统计量及其抽样分布01

1、**总体**：我们把研究对象的全体称为总体，构成总体的每个成员称为个体

总体的三层含义：（1）研究对象的全体

（2）数据

（3）总体是一个分布

为了了解总体的分布，我们从总体中随机抽取n个个体，记其指标值为$x_1, x_2, \dots x_n$称为总体的一个样本，n称为**样本容量，或简称样本量**，样本中的个体称为样品。

样本具有两重性：

（1）一方面，由于样本是从总体中随机抽取的，抽取前无法预知他们的数值，因此样本是随机变量，用大写字母$X_1, X_2, \dots X_n$表示

（2）另一方面，样本在抽取以后经观测就有确定的观测值，因此，样本又是一组数值。此时用小写字母$x_1, x_2, \dots x_n$表示是恰当的。

当样本观测值没有具体的数值，只有一个范围时，这样的样本就做分组样本。

2、简单随机抽样的两个要求

（1）随机性：总体中每一个个体都有同等机会被选入样本，这意味着x_i与总体X有相同的分布

（2）独立性：样本中每一样品的取值不影响其他样品的取值，这意味着$x_1, x_2, \dots, x_n$相互独立

用简单随机抽样方法得到的样本称为简单随机样本，也简称样本，于

是，样本$x_1, x_2, \dots x_n$可以看成是独立同分布的随机变量，其共同分布即为总体分布

总体分为有限总体和无限总体：实际中总体中的个体数大多是有限的。当个体数充分大时，将有限总体看作无限总体是一种合理的抽象。对无限总体，随机性与独立性容易实现，困难在于排除有意或无意的人为干扰。对有限总体，只要总体所含个体数很大，特别是与样本量相比很大，则独立性也可基本得到满足。

3、经验分布函数

设$x_1, x_2, \dots x_n$是取自总体分布函数为$F(X)$的样本，若将样本观测值由小到大进行排列，为$x_{(1)}, x_{(2)}, \dots x_{(n)}$，则称$x_{(1)}, x_{(2)}, \dots x_{(n)}$为有序样本。用有序样本定义如下函数

$$F_n(x) = \begin{cases} 0, x < x_{(1)} \\ \frac{k}{n}, x_{(k)} < x < x_{(k+1)}, k = 1,2, \dots, n-1 \\ 1, x_{(x)} \le x \end{cases}$$

则$F_n(x)$是一非减右连续函数，且满足$F_n(-\infty) = 0, F_n(+\infty) = 1$，由此可见$F_n(x)$是一个分布函数，并称$F_n(x)$为经验分布函数。

定理（格里纹科定理）设$x_1, x_2, \dots x_n$是取自总体分布函数为$F(X)$的样本，$F_n(x)$是其经验分布函数，当$n \to \infty$时，有$P\left(\sup_{-\infty<x<\infty} |F_n(x) - F(x)| \to 0\right) = 1$，格里纹科定理表明：当$n$相当大时，经验分布函数是总体分布函数$F(X)$的一个良好的近似。经典的统计学中一切统计推断都以样本为依据，其理由就在于此。

4、频数分布表

（1）对样本进行分组：首先确定组数，作为一般性的原则，组数通常在 5-20 个，对容量较小的样本；通常 5-6 组，容量为 100 左右的样本可以分为 7-10 组，容量为 200 左右的样本分为 9-13 组，容量为 300 及以上的样本分为 12-20 组，目的是使用足够的组来表示数据的差异。

（2）确定每组组距：近似公式为：组距 $\mathrm{d} = (最大观测值 - 最小观测值)/组数$

（3）确定每组的组限：各组区间端点为$a_0, a_0 + d = a_1$，$a_0 + 2d = a_2, \dots, a_0 + kd = a_k$，形成如下的分组区间：

$$(a_0, a_1], (a_1, a_2], \dots, (a_{k-1}, a_k]$$

其中，a_0略小于最小观测值，a_k略大于最大观测值。通常选取每组的组中值来代表该组的变量取值，组中值 $= (组上线 + 组下线)/2$

（4）统计样本数据落入每个区间的个数——频数，并列出其频数频率分布表。

5、样本数据的图形展示

（一）直方图

（二）茎叶图，比较两组样本时，可画出他们的背靠背的茎叶图

6、统计量与抽样分布

样本来自总体，因此样本包含有总体各方面的信息，当人们需要从样本获得对总体各种参数的认识时，最好的方法是构造样本的函数，不同的函数反应总体的不同特征。

定义：设 $x_1, x_2, \dots x_n$ 是取自某总体的样本，若样本函数 $T =$

$T(x_1, x_2, \ldots x_n)$中**不含有任何未知参数**，则称T为统计量，统计量的分布称为抽样分布。

设$X_1, X_2, \ldots X_n$是从某总体X中抽取的一个样本，则

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 都是统计量，而 $\sum_{i=1}^{n}\left(X_i - E(X)\right)^2, [X_i - E(X)]/D(X)$都不是统计量。这是因为其中$E(X)$和$D(X)$都是依赖于总体分布的未知参数。注：尽管统计量不依赖于未知参数，但是它的分布一般是依赖于位置参数的。

7、样本均值及其抽样分布

设$x_1, x_2, \ldots x_n$是从某总体x中抽取的一个样本，其算数平均值称为样本均值，一般用$\bar{x}$表示，即$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{k} x_i$，在分组场合$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{n}$

样本均值的基本性质：

定理 1：若把样本中的数据与样本均值之差成为偏差，则样本所有偏差之和为 0，即$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

定理 2：数据观测值与均值的偏差平方和最小，即在形如$\sum(x_i - c)^2$的函数中，c为任意给定的常数，$c = \bar{x}$

定理 3：设$x_1, x_2, \ldots x_n$是来自某个总体的样本，$\bar{x}$为样本均值

（1）若总体分布为$N(\mu, \sigma^2)$，则$\bar{x}$的精确分布为$N(\mu, \frac{\sigma^2}{n})$

（2）若总体分布位置或不是正态分布，但$E(X) = \mu, Var(x) = \sigma^2$，则$n$较大时，$\bar{x}$的渐进分布为$N(\mu, \frac{\sigma^2}{n})$，常记$\bar{x} \dot{\sim} N(\mu, \frac{\sigma^2}{n})$，这里渐进分布是指$n$较大时的近似分布

8、样本方差与样本标准差

设$x_1,x_2,\dots x_n$是来自某个总体的样本,则它关于样本均值的平均偏差平方和$s_n{}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$称为样本方差(有偏方差)，其算术平方根$s_n=\sqrt{s_n{}^2}$称为样本标准差。在 n 不大时，常用$s_n{}^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$作为样本方差（也称无偏方差)，其算术平方根$s_n=\sqrt{s_n{}^2}$称为样本标准差。

在定义中，$\sum_{i=1}^{n}(x_i-\bar{x})^2$称为偏差平方和，$n-1$称为偏差平方和和自由度。其含义是：在$\bar{x}$确定后，$n$个偏差$x_1-\bar{x},x_2-\bar{x},\dots,x_n-\bar{x}$中只有$n-1$个偏差可以自由变动，而第$n$个则不能自由取值，因为$\sum_{i=1}^{n}(x_i-\bar{x})=0$

样本偏差平方和有三个不同的表达式：

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

分组场合的样本方差为：$s^2=\frac{1}{n-1}\sum_{i=1}^{k}f_i(x_i-\bar{x})=\frac{1}{n-1}(\sum_{i=1}^{k}f_i x_i{}^2-n\bar{x}^2)$

定理：设总体X具有二阶矩，即$E(x)=\mu$，$Var(x)=\sigma^2<\infty$，$x_1,x_2,\dots x_n$是来自某个总体的样本，$\bar{x}$和s^2分别是样本均值和样本方差，则$E(\bar{x})=\mu$，$Var(\bar{x})=\frac{\sigma^2}{n}$，$E(s^2)=\sigma^2$

9、样本矩及其函数

定义：设$x_1,x_2,\dots x_n$是来自某个总体的样本，k为正整数，则统计量$a_k=\frac{1}{n}\sum_{i=1}^{n}x_i{}^k$，称为样本$k$阶原点矩，特别，样本一阶原点矩就是样本均值。统计量$b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k$称为$k$阶中心矩。特别的，样本二阶中心矩为方差。

当样本不对称时，只用均值和方差显得很不够，为此，需要一些

刻画分布形状的统计量，如样本偏度和样本峰度，他们都是样本中心矩的函数。

10、次序统计量及其分布

定义：设$x_1,x_2,\dots x_n$是来自总体X的样本，$X_{(i)}$称为该样本的第i个次序统计量，它的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$X_{(1)}=\min\{x_1,x_2,\dots x_n\}$称为该样本的最小次序统计量，称$X_{(n)}=\max\{x_1,x_2,\dots x_n\}$为该样本的最大次序统计量。

在同一样本中，$x_1,x_2,\dots x_n$是独立同分布的，而次序统计量$X_{(1)},X_{(2)},\dots,X_{(n)}$则既不独立，分布也不相同。

单个次序统计量的分布

定理：设总体X的**密度函数**为$p(x)$，分布函数$F(x)$，$x_1,x_2,\dots x_n$为样本，则第k个次序统计量$X_{(k)}$的密度函数为

$$p_k(x)=\frac{n!}{(k-1)!\,(n-k)!}\big(F(x)\big)^{k-1}\big(1-F(x)\big)^{n-k}p(x)$$

例：设总体密度函数为$p(x)=3x^2,0<x<1$，从该总体抽得一个容量为 5 的样本，试计算$P(x_{(2)}<1/2)$

解：首先求出$x_{(2)}$的分布。由总体密度函数不难求出总体分布函数为

$$F(x)=\begin{cases}0,\ x\le 0\\ x^3,0<x<1\\ 1,x\ge 1\end{cases}$$

由公式可以得出$x_{(2)}$的密度函数$p_2(x)=\frac{5!}{(2-1)!(5-2)!}\big(F(x)\big)^{2-1}\big(1-F(x)\big)^{5-2}p(x)=20x^3*(1-x^3)^3*3x^2=60x^5(1-x^3)^3,0<x<1$，于是

$$P\left(x_{(2)}<\frac{1}{2}\right)=\int_0^{1/2}60x^5(1-x^3)^3\,dx=\int_0^{1/8}20y(1-y^3)\,dy=0.1207$$

多个次序统计量及其函数的分布

定理：次序统计量$\left(x_{(i)},x_{(j)}\right),(i,j)$的联合部分密度函数为

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

次序统计量的函数在实际中经常用到，如样本极差$R_n=x_{(n)}-x_{(1)}$

例题：设总体分布$U(0,1)$, $x_1,x_2,\dots,x_n$为样本，则$(Y,Z)=\left(x_{(1)},x_{(n)}\right)$的联合密度函数为$p(y,z)=n(n-1)(z-y)^{n-2},0<y<z<1$

11、样本分位数与中位数

样本中位数也是一个很常见的统计量，他也是次序统计量的函数，通常如下定义：

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

更一般地，样本p分位数m_p可如下定义：

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

定理：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$p\left(p(x)\right)>0$，则当$n\to\infty$时样本p分位数m_p的渐进分布为

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

特别，对样本中位数，当$n\to\infty$时近似地有

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

通常，样本均值在概括数据方面具有一定的优势。但当数据中含有极端值时，使用中位数比使用均值更好，中位数的这种抗干扰性在统计

中称为具有稳健性。

12、三大抽样分布

（一）χ^2分布

定义：设随机变量$x_1, x_2, \dots x_n$相互独立，且$X_i(i=1,2,\dots,n)$服从标准正态分布$N(0,1)$，则他们的$\sum_{i=1}^{n} {X_i}^2$服从自由度为n的χ^2分布，记为$\chi^2 \sim \chi^2(n)$

χ^2分布的性质：

① χ^2分布的数学期望为：$E(\chi^2)=n$

② χ^2分布的方差为：$D(\chi^2)=2n$

③ χ^2分布具有可加性，即若${\chi_1}^2 \sim \chi^2(n_1), {\chi_2}^2 \sim \chi^2(n_2)$，且相互独立，则${\chi_1}^2+{\chi_2}^2 \sim \chi^2(n_1+n_2)$

（二）t分布

定义：设随机变量$X_1 \sim N(0,1), X_2 \sim \chi^2(n)$，且$X_1$与$X_2$独立，则$t=\frac{X_1}{\sqrt{X_2/n}}$，其分布成为$t$分布，记为$t \sim t(n)$，其中，$n$为自由度。

t分布的性质：

① 当$n \geq 2$时，t分布的数学期望$E(t)=0$

② 当$n \geq 3$时，t分布的方差$D(t)=\frac{n}{n-2}$

③ 当自由度较大（如$n \geq 30$）时，t分布可以用正态分布$N(0,1)$近似

④ 如果随机变量X服从$t(n)$分布，X^2服从$F(1,n)$的F分布

（三）F分布

定义：设随机变量X_1与X_2相互独立，且X_1与X_2分别服从自由度为m和n的χ^2分布，随机变量X有如下表达式$F=\frac{X_1/m}{X_2/n} \sim F(m,n)$，则称$F$服从

第一自由度为m，第二自由度为n的F分布，记为$F(m,n)$，简记为

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

F分布的P分位数$F_p(m,n)$可查F分布表获得，且$F_p(\boldsymbol{m},\boldsymbol{n})=\frac{1}{F_{1-p}(\boldsymbol{n},\boldsymbol{m})}$，因此，$F$分布中两个自由度的参数不可以互换。

13、充分统计量

为研究某个运动员的打靶命中率，我们对该运动员进行测试，观测其10次，发现除第三、六次未命中外，其余8次都命中，这样的观测结果包含了两种信息。

（1）打靶10次命中8次

（2）2次命中分别出现在第3次和第6次打靶上

第二种信息对了解该运动员的命中率是没有什么帮助的。一般地，设我们对该运动员进行n次观测，得到$x_1,x_2,\dots x_n$，每个x_j取值非0即1，命中为1，不命中为0．令$T=x_1+x_2+x_n$，T为观测到的命中次数。在这种场合仅仅记录使用T不会丢失任何与命中率ϑ有关的信息。统计上将这种“样本加工不损失信息”称为“充分统计量”

样本$x=(x_1,x_2,\dots x_n)$有一个样本分布$F_\theta(x)$，这个分布包含了样本中一切有关θ的信息。统计量$T=T(x_1,x_2,\dots x_n)$也有一个抽样分布${F_\theta}^T(t)$像$F_\theta(x)$一样概括了有关θ的一切信息，这即是说在统计量T的取值为t的情况下样本x的条件分布$F_\theta(x|T=t)$已不含θ的信息，这正是统计量具有充分性的含义。

定义：设$x_1,x_2,\dots x_n$是来自某个总体的样本，总体分布函数为$F(x;\theta)$，统计量$T=T(x_1,x_2,\dots x_n)$称为θ的充分统计量，如果再给定T的取值

后，$x_1, x_2, \dots x_n$的条件分布与θ无关

12、因此分解定理

充分性原则：在统计学中有一个基本原则——在充分统计量存在的场合，任何统计推断都可以基于充分统计量进行，这可以简化统计推断的程序。

定理：设总体概率函数为$p(x;\theta)$，$X_1, \dots, X_n$为样本，则$T = T(x_1, x_2, \dots x_n)$为充分统计量的充分必要条件是：存在两个函数$g(t;\theta)$和$h(x_1, x_2, \dots x_n)$，使得对任意的$\theta$和任一组观测值$x_1, x_2, \dots x_n$有

$$p(x_1, x_2, \dots x_n;\theta) = g\big(T(x_1, x_2, \dots x_n;\theta)\big)h(x_1, x_2, \dots x_n)$$

其中，$g(t,\theta)$是通过统计量T的取值而依赖于样本的

点估计、区间估计、假设检验，通 x 过样本具体的统计量来研究总体的参数