

Industrial Big Data 工业大数据

第二部分：在质量控制与改进中有
用的统计方法

第三章：过程质量建模

彭世喆

数字经济系

经济与管理学院

长沙理工大学



大纲

- 构造并解读可视化数据展示，包括茎叶图、直方图和箱线图
- 计算并解读样本均值、样本方差、样本标准差和样本极差
- 解释随机变量和概率分布的概念
- 理解并解读概率分布的均值、方差和标准差
- 根据概率分布确定概率
- 理解每种离散概率分布的前提假设
- 理解每种连续概率分布的前提假设
- 在特定应用中选用合适的概率分布
- 使用概率图
- 对某些超几何分布和二项分布使用近似方法

统计学

- 统计学是一种基于对样本中包含的信息进行分析，从而对过程或总体做出决策的技术集合
- 统计方法在质量控制与改进中发挥着至关重要的作用。它们是产品抽样、测试和评估的主要手段，并且这些数据中的信息可被用于控制及改进过程和产品
- 没有任何两个由同一工艺生产的产品是完全相同的！没有任何两项服务活动是完全相同的！
- 统计学是一门分析数据并得出推断的科学，它会考虑数据中的变异。数据由 $x_1, x_2, \dots, x_n$ 表示



描述变异——茎叶图

- 我们将每个数字 x_i 分为两部分：茎（由一个或多个前导数字组成）和叶（由其余数字组成）。
 - 如将数值76拆分为茎7和叶6
 - 在每个茎旁边，按数值大小列出对应的叶
 - 茎左侧的列给出了一个累积计数，对于较小的茎，它表示该茎及以下的数据观测数；对于较大的茎，它表示该茎及以上的数据观测数。对于中间的茎，括号内的数字表示该茎所包含的观测数量

Cycle Time in Days to Pay Employee Health Insurance Claims

Claim	Days	Claim	Days	Claim	Days	Claim	Days
1	48	11	35	21	37	31	16
2	41	12	34	22	43	32	22
3	35	13	36	23	17	33	33
4	36	14	42	24	26	34	30
5	37	15	43	25	28	35	24
6	26	16	36	26	27	36	23
7	36	17	56	27	45	37	22
8	46	18	32	28	33	38	30
9	35	19	46	29	22	39	31
10	47	20	30	30	27	40	17

Stem-and-Leaf Display: Days

Stem-and-leaf of Days

N = 40

Leaf Unit = 1.0

```

3    1    677
8    2    22234
13   2    66778
(8)  3    00012334
19   3    555666677
10   4    1233
6    4    56678
1    5
1    5    6
    
```

描述变异——茎叶图

- 查找数据的百分位数
 - 第一四分位数是排名为 $(0.25)(40)+0.5=10.5$ 的观测值（排在第十和第十一个观测值中间），即 $(26+27)/2=26.5$
 - 第三四分位数是排名为 $(0.75)(40)+0.5=30.5$ 的观测值（排在第三十和第三十一个观测值中间），即 $(37+41)/2=39$
 - 第一和第三四分位数分别用符号Q1和Q3表示，四分位数间距 $IQR=Q3-Q1$ 偶尔被用作衡量变异性的指标
- 还呈现了数据的形状、离散程度或变异性，以及数据的中心趋势或中间值（该值接近35）

Stem-and-Leaf Display: Days

Stem-and-leaf of Days

N = 40

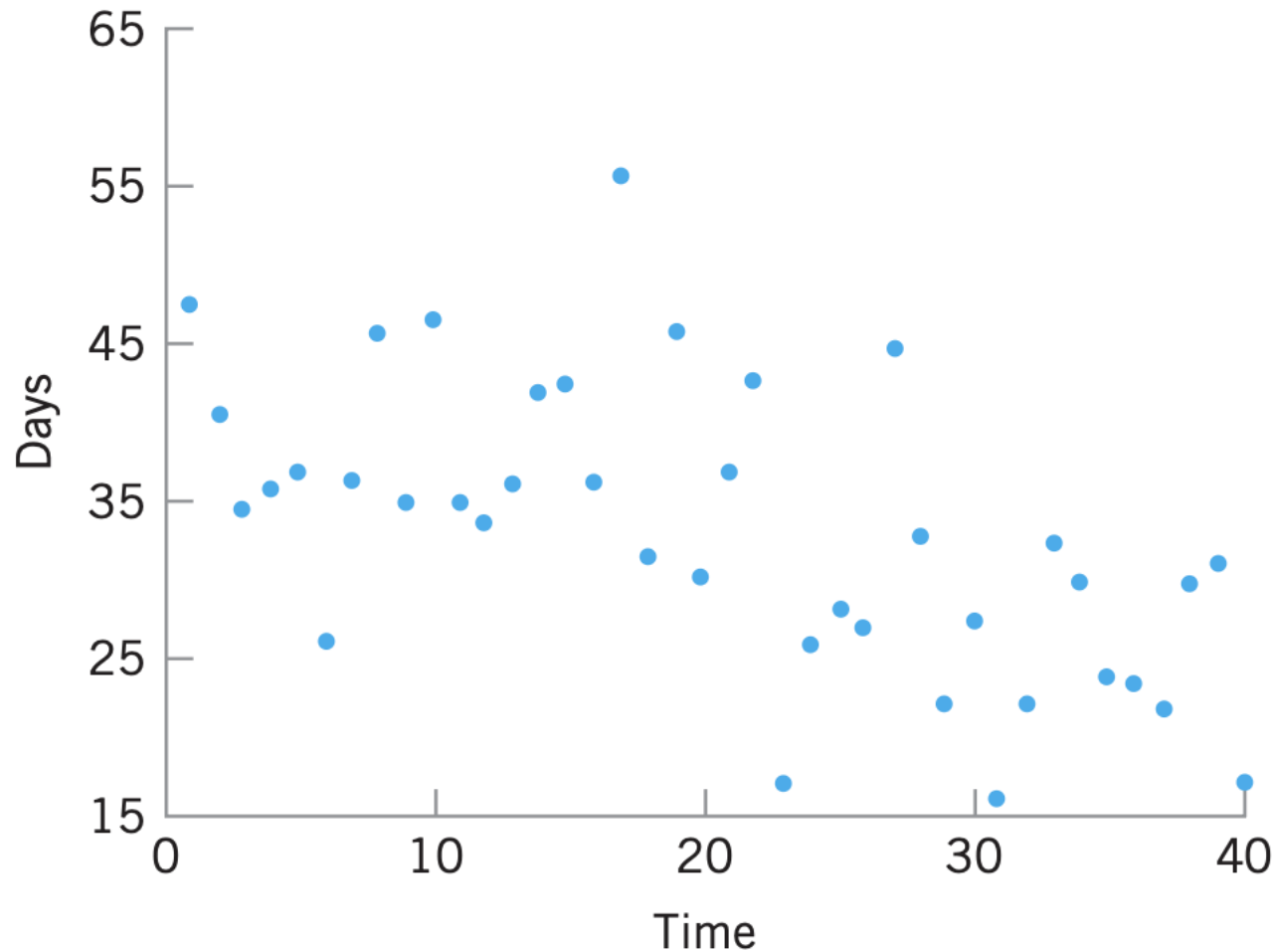
Leaf Unit = 1.0

3	1	677
8	2	22234
13	2	66778
(8)	3	00012334
19	3	555666677
10	4	1233
6	4	56678
1	5	
1	5	6



描述变异——时间序列图

- 若要考虑观测值的时间顺序



描述变异——直方图

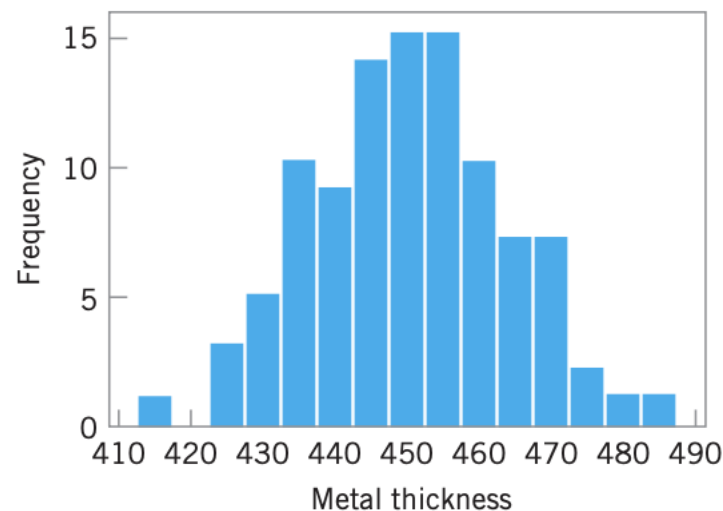
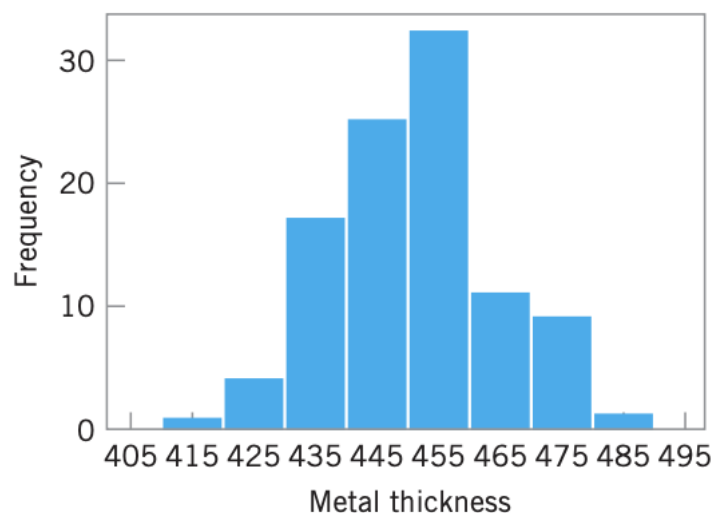
- 为连续数据构建直方图，我们必须将数据范围划分为区间，这些区间通常称为类区间、单元或箱
 - 在大多数情况下5到20个箱是令人满意的，并且箱数应随 n 增加
 - 在实践中，将箱数设置为约等于 $\sqrt{n}$ 往往会取得良好效果，或Sturges规则 ($1 + \log_2 n$)

Layer Thickness (Å) on Semiconductor Wafers

438	450	487	451	452	441	444	461	432	471
413	450	430	437	465	444	471	453	431	458
444	450	446	444	466	458	471	452	455	445
468	459	450	453	473	454	458	438	447	463
445	466	456	434	471	437	459	445	454	423
472	470	433	454	464	443	449	435	435	451
474	457	455	448	478	465	462	454	425	440
454	441	459	435	446	435	460	428	449	442
455	450	423	432	459	444	445	454	449	441
449	445	455	441	464	457	437	434	452	439

描述变异——直方图

- 直方图提供了测量值分布形状的视觉印象，以及数据内在变异性的相关信息
- 直方图对箱的数量和宽度选择相对敏感
- 从原始数据或茎叶图过渡到直方图时，我们在某种意义上丢失了一些信息，因为原始观测值并未在图中保留

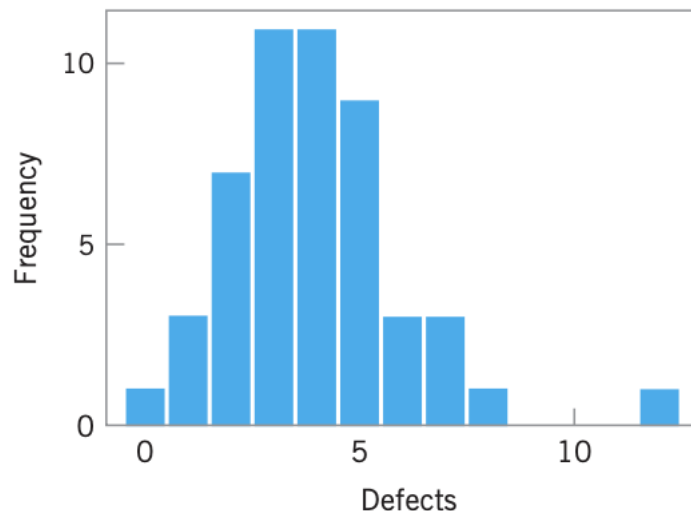


描述变异——直方图

- 频率分布和直方图也可以用于定性、分类或计数（离散）数据
 - 在某些应用中，类别有自然顺序（如大一、大二、大三、大四）
 - 而在其他应用中，类别的顺序则是任意的（如男性和女性）
 - 使用分类数据时，每个值对应一个组

Surface Finish Defects in Painted Automobile Hoods

6	1	5	7	8	6	0	2	4	2
5	2	4	4	1	4	1	7	2	3
4	3	3	3	6	3	2	3	4	5
5	2	3	4	4	4	2	3	5	7
5	4	5	5	4	5	3	3	3	12



数值汇总

- 样本中最重要的集中趋势度量是样本均值

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

- 样本数据的变异性由样本方差衡量，单位是原始数据单位的平方（这通常不方便且难以解释）

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

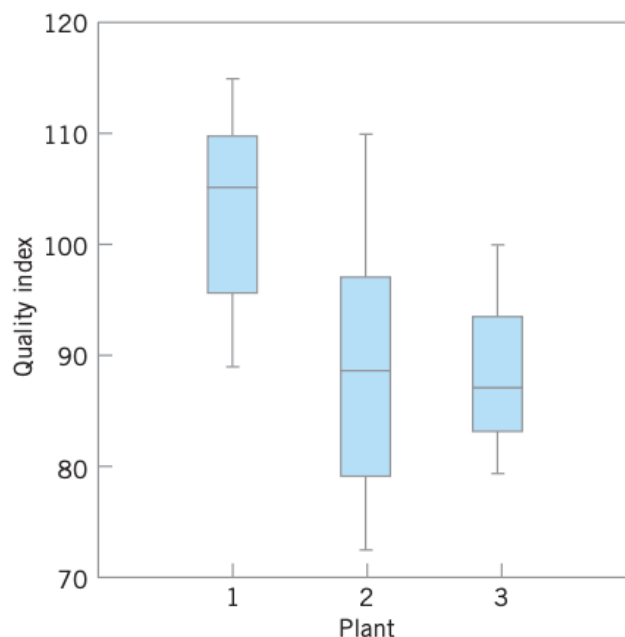
- 通常更喜欢使用 s^2 的平方根，称为样本标准差，作为变异性的度量，样本标准差的主要优点是它以原始测量单位表示

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

- 标准差并不反映样本数据的大小，只反映相对于均值的离散程度
- 样本数据同时加上一个数，标准差不变

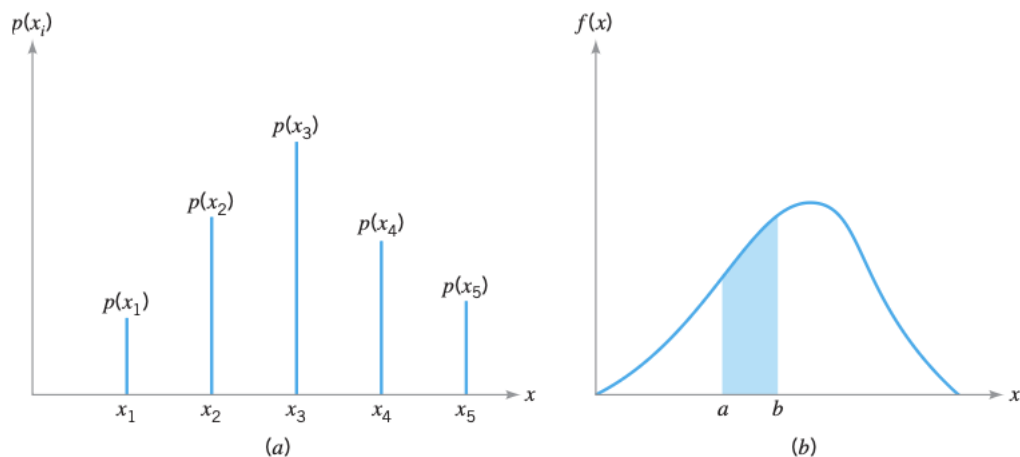
描述变异——箱线图

- 箱线图能同时显示数据的多个重要特征，例如位置或集中趋势、散布或变异性、围绕中位数的偏离对称性
- 箱线图在矩形箱上显示数据的三个四分位数（Q1，Q2，Q3）、最小值和最大值，两端各有一条线延伸到极值
- 须线最多只从箱体两端延伸 $1.5(Q3-Q1)$ 的距离，超出这些界限的观测值则被标记为潜在异常值



概率分布

- 概率分布是一种数学模型，它将变量的值与该值在总体中发生的概率联系起来
- 两种类型
 - 离散分布：当所测量的参数只能取某些特定值（例如整数0、1、2、...）时，其概率分布称为离散分布（如印刷电路板中的不合格数量或缺陷数的分布就是离散分布）
 - 连续分布：当所测量的变量在连续尺度上表示时，其概率分布称为连续分布（如金属层厚的概率分布是连续的）



例子

- 某制造过程每天生产数千个半导体芯片。平均而言，其中1%的芯片不符合规格。一名检验员每小时随机抽取25个芯片的样本，并将样本中的每个芯片归类为合格或不合格。如果用 x 表示样本中，不合格芯片数量的随机变量，则 x 的概率分布为二项分布

$$p(x) = \binom{25}{x} (0.01)^x (0.99)^{25-x} \quad x = 0, 1, 2, \dots, 25$$

- 假设 x 是一个表示一磅咖啡豆袋的实际含量（以盎司为单位）的随机变量

$$f(x) = \frac{1}{1.5} \quad 15.5 \leq x \leq 17.0$$

均值和方差

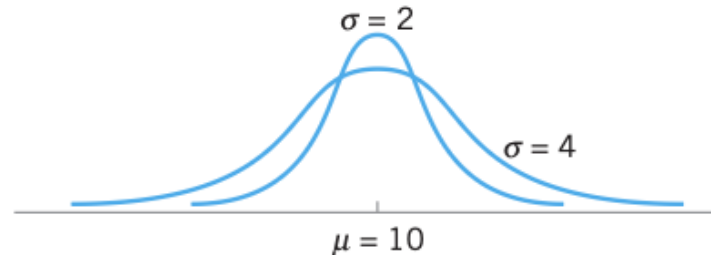
- 概率分布的均值是衡量该分布集中趋势或位置的指标

$$\mu = \begin{cases} \int_{-\infty}^{\infty} xf(x) dx, x \text{ continuous} \\ \sum_{i=1}^{\infty} x_i p(x_i), x \text{ discrete} \end{cases}$$

- 分布中的离散程度、散布度或变异程度由方差表示

$$\sigma^2 = \begin{cases} \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx, x \text{ continuous} \\ \sum_{i=1}^{\infty} (x_i - \mu)^2 p(x_i), x \text{ discrete} \end{cases}$$

- 通常使用方差的平方根，称为标准差



重要的离散分布——超几何分布

- 假设有一个包含 N 个物品的有限总体。其中有一定数量 D 的物品属于某个感兴趣的类别。从总体中不放回地随机抽取 n 个物品的样本，并观察样本中属于该兴趣类别的物品数量（记为 x ）。那么 x 是一个超几何随机变量，其概率分布定义如下。

Definition

The **hypergeometric probability distribution** is

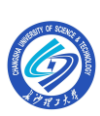
$$p(x) = \frac{\binom{D}{x} \binom{N-D}{n-x}}{\binom{N}{n}} \quad x = 0, 1, 2, \dots, \min(n, D) \quad (3.8)$$

The mean and variance of the distribution are

$$\mu = \frac{nD}{N} \quad (3.9)$$

and

$$\sigma^2 = \frac{nD}{N} \left(1 - \frac{D}{N}\right) \left(\frac{N-n}{N-1}\right) \quad (3.10)$$



重要的离散分布——二项分布

- 考虑一个由 n 次独立试验组成的序列过程。当每次试验的结果要么是“成功”要么是“失败”时，这些试验被称为伯努利试验。如果每次试验中“成功”的概率（设为 p ）保持不变，那么在 n 次伯努利试验中“成功”次数 x 服从参数为 n 和 p 的二项分布，定义如下：

Definition

The **binomial distribution** with parameters $n \geq 0$ and $0 < p < 1$ is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \dots, n \quad (3.11)$$

The mean and variance of the binomial distribution are

$$\mu = np \quad (3.12)$$

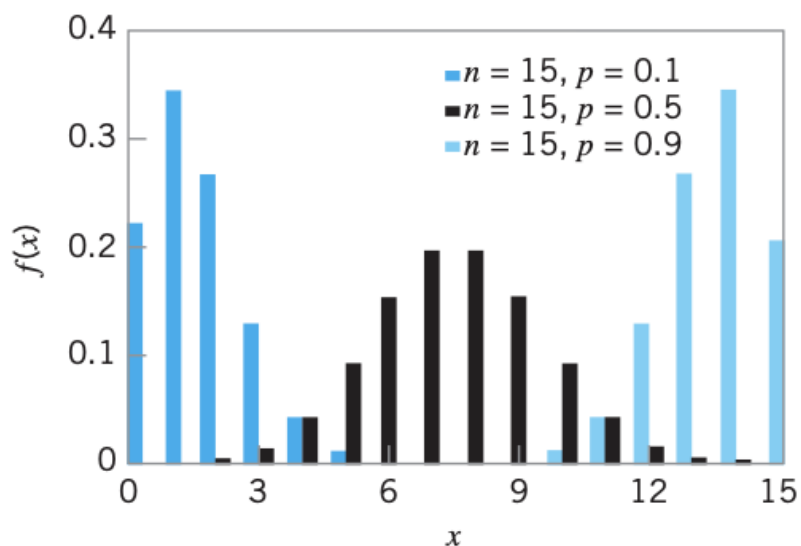
and

$$\sigma^2 = np(1-p) \quad (3.13)$$

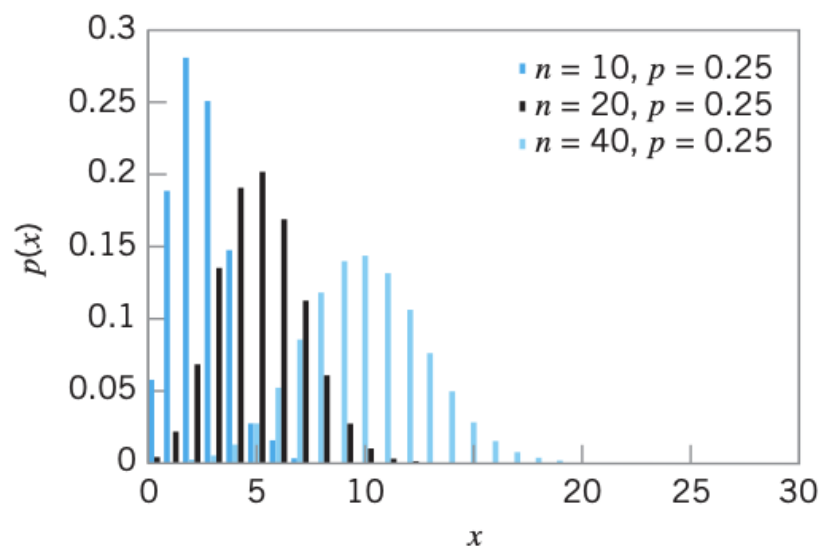


重要的离散分布——二项分布

- 对于固定的 n ，当 p 从0增大到0.5或从1减小到0.5时，分布变得更对称
- 对于固定的 p ，随着 n 增大，分布也变得更对称



(a) Binomial distributions for different values of p with $n = 15$.



(b) Binomial distributions for different values of n with $p = 0.25$.

重要的离散分布——二项分布

- 统计质量控制中经常出现的一个随机变量是

$$\hat{p} = \frac{x}{n}$$

其中 x 服从参数为 n 和 p 的二项分布

- 通常 $\hat{p}$ 是样本中观察到的缺陷项或不合格项数 x 与样本容量 n 的比值，这通常被称为样本缺陷率或样本不合格率

$$P\{\hat{p} \leq a\} = P\left\{\frac{x}{n} \leq a\right\} = P\{x \leq na\} = \sum_{x=0}^{[na]} \binom{n}{x} p^x (1-p)^{n-x}$$

这里 $[na]$ 表示小于或等于 na 的最大整数

重要的离散分布——泊松分布

- 任何以每单位（或每单位面积、每单位体积、每单位时间等）为基础发生的随机现象，通常都可以用泊松分布很好地近似

Definition

The **Poisson distribution** is

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, \dots \quad (3.15)$$

where the parameter $\lambda > 0$. The **mean** and **variance** of the Poisson distribution are

$$\mu = \lambda \quad (3.16)$$

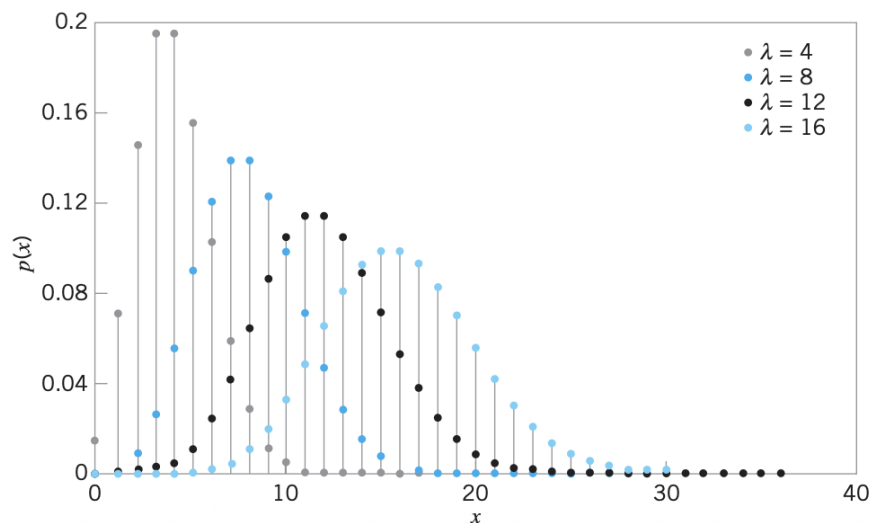
and

$$\sigma^2 = \lambda \quad (3.17)$$



重要的离散分布——泊松分布

- 该分布是偏态的，有一个向右的长尾
 - 随着参数 λ 增大，泊松分布在视觉上变得对称
- 可以将泊松分布推导为二项分布的一种极限形式
 - 在一个参数为 n 和 p 的二项分布中，如果让 n 趋近于无穷大， p 趋近于零，且使得 $np = \lambda$ 为一个常数，那么就会得到泊松分布
 - 也可以通过纯粹的概率论证来推导泊松分布



重要的离散分布——负二项分布与几何分布

- 考虑一系列独立的试验，每次试验的成功概率为 p ，令 x 表示第 r 次成功发生的试验序号。那么 x 是一个负二项随机变量，其概率分布定义如下

Definition

The **negative binomial distribution** is

$$p(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \quad x = r, r+1, r+2, \dots \quad (3.18)$$

where $r \geq 1$ is an integer. The **mean** and **variance** of the negative binomial distribution are

$$\mu = \frac{r}{p} \quad (3.19)$$

and

$$\sigma^2 = \frac{r(1-p)}{p^2} \quad (3.20)$$

respectively.

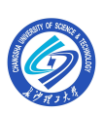


重要的离散分布——负二项分布与几何分布

- 二项分布和负二项分布之间存在一种重要的对偶关系
 - 在二项分布中，我们固定样本量（伯努利试验次数）并观察成功次数
 - 而在负二项分布中，我们固定成功次数并观察达到这些成功所需的样本量（伯努利试验次数）
- 性质
 - 当 λ 取实数值时，这种负二项分布的变体称为波利亚分布
 - 负二项分布的一个有用特例是当 $r = 1$ 时，此时我们得到几何分布。它是在第一次成功之前伯努利试验次数的分布

$$p(x) = (1 - p)^{x-1}p, \quad x = 1, 2, \dots$$

- 几何分布的无记忆性：由于伯努利试验序列是独立的，从任意位置开始计数直到下一次成功所需的试验次数，其概率分布不会改变。意味着被建模的系统不会因为疲劳或累积应力的损耗而失效
- r 个参数都为 p 的几何随机变量之和是一个参数为 p 和 r 的负二项随机变量



重要的连续分布——正态分布

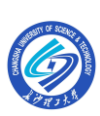
- 正态分布是统计学理论和应用中最重要的分布（如解决客户投诉的时间、某磁盘驱动器中的金属轴直径）

Definition

The **normal distribution** is

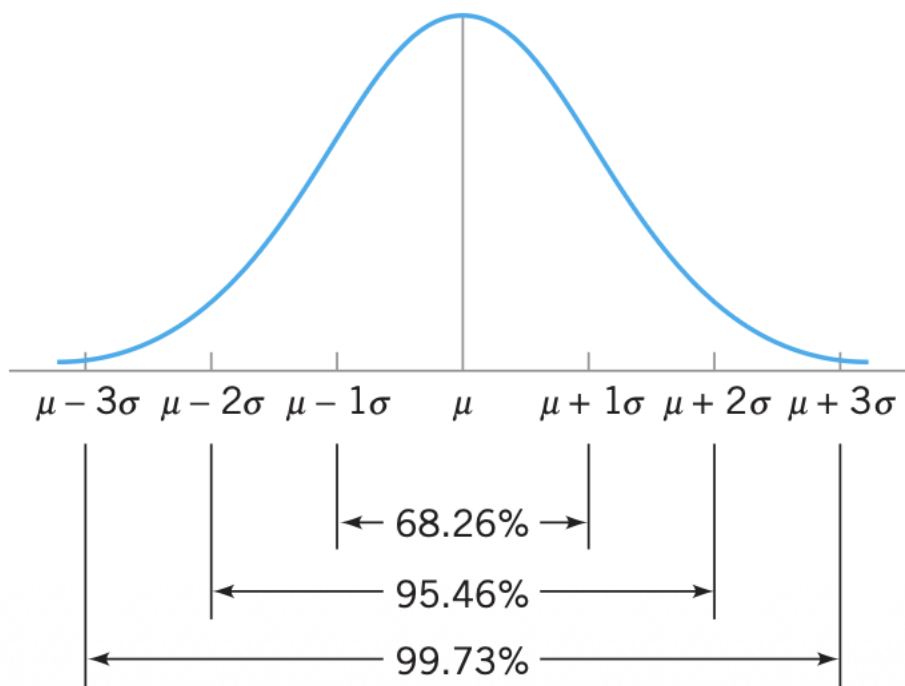
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty \quad (3.21)$$

The mean of the normal distribution is μ ($-\infty < \mu < \infty$) and the variance is $\sigma^2 > 0$.



重要的连续分布——正态分布

- 正态分布的图像是一条对称的、单峰的钟形曲线



- 标准化

$$P\{x \leq a\} = P\left\{z \leq \frac{a - \mu}{\sigma}\right\} \equiv \Phi\left(\frac{a - \mu}{\sigma}\right)$$

重要的连续分布——正态分布

- 中心极限定理常常是近似正态性的一个依据

Definition

The Central Limit Theorem If $x_1, x_2, \dots, x_n$ are independent random variables with mean μ_i and variance σ_i^2 , and if $y = x_1 + x_2 + \dots + x_n$, then the distribution of

$$\frac{y - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}$$

approaches the $N(0, 1)$ distribution as n approaches infinity.

重要的连续分布——对数正态分布

- 当 w 服从正态分布时， $x = \exp(w)$ 的分布称为对数正态分布
- 这一名称来源于变换 $\ln(x) = w$ ，即 x 的自然对数服从正态分布。

$$\begin{aligned} F(a) &= P[x \leq a] = P[\exp(w) \leq a] = P[w \leq \ln(a)] \\ &= P\left[z \leq \frac{\ln(a) - \theta}{\omega}\right] = \Phi\left[\frac{\ln(a) - \theta}{\omega}\right] \end{aligned}$$

Definition

Let w have a normal distribution mean θ and variance ω^2 ; then $x = \exp(w)$ is a **lognormal random variable**, and the lognormal distribution is

$$f(x) = \frac{1}{x\omega\sqrt{2\pi}} \exp\left[-\frac{(\ln(x) - \theta)^2}{2\omega^2}\right] \quad 0 < x < \infty \quad (3.29)$$

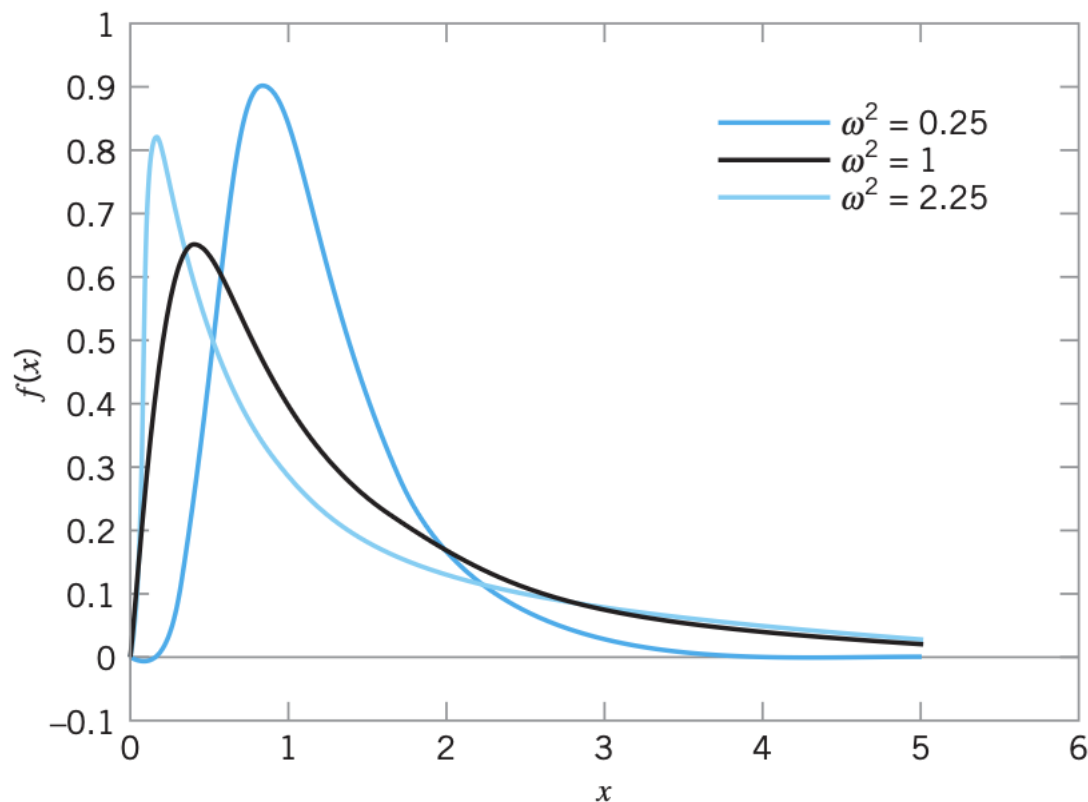
The mean and variance of x are

$$\mu = e^{\theta + \omega^2/2} \quad \text{and} \quad \sigma^2 = e^{2\theta + \omega^2} (e^{\omega^2} - 1) \quad (3.30)$$



重要的连续分布——对数正态分布

- 对数正态分布的参数是 θ 和 ω^2
- 但需要注意，这些参数实际上是正态随机变量 w 的均值和方差



重要的连续分布——指数分布

- 指数随机变量的概率分布定义如下：

Definition

The **exponential distribution** is

$$f(x) = \lambda e^{-\lambda x} \quad x \geq 0 \quad (3.31)$$

where $\lambda > 0$ is a constant. The **mean** and **variance** of the exponential distribution are

$$\mu = \frac{1}{\lambda} \quad (3.32)$$

and

$$\sigma^2 = \frac{1}{\lambda^2} \quad (3.33)$$

respectively.

$$\begin{aligned} F(a) &= P\{x \leq a\} \\ &= \int_0^a \lambda e^{-\lambda t} dt \\ &= 1 - e^{-\lambda a} \quad a \geq 0 \end{aligned}$$



重要的连续分布——指数分布

- 指数分布广泛应用于可靠性工程领域，用于模拟组件或系统失效的时间
 - 在这些应用中，参数 λ 被称为系统的失效率，而分布的均值 $1/\lambda$ 则被称为平均失效时间
- 指数分布与泊松过程之间存在重要关系
 - 如果将泊松分布视为某种事件在区间 $(0, t]$ 内发生次数的模型，如果事件的发生次数服从参数为 λ 的泊松分布，那么事件之间时间间隔的分布就是参数为 λ 的指数分布
- 指数分布具有无记忆性
 - 考虑两个时间点 t_1 和 $t_2 > t_1$ 。那么，事件发生在小于 $t_1 + t_2$ 但大于 t_2 的时间点的概率，就等于事件发生在小于 t_1 的时间点的概率（指数分布是唯一具有这种性质的连续分布）

重要的连续分布——伽马分布

Definition

The **gamma distribution** is

$$f(x) = \frac{\lambda}{\Gamma(r)} (\lambda x)^{r-1} e^{-\lambda x} \quad x \geq 0 \quad (3.36)$$

with **shape parameter** $r > 0$ and **scale parameter** $\lambda > 0$. The **mean** and **variance** of the gamma distribution are

$$\mu = \frac{r}{\lambda} \quad (3.37)$$

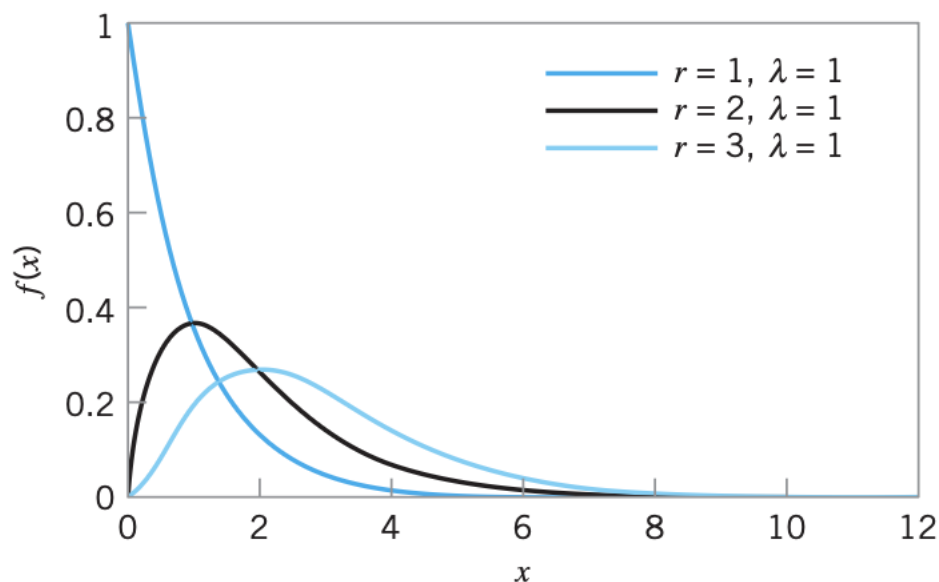
and

$$\sigma^2 = \frac{r}{\lambda^2} \quad (3.38)$$

respectively.³

重要的连续分布——伽马分布

- 伽马分布可以根据 r 和 λ 的不同取值呈现出许多不同的形状，这使得它成为各种连续随机变量的有用模型
 - 如果 $r = 1$ ，则伽马分布退化为参数为 λ 的指数分布
- 如果参数 r 为整数，那么伽马分布就是 r 个独立同分布的指数分布之和，每个指数分布的参数均为 λ
- 把伽马分布看作获得特定事件次数所需时间的模型



重要的连续分布——威布尔分布

- 威布尔分布非常灵活，通过适当选择参数 θ 和 β ，该分布可以呈现出多种多样的形状

Definition

The Weibull distribution is

$$f(x) = \frac{\beta}{\theta} \left(\frac{x}{\theta} \right)^{\beta-1} \exp \left[- \left(\frac{x}{\theta} \right)^{\beta} \right] \quad x \geq 0 \quad (3.41)$$

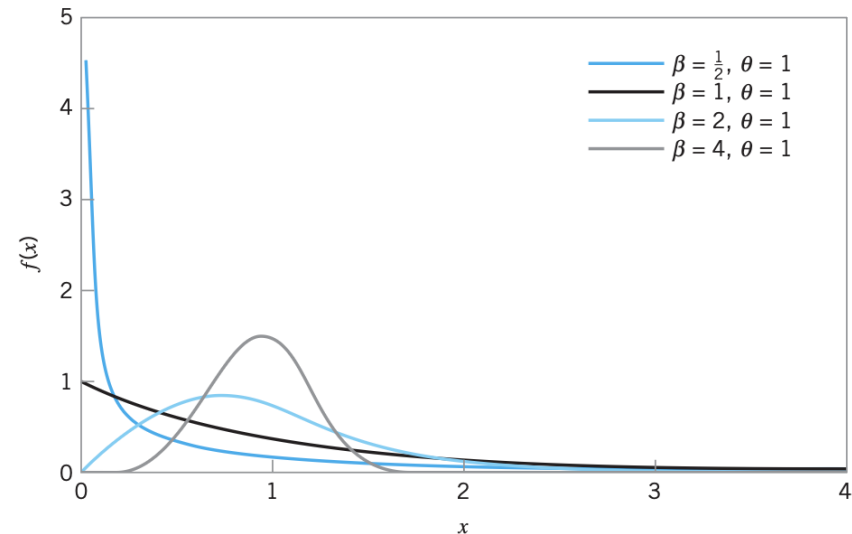
where $\theta > 0$ is the **scale parameter** and $\beta > 0$ is the **shape parameter**. The **mean** and **variance** of the Weibull distribution are

$$\mu = \theta \Gamma \left(1 + \frac{1}{\beta} \right) \quad (3.42)$$

and

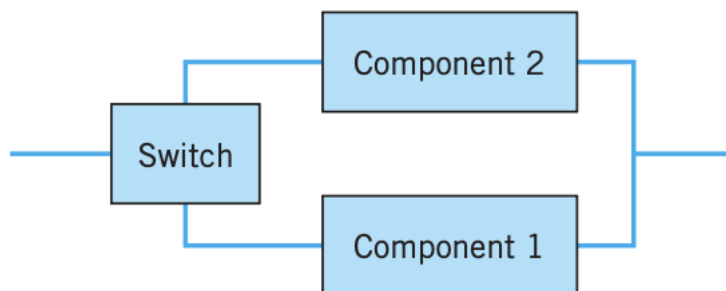
$$\sigma^2 = \theta^2 \left[\Gamma \left(1 + \frac{2}{\beta} \right) - \left\{ \Gamma \left(1 + \frac{1}{\beta} \right) \right\}^2 \right] \quad (3.43)$$

respectively.



例子

- 从 N 件产品（其中 D 件不合格或有缺陷）构成的批次中，不放回地随机抽取 n 件产品时适用的概率模型。
- 对无限大总体进行抽样时的适当概率模型，其中 p 代表总体中缺陷项或不合格项的比例
- 作为单个产品中出现的缺陷或不合格品数量的模型
- 单个产品中不合格项的出现次数
- 随时间退化的产品寿命
- 备用冗余系统



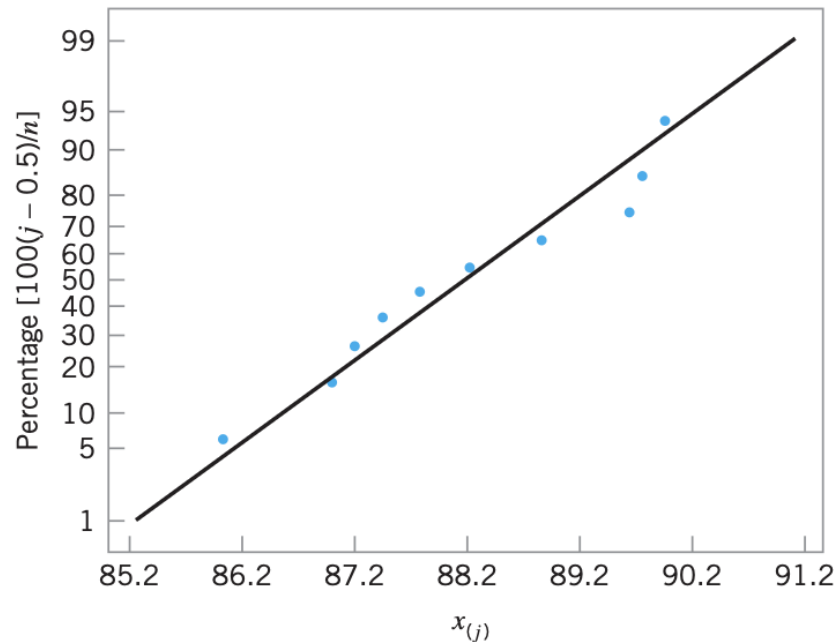
正态概率图

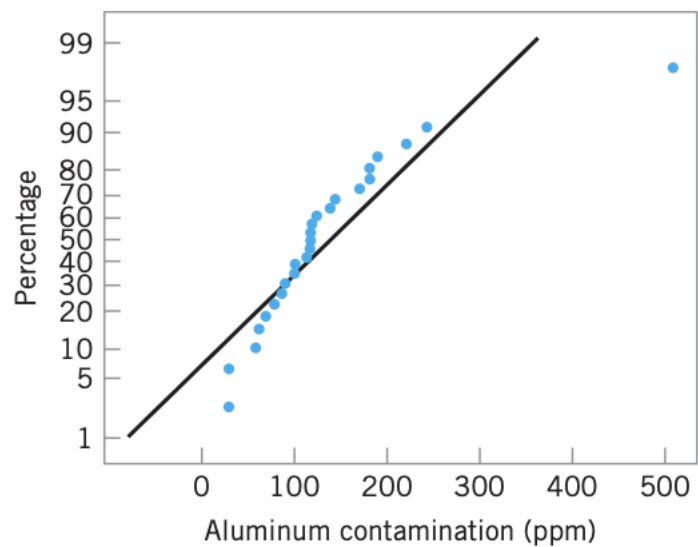
- 通过对数据进行主观的视觉检查，来确定样本数据是否符合假设的分布
- 步骤
 - 首先将样本中的观测值按从小到大的顺序排列
 - 然后将有序的观测值 $x_{(j)}$ 与观测到的累积频率 $\frac{j-0.5}{n}$ （或 $\frac{100(j-0.5)}{n}$ ）相对应地绘制出来
- 判断
 - 如果假设的分布能充分描述数据，则绘制的点将大致落在一条直线上
 - 如果绘制的点显著且系统地偏离一条直线，则假设的模型就不合适
 - 通常，判断数据是否绘成一条直线是主观的



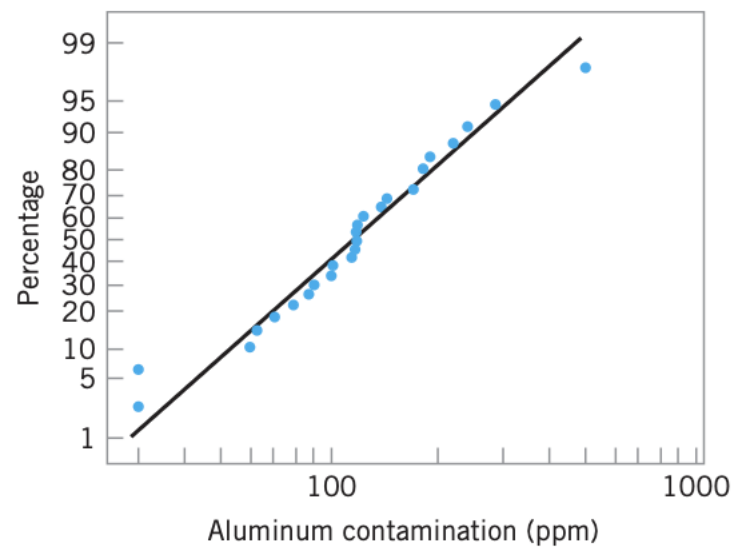
正态概率图

- 在绘制这条直线时，应该更多受图中中间部分点的影响，而不是极端点。一个好的经验法则是大致在第二十五和第七十五百分位数之间画线
- 在评估点相对于直线的系统偏差时，想象一支粗铅笔放在这条线上。如果所有点都被这支想象中的铅笔覆盖，那么正态分布就充分描述了数据。

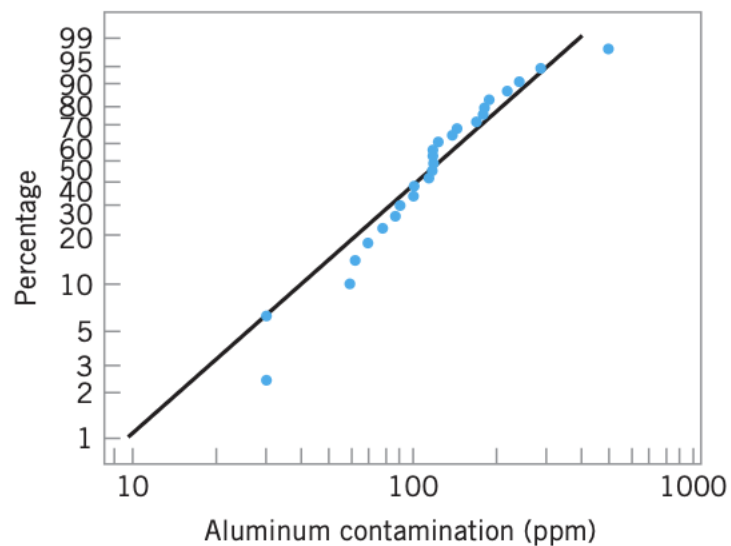




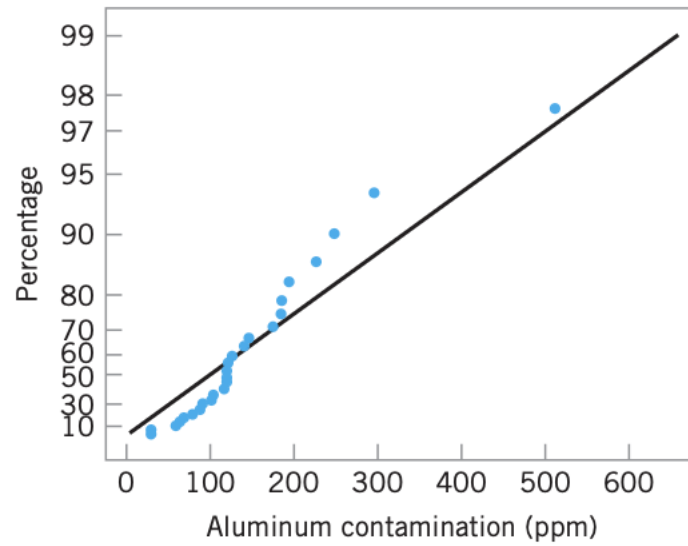
(a)



(b)



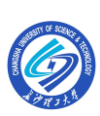
(c)



(d)

近似方法

- 二项分布对超几何分布的近似
 - 如果 n/N 的比率（通常称为抽样分数）很小——比如， $\frac{n}{N} \leq 0.1$ ——那么参数为 $p = D/N$ 和 n 的二项分布是超几何分布的良好近似
 - 对于较小的 n/N 值，近似效果更好
- 泊松分布对二项分布的近似
 - 对于 p 趋近于零且 n 趋近于无穷大且 $\lambda = np$ 保持常数的情况，泊松分布可作为二项分布的极限形式得到
 - 这意味着，当 p 较小且 n 较大时， $\lambda = np$ 的泊松分布可用于近似二项分布。当 n 较大且 $p < 0.1$ 时，该近似通常效果良好。 n 越大， p 越小，近似效果越好



近似方法

- 正态分布对二项分布的近似

$$\begin{aligned}P\{x = a\} &= \binom{n}{a} p^a (1-p)^{n-a} \\&= \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{1}{2}[(a-np)^2 / np(1-p)]}\end{aligned}$$

- 连续性校正

$$P\{x = a\} \cong \Phi\left(\frac{a + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

$$P\{a \leq x \leq b\} \cong \Phi\left(\frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

- 对于 p 大约为 $1/2$ 且 $n > 10$ 的情况，已知正态近似于二项分布是令人满意的。对于其他 p 值，需要更大的 n 值

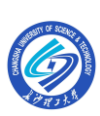


近似方法

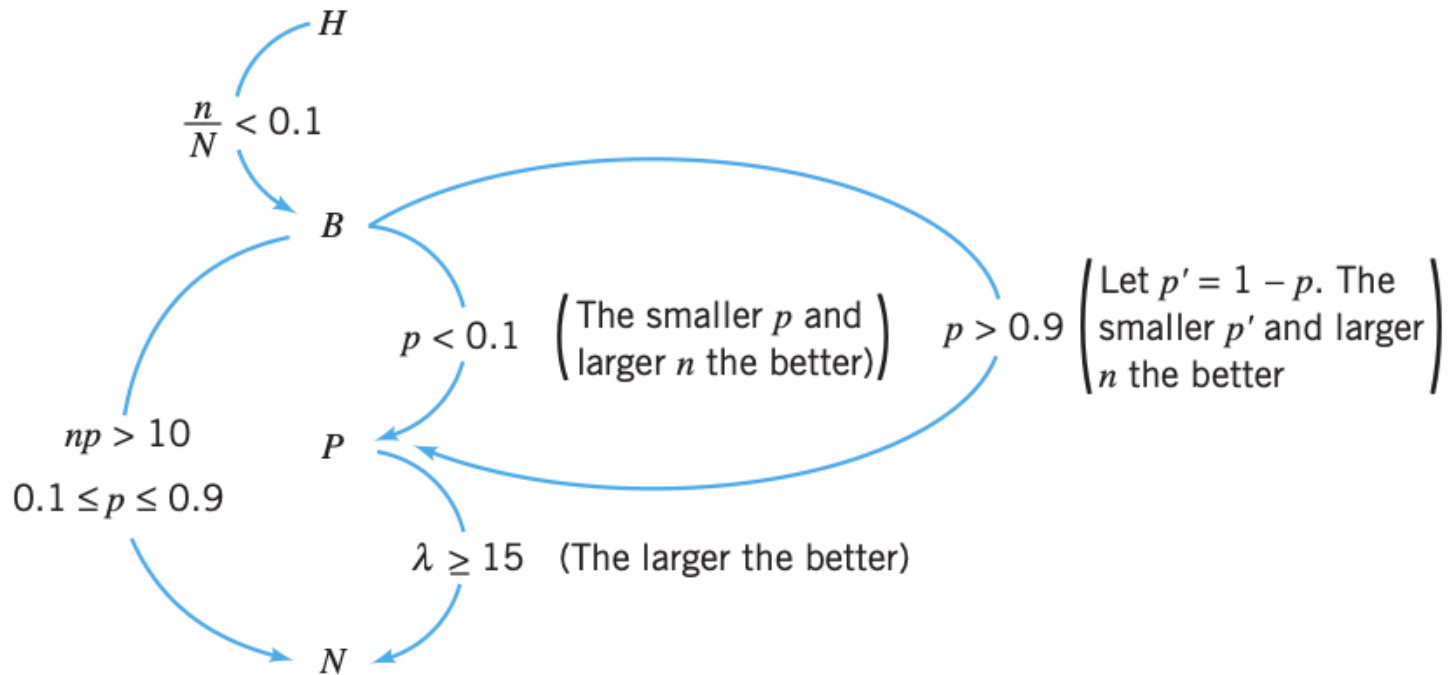
- 随机变量 $\hat{p} = x/n$ ，近似服从正态分布，均值为 p ，方差为 $\frac{p(1-p)}{n}$

$$P\{u \leq \hat{p} \leq v\} \cong \Phi\left(\frac{v-p}{\sqrt{p(1-p)/n}}\right) - \Phi\left(\frac{u-p}{\sqrt{p(1-p)/n}}\right)$$

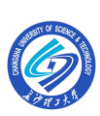
- 如果泊松分布的均值 λ 较大（例如至少**15**），那么均值为 $\mu = \lambda$ 且方差为 $\sigma^2 = \lambda$ 的正态分布就是一个令人满意的近似



近似方法



H、B、P和N分别代表超几何分布、二项分布、泊松分布和正态分布



总结

