第十章　直线回归与相关

张爱莲

新疆大学生命科学与技术学院　新疆生物资源基因工程重点实验室

zalxju@gmail. com

20101223

课前提问

1. 方差分析的用途
2. 方差分析的基本思路
3. 方差分析的出发点
4. 方差分析的步骤
5. 单因素方差分析中 SS_T， SS_A， SS_e 的含义及三者的关系。

本章内容

10.1 相关与回归的概念

10.2 直线回归分析

10.3 直线相关

10.1 相关与回归的概念

- 研究两个或两个以上变量间的关系有两类：

一类是变量间存在着完全确定性的一一对应关系，可以用精确的数学表达式来表示。（函数关系）

另一类是变量间不存在完全的确定性关系，不能用精确的数学公式来表示，统计学称之为相关关系，存在相关关系的变量称之为相关变量。

1. 相关

设有两个**随机变量X和Y**，对于任一随机变量的每一个可能值，另一个随机变量都有一个确定的分布与之相对应，则称这两个随机变量间存在相关关系。

2. 回归

如果对于**一般变量X**的每一个可能值，都有**随机变量Y的**一个确定分布相对应，则称随机变量Y对变量X存在回归关系，**X为自变量**，**Y为因变量**，回归是相关的特例。

通过回归方程，可以用X去预测Y。

一般不严格区分相关与回归，相关情况下，Y对X存在回归关系，同时X对Y也存在回归关系。

相关变量间的关系一般分为两种

- 因果关系

  一个变量受另一个或几个变量的影响.

- 平行关系

  几个变量互为因果或共同受到另外因素的影响.

- 回归分析

  研究呈因果关系的相关变量间的关系。

- 相关分析

  研究呈平行关系的相关变量之间的关系。

- “一因一果”，一元回归分析

  分为直线回归分析与曲线回归分析。

- “多因一果”，多元回归分析

  分为多元线性回归分析与多元非线性回归分析两种。

- 简单相关分析

  对两个变量间的直线关系进行相关分析，也叫直线相关分析。

- 复相关分析

  对多个变量进行相关分析时，研究一个变量与多个变量间的线性相关。

- 偏相关分析

  研究其余变量保持不变的情况下两个变量间的线性相关。

10.2 直线回归分析

10.2.1 直线回归方程的建立

10.2.2 回归方程计算

例如对小白鼠从出生后第六天起，每隔三天称一次体重，一直称到第18天，数据如表：

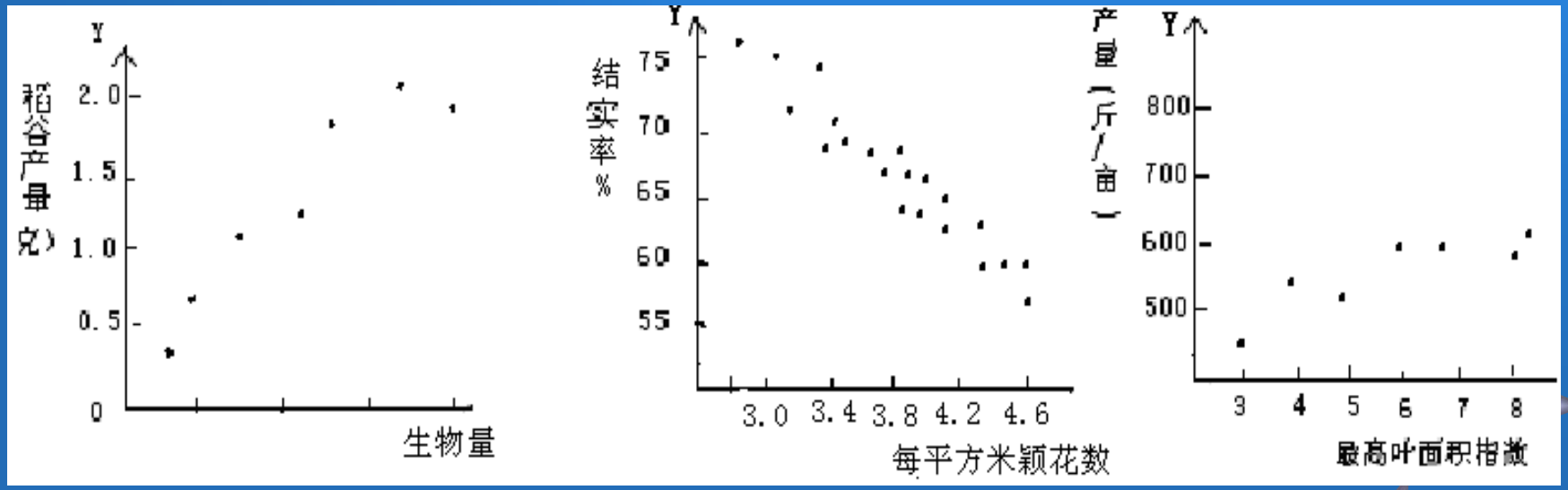
日龄（天）X	6	9	12	15	18
体重（克）Y	12	17	22	25	29

试估计出生后第10天的体重？第17天的体重？所估计体重的可靠程度是多少？

10.2.1 直线回归方程的建立

一、散点图:

- 通过实验获得两个变量的n对观测值：（x_1，y_1），（x_2，y_2），……，（x_n，y_n），为了直观地看出x和y间的变化趋势，将每一对观测值在平面直角坐标系描点，作出散点图。



- 从散点图可以看出：

  ①两个变量间有关或无关;若有关两个变量间的关系类型，是直线型还是曲线型;

  ②两个变量间直线关系的性质和相关程度;

- 散点图直观地、定性地表示了两个变量之间的关系。为了探讨它们之间的规律性，还必须根据观测值将其内在关系定量地表达出来。

根据回归的定义：

1. X 的每一个取值都有Y的一个正态分布与之对应。

2. 为了描述X与Y间的数量关系，必须找出一个能代表Y的值与 x_i 对应，这个代表值只能是当 $X = x_i$ 时，Y的平均数 $\mu_{Y.X=x_i}$ 。

3. $\mu_{Y.X=x_i}$ 称为Y的条件平均数

如何估计 $\mu_{Y.X=xi}$ 是直线回归所要解决的问题。

二、直线回归的数学模型

1. 借用 y=a+bx,

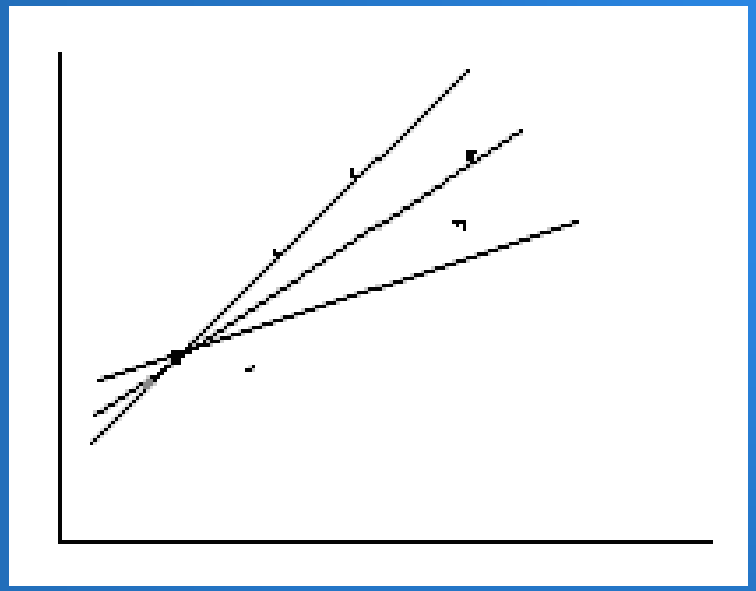
2. 得 $\mu_{Y.X} = \alpha + \beta x$,

表示 X 的每一个取值都有一个 Y 的分布与之对应，$\mu_{Y.X=xi}$

$$Y \sim N(\alpha + \beta x, \sigma^2)$$

总体回归线 $\mu_{Y.X} = \alpha + \beta x$

$\mu_{Y.x=xi}$、α、β 无法得到，需通过样本估计、推断。



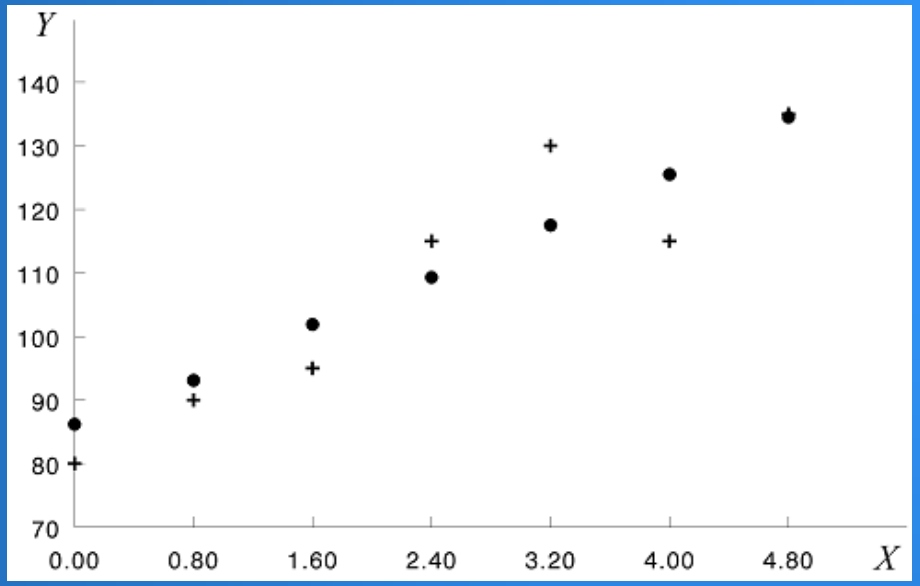
过两点可以做多条直线

$$\hat{Y} = a + bx_i$$

$$\Downarrow$$

$$\mu_{Y.X} = \alpha + \beta x$$

$\hat{Y}$ 叫做回归估计值，是当 x 在其研究范围内取某一个值时，Y 值平均数 $\alpha + \beta x$ 的估计值



多次重复的平均值所做的直线估计总体最理想

实际应用中并不设置重复，而是直接用n对观察值估计总体回归线。

$$\hat{Y} = a + bx_i$$

根据回归方程所画出的直线称为回归线，b是直线的斜率，称为回归系数。

怎样通过实际观测值得到总体回归α和β的最好点估计值a和b？

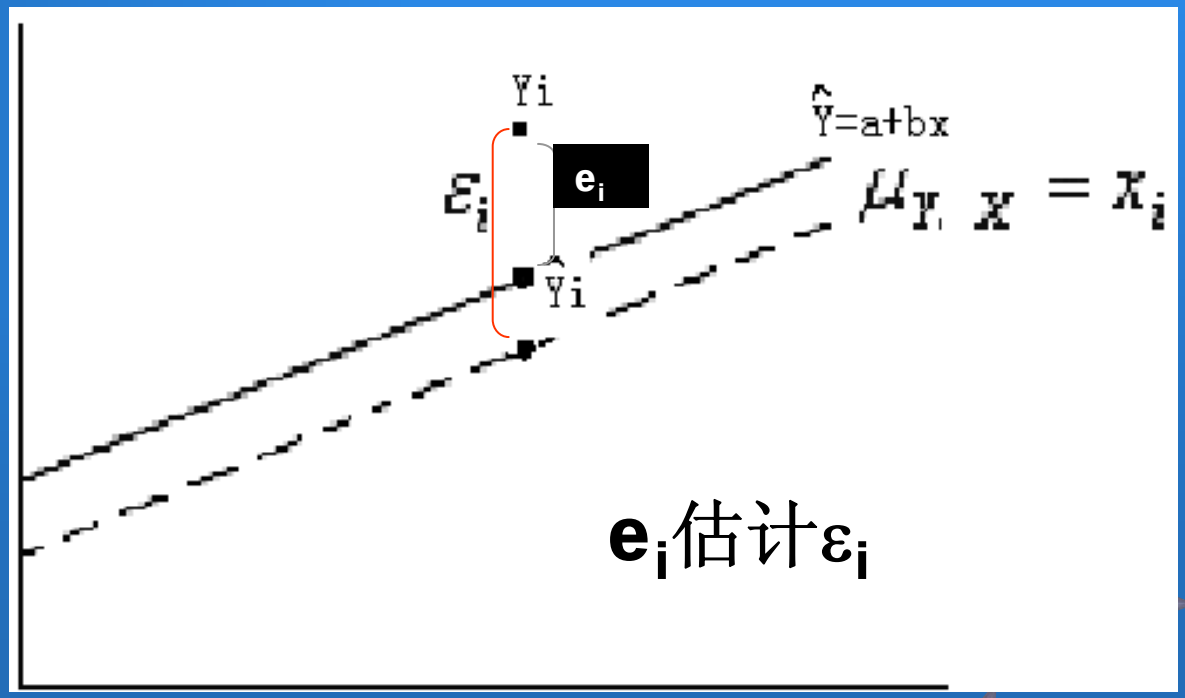
Question: 怎样确定$\hat{Y} = a + bx_i$ 这条直线？

n对观察值在x、y直角坐标平面上可以作出无数条直线，距离全部散点最近的直线，称为**回归直线**。

e_i最小的直线

$$\hat{Y} = a + bx_i$$

$$\mu_{Y.X} = \alpha + \beta x$$



a、***b***应使回归估计值 $\hat{y}$ 与实际观测值 **y** 的离差平方和最小，即： **ei最小的直线**

$$Q=\sum(y_i-\hat{y}_i)^2=\sum(y_i-a-bx_i)^2=\text{最小}$$

根据微积分学中的求极值的方法，令 **Q**对***a***、 ***b*** 的一阶偏导数等于**0**，即：

$$\frac{\partial Q}{\partial a}=-2\sum(y-a-bx)=0$$

$$\frac{\partial Q}{\partial b}=-2\sum(y-a-bx)x=0$$

这种使误差平方和最小的原则习惯上称为**最小二乘法**（**Least Square Method**）

整理得关于**a**、**b**的正规方程组：

$$\begin{cases} an + b\sum x = \sum y \\ a\sum x + b\sum x^2 = \sum xy \end{cases}$$

解正规方程组，得：

$$\begin{cases} b = \dfrac{\sum xy - (\sum x)(\sum y)/n}{\sum x^2 - (\sum x)^2/n} = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \dfrac{S_{xy}}{S_{xx}} \\ a = \bar{y} - b\bar{x} \end{cases}$$

$$b=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}=\frac{S_{xy}}{S_{xx}} \qquad a=\bar{y}-b\bar{x}$$

- $\sum(x-\bar{x})(y-\bar{y})$ 简称乘积和，记作S_{XY}
- $\sum(x-\bar{x})^2$ 是 x 的平方和 记作S_{XX}。
- a 叫做样本回归截距，当x=0时，$\hat{y}=a$；
- b 叫做样本回归系数，表示 x 改变一个单位，y 平均改变的数量；b 的符号反映了x影响y的性质，b的绝对值大小反映了 x 影响 y 的程度；

$$\hat{Y} = a + bX$$

$$a = \bar{y} - b\bar{x} \qquad b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

回归方程的基本性质：

性质**1** $Q = \sum (y-\hat{y})^2 =$ **最小；**

性质**2** $\sum (y-\hat{y}) = 0$ ；

性质**3** 回 归 直 线 通 过 点 $(\bar{x}, \bar{y})$ 。

10.2.2 回归方程计算

【例10.1】根据表10-1中的数据，计算干物重在Nacl含量上的回归方程。

	X	x'= x-2.4	x'^{2}	Y	Y'=Y-10	Y'^{2}	X'Y'
	0	-2.4	5.76	80	-30	900	72
	0.8	-1.6	2.56	90	-20	400	32
	1.6	-0.8	0.64	95	-15	225	12
	2.4	0	0	115	5	25	0
	3.2	0.8	0.64	130	20	400	16
	4.0	1.6	2.56	115	5	25	8
	4.8	2.4	5.76	135	25	625	60
和	16.8	0	17.92	760	-10	2600	200
平均	2.4						

1. 建立回归方程

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 17.92 - \frac{0}{7} = 17.92$$

$$S_{xy} = \sum_{1}^{n} xy - \frac{\sum x \sum y}{n} = 200 - \frac{0 \times (-10)}{7} = 200$$

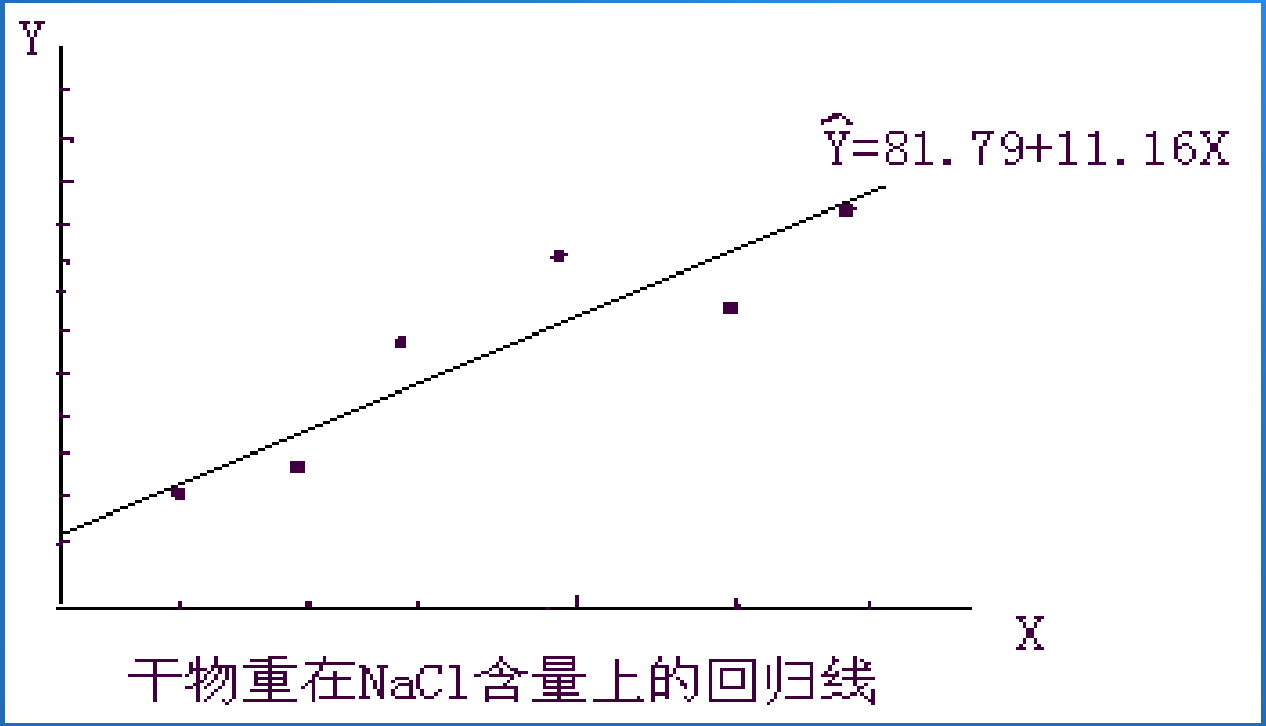
$$b = \frac{200}{17.92} = 11.16$$

$$a = \overline{y} - b\overline{x} = 108.57 - 11.16 \times 2.4 = 81.79$$

$$\hat{Y} = 81.79 + 11.16X$$

2.做散点图

将散点和回归直线做图，看出各个点子在回归直线两侧的分布情况。做直线时过点 $(\bar{x},\bar{y})$ 和（0，a）点即可。



$$\hat{Y} = 81.79 + 11.16X$$

b的生物学意义：当土壤Nacl含量增加1g时，植物单位叶面积干物重将增加11.16g。

a的生物学意义：当土壤Nacl含量为0时，植物单位叶面积干物重是81.79g。

这个回归方程**只是通过7对观察值对总体所做的最好估计**。

这种预测只适用于土壤含NaCl在（0，4.8）克的范围，如NaCl含量3g时，干重为115.27，不能任意外延。

【例10.2】一些夏季害虫盛发期的迟早和春季温度高低有关，江苏武进县测定了1956-1964年9年的3月下旬至4月中旬平均温度累积值（x，旬•度）和一代三化螟蛾盛发期（y，以5月10日为起算日0），试计算回归方程，预报害虫的盛发期。

年代	累积温X	盛发期Y	XY	X^2
1956	35.5	12	426	1260.25
1957	34.1	16	545.6	1162.81
1958	31.7	9	285.3	1004.89
1959	40.3	2	80.6	1624.09
1960	36.8	7	257.6	1354.24
1961	40.2	3	120.6	1616.04
1962	31.7	13	412.1	1004.89
1963	39.2	9	352.8	1536.64
1964	44.2	-1	-44.2	1953.64
和	333.7	70	2436.4	12517.49
平均	37.08	7.78		

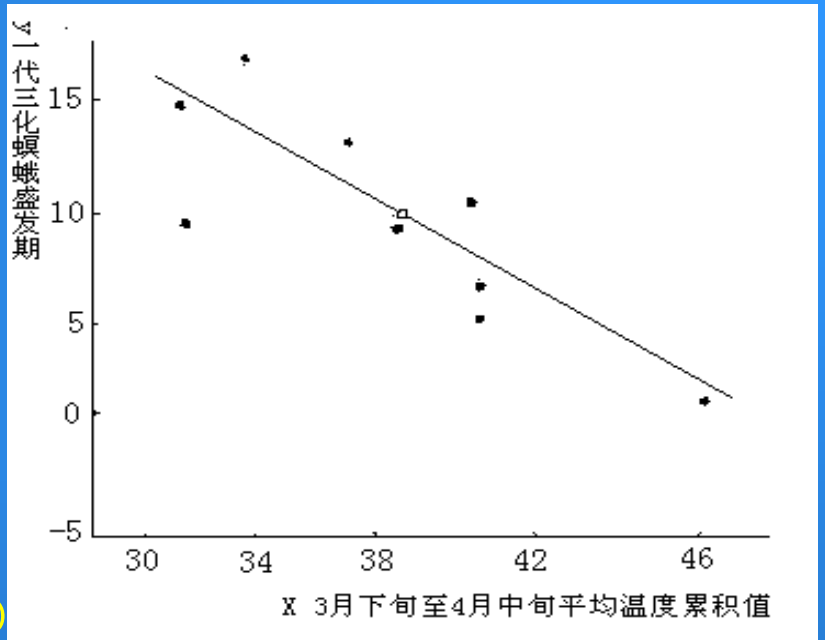
$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = -159.0444$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 144.6356$$

$$b = \frac{-159.0444}{144.6356} = -1.0996(\text{天}/\text{旬、度})$$

$$a = 7.78 - (-1.0996 \times 37.078) = 48.55(\text{天})$$

$$\hat{y} = 48.55 - 1.0996x = 48.5 - 1.1x$$



旬平均温度累积值和一代三化螟蛾盛发期的关系

回归方程的意义：当3月下旬至4月中旬的积温每提高1旬•度时，一代三化螟蛾的盛发期平均将提早1.1天，若积温为0，则一代三化螟蛾的盛发期在6月27-28日。

由于X的实测区间为[31.7，44.2]，在此区间外y的变化是否还符合上述回归方程无任何信息。

特别要指出的是：利用直线回归方程进行预测或控制时，一般 只适用于原来研究的范围，不能随意把范围扩大，因为在研究的范围内两变量是直线关系，这并不能保证在这研究范围之外仍然是直线关系。若需要扩大预测和控制范围，则要有充分的理论依据或进一步的实验依据。利用直线回归方程进行预测或控制， 一 般只能内插，不要轻易外延。

10.3 直线相关

直线相关分析：

根据 x、y 的观测值，计算表示两个相关变量x、y间线性相关程度和性质的统计量——相关系数 r

10.3.1 相关系数

- 相关系数

表示变量X和Y相关密切程度及其性质的量。

- 总体相关系数ρ：

设一个双变量总体X、Y，有N对（X，Y）

则总体相关系数ρ

$$\rho=\frac{1}{N}\sum_{1}^{N}\left[\frac{(X-\mu_x)}{\sigma_x}\cdot\frac{(Y-\mu_Y)}{\sigma_Y}\right]=\frac{\sum(X-\mu_x)(Y-\mu_Y)}{\sqrt{\sum(X-\mu_X)^2\sum(Y-\mu_Y)^2}}$$

样本相关系数 r：

$$r=\frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum\left(x_i-\bar{x}\right)^2\sum\left(y_i-\bar{y}\right)^2}}=\frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}}$$

10.3.2 相关系数的性质

相关系数r的性质：

（1）r=0时，点子随机无规律地分布，X与Y无相关

（2）r=1时，点子分布在一条直线上，相关程度最高

（3）r ≤1时，点子逐渐汇集到一条直线上

（4）r>0为正相关　r<0为负相关

5. r值很小，说明X与Y之间没有线性相关关系，但并不意味着X与Y之间没有其它关系，如很强的非线性关系。

6. 直线相关系数 r 一般只适用与测定变量间的线性相关关系，若要衡量非线性相关时，一般应采用相关指数r^2。为避免混淆，也记为R^2

$$R^2 = 1 - \frac{SS_{剩余}}{Syy}$$

10.3.3 相关系数的计算

外界气压的变化和对应的成年男子肺泡气体内CO_2分压的样本数据如下表，计算相关系数.

外界气压 10cmHg	肺泡气体内 CO_2分压mmHg				
X	Y	Y'=Y−40	X^2	Y'^2	XY'
5	39	−1	25	1	−5
6	39	−1	36	1	−6
8	40	0	64	0	0
11	38	−2	121	4	−22
13	40	0	169	0	0
16	41	1	256	1	16
18	39	−1	324	1	−18
23	39	−1	529	1	−23
28	42	2	984	4	56
和 128		−3	2308	13	−2

$$r=\frac{\sum XY-\frac{\sum X\sum Y}{n}}{\sqrt{\left[\sum x^2-\frac{\left(\sum x\right)^2}{n}\right]\left[\sum y^2-\frac{\left(\sum y\right)^2}{n}\right]}}=\frac{-2-\frac{128\times(-3)}{9}}{\sqrt{\left[2308-\frac{128^2}{9}\right]\left[13-\frac{-32}{9}\right]}}=0.531$$

即外界大气压与男子肺泡体内CO_2分压呈正相关，r=0.531，r可能是显著的，也可能是不显著，应进行显著性检验。

10.3.4相关系数与回归系数的关系

$$r = \sqrt{b_{yx} \cdot b_{xy}}$$

- 直线相关分析与回归分析的研究对象都是呈直线关系的相关变量。
- 直线回归分析将二个相关变量区分为自变量和依变量，侧重于寻求它们之间的联系形式——直线回归方程；
- 直线相关分析不区分自变量和依变量，侧重于揭示它们之间的联系程度和性质——计算出相关系数。

10.3.5 应用直线回归与相关的注意事项

直线回归分析与相关分析在生物科学研究领域中已得到了广泛的应用，但在实际工作中却很容易被误用或作出错误的解释。为了正确地应用直线回归分析和相关分析这一工具，必须注意以下几点：

1、变量间是否存在相关

直线回归分析和相关分析毕竟是处理变量间关系的数学方法，在将这些方法应用于生物科学研究时要考虑到生物本身的客观实际情况，譬如变量间是否存在直线相关以及在什么条件下会发生直线相关，求出的直线回归方程是否有意义，某性状作为自变量或依变量的确定等等，都必须由生物科学相应的专业知识来决定，并且还要用到生物科学实践中去检验。如果不以一定的生物科学依据为前提，把风马牛不相及的资料随意凑到一块作直线回归分析或相关分析，那将是根本性的错误。

2、其余变量尽量保持一致

由于自然界各种事物间的相互联系和相互制约，一个变量的变化通常会受到许多其它变量的影响，因此，在研究两个变量间关系时，要求其余变量应尽量保持在同一水平，否则，回归分析和相关分析可能会导致完全虚假的结果。例如研究人的身高和胸围之间的关系，如果体重固定，身高越高的人，胸围越小，但当体重在变化时，其结果也就会变化。

3、观测值要尽可能的多

在进行直线回归与相关分析时，两个变量成对观测值应尽可能多一些，这样可提高分析的精确性，一般至少有5对以上的观测值。同时变量x的取值范围要尽可能大一些，这样才容易发现两个变量间的变化关系。

4、外推要谨慎

直线回归与相关分析一般是在一定取值区间内对两个变量间的关系进行描述，超出这个区间，变量间关系类型可能会发生改变，所以回归预测必须限制在自变量x的取值区间以内，外推要谨慎，否则会得出错误的结果。

复 习

1. 相关与回归的概念
2. 回归方程的建立方法
3. 回归系数和回归截距的计算
4. 相关系数的计算