# 心理统计
# 第十三讲：相关

**严超赣**

Chao-Gan Yan, Ph.D.

yancg@psych.ac.cn
http://rfmri.org/yan

Institute of Psychology, Chinese Academy of Sciences



1

## 主要内容





2

## 相关的概念

- 相关是度量和描述两个变量之间关系的一种统计技术.
- 数据要求：一定要有至少两个变量，两组分数。

3

## 应用相关的研究情境

- 预测 – 如果两个变量间有强相关，我们就可以根据一个变量的值，预测另一个变量的值。
  - 如，如果知道某些人格特征，可以预测员工绩效
- 相容效度 – 如果发明新的心理测验（测验A），想知道它是否测量了X，就需要知道测验A 是否与X相关.
- 效标关联效度 – 如果发明新的量表，管理潜能量表来预测晋升所需时间，这个量表分数应当与晋升所需时间相关。
- 重测信度 – 如果对同一组被试两次用相同的测验，将两组分数做相关. 如果测验是可信的，两次测验应当得到相似的结果，产生高相关
- 理论验证 – 比如验证社交技能与焦虑的相关

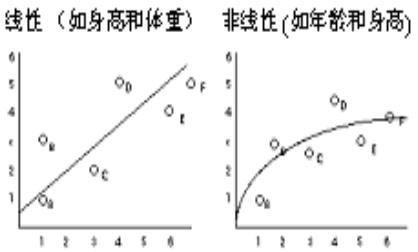
4

## 相关表明变量 X 和 Y之间关系的3个特征

- 1) 关系的方向
  - 正相关（正数）意味着两个变量向相同的方向变化. 亦即，一个变量增加，另一个变量也增加.
  - 负相关（负数）意味着两个变量向相反的方向变化. 亦即，一个变量增加，另一个变量反而减少.
- 2) 关系的形式
  - 本课集中讨论线性(直线)相关，但两变量的关系也有其他形式
- 3) 关系的程度
  - 相关也度量了X 和 Y间关系的强度. 相关系数的值 在-1 和+1之间. 0相关意味着没有关系. +1 意味着"完全的正相关"之间 两个，-1 意味着完全的负相关.

5

## 线性(直线)相关与非线性相关



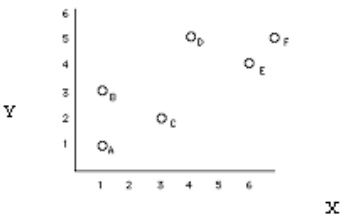

6

## 相关的数据和散点图

数据

| Person | X | Y |
|---|---|---|
| A | 1 | 1 |
| B | 1 | 3 |
| C | 3 | 2 |
| D | 4 | 5 |
| E | 6 | 4 |
| F | 7 | 5 |

散点图




7

## 散点图示例

| x | y |
|---|---|
| 1 | 2 |
| 2 | 4 |
| 3 | 1 |
| 4 | 5 |
| 5 | 3 |
| 6 | 9 |
| 7 | 10 |
| 8 | 7 |
| 9 | 8 |
| 10 | 6 |

plot(X(:,1),X(:,2),'o')

lsline

[r p]=corrcoef(X(:,1),X(:,2))

8

8

## 4）解释关系强度应考虑 $r^2$, 不只是 r.

- 它表明了：一个变量的方差中，由X和Y间的相关解释的方差的比例
- 当r=0.7时，Y变异的一部分能由X推出，$r^2$=0.49，即Y 49%的变异能够由X推出。


9

## 积差相关的效应大小

- Cohen's Convention：
  - r=.10 小的效应
  - r=.30 中等效应
  - r=.50 大的效应

10

## 积差相关的相关系数与统计效力换算表

| 双尾 | .10 | .30 | .50 |
|---|---|---|---|
| N | | | |
| 10 | .06 | .13 | .33 |
| 20 | .07 | .25 | .64 |
| 30 | .08 | .37 | .83 |
| 40 | .09 | .48 | .92 |
| 50 | .11 | .57 | .97 |
| 100 | .17 | .86 | 1.0 |

11

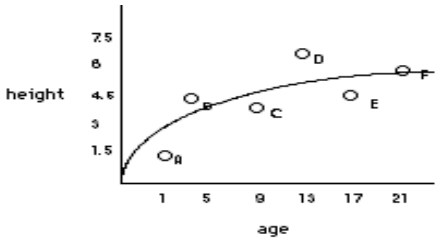
## 5）相关描述两个变量之间的关系, 但并不能解释变量相关的原因

- 最基本的原因是相关研究的性质。在相关研究中, 研究者没有操纵一个 (或几个) 变量而保持其他变量不变。因此，相关计算绝不能得到因果性推论。
- 伪相关（spurious correlation）
  - *一位研究者发现某月酒咖啡的次数和空难次数呈强的正相关.*
  - *一位研究者发现警察局的规模X和犯罪量Y呈强的正相关*
  - 这里我们发现另一个变量 Z, 同时导致 X 和 Y, X 和 Y 其实并不是因果关系

12

## 6）数据中的分数范围对相关有非常大的影响

- 年龄和身高之间的相关




13

## 6）数据中的分数范围对相关有非常大的影响

- 研究者A随机抽取了n=3000的企业人员样本，发现管理素质量表分数与管理绩效指标的相关是 .62。研究者B随机抽取了n=80的银行支行经理，发现同样的管理素质量表分数与同样的管理绩效指标的相关仅为 .19。

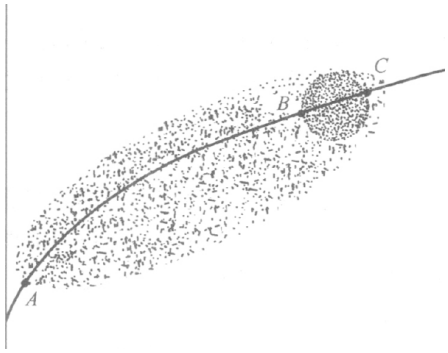
A）研究者B的样本小，相关就低

B）研究者B的发现是抽样误差

C）银行支行经理的职位都比较高，因此管理素质量表分数差别不大

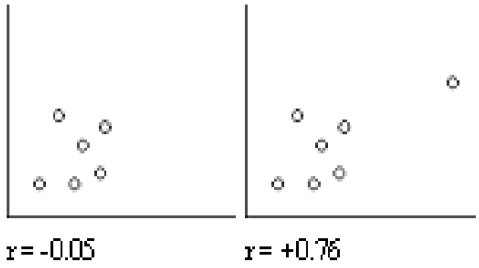
D）研究者B的样本中管理绩效指标的指标都分布在高分范围

14




15

## 7) 极端的分数（outlier）对相关有非常大的影响




16

## 将相关的概念数量化

- 我们主要讨论两种相关，Pearson 积差相关，Spearman 等级相关.
  - r = X 和 Y 共同变化的程度 / X 和 Y各自变化的程度 = X和Y的协方差 / X和Y各自的方差
  - 共变意味着随着X 变化，Y 也变化.
  - r = 1.0（或 -1.0）即"完全的相关".意味着分子分数等于分母分数。

17

## 离差的乘积和(SP)

- 定义公式：SP = $\sum(X-\overline{X})(Y-\overline{Y})$
- 对于每一点与X 和 Y的平均值的差，即离差，求两个离差的乘积，再求和
- SP的计算公式：$\sum XY - \frac{\sum X \sum Y}{n}$

18

## 乘积和(SP)与和方(SS)公式

- 非常相似，其区别是 SS只有一个变量 (X)，SP 有两个变量 (X 和 Y).

| | 和方 (SS) | 乘积和 (SP) |
|---|---|---|
| 定义公式 | $\sum(X-\overline{X})^2$ | $\sum(X-\overline{X})(Y-\overline{Y})$ |
| 计算公式 | $\sum X^2-\frac{(\sum X)^2}{n}$ | $\sum XY-\frac{\sum X\sum Y}{n}$ |

19

## 用计算公式计算SP

| X | Y | XY |
|---|---|---|
| 0 | 1 | 0 |
| 10 | 3 | 30 |
| 4 | 1 | 4 |
| 8 | 2 | 16 |
| 8 | 3 | 24 |
| 30 | 10 | 74 |

$SP=\sum XY-\frac{\sum X\sum Y}{n}$

$=74-\frac{(30)(10)}{5}$

$=74-60=14$

20

## Pearson 相关的计算

- 也称积差相关（product-moment correlation）
- $r=\frac{SP}{\sqrt{SSx\,SSy}}$
- 分子SP是X和Y协方差的指标. 分母是X和Y各自的变异

21

## 例2：计算以下两列数据的积差相关

| 被试 | X | Y |
|---|---|---|
| A | 0 | 4 |
| B | 2 | 1 |
| C | 8 | 10 |
| D | 6 | 9 |
| E | 4 | 6 |

22

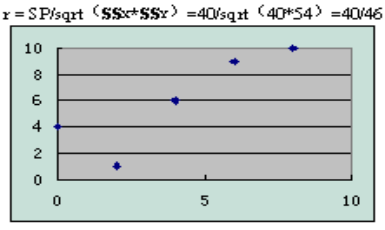
| X | Y | $X^2$ | $Y^2$ | XY |
|---|---|---|---|---|
| 0 | 4 | 0 | 16 | 0 |
| 2 | 1 | 4 | 1 | 2 |
| 8 | 10 | 64 | 100 | 80 |
| 6 | 9 | 36 | 81 | 54 |
| 4 | 6 | 16 | 36 | 24 |
| $\Sigma X=20$ | $\Sigma Y=30$ | $\Sigma X^2=120$ | $\Sigma Y^2=234$ | $\Sigma XY=160$ |

$SS_X=\Sigma X^2-(\Sigma X)^2/n=120-20^2/5=120-80=40$

$SS_Y=\Sigma Y^2-(\Sigma Y)^2/n=234-30^2/5=234-180=54$

$SP=\Sigma XY-\Sigma X\,\Sigma Y/n=160-20*30/5=160-120=40$

$r=SP/\mathrm{sqrt}(SS_X*SS_Y)=40/\mathrm{sqrt}(40*54)=40/46.48=.861$




23

## 相关系数的显著性检验

- 总体参数 ρ，样本统计量 r.
- 虚无假设和备择假设
  - 双侧:
    $H_0$: ρ = 0, X和Y 之间无相关
    $H_1$: ρ ≠0
  - 单侧:
    没有正相关. $H_0$: ρ ≤ 0 $H_1$: ρ > 0
    df = n − 2
- 查表或 $t=\frac{r}{\sqrt{\frac{1-r^2}{N-2}}}$

24

## 在论文中报告相关

- 对数据的相关分析显示受教育年限与年收入有显著相关，r(474)=+.66，p＜.01，双尾。

25

## 等级相关 (Spearman相关, 斯皮尔曼相关)

- 一种非参数检验
- 用于顺序型数据，非线性数据
- $r_s = 1 - 6\Sigma D^2/n(n^2-1)$
  - D – 各自排序后的等级差

26

## 计算下列变量的等级相关

| X | Y |
|---|---|
| 2 | 4 |
| 5 | 3 |
| 6 | 5 |
| 9 | 8 |
| 14 | 10 |

27

| X | Y | X的等级 | Y的等级 | D | $D^2$ |
|---|---|---|---|---|---|
| 2 | 4 | 1 | 2 | 1 | 1 |
| 5 | 3 | 2 | 1 | −1 | 1 |
| 6 | 5 | 3 | 3 | 0 | 0 |
| 9 | 8 | 4 | 4 | 0 | 0 |
| 14 | 10 | 5 | 5 | 0 | 0 |
| | | | | | $\Sigma D^2=2$ |

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6\times 2}{5(25-1)} = 0.9$$

28

## 计算下列变量的等级相关

| X | Y |
|---|---|
| 2 | 4 |
| 5 | 3 |
| 6 | 5 |
| 9 | 8 |
| 14 | 10 |

```
[rho pval]=corr(X,'Type','Spearman')

r = tiedrank(X);
[rho pval]=corr(r,'Type','Pearson')
```

29

## 相关系数---肯德尔和谐系数

- 等级相关的一种
- 适用资料：
  - 适用于k个评价者，评价多个事物的等级变量资料
  - 多用于评分者信度分析

$$W = \frac{\sum R_i^2 - \frac{(\sum R_i)^2}{N}}{\frac{1}{12}K^2(N^3-N)}$$

30

例如，有 4 名评分者，对 6 份答卷进行评分，所评等级如下：

| 评分者 | 答卷编号 | | | | | |
|---|---|---|---|---|---|---|
| | 一 | 二 | 三 | 四 | 五 | 六 |
| 甲 | 4 | 3 | 1 | 2 | 5 | 6 |
| 乙 | 5 | 3 | 2 | 1 | 4 | 6 |
| 丙 | 4 | 1 | 2 | 3 | 5 | 6 |
| 丁 | 6 | 4 | 1 | 2 | 3 | 5 |
| $R_i$ | 19 | 11 | 6 | 8 | 17 | 23 |

求肯德尔和谐系数

31

例如，有 4 名评分者，对 6 份答卷进行评分，所评等级如下：

| 评分者 | 答卷编号 | | | | | |
|---|---|---|---|---|---|---|
| | 一 | 二 | 三 | 四 | 五 | 六 |
| 甲 | 4 | 3 | 1 | 2 | 5 | 6 |
| 乙 | 5 | 3 | 2 | 1 | 4 | 6 |
| 丙 | 4 | 1 | 2 | 3 | 5 | 6 |
| 丁 | 6 | 4 | 1 | 2 | 3 | 5 |
| $R_i$ | 19 | 11 | 6 | 8 | 17 | 23 |

可求得

$$\sum R_i = 19 + 11 + 6 + 8 + 17 + 23 = 84$$

$$\sum R_i^2 = 19^2 + 11^2 + 6^2 + 8^2 + 17^2 + 23^2 = 1400$$

$$S = 1400 - 84^2/6 = 224$$

$$W = \frac{224}{\frac{1}{12} \times 4^2 \times (6^3 - 6)} = 0.80$$

32

| | | | | | |
|---|---|---|---|---|---|
| 4 | 3 | 1 | 2 | 5 | 6 |
| 5 | 3 | 2 | 1 | 4 | 6 |
| 4 | 1 | 2 | 3 | 5 | 6 |
| 6 | 4 | 1 | 2 | 3 | 5 |

W = y_kendallW(X')

```
r = tiedrank(Data);
n = size(r,1);
m = size(r,2);
Ri = sum(r,2);
Rbar = mean(Ri,1);
S = squeeze(sum((Ri - repmat(Rbar,n,1)).^2,1));

W = 12*S/m^2/(n^3-n);
```

33

## 相关系数---点二列相关

- 适用资料：
  - 一列为正态等距变量
  - 另一列为二分命名变量
  - 常用于试卷的信度分析
- 公式：

$$r_{pb} = \frac{\bar{x}_p - \bar{x}_q}{S_t} \cdot \sqrt{pq}$$

34

被试 测验得分 性别

| | | |
|---|---|---|
| 1 | 20 | 1 |
| 2 | 19 | 1 |
| 3 | 17 | 1 |
| 4 | 8 | 0 |
| 5 | 9 | 0 |
| 6 | 5 | 0 |
| 7 | 18 | 1 |
| 8 | 16 | 1 |
| 9 | 15 | 1 |
| 10 | 14 | 1 |
| 11 | 8 | 0 |
| 12 | 9 | 0 |

问测验得分与性别是否相关？

（1为男，0为女）

35

被试 测验得分 性别

| | | |
|---|---|---|
| 1 | 20 | 1 |
| 2 | 19 | 1 |
| 3 | 17 | 1 |
| 4 | 8 | 0 |
| 5 | 9 | 0 |
| 6 | 5 | 0 |
| 7 | 18 | 1 |
| 8 | 16 | 1 |
| 9 | 15 | 1 |
| 10 | 14 | 1 |
| 11 | 8 | 0 |
| 12 | 9 | 0 |

问测验得分与性别是否相关？

（1为男，0为女）

```
St = 4.88
p = 7/12 = 0.583
q = 5/12 = 0.417
Xp = 17
Xq = 7.8
rpb = (17-7.8)
/4.88*sqrt(.417*.583)
    = 0.93
```

若作独立样本t检验

```
Sx1-x2=1.15
tobs=7.98
```

36

## $r_{pb}$与t之间的关系

- $r^2pb=t^2/(t^2+df)$，df=N1+N2-2

37

## 作业

1. 一个极端分数能对皮尔逊相关有很大的影响，尤其在小样本中

| X | Y |
|---|---|
| 1 | 4 |
| 2 | 1 |
| 3 | 2 |
| 4 | 3 |

a) 作散点图。计算皮尔逊相关。

b) 假定样本中加入一个新的个体X=10，Y=10
作散点图，计算新的皮尔逊相关。

38

38

## 作业

2. 如果只在X分数（Y分数）的部分范围中计算皮尔逊相关，得到的相关系数可能与总体相关系数完全不同。

| X | Y |
|---|---|
| 1 | 2 |
| 2 | 4 |
| 3 | 1 |
| 4 | 5 |
| 5 | 3 |
| 6 | 9 |
| 7 | 10 |
| 8 | 7 |
| 9 | 8 |
| 10 | 6 |

a. 作散点图。

b. 计算这个分布的皮尔逊相关。

c. 计算前5个分数的皮尔逊相关。

d. 计算后5个分数的皮尔逊相关。

e. 解释c）与d）的结果为何与总体相关差异很大。

39

39

## 作业

3. 为考察焦虑水平与考试成绩的关系，一位心理学家得到n=6的大学生样本。要求学生们提前15分钟到达实验室，测量了他们焦虑水平的心理指标，然后学生参加考试，焦虑水平与考试成绩如下表，计算二者的皮尔逊相关。

| 学生 | 焦虑水平 | 考试分数 |
|---|---|---|
| A | 5 | 80 |
| B | 2 | 88 |
| C | 7 | 80 |
| D | 7 | 79 |
| E | 4 | 86 |
| F | 5 | 85 |

40

40

## 作业

4.一位动物心理学家感兴趣动物的脑重和学习能力的关系，但他不知道此关系是否线性的。他选取n=10种动物得到数据如下。用适当的统计方法分析动物的脑重和学习能力的关系。

| 脑重 | 学习分数 |
|---|---|
| 1.04 | 1.5 |
| 2.75 | 1.8 |
| 4.14 | 1.9 |
| 7.81 | 1.6 |
| 8.11 | 2.1 |
| 8.35 | 4.5 |
| 8.50 | 4.2 |
| 8.73 | 6.2 |
| 8.81 | 10.3 |
| 8.97 | 14.7 |

41

41