

第二章 统计学

2.2 样本的数字特征

我们通常用样本的数字特征来估计总体的数字特征, 例如, 我们常用样本平均数 $\bar{y}$ 去估计总体平均数 $\bar{Y}$, 用样本频率来估计总体的概率分布. 我们在初中还学习了**众数**、**中位数**、**方差**¹等数字特征, 他们可以被用来描述数据的集中趋势或离散程度, 下面介绍几个新的数字特征.

定义 2.2.1. **样本标准差**是样本方差的算术平方根, 记作 s .

一组数据中最大值与最小值的差称为**极差**.

一组数据的**第 p 百分位数**是这样一值, 它使得这组数据中至少有 $p\%$ 的数据小于或等于这个值, 且至少有 $(100 - p)\%$ 的数据大于或等于这个值.

中位数就是第 50 百分位数, 在实际应用中, 常用的分位数还有**第 25 百分位数**, **第 75 百分位数**. 这两个分位数也被称为**四分位数**, 其中第 25 百分位数被称为**第一四分位数**或**下四分位数**, 第 75 百分位数被称为**第三四分位数**或**上四分位数**.

我们可以通过下面的步骤计算一组 n 个数据的第 p 百分位数:

Step 1: 按从小到大排列原始数据.

Step 2: 计算 $i = n \times p\%$.

Step 3: 若 i 不是整数, 则第 p 百分位数为第 $[i]$ 项数据; 若 i 是整数, 则第 p 百分位数为第 i 项与第 $(i + 1)$ 项数据的平均数.

¹ 此处不再赘述.

第二章 统计学

2.2 样本的数字特征

习 题

A 组

- 【2014 四川文 2】在“世界读书日”前夕，为了了解某地 5000 名居民某天的阅读时间，从中抽取了 200 名居民的阅读时间进行统计分析. 在这个问题中，5000 名居民的阅读时间的全体是
(A) 总体 (B) 个体
(C) 样本的容量 (D) 从总体中抽取的一个样本
- 【新教材 A 版选修二习题 5.3 第 9 题】设 $x_1, x_2, \dots, x_n$ 为样本数据，令 $f(x) = \sum_{i=1}^n (x_i - x)^2$ ，则 $f(x)$ 的最小值点为
(A) 样本众数 (B) 样本中位数 (C) 样本标准差 (D) 样本平均数
- 【2020 全国 III 卷理 3】在一组样本数据中，1, 2, 3, 4 出现的频率分别为 p_1, p_2, p_3, p_4 ，且 $\sum_{i=1}^4 p_i = 1$ ，则下面四种情形中，对应样本的标准差最大的一组是
(A) $p_1 = p_4 = 0.1, p_2 = p_3 = 0.4$ (B) $p_1 = p_4 = 0.4, p_2 = p_3 = 0.1$
(C) $p_1 = p_4 = 0.2, p_2 = p_3 = 0.3$ (D) $p_1 = p_4 = 0.3, p_2 = p_3 = 0.2$
- 【2012 江西 9】样本 $(x_1, x_2, \dots, x_n)$ 的平均数为 $\bar{x}$ ，样本 $(y_1, y_2, \dots, y_n)$ 的平均数为 $\bar{y} (\bar{x} \neq \bar{y})$. 若样本 $(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n)$ 的平均数 $\bar{z} = a\bar{x} + (1-a)\bar{y}$ ，其中 $0 < a < \frac{1}{2}$ ，则 n, m 的大小关系为
(A) $n < m$ (B) $n > m$ (C) $n = m$ (D) 不能确定
- 【2024 新高考 II 卷 4】某农业研究部门在面积相等的 100 块稻田上种植一种新型水稻，得到各块稻田的亩产量（单位：kg）并整理得下表：

亩产量	[900, 950)	[950, 1000)	[1000, 1050)
频数	6	12	18
亩产量	[1050, 1100)	[1100, 1150)	[1150, 1200)
频数	30	24	10

据表中数据，下列结论中正确的是

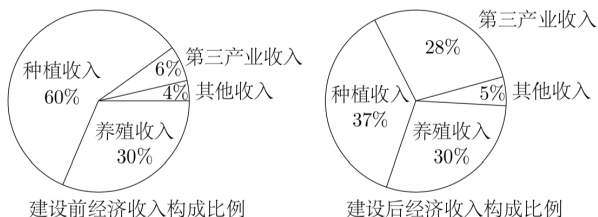
- (A) 100 块稻田亩产量的中位数小于 1050 kg

(B) 100 块稻田亩产量低于 1100 kg 的稻田所占比例超过 80%

(C) 100 块稻田亩产量的极差介于 200 kg 至 300 kg 之间

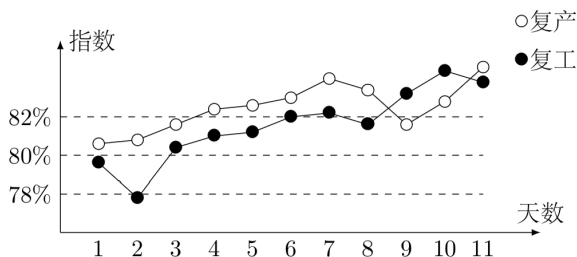
(D) 100 块稻田亩产量的平均值介于 900 kg 至 1000 kg 之间

6. 【2018 全国 I 卷理 3】某地区经过一年的新农村建设,农村的经济收入增加了一倍,实现翻番.为更好地了解该地区农村的经济收入变化情况,统计了该地区新农村建设前后农村的经济收入构成比例,得到如下饼图:



则下面结论中不正确的是

- (A) 新农村建设后, 种植收入减少
 (B) 新农村建设后, 其他收入增加了一倍以上
 (C) 新农村建设后, 养殖收入增加了一倍
 (D) 新农村建设后, 养殖收入与第三产业收入的总和超过了经济收入的一半
7. 【2009 上海文 18 理 17】在发生某公共卫生事件期间, 专业机构认为该事件在一段时间内没有发生大规模群体感染的标志为“连续 10 日, 每天新增疑似病例不超过 7 人”. 根据过去 10 天甲、乙、丙、丁四地新增疑似病例数据, 一定符合该标志的是
- (A) 甲地: 总体均值为 3, 中位数为 4
 (B) 乙地: 总体均值为 1, 总体方差大于 0
 (C) 丙地: 中位数为 2, 众数为 3
 (D) 丁地: 总体均值为 2, 总体方差为 3
8. 【2020 新高考 II 卷 9】(多选) 我国新冠肺炎疫情防控进入常态化, 各地有序推动复工复产. 下面是某地连续 11 天的复工、复产指数折线图.



根据该折线图

- (A) 这 11 天复工指数和复产指数均逐日增加
 (B) 在这 11 天期间, 复产指数的增量大于复工指数的增量

- (C) 第 3 天至第 11 天, 复工指数和复产指数都超过 80%
 (D) 第 9 天至第 11 天, 复产指数的增量大于复工指数的增量

9. 【202 新高考 I 卷 9】(多选) 有一组样本数据 $x_1, x_2, \dots, x_6$, 其中 x_1 是最小值, x_6 是最大值, 则

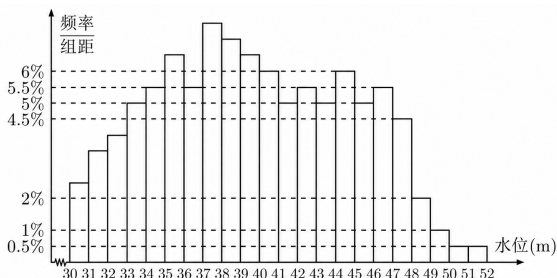
- (A) x_2, x_3, x_4, x_5 的平均数等于 $x_1, x_2, \dots, x_6$ 的平均数
 (B) x_2, x_3, x_4, x_5 的中位数等于 $x_1, x_2, \dots, x_6$ 的中位数
 (C) x_2, x_3, x_4, x_5 的标准差不小于 $x_1, x_2, \dots, x_6$ 的标准差
 (D) x_2, x_3, x_4, x_5 的极差不大于 $x_1, x_2, \dots, x_6$ 的极差

10. 【2021 新高考 II 卷 9】(多选) 下列统计量中可用度量样本 $x_1, x_2, \dots, x_n$ 离散程度的有

- (A) $x_1, x_2, \dots, x_n$ 的标准差 (B) $x_1, x_2, \dots, x_n$ 的中位数
 (C) $x_1, x_2, \dots, x_n$ 的极差 (D) $x_1, x_2, \dots, x_n$ 的平均数

11. 【2026 “fiddle” 模拟考 10】(多选) 根据某河流观测点的若干年水位数据, 得到频率分布直方图 (如图). 若 p 是河流水位大于 x 米的概率, $T = \frac{1}{p}$ (单位: 年), 则称 T 为 x 米水位的重现期. 洪水等级与重现期的关系如下表.

洪水等级	重现期 T
小洪水	$T < 5$
中洪水	$5 \leq T < 20$
大洪水	$20 \leq T < 50$
特大洪水	$T \geq 50$



用频率估计概率, 估计该观测点处

- (A) 47 米水位属于中洪水 (B) 特大洪水的最低水位是 48 米
 (C) 50 米水位的重现期是 100 年 (D) 水位的中位数位于区间 (39,40) (单位: 米)

12. 求样本数据 2, 8, 14, 16, 20 第 80 分位数.

13. 【2013 湖北文 12】求样本 7, 8, 7, 9, 5, 4, 9, 10, 7, 4 的标准差.

14. 【2014 安徽理 6】已知样本数据 $x_1, x_2, \dots, x_{10}$ 的标准差为 8, 求数据 $2x_1 - 1, 2x_2 - 1, \dots, 2x_{10} - 1$ 的标准差.

15. 【2020 上海 6】已知 $a, b, 1, 2$ 的中位数为 3, 平均数为 4, 求 ab .

16. 【2008 上海文 10 理 9】已知总体的各个体的值由小到大依次为 2, 3, 3, 7, $a, b, 12, 13.7, 18.3, 20$, 且总体的中位数为 10.5. 若要使该总体的方差最小, 求 a, b 的值.

17. 某班级有 30 名男生和 20 名女生, 调查学生周末在家学习时长 (单位: 小时), 得到男生样本数据的平均值为 8, 方差为 2, 女生样本数据的平均值为 10.5, 方差为 0.75, 求该班级全体学生周末在家学习时长的平均值 $\bar{x}$ 与方差 s^2 .

18. 【2025 集英苑 5 月（网络联考）13】已知样本数据 $1, 4, a, a, b, 6, 7$ 的众数为 0, 方差为 $\frac{48}{7}$, 求该样本数据的中位数.
19. 已知 $n(n \in \mathbb{N}^*)$ 个数据 $x_1, x_2, \dots, x_n$ 的平均数和标准差分别为 $\bar{x}, S$, 设 $n+1$ 个数据 $x_1, x_2, \dots, x_n, \bar{x}$ 的标准差为 S' , 判断 S 与 S' 的大小关系.
20. 【2013 辽宁文 16 理 16】为了考察某校各班参加课外书法小组的人数, 在全校随机抽取 5 个班级, 把每个班级参加该小组的人数作为样本数据. 已知样本平均数是 7, 样本方差是 4, 且样本数据互不相同, 求样本数据中的最大值.
21. 【2010 安徽文 14】某地有居民 100000 户, 其中普通家庭 99000 户, 高收入家庭 1000 户. 从普通家庭中以简单随机抽样方式抽取 990 户, 从高收入家庭中以简单随机抽样方式抽取 100 户进行调查, 发现共有 120 户家庭拥有 3 套或 3 套以上住房, 其中普通家庭 50 户, 高收入家庭 70 户. 依据这些数据并结合所掌握的统计知识, 你认为该地拥有 3 套或 3 套以上住房的家庭所占比例的合理估计是多少?
22. 某地区的公共卫生部门为调查本地区中学生的吸烟情况, 随机抽样了 200 名学生, 占总学生数的 1%. 调查中使用了两个问题:
- 问题 1: 你父亲的公历生日日期是不是奇数?
- 问题 2: 你是否经常吸烟?
- 每个被调查者随机从一个装有大小、形状和质量完全一样的 50 个白球和 50 个红球的袋子中摸取 1 个球 (摸出的球重新放回袋中). 摸到白球的学生如实回答第一个问题, 摸到红球的学生如实回答第二个问题, 回答“是”的人往一个盒子中放一个小石子, 回答“否”的人什么都不要做. 调查结束后, 盒子中放有 58 个小石子.
- 根据以上材料, 估计该地区吸烟的中学生人数, 并解释以上调查方法的好处.
23. 某大学计算机专业的考研人数连创新高. 今年报名刚结束, 某考生想知道报考人数并计算“上岸”概率. 考生的考号按“0001, 0002, ...”的顺序从小到大排列. 这位考生随机了解的 50 个考生的考号如下:

0400 0904 0747 0090 0636 0714 0017 0432 0403 0276

0986 0804 0697 0419 0735 0278 0358 0434 0946 0123

0647 0349 0105 0186 0079 0434 0960 0543 0495 0974

0219 0380 0397 0283 0504 0140 0518 0966 0559 0910

0658 0442 0694 0065 0757 0702 0498 0156 0225 0327

请给出至少 2 种方法, 根据这 50 个随机抽取得考号估计报考人数.

24. 叙述并证明分层 (2 层) 取样的平均数与方差公式, 并推广到 L 层取样方差公式, 即已知样本共分为 L 层、各层样本占比分别为 $w_1, w_2, \dots, w_L$ 、各层平均数分别为 $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_L$ 、各层方差分别为 $S_1^2, S_2^2, \dots, S_L^2$, 求样本平均数 $\bar{x}$ 与样本方差 s^2 .
25. 解答下述问题:

- (1) 【2009 上海文 18 理 17】在发生某公共卫生事件期间, 专业机构认为该事件在一段时间内没有发生大规模群体感染的标志为“连续 10 日, 每天新增疑似病例不超过 7 人”. 过去 10 日, 判断下述四地新增疑似病例数据是否符合该标志:

甲地: 总体均值为 3, 中位数为 4 乙地: 总体均值为 1, 总体方差大于 3

丙地: 中位数为 2, 众数为 3 丁地: 总体均值为 2, 总体方差为 3

- (2) 四名同学各掷骰子 5 次, 分别记录每次骰子出现的点数. 根据四名同学的统计结果, 可以判断一定没有出现点数 6 的是

同学甲: 平均数为 3, 中位数为 2 同学乙: 中位数为 3, 众数为 2

同学丙: 平均数为 2, 方差为 2.4 同学丁: 中位数为 3, 方差为 2.8

26. 上一年度某公司的 50 名员工中, 最高年收入达到了 400 万元, 员工年收入的平均数是 20 万元. 而你的预期是获得 18 万元年薪, 正犹豫是否受聘.

- (1) 是否能判断年薪为 18 万元的员工在这家公司算高收入者? 并说明理由;

- (2) 已知该公司员工的最低年收入是 6 万, 员工收入的第一四分位数为 9 万, 第三四分位数为 19 万,

- i. 求员工年收入中位数的取值范围并解释其小于平均数的原因;
- ii. 你将如何使用上述信息做出是否受聘的决定?

1. A.2. D. 方法一: 当 $f(x)$ 取最小值时, $f'(x) = 0$, 即

$$\sum_{i=1}^n 2(x - x_i) = 0$$

即 $2nx - 2 \sum_{i=1}^n x_i = 0$, 所以 $x = \frac{1}{n} \sum_{i=1}^n x_i$, 故最小值点为样本平均数.

方法二: 因为 $f(x)$ 是关于 x 的二次函数, 配方得

$$\begin{aligned} f(x) &= \sum_{i=1}^n (x_i^2 - 2x_i x + x^2) = nx^2 - 2(x_1 + \cdots + x_n)x + \sum_{i=1}^n x_i^2 \\ &= n \left(x - \frac{x_1 + \cdots + x_n}{n} \right)^2 - \frac{(x_1 + \cdots + x_n)^2}{n} + \sum_{i=1}^n x_i^2 \end{aligned}$$

所以当 $x = \frac{x_1 + \cdots + x_n}{n}$ (即为样本平均数) 时, $f(x)$ 取最小值.

3. B. 四个情形的均值都是 2.5, 而代入标准差的计算公式, 可知 A、B、C、D 选项的标准差分别为 0.65, 2.5, 2.05, 1.05. 故选 B.

4.

5.

6.

7.

8. D. 甲地可能不符合, 反例为 0, 0, 0, 0, 4, 4, 4, 4, 10. 乙地可能不符合, 反例: 0, 0, 0, 0, 0, 0, 0, 1, 9. 丙地可能不符合, 反例: 0, 0, 0, 1, 1, 3, 3, 3, 100. 丁地符合, 证明如下: 设 10 天数据为 $\{x_k\}_{k=1}^{10}$ (按疑似病例数从小到大排列), 则 $\sum_{k=1}^{10} x_k = 20$. 方差

$$3 = \frac{1}{10} \sum_{k=1}^{10} (x_k - 2)^2 \geq \frac{1}{10} (x_{10} - 2)^2$$

, 则 $x_{10} \leq 2 + \sqrt{30}$. 由于 x_{10} 是整数, 所以 $x_{10} \leq 7$. 即疑似病例数最多不超过 7.

9.

10.

11. AC. 统计中, 反映随机变量离散程度的有: 极差、方差、标准差; 刻画集中趋势的有平均数、中位数、众数.

12. ACD. 扩充表格, 把重现期对应的概率取值范围也列出来.

洪水等级	重现期 T	概率 $p = T^{-1}$
小洪水	$T < 5$	$p > 0.2$
中洪水	$5 \leq T < 20$	$0.05 \leq p < 0.2$
大洪水	$20 \leq T < 50$	$0.02 \leq p < 0.05$
特大洪水	$T \geq 50$	$p \leq 0.02$

对于 A 选项, 水位大于 47 米的频率是 $4.5\% + 2\% + 1\% + 0.5\% + 0.5\% = 8.5\% \in [0.05, 0.2]$, 所以 47 米水位属于中洪水. A 选项正确.

对于 B 选项, 水位大于 48 米的频率是 $2\% + 1\% + 0.5\% + 0.5\% = 4\% = 0.04 \in [0.02, 0.05]$, 所以估计 48 米的水位属于大洪水, B 选项错误. 实际上可以计算出属于特大洪水的最低水位是 49 米.

对于 C 选项, 根据直方图, 估计河流水位大于 50 米的概率是 $0.5\% + 0.5\% = 1\%$, 所以 50 米的河流水位的重现期是 100 年 (即所谓的 “百年一遇”). C 选项正确.

对于 D 选项, 水位大于 40 米的频率是 46.5%, 水位在 (39, 40) 中的频率大于 6%, 所以水位大于 39 米的频率是 52.5%. 因此水位的中位数位于区间 (39, 40). D 选项正确.

Fiddie 评: 本题改编自【2007 湖南文 7】, 但原题题干不够严谨, 且没有提及 “重现期”、“洪水等级” 的概念, 这些新概念取自鄂教版必修第四册的 “阅读与讨论”. 问题源于实际, 洪水等级与重现期的对应关系见于《水文情报预报规范》, 这是真实的规范文件, 可在网上查到.

13.

14.

15.

16. 7;2.

17. 10.5, 10.5. 由条件, $a+b=21$, 则这组数据的平均数是 10, 所以方差为 $s^2 = \frac{1}{10}[C + (a-10)^2 + (b-10)^2] = \frac{a^2+b^2}{10} + C'$, 其中 C, C' 是常数. 利用不等式 $2(a^2+b^2) \geq (a+b)^2$, 等号成立当且仅当 $a=b$, 可得 a, b 取值都是 10.5 时总体方差最小.

18.

19.

20.

21. 10. 设 5 个班级参加书法小组的人数为 x_1, x_2, x_3, x_4, x_5 . 不妨设 $x_1 < x_2 < x_3 < x_4 < x_5$, 则 $\frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5) = 7$, 且 $\frac{1}{5} \sum_{k=1}^5 (x_k - 7)^2 = 4$. 再次由样本数据互不相同, 从而 $x_k + 1 \leq x_{k+1}$ (即相邻两数相差至少为 1). 所以 $x_5 \geq x_1 + 4, x_5 \geq x_2 + 3, x_5 \geq x_3 + 2, x_5 \geq x_4 + 1$. 所以 $35 = x_1 + x_2 + x_3 + x_4 + x_5 \leq -10 + 5x_5 \Rightarrow x_5 \geq 9$. 又由于 $(x_5 - 7)^2 < \sum_{k=1}^5 (x_k - 7)^2 = 16$, 从而 $x_5 < 11$. 则 $x_5 \in \{9, 10\}$.

(1) 如果 $x_5 = 10$, 则 $x_1 + x_2 + x_3 + x_4 = 25$, 为奇数, 且 $\sum_{k=1}^4 (x_k - 7)^2 = 11$, 而 11 写成四个数的平方和只可能为 $11 = 0^2 + 1^2 + 1^2 + 3^2$, 所以 $x_1 = 4, x_2 = 6, x_3 = 7, x_4 = 8$.

(2) 如果 $x_5 = 9$, 则 $x_1 + x_2 + x_3 + x_4 = 26$ 为偶数, 且 $\sum_{k=1}^4 (x_k - 7)^2 = 16$, 而 16 写成 4 个数的平方和只可能为 $16 = 2^2 + 2^2 + 2^2 + 2^2$ 或 $16 = 0^2 + 0^2 + 0^2 + 4^2$, 这两种情况都会产生相同样本数据, 矛盾.

22.

23.

24.

25.

26.

27.

28.

第二章 统计学

2.2 样本的数字特征

第二章 统计学

2.2 样本的数字特征

第二章 统计学

2.2 样本的数字特征