



第四部分

统计



一、统计学★★★

分支	内容	
描述统计	含义	描述统计是研究数据 收集 、 整理 和 描述 的统计学方法。
	内容	①如何取得所需要的数据； ②如何用图表或数学方法对数据进行整理和展示； ③如何描述数据的一般性特征。
推断统计	含义	推断统计是研究如何利用样本数据来推断总体特征的统计学方法。
	内容	① 参数估计 ：利用样本信息推断总体特征； ② 假设检验 ：利用样本信息判断对总体的假设是否成立。



二、变量和数据★★★

1.变量	内容	
定量变量 (数量变量)	变量的取值是数量。 如企业销售额、注册员工数量等。	
定性变量	分类变量	变量的取值表现为类别，如企业所属行业。
	顺序变量	变量的取值表现为类别且有一定顺序，如员工受教育水平。



二、变量和数据★★★

2. 数据	概念	案例
分类数据	对 分类变量 的观测结果，表现为类别，一般用 文字 来表述，也可用 数值代码 表示。	比如用1表示“男性”，用2表示“女性”。对分类数据可以计算出各类别的频率。
顺序数据	对 顺序变量 的观测结果，表现为类别，一般用 文字 表述，也可用 数值代码 表示。	比如用1表示“硕士及以上”，用2表示“本科”，用3表示“大专及以下”。
数值型数据	对 定量变量 的观测结果，表现为 具体大小或数值 。	对数值型数据可以进行数学运算，比如计算均值和方差等统计量。



三、数据的来源★★

来源	分类	内容	举例
收集方法	观测数据	通过直接调查或测量而收集到的数据。	几乎所有与 社会经济现象 有关的统计数据都是观测数据，如GDP、CPI、房价等。
	实验数据	通过在 实验 中控制实验对象以及其所处的实验环境收集到的数据。	自然科学领域 的数据大多都是实验数据。如：一种新产品使用寿命的数据；一种新药疗效的数据。
使用者的角度	直接数据 (一手数据)	来源于 直接的调查和科学实验的数据 。 对数据使用者来说是 直接来源 （调查或观察、实验）。	
	间接数据 (二手数据)	来源于 别人的调查或实验的数据 。 对数据使用者来说是 间接来源 。	



四、统计调查的方式★★★

1. 统计报表

统计报表是按照国家有关法规的规定，自上而下地统一布置、自下而上地逐级提供基本统计数据的一种调查方式。

2. 普查【口诀】范围窄，一定准

(1) 通常是一次性的或周期性的。

(2) 一般需要规定统一的标准调查时间，以避免调查数据的重复或遗漏，保证普查结果的准确性。

(3) 数据一般比较准确，规范化程度也较高，可以为抽样调查或其他调查提供基本依据。

(4) 使用范围比较窄，只能调查一些最基本及特定的现象。



四、统计调查的方式★★★

普查	年份	具体时间
经济普查	每10年进行2次， 年份末尾数3、8。	1月1日0时
人口普查	每10年进行1次， 逢“0”的年份进行。	前四次：7月1日0时， 第五次、第六次：11月1日0时
农业普查	每10年进行1次， 逢“6”的年份进行	1月1日0时



四、统计调查的方式★★★★

3. 抽样调查（应用最广泛）

- （1）经济性（最显著优点）
- （2）时效性强
- （3）适应面广
- （4）准确性高

4. 重点调查

选择重点单位，调查的标志值在总体中占绝大比重。

以较少的投入、较快的速度取得某些现象主要标志的基本情况或变动趋势。

5. 典型调查

有意识地选择若干具有典型意义的或有代表性的单位进行的调查。



五、数据科学与大数据★★★

项目	内容		
大数据	数据量大、数据多样化、价值密度低、数据的产生和处理速度快、真实性。 【口诀】大多低快真		
数据挖掘	监督学习	分类	逻辑斯特回归、决策树、随机森林和支持向量机 【口诀】逻辑支持随机决策
		回归	线性回归、非线性回归和分位数回归等。
	无监督学习	聚类	基于划分的方法、 基于分层的方法、基于密度的方法、 基于网格的方法、基于模型的方法
		降维	主成分分析法、因子分析法
	半监督学习	半监督分类、半监督回归、半监督聚类	



六、数据分布特征的测度★★★

1. 集中趋势的测度

项目	均值	中位数	众数
适用范围	适用于定量变量。	适用于顺序变量和定量变量，特别是分布不对称的数据。	适用于分类和顺序变量。
优点	能够充分利用数据的全部信息，比较稳定。	不受极端值的影响。	不受极端值的影响。
缺点	易受极端值的影响。	没有充分利用数据的全部信息，稳定性差于均值，优于众数。	没有充分利用数据的全部信息，缺乏稳定性，而且可能不唯一。



六、数据分布特征的测度★★★

2.离散程度的测度			
项目	方差	标准差	离散系数
内涵	方差是数据组中各数值与其均值离差平方的平均数。	标准差即方差的平方根。 $s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$	离散系数也称变异系数或标准差系数，即标准差与均值的比值，记为CV。 $CV = \frac{s}{\bar{X}}$
注意	方差的单位是原数据单位的平方，没有解释意义。	标准差不仅能度量数值与均值的平均距离，还与原始数值具有相同的计量单位。	主要用于不同类别数据离散程度的比较。
	标准差与方差计算简单、有较好的数学性质，是应用最广泛的统计离散程度的测度方差。		



六、数据分布特征的测度★★★

3. 分布形态的测度		
偏态系数	0	数据的分布是对称的
	正值	0 ~ 0.5 轻度右偏；0.5 ~ 1 中度右偏； > 1 严重右偏
	负值	0 ~ -0.5 轻度左偏；-0.5 ~ -1 中度左偏； < -1 严重左偏
标准分数 (Z 分数)	$Z_i = \frac{X_i - \bar{X}}{s}$ ① 约有 68% 的数据与平均数的距离在加减 1 个标准差的范围之内； ② 约有 95% 的数据与平均数的距离在加减 2 个标准差的范围之内； ③ 约有 99% 的数据与平均数的距离在加减 3 个标准差的范围之内。 【口诀】168，295，399	



七、相关系数★★★★

相关系数的取值范围在+1和-1之间，即 $-1 \leq r \leq 1$ 。

相关系数	变量X和Y之间的关系	
$0 < r \leq 1$	正线性相关关系	
$-1 \leq r < 0$	负线性相关关系	
$r = 1$	完全正线性相关	当 $ r = 1$ 时，变量Y的取值完全依赖于X
$r = -1$	完全负线性相关	
$r = 0$	不存在线性相关关系，但并不说明变量之间没有任何关系，应结合散点图合理解释。	



七、相关系数★★★

相关程度【口诀】358

$ r $ 的取值	两变量之间的相关程度
$ r < 0.3$	相关程度极弱，可视为无线性相关关系
$0.3 \leq r < 0.5$	可视为低度相关
$0.5 \leq r < 0.8$	可视为中度相关
$ r \geq 0.8$	可视为高度相关