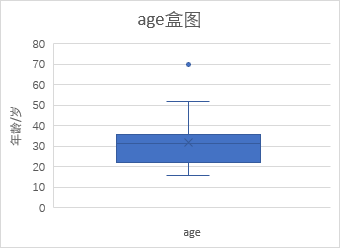
第一次作业

第一部分（20分）

1. 习题2.2
2. 均值为29.96；中位数为25；
3. 众数是25和35；该数据是一个双峰的分布，即是二模；
4. 中列数为（70+13）/2=41.5；
5. 该数据共有27个；27/4约为7，则Q1=20；7\*3=21，则Q3=35；
6. 五数分别为最小值=13；Q1=20；中位数=25；Q3=35；最大值=70；
7. IQR=Q3-Q1=15;盒图如下：



1. 分位数-分位数图是反映了同一个属性的不同样本的数据分类情况，使得用户可以很方便的比较两个样本之间的区别或者相同之处；而分位数图是观察单变量数据分布的简单有效方法，它显示给定属性的所有数据的分布情况，还绘出了分位数信息。
2. 习题2.3

总频数=200+450+300+1500+700+44=3194；3194/2=1597；

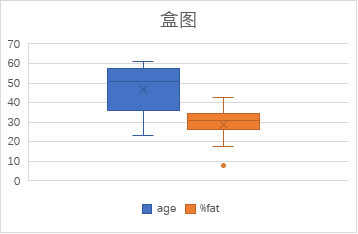
可以得出中位数应该在21-50之间；则中位数=21+（（1597-950）/1500））\*30=33.94；

1. 习题2.4

（a）:age 均值=46.44，中位数=51，标准差=12.85；

%fat 均值=28.79，中位数=30.8，标准差=8.997；

（b）盒图如下



（c）散点图如下：

(4) 习题2.6

(a) 欧几里得距离 d==6.7;

(b) 曼哈顿距离 d==11;

(c) 闵可夫斯基距离q=3, d==6.15;

(d) 上确界距离=42-36=6；

第二部分（40分）

1.（1）Q1:标称数据 Q2：比率数值 Q3：比率数值 Q4：序数 Q5：标称

Q6：比率数值 Q7：比率数值 Q8：二元 Q9：标称 Q10：标称

Q11：区间标度 Q12：序数 Q13：序数 Q14：序数 Q15：序数

Q16：标称

（2）Q1：

由上表可以看出，Q1比较集中，中心趋势集中出现在car上，同时SUV运动型实用汽车出现次数也比较多。整体数据比较集中。

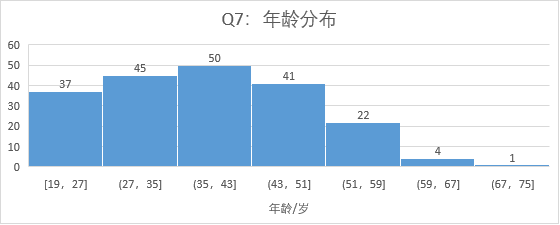
Q2：

从上表可以看出，过去两年违章的次数众数的是0次，约占有1半；数据的中位数为1，去除10次以上的数据后，数据平均值为1.39，集中在1左右；

去除10次以上的数据后，数据的标准偏差值为1.66，方差为2.73，总体而言数据比较靠近均值。

从上图可以看出，过去两年内出现交通事故的概率约为25%，最大值为5，说明有人出过5次交通事故，有75%的人没有发生过交通事故，平均值为0.35，众数为0，中位数为0，中列数为2.5；计算可得，标准偏差为0.73，方差为0.53，可以得出数据的分散程度。

从上图可以看出，众数为每天，每天占比52%，最少的为每月一次，占8%。



从上图可以看出，平均年龄为38岁，其中分布在35-43的人最多，占25%，19-59岁的人数有173个，占据大多数。

从上表可以看出收入在10万元以下的有90人，20万以下的有158人，占大多数，50万以上的只有5人，占2.5%；

（3）Q1

由于Q1属性中，car及suv占据90%，其他样本数据太小，不具有代表性。从以上两个表可以看出，SUV有罚单的概率为73%，car有罚单的概率为55.2%，显然suv发生的概率高于car.排除意外因素，应该跟suv的使用环境及车速等有关。

Q3

绘制上图时，为了便于发现规律，将10次以上的9999记为10。如图所示，可以看出，Q2与Q3的重叠概率很高，可见发生事故次数多的人拿到罚单的概率就很高，两者之间有正相关性。

Q4

违章概率=64.4%

违章概率=71.1%

违章概率=50%

违章概率=37.5%

违章概率=9.1%

由此来看，使用频率为一周几次的人出现违章的概率最高，整个呈抛物线趋势，每天开车的反而低于一周几次的，推测可能是因为每天的熟练程度更高。

Q7

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **列1** | **19-20** | **21-30** | **31-40** | **41-50** | **51-60** | **61-70** |
| 违章比例 | 0.2 | 0.314 | 0.683 | 0.627 | 0.68 | 0.6 |

从上图及表可以看出，随着年龄的增长，违章比例有一定的增长趋势，41-50岁的司机比较熟练而且是开车的主力军，违章比例略有下降，至于60以上，样本虽少，但也能反映问题，不开车的人比较多，所以开车违章的比例必然高于0.6，可见整体来说违章比例比较高。

Q11

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **列1** | **不到10万元** | **10万元以上20万元以下** | **20万元以上30万元以下** | **30万元以上50万元以下** | **50万元以上100万元以下** | **100万元以上** |
| 违章比例 | 0.467 | 0.559 | 0.69 | 0.627 | 0.889 | 1 |

从上表可以得出很明显的结论，随着收入的增加，违章比例也在明显上升，两者成正相关的关系，尤其是100万元以上的竟然达到了1，因此在设计罚款金额，可以根据收入逐层递加，以此达到减少违章概率的目的。

第三部分：2.

（1）a.停车线周期流量

|  |  |  |  |
| --- | --- | --- | --- |
| **列1** | **相位4-1车道** | **相位4-2车道** | **相位4-3车道** |
| 平均值 | 19 | 20 | 16 |
| 中位数 | 10 | 11 | 10 |
| 众数 | 1 | 2 | 1 |

上表是相位4的停车线周期流量的中心趋势。三个车道的平均值为18，平均值和为55。可以得到三个车道的中位数分别为10，11，10，在10左右。众数为1，2，1，说明停车周期流量出现最多的次数在1左右，中位数明显低于平均值，说明极差应该比较大。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **列1** | **相位4-1车道** | **相位4-2车道** | | **相位4-3车道** |
| 极差 | 59 | 53 | 47 | |
| 四分位数 | 1 | 3 | 2 | |
| 方差 | 359.55 | 315.44 | 242.15 | |
| 标准差 | 18.96 | 17.76 | 15.56 | |
| 四分位数极差 | 18.04 | 20.24 | 17.44 | |

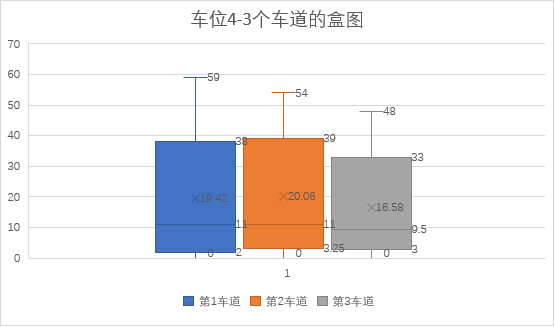
上表是相位4的停车线周期流量的分散趋势，从图中可以看出，车道1的分散程度最大，极差和方差都最大，但2车道的四分位数极差最大，盒图较长。车道3的极差，方差都最小，相对来说分散趋势最小，比较集中。

b.排队长度

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1车道 | 2车道 | 3车道 |
| 平均值 | 19.42 | 20.06 | 16.58 |
| 中位数 | 11 | 11 | 9.5 |
| 众数 | 2 | 3 | 2 |

由上表可以看出，1，2，3的均值与中位数差距较大，均值偏大，说明有较大的值的出现，拔高了平均值。3个车道的平均值为19，众数分别为2，3，2，说明出现的排队长度大多数时间在2左右，并不是很长。

|  |  |  |  |
| --- | --- | --- | --- |
| **列1** | **1车道** | **2车道** | **3车道** |
| 极差 | 59 | 54 | 48 |
| 四分位数 | 2 | 3.75 | 3 |
| 方差 | 358.72 | 321.62 | 237.68 |
| 标准差 | 18.94 | 17.93 | 15.42 |
| 四分位数极差 | 36 | 35.25 | 30 |

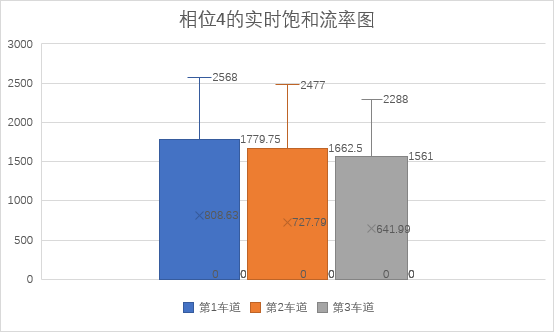


从上面的盒图和表可以看出，排队长度的分散趋势，相对而言，第三车道相对集中，集中在3-33之间，可以看出中位数偏向于底部，不过由于极差的原因，导致平均值大于中位数。

C.实时饱和流率

|  |  |  |  |
| --- | --- | --- | --- |
| **列1** | 车道1 | 车道2 | 车道3 |
| 平均值 | 808.63 | 727.79 | 641.99 |
| 中位数 | 0 | 0 | 0 |
| 众数 | 0 | 0 | 0 |

从上表可以看出，整体的中位数及众数都为0，说明有一半以上的时间处于0的状态，不过由于在一些不算长的时间里，饱和流率极大，仍然导致了平均值很大。



从上图可以看出，并不是很分散，最小值和四分位数都为0，结合上表中的中位数可知，由超过一半的集中在0点，但是最大值和四分之三分位数仍然很大，说明有离群点的存在，分布在四分之3位数以上。

（2）第8相位

a.停车线周期流量

|  |  |  |  |
| --- | --- | --- | --- |
| **列1** | **1**车道 | **2**车道 | **3**车道 |
| 平均值 | 18.71 | 18.60 | 16.98 |
| 标准差 | 17.74 | 17.33 | 16.60 |

从上表可以看出总体统计特征，平均值为在18左右，标准差在17左右；一天内的随时间变化规律，如上图所示，横轴表示时间，为2018-8-30日一天的变化情况，可见在该交叉口5：17分之前周期流量都很低，在此之后一路上到第一个高峰点，在7：20左右达到峰值，然后趋于下降，而后周期性的上升和下降，在35左右徘徊，直到16点-18点30分之间又达到一个在43左右徘徊的高峰期，从19点开始，整个周期流量以锯齿形下降，知道夜晚11点达到低谷，并趋于稳定。

b.排队长度

|  |  |  |  |
| --- | --- | --- | --- |
| **列1** | **1车道** | **2车道** | **3车道** |
| 平均值 | 17.92 | 17.85 | 15.58 |
| 中位数 | 5.5 | 6 | 5 |
| 众数 | 2 | 2 | 2 |
| 极差 | 52 | 53 | 52 |
| 标准差 | 17.78 | 17.42 | 16.24 |
| 四分位数 | 2 | 3 | 2 |

从上表可以看出相位8的总体统计特征，平均值在17，中位数为5.5，6，5；众数均为2，说明极差很大，有很大的数值拉高了整个平均值。极差接近52，四分位数为2，3，2说明在6以下2以上的数占据了接近1/4.

从上图可以看出，与周期流量类似，在5：17分之前，排队长度持续处在低水平，此后开始升高，不同于周期流量的是，排队长度的变化幅度比较大，有从2跨度到50多的大幅度升降，两个高峰分别出现在7：29和接近5：00。可以看出排队长度整体而言，在白天，变化幅度比较大，且出现过数次的大幅升降，与此同时，排队长度夜间下降的时间相对于周期流量来说，发生了前移，在10点半，整体的排队长度都处在了10以下。

c.实时饱和流率

|  |  |  |  |
| --- | --- | --- | --- |
| **列1** | **1**车道 | **2**车道 | **3**车道 |
| 平均值 | 597.76 | 533.20 | 472.39 |
| 中位数 | 0 | 0 | 0 |
| 众数 | 0 | 0 | 0 |
| 极差 | 2566 | 2367 | 2411 |
| 方差 | 817.58 | 795.55 | 756.29 |

从上表可以看出，中位数为0，说明有过半的时间，实时饱和流率为0，但平均值为4-500多，足以显示出有部分极大值，带来了平均值比较高的情况。极差在本例中同时也是最大值，从方差来看，整体数据的离散情况相对来说比较大，但同时0附近又比较密集。

从上图可以看出，在夜间10：30到凌晨6：00，实时饱和流率均为0；在白天饱和流率有一个较大的变化，饱和流率的单位为辆/每车道\*小时，可以看出当一个周期内车流比较大时，折换成上述单位后就会比较大，所以在0-几千之间的变化比较多。从中可以看出在下午3：13-6：38之间，整体流量很少出现为0的情况，说明这一段时间内为车流的高峰期，车流来往很密切。

（3）a.周期流量

从上述两个图的对比来看，两者在凌晨0点到5：30之间，周期流量都比较低。可以看出在早高峰上，第8相位出现时间要比第4相位延迟几分钟。整体来看，两者的高峰和低峰出现时间几乎一致，但第4相位的高峰相对要比第8相位高一些。

b.排队长度

从上面两图来看，8相位排队长度一天之内变化比较频繁，相对而言，4相位排队长度变化次数比较少。第4相位的排队长度在高峰期要高于第8相位。两者在11：44-12：38之间都出现了一个明显的低谷，然后又上升到高峰。两个都可以算作是二模数据，有两个维持时间比较长的高峰。

从上面两图可以看出，第8相位的实时饱和流率相对第4相位明显变化频繁，在0-2000之内来回波动频繁，第4相位在18：34分之后，与第8相位又接近相同的变化趋势。两者的饱和流率为0的时间段接近完全一致，整体来说，第4相位的实时饱和流率要略高于第8相位。