第 3 章

基础统计分析

本章将简要讲解统计学的基本概念和分析方法。作为数据分析的第一步，基础的统计分析有助于研究者对数据有一个初步的认知和整体把握，是进行更复杂的数据分析处理的基石。接下来，我们从对数据进行实际处理的角度出发，首先介绍统计学的基本概念和重要的统计指标，再通过可以呈现数据结构的基本可视化图表，全面地展现数据可视化分析的基础环节。

3.1 统计学的基本概念

统计学是概率论的基础，两者又相互依存。概率论是定量地描述某个事件发生的可信度。比如我们多次随意抛出一枚硬币在桌子上，待硬币静止不动之后，我们会发现基本上只有两种情况发生，即硬币的正面朝上或者是背面朝上。假如我们连续进行 1 000 次这样的实验(相信很少有人有耐心这样做)，我们会发现，在这 1 000 次的实验里，出现正面的次数和出现反面的次数差距不大,并且几乎不可能出现硬币能自己立在桌面上的第三种情况。因此，我们可以归纳出如下结论：抛硬币可能出现两种结果，每种结果发生的可能性是相等的。如果用定量的方式去描述，显而易见，用我们已经熟悉的分数形式，将所有可能性作为分母，每种情况发生的可能性作为分子，从而可以得出我们日常使用的结论，即“硬币出现正面或反面的概率是 1/2”。

得出上述“硬币出现正面或反面的概率是 1/2”的结论所使用的方法，其实正是统计学的方法。统计学是一门基于实验的科学。“没有调查就没有发言权”，这句话用在统计学里再合适不过了。利用统计学所得出的种种结论，都离不开对所研究对象开展的观测和记录，并且所获得数据的数量或规模要有一定的保证。“调查”的重要性直接影响最终结论

的真实性和可信度。值得注意的是，所调查的对象本身可以是实验模拟出来的，也可以是实际生产生活中真实发生的事件。比如上面我们研究的“抛硬币”问题，就可以按照反复测试的实验方式，获取实验观测的数据。

如今，统计学作为一门独立的学科，在国内外众多高校都开设了。大数据时代背景下，人们对统计学的关注度较前些年又得到进一步的提升，凸显了统计学重要的社会价值。

对于要研究的某个具体问题来说，一个收集数据的最小对象被称为个体，而问题涉及的全部个体的集合被称为总体。假设要研究的问题是便利店货物销售的规律性，那么该便利店在某天货物的销量就可作为个体，将过去某一个时间段的销售记录作为总体。进行研究时，从总体中抽取的可测量个体的集合被称为样本。另外，还需要对样本进行标记。具体而言，假设使用整个一年的销售记录作为样本进行研究，不仅要记录下每天的薯片销量、汽水销量，同时还要标记当天的信息，比如是周几，以及这一天是否是节假日。这些标记能反映出个体特征的研究指标，被称为变量。数据表格中一般以个体为一行，以一个变量为一列，如图 3.1 所示。

	变量*a*	变量*b*	变量*c*
个体 1			
个体 2			
个体 3			

图 3.1　变量与个体的关系

表格中的内容就是具体的数值，根据数据类型的不同可以把变量分为不同类型。连续变量的取值是在实数域上的，例如，销售量和销售额这样用数值描述的数据，或者是在一个区间内的任意取值。分类变量，比如星期几和是否是节假日（只有“是”和“否”这两个值），是无法用数值描述的，包括性别、国籍这样无序的变量，也可能是优、良、中、差这样有序的变量。

3.2　连续变量的统计描述

通过抽样调查收集到数据之后，为了便于理解，对数据进行汇总的过程叫做统计描述。

➲ 3.2.1　频数

所谓频数，是指对数据进行分组后，每个分组中出现的数据次数。此外，还可以利用频数与数据总个数的比，也就是用频率来替换频数。频率的好处是可以平衡不同数据之间数量级的不同。下面我们以图 3.2 所示的 2019 年各省份国内生产总值（GDP）数据为例。

地区	2019年
北京市	35 371.28
天津市	14 104.28
河北省	35 104.52
山西省	17 026.68
内蒙古自治区	17 212.53
辽宁省	24 909.45
吉林省	11 726.82
黑龙江省	13 612.68
上海市	38 155.32
江苏省	99 631.52
浙江省	62 351.74
安徽省	37 113.98
福建省	42 395.00
江西省	24 757.50
山东省	71 067.53
河南省	54 259.20
湖北省	45 828.31
湖南省	39 752.12
广东省	107 671.07
广西壮族自治区	21 237.14
海南省	5 308.93
重庆市	23 605.77
四川省	46 615.82
贵州省	16 769.34
云南省	23 223.75
西藏自治区	1 697.82
陕西省	25 793.17
甘肃省	8 718.30
青海省	2 965.95
宁夏回族自治区	3 748.48
新疆维吾尔自治区	13 597.11

图 3.2　2019 年各省份 GDP[①]

以 10 000 亿元为间隔，可以将这 31 个数据分到 11 组中，进而得到一张频数表。这里的 10 000 亿元被称为组宽，11 被称作组数，如表 3.1 所示。

表 3.1　各省 GDP 频数表

GDP/亿元	频数	频率
0～10 000	5	0.16
10 000～20 000	7	0.23
20 000～30 000	6	0.19
30 000～40 000	5	0.16
40 000～50 000	3	0.10
50 000～60 000	1	0.03
60 000～70 000	1	0.03
70 000～80 000	1	0.03
80 000～90 000	0	0
90 000～100 000	1	0.03
100 000～110 000	1	0.03

也可以将频数表画成图，也就是直方图，如图 3.3 所示。

频数观察是一种简单直观但是粗糙的观察方法，观察的效果一般受到组数和组宽的影响，所以可以多次选择组数和组宽进行观察。从直方图中可以观察出数据的集中趋势、离散趋势和整体的分布形态，不过想要更精确地描述数据，需要进一步使用描述指标。

① 数据来自国家数据网 https://data.stats.gov.cn/.

(单位：亿元)

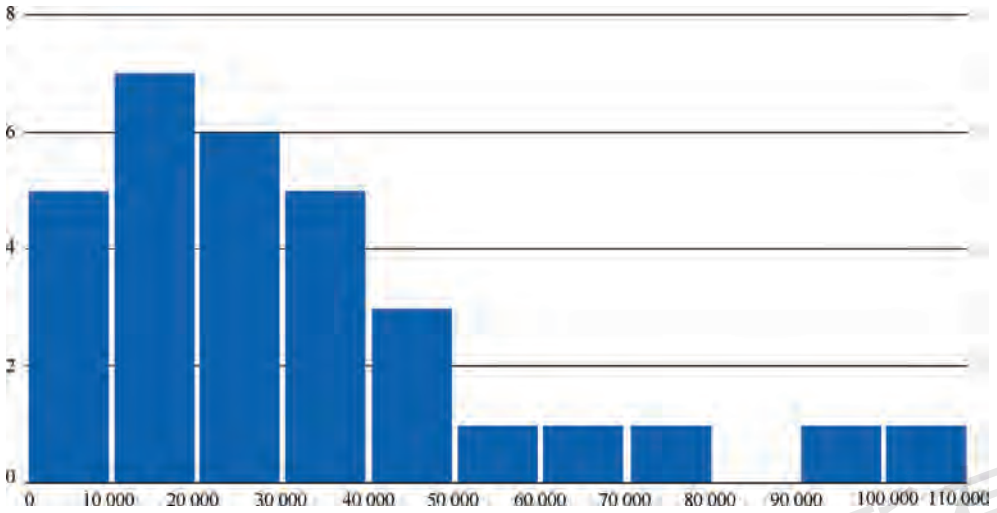

图 3.3　各省 GDP 直方图

3.2.2　集中趋势描述指标

从 3.2.1 小节的图 3.3 可以看出，GDP 在[10 000，20 000]这个区间的省份是最多的，频数进而向两边很自然地滑落，绝大多数省份的 GDP 都位于[0, 50 000]的区间内。也就是说，数据存在一个集中在 20 000 亿元的趋势。类似这种趋势普遍存在于多种数据之中，我们可以通过使用一些常用的指标来对这类趋势进行描述。

最常用的描述指标是均值，也称为算术平均数。均值的计算非常简单，将一组数据进行求和，再除以这组数据的个数即可。均值的本质是将全部个体的差异抽离出来相互抵消，得到一个数据集中的位置。根据均值的计算方法，可以很快算出图 3.2 中 2019 年各省的平均 GDP 为 31 784.94 亿元。均值能反映数组中每一个数据的变化，因此保有了一定程度的信息量。

同时，需要注意到，均值和图 3.3 中频数最高的区间相近，但是仍有一定差距。这一差距产生的原因，是因为个别极端的数值影响了整体数组的均值，比如远高于均值的 107 671.07 亿元(广东省)以及数值较低的 1697.82 亿元(西藏自治区)。由此可见，受到极端值的影响，均值一定程度上掩盖了内部的差异。

为了减少极端值对均值的影响，可以计算截尾均值，即先去掉最大值和最小值各 5%的数据，再使用中间 90%的数据计算均值。按照此方法，能够由上述数据算出 2019 年各省 GDP 的截尾均值为 30 205.66 亿元。另一种思路是选取出现最多的值，或者用一定精度的参考值作为全部值的代表，即众数。按照这个方法计算，可以得出 2019 年各省 GDP 的众数为 20 000 亿元。但显然，众数具有很大的偶然性，尤其是组宽比较小的时候。

除此之外，能够描述集中趋势的是中位数。中位数是所在数组中数值大小位于中间的那个数据值。中位数本质上是位置平均数，基本不受极端值影响。然而，中位数损失了大部分数据的数值变化信息，一旦数据量较少就很不稳定。上述数据中，2019 年各省 GDP 的中位数是江西省的 24 757.50 亿元。

3.2.3　离散趋势描述指标

从图 3.3 的直方图还可以看出，不同省份之间的数值差距非常大。如果只关注上述集中趋势指标，这种差距就会被忽略。为了能够对数据有更全面的描述，还需要关注离散趋势的描述。

最简单直观的离散趋势指标叫全距，或者叫极差，指的是一组数据中最大值和最小值之差。例如在上述数据中，2019 年各省 GDP 的极差是 107 671.07 亿元(广东省)和 1697.82 亿元(西藏自治区)的差值 105 973.25 亿元。更为常用的是方差：

$$\sigma^2 = \sum \frac{(x-\mu)^2}{n}$$

其中，x 是这组数据中的每一个取值，而 μ 表示的是整个数组的均值。方差和数据本身的单位不一致，可以开方得到标准差：

$$\sigma = \sqrt{\sum \frac{(x-\mu)^2}{n}}$$

同样的，可以计算出上述数据中，2019 年各省 GDP 的标准差为 837.07 亿元。此外，如果想要跨量纲地进行比较，比如研究不同省份之间人口和 GDP 的差异，这些指标可以使用变异系数：

$$CV = \frac{\sigma}{\mu}$$

方差、标准差和变异系数是比较常用的离散趋势描述指标，同均值一样，它们一般在对称分布中使用。

一个更广泛适用的指标是百分位数。它是一种位置指标，用 *Px* 表示。一个百分位数 *Px* 将数据分为两个部分，使得 *x*%的数据比它小，(100–*x*)%的数据比它大。百分位数需要多个组合使用，最常用的组合是四分位数，即 *P*25(下四分位数)、*P*50(中位数)和 *P*75(上四分位数)。四分位数将数据四等分，排除了两端的极端值影响。在上述数据中，2019 年各省 GDP 的四分位数是黑龙江省的 13 612.68 亿元,江西省的 24 757.50 亿元和湖南省的 39 752.12 亿元。

3.3　分类变量的统计描述

对于分类变量来说，频数是该变量每一取值出现的次数。以一组汽车的参数数据为例，变量“汽缸数”的取值有“四缸”“六缸”和“八缸”三种。可以做频数表如表 3.2 所示(由于各自保留两位有效数字，频率之和并不为 1)。

表 3.2　汽缸数参数频数表

汽缸数	频数	频率
四缸	199	0.52
六缸	83	0.22
八缸	103	0.27

将分类变量的频数表画成图，得到的是条形图，或者叫作柱状图(见图 3.4)。柱状图的柱子之间是相互分隔的，因为柱状图的不同柱子源于变量自身的分类，而非对数轴的切割分组。

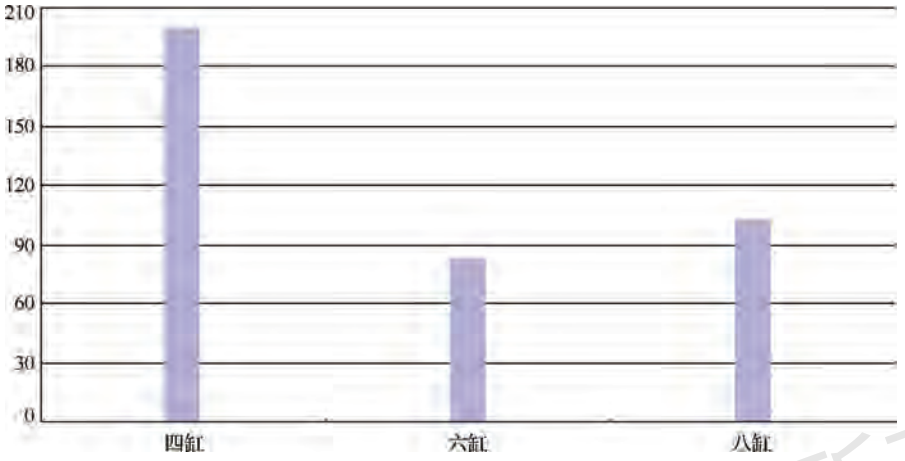

图 3.4　汽缸数参数柱状图

当然，要想强调各个分类在总体中的占比，可以使用饼图(见图 3.5)。

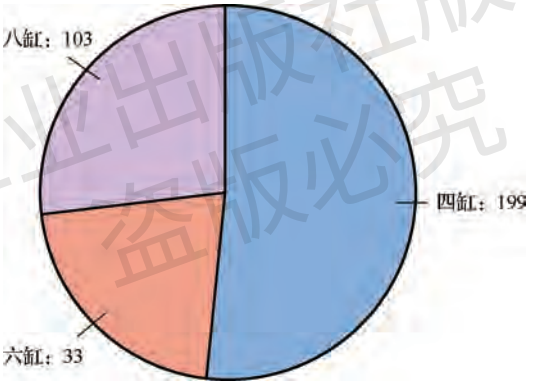

图 3.5　汽缸数参数饼图

分类变量的描述指标通常是根据既有指标，通过比值来定义新的指标——既可以是同一变量的两个分组，也可以是交叉不同的变量。要根据变量的具体含义，对实际情况有指导意义。比如在研究离婚率的问题中，离婚作为一个具有较长时间跨度的事件，在进行数据统计的时候，需要考虑对抽样的人群进行长期的追踪观察，记录每一年中有多少个体发生了离婚，总时间跨度可以是几十年。

3.4　常用统计图

从前两节的内容我们可以看出，在对变量进行统计描述时，图往往有比文字和数字更强的表达效果。因此，本节将对一些常用的图进行讲解。

3.4.1　饼图

饼图是将圆划分为几个扇形的圆形图(见图 3.6)，每个扇形表示一个分类，扇形圆心角与分类数量成正比。饼图本质上是在表现构成比，即各部分占总体的比例。饼图的表现效果直观，易于接受和理解，但是不宜有太多分类，以免杂乱。

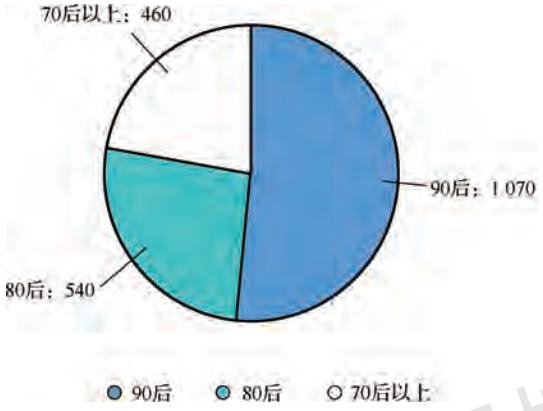

图 3.6　饼图

饼图的缺点是只有构成比，所以不同饼图之间的比较不太有意义。另外圆形设计对空间的利用率比较低，可以空出中心区域来填入其他信息，即甜甜圈图(见图 3.7)。

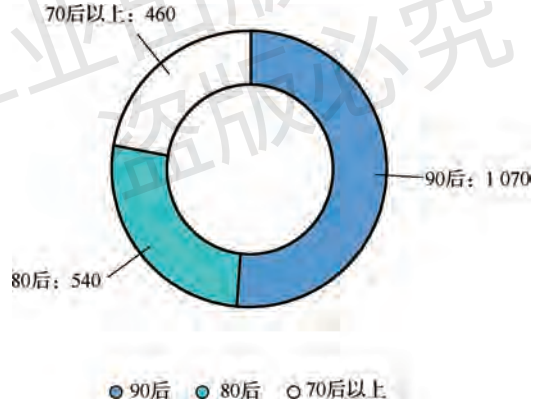

图 3.7　甜甜圈图

水球图也是饼图的变体，它用更多的空间和突出的视觉效果来强调最关键的信息(见图 3.8)。

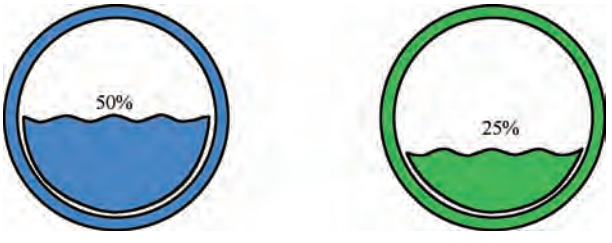

图 3.8　水球图

3.4.2 柱状图

柱状图是使用矩形长条对比分类变量的统计图(见图 3.9)，每个矩形表示一个分类，矩形长度与分类频数成正比。通过修改纵轴的起点可以突出条形之间的差异，但是容易产生对数据差异的夸大化。

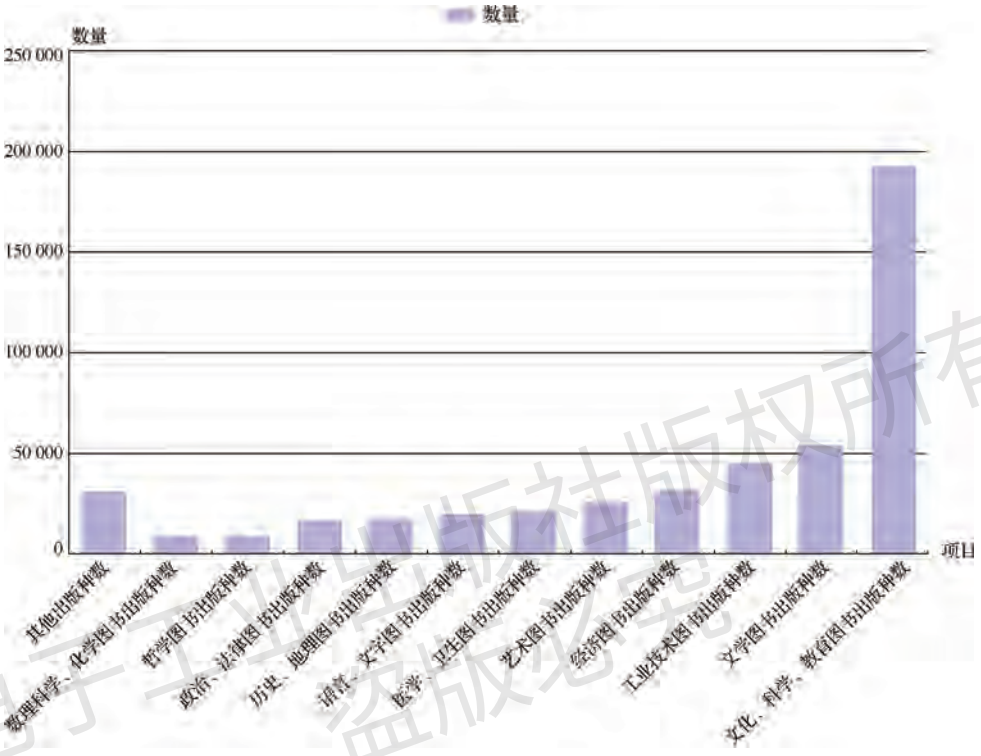

图 3.9 柱状图

柱状图也可以画在极坐标上来获得不同的视觉效果，即玉珏图(见图 3.10)。

图 3.10 玉珏图

如果将两个变量交叉分组，其中一个变量是二元的，那就可以绘制左右对称的旋风图（见图 3.11）。

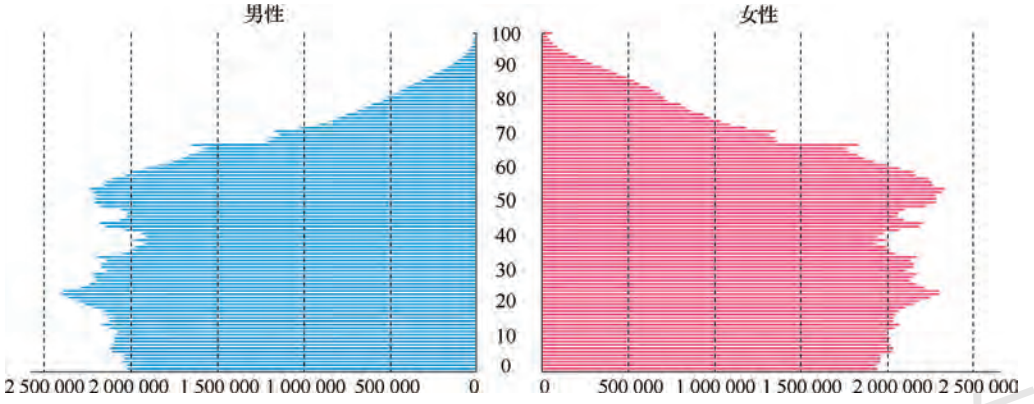

图 3.11　旋风图

更普遍的两个变量交叉分组可以绘制为堆叠柱状图，也就是在分组的基础上再分组，如图 3.12 所示。

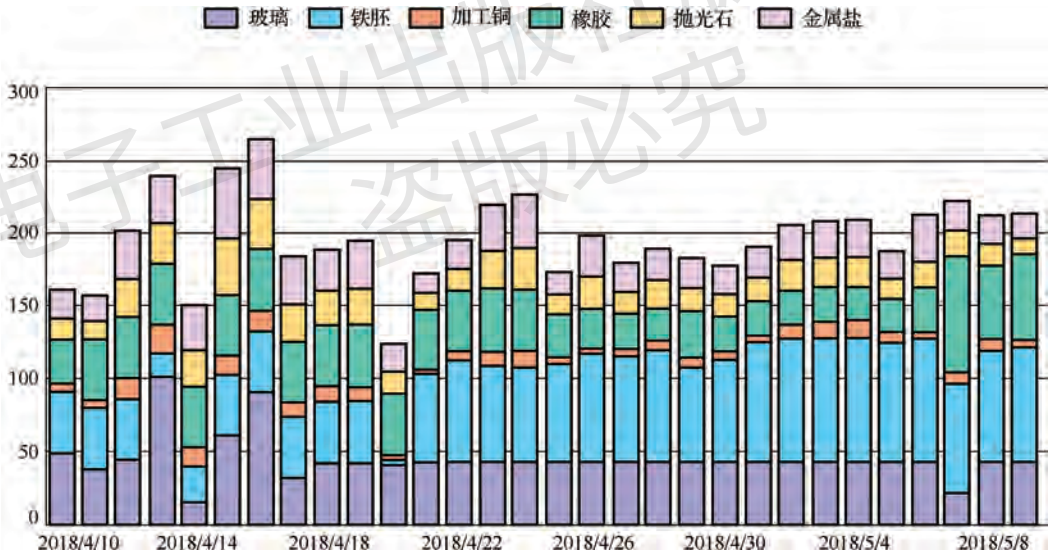

图 3.12　堆叠柱状图

柱状图的变体也可以用于分组的连续变量，区间柱状图是把各组的最值绘制为矩形的上下边缘，用整个矩形表示该组的分布区间，如图 3.13 所示。

误差柱状图则使用矩形长度来表示分组中数据的均值，还可以增加误差线来表示标准误差，如图 3.14 所示。

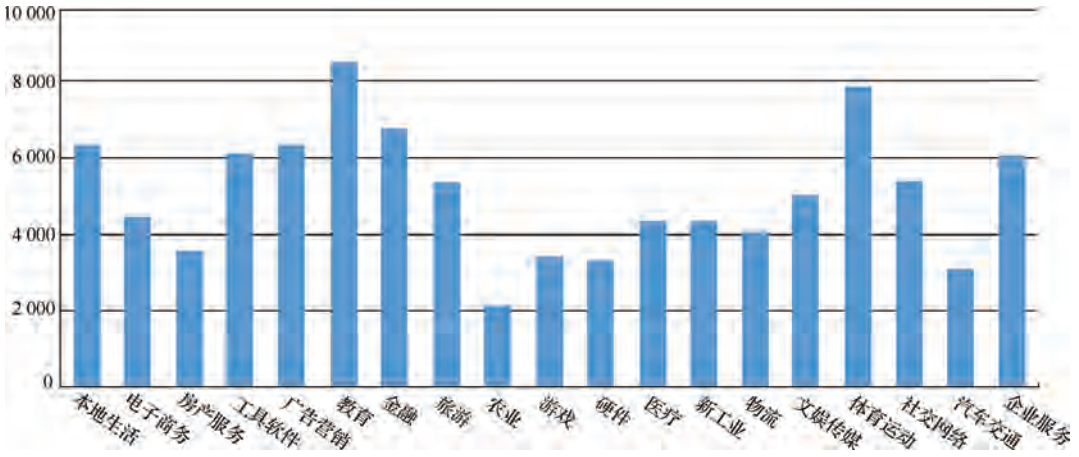

图 3.13　区间柱状图

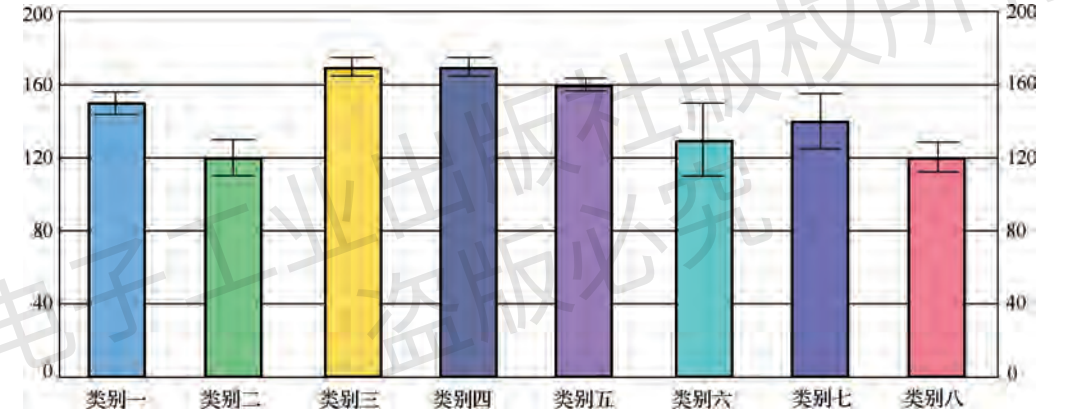

图 3.14　误差柱状图

3.4.3　散点图

散点图是同时呈现两个连续变量的图（见图 3.15），作直角坐标系，将两个变量分别映射到两个轴，描出所有数据点，大量的数据点呈现出了整体的分布趋势。散点图可以显示两个变量之间的相关性，但是相关性不代表因果性，也许两个变量之间不存在直接关系，而是同时受到另一个外部变量的影响。散点图可以画在三维坐标系中，还可以增加数据以反映大小、颜色、形状等多种因素的映射，从而同时呈现多个变量之间的关系。但显然，可视因素增多时会影响观察的效果。

可以对数据进行回归分析，也就是通过数学建模做一条光滑的曲线来模拟变量之间的因果关系，利用散点图可以观察回归分析的效果，如图 3.16 所示。

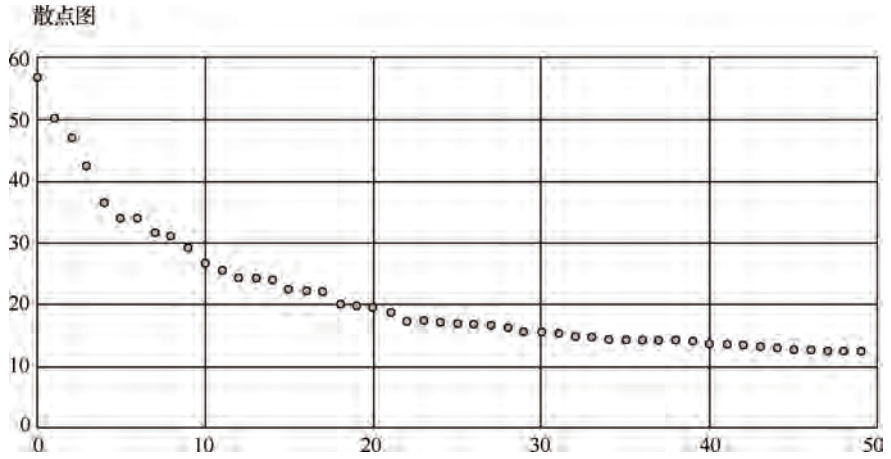

图 3.15　散点图

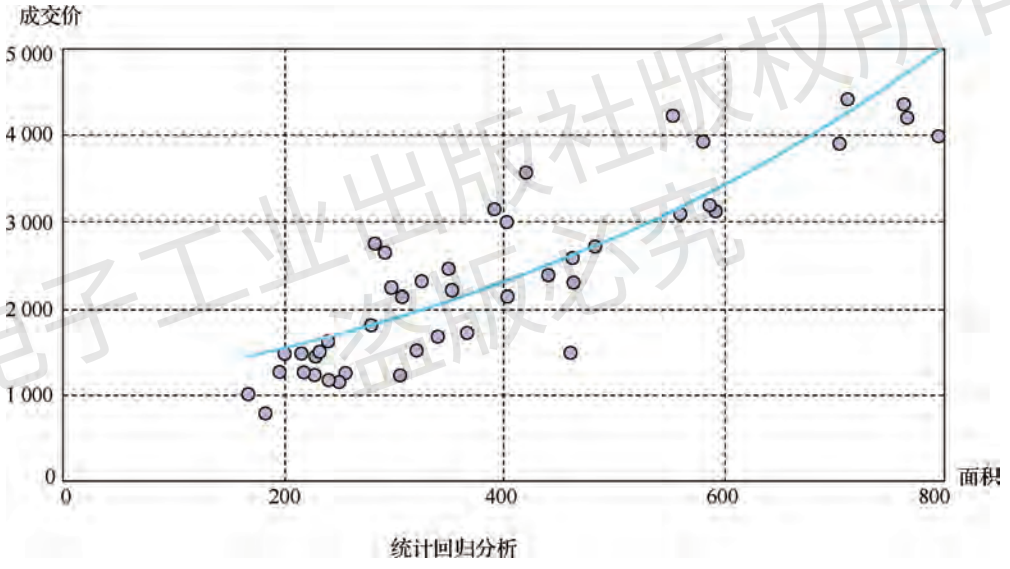

图 3.16　从散点图上观察回归曲线

3.4.4　折线图

折线图是作直角坐标系，用横轴表示有序分类变量(如时间序列变量)，纵轴表示连续变量，将数据点描出，并使用线段连接相邻数据点以后得到的图表(见图 3.17)。折线图可以显示连续变量随有序变量而变化的趋势。

阶梯折线图将折线变为直角阶梯状，用以呈现频率低而效果显著的变化，如税率的变化等，如图 3.18 所示。

当连续变量有积分的意义时，可以使用面积图来凸显累积的效果，如降雨量等，如图 3.19 所示。

图 3.17　折线图

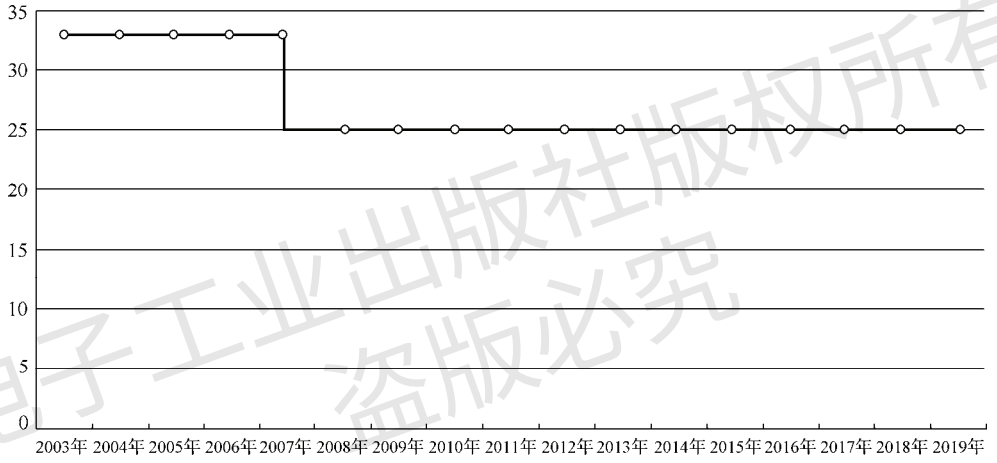

图 3.18　阶梯折线图

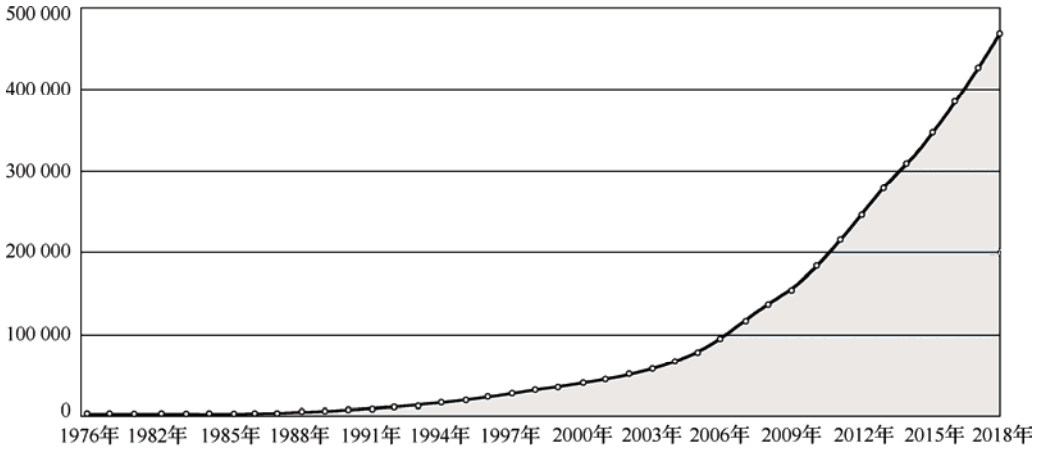

图 3.19　面积图

3.5　统计分析应用示例

本节我们通过一个人口统计的简单应用来熟悉本章所讨论的一些统计知识。

我国于 2020 年度进行了第七次人口普查。作为最精准的中国统计数据之一，人口普查可以为国家经济建设和制定经济社会发展规划、推动经济高质量发展提供准确的统计信息支持。

人口普查不同于一般的统计抽样调查，调查的对象是全部的样本，即中国大陆地区生活的常住人口。在人口普查之间的年份，国家统计局会根据统计方法，结合最近一次的人口普查结果和抽样调查的数据，给出相关的人口估算数据。那么，就让我们利用前面介绍的相关统计知识，通过国家统计局的官方数据[①]来看一看第七次人口普查前，我们能够获知的人口统计概况。这样，当第七次人口普查数据公布后，我们可以就两者的数据差异进行一个初步的比较。

3.5.1　人口变化总趋势

研究统计数据，首先就要明确统计对象，以及需要关注的维度，也就是统计对象的变量。对于人口而言，最基本的指标就是每个年度的人口数是多少。

我们可以从表 3.3 中直接看出中国人口近年来的增长变化，但具体的增长趋势和其他的人口结构却并不能从表格数据中看出。如果想了解更多的人口相关信息，我们就需要增加对人口数据的描述。有了丰富的人口数据，就便于我们进行多角度的图表构建，进行对人口的综合考量。

表 3.3　中国人口总数 2010—2019 年

年末总人口（万人）	年度
140 005	2019
139 538	2018
139 008	2017
138 271	2016
137 462	2015
136 782	2014
136 072	2013
135 404	2012
134 735	2011
134 091	2010

首先，我们将表中的数据绘制成总人口的折线图（见图 3.20）。可以看到，过去 10 年，中国人口呈现稳定增长的趋势，增长速度变化不大。在 2017 年左右，人口增速有明显的变缓趋势。当然，由于 2017—2019 年并非人口普查年，因此人口数据是根据模型估算出来的，我们不能确定这种人口增速的变化是不是由模型的改变导致的。

① 数据来自国家数据网 https://data.stats.gov.cn/.

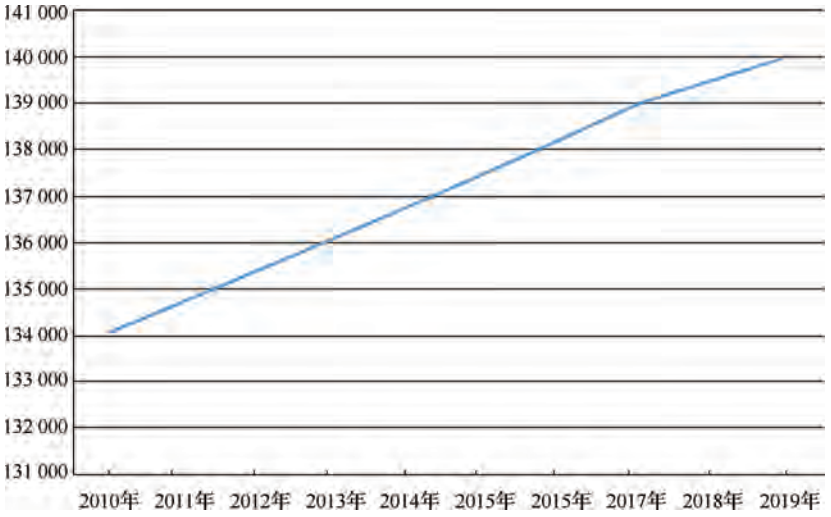

图 3.20　中国 2010—2019 年总人口折线图

3.5.2　人口结构变化

为了进一步了解人口结构的变化，我们用饼图查看城镇和乡村人口在起始年份和终端年份的变化，如图 3.21、图 3.22 所示。

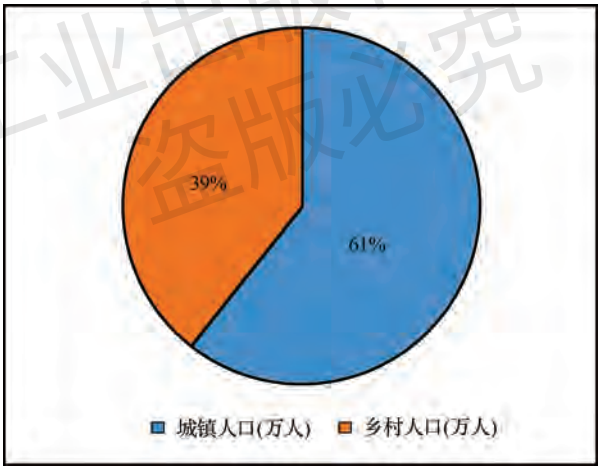

图 3.21　中国 2019 年城镇、乡村人口比例结构

不难看出，从 2010—2019 年，中国的城镇化总体上增长了 11%。显然，中国的城镇化进程还有很大的进步空间。

图 3.23 的条形图给出了 2011 年和 2019 年的人口结构对比。和 2011 年相比，2019 年里 60 岁以上人口占总人口的比例增多，而中年人的人口比例缩小，新生儿的比例持平。也就是说，中国的人口老龄化问题日渐凸显，而出生率并没有在过去几年得到显著提高，这导致了人口增长速度的变缓。同时，由于适龄劳动力人口的比例(15～64 岁)下降，直接影响了从农村转移到城市的劳动力人口数量。

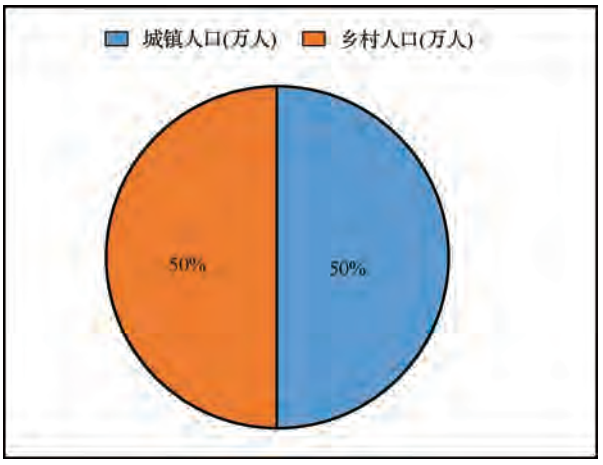

图 3.22　中国 2010 年城镇、乡村人口比例结构

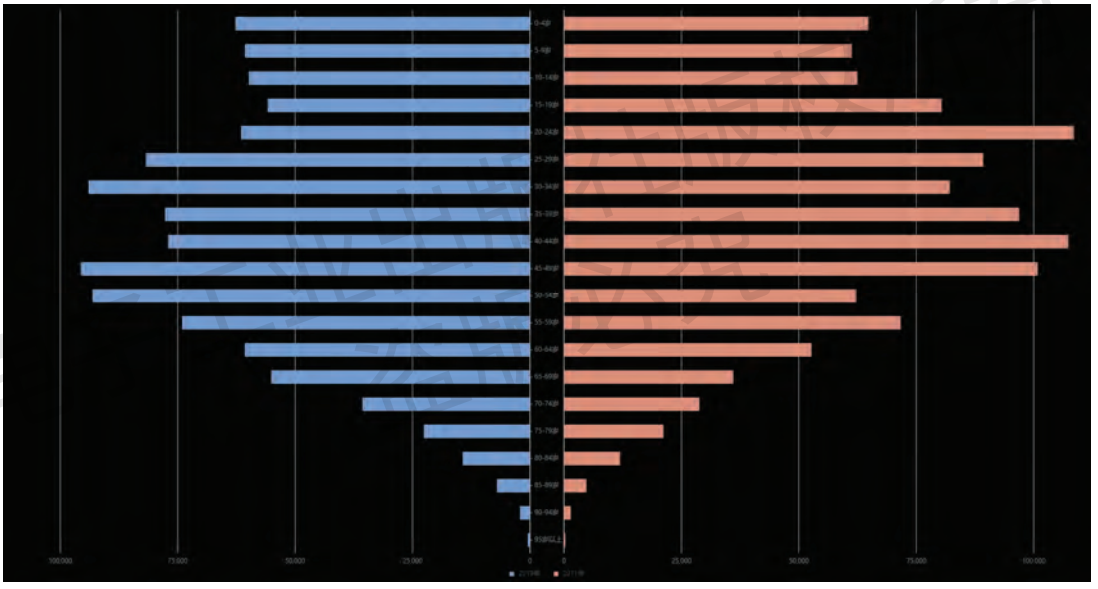

图 3.23　中国 2011 年、2019 年人口按年龄分比例结构

3.5.3　二胎与生育率

从人口数据也能够看出社会对二胎问题的反应。如图 3.24、图 3.25 所示，我们选取了孕龄妇女从 2011—2015 年的一胎、二胎生育率的变化。

总体而言，一胎生育率最高的妇女年龄段在 20～29 岁，并在 30 岁之后显著下降。从时间上来看，2015 年的一胎生育率较 2011 年降低很多。

从二胎的图结构来看，二胎分布的范围较一胎更均匀，集中在 20～35 岁的适孕妇女中，并且近年有所提高。然而整体上二胎的生育率要远低于一胎的生育率。虽然数据有限只记录到 2015 年，但是从人口分布结构图来看，生育率整体上在 2019 年并没有大幅度的提高。

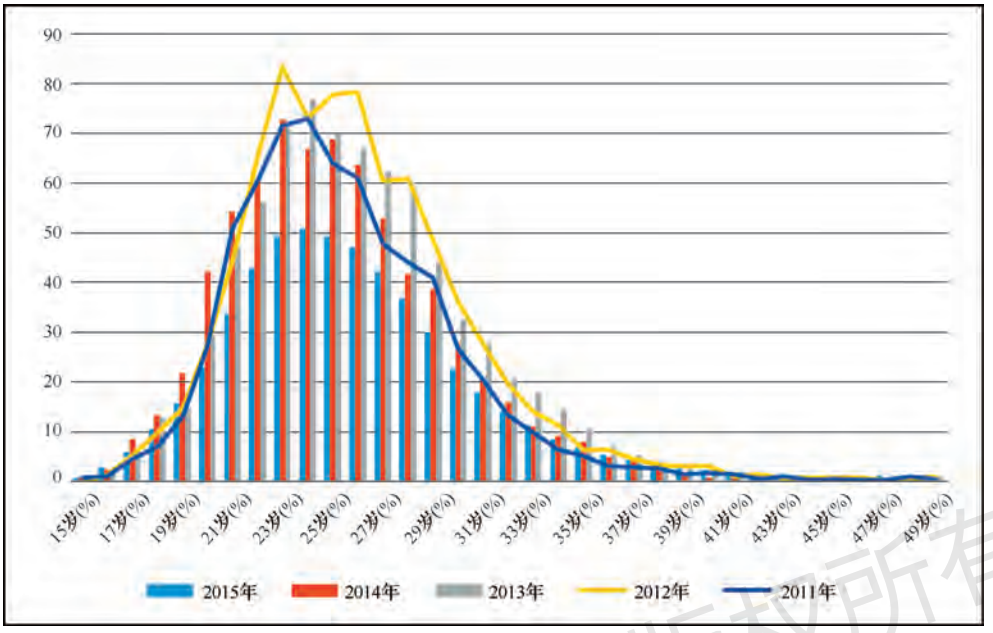

图 3.24　中国 2011—2015 年育龄妇女一胎生育率变化

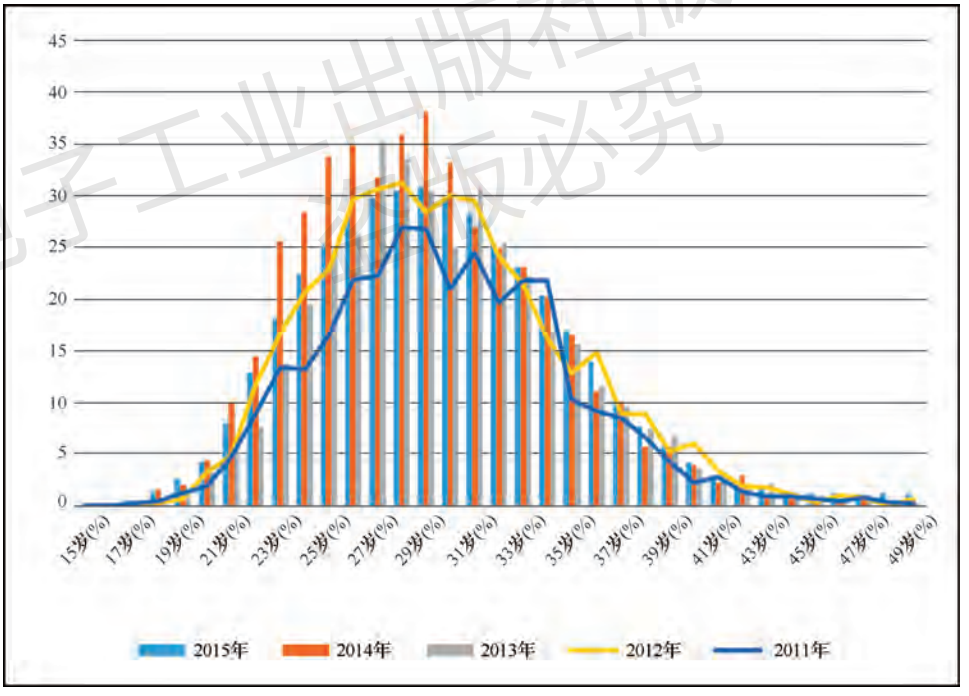

图 3.25　中国 2011—2015 年育龄妇女二胎生育率变化

综上，我们以中国人口数据为背景，用一些统计学的基本知识，讨论了和人口相关的几个社会问题。我们鼓励读者尝试应用统计学知识和统计图表，对其他问题进行分析和探索。