

某种事物的看法也是一个顺序变量,其变量值可以为“同意”、“保持中立”、“反对”等。

▮ 定义 1.19 说明事物数字特征的一个名称,称为数值型变量(metric variable)。

数值型变量的取值就是数值型数据。如“产品产量”、“商品销售额”、“零件尺寸”、“年龄”、“时间”等都是数值型变量,这些变量可以取不同的数值。数值型变量根据其取值的不同,又可以分为离散变量和连续变量。

▮ 定义 1.20 只能取可数值的变量,称为离散型变量(discrete variable)。

离散型变量只能取有限个值,而且其取值都以整位数断开,可以一一列举,如“企业数”、“产品数量”等就是离散变量。

▮ 定义 1.21 可以在一个或多个区间中取任何值的变量,称为连续型变量(continuous variable)。

连续型变量的取值是连续不断的,不能一一列举,如“年龄”、“温度”、“零件尺寸的误差”等都是连续变量。在对社会和经济问题的研究中,当离散变量的取值很多时,也可以将离散变量当作连续变量来处理。

变量这一概念以后经常要用到,但多数情况下所说的变量主要是指数值型变量,大多数统计方法所处理的也都是数值型变量。当然,也可以从其他角度对变量进行分类,比如随机变量(有关随机变量的含义将在第5章中再作介绍)和非随机变量、经验变量(empirical variables)和理论变量(theoretical variables)等。经验变量所描述的是在周围环境中可以观察到的事物。理论变量则是由统计学家用数学方法所构造出来的一些变量,比如后面的有些章节中将要用到的 z 统计量、 t 统计量、 χ^2 统计量、 F 统计量等都是理论变量。



思考与练习

思考题

- 1.1 什么是统计学? 怎样理解统计学与统计数据的关系?
- 1.2 统计数据可分为哪几种类型? 不同类型的数据各有什么特点?
- 1.3 简述分类数据、顺序数据和数值型数据的含义。
- 1.4 举例说明总体、样本、参数、统计量、变量这几个概念。
- 1.5 什么是有限总体和无限总体? 举例说明。

 人物传记^①

Adolphe Quetelet



Lambert Adolphe Jacques Quetelet (1796—1874) 出生于比利时的 Ghent。他早期对美术感兴趣(他画画、写诗,甚至创作歌剧),但数学的吸引力很快掩盖了他的这一兴趣。在新建的根特大学,他的博士论文是首篇有关解析几何的论文,由于对解析几何做出了独创性贡献,该论文得到了广泛称赞。也正因为这篇论文,Quetelet 在他 24 岁的时候入选布鲁塞尔皇家学会会员,并在布鲁塞尔雅典娜神庙教授数学、物理和天文学。他是一位伟大的教师(来自欧洲各地的学生与访问学者蜂拥而至),同时他又是一位伟大的作家(他创作了大批散文和书籍,并主编了一本重要杂志——*Correspondence Mathématique et Physique*(数学和物理学学报))。然而,Quetelet 在事业上所倾注的、看起来似乎永不倦怠的精力反而改变了他事业的方向。他对天文学的热爱使得他最终在布鲁塞尔兴建了一座天文台,并由他本人进行管理,而这些也让 Quetelet 有机会接触到许多杰出的法国数学家,如 Fourier, Laplace 和 Poisson。他们对概率论以及概率论在社会现象中的应用的兴趣使 Quetelet 感到兴奋。随后,他对收集社会经验数据的积极鼓励促成了比利时与荷兰的第一次全国人口普查(1829 年),促成了伦敦统计学会(现命名为皇家统计学会)的建立(1834 年),并促成了比利时中央统计委员会的组建(1841 年),后者是负责收集统计数据的主要中介。作为该委员会的会长,Quetelet 做了大量工作以激励欧洲各地统计局的建立,并不辞辛劳地推动数据收集与表示的国际统一方法和术语的形成。在他的领导下,国际统计大会第一次会议于 1853 年在布鲁塞尔举行。

① 资料来源:参考文献(Heinz, 1998, P. 28-29)。



思考与练习

思考题

- 2.1 简述普查和抽样调查的特点。
- 2.2 统计数据的具体收集方法有哪些？
- 2.3 简述调查方案的基本结构。
- 2.4 什么是问卷？它由哪几部分组成？
- 2.5 设计问卷的提问项目应注意哪些问题？
- 2.6 封闭型问题答案的设计有哪些主要方法？
- 2.7 问卷中问题顺序的设计应注意哪些问题？

练习题

- 2.1 假定我们要研究在校大学生的生活时间分配状况，请写出你的研究思路。
- 2.2 某家用电器生产厂家想通过市场调查了解以下问题
 - a. 企业产品的知名度
 - b. 产品的市场占有率
 - c. 用户对产品质量的评价及满意程度
 - (1) 请你设计出一份调查方案。
 - (2) 你认为这项调查采取哪种调查方法比较合适？
 - (3) 设计出一份调查问卷。



人物传记^①

William G. Cochran

W. G. Cochran(1910—1980)生于英国苏格兰的一个铁路雇员家庭，他自小聪颖好学，成绩斐然，曾多次获奖，深受师长喜爱。自幼所形成的良好学习习惯和坚实的知识基础为他以后获得格拉斯哥及

① 资料来源：倪加勋. 近代统计学家——W. G. 科克伦. 统计, 1987, 5。

表 3.23 某城市居民关注广告类型的频数分布

广告类型	人数(人)	频率(%)
商品广告	112	56.0
服务广告	51	25.5
金融广告	9	4.5
房地产广告	16	8.0
招生招聘广告	10	5.0
其他广告	2	1.0
合计	200	100

表 3.24 某城市居民关注广告类型的频数分布

广告类型	人数(人)	频率(%)
商品广告	112	56.0
服务广告	51	25.5
金融广告	9	4.5
房地产广告	16	8.0
招生招聘广告	10	5.0
其他广告	2	1.0
合计	200	100

按照你的喜欢程度对这 3 个表格作一个排序,你会得出什么结论? 是表 3.24、表 3.23、表 3.22 吗? 应该是! 表 3.22 有两点不足: 一是表中所用的线太多,有点儿喧宾夺主,线成了主角儿,数据成了配角儿;二是表格中列的间距太宽,这样会使你看数据时,眼睛从一侧到另一侧,距离有些远。表 3.23 所用的线就少多了,列与列之间的间隔也较合适,但那两条竖线就像两堵墙,妨碍你的眼睛从一列移动到另一列,总体来说还算可以。最简明的要算是表 3.24,优点不言而喻。

许多人使用统计表时还会犯的一个毛病就是丢三落四。比如,统计表没有表头,没有编号,不注明数据来源,表格中数据小数点的位数不统一,不标出计量单位,等等。这些都是不好的习惯。

思考与练习

思考题

3.1 数据的预处理包括哪些内容?

续表

月份	北京	长春	南京	郑州	武汉	广州	成都	昆明	兰州	西安
7	69	70	82	74	81	86	84	84	58	62
8	74	79	82	71	73	84	78	74	57	55
9	68	66	71	67	71	81	75	77	55	65
10	47	59	75	53	72	80	78	76	45	65
11	66	59	82	77	78	72	78	71	53	73
12	56	57	82	65	82	75	82	71	52	72

资料来源:《中国统计年鉴 1998》,北京:中国统计出版社,1998,第 10 页。



人物传记^①

John W. Tukey



John W. Tukey(1915—)起初是一位化学家,之后成为一位数学家,在第二次世界大战期间,通过他称做“真正的问题体验和真正的数据体验”的职业,转而学习统计。1937 年 Tukey 来到普林斯顿大学学习化学,却在 1939 年获得了数学博士学位。战争期间,他研究有关测距仪和轰炸机轰炸的精确度问题。战后,他把时间在普林斯顿大学和附近的贝尔实验室之间进行了分配,后者是当时世界一流的工业研究团队。

Tukey 把大量的精力投入到对那些数据复杂的零乱问题的统计研究中:医院里医生对病人使用的麻醉剂的安全问题,对人类性行为的全塞研究问题,执行核试验禁令的监控问题,以及空气质量和环境污染问题。

从这些“真正的问题体验和真正的数据体验”中,Tukey 发展了探索性数据分析。他发明了很多工具,例如箱线图和茎叶图。更重要的是,他发展了数据分析的哲学,改变了统计学家们的思维方式,因而被称为数据分析的哲学家。

^① 资料来源:参考文献(David,1999,p.78)。

表 4.9 Excel 中的描述统计函数

函数名	语 法	功 能
AVEDEV	AVEDEV (number1,number2,...)	计算平均差
AVERAGE	AVERAGE (number1,number2,...)	计算平均数
GEOMEAN	GEOMEAN(number1,number2,...)	计算几何平均数
HARMEAN	HARMEAN(number1,number2,...)	计算简单调和平均数
KURT	KURT(number1,number2,...)	计算峰态系数
MODE	MODE(number1,number2,...)	计算众数
MEDIAN	MEDIAN(number1,number2,...)	计算中位数
QUARTILE	QUARTILE(array,quart)	计算四分位数
SKEW	SKEW(number1,number2,...)	计算偏态系数
STDEV	STDEV(number1,number2,...)	计算样本标准差
STDEVP	STDEVP(number1,number2,...)	计算总体标准差
TRIMMEAN	TRIMMEAN(array,percent)	计算切尾均值

下面的图 4.5 总结了数据的分布特征和适用的描述性统计量。

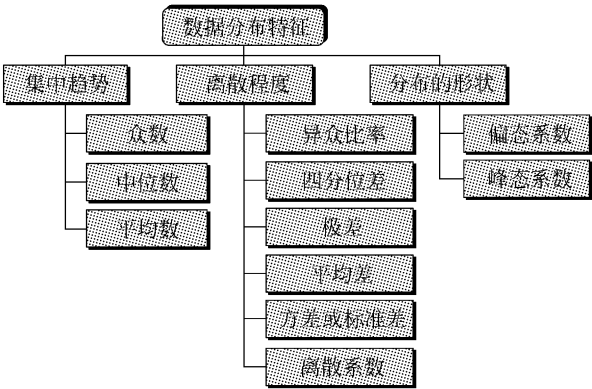
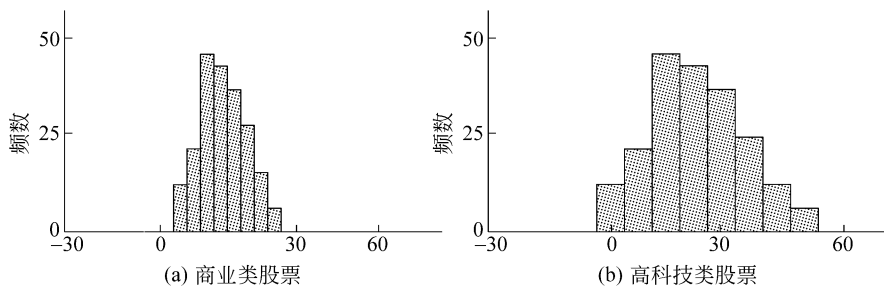


图 4.5 数据分布的特征与适用的描述统计量

思考与练习

思考题

- 4.1 一组数据的分布特征可以从哪几个方面进行测度？
- 4.2 怎样理解均值在统计学中的地位？
- 4.3 简述四分位数的计算方法。



人物传记^①

Pafnuty Lvovich Chebyshev



P. L. Chebyshev (1821—1894) 出生于俄国。他的父母属于贵族阶层,他接受了良好的私人教育;他很快迷上了数学,并最终在莫斯科大学攻读数学和物理。还在学生时代,Chebyshev 就因为他的一篇有关方程根计算问题的论文而获得了银质奖章,这篇论文现在很有名。他在圣彼得堡大学教授数学并开始对机械工程学产生强烈兴趣时撰写了许多杰出的论文,前者只是其中的一篇(在其他方面,他主要致力于弹道学的研究,带来了炮弹的多种创新,并发明了一台计算机)。他始终强调理论与实践的统一,他说:“远古时代数学就受到了人们特别的关注;如今,由于数学对工业和艺术的影响,它引起了人们更多的兴趣。理论与实践的统一能够带来十分有益的结果;并不只有实践这一方受益;在它的影响下科学也将前进,因为它能为科学发现新的研究对象,发现那些长期以来熟知的事物可供挖掘的新的方面。”

① 资料来源:参考文献(Heinz,1998,P.103-104)。

的时间间隔服从指数分布。

在本节结束之前将上面讨论过的三种常用连续型概率分布的数学期望和方差总结如下,由于其间涉及较复杂的积分运算,所以表 5.12 直接给出了各类概率分布数学期望与方差的计算结果,具体步骤读者可参看相关统计教程。

表 5.12 几种连续型概率分布的数学期望与方差

概率分布	数学期望 μ	方差 σ^2
正态分布	μ	σ^2
均匀分布	$\frac{a+b}{2}$	$\frac{1}{12}(b-a)^2$
指数分布	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$



思考与练习

思考题

- 5.1 简述试验与事件的含义。
- 5.2 简述样本空间的含义。
- 5.3 简述概率与频率的关系。
- 5.4 什么是互斥事件? 互斥事件的加法规则是什么?
- 5.5 概率具有哪些基本性质?
- 5.6 何谓条件概率?
- 5.7 两个事件独立意味着什么?
- 5.8 什么是随机变量? 它有哪些类型?
- 5.9 什么是离散型随机变量的概率分布?
- 5.10 两点分布与二项分布有什么不同?
- 5.11 简述 n 重伯努利试验的条件。
- 5.12 超几何分布的适用场合是什么?
- 5.13 简述概率密度函数和分布函数的含义。
- 5.14 判断数据正态性的方法主要有哪些?
- 5.15 简述标准正态概率图的绘制步骤。
- 5.16 在什么条件下用正态分布近似计算二项分布的概率效果比较好?

5.16 设 X 服从 $\lambda=0.5$ 的指数分布。求：

(1) $P(X>2)$ 。(2) $P(X\leq 3)$ 。

5.17 某电话室公用电话每次的通话时间长度(单位: min)服从如下的概率分布：

$$f(x) = \begin{cases} \frac{1}{5}e^{-\frac{1}{5}x}, & x \geq 0, \\ 0, & \text{其他。} \end{cases}$$

当走进电话室时,若恰好有人开始打电话,计算下列几个事件发生的概率：

(1) 等待时间不超过 2min。

(2) 等待时间为 3~5min。

5.18 某公司决定对职员增发“销售代表”奖,计划根据过去一段时期内的销售状况对月销售额最高的 5% 的职员发放该奖金。已知这段时期每人每个月的平均销售额(单位: 元)服从均值为 40000、方差为 360000 的正态分布,那么公司应该把“销售代表”奖的最低发放标准定为多少元?



人物传记^①

James Bernoulli



James Bernoulli (1654—1705) 也被称为 Jacques 或 Jakob, 出生于瑞士巴塞尔, 他的家族由于具有新教徒信仰而在几个世纪前逃亡到安特卫普, 这也使得他的家族在三代人之中产生了 9 个一流的数学家。James 本人被安排攻读神学并完成了学业, 但是, 他很快违背了父亲的意愿, 转而研究数学和天文学。在途经荷兰、法国和英国的旅行中, 他熟识了这些科学的现状, 并带回了他人有关彗星和地球引力的理论。不久以后, 他被委派到巴塞尔大学担任数学教授。受到 G. W. Leibniz 有关无穷小微积分学研究的启示, James 用它

^① 资料来源: International Encyclopedia of Statistics, New York: The Free Press, 1978(1): 18-19。

第2步: 在B3单元格输入公式“=FINV(\$B\$1,B\$2,\$A3)”, 然后将其向下、向右复制即可得到 χ^2 分布的临界值表, 结果如图6.14所示。

B3		=FINV(\$B\$1,B\$2,\$A3)									
	A	B	C	D	E	F	G	H	I	J	K
1	$\alpha =$	0.05									
2	df2/df1	1	2	3	4	5	6	7	8	9	10
3	1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
4	2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
5	3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
6	4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
7	5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
8	6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
9	7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
10	8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
11	9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
12	10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
13	11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
14	12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
15	13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
16	14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
17	15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
18	16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
19	17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
20	18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
21	19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
22	20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348

图 6.14 F 分布的临界值表

除上面介绍的 χ^2 分布和F分布外, 在参数估计和假设检验中还用到另一个重要的小样本分布, 即t分布。有关它的知识将在参数估计一章中作介绍。



思考与练习

思考题

- 解释总体分布、样本分布和抽样分布的含义。
- 解释中心极限定理的含义。
- 重复抽样和不重复抽样相比, 抽样均值抽样分布的标准差有何不同?
- 样本均值的分布与总体分布的关系是什么?
- 样本方差和两个样本的方差比各服从什么分布?
- χ^2 分布和F分布的图形各有什么特点?

练习题

- 从均值为200、标准差为50的总体中, 抽取 $n=100$ 的简单随机样