第三节　抽样分布（上）

统计量

经验分布函数

一、统计量

统计量

- 由样本值去推断总体情况，需要对样本值进行“加工”，这就要构造一些样本的函数，它把样本中所含的（某一方面）的信息集中起来.
- 这种不含任何未知参数的样本的函数称为统计量. 它是完全由样本决定的量.

定义

设$X_1, X_2, \cdots, X_n$是来自总体X的一个样本，

$g(X_1, X_2, \cdots, X_n)$是$X_1, X_2, \cdots, X_n$的函数，

若g中不含未知参数，则称$g(X_1, X_2, \cdots, X_n)$

是一个**统计量**。

设$X_1, X_2, \cdots X_n$是来自总体X的一个样本，

$x_1, x_2, \cdots x_n$是一个样本的观察值，则$g(x_1, x_2, \cdots x_n)$

是**统计量**$g(X_1, X_2, \cdots X_n)$**的观察值**。

例 设$X_1, \cdots, X_n$为来自总体 $X \sim N(\mu, \sigma^2)$ 的一个样本，其中μ 未知，σ^2已知，问下列随机变量中哪些是统计量?

[1] $\frac{1}{n}\sum_{i=1}^{n} X_i$ √

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

[3] $\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$ √

[4] $\frac{1}{n}\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$

[5] $X_1^2 + X_2^2 + \sigma^2$ √

[6] $2\mu X_1 X_2 \cdots X_n$

例 设$X_1, \cdots, X_n$为来自总体 $X \sim N(\mu, \sigma^2)$ 的一个样本，其中 μ 已知，σ^2未知 问下列随机变量中哪些是统计量?

[1] $\frac{1}{n}\sum_{i=1}^{n} X_i$ √

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

[3] $\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$ √

[4] $\frac{1}{n}\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$

[5] $X_1^2 + X_2^2 + \sigma^2$

[6] $2\mu X_1 X_2 \cdots X_n$ √

设 $(X_1, X_2, \cdots, X_n)$ 是抽自总体 $X \sim F(x)$ 的样本，将它们按从小到大的次序排列为 $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$，则称 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 为由样本 $X_1, X_2, \cdots, X_n$ 生成的**顺序统计量**，$X_{(k)}$ 称为**第 k 个顺序统计量**.

最大顺序统计量 $X_{(n)} = \max\{X_1, X_2, \cdots, X_n\}$

最小顺序统计量 $X_{(1)} = \min\{X_1, X_2, \cdots, X_n\}$

极差 $R_n = X_{(n)} - X_{(1)}$

样本中位数

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

【几个常见统计量】

它反映了总体均值的信息

样本平均值 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

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

它反映了总体方差的信息

样本方差

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

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

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

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

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

【几个常见统计量】

它反映了总体均值的信息

样本平均值 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

它反映了总体方差的信息

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

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

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

【几个常见统计量】

它反映了总体k阶矩的信息

样本 k 阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, 2, \ldots$

它反映了总体k 阶中心矩的信息

样本 k 阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k, \quad k = 2, 3, \ldots$

【辨析】 $B_2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$ $\qquad$ $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$

【统计量的观察值】

将样本观察值代入，便可求得统计量的观察值。

样本平均值 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

样本方差

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

样本标准差

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

样本k 阶原点矩

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

样本k阶中心矩

$$b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k,$$

$$k=2,3,\cdots$$

学校英语提高班采用小班教学，每班15人．现有A、B两个班参加统一的口语测试，成绩如下表所示：

A班	67	72	93	69	86	84	45	77	88	91	81	76	84	90	63
B班	78	96	56	83	86	48	98	67	62	70	64	97	96	79	86

试问哪个班的成绩较好些？

【演示】用excel计算样本均值和方差.xlsx

若总体X的k阶矩$E(X^k)=\mu^k$存在，则当$n\to\infty$时，

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^{\,k}\xrightarrow{\;p\;}\mu^k\qquad k=1,2,\cdots.$$

事实上 由$X_1,X_2,\cdots,X_n$独立且与X同分布，有$X_1^{\,k},X_2^{\,k},\cdots,X_n^{\,k}$独立且与$X^k$同分布,$E(X_i^k)=\mu^k\ \ k=1,2,\cdots,n$ 再由辛钦大数定律可得上述结论.

再由依概率收敛性质知，可将上述性质推广为

$$g(A_1,A_2,\cdots,A_k)\xrightarrow{\;p\;}g(\mu_1,\mu_2,\cdots,\mu_k)$$

其中g为连续函数.

这就是矩估计法的理论根据（详见7.1）。

【辨析】样本方差与总体方差

对总体而言，随机变量X的数学期望 μ 是已知的，方差 σ^2 未知，

则　　总体的期望 E(X) = μ，总体的方差D(X) = σ^2

$$D(X)=E[(X-E(X))^2]=E[(X-\mu)^2]=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

样本均值的期望

$$E(\overline{X})=E(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\mu$$

样本均值的方差

$$D(\overline{X})=D(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{1}{n}\sigma^2$$

【辨析】样本方差与总体方差

但对样本而言，随机变量X的数学期望 μ 是未知的，故 样本的方差 应当是

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad \text{（无偏估计）}$$

而不是

$$B_2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad \text{（有偏估计）}$$

因为

$$E(S^2)=\sigma^2, E(B_2)=\frac{n-1}{n}\sigma^2,$$

- 当样本容量n充分大时，两者是近似相等的。

【样本方差S^2的期望】

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

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

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

$$\because E(X_i^2)=D(X_i)+(EX_i)^2=\sigma^2+\mu^2$$

$$E[(\bar{X})^2]=D(\bar{X})+[E(\bar{X})]^2=\frac{\sigma^2}{n}+\mu^2$$

$$\therefore E(S^2)=\frac{1}{n-1}[n(\sigma^2+\mu^2)-n(\frac{\sigma^2}{n}+\mu^2)]=\sigma^2$$

【辨析】

$$B_2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$E(B_2)=\frac{n-1}{n}\sigma^2$$

【辨析*】样本方差与总体方差（自由度）

- 自由度指的是等式中能够自由取值的变量的个数，如果有**n**个数能够自由取值，那么自由度就为**n**。
- 样本方差计算公式中，X_i有**n**个可取的值，所以X_i的自由度为**n**，但是，它接着还减去了样本的平均值$\bar{X}$。这其实相当于增加了一个限制条件，原来的自由度要减去**1**，得**n-1**。（因为如果自由度仍为**n**，那么**n**个数可以随意取值的情况下，是不能得到一个确定的均值的；或者说，一堆数，如果知道了均值，那么其实只需要知道另外的**n-1**个数，这堆数中的每个数都确定了）

【类比*】样本方差与总体方差（自由度）

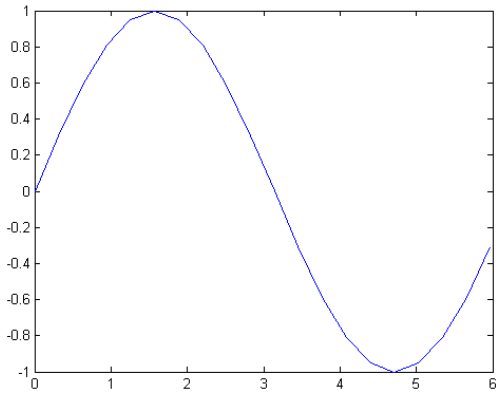
- 如果现已知期望值，比如测水的沸点，那么测量10次，测量值和期望值之间是独立的（期望值不依测量值而改变，温度计坏了也好，看反了也好，总之期望值应该是100度），那么$E[(X-期望)^2]$，就有10个自由度。所以叫做有偏估计，测量结果偏于那个“已知的期望值”。
- 如果现在往水里撒把盐，水的沸点未知了，该怎么办？只能以样本的平均值，来代替原先那个期望100度。 同样的过程，但原先的（X-期望），被（X-均值）所代替。（X_i-均值）的方差，它不再等于X_i的方差， 而是有一个协方差，因为均值中，有一项X_i/n是和X_i相关的，这就是那个“偏”的由来。自由度就只剩下9了。

【辨析*】样本方差与总体方差（启发式思考）

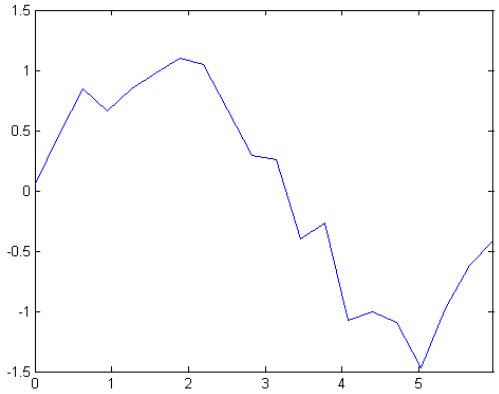
- 对于样本方差来说，假如从总体中只取一个样本，即**n=1**，那么样本方差公式的分子分母都为**0**——方差完全不确定。这很好理解，因为样本方差是用来估计总体中个体之间的变化大小，只拿到一个个体，当然完全看不出变化大小。反之，如果公式的分母不是**n-1**而是**n**，计算出的方差就是**0**——这是不合理的，因为不能只看到一个个体就断定总体的个体之间变化大小为**0**。
- 对于总体方差来说，假如总体中只有一个个体，即**N=1**，那么方差，即个体的变化，当然是**0**。如果分母是**N-1**，总体方差为**0/0**，即不确定，却是不合理的——总体方差不存在不确定的情况。
- 从这个启发式思考可以从直觉上看出，样本方差分母为**n-1**，总体方差分母为**n**，才是合理的。

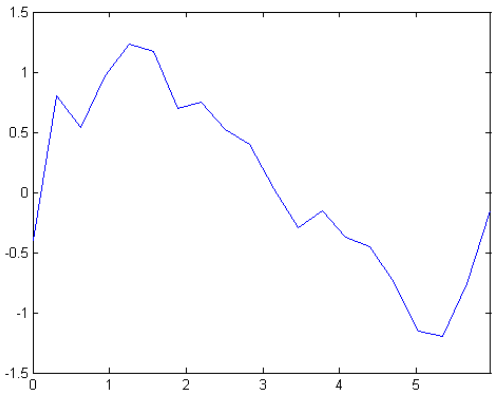
样本容量对统计量的影响

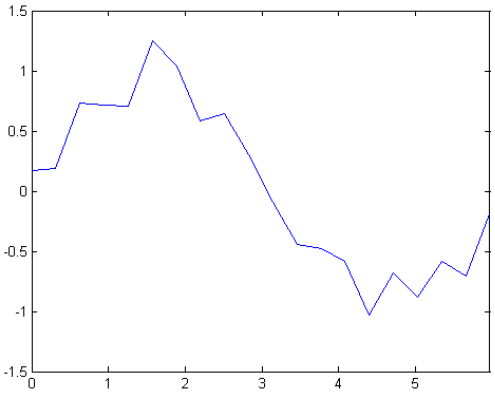
- 理想信号

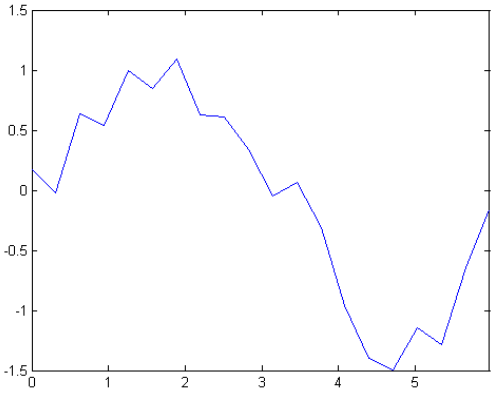


- 测量信号

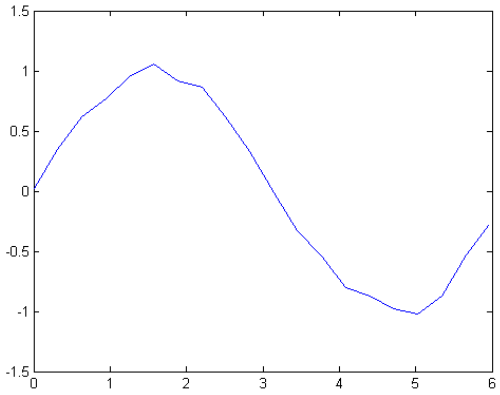




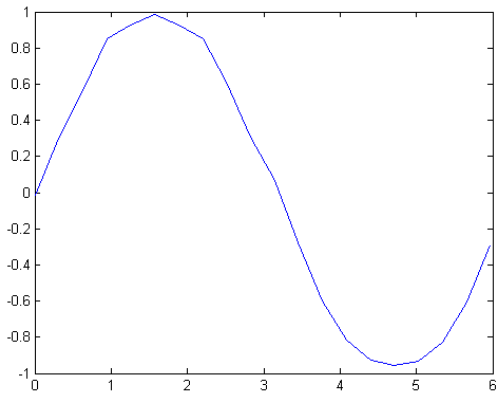




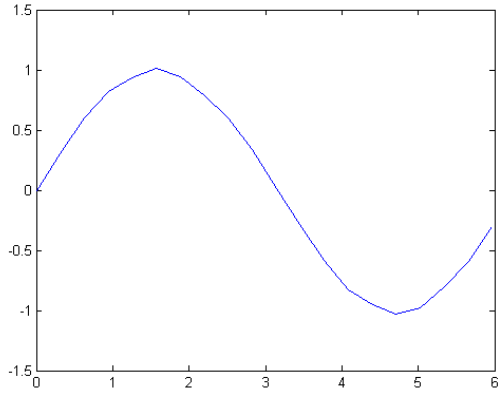
- 连续测量**20**次后求样本均值



- 连续测量**50**次后求样本均值



- 连续测量**200**次后求样本均值



【演示】

314抽样分布的计算机实验1.xls

二、经验分布函数

定义

设$X_1, X_2, \cdots, X_n$是总体F的一个样本，用$s(x), -\infty < x < \infty$表示$X_1, X_2, \cdots, X_n$中不大于$x$的随机变量的个数。

经验分布函数$F_n(x)$为

$$F_n(x) = \frac{1}{n} s(x) \quad -\infty < x < \infty$$

对于一个样本，经验分布函数 $F_n(X)$的观察值（仍以 $F_n(X)$表示）是很容易得到的。

设 (3, 1, 7, 3, 3, 1) 是来自总体 X 的一个样本值，

则**经验分布函数**为

$$F_6(x)=\begin{cases}0, & x<1\\ \frac{2}{6}, & 1\le x<3\\ \frac{5}{6}, & 3\le x<7\\ 1, & x\ge 7\end{cases}$$

$$S(x)=\begin{cases}0, & x<1\\ 2, & 1\le x<3\\ 5, & 3\le x<7\\ 6, & x\ge 7\end{cases}$$

一般，设$x_1, x_2, \cdots, x_n$是总体的一个容量为 n 的样本值.

将它们按大小次序排列如下：$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$

则经验分布函数$F_n(x)$的观察值为

$$F_n(x)=\begin{cases} 0, & \text{若 } x < x_{(1)} \\ \dfrac{k}{n}, & \text{若 } x_{(k)} \le x < x_{(k+1)}, (k=1,2,\cdots,n-1) \\ 1, & \text{若 } x \ge x_{(n)} \end{cases}$$

计算经验分布函数观察值的步骤：

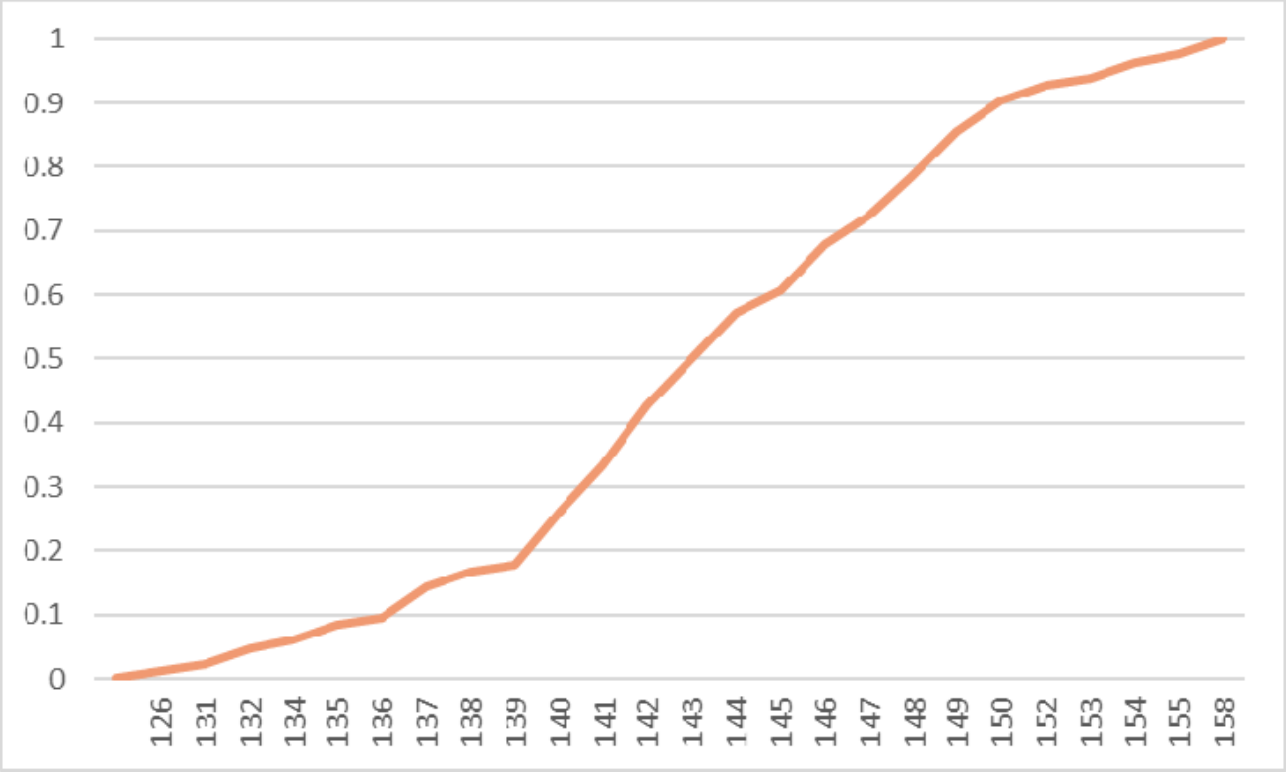
1. **对样本数据从小到大进行排序，合并相同数据，并统计频数；**
2. **用频数除以总数计算频率值；**
3. **计算累积频率。**

【演示】经验分布函数的图形绘制.xls

例 下面给出了84个伊特拉斯坎（Etruscan）人男子的头颅的最大宽度（mm）。

141	148	132	138	154	142	150	146	155	158	150	140
147	148	144	150	149	145	149	158	143	141	144	144
126	140	144	142	141	140	145	135	147	146	141	136
140	146	142	137	148	154	137	139	143	140	131	143
141	149	148	135	148	152	143	144	141	143	147	146
150	132	142	142	143	153	149	146	149	138	142	149
142	137	134	144	146	147	140	142	140	137	152	145

数据	频数	频率	累积频率	区间
			0	x<128
126	1	0.011905	0.011905	126≤x<131
131	1	0.011905	0.02381	131≤x<132
132	2	0.02381	0.047619	132≤x<134
134	1	0.011905	0.059524	134≤x<135
135	2	0.02381	0.083333	135≤x<136
136	1	0.011905	0.095238	136≤x<137
137	4	0.047619	0.142857	137≤x<138
138	2	0.02381	0.166667	138≤x<139
139	1	0.011905	0.178571	139≤x<140
140	7	0.083333	0.261905	140≤x<141
141	6	0.071429	0.333333	141≤x<142
142	8	0.095238	0.428571	142≤x<143
143	6	0.071429	0.5	143≤x<144
144	6	0.071429	0.571429	144≤x<145
145	3	0.035714	0.607143	145≤x<146
146	6	0.071429	0.678571	146≤x<147
147	4	0.047619	0.72619	147≤x<148
148	5	0.059524	0.785714	148≤x<149
149	6	0.071429	0.857143	149≤x<150
150	4	0.047619	0.904762	150≤x<152
152	2	0.02381	0.928571	152≤x<153
153	1	0.011905	0.940476	153≤x<154
154	2	0.02381	0.964286	154≤x<155
155	1	0.011905	0.97619	155≤x<158
158	2	0.02381	1	158≤x


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
126
131
132
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
152
153
154
155
158

经验分布函数图

【格里汶科（Ghivenko）定理】

对于任一实数 x，当 $n \to \infty$ 时，$F_n(x)$ 以概率 1 一致收敛于分布函数 $F(x)$，即

$$P\left\{\lim_{n\to\infty} \sup_{-\infty<x<+\infty} \left|F_n(x)-F(x)\right| = 0\right\} = 1.$$

对于任一实数 x 当 n 充分大 时，经验分布函数的任一个观察值 $F_n(x)$ 与总体分布函数 $F(x)$ 只有微小的差别，从而在实际上可当作 $F(x)$ 来 使用.

（1） 对每一组样本观测值 $(x_1, x_2, \cdots, x_n)$，经验分布函数$F_n(x)$是一个分布函数。

（2） 对于固定的 x （$-\infty < x < \infty$），经验分布函数$F_n(x)$是样本$(X_1, X_2, \cdots, X_n)$的函数，从而是统计量（随机变量）。

（3）当样本容量 n 足够大时，总体的经验分布函数是它的理论分布函数很好的近似。