相关分析

主讲人：刘宏志

liuhz@ss.pku.edu.cn

北京大学

两种现象间的依存关系

出租汽车费用与行驶里程

函数关系
（确定性关系）

总费用=行驶里程 × 每公里单价

家庭收入与恩格尔系数

相关关系
（非确定性关）

家庭收入高，则恩格尔系数低

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。

因变量 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*因变量*，通常用*Y*表示
  - 它是自变量变化的结果，是不确定的值。

研究居民收入水平与储蓄存款余额的关系

居民收入水平是自变量，储蓄存款余额是因变量

工业产值与工业贷款额的关系

如果研究工业生产规模对工业贷款额的需求量的影响，工业产值是自变量，工业贷款就是因变量；

如果研究贷款量对工业生产规模的影响情况，工业贷款额是自变量，工业产值是因变量。

有时相关关系表现的因果关系不明显，要根据研究目的来确定

相关关系的种类

相关关系的种类

- 按涉及变量的多少分为
  - 一元相关
  - 多元相关
- 按照表现形式不同分为
  - 直线相关
  - 曲线相关
- 按照变化方向不同分为
  - 正相关
  - 负相关

北京大学

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

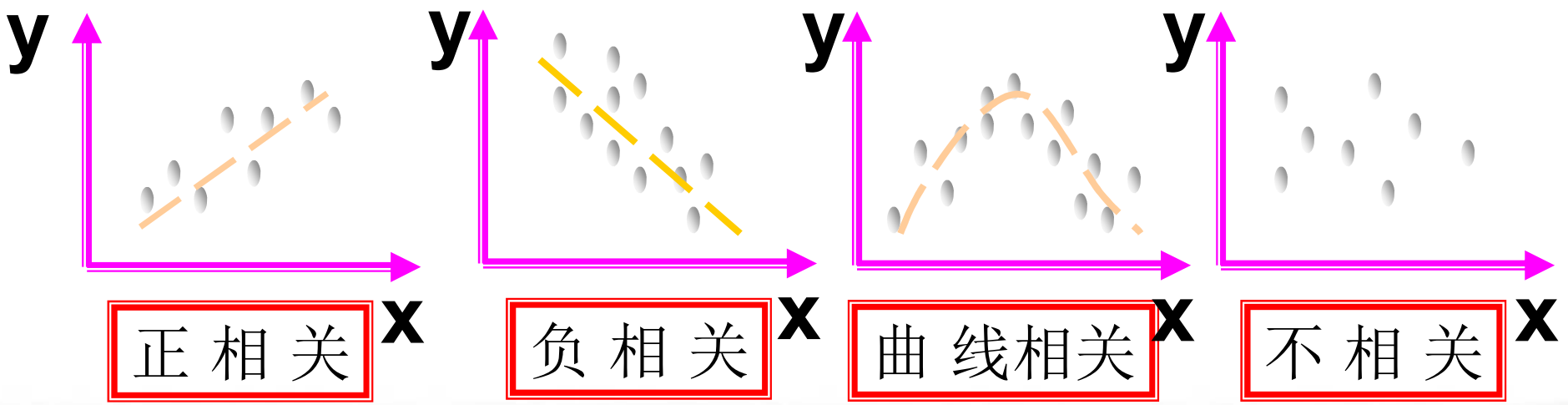
相关关系的测定

- 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

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



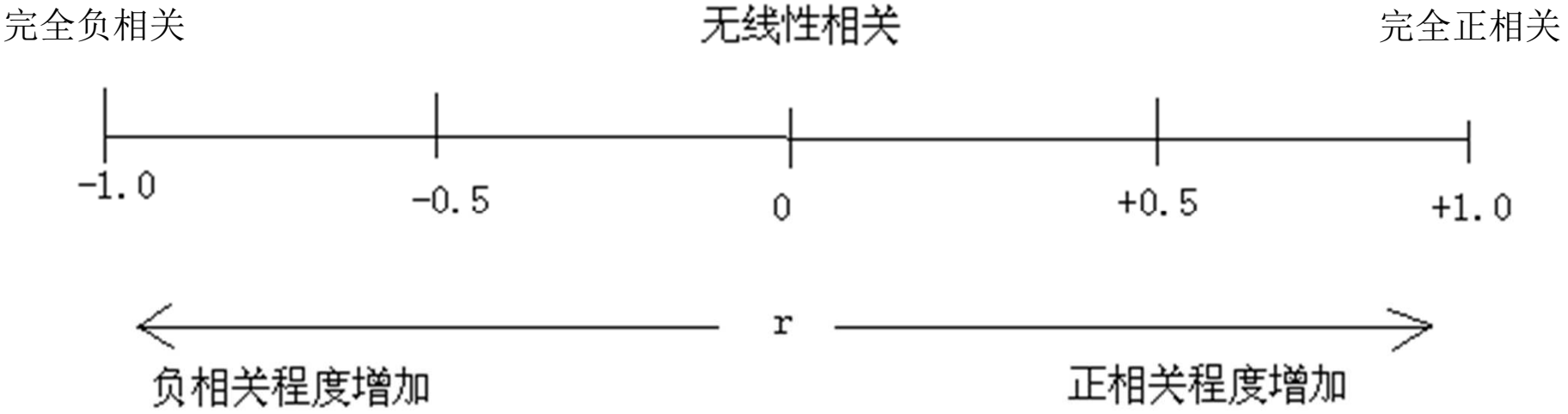
北京大学

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**

相关系数的性质

- |r|表明两变量间相关的程度
- r>0表示正相关
- r<0表示负相关
- r=0表示零相关



相关系数的性质

- |r|越接近于1，表明两变量相关程度越高

|r|的取值与相关程度

|r|的取值范围	|r|的意义
0.00-0.19	极低相关
0.20-0.39	低度相关
0.40-0.69	中度相关
0.70-0.89	高度相关
0.90-1.00	极高相关

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

Pearson相关系数：计算公式

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

$$= \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

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

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

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

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

示例：血铅 vs. 尿铅

- 测得某地15名正常成年人的血铅X和24小时的尿铅Y，试分析X与Y之间是否直线相关

编号	X	Y	编号	X	Y
1	0.11	0.14	9	0.23	0.24
2	0.25	0.25	10	0.33	0.30
3	0.23	0.28	11	0.15	0.16
4	0.24	0.25	12	0.04	0.05
5	0.26	0.28	13	0.20	0.20
6	0.09	0.10	14	0.34	0.32
7	0.25	0.27	15	0.22	0.24
8	0.06	0.09			

$\sum X=3.00$ $\sum Y=3.17$ $\sum X^2=0.7168$

$\sum Y^2=0.7681$ $\sum XY=0.7388$ $n=15$

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

$$=\frac{n\sum XY-\sum X\sum Y}{\sqrt{n\sum X^2-(\sum X)^2}\sqrt{n\sum Y^2-(\sum Y)^2}}=0.9787$$

相关系数的假设检验

- 上例中的相关系数*r*等于0.9787，说明了15例样本中血铅与尿铅之间存在相关关系。
- 但这15例只是总体中的一个样本，由此得到的相关系数会存在抽样误差。因为，总体相关系数（ρ）为零时，由于抽样误差，从总体抽出的15例，其r可能不等于零。
- 所以，要判断该样本的r是否有意义，需与总体相关系数ρ=0进行比较，看两者的差别有无统计学意义。这就要对r进行假设检验，判断r不等于零是由于抽样误差所致，还是两个变量之间确实存在相关关系。

相关系数的假设检验

1.提出假设　　H_0：　$\rho=0$　　无关

H_1：　$\rho\neq0$　　相关

2.确定显著性水平 $\alpha=0.05$

如果从相关系数ρ=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值的很可能是从此总体中取得的。因此判断两变量间无显著关系；

如果取得r值的概率P≤0.05或P≤0.01，我们就在α=0.05或α=0.01水准上拒绝检验假设，认为该r值不是来自ρ=0的总体，而是来自ρ≠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。

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

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

示例

1. H_0：　ρ=0　　无关

   H_1：　ρ≠0　　相关　　α=0.05

2. r=0.9787, n=15, 代入公式$t_r = \frac{|r-0|}{\sqrt{\frac{1-r^2}{n-2}}} = 17.189$

3. v=15-2=13，查t界值表, P<0.001，拒绝H_0，认为血铅与尿铅之间有正相关关系。

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

Spearman秩相关检验：基本思想

- 将n对观察值X_i、$Y_i(i=1,2,\ldots,n)$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$n(n^2-1)/3$

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 ρ_s 的估计值。

检验 ρ_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

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

r_s 界值表

n	$r_{s0.05}$	$r_{s0.01}$	n	$r_{s0.05}$	$r_{s0.01}$
6	0.886	1.000	29	0.368	0.475
7	0.786	0.929	30	0.362	0.467
8	0.738	0.881	31	0.356	0.459
9	0.700	0.833	32	0.350	0.452
10	0.648	0.794	33	0.345	0.446
11	0.618	0.755	34	0.340	0.439
12	0.587	0.727	35	0.335	0.433
13	0.560	0.703	36	0.330	0.427
14	0.538	0.679	37	0.325	0.421
15	0.521	0.654	38	0.321	0.415
16	0.503	0.635	39	0.317	0.410
17	0.485	0.615	40	0.313	0.405
18	0.472	0.600	41	0.309	0.400
19	0.460	0.584	42	0.305	0.395
20	0.447	0.570	43	0.301	0.391
21	0.435	0.556	44	0.298	0.386
22	0.425	0.544	45	0.294	0.382
23	0.415	0.532	46	0.291	0.378
24	0.406	0.521	47	0.288	0.374
25	0.398	0.511	48	0.285	0.370
26	0.390	0.501	49	0.282	0.366
27	0.382	0.491	50	0.297	0.363
28	0.375	0.483			

示例

- 某省调查了1995年到1999年当地居民18类死因的构成以及每种死因导致的潜在工作损失年数WYPLL的构成。
- 以死因构成为X，WYPLL构成为Y，作等级相关分析。

某省 1995 年到 1999 年居民死因构成与 WYPLL 构成

类别	死因构成（%）		WYPLL 构成（%）		d	d^2	PQ
（1）	X（2）	P（3）	Y（4）	Q（5）	(6)=(3)-(5)	(7)=(6)^2	(8)=(3)(5)
1	0.03	1	0.05	1	0	0	1
2	0.14	2	0.34	2	0	0	4
3	0.20	3	0.93	6	-3	9	18
4	0.43	4	0.69	4	0	0	16
5	0.44	5	0.38	3	2	4	15
6	0.45	6	0.79	5	1	1	30
7	0.47	7	1.19	8	-1	1	56
8	0.65	8	4.74	12	-4	16	96
9	0.95	9	2.31	9	0	0	81
10	0.96	10	5.95	14	-4	16	140
11	2.44	11	1.11	7	4	16	77
12	2.69	12	3.53	11	1	1	132
13	3.07	13	3.48	10	3	9	130
14	7.78	14	5.65	13	1	1	182
15	9.82	15	33.95	18	-3	9	270
16	18.93	16	17.16	17	-1	1	272
17	22.59	17	8.42	15	2	4	255
18	27.96	18	9.33	16	2	4	288
合 计	—	171	—	171	—	92	2063

H_0：$\rho_s = 0$，即死因构成和 WYPLL 构成之间无直线相关关系

H_1：$\rho_s \neq 0$，即死因构成和 WYPLL 构成之间有直线相关关系

α=0.05

将两变量 X、Y 的实测值分别从小到大编秩(即秩变换)，用 P_i 和 Q_i 表示。每个变量中若有观察值相同则取平均秩。求每对秩的差值 d、d^2、$\sum d^2$，计算统计量 r_s。

$$r_s = 1 - \frac{6(92)}{18^3 - 18} = 0.905$$

本例 $n = 18$，查 r_s 界值表，得 P<0.01。按 $\alpha = 0.05$ 水准拒绝 H_0，接受 H_1，可认为当地居民死因的构成和各种死因导致的潜在工作损失年数 WYPLL 的构成存在正相关关系。

相同秩较多时r_s的校正

- 对X与Y分别排秩时，若相同秩较多，宜用下式计算校正

$$r_s' = \frac{[(n^3 - n)/6] - (T_X + T_Y) - \sum d^2}{\sqrt{[(n^3 - n)/6] - 2T_X}\sqrt{[(n^3 - n)/6] - 2T_Y}}$$

- T_x(或T_Y)＝Σ(t^3－t)/12，t为X(或Y)中相同秩的个数
- 当T_x＝T_Y＝0时，r_s'与r_s相等

Kendall-τ秩相关系数

- 从二维随机样本是否协同一致来检验两变量之间是否存在相关性
- 协同: 对样本$(x_i, y_i)^T, i = 1,2,\cdots,n$
  - 若$(x_j - x_i)(y_j - y_i) > 0, \forall j > i \ \ (i,j = 1,2,\cdots,n)$

    则称数对(x_i, y_i)与(x_j, y_j)满足协同性

    即第j对观测值的两个分量同时比第i对观测值的两个分量大（或小）
  - 若$(x_j - x_i)(y_j - y_i) < 0, \forall j > i \ \ (i,j = 1,2,\cdots,n)$

    则称数对(x_i, y_i)与(x_j, y_j)不协同

Kendall-τ秩相关系数

N_c表示协同数对的数目

N_d表示不协同数对的数目

$$N_c + N_d = C_n^2$$

无结点时, Kendall相关系数定义为

$$\tau = \frac{N_c - N_d}{\frac{n(n-1)}{2}}$$

若所有数对全部协同，$N_c = \frac{n(n-1)}{2}$，$N_d = 0$，$\tau = 1$

若所有数对全部不协同，$N_c == 0$，$N_d \frac{n(n-1)}{2}$，$\tau = -1$

$$-1 \le \tau \le 1$$

Kendall-τ秩相关系数

定义

$$\text{sign}((X_1 - X_2)(Y_1 - Y_2)) = \begin{cases} 1, & (X_1 - X_2)(Y_1 - Y_2) > 0 \\ 0, & (X_1 - X_2)(Y_1 - Y_2) = 0 \\ -1, & (X_1 - X_2)(Y_1 - Y_2) < 0 \end{cases}$$

则

$$\tau = \frac{2}{n(n-1)} \sum_{1 \leqslant i < j \leqslant n} \text{sign}((x_i - x_j)(y_i - y_j))$$

1.样本数较小时，用检验临界值表

2.样本数较大时，利用正态分布的近似

3.当有结点时,我们用平均秩计算秩 得到统计量 τ^*