构造两个统计量

$\hat{\theta}_1 = (X_1, X_2, \mathrm{L}, X_n)$ 和 $\hat{\theta}_2 = (X_1, X_2, \mathrm{L}, X_n)$ $(\hat{\theta}_1 < \hat{\theta}_2)$

用区间 $(\hat{\theta}_1, \hat{\theta}_2)$ 来估计未知参数 θ 的可能取值范围，要求 θ 落在区间 $(\hat{\theta}_1, \hat{\theta}_2)$内的概率尽可能大.

通常，我们事先给定一个很小的数α（$0 < \alpha < 1$，常取5%或1%），按概率1-α估计总体参数 θ可能落在区间 $(\hat{\theta}_1, \hat{\theta}_2)$ 内的概率. 1-α称为置信度或置信水平，α称为检验水平（估计不成功的概率），区间 $(\hat{\theta}_1, \hat{\theta}_2)$ 称为置信度为1-α的置信区间.