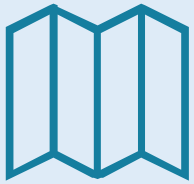


带你走入统计学

统计与数据科学学院 马俊玲

专题四
数据的初步分析

专题四 数据的初步分析

一、数据图形可视化

二、数据描述性分析

三、玫瑰图案例

四、案例1：知识图谱

五、案例2：地理信息系统

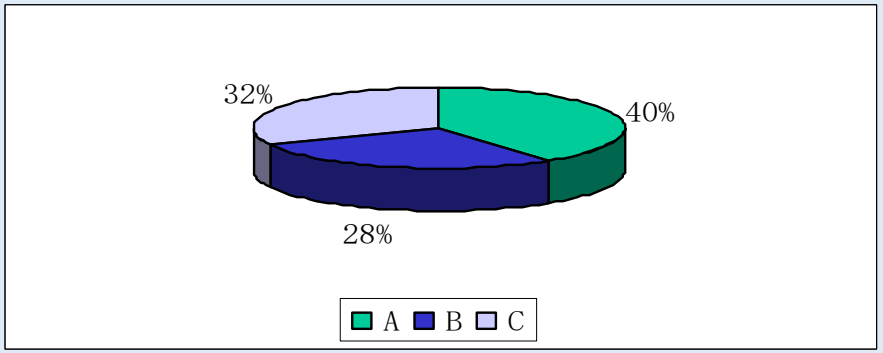
一、 数据图形可视化

饼图、条形图、直方图、折线图、曲线图、散点图、箱线图等。

饼图

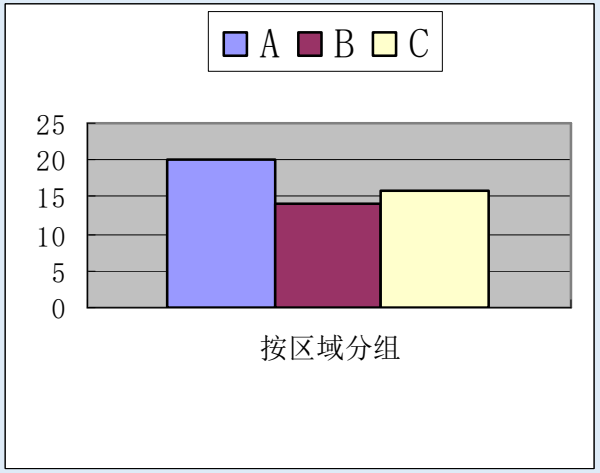
饼图是以整个圆的360度代表全部数据的总和，按照各类组所占的百分比（频率），把一个“饼”切割为各个扇形。适用于定性数据。

50家门店按区域分组的饼图

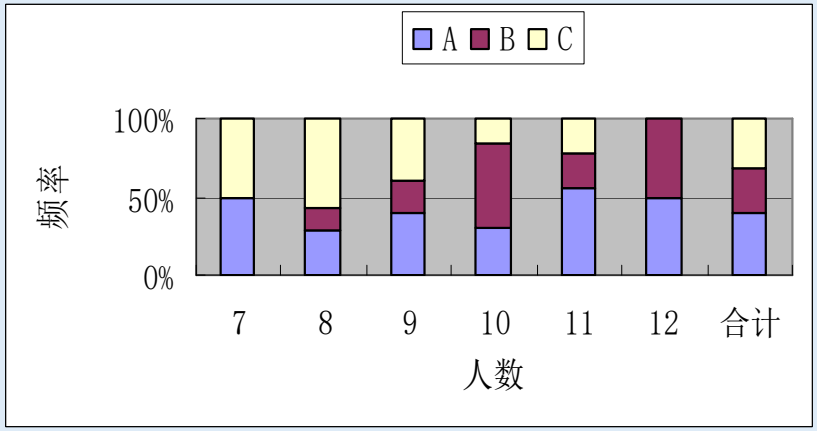


条形图

条形图中，每一分类组表示成一个条，条的长度代表了这个组中所含数据的频数或频率。适用于定性数据。



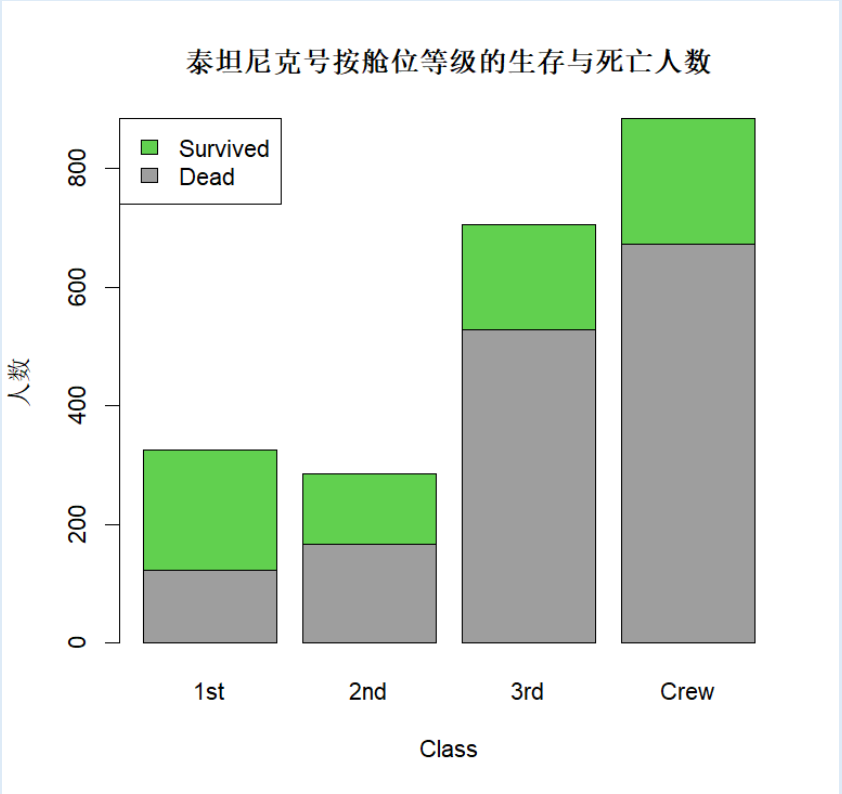
50家门店数按区域分组的条形图



50家门店按区域并按人数分组的分段比例条形图

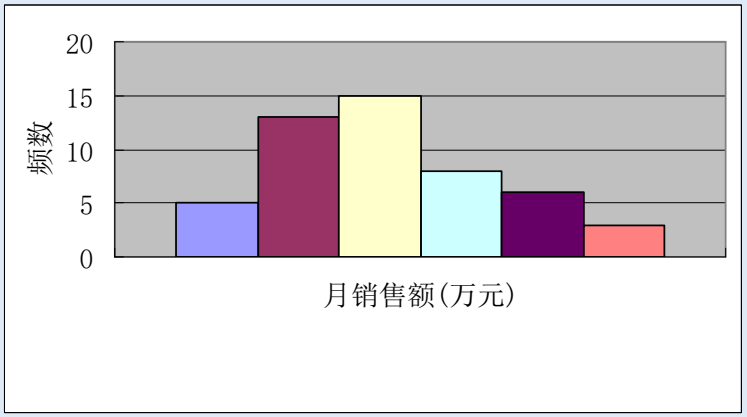
泰坦尼克号生存和死亡人数条形图

Survived	Class	Freq
No	1st	122
Yes	1st	203
No	2nd	167
Yes	2nd	118
No	3rd	528
Yes	3rd	178
No	Crew	673
Yes	Crew	212
	人数总计	2201



直方图

直方图是在每个分组区间上绘制一个长条形而产生的图形，它可以用来描述已表示成频数或频率的数据。适用于定量数据。

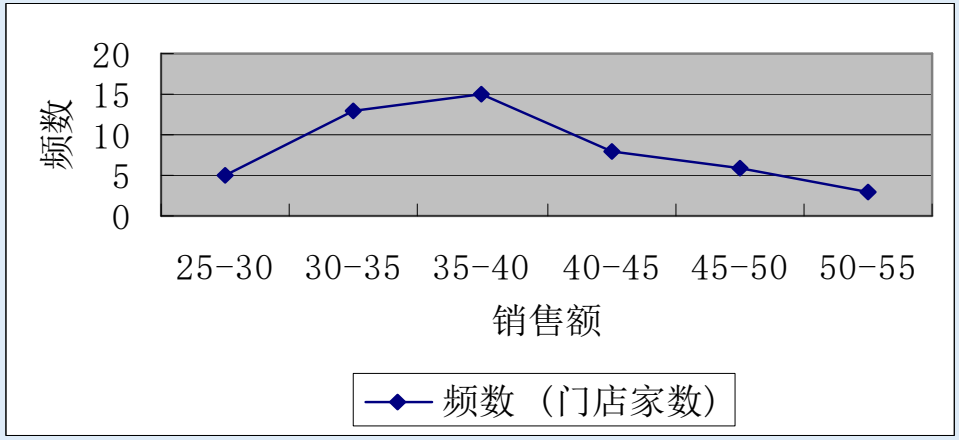


根据50家门店分组的频数分布表 绘制的直方图

★对于异距数列，以组距为宽，以频数密度为高来绘制直方图。

$$\text{频数密度} = \frac{\text{某组的频数}}{\text{该组的组距}}$$

折线图可以在直方图基础上，将每个长方形的顶端中点用折线连接而成，或用组中值与频数（或频率）求坐标点连接而成 。



根据直方图绘制的折线图

曲线图当变量的取值非常多，变量数列的组数无限增多时，折线便趋于一条平滑的曲线，这是一种概括描述变量数列分布特征的理论曲线。

散点图

箱线图是基于五数概括作出的图，即用五个数来概括一批数据的分布特征。（下文介绍）

二、数据的描述性分析

（一）集中趋势的测定

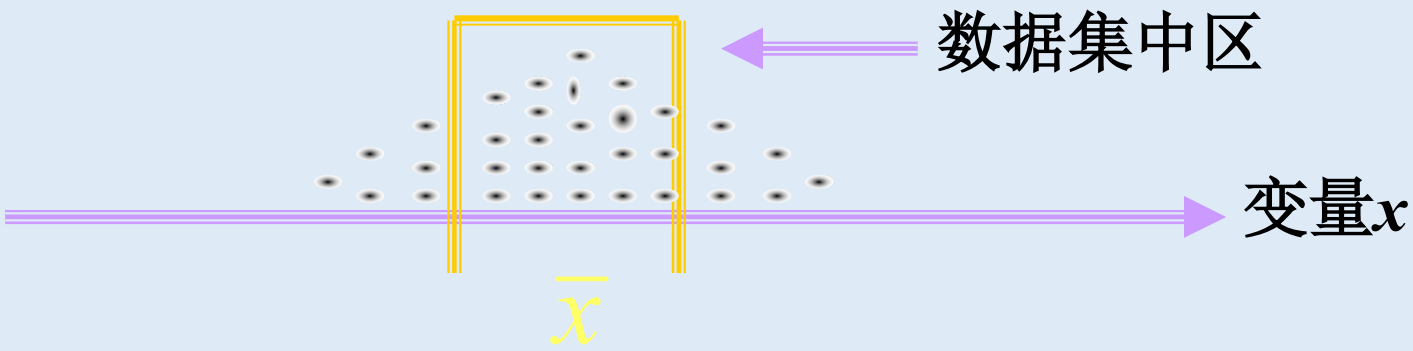
（二）离散趋势的测定

（三）数据的形态测定

（一）集中趋势的测定

概念

表明同类现象在一定时间、地点、条件下所达到的一般水平，是总体内某个变量大小各异的观察值的代表性数值。也是对变量分布集中趋势的测定。



常用的几种集中趋势指标

	概　念	特　　点
1. 算术平均数 $\bar{x}$	一个变量的所有观察值相加，再除以观察值的个数	优点： ①容易理解，便于计算 ②灵敏度高 ③稳定性好 缺点： ①易受极值影响 ②在偏斜分布和U形分布中，不具有代表性

常用的几种集中趋势指标

	概　念	特　　点
2. 中位数（M_e）	是一种位置平均数，数据按大小顺序排列，处于数据序列中间位置的数值就是中位数	优点： ①容易理解 ②不受极值影响 缺点： 灵敏度和计算功能差

常用的几种集中趋势指标

	概　念	特　　点
3. 众数（M_o）	是一种位置平均数，是一批数据中出现次数最多的那个数值. 通常只用于定性数据或离散型的定量数据。	优点： ①容易理解 ②不受极值影响 缺点： ①灵敏度和计算功能差 ②稳定性差 ③具有不唯一性

几种平均数的比较与联系

1.众数适用于所有的定性数据和定量数据
中位数适用于定性数据中的定序数据和定量数据
算术平均数只适用于定量数据

2.定量数据:若是钟形分布，三种集中趋势指标一般都可适用。

3. 在确定集中趋势指标的过程中，算术平均数比中位数和众数使用了更多的数据信息 。

（二）离散趋势的测定

- 概念

  标志变异指标是反映变量分布离散趋势、与平均指标相匹配的指标。

- 作用

  （1）反映变量分布的离散趋势；

  （2）是对平均数的代表性程度的量度；

  （3）是对事物发展均衡性的量度。

常用的几种标志变异指标

	概　念	计 算方法	特　点
1．异众比率	是非众数组所占比重	如某便利超市公司50家门店按区域划分的众数是A区域，该组的次数是20家，所以异众比率为60%，这说明50家门店按区域划分的离散程度比较大，众数的代表性较差。	异众比率是反映定性数据离散趋势的唯一指标，这个指标越小，说明数据的离散程度越小，集中程度越大

常用的几种标志变异指标

	概　念	计　算	特　点
2．极差（*R*）	数列中最大值与最小值之差	*R*=最大值-最小值 *R*=最大组的上限-最小组的下限	优点：容易理解，计算方便 缺点：不能反映全部数据分布状况
3．四分位差	是一批数据中的第三四分位数与第一四分位数之差的二分之一	$(M_3- M_1)/2$	在反映数据的离散程度方面比全距较为准确，但仍显粗略

常用的几种标志变异指标

	概　念	计　算	特　点
4．方差（σ^2 s^2）和 标准差(σ s)	所有观察值与平均数离差平方平均数的平方根，亦称均方差。标准差的平方即为方差。	简单：$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N}$ $S^2 = \frac{\Sigma(x - \bar{x})^2}{n-1}$ $\sigma = \sqrt{\sigma^2}$ $s = \sqrt{s^2}$	优点：反映全部数据分布状况，数字上合理。 缺点：受计量单位和平均水平影响，不便于比较

常用的几种标志变异指标

	概　念	计　算	特　点
5．变异系数（V_σ）	标准差与均值之商，是无量纲的	$V_\sigma = \frac{\sigma}{\mu}$ $V_S = \frac{S}{\overline{X}}$	两列数据的分布进行离散程度的比较，当它们的平均数不等、计量单位不同时则应消除平均数不同和计量单位不可比的影响。此时就需要用离散系数这种相对数来是测定离散趋势

方差（σ^2）和标准差（σ）是应用最广的标志变异指标

（三）数据的形态测定

- **偏度:是测定数据分布的偏斜程度的指标。**

偏度系数 $\alpha = \frac{m_3}{\sigma^3}$

<0负偏态
=0对称分布
>0正偏态

- 峰度：是用来反映数据分布曲线顶端的尖峭或扁平程度的指标。

峰度系数 $\beta = \frac{m_4}{\sigma^4}$

<3平顶曲线
=3正态曲线
>3尖顶曲线

注：在软件中很多是在上述公式基础上再减3。故利用软件时要注意。如R、EXCL中的峰度均是减3的数值。

- 五数概括 ：即最小值x_{min} 、最大值x_{max} 、第一四分位数M_1、中位数Me和第三四分位数M_3

五个数之间的关系，确定数据分布形态的方法：

- 数据是完全对称 ：

  最小值x_{min}到中位数的距离等于中位数到最大值x_{max}的距离。

  从x_{min}到M_1的距离等于M_3到x_{max}的距离。

- 数据是不对称 ：

右偏分布

从x_{max}到中位数的距离大于中位数到x_{min}的距离。

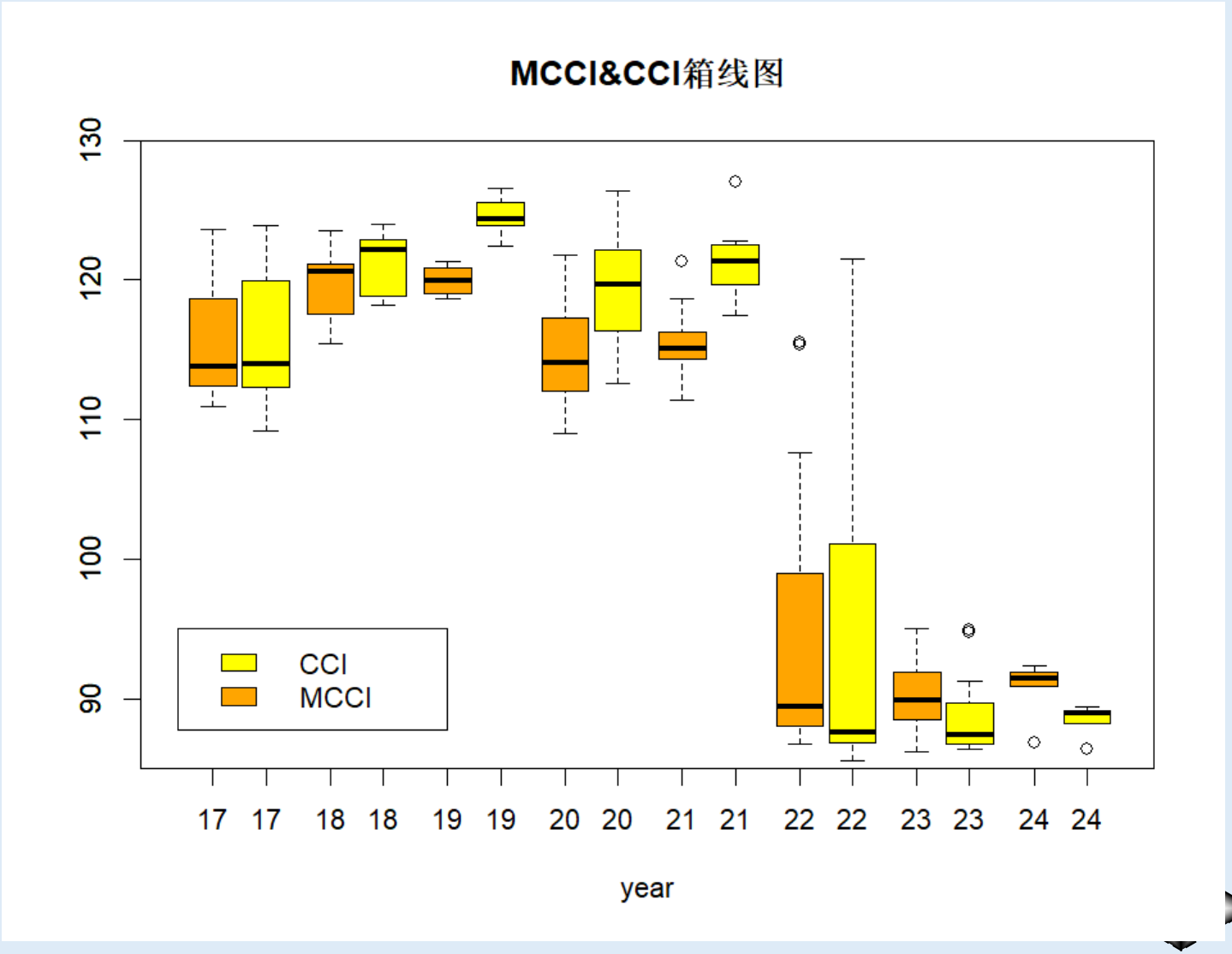
从M_3到x_{max}的距离大于从从x_{min}到M_1的距离。

左偏分布

从x_{min}到中位数的距离大于中位数到x_{max}的距离。

从x_{min}到M_1的距离大于M_3到x_{max}的距离。

- 箱线图:是基于五数概括的图示方式，使得集中趋势、离散趋势和偏态更为直观。


MCCI&CCI箱线图
130
120
110
100
90
CCI
MCCI
17 17 18 18 19 19 20 20 21 21 22 22 23 23 24 24
year


乌克兰
ROMANIA
欧
洲
SERBIA
BULGARIA
MACEDONIA
GREECE
Odessa
Rostov-on-Don
Sea of Azov
俄罗斯
克里米亚半岛
Kerch
Sevastopol
Novorossiysk
Constanta
Yalta
塞瓦斯托波尔
Varna
Burgas
黑 海
Sochi
亚
洲
GEORGIA
Batumi
Istanbul
Zonguldak
Samsun
Trabzon
Sea of Marmara
Aegean Sea
土耳其
Athens
Mediterranean Sea
NORTH
0
125
250
500
Kilometres

半岛，
里木半岛”。

6

-1856，克亚战争
国、法国、其和俄国）

在克里米亚战争中，战地护士南丁格尔在统计英国士兵的死亡情况后发现伤士死亡率高达42%，而由于医疗卫生条件恶劣导致的死亡人数大大超出直接阵亡人数。1855年卫生委员会来到医院改善整体的卫生环境后，死亡率才戏剧性地降至 2.5%。当时的南丁格尔注意到这件事，认为政府应该改善战地医院的条件来拯救更多年轻的生命。

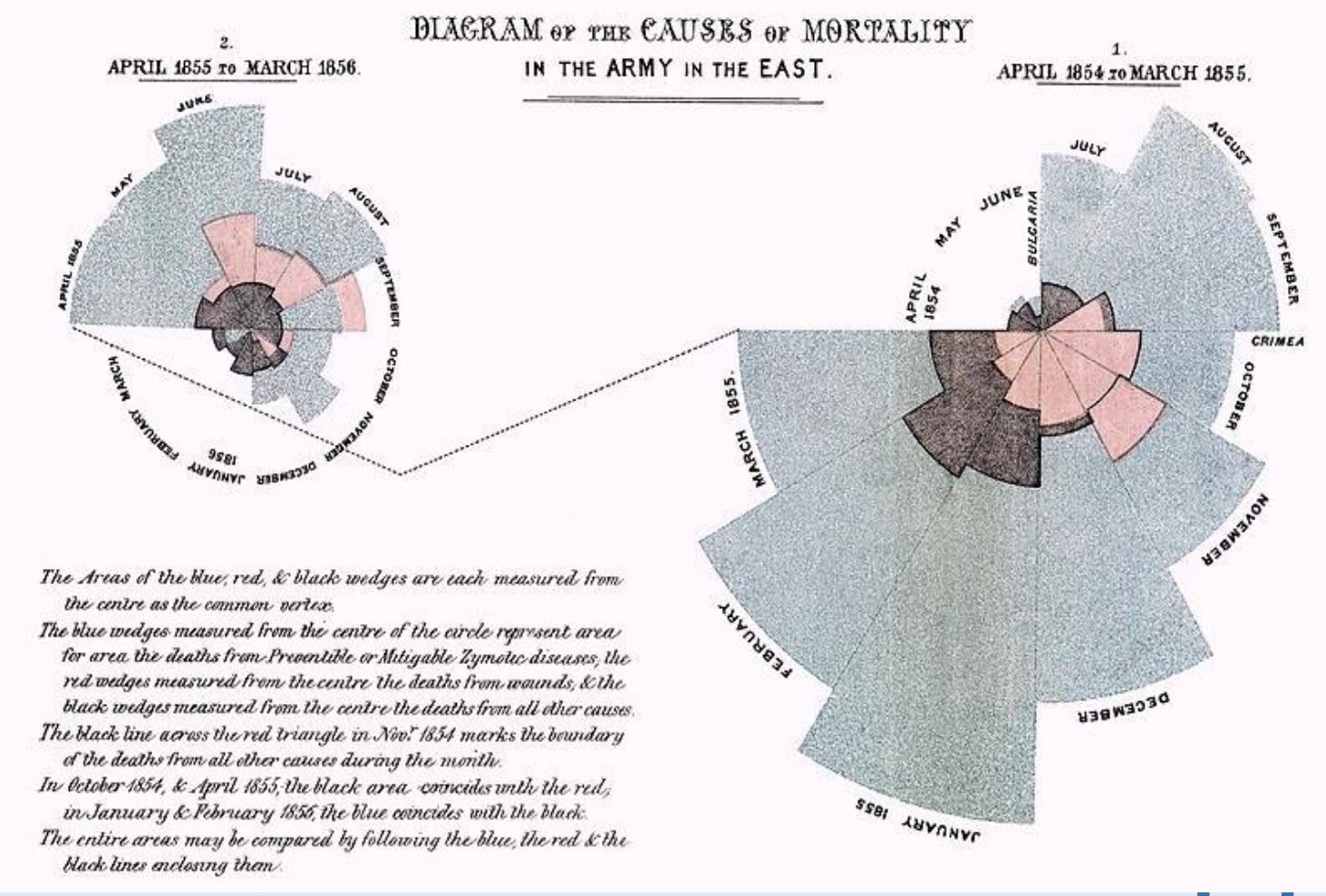
南丁格尔根据父亲传授的统计知识，将“战斗死亡”和“非战斗死亡”两类死亡人数绘制成统计图表，一目了然地显示了两者人数的悬殊对比，英国政府在社会的强烈反响下迅速建立了世界上第一座野战医院，挽救了无数战士的生命。

英国护士和统计学家弗罗伦斯·南丁格尔（1820.5.12－1910.8.13）

南丁格尔绘制的这份类似饼干的图表就是世界历史上的第一份“极区图”，也叫南丁格尔玫瑰图，是古典统计学中的经典图表之一。

南丁格尔玫瑰图将柱图转化为更美观的饼图形式，是极坐标化的柱图，其夸大了数据之间差异的视觉效果，适合展示数据原本差异小的数据。

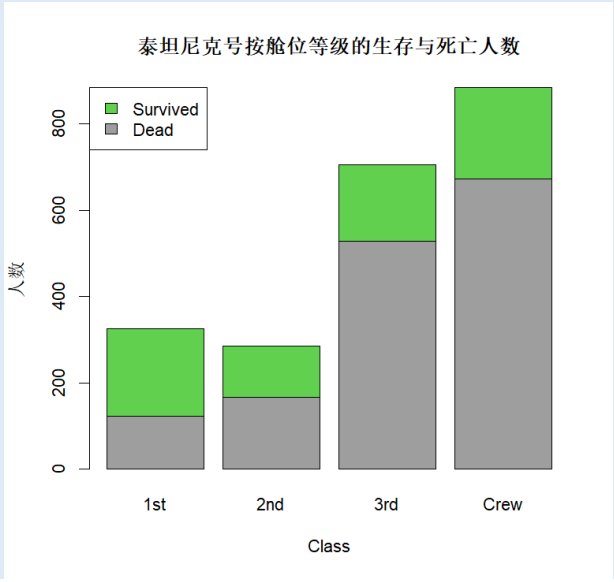
不同于饼图用角度表现数值或占比，南丁格尔玫瑰图使用扇形的半径表示数据的大小，各扇形的角度则保持一致。

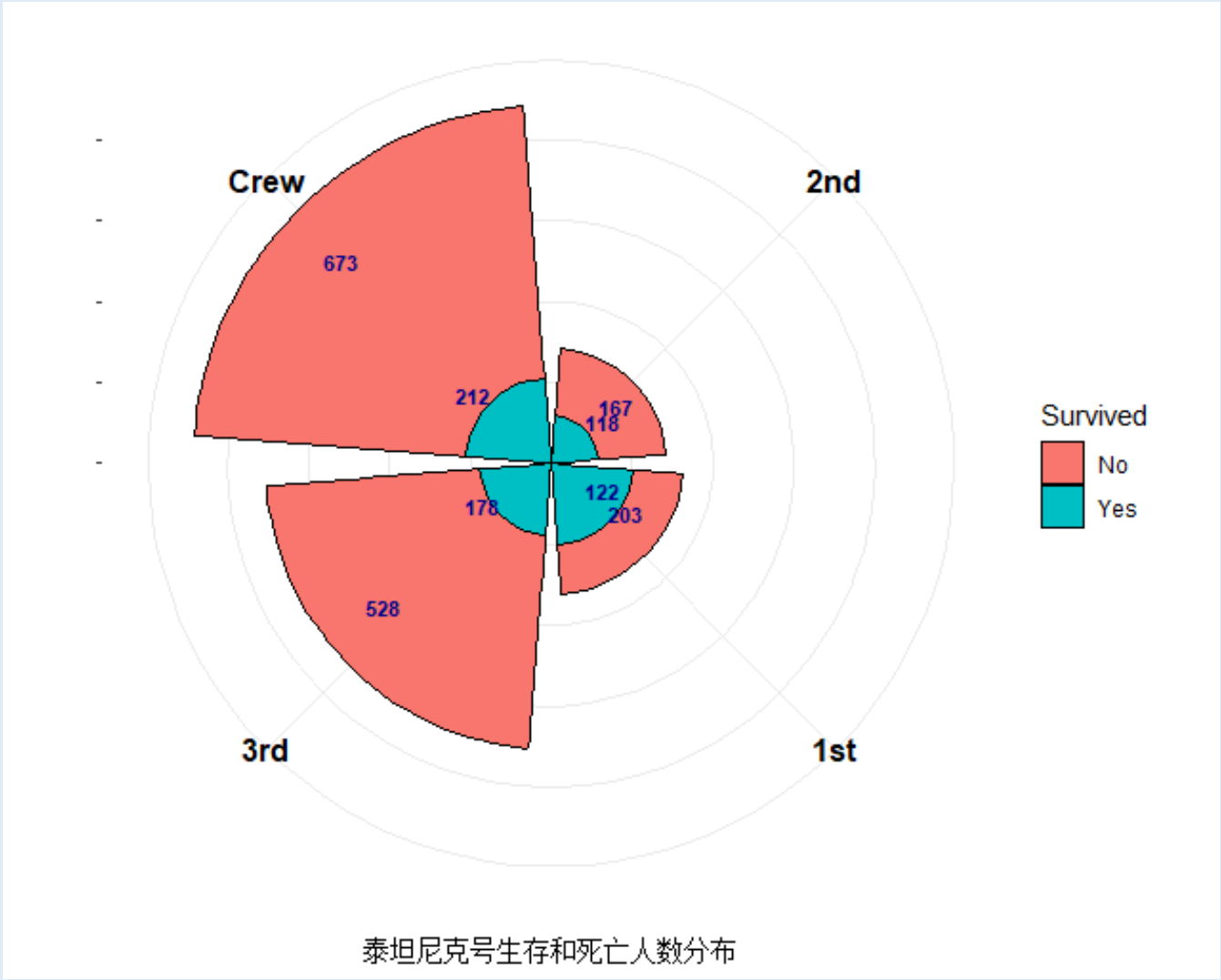


- 各色块圆饼区均由圆心往外的面积来表现数字
- 蓝色区域：死于原本可避免的感染的士兵数
- 红色区域：因受伤过重而死亡的士兵数
- 黑色区域：死于其它原因的士兵数

泰坦尼克号生存和死亡人数条形图

Survived	Class	Freq
No	1st	122
Yes	1st	203
No	2nd	167
Yes	2nd	118
No	3rd	528
Yes	3rd	178
No	Crew	673
Yes	Crew	212
	人数总计	2201





四、知识图谱

- 除去经典可视化方法之外，近年来出现一种新的数据表示方法：知识图谱。这可能是一个新颖概念，但其实知识图谱无处不在，这源于树状结构信息的广泛存在性。

- 例：谷歌搜索

当你在谷歌上搜索一个名人的名字时，你可能会注意到搜索结果页面的右侧会出现一个信息框，显示该名人的照片、基本信息、相关人物等。

这就是知识图谱的一个实际应用。谷歌通过知识图谱将不同的信息点连接在一起，提供更全面和直观的信息。

四、知识图谱

Google elon musk 登录

Wikipedia
https://en.wikipedia.org › wiki › Elon_Musk

Elon Musk

Elon Reeve Musk FRS is a businessman and investor known for his key roles in the space company SpaceX and the automotive company Tesla, ...

Wealth of Elon Musk · Family of Elon Musk · Elon Musk (disambiguation) · Grimes

X
https://twitter.com › elonmusk · 翻译此页

Elon Musk (@elonmusk) / X

ELON: I LIVED IN PENN—IT WAS DANGEROUS THEN, BUT IT'S EVEN WORSE NOW! "I lived in Pennsylvania for three years, went to school here, so I know the state well.

维基百科
https://zh.wikipedia.org › wiki · 转为简体网页

伊隆·馬斯克- 維基百科，自由的百科全書

伊隆·里夫·馬斯克（英語：Elon Reeve Musk，/ˈiːlɒn/ EE-lon；1971年6月28日— ），FRS，曾取漢名馬誼郎於臺灣作為公司登記使用，是一名企業家、商業大亨、英國皇家學會會士、 ...

Tesla
https://www.tesla.com › en_hk › elon-musk · 翻译此页

Elon Musk | Tesla Hong Kong

收听

Spotify　Apple Music

简介

埃隆·里夫·马斯克，FRS，曾取汉名马谊郎于台湾作为公司登记使用，是一名企业家、商业大亨、英国皇家学会会士、美国工程院院士。 维基百科

出生信息： 1971 年 6 月 28 日（53 岁），南非比勒陀利亚

资产净值： 2478 亿美元（2024 年） 福布斯

子女： 薇薇安·詹娜·威尔逊、 Tau Techno Mechanicus Musk, 等等

配偶： 妲露拉·莱莉 (结婚时间：2013 年–2016 年), 等等

父母： 埃罗尔·马斯克、 梅耶·马斯克

兄弟姐妹： 金巴尔·马斯克、 托斯卡·马斯克、 亚娜·贝祖伊登霍特、 亚历山德拉·马斯克、 阿莎·马斯克

反馈

四、知识图谱

- 知识图谱的严格定义：
  知识图谱（Knowledge Graph）是一种用于表示知识的结构化方式，它通过节点（实体）和边（关系）来描述现实世界中的事物及其相互关系。知识图谱的目标是将信息转化为知识，使得计算机能够理解和推理。
- 关键概念：

- 实体（Entity）：知识图谱中的节点，代表具体的事物，如人、地点、事件等。
- 关系（Relation）：连接实体的边，描述实体之间的关系，如“是朋友”、“位于”等。
- 属性（Attribute）：实体的特征或信息，如人的出生日期、地点的面积等。

四、知识图谱

- 知识图谱在许多领域都有应用，包括但不限于：
  - 搜索引擎：提高搜索结果的相关性和信息的直观展示。
  - 推荐系统：通过分析用户与实体的关系，提供个性化推荐。
  - 问答系统：通过知识图谱理解用户问题并提供准确答案。

- 有多种工具可以用来构建和操作知识图谱，例如：
  - Neo4j：一个流行的图数据库，支持图形化界面和Cypher查询语言。
  - GraphDB：一个支持RDF和SPARQL的图数据库，适合语义网应用。

四、知识图谱

- Neo4j 简单操作示例：

1. 安装 Neo4j Desktop：下载并安装Neo4j Desktop。
2. 创建数据库：在Neo4j Desktop中创建一个新的数据库。
3. 导入数据：使用Cypher语言导入数据，例如：

```
CREATE (a:Person {name: 'Alice'})
    CREATE (b:Person {name: 'Bob'})
    CREATE (a)-[:KNOWS]->(b)
```

4. 查询数据：使用Cypher语言查询数据，例如：

```
MATCH (a:Person)-[:KNOWS]->(b:Person)
RETURN a.name, b.name
```

四、知识图谱

- 执行查询后，Neo4j会在界面中显示一个图形化的结果。
  你会看到两个节点（Alice和Bob），以及一条连接它们的边（KNOWS关系）。
  Alice和Bob节点会以圆形表示，KNOWS关系会以箭头表示，指向Bob。

- 可视化示例
  节点：Alice和Bob会显示为两个圆圈，内部标注有各自的名字。
  关系：从Alice指向Bob的箭头，标注为“KNOWS”。

- 通过这种可视化，用户可以直观地看到数据中的实体及其关系。这种图形化展示是Neo4j的一个强大功能，帮助用户更好地理解和分析数据。

四、知识图谱

- 在日常学习中的应用
  - 项目研究：使用知识图谱整理和分析研究资料，发现新的研究方向。
  - 学习笔记：将学习内容以知识图谱的形式组织，帮助理解和记忆。
  - 跨学科联系：通过知识图谱发现不同学科之间的联系，促进综合学习。

五、地理信息系统

- 案例：在线地图

当你使用在线地图软件（高德地图、谷歌地图）查找路线或查看某个地区的卫星图像时，你就在使用地理信息系统（GIS）。在线地图软件通过整合地理数据和地图信息，帮助用户进行导航、查看交通状况、探索地点等。

- 地理信息系统的概念

地理信息系统（GIS）是一种用于捕获、存储、分析、管理和展示地理空间数据的系统。GIS结合了地理学、统计学、计算机科学和信息技术，能够处理和分析与地理位置相关的信息。

五、地理信息系统

- 关键概念：

空间数据：与地理位置相关的数据，如坐标、地形、土地使用等。
属性数据：描述空间数据特征的信息，如人口密度、气候数据等。
图层：GIS中的基本单位，每个图层包含特定类型的地理信息，如道路、河流、建筑物等。

- 地理信息系统可广泛的运用于经济研究中。
- 环境经济学：

资源管理：GIS用于分析自然资源的分布和利用情况，帮助制定可持续的资源管理政策。
环境影响评估：评估经济活动对环境的影响，支持环境保护和政策制定。

五、地理信息系统

- 农业经济学

精准农业：通过GIS分析土壤、气候、作物生长等数据，优化农业生产，提高产量和效率。

土地利用规划：帮助制定合理的土地利用政策，促进农业可持续发展。

- 区域经济发展

区域竞争力分析：GIS可以用于分析不同区域的经济活动、基础设施、劳动力市场等，帮助制定区域发展战略。

投资决策支持：通过空间分析，识别具有投资潜力的区域和行业。

五、地理信息系统

- 商业案例：沃尔玛的选址和供应链优化
- 背景：

沃尔玛是全球最大的零售商之一，其成功的一部分归功于其高效的选址策略和供应链管理。为了保持竞争优势，沃尔玛利用GIS技术来优化其商店选址和供应链网络。

- 应用：
- 商店选址：

1. 人口分析：沃尔玛使用GIS分析人口密度、人口增长趋势、收入水平等数据，以确定新店的最佳位置。
2. 竞争分析：通过GIS，沃尔玛可以分析竞争对手的分布和市场份额，选择具有战略优势的地点。
3. 交通流量：GIS帮助沃尔玛评估交通流量和可达性，确保新店能够吸引足够的顾客。

五、地理信息系统

- 供应链优化：

1. 配送中心选址：GIS用于分析地理位置、运输成本、供应商位置等因素，以优化配送中心的选址。
2. 物流路径优化：通过GIS分析交通网络和运输路线，沃尔玛能够设计最优的物流路径，降低运输成本，提高配送效率。

- 市场分析：

1. 客户行为分析：GIS帮助沃尔玛分析客户的购物习惯和偏好，调整产品供应和市场营销策略。
2. 区域市场潜力评估：通过GIS，沃尔玛可以评估不同区域的市场潜力，制定区域发展战略。

五、地理信息系统

- 结果：

通过GIS技术，沃尔玛能够更准确地进行市场分析和选址决策，优化供应链管理。这不仅提高了运营效率，还增强了市场竞争力，帮助沃尔玛在全球范围内保持领先地位。

- 这个案例展示了GIS如何在复杂的经济管理中发挥关键作用，帮助企业做出数据驱动的决策。

思考与练习

1. 获取数据练习对数据作图，注意思考选择合适的图形。
2. 进一步查阅Neo4j图关系数据库相关资料，并思考如何将知识图谱运用到日常工作或学习中去。

THANK YOU!