

# Graphical descriptive techniques

Applied Statistics

Fall 2025

## 目录

<b>1 分类数据 Categorical Data</b>	<b>4</b>
1.1 定义 Definition	4
1.2 两种类型 Two Types	4
1.2.1 名义数据 Nominal Data	4
1.2.2 顺序数据 Ordinal Data	4
1.2.3 虚拟变量 Dummy Variable	5
<b>2 数值数据 Numerical Data</b>	<b>5</b>
2.1 定义 Definition	5
2.2 两种类型 Two Types	5
2.2.1 离散数据 Discrete Data	5
2.2.2 连续数据 Continuous Data	6
2.3 数据类型判断示例 Example of Determining Data Types	6
2.4 关于离散与连续数据的补充说明 More on Discrete and Continuous Data	7
2.4.1 离散数据 Discrete Data	7
2.4.2 连续数据 Continuous Data	7
2.4.3 统计工具选择 Statistical Tools	7
2.4.4 离散数据近似为连续数据 Approximating Discrete as Continuous	7
<b>3 数据分析前的数据清洗 Data Cleaning before We Start</b>	<b>8</b>
3.1 检查缺失值 Examine Missing Variables	8
3.2 检查“错误”值 Examine “Wrong” Values	8
3.3 图表与摘要统计 Graphs and Summary Statistics	8
<b>4 表格与图表 Tables and Graphs</b>	<b>8</b>
4.1 作用 Purpose	8
4.2 关注三个方面 Focus on Three Aspects	8

4.3	分类数据的表格与图表 Tables and Graphs for Categorical Data . . . . .	9
4.3.1	频数分布表 Frequency Distribution Table . . . . .	9
4.3.2	常用图表 Common Graphs . . . . .	9
4.3.3	示例: GSS 2012 调查中的工作状态 Example: Work Status in the GSS 2012 Survey . . . . .	9
4.3.4	图表选择建议 Notes on Chart Selection . . . . .	12
4.4	数值数据的表格与图表 Tables and Graphs for Numerical Data . . . . .	12
4.4.1	直方图 Histogram . . . . .	12
4.4.2	频数表 Frequency Table . . . . .	12
4.4.3	类比与构建 Analogy and Construction . . . . .	13
4.4.4	示例: GSS 2012 年龄分布 Example: Age Distribution in GSS 2012	13
4.4.5	区间数选择指南 Guidelines on the Number of Intervals . . . . .	13
4.4.6	直方图与密度 Histogram and Density . . . . .	15
<b>5</b>	<b>解读直方图 Interpret the Histogram</b>	<b>15</b>
5.1	中心: 众数区间 Center: Modal Interval . . . . .	15
5.2	直方图的形状 Shapes of Histograms . . . . .	15
5.2.1	对称性 Symmetry . . . . .	16
5.2.2	偏度 Skewness . . . . .	16
5.3	累积相对频数分布 Cumulative Relative Frequency Distribution . . . . .	16
5.4	累积频数图与多边形图 Cumulative Frequency Graph and Polygon . . . . .	17
<b>6</b>	<b>时间序列数据与横截面数据 Time-series data and Cross-sectional data</b>	<b>17</b>
6.1	横截面数据 Cross-sectional Data . . . . .	17
6.2	时间序列数据 Time-series Data . . . . .	17
6.3	趋势图 Trend Graph . . . . .	18
6.4	练习: 识别数据类型 Practice: Time-series data and Cross-sectional data	18
6.4.1	问题 Questions . . . . .	18
6.4.2	答案 Answers . . . . .	18
<b>7</b>	<b>两个分类变量之间的关系 Relationship between two categorical variables</b>	<b>19</b>
7.1	交叉分类表 Cross-classification Table . . . . .	19
7.1.1	示例: 工作状态与性别的交叉分类表 . . . . .	19
7.2	条件分布 Conditional Distributions . . . . .	20
7.2.1	条件分布表示例 . . . . .	20
<b>8</b>	<b>两个数值变量之间的关系 Relationship between Two Numerical Variables</b>	<b>20</b>
8.1	散点图 Scatter Plot . . . . .	20

8.1.1	解读散点图 Examining a Scatter Plot . . . . .	21
8.1.2	注意事项与改进方法 Some Caveats and Improvements . . . . .	23
<b>9</b>	<b>总结 Summary</b>	<b>23</b>
<b>10</b>	<b>常见错误与注意事项 Common Mistakes to Avoid</b>	<b>23</b>
10.1	图表制作规范 . . . . .	23
10.2	尺度操纵 Scale Manipulation . . . . .	23
10.3	先看图再解读 Figures before Interpretation . . . . .	23
<b>11</b>	<b>综合练习 Comprehensive Practice</b>	<b>25</b>
11.1	调查问题与数据类型 Survey Questions and Data Types . . . . .	25
11.2	练习 Practice . . . . .	25
11.2.1	血糖水平分布形状 Shape of Blood Glucose Levels Distribution . . . . .	25

# 1 分类数据 Categorical Data

## 1.1 定义 Definition

- **中文:** 数据值为类别, 如性别、婚姻状况等。可用数字编码, 但数字本身无内在意义。
- **English:** The values are categories, such as gender, marital status, etc. Can be coded as numbers, but the numbers do not have intrinsic meanings.
- **示例 Example:**
  - 1 = 女性 (female), 0 = 男性 (male)
  - 0 = 男性 (male), 1 = 女性 (female)

## 1.2 两种类型 Two Types

### 1.2.1 名义数据 Nominal Data

- **中文:** 类别间无特定顺序
- **English:** No particular order across categories
- **示例 Examples:**
  - 出生月份 Birth month
  - 电话号码 Phone number
  - 性别 Gender

### 1.2.2 顺序数据 Ordinal Data

- **中文:** 类别有顺序
- **English:** Ordered categories
- **示例 Example:**
  - 顾客满意度 1-5 分 Consumer satisfaction from 1 to 5
  - 1 = 非常满意 (most satisfied), 5 = 非常不满意 (least satisfied)

### 1.2.3 虚拟变量 Dummy Variable

- **中文:** 只有两个类别的分类数据, 常用 0/1 编码。
- **English:** Data with only two categories, often coded as 0/1.
- **示例 Example:**
  - 性别变量 (1 = 女性, 0 = 男性)  
Female variable (1 = female, 0 = male)

## 2 数值数据 Numerical Data

### 2.1 定义 Definition

- **中文:** 数据值为实数, 来自计数或测量。
- **English:** Data values are real numbers from counting or measuring something.
- **示例 Examples:**
  - 消费金额 Spending
  - 身高 Height
  - GDP
- **数值有意义 Meaningful Numbers:** 刻度单位之间的差异是相等的。The differences between scale units are equal.

### 2.2 两种类型 Two Types

#### 2.2.1 离散数据 Discrete Data

- **中文:** 可数的不同值
- **English:** Countable number of distinct values
- **示例 Example:**
  - 每小时顾客到访次数 How many customers visit the store per hour

### 2.2.2 连续数据 Continuous Data

- **中文:** 区间内无限个可能值
- **English:** Countless values within an interval
- **示例 Example:**
  - 两地之间的距离 Distance between two locations

## 2.3 数据类型判断示例 Example of Determining Data Types

表 1: 数据类型示例 Examples of Data Types

数据项 Data Item	数据类型 Data Type
教育程度 (1= 高中, 2= 大学, 3= 研究生) Educational level: 1=High School, 2=College, 3=Graduate Degree	顺序分类数据 Ordinal Categorical
推荐餐厅的可能性 (1-5 分, 1= 最不可能) How likely would you recommend this restaurant? From 1 to 5, 1=least likely and 5=most likely	顺序分类数据 Ordinal Categorical
每周锻炼次数 How many times do you exercise each week	离散数值数据 Discrete Numerical
出生月份 Your birth month	名义分类数据 Nominal Categorical
电话号码 Phone number	名义分类数据 Nominal Categorical
股票价格 Stock prices	连续数值数据 Continuous Numerical

## 2.4 关于离散与连续数据的补充说明 More on Discrete and Continuous Data

### 2.4.1 离散数据 Discrete Data

- **中文:** 通常来自计数, 多为整数。
- **English:** Often from counting and often integers.

### 2.4.2 连续数据 Continuous Data

- **中文:** 常以“离散”报告单位记录, 如价格、身高、体重、温度等。
- **English:** Often recorded in “discrete” reporting units, e.g. retail prices, height and weight, temperature.

### 2.4.3 统计工具选择 Statistical Tools

- **中文:** 离散数据通常取值范围较小, 适用的统计工具可能与连续数据不同 (如每个值作为一个区间的直方图)。
- **English:** Discrete data usually have relatively few potential values so the appropriate statistical tools used may be different from continuous data; e.g. histogram with each value as an interval.

### 2.4.4 离散数据近似为连续数据 Approximating Discrete as Continuous

- **中文:** 若取值范围大 (可能值多), 有时可将离散数据视为连续数据处理。
- **English:** We sometimes treat discrete data as continuous if the range is large (many potential values).
- **示例 Examples:**
  - **中文:** 高端美容院每小时顾客数 ( $< 5$  人)  $\rightarrow$  适合泊松分布
  - **English:** High-end beauty salon with limited capacity (like  $< 5$  customers at the same time)  $\rightarrow$  Poisson distribution would be appropriate.
  - **中文:** 繁忙商场每小时顾客数 (数百人)  $\rightarrow$  适合正态分布
  - **English:** Busy mall with hundreds of customers  $\rightarrow$  using a normal distribution is more than enough.

### 3 数据分析前的数据清洗 Data Cleaning before We Start

#### 3.1 检查缺失值 Examine Missing Variables

- **中文:** 区分逻辑缺失（调查对象不适用）和无应答。
- **English:** Distinguish between missing by logic (universe matters a lot) and non-response.

#### 3.2 检查”错误”值 Examine ”Wrong” Values

- **中文:** 排序数据很有用。
- **English:** Ordering the data is very useful.
- **示例 Example:**
  - **中文:** 年龄变量中出现 -1 或 999，可能是”缺失”或”拒绝回答”的编码。
  - **English:** In an age variable, a value of -1 or 999 is likely a code for ”missing” or ”refused to answer,” not a real age.

#### 3.3 图表与摘要统计 Graphs and Summary Statistics

- **中文:** 可用于判断数据可靠性，注意异常值。
- **English:** Can also indicate the reliability of data (when you plot or summarize the data, pay attention to abnormalities).

### 4 表格与图表 Tables and Graphs

#### 4.1 作用 Purpose

- **中文:** 以简洁方式组织、探索和总结数据。
- **English:** Organize, explore, and summarize data in a succinct way.

#### 4.2 关注三个方面 Focus on Three Aspects

1. **中心 Center:** 数据集中在哪里？典型值或中位数。  
Where are the data values concentrated? The typical or middle data values.
2. **离散程度 Variability/Dispersion:** 数据分散程度如何？是否有异常值？  
How spread out are the data? Are there outliers?



### 3. 分布形状 Shape: 是否对称? 偏态? 峰度如何?

Symmetrically around the center or skewed? How peaked is the distribution?

## 4.3 分类数据的表格与图表 Tables and Graphs for Categorical Data

### 4.3.1 频数分布表 Frequency Distribution Table

- **中文:** 列出类别及其观测数 (绝对频数) 和/或相对频数 (百分比)。
- **English:** A table that presents the categories and the number of observations (absolute frequency) and/or relative frequency (percentage).

### 4.3.2 常用图表 Common Graphs

- **饼图 Pie Chart:**
  - **中文:** 用扇形面积表示各类别所占比例。
  - **English:** Show the relative frequency in a pie; the area represents the share.
- **条形图 Bar Chart:**
  - **中文:** 用条形高度表示各类别的频数。
  - **English:** Drawing a rectangle representing each category to show the frequency distribution; the height of the bar shows the frequency.

### 4.3.3 示例: GSS 2012 调查中的工作状态 Example: Work Status in the GSS 2012 Survey

#### 调查问题 Survey Question

- **中文:** " 上周您是全职工作、兼职工作、在校学习、料理家务, 还是其他? "
- **English:** "Last week were you working full-time, part-time, going to school, keeping house, or what?"

表 2: GSS 2012 工作状态频数分布 Frequency Distribution of Work Status in GSS 2012

工作状态 Work Status	代码 Code	频数 Frequency	相对频数 (%) Relative Frequency
全职工作 Working full-time	1	912	46.2
兼职工作 Working part-time	2	226	11.5
暂时未工作 Temporarily not working	3	40	2.0
失业/下岗 Unemployed, laid off	4	104	5.3
退休 Retired	5	357	18.1
在校 School	6	70	3.5
料理家务 Keeping house	7	210	10.6
其他 Other	8	54	2.7
总计 Total		1973	100

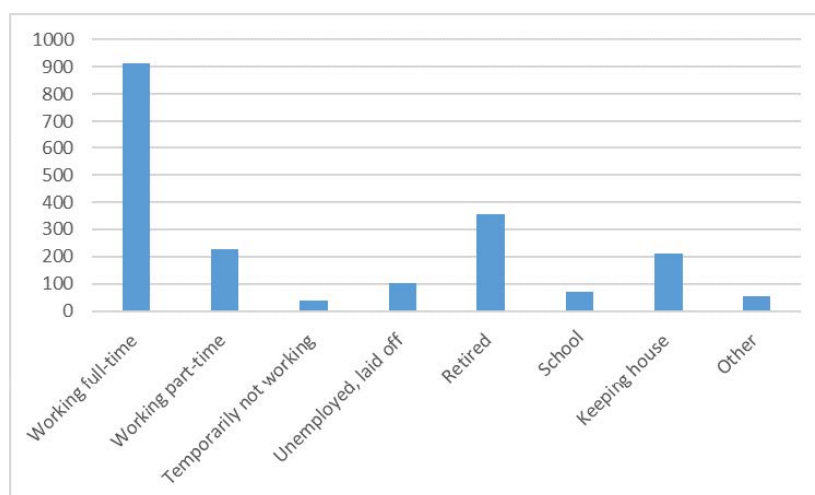


图 1: Bar Chart Example

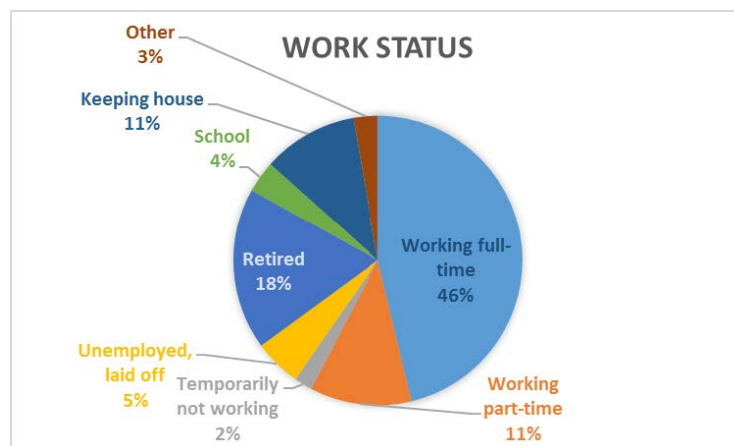


图 2: Example Pie Chart

#### 4.3.4 图表选择建议 Notes on Chart Selection

##### 条形图方向 Bar Chart Orientation

- **中文:** 若类别标签较长, 横向条形图更美观。
- **English:** When the labels of the categories are longer, it might be visually more appealing to arrange the bar horizontally.

##### 类别顺序 Category Order

- **中文:** 对于顺序数据, 条形图中应保持类别顺序。
- **English:** Order of the categories matters in bar charts when dealing with ordinal data, e.g. education level.

##### 饼图 vs 条形图 Pie Chart vs Bar Chart

- **饼图 Pie Chart:** 直观显示各类别占比。若关注比例, 饼图更直观。  
Visually shows the proportion of each category.
- **条形图 Bar Chart:** 清晰比较各类别频数差异。类别多时, 条形图更清晰; 频数差异小时, 条形图更能展示差异。  
Clearly compares frequency differences across categories.

#### 4.4 数值数据的表格与图表 Tables and Graphs for Numerical Data

##### 4.4.1 直方图 Histogram

- **定义:** 数值数据最常用的图表, 是频数分布的可视化展示。
- **English:** Most common graph for numerical data, a visual display of frequency distribution.

##### 4.4.2 频数表 Frequency Table

- **定义:** 将数据总结为相对较少区间的表格。
- **English:** A table of data summarized into a relatively small number of intervals.

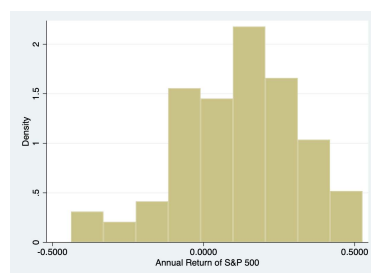


图 3: Histogram Example

#### 4.4.3 类比与构建 Analogy and Construction

- **类比:** 类似于分类数据, 我们创建区间 (类别) 并显示每个区间内数据值的频数。
- **English:** A analogy to categorical data: we create intervals (categories) and show the frequency of data values within each interval.
- 1. 选择区间数  $k$ , 根据数据范围确定宽度。
- 2. **相对频数 Relative Frequency:** 绝对频数除以总观测数。
- 3. **累积频数 Cumulative Frequency:** 累加相对频数。

#### 4.4.4 示例: GSS 2012 年龄分布 Example: Age Distribution in GSS 2012

#### 4.4.5 区间数选择指南 Guidelines on the Number of Intervals

##### 关键原则 Key Principle

- 绘制信息丰富的直方图关键在于选择正确的区间数  $k$ 。
- The key for an informative histogram is to pick the right number of intervals  $k$ .
- 区间太多, 某些区间数据稀疏甚至为空。
- Too many classes, some will be sparsely populated or even empty.
- 区间太少, 不同的数据值会被混在一起 (想象极端情况, 只有一个区间)。
- Too few classes, dissimilar data values are lumped together (imagine the extreme, only one class).

#### 基于观测数的建议 Approximate Number of Classes Based on Observations

##### 软件自动计算 Software Automatic Calculation

- 更高级的软件通常会自动计算区间并创建直方图。

表 3: GSS 2012 年龄分布频数表 Age Distribution Frequency Table in GSS 2012

年龄组 Age Class	频数 Frequency	相对频数 Relative Frequency	累积百分比 Cumulative
15-19	51	2.59%	2.59%
20-24	150	7.62%	10.21%
25-29	177	8.99%	19.20%
30-34	190	9.65%	28.85%
35-39	182	9.24%	38.09%
40-44	176	8.94%	47.03%
45-49	183	9.29%	56.32%
50-54	189	9.60%	65.92%
55-59	143	7.26%	73.18%
60-64	172	8.74%	81.92%
65-69	103	5.23%	87.15%
70-74	97	4.93%	92.08%
75-79	68	3.45%	95.53%
80-84	46	2.34%	97.87%
85-89	42	2.13%	100.00%

表 4: 频数分布中区间数的近似指南

Approximate Number of Classes in Frequency Distributions

观测数 Number of Observations	区间数 Number of Classes
Less than 50	5-7
50-200	7-9
200-500	9-10
500-1,000	10-11
1,000-5,000	11-13
5,000-50,000	13-17
More than 50,000	17-20

- More advanced software usually calculates the classes and creates the histogram automatically.

- 例如，Stata 使用公式：

$$k = \min\{\sqrt{N}, 10 \times \ln(N)/\ln(10)\}$$

#### 4.4.6 直方图与密度 Histogram and Density

##### 密度曲线 Density Curve

- 密度曲线下的面积等于 1，因此 y 轴单位会调整以适应此定义。
- The area under the density curve equals one, so the y-axis unit is adjusted to fit this definition.
- 密度图是直方图的平滑版本，用于估计概率密度函数。

## 5 解读直方图 Interpret the Histogram

1. 中心 Center: 数据集中在哪里。
2. 变异性 Variability: 数据分散程度如何，是否有异常值？数据的范围是多少？
3. 形状 Shape: 分布是对称还是偏斜？

### 5.1 中心：众数区间 Center: Modal Interval

- 众数区间 Modal Interval: 直方图中频数最高的区间。
- 只有一个峰值时，分布为单峰 unimodal。
- 数据集可能有多个众数区间：双峰 bimodal 等。

### 5.2 直方图的形状 Shapes of Histograms

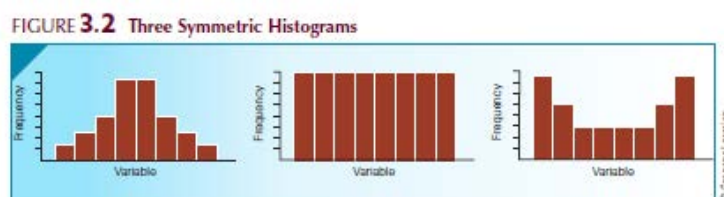


图 4: Three Symmetric Histogram

### 5.2.1 对称性 Symmetry

- 如果沿直方图中心画一条垂直线，两侧形状和大小相同，则称直方图对称。
- A histogram is said to be symmetric if, when we draw a vertical line down the center of the histogram, the two sides are identical in shape and size.
- 最常见的对称形状是钟形 bell-shaped。

### 5.2.2 偏度 Skewness

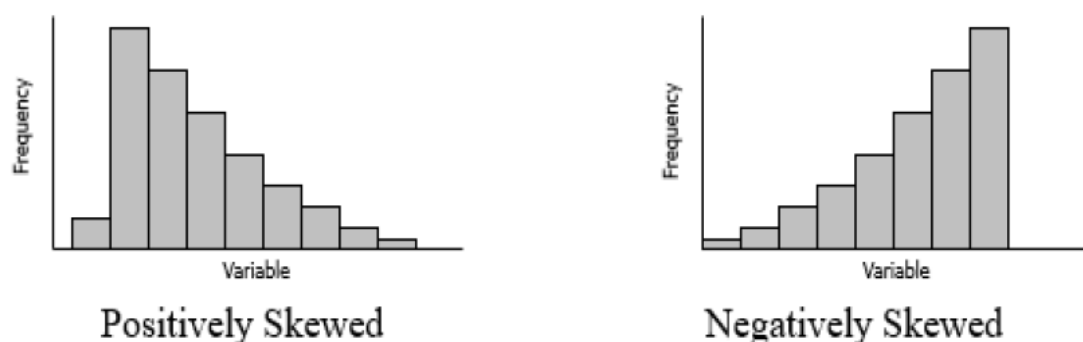


图 5: Skewed Example

- 偏斜直方图 **Skewed Histogram**: 有一个长尾延伸至右侧或左侧。
- 右偏/正偏 **Right-skewed / Positively Skewed**: 长尾在右侧。例如，一国个人收入分布通常右偏；大多数人收入中低，少数人收入极高，形成右长尾。
- 左偏/负偏 **Left-skewed / Negatively Skewed**: 长尾在左侧。

### 5.3 累积相对频数分布 Cumulative Relative Frequency Distribution

- 可以绘制累积相对频数分布图。
- 中位数和其他百分位数可以在累积频数图中轻松识别。
- 可以用连接直方图区间中点的线图代替条形。
- 用于比较多个数据集的分布。



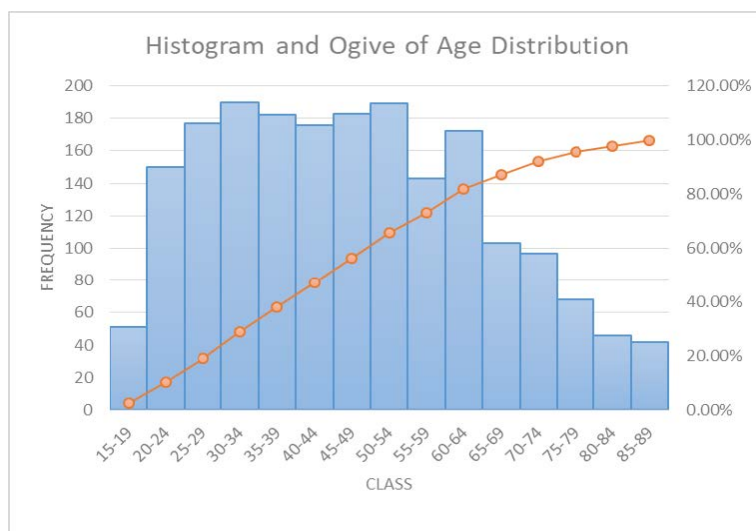


图 6: Cumulative frequency graph (相对频数分布图)

#### 5.4 累积频数图与多边形图 Cumulative Frequency Graph and Polygon

- 累积频数图 (Ogive): 显示累积频数或累积相对频数。
- 频数多边形 Frequency Polygon: 连接直方图各柱顶部中点形成的折线图, 可用于比较多个分布。

## 6 时间序列数据与横截面数据 Time-series data and Cross-sectional data

### 6.1 横截面数据 Cross-sectional Data

- 定义: 在同一时间点测量的观测值。
- **English:** The observations are measured at the same time.
- 示例: 2012 年 GSS 调查数据。

### 6.2 时间序列数据 Time-series Data

- 定义: 在连续时间点上测量的观测值。
- **English:** Measurements at successive points in time.
- 示例: 人均 GDP 年度数据。

### 6.3 趋势图 Trend Graph

- **定义:** 用于描述时间序列数据的折线图。
- **English:** A line chart to describe the time-series data.
- **作图方法:** 将观测值绘制在 y 轴, 时间周期绘制在 x 轴。

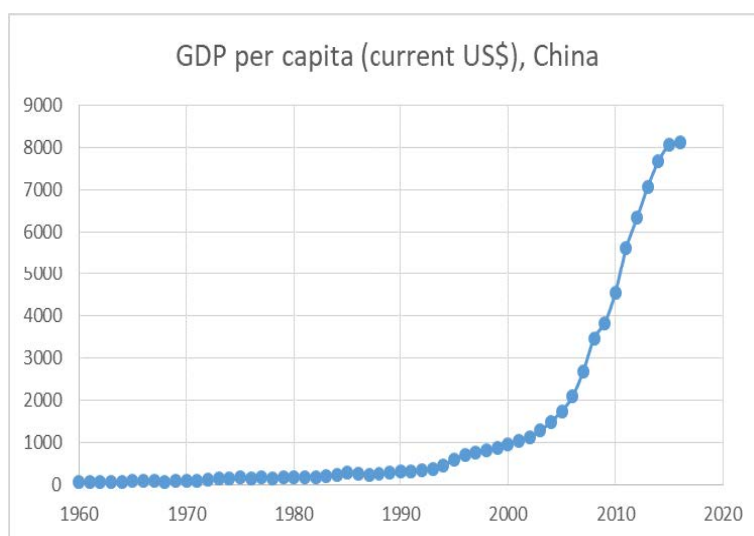


图 7: Trend graph example (趋势图)

### 6.4 练习: 识别数据类型 Practice: Time-series data and Cross-sectional data

#### 6.4.1 问题 Questions

判断以下变量各属于哪种数据类型:

1. 中国过去 10 个季度的 GDP。
2. 去年年底中国 27 个省份的失业率。
3. 中国过去 10 年年末的失业率。
4. 当前中国 10 个最大城市的平均住房价值。

#### 6.4.2 答案 Answers

1. 中国过去 10 个季度的 GDP: **时间序列数据 Time-series**
2. 去年年底中国 27 个省份的失业率: **横截面数据 Cross-sectional**

3. 中国过去 10 年年末的失业率：时间序列数据 Time-series
4. 当前中国 10 个最大城市的平均住房价值：横截面数据 Cross-sectional

## 7 两个分类变量之间的关系 Relationship between two categorical variables

### 7.1 交叉分类表 Cross-classification Table

- 定义：列出两个变量值组合的频数的表格。
- **English:** Lists the frequency of each combination of the values of the two variables.
- 术语：
  - 行变量 Row variable
  - 列变量 Column variable
  - 单元格 Cell：两个变量值的每个组合。

#### 7.1.1 示例：工作状态与性别的交叉分类表

表 5: 工作状态与性别的交叉分类表（频数）

Cross-Classification Table of Frequencies: Work Status by Sex

工作状态 Work Status	男性 Male	女性 Female	总计 Total
全职工作 Working full-time	495	417	912
兼职工作 Working part-time	83	143	226
暂时未工作 Temporarily not working	11	29	40
失业/下岗 Unemployed, laid off	52	52	104
退休 Retired	164	193	357
在校 School	29	41	70
料理家务 Keeping house	22	188	210
其他 Other	30	24	54
总计 Total	886	1,087	1,973

## 7.2 条件分布 Conditional Distributions

- **概念:** 描述分类变量之间关系的方法: 以一个变量的值为条件, 计算另一个变量的分布。
- **English:** Relationships among categorical variables are described by conditional distributions: condition on the value of one variable and calculate the distribution of the other variable.
- 需要区分自变量  $x$  和因变量  $y$ 。
- 我们比较在  $x$  特定值下  $y$  的条件分布。
- 如果两个变量无关, 则条件分布中表现出的模式应大致与  $y$  变量的分布相同。

### 7.2.1 条件分布表示例

表 6: 工作状态与性别的条件分布 (百分比)

Cross-Classification Table: Conditional Distributions (Percentages)

工作状态 Work Status	男性 Male	女性 Female	总计 Total
全职工作 Working full-time	55.87%	38.36%	46.22%
兼职工作 Working part-time	9.37%	13.16%	11.45%
暂时未工作 Temporarily not working	1.24%	2.67%	2.03%
失业/下岗 Unemployed, laid off	5.87%	4.78%	5.27%
退休 Retired	18.51%	17.76%	18.09%
在校 School	3.27%	3.77%	3.55%
料理家务 Keeping house	2.48%	17.30%	10.64%
其他 Other	3.39%	2.21%	2.74%
总计 Total	100.00%	100.00%	100.00%

## 8 两个数值变量之间的关系 Relationship between Two Numerical Variables

### 8.1 散点图 Scatter Plot

- **定义:** 展示在同一观测个体上测量的两个数值变量之间关系的图表。

- **English:** Shows the relationship between two numerical variables measured on the same individuals.
- **构成:** 将  $n$  对观测值以点的形式绘制在 X-Y 坐标系中。
- **坐标轴:**
  - X 轴 (自变量 Independent variable): 通常放置在水平轴。
  - Y 轴 (因变量 Dependent variable): 放置在垂直轴。

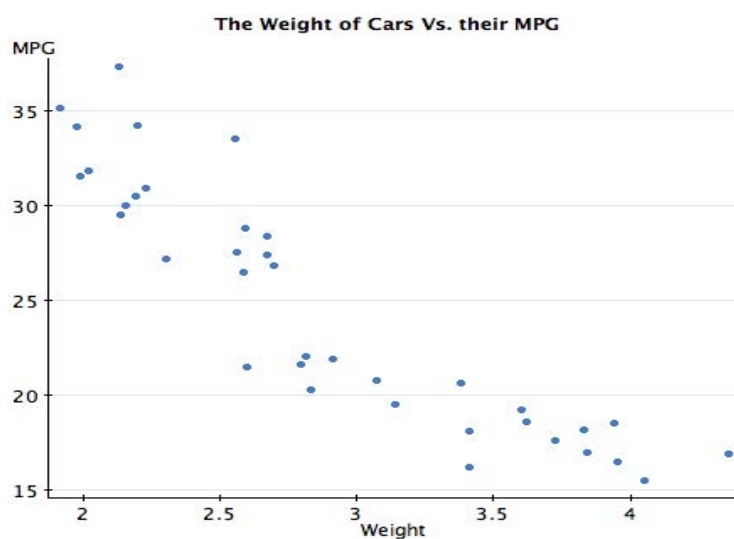


图 8: Scatter plot example (散点图示例)

### 8.1.1 解读散点图 Examining a Scatter Plot

#### 整体模式与异常 Overall Pattern and Deviations

- 观察数据的整体模式以及与该模式的显著偏离。
- 散点图的整体模式: 线性、方向和相关强度。
- **异常值 Outlier:** 落在关系整体模式之外的个别值。
- **相关不等于因果 Correlation is not causation:** 两个变量相关并不意味着一个导致另一个。

#### 方向 Direction

- **正线性关系 Positive Linear Relationship:** 随着 X 增加, Y 也增加。
- **负线性关系 Negative Linear Relationship:** 随着 X 增加, Y 减少。

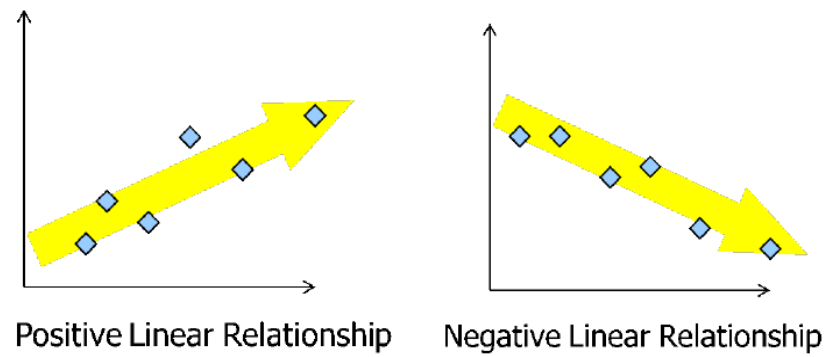


图 9: Linear relationship example

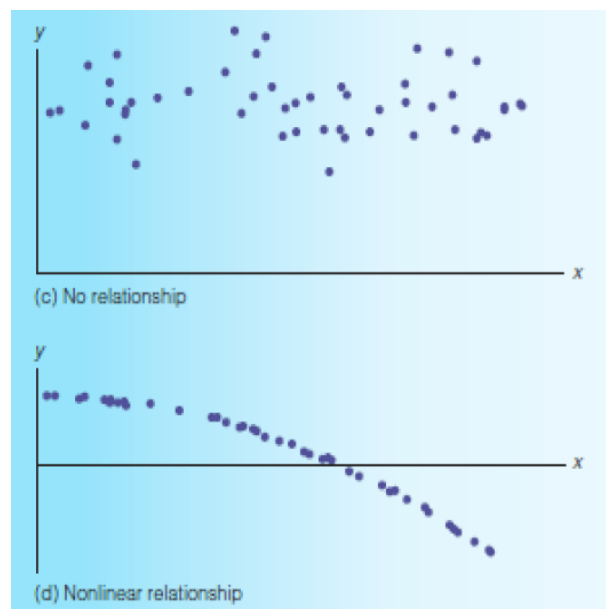


图 10: Nonlinear example

线性关系 Linearity

- 散点图中的点是否大致沿着一条直线分布。

8.1.2 注意事项与改进方法 Some Caveats and Improvements

散点图的适用性

- 适用于观测数量较少且关联较强的情况。
- 在社会科学中，这些条件很少满足（例如收入与年龄的关系）。

改进方法：绘制条件均值折线图

- 可以绘制按年龄组的平均收入折线图。
- 这实际上将个体数据点的散点图转换为一个变量给定另一个变量时的条件均值的清晰图像。

9 总结 Summary

表 7: 数据类型与图表选择总结Summary of Data Types and Graph Selection		
分析类型 Analysis Type	数值数据 Numerical Data	分类数据 Categorical Data
单变量数据集 Single Set of Data	直方图 Histogram	频数与相对频数表 Frequency and Relative Frequency Tables
		条形图与饼图 Bar and Pie Charts
		交叉分类表 Cross-classification Table
两个变量之间 Relationship Between Two Variables	散点图 Scatter Diagram	条形图 Bar Charts

## 10 常见错误与注意事项 Common Mistakes to Avoid

### 10.1 图表制作规范

- 确保标注坐标轴和图表标题。
- 选择并标注适当的尺度。

### 10.2 尺度操纵 Scale Manipulation

- 问题: 通过改变坐标轴尺度来夸大或掩盖数据变化。
- 示例: 销售增长图。
  - 左图: 使用完整尺度, 显示稳定增长。
  - 右图: 放大 y 轴范围, 夸大季度间微小波动。

### 10.3 先看图再解读 Figures before Interpretation

- 原则: 先阅读图表本身, 然后再看解释, 避免被误导。
- 示例: 利率走势图。
  - 图 (a) 标题: ”利率终于开始下降”, 但图表显示长期上升趋势中仅有微小回落。
  - 图 (b) 标题: ”上周利率上升趋势暂时得到缓解”, 更准确地描述了短期波动。

## 11 综合练习 Comprehensive Practice

### 11.1 调查问题与数据类型 Survey Questions and Data Types

金禾中心计划在毕业生毕业一年后进行一项调查, 询问以下问题。请确定每个问题的数据类型, 并说明应创建何种图表。

1. 您的职业是什么? What's your occupation?
  - 数据类型: 名义分类数据 Nominal Categorical
  - 建议图表: 条形图 Bar Chart
2. 您的收入是多少? What's your income?
  - 数据类型: 连续数值数据 Continuous Numerical



- **建议图表:** 直方图或箱线图 Histogram or Box Plot

3. 您获得的学位是什么? What degree did you obtain?

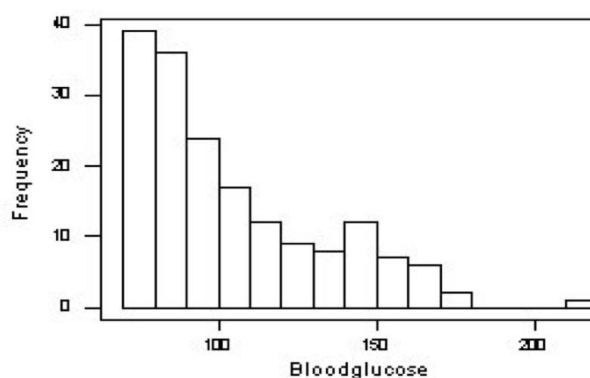
- **数据类型:** 名义分类数据 Nominal Categorical
- **建议图表:** 条形图或饼图 Bar Chart or Pie Chart

4. 您如何评价课程质量? (优秀、很好、好、一般、差)

How would you rate the quality of the program? (excellent, very good, good, fair, and poor)

- **数据类型:** 顺序分类数据 Ordinal Categorical
- **建议图表:** 条形图 (按正确顺序排列类别)  
Bar Chart (with categories in correct order)

## 11.2 练习 Practice



### 11.2.1 血糖水平分布形状 Shape of Blood Glucose Levels Distribution

- **正确答案:** 右偏 Right-skewed (Positively Skewed)
- **解释:** 大多数个体血糖水平较低, 较高值的尾部向右延伸, 这在生物和健康数据中常见。