第3章

概率与分布

我们通常只能获得部分数据和信息，很少可以得到完全的信息。一个经验事实是大多数实验和观测不能完美地重现，无法复制的程度可能千差万别。然而，如果我们把数据视为来自于一个统计分布，那么这个观点将有助于我们对问题和统计方法的理解。这使得我们必须了解随机性和概率这两个统计学中的核心概念。

3.1 随机性和规律性

当不能预测一件事情的结果时，这件事就和随机性联系起来了。随机性和规律性是事物的正反面，是相对统一的。单个的事情可能具有随机性。例如，掷硬币时，我们不能确定硬币将正面朝上还是反面朝上。但是，当把大量的随机事件放在一起时，它们会表现出令人惊奇的规律性。

例 3.1 青原博士模拟扔硬币，了解随机性和规律性的相对统一性。

分别扔硬币 10 次、100 次、10 000 次可以发现，扔硬币的次数越多，正面和反面朝上的次数越接近，这个结果相当稳定。但是规律也表现出某种随机性。实际上如果重复扔 10 次硬币，我们会发现大部分时候并不能得到和上次观察一模一样的结果。这表明了统计的一个重要的本质特征。

```
> a = sample(c("H","T"),10,replace = T)
> table(a)
a
H T
7 3
> a = sample(c("H","T"),100,replace = T)
> table(a)
a
H  T
```

```
57 43
> a = sample(c("H","T"),10000,replace = T)
> table(a)
a
H    T
5007  4993
```

通过对看起来随机的现象进行统计分析，统计知识能够帮助我们把随机性归纳到可能的规律性中。统计从如何观察事物和事物本身如何真正发生这两个方面帮助我们理解随机性和规律性的重要性。因此，统计可以看作一项对随机性中的规律性的研究。

3.2　概率

因为涉及随机性，所以我们需要了解概率。**概率**(probability)是一个 0～1 之间的数，它告诉我们某一事件发生的机会有多大。概率为统计推断奠定了基础。在学习完第 4、5 章后会知道我们可能永远也不能确定两个数字的差异是否超出了随机性本身所能解释的范围，但是可以确定这种差异发生的概率是大还是小。因此，在很多情况下，仍然可以得出关于我们所处的这个世界的重要结论。

用 R 软件很容易计算概率。考虑无放回抽样的情形，如 sample(1:100,5)，得到一个给定数字作为第一个样本的概率是 1/100，第二个则是 1/99，以此类推。那么给定一个样本的概率就是 1/(100×99×98×97×96)。在 R 软件中使用 prod 函数计算一串数字的乘积：

```
> 1/prod(100:96)
[1] 1.106868e-10
```

要注意，这是一个在给定顺序下获得给定数字的概率。然而，我们更感兴趣的是正确猜出 5 个给定的数字集合的概率。因此，需要包含那些有同样的数字但是顺序不同的情形。注意，所有 5 个数字的集合必须有相同的概率。所以，我们所要做的就是计算从 100 个数字中选取 5 个数字的所有可能数①。在 R 软件中，可以用 choose()函数来计算这个数字，所以上述概率可以写成：

```
> 1/choose(100,5)
[1] 1.328241e-08
```

有其他方法得到同样的结果。很明显，由于每种情况的概率都是相同的，我们需要做的就是输出一共有多少种这样的情形。第一个数字有 5 种可能，对其中每一种情况的第二个数字又有 4 种可能，以此类推，从而可能的种数有 5×4×3×2×1，即 5!(5 的阶乘)。所以选取这 5 个数的概率为

① 组合数 $\binom{100}{5}=\frac{100!}{5!95!}=75\,287\,520$。

```
> prod (5:1) /prod (100:96)
[1] 1.328241e - 08
```

3.3 变量的分布

在扔硬币的活动中，结果总是随机出现的，因此把得到的结果称为**随机变量**(random variable)。在定义了一个感兴趣的随机变量之后，对实验结果的概率分析也就转化为对该随机变量各种可能取值的概率分析。随机变量取一切可能值或范围的概率或概率的规律称为**概率分布**(probability distribution)，简称分布。一般表示为累积分布函数 $F(x)=P(X\leqslant x)$，该函数描述的是对一个给定分布小于或等于 x 的分布的概率。例如，一个班级学生成绩可以很好地用均值 85、标准差 10 的正态分布(下面即将介绍该分布)来表示，那么，如果一个学生的成绩为 90 分，则这个班级中只有 30.85%的学生是这个分数或者比这个分数更高。R 软件的计算过程如下：

```
> 1 - pnorm(90,mean = 85,sd = 10)
[1] 0.3085375
```

在得不到真实的累积分布函数时，可以考虑经验累积分布函数，其定义为小于或等于 x 的数据占全部数据的比例。也就是说，如果 x 是第 k 小的观测值，那么小于或等于 x 的数据的比例为 k/n。我们可以作出一个经验累积分布函数图，结果如图 3-1 所示。①

```
> x = rnorm(100);n = length(x)
> plot(sort(x),(1:n)/n,type = "s",ylim = c(0,1))
```

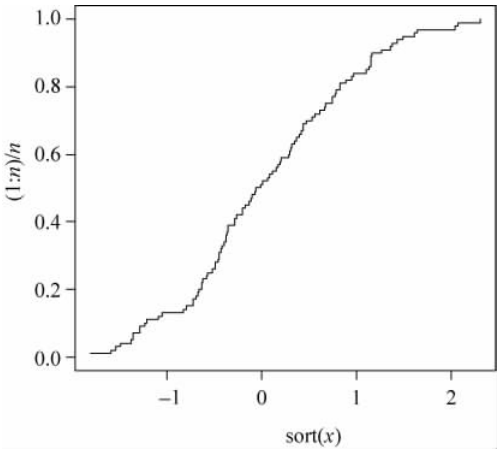

图 3-1 经验累积分布函数

① 或者使用下面的语句：

```
F100 <- ecdf(rnorm(100));plot(F100);plot(F100, verticals = TRUE, do.points = FALSE)
```

3.3.1　离散型分布

离散随机变量的所有可能取值是有限个或可列个数值，比如离散随机变量 X 取值为 $x_1, x_2, \cdots, x_n$，那么事件 $X=x_i$ 发生概率 $p(x_i)$ 的全体就是离散型概率分布，也称为概率分布列。随机变量 X 具有概率分布，可以用点概率 $p(x)=P(X=x)$ 或累积分布函数 $F(x)=P(X\leqslant x)$ 描述，还可以采用表格的形式展示概率分布列。离散型概率分布必须满足

$$\sum_{i=1}^{n} p(x_i) = 1, \quad 0 \leqslant p(x_i) \leqslant 1; \quad i = 1, 2, \cdots, n$$

在现实中有许多广泛应用的离散型概率分布，它们可以使用一般公式来表达，只要给定随机变量的任意一个取值，就可以直接计算出概率。

当观察一个独立重复二项实验时，通常对每次实验的成功或失败并不感兴趣，更感兴趣的是成功（或失败）的总数，此时就是**二项分布**（binomial distribution）。在上述二项分布情形下，分布可以用点概率来得到：

$$f(k) = \binom{n}{x} p^x (1-p)^{n-x}$$

这就是已知的二项分布，$\binom{n}{x}$ 称为二项系数。参数 p 是一次独立实验中成功的概率。

在 R 软件中使用下面的命令可得到如图 3-2 所示的 $n=100, p=0.33$ 时的二项分布图形。

```
> x = 0:100
> plot(x,dbinom(x,size = 100,prob = .33),type = "h") # type = "h"画出针形图
```

注意，在画出二项分布的点概率图时，除了 x，还要指定实验次数 n 和概率 p。以上画出的分布可以理解为投掷一个公平的骰子 100 次，出现 1 点或 2 点的次数。

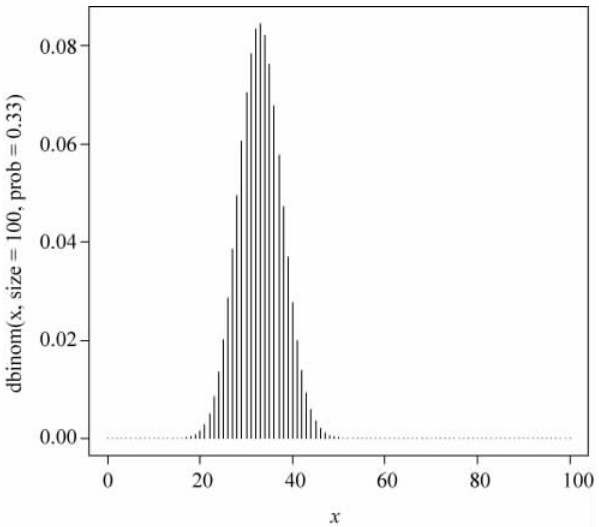

图 3-2　二项分布的点概率

几何分布(geometric distribution)类似二项分布,只是它记录的是第一次成功之前失败发生的次数。具体来说,几何分布的定义为:在 n 次独立重复二项实验中,实验 k 次才得到第一次成功的概率。即前 $k-1$ 次都失败,第 k 次成功的概率为

$$f(k) = p(1-p)^{k-1}$$

图 3-3 是 $p=0.33$ 的几何分布图形。

```
> x = 0:100
> plot(x,dgeom(x,prob = .33),type = "h")
```

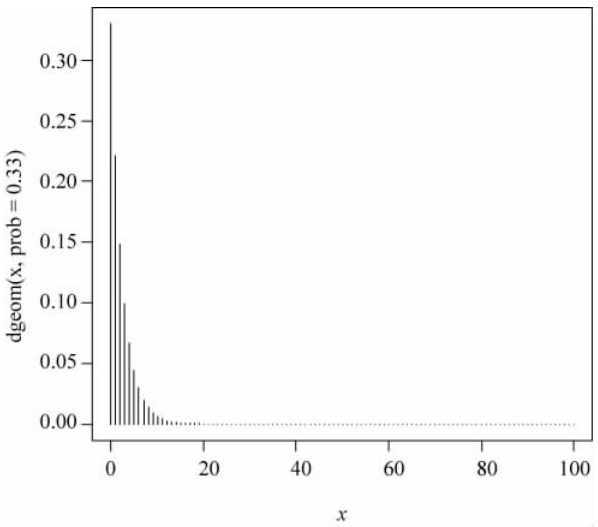

图 3-3 参数为 0.33 的几何分布点概率

泊松分布(poisson distribution)描述的是在特定区间内某种事件发生的次数,区间可以是时间、距离、面积或者体积。泊松分布的可能取值范围为所有非负整数。参数为 λ 的泊松分布变量的概率分布为($p(k)$表示泊松变量等于 k 的概率)

$$p(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0,1,2,\cdots$$

式中,λ 为特定区间内事件发生(成功)的均值;e 为常数 2.718 28…。泊松分布的均值和方差都等于 λ。与二项分布一样,泊松分布也是一个分布族,族中不同成员的区别在于事件出现次数的均值 λ 不一样。当事件发生的概率很小或者 n 值很大时,泊松分布是一种有限的二项分布。图 3-4 是 $\lambda=10$ 时的泊松分布图形。

```
> x = 0:100
> plot(x,dpois(x,10),type = "h")
```

3.3.2 连续型分布

有些数据来自于对连续尺度的测量,比如温度、浓度等。连续随机变量是可以取某一个或若干区间内任意数值的随机变量,因此是不可数的,这使得任何特定值的概率是零,所以

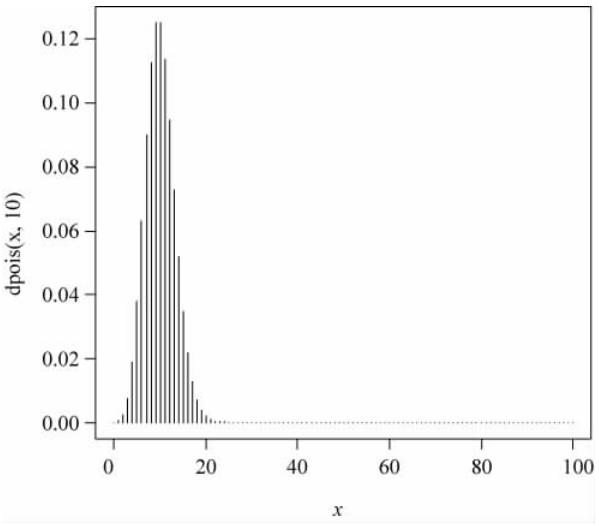

图 3-4　参数为 10 的泊松分布点概率

这里没有像离散型随机变量那样的点概率的说法，必须在某一区间内考虑相应的概率问题，因此取而代之的是密度的概念。它是指 x 的一个小邻域的无穷小概率除以区域的大小。此时，累积分布函数如下所示：

$$F(x)=\int_{-\infty}^{x} f(x)\mathrm{d}x \text{ ①}$$

常见的连续分布包括均匀分布和正态分布。**均匀分布**(uniform distribution)是最简单的连续型分布，记为 $U(a,b)$，表示定义在一个比如(a,b)这样特定的区间(通常为$[0,1]$)，均匀分布在该区间上有常数密度$\frac{1}{b-a}$。均匀分布的密度函数为

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant x\leqslant b\\ 0, & \text{其他}\end{cases}$$

在 R 软件中使用下面的命令可得到图 3-5 所示的是均匀分布的密度曲线图：

```
> x = runif(100)
> plot(x,dunif(x),type = "l")
```

在实际生活中，最常用的连续型变量的分布是**正态分布**(normal distribution)，又称**高斯分布**(Gaussian distribution)，它的密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

其中：μ 为均值；σ 为标准差。一般把正态分布记为 $N(\mu,\sigma^2)$。改变 μ 和 σ，密度曲线会平移和放缩。特别地，当 $\mu=0,\sigma=1$ 时，正态分布为标准正态分布 $N(0,1)$。正态分布密度曲线为

① $\int_{-\infty}^{x} f(x)\mathrm{d}x$ 表示所有小于或等于 x 的值出现的概率之和。

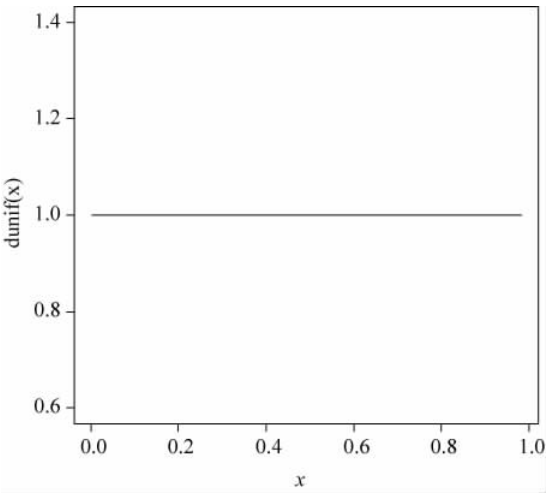

图 3-5　均匀分布 $U(0,1)$的密度曲线

钟型曲线。在 R 软件中使用下面的命令可得到如图 3-6 所示的标准正态分布密度曲线：①

```
> x = seq( - 4,4,0.1) # 函数 seq 用于产生等距数值,这里是从 - 4 到 4,步长为 0.1
> plot(x,dnorm(x),type = "l") # type = "l"表示函数在点与点之间画线而不只是画出点本身来
# dnorm 还有其他参数,即均值和标准差,通常默认为 0 和 1,即默认为标准正态分布
```

或者使用下面的作图方法，但它需要 y 值可以通过 x 的简单函数表达式表示出来。

```
> curve(dnorm(x),from = - 4,to = 4)
```

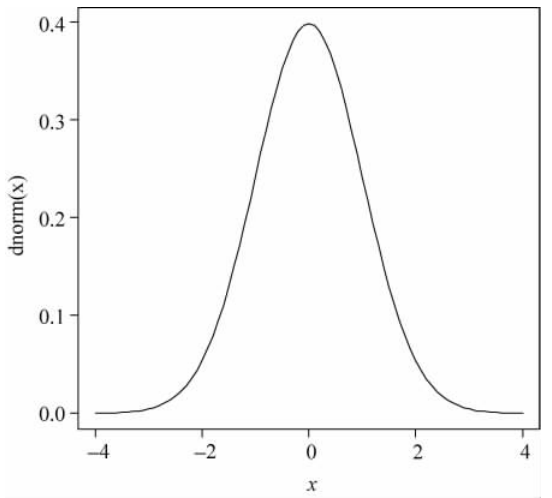

图 3-6　标准正态分布 $N(0,1)$的密度曲线

① R 软件中的密度函数以 d(density)开头。类似地，R 软件中的累积分布函数、分位数和随机数分别以 p(probability)、q(quantile)和 r(random)开头。

3.4　中心极限定理和抽样分布

3.4.1　中心极限定理

中心极限定理(central limit theorem,CLT)是概率论最重要的定理之一。中心极限定理的准确叙述如下：若给定样本量的所有样本来自任意总体，则样本均值的抽样分布近似服从正态分布，且样本量越大，近似性越强。

中心极限定理指出，对于大容量的随机样本，其样本均值的抽样分布形态近似于一个正态概率分布。这是统计学中非常有用的结论之一。可以在对样本来源分布形态一无所知的情况下，推断样本均值的分布。根据中心极限定理可知，样本均值作为随机变量有如下性质(注意，这里并没有假定 X 的分布)。

(1) 如果能够选择给定总体的特定容量的所有可能样本，那么，样本均值的抽样分布的均值将恰好等于总体均值，即 $\mu=\mu_{\bar{x}}$，即使不能得到所有样本，也可以预计样本均值分布的均值会接近于总体均值。

(2) 样本均值的抽样分布的离散程度小于总体分布。若总体标准差是 σ，则样本均值 $\bar{x}$ 的抽样分布的标准差为 $\sigma/\sqrt{n}$。当样本量增大时，$\sigma/\sqrt{n}$ 值将变小，即 $\bar{x}$ 的集中程度变大。

若把 σ 换成样本标准差 s，得到的 $s/\sqrt{n}$ 就是**均值的标准误**(standard error of mean)，它是 $\sigma/\sqrt{n}$ 的一个近似。为什么样本均值的波动会比总体的波动小呢？这是由于样本是把 N 个数据取均值，而这 N 个数据里总是更可能有大有小，因而平均起来就会相互抵消，造成的结果就是波动范围变小。而且，N 越大，这种相互之间的“拉平”作用越明显，从而波动(标准差)就减小得更多。

(3) 即使 X 不是正态分布变量，在很一般的条件下，当样本量增加时，$\bar{x}$ 的分布趋近于正态分布 $N(\mu,\sigma^2/n)$。

现在，从 $U(0,1)$ 分布对于三种样本量大小 $n=1$，3，100 分别取 1000 个样本，对每个样本算出均值。这样对每一种样本量都有 1000 个均值。用这些均值画直方图，如图 3-7 所示。图中的曲线是对这 1000 个均值的密度估计。下面小的短线标出了这 1000 个均值的实际位置。可以看出，样本量越大，均值的直方图越像正态变量的直方图，而且数据的分散程度也越小，数据越集中。

```
> a = NULL;for(i in 1:1000)a = c(a,runif(1))
> b = NULL;for(i in 1:1000)b = c(b,mean(runif(3)))
> c = NULL;for(i in 1:1000)c = c(c,mean(runif(100)))
>> unif = cbind(a,b,c);par(mfrow = c(1,3))
> hist(unif[,1],freq = F,xlab = "",main = expression(paste(U(0,1),", n = 1")))
> lines(density(a));rug(a)
> hist(unif[,2],freq = F,xlab = "",main = expression(paste(U(0,1),", n = 3")))
> lines(density(b));rug(b)
> hist(unif[,3],freq = F,xlab = "",main = expression(paste(U(0,1),", n = 100")))
```

```
> lines(density(c));rug(c)
> par(mfrow = c(1,1))
```

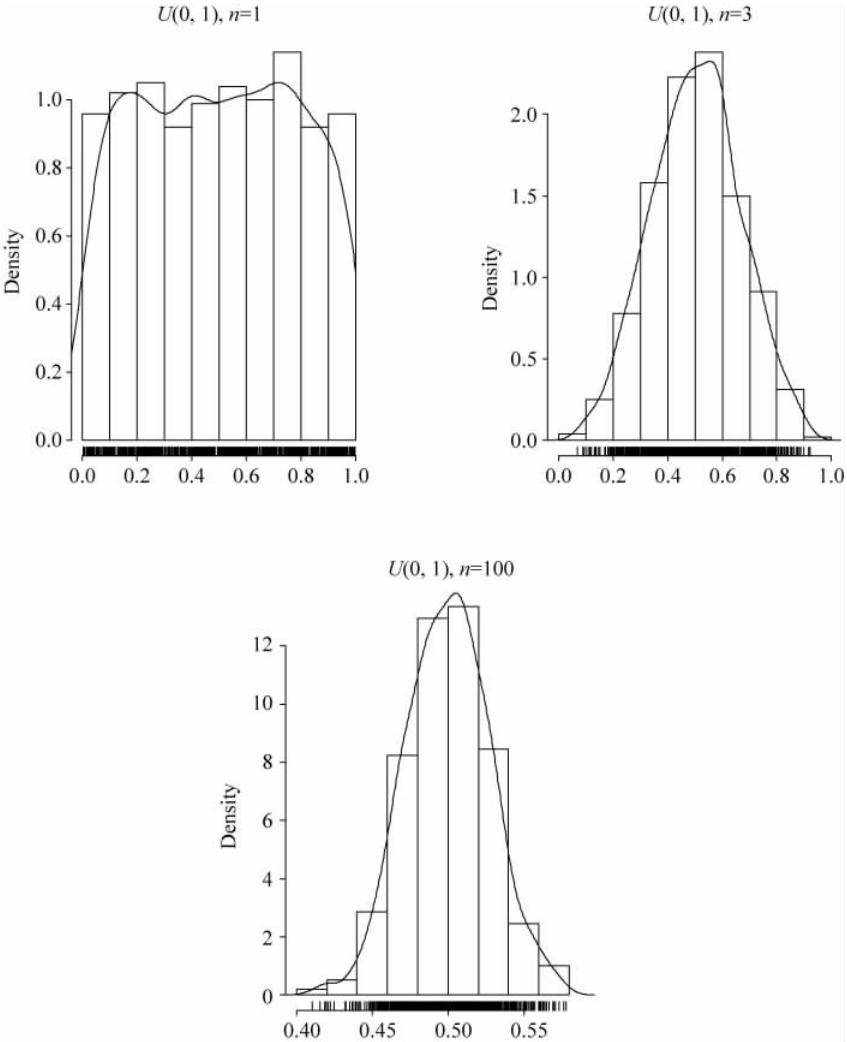

图 3-7　对 $U(0,1)$ 分布按照样本量 $n=1,3,100$ 分别取 1000 个样本计算出均值，得到的直方图

在实际的抽样问题中，我们常常希望对总体进行评价，但往往又缺少总体信息。此时，中心极限定理就能发挥效力。假定总体均值 μ 和总体标准差 σ 都是未知的，而通常主要对总体均值 μ 感兴趣。假定一个连续分布变量 X 的 n 个观测值组成一个样本，我们可以计算样本均值 $\bar{x}$ 和样本标准差 s。可以用样本均值来估计 μ 的值，这种估计的好坏取决于样本均值的抽样分布。我们知道，对任何形态的总体分布，如果抽取一个容量足够大的样本，那么样本均值的抽样分布将服从正态分布。统计理论证明，只要样本量大于 30，就有理由相信均值的抽样分布服从正态分布。

接下来将介绍几种常用的抽样分布。

3.4.2　抽样分布

我们希望利用样本，特别是作为样本函数的样本统计量来了解总体，对总体参数进行推断。这些样本统计量包括前面提到过的样本均值、样本中位数、样本标准差以及由它们组成的函数。利用样本结果估计总体参数会产生抽样误差，那么，如何基于样本信息对感兴趣的目标进行估计或预测呢？为回答该问题，我们来考察样本统计量的分布。相同样本量的样本统计量会随着样本的不同而不同，即样本统计量作为随机样本的函数，也是随机的，也有自己的分布，这些分布就称为**抽样分布**(sampling distribution)。

1. χ^2 分布

由正态变量导出的分布之一是**χ^2 分布**(chi-square distribution，卡方分布)。如果 x_1，x_2，…，x_n 是独立的标准正态分布变量，则 $\sum_{i=1}^{n} x_i^2$ 服从自由度为 n 的 χ^2 分布，记为 $\chi^2(n)$。这里的自由度 n 指包含的独立变量个数。更一般地，若干独立的 χ^2 分布变量的和也服从 χ^2 分布，其自由度等于那些 χ^2 分布变量自由度之和。由于 χ^2 分布变量为正态变量的平方和，所以它不会取负值。χ^2 分布也是一族分布，由该族成员的自由度来区分。

图 3-8 为三个不同自由度的 χ^2 分布密度曲线图。从图中可以看出，随着自由度的增大，图像趋于对称。

```
> x = c(seq(0,20,length = 1000)); y2 = dchisq(x,2);y3 = dchisq(x,3);y9 = dchisq(x,9)
> plot(x,y2,type = "l",xlab = "",ylab = "",lty = 1,main = expression(paste(chi ^2," 分布")))
> lines(x, y3, type = "l", xlab = "", ylab = "",lty = 2)
> lines(x,y9,type = "l",xlab = "",ylab = "",lty = 3)
> labels = c("df = 2","df = 3","df = 9"); atx = c(2,4,10) ; aty = c(0.45,0.2,0.12)
> text(atx, aty, labels = labels)
```

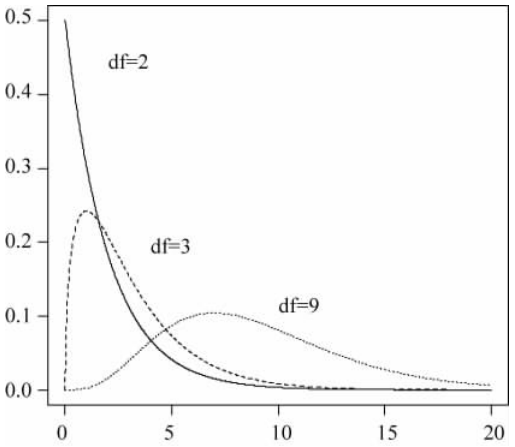

图 3-8　自由度为 2，3，9 的 χ^2 分布密度曲线

2. t 分布

在统计推断中往往希望利用样本均值减去总体均值再除以均值的总体标准差来得到标准正态变量 $z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$。但是在实际应用中，$\sigma$ 往往未知。因此在这个变换中，常常用样本标准差 s 来代替未知的总体标准差 σ，这时得到的变量 $t=\frac{\bar{x}-\mu}{s/\sqrt{n}}$ 就不再服从标准正态分布。它的密度曲线看上去有些像标准正态分布，但是中间瘦一些，而且尾巴厚一些。这种分布称为**t 分布**(t-distribution，或学生 t 分布(Student's t)。严格地说，假定有一个正态分布 $N(\mu,\sigma^2)$的样本，样本标准差为 s，样本均值为$\bar{x}$，样本量为 n，那么 $t=\frac{\bar{x}-\mu}{s/\sqrt{n}}$ 就服从自由度为$(n-1)$的 t 分布，记为 $t(n-1)$。不同的样本量通过标准化所产生的 t 分布也不同，这样就形成了一族分布。t 分布族中的成员是以自由度来区分的。有 k 个自由度的 t 分布用 $t(k)$表示，也有用 $t_{(k)}$或 t_k 表示的。

t 分布还可以根据 χ^2 分布导出：如果 X 是 $N(0,1)$变量，Y 是 $\chi^2(n)$变量，且 X 和 Y 独立，那么 $t=\frac{X}{\sqrt{Y/n}}$为有 n 个自由度的 t 分布，记为 $t\sim t(n)$。

图 3-9 展示了标准正态分布 $N(0,1)$和自由度分别为 1 和 10 的 t 分布的密度函数曲线。可以看出，t 分布两边尾巴比较长。但是当自由度增加时，它的分布就逐渐接近标准正态分布了。在 t 分布中，如果自由度趋于无穷，那么 t 分布就是标准正态分布。因此，在大样本情况下，可以用标准正态分布来近似 t 分布。

```
> x = seq( - 5,5,by = .1)
> plot(x,dnorm(x),type = 'l',xlab = "",ylab = "")
> lines(x,dt(x,df = 1),lty = 2)
> lines(x,dt(x,df = 10),lty = 3)
> labels = c("N(0,1)","t(1)","t(10)")
> atx = c(1.5, - 0.6, - 1.2) ; aty = c(0.35,0.16,0.3)
> text(atx, aty, labels = labels)
```

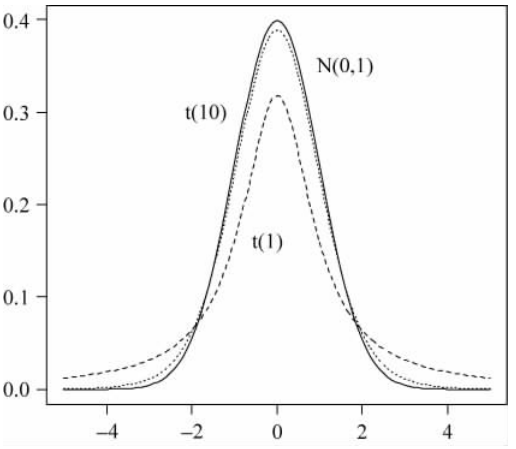

图 3-9　自由度分别为 1 和 10 的 t 分布和标准正态分布的密度曲线(虚线为 t 分布)

3. F 分布

两个独立 χ^2 分布变量在除以它们各自自由度之后的比为 F 分布变量。如果 X 是 $\chi^2(m)$变量，Y 是 $\chi^2(n)$变量，且 X 和 Y 独立，那么 $F=\frac{X/m}{Y/n}$服从自由度为(m,n)的 **F 分布**，记为 $F(m,n)$。F 分布有一个重要的性质：如果 F 变量服从 $F(m,n)$分布，那么$\frac{1}{F}$服从 $F(n,m)$分布，因为$\frac{1}{F}=\frac{Y/n}{X/m}$。

图 3-10 为不同自由度组合情况下的 F 分布密度曲线图。可以看出，当第二个自由度相同时，第一个自由度越小，波峰越靠近左边。F 分布不以正态分布为其极限，总为正偏分布。

```
> x = c(seq(0,5,length = 1000)); y2 = df(x,2,30);y3 = df(x,10,30);y4 = df(x,20,30);
> plot(x,y2,type = "l",xlab = "",ylab = "",lty = 1,main = "")
> lines(x,y3,type = "l",xlab = "",ylab = "",lty = 2)
> lines(x, y4, type = "l", xlab = "", ylab = "",lty = 3)
> labels = c("F(2,30)", "F(10,30)", "F(20,30)")
> atx = c(0.2, 1,1.6) ; aty = c(0.8,0.7,0.95)
> text(atx, aty, labels = labels)
```

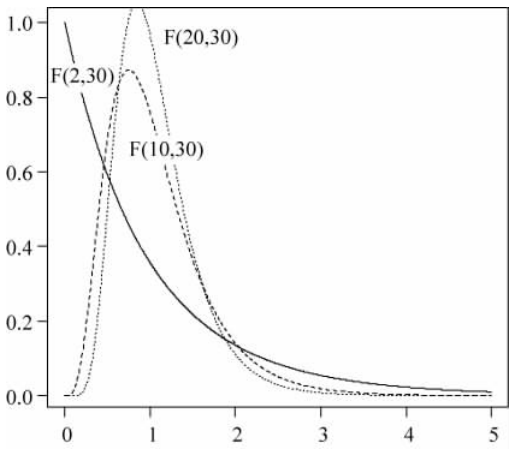

图 3-10　自由度分别为(2,30),(10,30),(20,30)的 F 分布的密度曲线

3.5　分位数

分位数函数是累积分布函数的反函数。p-分位数是具有这样性质的一个值：得到小于或等于它的概率为 p。根据定义，中位数即 50%分位数。分位数通常用于置信区间的计算，以及与设计实验有关的势函数的计算。下面给出一个置信区间计算的简单例子。如果 $\sigma=12$，测量了 $n=5$ 个观测值，得到均值，那么可以计算相关的分位数如下：

```
> xbar = 83
> sigma = 12
> n = 5
> sem = sigma/sqrt(n) #标准误
> sem
[1] 5.366563
> xbar + sem * qnorm(0.025)
[1] 72.48173
> xbar + sem * qnorm(0.975)
[1] 93.51827
```

因此，得到 μ 的一个置信度为 95%的置信区间。由于正态分布的对称性，有 $z_{0.025}=-z_{0.975}$，所以，通常把置信区间写成 $x\pm\sigma/\sqrt{n}\times z_{0.025}$。更多的置信区间知识见第 4 章。

练习题

1. 在例 3.1 中的模拟投币实验中用 rbinom 代替 sample，结果会如何变化？

2. 一条食品生产线每 8 小时一班中出现故障的次数服从平均值为 1.5 的泊松分布。求：

(1) 晚班期间恰好发生两次事故的概率。

(2) 下午班期间发生少于两次事故的概率。

(3) 连续三班无故障的概率。

3. 已知学生的统计考试成绩服从均值为 72、标准差为 8 的正态分布，求学生成绩不及格的概率和处于 65～80 的概率。

4. 对于一种疾病，已知有术后并发症发生的频率为 20%。10 位病人全部手术成功没有并发症的概率是多大？

5. 已知某地区人口的性别比例(男性人数：女性人数)为 106：100，而且有 5%的男人和 0.4%的女人是色盲。该地区任一人是色盲的概率有多大？

6. 消费者协会经过调查发现，某品牌空调有重要缺陷的产品数 X(单位：台)出现的概率分布如表 3-1 所示。

表 3-1　某品牌空调有重要缺陷的产品概率

X	0	1	2	3	4	5	6	7	8	9	10
p	0.041	0.130	0.209	0.223	0.178	0.114	0.061	0.028	0.011	0.004	0.001

根据这些数值，分别计算：

(1) 有 2～5 个(包括 2 个与 5 个在内)空调出现重要缺陷的可能性。

(2) 只有不到 2 个空调出现重要缺陷的可能性。

(3) 有超过 5 个空调出现重要缺陷的可能性。

7. 某厂生产的螺栓的长度服从均值为 10cm，标准差为 0.05cm 的正态分布。按质量标

准规定，长度在 9.9～10.1cm 范围内的螺栓为合格品。该厂螺栓的不合格率是多少？

8. 某企业生产的某种电池寿命近似服从正态分布，且均值为 200 小时，标准差为 30 小时。若规定寿命低于 150 小时为不合格品。

① 该企业所生产电池的合格率是多少？

② 电池寿命在 200 小时左右多大范围内的概率不小于 0.9？

9. 一个药厂声称他们生产的某种药的疗效达到 80%，但是 100 个人使用该药后只有 40 个人说有效。那么，药厂的说法对吗？这是不是小概率事件？（提示：计算小于或等于 40 人有效的概率）

10. 我们得到 5 个学生的身高数据（单位：cm）：177，180，165，166，170。现在要求对这 5 个数据进行重复抽样，每次取 3 个数据，重复 1000 次，写出样本均值的概率分布，画出直方图。