



# 数据分析基础

主讲人：刘宏志

liuhz@ss.pku.edu.cn



北京大学



# 导论



北京大学



# Data Analysis

- A process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. –Wiki
- Varieties of data analysis:
  - **Data mining** focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes.
  - **Business intelligence** focuses on business information and relies heavily on aggregation.
  - **Text analytics** applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data.

.....



北京大学



# 数据分析流程



北京大学



# 概率与分布



北京大学



# 样本空间

- 随机试验
  - 随机现象的实现和对它的观察
  - 例：抛一颗骰子，观察出现的点数
- 随机事件
  - 随机试验的每一可能结果称为一个基本事件
  - 一个或一组基本事件统称随机事件，或简称事件
- 样本空间
  - 一个随机试验的所有可能试验结果组成的集合称为该随机试验的样本空间，记为 $\Omega$ ，其元素记为 $\omega$
  - 样本空间又称为基本事件空间
  - 例：抛一颗骰子，出现点数的样本空间为
$$\Omega = \{1, 2, 3, 4, 5, 6\}$$



北京大学



# 概率测度

- 概率

  - 对随机事件发生的可能性的度量

  - 定义在样本空间 $\Omega$ 子集上的实函数

- 公理

  - $P(\Omega) = 1$

  - 如果 $A \subset \Omega$ ， 则 $P(A) \geq 0$

  - 如果  $A \cap B = \emptyset$ , 则 $P(A \cup B) = P(A) + P(B)$

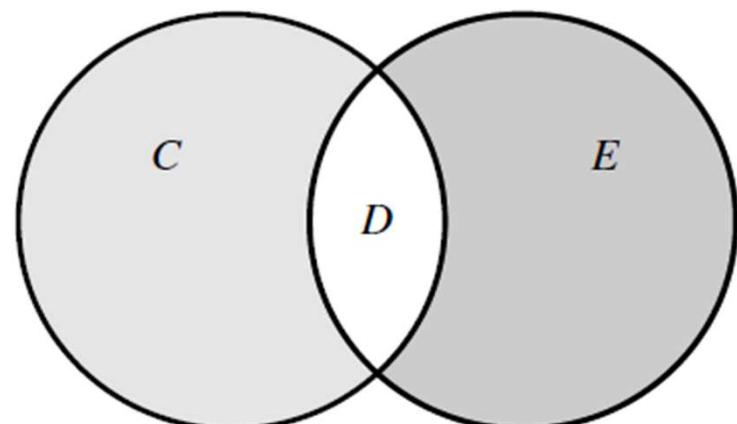


北京大学



## 概率测度的基本性质

- $P(A^c) = 1 - P(A)$ 
  - 因为:  $A \cap A^c = \emptyset$  且  $A \cup A^c = \Omega$
- $P(\emptyset) = 0$ 
  - 因为:  $\emptyset = \Omega^c$  且  $P(\Omega) = 1$
- $A \subset B \Rightarrow P(A) < P(B)$ 
  - 因为:  $B = A + (A^c \cap B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$





## 概率计算：计数法

- 有限样本空间:  $\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}$
- 事件A发生的概率 = A包含的基本事件 $\omega_i$ 发生概率的加和
- 例: 一次掷色子的点数小于3的概率是多少?
  - $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - $A = \{1, 2\}$ ,  $P(A) = 1/3$
- 若 $\Omega$ 中的元素具有等概率性, 事件A通过多个互斥途径中的任一种方式方法, 则

$$P(A) = \frac{\text{导致 } A \text{发生的方式个数}}{\text{所有试验结果个数}}$$

注意: 此公式仅当所有实验结果是等可能发生时才成立



北京大学



# 案例：抽奖游戏

- 游戏1

- 一黑盒中装有5个红球和6个绿球，
- 一白盒中装有3个红球和4个绿球。
- 允许你选择一个盒子并随机从中选一个球，
- 如果选中红球，则会有一份奖品。
- 你会选择从哪个盒子中取球？

- 游戏2

- 一个黑盒中装有6个红球和3个绿球，
- 一个白盒中装有9个红球和5个绿球。
- 规则不变，你会选择从哪个盒子中取球？

- 游戏3

- 将游戏2中黑盒和白盒中球倒入游戏1中对应颜色的盒中。
- 规则不变，你会选择从哪个盒子中取球？

**辛普森悖论：**在某个条件下的两组数据，分别讨论时都会满足某种性质，可是一旦合并考虑，却可能导致相反的结论。



北京大学



## 案例：野生动物总数估算

- 假设捕捉10个动物，将它们做上标记后释放。这之后，再捕捉20个动物，发现有4个带有标记。动物的总数是多少？
- 设动物总数为 $n$ ，则第2次捕捉中含4个带标记动物的概率为：

$$\frac{\binom{10}{4} \binom{n-10}{16}}{\binom{n}{20}}$$

- 观察结果发生的概率是待估参数 $n$ 的函数，称为似然
- 最大似然估计：将使观测结果出现可能性最大的 $n$ 作为估计值

**标记重捕法：**假设（1）调查期间**数量稳定**；（2）标记个体**均匀分布**在全部个体之中；（3）标记操作**不影响**动物的行为和死亡。



北京大学



# 条件概率

- 事件A在另一事件B已经发生条件下发生的概率

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

- 乘法定律:  $P(A \cap B) = P(A|B)P(B)$
- 全概率定律:

令 $B_1, B_2, \dots, B_n$ 满足 $\bigcup_{i=1}^n B_i = \Omega$ ,  $B_i \cap B_j = \emptyset$ ,  
 $i \neq j$ , 且对所有的*i*,  $P(B_i) > 0$ 。则对任意A,

$$P(A) = \sum_{i=1}^n P(A|B_i) P(B_i)$$



北京大学



# 条件概率

- 贝叶斯公式：

令 $A$ 和 $B_1, B_2, \dots, B_n$ 是事件，其中 $\bigcup_{i=1}^n B_i = \Omega$ ,  
 $B_i \cap B_j = \emptyset, i \neq j$ , 且对所有的 $i$ ,  $P(B_i) > 0$ 。则

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^n P(A|B_i)P(B_i)}$$



北京大学



# 事件独立性

- 两事件相互独立：
  - A, B是两事件， $P(AB)=P(A)P(B)$
- 两两独立：
  - A,B,C三个事件， $P(AB)=P(A)P(B)$ ,  
 $P(AC)=P(A)P(C)$ ,  $P(BC)=P(B)P(C)$
- 事件集相互独立：
  - 事件集 $A_1, A_2, \dots, A_n$ , 任意事件子集 $A_{i1}, A_{i2}, \dots, A_{im}$ ,  
满足 $P(A_{i1}A_{i2}\dots A_{im})=P(A_{i1})P(A_{i2})\dots P(A_{im})$



北京大学



# 随机变量

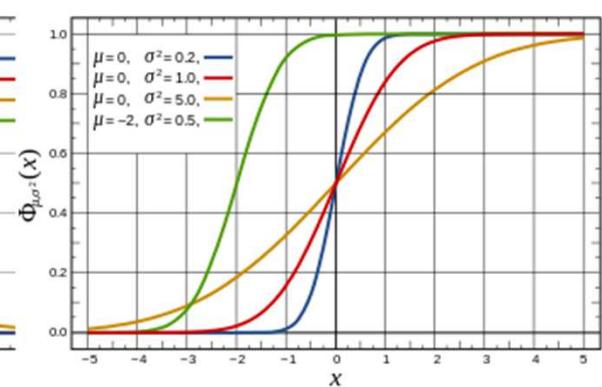
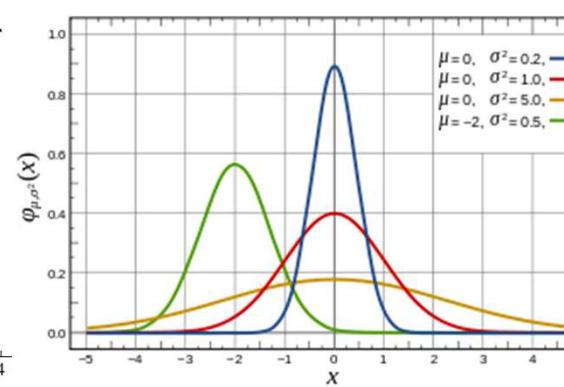
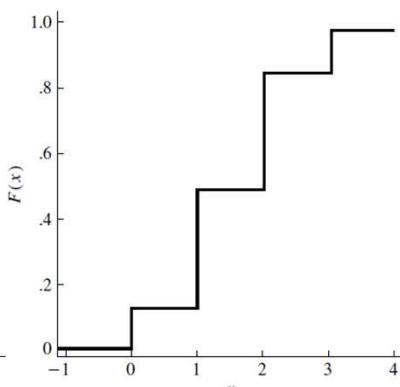
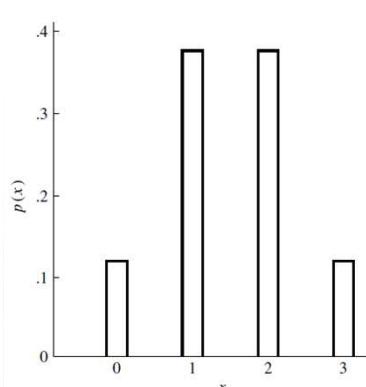


北京大学



# 随机变量

- 一个定义在样本空间 $\Omega$ 上的**实值函数**
  - 定义了每个样本点 $\omega \in \Omega$ 取值实数的法则
- 本质上是一个**随机数**
  - $\Omega$ 中试验结果的发生是随机的，相应的取值也是随机的
- 概率函数
  - 质量函数（PMF）： $p(x_i) = P(X=x_i)$  （离散型随机变量）
  - 密度函数（PDF）： $f(x)$  （连续型随机变量）
- 累积分布函数（CDF）： $F(x) = P(X \leq x)$ 
  - 随机变量小于或者等于某个数值的概率 $P(X \leq x)$





# 常见分布

- 离散分布：
  - 伯努利分布
  - 二项分布
  - 几何分布
  - 泊松分布

- 连续分布：
  - 均匀分布
  - 指数分布
  - 伽马分布
  - 正态分布
  - 贝塔分布



北京大学



# 期望和方差



北京大学



# 期望的定义

- 离散型随机变量

➤如果X是频率函数为 $p(x)$ 的离散型随机变量，且满足 $\sum_i |x_i| p(x_i) < \infty$ ，则X的期望为：

$$E(X) = \sum_i x_i p(x_i)$$

➤如果和式发散，则期望无定义

- 连续型随机变量

➤如果X是密度函数为 $f(x)$ 的连续型随机变量，且满足 $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$ ，则X的期望为：

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

➤如果积分发散，则期望无定义



北京大学



## 马尔可夫不等式

- 定理：如果随机变量 $X$ 满足 $P(X \geq 0) = 1$ ，且 $E(X)$ 存在，则对任意 $t > 0$ ,

$$P(X \geq t) \leq E(X)/t .$$

- 证明：
$$E(X) = \sum_x xp(x) = \sum_{x < t} xp(x) + \sum_{x \geq t} xp(x)$$

$$E(X) \geq \sum_{x \geq t} xp(x) \geq \sum_{x \geq t} tp(x) = tP(X \geq t)$$

- 变形：令 $t = kE(X)$ ，则 $P(X \geq kE(X)) \leq 1/k$
- 应用：

➤ 不超过 $1/5$ 的人口会有超过 $5$ 倍于人均收入的收入



北京大学



# 随机变量函数的期望

- 假设  $Y=g(X)$ .

如果  $X$  是频率函数为  $p(x)$  的离散型随机变量，且满足  $\sum |g(x)|p(x) < \infty$ ，则：

$$E(Y) = \sum_x g(x)p(x)$$

如果  $X$  是密度函数为  $f(x)$  的连续型随机变量，且满足  $\int_{-\infty}^{\infty} |g(x)|f(x) dx < \infty$ ，则：

$$E(Y) = \int_{-\infty}^{\infty} g(x)f(x) dx$$

注意：  $E[g(X)] \neq g[E(X)]$ ，即函数的期望不等于期望的函数



北京大学



# 随机变量线性组合的期望

- 定理：如果  $X_1, \dots, X_n$  是具有期望  $E(X_i)$  的联合分布随机变量，  
 $Y$  是  $X_i$  的的线性函数  $Y = a + \sum_{i=1}^n b_i X_i$ ，则：

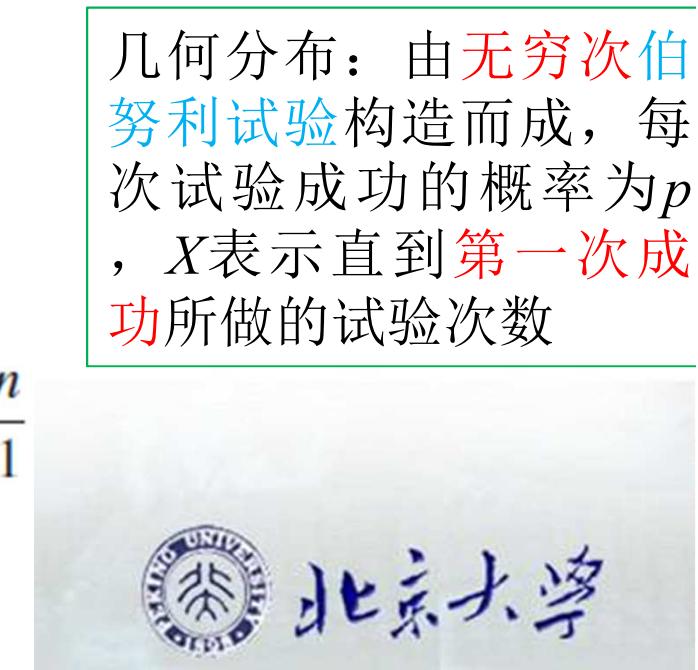
$$E(Y) = a + \sum_{i=1}^n b_i E(X_i)$$

- 例：假设你收集购物券，共有  $n$  种不同类型的购物券，且每次试验你都可以等可能地得到任一类型的购物券。你期望多少次试验才能收集到完整的券集？

- 令  $X_{i+1}$  表示在收集了  $i$  种购物券后出现第  $i+1$  类购物券所需的次数
- $X_r$  的分布为：几何分布
- 每次试验成功的概率为：  $p = (n-r+1)/n$
- $E(X_r) = 1/p = n/(n-r+1)$

$$\begin{aligned} E(X) &= \sum_{r=1}^n E(X_r) \\ &= \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1} \\ &= n \sum_{r=1}^n \frac{1}{r} \end{aligned}$$

几何分布：由无穷次伯努利试验构造而成，每次试验成功的概率为  $p$ ， $X$  表示直到第一次成功所做的试验次数





# 方差和标准差

- 随机变量的标准差描述分布关于中心的发散程度，度量随机变量偏离期望的平均幅度
- 定义：如果 $X$ 是具有期望 $E(X)$ 的随机变量，只要下述期望存在，则 $X$ 的方差为：

$$\text{Var}(X) = E\{[X - E(X)]^2\}$$

- $X$ 的标准差是方差的平方根
  - 定理1：如果 $\text{Var}(X)$ 存在， $Y = a + bx$ , 则
- $$\text{Var}(Y) = b^2 \text{Var}(X)$$
- 定理2：如果 $X$ 的方差存在，也可计算如下：

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$



北京大学



## 切比雪夫不等式

- 定理：令 $X$ 是均值为 $\mu$ , 方差为 $\sigma^2$ 的随机变量，则对任意 $t>0$ ,  $P(|X-\mu|\geq t)\leq \sigma^2/t^2$ .
- 设定 $t=k\sigma$ , 不等式变为

$$P(|X-\mu|\geq k\sigma) \leq 1/k^2$$

- 推论：如果 $\text{Var}(X)=0$ , 则 $P(X=\mu)=1$
- 应用：在所有数据中
  - 至少有 $3/4$ （或 $75\%$ ）的数据位于期望值2个标准差范围内
  - 至少有 $8/9$ （或 $88.9\%$ ）的数据位于期望值3个标准差范围内
  - 至少有 $24/25$ （或 $96\%$ ）的数据位于期望值5个标准差范围内



北京大学



# 协方差

- 两个随机变量的协方差 (covariance) 是它们联合变异性 的度量，或是它们关联性的度量
- 定义：如果  $X$  和  $Y$  是分别具有期望  $\mu_X$  和  $\mu_Y$  的随机变量，只要下述期望存在，则  $X$  和  $Y$  的协方差为

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

- 性质：
  - 两个随机变量的关联是正向的，则协方差为正
  - 两个随机变量的关联是反向的，则协方差为负
  - 两个随机变量是独立的，则协方差为0 (无关的)
  - $\text{Cov}(X, X) = \text{Var}(X)$



北京大学



# 相关系数

- 定义：如果 $X$ 和 $Y$ 的方差和协方差都存在，且方差非零，则 $X$ 和 $Y$ 的相关系数为：

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

- 性质：

➤  $-1 \leq \rho \leq 1$ .  $\rho = \pm 1$  当且仅当  $P(Y = a + bX) = 1$ , 其中  $a$  和  $b$  为常数

➤ 相关系数无量纲

➤ 如果 $X$ 和 $Y$ 进行线性变换，相关系数保持不变



北京大学



# 参数估计

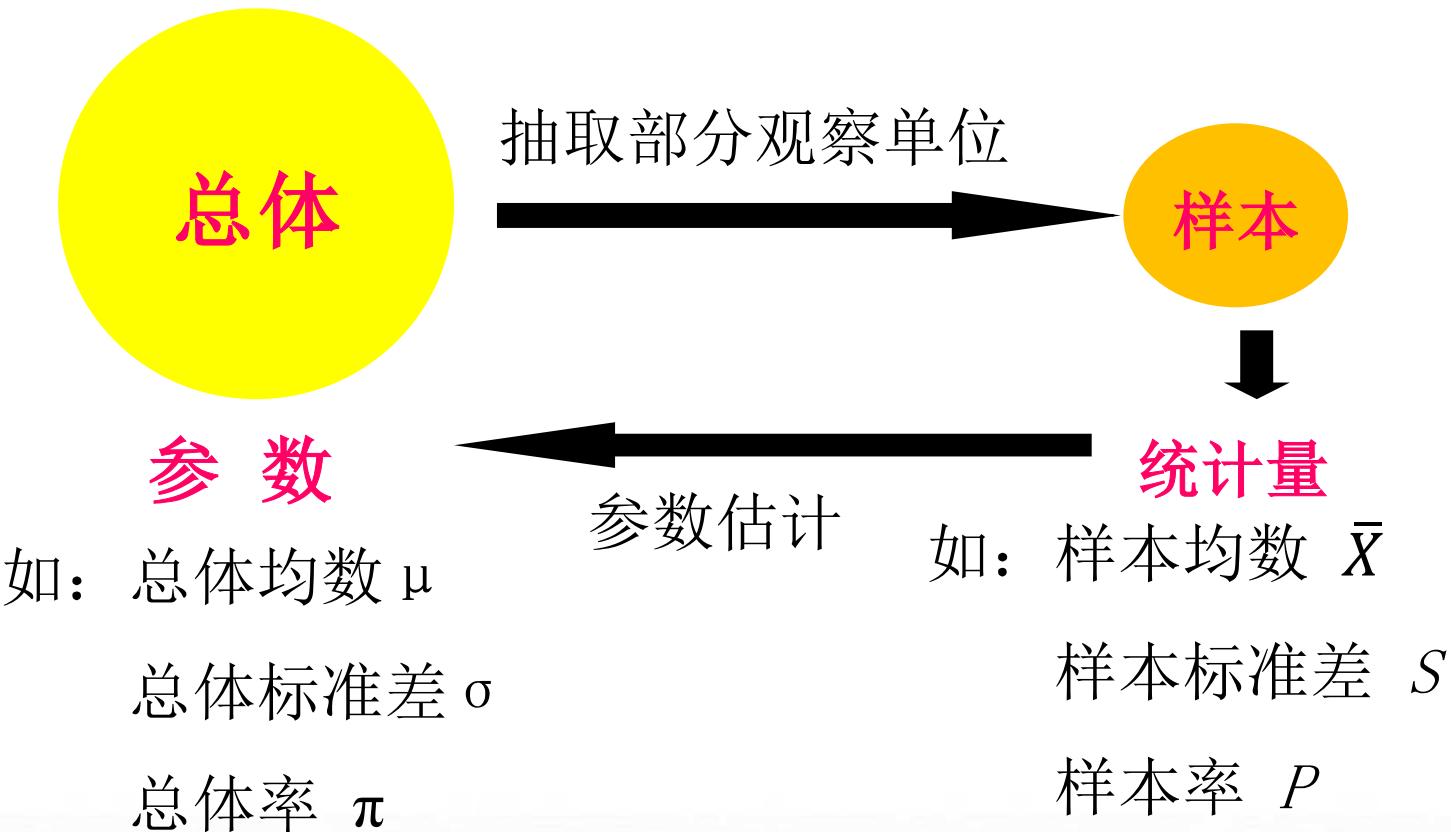


北京大学



# 参数估计的基本思想

- 用所获得的样本值去估计参数取值

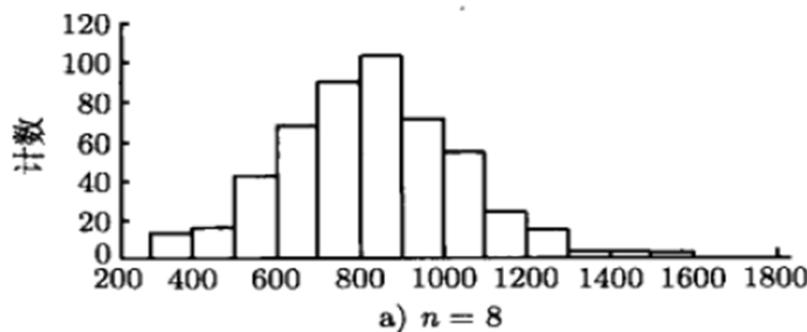


北京大学

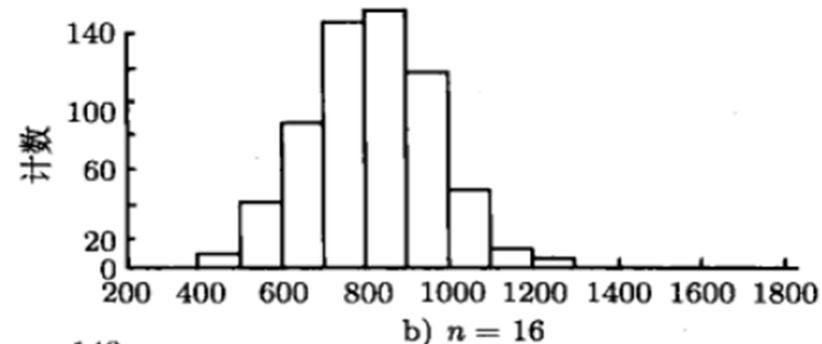


# 样本统计量

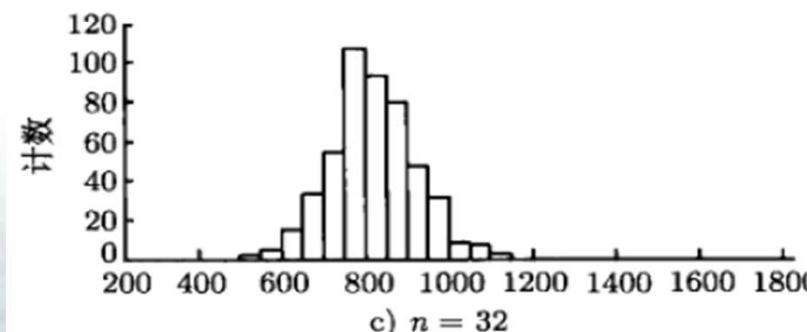
- 样本统计量是样本的函数，是随机变量，其分布称为抽样分布
  - 随机样本得到的数值或统计量都是随机的
  - 样本均值： $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
  - 样本总数： $T=N\bar{X}$
  - 样本方差： $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
- 例：以393个医院为总体，利用容量为 $n$ 的样本进行参数估计



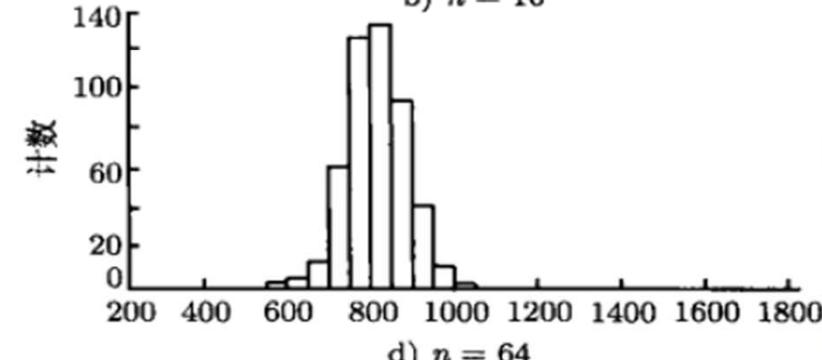
a)  $n = 8$



b)  $n = 16$



c)  $n = 32$



d)  $n = 64$

图 7.2 自 393 个医院总体中进行 500 次简单随机抽样时，不同样本容量下出院人数均值的直方图



## 参数估计方法分类

参数估计 { 点估计  
区间估计 }

用某一数值作为  
参数的近似值

在要求的精度范围内  
指出参数所在的区间

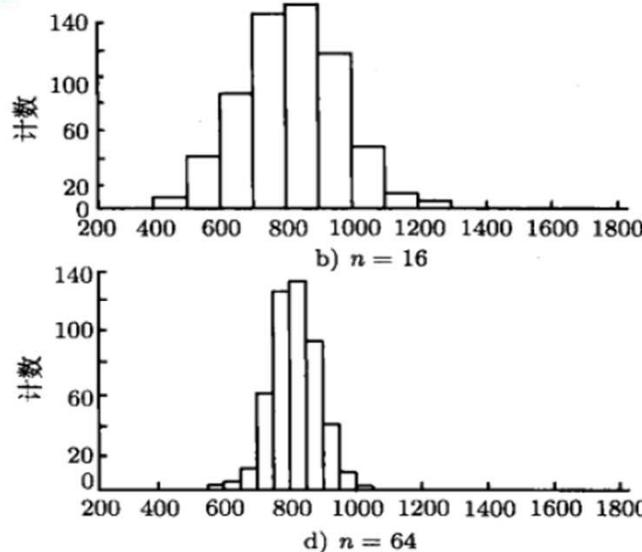
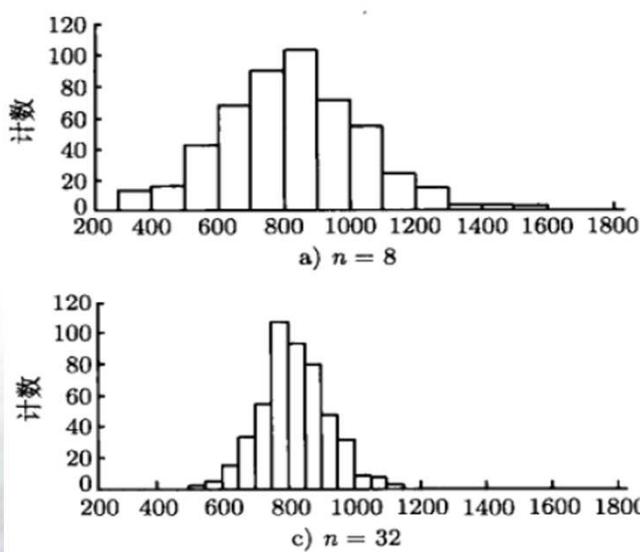


图 7.2 自 393 个医院总体中进行 500 次简单随机抽样时，不同样本容量下出院人数均值的直方图

北京大学



# 点估计

- 设 $\theta$ 是总体 $X$ 的未知参数
- 利用样本 $X_1, X_2, \dots, X_n$ 构造一个统计量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 来估计 $\theta$
- $\hat{\theta}(X_1, X_2, \dots, X_n)$ 称为 $\theta$ 的点估计量，是一个随机变量
- 将样本观测值( $x_1, x_2, \dots, x_n$ )代入估计量 $\hat{\theta}(X_1, X_2, \dots, X_n)$ ，得到一个具体数值 $\hat{\theta}(x_1, x_2, \dots, x_n)$ ，称为 $\theta$ 的点估计值
- 如果总体 $X$ 有 $m$ 个未知参数需要估计，则需要构造 $m$ 个统计量分别作为对每个参数的估计
- 常用方法：矩估计法、最大似然法



北京大学



# 矩估计法

- 基本思想：简单的“替换”
  - 用样本矩估计总体矩
- 原点矩
  - 定义：随机变量 $X$ 的 $k$ 次幂的数学期望( $k$ 为正整数)称为随机变量 $X$ 的 $k$ 阶原点矩： $\mu_k(X) = E(X^k)$
  - 一阶原点矩即为数学期望： $\mu_1(X) = E(X)$
- 中心矩
  - 定义：随机变量 $X$ 的离差的 $k$ 次幂的数学期望( $k$ 为正整数)称为随机变量 $X$ 的 $k$ 阶中心矩： $\gamma_k(X) = E\{[X - E(X)]^k\}$
  - 一阶中心矩恒等于零： $\gamma_1(X) = 0$
  - 二阶中心矩即为方差： $\gamma_2(X) = D(X)$
  - 三阶中心矩 $E\{[X - E(X)]^3\}$ 主要衡量随机变量的分布是否有偏



北京大学



# 最大似然估计

- 基本思想：  
根据样本值来选择参数，使该样本发生的概率最大
- **案例1：**某位同学与一位猎人一起外出打猎。一只野兔从前方窜过。只听一声枪响，野兔应声倒下。如果要你推测，是谁打中的？你会如何想？



**分析：**只发一枪便打中，猎人命中的概率大于这位同学命中的概率。

**结论：**看来这一枪是猎人射中的。





# 似然函数

定义：设总体  $X$  的分布类型已知，但含有未知参数  $\theta$ .

(1) 设离散型总体  $X$  的概率分布律为  $p(x; \theta)$ ，则样本  $(X_1, X_2, \dots, X_n)$  的联合分布律

$$p(x_1; \theta)p(x_2; \theta)\cdots p(x_n; \theta) = \prod_{i=1}^n p(x_i; \theta)$$

称为似然函数，并记之为  $L(\theta) = L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n p(x_i; \theta)$ .

(2) 设连续型总体  $X$  的概率密度函数为  $f(x; \theta)$ ，则样本  $(X_1, X_2, \dots, X_n)$  的联合概率密度函数

$$f(x_1; \theta)f(x_2; \theta)\cdots f(x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

仍称为似然函数，并记之为  $L(\theta) = L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$ .



北京大学



# 极大似然估计值/量

定义：设总体的分布类型已知，但含有未知参数  $\theta$ .

(1) 设  $(x_1, x_2, \dots, x_n)$  为总体  $X$  的一个样本观察值，若似然函数  $L(\theta)$  在  $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$  处取到最大值，则称  $\hat{\theta}(x_1, x_2, \dots, x_n)$  为  $\theta$  的 极大似然估计值.

(2) 设  $(X_1, X_2, \dots, X_n)$  为总体  $X$  的一个样本，若  $\hat{\theta}(x_1, x_2, \dots, x_n)$  为  $\theta$  的极大似然估计值，则称  $\hat{\theta}(X_1, X_2, \dots, X_n)$  为参数  $\theta$  的 极大似然估计量.





# 最大似然估计

一般步骤：

(1) 求似然函数  $L(\theta)$ ；

(2) 求出  $\ln L(\theta)$  及方程  $\frac{d}{d\theta} \ln L(\theta) = 0$ ；

(3) 解上述方程得到极大似然估计值

$$\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n).$$

(4) 解上述方程得到极大似然估计量

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n).$$



北京大学



## 案例：池塘鱼数估计

- 池中有许多鱼，捉住500条，做上标记再放入池中，待充分混合后，再捉1000条鱼，其中100条鱼带记号。试估计池中有多少条鱼？
- 分析：设池中鱼的总条数（待估计量）为 $N$ ，其中 $r$ 条鱼有标记，随机捉 $s$ 条鱼发现有 $x$ 条带标记；用 $X$ 记捉住 $s$ 条鱼中带标记的鱼数

$$P(X = x) = \binom{N - r}{s - x} \binom{r}{x} / \binom{N}{s}$$

- 似然函数 $L(N)=P(X=x)$

$$\frac{L(N)}{L(N-1)} = \frac{(N-s)(N-r)}{(N+x-r-s)N} = \frac{N^2 - (r+s)N + rs}{N^2 - (r+s)N + xN}$$

- 当 $rs > xN$ 时， $L(N) > L(N-1)$ ；当 $rs < xN$ 时， $L(N) < L(N-1)$
- 故 $N$ 的极大似然估计量为 $\hat{N} = \left[ \frac{rs}{x} \right]$ ，代入数值的 $\hat{N} = 5000$



北京大学



# 估计量的评选标准

- 动机：
  - 对同一参数，用不同方法可得到不同的估计量
  - 同一参数的多个估计量，哪个更好？
- 评价标准：估计量的统计性质
  - 无偏性
  - 有效性
  - 一致性



北京大学



# 无偏性

- 动机：
  - 估计量是随机变量，不同样本值会得到不同估计值
  - 希望估计值在真值附近波动，且其期望等于未知参数的真值
- 定义：设 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 为 $\theta \in \Theta$ 的估计量。  
若  $E[\hat{\theta}(X_1, X_2, \dots, X_n)] = \theta, \forall \theta \in \Theta$   
则称 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 为 $\theta$ 的无偏估计. 否则称为有偏的
- 含义： $E(\hat{\theta}) = \theta$ 
  - $\hat{\theta}$ 的平均值恰好等于 $\theta$ 的真值，即要求没有系统误差
  - 用一台秤去称物体，误差源有两个：一是秤本身制作结构上的问题，属于系统误差；另一种是操作或其它随机因素的干扰，属于随机误差



北京大学



# 有效性

- 动机:

- 无偏性只保证了估计量取值在参数真值周围波动
- 但未考虑**波动幅度**的大小
- 一个参数的无偏估计量**不是唯一的**
- 希望波动的幅度**越小越好**
- **方差**是随机变量取值与其期望的偏离程度的度量

- 定义:

设  $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \dots, X_n)$  与  $\hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \dots, X_n)$

都是  $\theta$  的**无偏估计量**, 若对  $\forall \theta \in \Theta$  有  $D(\hat{\theta}_1) \leq D(\hat{\theta}_2)$ , 且至少有一个  $\theta \in \Theta$  使不等式成立, 则称  $\hat{\theta}_1$  比  $\hat{\theta}_2$  有**较高的效率**, 简称  $\hat{\theta}_1$  比  $\hat{\theta}_2$  有效.



北京大学



# 一致性

- 动机：
  - 无偏性和有效性都是在样本容量固定的前提下提出的
  - 希望随着样本容量增大，估计值稳定于待估参数的真值

- 定义：

设总体 $X$ 有概率函数 $p(x; \theta)$ ,  $\theta \in \Theta$ 为待估参数,

$\hat{\theta}_n(X_1, X_2, \dots, X_n)$ 为 $\theta$ 的估计量. 若对于任意 $\varepsilon > 0$ , 总有

$$\lim_{n \rightarrow \infty} P\{|\hat{\theta}_n - \theta| > \varepsilon\} = 0$$

则称 $\hat{\theta}_n(X_1, X_2, \dots, X_n)$ 为参数 $\theta$ 的一致估计.



北京大学



# 区间估计

- 动机：点估计不能反映估计的误差和精确程度
  - 区间估计利用样本统计量和抽样分布估计总体参数的可能区间
- 抽样误差：
  - 一个无偏估计与其对应的总体参数之差的绝对值。
  - 例：均值的抽样误差  $e = |\bar{x} - \mu|$  (实际未知)
  - 区间估计的关键是对抽样误差  $e$  进行求解
  - 若  $e$  已知，则均值的区间可表示为：  $[\bar{x} - e, \bar{x} + e]$
- 正态分布的标准方差  $\sigma$  已知时，均值的区间估计：  $\bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ 
  - $(1 - \alpha)$  为置信系数
  - $Z_{\alpha/2}$  为在标准正态分布的右侧尾部所提供的面积为  $\alpha/2$  的  $Z$  值
- 标准方差  $\sigma$  未知时，均值的区间估计为：  $\bar{x} \pm Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$



北京大学



# 假设检验



北京大学



# 假设检验

- 根据样本信息检验关于总体的某个假设是否正确
- 逻辑上运用反证法，统计上依据小概率原理
- 小概率：
  - 在一次试验中，一个几乎不可能发生的事件发生的概率
  - 概率是0~1之间的一个数，因此小概率就是接近0的一个数
  - R. Fisher 把 $1/20$ 作为标准，即比 $0.05$ 小的概率被认为是小概率
- 小概率原理：
  - 小概率事件在一次实验中几乎是不可能发生的





# 假设检验

- 基本思想：
  - 在一次试验中小概率事件发生，就有理由拒绝原假设
- 一般步骤：
  - 首先假设总体参数某项取值为真，即其发生可能性很大
  - 然后抽取一个样本进行观察
  - 如果样本信息显示出与事先假设相反的结果且差别很大，则说明原来假定的小概率事件在一次实验中发生了
  - 这是一个违背小概率原理的不合理现象，因此有理由怀疑和拒绝原假设
  - 否则不能拒绝原假设

不能拒绝原假设是否意味原假设为真？



北京大学



# 假设检验

假设检验 {

参数假设检验

非参数假设检验

总体分布已知，检验关于未知参数的某个假设

总体分布未知时的假设检验问题



北京大学



# 假设检验的形式

## 原假设(null hypothesis)

1. 研究者想收集证据予以反对的假设
2. 又称“0假设”
3. 总是有符号 $=$ ,  $\leq$  或  $\geq$
4. 表示为  $H_0$

例如,  $H_0: \mu = 10\text{cm}$

## 备择假设(alternative hypothesis)

1. 研究者想收集证据予以支持的假设
2. 也称“研究假设”
3. 总是有符号 $\neq$ ,  $<$ 或 $>$
4. 表示为  $H_1$

例如,  $H_1: \mu \neq 10\text{cm}$

注意: 原假设总是有等号:  $=$ 或 $\leq$ 或 $\geq$ 。



北京大学



# 假设检验的两类错误

第一类错误： 拒绝正确的原假设，简称“**拒真”**；

第二类错误： 接受错误的原假设，简称“**纳伪”**

		总体情况	
		$H_0$ 正确	$H_0$ 错误
结论	接受 $H_0$	正确结论	第二类错误
	拒绝 $H_0$	第一类错误	正确结论

两类错误发生的概率表示如下：

$\alpha$  —— 第一类错误发生的概率，被称为**显著性水平**

$\beta$  —— 第二类错误发生的概率；

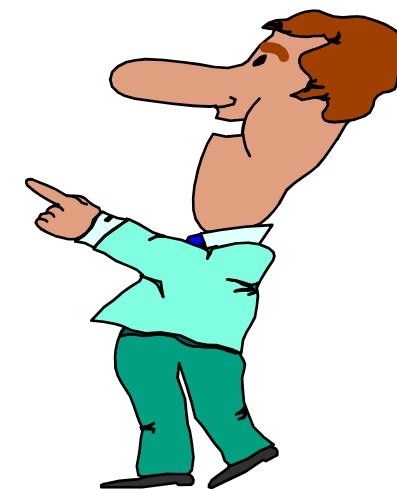
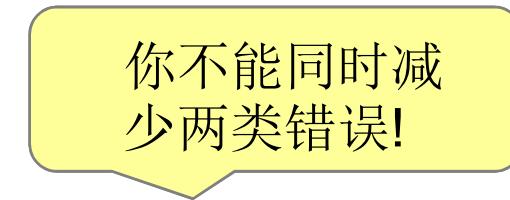
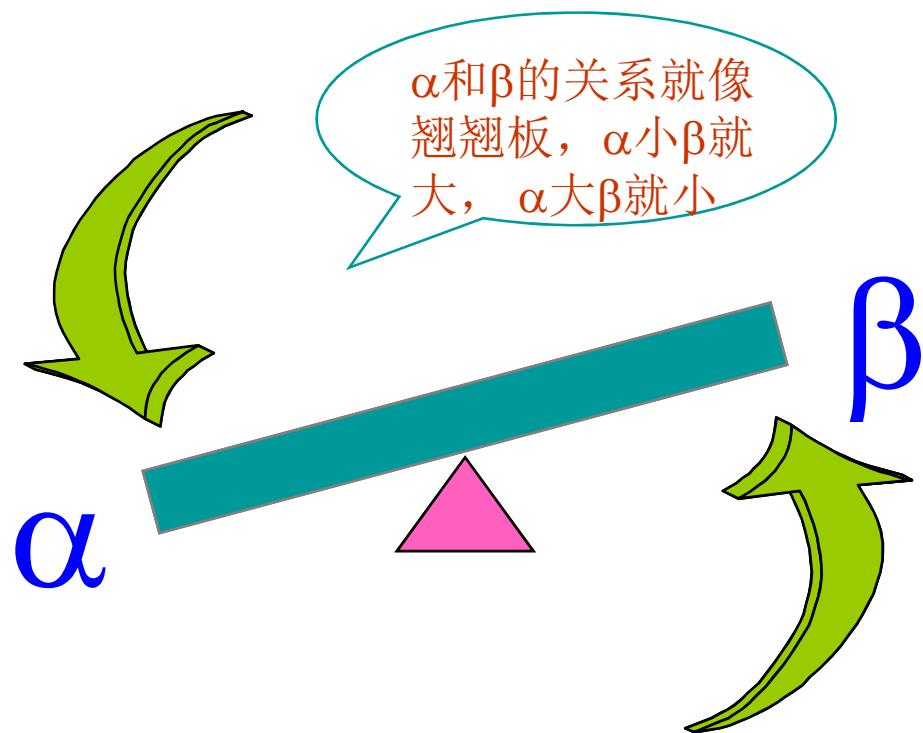
显著性水平  $\alpha$  取值由研究者**事先确定**，常见取值： 0.01, 0.05, 0.10



北京大学



# $\alpha$ 错误和 $\beta$ 错误的关系



样本容量  $n$  一定的情况下，假设检验不能同时做到犯  $\alpha$  和  $\beta$  两类错误的概率都很小。要使  $\alpha$  和  $\beta$  同时变小，只有增大样本容量，但会受人力、经费、时间等多因素的限制。



北京大学



# 假设检验的基本步骤

1. 提出原假设 $H_0$ ，确定备择假设 $H_1$
2. 构造分布已知的合适的检验统计量
3. 由给定的检验水平 $\alpha$ ，求出在 $H_0$ 成立的条件下的临界值（上侧 $\alpha$ 分位数，或双侧 $\alpha$ 分位数）
4. 根据样本计算检验统计量的样本观测值
5. 将检验统计量值与临界值比较，如果落在拒绝域内，则拒绝原假设，否则，接受原假设



北京大学



# 假设检验：p值法

假设检验方法



例1 设总体  $X \sim N(\mu, \sigma^2)$ ,  $\mu$  未知,  $\sigma^2 = 100$ , 现有样本  $x_1, x_2, \dots, x_{52}$ , 算得  $\bar{x} = 62.75$ .

现在来检验假设

$$H_0: \mu \leq \mu_0 = 60, \quad H_1: \mu > 60.$$

采用Z检验法, 检验统计量为



北京大学



定义 假设检验问题的  $p$  值 (*probability value*) 是由检验统计量的样本观察值得出的原假设可被拒绝的 **最小显著性水平**.

任一检验问题的  $p$  值可以根据检验统计量的样本观察值的以及检验统计量在  $H_0$  下一个特定的参数值(一般是  $H_0$  与  $H_1$  所规定的参数的分界点)对应的分布求出.



北京大学



# 三大抽样分布



北京大学



## 卡方分布

- 若 $Z$ 是标准正态分布随机变量，即 $Z \sim N(0, 1)$ ，则 $U = Z^2$ 的分布称为自由度为1的卡方分布，记作 $\chi_1^2$
- 概率密度函数： $f(u) = \frac{u^{-1/2}}{\sqrt{2\pi}} e^{-u/2}$ ,  $u > 0$
- $\chi_1^2 \sim \text{Ga}(1/2, 1/2)$
- 期望： $E(U) = 1$ ; 方差： $\text{Var}(U) = 2$

$$g(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x \geq 0$$
$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du, x > 0$$
$$E(X) = \alpha/\lambda; \quad D(X) = \alpha/\lambda^2$$

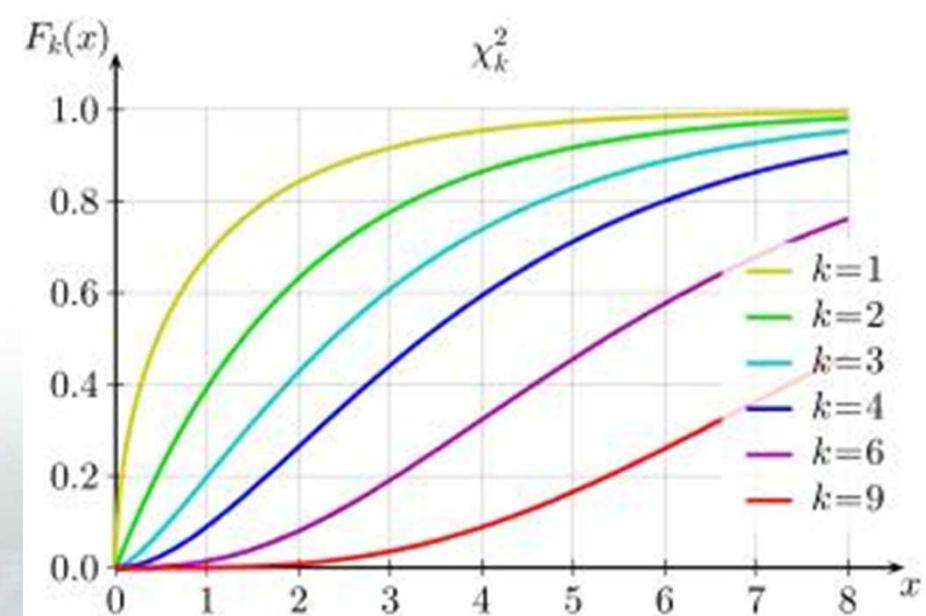
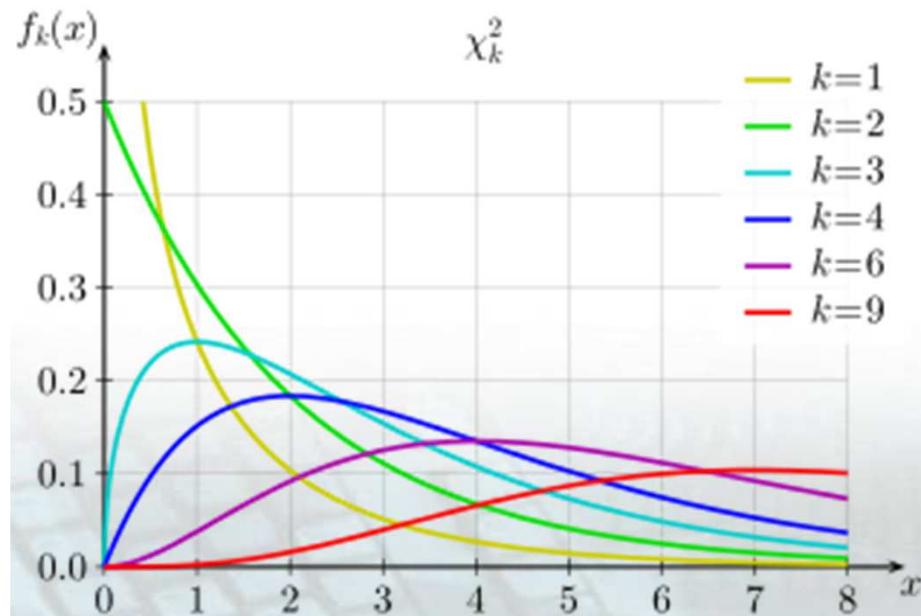


北京大学



# 卡方分布

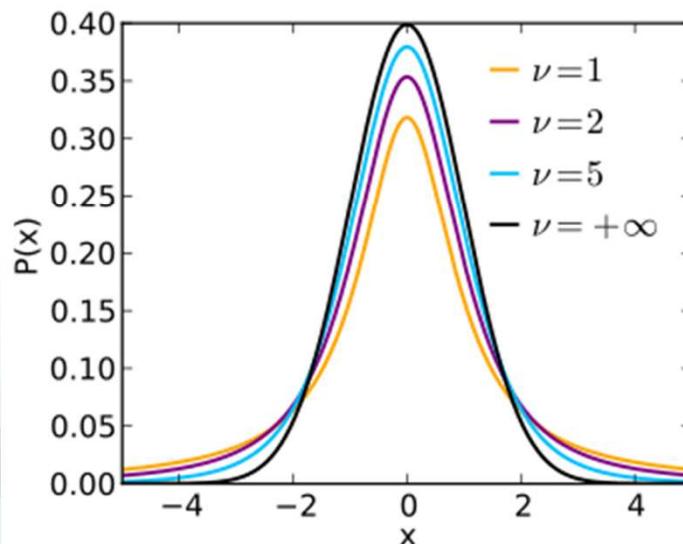
- 若  $X_1, X_2, \dots, X_n$  是相互独立的且服从标准正态分布的随机变量，则称  $\chi_n^2 = X_1^2 + X_2^2 + \dots + X_n^2$  所服从的分布为自由度为  $n$  的卡方分布
- 若  $U_1, U_2, \dots, U_n$  是相互独立的自由度为 1 的卡方随机变量，则称  $V = U_1 + U_2 + \dots + U_n$  所服从的分布为自由度为  $n$  的卡方分布，记作  $\chi_n^2$
- 概率密度函数:  $f(v) = \frac{v^{(n/2)-1}}{2^{n/2}\Gamma(n/2)} e^{-v/2}, v > 0$
- 性质:  $\chi_n^2 \sim \text{Ga}(n/2, 1/2)$ 
  - 期望:  $E(V) = n$ ; 方差:  $\text{Var}(V) = 2n$
  - 可加性:  $U \sim \chi_n^2, V \sim \chi_m^2$ , 且  $U$  和  $V$  独立, 则  $U + V \sim \chi_{m+n}^2$





# *t*分布

- 若  $Z \sim N(0, 1)$ ,  $U \sim \chi_n^2$ , 且  $Z$  和  $U$  独立, 则称随机变量  $T = Z / \sqrt{U/n}$  服从的分布是自由度为  $n$  的  $t$  分布
- 密度函数:  $f(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\Gamma(n/2)} (1 + \frac{t^2}{n})^{-(n+1)/2}$
- 性质:
  - 对称性:  $t$  分布关于零点对称,  $f(t)=f(-t)$ ,  $E(T)=0$
  - 当自由度趋向无穷时( $n>30$ ),  $t$  分布趋向于标准正态分布

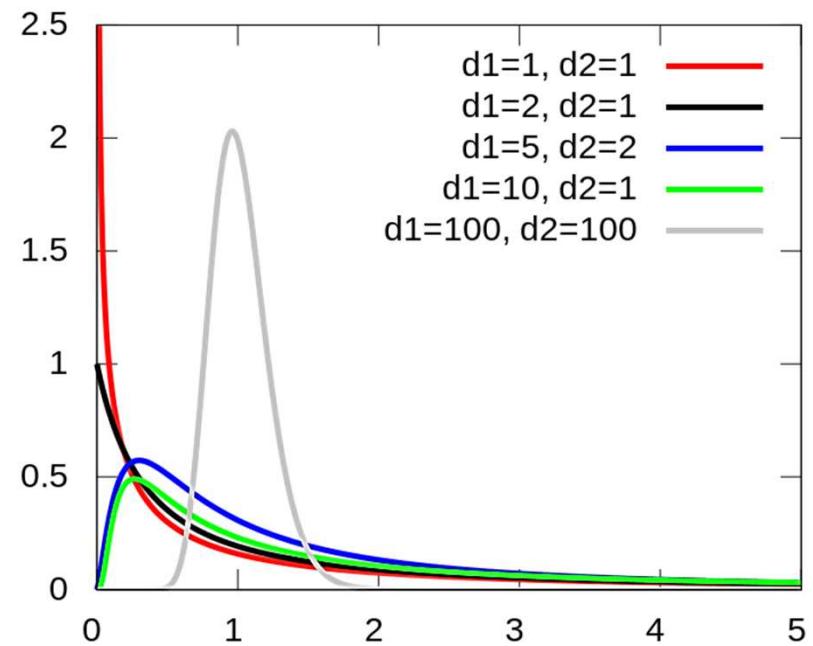


# $F$ 分布

- 若  $U \sim \chi_m^2$ ,  $V \sim \chi_n^2$ , 且  $U$  和  $V$  相互独立, 则称  $W = \frac{U/m}{V/n}$  服从的分布是自由度为  $m$  和  $n$  的  $F$  分布, 记作  $F_{m,n}$
- 密度函数:

$$f(w) = \frac{\Gamma\left[\frac{(m+n)}{2}\right]}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} w^{\frac{m}{2}-1} \left(1 + \frac{m}{n}w\right)^{-(m+n)/2}$$

- 性质:
  - 当  $n > 2$  时,  $E(W) = n/(n-2)$
  - 若  $W \sim F_{m,n}$ , 则  $W^{-1} \sim F_{n,m}$



设  $X_1, X_2, \dots, X_n$  是总体  $N(\mu, \sigma^2)$  的样本,  
 $\bar{X}, S^2$  分别是样本均值和样本方差, 则有

- 结论
- (1)  $\bar{X} \sim N(\mu, \sigma^2 / n)$ . 即  $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$
  - (2)  $\frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t(n-1)$
  - (3)  $\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \sim \chi^2(n)$
  - (4)  $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$
  - (5)  $\bar{X}$  与  $S^2$  相互独立



# 数据汇总



北京大学



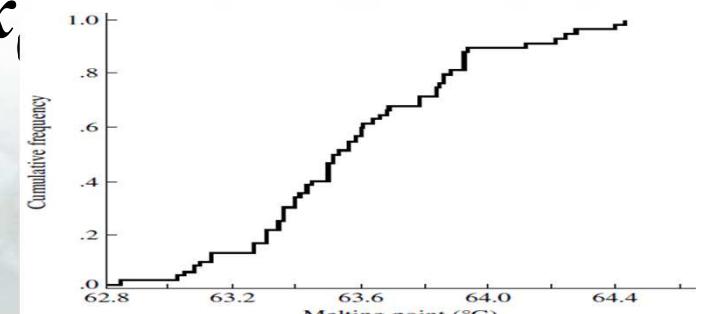
# 经验累积分布函数

$$(X_1, X_2, \dots, X_n) \longrightarrow (x_1, x_2, \dots, x_n)$$



$$x_{(1)}, x_{(2)}, \dots, x_{(n)}$$

定义函数  $F_n(x) = \begin{cases} 0, & \text{当 } x < x_{(1)} \\ \vdots & \vdots \\ k/n, & \text{当 } x_{(k)} \leq x < x_{(k+1)} \\ \vdots & \vdots \\ 1, & \text{当 } x \geq x_{(n)} \end{cases}$





# 生存时间函数

- 动机：
  - 医疗上，需要考虑各种药物或疗法的效果
  - 保险公司需要评估各种人群的寿命，以制定投保方案
  - 工程上，需要考虑材料（原件、设备等）的寿命
  - .....
- 生存分析：研究生存时间的分布规律以及生存时间和相关因素之间关系的一种统计分析方法
- 生存时间：从某起始事件到某终止事件经历的时间跨度
- 生存时间函数：描述生存时间分布规律的函数
  - 例如，生存函数、死亡函数、死亡密度函数、风险函数



北京大学



# 生存时间数据类型

- 完全数据
  - 提供的关于生存时间的信息是完整确切的
  - 准确地度量了观察对象实际生存的时间
- 截尾数据
  - 提供的关于生存时间的信息是不完整不确切的
  - 没有准确地度量观察对象实际生存的时间
    - 在随访过程中某些观察对象失访
    - 死于其它原因
    - 在规定的研究过程结束时观察对象的终止事件还未发生



北京大学



# 生存时间函数的估计

- 生存函数:  $S(t) = P(T > t)$

➤ 观察对象的生存时间T大于某时刻t的概率

➤ 性质:  $S(0)=1, S(\infty)=0$ , 且  $0 \leq S(t) \leq 1$

$$\hat{S}(t) = \frac{t\text{时刻尚生存的观察对象数}}{\text{观察对象总数}}$$

- 死亡函数:  $F(t) = P(T \leq t)$

➤ 观察对象的生存时间T不大于某时刻t的概率

➤ 性质:  $F(0)=0, F(\infty)=1$ , 且  $0 \leq F(t) \leq 1$

$$\hat{F}(t) = 1 - \hat{S}(t)$$



北京大学



# 生存时间函数的估计

- 死亡密度函数:  $f(t) = \lim_{\delta \rightarrow 0} \frac{F(t+\delta) - F(t)}{\delta}$

► 观察对象在某时刻  $t$  的瞬时死亡率

$$\hat{f}(t) = \frac{\text{观察对象在时间区间 } [t, t + \Delta t] \text{ 内死亡数}}{\text{观察对象总数} \times \text{区间 } [t, t + \Delta t] \text{ 所含单位时间数}}$$

- 风险函数:  $h(t) = \frac{f(t)}{S(t)}$

► 生存到时刻  $t$  的观察对象在时刻  $t$  的瞬时死亡率

$$\hat{h}(t) = \frac{\hat{f}(t)}{\hat{S}(t)} = \frac{\text{观察对象在时间区间 } [t, t + \Delta t] \text{ 内死亡数}}{t \text{ 时刻生存者数量} \times \text{区间 } [t, t + \Delta t] \text{ 所含单位时间数}}$$



北京大学



# 分位数-分位数图 (Q-Q图)

- 如果 $X$ 是具有严格单增分布函数 $F$ 的连续型随机变量，该分布的第 $p$ 分位数为满足下式的 $x$ 值：

$$F(x)=p$$

- $x_p=F^{-1}(p)$ : 连续分布函数中的一点，该点的一侧对应概率 $p$
- Q-Q图：比较两个分布函数
  - 两组容量为 $n$ 的数据，顺序统计量分别为 $X_{(1)}, \dots, X_{(n)}$ 和 $Y_{(1)}, \dots, Y_{(n)}$
  - 利用点对 $(X_{(i)}, Y_{(i)})$ 简单构造Q-Q图
- 性质：
  - 如果两个分布相似，则对应的Q-Q图趋近直线 $y=x$
  - 如果两个分布线性相关，则对应的Q-Q图趋近一条直线

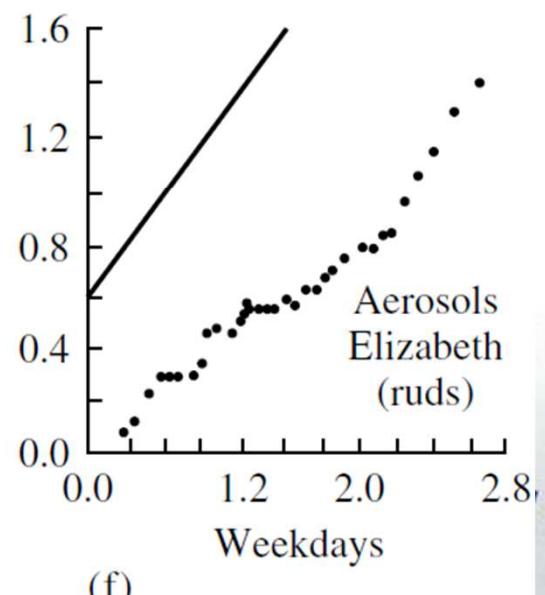
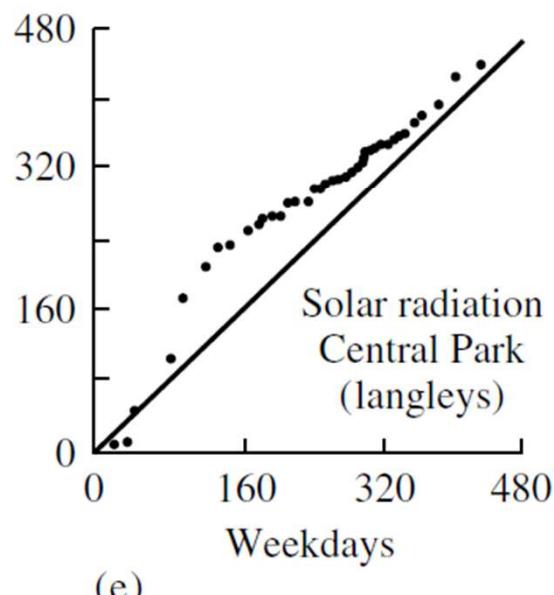
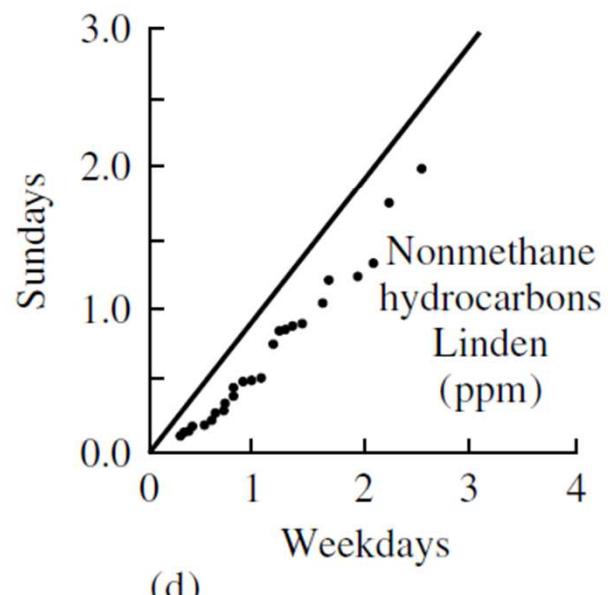
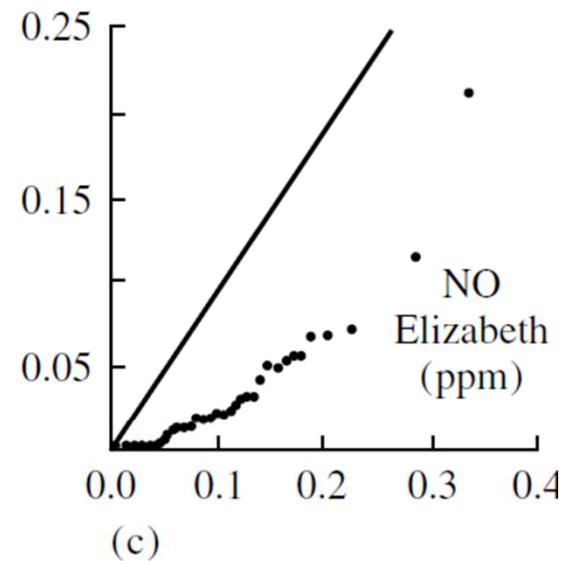
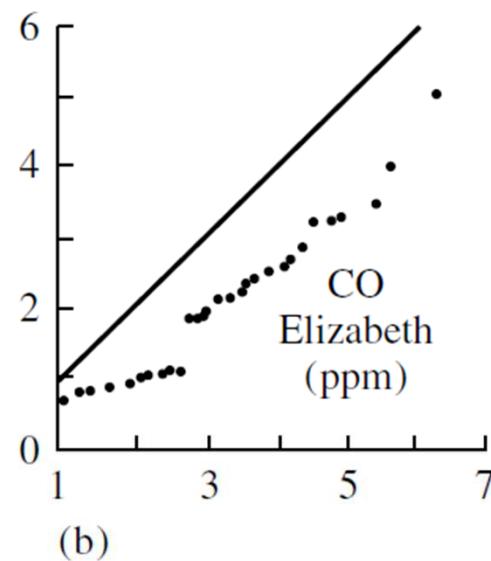
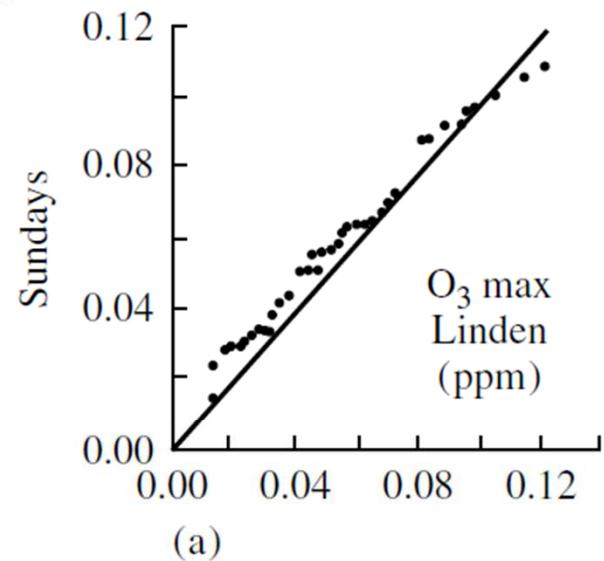
给定 $n$ 个观测，顺序统计量为 $X_{(1)}, \dots, X_{(n)}$ ，数据的 $(k-0.5)/n$ 分位数分配给 $X_{(k)}$



北京大学



# 案例：周日和平日空气污染对比





# 直方图

- 直方图：

- 将数据区域划分成几个区间或频带，画出落入每个频带的观测数或比例

- 常用于显示没有任何随机模型假设的数据图形

- 展示数据分布形状的方式类似于密度函数显示概率

- 基本步骤：

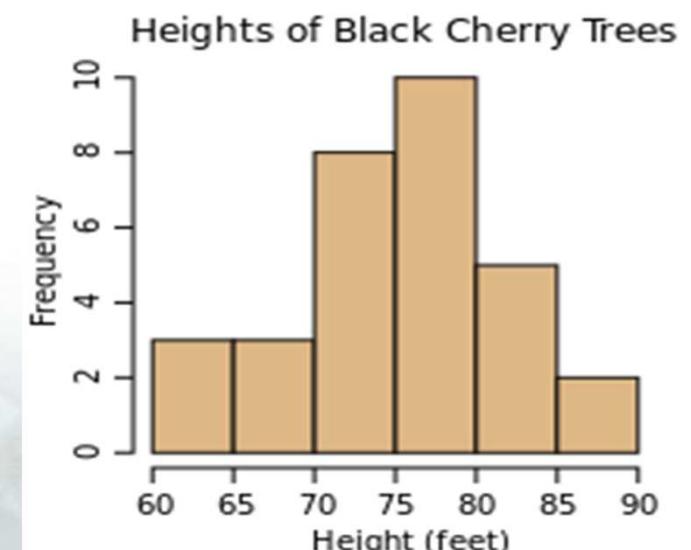
- 计算最大值与最小值的差(确定变动范围)

- 决定频带宽度与频带数(将数据分组)

- 决定分点

- 列出频率分布表

- 画出频率分布直方图



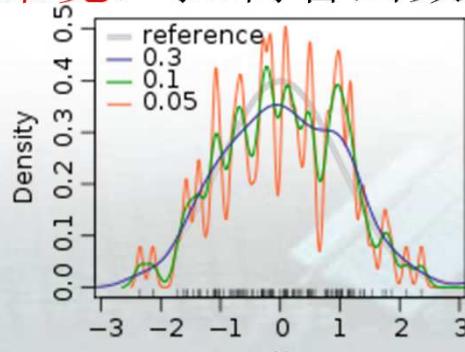


# 核概率密度估计

- 基本思想：
  - 如果某一个数在观察中出现了，可认为这个数的**概率密度较大**
  - 和这个数**较近**的数的**概率密度也会较大**，而那些**远离**这个数的数的**概率密度会较小**
- 估计方法：
  - 针对观察中的每个数，以 $K_h(x - x_i)$  拟合想象中的那个**远小近大**概率密度
  - 针对每个观察中出现的数拟合出多个概率密度分布函数，**取平均**

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right),$$

- $K(\cdot)$ 为核函数，通常取**标准正态分布密度函数**
- 参数 $h$ 为估计函数的**带宽**，控制着函数的**光滑性**



北京大学



# 位置度量

- 位置度量是一组**数据中心**的测量值
- 基本思想：
  - 如果数据是同一个量不同的测量结果，利用**位  
置度量**来代替单个观测值，更精确地表示测量  
尺寸
- 常用的位置度量：
  - 算术平均、中位数、截尾均值、M估计



北京大学



# 位置度量

- 算术平均:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- 中位数:
  - 如果样本容量是奇数, 中位数为顺序观测的中间值
  - 如果样本容量为偶数, 中位数为两个中间值的平均
- 截尾均值:
  - 丢掉最小的 $100\alpha\%$ 和最大的 $100\alpha\%$ 观测数据
  - 计算剩余数据的算术平均
- M估计:
  - 当标的分布为正态时, 样本均值是位置参数 $\mu$ 的最大似然估计



北京大学



## 位置度量：自助法

- 设位置参数为 $\hat{\theta}$ 
  - 随机变量 $X_1, X_2, \dots, X_n$ 的函数，抽样分布由 $n$ 和 $F$ 确定
- 若 $F$ 已知，可利用模拟来计算 $\hat{\theta}$ 的概率分布
  - 由 $F$ 生成很多容量为 $n$ 的样本
  - 利用每一个样本计算 $\hat{\theta}$
  - 结果值 $\theta_1^*, \theta_2^*, \dots$ 的经验分布是 $\hat{\theta}$ 分布函数的近似
- 若 $F$ 未知，将经验cdf  $F_n$ 视作 $F$ 的近似，从 $F_n$ 中抽样





# 散度度量

- 散度或规模度量给出了一组数据“分散状态”的数值表示
- 常用散度度量

➤ 样本标准差：样本方差的平方根

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

➤ 四分位差（IQR）：

两个样本四分位数（第25和75分位数）的差

➤ 中位数绝对偏差（MAD）：

若数据  $x_1, \dots, x_n$  具有中位数  $\tilde{x}$ , MAD为数值  $|x_i - \tilde{x}|$  的中位数

➤ IQR和MAD分别除以1.35和0.675得到正态分布的 $\sigma$ 的估计

铂数据的3个散度度量：

$$s = 4.45$$

$$\text{IQR}/1.35 = 1.26$$

$$\text{MAD}/0.675 = 0.934$$

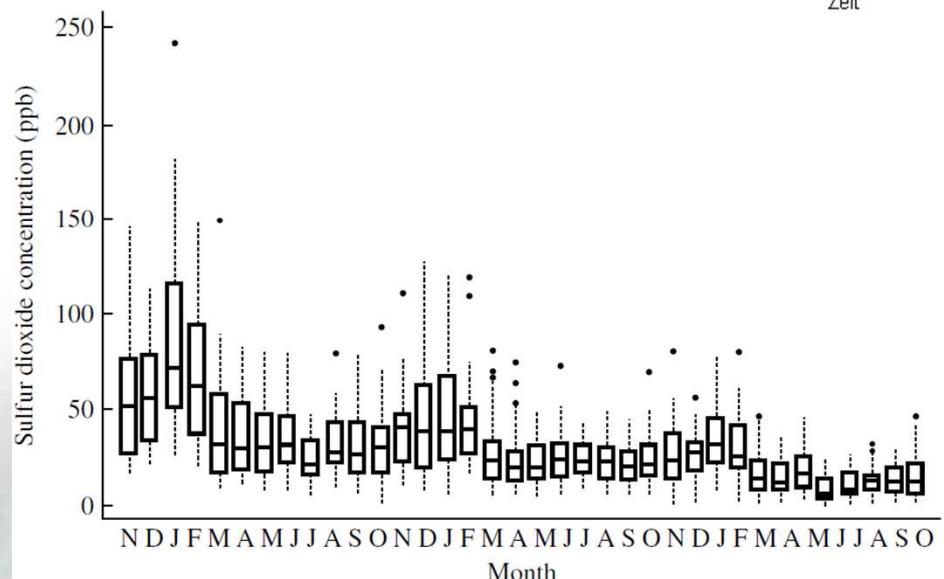
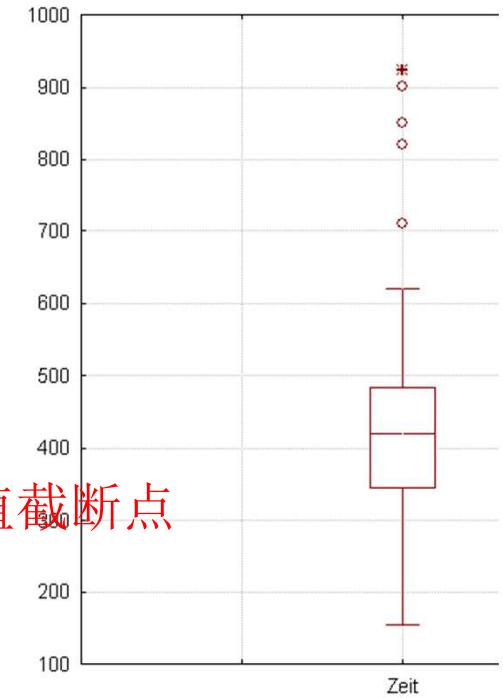


北京大学



# 箱形图

- 一种用作显示数据分散情况的统计图
  - 位置度量（中位数）、散度度量（四分位差）
  - 可能出现的离群点
- 构造过程
  - 画数轴，起点比最小值稍小，长度比该数据全距稍长
  - 画矩形盒，两端边的位置对应数据的上下四分位数
  - 在矩形盒内部中位数位置画一条线段为中位线
  - 在 $Q3 + 1.5 \text{IQR}$ 和 $Q1 - 1.5 \text{IQR}$ 处画两条线段，为异常值截断点
  - 从矩形盒两端边向外各画一条线段直到异常值截断点
  - 用星号或点（\*或·）标出异常值
- 箱形图的作用
  - 识别数据异常值
  - 判断数据偏态和尾重
  - 比较几批数据的形状





# 分布拟合检验

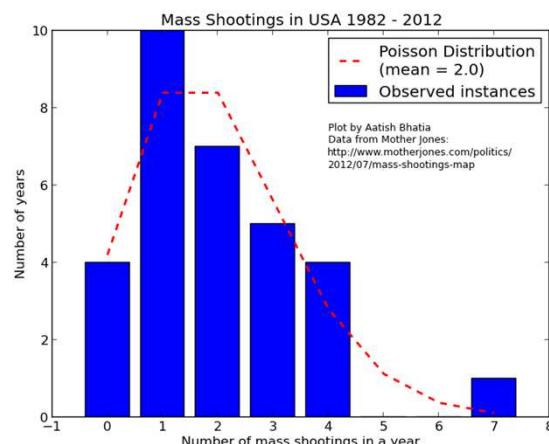


北京大学

# $\chi^2$ 检验



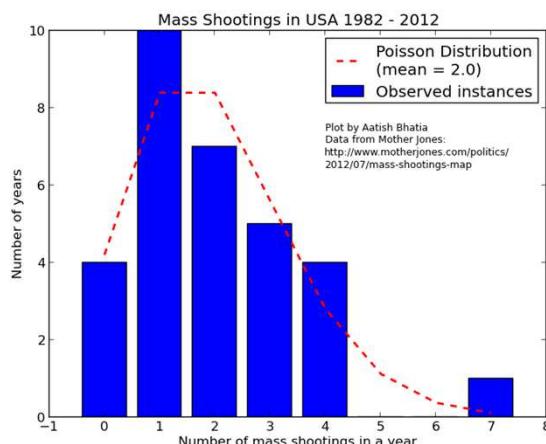
- 基本思想：
  - 总体分布未知，根据样本检验关于总体分布的假设
  - 先提出原假设： $H_0$ ：总体X的分布函数为 $F(x)$
  - 然后根据样本的经验分布和所假设的理论分布之间的吻合程度来决定是否接受原假设
- 这种检验通常被称作拟合优度检验，是一种非参数检验
- 由统计学家K.皮尔逊在1900年提出，被视为近代统计学的开端





# $\chi^2$ 检验：基本原理和步骤

1. 提出原假设:  $H_0$ : 总体 $X$ 的分布函数为 $F(x)$
2. 将总体 $X$ 的取值范围划分为 $k$ 个互不重迭的区间,  $(a_0, a_1], \dots, (a_{k-1}, a_k)$ , 记作 $A_1, A_2, \dots, A_k$
3. 把落入第 $i$ 个区间 $A_i$ 的样本值的个数记作 $f_i$ , 称为实测频数
4. 根据所假设的理论分布算出总体 $X$ 的值落入每个 $A_i$ 的概率 $p_i$ , 则 $np_i$ 为落入 $A_i$ 的样本值的理论频数



北京大学



实测频数

理论频数

$$f_i - np_i$$

皮尔逊引进如下**统计量**表示经验分布与理论分布间的**差异**:

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i}$$

在理论分布  
已知的条件下，  
 $np_i$ 是常量

或       $\chi^2 = \sum_{i=1}^k \frac{n}{p_i} \left( \frac{f_i}{n} - p_i \right)^2 = \sum_{i=1}^k \frac{f_i^2}{np_i} - n$



北京大学



5. 对给定的显著性水平 $\alpha$ , 通过查 $\chi^2$ 分布表确定 $I$ 值, 使 $P\{\chi^2 > I\} = \alpha$ , 得到拒绝域:  $\chi^2 > I$
6. 根据所给的样本值 $x_1, x_2, \dots, x_n$ 计算统计量 $\chi^2$
7. 若 $\chi^2$ 的实测值落入拒绝域, 则拒绝原假设 $H_0$ , 否则就认为差异不显著而接受原假设 $H_0$



北京大学



## 示例：骰子检查

- 将一颗骰子掷120次，所得数据如下表：

点数 $i$	1	2	3	4	5	6
出现次数 $f_i$	23	26	21	20	15	15

问这颗骰子是否均匀、对称（取 $\alpha=0.05$ ）？

- 解：若这颗骰子是均匀、对称的，则1~6点中每点出现的可能性相同，都为 $1/6$ . 如果用 $A_i$ 表示第*i*点出现，则待检假设：

$$H_0: P(A_i) = 1/6, \quad i=1,2,\dots,6$$

- 在 $H_0$ 成立的条件下，理论概率 $p_i = p(A_i) = 1/6$
- 由 $n=120$ 得频率  $np_i=20$



北京大学



$i$	$f_i$	$p_i$	$np_i$	$(f_i - np_i)^2 / (np_i)$
1	23	1/6	20	9/20
2	26	1/6	20	36/20
3	21	1/6	20	1/20
4	20	1/6	20	0
5	15	1/6	20	25/20
6	15	1/6	20	25/20
合计	120			4.8

此分布不含未知参数， $k=6$ ,  $\alpha=0.05$ , 查表得:  $\chi^2_{\alpha}(k-1) = \chi^2_{0.05}(5) = 11.071$

上表:  $\chi^2 = \sum_{i=1}^6 \frac{(f_i - np_i)^2}{np_i} = 4.8 < 11.071$ , 故接受  $H_0$ , 即骰子是均匀对称的

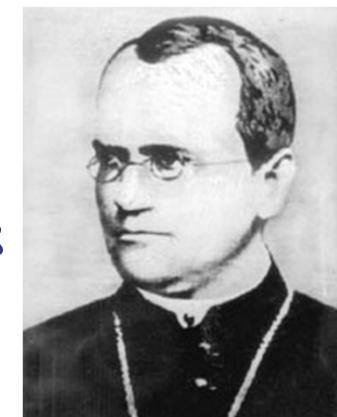
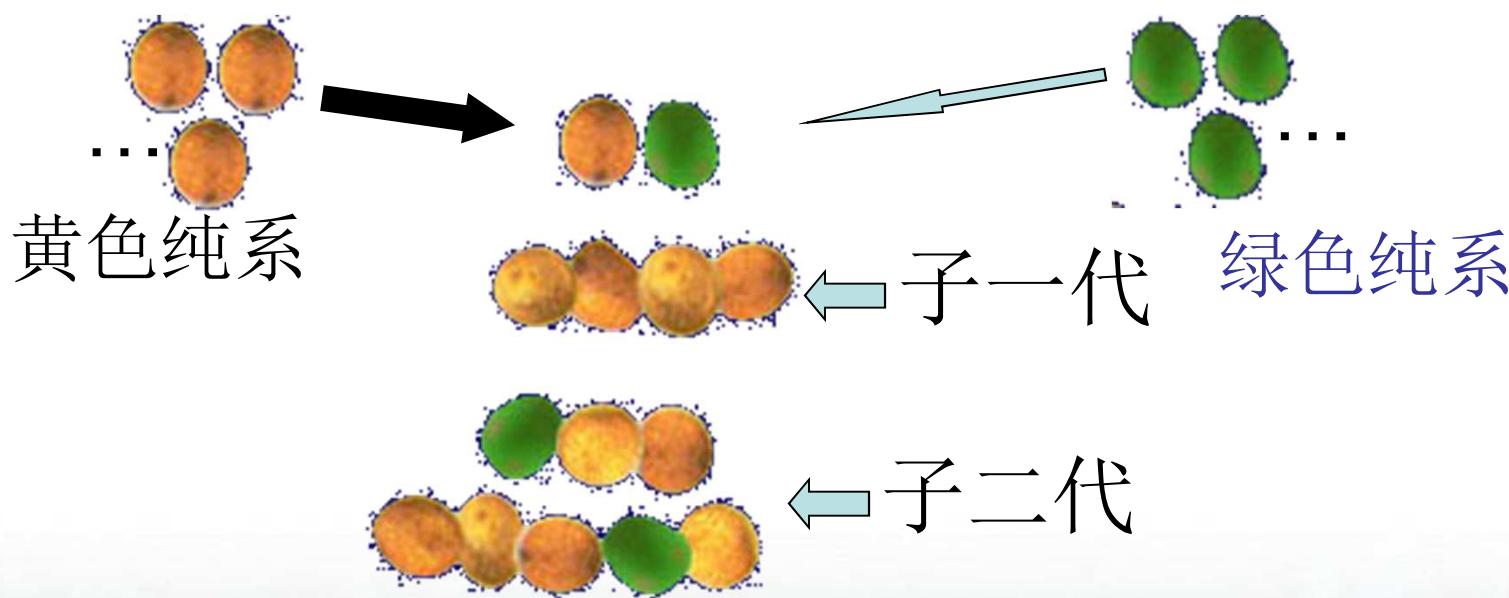


北京大学



## 示例：遗传规律

- 生物学家孟德尔进行了八年的豌豆杂交试验，并根据试验结果，运用他的数理知识，发现了遗传的基本规律
- 理论：黄、绿豌豆杂交，子二代中，黄、绿之比为3:1
- 一组观察结果为： 黄：70、绿：27，是否符合理论？



孟德尔



北京大学

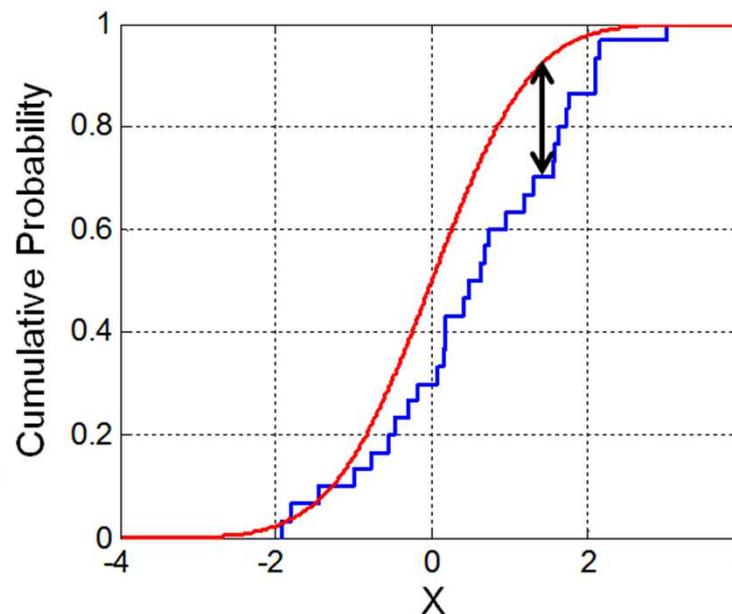


# Kolmogorov–Smirnov检验

1. 提出原假设 $H_0$ : 总体 $X$ 的分布函数为 $F(x)$
2. 计算样本累计频率与理论分布累计概率的绝对差，令最大的绝对差为 $D_n$ :

$$D_n = \max_{1 \leq i \leq n} \{ |F(x_i) - F_n(x_i)| \}$$

3. 用样本容量 $n$ 和显著水平 $\alpha$ 查表得临界值 $D_n^{\alpha}$
4. 通过 $D_n$ 与 $D_n^{\alpha}$ 的比较做出判断，若 $D_n < D_n^{\alpha}$ ，则认为拟合是满意的



北京大学

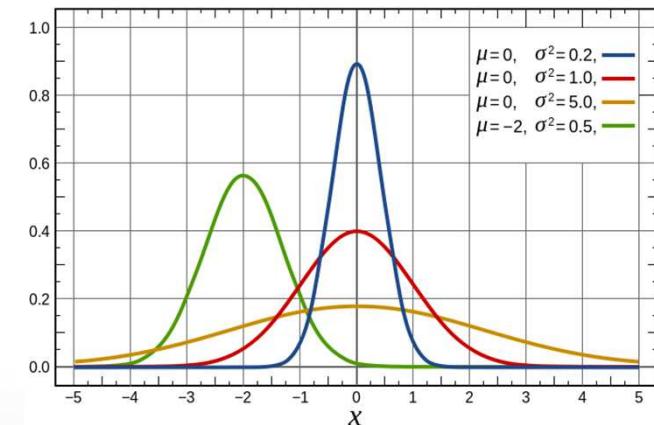


# 正态性检验：J-B检验

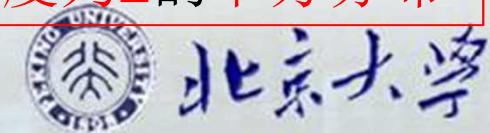
- 正态分布的性质：
  - 偏度（三阶中心矩）： $S=0$ ； 峰度（四阶中心矩）： $K=3$
- 基本思想：
  - 若样本来自正态总体，则其偏度和峰度应该在0, 3附近
- J-B统计量： $JB = \frac{n}{6} [S^2 + \frac{(K-3)^2}{4}]$ ,  $n$ 为样本容量

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2}$$



在正态分布的假设下，JB统计量渐进地服从自由度为2的卡方分布





# 两样本比较



北京大学



## z检验和t检验

*z* 检验的应用条件:

- (1) 样本来自正态总体
- (2a) 样本含量  $n$  较大, 或
- (2b)  $n$  虽小但总体标准差  $\sigma$  已知

*t* 检验的应用条件:

- (1) 总体标准差  $\sigma$  未知;
- (2) 样本含量  $n$  较小;
- (3) 样本来自正态总体;
- (4) 两样本均数比较时方差齐,  
即  $\sigma_1^2 = \sigma_2^2$



北京大学



## 配对 $t$ 检验(paired t-test)

- 配对设计:

- 两组观察对象除了研究因素不同外，其它的可能影响研究结果的因素相同或相似

- 四种情况:

- 两个同质受试对象分别接受两种不同的处理
  - 同一受试对象分别接受两种不同的处理
  - 同一受试对象接受某种处理的前后数据
  - 同一受试对象的两个不同部位的数据



北京大学



## 两独立样本均数的比较 (two-sample test)

两样本均为随机抽样得到的样本  
或 采用随机分组得到的样本



北京大学



# *t* 检验

目的：推断两样本均数分别代表的总体

均数  $\mu_1$  与  $\mu_2$  有无差别

适用条件：

- 随机抽样的小样本 ( $\sigma$  未知)
- 两样本来自正态总体
- 两样本的总体方差齐同 ( $\sigma_1^2 = \sigma_2^2$ )



北京大学



注：方差齐性的经验判断方法

若  $s_1^2 / s_2^2 \geq 3$  } 可怀疑两样本总体方差不等  
或  $s_1 / s_2 \geq 2$  }

$s_1^2 / s_2^2 \geq 5$  可认为两样本总体方差不等

否则可认为两总体方差相等



北京大学



# 方差齐性检验

方差齐性检验的计算公式为：

$$F = \frac{s_1^2 \text{ (较大)}}{s_2^2 \text{ (较小)}}$$

$$v_1 = n_1 - 1$$

$$v_2 = n_2 - 1$$

若两样本是来自同一个正态总体，则它们的方差不应相差过大，其  $F \geq 1$ 。

由于抽样误差的存在，其  $F$  可能偏离于 1，当其偏离过大，超出抽样误差所能引起的范围，则表明方差不齐



北京大学



## 两总体均数比较



## 方差齐性检验



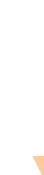
方差齐



方差不齐



$t$  检验、 $z$ 检验



$t'$ 检验

前提：来自正态总体



北京大学



# $t'$ 检验 — 近似t检验

- 基本思想：
    - 在方差不齐的情况下进行比较
    - 样本均数的分布曲线由  $t$  分布转化为  $t'$  分布
    - $t'$  分布较复杂，故用  $t$  分布的临界值估计  $t'$  分布的临界值，即对临界值校正后依  $t$  检验进行分析
  - Cochran & cox 法： 对临界值校正
  - Satterthwaite 法
  - Welch 法
- } 对自由度校正



北京大学



# 两独立样本差别的秩和检验

## Wilcoxon rank sum test

两独立样本秩和检验计算表

A样本		B样本	
观察值	秩次	观察值	秩次
7	4	3	1
14	6	5	2
22	10	6	3
36	11	10	5
40	13	17	7
48	14	18	8
63	15	20	9
98	16	39	12
$n_1=8$	秩和 $R_1=89$	$n_2=8$	秩和 $R_2=47$

假定：两组样本的总体分布形状相同

基本思想

如果两  
总体分  
布相同  
→ 两样本来自同一总体  
↓  
任一组秩和不应太大或太小  
↓  
 $T$  与平均秩和  $n_0(1 + N) / 2$  应相差不大

$$T = \begin{cases} \text{较小例数组的秩和, } n_1 \neq n_2 \\ \min(R_1, R_2), n_1 = n_2 \end{cases}$$

$$N = n_1 + n_2$$

$$n_0 = \min(n_1, n_2)$$



北京大学



- (1) 提出假设  $H_0$ : 两样本来自相同总体;  
 $H_1$ : 两样本来自不同总体（双侧） 或  $H_1$ : 样本A高于样本B（单侧）
- (2) 编秩: 两样本混合编秩次，求得  $R_1$ 、 $R_2$ 、 $T$ 。  
相同观察值（即相同秩，ties），不同组---平均秩次。
- (3) 确定P值作结论:
- ①小样本: **查表法**（威尔科克森和曼恩-惠特尼检验临界值表）  
如果  $T$  位于检验界值区间内， $P > \alpha$ ，不拒绝  $H_0$ ；否则，拒绝  $H_0$   
本例  $T = 47$ ，取  $\alpha = 0.05$ ，查表得双侧检验界值区间  $(49, 87)$ ， $T$  位于区间外， $P < 0.05$ ，因此在  $\alpha = 0.05$  的水平上，拒绝  $H_0$ ，接受  $H_1$ 。

②大样本: **正态近似法**

$$u = \frac{|T - n_0(N + 1)/2|}{\sqrt{n_1 n_2 (N + 1)/12}}$$

本例  $u = 2.205 > \mu_{0.05/2} = 1.96$



北京大学



# 配对设计资料的秩检验 (Wilcoxon signed rank test)

家兔号	A照射	B照射	B-A	秩次
1	39	55	16	10
2	42	54	12	9
3	51	55	4	3
4	43	47	4	3
5	55	53	-2	-1
6	45	63	18	11
7	22	52	30	12
8	48	44	-4	-3
9	40	48	8	6
10	45	55	10	8
11	40	32	-8	-6
12	49	57	8	6
合计				$T=10$ (68)

- $H_0$ : 差值的总体中位数=0,  
 $H_1$ : 差值的总体中位数 $\neq 0$ ;  
 $\alpha=0.05$
- 求差值; 绝对值从小到大编秩次
  - 绝对值相等者取平均秩次;
  - 将差值的正负标在秩次之前;
  - 零差值时不参与编秩
- 分别求正负秩次之和, 以绝对值较小者为 $R$ 值
- 根据统计量 $R$ 确定对应的 $P$ 值
  - 小样本时, 查表
  - 大样本时, 正态近似



北京大学



# t检验 vs. 秩和检验

## 适用条件

t检验:

- a. 样本所在总体呈正态分布
- b. 各总体方差要齐

秩和检验:

- a. 不满足正态和方差齐性条件的小样本资料
- b. 总体分布类型不明的小样本资料
- c. 单向有序列联表资料
- d. 各种资料的初步分析



北京大学



# 方差分析



北京大学



# 方差分析简介

- 方差分析，简称ANOVA (Analysis of Variance)
- R.A. Fisher于1923年提出的一种统计方法
- 为纪念Fisher，以F命名，故方差分析又称F检验
- 检验多个总体的均值是否存在显著差异
- 应用条件：
  - 各样本都来自正态总体
  - 各个总体方差相等（方差齐同）
  - 各样本是相互独立的随机样本



Ronald Aylmer Fisher  
(1890-1962)



# 方差分析中的基本概念

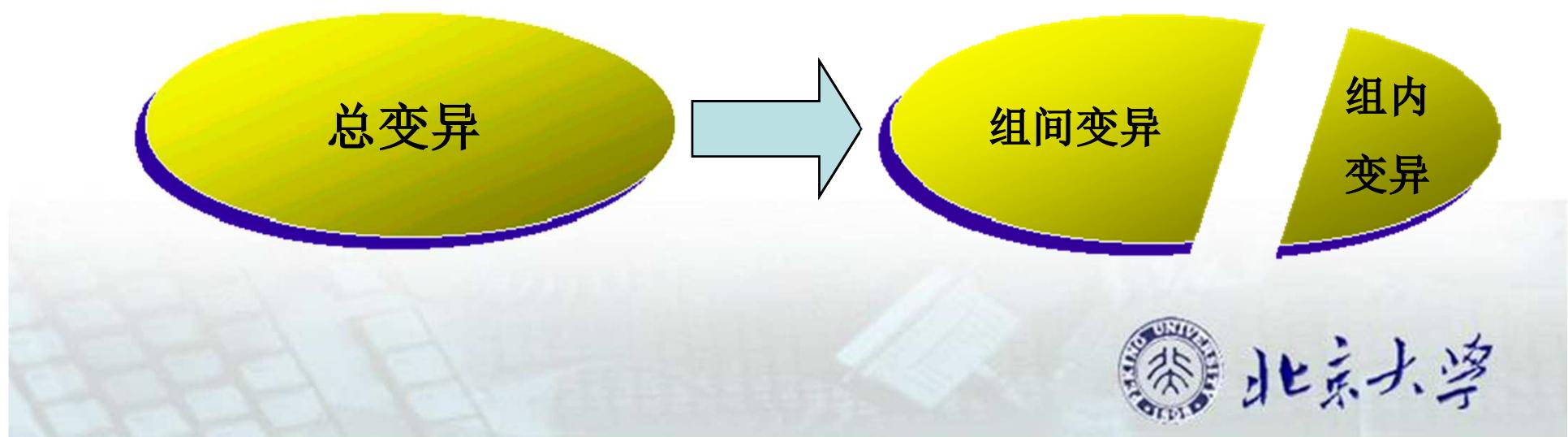
- 因素或因子(Factor): 所要检验的对象
  - 例如: 要分析城市对推广策略是否有影响, 城市是要检验的因素或因子
- 水平或处理(Treatment): 因子的不同表现
  - 例如: 北京、广州、上海、深圳就是因子的水平
- 观察值: 在每个因素水平下得到的样本数据
  - 例如: 每个城市的周销售量就是观察值





# 方差分析的基本原理和方法

- 样本数据波动的来源：
  - 因素中的不同水平
  - 抽选样本的随机性
- 波动的度量：
  - 组间方差：水平之间的方差
  - 组内方差：水平内部的方差





# 方差分析的基本原理和方法

- 组间方差反映不同因子对样本波动的影响
- 组内方差则是不考虑组间方差的纯随机影响
- 若不同水平对结果无影响，则组间方差中仅有随机因素的差异，而无系统性的差异，它与组内方差就应该近似，两个方差比值应接近于1
- 反之，两个方差的比值就会显著地大于1，当这个比值大到某个程度，或者说达到某临界点，就可以判断出不同水平之间存在着显著性差异
- 小概率原理是方差分析的指导思想

组内方差仅包括随机性因素

组间方差既包括系统性因素，也包括随机性因素



北京大学



# 方差分析的基本原理和方法

- 产生方差的**独立变量的个数**对方差大小有影响
  - 独立变量个数越多，方差就有可能越大
  - 独立变量个数越少，方差就有可能越小
- 使用均方差来消除独立变量个数对方差的影响
  - **均方差（Mean Square）** 等于方差除以独立变量个数
- 引起方差的独立变量的个数，称为**自由度**





# 方差分析的基本原理和方法

检验因子影响是否显著通常用如下F统计量：

$$F = \frac{\text{组间均方差}}{\text{组内均方差}}$$

- F统计量越大，越说明组间方差是主要方差来源，因子影响越显著
- F越小，越说明组内方差（随机方差）是主要的方差来源，因子的影响越不显著



北京大学



# 单因素方差分析的数据结构

观测值 ( $j$ )	因素 ( $A$ ) $i$			
	水平 $A_1$	水平 $A_2$	...	水平 $A_K$
1	$X_{11}$	$X_{21}$	...	$X_{k1}$
2	$X_{12}$	$X_{22}$	...	$X_{k2}$
:	:	:	:	:
:	:	:	:	:
$n$	$X_{1n}$	$X_{2n}$	...	$X_{kn}$



北京大学



# 单因素条件下离差平方和的分解

- 总离差平方和  $SS_T = SS_E + SS_A$
- 总离差平方和  $SS_T$  反映了离差平方和的总体情况

$$SS_T = \sum \sum (X_{ij} - \bar{X})^2$$

- 误差项离差平方和  $SS_E$  反映的是水平内部，或组内观察值的离散状况

$$SS_E = \sum \sum (X_{ij} - \bar{X}_{i.})^2$$

- 水平项离差平方和  $SS_A$  反映的是组间差异

$$SS_A = \sum \sum (\bar{X}_{i.} - \bar{X})^2 = \sum n \cdot (\bar{X}_{i.} - \bar{X})^2$$



北京大学



# 因素作用显著性的检验

- $SS_T$ 的自由度为 $nk-1$  
$$SS_T = \sum \sum (X_{ij} - \bar{X})^2$$
  - 方差是由于变量波动引起的，但所有的 $nk$ 个变量并不独立，它们满足一个约束条件，真正独立的变量只有 $nk-1$ 个
- $SS_A$ 的自由度为 $k-1$ 
  - $SS_A$ 是因子在不同水平上的均值变化而产生的方差，但 $k$ 个均值并不独立，它们满足一个约束条件，因此也丢失一个自由度
- $SS_E$ 的自由度为 $nk-k$ 
  - $SS_E$ 是由所有在各因素水平上围绕均值的波动产生，它们满足的约束条件一共 $k$ 个，失去了 $k$ 个自由度
- $SS_T$ 、 $SS_A$ 和 $SS_E$ 的自由度满足如下关系  
$$nk-1 = (k-1) + (nk-k)$$



北京大学

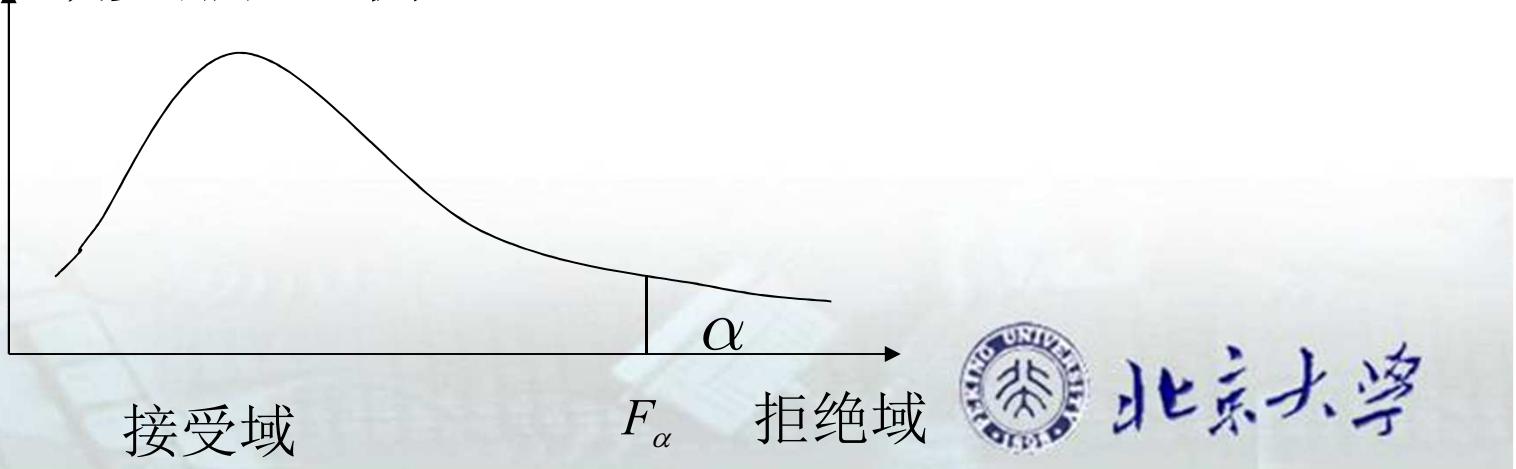


# 因素作用显著性的检验

检验统计量:  $F = \frac{MS_A}{MS_E} \sim F(k-1, n \cdot k - k)$

其中,  $MS_A = \frac{SS_A}{k-1}$ ,  $MS_E = \frac{SS_E}{n \cdot k - k}$

- F值越大, 越说明总的方差波动中, 组间方差是主要部分, 有利于拒绝原假设接受备选假设
- F值越小, 越说明随机方差是主要的方差来源, 有利于接受原假设, 有充分证据说明待检验的因素对总体波动有显著影响
- 检验的拒绝域安排在右侧





## 关系强度的测量

- 拒绝原假设表明因素(自变量)与观测值之间有关系
- 组间平方和( $SS_A$ )度量了自变量(城市)对因变量(销售量)的影响效应：
  - 只要组间平方和 $SS_A$ 不等于0，就表明两个变量之间有关系(只是是否显著的问题)
  - 当组间平方和 $SS_A$ 比组内平方和 $SS_E$ 大，且大到一定程度时，就意味着两个变量之间的关系显著，大得越多，表明它们之间的关系就越强。反之，小得越多，表明它们之间的关系就越弱



北京大学



## 关系强度的测量

- 变量间关系的强度可用自变量平方和 $SS_A$  占总平方和 $SS_T$  的比例大小来度量
- 自变量平方和占总平方和的比例记为 $R^2$ ,即

$$R^2 = \frac{SS_A}{SS_T}$$

- 其平方根 $R$  可用来测量两个变量之间的关系强度





# 双因素方差分析

- 动机:
  - 单因素方差分析只考虑一个自变量对数值型因变量的影响
  - 实际中，有时需要考虑**两个或多个因素**对实验结果的影响
- 当方差分析中涉及两个自变量时，称为双因素方差分析  
**(Two-way Analysis of Variance)**
- 获取数据时，将一个因素安排在“行”的位置，称为行因素，另一因素安排在“列”的位置，称为列因素
- 前提:
  - 每个总体都服从正态分布、各总体的方差齐同、各观察值独立
- 类型:
  - **无交互作用**的双因素方差分析
  - **有交互作用**的双因素方差分析





# 秩方差分析



北京大学



# Kruskal-Wallis秩和检验

- 简称K-W检验，也称H检验
- 由Kruskal和Wallis二人在1952年提出
- 一种单因素方差分析方法
- 将两个独立样本的秩和检验推广到3个或多组的检验
- 基本原理：
  - 若各组的处理效应相同，则混合编秩后，各组的平均秩应近似相等



北京大学



# 单因素方差分析的数据结构

		因素( $A$ )			
		水平 $A_1$	水平 $A_2$	...	水平 $A_K$
重复 观察 量		$x_{11}$	$x_{21}$	...	$x_{k1}$
		$x_{12}$	$x_{22}$	...	$x_{k2}$
	:	:	:	:	:
	:	:	:	:	:
		$x_{1n_1}$	$x_{1n_2}$	...	$x_{1n_k}$





# 单因素方差分析的秩矩阵

		因素( $A$ )			
重 复 观 察 量 的 秩	水平 $A_1$	水平 $A_2$	...	水平 $A_K$	
	$R_{11}$	$R_{21}$	...	$R_{k1}$	
	$R_{12}$	$R_{22}$	...	$R_{k2}$	
	:	:	:		:
	:	:	:		:
	$R_{1n_1}$	$R_{2n_2}$	...	$R_{kn_k}$	
	秩和	$R_{1\cdot}$	$R_{2\cdot}$	...	$R_{k\cdot}$

各组的平均秩:  $\bar{R}_{i\cdot} = R_{i\cdot}/n_i$

混合后的平均秩:  $\bar{R}_{..} = (N + 1)/2$



北京大学



# 检验统计量：H值

$$H = \sum_{i=1}^k \frac{(R_{i\cdot} - n_i \bar{R}_{..})^2}{n_i S^2}$$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (R_{ij} - \bar{R}_{..})^2 = \frac{1}{N-1} \sum_{i=1}^k \sum_{j=1}^{n_i} R_{ij}^2 - N \bar{R}_{..}^2$$

如果没有同秩(tie)现象，则有

$$S^2 = \frac{1}{N-1} \left[ \frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4} \right] = \frac{N(N+1)}{12}$$

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$$

当混合编秩过程中出现同秩现象时，需要对H统计量进行校正



北京大学



# K-W检验：一般步骤

## 1. 建立假设、确定检验水准

$H_0$ ：各样本分布位置相同

$H_1$ ：各样本分布位置不全相同

$\alpha=0.05$

## 2. 选择检验方法、计算统计量

(1) 混合编秩

(2) 求秩和( $R_i$ )和统计量 $H$ 值：

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$$



北京大学



### 3.确定P值，作出推断结论

#### (1)小样本情况：

当组数 $k \leq 3$ ，且 $n_i < 5$ 时，查H界值表，确定P值。

如果 $H > H_\alpha$ ，则 $P < \alpha$ ；反之， $P > \alpha$ 。

#### (2)大样本情况：

若 $k > 3$ 或 $n_i > 5$ 时，理论上，H近似服从自由度为 $k-1$ 的 $\chi^2$ 分布，可查 $\chi^2$ 界值表确定P值。

如果 $H > H_\alpha$ ，则 $P < \alpha$ ；反之， $P > \alpha$ 。



北京大学



# 相关分析



北京大学



# 两种现象间的依存关系

函数关系

现象间具有严格的确定性的依存关系

相关关系

客观现象间存在关系，但数量上不是严格对应的依存关系

函数关系与相关关系之间并无严格的界限：

有函数关系的变量间，由于有测量误差及各种随机因素的干扰，可表现为相关关系；

具有相关关系的变量有深刻了解之后，相关关系有可能转化为或借助函数关系来描述。



北京大学



# 应变量 vs. 自变量

- 现象之间的相互联系，常表现为一定的因果关系
- 自变量：
  - 起着影响作用的变量称为自变量，通常用X表示
  - 是引起另一现象变化的原因，是可控制、给定的值
- 应变量：
  - 受自变量影响的变量称为应变量，通常用Y表示
  - 它是自变量变化的结果，是不确定的值。





## 相关分析的内容

- 分析现象之间相互关系的方向和程度
- 主要内容：
  - 确定现象之间是否存在相关关系以及相关关系的表现形式
  - 确定相关关系的密切程度
  - 确定相关关系的数学表达式，即回归方程式
  - 检验估计值的误差





# 相关关系的测定

- 定性分析:

- 依据研究者的理论知识和实践经验，对客观现象之间是否存在相关关系以及何种关系作出判断

- 定量分析:

- 在定性分析的基础上，通过绘制相关图、计算相关系数与判定系数等方法，来判断现象之间相关的方向、形态及密切程度



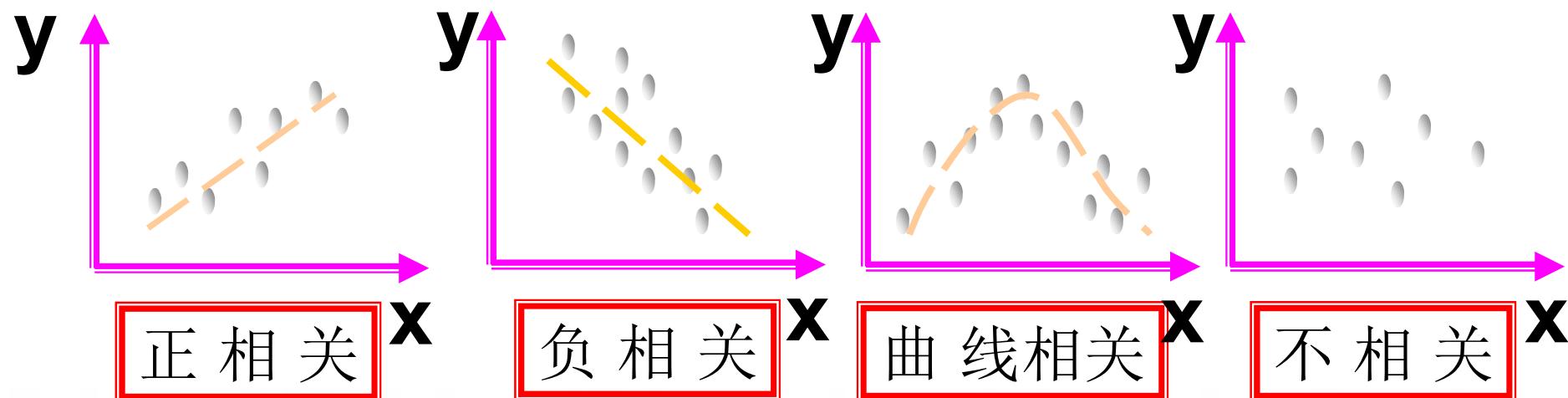
北京大学



# 相关图

又称**散点图**，用以表明相关点分布状况的图形

- 用x轴代表自变量，y轴代表因变量
- 将两变量间相对应的变量值用坐标点的形式描绘出来





# 相关系数

- 在直线相关的条件下，用以反映两变量间线性相关密切程度的统计指标，用 $r$ 表示
- 变量的取值区间越大，观测值个数越多，相关系数受抽样误差的影响越小，结果越可靠
- 如果数据较少，本不相关的两列变量，计算的结果可能相关
- 相关系数取值:  $-1 < r < 1$



北京大学



## Pearson相关系数：适用条件

- 两变量均应由测量得到的连续变量
- 两变量所来自的总体都应是正态分布，或接近正态的单峰对称分布
- 变量必须是成对的数据
- 两变量间为线性关系



北京大学



# Pearson相关系数：计算公式

$$r = \frac{S_{XY}^2}{S_X S_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})/n}{\sqrt{\sum (X - \bar{X})^2/n} \sqrt{\sum (Y - \bar{Y})^2/n}}$$
$$= \frac{n \sum XY - \sum X \sum Y}{\sqrt{n \sum X^2 - (\sum X)^2} \sqrt{n \sum Y^2 - (\sum Y)^2}}$$

$X$  的离均差平方和： $S_X = \sum (X - \bar{X})^2$

$Y$  的离均差平方和： $S_Y = \sum (Y - \bar{Y})^2$

$X$ 与 $Y$ 间的离均差积和： $S_{XY} = \sum (X - \bar{X})(Y - \bar{Y})$



北京大学



# 判定系数

- 相关系数的平方，用 $r^2$ 表示
- 用来衡量回归方程对 $y$ 的解释程度
- 判定系数取值范围： $0 \leq r^2 \leq 1$
- $r^2$ 越接近于1，表明 $x$ 与 $y$ 之间的相关性越强
- $r^2$ 越接近于0，表明两个变量之间几乎没有直线相关关系



北京大学



# 相关系数的假设检验

## 1. 提出假设

$$H_0 : \rho = 0 \quad \text{无关}$$

$$H_1 : \rho \neq 0 \quad \text{相关}$$

## 2. 确定显著性水平 $\alpha=0.05$

如果从相关系数 $\rho=0$ 的总体中取得某 $r$ 值的概率 **$P>0.05$** ，我们就接受假设，认为此 $r$ 值的很可能从此总体中取得。因此判断两变量间**无显著关系**；

如果取得 $r$ 值的概率 **$P\leq 0.05$ 或 $P\leq 0.01$** ，我们就在 $\alpha=0.05$ 或 $\alpha=0.01$ 水准上拒绝检验假设，认为该 $r$ 值不是来自 $\rho=0$ 的总体，而是来自 $\rho\neq 0$ 的另一个总体，因此就判断两变量间**有显著关系**。

## 3. 计算检验统计量，查表得到P值。拒绝 $H_0$ ，则两变量相关。否则，两变量无关。



北京大学



# 相关系数的假设检验

t检验法 计算检验统计量 $t_r$ , 查 $t$ 界值表, 得到 $P$ 值

$$t_r = \frac{|r - 0|}{\sqrt{\frac{1 - r^2}{n - 2}}} \quad v = n - 2$$



北京大学



## 小结：如何判断两变量的相关性

- (1) 找出两个变量的正确**相应**数据
- (2) 画出它们的**散布图** (**散点图**)
- (3) 通过**散布图****判断**它们的相关性
- (4) 给出**相关** ( $r$ ) 的解答
- (5) 对结果进行评价和**检验**





## 注意事项

- 线性相关的前提条件是X、Y都服从正态分布（双变量正态分布）
- 当散点图有线性趋势时，才可进行线性相关分析
- 必须在假设检验认为相关的前提下才能以 $r$ 的大小判断相关程度
- 相关关系并不一定是因果关系，有可能是伴随关系





# 秩相关

- 用双变量计量或等级数据作直线相关分析
- 对原变量分布不作要求，为**非参数统计方法**
- 适用范围：
  - 不服从双变量正态分布
  - 总体分布类型未知或有“超限值”（如 $X < 0.01$ ）
  - 原始数据用等级表示



北京大学



## Spearman秩相关检验：基本思想

- 将 $n$ 对观察值 $X_i$ 、 $Y_i(i=1,2,\dots,n)$ 分别由小到大编秩， $P_i$ 表示 $X_i$ 的秩， $Q_i$ 表示 $Y_i$ 的秩
- 其中每对 $P_i$ 、 $Q_i$ 可能相等，也可能不等
- 用 $P_i$ 与 $Q_i$ 之差来反映 $X$ 、 $Y$ 两变量秩排列一致性的情况，令 $d_i = P_i - Q_i$
- $n$ 为一定时，当每对 $X_i$ 、 $Y_i$ 的秩完全相等（完全正相关）， $\sum d_i^2$ 有最小值0
- 当每对 $X_i$ 、 $Y_i$ 的秩完全相反（完全负相关）， $\sum d_i^2$ 有最大值 $n(n^2-1)/3$



北京大学



# Spearman秩相关检验：基本思想

$\sum d_i^2$  从 0 到  $n(n^2 - 1)/3$  间变化，反映了  $X$ 、 $Y$  两变量的相关程度。

按以下公式计算 Spearman 等级相关系数

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

$r_s$  值界于 -1 与 1 之间， $r_s$  为正表示正相关， $r_s$  为负表示负相关， $r_s$  等于零为零相关。

样本等级相关系数  $r_s$  是总体等级相关系数  $\rho_s$  的估计值。

检验  $\rho_s$  是否不为零可用查表法，当  $n > 50$  时，可用计算法，计算检验统计量  $u$

$$u = r_s \sqrt{n-1}$$



北京大学



# 回归分析



北京大学



# 回归分析的起源

- F.Galton和K.Pearson收集了1078个家庭的身高记录
- 寻找儿子们身高与父亲们身高之间关系的具体形式
- 下图是根据1078个家庭的调查所作的散点图

父亲们身高：

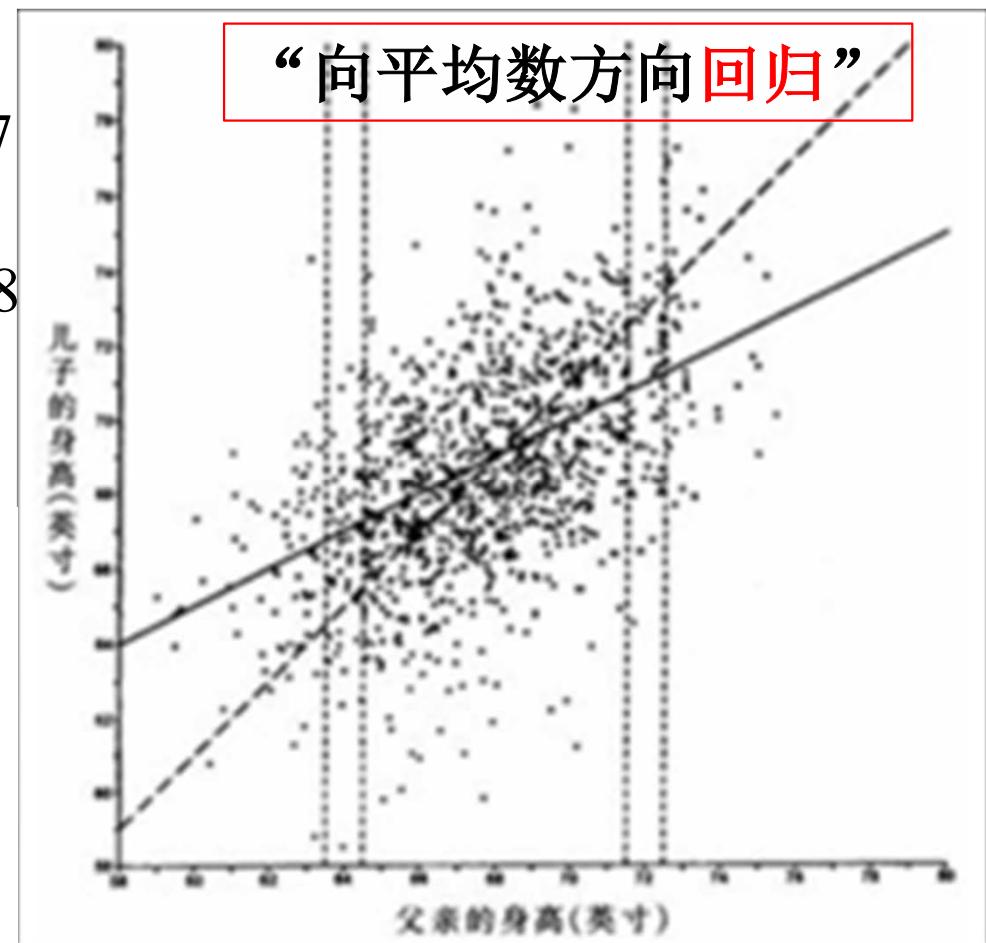
平均值： $\bar{X} \approx 68$ ; 标准差： $S_x \approx 2.7$

儿子们身高：

平均值： $\bar{Y} \approx 69$ ; 标准差： $S_y \approx 2.8$

斜虚线：依平均身高推测的关系线

斜实线（回归线）：线上的点是当给定某一 $X_i$ 值（父亲身高值）时，对应的若干 $Y_i$ 值（儿子的身高值）与之（直线上点 $Y$ 值）离差平方和最小的直线





# 回归分析的主要内容

- 根据研究目的和现象之间的内在联系，确定自变量和应变量
- 确定回归模型的类型及数学表达式（曲线拟合）
- 对回归分析模型进行评价与诊断（统计检验）
- 根据给定的自变量值确定应变量的值（预测）



北京大学



## 常见的曲线拟合方法

(1) 使偏差绝对值之和最小, 即

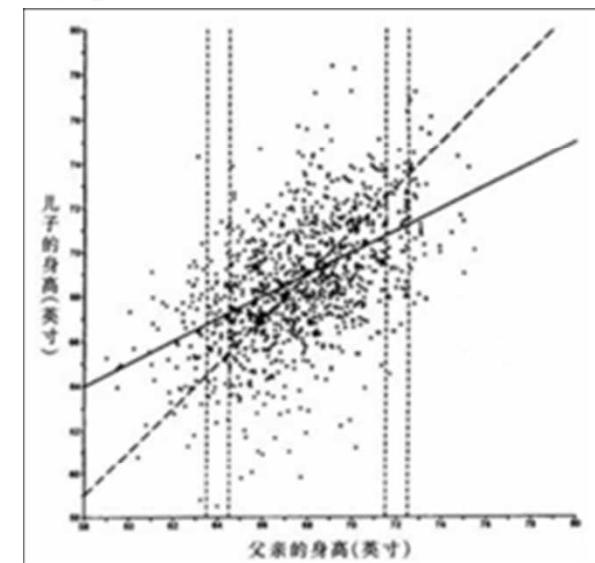
$$\min_{\varphi} \sum_{i=1}^m |\delta_i| = \sum_{i=1}^m |\varphi(x_i) - y_i|$$

(2) 使偏差绝对值最大的最小, 即

$$\min_{\varphi} \max_i |\delta_i| = |\varphi(x_i) - y_i|$$

(3) 使偏差平方和最小, 即

$$\min_{\varphi} \sum_{i=1}^m \delta_i^2 = \sum_{i=1}^m (\varphi(x_i) - y_i)^2$$



按偏差平方和最小的原则选取拟合曲线的方法, 称为最小二乘法。



# 回归模型的类型

一个自变量

回归模型

两个及两个以上自变量

一元回归

多元回归

线性  
回归

非线性  
回归

线性  
回归

非线性  
回归

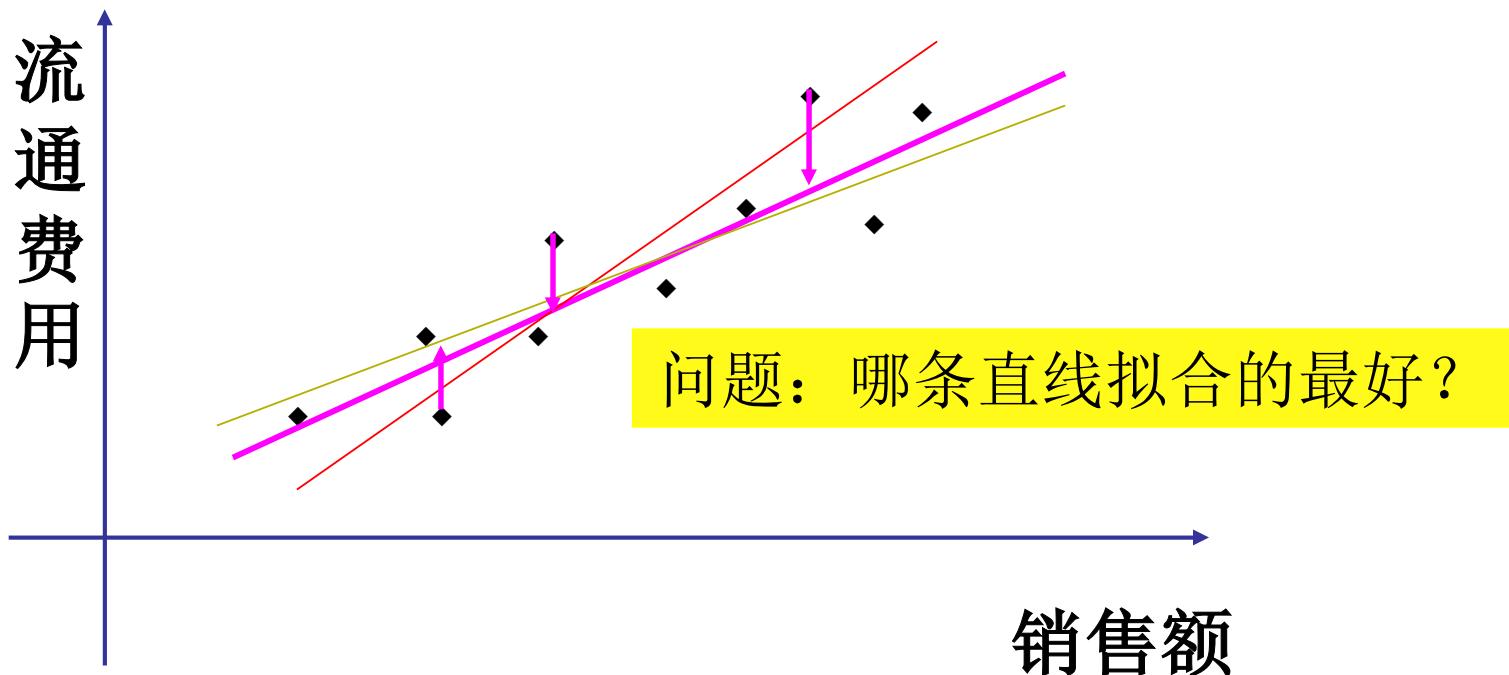
$$y = \varphi(x_1, x_2, x_3, \dots, x_p) + \varepsilon$$



北京大学



# 参数估计（曲线拟合）



思路：离差的平方和最小

最小二乘法



北京大学



# 判定系数

- 回归变差占总变差的比例，称为判定系数

$$R^2 = \frac{SSR}{SST} = \frac{\sum (y_c - \bar{y})^2}{\sum (y - \bar{y})^2} = 1 - \frac{\sum (y_c - y)^2}{\sum (y - \bar{y})^2}$$

- 反映回归直线的拟合优度的统计指标，取值范围为  $[0, 1]$
- $R^2 \rightarrow 1$ , 说明回归方程拟合得越好
- $R^2 \rightarrow 0$ , 说明回归方程拟合越差。
- 在一元线性模型中，判定系数等于相关系数的平方 ( $R^2 = r^2$ )



北京大学



# 回归方程的统计检验： 参数显著性的检验

- 判断每个自变量对于回归模型是否必要的
- 一元线性回归模型截距和斜率的显著性检验 ( $t$ 检验)

$$t = \frac{a}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x - \bar{x})^2}}} \propto t(n-2)$$

$$t = \frac{b}{\hat{\sigma} \sqrt{\frac{1}{\sum (x - \bar{x})^2}}} \propto t(n-2)$$

其中  $\hat{\sigma} = \sqrt{\frac{\sum (y - y_c)^2}{n - 2}} = S_{yx}$



北京大学



# 回归分析的模型检查

- 前提假设:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i=1, 2, \dots, n$$

- 其中  $\varepsilon_i \sim N(0, \sigma^2)$
- $\varepsilon_i$  是相互独立且它们的方差相同

- 模型检查: 检查对模型所做的假设是否成立
  - $\varepsilon_i$  是相互独立的随机变量序列的检查
  - $\varepsilon_i$  是方差齐性的检查



北京大学



## 残差图

- 残差：观测值与理论值的差

$$\gamma_i = y_i - \hat{y}_i$$

- 标准化残差：

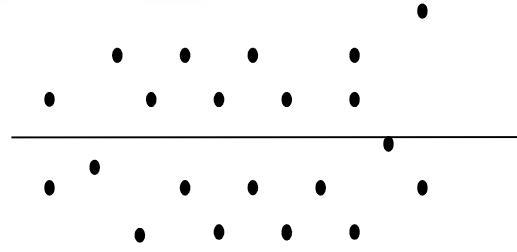
$$e_i = \frac{\gamma_i}{\sqrt{D(\gamma_i)}}$$

- 残差图：

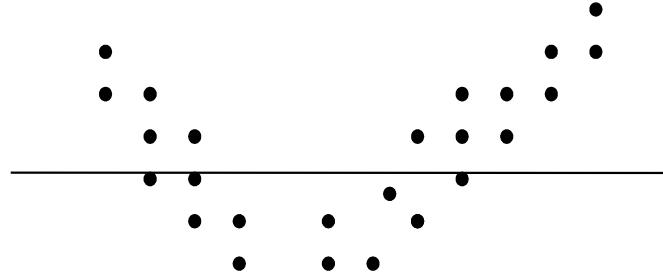
- 以 $x$ 为坐标横轴，残差 $e$ 为坐标纵轴，由所有点 $(x_i, e_i)$ 构成
- 可检验随机变量序列 $\varepsilon_i$ 的**独立性**, **正态性**和**方差整齐性**
  - 理论上可证明 $e_1, e_2, \dots, e_n$ 应**相互独立**且近似的**服从 $N(0, 1)$**
  - 正常残差图：残差图中的点随机分布在-2到+2之间的**子带**里



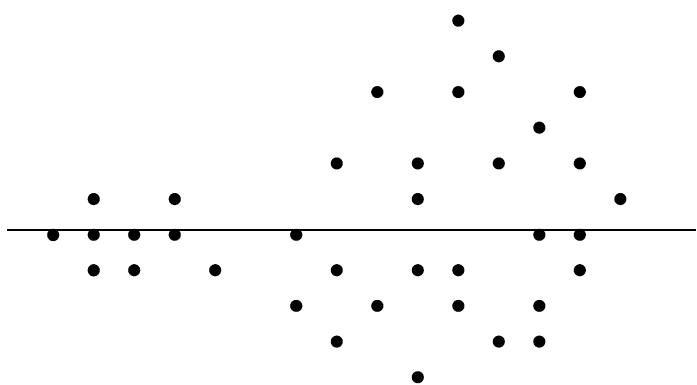
北京大学



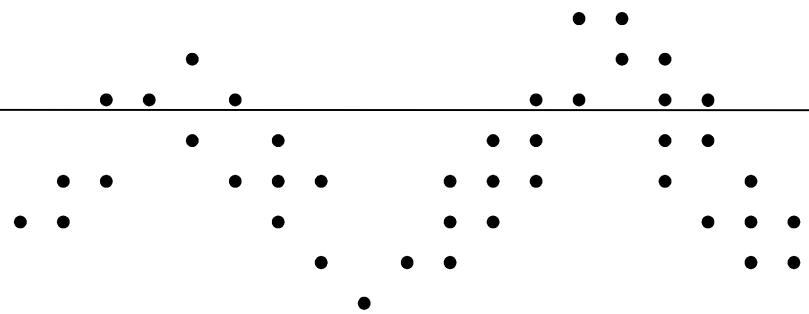
1. 正常的残差图



2. 直线回归模型不合适



3. 方差齐性不成立



4. 误差项不独立



北京大学



# 方差齐性的诊断及修正方法

- 误差方差非齐性时, 残差图不正常
- 可通过对**应变量**作适当的**变换**
  - 令 $z=f(y)$ 使得回归分析中误差的方差接近于齐性
  - 变换后重新做回归及残差图, 如残差图有改善或已属正常, 则该变换是合适的; 否则改变变换函数计算直到找到合适的变换为止
- 常用变换有:

$$Z = \ln(Y) \quad Z = \sqrt{Y} (Y > 0) \quad Z = \frac{1}{Y} (Y \neq 0)$$



北京大学



# 自变量的选择

- 动机
  - 若漏掉显著的自变量，实际预测时会产生较大的偏差
  - 若包括了不显著的自变量，也会影响到预测的精度
- 最优回归方程
  - 对y的作用有统计学意义的自变量，全部选入回归方程
  - 对y的作用没有统计学意义的自变量，一个也不引入
- 问题描述：
  - 从自变量集 $\{x_1, x_2, \dots, x_p\}$ 中选出子集 $A = \{x_{i_1}, x_{i_2}, \dots, x_{i_l}\}$
  - 使得建立应变量  $y$  与子集的回归方程 “最优”



北京大学



## 自变量选择：评价标准

- 均方误差  $S^2$  最小：

$$S^2(A) = SSE(A)/(n - l - 1)$$

其中  $SSE(A)$  是残差平方和,  $l$  是子集中自变量的个数

- 赤池信息准则 (Akaike information criterion, AIC)

$$AIC = n \ln(SSE) + 2l$$

- 贝叶斯信息准则 (Bayesian Information Criterion, BIC)

$$BIC = n \ln(SSE) + l \ln(n)$$

- 修正的  $R^2$  准则

$$R'^2 = 1 - \frac{n - i}{n - l} (1 - R^2)$$



北京大学



# 自变量选择：选择方法

- 全局择优法：
  - 从所有可能的回归模型选取最优者
- 向后剔除法 (backward selection)
- 向前引入法 (forward selection)
- 逐步回归法 (stepwise regression)



北京大学



# 非线性回归

- 可化为线性的回归模型

1、指数曲线模型

$$y_c = ab^x \Rightarrow \ln y_c = \ln a + x \ln b$$

2、对数曲线模型

$$y_c = a + b \ln x, \text{令} x' = \ln x$$

3、双曲线模型

$$\frac{1}{y_c} = a + b \frac{1}{x}, \text{令} y'_c = \frac{1}{y_c}, x' = \frac{1}{x}$$

4、幂函数曲线模型

$$y_c = ax^b \Rightarrow \lg y_c = \lg a + b \lg x$$

5、抛物线模型

$$y_c = a + bx + cx^2, \text{令} x_1 = x, x_2 = x^2$$



北京大学



# 期末考试

- 考试范围：
  - PPT + 课后习题
- 题型：
  - 名词解释
  - 基本公式
  - 简答
  - 计算



北京大学