§13.3 统计初步

13.3.1 相关概念

统计初步

众数：一组数据中出现次数最多的那个数据，叫做这组数据的**众数**。

中位数：令 $x_1, x_2, \cdots, x_n$ 为由小到大排列的一组数，则 x_{k+1}（$n=2k+1$）或 $\frac{x_k + x_{k+1}}{2}(n=2k)$ 叫做这组数据的**中位数**。

百分位数：如果一组数据至少有 $p\%(p \in (0,100))$ 的数据 $\leq$ 该值，且至少有 $(100-p)\%$ 的数据 $\geq$ 该值，则称该值为这组数据的第 p **百分位数**。直观来讲，一组数据的第 p 百分位数，就是将这组数据由小到大排列后，处于 $p\%$ 位置的数，例如，中位数就是一个第 50 百分位数。显然，按定义所得的第 p 百分位数可能不唯一。

为了方便，我们按如下方式确定第 p 百分位数：设一组数按照由小到大排列后为 $x_1, x_2, \cdots, x_n$，计算 $i=np\%$ 的值，如果 i 不是整数，设 i_0 为大于 i 的最小整数，取 x_{i_0} 为第 p 百分位数；如果 i 是整数，取 $\frac{x_i + x_{i+1}}{2}$ 为第 p 百分位数；特别地，规定：第 0 百分位数是 x_1（即最小值），第 100 百分位数是 x_n（即最大值）。

实际应用中，经常使用的是第 25 百分位数（简称为**第一四分位数，**也称为**下四分位数**）与第 75 百分位数（简称为**第三四分位数，**也称为**上四分位数**）

例 1：求 $1,2,3,4,5,6,7,8,9,10$ 的第 25 百分位数，第 75 百分位数和第 90 百分位数。

【解】(1) $10\times25\%=2.5$，2.5 不是整数，但大于 2.5 的最小整数为 3，故 $x_3=3$ 即为第 25 百分位数。

(2) $10\times75\%=7.5$，7.5 不是整数，但大于 7.5 的最小整数为 8，故 $x_8=8$ 即为第 75 百分位数。

(3) $10\times90\%=9$，9 整数，故 $\frac{x_9 + x_{10}}{2}=9.5$ 即为第 90 百分位数。

平均数：样本数据的算数平均数，即 $\bar{x}=\frac{1}{n}(x_1+x_2+\cdots+x_n)$

极差：一组数据的极差指的是这组数的最大值减最小值所得的差。

方差和标准方差：对于一组数据 $x_1, x_2, \cdots, x_n$，我们称

$$s^2=\frac{1}{n}[(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2]$$

为该组数据的**方差**，$\sigma=\sqrt{s^2}$ 称该组数据的**标准方差**，简称标准差。其中 $\bar{x}$ 为 $x_1,x_2,\cdots,x_n$ 的平均数。

显然，如 a,b 为常数，由方差的公式知：数据 $ax_1+b,ax_2+b,\cdots,ax_n+b$ 的方差为 a^2s^2 。

简单随机抽样：从总体中不加任何分组、划类和排队等，完全随机地抽取个体的方法叫做**简单随机抽样**。当总体中的个体之间差异程度较小和总体中个体数目较少时，常采用这种方法。常见的简单随机抽样有**抽签法**和**随机数表法**。

随机数表抽样的步骤：

(1) 对总体进行编号；

(2) 在随机数表中任意指定一个选取的位置；

(3) 按照一定的规则选取编号。例如，编号为两位，规则可以是每次从左往右选取两个数字，也可以是每次只选取每一组的前两个数字，还可以是每次只选取下面一行同一位置对应的两个数字；选取过程中，遇到超过编号范围或已经选取了的数字，应该舍取。例如：我们要从 90 个节能灯中抽取 5 个样本。采用随机数表抽样法的步骤如下

(1) 跟 90 个节能灯编号，分别为 $01,02,03,\cdots,89,90$

(2) 用随机数生成器生成一个随机数表，下面是该表的一部分。

(3) 我们规定从该表第五行第一组的第一个数字开始，每次只取每一组的前两个数字。

48628	50089	38155	69882	27761	73903	53014	98720	41571	79413
53666	08912	48395	32616	34905	63640	57931	72328	49195	17699
00620	79613	29901	92364	38659	64526	20236	29793	09063	99398
01114	19048	00895	91770	95934	31491	72529	39980	45750	14155
98246	18957	91965	13529	97168	97299	68402	68378	89201	67871
41410	51595	89983	82330	96809	93877	92818	84275	45938	48490

则抽取的结果为：$18,13,68,89,67$ 。很明显，操作过程中，我们实际选了 10 次数。

分层抽样：若总体由差异明显的几部分组成，抽样时，先将总体分成互不交叉的层，然后按照一定的比例，从各层独立地抽取一定数量的个体，再将各层取出的个体合在一起作为样本

例如：某社区有 5000 人，其中 18 岁以下的有 1000 人，19 到 60 岁的有 3000 人，61 岁以上的有 1000 人，现采用分层抽样的方式从中随机抽取 50 人，调查其对春晚的满意度，则应从 18 岁以下的人中抽取 $50\times\frac{1000}{5000}=10$ 人，从 19 到 60 岁的人中抽取 $50\times\frac{3000}{5000}=30$ 人，从 61 岁以上的人中

抽取 $50\times\frac{1000}{5000}=10$ 。

茎叶图: 先看如下例子，某赛季甲、乙两名运动员每场比赛的得分情况如下

甲：12　15　24　25　31　31　36　36　37　39　44　49　50;

乙：8　13　13　14　16　23　26　29　33　35　38　39　51.

这两组数据可以用如下的图来表示

甲		乙
	0	8
5 2	1	3 3 4 6
5 4	2	3 6 9
9 7 6 6 1 1	3	3 5 8 9
9 4	4	
0	5	1

图中，中间的数字表示两位运动员得分的十位数，两边的数字表示各自得分的个位数，这种表示数据的图叫做**茎叶图。**

一般说来：茎叶图中，所有的茎都竖着排列，而叶沿水平方向排列，茎叶图也可以只表示一组数据（只有一边有叶）。

显然，将一组数据整理成茎叶图后，如果每一行的数据都是按由小到大（或由大到小）的顺序排列的，则从中可以方便的获取这组数的最值、中位数、众数等数字特征。

例 2.为了快速了解某学校学生体重（单位：Kg）的大致情况，随机抽取 10 名学生称重，得到的数据整理成如图所示的茎叶图，请估计该校学生体重的平均数和方差。

【解】将所有数据都减去 50，得

-5，-1，　-3，-1，-4，-4，1，8，9，10

因此，这组数的平均数为 $\frac{-5-1-3-1-4-4+1+8+9+10}{10}=1$

方差为 $\frac{6^2+2^2+4^2+2^2+5^2+5^2+0^2+7^2+8^2+9^2}{10}=30.4$

故，该校学生体重的平均数为 51，方差为 30.4。

4	5 9 7 9 6 6
5	1 8 9
6	0

统计图表与频率分布直方图:（其概念在初中学过，此处略）

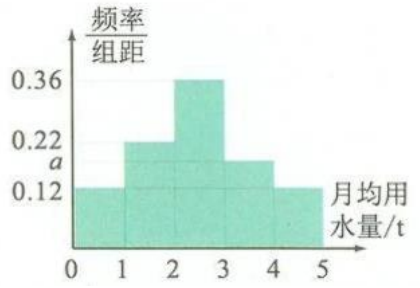
注意:频率分布直方图是以图形面积的形式反映数据落在各个小组内的频率大小。

① 频率=$\dfrac{\text{频数}}{\text{样本容量}}$　　小长方形面积=组距×$\dfrac{\text{频数}}{\text{组距}}$=频率

② 所有小长方形面积的和=各组频率和=1

③ 直方图的纵轴(小矩形的高)一般是频数除以组距的商(而不是频率)，横轴一般是数据的大小，小矩形的面积表示频率。

例 3.我国是世界上严重缺水的国家之一，某市为制定合理的节水方案，对家庭用水情况进行了调查，通过抽样，获得了某年 100 个家庭的月均用水量（单位：t），将数据按照 $[0,1),[1,2),[2,3),[3,4),[4,5]$ 分成 5 组，制成了如图所示的频率分布直方图，



（1）求图中 a 的值；

（2）设该市有 10 万个家庭，估计全市月均用水量不低于 3t 的家庭数；

（3）假设同组中的每个数据都用该组区间的中点值代替，估计全市家庭月均用水量的平均数。

【解】(1) 由于频率分布直方图中各矩形的面积和为 1，故

$(0.12+0.22+0.36+a+0.12)\times 1=1$，解得 $a=0.18$。

(2) 抽取的样本中，月均用水量不低于 3t 的家庭所占的比例为 $(a+0.12)\times 1=0.3=30\%$

因此，所求家庭数为：$100000\times 30\%=30000$

(3) 因 $0.12\times 0.5+0.22\times 1.5+0.36\times 2.5+0.18\times 3.5+0.12\times 4.5=2.46$

因此，估计全市家庭月均用水量的平均数为 2.46 t。

线性回归：一般地，已知变量 x（普通变量）与 y（随机变量）的 n 对数据 $(x_i,y_i)(i=1,2,\cdots,n)$，如果将其看成平面直角坐标系中 n 个点的坐标，且这些点紧密分布在一条直线附近，即变量 x 与 y 的关系可以近似地用一条直线来刻画，则称 x 与 y 线性相关，如果其中一个变量的值增大，另一个变量的值也增大，则称这两个变量正相关；如果一个变量的值增大，另一个变量的值反而减小，则称这两个变量负相关。为了合理描述 x 与 y 的相关性情况，我们称下面的数

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

为**相关系数** ：其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\ \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$

如果我们构造两个向量 $\vec{a}=(x_1-\bar{x},x_2-\bar{x},\cdots,x_n-\bar{x})$ ，$\vec{b}=(y_1-\bar{y},y_2-\bar{y},\cdots,y_n-\bar{y})$，则 $r=\frac{\vec{a}\bullet\vec{b}}{|\vec{a}||\vec{b}|}=\cos\alpha$ ，其中，α 为向量 $\vec{a},\vec{b}$ 的夹角。

显然，$|r|=|\cos\alpha|\le 1$，且 $|r|$ 越接近于 1，相关程度越大；$|r|$ 越接近于 0，相关程度越小；$r>0$ 表示正相关，$r<0$ 表示负相关。

回归直线方程 ：如果变量 x 与 y 线性相关，那它们之间的关系可以用一条直线来刻画，这条直线我们称为变量 x 与 y 的**线性回归直线**，对应的直线方程 $y=bx+a$ 我们称为 x 与 y 的**线性回归方程**。

如果我们现在获得了关于 x 与 y 的 n 对数据 $(x_i,y_i)(i=1,2,\cdots,n)$，怎样来获取线性回归方程 $y=bx+a$ 呢？很明显，只需确定参数 b 和 a 。

获取回归方程 $y=bx+a$ 的方法很多，中学阶段，我们利用**最小二乘法**来获取。

首先，我们将上面的数据 y_i 称为观察值，将还未得到的数据 $y_i=bx_i+a$ 称为预测值，将 $e_i=y_i-y_i$ 称为**残差**，我们的方法是：让**残差平方和** $\sum_{i=1}^{n}e_i^{\,2}=\sum_{i=1}^{n}(y_i-y_i)^2$ 最小。

不难得到：$\sum_{i=1}^{n}(y_i-y_i)^2=\sum_{i=1}^{n}[(y_i-\bar{y})-b(x_i-\bar{x})]^2+n(\bar{y}-b\bar{x}-a)^2$

显然，要让上式最小，首先需 $n(\bar{y}-b\bar{x}-a)^2=0$，也即 $a=\bar{y}-b\bar{x}$；

其次，需 $\sum_{i=1}^{n}[(y_i-\bar{y})-b(x_i-\bar{x})]^2$ 最小，易知

$\sum_{i=1}^{n}[(y_i-\bar{y})-b(x_i-\bar{x})]^2=b^2\sum_{i=1}^{n}(x_i-\bar{x})^2-2b\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})+\sum_{i=1}^{n}(y_i-\bar{y})^2$ 为关于 b 的一元二次函数，易知其最小值在 $b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ 时取得。将其带入 $a=\bar{y}-b\bar{x}$ 又可得到 a ，参数 a,b 确定后，令 $y=bx+a$ ，这便是 x 与 y 的**线性回归方程**，其中，显然，其图像经过点 $(\bar{x},\bar{y})$。

列联表与独立性检验：为了解某校高三学生是否喜欢长跑，我们随机抽取了 2 个班共 110 人进行调查，其中女生 50 人，男生 60 人，女生中有 20 人喜欢长跑，30 人不喜欢长跑，男生中有

40 人喜欢长跑，20 人不喜欢长跑，我们制成下面的表。

	喜欢长跑	不喜欢长跑	总计
女生	20	30	50
男生	40	20	60
总计	60	50	110

很明显，核心数据是中间的四个格子，这样的表，我们称其为2×2列联表。

如果用A表示事件“喜欢长跑”，B表示事件“是女生”，从上面的列联表容易得到A发生的概率的估计值为$P(A)=\frac{60}{110}=\frac{6}{11}$，$B$发生的概率的估计值为$P(B)=\frac{50}{110}=\frac{5}{11}$，“既是女生又喜欢长跑”发生的概率的估计值为$P(AB)=\frac{20}{110}=\frac{2}{11}$。需要注意的是：由于这里的$P(A),P(B),P(AB)$都是根据样本数据得到的估计值，因此，我们不能通过$P(AB)$是否等于$P(A)P(B)$来判断事件$A$与$B$是否独立。

但是，如果A与B独立，那么$P(AB)$与$P(A)P(B)$应该很接近，因此可用$P(A)P(B)$作为$P(AB)$的估计值，也就是说，喜欢长跑的女生的估计值为$110P(A)P(B)$【实际是 20 人，即$110P(AB)$】，因此，数

$$\frac{[110P(AB)-110P(A)P(B)]^2}{110P(A)P(B)}$$

应该很小，同理，考虑$\overline{A}$与B，A与$\overline{B}$，$\overline{A}$与$\overline{B}$，我们会得到下面的三个数

$$\frac{[110P(\overline{A}B)-110P(\overline{A})P(B)]^2}{110P(\overline{A})P(B)},\ \frac{[110P(A\overline{B})-110P(A)P(\overline{B})]^2}{110P(A)P(\overline{B})},\ \frac{[110P(\overline{A}\overline{B})-110P(\overline{A})P(\overline{B})]^2}{110P(\overline{A})P(\overline{B})}$$

也应很小。

我们将上面的四个数的和记为χ^2（读作“卡方”），带入相关数据，我们得到$\chi^2\approx 7.8$

在概率论中，如果A与B独立，则$\chi^2\geq 6.635$的概率只有1%，即$P(\chi^2\geq 6.635)=1\%$，这是一个小概率事件，由于我们这里算得的$\chi^2=7.8\geq 6.635$，因此，如果A与B独立（即“喜欢长跑”与“是女生”对立），那么我们就观察到了一个小概率事件（概率不超过1%），换句话说：我们有99%的把握认为“喜欢长跑”与“是女生”不独立，也就是说：我们有99%的把握认为“喜欢长跑”与性别有关。

上面的1%称为显著性水平，与之对应的6.635称为显著性水平1%所对应的分位数。

一般地，可以用类似的方法检验两个随机事件是否独立。

如果随机事件 A 和 B 的样本数据的 2×2 列联表如下：

	A	$\overline{A}$	总计
B	a	b	$a+b$
$\overline{B}$	c	d	$c+d$
总计	$a+c$	$b+d$	$a+b+c+d$

记 $n=a+b+c+d$，则

$$\chi^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

如果随机变量 A,B 相互独立，一般说来，上面的数 χ^2 都不会太大。另一方面，对于给定的数 α（即显著性水平，通常取 $0.05,\ 0.01$ 等），可以找到满足条件 $P(\chi^2\geq K)=\alpha$ 的 K（即显著性水平对应的分位数）。如果我们根据样本数据算得的 $\chi^2\geq K$ 成立，就称在犯错误的概率不超过 α 的前提下，认为 A,B 不对立（即 A 与 B 有关）；或则说：我们有 $1-\alpha$ 的把握认为 A 与 B 有关。如果算得的 $\chi^2<K$，就称不能得到前面的结论。这一过程称为独立性检验。

例 4.为了了解阅读量多少与幸福感强弱之间的关系，一调查机构得到了如下的调查数据

	幸福感强	幸福感弱	总计
阅读量多	54	18	72
阅读量少	36	42	78
总计	90	60	150

根据调查数据回答：在犯错误的概率不超过 1% 的前提下，可以认为阅读量多少与幸福感强弱有关吗?

【解】由题意知：$\chi^2=\dfrac{150(54\times42-18\times36)^2}{72\times78\times79\times60}=\dfrac{675}{52}\approx12.981$，

又，查表得 $P(\chi^2\geq6.635)=0.01$，而 $12.981>6.635$，

故，在犯错误概率不超过 1% 的前提下，可以认为阅读量多少与幸福感强弱有关。

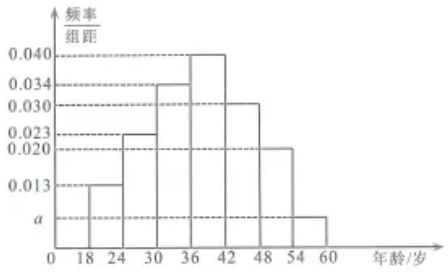
13.3.2 典型例题

例 1（1）为了解某地区的中小学生的视力情况，拟从该地区的中小学生中抽取部分学生进行调查，事先已了解到该地区小学、初中、高中三个学段学生的视力情况有较大差异，而男女生视力情况差异不大．在下面的抽样方法中，最合理的抽样方法是(　　)．

A．简单随机抽样　　B．按性别分层抽样　　C．按学段分层抽样　　D．系统抽样

（2）右图为某地新冠病毒疫苗接种年龄样本的频率分布直方图，估计接种年龄的中位数为（　）

A. 40　　B. 39　　C. 38　　D. 40



【解】（1）题目中出现了小学、初中、高中学段，不同学段的学生，近视程度肯定不一样，因此，因此应采用分层抽样。选 C。

（2）易知前三个小矩形的面积为 $6(0.013+0.023+0.034)=0.42$，

令 $0.040x=0.08$，得 $x=2$，因此，中位数为 $36+2=38$，选 C。

例 2.下列说法错误的是(　　)

A．在回归模型中，预报变量 y 的值不能由解释变量 x 唯一确定

B．若变量 x,y 满足关系 $y=-0.1x+1$，且变量 y 与 z 正相关，则 x 与 z 也正相关

C．在残差图中，残差点分布的带状区域的宽度越狭窄，其模型拟合的精度越高

D．以模型 $y=ce^{kx}$ 去拟合一组数据时，为了求出回归方程，设 $z=\ln y$，将其变换后得到线性方程 $z=0.3x+4$，则 $c=e^4,k=0.3$。

【解】对于 A，在回归模型中，预报变量 y 的值由解释变量 x 和随机误差 e 共同确定，即 x 只能解释部分 y 的变化，所以 A 正确；

对于 B，由回归方程知变量 y 与 z 正相关，则 x 与 z 负相关，所以 B 错误；

对于 C，在残差图中，残差点分布的带状区域的宽度越狭窄，其模型拟合的精度越高，C 正确；

由回归分析的意义知 D 正确.

综上，选 B.

例 3（1）若样本数据 $x_1,x_2,\cdots,x_{10}$ 的标准差为 8，则数据 $2x_1-1,2x_2-1,\cdots,2x_{10}-1$ 的标准差为(　　)

A.8　　B.15　　C.16　　D.32

（2）设 X 为随机变量，则 $D(\dfrac{X-EX}{\sqrt{DX}})=$_____

【解】⑴由题意知 $D(X)=64$，故 $D(2X-1)=4D(X)=4\times64$

故 $\sqrt{D(2X-1)}=\sqrt{4\times64}=16$，选 C。

⑵因 $\frac{X-EX}{\sqrt{DX}}=\frac{X}{\sqrt{DX}}-\frac{EX}{\sqrt{DX}}$，所以 $D(\frac{X-EX}{\sqrt{DX}})=(\frac{1}{\sqrt{DX}})^2\times DX=1$。

例 4．重庆市 2020 年各月的平均气温(℃)数据的茎叶图如下：

0	8	9			
1	2	5	8		
2	0	0	3	3	8
3	1	2			

则这组数据的中位数是(　　)

A.19　　B.20　　C.21.5　　D.23

【解】从茎叶图知：所有数据由小到大排列为：

8，9，12，15，18，20，20，23，23，28，31，32，

一共 12 个数，故中位数为 $\frac{x_6+x_7}{2}=20$，选 B

例 5．为了解某社区居民的家庭年收入与年支出的关系，随机调查了该社区 5 户家庭，得到如下统计数据表：

收入 x (万元)	8.2	8.6	10.0	11.3	11.9
支出 y (万元)	6.2	7.5	8.0	8.5	9.8

根据上表可得回归直线方程 $y=\hat{b}x+a$，其中 $\hat{b}=0.76,\ a=\bar{y}-\hat{b}\bar{x}$.据此估计，该社区一户年收入为 15 万元家庭的年支出为(　　)

A.11.4 万元　　B.11.8 万元　　C.12.0 万元　　D.12.2 万元

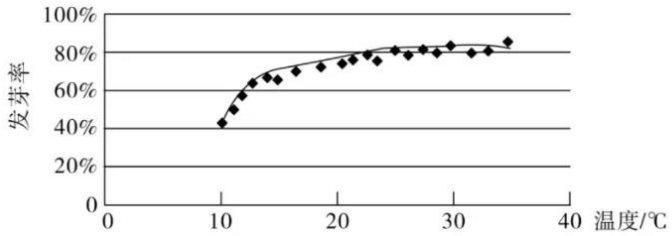
【解】易知 $\bar{x}=10,\ \bar{y}=8$，又因 $\hat{b}=0.76$，

故 $a=\bar{y}-\hat{b}\bar{x}=8-0.76\times10=0.4$，故回归方程为：$y=0.76x+0.4$，

故，$x=15$时，$y=0.76\times15+0.4=11.8$（万元）

故选 B.

例 6．某校一个课外学习小组为研究某作物种子的发芽率 y 与温度 x（单位：^0C）的关系，在 20 个不同的温度条件下进行种子发芽实验，由实验数据 $(x_i,y_i)(i=1,2,\cdots,20)$ 得到下面的散点图



由此散点图，在10^0C至40^0C之间，下面四个回归方程类型中最适宜作为发芽率y和温度x的回归方程类型的是（　　）

A．$y=a+bx$　　B．$y=a+bx^2$　　C．$y=a+be^x$　　D．$y=a+b\ln x$

【解】很明显，这些散点集中在某一条对数函数的图像附近，故选 D。

例 7（1）根据如下样本数据

x	3	4	5	6	7	8
y	4.0	2.5	−0.5	0.5	−2.0	−3.0

得到的回归方程为$y=bx+a$，则(　　)

A. $a>0,b>0$　　B. $a>0,b<0$　　C. $a<0,b>0$　　D. $a<0,b<0$

（2）判断正误：离散型随机变量的数学期望和方差都是一个数值，它们不随试验的结果而变化（　　）

【解】（1）从表格数据知：随着x的增大，y在减小，因此x,y负相关，从而知：$b<0$；易知：$\bar{x}=5.5,\ \bar{y}=0.3$，因回归直线过点$(\bar{x},\bar{y})$，故$0.3=5.5b+a$，考虑到$b<0$，所以$a>0$。选 B。

（2）正确。随机变量的数学期望和方差都是客观存在的常数，不随抽样样本的变化而变化；而样本均值和样本方差都是随机变量，都随试验的变化而变化。注意二者之间的关系。

例 8.对一个容量为N的总体抽取容量为n的样本，当选取简单随机抽样、系统抽样和分层抽样三种不同方法抽取样本时，总体中每个个体被抽中的概率分别为p_1,p_2,p_3，则(　　)

A. $p_1=p_2<p_3$　　B. $p_2=p_3<p_1$　　C. $p_1=p_3<p_2$　　D. $p_1=p_2=p_3$

【解】因为采取简单随机抽样、系统抽样和分层抽样时，总体中每个个体被抽中的概率相等，故选 D

例 9.“大众创业，万众创新”是李克强总理在政府工作报告中向全国人民发出的口号．某生产企业积极响应号召，大力研发新产品，为了对新研发的一批产品进行合理定价，将该产品按事先拟定的价格进行试销，得到一组销售数据(x_i,y_i)（$i=1,2,\cdots,6$），如表所示：

试销单价x（元）	4	5	6	7	8	9

产品销量 y（件）	q	84	83	80	75	68

已知 $\overline{y}=\frac{1}{6}\sum_{i=1}^{6}y_i=80$.

（Ⅰ）求出 q 的值;

（Ⅱ）已知变量 x，y 具有线性相关关系，求产品销量 y（件）关于试销单价 x（元）的线性回归方程 $y=\hat{b}x+a$;

（Ⅲ）用 y_i 表示用（Ⅱ）中所求的线性回归方程得到的与 x_i 对应的产品销量的估计值．当销售数据 (x_i,y_i) 对应的残差的绝对值 $|y_i-y_i|\leq 1$ 时，则将销售数据 (x_i,y_i) 称为一个“好数据”．现从 6 个销售数据中任取 3 个，求“好数据”个数 Z 的分布列和数学期望 $E(Z)$.

（参考公式：线性回归方程中 $\hat{b}$，a 的最小二乘估计分别为 $\hat{b}=\frac{\sum_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}$，$a=\overline{y}-\hat{b}\overline{x}$）

【解】（Ⅰ）由 $\overline{y}=\frac{1}{6}(q+84+83+80+75+68)=80$，得 $q=90$.

（Ⅱ）易知 $\overline{x}=6.5$，又 $\overline{y}=80$，故 $\hat{b}=\frac{\sum_{i=1}^{6}x_iy_i-n\overline{x}\,\overline{y}}{\sum_{i=1}^{6}x_i^2-n(\overline{x})^2}=\frac{3050-6\times6.5\times80}{271-253.5}=-\frac{70}{17.5}=-4$

所以，$a=\overline{y}-\hat{b}\overline{x}=80+4\times6.5=106$，

故，所求的线性回归方程为 $y=-4x+106$.

（Ⅲ）由（Ⅱ）知 $y=-4x+106$，故

$y_1=90$，$y_2=86$，$y_3=82$，$y_4=78$，$y_5=74$，$y_6=70$.

显然，满足 $|y_i-y_i|\leq 1$（$i=1,2,\cdots,6$）的共有 3 个“好数据”：$(4,90)$、$(6,83)$、$(8,75)$.

于是 Z 的所有可能取值为 0，1，2，3.

$P(Z=0)=\frac{C_3^3}{C_6^3}=\frac{1}{20}$；$P(Z=1)=\frac{C_3^1C_3^2}{C_6^3}=\frac{9}{20}$；$P(Z=2)=\frac{C_3^2C_3^1}{C_6^3}=\frac{9}{20}$；

$P(Z=3)=\dfrac{C_3^3}{C_6^3}=\dfrac{1}{20}$，

$\therefore Z$ 的分布列为：

Z	0	1	2	3
P	$\dfrac{1}{20}$	$\dfrac{9}{20}$	$\dfrac{9}{20}$	$\dfrac{1}{20}$

于是 $E(Z)=0\times\dfrac{1}{20}+1\times\dfrac{9}{20}+2\times\dfrac{9}{20}+3\times\dfrac{1}{20}=\dfrac{3}{2}$.

例 10.随着移动支付的普及，中国人的生活方式正在悄然发生改变，带智能手机而不带钱包出门，渐渐成为中国人的新习惯。2020 年我国的移动支付迅猛增长，据统计，某平台 2020 年移动支付的笔数占总支付笔数的 80%。

（Ⅰ）从该平台的 2020 年的所有支付中任取 10 笔，求移动支付笔数的期望和方差；

（Ⅱ）现有 500 名使用移动支付平台的用户，其中 300 名是城市用户，200 名是农村用户，调查他们 2020 年个人移动支付的比例是否达到 80%，得到 2×2 列联表如下：

	个人移动支付比例达到了 80%	个人移动支付比例未达到 80%	合计
城市用户	270	30	300
农村用户	170	30	200
合计	440	60	500

根据上表数据，问是否有 95% 的把握认为 2020 年个人支付比例达到了 80% 与该用户是否是城市用户还是农村用户有关？

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

$P(\chi^2\geq K)$	0.050	0.010
K	3.841	6.635

【解】（Ⅰ）设所取 10 笔支付中，移动支付的笔数为 X，由题意知：某一笔支付为移动支付的概率为 0.8，故

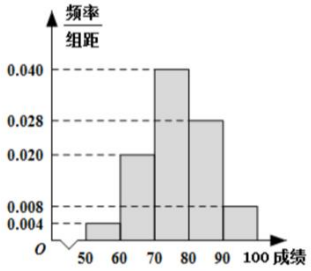
$X\sim B(10,0.8)$，所以 $EX=10\times0.8=8, DX=10\times0.8\times0.2=1.6$.

（Ⅱ）因为 $\chi^2=\dfrac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}=\dfrac{500\times(270\times30-170\times30)^2}{440\times60\times300\times200}\approx2.841<3.841$

所以没有 95%的把握认为 2020 年个人移动支付比例达到了 80%与该用户是否是城市用户还

是农村用户有关.

例 11. 某大学有国防生 50 名，学校在关注国防生文化素养的同时也非常注重他们的身体素质，要求每月至少参加一次野营拉练活动(下面简称“活动”)并记录成绩。10 月份某次活动中同学们的成绩统计如图所示:



⑴根据图表，估算学生在活动中取得成绩的中位数(精确到 0.1);

⑵根据成绩从[50,60)、[90,100)两组学员中任意选出两人为一组，若选出成绩分差大于 10，则称该组为“帮扶组”，试求选出两人为“帮扶组”的概率.

【解】（1）成绩在区间$[50,60)$的频率为$10\times0.004=0.04$，

成绩在区间$[60,70)$的频率为$10\times0.02=0.2$，

$0.5-0.04-0.2=0.26$，设中位数为x，

则$(x-70)\times0.04=0.26, x-70=\dfrac{0.26}{0.04}=6.5$，解得$x=76.5$

(2) 成绩在$[50,60)$和$[90,100)$的人数分别为$0.04\times50=2$，$0.008\times10\times50=4$

因此，从两组学员中选出2人，其成绩分差大于10（也即该两组为“帮扶组”）的概率为$\dfrac{C_2^1C_4^1}{C_6^2}=\dfrac{8}{15}$

例 12.为帮助乡村脱贫，某勘探队计划了解当地矿脉某金属的分布情况，测得了平均金属含量y (单位:g/m)与样本对原点的距离x (单位:m)的数据，并作了初步处理，得到了下面的一些统计理的值.（表中$u_i=\dfrac{1}{x_i}, \bar{u}=\dfrac{1}{9}\sum_{i=1}^{9}u_i$）

$\bar{x}$	$\bar{y}$	$\bar{u}$	$\sum_{i=1}^{9}(x_i-\bar{x})^2$	$\sum_{i=1}^{9}(u_i-\bar{u})^2$	$\sum_{i=1}^{9}(y_i-\bar{y})^2$	$\sum_{i=1}^{9}(x_i-\bar{x})(y_i-\bar{y})$	$\sum_{i=1}^{9}(u_i-\bar{u})(y_i-\bar{y})$
6	97.90	0.21	60	0.14	14.12	26.13	−1.40

⑴利用样本相关系数的知识，判断$y=a+bx$与$y=c+\dfrac{d}{x}$哪一个更适宜作为平均金属含量y关于样本对原点的距离x的回归方程类型?

⑵根据⑴的结果回答下列问题:

⑴建立 y 关于 x 的回归方程;

⑵ 样本对原点的距离 $x=20$ 时，金属含量的预报值是多少?

⑶ 已知该金属在距离原点 x**m** 时的平均开采成本 W（单位:元）与 x,y 关系为 $W=1000\left(y-\ln x\right)\left(1\le x\le 100\right)$，根据（2）的结果回答， x 为何值时，开采成本最大?

附：对于一组数据 $\left(t_1,s_1\right),\left(t_2,s_2\right),\cdots,\left(t_n,s_n\right)$，其线性相关系数 $r=\dfrac{\sum_{i=1}^{n}\left(t_i-\bar{t}\right)\left(s_i-\bar{s}\right)}{\sqrt{\sum_{i=1}^{n}\left(t_i-\bar{t}\right)^2\sum_{i=1}^{n}\left(s_i-\bar{s}\right)^2}}$

其回归直线 $s=\alpha+\beta t$ 的斜率和截距的最小二乘估计分别为： $\beta=\dfrac{\sum_{i=1}^{n}\left(t_i-\bar{t}\right)\left(s_i-\bar{s}\right)}{\sum_{i=1}^{n}\left(t_i-\bar{t}\right)^2},\alpha=\bar{s}-\beta\bar{t}$

【解】（1）由题意知： $\left(x_i,y_i\right)$ 的线性相关系数为 $r_1=\dfrac{\sum_{i=1}^{9}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{9}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{9}\left(y_i-\bar{y}\right)^2}}\approx 0.898$，

$\left(u_i,y_i\right)$ 的线性相关系数为 $r_2=\dfrac{\sum_{i=1}^{9}\left(u_i-\bar{u}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{9}\left(u_i-\bar{u}\right)^2\sum_{i=1}^{9}\left(y_i-\bar{y}\right)^2}}\approx -0.996$

因为 $\left|r_1\right|<\left|r_2\right|$，故 $y=c+\dfrac{d}{x}$ 更适宜作为平均金属含量 y 关于样本对原点的距离 x 的回归方程类型

（i）令 $u=\dfrac{1}{x}$，先建立 y 关于 u 的线性回归方程，由于 $d=\dfrac{\sum_{i=1}^{9}\left(u_i-\bar{u}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^{9}\left(u_i-\bar{u}\right)^2}=\dfrac{-1.4}{0.14}=-10$

$c=\bar{y}-d\bar{u}=97.90+10\times 0.21=100$

所以 y 关于 u 的线性回归方程 $y=100-10u$，

因此 y 关于 x 的线性回归方程 $y=100-\dfrac{10}{x}$

（ii）由（i）知， $x=20$ 时，金属含量的预报值是 $y=100-\dfrac{10}{20}=99.5$

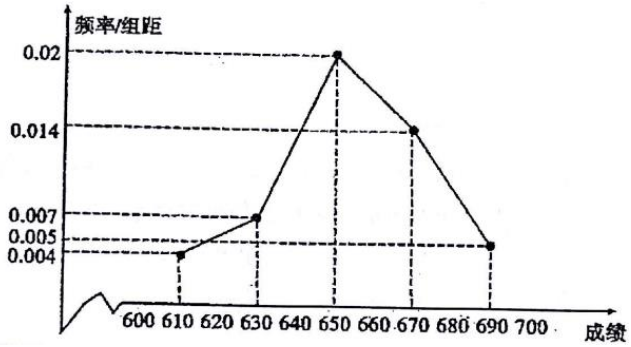
(3) 由（2）知，平均开采成本

$$W=W(x)=1000(y-\ln x)=1000\left(100-\frac{10}{x}-\ln x\right)(1\le x\le 100)$$

故$W'(x)=1000\left(\frac{10}{x^2}-\frac{1}{x}\right)=1000\cdot\frac{10-x}{x^2}(1\le x\le 20)$

很明显，$x=10$为$W(x)$的最大值点，故$x=10$时，开采成本最大。

例 13.某中学在 2020 年高考分数公布后对高三年级各班的成绩进行分析。经统计，某班有 50 名同学，总分都在区间$(600,700]$内，将得分区间平均分成 5 组，统计频数、频率后得到了如图所示的频率分布折线图。



（1）请根据频率分布折线图画出频率分布直方图；并根据频率分布直方图估计该班级的平均分；

（2）经过相关部门的计算，本次高考总分大于等于 680 可以获得高校 T 的“强基计划”入围资格。高校 T 的“强基计划”校考分为两轮。第一轮为笔试，所有入围同学都要参加，考试科目为数学和物理，每科的笔试成绩由高到低，依次有A^+,A,B,C四个等级。两科中至少有一科得A^+，且两科均不低于B，才能进入第二轮，第二轮得到“通过”的同学将被 T 高校提前录取。已知入围的同学参加第一轮笔试时，总分高于 690 分的同学在每科笔试中取得A^+,A,B,C的概率$\frac{2}{3},\frac{1}{6},\frac{1}{12},\frac{1}{12}$；总分不超过 690 分的同学在每科笔试中取得$A^+,A,B,C$的概率分别为$\frac{1}{3},\frac{1}{4},\frac{1}{6},\frac{1}{4}$；进入第二轮的同学，若两科笔试成绩均为$A^+$，则免面试，并被高校 T 提前录取；若两科笔试成绩只有一个A^+，则参加面试，总分高于 690 的同学面试“通过”的概率为$\frac{2}{3}$，总分不超过 690 的同学面试“通过”的概率为$\frac{2}{5}$，面试“通过”的同学也将被高校 T 提前录取。若该班级 680 分以上的同学都报考了高校 T 的“强基计划”，且恰有 2 人成绩高于 690 分，求

①总分高于 690 分的某位同学没有进入第二轮的概率 p_1

②该班恰有 2 名同学通过“强基计划”被高校 T 提前录取的概率 p_2

【解】（1）频率分布直方图如图所示，平均分

$$\bar{x}=(610\times0.004+630\times0.007+650\times0.02+670\times0.014+690\times0.005)\times20=653.6$$

（2）总分大于等于 680 分的同学有 $50\times0.005\times20=5$ 人，

由已知，其中有 3 人小于等于 690 分，2 人大于 690 分；

① $P_1=1-P\left(A^+A^++A^+A+A^+B\right)=1-\left(\frac{2}{3}\right)^2-C_2^1\times\frac{2}{3}\times\frac{1}{6}-C_2^1\times\frac{2}{3}\times\frac{1}{12}=\frac{2}{9}$，

②设高于 690 分的同学被高校 T 提前录取为事件 M，不超过 690 分的同学被高校 T 提前录取为事件 N

则 $P(M)=P\left(A^+A^+\right)+\frac{2}{3}P\left(A^+A+A^+B\right)=\left(\frac{2}{3}\right)^2+\frac{2}{3}\left(C_2^1\times\frac{2}{3}\times\frac{1}{6}+C_2^1\times\frac{2}{3}\times\frac{1}{12}\right)=\frac{2}{3}$,

$P(N)=P\left(A^+A^+\right)+\frac{2}{5}P\left(A^+A+A^+B\right)=\left(\frac{1}{3}\right)^2+\frac{2}{5}\left(C_2^1\times\frac{1}{3}\times\frac{1}{4}+C_2^1\times\frac{1}{3}\times\frac{1}{6}\right)=\frac{2}{9}$,

$p_2=\left(\frac{2}{3}\right)^2\times\left(\frac{7}{9}\right)^3+\left(\frac{1}{3}\right)^2\times C_3^2\left(\frac{7}{9}\right)\times\left(\frac{2}{9}\right)^2+C_2^1\times\frac{1}{3}\times\frac{2}{3}\times C_3^2\times\frac{2}{9}\times\left(\frac{7}{9}\right)^2=\frac{2632}{6561}$

例 14. 新型冠状病毒的传染主要是人与人之间进行传播，感染人群年龄大多数是 50 岁以上人群。该病毒进入人体后有潜伏期，潜伏期是指病原体侵入人体至最早出现临床症状的这段时间，潜伏期越长，感染到他人的可能性越高。现对 400 个病例的潜伏期(单位：天)进行调查，统计发现潜伏期平均数为 7.2，方差为 2.25^2。如果认为超过 8 天的潜伏期属于“长潜伏期”，按照年龄统计样本，得到下面的列联表：

年龄/人数	长期潜伏	非长期潜伏
50 岁以上	60	220
50 岁及 50 岁以下	40	80

⑴是否有 95%的把握认为“长期潜伏”与年龄有关；

⑵假设潜伏期 X 服从正态分布 $N\left(\mu,\sigma^2\right)$，其中 μ 近似为样本平均数 $\bar{x}$，σ^2 近似为样本方差 s^2。

(i)现在很多省市对入境旅客一律要求隔离 14 天，请用概率的知识解释其合理性；

(ii)以题目中的样本频率估计概率，设 1000 个病例中恰有$k\left(k\in N^*\right)$个属于“长期潜伏”的概率是$p\left(k\right)$，当k为何值时，$p\left(k\right)$取得最得最大值。

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

$P\left(K^2\geq k_0\right)$	0.1	0.05	0.010
k_0	2.706	3.841	6.635

若$\xi\sim N\left(\mu,\sigma^2\right)$，则$P\left(\mu-\sigma<\xi<\mu+\sigma\right)=0.6862,P\left(\mu-2\sigma<\xi<\mu+2\sigma\right)=0.9544$，$P\left(\mu-3\sigma<\xi<\mu+3\sigma\right)=0.9974$

【解】（1）依题意有$K^2=\dfrac{400\left(60\times80-220\times40\right)^2}{280\times120\times100\times300}\approx6.35$，

由于$6.35>3.841$，故有 95%的把我认为“长期潜伏”与年龄有关；

（2）（i）若潜伏期$X\sim N\left(7.2,2.25^2\right)$，由$P\left(X\geq13.95\right)=\dfrac{1-0.9974}{2}=0.0013$，

得知潜伏期超过 14 天的概率很低，因此隔离 14 天是合理的；

（ii）由于 400 个病例中有 100 个属于长期潜伏期，若以样本频率估计概率，一个患者属于“长潜伏期”的概率为$\dfrac{1}{4}$，

于是$p\left(k\right)=C_{1000}^k\cdot\left(\dfrac{1}{4}\right)^k\cdot\left(\dfrac{3}{4}\right)^{1000-k}$，

则$\dfrac{p\left(k\right)}{p\left(k-1\right)}=\dfrac{C_{1000}^k\cdot\left(\dfrac{1}{4}\right)^k\cdot\left(\dfrac{3}{4}\right)^{1000-k}}{C_{1000}^{k-1}\cdot\left(\dfrac{1}{4}\right)^{k-1}\cdot\left(\dfrac{3}{4}\right)^{1001-k}}=\dfrac{C_{1000}^k}{3C_{1000}^{k-1}}=\dfrac{1}{3}\cdot\dfrac{\left(k-1\right)!\left(1001-k\right)!}{k!\left(1000-k\right)!}=\dfrac{1}{3}\left(\dfrac{1001}{k}-1\right)$

当$0<k<\dfrac{1001}{4}$时，$\dfrac{p\left(k\right)}{p\left(k-1\right)}>1$；当$\dfrac{1001}{4}<k<1000$时，$\dfrac{p\left(k\right)}{p\left(k-1\right)}<1$

$\therefore p\left(1\right)<p\left(2\right)<\cdots<p\left(250\right)$，$p\left(250\right)>p\left(251\right)>\cdots>p\left(1000\right)$

故当$k=250$时，$p\left(k\right)$取得最大值。