概率统计模块教学的一点总结

数学刘俊华　赵佐伦

一、A、B两版教材的比对

1、条件概率部分：基本一致.

2、随机变量及其分布列：教材内容顺序不同，A版后讲“二项分布与超几何分布”，其中在介绍准确.

问题 已知100件产品中有8件次品，分别采用有放回和不放回的方式随机抽取4件. 设抽取的4件产品中次品数为 X，求随机变量 X 的分布列.

我们知道，如果采用有放回抽样，则每次抽到次品的概率为0.08，且各次抽样的结果相互独立，此时 X 服从二项分布，即 $X\sim B(4,\ 0.08)$.

采用不放回抽样，虽然每次抽到次品的概率都是0.08，但每次抽取不是同一个试验，而且各次抽取的结果也不独立，不符合 n 重伯努利试验的特征，因此 X 不服从二项分布.

可以根据古典概型求 X 的分布列. 由题意可知，X 可能的取值为0，1，2，3，4. 从100件产品中任取4件，样本空间包含 C_{100}^4 个样本点，且每个样本点都是等可能发生的. 其中4件产品中恰有 k 件次品的结果数为 $C_8^kC_{92}^{4-k}$. 由古典概型的知识，得 X 的分布列为

$$P(X=k)=\frac{C_8^kC_{92}^{4-k}}{C_{100}^4},\ k=0,\ 1,\ 2,\ 3,\ 4.$$

计算结果数时，考虑抽取的次序和不考虑抽取的次序，对分布列的计算有影响吗？为什么？

例4 袋中有8个白球、2个黑球，从中随机地连续抽取3次，每次取1个球.

(1) 若每次抽取后都放回，设取到黑球的个数为 X，求 X 的分布列；

(2) 若每次抽取后都不放回，设取到黑球的个数为 Y，求 Y 的分布列.

解 (1) 若每次抽取后都放回，则每次抽到黑球的概率均为 $\frac{2}{8+2}=\frac{1}{5}$.

而3次取球可以看成3次独立重复试验，因此 $X\sim B(3,\ \frac{1}{5})$，所以

$$P(X=0)=C_3^0\times\left(\frac{1}{5}\right)^0\times\left(\frac{4}{5}\right)^3=\frac{64}{125},$$

$$P(X=1)=C_3^1\times\left(\frac{1}{5}\right)^1\times\left(\frac{4}{5}\right)^2=\frac{48}{125},$$

$$P(X=2)=C_3^2\times\left(\frac{1}{5}\right)^2\times\left(\frac{4}{5}\right)^1=\frac{12}{125},$$

$$P(X=3)=C_3^3\times\left(\frac{1}{5}\right)^3\times\left(\frac{4}{5}\right)^0=\frac{1}{125}.$$

因此 X 的分布列为

X	0	1	2	3
P	$\frac{64}{125}$	$\frac{48}{125}$	$\frac{12}{125}$	$\frac{1}{125}$

(2) 若每次抽取后都不放回，则随机抽取3次可看成随机抽取1次，但1次抽取了3个，因此黑球数 Y 服从参数为10，3，2的超几何分布，即

$Y\sim$ [7] ，

因此

$$P(Y=0)=\frac{C_2^0C_8^3}{C_{10}^3}=\frac{7}{15},$$

$$P(Y=1)=\frac{C_2^1C_8^2}{C_{10}^3}=\frac{7}{15},$$

$$P(Y=2)=\frac{C_2^2C_8^1}{C_{10}^3}=\frac{1}{15}.$$

3、正态分布：区别较大，A版延续老版教材用频率分布直方图作为引入，B版教材用的是二项分布，并且B版把标准正态分布作为知识点进行讲解。

事实上，很多服从二项分布的随机变量分布列的直观图都具有类似的特点. 例如，若 $X\sim B(50,\frac{1}{2})$，则其分布列可用图 4-2-13 (1) 表示；若 $X\sim B(100,\frac{1}{2})$，则其分布列可用图 4-2-13 (2) 表示.

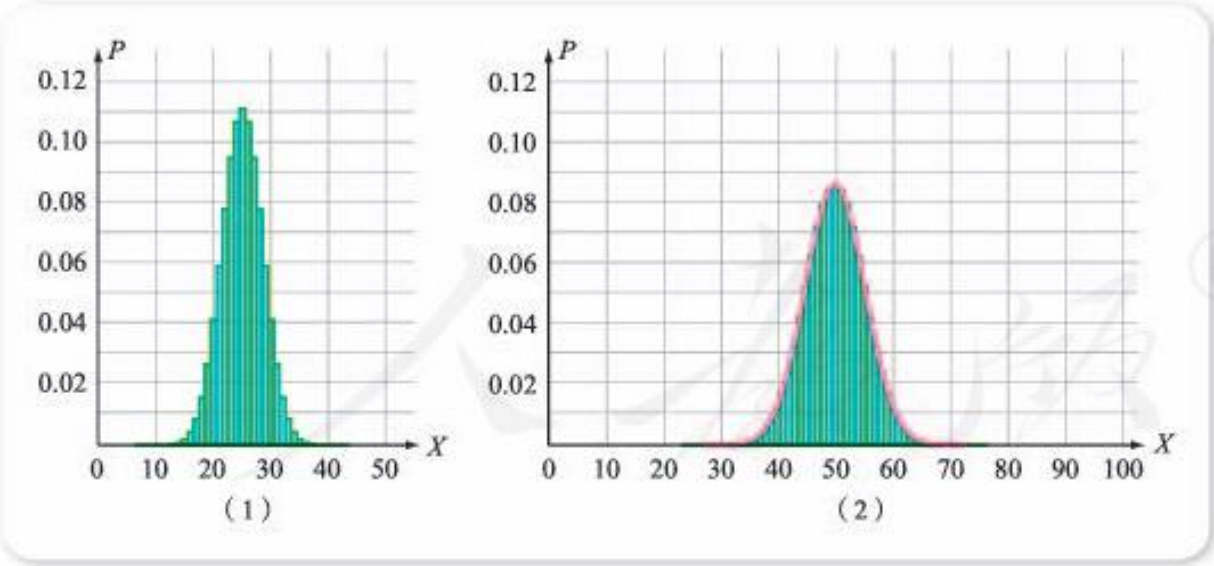


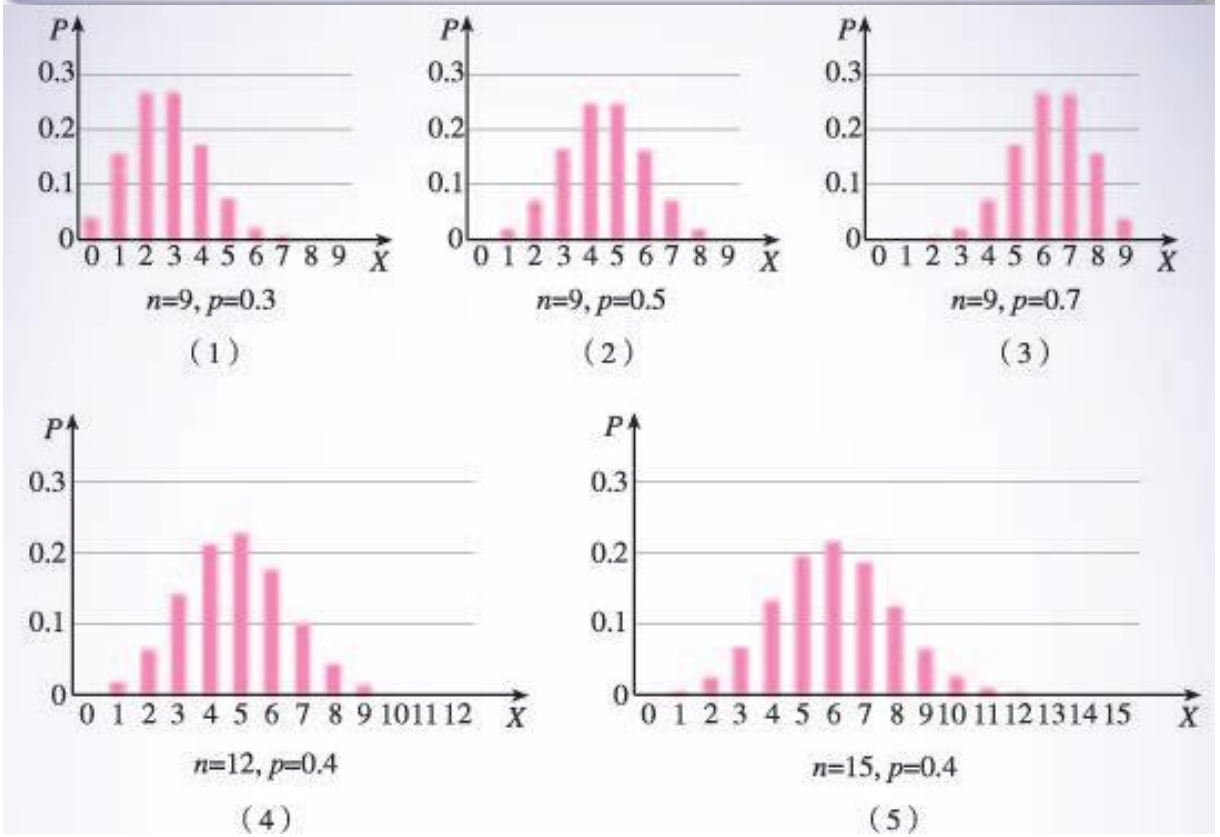
图 4-2-13

二项分布的性质

设随机变量 $X\sim B(n,p)$，则 X 的分布列为

$$P(X=k)=C_n^k p^k(1-p)^{n-k},\ k=0,1,\cdots,n.$$

对不同的 n 和 p 的值，绘制的概率分布图如图 1 所示.



现实中，除了前面已经研究过的离散型随机变量外，还有大量问题中的随机变量不是离散型的，它们的取值往往充满某个区间甚至整个实轴，但取一点的概率为 0，我们称这类随机变量为**连续型随机变量**（continuous random variable）. 下面我们看一个具体问题.

问题 自动流水线包装的食盐，每袋标准质量为 400 g. 由于各种不可控制的因素，任意抽取一袋食盐，它的质量与标准质量之间或多或少会存在一定的误差(实际质量减去标准质量). 用 X 表示这种误差，则 X 是一个连续型随机变量. 检测人员在一次产品检验中，随机抽取了 100 袋食盐，获得误差 X(单位：g)的观测值如下：

−0.6	−1.4	−0.7	3.3	−2.9	−5.2	1.4	0.1	4.4	0.9
−2.6	−3.4	−0.7	−3.2	−1.7	2.9	0.6	1.7	2.9	1.2
0.5	−3.7	2.7	1.1	−3.0	−2.6	−1.9	1.7	2.6	0.4
2.6	−2.0	−0.2	1.8	−0.7	−1.3	−0.5	−1.3	0.2	−2.1
2.4	−1.5	−0.4	3.8	−0.1	1.5	0.3	−1.8	0.0	2.5
3.5	−4.2	−1.0	−0.2	0.1	0.9	1.1	2.2	0.9	−0.6
−4.4	−1.1	3.9	−1.0	−0.6	1.7	0.3	−2.4	−0.1	−1.7
−0.5	−0.8	1.7	1.4	4.4	1.2	−1.8	−3.1	−2.1	−1.6
2.2	0.3	4.8	−0.8	−3.5	−2.7	3.8	1.4	−3.5	−0.9
−2.2	−0.7	−1.3	1.5	−1.5	−2.2	1.0	1.3	1.7	−0.9

(1) 如何描述这 100 个样本误差数据的分布？

(2) 如何构建适当的概率模型刻画误差 X 的分布？

根据已学的统计知识，可用频率分布直方图描述这组误差数据的分布，如图 7.5-1 所示. 频率分布直方图中每个小矩形的面积表示误差落在相应区间内的频率，所有小矩形的面积之和为 1.

观察图形可知：误差观测值有正有负，并大致对称地分布在 $X=0$ 的两侧，而且小误差比大误差出现得更频繁.

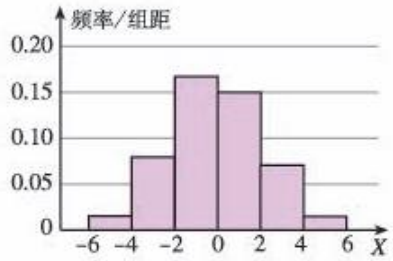


图 7.5-1

随着样本数据量越来越大，让分组越来越多，组距越来越小，由频率的稳定性可知，频率分布直方图的轮廓就越来越稳定，接近一条光滑的钟形曲线，如图 7.5-2 所示.

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x\in\mathbf{R}.$$

其中 $\mu\in\mathbf{R}$，$\sigma>0$ 为参数.

显然，对任意的 $x\in\mathbf{R}$，$f(x)>0$，它的图象在 x 轴的上方. 可以证明 x 轴和曲线之间的区域的面积为 1. 我们称 $f(x)$ 为**正态密度函数**，称它的图象为**正态密度曲线**，简称**正态曲线**，如图 7.5-4 所示.若随机变量 X 的概率分布密度函数为 $f(x)$，则称随机变量 X 服从**正态分布**(normal distribution)，记为 $X\sim N(\mu,\sigma^2)$. 特别地，当 $\mu=0$，$\sigma=1$ 时，称随机变量 X 服从**标准正态分布**.

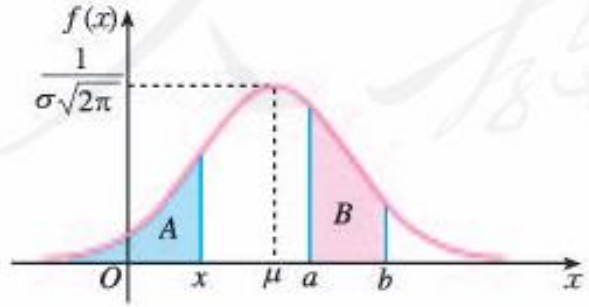


图 7.5-4

若 $X\sim N(\mu,\sigma^2)$，则如图 7.5-4 所示，X 取值不超过 x 的概率 $P(X\leqslant x)$ 为图中区域 A 的面积，而 $P(a\leqslant X\leqslant b)$ 为区域 B 的面积.

4、一元线性回归模型：A版内容比B版更丰富。

（1）A版把成对数据的相关性单列一节，介绍了成对数据的相关性和相关系数，第二节讲解一元线性回归模型。

（2）A版重视讲解公式的由来及其推导，比如相关系数r、最小二乘法、决定系数R^2。

（3）A版把残差作为一个重点内容，涉及残差分析和残差图，而B版只是提到了残差及其在最小二乘法中的作用。

为了消除度量单位的影响，需要对数据作进一步的“标准化”处理. 我们用

$$s_x=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2},\ s_y=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

分别除 $x_i-\overline{x}$ 和 $y_i-\overline{y}(i=1, 2, \cdots, n)$，得

$$\left(\frac{x_1-\overline{x}}{s_x}, \frac{y_1-\overline{y}}{s_y}\right), \left(\frac{x_2-\overline{x}}{s_x}, \frac{y_2-\overline{y}}{s_y}\right), \cdots, \left(\frac{x_n-\overline{x}}{s_x}, \frac{y_n-\overline{y}}{s_y}\right).$$

为简单起见，把上述“标准化”处理后的成对数据分别记为

$$(x_1', y_1'), (x_2', y_2'), \cdots, (x_n', y_n'),$$

仿照 L_{xy} 的构造，可以得到

$$r=\frac{1}{n}(x_1'y_1'+x_2'y_2'+\cdots+x_n'y_n')$$

$$=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}. \quad (1)$$

我们称 r 为变量 x 和变量 y 的**样本相关系数**（sample correlation coefficient）.

观察 r 的结构，联想到二维（平面）向量、三维（空间）向量数量积的坐标表示，我们将向量的维数推广到 n 维，n 维向量 $\boldsymbol{a}$，$\boldsymbol{b}$ 的数量积仍然定义为

$$\boldsymbol{a}\cdot\boldsymbol{b}=|\boldsymbol{a}||\boldsymbol{b}|\cos\theta,$$

其中 θ 为向量 $\boldsymbol{a}$，$\boldsymbol{b}$ 的夹角. 类似于平面或空间向量的坐标表示，对于向量 $\boldsymbol{a}=(a_1, a_2, \cdots, a_n)$ 和 $\boldsymbol{b}=(b_1, b_2, \cdots, b_n)$，我们有

$$\boldsymbol{a}\cdot\boldsymbol{b}=a_1b_1+a_2b_2+\cdots+a_nb_n.$$

因为 $|\boldsymbol{x}'|=|\boldsymbol{y}'|=\sqrt{n}$，所以样本相关系数

$$r=\cos\theta,$$

其中 θ 为向量 $\boldsymbol{x}'$ 和向量 $\boldsymbol{y}'$ 的夹角.

由 $-1\leqslant\cos\theta\leqslant1$，可知

$$-1\leqslant r\leqslant1.$$

用 x 表示父亲身高，Y 表示儿子身高，e 表示随机误差. 假定随机误差 e 的均值为 0，方差为与父亲身高无关的定值 σ^2，则它们之间的关系可以表示为

$$\begin{cases} Y=bx+a+e, \\ E(e)=0,\ D(e)=\sigma^2. \end{cases} \tag{1}$$

为什么假设 $E(e)=0$，而不假设其为某个不为 0 的常数？

我们称（1）式为 Y 关于 x 的**一元线性回归模型**（simple linear regression model）. 其中，Y 称为**因变量**或**响应变量**，x 称为**自变量**或**解释变量**；a 和 b 为模型的未知参数，a 称为截距参数，b 称为斜率参数；e 是 Y 与 $bx+a$ 之间的随机误差. 模型中的 Y 也是随机变量，其值虽然不能由变量 x 的值确定，但是却能表示为 $bx+a$ 与 e 的和（叠加），前一部分由 x 所确定，后一部分是随机的. 如果 $e=0$，那么 Y 与 x 之间的关系就可用一元线性函数模型来描述.

一元线性回归模型的两个变量的 n 对样本数据为 (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n)，由 $y_i=bx_i+a+e_i(i=1, 2, \cdots, n)$，得 $|y_i-(bx_i+a)|=|e_i|$. 显然 $|e_i|$ 越小，表示点 (x_i, y_i) 与点 (x_i, bx_i+a) 的“距离”越小，即样本数据点离直线 $y=bx+a$ 的竖直距离越小，如图 8.2-5 所示. 特别地，当 $e_i=0$ 时，表示点 (x_i, y_i) 在这条直线上.

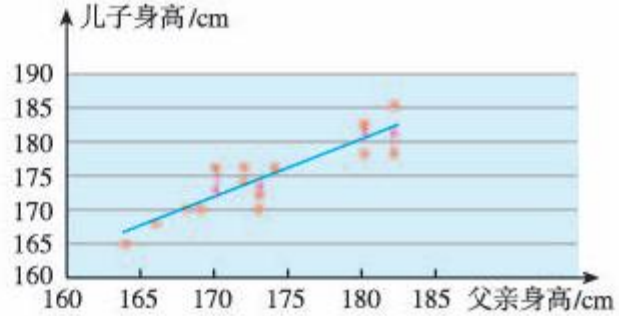


图 8.2-5

因此，可以用这 n 个竖直距离之和

$$\sum_{i=1}^{n}|y_i-(bx_i+a)|$$

来刻画各样本观测数据与直线 $y=bx+a$ 的“整体接近程度”.

在实际应用中，因为绝对值使得计算不方便，所以人们通常用各散点到直线的竖直距离的平方之和

$$Q=\sum_{i=1}^{n}(y_i-bx_i-a)^2$$

来刻画“整体接近程度”.

$$\begin{aligned} Q(a, b)&=\sum_{i=1}^{n}(y_i-bx_i-a)^2 \\ &=\sum_{i=1}^{n}[y_i-bx_i-(\overline{y}-b\overline{x})+(\overline{y}-b\overline{x})-a]^2 \\ &=\sum_{i=1}^{n}[(y_i-\overline{y})-b(x_i-\overline{x})+(\overline{y}-b\overline{x})-a]^2 \\ &=\sum_{i=1}^{n}[(y_i-\overline{y})-b(x_i-\overline{x})]^2+2\sum_{i=1}^{n}[(y_i-\overline{y})-b(x_i-\overline{x})]\times \\ &\quad [(\overline{y}-b\overline{x})-a]+n[(\overline{y}-b\overline{x})-a]^2, \end{aligned}$$

注意到

$$\begin{aligned} &\sum_{i=1}^{n}[(y_i-\overline{y})-b(x_i-\overline{x})](\overline{y}-b\overline{x}-a) \\ &=(\overline{y}-b\overline{x}-a)\sum_{i=1}^{n}[(y_i-\overline{y})-b(x_i-\overline{x})] \\ &=(\overline{y}-b\overline{x}-a)[\sum_{i=1}^{n}(y_i-\overline{y})-b\sum_{i=1}^{n}(x_i-\overline{x})] \\ &=(\overline{y}-b\overline{x}-a)[(n\overline{y}-n\overline{y})-b(n\overline{x}-n\overline{x})] \\ &=0, \end{aligned}$$

所以

$$Q(a, b)=\sum_{i=1}^{n}[(y_i-\overline{y})-b(x_i-\overline{x})]^2+n(\overline{y}-b\overline{x}-a)^2.$$

上式右边各项均为非负数，且前 n 项与 a 无关. 所以，要使 Q 取到最小值，后一项的值应为 0，即 $a=\overline{y}-b\overline{x}$. 此时

$$\begin{aligned} Q(a, b)&=\sum_{i=1}^{n}[(y_i-\overline{y})-b(x_i-\overline{x})]^2 \\ &=b^2\sum_{i=1}^{n}(x_i-\overline{x})^2-2b\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})+\sum_{i=1}^{n}(y_i-\overline{y})^2. \end{aligned}$$

上式是关于 b 的二次函数，因此要使 Q 取得最小值，当且仅当 b 的取值为

$$\hat{b}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}.$$

在一般情况下，直接比较两个模型的残差比较困难，因为在某些散点上一个模型的残差的绝对值比另一个模型的小，而另一些散点的情况则相反. 可以通过比较残差的平方和来比较两个模型的效果. 由

$$Q_1=\sum_{i=1}^{8}(\hat{e}_i)^2\approx 0.669,\quad Q_2=\sum_{i=1}^{8}(\hat{u}_i)^2\approx 0.004,$$

可知Q_2小于Q_1. 因此在残差平方和最小的标准下，非线性回归模型

$$\begin{cases}Y=c_2\ln(t-1\,895)+c_1+u,\\ E(u)=0,\ D(u)=\delta^2\end{cases}$$

的拟合效果要优于一元线性回归模型的拟合效果.

也可以用决定系数R^2来比较两个模型的拟合效果，R^2的计算公式为

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}.$$

在R^2表达式中，$\sum_{i=1}^{n}(y_i-\overline{y})^2$与经验回归方程无关，残差平方和$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$与经验回归方程有关. 因此$R^2$越大，表示残差平方和越小，即模型的拟合效果越好；$R^2$越小，表示残差平方和越大，即模型的拟合效果越差.

对于响应变量Y，通过观测得到的数据称为**观测值**，通过经验回归方程得到的$\hat{y}$称为**预测值**，观测值减去预测值称为**残差**. 残差是随机误差的估计结果，通过对残差的分析可以判断模型刻画数据的效果，以及判断原始数据中是否存在可疑数据等，这方面工作称为残差分析.

观察图 8.2-8 中四幅残差图，你认为哪一个残差满足一元线性回归模型中对随机误差的假定？

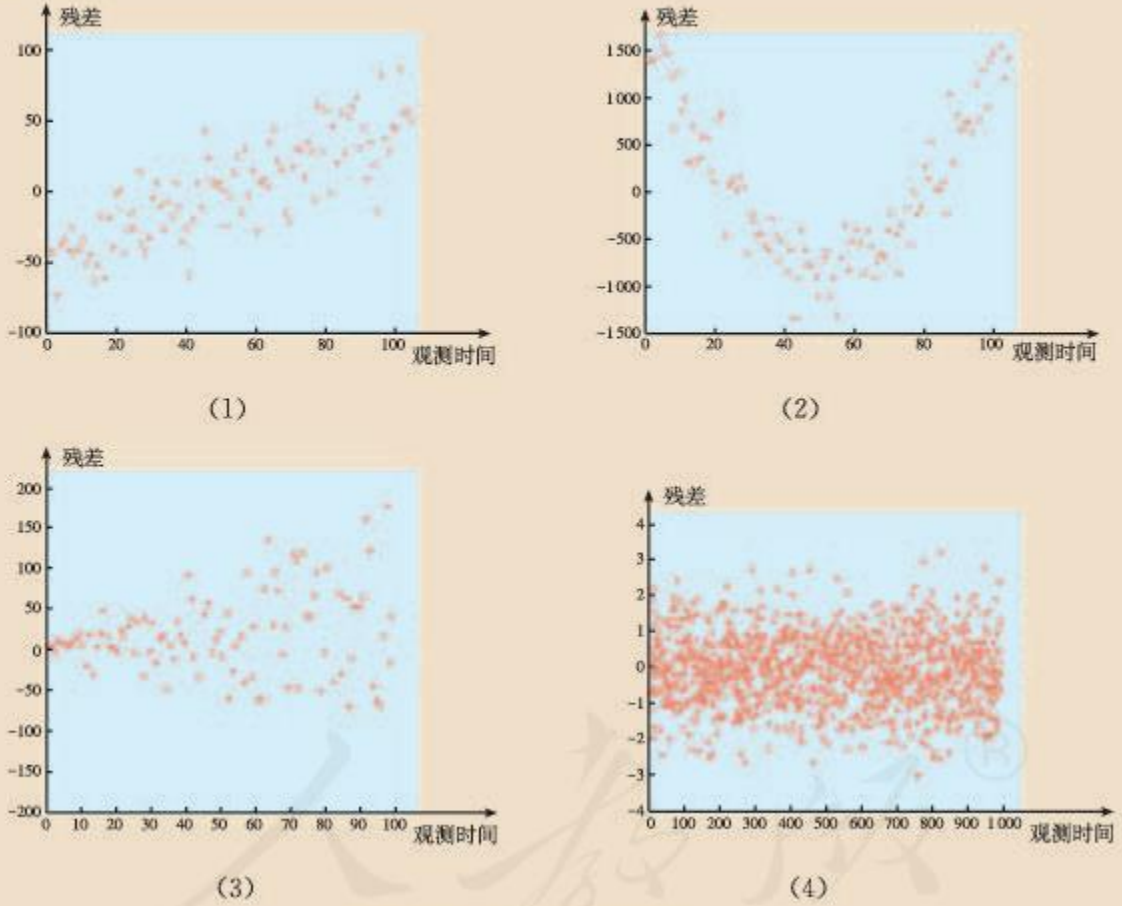
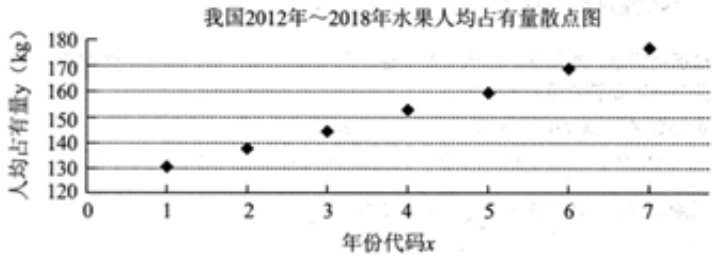


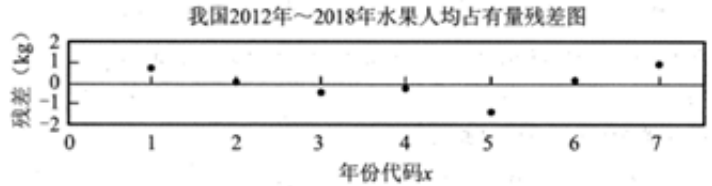
图 8.2-8　四种类型的残差图

根据一元线性回归模型中对随机误差的假定，残差应是均值为 0、方差为σ^2的随机变量的观测值. 在图 8.2-8 中，图（1）显示残差与观测时间有线性关系，应将时间变量纳入模型；图（2）显示残差与观测时间有非线性关系，应在模型中加入时间的非线性函数部分；图（3）说明残差的方差不是一个常数，随观测时间变大而变大；图（4）的残差比较均匀地分布在以取值为 0 的横轴为对称轴的水平带状区域内. 可见，在图 8.2-8 中，只有图（4）满足一元线性回归模型对随机误差的假设.

2019年12月山东省联考

20．下面给出了根据我国2012年~2018年水果人均占有量y（单位：kg）和年份代码x绘制的散点图和线性回归方程的残差图（2012年~2018年的年份代码x分别为1~7）．





（1）根据散点图分析y与x之间的相关关系；

（2）根据散点图相应数据计算得$\sum_{i=1}^{7} y_i = 1074, \sum_{i=1}^{7} x_i y_i = 4517$，求$y$关于$x$的线性回归方程；

（3）根据线性回归方程的残差图，分析线性回归方程的拟合效果．（精确到0．01）

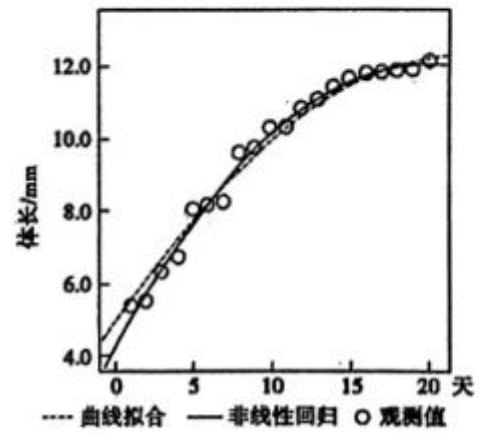
附：回归方程$\hat{y} = \hat{a} + \hat{b}x$中斜率和截距的最小二乘估计公式分别为：

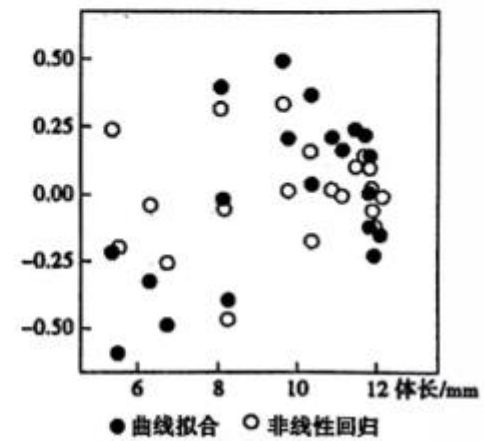
$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \hat{a} = \bar{y} - \hat{b}\bar{x}．$$

（3）由残差图可以看出，残差对应的点均匀地落在水平带状区域内，且宽度较窄，说明拟合效果较好．

2023年烟台一模

19．（12分）黄河鲤是我国华北地区的主要淡水养殖品种之一，其鳞片金黄、体形梭长，尤以色泽鲜丽、肉质细嫩、气味清香而著称．为研究黄河鲤早期生长发育的规律，丰富黄河鲤早期养殖经验，某院校研究小组以当地某水产养殖基地的黄河鲤仔鱼为研究对象，从出卵开始持续观察20天，试验期间，每天固定时段从试验水体中随机取出同批次9尾黄河鲤仔鱼测量体长，取其均值作为第t_i天的观测值y_i（单位：mm），其中$t_i = i$，$i = 1,2,3,\cdots,20$．根据以往的统计资料，该组数据(t_i, y_i)可以用Logistic曲线拟合模型$y = \frac{1}{\frac{1}{u} + ab^t}$或Logistic非线性回归模型$y = \frac{u}{1+e^{a-bt}}$进行统计分析，其中$a,b,u$为参数．基于这两个模型，绘制得到如下的散点图和残差图：





（1）你认为哪个模型的拟合效果更好？分别结合散点图和残差图进行说明；

19．解：（1）Logistic非线性回归模型$y = \frac{u}{1+e^{a-bt}}$拟合效果更好．……1分

从散点图看，散点更均匀地分布在该模型拟合曲线附近；从残差图看，该模型下的残差更均匀地集中在以残差为0的直线为对称轴的水平带状区域内．……3分

2023年潍坊一模

20.（12分）

某学校研究性学习小组在学习生物遗传学的过程中，为验证高尔顿提出的关于儿子成年后身高 y（单位：cm）与父亲身高 x（单位：cm）之间的关系及存在的遗传规律，随机抽取了5对父子的身高数据，如下表：

父亲身高 x	160	170	175	185	190
儿子身高 y	170	174	175	180	186

（1）根据表中数据，求出 y 关于 x 的线性回归方程，并利用回归直线方程分别确定儿子比父亲高和儿子比父亲矮的条件，由此可得到怎样的遗传规律？

（2）记 $\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{b}x_i - \hat{a}(i=1,2,\cdots,n)$，其中 y_i 为观测值，$\hat{y}_i$ 为预测值，$\hat{e}_i$ 为对应 (x_i, y_i) 的残差. 求（1）中儿子身高的残差的和，并探究这个结果是否对任意具有线性相关关系的两个变量都成立？若成立加以证明；若不成立说明理由.

参考数据及公式：$\sum_{i=1}^{5} x_i = 880$，$\sum_{i=1}^{5} x_i^2 = 155450$，$\sum_{i=1}^{5} y_i = 885$，$\sum_{i=1}^{5} x_i y_i = 156045$，

$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

$\hat{a} = \bar{y} - \hat{b}\bar{x} = 177 - 0.5 \times 176 = 89$，

所以回归直线方程为 $y = 0.5x + 89$，…… 4分

令 $0.5x + 89 - x > 0$ 得 $x < 178$，即 $x < 178$ 时，儿子比父亲高；

令 $0.5x - 89 - x < 0$ 得 $x > 178$，即 $x > 178$ 时，儿子比父亲矮，…… 5分

可得当父亲身高较高时，儿子平均身高要矮于父亲，即儿子身高有一个回归，回归到全种群平均高度的趋势.（意思对即可） …… 6分

（2）$\hat{y}_1 = 169$，$\hat{y}_2 = 174$，$\hat{y}_3 = 176.5$，$\hat{y}_4 = 181.5$，$\hat{y}_5 = 184$，所以 $\sum_{i=1}^{5} \hat{y}_i = 885$，

又 $\sum_{i=1}^{5} y_i = 885$，所以 $\sum_{i=1}^{5} \hat{e}_i = 0$，…… 8分

结论：对任意具有线性相关关系的变量 $\sum_{i=1}^{n} \hat{e}_i = 0$，…… 9分

证明：$\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{b}x_i - \hat{a})$

$= \sum_{i=1}^{n} y_i - \hat{b}\sum_{i=1}^{n} x_i - n\hat{a} = n\bar{y} - n\hat{b}\bar{x} - n(\bar{y} - \hat{b}\bar{x}) = 0.$ …… 12分

拓展阅读

“回归”一词的由来

《现代汉语词典（第7版）》中，“回归”的解释是“回到（原来的地方）”；地理学中，“回归线”是指地球赤道南北各23°26′处的纬线，太阳直射点在南回归线与北回归线之间来回移动. 看了这些，你是不是感觉到回归直线方程中的“回归”与上面这些说法相差很大？

统计学中的“回归”一词，是统计学家高尔顿引入的. 早在19世纪80年代，高尔顿就开始了亲代与子代（即父母亲与子女）之间相似特征（身高、性格等）的研究. 他收集了一些亲代的身高 x 与子代的身高 y 的成对数据，并作出了散点图，发现 y 与 x 的关系可以借助一次函数来近似表示，而且总体上亲代的身高增加时，子代的身高也增加.

但是，高尔顿在研究过程中，发现了一个有趣的现象. 他收集的数据显示，总体上亲代的平均身高为68英寸（约为172.72 cm），子代的平均身高为69英寸，子代的平均身高比亲代的平均身高大1英寸（约为2.54 cm）. 于是，一个自然的推想是：平均身高为63英寸的亲代，其子代的平均身高应约为64英寸；平均身高为72英寸的亲代，其子代的平均身高应约为73英寸. 但实际数据显示：平均身高为63英寸的亲代，其子代的平均身高为67英寸，增加量为4英寸；平均身高为72英寸的亲代，其子代的平均身高为71英寸，增加量为−1英寸. 也就是说，平均身高不同的亲代，其子代的平均身高增加量并不相等，但子代的平均身高有回归于中心（即总体平均值）的趋势.

正是由于这种现象的存在，高尔顿引入了“回归”一词. 虽然不是所有相关关系中都会发生类似的现象，但从那以后，“回归”就成了相关关系讨论中一个约定俗成的词了.

小阅读：相关关系

自行选择标准，将下列变量之间的关系分为两类，并分别阐述每一类中变量关系的特点：

(1) 圆的面积 S 与半径 r 之间的关系；

(2) 16 岁学生的体重 w 与身高 h 之间的关系；

(3) 商品销售量 Q 与销售价格 P 之间的关系；

(4) 匀速运动的物体，其运动的路程 S 与时间 t 之间的关系；

(5) 平均学习时间 t 与学习成绩 f 之间的关系；

(6) 科技创新能力 y 与人才培养近亲繁殖率 x 之间的关系.

问题 1：上述的两个变量间的关系有那几个是我们熟悉的函数关系？

问题 2：上述的两个变量间的关系不是函数关系的，它们之间有关系吗？

新概念：

例如，一般情况下，已知一名 16 岁学生的身高 h，不能确定其体重 w，但身高越高的人体重可能越重，不过同样身高的人体重往往存在差异；商品的销售价格 P 越低，买这种商品的人可能会越多，从而会导致销售量 Q 增长，但同样的销售价格可能会有不同的销售量；平均学习时间 t 越长，学习成绩 f 可能越好，但学习时间相同不能确保学习成绩相同；人才培养近亲繁殖率 x 越大，科技创新能力 y 可能越弱，但繁殖率确定时，创新能力并不是确定的. 这些两个变量之间的关系，统计学上都称为**相关关系**.

问题 3：你能再举几个生活中两个变量之间有相关关系的实例吗？

问题 4：你能找到直观的方法描述两个变量之间存在相关关系吗？
（可以回忆学习函数时，函数的那种表示方法可以直观的反映出两个变量间的关系）

问题 5：尝试分析下列数据，确定变量 x 与变量 y 是否有相关关系？你还能做出哪些猜想？

某地区从某一年开始进行了环境污染整治，得到了如下数据：

第 x 年	1	2	3	4	5	6	7
污染指数 y	6.1	5.2	4.5	4.7	3.8	3.4	3.1

景问题和
解对这些
，然后就
新情景问
材，只是
织成阅读
，以“相
查形式应
成研究性
例

前面我们已掌握：

变量 X、Y 满足 $Y=aX+b$，则 $E(Y)=$_____，$D(Y)=$_____，$\sqrt{D(Y)}=$_____.

现在我们学习了正态分布：$X\sim N(\mu,\sigma^2)$，其中 μ 为变量 X 的_____；σ 为变量 X 的_____.

问题：能否借鉴之前学过的知识将正态分布 $X\sim N(\mu,\sigma^2)$ 变换为标准正态分布 $Y\sim N(0,1)$？

……**（提示：尝试构造 X 与 Y 的关系式）**

标准正态分布的应用：

如果 $X\sim N(0,1)$，那么对于任意 a，通常记

$$\Phi(a)=P(X<a),$$

也就是说 $\Phi(a)$ 表示 $N(0,1)$ 对应的正态曲线与 x 轴在区间 $(-\infty,a)$ 内所围的面积，如图 4-2-15 所示.

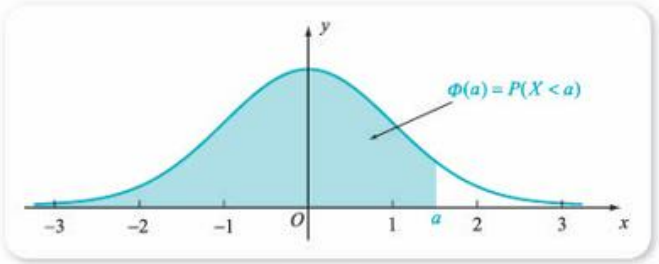


图 4-2-15

为了方便起见，人们将 $a\geqslant 0$ 时部分 $\Phi(a)$ 的值制成了专门的表格，可供查询，下表是部分数据.

$\Phi(a)=P(X<a)$

a	0	1	2	3	4	5	6	7	8	9
0.0	.500 0	.504 0	.508 0	.512 0	.516 0	.519 9	.523 9	.527 9	.531 9	.535 9
0.1	.539 8	.543 8	.547 8	.551 7	.555 7	.559 6	.563 6	.567 5	.571 4	.575 3
0.2	.579 3	.583 2	.587 1	.591 0	.594 8	.598 7	.602 6	.606 4	.610 3	.614 1
0.3	.617 9	.621 7	.625 5	.629 3	.633 1	.636 8	.640 6	.644 3	.648 0	.651 7
0.4	.655 4	.659 1	.662 8	.666 4	.670 0	.673 6	.677 2	.680 8	.684 4	.687 9
0.5	.691 5	.695 0	.698 5	.701 9	.705 4	.708 8	.712 3	.715 7	.719 0	.722 4

例如，从上表中可以查出，$\Phi(0.16)=0.563\,6$，$\Phi(0.58)=0.719\,0$.

例 4 已知 $X\sim N(0,1)$，利用上述表格求以下概率值：

(1) $P(X<0.28)$；　　(2) $P(X<-0.36)$；

(3) $P(0.18\leqslant X<0.57)$.

二、高考题中的概率统计问题

2005-2017年山东省自主命制，概率统计解答题一直较稳定，主要考察古典概型、事件的独立性、分布列等，模型也较简单，如比赛、选人、答题等。自2018年使用全国卷，感觉难度直接提升了一大截，模型复杂，综合性强。

2018年全国卷

20. 某工厂的某种产品成箱包装，每箱 200 件，每一箱产品在交付用户之前要对产品作检验，如检验出不合格品，则更换为合格品．检验时，先从这箱产品中任取 20 件作检验，再根据检验结果决定是否对余下的所有产品作检验，设每件产品为不合格品的概率都为 p $(0<p<1)$，且各件产品是否为不合格品相互独立．

(1) 记 20 件产品中恰有 2 件不合格品的概率为 $f(p)$，求 $f(p)$ 的最大值点 p_0；

(2) 现对一箱产品检验了 20 件，结果恰有 2 件不合格品，以 (1) 中确定的 p_0 作为 p 的值．已知每件产品的检验费用为 2 元，若有不合格品进入用户手中，则工厂要对每件不合格品支付 25 元的赔偿费用．

① 若不对该箱余下的产品作检验，这一箱产品的检验费用与赔偿费用的和记为 X，求 EX；

② 以检验费用与赔偿费用和的期望值为决策依据，是否该对这箱余下的所有产品作检验？

2019年全国卷

21. 为治疗某种疾病，研制了甲、乙两种新药，希望知道哪种新药更有效，为此进行动物试验．试验方案如下：每一轮选取两只白鼠对药效进行对比试验．对于两只白鼠，随机选一只施以甲药，另一只施以乙药．一轮的治疗结果得出后，再安排下一轮试验．当其中一种药治愈的白鼠比另一种药治愈的白鼠多 4 只时，就停止试验，并认为治愈只数多的药更有效．为了方便描述问题，约定：对于每轮试验，若施以甲药的白鼠治愈且施以乙药的白鼠未治愈则甲药得 1 分，乙药得 -1 分；若施以乙药的白鼠治愈且施以甲药的白鼠未治愈则乙药得 1 分，甲药得 -1 分；若都治愈或都未治愈则两种药均得 0 分．甲、乙两种药的治愈率分别记为 α 和 β，一轮试验中甲药的得分记为 X．

(1) 求 X 的分布列；

(2) 若甲药、乙药在试验开始时都赋予 4 分，p_i $(i=0,\ 1,\ \cdots,\ 8)$ 表示“甲药的累计得分为 i 时，最终认为甲药比乙药更有效”的概率，则 $p_0=0$，$p_8=1$，$p_i=ap_{i-1}+bp_i+cp_{i+1}$ $(i=1,\ 2,\ \cdots,\ 7)$，其中 $a=P(X=-1)$，$b=P(X=0)$，$c=P(X=1)$．假设 $\alpha=0.5$，$\beta=0.8$．

① 证明：$\{p_{i+1}-p_i\}$ $(i=0,\ 1,\ 2,\ \cdots,\ 7)$ 为等比数列；

② 求 p_4，并根据 p_4 的值解释这种试验方案的合理性．

其实全国卷的概率统计一直都较难，只是在18、19年达到顶点。

2017年全国卷

19. 为了监控某种零件的一条生产线的生产过程，检验员每天从该生产线上随机抽取 16 个零件，并测量其尺寸（单位：cm）. 根据长期生产经验，可以认为这条生产线正常状态下生产的零件的尺寸服从正态分布 $N(\mu,\sigma^2)$.

(1) 假设生产状态正常，记 X 表示一天内抽取的 16 个零件中其尺寸在 $(\mu-3\sigma,\mu+3\sigma)$ 之外的零件数，求 $P(X\geqslant 1)$ 及 X 的数学期望；

(2) 一天内抽检零件中，如果出现了尺寸在 $(\mu-3\sigma,\mu+3\sigma)$ 之外的零件，就认为这条生产线在这一天的生产过程可能出现了异常情况，需对当天的生产过程进行检查.

① 试说明上述监控生产过程方法的合理性；

② 下面是检验员在一天内抽取的 16 个零件的尺寸：

9.95	10.12	9.96	9.96	10.01	9.92	9.98	10.04
10.26	9.91	10.13	10.02	9.22	10.04	10.05	9.95

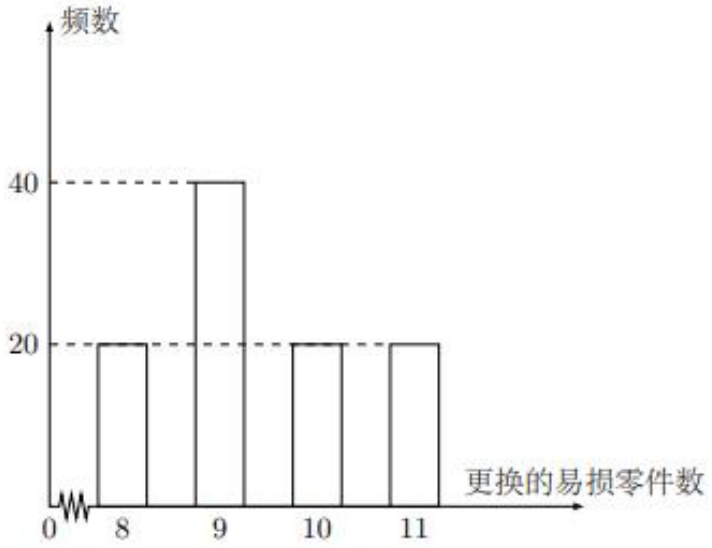
经计算得 $\overline{x}=\frac{1}{16}\sum_{i=1}^{16}x_i=9.97$，$s=\sqrt{\frac{1}{16}\sum_{i=1}^{16}(x_i-\overline{x})^2}=\sqrt{\frac{1}{16}\left(\sum_{i=1}^{16}x_i^2-16\overline{x}^2\right)}\approx 0.212$，其中 x_i 为抽取的第 i 个零件的尺寸，$i=1,2,\cdots,16$.

用样本平均数 $\overline{x}$ 作为 μ 的估计值 $\hat{\mu}$，用样本标准差 s 作为 σ 的估计值 $\hat{\sigma}$，利用估计值判断是否需对当天的生产过程进行检查？剔除 $(\hat{\mu}-3\hat{\sigma},\hat{\mu}+3\hat{\sigma})$ 之外的数据，用剩下的数据估计 μ 和 σ（精确到 0.01）.

附：若随机变量 Z 服从正态分布 $N(\mu,\sigma^2)$，则 $P(\mu-3\sigma<Z<\mu+3\sigma)=0.9974$，$0.9974^{16}\approx 0.9592$，$\sqrt{0.008}\approx 0.09$.

2016年全国卷

19. 某公司计划购买 2 台机器，该种机器使用三年后即被淘汰，机器有一易损零件，在购进机器时，可以额外购买这种零件作为备件，每个 200 元. 在机器使用期间，如果备件不足再购买，则每个 500 元. 现需决策在购买机器时应同时购买几个易损零件，为此搜集并整理了 100 台这种机器在三年使用期内更换的易损零件数，得下面柱状图：



以这 100 台机器更换的易损零件数的频率代替 1 台机器更换的易损零件数发生的概率，记 X 表示 2 台机器三年内共需更换的易损零件数，n 表示购买 2 台机器的同时购买的易损零件数.

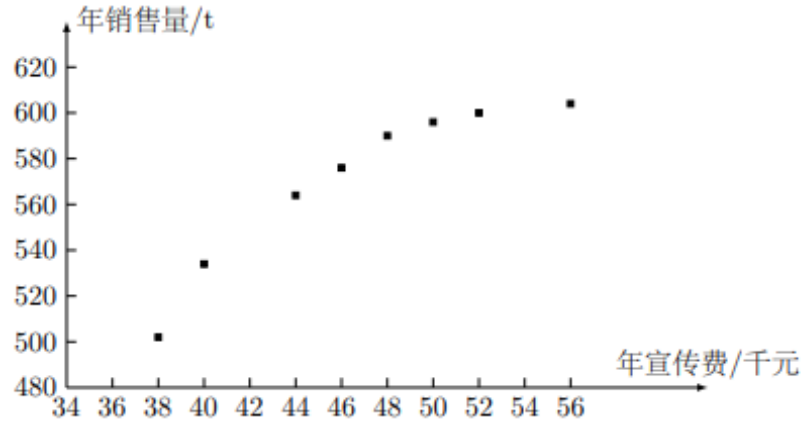
(1) 求 X 的分布列；

(2) 若要求 $P(X\leqslant n)\geqslant 0.5$，确定 n 的最小值；

(3) 以购买易损零件所需费用的期望值为决策依据，在 $n=19$ 与 $n=20$ 之中选其一，应选用哪个？

2015年全国卷

19. 某公司为确定下一年度投入某产品的宣传费, 需了解年宣传费 x (单位: 千元) 对年销售量 y (单位: t) 和年利润 z (单位: 千元) 的影响. 对近 8 年的年宣传费 x_i 和年销售量 y_i $(i=1,\ 2,\ \cdots,\ 8)$ 数据作了初步处理, 得到下面的散点图及一些统计量的值.



$\overline{x}$	$\overline{y}$	$\overline{w}$	$\sum_{i=1}^{8}(x_i-\overline{x})^2$	$\sum_{i=1}^{8}(w_i-\overline{w})^2$
46.6	563	6.8	289.8	1.6

$\sum_{i=1}^{8}(x_i-\overline{x})(y_i-\overline{y})$	$\sum_{i=1}^{8}(w_i-\overline{w})(y_i-\overline{y})$
1.469	108.8

表中 $w_i=\sqrt{x_i}$, $\overline{w}=\dfrac{1}{8}\sum_{i=1}^{8}w_i$.

(1) 根据散点图判断, $y=a+bx$ 与 $y=c+d\sqrt{x}$ 哪一个适宜作为年销售量 y 关于年宣传费 x 的回归方程类型? (给出判断即可, 不必说明理由)

(2) 根据 (1) 的判断结果及表中数据, 建立 y 关于 x 的回归方程;

(3) 已知这种产品的年利润 z 与 x, y 的关系为 $z=0.2y-x$. 根据 (2) 的结果回答下列问题:

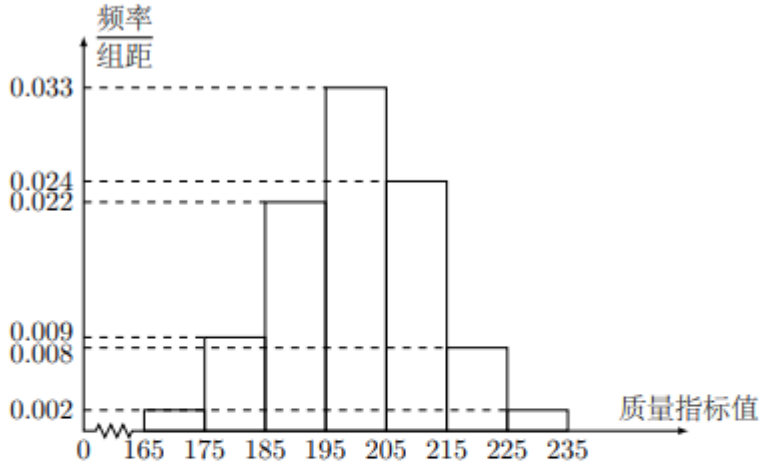
① 年宣传费 $x=49$ 时, 年销售量及年利润的预报值是多少?

② 年宣传费 x 为何值时, 年利润的预报值最大?

附: 对于一组数据 $(u_1,v_1),(u_2,v_2),\cdots,(u_n,v_n)$, 其回归直线 $v=\alpha+\beta u$ 的斜率和截距的最小二乘估计分别为 $\hat{\beta}=\dfrac{\sum_{i=1}^{n}(u_i-\overline{u})(v_i-\overline{v})}{\sum_{i=1}^{n}(u_i-\overline{u})^2}$, $\hat{\alpha}=\overline{v}-\hat{\beta}\overline{u}$.

2014年全国卷

18. 从某企业生产的某种产品中抽取 500 件, 测量这些产品的一项质量指标值, 由测量结果得如下频率分布直方图:



(1) 求这 500 件产品质量指标值的样本平均数 $\overline{x}$ 和样本方差 s^2 (同一组数据用该区间的中点值作代表);

(2) 由频率分布直方图可以认为, 这种产品的质量指标值 Z 服从正态分布 $N(\mu,\sigma^2)$, 其中 μ 近似为样本平均数 $\overline{x}$, σ^2 近似为样本方差 s^2.

① 利用该正态分布, 求 $P(187.8<Z<212.2)$;

② 某用户从该企业购买了 100 件这种产品, 记 X 表示这 100 件产品中质量指标值位于区间 $(187.8,212.2)$ 的产品件数, 利用①的结果, 求 EX.

附: $\sqrt{150}\approx 12.2$, 若 $Z\sim N(\mu,\sigma^2)$, 则 $P(\mu-\sigma<Z<\mu+\sigma)=0.6826$, $P(\mu-2\sigma<Z<\mu+2\sigma)=0.9544$.

年份	问题的实际背景	考查内容
2014年	某种产品的质量指标值	频率分布直方图（求均值、方差）、正态分布、二项分布
2015年	产品的宣传费对销量的影响	依据散点图选拟合模型、非线性回归方程、预测问题、最值问题
2016年	机器易损件的更换	分布列（事件的独立性）、最值问题、决策问题（期望）
2017年	生产线生产过程的监控	正态分布、二项分布、概率的实际意义（说明监控方法的合理性）、数据处理（剔除部分数据后求均值、方差）
2018年	产品的检验问题	二项分布、利用导数求最值、决策问题（期望）
2019年	新药的动物试验	分布列（事件的独立性）、递推公式（数列）、概率的实际意义（解释试验的合理性）
2020年	羽毛球比赛（3人淘汰）	事件的独立性、对赛制的分析
2021年	新旧设备生产的产品指标比较	求离散数据的均值和方差、比较大小（判断）
2022年	林区某种树木的总材积量	均值、求相关系数（公式的变形）、估测问题

2020年至今新高考，概率统计题的难度有所降低，但随着加入新高考的省越来越多，命题方向应该有回归全国卷的趋势。

2020年新高考

19. 为加强环境保护，治理空气污染，环境监测部门对某市空气质量进行调研，随机抽查了 100 天空气中的 PM2.5 和 SO_2 浓度（单位：$\mu g/m^3$），得下表：

PM2.5 \ SO_2	$[0,50]$	$(50,150]$	$(150,475]$
$[0,35]$	32	18	4
$(35,75]$	6	8	12
$(75,115]$	3	7	10

(1) 估计事件“该市一天空气中 PM2.5 浓度不超过 75，且 SO_2 浓度不超过 150”的概率；

(2) 根据所给数据，完成下面的 2×2 列联表：

PM2.5 \ SO_2	$[0,150]$	$(150,475]$
$[0,75]$		
$(75,115]$		

(3) 根据 (2) 中的列联表，判断是否有 99% 的把握认为该市一天空气中 PM2.5 浓度与 SO_2 浓度有关？

附：$K^2=\dfrac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$，

$P(K^2\geqslant k)$	0.050	0.010	0.001
k	3.841	6.635	10.828

2020年全国卷

19. 甲、乙、丙三位同学进行羽毛球比赛，约定赛制如下：累计负两场者被淘汰；比赛前抽签决定首先比赛的两人，另一人轮空；每场比赛的胜者与轮空者进行下一场比赛，负者下一场轮空，直至有一人被淘汰；当一人被淘汰后，剩余的两人继续比赛，直至其中一人被淘汰，另一人最终获胜，比赛结束．经抽签，甲、乙首先比赛，丙轮空．设每场比赛双方获胜的概率都为 $\dfrac{1}{2}$．

(1) 求甲连胜四场的概率；

(2) 求需要进行第五场比赛的概率；

(3) 求丙最终获胜的概率．

2021年新高考

18．某学校组织“一带一路”知识竞赛，有A，B两类问题，每位参加比赛的同学先在两类问题中选择一类并从中随机抽取一个问题回答，若回答错误则该同学比赛结束；若回答正确则从另一类问题中再随机抽取一个问题回答，无论回答正确与否，该同学比赛结束．A类问题中的每个问题回答正确得20分，否则得0分；B类问题中的每个问题回答正确得80分，否则得0分，已知小明能正确回答A类问题的概率为0.8，能正确回答B类问题的概率为0.6，且能正确回答问题的概率与回答次序无关．

（1）若小明先回答A类问题，记X为小明的累计得分，求X的分布列；

（2）为使累计得分的期望最大，小明应选择先回答哪类问题？并说明理由．

2021年全国卷

17. 某厂研制了一种生产高精产品的设备，为检验新设备生产产品的某项指标有无提高，用一台旧设备和一台新设备各生产了 10 件产品，得到各件产品该项指标数据如下：

旧设备	9.8	10.3	10.0	10.2	9.9	9.8	10.0	10.1	10.2	9.7
新设备	10.1	10.4	10.1	10.0	10.1	10.3	10.6	10.5	10.4	10.5

旧设备和新设备生产产品的该项指标的样本平均值分别记为 $\bar{x}$ 和 $\bar{y}$，样本方差分别记为 s_1^2 和 s_2^2.

(1) 求 $\bar{x}$, $\bar{y}$, s_1^2, s_2^2;

(2) 判断新设备生产产品的该项指标的均值较旧设备是否有显著提高（如果 $\bar{y}-\bar{x} \geqslant 2\sqrt{\dfrac{s_1^2+s_2^2}{10}}$，则认为新设备生产产品的该项指标的均值较旧设备有显著提高，否则不认为有显著提高）.

2022年新高考

20. 一医疗团队为研究某地的一种地方性疾病与当地居民的卫生习惯（卫生习惯分为良好和不够良好两类）的关系，在已患该疾病的病例中随机调查了 100 例（称为病例组），同时在未患该疾病的人群中随机调查了 100 人（称为对照组），得到如下数据：

	不够良好	良好
病例组	40	60
对照组	10	90

(1)能否有 99%的把握认为患该疾病群体与未患该疾病群体的卫生习惯有差异？

(2)从该地的人群中任选一人，A 表示事件“选到的人卫生习惯不够良好”，B 表示事件“选到的人患有该疾病”，$\frac{P(B|A)}{P(\bar{B}|A)}$ 与 $\frac{P(B|\bar{A})}{P(\bar{B}|\bar{A})}$ 的比值是卫生习惯不够良好对患该疾病风险程度的一项度量指标，记该指标为 R.

（i）证明：$R=\frac{P(A|B)}{P(\bar{A}|B)}\cdot\frac{P(\bar{A}|\bar{B})}{P(A|\bar{B})}$；

（ii）利用该调查数据，给出 $P(A|B),P(A|\bar{B})$ 的估计值，并利用（i）的结果给出 R 的估计值.

2022年全国卷

19. 某地经过多年的环境治理，已将荒山改造成了绿水青山．为估计一林区某种树木的总材积量，随机选取了 10 棵这种树木，测量每棵树的根部横截面积（单位：m^2）和材积量（单位：m^3），得到如下数据：

样本号 i	1	2	3	4	5	6	7	8	9	10	总和
根部横截面积 x_i	0.04	0.06	0.04	0.08	0.08	0.05	0.05	0.07	0.07	0.06	0.6
材积量 y_i	0.25	0.40	0.22	0.54	0.51	0.34	0.36	0.46	0.42	0.40	3.9

并计算得 $\sum_{i=1}^{10}x_i^2=0.038,\sum_{i=1}^{10}y_i^2=1.6158,\sum_{i=1}^{10}x_iy_i=0.2474$．

(1)估计该林区这种树木平均一棵的根部横截面积与平均一棵的材积量；

(2)求该林区这种树木的根部横截面积与材积量的样本相关系数（精确到 0.01）；

(3)现测量了该林区所有这种树木的根部横截面积，并得到所有这种树木的根部横截面积总和为 $186m^2$．已知树木的材积量与其根部横截面积近似成正比．利用以上数据给出该林区这种树木的总材积量的估计值．

附：相关系数 $r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$，$\sqrt{1.896}\approx1.377$．

从2021年起新高考分一二卷，一二卷试题的对比成为猜测命题方向的方式之一.

2021年新高考（1卷）

18．某学校组织“一带一路”知识竞赛，有A，B两类问题，每位参加比赛的同学先在两类问题中选择一类并从中随机抽取一个问题回答，若回答错误则该同学比赛结束；若回答正确则从另一类问题中再随机抽取一个问题回答，无论回答正确与否，该同学比赛结束．A类问题中的每个问题回答正确得20分，否则得0分；B类问题中的每个问题回答正确得80分，否则得0分，已知小明能正确回答A类问题的概率为0.8，能正确回答B类问题的概率为0.6，且能正确回答问题的概率与回答次序无关．

（1）若小明先回答A类问题，记X为小明的累计得分，求X的分布列；

（2）为使累计得分的期望最大，小明应选择先回答哪类问题？并说明理由．

2021年新高考（2卷）

21．一种微生物群体可以经过自身繁殖不断生存下来，设一个这种微生物为第0代，经过一次繁殖后为第1代，再经过一次繁殖后为第2代……，该微生物每代繁殖的个数是相互独立的且有相同的分布列，设X表示1个微生物个体繁殖下一代的个数，

$P(X=i)=p_i(i=0,1,2,3)$．

（1）已知$p_0=0.4, p_1=0.3, p_2=0.2, p_3=0.1$，求$E(X)$；

（2）设p表示该种微生物经过多代繁殖后临近灭绝的概率，p是关于x的方程：$p_0+p_1x+p_2x^2+p_3x^3=x$的一个最小正实根，求证：当$E(X)\le 1$时，$p=1$，当$E(X)>1$时，$p<1$；

（3）根据你的理解说明（2）问结论的实际含义．

2022年新高考（1卷）

20．一医疗团队为研究某地的一种地方性疾病与当地居民的卫生习惯（卫生习惯分为良好和不够良好两类）的关系，在已患该疾病的病例中随机调查了 100 例（称为病例组），同时在未患该疾病的人群中随机调查了 100 人（称为对照组），得到如下数据：

	不够良好	良好
病例组	40	60
对照组	10	90

(1)能否有 99%的把握认为患该疾病群体与未患该疾病群体的卫生习惯有差异？

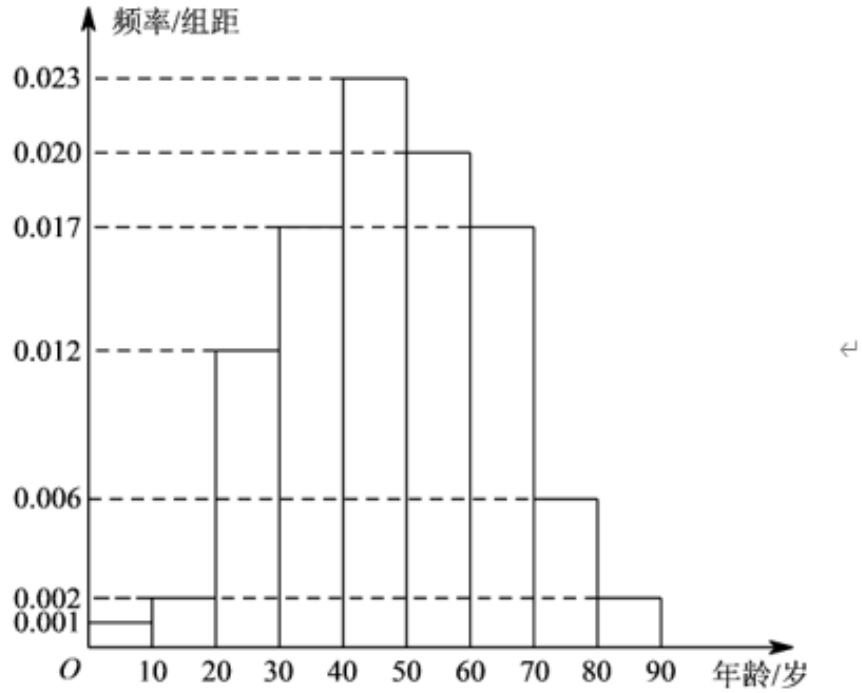
(2)从该地的人群中任选一人，A 表示事件“选到的人卫生习惯不够良好”，B 表示事件“选到的人患有该疾病”．$\frac{P(B|A)}{P(\bar{B}|A)}$ 与 $\frac{P(B|\bar{A})}{P(\bar{B}|\bar{A})}$ 的比值是卫生习惯不够良好对患该疾病风险程度的一项度量指标，记该指标为 R．

（i）证明：$R=\frac{P(A|B)}{P(\bar{A}|B)}\cdot\frac{P(\bar{A}|\bar{B})}{P(A|\bar{B})}$；

（ii）利用该调查数据，给出 $P(A|B),P(A|\bar{B})$ 的估计值，并利用（i）的结果给出 R 的估计值．

2022年新高考（2卷）

19．在某地区进行流行病学调查，随机调查了 100 位某种疾病患者的年龄，得到如下的样本数据的频率分布直方图：



(1)估计该地区这种疾病患者的平均年龄（同一组中的数据用该组区间的中点值为代表）；

(2)估计该地区一位这种疾病患者的年龄位于区间 $[20,70)$ 的概率；

(3)已知该地区这种疾病的患病率为 0.1%，该地区年龄位于区间 $[40,50)$ 的人口占该地区总人口的16%.从该地区中任选一人，若此人的年龄位于区间 $[40,50)$，求此人患这种疾病的概率．（以样本数据中患者的年龄位于各区间的频率作为患者的年龄位于该区间的概率，精确到 0.0001）．

年份	问题的实际背景	考查内容
2020年	空气质量指标检测	用频率估计概率、独立性检验（2*2列联表、卡方统计量、相关性的判断）
2021年	答题问题	分布列（事件的独立性）、决策问题（期望）
	微生物繁殖（代数、数量）	分布列（不用求）、期望、三次方程的实数解（结合参数的实际意义给出证明）、说明结论的实际意义
2022年	地区性疾病与居民卫生习惯的关系	独立性检验、条件概率和乘法公式（证明）、用频率估计概率（条件概率）
	流行病学调查，患者年龄分布	频率分布直方图（求均值）、频率估计概率、条件概率和乘法公式（对已知概率事件的理解）

统计概率的命题基于实际问题，对学生的阅读、分析转化问题、提炼和处理数据等方面的能力要求较高，重视学生在平时学习过程中情境模型的归类整理，形成相对系统的概率模型和统计模型。

与概率统计相关的复杂情境题型：

1、模拟类：根据实际情境中出现的问题提出合理的假设，并利用假设提出几种方案和解决方案；

2、分析类：分析实际情境中出现的问题，找出关联关系，分辨变量间关系，对多变量题型提出解题思路；

3、推断类：利用已知条件及实际情况推断出结论，运用统计分析的方法进行评价，模拟不同的情况以论证结论的正确性和可行性。

2023泰安一模

20.(12分)

某公司为活跃气氛提升士气，年终拟通过抓阄兑奖的方式对所有员工进行奖励.规定：每位员工从一个装有4个标有面值的阄的袋中一次性随机摸出2个阄，阄上所标的面值之和为该员工获得的奖励金额.

(1)若袋中所装的4个阄中有1个所标的面值为800元，其余3个均为200元，求

①员工所获得的奖励为1000元的概率；

②员工所获得的奖励额的分布列及数学期望；

(2)公司对奖励额的预算是人均1000元，并规定袋中的4个阄只能由标有面值200元和800元的两种阄或标有面值400元和600元的两种阄组成.为了使员工得到的奖励总额尽可能符合公司的预算且每位员工所获得的奖励额相对均衡，请对袋中的4个阄的面值给出一个合适的设计，并说明理由.

尝试对教材中情境问题的改编（以超几何与二项分布为例）

问题 已知 100 件产品中有 8 件次品，分别采用有放回和不放回的方式随机抽取 4 件. 设抽取的 4 件产品中次品数为 X，求随机变量 X 的分布列.

我们知道，如果采用有放回抽样，则每次抽到次品的概率为 0.08，且各次抽样的结果相互独立，此时 X 服从二项分布，即 $X\sim B(4, 0.08)$.

采用不放回抽样，虽然每次抽到次品的概率都是 0.08，但每次抽取不是同一个试验，而且各次抽取的结果也不独立，不符合 n 重伯努利试验的特征，因此 X 不服从二项分布.

可以根据古典概型求 X 的分布列. 由题意可知，X 可能的取值为 0，1，2，3，4. 从 100 件产品中任取 4 件，样本空间包含 C_{100}^4 个样本点，且每个样本点都是等可能发生的. 其中 4 件产品中恰有 k 件次品的结果数为 $C_8^k C_{92}^{4-k}$. 由古典概型的知识，得 X 的分布列为

$$P(X=k)=\frac{C_8^k C_{92}^{4-k}}{C_{100}^4},\ k=0, 1, 2, 3, 4.$$

计算结果数时，考虑抽取的次序和不考虑抽取的次序，对分布列的计算有影响吗？为什么？

1、已知 7 件产品中有 3 件次品，分别采用有放回和无放回的方式随机抽取 3 件.

（1）分别求两种抽取方式每次抽到次品的概率，你能得到什么结论？

（2）设抽取的 3 件产品中次品数为 X，分别求两种抽样方式中 X 的期望，你有什么发现，简单说明原因（合理即可）.

感谢倾听，不当之处请多多指教.