

---

## 第二章 统计学

## 第二章 统计学

### 总复习题

1. 【2022 全国乙卷理 19】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:  $\text{m}^2$ ) 和材积量 (单位:  $\text{m}^3$ ), 得到如下数据:

| 样本号 $i$      | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   |
|--------------|------|------|------|------|------|------|------|------|------|------|
| 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 |
| 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 |

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

- (1) 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);
- (2) 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  $186 \text{ m}^2$ . 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.

参考数据:  $\sqrt{1.896} \approx 1.377$

2. 【2018 全国 II 卷理 18】下图是某地区 2000 年至 2016 年环境基础设施投资额  $y$  (单位: 亿元) 的折线图. 为了预测该地区 2018 年的环境基础设施投资额, 建立了  $y$  与时间变量  $t$  的两个线性回归模型. 根据 2000 年至 2016 年的数据 (时间变量  $t$  的值依次为  $1, 2, \dots, 17$ ) 建立模型 (a):  $\hat{y} = -30.4 + 13.5t$ ; 根据 2010 年至 2016 年的数据 (时间变量  $t$  的值依次为  $1, 2, \dots, 7$ ) 建立模型 (b):  $\hat{y} = 99 + 17.5t$ . 你认为用哪个模型得到的预测值更可靠? 并说明理由.
3. 【2017 全国 I 卷理 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:  $\text{cm}$ ). 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  $P(X \geq 1)$  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.

- 试说明上述监控生产过程方法的合理性;
- 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

$$\text{经计算得 } \bar{x} = \frac{1}{16} \sum_{i=1}^{16} x_i = 9.97, s = \sqrt{\frac{1}{16} \sum_{i=1}^{16} (x_i - \bar{x})^2} = \sqrt{\frac{1}{16} \left( \sum_{i=1}^{16} x_i^2 - 16\bar{x}^2 \right)} \approx 0.212, \text{ 其中 } x_i \text{ 为抽取的第 } i \text{ 个零件的尺寸, } i = 1, 2, \dots, 16. \text{ 用样本平均数 } \bar{x} \text{ 作为 } \mu$$

的估计值  $\hat{\mu}$ , 用样本标准差  $s$  作为  $\sigma$  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  $\mu$  和  $\sigma$  (精确到 0.01) .

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

4. 【2023 新高考 II 卷 19】某研究小组经过研究发现某种疾病的患病者与未患病者的某项医学指标有明显差异, 经过大量调查, 得到如下的患病者和未患病者该指标的频率分布直方图: 利用该指标制定一个检测标准, 需要确定临界值  $c$ , 将该指标大于  $c$  的人判定为阳性, 小于或等于  $c$  的人判定为阴性. 此检测标准的漏诊率是将患病者判定为阴性的概率, 记为  $p(c)$ ; 误诊率是将未患病者判定为阳性的概率, 记为  $q(c)$ . 假设数据在组内均匀分布. 以事件发生的频率作为相应事件发生的概率.

- (1) 当漏诊率  $p(c) = 0.5\%$  时, 求临界值  $c$  和误诊率  $q(c)$ ;
- (2) 设函数  $f(c) = p(c) + q(c)$ . 当  $c \in [95, 105]$  时, 求  $f(c)$  的解析式, 并求  $f(c)$  在区间  $[95, 105]$  的最小值.

5. 【2024 北京 18】某保险公司为了了解该公司某种保险产品的索赔情况, 从合同保险期限届满的保单中随机抽取 1000 份, 记录并整理这些保单的索赔情况, 获得数据如下表:

| 索赔次数 | 0   | 1   | 2  | 3  | 4  |
|------|-----|-----|----|----|----|
| 保单份数 | 800 | 100 | 60 | 30 | 10 |

假设一份保单的保费为 0.4 万元; 前三次索赔时, 保险公司每次赔偿 0.8 万元; 第四次索赔时, 保险公司赔偿 0.6 万元. 假设不同保单的索赔次数相互独立, 用频率估计概率.

- (1) 估计一份保单索赔次数不少于 2 的概率;
  - (2) 一份保单的毛利润定义为这份保单的保费与赔偿总金额之差.
    - i. 记  $X$  为一份保单的毛利润, 估计  $X$  的数学期望  $EX$ ;
    - ii. 如果无索赔的保单的保费减少 4%, 有索赔的保单的保费增加 20%, 试比较这种情况下一份保单毛利润的数学期望估计值与 (i) 中  $EX$  估计值的大小.
6. 【2019 北京理 17】为了解某校学生上个月  $A, B$  两种移动支付方式的 usage 情况, 从全校学生中随机抽取了 100 人, 发现样本中  $A, B$  两种支付方式都不使用的有 5 人, 样本中仅使用  $A$  和仅使用  $B$  的学生的支付金额分布情况如下:

| 支付方式   | 支付金额 (元)  |              |         |
|--------|-----------|--------------|---------|
|        | (0, 1000] | (1000, 2000] | 大于 2000 |
| 使用 $A$ | 18 人      | 9 人          | 3 人     |
| 使用 $B$ | 10 人      | 14 人         | 1 人     |

已知上个月样本学生的支付方式在本月没有变化. 现从样本仅使用  $A$  的学生中, 随机抽查 3 人, 发现他们本月的支付金额都大于 2000 元. 根据抽查结果, 能否认为样本仅使用  $A$  的学生中本月支付金额大于 2000 元的人数有变化? 说明理由.

7. 【2022 新高考 I 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），同时从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 |

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  $R$ .

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

(2) 利用该调查数据，给出  $P(A|B), P(A|\bar{B})$  的估计值，并利用 (1) 的结果给出  $R$  的估计值.

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

11. (1)

$$\begin{aligned}
 R &= \frac{P(B|A)}{P(B|\bar{A})} \cdot \frac{P(\bar{B}|A)}{P(\bar{B}|\bar{A})} = \frac{P(B|A)P(A)}{P(B|\bar{A})P(\bar{A})} \cdot \frac{P(\bar{B}|A)P(A)}{P(\bar{B}|\bar{A})P(\bar{A})} \\
 &= \frac{P(AB)}{P(\bar{A}B)} \cdot \frac{P(A\bar{B})}{P(\bar{A}\bar{B})} = \frac{P(A|B)}{P(\bar{A}|B)} \cdot \frac{P(\bar{A}|\bar{B})}{P(A|\bar{B})}.
 \end{aligned}$$

(2) 由该调查数据得, 病例组中卫生习惯不够良好的频率为  $\frac{40}{100} = 0.4$ , 对照组中卫生习惯不够良好的频率为  $\frac{10}{100} = 0.1$ , 所以  $P(A|B)$  的估计值为 0.4,  $P(A|\bar{B})$  的估计值为 0.1.  $P(\bar{A}|B)$  的估计值为 0.6,  $P(\bar{A}|\bar{B})$  的估计值为 0.9, 利用 (1) 的结果可得  $R$  的估计值为  $\frac{0.4}{0.6} \times \frac{0.9}{0.1} = 6$ .

评: 此题释放了一个明确的信号: 概率统计里面的公式推导非常重要. 尤其是条件概率在高中的要求提升了, 增加了全概率公式和贝叶斯公式。

实测数据: (福建) 阅卷数据如下:

| 类别        | 平均分  | 难度   | 区分度  | 零分率  |
|-----------|------|------|------|------|
| 物理组 20(2) | 1.51 | 0.19 | 0.41 | 0.59 |
| 历史组 20(2) | 0.76 | 0.10 | 0.23 | 0.75 |
| 总体 20(2)  | 1.29 | 0.16 | 0.37 | 0.64 |

---

## 第二章 统计学

---

## 第二章 统计学

---

## 第二章 统计学