# 业务中的描述统计学

*<2019年1月 5 日更新 V2.0>*

# 修订记录

| 日期 | 版本 | 修改人 | 修改原因 |
| --- | --- | --- | --- |
| 2019年 1月5日 | V2.0 | Kylie | ➔ 校对和增加内容 |
| 2019年 1月1日 | V1.0 | Alvin, Kylie | ➔ 创建“业务中的统计学”文档v1.0版本 |

# 一. 数据与业务

工作中会产生非常多的数据，数据的收集和分析、解读对我们的业务工作起到指导作用。数据是指不同的信息片段，不仅仅是表格上的简单数据，还有可能以多种多样的形式存在 ：比如音频图像等。

数据业务流程：一般我们接到数据业务需求后，会跟业务方明确数据维度，当然有可能业务方自己也并不清楚自己想要的业务数据，这样我们就需要很好的挖掘一下业务方的核心痛点，所以说根据不同业务，我们所需要的维度或数据也是不一样的，随后根据这些数据维度去收集相应的数据，清洗整理数据后，根据数据展开探索与挖掘，运用统计学在数据中发现一些规律，从而运用到指导业务中去。


## 1. 数据来源

内部数据：一般来说，大部分公司会有自己的CRM（客户关系管理），OA（办公自动化）系统等来存储业务数据；很多公司会使用 SQL server 等数据库存储的方式储存关系型数据；当然还有工作中日常产生的许多数据，比如用户调查、财务数据等，会采用电子表格的形式存储数据。

**外部数据**：公开平台上的开源数据；爬虫获取网站的数据。

## 2. 数据在业务中起到的作用

数据在业务中起到的作用不言而喻，对数据驱动型公司甚至有指导决策的意义，当然对于大多数公司不同职位的朋友来说都有很多的帮助：

**普通员工**：不同职位的普通员工都可以利用数据帮助自己进行一些业务决策，比如销售人员可以根据数据优化营销策略、产品经理可以根据数据优化产品，寻找用户痛点；

**中层管理**：使用数据进行KPI的追踪，统一统计绩效的口径；进行商业预测，提出下一步的经营建议；

**高层管理**：通过数据看整个公司的运营和管理；或展示商业回报，为融资做足数据准备等。

# 二. 工作中的描述统计学

数据在我们的工作和生活中无处不在。我们接下来将通过一些例子来解释课程中学到的描述统计学的一些概念。

## 1. 数据类型

| | | |
|---|---|---|
| 数值: | 连续 | 离散 |
| | 身高、年龄、收入 | 书中的页数、院子里的树、咖啡店里的狗 |
| 分类: | 定序 | 定类 |
| | 字母成绩等级、调查评级 | 性别、婚姻状况、早餐食品 |

在开始着手对数据的分析之前，我们需要了解手上的数据是什么类型，然后再根据不同类型的数据采取相应的分析方式或可视化图形。

**数值数据**：数值型数据是按数字尺度测量的观察值，其结果表现为具体的数值。针对于数值型数据，我们一般分析其四个主要方面：集中趋势、离散程度、形状和异常值。

**分类数据**：分类型数据指反映事物类别的数据。如人按性别分为男、女两类。对于分类数据一般探讨的比较少，分析方法多采用查看每个组的独立个体的数量或比例。

## 2. 集中趋势

我们着重探讨数值型数据，当然我们最常分析的就是数据集的集中趋势和离散情况，这里先着重探讨一下集中趋势：平均数、中位数和众数。

首先我们需要明确的是这三个描述集中趋势的值以及值的大小并没有优劣之分，只是我们需要结合不同的业务场景，采用不同的值来描述手上的数据集比较合适。

**均值（平均值）**：比如下图的折线图，本月业务线日销售额，我们想要了解一下某天销售额的情况，取每月各业务线的销售均值进行对比就非常直观了。

如果与之前的任意一天进行对比都不够客观。


**中位数**：对于中位数我们一般在数据集中有明显极端值的时候用的比较多，比如组内成员的业务能力相对表平均，但是有一位同学业务能力超强，会将组内销售额水平整体拖高，当然汇报的时候采用均值会相对好看，但是如果指定KPI，还是中位数要相对公平客观一点。

**众数**：众数的场景相对简单，当我们需要了解手上数据表现出来的高频，多采用众数，比如了解自己的核心用户、查看用户使用的最高频词汇等。


## 3. 离散程度

说完了数据集的集中趋势，我们再来看离散程度，在离散程度中我们一般需要了解四个值，分别是极差、四分位差、标准差、方差。这四个值中最常用的是标准差。标准差的定义是：每个观察值与均值之间的平均差异最常见的数据离散程度度量之一。。

此外，课程中还提供了一个五数概括法来更准确地描述数据集的离散程度。下图就是刚刚提到的五数概括法：我们会用极大极小值，上下四分位数以及中位数来描述数据集，如果我们想比较两个数据集的离散程度，虽然直方图还算直观，但是每次都展示直方图或者计算五数，可以说相当麻烦，如果用标准差，那么只用一个数值就能比较两个数据集的离散程度了；而且标准差具有推论统计学方面的优势，推论统计学，大家学到数据分析课程的时候会详细讲解，这里就不赘述了；当然非常重要的一点就是标准差与原始数据相同的单位，这也是标准差用的比方差多的多的原因。

**标准差栗子1**：说了这么多，口说无凭，我们来举个栗子，这是一个标准差越小越好的栗子，业务背景：该公司对业务的需求不是让大家都变得业务水平很高，而是尽量拉平大家的业务水平；业务需求是，洞察A组同学的业务知识水平，并予以统一。所以我们明确一下数据集中是值：明确不清楚的业务点，对症下药，围绕展开培训 ；数据离散是指：大家掌握参差不齐，很难找重点予以培训。首先了解同学们的业务离散程度我们用标准差就很方便，更方便的是可以对比培训前后标准差的大小，查看离散程度有没有降低，或者比较这个月前后的标准差，查看离散程度。

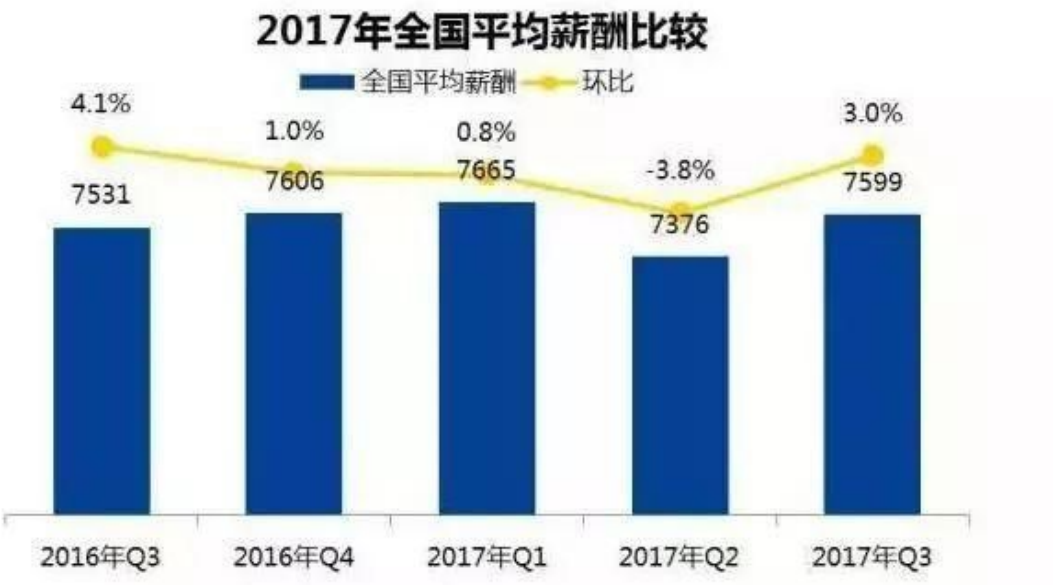
**标准差栗子2**：下面再说一个标准差越大越好的栗子，当然下面这个栗子只是一般情况，还有很多特殊情况不在我们这次探讨的范围内。对于公司的薪资来说，不管是相同岗位还是不同岗位，我们站在市场的角度，都希望其离散程度越大越好，为什么这么说呢，首先对于相同职位如果工资离散程度比较大， 说明大家同工不同酬，大家凭实力说话，更容易提高积极性，拉开距离。其次对于不同职位如果工资离散程度比较大说明了该公司职位很多，层级明显，侧面反映了公司的规模很大。

所以在现实生活工作中，我们对标准差的运用还是非常多的，大家要根据实际情况判断我们要怎么使用手头的统计学工具。

## 三. 统计和非事实

有的时候统计得出的结果和事实自我感觉并不一致，难道是统计出了问题么，这里我们来找具体例子探讨一下：

# 1. 我的工资被平均了




相信大家每年年末总会看到年末发出来的各城市的平均工资，相信大家的感觉是蒙的，这个平均数怕是不准吧，我的工资什么时候那么高了，我是不是拖了后腿，其实并不是平均数的锅，而是用错了统计方法，要是使用中位数会好很多。像前面说的，要是数据集里面有极端值，那么这个平均数很可能是没有参考价值的，比如，A房间里有小明和姚明，平均身高是1米8，那么小明要不要心里美滋滋呢，答案很明显哒，我们处理自己数据集的时候一定要注意排除极端值~

# 2. 我为什么被抽样代表了

很多同学都质疑抽样调查是不是不准，很多时候感觉抽样的结果与自己的认知相差甚远，其实好的抽样调查，不是你想象的那么简单，一个合格的抽样分析，要满足很多条件：

首先我们采用抽样调查，是因为很多事情我们不能采用全样本分析的办法，比如了解全国学生的视力情况，权衡了对全样本操作的成本和抽样的成本后，我们采取后者；

样本量经过科学的计算，不是随口来说，我们会去计算最小样本量，用最经济科学的样本量去预估总体的情况；

抽样的方法繁多，我们总会根据具体的实例采用不同的抽样方法，比如随机抽样、分层抽样等等。

不管采用哪种抽样方法，都要保证样本的随机性，排除人的影响是我们始终需要注意的事情
抽样的同时，我们还会限定置信区间，就是是常说的置信度，是平常事件的95%还是医疗情况的99%置信度，都尽量使得我们得到的结果落在一个区间范围内，结果可控且能量化，更有助于我们做出决策

# 3. 美味的冰淇淋会导致溺水

一个统计机构发现吃冰淇淋的人数增多，会导致游泳溺亡的人数增多。所以得出了结论：大量吃冰淇淋会导致溺水。

这当然是不对的，我们需要明确的是，一般来说，我们用统计学探究的都是相关性，而不是因果性，因果性的探究要严谨很多很多，比如我们在医学上探究一个疾病与药效的相关性就需要采用大样本随机双盲实验，一个实验下来基本都是几年的光景，所以需要注意我们以后的报告如果是通过统计而不是试验得出的结论，尽量都说明探究的是相关性而不是因果性。

相关性并非属于描述统计学的范畴了，而是推论统计学。**大家请在接下来的课程中继续学习。**

**友情提示：推论统计学课程难度比描述统计学难度要大许多，学有余力的同学建议可以在学习本课程内容之后提前学习哦！**