概率论与数理统计

第八章

应用回归分析

回归分析的研究对象

现实世界中变量之间的关系并不总是可以用函数关系(*自变量确定,因变量唯一*)来表示的
比如:

1) 家庭收入与家庭支出的关系
2) 父母身高与子/女身高的关系
3) 平时作业成绩与最后的考试成绩的关系
4) 银行利率与股票指数的关系

- 统计相关关系:

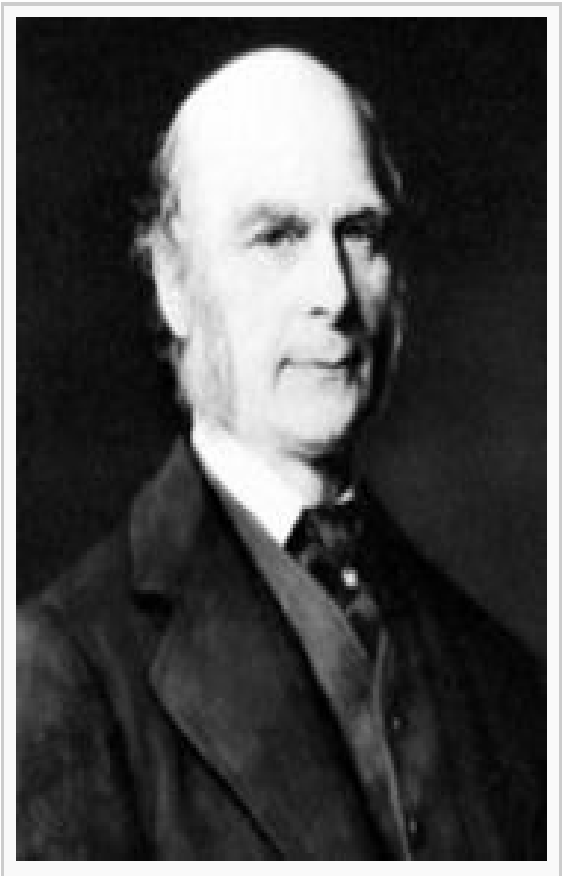
经验和统计数据表明某些变量的取值相互之间是有关系的(不是完全无关的),这种关系称为**统计相关关系**

- 回归分析及回归方程:

回归分析就是研究变量间的统计相关关系一种统计方法.

根据变元的统计数据,用一个函数来近似变元间的统计相关关系,这个函数叫**回归方程**或**回归函数**

法兰西斯·高尔顿（Francis Galton，英国,1822－1911）



“维多利亚女王时代最博学的人”

优生学家、人类学家、探险家、
地理学家、发明家、气象学家、
统计学家、心理学家、 遗传学家

“种族主义者和法西斯的鼻祖和精神领袖”

统计学上的贡献:

首次提出相关性概念

建立回归分析的方法

1886 年，高尔顿发表论文《遗传中向平均身高回归的现象》。高尔顿与皮尔逊合作，一起研究这个课题。他们收集了 1078 对父亲和儿子身高的数据：

父身高 子身高

(x_i, y_i)，$i = 1, 2, \cdots, 1078$

得到直线的方程为

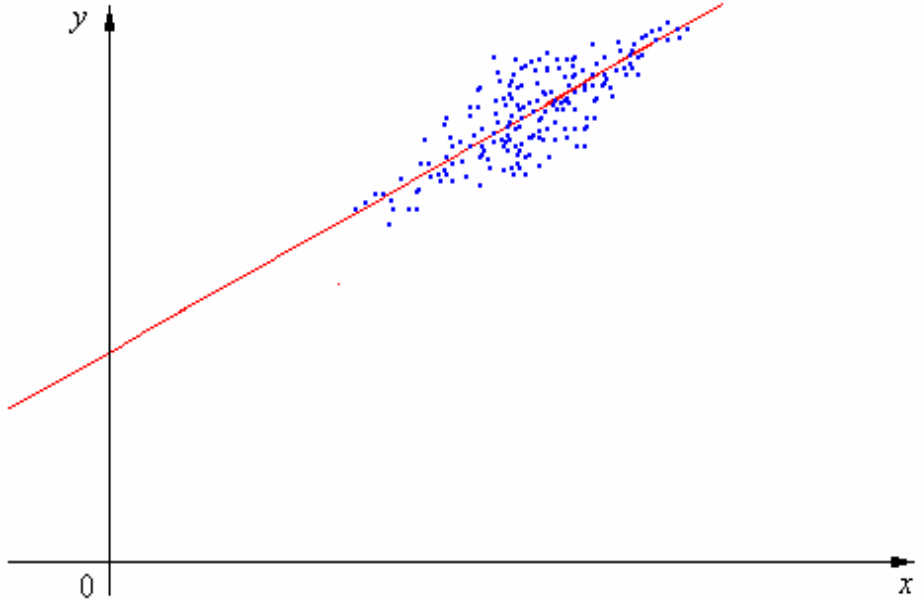
$\hat{y} = 0.8567 + 0.516x$ （单位：米）

例：$x = 1.900 \rightarrow \hat{y} = 1.837$

$x = 1.837 \rightarrow \hat{y} = 1.805$

$x = 1.600 \rightarrow \hat{y} = 1.682$

$x = 1.682 \rightarrow \hat{y} = 1.725$



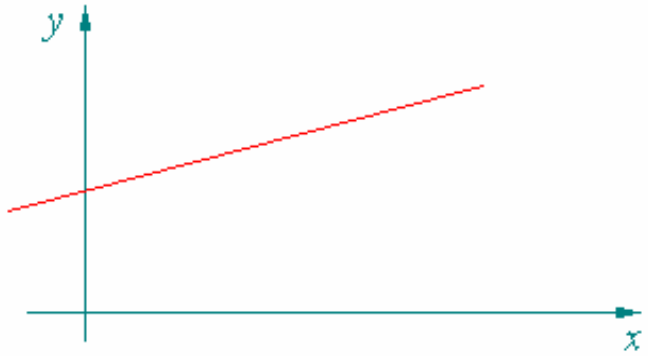
本例中，父亲身高与儿子身高的关系就是*统计相关关系*

上述高尔顿得到的近似直线方程就是*回归方程*

回归方程，可以是线性的，也可以是非线性的，当回归方程为线性时，称为**线性回归**（Linear Regression），当回归方程为非线性时，称为**非线性回归**（Nonlinear Regression）。

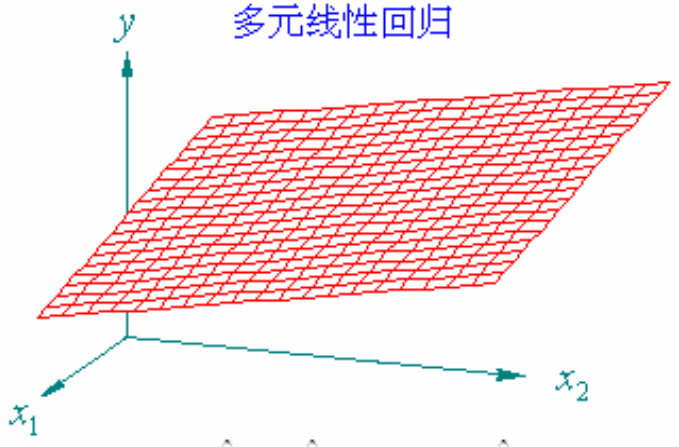
在回归方程中，可以只有一个自变量，也可以有多个自变量，只有一个自变量的回归称为**一元回归**（Simple Regression），有多个自变量的回归称为**多元回归**（Multiple Regression）。

一元线性回归



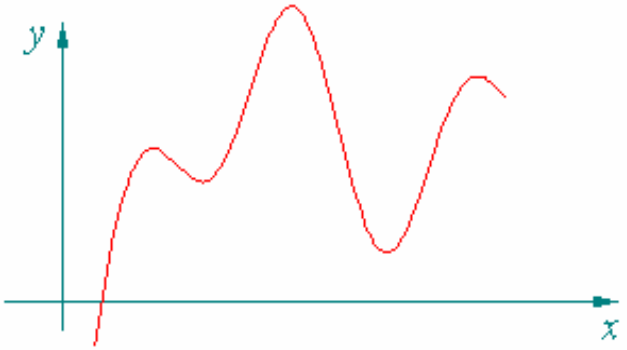
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

多元线性回归



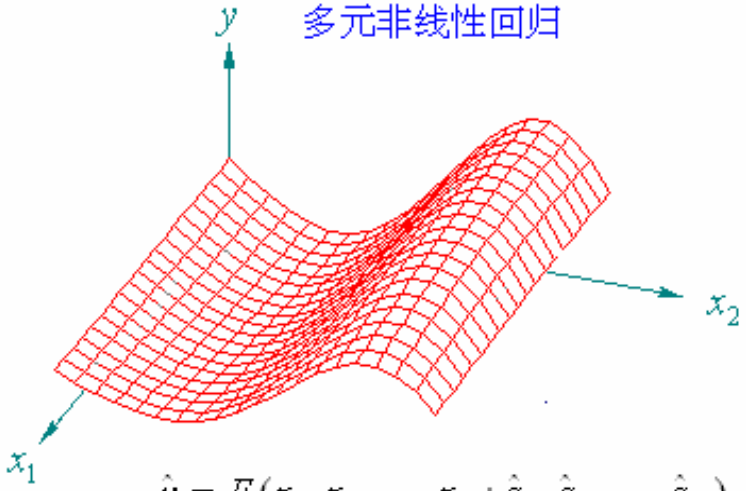
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m$$

一元非线性回归



$$\hat{y} = F(x; \hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p)$$

多元非线性回归



$$\hat{y} = F(x_1, x_2, \cdots, x_m; \hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p)$$

§8.1 一元线性回归

- 一元线性回归的模型:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

其中,X为确定性变量,它是可以测量和控制的,也称解释变量或自变量;

Y为被解释变量或响应变量

β_0和β_1为未知的待估计参数

ε为误差项,它表示X与Y间不能用线性关系解释的因素

根据变元（X，Y）的一组观测值 (x_i, y_i)，（i=1,2,...,n）代入上述一元线性回归模型，得：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (i = 1,2,...,n)$$

为了从上述这组等式中解出未知参数 β_0 和 β_1，及判断他们具有哪些性质，通常我们要求随机项 ε_i 满足下述三个前提条件：

1) 正态性： $\varepsilon_i \sim N(0,\sigma^2)$

2) 独立性：ε_i 相互独立

3) 方差齐性：ε_i 的方差相同与i无关

回到我们的一元线性回归模型：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (i = 1, 2, \ldots, n)$$

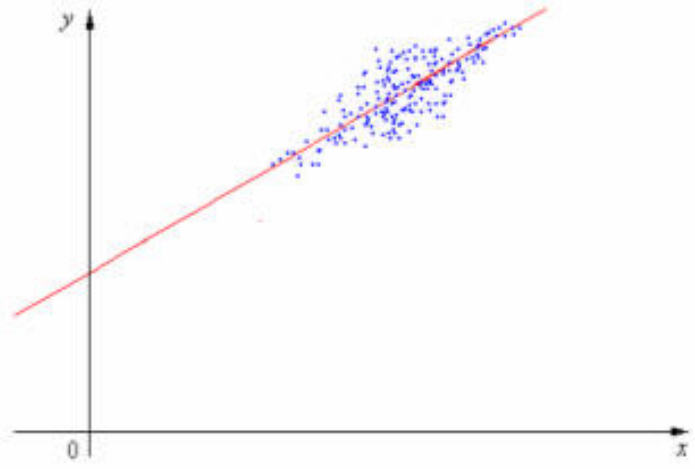
其中误差项满足：

1) 正态性： $\varepsilon_i \sim N(0, \sigma^2)$

2) 独立性：ε_i相互独立

3) 方差齐性：ε_i的方差相同与i无关

观测值（x_i, y_i）即散点图中的各个点，
如果没有随机误差项ε_i，这些点都将落
在直线（回归方程）上，因为ε_i的不同
取值，才导致了y_i可能偏离了回归直线。
因为ε_i是随机变量，因此
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ 也都是随机变量



由 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ （$\varepsilon_i \sim N(0, \sigma^2)$，易知：

$$y_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$$

于是，（y_1，y_2，...，y_n)的联合密度函数为：

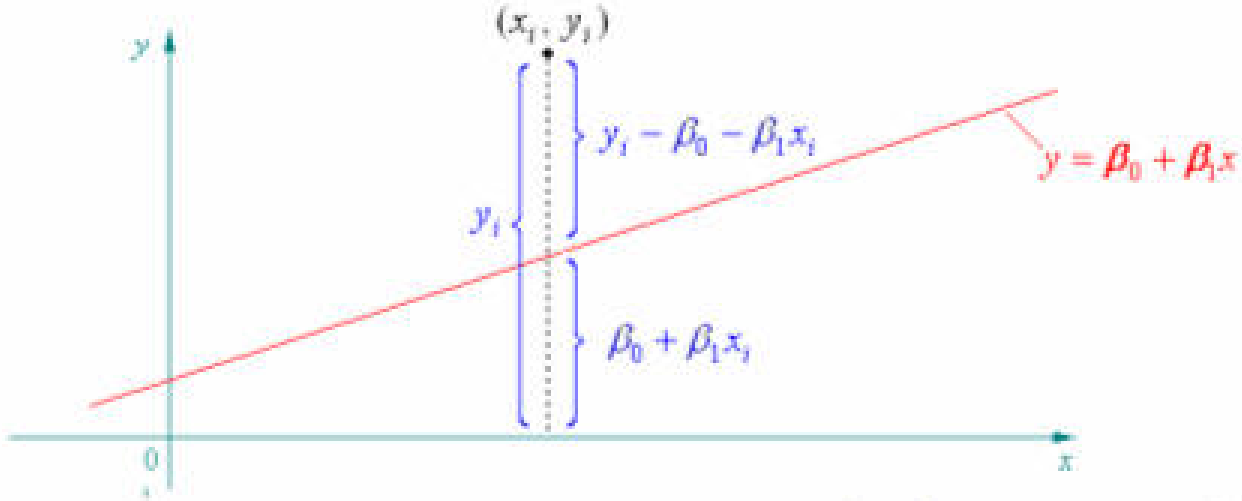
$$p(y_1,\ y_2,\ \ldots,\ y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}$$

其中，β_0，β_1，σ^2均为未知参数。根据极大似然估计的方法：

取似然函数 $L(\beta_0,\ \beta_1,\ \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}$

$$\text{由}\begin{cases}\dfrac{\partial \ln L}{\partial \beta_0} = 0\\ \dfrac{\partial \ln L}{\partial \beta_1} = 0\\ \dfrac{\partial \ln L}{\partial \sigma^2} = 0\end{cases} \Rightarrow \begin{cases}\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \dfrac{L_{xy}}{L_{xx}}\\ \hat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2\end{cases}$$

我们导出了参数的极大似然估计，但是，历史上高尔顿是用我们高等数学中所学过的最小二乘法导出的，因此，一般称之为最小二乘估计



问题 已知 (x_i, y_i)，$i = 1, 2, \cdots, n$，求常数 β_0, β_1 的估计 $\hat{\beta}_0, \hat{\beta}_1$，使得当 $\beta_0 = \hat{\beta}_0$，$\beta_1 = \hat{\beta}_1$ 时，$Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ 达到最小。

推导：
$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = 0 \end{cases} \Rightarrow \begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{L_{xy}}{L_{xx}} \end{cases}$$

如果我们把求出的参数$\hat{\beta}_0$，$\hat{\beta}_1$代入Q，得：

$$Q_{\min}=\sum_{i=1}^{n}[y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)]^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \triangleq SSE$$ ——称为残差平方和

显然，SSE越小，表示观测值距回归直线越近，特别地：

当 SSE=0 时，表示所有观测值的点都在回归直线上。

注意到我们已经证明：误差项$\varepsilon_i \sim N(0,\sigma^2)$中方差$\sigma^2$的极大似然估计为

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2=\frac{SSE}{n}$$

但这个估计不是无偏的，可以证明σ^2的无偏估计为$\frac{SSE}{n-2}$，

因此称 $\hat{\sigma}=\sqrt{\frac{SSE}{n-2}}$ 为一元回归的估计标准差

估计标准差越小，即SSE越小，它也表示回归效果越好

除残差平方和SSE，估计标准差$\hat{\sigma}$可以表示回归效果外，

我们还可以用样本相关系数来表示回归的效果

变元X与Y的相关系数的定义是：$$R=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

对比我们曾学过的随机变量X与Y的相关系数

$$r=\frac{\mathrm{cov}(X,Y)}{\sqrt{DX\cdot DY}}=\frac{E(X-EX)(Y-EY)}{\sqrt{DX\cdot DY}}$$

会发现他们形式上很象。事实上:

样本相关系数R 是把变元X与Y视为随机变量时X与Y的相关系数 r 的矩法估计;

样本相关系数R也是把(X， Y）视为服从二维正态分布时，其相关系数ρ的极大似然估计

变元X与Y的相关系数 $R=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$ 与随机变量的相关系数也有类似的性质：

1) $-1 \le R \le 1$

2) |R|越大，表示变元X与Y线性关系越强，反之，则表示线性关系越弱

3) R>0表示变元X与Y是正统计相关关系，即X越大则大体上Y也越大

R<0表示变元X与Y是负统计相关关系，即X越大而大体上Y会越小

如果记： $SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$ ---总离差平方和

$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ ---回归平方和

及前面讲到的：

$SSE=\sum_{i=1}^{n}(y_i-\hat{y})^2$ ---残差平方和

则可以证明：

$SST=SSR+SSE$ ---离差分解公式

离差分解公式　　ＳＳＴ＝ＳＳＲ＋ＳＳＥ

我们称回归平方和与总离差平方和的比值 $\frac{SSR}{SST}$ 为可决系数

或判定系数(coefficient of determination)，记为：$R^2=\frac{SSR}{SST}$

注：

1） 可以证明可决系数$\frac{SSR}{SST}$一定等于变元X与Y相关系数R 的平方，

因此，可记 $R^2=\frac{SSR}{SST}$ （证明略，提示利用正规方程）

2）离差分解公式中，SSR=$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$表示回归方程 $\hat{y}_i$的离差平方和（$\hat{y}_i$的均值等于$\bar{y}$），

SSE是由随机误差造成的，ε_i的方差σ^2越大则SSE会越大，（$\frac{SSE}{n-2}$是σ^2的无偏估计）

3）上述一元回归的离差分解公式，及可决系数的定义可直接推广到多元线性回归

例 1 测量上海市 1～3 岁男孩的平均体重，得到数据如下：

年龄 x_i（岁）	1.0	1.5	2.0	2.5	3.0
体重 y_i（kg）	9.75	10.81	12.07	12.88	13.74

设 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ， $\varepsilon_i \sim N(0, \sigma^2)$ ， $i = 1, 2, \cdots, 5$ ， $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_5$ 相互独立。

求： （1） β_0, β_1 的最小二乘估计 $\hat{\beta}_0, \hat{\beta}_1$ ；

（2） 残差平方和 SS_e ，估计的标准差 $\hat{\sigma}$ ，样本相关系数 r 。

解 $n = 5$ ， $\bar{x} = 2$ ， $L_{xx} = 2.5$ ， $\bar{y} = 11.85$ ， $L_{yy} = 10.173$ ， $L_{xy} = 123.525 - 5 \times 2 \times 11.85 = 5.025$ 。

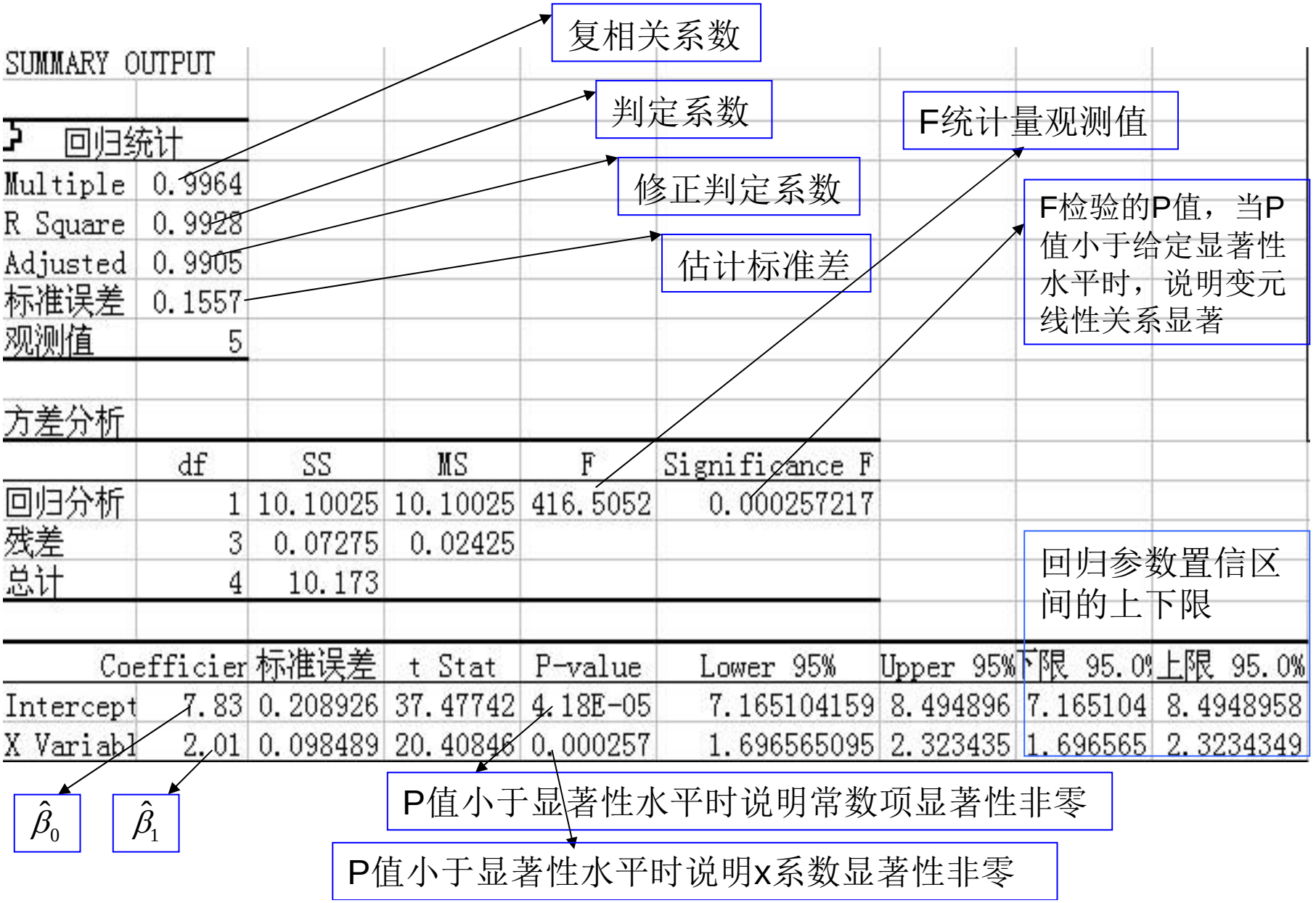
（1） $\hat{\beta}_1 = \dfrac{L_{xy}}{L_{xx}} = \dfrac{5.025}{2.5} = 2.01$ ， $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 11.85 - 2.01 \times 2 = 7.83$ 。

所以，回归方程为 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 7.83 + 2.01x$ 。

（2） $SS_e = L_{yy} - \hat{\beta}_1 L_{xy} = 10.173 - 2.01 \times 5.025 = 0.07275$ ， $\hat{\sigma} = \sqrt{\dfrac{SS_e}{n-2}} = \sqrt{\dfrac{0.07275}{5-2}} = 0.1557$ ，

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} = \frac{5.025}{\sqrt{2.5 \times 10.173}} = 0.9964 \text{ 。}$$

对例1中数据的EXCEL回归分析结果：



例2：恩格尔系数（食品支出与收入之比）的估算

已知人均月收入X与人均食品月支出Y的15组抽样数据如下，求恩格尔系数：

X	1020	960	970	1020	910	1580	540	830	1230	1060	1290	1380	810	920	640
Y	270	260	250	280	270	360	190	260	310	310	340	380	270	280	200

分析：根据给定数据，先找出X，Y的回归函数，再根据回归函数来估计恩格尔系数

解：利用EXCEL进行回归分析，得：

1020	270	SUMMARY OUTPUT							
960	260								
970	250	回归统计							
1020	280	Multiple	0.94145						
910	270	R Square	0.886328						
1580	360	Adjusted	0.877584						
540	190	标准误差	18.28581						
830	260	观测值	15						
1230	310								
1060	310	方差分析							
1290	340		df	SS	MS	F	ignificance F		
1380	380	回归分析	1	33893.18	33893.18	101.364	1.66314E-07		
810	270	残差	13	4346.82	334.3707				
920	280	总计	14	38240					
640	200								
			Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
		Intercept	99.87161	18.69586	5.341911	0.00013	59.48167559	140.2615	59.48168	140.2615
		X Variabl	0.180206	0.017899	10.06797	1.7E-07	0.141537857	0.218875	0.141538	0.218875

于是，得X，Y的回归方程为 $\hat{Y}=99.8716+0.1802X$

即： $\frac{\hat{Y}}{X}=\frac{99.8716}{X}+0.1802$

即恩格尔系数约为0.1802,且恩格尔系数会随收入的增大而变小

试卷长啥样?

难度有多大?

重点是什么?

做自测题, 你自己就能 猜 出来(这也是统计推断)

要做几套? 多多益善, 越做越顺手, 得分越高

(你最后自测的分数, 基本就是你的考试分数)

如下是习题册中的一套自测题

备用数据： $\Phi(\sqrt{2})=0.9225$，$\Phi(1.282)=0.9$，$\chi^2_{0.9}(16)=23.54$ $\chi^2_{0.9}(15)=22.31$，

$\chi^2_{0.025}(11)=3.816$，$\chi^2_{0.975}(11)=21.92$，$t_{0.995}(11)=3.1058$，$t_{0.995}(12)=3.0545$

一． 计算题

1.（11 分）设随机变量 ξ 与 η 相互独立，且分别服从区间[-1，1]上的均匀分布和参数 $\lambda=1$ 的指数分布，求（1） $Z=\xi+\eta$ 的密度函数；（2） ξ 和 z 的相关系数

解：（1） $p_z(z)=\int_{-\infty}^{+\infty}p_{\xi,\eta}(x,z-x)dx=\int_{-\infty}^{+\infty}p_\xi(x)p_\eta(z-x)dx$

$$=\begin{cases}0 & z<-1\\ \int_{-1}^{z}\frac{1}{2}e^{-(z-x)}dx & -1\le z<1\\ \int_{-1}^{1}\frac{1}{2}e^{-(z-x)}dx & 1\le z\end{cases}=\begin{cases}0 & z<-1\\ \frac{1}{2}(1-e^{-z-1}) & -1\le z<1\\ \frac{1}{2}e^{-z}(e-e^{-1}) & 1\le z\end{cases}\quad -- \ 6''$$

(2) $$\rho=\frac{\operatorname{cov}(\xi,Z)}{\sqrt{D\xi}\sqrt{DZ}}=\frac{\operatorname{cov}(\xi,\xi)+\operatorname{cov}(\xi,\eta)}{\sqrt{D\xi}\sqrt{D\xi+D\eta}}=\frac{D\xi+0}{\sqrt{D\xi}\sqrt{D\xi+D\eta}}$$

$$=\frac{4/12}{\sqrt{4/12}\sqrt{4/12+1}}=\frac{1}{4}$$

--5"

2. （10 分）据调查顾客在淘宝网上购买小件物品的消费额（单位：元）服从[50，150]上的均匀分布。而要想依靠在淘宝开一个网店来谋生，每月网店的销售额不能少于 3000 元。假设顾客各次光顾网店的消费额是相互独立的。试用中心极限定理估计一个网店每月要成功销售多少次物品才能以 90%的把握来保证当月的生计？

解：设需要卖出 n 次物品才能满足要求，而 ξ_i 为第 i 次卖出物品的销售额

$\xi_i \sim U(50,150)$， $E\xi_i=100$，$D\xi_i=10000/12$

n 次的累计销售额 $\xi=\xi_1+\xi_2+\ldots+\xi_n$

$E\xi=100n$，$D\xi=10000n/12$，由题意得： ---2”

$$0.9 \le P\{\xi \ge 3000\} = P\{\frac{\xi - E\xi}{\sqrt{D\xi}} \ge \frac{3000-100n}{\sqrt{10000n/12}}\} \approx 1-\Phi(\frac{3000-100n}{\sqrt{10000n/12}})$$

故 $0.1 \ge \Phi(\frac{3000-100n}{\sqrt{10000n/12}})$， ---4”

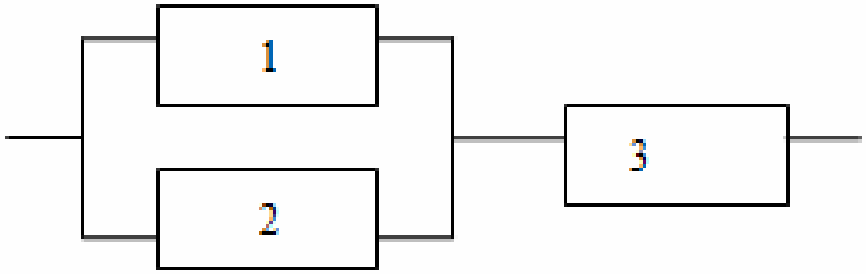
即 $\frac{3000-100n}{\sqrt{10000n/12}} \le -1.282$ ---2"

解得（可令 n=12x 来简化计算） $n \ge 32.097$，

答：每月需卖出 33 次物品才满足要求。 ---2"

[注：本题若由 $\frac{3000-100n}{\sqrt{10000n/12}} \ge 1.282$ 解得 n=28.04 为错，扣 4 分]

3.（8 分）如图：设系统中每个元件的工作相互独立，且他们损坏（不能正常工作）的概率均为 p



（1） 求整个系统的可靠性；（本小题 4 分）

（2） 若已知整个系统不能正常工作，求此时元件 3 损坏的概率（本小题 4 分）

解：（1） 设 A_i 为元件 i 损坏，i=1，2，3，则整个系统正常工作 C 的概率为：

$$P(C)=P\{\overline{A_1A_2\cup A_3}\}=1-P\{A_1A_2\cup A_3\}=1-(P\{A_1A_2\}+P\{A_3\}-P\{A_1A_2A_3\})$$

$$=1-p-p^2+p^3$$ ---4”

$$(2)\ P(A_3\mid\bar{C})=\frac{P(A_3\bar{C})}{P(\bar{C})}=\frac{P(A_3)}{1-P(C)}=\frac{p}{p+p^2-p^3}=\frac{1}{1+p-p^2}$$ -------- 4"

4．（10 分）已知（2，4，3，8，5，2，4）为取自总体 ξ 的一组样本观测值，ξ 的概率密度函数为

$$p_\xi(x)=\begin{cases} ba^b x^{-1-b} & a \le x \\ 0 & \text{其他} \end{cases} \qquad (a>0,\ b>0)$$

(1) 当已知 b=2 时，求参数 a 的极大似然估计；（本小题 5 分）
(2) 当已知 a=2 时，求参数 b 的矩法估计。（本小题 5 分）

解：(1) 参数a的似然函数为:

$$L(a)=\prod_{i=1}^{7} p(x_i)=\begin{cases} 2^7 a^{2\times 7} 2^{-3\times 2} 4^{-3\times 2} 3^{-3} 5^{-3} 8^{-3} & a \le 2 \\ 0 & \text{其他} \end{cases}$$

当 $a \le 2$ 时，由 $\frac{dL(a)}{da}=0$，解得 a=0，这与 a>0 的条件矛盾。

但由 L(a)表达式易见：a 越大 L(a)也越大（L(a)导数大于零），
故当 a 取其上限，即 a=2 时 L(a)取到极大值。
故 a 的极大似然估计值为 $\hat{a}=2$ ---5”

(2) 由于 $E\xi=\int_2^{+\infty} xb2^b x^{-1-b}dx = b2^b\int_2^{+\infty} x^{-b}dx = \frac{b2^b}{1-b}x^{1-b}\Big|_2^{+\infty} = \frac{2b}{b-1}$ （b>1）

因样本均值 $\overline{X}=(2+4+3+8+5+2+4)/7=4$

令 $E\xi=\overline{X}$ 得，b 的矩法估计值为 $\hat{b}=2$ ---5”

5．（11 分）已知随机变量 X，Y 相互独立，且 $P\{X+Y=0\}=1/24$，$P\{X+Y=2\}=3/8$，$EX=1/4$，求（X，Y）的联合分布和边际分布（注：只需完成下表的各空格，每空 1 分）

Y / X	-1	2	3	$P\{X=x_i\}$
0				
1				
$P\{Y=y_j\}$				1

解：

Y / X	-1	2	3	$P\{X=x_i\}$
0	1/8	3/8	1/4	3/4
1	1/24	1/8	1/12	1/4
$P\{Y=y_j\}$	1/6	1/2	1/3	1

6．（10 分）已知初生婴儿的体重服从正态分布$\xi \sim N(\mu, \sigma^2)$，随机抽取 12 名婴儿，测得体重（单位：克）如下：3100，2520，3000，3600，3000，2540，3160，3560，3320，2880，2600，3400

数据的描述性统计	
平均	3056.667
标准误	108.3438
中值	3050
模式	3000
标准差	375.314
样本方差	140860.6
峰值	-1.10406
偏斜度	-0.09444
区域	1080
最小值	2520
最大值	3600
求和	36680
计数	12

试问：（1）在显著性水平 $\alpha = 0.01$ 情况下，能否认为初生婴儿体重的均值为 3000 克？（5 分）

（2）在显著性水平 $\alpha = 0.05$ 情况下，能否认为初生婴儿体重的标准差为 350 克？（5 分）

解：（1）H_0：$\mu = 3000$；　H_1：$\mu \neq 3000$

$T = \frac{\overline{X} - \mu}{S_{n-1}}\sqrt{n} \sim t(n-1)$，代入观测值得：

$T = \frac{3056.667 - -3000}{375.314}\sqrt{12} = 0.523$

H_0的接受域为$W_0 = [-t_{1-\frac{\alpha}{2}}(n-1),\ t_{1-\frac{\alpha}{2}}(n-1)] = [-3.1058, 3.1058]$

$T \in W_0$，故接受H_0，即可以认为婴儿平均体重为 3000 克。

（2）H_0：$\sigma = 350$；　H_1：$\sigma \neq 350$

$\chi^2 = \frac{(n-1)\ S_{n-1}^2}{\sigma^2} \sim \chi^2(n-1)$，代入观测值得：

$\chi^2 = \frac{(12-1)140860.6}{350^2} = 12.6487$

H_0的接受域为$W_0 = [\chi^2_{\frac{\alpha}{2}}(n-1),\ \chi^2_{1-\frac{\alpha}{2}}(n-1)] = [3.816, 21.920]$

$\chi^2 \in W_0$，故接受H_0，即可以认为婴儿体重的标准差为 350 克。

二. 填空题（每孔 3 分，共 24 分）

1. 设二维随机变量 $(\xi, \eta) \sim N(1,1,1,1,0)$，则：

$P\{|\xi-\eta|>2\} = \underline{2\Phi(\sqrt{2})-1} = \underline{0.845}$ (可算到可查表为止)；

若用切比雪夫不等式来估计，则 $P\{|\xi-\eta|>2\} \le \underline{0.5}$

2. 我校理学院共有 20 个班级(各班人数均为 30 人)，从各班分别随机抽取 n_i 人参加体能测试 (其中 $n_i \le 30$，且 $n_1+n_2+\ldots+n_{20}=100$)，若每个人能否通过体能测试相互独立，通过测试的概率都是 0.9，则：

第 i 个班级选取的 n_i 人中，通过体能测试的人数 $X_i \sim \underline{B(n_i,0.9)}$；

所有选取的 100 人中，通过体能测试的人数 X 的数学期望 $EX=\underline{90}$.

3. 设随机变量 ξ 的概率密度为 $p(x)=ce^{-\frac{x^2-2x+1}{4}}$ $(-\infty<x<+\infty)$，则：

参数 $c = \underline{\dfrac{1}{2\sqrt{\pi}}}$；

3． 设随机变量ξ的概率密度为 $p(x)=ce^{-\frac{x^2-x+1}{4}}$ $(-\infty<x<+\infty)$, 则:

参数$c=$ $\frac{1}{2\sqrt{\pi}}$;

令ξ^*为ξ的标准化随机变量, 则 $P\{\xi^*\le 0\}=$ 0.5 .

4． 设总体$\xi\sim N(\mu, 3^2)$，$(X_1, X_2, \ldots, X_{16})$为取自总体的样本， 样本均值为 $\overline{X}$ ， 样本方差为 S_{n-1}^2 . （附 $\Phi(1.96)=0.975$ ， $\chi_{0.9}^2(16)=23.54$ ，$\chi_{0.9}^2(15)=22.31$)，则:

总体均值μ的置信水平为95%的置信区间的长度为 2.94 ;

$P\{\frac{15S_{n-1}^2}{3^2}\le$ 22.31 $\}=0.9$.

三. 选择题（每小题 4 分，共 16 分）

1．设 $\xi \sim B(100, 0.05)$，则下列 4 个选项错误的是（D）。

(A) ξ 近似服从泊松分布 P(5)

(B) ξ 近似服从正态分布 N(5, 4.75)

(C) 若 $\xi_i \sim B(1, 0.05)$ 且相互独立，则 ξ 可表示为 $\xi = \xi_1 + \xi_2 + \ldots + \xi_{100}$

(D) 对于 $0 \le k \le 100$，若 k 的取值越大，概率 $P\{\xi = k\}$ 就越小.

2．下列函数中，是分布函数的是（B）。

(A) $F(x) = \begin{cases} 1/10 & -\infty < x < -1 \\ 2/10 & -1 \le x < 0 \\ 3/10 & 0 \le x < 1 \\ 4/10 & 1 \le x < +\infty \end{cases}$

(B) $F(x) = \begin{cases} 0 & -\infty < x < -1 \\ 1/3 & -1 \le x < 0 \\ 2/3 & 0 \le x < 1 \\ 1 & 1 \le x < +\infty \end{cases}$

(C) $F(x) = \begin{cases} 0 & x \le 1 \\ & \\ 1 & 1 < x \end{cases}$

(D) $F(x) = \begin{cases} 0 & x < -1 \\ 1/2 & -1 \le x < 0 \\ 1/3 & 0 \le x < 1 \\ 1 & 1 \le x \end{cases}$

3．对于任意的总体ξ，若其期望和方差（μ和σ^2）都存在，则（ A ）。

(A) 样本均值$\overline{X}$一定是总体期望μ的无偏估计；

(B) 样本的中位数 Me 一定是总体期望μ的无偏估计；

(C) 样本的众数 Mod 一定是总体期望μ的无偏估计；

(D) 样本二阶原点矩$\overline{X^2}$一定是总体二阶原点矩μ^2的无偏估计；

4．设X_i相互独立，且服从参数为i的指数分布：$X_i \sim E(i)$ $i=1,2,...,10$，则有

$P\{\max(X_1, X_2, ..., X_{10}) \geq 1\}$等于（ C ）。

(A) $\prod_{i=1}^{10} ie^{-i}$　　(B) $\prod_{i=1}^{10}(1-e^{-i})$

(C) $1-\prod_{i=1}^{10}(1-e^{-i})$　　(D) $1-\prod_{i=1}^{10} ie^{-i}$

最后，谢谢大家一个学期的配合，并：

预祝你考出好成绩

By KpZhu