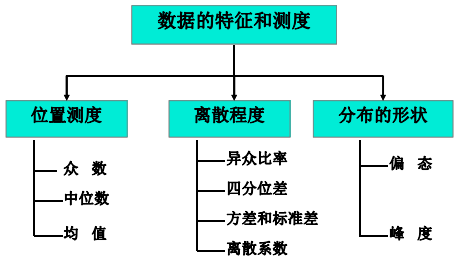
数据分布的特征和测度



数据类型与位置测度值

表a 数据类型和所适用的位置测度值

数据类型	定类数据	定序数据	定距数据	定比数据
适用的测度值	※众数	※中位数	※均值	※均值
	—	四分位数	众数	调和平均数
	—	众数	中位数	几何平均数
	—	—	四分位数	中位数
	—	—	—	四分位数
	—	—	—	众数

数据类型与离散程度测度值

表b 数据类型和所适用的离散程度测度值

数据类型	定类数据	定序数据	定距数据或定比数据
适用的测度值	※异众比率	※四分位差	※方差或标准差
	—	异众比率	※离散系数（比较时用）
	—	—	极差
	—	—	四分位差
	—	—	异众比率

第三章 概率和分布

事件

- 随机事件
- 事件的相互关系
  1、和事件（$A+B$）
  2、积事件（$A\cdot B$）
  3、互斥事件与对立事件（$A\cdot B=\emptyset$）
  4、独立事件
  5、完全事件系

概率（probability）

概率是0和1之间的一个数目，表示某个事件发生的可能性或经常程度。

$$0 \le P(A) \le 1$$

事件（event）相当于集合论中的集合（set）。
而概率则是事件的某种函数

小概率事件（small probability event）

§3.1 得到概率的几种途径

1、古典概型（先验概率）

1）随机试验的全部可能的结果是有限的；
2）基本事件间是互不相容且等可能的

事件A的概率是A中所包含的基本事件数（m）与基本事件总数（n）的比值

$$P(A)=\frac{m}{n}$$

2、统计定义（后验概率）

事件出现的频数 k 除以重复试验的次数 n，该比值 k/n 称为相对频数（relative frequency）或频率。

很多事件无法进行长期重复试验。因此这种通过相对频数获得概率的方法也并不是万能的。

3、主观概率

主观概率**(subjective probability)**是一次事件的概率。或为基于所掌握的信息，某人对某事件发生的自信程度。

§3.2 概率的运算

1.互补事件的概率

- 如果一个不出现，则另一个肯定出现的两个事件称为**互补事件（complementary events，或者互余事件或对立事件）。**

$$P(A)+P(A^C)=1 \quad 或 \quad P(A^C)=1-P(A)$$

优势或赔率（odds）常用来形容输赢的可能。

$$P(A)/P(A^C)=P(A)/[1-P(A)]$$

2.概率的加法

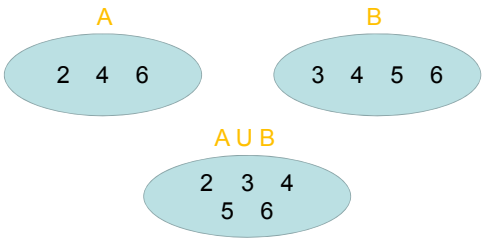
- 如果两个事件不可能同时发生，那么其和事件的概率为这两个概率的和。

$$P(A+B) = P(A) + P(B)$$

例：假定掷骰子时，一个事件A为“得到偶数点”，另一个事件B为“得到大于或等于3点”；

P(A)=1/2　　P(B)=2/3。

“得到大于或等于3点或者偶数点”的事件的概率是？



$$P(A+B)=P(A)+P(B)-P(A\cdot B)$$

3.概率的乘法

例：三个人抽签，而只有一个人能够抽中，因此每个人抽中的机会是1/3。假定用A_1、A_2和A_3分别代表这三个人抽中的事件，那么，

$P(A_1) = P(A_2) = P(A_3) = 1/3$

条件概率

如第一个人抽到（事件A_1），则

$P(A_2|A_1)=P(A_3|A_1)=0$。

如第一个人没抽到（事件A_1^C），那么

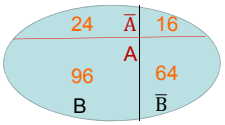
$P(A_2|A_1^C)=P(A_3|A_1^C)=1/2$。

在事件B发生的情况下，事件A发生的条件概率定义为

$$P(A|B)=\frac{P(A\cdot B)}{P(B)},P(B)\neq 0$$

表2-1 使用两种药物杀灭蜈虫效果

	死亡（A）	存活（$\bar{A}$）	和
甲药物（B）	96	24	120
乙药物（$\bar{B}$）	64	16	80
和	160	40	200



$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

接受甲药物后虫子死亡的概率 $=\frac{96}{120}$

死亡的虫子，接受了甲药物的概率是？

$$P(A|B)=\frac{P(A\cdot B)}{P(B)},P(B)\neq 0$$

$$P(B|A)=\frac{P(A\cdot B)}{P(A)},P(B)\neq 0$$

$$\Rightarrow P(A\cdot B)=P(B)P(A|B)=P(A)P(B|A)$$

概率乘法法则

- 若事件A的发生，并不影响事件B发生的概率，即

$$P(B|A)=P(B)\ 或\ P(A|B)=P(A)$$

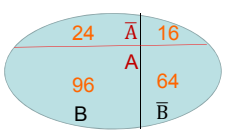
此时称A和B是独立事件（independent event）

$$P(A\cdot B)=P(B)P(A|B)=P(A)P(B|A)$$

$$\Rightarrow P(A\cdot B)=P(A)P(B)$$

表2-1 使用两种药物杀灭蜈虫效果

	死亡（A）	存活（$\bar{A}$）	和
甲药物（B）	96	24	120
乙药物（$\bar{B}$）	64	16	80
和	160	40	200



$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

接受甲药物后虫子死亡的概率 $=\frac{96}{120}=0.8$

$$P(A)=\frac{160}{200}=0.8$$

说明：死亡与否不受是否接受甲药物的影响。

4.贝叶斯公式

$$P(A|B)=\frac{P(A\cdot B)}{P(B)},P(B)\neq 0$$

- 若事件B能且只能与$A_1,A_2,\ldots,A_n$之一同时发生，那么，在事件B已发生的条件下，

$$P(\mathrm{A_i}|B)=\frac{P(\mathrm{A_i})P(B|\mathrm{A_i})}{\sum_{j=1}^{k}P(\mathrm{A_j})P(B|\mathrm{A_j})}$$

全概率公式

例：假定一中年男性群体，肥胖者占20%，标准体重占50%，低体重的占30%。这3类人出现动脉硬化的概率分别为30%，10%和1%。现从这个群体中随机抽出一人，他恰恰是动脉硬化的患者，问这个人来自肥胖组、标准体重和低体重组的概率各是多少？

解：用B表示抽到动脉硬化患者的事件；
分别用A_1、A_2、A_3表示抽到肥胖者、标准体重者、低体重者的事件

$$P(A_i|B)=\frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{3}P(A_j)P(B|A_j)}$$

离散型随机变量与连续型随机变量

试验	随机变量	可能的取值
抽查100个产品	取到次品的个数	0,1,2,…,100
一家餐馆营业一天	顾客数	0,1,2,…
抽查一批电子原件	使用寿命	X≥0
新建一座住宅楼	半年完成工程的百分比	0≤X ≤100

概率分布

- 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)**。
- 概率分布可以用各种图或表来表示；一些可以用公式来表示。
- 概率分布是关于总体的概念。有了概率分布就等于知道了总体。

§3.3 离散变量的分布

一般来说，某离散随机变量的每一个可能取值 x_i 都相应于取该值的概率 $p(x_i)$，这些概率应该满足关系

$$\sum_i p(x_i)=1,\ p(x_i)\geq 0$$

离散型随机变量的概率分布也可用表格的形式表示

§3.3.1 离散型随机变量的数学期望和方差

离散型随机变量的数学期望 (expected value)

1. 离散型随机变量X的所有可能取值x_i与其取相对应的概率p_i乘积之和
2. 描述离散型随机变量取值的集中程度
3. 记为μ或$E(X)$
4. 计算公式为

$$\mu = E(X) = \sum_i x_i p_i$$

离散型随机变量的方差 (variance)

1. 随机变量X的每一个取值与期望值的离差平方和的数学期望，记为σ^2或$D(X)$
2. 描述离散型随机变量取值的分散程度
3. 计算公式为

$$\sigma^2 = D(X) = E\{[X - E(X)]^2\} = \sum_i (x_i - \mu)^2 \cdot p_i$$

4. 方差的平方根称为标准差，记为σ或 $\sqrt{D(X)}$

离散型数学期望和方差 (例题分析)

【例】一家电脑配件供应商声称，他所提供的配件100个中拥有次品的个数及概率如下表

每100个配件中的次品数及概率分布

次品数$X = x_i$	0	1	2	3
概率$P(X=x_i)=p_i$	0.75	0.12	0.08	0.05

求该供应商次品数的数学期望和标准差

$$\mu = \sum_i x_i p_i = 0 \times 0.75 + 1 \times 0.12 + 2 \times 0.08 + 3 \times 0.05 = 0.43$$

$$\sigma^2 = D(X), \quad \sum_i (x_i - \mu)^2 \cdot p_i = 0.7051, \sigma = 0.8397$$

随机变量平均数的法则

$$\mu_{X \pm Y} = \mu_X \pm \mu_Y$$

$$\mu_{a+bY} = a + b\mu_Y$$

随机变量方差的法则

$$\sigma^2_{a+bY} = b^2\sigma^2_Y$$

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y$$

§3.3.2 二项分布

仅有两种结果的重复独立试验被称为**伯努利试验（Bernoulli trials）**。

如果进行n次伯努利试验，每次成功的概率为p，那么成功的次数为k概率$p(k)$服从

二项分布(binomial distribution)

用符号 B(*n*,*p*) 或 Bin(*n*,*p*) 表示。

二项分布公式

$$P(k) = C_n^k \cdot p^k \cdot (1-p)^{n-k}, \quad k = 0,1,\dots,n$$

其中 $C_n^k = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$

$$F(k) = \sum_{x=0}^{k} P(x) \qquad k = 0,1,\dots,n$$

二项分布公式

若 $Y \sim B(n,p)$

$$\mu = E(Y) = \sum Y_i P(Y_i) = np$$

$$\sigma^2 = E(Y - EY)^2 = np(1-p)$$

例：从雌雄各半的100只动物中，随机抽取一只记下性别后放回，再做第二次抽取。按这种方式共抽取十次，抽到雄性动物的次数服从什么分布？

放回式抽样

若抽取后不放回，连续随机抽取十次，抽到雄性动物的次数是否服从二项分布？

成功概率p不同！

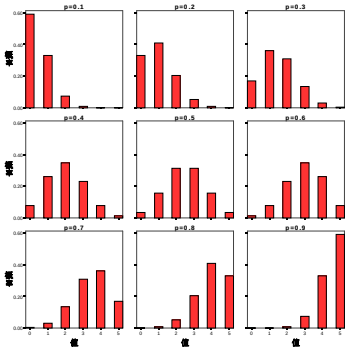
在制定实验计划时，首先要以指定的概率求出所需的样本含量。

例：用棕色正常毛(bbRR)的家兔和黑色短毛兔(BBrr)杂交。F_1代为黑色正常毛(BbRr)，近亲交配，F_2代期望产生1/16棕色短毛(bbrr)的家兔。问最少需多少F_2代的家兔，才能以99%的概率至少得到一只棕色短毛兔？

解：F_2代为bbrr家兔的概率为p=1/16。抽取n只F_2代家兔，非bbrr型的只数为n的概率为1%

$$C_n^0 \cdot p^0 \cdot (1-p)^{n-0} = (1-p)^n = \left(\frac{15}{16}\right)^n \leq 0.01$$

图3.1 二项分布B(5,p) (p＝0.1到0.9)的概率分布图



§ 3.3.3 泊松分布

“泊松分布”（Poisson分布）是描述在一定时空内某种事件出现次数分布的理想化模型。

$p(k)$表示随机变量等于 k 的概率，

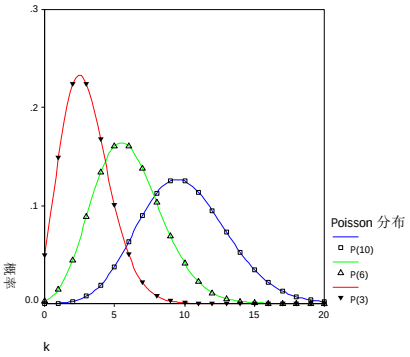
$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \mathrm{k} = 0, 1, 2, \ldots$$

参数 λ 表示一定时空内某事件出现的平均次数

$\lambda = np$

$\mu = E(Y) = \lambda$　　$\sigma^2 = E\left((Y - EY)^2\right) = \lambda$

图3.2 参数为3、6、10的泊松分布



例：麦田内平均每10m²有1株杂草，问每100m²麦田中有0株杂草，1株杂草，2株杂草…的概率是多少？

解：每100m²麦田中，平均杂草数

$$\lambda = 10$$

由此可得，每100m²麦田中有 y 株杂草的概率：

$$p(y) = \frac{10^y}{y!}e^{-10}$$

例：用显微镜检查某食品样本内结核菌的数目，对在某些视野内各小方格中的细菌数目加以计数，然后记录含不同细菌数目的实际格子数目N，结果如下表所示，试求各种细菌数的理论格子数n。

细菌数(x)	0	1	2	3	4	5	6	7	8	9	合计
N	5	19	26	26	21	13	5	1	1	1	118
P(x)											
n											

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1}{118}(0\times 5 + 1\times 19 + \cdots + 9\times 1) = 2.9831$$

$x \sim P(2.9831)$　　$n(x) = \sum f \cdot P(x)$

例：用显微镜检查某食品样本内结核菌的数目，对在某些视野内各小方格中的细菌数目加以计数，然后记录含不同细菌数目的实际格子数目N，结果如下表所示，试求各种细菌数的理论格子数n。

细菌数(x)	0	1	2	3	4	5	6	7	8	9	合计
N	5	19	26	26	21	13	5	1	1	1	118
P(x)	0.0506	0.1511	0.2253	0.2240	0.1671	0.0997	0.0496	0.0211	0.0079	0.0026	0.9990
n	5.97	17.83	26.59	26.43	19.72	11.76	5.85	2.49	0.93	0.31	117.88

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1}{118}(0\times 5 + 1\times 19 + \cdots + 9\times 1) = 2.9831$$

$x \sim P(2.9831)$　　$n(x) = \sum f \cdot P(x)$

§3.3.4 超几何分布 (hypergeometric distribution)

$$p(y) = \frac{C_K^y C_{N-K}^{n-y}}{C_N^n}, \qquad y = 0, 1, 2, \ldots, n$$

N：总体中的个体数

K：两种类型中某一种类型的个体数

y：n次抽样中抽中某一类型的个体数

$$\mu = \frac{nK}{N} \qquad \sigma^2 = \frac{nK(N-K)(N-n)}{N^2(N-1)}$$

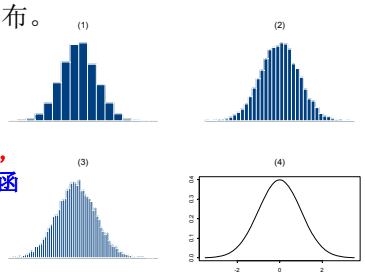
例：在野生动物考察时，了解野生动物群体大小的一种方法是：先捕捉一定数目(*K*)的动物，做上标记，把它们放回到群体中，然后再捕捉第二个样本(样本含量*n*)，计数其中有标记的动物数(*y*)，由此估计群体的大小。

在捕捉第二个样本时，其中有标记的动物数是一个服从超几何分布的随机变量。

$$\mu = \frac{nK}{N} \quad \Longrightarrow \quad \hat{N} = \frac{nK}{y}$$

§3.4 连续变量的分布

- 随机变量如果能够在一区间内取任意值，则该变量在此区间内是连续的，其分布为连续型概率分布。



概率密度函数(probability density function，pdf)，简称密度函数或概率密度

- 连续变量密度函数曲线（这里用*f*表示）下面覆盖的总面积为1，即

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

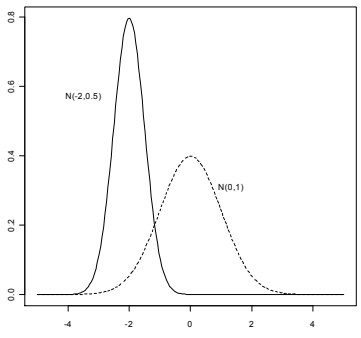
- 取某个特定值的概率都是零，而只有变量取值于某个（或若干个）区间的概率才可能大于0。

§3.4.1 正态分布(normal distribution）

- 又叫**高斯分布（Gaussian distribution)**
- 一个正态分布用N(μ,σ)表示，其中μ为均值，而σ为标准差；也常表示成N(μ,σ²)。

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

两条正态分布的密度曲线。左边是N(-2,0.5)分布，右边是N(0, 1)分布



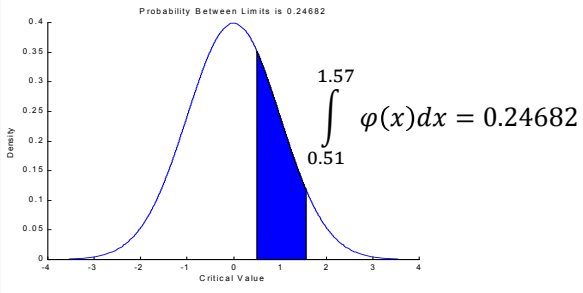
- **标准差为1的正态分布N(0，1)称为标准正态分布(standard normal distribution)**，其概率密度用$\varphi(x)$表示。

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

- 任何具有正态分布N(μ,σ)的随机变量X都可以用简单的变换成为标准正态随机变量。

$$Z = (X - \mu) / \sigma$$

标准正态变量在区间(0.51, 1.57)中取值的概率



$$\int_{0.51}^{1.57} \varphi(x)dx = 0.24682$$

单侧临界值

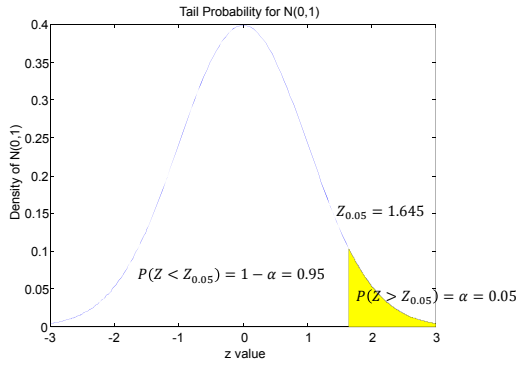
对于连续型随机变量X，α下分位数（又称为α分位数，α-quantile）定义为数x_α，它满足

$$P(X \le x_\alpha) = \alpha$$

α上分位数（α-upper quantile）定义为数x_α，它满足

$$P(X \ge x_\alpha) = \alpha$$

N(0,1)分布右侧尾概率P(z>z_α)=α的示意图



x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9648	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767

例：假定一批莲子粒重Y服从N(1.001,0.164²)。求Y在0.97和1.04g之间的概率。

解：

$$Z = \frac{Y - 1.001}{0.164} \sim N(0,1)$$

$$p(0.97 < Y < 1.04)$$
$$= p\left(\frac{0.97 - 1.001}{0.164} < Z < \frac{1.04 - 1.001}{0.164}\right)$$
$$= \Phi(0.24) - \Phi(-0.19)$$
$$= 0.5948 - 0.4247 \doteq 0.17$$

一些常用值

$$\int_{-1}^{1} \varphi(x)dx = 0.6827 \qquad \int_{-2}^{2} \varphi(x)dx = 0.9543$$

$$\int_{-3}^{3} \varphi(x)dx = 0.9973$$ 3σ法则

$$\int_{-1.96}^{1.96} \varphi(x)dx = 0.95 \qquad \int_{-2.576}^{2.576} \varphi(x)dx = 0.99$$

§3.4.2 指数分布 (exponential distribution)

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$$\mu = E(X) = \frac{1}{\lambda} \qquad \sigma = \frac{1}{\lambda}$$

- 无记忆性：

$$P(x > s + t | x > s) = \frac{P(x > s + t)}{P(x > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}$$

§3.6 大数定律与中心极限定理

一、大数定律：阐述大量随机变量的平均结果具有稳定性的一系列定律的总称。

1、辛钦大数定律（独立同分布大数定律）：
若$x_1, x_2, \ldots, x_n$符合i.i.d.,存在有限的数学期望和方差，对任意小的正数ε，有$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right| < \varepsilon\right\} = 1$

独立同分布大数定律：提供了用样本平均数估计总体平均数的理论依据

伯努利大数定律

2、伯努利大数定律：
设m是n次独立随即实验中事件A发生的次数，
p是事件A在每次试验中发生的概率，则
对任意小的正数ε，有$\lim_{n\to\infty} P\left\{\left|\frac{m}{n}-p\right|<\varepsilon\right\}=1$

伯努利大数定律：提供了用频率代替概率的理论依据

中心极限定理

二、中心极限定理：阐述大量随机变量之和的极限分布是正态分布的一系列定理的总称。

若$x_1, x_2, \ldots, x_n$符合i.i.d.,存在有限的数学期望μ和方差σ^2，当$n\to\infty$时，随机变量的总和$\sum x_i \sim N(n\mu, n\sigma^2)$或其算术平均数$\bar{x}\sim N(\mu, \frac{\sigma^2}{n})$

德莫佛—拉普拉斯中心极限定理

设m是n次独立随即实验中事件A发生的次数，
p是事件A在每次试验中发生的概率，
则X服从二项分布B(n,p)
当$n\to\infty$时，$X\sim N(np, npq)$

该定理提供了用正态分布近似计算二项分布概率的方法。

例：对于一个学生而言，来参加家长会的家长人数是一个随机变量，设一个学生无家长、1名家长、2名家长来参加会议的概率分别为0.05、0.8、0.15。若学校共有400名学生，设各学生参加会议的家长数相互独立，且服从同一分布。（1）求参加会议的家长数X超过450的概率；（2）求有1名家长来参加会议的学生数不多于340的概率。

解（1）以X_k(k=1,2,…,400)记第k个学生来参加会议的家长数，则X_k的分布律为

X_k	0	1	2
p_k	0.05	0.8	0.15

易知$E(X_k)=1.1$，$D(X_k)=0.19$ k=1,2,…,400,而 $X=\sum_{k=1}^{400} X_k$

$$\frac{\sum_{k=1}^{400} X_k - 400\times 1.1}{\sqrt{400}\sqrt{0.19}} = \frac{X-400\times 1.1}{\sqrt{400}\sqrt{0.19}}$$

近似服从正态分布$N(0,1)$,

于是，$P(X>450)=P\left\{\frac{X-400\times1.1}{\sqrt{400}\sqrt{0.19}}>\frac{450-400\times1.1}{\sqrt{400}\sqrt{0.19}}\right\}$

$$=1-P\left\{\frac{X-400\times1.1}{\sqrt{400}\sqrt{0.19}}\le 1.147\right\}$$
$$\approx 1-\emptyset(1.147)=0.1357$$

(2)以Y记有一名家长来参加会议的学生人数，则Y~b(400,0.8)，

$\mu = np = 400\times 0.8$=320

$\sigma^2 = np(1-p) = 400\times 0.8\times 0.2$=64

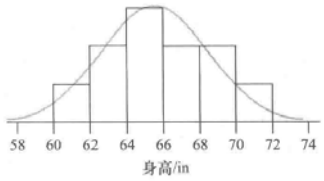
由中心极限定理得

$$P(Y\le 340)=P\left\{\frac{Y-320}{\sqrt{64}}\le\frac{340-320}{\sqrt{64}}\right\}$$

$$=P\left\{\frac{Y-320}{8}\le 2.5\right\}\approx\emptyset(2.5)=0.9938$$

§3.7 正态性评估

例：11位女性身高（以in为单位）：61, 62.5, 63, 64, 64.5, 65, 66.5, 67, 68, 68.5, 70.5



$\mu = 65.5$

$\sigma = 2.9$

正态概率图（normal probability plot）

Step 1：将数据排序，计算样本的百分位数

11 位女性身高的指标计算和百分位数

i	1	2	3	4	5	6	7	8	9	10	11
观察身高	61.0	62.5	63.0	64.0	64.5	65.0	66.5	67.0	68.0	68.5	70.5
百分位数 100（i/11）	9.09	18.18	27.27	36.36	45.45	54.55	63.64	72.73	81.82	90.91	100.00
调整的百分位数 100（i-1/2）/n	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.36	95.45

正态概率图（normal probability plot）

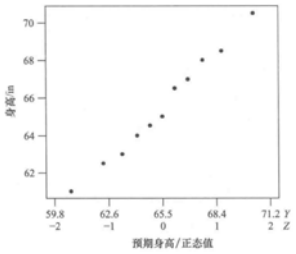
Step 2：计算理论身高 $\mu + Z_\alpha \sigma$

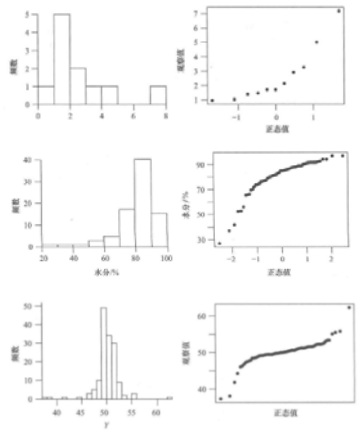
11 位女性身高的理论 Z 值计算

i	1	2	3	4	5	6	7	8	9	10	11
观察身高	61.0	62.5	63.0	64.0	64.5	65.0	66.5	67.0	68.0	68.5	70.5
调整的百分位数 100（i-1/2）/n	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.36	95.45
Z	−1.69	−1.10	−0.75	−0.47	−0.23	0.00	0.23	0.47	0.75	1.10	1.69
理论身高	60.6	62.3	63.4	64.1	64.8	65.5	66.2	66.7	67.6	68.7	70.4

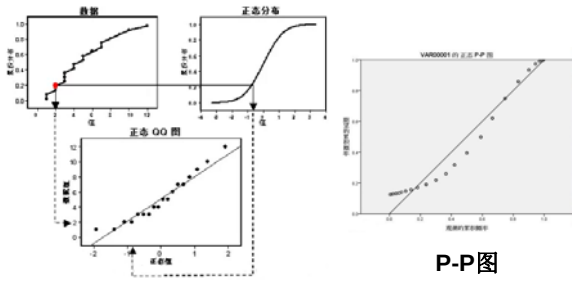
正态概率图（normal probability plot）

Step 3：绘制观测值对理论值的散点图





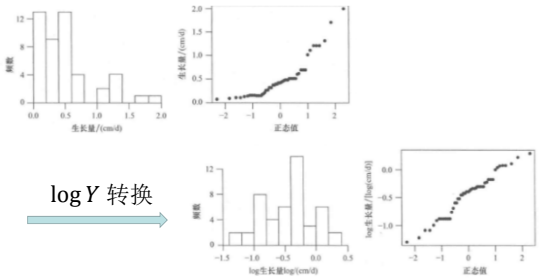
Q-Q图与P-P图



Q-Q图

P-P图

非正态数据的转换



$\log Y$ 转换

$\sqrt{Y}$、$\log Y$、$1/\sqrt{Y}$、$1/Y$效果依次增强

The End