**利用概率统计知识分析数据集——以成都滴滴平台订单数据为例**

**韩森宇 519021911288**

**[摘要]**在万物联网的当下，我们每日的出行、购物、餐饮等行为均会在互联网上留下一条条的记录。由无数人的行为产生的大数据对学术研究具有重要意义，而数据持有方往往会面向学术界开放部分数据集，推进学术研究与实际产品、成果的转化，为社会发展创造更大的价值。本文从生活中常见的“网约车”平台所产生的大数据入手，利用课程所学的概率统计相关知识分析平台提供的数据集，提出一系列假设并加以验证，以此达到应用课程知识解决实际问题的目的。

**[关键词]**概率统计，大数据，网约车

本文的数据来源为滴滴数据开放计划KDD CUP 2020数据集[1]，涵盖成都市2016年11月1日至2016年11月15日快专车订单的2 939 820条数据记录，包含订单起止时间、费用等字段。利用《问题求解与实践》课程大作业制作的数据分析与可视化程序与Excel、Origin等软件进行数据处理。

一、数据预处理

由于数据量庞大，为便于分析，选取较为典型的数据作为总体。数据集将成都市区划分为100个栅格，本文选取以成都市金牛区所在栅格为起点或终点的订单作为研究对象。

金牛区是成都市较为繁华的地区，数据集显示从该地所在栅格出发的订单共有540 569条，是所有栅格中出发订单数最多的，占所有订单的18.5%。选取该栅格研究可以较好地模拟成都市区内的快专车交通情况。根据研究需要的不同，可以进一步限制订单的起止时间、订单目的地等。

二、泊松分布验证与估算——早高峰期的订单数量

当一个随机事件以固定的平均瞬时速率随机且独立地出现时，那么这个事件在一段时间内出现的次数应近似服从泊松分布[2]。为验证这一规律，此例中，我们选取数据集内2016年11月1日（星期二）8:00—9:00一小时内从金牛区所在栅格出发的订单共1 786条记录。

设为每分钟平台接收到的订单数量。早高峰期间，用户叫车需求稳定，则理论上应服从泊松分布。取为样本，其中是从8:00起第分钟内的订单数量，数据如下（从至顺序排列）：

31 25 28 25 25 24 22 40 35 29 25 32 27 29 24

31 34 34 26 33 29 24 25 31 40 28 28 29 29 36

32 35 27 33 27 33 32 31 28 29 25 36 38 32 19

31 36 33 26 24 22 34 29 28 29 32 33 30 35 29

假设：。首先进行参数估计。由参数估计的相关知识知，估计量是的无偏、一致、有效估计量，故选取作为参数的估计值。依非参数检验，将的取值范围分为个子区间，计算结果如下表1所示。

**表 1 泊松分布检验计算表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 事件 | 实际频数 | 理论概率 |  |  |  |
|  | 3 | 0.123 | 7.38 | -4.38 | 2.60 |
|  | 12 | 0.158 | 9.48 | 2.52 | 0.67 |
|  | 17 | 0.211 | 12.66 | 4.34 | 1.49 |
|  | 11 | 0.207 | 12.42 | -1.42 | 0.16 |
|  | 11 | 0.153 | 9.18 | 1.82 | 0.36 |
|  | 6 | 0.148 | 8.88 | -2.88 | 0.93 |

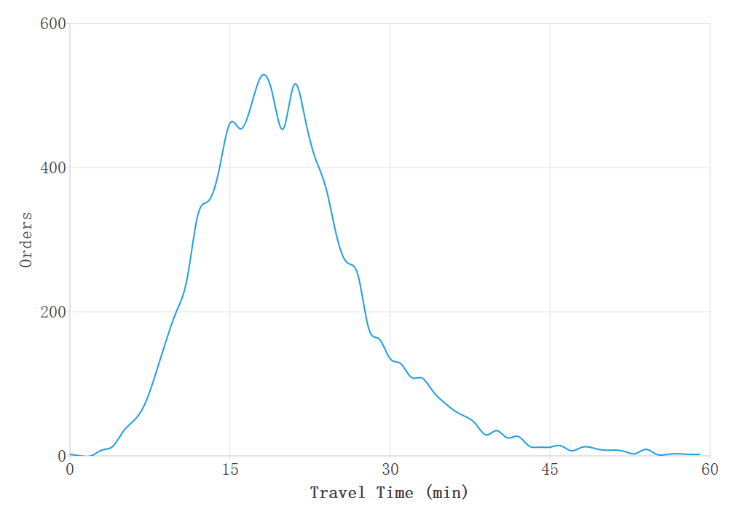
所需估计参数个数。依Pearson定理，检验统计量，取显著性水平，此值下的拒绝域为

计算得的样本观测值

查表得，因此接受假设，即认为早高峰期平台每分钟接收到的从金牛区出发的订单数量服从参数为29.77的泊松分布。

三、正态分布的参数估计与假设检验——工作日与双休日的订单用时分布

首先使用可视化程序，选取以武侯区所在栅格为起点、金牛区所在栅格为终点的，出发时间位于2016年11月9日（星期二）、10日（星期三）之中工作日内的共8 776条订单记录，绘制其订单用时的分布（下图1）：



**图 1 工作日期间武侯区—金牛区订单所用时间分布**

可以看出，该分布图像与正态分布较为接近。实际上，对大数据整体进行订单用时的分布绘图时，其展现的图像明显为正偏态分布而非正态分布。为得到正态分布的曲线，本例限制了出发点与终点，减少了长途订单的干扰，使订单用时更为集中。依驾车通勤、乘车旅游等生活经验，在固定时间段内从某地出发前往另一地所用的时间应当服从正态分布，因此也可以假设工作日期间以武侯区为起点、金牛区为终点的订单所用时间也服从正态分布。

设订单所用时间为。略去正态分布的检验步骤，假设订单所用时间。从记录中随机选取个样本，数据如下：

1609 525 765 1478 1080 964 1723 1035 1086 943

1166 1144 1277 1447 766 1332 2109 1308 1867 144

1748 1119 589 1394 2030 1228 1141 1067 976 2259

1599 1500 1233 589 1771 1731 907 835 1458 1481

1591 1621 1210 1436 527 468 434 1614 1242 1890

1025 1347 1693 1391 1632 1788 1112 1597 1099 1003

计算得。二者分别为的无偏统计量。

接下来计算参数的置信区间。依区间估计相关知识，取置信度，由于均未知，对参数，选取为枢轴量，有置信区间

近似取，代入数值得置信区间。

对参数，选取为枢轴量，有置信区间

近似取，代入数值得置信区间。

由上述的计算可知，工作日期间从武侯区出发乘快专车到达金牛区所用的平均时间在20分钟到23分钟之间。我们可以提出这样一个问题：双休日期间，从武侯区到金牛区的订单用时是否也符合上述分布呢？

假设工作日期间，从武侯区到金牛区的订单所用时间确实满足。接下来检验双休日期间的订单用时是否显著不符合这一分布规律。同样地，从出发时间位于2016年11月5日（星期六）、6日（星期日）之内的，以武侯区所在栅格为起点、金牛区所在栅格为终点的7 972条订单记录中随机选取个样本，数据如下：

1486 1083 891 668 1141 1183 811 942 954 1918

780 1412 1459 1152 428 977 2116 1374 1015 2190

2194 1333 1700 1618 2319 1324 1071 2529 1512 2618

1724 1104 1738 1650 1266 863 899 1489 739 972

1800 3629 508 1477 1364 2191 891 1623 769 1713

1215 1010 1289 823 878 1282 1047 1242 1098 564

计算得。

假设：双休日的订单用时分布也服从正态分布，即。依假设检验相关知识，取显著性水平，双休日的订单用时均值、方差均未知。

先检验均值。构造检验统计量，欲检验，则其拒绝域为

计算得，不能拒绝假设中的均值假设。

再检验方差。构造检验统计量，欲检验，则其拒绝域为

计算得，故拒绝假设中的方差假设。

综上，有理由认为双休日期间从武侯区到金牛区的订单用时与工作日期间的订单用时有显著差别，具体表现在订单用时的方差增大，更为均匀。可以猜测：双休日期间，人们前往金牛区这类的繁华地段娱乐的需求增加，相对更远的订单数量增加；或者双休日期间高峰期的拥堵程度要高于工作日期间。

四、二维变量的分布——订单费用与订单用时的关系

快专车的订单费用除了与里程有关，往往与订单的用时也有关系。由于数据集中仅给出了订单起终点的经纬坐标而未给出具体里程，故我们无法分析订单费用和里程的关系，但仍可以探究订单费用和用时的关系。

此例中，我们选取2016年11月15日的、以金牛区所在栅格为起点的、起止时间位于8:00—20:00之间的共44 468条记录。从订单记录中随机选取300个样本（样本数据冗长，在此省略），设为订单所用时间（秒），为订单费用[[1]](#footnote-1)，将视作样本点。利用Origin软件绘制这300个样本点的散点图（下图2）：



**图 2 订单用时与订单费用的散点图**

可以看到，订单用时小于600秒（10分钟）时，订单的费用几乎都集中在1.8附近；之后随订单用时增加，订单的费用也逐渐上升。这与我们的常识相符：出租车或快专车以起价计费，超过起价里程后价格随里程与时间的增加而增加。有理由认为成都地区滴滴快专车的起价即为1.8。

接下来讨论用时大于600秒的订单，其订单用时与订单费用的关系。对筛选后剩下的240个样本点作一拟合直线表示趋势（下图3）：



**图 3 筛选后作趋势线的散点图**

从图3可以看出，订单用时与费用确实存在某种关系，越大，往往也就越大。故可以计算的相关系数，以评价与之间线性关系的强弱。利用Origin的拟合分析，计算可得，说明与之间确实有一定的正相关性。

进一步可以得到拟合直线。由于并不接近于1，且订单费用仍受到里程数等其他因素影响，故该拟合直线表达式的意义有限，但还是能作为粗略的规律。

五、总结

通过上述三个案例的分析，我们可以得到成都市滴滴快专车订单在数量、用时、费用等方面的规律。并且，依照时间段与星期等额外条件的不同，往往可以得到与其他条件下的分析有所差异的结果。事实上，庞大的大数据中可以挖掘的信息是无限的，本文仅是利用课程所学的概率统计知识对部分数据作一些粗浅的讨论，这份由滴滴平台提供的订单信息仍有着相当的数据潜力没有被挖掘出来。在未来的学习中，我将进一步学习数据挖掘等与人工智能相关的知识，届时我将具备更优秀的能力来处理这样庞大的数据集。

另外要指出的是，由我编写的可视化程序在本文的数据筛选过程中起到了重要的作用；Excel、Origin等软件则为绘图、统计量的计算提供了相当便利。诸如MATLAB等软件的开发与应用广泛而深远地在数学领域的分析、统计、计算方面产生影响，也提醒我们要学会使用数学软件与编程语言，按自己的分析意愿，高效而准确地处理数据。

参考文献

[1] 滴滴盖亚数据开放计划. https://outreach.didichuxing.com/app-vue/dataList.

[2] 卫淑芝,熊德文,皮玲,等. 大学数学概率论与数理统计：基于案例分析[M]. 北京：高等教育出版社,2020.8.

附录

本文的选题基于我在本学期参加的课程CS241：《问题求解与实践》的课程大作业：Analysis and Visualization of Online Ride-Hailing Order Data。该大作业要求分析滴滴平台提供的订单数据，并编写一个分析与可视化程序对庞大的数据进行分析。

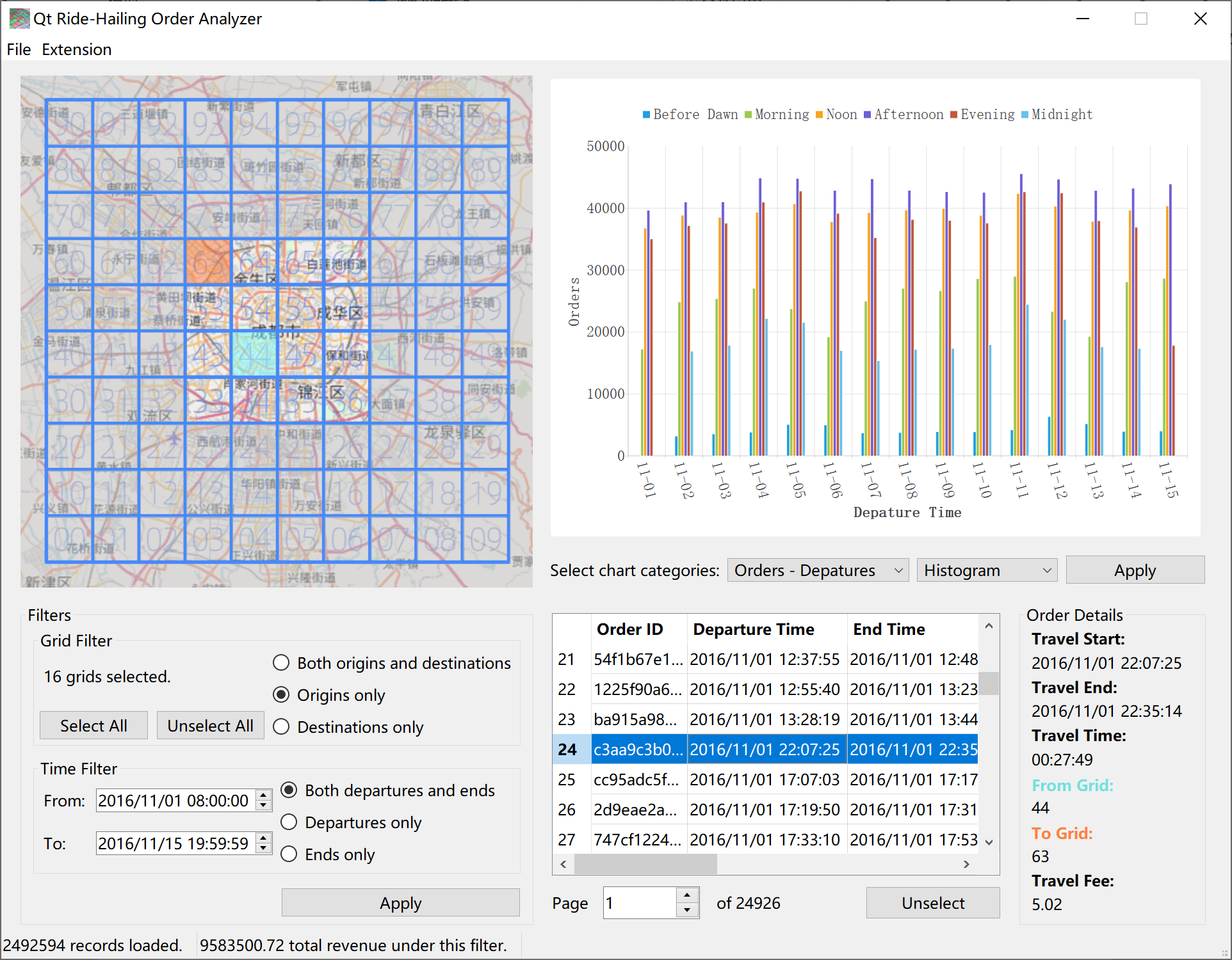
一、订单数据

学生获得的订单数据包含以下字段：订单编号、起始时间、结束时间、起点经纬度、终点经纬度、订单费用。订单编号与费用经过了一定的脱敏处理。另外，按照课程要求，一系列的经纬度信息将成都市区划分为了100个栅格。从经纬度信息可以得到对应的栅格，时间戳格式的起始时间与结束时间经过变换后可以得到实际的时间，并获得订单用时。

二、可视化程序

下图为我编写的可视化程序的主界面，该程序是《问题求解与实践》课程的项目作业。它可以筛选、查询指定时间地点的订单信息，并将出发时间、订单用时、订单费用等数据绘制成多种图像直观显示出来，此外还具有基于同类数据对订单起终点、用时与费用进行预测的功能。

程序代码可于2020年12月28日（《问题求解与实践》课程大作业截止提交）后在<https://github.com/cnlnpjhsy/CS241_CS2309-Principles-and-Practice-of-Problem-Solving>处获得。



**附图 可视化程序主界面**

1. 由于滴滴平台对数据集进行了脱敏处理，订单费用的实际值被隐藏。故本文中的订单费用不加单位。 [↑](#footnote-ref-1)