第 7 讲*

随机抽样

麻省理工学院 14.30 2006 年春季

Herman Bennett

17 定义

17.1 随机样本

若 $X_1,...,X_n$ 是相互独立的随机变量，那么，对 $\forall\ i \neq j$，有 $f_{X_i}(x) = f_{X_j}(x)$，记作 $f_{X_i}(x) = f(x)$。那么，集合 $X_1,...,X_n$ 称为总体 $f(x)$ 的一个样本容量为 n 的随机样本。

例子：

---掷 1 枚骰子 n 次。

---挑选 10 个麻省理工的学生，测量他们的身高。

- 放回和不放回抽样：从一个大样本总体中抽样（“近似相互独立”）。

- 另外，$X_1,...,X_n$ 这一集合（或抽样）也被称为概率质量/密度函数为 $f(x)$ 的独立同分布随机变量，或者简称为独立同分布样本。

- 注意仍然要区别 X 和 x（我们继续研究随机变量）。

注意：这些讲义不一定是自封的。它们只是对讲座的一种补充而不是替代。

17.2 统计量

随机变量 $X_1, X_2, ..., X_n$ 是来自总体 $f(x)$ 中样本容量为 n 的随机样本。那么，任何实值函数 $T = r(X_1, X_2, ..., X_n)$ 称为一个<u>**统计量**</u>。

- 记住 $X_1, X_2, ..., X_n$ 是随机变量，因此 T 也是一个随机变量，其概率质量/密度函数 $f_T(t)$，它可以取任何实值 t。

17.3 样本均值

<u>样本均值</u>，记为 $\overline{X}_n$，定义为一个样本容量为 n 的随机样本的算术平均值统计量。

$$\bar{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \textbf{(52)}$$

17.4 样本方差

<u>样本方差</u>，记为 $S_n^{\ 2}$，定义为如下统计量：

$$S_n^{\ 2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \textbf{(53)}$$

样本标准差统计量定义为 $S_n = \sqrt{S_n^{\ 2}}$。[1]

- 记住，统计量的观测值用小写字母来表示。因此，$\bar{x}, s^2$, 和 s 就是随机变量 $\bar{X}, S^2$和S的观测值。

---

[1] 样本方差和样本标准差有时又分别用 $\hat{\sigma}^2$ 和 $\hat{\sigma}$ 来表示。

18 样本均值分布和样本方差分布的重要性质

18.1 $\bar{X}$ 和 S^2 的均值和方差

$X_1,\ldots,X_n$ 是来自总体 $f(x)$ 的样本容量为 n 的随机样本，总体均值为 μ（有限的），方差为 σ^2（有限的）。那么，

$$E(\bar{X})=\mu, E(S^2)=\sigma^2, Var(\bar{X})=\frac{\sigma^2}{n}, \text{且} \mathrm{Var}_{n\to\infty}(S^2)\to 0. \qquad \textbf{(54)}$$

- 标准差为：$\sqrt{\mathrm{Var}(\bar{X})}$

例 18.1. 证明式（54）的前三个命题。

18.2 正态总体随机样本的特殊情况

$X_1,..., X_n$ 是一个来自总体 $N(\mu,\sigma^2)$ 的样本容量为 n 的随机样本。那么，

a. $\bar{X}$ 和S^2 是独立的随机变量。 **(55)**

b. $\bar{X}$ 服从 $N\left(\mu,\sigma^2/n\right)$ 分布。 **(56)**

c. $\dfrac{(n-1)S^2}{\sigma^2}$ 服从 $\chi^2_{(n-1)}$ 分布。 **(57)**

例 **18.2.** 证明式**(56)**。

18.3 极限结论（$n\to\infty$）

这些概念被广泛运用于经济学中。

18.3.1（弱）大数法则

若 $X_1,..., X_n$ 为独立同分布（随机抽样）的随机变量，均值 $E(X_i)=\mu$（有限），方差为$Var(X_i)=\sigma^2$（有限）。定义 $\bar{\bar{X}}_n=\frac{1}{n}\sum_{i=1}^n X_i$ 。那么，对任意 $\varepsilon>0$，都有

$$\lim_{n\to\infty} P(\left|\bar{X}_n-\mu\right|<\varepsilon)=1. \quad \textbf{(58)}$$

这个结论也可表示为，

$$\bar{X}_n \xrightarrow{p} \mu \text{（} \bar{X}_n \text{ 以概率收敛于 } \mu \text{）} \quad \textbf{(59)}$$

例 18.3. 利用切比雪夫不等式，证明式（**58**）。注意 $S^2 \xrightarrow{p} \sigma^2$ 可用类似方法加以证明。

18.3.2 中心极限定理 （CLT）

$X_1,\ldots,X_n$ 是独立同分布（ $i.i.d$ ）的随机变量，均值 $E(X_i)=\mu$ （有限），方差为 $Var(X_i)=\sigma^2$ （有限）。设 $\bar{\bar{X}}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ 。那么，对任意 $x \in (-\infty,+\infty)$ ，都有

$$\lim_{n\to\infty} p\left(\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} < x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = \Phi(x) \qquad \textbf{(60)}$$

其中 $\Phi()$ 是标准正态分布的累积分布函数。

用文字解释： 由式**(56)**可知，如果 X_i 服从正态分布，则样本均值统计量 $\bar{X}_n$ 也将服从正态分布。式（**60**）表明：如果 $n\to\infty$ ，则无需考虑这些 X_i 的分布，样本均值统计量的函数 $\frac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma}$ 都将服从正态分布。

实际应用（1）…如果 n 足够大，即使不知道随机样本 $f_{X_i}(x)$ 的分布情况，也可以推测出 $\bar{X}_n$ 的函数， $\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma}$ 的分布。［这是一个很强的结果］

实际应用（2）…若 $Z=\frac{\sqrt{n}(\bar{x}_n-\mu)}{\sigma}$ 且 n 足够大，则

$$F_Z(\frac{\sqrt{n}(\bar{x}_n-\mu)}{\sigma})\approx\Phi(\frac{\sqrt{n}(\bar{x}_n-\mu)}{\sigma}) \tag{61}$$

$$\Downarrow$$

$$\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma}\overset{a}{\sim}N(0,1)\text{或者}\ \ \bar{X}_n\overset{a}{\sim}N(\mu,\sigma^2/n)\ \ (a:\text{近似地}) \tag{62}$$

…无需考虑概率质量/密度函数 $f_{X_i}(x)$ 的分布！

- n 的取值越大，近似值效果越好。但是，n 的值多大是“足够大”？对此没有严格的事先限定，它取决于 $f_{X_i}(x)$ 的基本（总体）分布。$f_{X_i}(x)$ 的钟形分布曲线越窄，说明 n 的值越大。根据过去经验，对于 n 的取值，一些学者遵循经验法则：$n\geq 30$。

- 放大镜（详见模拟）

例 **18.4.** 一个天文学者爱好测量从他的天文台到一颗遥远的星星的距离（单位光年）。由于不断变化的大气状况和测量误差，每一次测量的结果都不能作为准确的距离。所以，该天文学者计划进行多次测量，然后用它们的平均值作为估计的距离。他相信测量值是独立同分布的，均值为 d（实际距离），方差为 4（光年）。那么，需要进行多少次测量才可能保证估计的距离与准确距离之间相差 ± 0.5 光年？