《大数据分析与处理》第1次作业

姓名： 李燕琴 学号： 20195633 日期： 2022.02.27

*说明：作业提交电子版，绘图部分，提交绘图源代码（直接粘贴到word）。*

**1. 简要介绍标称属性、二元属性、序数属性、数值属性的特点和区别，并给出相应的例子。（25分）**

标称属性：其值是一些符号或者事物的名称。 如头发颜色= {黑色，棕色，灰色，白色}。

二元属性：又叫布尔（bool）属性，是一种标称属性，只有两个状态。根据两种状态的价值，分为**对称的**（两种状态具有同等价值，携带相同权重，一般用独热编码表示，如性别）、**非对称的**（其状态的结果不是同样重要。如：艾滋病毒的阳性和阴性结果。一般对重要的结果用1编码，另一个用0编码）。

序数属性：其可能的值之间具有有意义的序或者秩评定(ranking)， 但是相继值之间的差是未知的，其中心趋势可以用它的众数和中位数表示，但不能定义均值。如成绩={优，良，中，差}。

数值属性：分为**区间标度属性**（使用相等的单位尺度度量。值有序，可以评估值之间的差，不能评估倍数或倍数没有实际的意义。没有绝对的零点。如:摄氏温度，华氏温度，日期等）和**比率标度属性**（具有固定零点的数值属性，值有序，可以评估值之间的差，也可以说一个值是另一个的倍数。如：开式温温标(K)，重量，高度，速度等）

2. 假设所分析的数据包括属性age，它在数据元组中的值（以递增序）为13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70。（25分）

(a) 该数据的均值是多少?中位数是什么?

均值：29.963，中位数：25

(b) 该数据的众数是什么?讨论数据的模态(即双峰、三峰等)。

众数：25

分布的模态可以是单峰或双峰，具体取决于出现值的频率。单峰分布是单峰分布，其中一个值出现的频率最高。它是一个具有单个清晰可见的峰值或单个最频繁值的分布。另一方面，双峰分布是两个值以最大频率出现的分布，这意味着两个频繁值之间有一个间隔。该数据分布图如图 1所示，可知该数据的分布模态为双峰分布。

图表, 直方图

描述已自动生成

图 1 该数据分布图

(c) 该数据的中列数是多少?

中列数 = (max-min)/2 = (70-13)/2 = 28.5.

(d) 你能(粗略地)找出该数据的第一个四分位数(Q1)和第三个四分位数(Q3)吗?

Q1 = 20.5, Q3=35.

(e) 给出该数据的五数概括。

[最小值，最大值，第一四分位数，第三四分位数，中位数] = [13, 70, 20.5, 35, 25].

(f) 绘制该数据的盒图。

四分位数极差：IQR = Q3-Q1 = 35-20.5 = 15.5

上边界：Q3+1.5\*IQR = 58.25, 可知数字70离群。

下边界：Q1-1.5\*IQR = -2.75。

盒图如下：

图表, 箱线图

描述已自动生成

(g) 分位数-分位数图（q-q图）与分位数图有何不同?

分位数图：显示给定数据的分位属性，其中(x,y)指给定数据中的y在数据中所占百分比为x。

q-q图：主要用于两个数据分布是否相同，其中(x,y)中，x和y分别表示不同的数据，当两个观测集的值的个数不一致时，不是所有的值都能被表示。

代码如下：

d = np.array([13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70])

c = Counter(d)

**print**('mean=%.3f, median=%.3f, mode = %.3f'%(np.mean(d),np.median(d),stats.mode(d)[0][0]))

title = "该数据分布图"

plt.plot(list(c.keys()),list(c.values()),'b')

plt.plot(list(c.keys()),list(c.values()),'r\*')

plt.xlabel("data")

plt.ylabel("count")

plt.title(title)

plt.savefig(title+'.png', dpi=300, bbox\_inches="tight", pad\_inches=0.1)

plt.show()

**print**('max=%.3f, min=%.3f, max\_min\_mean=%.3f'%(np.max(d),np.min(d),(np.max(d)-np.min(d))/2))

**print**('Q1=%.3f, Q3=%.3f'%(np.percentile(d,25),np.percentile(d,75)))

title= "该数据的盒图"

plt.boxplot([d])

plt.title(title)

plt.savefig(title+'.png', dpi=300, bbox\_inches="tight", pad\_inches=0.1)

plt.show()

3. 假设医院对18个随机挑选的成年人检查年龄和体脂率，得到结果如下：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 年龄 | 23 | 23 | 27 | 27 | 39 | 41 | 47 | 49 | 50 |
| 体脂率 | 9.5 | 26.5 | 7.8 | 17.8 | 31.4 | 25.9 | 27.4 | 27.2 | 31.2 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 年龄 | 52 | 54 | 54 | 56 | 57 | 58 | 58 | 60 | 61 |
| 体脂率 | 34.6 | 42.5 | 28.8 | 33.4 | 30.2 | 34.1 | 32.9 | 41.2 | 35.7 |

根据上述数据，完成如下计算任务（25分）：

1. 计算年龄和体脂率的均值、中位数和标准差。

(b) 绘制年龄和体脂率的盒图。

(c) 绘制基于这两个变量的散点图和q-q图。

绘图参考提示：<https://zhuanlan.zhihu.com/p/346467939>

1. 年龄：mean=46.444, median=51.000, std=12.846

体脂：mean=28.783, median=30.700, std=8.994

1. 二者的盒图如下：

图表, 箱线图

描述已自动生成

1. 散点图

图表, 散点图

描述已自动生成

QQ图

图表, 折线图

描述已自动生成

代码如下：

age = np.array([23,23,27,27,39,41,47,49,50,52,54,54,56,57,58,58,60,61])

fat = np.array([9.5,26.5,7.8,17.8,31.4,25.9,27.4,27.2,31.2,34.6,42.5,28.8,33.4,30.2,34.1,32.9,41.2,35.7])

title = "年龄和体脂 盒图"

**print**("mean=%.3f, median=%.3f, std=%.3f"%(np.mean(age),np.median(age),np.std(age)))

**print**("mean=%.3f, median=%.3f, std=%.3f"%(np.mean(fat),np.median(fat),np.std(fat)))

plt.boxplot([age,fat],labels=['年龄','体脂'])

plt.title(title)

plt.savefig(title+'.png', dpi=300, bbox\_inches="tight", pad\_inches=0.1)

plt.show()

title = "年龄-体脂 散点图"

s\_age= np.sort(age)

s\_fat= np.sort(fat)

sns.regplot(s\_age,s\_fat,ci=None,color='b',line\_kws={'color':'r'})

plt.xlabel("年龄")

plt.ylabel("体脂")

plt.title(title)

plt.savefig(title+'.png', dpi=300, bbox\_inches="tight", pad\_inches=0.1)

plt.show()

title = "年龄-体脂 Q-Q图"

s\_age = sorted([np.percentile(age,i) **for** i **in** np.linspace(1,100,500)], reverse=True)

s\_fat = sorted([np.percentile(fat,i) **for** i **in** np.linspace(1,100,500)], reverse=True)

sns.regplot(s\_age,s\_fat,ci=None,color='b',line\_kws={'color':'r'})

plt.xlabel("年龄")

plt.ylabel("体脂")

plt.title(title)

plt.savefig(title+'.png', dpi=300, bbox\_inches="tight", pad\_inches=0.1)

plt.show()

4. 简要概述如何计算如下属性描述的对象的相异性（25分）：

(a) 标称属性

对象和对象j之间的相异性可用不匹配率表示，即

1. 非对称二元属性

根据两个对象的频数表，得到

表格

描述已自动生成

1. 数值属性

数值属性一般使用闵可夫斯基距离表示，公式如下：

黑色的钟表

低可信度描述已自动生成

1. 序数属性

设属性共有个，排位，先通过进行归一化，再通过闵可夫斯基公式计算。

1. 词频向量

词频向量通常很长，稀疏的，使用余弦相似性作为度量：