

1、（简答题）查询资料，简述某个特定大语言模型预训练时采用的数据及其特点。

2、（编程题）给定总体数量  $m$  和抽样率  $a$ ，编程实现等距抽样；给定总体数量  $m$ ，抽样率  $a$  和层数  $k$ ，编程实现分层抽样。并分别给出以下数据的抽样结果：（ $m=15$ ， $a=1/3$ ， $k=4$ ）

50,50.3,49.8,49.9,47.6,47,47.5,43,43.2,43.3, 43.1,43.3,43,43.5,0

（此处截图给出抽样结果，源码另附）

3、（证明题）任意正态分布  $N(\mu, \sigma^2)$  都可以通过如下线性变换  $Z = \frac{X-\mu}{\sigma}$  转换为标准正态分布  $N(0,1)$ 。试证明这一结论。（提示：从概率分布定义出发）

4、（解答题）已知某工厂生产的零件尺寸服从正态分布  $N(\mu, \sigma^2)$ ，其中  $\mu$  和  $\sigma^2$  为未知参数。现从一批零件中随机抽取 5 个样本，测得其尺寸（单位：mm）分别为：10.2、9.8、10.0、10.1、9.9。请使用极大似然法估计该正态分布的未知参数  $\mu, \sigma^2$ 。

5、（解答题）设总体  $X$  的概率密度函数是

$$f(x; \alpha) = \begin{cases} \alpha^2 x^{-\alpha+10}, & 0 < x < 1 \\ 0, & \text{else} \end{cases}$$

求参数  $\alpha$  的极大似然估计。

6、（解答题）经测量，我校 2020 年男同学平均身高为 175.2cm，身高标准差为 0.025，女同学平均身高为 162.1cm，标准差为 0.03。2021 年未进行全部同学的身高测量，而是随机选了男女同学各 200 个进行测量，结果男同学平均身高为 176.2cm，女同学平均身高为 163.6cm，问我校同学 2021 年的平均身高与 2020 年有无明显差异。

(显著性水平为 0.05, 男女同学分别分析)

作业要求:

(1) 准备单线本答题。

源码和截图内容通过电子版提交, 建议用word格式, 命名为数理统计+姓名, 提交到网络教学平台。

(2) 严禁抄袭!

(3) 截止时间: 2024年10月22日