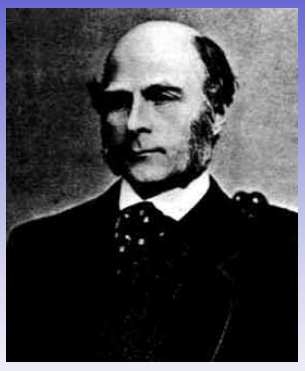
《医学统计学》第十一次课

线性相关分析

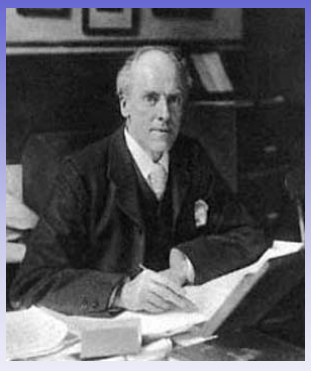
Department of Health Statistics

相关（correlation）的由来



Francis Galton

“遗传学研究”



Karl Pearson

子与父身高的关系——“回归”方程

兄弟与姐妹身高的关系？？

相关关系与确定性关系

- 确定性关系：两变量间的函数关系。

  圆的周长与半径的关系： $C=2\pi R$

  速度、时间与路程的关系：$L=ST$

  X与Y的函数关系： $Y=a+bX$

- 非确定性关系（相关关系）：两变量在宏观上存在关系，但无需或不能用确定的函数关系来表达。

  食盐摄入量与血压的关系；

  年龄与血脂的关系；

  吸烟引起的结核病与期望寿命的关系；

  肝脏中胆固醇含量与锰含量的关系；……

第一节

线性相关的概念

一、散点图

例13-1　为研究中年女性体重指数和收缩压之间的关系，随机测量了16名40岁以上女性的体重指数和收缩压，见表13-1，试绘制散点图。

表13-1 16名中年女性的体重指数（kg/m²）和收缩压（kPa）

编号	体重指数(X)	收缩压(Y)
（1）	（2）	（3）
1	28.6	18.00
2	34.1	18.93
3	36.2	20.00
4	32.0	17.60
…	…	…
15	33.3	19.87
16	37.6	21.07
合计	565.0	314.68

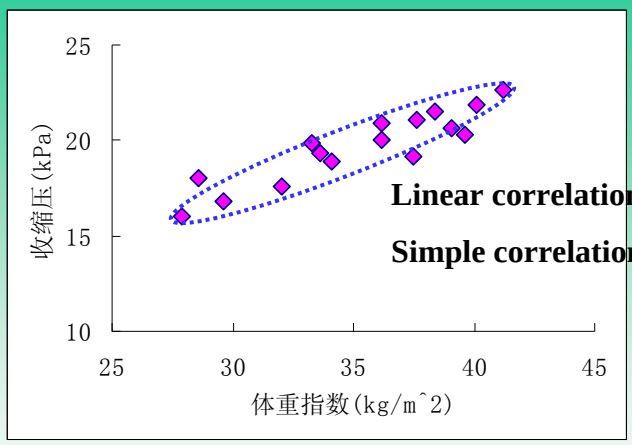


图13-1 16名中年女性体重指数和收缩压的散点图

二、线性相关

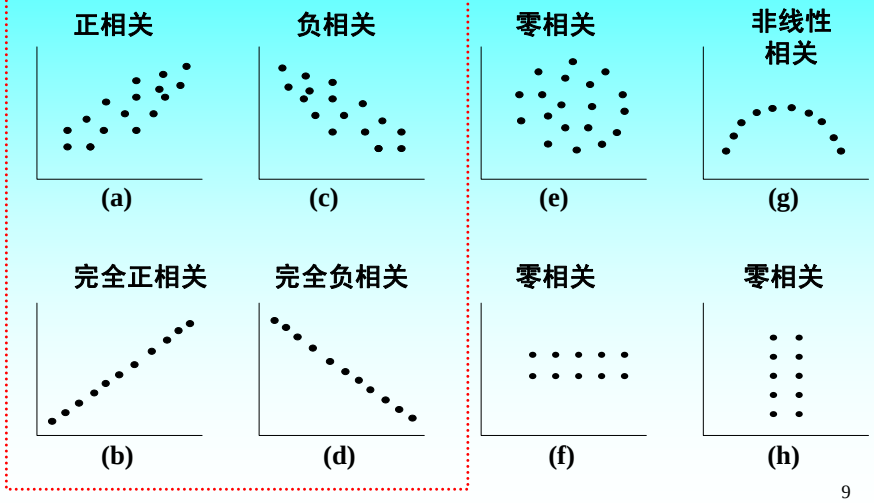
线性相关分析:

描述两变量间是否有直线关系以及直线关系的方向和密切程度的分析方法。

条件:

两变量（x, y）都是来自正态分布的随机变量。

相关关系示意图



第二节

线性相关系数

一、线性相关系数（linear correlation coefficient）：

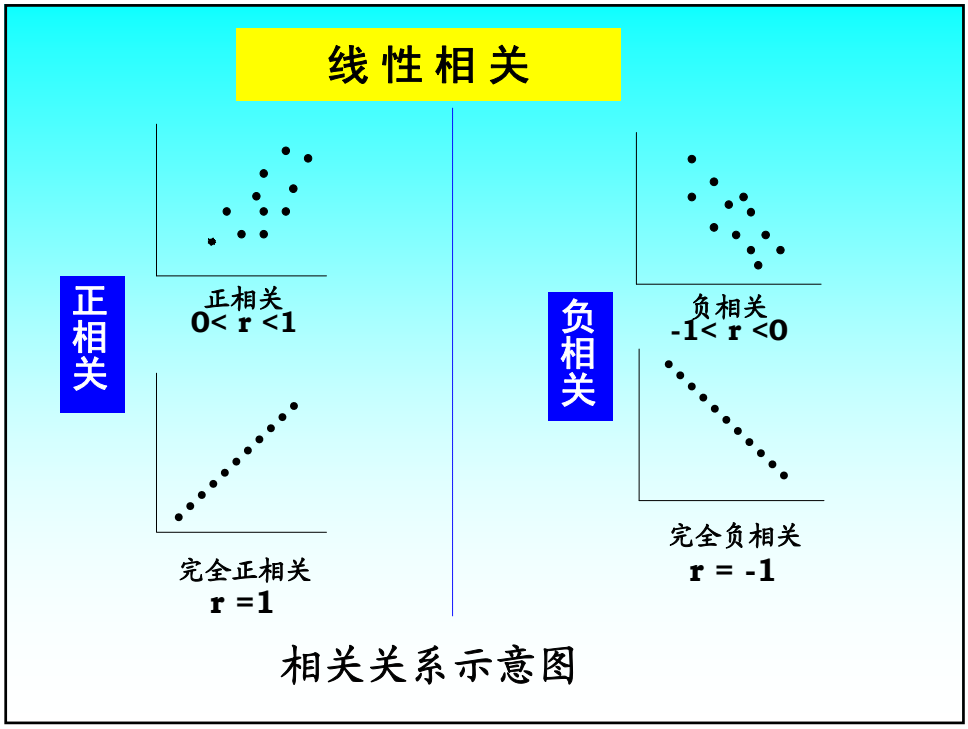
又称积差相关系数，简称相关系数或Pearson相关系数，用以描述两个随机变量间线性相关关系的密切程度与相关方向的统计指标。

符号： r　ρ

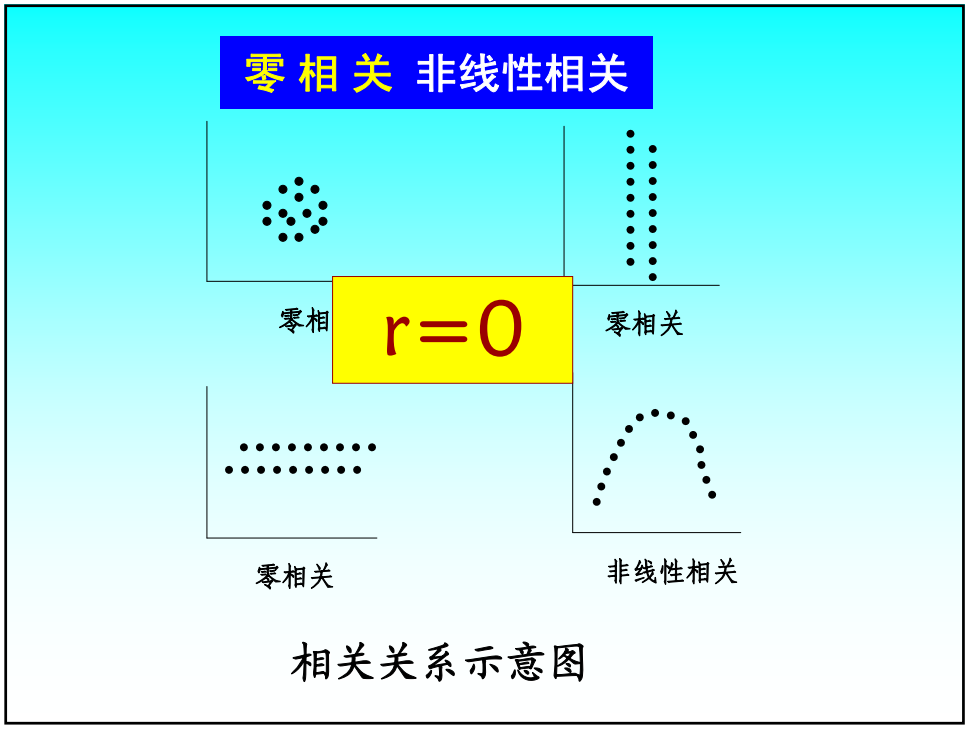
二、计算

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} = \frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}$$

$$l_{xy} = \sum (x-\bar{x})(y-\bar{y}) = \sum xy - \frac{\sum x \sum y}{n}$$



相关关系示意图



相关关系示意图

例13-2 （续前例13-1） 计算表13-1中体重指数和收缩压的相关系数。

1、绘制散点图：

2、计算：

$l_{xx}=263.50$

$l_{yy}=51.8001$

$l_{xy}=106.441$

$r = \ldots=0.9110$

第三节 相关系数的假设检验

总体相关系数的假设检验 H_0：$\rho=0$

t检验法 查表法

总体相关系数的假设检验　H_0：$\rho=0$

1、t检验法

$$t=\frac{r-0}{s_r}=\frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

$\nu=n-2$

例13-3　（续前例13-1）　根据体重指数和收缩压间样本相关系数r=0.91，对总体相关系数$\rho=0$进行假设检验。

（1）建立假设，确定检验水准

H_0：$\rho=0$ 即变量间不存在线性相关关系

H_1：$\rho\neq0$即变量间存在线性相关关系

$\alpha=0.05$

（2）计算检验统计量

r=0.91，n=16，代入公式 计算得

t=...=8.2653

（3）查t界值表，确定P值，下结论

根据ν=16-2=14查t界值表得P<0.05，按α=0.05的检验水准，拒绝H_0，接受H_1。可认为体重指数和收缩压之间存在正相关关系。

相关系数的假设检验　H_0：ρ=0

2、查表法

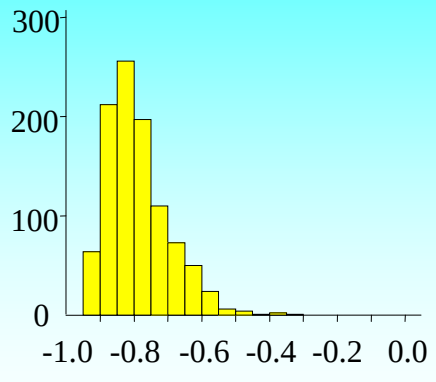
根据r值及ν=n-2，查附表13（相关系数r界值表）确定P值:

$r > r_{\alpha/2,\nu}$ => $P < \alpha$。

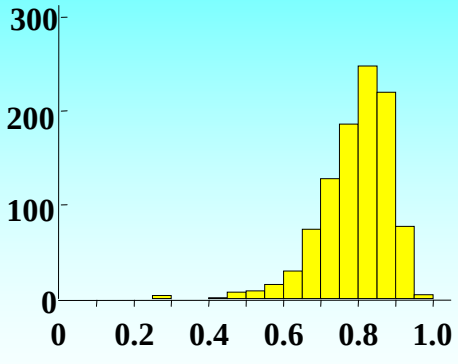
本例，r=0.91，ν=14查r界值表得$r_{0.05/2,14}$= 0.497 ,所以P < 0.05 。按α=0.05的检验水准，拒绝H_0，接受H_1，可认为体重指数和收缩压之间存在正相关关系。

第四节
相关系数的可信区间

相关系数的抽样分布($|\rho| = 0.8$)



$\rho = -0.8$



$\rho = 0.8$

R.A. Fisher(1921) 的 z 变换

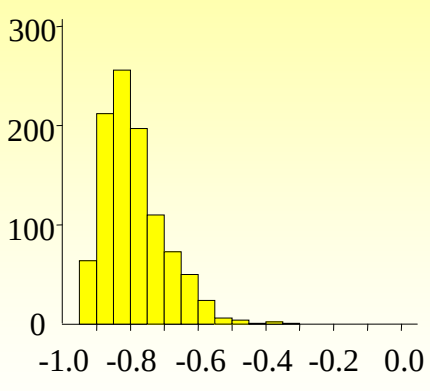
$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

$$r = \frac{e^{2z}-1}{e^{2z}+1}$$

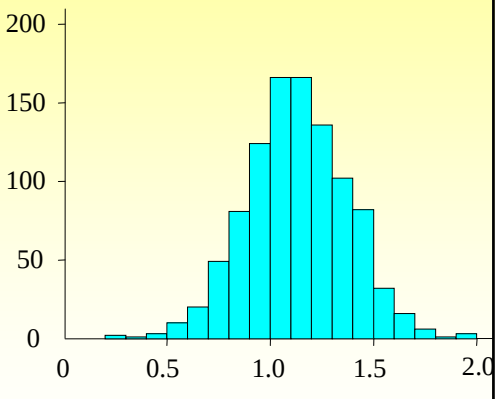
z 近似服从均数为 $\frac{1}{2}\ln[(1+r)/(1-r)]$，

标准差为 $1/\sqrt{n-3}$ 的正态分布。

相关系数的z变换值的抽样分布(ρ = - 0.8)

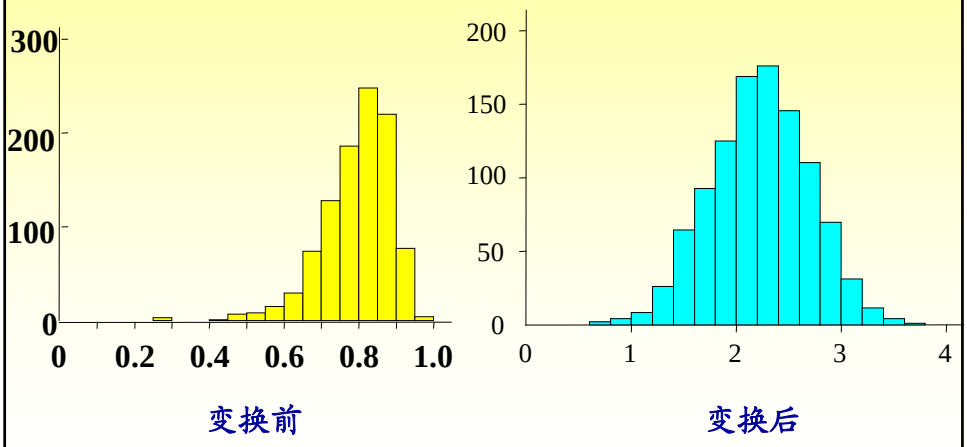


变换前



变换后

相关系数的z变换值的抽样分布(ρ=0.8)



变换前　　变换后

相关系数的可信区间估计

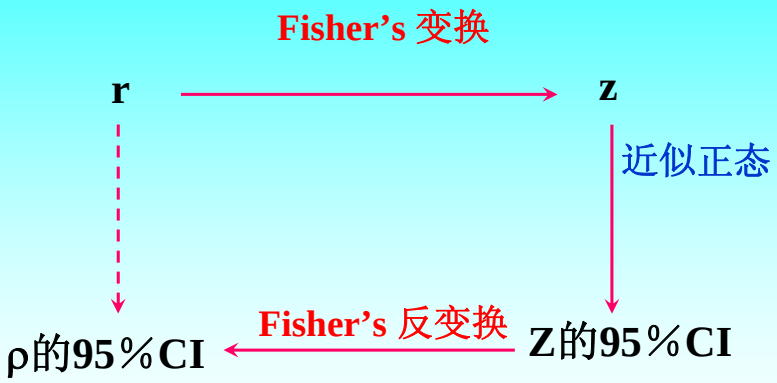
(1) 将 r 变换为 z；

(2) 根据 z 服从正态分布，估计 z 的可信区间：

$$z \pm u_{\alpha} s_z = z \pm u_{\alpha} \frac{1}{\sqrt{n-3}}$$

(3) 再将 z 变换回 r。

相关系数的可信区间估计



例13-4 （续前例13-1） 根据体重指数和收缩压间样本相关系数r=0.91，求总体相关系数ρ的95%可信区间。

r=0.91

$$Z=\tanh^{-1}r=\frac{1}{2}\ln\frac{1+r}{1-r}=1.5334$$ （Fisher's 变换）

$$Z \pm u_{\alpha/2} / \sqrt{n-3}$$

$$= 1.5334 \pm 1.96 / \sqrt{16-3}$$

$$= 0.9898 \sim 2.0770$$

$$r = \tanh Z = \frac{e^{2z}-1}{e^{2z}+1}$$ （Fisher's 反变换）

tanh0.9898 ~ tanh2.0770=0.76~0.97

结论：总体相关系数ρ的95%CI：0.76~0.97

第五节
相关系数应用的注意事项

相关系数应用的注意事项

1. 相关分析一定要有*实际意义*，相关关系*不一定是因果*关系；

案例一:

当样本足够大时，身高Y与家庭中的每月用电量X的线性相关性具有统计学意义（相关系数的假设检验P<0.05）。

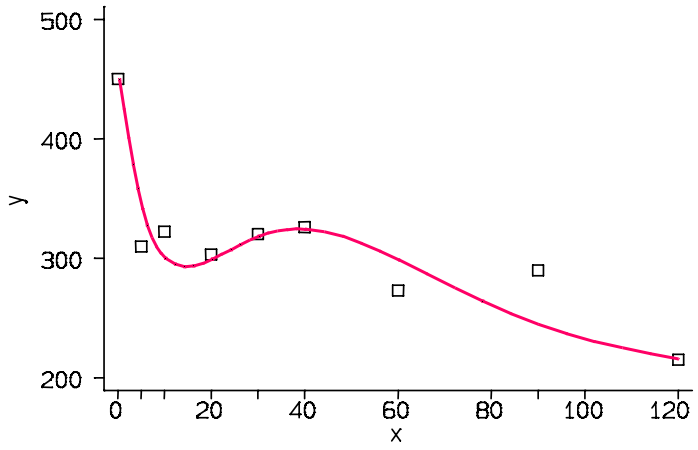
案例二:

一项研究抽取了5000个美国人组成样本，结果发现自我报告工作满意度与期望寿命之间r=0.7(P=0.01)。因此，“要想长寿，就应当喜欢你的工作。”

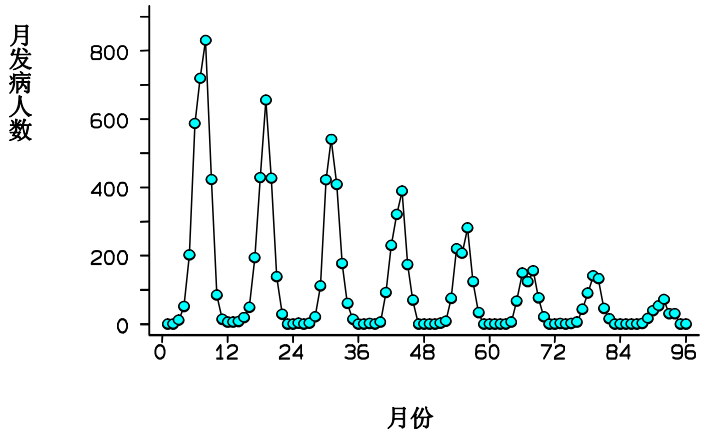
相关系数应用的注意事项

2. 进行线性相关分析前要先绘制*散点图*，从散点图的趋势判断是否可以作线性相关分析；

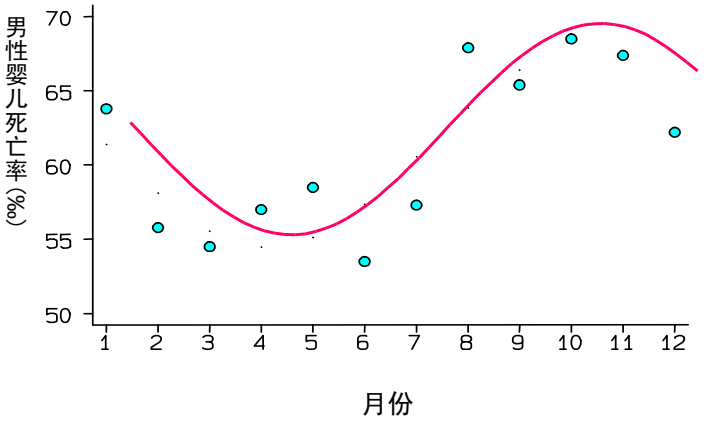
紫外光对新生小鼠背皮ATP酶阳性的郎格汉斯细胞(LC)照射不同时间的细胞密度(个/mm^3)



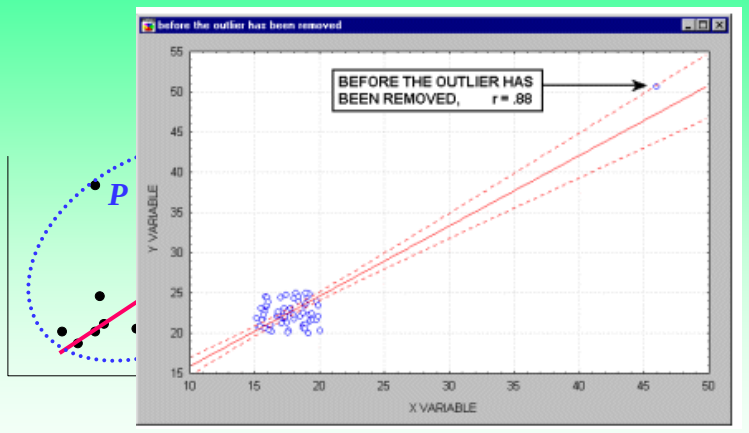
建湖县1978～1985年疟疾逐月发病数



我国1940～1988年间不同月份的男性婴儿死亡率(‰)的季节性分析



- 通过散点图识别离群值：



离群值对相关的影响

相关系数应用的注意事项

3. 必须对总体相关系数ρ进行假设检验；

当样本量较大（n>100），并对ρ进行假设检验，有统计学意义时：

|r|>0.7 ← 两变量高度相关；

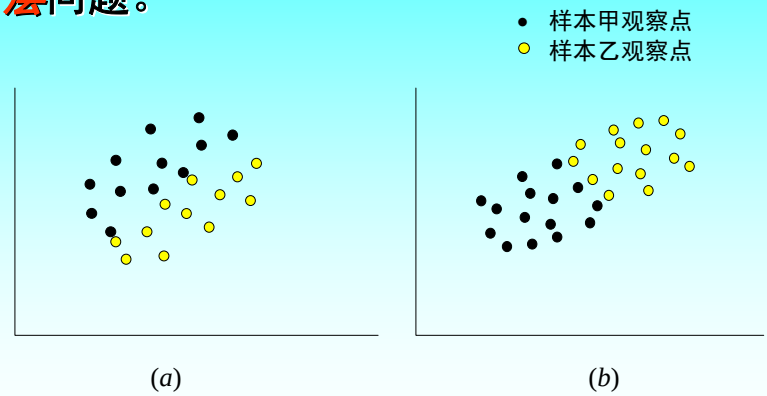
0.4<|r|≤0.7 ← 两变量中度相关；

0.2<|r|≤0.4 ← 两变量低度相关。

4. 同一个观察指标的两次重复测量结果间的相关系数表示测量结果的可靠性（P206 § 6）；

相关系数应用的注意事项

5. 应审慎对待相关分析的样本的*合并与分层*问题。



样本的合并可能对相关性造成的误导

线性相关与回归的区别

1.意义：

相关：两变量的相互关系，两个变量中，任何一个的变化都会引起另一个的变化，是一种双向变化的关系。

回归：两个变量的依存关系，一个变量的改变会引起另一个变量的变化，是一种单向变化的关系。

2.应用：

相关分析：研究两个变量的相互关系

回归分析：研究两个变量的依存关系

线性相关与回归的区别

3. 研究性质：

相关：对两个变量之间的关系进行描述，看两个变量是否有关，关系是否密切，是正相关还是负相关。

回归：对两个变量定量描述，研究两个变量的数量关系，已知一个变量值可以预测出另一个变量值，可以得到定量结果。

线性相关与回归的区别

4. 相关系数r与回归系数b的绝对值反映的意义不同:

r的绝对值越大，散点图中的点越趋向于一条直线，表明两变量的关系越密切，相关程度越高;
b的绝对值越大，回归直线越陡，说明当X变化一个单位时，Y的平均变化就越大。

线性相关与回归的联系

关系:
能进行回归分析的变量之间必然存在相关关系。所以，对于两个研究变量应先做散点图，求出它们的相关系数，对于确有相关关系的变量再进行回归分析，求出回归方程。

相关系数r与回归系数b :
1、符号一致;
2、假设检验结果一致 ，可用r的显著检验代替b的显著性检验。

第六节
Spearman秩相关
（P222）

总体Spearman秩相关系数：ρ_s，

样本Spearman秩相关系数：r_s。

反应两个变量X、Y之间的相关性而不依赖于X、Y的分布。

Spearman秩相关系数的计算：

- 将两变量X,Y分别编秩$P=R_X$, $Q=R_Y$;
- 计算P与Q的Pearson相关;
- 所得结果即为Spearman秩相关系数r_S。
- $-1 \leqslant r_s \leqslant 1$

如果n个X值互不相等（不同秩），n个Y值也互不相等，可利用简化公式：

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} \qquad (14-9)$$

d=P-Q

总体秩相关系数的假设检验 H_0：$\rho_s = 0$

1、查表法： 计算出r_s后，查附表14（r_s界值表），得到P值。**（注意与附表13不同，表14中首列为对子数n）**

2、t检验： 当n超出附表14范围时（n>50），用公式（14-10）检验总体秩相关系数。

$$t = \frac{r_s}{\sqrt{\frac{1-r_S^2}{n-2}}}, \quad \upsilon = n-2 \qquad (14-10)$$

例14-4 调查了某地区10个乡的钉螺密度与血吸虫感染率(%)数据如表14-7。试分析该地区钉螺密度与感染率之间有无相关关系？

表14-7 10个乡的钉螺密度与血吸虫感染率(%)

乡编号	钉螺密度,X	感染率,Y	X的秩,V	Y的秩,W	d
1	33	17	3	2	1.0
2	52	24	10	8.5	1.5
3	22	13	1	1	0.0
4	42	27	6	10	-4.0
5	35	19	4	5	-1.0
6	49	23	9	7	2.0
7	31	18	2	3.5	-1.5
8	39	18	5	3.5	1.5
9	45	24	8	8.5	-0.5
10	43	20	7	6	1.0
			l_{VV}=82.5	l_{WW}=81.5	Σd^2=30.0
				l_{VW}=67.0	

解：由于本例数据涉及感染率，而率一般不服从正态分布，故计算Spearman秩相关系数。本例数据中变量Y的值出现了同秩，因此，利用公式（14-8）得

$$r_s = \frac{l_{VW}}{\sqrt{l_{VV}l_{WW}}} = \frac{67.0}{\sqrt{82.5\times 81.5}} = 0.8171$$

对求得的Spearman秩相关系数进行检验：

1. 建立检验假设，确定检验水准。

H_0：$\rho_s = 0$（钉螺密度与血吸虫感染率无关）

H_1：$\rho_s \neq 0$（钉螺密度与血吸虫感染率有关）

$\alpha = 0.01$

2. 查表法

由附表14，$r_{10,0.01}=0.794<0.8171$, $P<0.01$。

3. t检验法

将$r_s = 0.8171$代入公式（14-10）得

$$t = \frac{r_s}{\sqrt{\frac{1-r_S^2}{n-2}}} = \frac{0.8171}{\sqrt{\frac{1-(0.8171)^2}{10-2}}} = 4.01$$

$$\upsilon = n-2 = 8$$

查附表2，得$t_{0.01/2,8}=3.355 < t = 4.01$, $P<0.01$。

4. 下结论：在$\alpha = 0.01$水准上，拒绝H_0，接受H_1，认为该地区钉螺密度与感染率之间有秩相关关系。

秩（等级）相关的含义

- 秩（等级）相关反映的是两变量等级间的相关，并不反映两变量间的数值关系。

例1		例2		例3		例4	
X	Y	X	Y	X	Y	X	Y
1	1	1	1	1	1	1	1
2	2	2	4	2	1.1	2	10
3	3	3	9	3	1.2	3	100
4	4	4	16	4	1.3	4	1000
5	5	5	25	5	1.4	5	10000

Chapter Summary

一、相关分析的用途与步骤，及其与线性回归分析的区别与联系；

二、相关系数r的计算和意义；

三、总体相关系数ρ的假设检验；

四、不满足正态分布的变量其相关性用Spearman秩相关系数衡量。

SAS程序 直线相关分析

例7.1 某地方病研究所调查了8名正常儿童的尿肌酐含量（mmol/24h）如表7.1，试分析尿肌酐含量（mmol/24h）与年龄（岁）之间的相关关系。

表7.1 8名正常儿童的年龄（岁）与尿肌酐含量（mmol/24h）

编号	1	2	3	4	5	6	7	8
年龄	13	11	9	6	8	10	12	7
尿肌酐含量	3.54	3.01	3.09	2.48	2.56	3.36	3.18	2.65

```
data prg7_1;
 input x y @@;
cards;
13 3.54 11 3.01 9 3.09 6 2.48 8 2.56 10 3.36 12 3.18 7 2.65
;
proc corr;
  var x y;
run;
```

The SAS System

The CORR Procedure

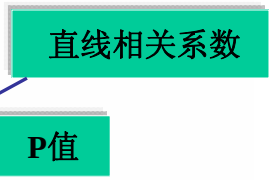
2 Variables: x y

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
x	8	9.50000	2.44949	76.00000	6.00000	13.00000
y	8	2.98375	0.38659	23.87000	2.48000	3.54000

Pearson Correlation Coefficients, N = 8
Prob > |r| under H0: Rho=0

	x	y
x	1.00000	0.88177 0.0038
y	0.88177 0.0038	1.00000



- 结论：

本例Pearson相关系数为r=0.88177，所对应的P=0.0038<0.05，说明两个变量之间存在正相关关系，即一个变量的值增大时，另一个变量的值也相应地增大。

SAS程序 秩相关

例7.4 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，结果见表7.3。以死因构成为x，WYPLL构成为y，试作秩相关分析。

表7.3 某省1995年到1999年居民死因构成与WYPLL构成

死因类别	各死因构成(%)	WYPLL构成	死因类别	各死因构成(%)	WYPLL构成
1	0.03	0.05	10	0.96	5.95
2	0.14	0.34	11	2.44	1.11
3	0.20	0.93	12	2.69	3.53
4	0.43	0.69	13	3.07	3.48
5	0.44	0.38	14	7.78	5.65
6	0.45	0.79	15	9.82	33.95
7	0.47	1.19	16	18.93	17.16
8	0.65	4.74	17	22.59	8.42
9	0.95	2.31	18	27.96	9.33

```
data prg7_4;
input x y @@;
cards;
0.03 0.05 0.14 0.34 0.20  0.93  0.43  0.69  0.44 0.38  0.45 0.79
0.47 1.19 0.65 4.74 0.95  2.31  0.96  5.95  2.44 1.11  2.69 3.53
3.07 3.48 7.78 5.65 9.82 33.95 18.93 17.16 22.59 8.42 27.96 9.33
;
proc corr spearman;
var x y;
run;
```

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
x	18	5.55556	8.66889	0.95500	0.03000	27.96000
y	18	5.55556	8.31028	2.89500	0.05000	33.95000

Spearman Correlation Coefficients, N = 18
Prob > |r| under H0: Rho=0

	x	y
x	1.00000	0.90506 <.0001
y	0.90506 <.0001	1.00000

spearman 相关系数

作　业

P212　三-2

P227　三-3


谢谢