Lec1: 总体,样本和统计量

张伟平

2012 年 2 月 14 日

1 总体,样本

首先我们看看统计(Statistics)的定义:

> “A branch of mathematics dealing with the collection, analysis, interpretation and presentation of masses of numerical data” —Webster's New Collegiate Dictionary
>
> “The branch of the scientific method which deals with the data obtained by counting or measuring the properties of populations” — Fraser (1958)
>
> “The entire science of decision making in the face of uncertainty” — Freund and Walpole (1987)
>
> “The technology of the scientific method concerned with (1) the design of experiments and investigations, (2) statistical inference” — Mood, Graybill and Boes (1974)

所有的定义都意味着: 统计是一个以知识推断为目的的理论.

1.1 总体

简单的说, 总体是我们所感兴趣的那些个体组成的集合. 比如

例1.1. *假定一批产品有10000件, 其中有正品也有废品, 为估计废品率, 我们往往从中抽取一部分, 如100件进行检查. 此时每件产品即为一个个体, 这批10000件产品称为总体 (Population), 而这批产品的数量10000称为总体容量或者总体大小.*

若总体中个体的数目为有限个, 则称为有限总体 (Finite population), 否则称为无限总体 (Infinite population).

总体的子集称为子总体 (Subpopulation). 如果不同的子总体有着不同的特征(或者性质), 那么将这些不同的子总体在研究开始时便区分开来能够更好的了解总体. 比如某种特定的药品

对特定的子总体有着不同的效应, 那么如果不加区分这些不同的子总体, 则该药品的效应可能就会被混淆而不能辨别出来.

另外, 区别出来这些不同的子总体也可以获得更准确的估计(推断). 比如要研究人们身高的分布, 则按照性别,人种等把人群分为一些子总体, 能够作出更准确的推断.

从而, 按照总体是否有着性质迥异的子总体, 可以把总体分为“ 同质” (Homogeneity)和“ 异质” (Heterogeneity)总体. 绝大部分统计方法都是要求总体同质的. 这里我们仅考虑同质总体.

进一步, 人们真正所关心的不是总体内个体的本身, 而是关心个体上的一项(或几项) 数量指标, 如日光灯的寿命, 零件的尺寸. 在上例中若产品为正品用0表示, 若产品为废品用1表示, 我们关心的个体取值是0还是1. 因此我又可获得总体的如下定义:

总体可以看成有所有个体上的某种数量指标构成的集合, 因此它是数的集合.

由于每个个体的出现是随机的, 所以相应的个体上的数量指标的出现也带有随机性. 从而可以把此种数量指标看成随机变量(Random variable,简记为$r.v.$), 而该数量指标在总体中的分布就是此随机变量的分布. 以上例来说明, 假定10000只产品中废品数为100件, 其余的为正品, 废品率为0.01. 我们定义随机变量X如下:

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

其概率分布为0–1分布$B(1, 0.01)$. 因此特定个体上的数量指标是$r.v.$ X的观察值(Observation). 这样一来, 总体可以用一个随机变量及其分布来描述, 获得如下定义:

定义 1.1. *总体是一个概率分布.*

既然总体可以视为是一个概率分布, 因此若此概率分布为“ xx分布”, 也经常称总体为“ xx总体”. 比如若为正态分布, 则称为“ 正态总体”; 若为指数分布, 则称为“ 指数总体”等.

1.2 样本

样本(Sample)是由总体中(按某种原则)抽取的一部分个体组成的集合. 样本所包含的个体数目即称为“ 样本容量” 或者“样本大小” (sample size). 个体被取出来组成样本过程称为抽样(sampling), 抽样的方式有两种: 概率抽样(probability sampling) 和非概率抽样(nonprobability sampling).

- *概率抽样* 所谓概率抽样, 是指总体中的每个个体都以一个事先可以确定准确的概率被抽取出来作为样本. 概率抽样的方式有: 简单随机抽样, 等距抽样, 分层抽样,多阶段抽样等等(详细请参看抽样调查). 其中常用的就是简单随机抽样. 这种抽样方式有两个特征:

  (1) 每个个体都以相同的概率被抽取. 这意味着每个个体都具有代表性.

  (2) 个体的取值之间相互独立.

  因此, 若记样本为$X_1, \cdots, X_n$, 总体为X, 则在简单随机抽样方式下, 样本$X_1, \cdots, X_n$与总体X是独立同分布的, 常记为

  $$X_1, \cdots, X_n \ i.i.d \ X$$

  由简单随机抽样获得的样本$(X_1, \cdots, X_n)$称为*简单随机样本*. 用数学语言将这一定义叙述如下:

定义 1.2. 设有一总体F, $X_1, \cdots, X_n$为从F中抽取的容量为n的样本, 若

(i) $X_1, \cdots, X_n$相互独立,

(ii) $X_1, \cdots, X_n$相同分布, 即同有分布F,

则称$X_1, \cdots, X_n$为简单随机样本, 有时简称为简单样本或随机样本.

设总体为F, $X_1, \cdots, X_n$为从总体中抽取的简单随机样本, 则$X_1, \cdots, X_n$的联合分布为

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

若F有密度f,则其联合密度为

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

当总体容量较小时,只有进行有放回抽样才能获得简单随机样本. 当总体容量较大或所抽样本在总体中所占比例较小时, 可以近似认为无放回抽样获得的样本是简单随机样本.

这里我们仅考虑简单随机样本, 以下简称样本.

- 非概率抽样 这是指当总体中的某些个体没有机会被抽出或者个体被抽出的概率不能准确确定时的抽样方法. 因此个体被抽出的标准是基于感兴趣的总体的一些假设而作出的. 由于个体是非随机的被抽出, 因此非概率抽样不能估计抽样误差. 非概率抽样方式有偶然抽样(Accidental sampling), 配额抽样(Quota sampling)和目的抽样(Purposive sampling)等.

在个体被抽取之前, 准备抽取大小为n的样本就可以视为是随机变量, 因此常记为$X_1, \cdots, X_n$; 而当个体被抽取后, 样本就表现为具体的数值$x_1, \cdots, x_n$(称为样本的观测值或者样本的一组取值, 样本的可能取值范围常称为样本空间, 记为$\mathcal{X}$. 因此样本既可以视为是随机变量(抽取前), 又可以视为是具体的数值(抽取后).

1.3 抽样分布

样本可以视为是随机变量, 从而样本的概率分布称为抽样分布. 要决定抽样分布, 就要根据观察值的具体指标的性质(这往往涉及有关的专业知识) , 以及对抽样方式和对试验进行的方式的了解, 此外常常还必须加一些人为的假定. 下面看一些例子:

例1.2. 一大批产品共有N件, 其中废品M件, N已知, 而M未知. 现在从中抽出n个检验其中废品的件数, 用以估计M或废品率$p = M/N$. 抽样方式为: 不放回抽样, 一次抽一个, 依次抽取, 直到抽完n个为止. 求抽样分布.

先将问题数量化. 设X_i表示第i次抽出的样本, 令

$$X_i = \begin{cases} 1 & \text{抽出的为废品} \\ 0 & \text{抽出的为合格品}, \end{cases}$$

样品$X_1, \cdots, X_n$中的每一个都只能取$0, 1$值. 给定一组样本$x_1, \cdots, x_n$, 每个x_i为0或1. 我们所求的抽样分布为$P(X_1 = x_1, \cdots, X_n = x_n)$. 若记事件$A_i = \{X_i = x_i\}$, 利用概率乘法公式

$$P(A_1 \cdots A_n) = P(A_1)P(A_2|A_1) \cdots P(A_n|A_1A_2 \cdots A_{n-1})$$

不难求出抽样分布. 为便于讨论, 先看$n=3$.设$x_1=1, x_2=0, x_3=1$,则

$$\begin{aligned}&P(X_1=1, X_2=0, X_3=1)\\&=P(X_1=1)P(X_2=0|X_1=1)P(X_3=1|X_1=1, X_2=0)\\&=\frac{M}{N}\cdot\frac{N-M}{N-1}\cdot\frac{M-1}{N-2}=\frac{M}{N}\cdot\frac{M-1}{N-1}\cdot\frac{N-M}{N-1}\end{aligned}$$

对一般情形, 记$\sum_{i=1}^{n}x_i=a$,利用概率乘法公式易求

$$\begin{aligned}&P(X_1=x_1, X_2=x_2, \cdots, X_n=x_n)\\&=\frac{M}{N}\cdot\frac{M-1}{N-1}\cdots\frac{M-a+1}{N-a+1}\cdot\frac{N-M}{N-a}\cdots\frac{N-M-n+a+1}{N-n+1},\\&\text{当 } x_1, \cdots x_n\text{都为 0或 1, 且 } \sum_{i=1}^{n}x_i=a \text{ 时为上述结果(其余情形为 0) .}\end{aligned}\tag{1.1}$$

由上述计算可见样本$X_1, \cdots, X_n$不是相互独立的, 抽样分布是利用乘法公式, 通过条件概率计算出来的.

例1.3. *仍以上例为例, 抽样方式改为有放回抽样, 即每次抽样后记下结果, 然后将其放回去, 再抽第二个, 直到抽完n个为止, 求抽样分布.*

在有放回抽样情形, 每次抽样时, N个产品中的每一个皆以$1/N$的概率被抽出, 此时$P(X_i=1)=M/N,\ P(X_i=0)=(N-M)/N$, 故有

$$\begin{aligned}&P(X_1=x_1, \cdots, X_n=x_n)=\left(\frac{M}{N}\right)^a\left(\frac{N-M}{N}\right)^{n-a},\\&\text{当 } x_1, \cdots x_n\text{都为 0或 1, 且 } \sum_{i=1}^{n}x_i=a \text{ 时为上述结果(其余情形为 0).}\end{aligned}\tag{1.2}$$

可见此例比上例要简单, 因为本例中样本$X_1, \cdots, X_n$是独立同分布的, 而上例中$X_1, \cdots, X_n$不独立. 当n/N 很小时, (1.1)和(1.2)差别很小. 因而当n/N 很小时,可把上例中的无放回抽样当作有放回抽样来处理.

1.4 统计推断

从总体中抽取一定大小的样本去推断总体的概率分布的方法称为统计推断(Statistical Inference)

当总体分布F的形式已知, 只是含有有限个未知参数时, 要研究的问题常常表现为对参数的某种推断. 比如

例1.4. *假设误差服从正态分布$N(0, \sigma^2)$, 将某物体称重n次, 而样本$X_1, \cdots, X_n$的抽样分布很容易得到*

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

对此试验的统计推断可以是物体的重量的估计(用$\bar{X}$来估计), 或者称重精度界限等等. 因此这类问题称为是参数统计.

而当总体分布形式未知时所进行的统计推断称为非参数统计推断, 非参数统计推断的主要目的是对总体分布作出推断.

统计推断 包括下列三方面内容: (1) 提出种种统计推断的方法. (2) 计算有关推断方法性能的数量指标, 如前述例子中用$\bar{X}$ 估计$N(a,\sigma^2)$中的a用$P(|\bar{X}-a|>c)$表示推断性能的数量指标. (3) 在一定的条件和优良性准则下寻找最优的统计推断方法, 或证明某种统计推断方法是最优的.

2 统计量

定义 2.1. 由样本算出的量是统计量 *(Statistic),* 或曰, 统计量 是样本的函数.

对这一定义我们作如下几点说明:

(1) 统计量只与样本有关, 不能与未知参数有关. 例如$X\sim N(a,\sigma^2)$, $X_1,\cdots,X_n$是从总体X中抽取的i.i.d.样本, 则$\sum_{i=1}^n X_i$和$\sum_{i=1}^n X_i^2$都是统计量, 当a和σ^2皆为未知参数时, $\sum_{i=1}^n(X_i-a)$和$\sum_{i=1}^n X_i^2/\sigma^2$都不是统计量.

(2) 由于样本具有两重性, 即样本既可以看成具体的数, 又可以看成随机变量; 统计量 是样本的函数, 因此统计量 也具有两重性. 正因为统计量 可视为随机变量(或随机向量), 因此才有概率分布可言, 这是我们利用统计量进行统计推断的依据.

二、若干常用的统计量

1. 样本均值

设$X_1,\cdots,X_n$是从某总体X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). 它分别反映了总体数学期望的信息.

2. 样本方差

设$X_1,\cdots,X_n$是从某总体X中抽取的样本,则称

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

为样本方差(Sample variance).它分别反映了总体方差的信息, 而S_n反映了总体标准差的信息.

3. 样本矩

设$X_1,\cdots,X_n$为从总体F中抽取的样本, 则称

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

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

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

为样本 k阶中心矩.特别$k=2$时,$m_{n,2}=S_n^2$,即样本方差. 样本的原点矩和中心矩统称为样本矩 (Sample moments).

4. 二维随机向量的样本矩

设$(X_1,Y_1),\cdots,(X_n,Y_n)$为从二维总体$F(x,y)$中抽取的样本, 则

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\qquad S_X^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$
$$\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}nY_i,\qquad S_Y^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$
$$S_{XY}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$$

分别称为X和Y的样本均值、样本方差及X 和Y的样本协方差 (Sample covariance).

5. 次序统计量及其有关统计量

设$X_1,\cdots,X_n$为从总体F中抽取的样本, 把其按大小排列为$X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)}$,则称$(X_{(1)},X_{(2)},\cdots,X_{(n)})$为次序统计量 (Order statistic), $(X_{(1)},\cdots,X_{(n)})$的任一部分也称为次序统计量 .

利用次序统计量可以定义下列统计量:

(1) 样本中位数:

$$m_{1/2}=\begin{cases}X_{(\frac{n+1}{2})} & \text{当n为奇数}\\ \frac{1}{2}[X_{(n/2)}+X_{(n/2+1)}] & \text{当n为偶数}\end{cases}\tag{2.1}$$

样本中位数 (Sample median)反映总体中位数的信息. 当总体分布关于某点对称时, 对称中心既是总体中位数又是总体均值, 故此时$m_{1/2}$也反映总体均值的信息.

(2) 极值: $X_{(1)}$和$X_{(n)}$称为样本的极小值和极大值,它们统称为样本极值 (Extreme values of sample). 极值统计量在关于灾害问题和材料试验的统计分析中是常用的统计量.

(3) 样本p分位数 $(0<p<1)$: 可定义为$X_{[(n+1)p]}$,此处$[a]$表示实数a的整数部分. 当$p=1/2$, n为奇数时, 此定义与(1)中的样本中位数相同. 样本p分位数(Sample p-fractile)反映了总体p分位数信息.

(4) 样本极差: $R=X_{(n)}-X_{(1)}$,称为样本极差 (Sample range), 它是反映总体分布散布程度的信息.

6. 样本变异系数

设$X_1,\cdots,X_n$为从总体F中抽取的样本,则称

$$\hat{V}=S_n/\bar{X}\tag{2.2}$$

为样本变异系数 (Sample coefficient of variation). 它反映了总体变异系数(Population coefficient of variation) c_ν的信息. 总体变异系数的定义是: $c_\nu=\sqrt{Var(X)}/E(X)$,它是衡量总体分布散布程度的量, 但这散布程度是以总体均值为单位来度量.

7. 样本偏度系数

设$X_1,\cdots,X_n$为从总体F中抽取的样本,则称

$$\hat{\beta}_1=\frac{m_{n,3}}{m_{n,2}^{3/2}}=\sqrt{n}\sum_{i=1}^{n}(X_i-\bar{X})^3\Big/\Big(\sum_{i=1}^{n}(X_i-\bar{X})^2\Big)^{3/2} \tag{2.3}$$

为样本偏度系数 (Sample skewness). 它反映了总体偏度系数的信息, 总体偏度系数(Population skewness)定义是: $\beta_1=\mu_3/\mu_2^{3/2}$,此处$\mu_i$ $(i=2,3)$是总体的i阶中心矩. β_s是反映总体分布的非对称性或“偏倚性”的一种度量. 正态分布$N(a,\sigma^2)$的偏度为零.

8. 样本峰度系数

设$X_1,\cdots,X_n$为从总体F中抽取的样本,则称

$$\hat{\beta}_2=\frac{m_{n,4}}{m_{n,2}^2}-3=n\sum_{i=1}^{n}(X_i-\bar{X})^4\Big/\Big(\sum_{i=1}^{n}(X_i-\bar{X})^2\Big)^2-3 \tag{2.4}$$

为样本峰度系数 (Sample kurtosis). 它反映了总体峰度系数β_k的信息. 总体峰度系数(Population kurtosis)定义是:$\beta_2=\mu_4/\mu_2^2-3$,其中μ_i $(i=2,4)$如前所述. β_k是反映总体分布密度曲线在众数附近的“峰”的尖峭程度的一种度量. 正态分布$N(a,\sigma^2)$的峰度为零.

三、经验分布函数

定义 2.2. *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$X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)}$,对任意实数$x$,称下列函数*

$$F_n(x)=\begin{cases}0 & x<X_{(1)}\\ \dfrac{k}{n} & X_{(k)}\le x<X_{(k+1)},\ k=1,2,\cdots,n-1\\ 1 & X_{(n)}\le x\end{cases} \tag{2.5}$$

为经验分布函数 (Empirical distibution function).

易见经验分布函数是单调非降右连续函数, 具有分布函数的基本性质. 它在$x=X_{(k)}$, $k=1,2,3,\cdots,n$处有间断, 它是在每个间断点跳跃的幅度为$1/n$的阶梯函数. 若记示性函数

$$I_{[A]}(x)=\begin{cases}1 & \text{当}x\in A\\ 0 & \text{其他,}\end{cases}$$

则$F_n(x)$可表为

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n}I_{[X_i\le x]}. \tag{2.6}$$

由定义可知$F_n(x)$是仅依赖于样本$X_1,X_2,\cdots,X_n$的函数, 因此它是统计量. 它可能取值为0, $1/n$, $2/n,\cdots,(n-1)/n$, 1.若记$Y_i=I_{[X_i\le x]}$, $i=1,2,\cdots,n$,则$P(Y_i=1)=F(x)$, $P(Y_i=0)=1-F(x)$, 且$Y_1,Y_2,\cdots,Y_n$, i.i.d. $\sim b(1,F(x))$,故$nF_n(x)=\sum\limits_{i=1}^{n}Y_i\sim b(n,F(x))$,因此有

$$P(F_n(x)=k/n)=P\Big(\sum_{i=1}^{n}Y_i=k\Big)=\binom{n}{k}[F(x)]^k[1-F(x)]^{n-k}$$

利用二项分布的性质可知$F_n(x)$具有下列大样本性质:

(1) 由中心极限定理, 则当$n \to \infty$时有

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1-F(x))}} \xrightarrow{\mathscr{L}} N(0,1).$$

(2) 由Benoulli大数定律, 则在$n \to \infty$时有

$$F_n(x) \xrightarrow{P} F(x)$$

(3) 由Borel强大数定律, 则在$n \to \infty$时有

$$P(\lim_{n\to\infty} F_n(x) = F(x)) = 1$$

(4) 更进一步, 有下列Glivenko-Cantelli Theorem (1933):

定理 2.1. 设$F(x)$为$r.v.\ X$的分布函数, $X_1, \cdots, X_n$为取自总体$F(x)$的简单随机样本, $F_n(x)$为其经验分布函数, 记$D_n = \sup\limits_{-\infty<x<\infty} |F_n(x) - F(x)|$,则有

$$P(\lim_{n\to\infty} D_n = 0) = 1.$$