HomeWork 3.3

邹翔宇 | 2410833001

第一题

1.数据的样本量是多少?（5 分）

样本量是指观测的个体数量。从表格中可以看到编号从 1 到 203，所以数据的样本量是 203

2.数据包含多少个变量?（5 分）

数据包含 G、X_1、X_2、X_3、X_4、X_5、X_6、X_7、X_8、X_9、X_{10}、X_{11}、X_{12}、X_{13}、X_{14}，共 15 个变量。

3.辨识数据类型，指出哪些是计量，计数，等级数据。（10 分）

- 计量数据: $X_2, X_5 - X_{11}$
- 计数数据: G, X_1, X_3, X_4, X_{14}
- 等级数据: X_{12}, X_{13}

第二题

某超市记录了连续 5 天的某种商品销售量（单位：件）如下：5, 5, 5, 10, 10。请计算以下统计量:

1. 算数平均数

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$\frac{5+5+5+10+10}{5} = 7$$

2. 几何平均数

$$G = \sqrt[n]{x_1 * x_2 * \ldots * x_n}$$

$$\sqrt[5]{5*5*5*10*10} \approx 6.60$$

3. 调和平均数

$$H = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

$$\sum_{i=1}^{n}\left(\frac{1}{x_i}\right) = \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{10} + \frac{1}{10} = 0.8$$

$$H = \frac{5}{0.8} = 6.25$$

4. 中位数

首先将数据从小到大排序：5,5,5,10,10
中位数为 5.0

5. 众数

众数是一组数据中出现次数最多的数据值。在 5,5,5,10,10 中，5 出现了 3 次，10 出现了 2 次，5 出现的次数最多。所以众数是 5.00。

第三题

1. 某班级学生的身高

- 适合用算术均数。理由：这组数据是连续型定量数据，且数据分布相对均匀，没有极端值的干扰，算术均数能很好地反映这组数据的平均水平

2. 某公司员工的月收入

- 适合用中位数。理由：数据中存在 `100000` 这样的极端值（异常值），如果使用算术均数，会受到这个极端值的较大影响，不能准确反映员工月收入的一般水平。而中位数是将数据排序后位于中间位置的数值，不受极端值的影响，能更合理地表示该组数据的集中趋势。

3. 某城市连续 5 天的气温变化率

- 适合用几何均数。理由：气温变化率是比率数据，各数据之间存在连乘关系（如计算一段时间内的总变化率），几何均数的计算公式为能更好地反映这类比率数据的平均变化情况。

4. 某超市某种商品的单价

- 适合用众数。理由：从数据中可以看出，`20` 出现的次数最多，众数是一组数据中出现次数最多的数据值，所以众数能直观地反映出该商品最常见的单价。

5. 某路段车辆的行驶速度

- 适合用算术均数。理由：这组数据是连续型定量数据，虽然数据有一定差异，但没有明显的极端值对整体数据产生极大干扰，算术均数能综合反映车辆行驶速度的平均水平。

第四题

1. 总体方差 CV

- 首先计算总体均数μ：
  - 总体均数$\mu = \sum(i, 1, n)\frac{x_i}{n}$
  - $\sum(i, 1, 10)x_i = 85 + 72 + 90 + 68 + 78 + 88 + 92 + 65 + 80 + 75 = 793$
  - 所以$\mu = \frac{793}{10} = 79.30$
  - 然后计算总体方差σ^2：
    - $\sum(i, 1, 10)(x_i - \mu)^2 = (85 - 78.30)^2 + (72 - 78.30)^2 + ... + (75 - 78.30)^2 = 790.1$
    - 总体方差$\sigma^2 = \frac{790.1}{10} = 79.01$

2. 标准差

- 标准差是总体方差的平方根，即$\sigma = \sqrt{\sigma^2}$。
- 由上一步得到$\sigma^2 = 79.01$，所以$\sigma = \sqrt{79.01} \approx 8.99$

3. 样本均数的标准误差

- 样本均数的标准误差$S_{\overline{x}} = \frac{s}{\sqrt{n}}$
- 样本标准差$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = 9.37$
- 已知$s \approx 9.37$，$n = 10$，则$S_{\overline{x}} = \frac{9.37}{\sqrt{10}} \approx 2.96$

4. 变异系数

- 变异系数公式为$= \frac{s}{\overline{x}} * 100\%$
- 已知$s \approx 9.37$，$\overline{x} = 79.30$，则$\mathrm{CV} = \frac{9.37}{79.30} * 100\% \approx 11.82\%$

5. 四分位数间距

- 首先将数据从小到大排序：$65, 68, 72, 75, 78, 80, 85, 88, 90, 92$
- $n = 10$，下四分位数Q_1的位置$L_{\mathrm{Q1}} = \frac{n+1}{4} = \frac{10+1}{4} = 2.75$
  - $Q_1 = 68 * 0.25 + 72 * 0.75 = 71$。
- 上四分位数Q_3的位置$L_{\mathrm{Q3}} = 3 * \frac{n+1}{4} = 3 * \frac{10+1}{4} = 8.25$
  - $Q_3 = 88 * 0.75 + 90 * 0.25 = 88.5$。
- 四分位数间距$\mathrm{IQR} = Q_3 - Q_1 = 88.5 - 71 = 17.5$

第五题

1. 计算均值

- 数学成绩 (X) 的均值 $\overline{X}$：

$$\frac{80 + 85 + 90 + 70 + 75}{5} = 80$$

- 物理成绩 (Y) 的均值 $\overline{Y}$：

$$\frac{75 + 80 + 85 + 65 + 70}{5} = 75$$

2. 计算标准差

- 数学成绩 (X) 的标准差 S_X：

首先计算离均差平方和 $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$ $(80-80)^2+(85-80)^2+(90-80)^2+(70-80)^2+(75-80)^2=0+25+100+100+25=250$ 根据标准差公式 $S_X=\sqrt{\sum_{i=1}^{n}\frac{\left(X_i-\overline{X}\right)^2}{n}}$，其中 $n=5$ $S_X=\sqrt{\frac{250}{5}}=\sqrt{50}\approx 7.07$

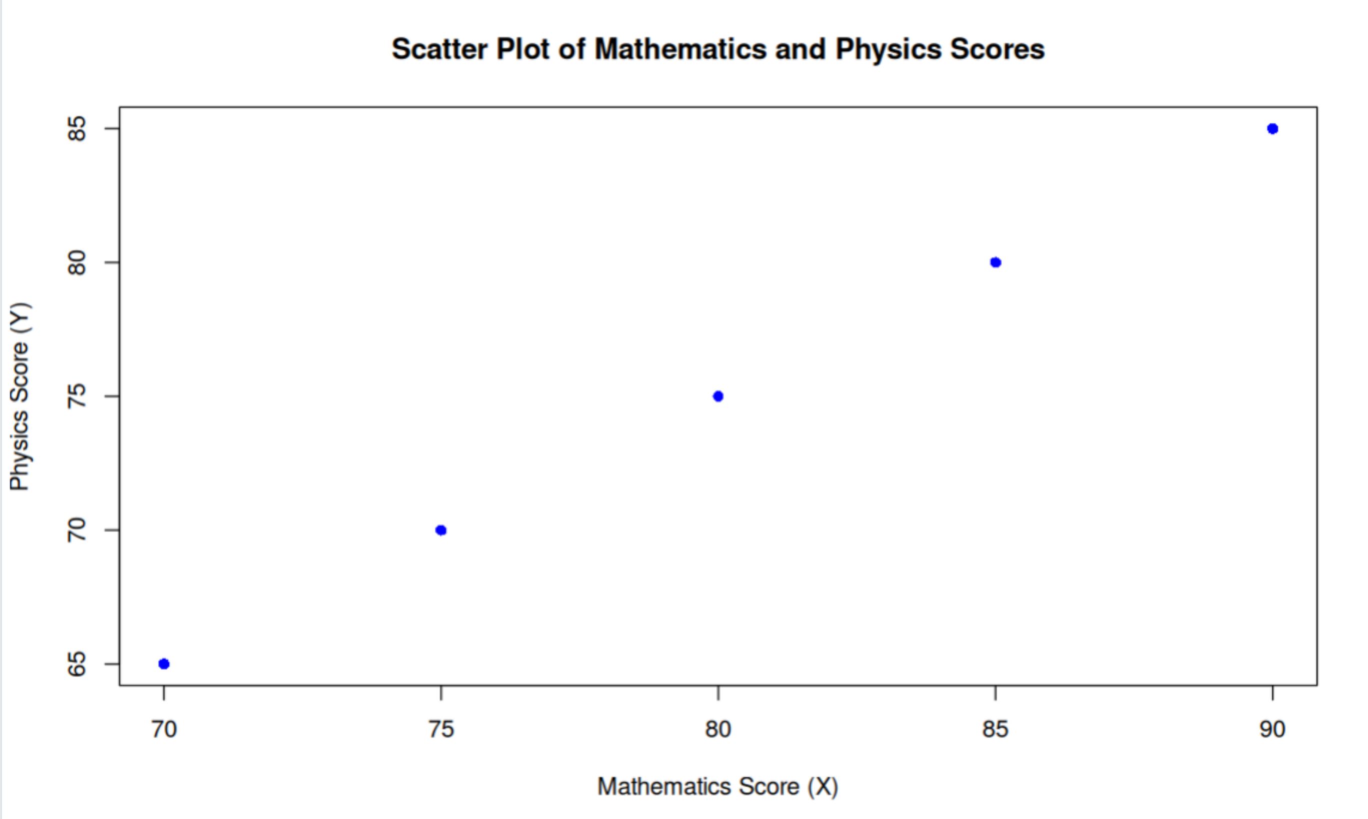
- 物理成绩 (Y) 的标准差 S_Y：

计算离均差平方和 $\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$ $(75-75)^2+(80-75)^2+(85-75)^2+(65-75)^2+(70-75)^2$ $=0+25+100+100+25=250$ 根据标准差公式 $S_Y=\sqrt{\sum_{i=1}^{n}\frac{\left(Y_i-\overline{Y}\right)^2}{n}}$，其中 $n=5$ $S_Y=\sqrt{\frac{250}{5}}=\sqrt{50}\approx 7.07$

3. 计算 Pearson 相关系数

首先计算 $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)$ $(80-80)(75-75)+(85-80)(80-75)+(90-80)(85-75)+(70-80)(65-75)+(75-80)(70-75)=0+25+100+100+25=250$ 根据 Pearson 相关系数公式 $r=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}$ 已知 $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=250$，$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=250$ $r=\frac{250}{\sqrt{250}\times\sqrt{250}}=\frac{250}{250}=1$

4. 绘制散点图



5. 描述变量关系

根据计算出的 Pearson 相关系数 $r=1$，说明数学成绩 (X) 和物理成绩 (Y) 之间存在完全正相关关系。即数学成绩越高，物理成绩也越高，并且两者呈现出非常严格的线性关系，所有数据点都恰好位于一条直线上。