次品.在抽验的 4 只产品中,发现有3 只是次品.现在问:如果根据这个抽验的结果判断这批产品中次品是否超过8 只,该怎样回答呢?更确切地说,如果规定:180 只为一批,一个合格批中次品不能超过8 只,那么这批产品是合格的,还是不合格的?

因为如果这批产品是合格的,那么根据上面的计算,抽验的 4 只产品中次品数超过1 的可能性很小,如今抽得的4 只产品中竟有3 只次品,可能性很小的事件居然出现了,这使我们有理由怀疑这批产品是合格的,因而可以做出这批产品是不合格的判断.

自然,做出这种判断是会犯错误的,假如这批产品真的只有 3 件次品,抽验的 4 只中碰巧包含了这 3 只次品,这时判断这批产品不合格,犯了把合格批断定为不合格批的错误.不过犯这种错误的概率很小,不超过 1%.称这种错误为第一类错误.允许犯第一类错误的概率用 α 表示.

我们可以根据这一想法为厂家制定一个产品检验方案:用 H 表示需要做出判断的命题(在统计中称为原假论).

H: 由 180 只产品组成的一批产品中次品数不超过 8.

抽验方案:在一批产品中抽取 4 只做检验,用 ξ 表示其中次品数.如果 $\xi \leqslant 1$,则接受原假设 H(即认为这批产品是合格的);如果 $\xi > 1$,则拒绝原假设 H (即认为这批产品不合格,次品数大于 8).

犯第一类错误的概率,即在原假设 H 为真的条件下拒绝原假设的概率 $\alpha \leqslant 0.01$.

当然也可能犯第二类错误,即在原假设不真时接受原假设.这里我们不去深入探讨.

本例属于数理统计中的假设检验问题,在第五章中将对这类问题做进一步的讨论.

例 1.2.4 (捕鱼问题) 从一个养鱼池里捕出 1000 尾鱼,涂以红点后放回鱼池,隔了一段时间当鱼充分混合后,又捕出 1000尾鱼,发现其中有 100 尾鱼涂有红点,根据这些数据,应该对池

中鱼的总数做出怎样的估计 (假定两次捕鱼中间池中鱼的总数未变，而且两次捕鱼都是在整个鱼池中随机捕捉的)?

现在我们用数理统计中的最大似然估计法来估计池中鱼的总数，并且通过这个例子介绍最大似然估计的思想.

解 为使问题一般化，我们令

$N=$ 池中鱼的总数(未知);

$M=$ 第一次捕出的鱼数 (涂以红点);

$n=$ 第二次捕出的鱼数;

$k=$ 第二次捕出的 n 条鱼中涂有红点的鱼数;

$q_k(N)=$ 第二次捕出的鱼中恰有 k 尾涂有红点鱼的概率.

由例 1.2.1 知，如果 N 已知，则

$$q_k(N)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}.$$

如今 M,k,n 均为已知，N 为未知，我们只知道前后两次共捕到过 $M+n-k$ 尾鱼，N 应该不小于这个数. 在我们的例子中 $M+n-k=1900$，假定 $N=1900$，那么由 $M=n=1000$，$k=100$ 易知

$$q_k(N)=\frac{\binom{1000}{100}\binom{900}{900}}{\binom{1900}{1000}}=\frac{(1000!)^2}{100!900!}.$$

由斯特林公式可以证明这个概率的数量级将是 10^{-430}，因此可以认为，这个假定 $(N=1900)$ 是不太合理的，同样如果我们假定 N 很大，例如 $N=1000000$，经计算 $q_{100}(N)$ 也是非常非常小的，因而

这种假定也不太合理. 那么什么样的 N 才合理呢? 自然我们想到要求使 $q_k(N)$ 达到最大值的那个 N, 也就是要求那样的 N, 使得我们捕捉的 n 条鱼中恰有 k 条带有红点的概率最大. 对于一组特定的观察值 M,n,k, 我们把使 $q_k(N)$ 达到最大值的那个 N 记作 $\hat{N}$, 称之为 N 的**最大似然估计**. 为求 $\hat{N}$, 考虑比值

$$\frac{q_k(N)}{q_k(N-1)} = \frac{(N-M)(N-n)}{(N-M-n+k)N}.$$

由简单的计算知, 当 $Nk<Mn$ 时这个比值大于1, 当 $Nk>Mn$ 时这个比值小于1. 因而当 $N<\dfrac{Mn}{k}$ 时 $q_k(N)>q_k(N-1)$, 当 $N>\dfrac{Mn}{k}$ 时, $q_k(N)<q_k(N-1)$, 当 $N=\left[\dfrac{Mn}{k}\right]$ ($[x]$ 表示不超过 x 的最大整数)时, $q_k(N)$ 最大. 即 N 的最大似然估计

$$\hat{N}=\left[\frac{Mn}{k}\right].$$

在我们的例子中 $\hat{N}=\dfrac{1000\times 1000}{100}=10000$, 即池中鱼的总数估计在10000尾左右.

例 1.2.5 在例1.2.1中从 N 个球中一次取 n 个, 相当于依次取 n 个, 每次取出后不放回. 假如每次取出一个球后放回袋中, 试求 n 次抽取中恰有 k 次抽到红球的概率.

解 从 N 个球中每次取一个, 共取 n 次, 一切可能的取法数为 N^n, 每种取法的可能性是相同的, 在 n 次抽取中选定 k 次有 $\binom{n}{k}$ 种不同的选法, 而在指定的 k 次抽到红球另外 $n-k$ 次抽到黑球

故 $I=\sqrt{2\pi}$，从而得证所需结论.

由定理 2.5.9 知，正态分布可以看作二项分布的极限，这可以看作是正态分布的理论来源.

在实践上，正态分布是一个有广泛应用的概率分布. 相当广泛的一类随机现象可以用正态分布或近似地用正态分布来刻画(参见 §5.6). 下面举两个例子来说明如何从观察数据来判断所观察的指标是服从正态分布的.

例 3.3.10 表3－2 中的数据是一个自动机床生产的 100 个零件的口径 (单位：mm)，最小值是 10.93，最大值是 11.08，将区间 [10.925，11.085] (左端点略小于 10.93，右端点略大于 11.08) 等分成 8 个小区间 [10.925，10.945)，[10.945，10.965]，…，[11.065，11.085)，然后统计出这 100 个数据在每个小区间中的个数——频数，算出频数与数据总数的比——频率，制成表 3－3.

表 3－2

11.02	10.99	10.93	11.01	10.98	10.94	11.02
11.01	10.99	11.00	11.07	10.98	10.97	10.99
10.96	11.02	10.97	10.98	11.00	10.97	11.05
10.95	11.00	11.02	10.99	10.98	11.00	10.98
11.05	11.00	10.98	10.97	11.01	11.07	11.06
10.97	10.94	10.99	11.01	11.03	10.95	11.00
10.95	11.00	11.00	11.00	10.99	10.97	11.02
10.93	10.99	11.02	11.01	11.08	11.00	10.95
11.01	10.93	11.00	10.99	11.01	10.99	10.99
11.04	11.00	11.00	10.97	11.02	11.01	11.04
10.96	10.96	10.99	10.96	10.98	11.08	11.06
10.99	11.00	10.97	11.02	10.98	10.98	11.00
10.99	11.00	11.00	10.99	10.99	10.98	10.97
10.97	10.99	11.03	11.02	11.06	11.04	11.03
11.05	10.96					

表 3－3

序　号	区　间	频　数 n_i	频　率 $\frac{n_i}{100}$	概　率
1	[10.925，10.945]	5	0.05	0.04
2	[10.945，10.965]	9	0.09	0.11
3	[10.965，10.985]	20	0.20	0.19
4	[10.985，11.005]	33	0.33	0.24
5	[11.005，11.025]	17	0.17	0.21
6	[11.025，11.045]	6	0.06	0.13
7	[11.045，11.065]	6	0.06	0.05
8	[11.065，11.085]	4	0.04	0.02

表 3－2 中杂乱无章的数据经过分组制成表 3－3 后，就能看出一点规律性的东西．例如多数数据集中在区间 [10.965，11.025) 中，小于 10.945 或大于 11.065 的都不多．还可以用图形将各小区间内数据的频率表示出来，取一个直角坐标系，如图 3－6，在横坐标轴上标出各小区间的端点：$a_1=10.925$，$a_2=10.945$，…，$a_9=11.085$ 等等，然后以每个小区间为底，以相应的 (频率)/(小区间长) 为高做出 8 个矩形，这样，100 个数据中出现在某个小区间中的频率，就由这个小区间上矩形的面积表示出来了，这使我们能更直观地看到 100 个数据在 x 轴上是怎样分布的．通常称图 3－6 为**频率直方图**，或简称为**直方图**．

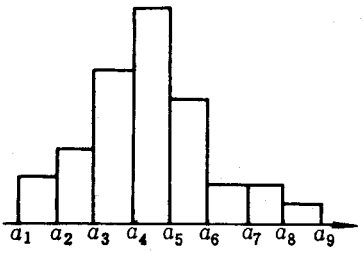

图 3－6

如果用 ξ 表示自动机床生产的一个零件的口径，那么表 3－2 中的 100 个数据就是取自 ξ 的 100 个值．ξ 在每个小区间内取值的频率，在一般情况下接近于 ξ 在那个区间中取值的概率．例如，ξ 在区间 [10.965，10.985] 中取值的频率一般接近于 $P(10.965 \leqslant \xi \leqslant 10.985)$，所以图 3－6 的直方图的外轮廓线一般接近于 ξ 的概率分布密度曲线．从直方图看，它有一个峰，比较对称，而且越远离中间值的数据出现的频率越小．所以我们有理由假设 ξ 的分布是正态分布或接近于正态分布．

为了进一步检查上述 100 个数据是否来自正态母体（分布）$\xi \sim N(\mu, \sigma^2)$，用 x_i，$i=1, 2, \cdots, 100$ 表示这 100 个数据．由 §5．2 知，可以用 $\overline{x}=\frac{1}{100}\sum_{i=1}^{100} x_i$，$s^2=\frac{1}{100}\sum_{i=1}^{100}(x_i-\overline{x})^2$ 作为 μ，σ^2 的估计值．经计算得知

$$\mu \approx \overline{x}=10.9972,\ \sigma \approx s=0.03268.$$

如果这100 个数据的确来自 $\xi \sim N(\mu, \sigma^2)$，那么表 3－3 中所列各区间上的频率应该近似于 $N(10.9972, (0.03268)^2)$ 的相应区间上的概率．例如 [10.965，10.985] 上的频率 0.20 应近似于

$$\int_{10.965}^{10.985} \frac{1}{\sqrt{2\pi}\times 0.03268}\exp\left\{-\frac{(x-10.9972)^2}{2(0.03268)^2}\right\}dx \approx 0.19155.$$

表 3－3 的最末一行列出了 $N(10.9972, (0.03268)^2)$ 的相应区间上的概率．比较这些频率与概率，可以看出它们还是相当接近的，所以可以认为这 100 个数据来自正态母体 $\xi \sim N(10.9972, (0.03268)^2)$．

例 3．3．11 表3－4 中的 157 个数据是 1979 年某地区

157 名学生高考中的化学成绩*),图 3－7 是这 157 个数据的频率直方图．若用 η 表示该地区一个考生的化学高考成绩，η 是一个随机变量．从直方图看，我们有理由假设 η 的分布是正态分布或接近于正态分布．

表 3－4

41.5	29.5	40.5	23.5	34.5	39.5	37
20.5	32.5	26	33	24	36	14.5
25	32	34.5	42	37	45	31
31	44.5	29.5	50.5	39	55	57
34	49.5	31.5	30	38	18	27
29.5	27	43.5	39	37.5	15	33
36.5	52.5	24	53	35	45.5	29.5
16	39.5	42.5	41.5	50.5	32.5	32.5
46.5	30.5	26	30.5	21.5	64	68
58	56	37.5	47.5	38	34.5	43.5
37	57	43	36.5	24	18.5	42.5
33.5	29.5	27.5	42	20	15.5	26
37.5	28.5	26.5	40.5	18	12.5	13
27.5	24.5	19.5	39	37.5	49.5	19
28.5	37.5	11.5	32.5	35	47.5	41.5
50.5	23.5	28	13.5	26	11	12
42.5	36.5	39.5	24	33.5	45.5	59.5
38	42.5	35.5	18	42.5	47.5	38.5
46.5	40	39.5	59.5	37.5	44.5	35
56.5	59.5	48.5	50.5	37.5	24	39.5
46.5	27.5	45	42.5	31	43.5	41.5
38.5	36	48.5	36	37.5	36.5	44.5
10.5	37	28				

*)这个材料取自叶佩华等编的《教育统计学》(人民教育出版社)．

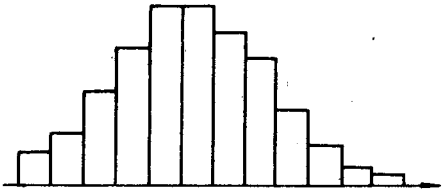

图 3-7

由于正态分布具有广泛的用途，而且我们经常需要计算概率 $P(a \leqslant \xi \leqslant b)$，但是由于不定积分

$$\int \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

不是初等函数，因而必须用数值计算来求积分的近似值．本书附表 2 给出了标准正态分布函数

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

的数值表．表中列出了对应于 $x \geqslant 0$ 的函数值．利用这个表可以计算一切正态随机变量取值于某一区间的概率的近似值．

例 3．3．12 若随机变量 $\xi \sim N(0, 1)$，则由附表 2 得

(1) $P(\xi \leqslant 2) = \Phi(2) = 0.97725$；

(2) $P(\xi > 2.34) = 1 - \Phi(2.34) = 1 - 0.99036 = 0.00964$；

(3) $P(\xi < -2) = P(\xi > 2) = 1 - \Phi(2)$

$= 1 - 0.97725 = 0.02275$，

这里利用了正态分布的密度函数关于原点的对称性；

(4) $P(-0.5 \leqslant \xi \leqslant 2.65) = P(\xi \leqslant 2.65) - P(\xi < -0.5)$