随机数学(二)

葛余博　编著

清华大学数学系

一九九九年九月

主要符号表

ps ： 概率空间　　df ： 分布函数

pdf ： 概率密度函数　　iid ： 独立同分布

rv ： 随机变量　　$r\vec{v}$ ： 随机向量

$B(n,p)$ ：二项分布　　$Ge(p)$ ：几何分布

$P(\lambda)$ ：Poisson （泊松）分布

$U_{(a,b)}$ ：均匀分布　　$Ex(\lambda)$ ：指数分布

$N(\mu,\sigma^2)$：正态分布

$\Gamma(r,\lambda)$ ：Gamma（伽玛）分布

μ_k ：k 阶矩（总体）　　$\mu(=\mu_1)$ ：数学期望（总体）　　σ^2 ：方差（总体）

M_k ：样本 k 阶矩　　$\bar{X}(=M_1)$：样本均值　　S^2 ：样本方差

S_n^2 ：样本二阶中心矩

z_α 、 $z_{1-\alpha}$ ：标准正态分布 $N(0,1)$ 的百分位点

$t_\alpha(n)$ 、 $t_{1-\alpha}(n)$ ：$t(n)$分布的百分位点

$\chi_\alpha^2(n)$ 、 $\chi_{1-\alpha}^2(n)$ ： $\chi^2(n)$分布的百分位点

$F_\alpha(n,m)$ 、 $F_{1-\alpha}(n,m)$：$F(n,m)$分布的百分位点

目　录

第五章　数理统计的基本概念

引言

概率论（基础）的任务是对随机现象的研究，在数学上建立概率的公理化体系，引入基本的概念、揭示常见各类随机现象的规律性，总结为基本的随机模型和分布律，并研究它们的性质及数字特征. 对大量随机因素综合影响的结果，以极限定理为内容作了介绍. 这样对随机现象的研究，我们已经有了基本的概念、思想方法和工具. 但当我们真地动手去研究并解决一个实际问题时， 我们会立即遇到下面问题：

1. 这个随机现象可以用什么样的分布律来刻划，这种分布律的选用合理吗？
2. 所选用的这一分布律的参数是多少？如何估计和确定这些参数？

我们对要研究并解决的这个实际问题往往所知甚少，这样我们只能求助于观测，合理地取得一些数据，据此作出统计上的推断，回答上述问题，增进对这一实际问题里的随机现象的了解与把握，从而着手去解决问题. 而这，就是数理统计的基本且主要的任务. 更准确地说，数理统计的主要的内容是：

1. 试验的设计和研究，即研究如何更合理、更有效地抽取样本，从而获得观测数据和资料的方法.
2. 统计推断，即如何利用一定的数据资料，对所关心的问题，作出尽可能精确且可靠的统计结论：

1） 估计——从局部观测资料的统计特征，推断所观测对象的总体的特征，包括总体分布与数字特征；

2）假设检验——依据抽样数据资料，对总体的某种假设作检验，从而决定对此假定是拒绝抑或接受.

§5.1 总体和样本

本节介绍两个基本概念：总体和样本，并讨论它们间的关系.

一、总体和样本的概念

总体：研究对象的全体，例如某灯具厂生产的一批荧光灯全体. 如果我们关心的是这批荧光灯的使用寿命，那么总体也就是这批荧光灯的使用寿命的全体. 常以X 记总体.

个体：组成总体的每个基本单元，上例中的每支荧光灯，或它的寿命（例如，使用的小时数）；

样本：总体中抽取出来作观测的个体；

样本容量：抽取的个体的数目.

也称总体为**母体**，样本为**子样**，而样本容量也叫作**样本大小**.

假定从该厂这批荧光灯中随机地抽取 5 支荧光灯，依序编号后作实际的使用寿命试验，得到如下寿命（小时）数据：725，520，683，992，742. 一般地记为 x_1，x_2，…，x_5,称为观察值（或观测值）. 如又随机抽取 5 支，可得另一组观察值 x_1'，x_2'，…，x_5'. 再抽 5 支，又有观察值 x_1''，x_2''，…，x_5''，如此可继续抽取. 一般地，各组观察值是彼此不同的. 并且，每组中的第一支荧光灯的寿命 x_1，x_1'，x_1''，…，也是彼此不同的. 这样，所抽取的第一支荧光灯的寿命应该是一个 *rv* 随机变量），记为 X_1. 同样，第二支荧光灯的寿命是 *rv*X_2 ，如此我们得到一组 *rv*，X_1，X_2，…，X_5 ，称为**大小为 5 的样本**. 易 5 为一般的 n，则有**大小（容量）为 n 的样本** X_1，X_2，…，X_n 。称 x_1，x_2，…，x_n 为**样本观察值，或样本的一个现实**.

上面这样抽取的样本，如能切实保证其随机性，那么 X_1，X_2，…，X_n 应该是彼此独立的，且能反映总体的随机规律性，即所有样本彼此独立且与总体同分布. 这样的样本，我们称之为**简单样本**. 这种抽样方法，叫**简单抽样。**注意，在有限总体（如该批产品数量十分有限）中，各观察结果可能不独立. 除此之外的抽样方法，还有分层抽样，序贯抽样等等，考察一个Ψ大的系统或设备的可靠性时，为节省时间和经费，常对部件作分层或序贯抽样. **本书只讨论简单抽样，因此说到样本，均指简单样本**. 对其它抽样有需要或有兴趣的读者，请看有关试验设计方面的书[7].

通过抽样观察，我们对要解决的且又所知不多的随机问题，取得了一批样本数据，从而有了大小为 n 的样本，n 个 *iid*（独立同分布）的 *rv*. 基于概率论基础提供的理论和方法，我们来看看样本能够“透露”出总体一些什么信息. 首先，我们来考察样本与总体的关系，这也包括由样本构造的简单方便又有明显概率意义的样本函数 $g(X_1, X_2, \cdots, X_n)$，并研究它们的包括分布和矩在内的随机规律性，以及与总体在随机规律性上的关系.

二、样本的数字特征与分布

最简单又方便的样本函数 $g(X_1, X_2, \cdots, X_n)$，当然是 X_i 们的一次和二次（以及有时需要的更高次的）的线性函数. 由于（简单）样本是“平等”的，因此在选用的样本函数中，它们应该有相等的权系数. 而一次的等权的线性函数，就是样本的算术平均值，它明显地可以减少随机抽样带来的波动性. 二次等权的线性函数也还常作中心化. 它们都有明显概率意义. 这样，引入下列概念.

定义 1.1 设 $X_1, X_2, \ldots, X_n$ 为总体X 的大小为 n 的样本，分别称

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad 和 \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{1.1}$$

为**样本的均值及样本的方差**。而依次称

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \;、\; N_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k \quad 和 \quad S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{1.2}$$

为**样本的 k 阶矩、样本的 k 阶中心矩及样本的二阶中心矩。** □

请注意，$\overline{X} = M_1$ 而 $S_n^2 = N_2$. 这里没有把 S_n^2 叫样本的方差，其中的原因在下一章估计量的评选标准中说明. 另外还要注意，样本的均值、方差及 k 阶矩等等都是 rv，并且因 n 有限而总是存在的. 但总体的期望、方差及 k 阶矩等等是作为一个 rv 的相应的矩来定义的，因此它们却不一定存在. 并且即便存在，它们也都只是实数值，而非 rv. 引入总体X 的 k 阶矩和 k 阶中心矩的记号，

$$\mu_k = E\mathbf{X}^k = \int_{-\infty}^{\infty} x^k \, dF_{\mathbf{X}}(x) \quad 和$$

$$\sigma_k = E(\mathbf{X} - E\mathbf{X})^k = \int_{-\infty}^{\infty} (x-\mu)^k \, dF_{\mathbf{X}}(x)$$

其中 $\mu = \mu_1$ 是期望，并注意这里 σ_2 实际是 σ^2. 在样本的上述矩中代入观察值时，可建立相应的矩的观察值的概念，记号改为对应的小写，例如样本均值的观察值为 $\bar{x}$，而 k 阶矩观察值为 m_k.

性质 1 如果总体 k 阶矩存在，则样本的 k 阶矩的数学期望等于总体的 k 阶矩，而当 n 趋于无穷时，样本的 k 阶矩以概率收敛到总体的 k 阶矩，即

$$EM_k = \mu_k \;,\; M_k \xrightarrow[n\to\infty]{P} \mu_k \;.$$

实际上后一收敛性尚可强化为几乎处处收敛。

证明 由 $X_1, X_2, \ldots, X_n$ 相互独立且与 X 同分布，

$$EM_k = \frac{1}{n}\sum_{i=1}^{n} EX_i^k = \mu_k \;.$$

利用（强）大数定理可得关于收敛性的结论。 □

由简单样本的定义，易证样本的分布的下一性质。

性质 2 $F_{X_j}(x) = F_X(x)$ 且 $F_{(X_1,X_2,\cdots,X_n)}(x_1,x_2,\cdots,x_n) = \prod_{j=1}^{n} F_X(x_j)$

上面两个性质告诉我们，只要知道样本的分布与矩，就可以求出总体的分布与矩，问题似乎已经解决了。但细想一下，样本的分布和样本矩的期望如何去求，仍然是个问题，后者又回到需先要知道总体的分布：

$$EM_k = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{1}{n}\sum_{j=1}^{n} x_j^{\,k}\, dF_{(X_1,X_2,\cdots,X_n)}(x_1,x_2,\cdots,x_n).$$

只有性质 1 的关于强收敛的结论，能帮助我们从几乎每一组观察值求得总体的矩：

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i^{\,k} = \mu_k$$

三、顺序统计量与经验分布函数

要求总体的分布，还得另想办法。这当然还得从样本和样本的观察值出发，为此我们引入经验 *df*(分布函数)的概念。

定义 1.2 设 x_1，x_2，…，x_n 是总体 X 的容量为 n 的样本观察值，将它们以从小到大为序重新排列，记为 $x_1^{(n)} \le x_2^{(n)} \le \ldots \le x_n^{(n)}$，$\forall x \in R^1$ 令

$$Fn^*(x;\ x_1,x_2,\ldots,x_n) = \begin{cases} 0, & \text{如} x < x_1^{(n)} \\ k/n, & \text{如} x_k^{(n)} \le x < x_{k+1}^{(n)},\ k = 1,2,\ldots,n-1, \\ 1, & \text{如} x_n^{(n)} \le x \end{cases} \tag{1.3}$$

并称为由 x_1，x_2，…，x_n 决定的**经验 df**。也简记为 $F_n^*(x)$。 □

容易验证 $F_n^*(x)$ 满足一元 *df* 的三个条件（非降、右连续、以及 *df* 的边界极端性质，即 x 趋向 $\pm\infty$ 时函数值分别是 1 和 0），因此它确为 *df*。以本节五支荧光灯的寿命数据为例，可画出经验分布函数图形如图 1.1。

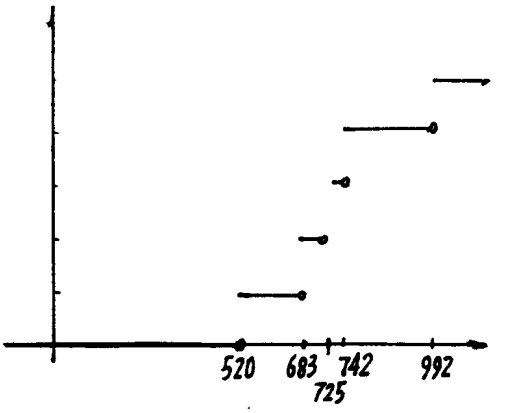

图 1-1 经验分布函数

我们知道 *rv* 实际是一个函数，因此类似高等数学中的函数复合，我们可建立如下概念。

定义 1.3 以 $X_1, X_2, \ldots, X_n$ 易 $x_1, x_2, \ldots, x_n$，我们从由样本观测值 $x_1, x_2, \ldots, x_n$ 决定的经验分布函数 $F_n^*(x;\ x_1, x_2, \ldots, x_n)$ 得到由样本决定的经验 *df*

$$F_n^*(x;\ X_1, X_2, \ldots, X_n),$$

简称为**样本经验 df**。 以从小到大为序重新排列的一组样本，称为**顺序统计量，专记为** $X_1^{(n)}, X_2^{(n)}, \cdots, X_n^{(n)}$。 □

样本经验 $df\ F_n^*(x;\ X_1, X_2, \ldots, X_n)$ 中，当 x 固定，它是 *rv*，固定ω而让 x 在实数域变化，则它是相对与某一组观察值的 *df*。

性质 3 经验 *df* 的 k 阶矩等于样本的 k 阶矩，即 $\int x^k dF_n^*(x) = M_k$.

证明 设 $x_1, x_2, \ldots, x_n$，为样本的一组观察值。 对实数轴 $(-\infty, \infty)$ 一个分割 $\{y_i\}$，当其足够细小时，在任意一个分段 $[x_j, x_{j+1})$ 上，只会有两种情形：要么它不含任一个观察值，此时样本观察值的经验 *df* $F_n^*(x;\ X_1, X_2, \ldots, X_n)$ 在此区间上的变差为 0，要么它只有一个观察值，例如是 x_j^*，此时 *df* 的变差为 1/*n*. 在求相应的 Stieljies 和式中的函数值就取为 x_j^* 。 从而由下面的推导可证得此性质。

$$\int x^k dF_n^*(x;\ x_1, x_2, \ldots, x_n) = \sum_{j=1}^n (x_j^*)^k [F_n^*(x_{j+1}^*) - F_n^*(x_j^*)]$$

$$= \sum_{j=1}^n (x_j^*)^k \cdot \frac{1}{n} = \frac{1}{n} \sum_{j=1}^n (x_j)^k = m_k$$

□

下面一个非常重要的定理保证，几乎由每一组观察值得到的经验分布函数，只要 n 足够大，都可作为总体 *df* 的近似，定理中一致收敛性和几乎处处收敛性，给了我们充分的自由. 从而由样本去找总体 *df* ，理论上有了一个完满的解决.

定理 1.1 (Genevinko) 设 $F^{(n)}(x, X) := F^{(n)}(x;\ X_1, X_2, \ldots, X_n)$ 是样本经验 *df*，则 $F^{(n)}(x, X)$ 以概率 1 对 x 一致地收敛到总体 *df*，即

$$P(\lim_{n\to\infty} \sup |F^{(n)}(x, X) - F_x(x)|) = 1$$

□

§5.2 直方图与概率纸

本节介绍如何依据抽样数据，从图形上初步估计总体的分布和矩.

设 $x_1, x_2, \cdots, x_n$ 是总体 X 的一个容量为 n 的样本观察值. 为了具体地叙述处理问题的

方法和步骤，我们通过下面的例题来进行.

例 2.1 从某厂生产的某种型号的铆钉中随机地抽取了 120 个，测得其直径的数据如下（单位：毫米）

13.40	13.39	13.52	13.37	13.62	13.48	13.40	13.35	13.44	13.54
13.47	13.29	13.53	13.50	13.32	13.51	13.45	13.48	13.34	13.26
13.46	13.57	13.58	13.80	13.14**	13.40	13.56	13.20	13.40	13.41
13.34	13.55	13.48	13.43	13.43	13.43	13.42	13.63	13.51	13.57
13.57	13.23	13.36	13.28	13.38	13.29	13.39	13.45	13.33	13.29
13.38	13.44	13.54	13.50	13.38	13.39	13.33	13.51	13.37	13.45
13.59	13.43	13.20	13.41	13.42	13.32	13.25	13.24	13.34	13.64
13.32	13.44	13.47	13.51	13.40	13.48	13.48	13.47	13.35	13.46
13.39	13.29	13.69**	13.44	13.28	13.49	13.40	13.31	13.52	13.51
13.40	13.52	13.48	13.29	13.46	13.56	13.44	13.50	13.38	13.46
13.60	13.31	13.50	13.35	13.28	13.53	13.48	13.30	13.55	13.62
13.58	13.62	13.51	13.42	13.48	13.45	13.32	13.43	13.31	13.38

标有**者，为最小值或最大值

试初步分析该厂生产的这种型号铆钉的直径X,服从什么分布？

分析　上面所列 120 个铆钉直径，是该厂生产的这种型号铆钉直径 X 的一个容量 $n=120$ 的样本观察值. 样本的容量相当大，应该能很好提供总体分布规律的信息.

初看起来给出的 120 个数是杂乱无章的，没有什么规律. 我们的任务是对这些数据进行科学的整理和归纳，揭示蕴藏在这批数据里的总体分布规律. 对于由试验或观察得来的数据，一般按下述步骤先进行整理，再画出直方图.

一、数据整理与直方图

1、数据整理的一般步骤

1) 找出样本的最小值和最大值，分别记为 $x_1^{(n)}$ 和 $x_n^{(n)}$，用以确定了样本的取值范围.

对于例 2.1，通过观察可以得到

$$x_1^{(n)} = \min_{1\le i\le 120} x_i = 13.14, \qquad x_n^{(n)} = \max_{1\le i\le 120} x_i = 13.69$$

一般地，当该厂生产的这种型号的铆钉足够多时，可以认为铆钉的直径可以取得包含[13.14, 13.69]在内的某个区间上的任何值. 即认为X是连续型 *rv*.

2). 确定分组组数 *K*.

对于连续型 *rv X*, 由于 P{*X*=*x*}=0, 所以只能考虑它落在某些区间上的概率. 对应地，由“频率的稳定性”，若 *n* 足够大时,可以用样本落在这些区间上的频率来近似总体落在这些区间上的概率. 因此，一般把样本的变化范围[$x_1^{(n)}$, $x_n^{(n)}$]等分为 *K* 个小区间，从而把观测值 $x_1, x_2, \cdots, x_n$ 分成 *K* 组，然后来计算样本值落在各个区间上的频率.

分组的组数K可以根据样本的容量按下列经验规律选取.

当 $n \leqslant 20$ 时, 取 $K=5\sim6$;

当 $n=20\sim60$ 时, 取 $K=6\sim8$;

当 $n=60\sim100$ 时, 取 $K=8\sim10$;

当 $n=100\sim500$ 时,取 $K=10\sim20$.

Sturges（司图格司）建议按以下公式选取 *K*:

$$K = 1 + 3.3\log_{10} n \tag{2.1}$$

按这个公式，$n=80$ 时 $K=7.28$, 选取分组数 *K* 为 7 或 8 ； $n=100$ 时 $K=7.6$, 选 *K* 为 8. 从实际应用情况看，此公式的分组数一般偏小. *n* 和 *K* 的这种关系不是完全定死的，允许有一定的灵活性. 一般说，每个组中至少应该有 1 个样本观测值，而以有 5 个以上观测值为好. 现在对此例，$n=120$, 取 $K=10$.

3). 定组距 $\Delta x = (b-a)/K$

一般取等间隔作组距 Δx，现在

$$\Delta x = \frac{13.69-13.14}{10} = \frac{0.55}{10} = 0.055$$

但我们取$\Delta x = 0.06$，一方面这可使计算简化，另一方面，考虑到样本的随机性，*X* 的实际取值既可能比 $x_1^{(n)}$ 更小点也可能比 $x_n^{(n)}$ 更大点. 取 $\Delta x = 0.06$ 时 $0.06\times10 = 0.60$, 比 0.55 大 0.05. 把多出的 0.05 分在区间的两端（可以均分，也可以某一端多分些），使分组的总范围[*a*, *b*]包含[$x_1^{(n)}$, $x_n^{(n)}$]. 现在取[*a*, *b*]=[$x_1^{(n)}$-0.01, $x_n^{(n)}$+0.04] = [13.13, 13.73], 并将[*a*, *b*]分成 10 个小区间，这些小区间是

[13.13, 13.19], (13.19, 13.25], (13.25, 13.31], ⋯ , (13.67, 13.73].

4). 列样本观察值的分组频率分布表 例 2.1 的分组频率分布表如表 2.1 所示.

表 2.1

组限 $(e_{i-1}, e_i]$	组中值 $\bar{x}_i = \frac{e_i + e_{i-1}}{2}$	频数 m_i	累积频数	频率 $f_i = m_i / n$	累积频率 $F^*(x_i)$
13.13~13.19	13.16	1	1	0.0083	0.0083
13.19~13.25	13.22	5	6	0.0417	0.0500

13.25~13.31	13.28	13	19	0.1083	0.1583
13.31~13.37	13.34	14	33	0.1167	0.2750
13.37~13.43	13.40	27	60	0.2250	0.5000
13.43~13.49	13.46	25	85	0.2083	0.7083
13.49~13.55	13.52	19	104	0.1583	0.8667
13.55~13.61	13.58	10	114	0.0833	0.9500
13.61~13.67	13.64	5	119	0.0417	0.9917
13.67~13.73	13.70	1	120	0.0083	1.0000
$\sum$		120		1.0000	

表中的 e_i 是第 i 个小区间的右端点，或称第 i 组的组上限. $\tilde{x}_i$ 是第 i 个小区间的中值，

$$\tilde{x}_i = \frac{e_i + e_{i-1}}{2}$$

在分析计算中有时用到.

一般说这种表可以大致给出样本的分布规律. 从表 2.1 不难发现频数和频率均呈“两头小，中间大，两边近似对称”的样子. 样本的频数和频率具有这种特征时，总体可能服从正态分布. 也可能是在下一章介绍的 t-分布. 但对样本数大于 30 时，可以认为这两个分布没有什么差别. 从这类表的频数和频率的走势，常能发现的分布，还有指数分布、下一章介绍的χ^2-分布，以及离散型的二项分布、Poisson 分布、几何分布及负二项分布（Pascal 分布）等等、由本节“三”介绍的利用专用的概率纸，可以对粗估的分布作进一步的检查. 而在第七章还会用更可靠的方法对这种初步的估计进行检验.

2、直方图

直方图可以更直观更形象地描绘出频数和频率的特征，帮助我们认识总体的分布.

直方图有两种：频数直方图和频率直方图.

在横坐标上标出组限，纵坐标上标出各组的频数 m_i，在 x 轴的上方，分别画出以组距为底，对应频数 m_i 为高的矩形，这样做出的图形称为**频数直方图**. 若纵坐标上标出的是频率 f_i/组距Δx，在 x 轴上方分别画出以组距为底，以对应的频率 f_i/组距Δx 为高的矩形，则这种图形称为**频率直方图**. 在纵坐标轴上采用两种适当的刻度，当然也可以把这两种直方图画成一张图. 注意频率直方图中所有矩形的面积之和等于 1，

$$\sum_{i=1}^{K} \frac{f_i}{\Delta x} \cdot \Delta x = \sum_{i=1}^{K} f_i = 1$$

例 2.1 的直方图如图 2-1 所示.

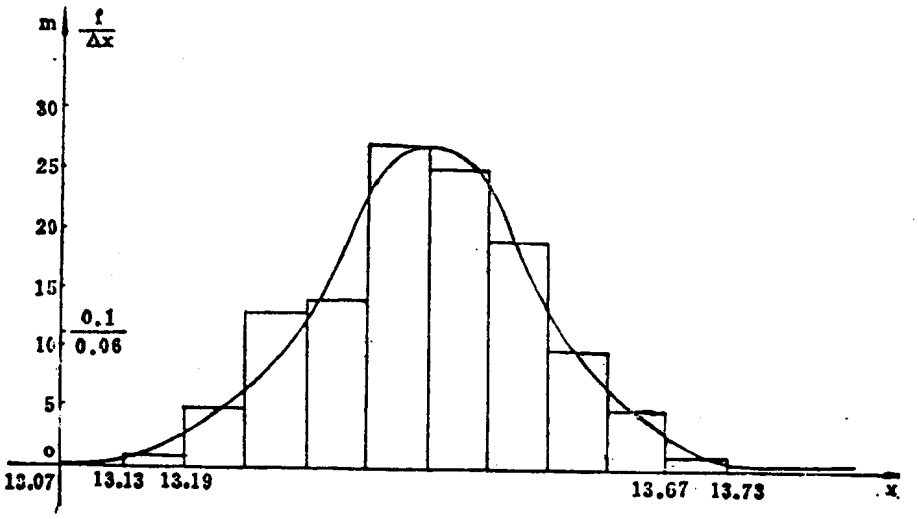

图 2-1 频率直方图

图 2-1 中直方图的特点是："两边低，中间高，单峰，左右近似对称". 直方图显示这一特点比表 2.1 表现得更为直观. 由于铆钉直径 X 是连续型 *rv*，因此，若加大样本容量，缩小组矩，直至 $n\to\infty$ 且每个矩形的宽趋于零时，频率直方图的上边缘将以光滑的曲线为极限. 这条光滑曲线就是总体的 *pdf*（概率密度函数）. 例 2.1 的这种近似光滑曲线在图 2-1 也已画出. 它的形状很象正态 *pdf* 曲线，这进一步说明总体很可能服从正态分布.

与几种常用分布接近的直方图走势见图 2-2.

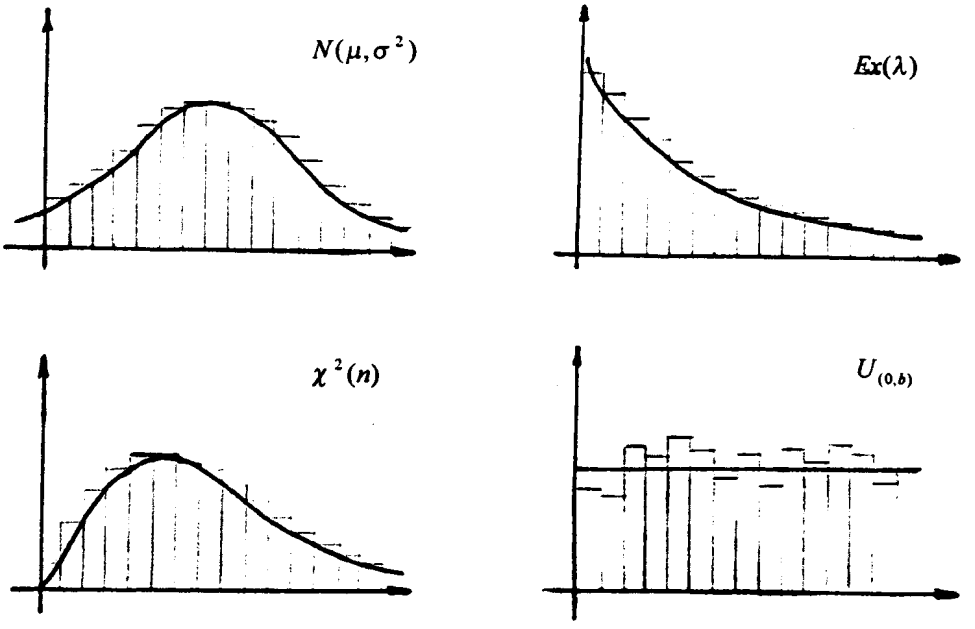

图 2-2 直方图轮廓

二、用概率纸检验总体的分布

直方图可以告诉我们总体的可能分布，但是不能肯定. 进一步确认的一个直观办法是概率纸法，更为科学的方法是将在第七章介绍的假设检验法.

每种概率纸都是一种特殊的专用坐标纸，可以用来检查总体的一种分布和进行相应分布的参数估算. 检查总体是否正态分布就用正态概率纸，检查总体是否指数分布就指数分布概率纸. 概率纸的种类很多，常用的有正态概率纸、单对数坐标纸和威布尔概率纸等. 其中尤以正态概率纸应用最广，由于概率纸的构造原理和使用方法类似，所有只介绍正态概率纸及其用法.

利用概率纸解决数理统计问题的缺点是准确性较低，但是它有直观，形象，步骤简单，计算量小等优点.

1、正态概率纸的构造

正态概率纸的横坐标轴是均匀刻度，纵坐标轴则是按正态分布的规律刻制，其具体方法如下.

如图 2-3 所示，先画一条直线，有均匀刻度作为 y' 尺，在此尺上每个刻度读数为 y' 的点的右方，按标准正态分布函数关系

$$\Phi(y') = y\,(\%) \qquad (2.2)$$

标出读数 y，作为 y 尺. 构成的坐标纸就是正态概率纸. 下面将说明，在正态概率纸中，标准正态分布是过点 (0, 50%) 斜率为 1 的一条直线，而且任何正态分布都是斜率为正数的直线.

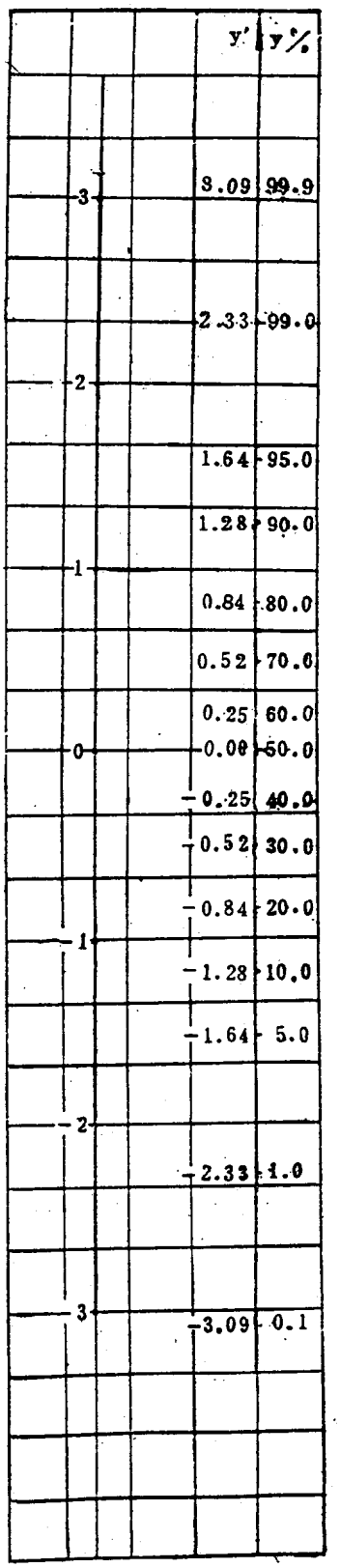

图 2-3 正态纸上的 y 尺于 y'尺

表 2.2

y(%)	y'	y(%)	y'
0.1	-3.09	60.0	0.25
1.0	-2.33	70.0	0.52
5,0	-1.64	80.0	0.84
10.0	-1.28	90.0	1.28
20.0	-0.84	95.0	1.64
30.0	-0.52	99.0	2.33
40.0	-0.25	99.9	3.09
50.0	-0.00		

2、正态概率纸的应用

〖判断总体是否服从正态分布〗

(1) 原理 如图 2-4 所示, 我们按着正态概率纸的构造原理,同时建立起了 xoy 和 xoy'(未画出)两个坐标系, 并且使得 y 轴和 y' 轴重合(两轴上的刻度不同). 由于 $y(\%)=\Phi(x)$是单调函数, 对于任意给定的 x, 可以通过

$$y(\%)=\Phi(x)$$

求得一个确定的 $y(\%)$, 对于这个 $y(\%)$, 可以通过

$$\Phi(y')=y(\%)$$

求得一个确定的 y'; 反之亦然. 可知 y' 和 x 一一对应, 且

$$y'=x.$$

这是 xoy' 坐标系中过原点, 斜率为 1 的一条直线, 因为 $\Phi(y')=y(\%)$ 所对应的 $y(\%)$ 和 y' 在纵坐标轴上是用同一点表示的, 所以$(x, y(\%))$点和 (x, y')点在坐标系中是同一点. x 和 y' 之间的关系是直线关系, x 和 $y(\%)$之间的关系也是直线关系, 即在 xoy 坐标系中 $y(\%)=\Phi(x)$是过点 (0, 50(%)), 斜率为 1 的直线.

同理, 在上述重合的两个坐标系中, 对于任意的正态分布 $N(\mu,\sigma^2)$的分布函数

$$y=F(x)=\Phi\left(\frac{x-\mu}{\sigma}\right)=\Phi(y')$$

并且, 由于 $y=F(x)$的单调性, 有

$$y'=\frac{x-\mu}{\sigma}$$

即在 xoy' 坐标系中，这是一条过点(μ,0)斜率为 $1/\sigma$ 的直线. 由于在纵坐标轴上，$y\%$和 y' 是同一点，所以在坐标平面上 $(x, y\%)$和 (x, y')是同一点. 从而在 xoy 坐标系中 $y\% = F(x)$是过点(0, 50%)，斜率为 $1/\sigma > 0$ 的一条直线. 上面讲的道理反过来也是对的. 即

若某个函数 $y\% = F(x)$，在正态概率纸上的图形是一条斜率大于零的直线，则 $F(x)$一定是正态分布函数.

(2) 判别方法

由以上所讲道理和 $\lim\limits_{n\to\infty} F_n^*(x) = F(x)$，其中 $F_n^*(x_i)$ 为累计频率函数，故 n 充分大时，样本的 df(分布函数) 近似总体 df，$F_n^*(x_i) \approx F(x)$，所以当总体为正态时，点 $(x_i, F_n^*(x_i))$，$i = 1, 2, \cdots, K$ 在正态概率纸上应当近似地在一条直线上. 反之，当 $(x_i, F_n^*(x_i))$在正态概率纸上近似呈一条斜率大于零的直线时，其总体服从正态分布.

例 2.2 查验例 2.1 中总体服从的分布，并估计分布的参数. .

对于例2.1，把表2.1中的诸点(x_i'，$F_n^*(x_i')$)标在图2-4中，可见它们近似呈一条直线，所以可认为该厂生产的这种铆钉的直径服从正态分布.

注意，由于样本有随机性，诸点(x_i'，$F_n^*(x_i')$)不可能全都在一条直线上，而是有偏离. 但是偏差不能太大，太大就应该认为总体不服从正态分布了. 画直线时，应该使 $F_n^*(x_i')$ 的30%～70%个对应的点到此直线距离尽量小，两头的，即10%以下及90%以上范围内对应的点到直线的距离允许大些，同时使直线两侧的点数大致相等. 各点单独地或成组地交错分布在直线的两侧.

〖求期望值 μ 和方差 σ^2 的估计值〗

当已经判明可以认为总体服从正态分布，而且已经配出了 $y\%=F(x)$在正态概率纸上的直线时，我们可从正态概率纸上估计 μ 和 σ^2，其方法如下.

因为 $\mu = E(X)$ 和 $F(\mu) = 0.5 = 50\%$，所以过正态概率纸上纵坐标为 50%的点作平行于 x 轴的直线，它和所配直线的交点的横坐标即 μ 的估计值. 对于例 2.2，μ 的估计值的求法如图 2-4 所示.

$$\mu \approx \hat{\mu} = 13.43$$

因为 $F(\mu - \sigma) = 15.9\%$*，故过正态概率纸上纵坐标为 15.9%的点作平行于 x 轴的直线，它和所配直线的交点的横坐标近似地等于 $\mu - \sigma$. 再利用上面求出的 $\hat{\mu}$ 就可以求出 σ 的估计值 $\hat{\sigma}$. 对于例 2.2，如图 2-4 所示，可以用此法求得

$$\hat{\mu} - \hat{\sigma} = 13.33$$

所以

$$\hat{\sigma} = 13.43 - 13.33 = 0.10$$

图 2-4 在本章末、习题前

〖用样本的平均值和均方差作为 μ 和 σ 的估计值〗

用正态概率纸求 μ 和 σ 的估计值，虽然方法简单而且不用作繁重的计算，但是这样求出的估计值有时准确度不高. 较好的方法是用样本观测值的平均值 $\bar{x}$ 作为 μ 的估计值，用 s 作为 σ 的估计值（参看 §6.1）. 对于例 2.1

$$\mu \approx \hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{120}\sum_{i=1}^{120} x_i$$

$$= \frac{1}{120}(13.39 + 13.34 + \cdots\cdots + 13.38) = 13.430 \approx 13.43$$

即取 $\mu \approx 13.43$. 这和用正态概率纸，求得的值一样.

$$\sigma \approx \hat{\sigma} = s_n$$

$$= \left(\frac{1}{120}[(13.39 - 13.43)^2 + (13.34 - 13.43)^2 + \cdots\cdots + (13.38 - 13.43)^2]\right)^{1/2}$$

$$= 0.107 \approx 0.11$$

〖求总体概率密度的近似表示式〗

我们已经判明总体服从正态分布，并且求出了 $\hat{\mu}$ 和 $\hat{\sigma}$ 后，便可以具体的写出总体 *pdf* 的近似估计式

$$f(x) \approx \frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}} \qquad x \in (-\infty, +\infty)$$

对于例 2.2, 有

$$f(x) = \frac{1}{0.11\sqrt{2\pi}} e^{-\frac{(x-13.43)^2}{2(0.11)^2}} \qquad x \in (-\infty, +\infty)$$

从上面的叙述和例 2.2, 可以看出：经过对原始数据(样本观察值)的整理、画直方图，可以粗估总体的分布类型，可以利用概率纸进一步判明它的分布. 当由此认定总体服从的分布时，还可以利用概率纸估计总体的数学期望值 μ 和方差 σ^2. 但是这种方法准确性不高，我们不能确切的说明什么叫“所有的点 $(x_i', F_n^*(x_i'))$，$i = 1, 2, \cdots, n$，近似地在一条直线上”. 由图上读出的 $\hat{\mu}$ 和 $\hat{\sigma}$ 也常常因人而异，因此有必要对问题作进一步研究. 这是后面两章的重要内容.

§5.3 抽样分布与统计量

从抽样数据对总体的分布或参数作估计和推断，以及从抽样数据对总体或其参数的某个假定作检验和推断，是数理统计的几项主要任务. 我们作统计推断的两个依据，一是抽样数据，或者抽象成 *rv*（随机变量）——（简单）样本，再一个是概率论提供的概率规律. 很自然地，我们要进一步研究从抽样样本去构造一些样本函数，看看它们是什么分布，依此设法去建立估计和检验的理论根据和方法. 这就是说，从总体X 独立抽取的大小（容量）为 n 的简单样本 $X_1, X_2, \cdots, X_n$ 已知，构造它们的简单方便又有明显概率意义的样本函数 $g(X_1, X_2, \cdots, X_n)$，并研究它们的包括分布和矩在内的随机规律性，作为对总体作统计推断的理论根据.

最简单方便又有明显概率意义的样本函数 $g(X_1, X_2, \cdots, X_n)$，当然是§5.1 介绍的一次的样本均值和二次的样本二阶中心矩，或者样本方差.

由于中心极限定理（§4.3），正态分布在概率论中有特殊的重要意义. 我们先来考察正态总体的常用的样本函数.

一、正态总体常用的样本函数

设总体X～ $N(\mu, \sigma^2)$. 则

1、样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$，从而 $Z := \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

证明 因为 $X_1, X_2, \cdots, X_n$ 为 *iid* $\sim N(\mu, \sigma^2)$，故 $(X_1, X_2, \cdots, X_n)$ 是 n 元正态分布，从而其分量的线性组合、特别 $\overline{X}$ 是一元正态分布. 又由期望的线性性及独立和的方差性质，有

$$E\overline{X} = \frac{1}{n}\sum_{i=1}^{n} EX_i = \mu \quad \text{和}$$

$$D\overline{X} = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

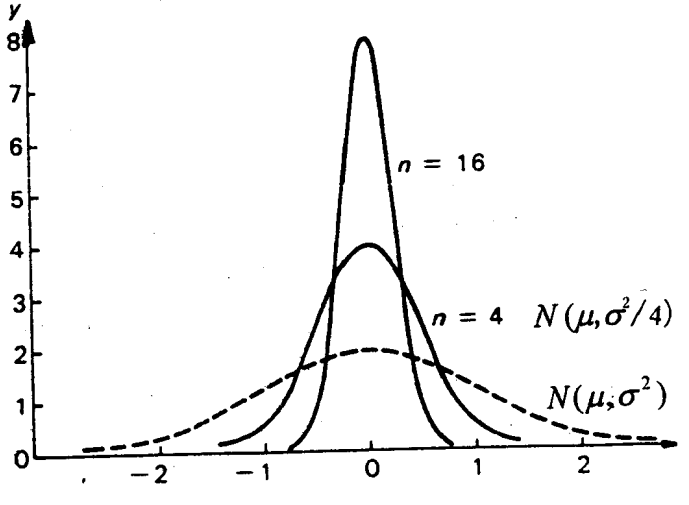

图 3-1 $\overline{X}$ 分布与正态分布的密度函数

2、 $K_n^2 := \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$ 的分布.