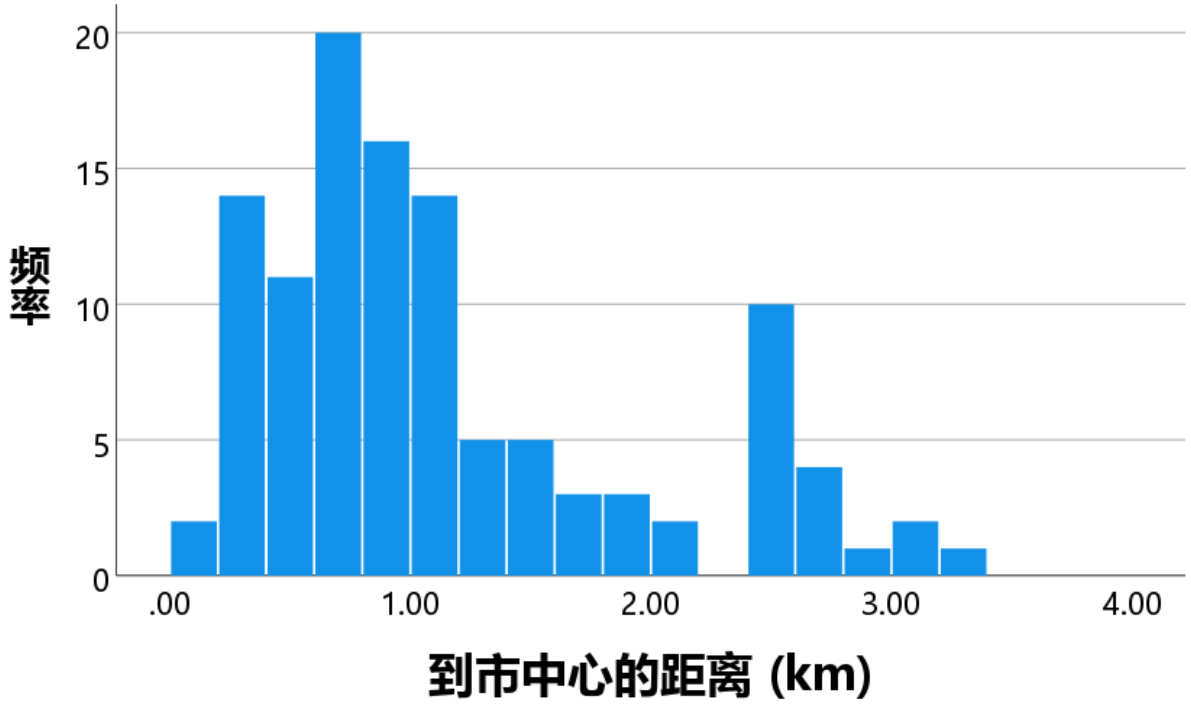


统计学基础

城市分析方法系列课程

苏州大学 王灿

课程政策

- 所有课件均会提供，请按需做笔记
- 如有问题，请务必及时提问
- 网站：methods.courses.wangc.net
- 成绩：
  - 日常出勤，10%
  - 课后小练习，30%
  - 期中：随堂小测试，20%（暂定）
  - 期末：课程报告，40%（暂定）

大纲

- 什么概率和分布?
- 什么是概率论与统计学?
- 变量有几种类型?
- 有哪些常见的描述性统计方法?
- 有哪些常见的统计图?

概率和分布

什么是概率?

16:00点下雨的概率是68%

右侧4位玩家中，没有狼的概率是16.45%

毕业生选择考研的概率是70%

什么是概率?

- 有些事情的机制是未知的，我们对它们的知识不足。
- 有些事情本身就是随机的，而非必然的。

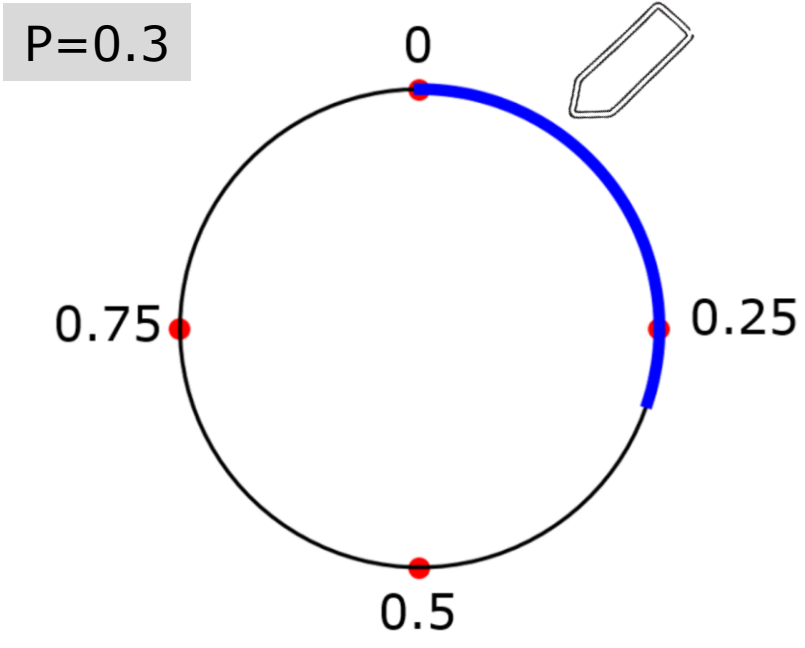
什么是概率?

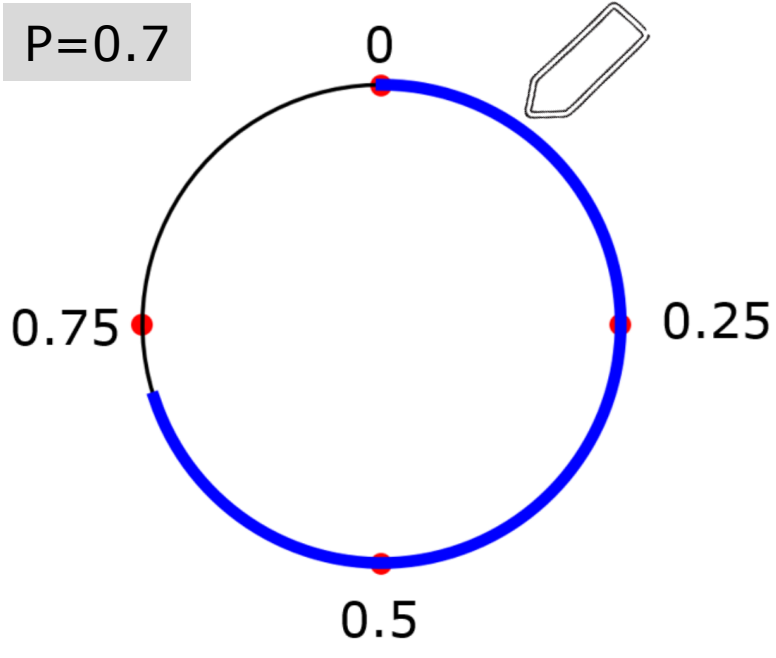
- **频率学派：概率是**一个事件在大量重复实验中出现的**频率**。

```
['平民', '狼人', '狼人', '平民', '平民', '狼人', '预言家', '女巫', '猎人', '平民']
右边没有狼!

['平民', '平民', '狼人', '狼人', '女巫', '平民', '狼人', '平民', '预言家', '猎人']

['女巫', '平民', '猎人', '平民', '狼人', '平民', '平民', '狼人', '狼人', '预言家']

['猎人', '平民', '狼人', '狼人', '平民', '平民', '狼人', '女巫', '平民', '预言家']

['预言家', '平民', '猎人', '女巫', '狼人', '平民', '平民', '狼人', '狼人', '平民']

['平民', '平民', '平民', '预言家', '女巫', '狼人', '平民', '狼人', '猎人', '狼人']

['平民', '平民', '狼人', '狼人', '预言家', '狼人', '平民', '女巫', '平民', '猎人']
右边没有狼!

['预言家', '平民', '狼人', '平民', '猎人', '平民', '平民', '女巫', '狼人', '狼人']
右置位不开狼的概率: 0.1645
```

什么是概率?

- **贝叶斯学派：概率**是对不确定性的主观度量，即个体对事件发生的**信心**程度。





概率和分布

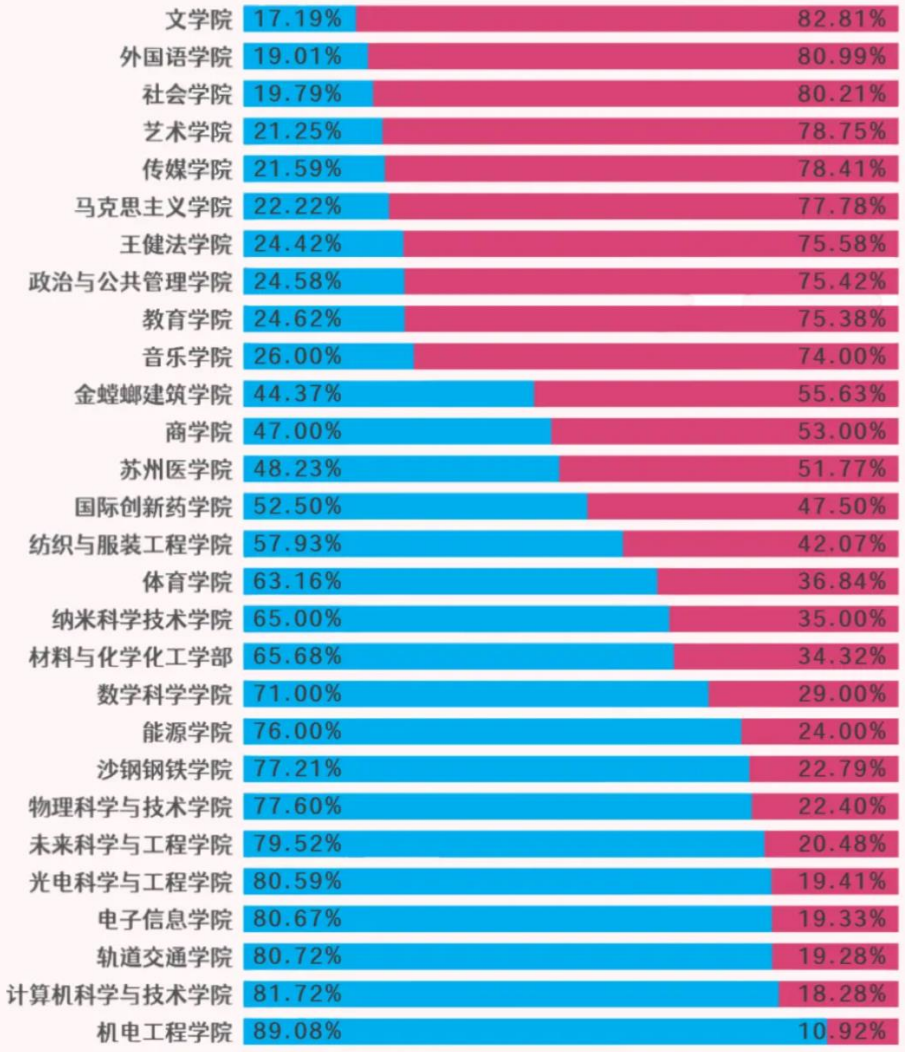
什么是分布?

- 随机变量所有可能的取值，以及每种取值发生的概率。

随机变量 X：新生的性别

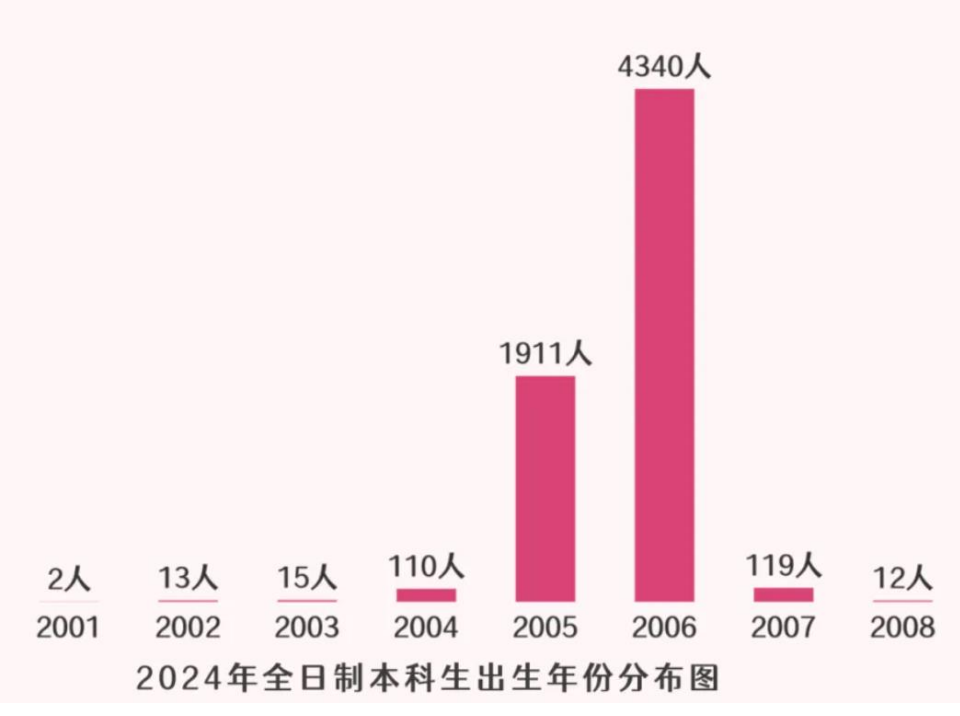
X 所有可能的取值：
x_1 =男，x_2=女

X 的概率分布：
$P(X = x_i) = p_i$



什么是分布?

- 随机变量所有可能的取值，以及每种取值发生的概率。



经验分布 & 理论分布

平均值：$\lambda = 4$

随机变量：某条街道路边停车的数量

泊松分布：$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

日期	停车量
8/29	3
8/30	5
8/31	4
9/1	6
9/2	3
9/3	2
9/4	5

停车量	频数	概率
0	0	0
1	0	0
2	**1**	**0.14**
3	**2**	**0.29**
4	**1**	**0.14**
5	**2**	**0.29**
6	**1**	**0.14**
7	0	0
…	0	0

经验分布

停车量	计算	概率
0	$4^0e^{-4}/0!$	0.02
1	$4^1e^{-4}/1!$	0.07
2	$4^2e^{-4}/2!$	**0.15**
3	$4^3e^{-4}/3!$	**0.20**
4	$4^4e^{-4}/4!$	**0.20**
5	$4^5e^{-4}/5!$	**0.16**
6	$4^6e^{-4}/6!$	**0.10**
7	$4^7e^{-4}/7!$	0.06
…	…	…

理论分布

概率和分布

经验分布

- 直接根据实际观测数据构建的分布。
- 反映了数据的实际情况。
- 没有假设任何具体的分布形式。

理论分布

- 根据概率模型推导出来的分布。
- 通常用于建模数据的产生机制。
- 有不同的分布形式，通常表现为相应的参数和公式。

常见的理论分布

离散分布

随机变量不可无限细分

泊松分布

离散均匀分布

连续分布

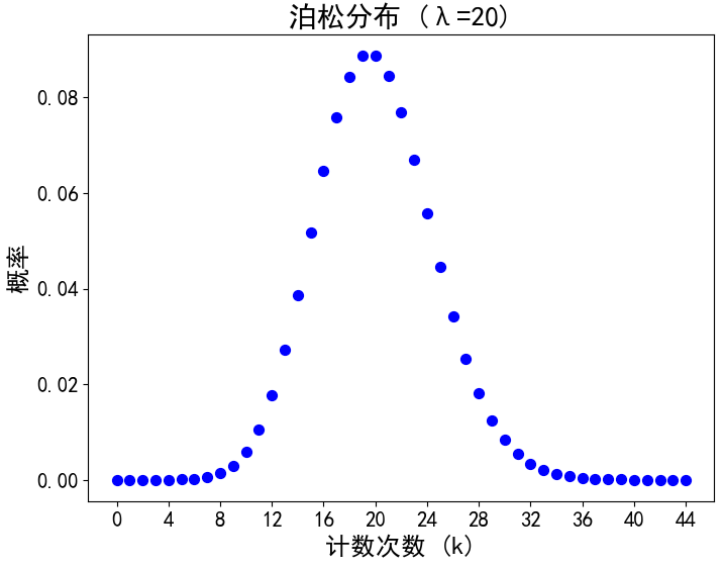
随机变量可以无限细分

正态分布

均匀分布

幂律分布

常见的理论分布：泊松分布

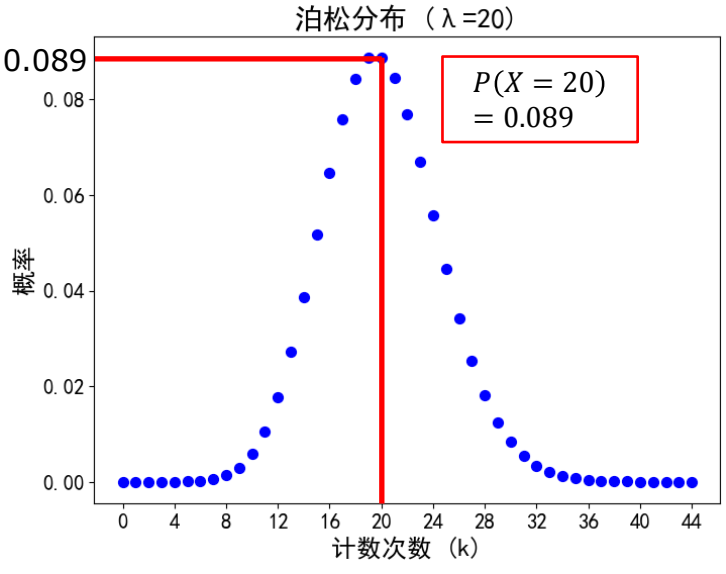


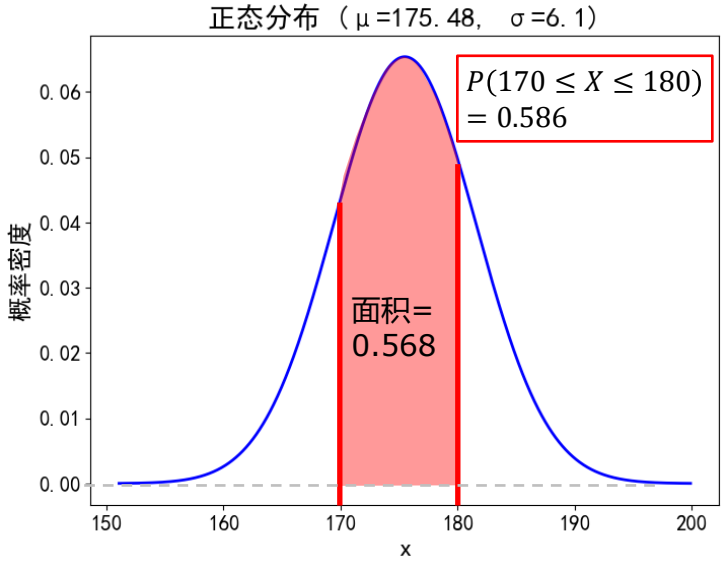
泊松分布 (Poisson distribution)

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

平均次数λ

常见的理论分布：泊松分布 & 正态分布





泊松分布 (Poisson Distribution)

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

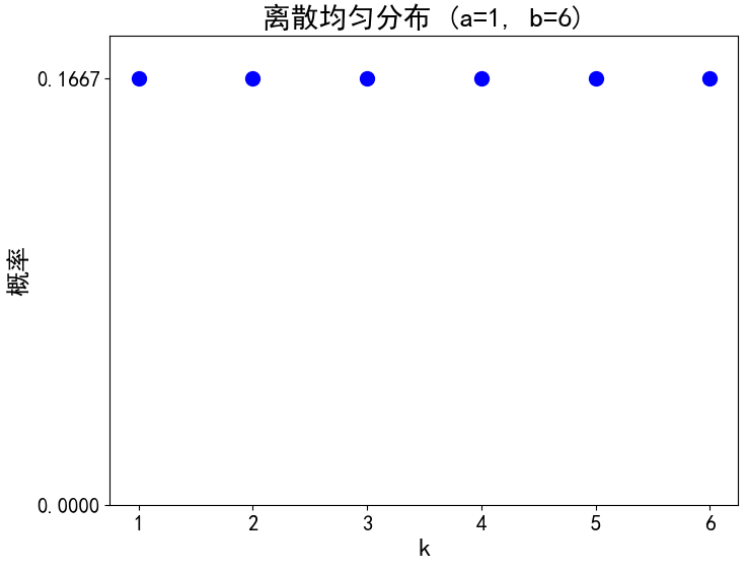
平均次数λ

正态分布 (Normal Distribution) / 高斯分布

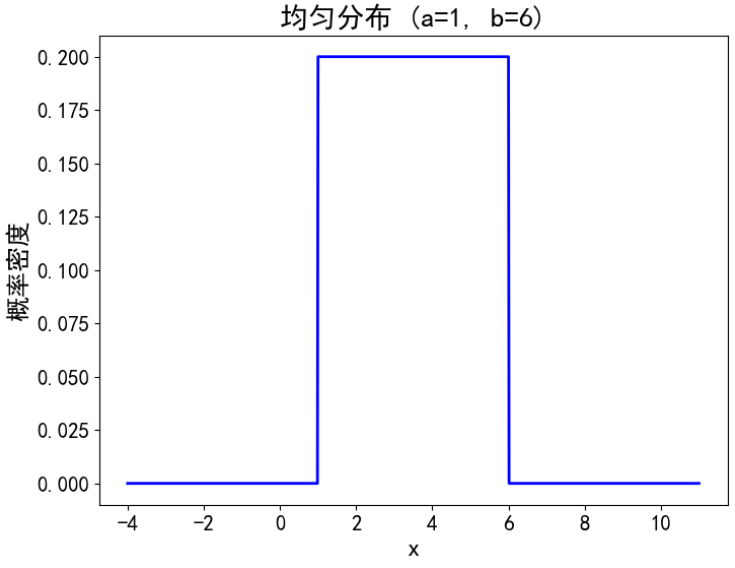
$$PDF(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

均值μ 标准差σ

常见的理论分布：均匀分布



离散均匀分布



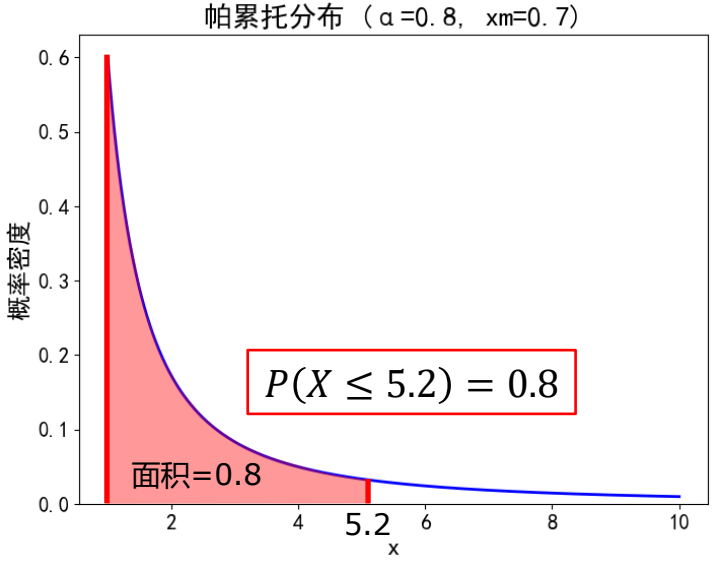
（连续）均匀分布 (Uniform Distribution)

$$P(X = k) = \frac{1}{b - a + 1}$$

最低值a 最高值b

$$PDF(x) = \begin{cases} \frac{1}{b-a}, a \le x \le b \\ 0, x < a \ or \ x > b \end{cases}$$

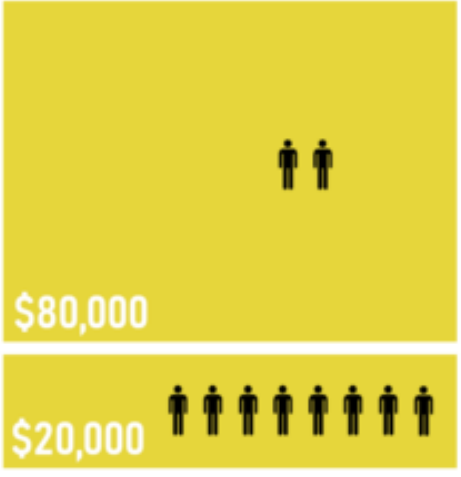
常见的理论分布：幂律分布



二八法则

(Pareto Principle,
80/20 rule,
the law of vital few)：

20%的要素影响了
80%的结果



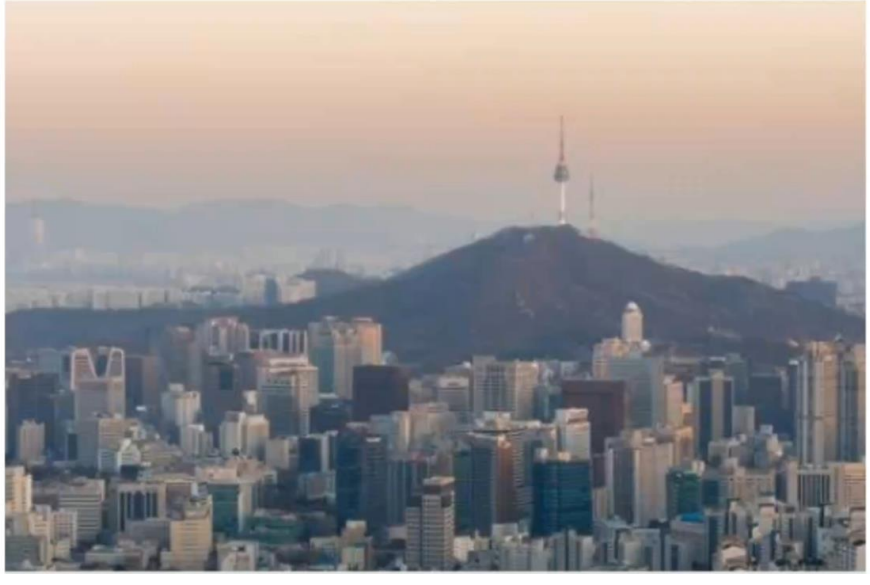
幂律分布（Power Law Distribution）的一种：帕累托分布（Pareto Distribution）

$$PDF(x) = \begin{cases} \dfrac{\alpha x_m^{\alpha}}{x^{\alpha+1}}, & x \geq x_m \\ 0, & x < x_m \end{cases}$$

形状参数α　　尺度参数x_m

常见的理论分布：幂律分布

道路长度的分布



建筑高度的分布

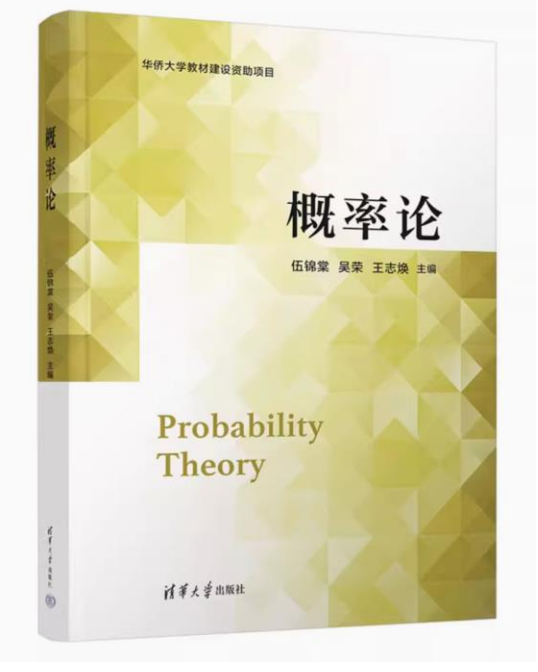
Bin Jiang, Ju-Tzu Huang. A New Approach to Detecting and Designing Living Structure of Urban Environments. 2021

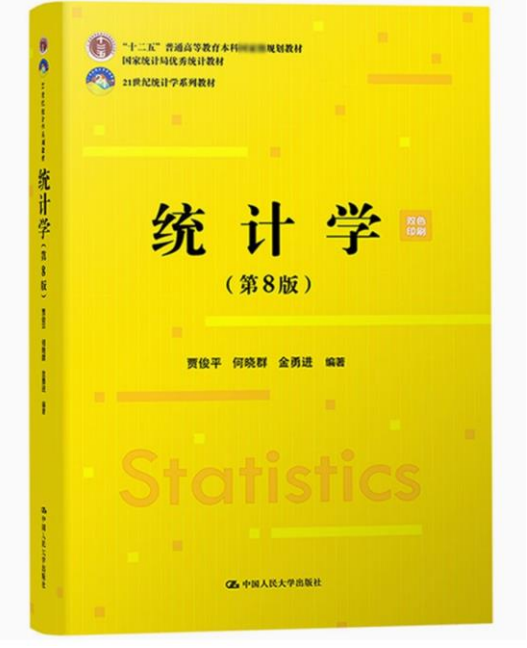
常见的理论分布

- 聚焦于平均水平的分布：正态分布、泊松分布
- 没有焦点的均质分布：均匀分布
- 不均衡的分布：幂律分布

概率论与统计学

概率论与统计学有什么不同？





概率论与统计学有什么不同?

- **概率论：已知概率模型，求一个事件的概率**
  - 一个**公平的骰子**，掷出偶数点的概率是多少?
  - 一局**随机发牌**的狼人杀游戏，右置位没有狼的概率是多少?

- **统计学：概率模型未知，从观察数据中建立模型**
  - 根据骰子每次掷出的**点数**，判断它是否灌铅?
  - 根据街道上每天的**停车数量**，判断其属于泊松分布还是均匀分布?

概率论的核心知识

- 不同形式的概率分布
- 期望（均值）和方差
- 独立性和条件概率

泊松分布

正态分布

均匀分布

幂律分布

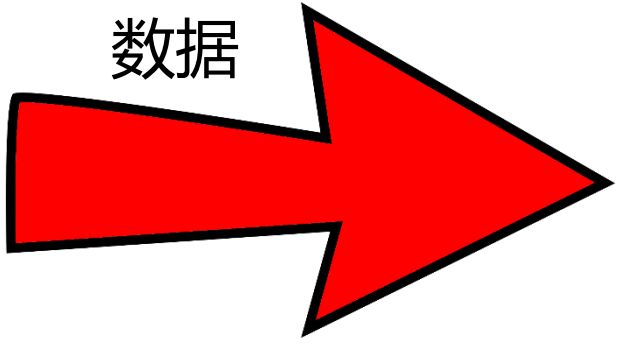
……

两事件独立：一件事情的结果对另一件事情的结果否毫无影响。

$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

$$P(AB) = P(A) \times P(B)$$

统计学知识



描述性统计

对数据的描述和总结

推断性统计

从样本推断总体

相关分析

两个变量之间有关联吗?

线性回归

自变量如何影响因变量?

Logistic回归

对分类因变量的分析

离散选择模型

对选择结果的分析

……

统计软件

操作极其简单

功能相当全面

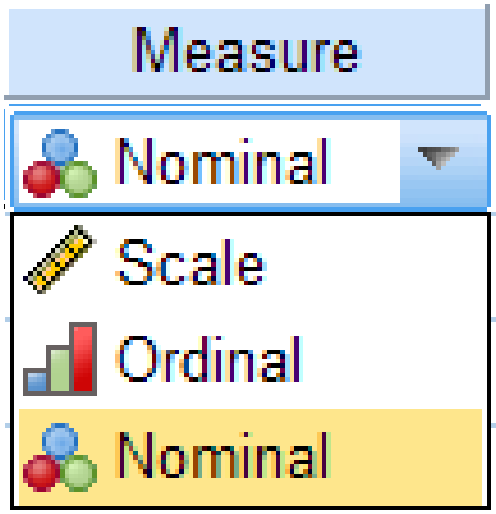
操作相对简单

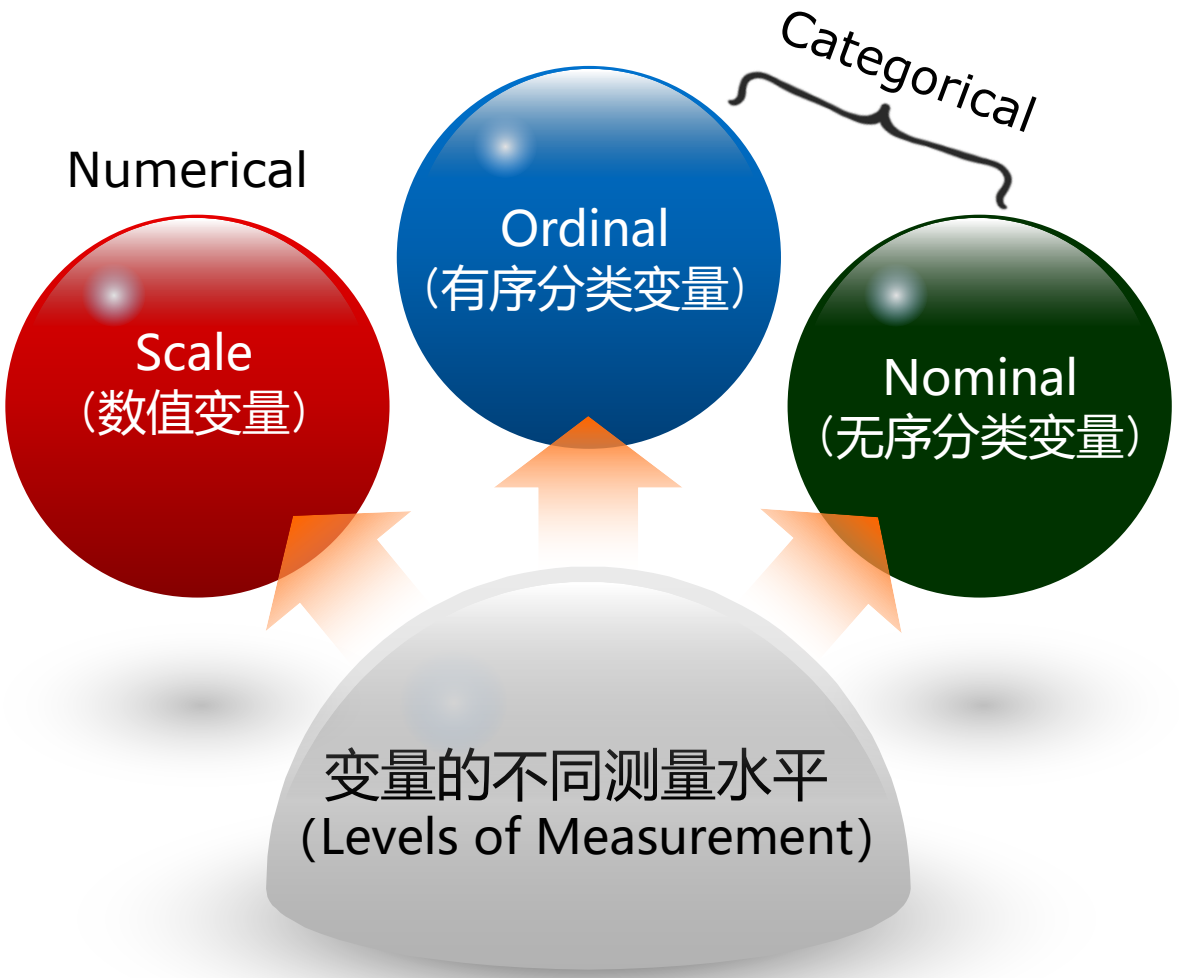
功能极其全面

- 每一行是一个样本
- 每一列是一个变量（variable）

	ID	新功能	用地面积	建筑面积	区位	道路条件	单位性质
1	1	创意产业园	.47	.57	内环内	支路	国有企业
2	2	创意产业园	1.77	3.34	内环内	支路	国有企业
3	3	创意产业园	3.34	3.96	内中环之间	干路	非国有企业
4	4	创意产业园	.83	1.87	内环内	支路	国有企业
5	5	创意产业园	.56	.82	内环内	支路	非国有企业
6	6	创意产业园	1.22	1.83	中外环之间	支路	国有企业
7	7	创意产业园	.10	.35	中外环之间	干路	国有企业
8	8	创意产业园	.39	1.42	内环内	干路	国有企业
9	9	创意产业园	.66	2.98	内环内	干路	国有企业
10	10	创意产业园	.45	1.44	内环内	干路	国有企业
11	11	创意产业园	2.11	1.80	内中环之间	支路	国有企业
12	12	创意产业园	.87	.82	内中环之间	干路	国有企业
13	13	创意产业园	.66	1.48	内中环之间	支路	国有企业

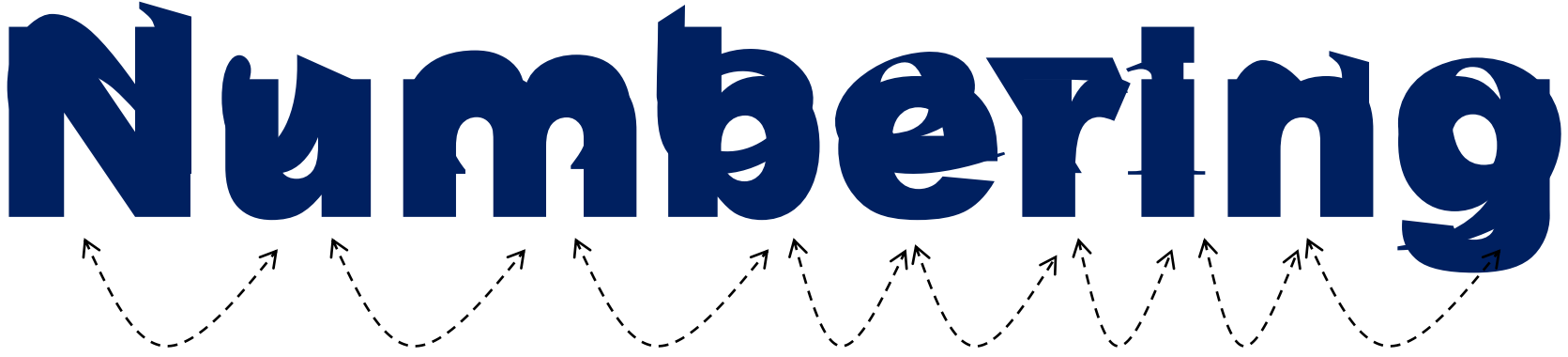
不同类型的变量


Measure
Nominal
Scale
Ordinal
Nominal


Categorical
Numerical
Ordinal
（有序分类变量）
Scale
（数值变量）
Nominal
（无序分类变量）
变量的不同测量水平
(Levels of Measurement)

不同类型的变量

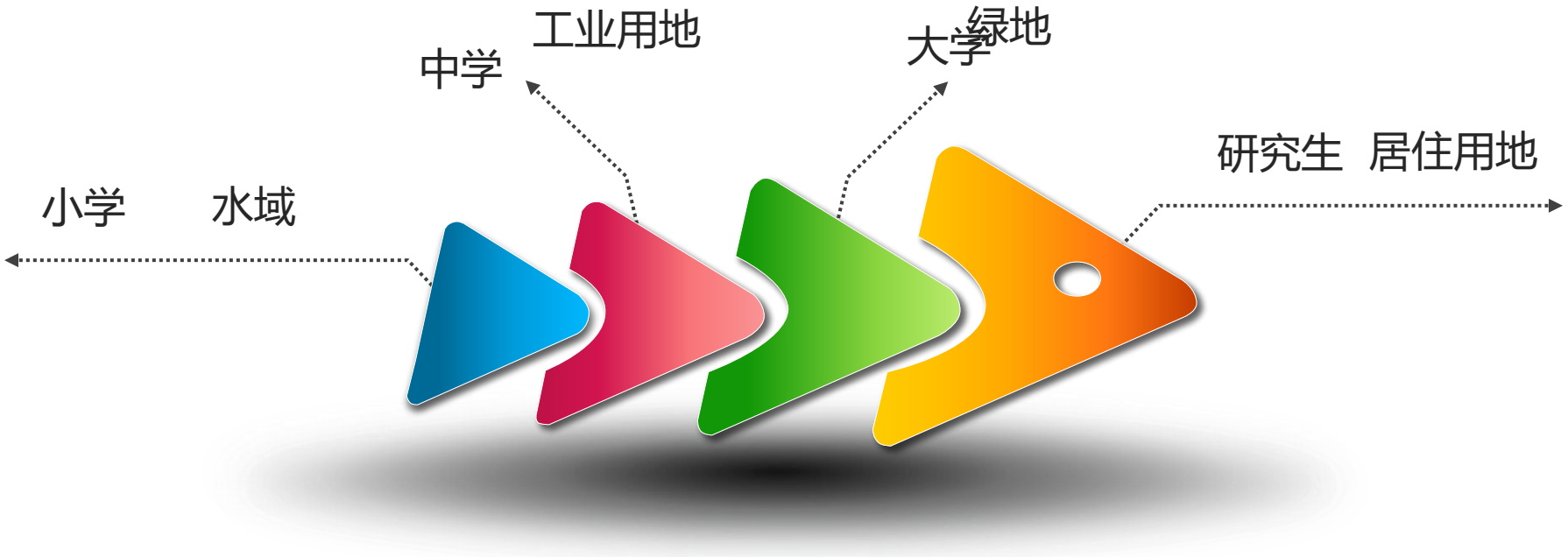
数值变量 Scale, Numerical



不同类型的变量

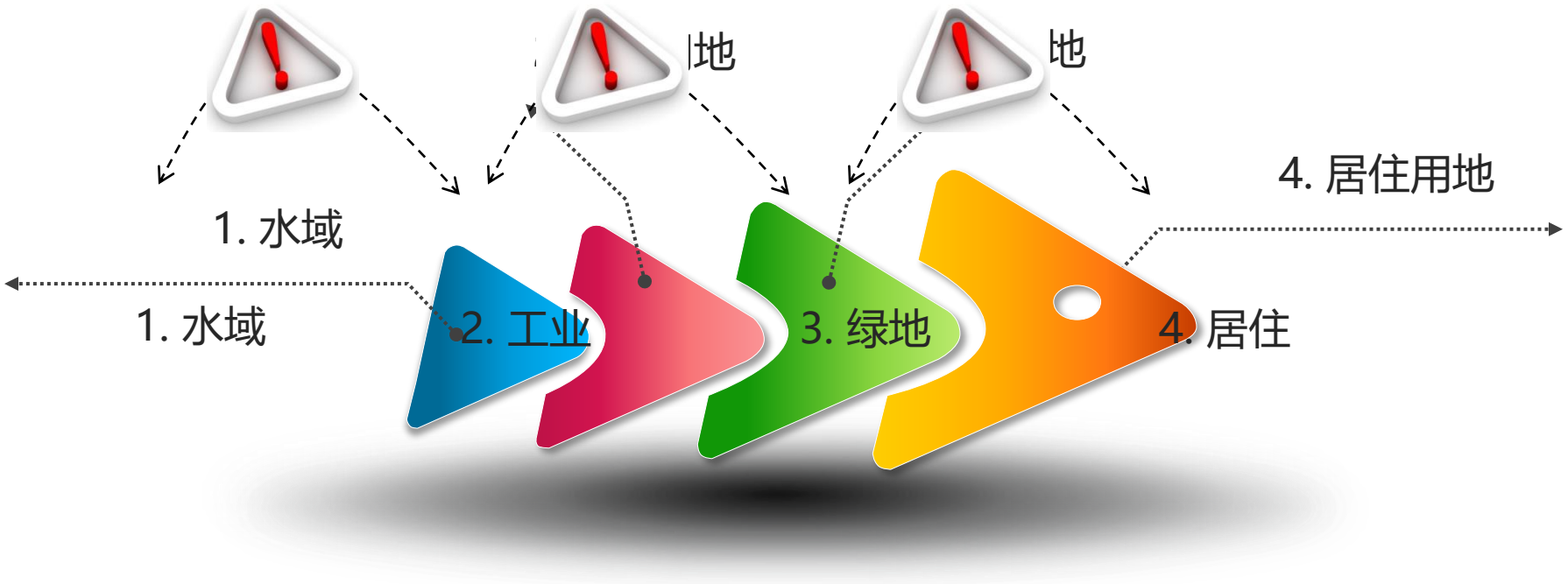
分类变量 Categorical

- 受教育程序是一个有序分类变量（ordinal）
- 用地性质是一个无序分类变量（nominal）



分类变量 Categorical

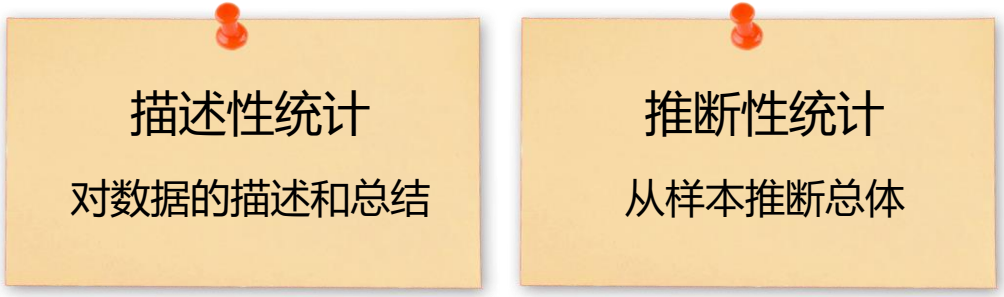
- 数字编码没有数值意义：1→2，2→3，3→4的效应可以完全不同。



大纲

- 什么概率和分布?
- 什么是概率论与统计学?
- 变量有几种类型?
- **有哪些常见的描述性统计方法?**
- **有哪些常见的统计图?**

描述性统计方法

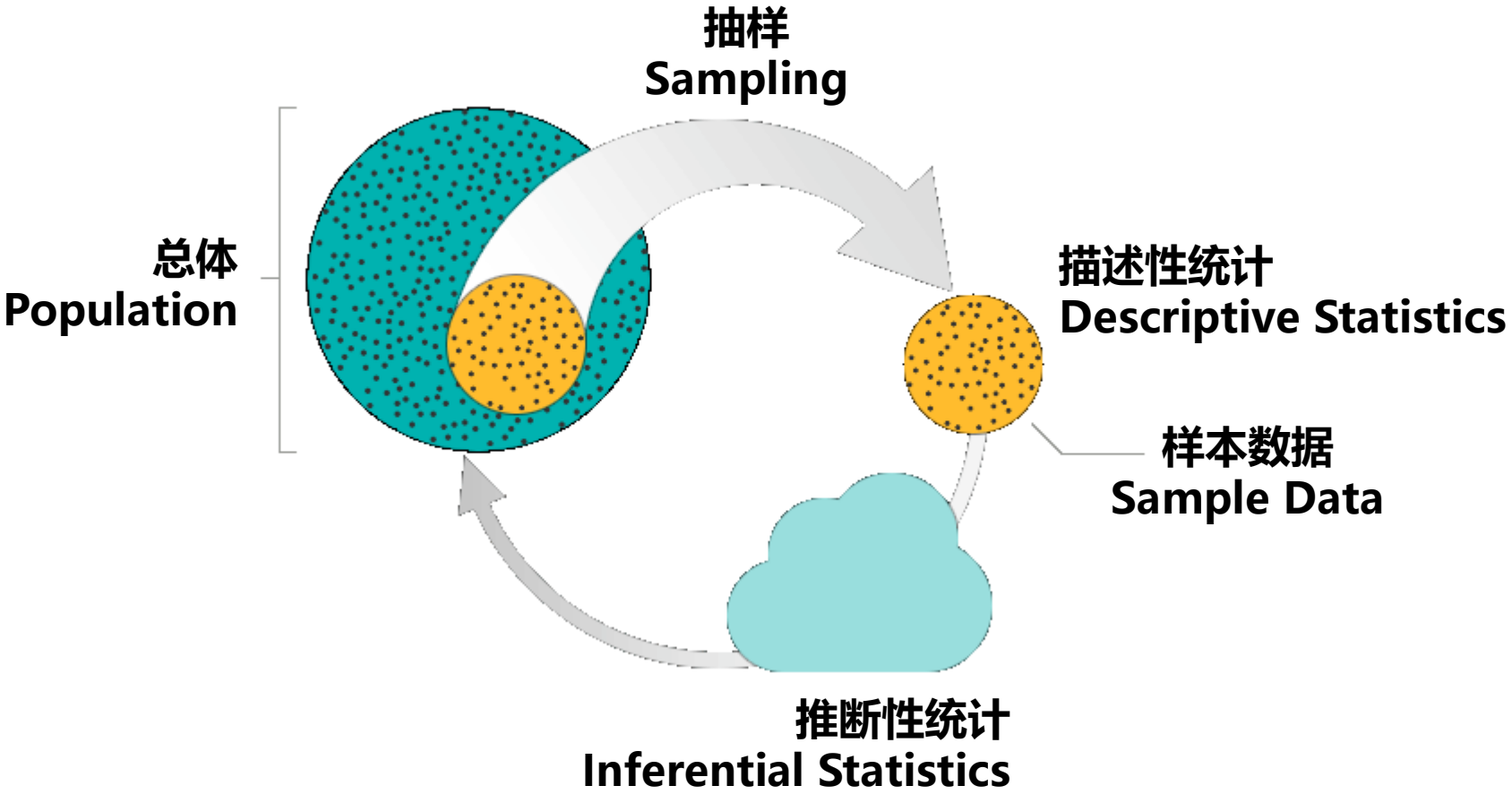


Descriptive Statistics

Inferential Statistics

统计学
Statistics

描述性统计方法


抽样
Sampling
总体
Population
描述性统计
Descriptive Statistics
样本数据
Sample Data
推断性统计
Inferential Statistics

描述性统计方法

分类变量：频数统计

更新功能的频数统计

	频数	频率
创意产业园	13	11.5%
商务办公	33	29.2%
商业服务	30	26.5%
生产物流	37	32.7%
全体	113	100%

区位的频数统计

	频数	频率	累积频率
内环内	63	55.8%	55.8%
内中环之间	22	19.5%	75.2%
中外环之间	28	24.8%	100.0%
全体	113	100%	

- 频数：每个类别出现的次数
- 频率 = 频数 / 样本数

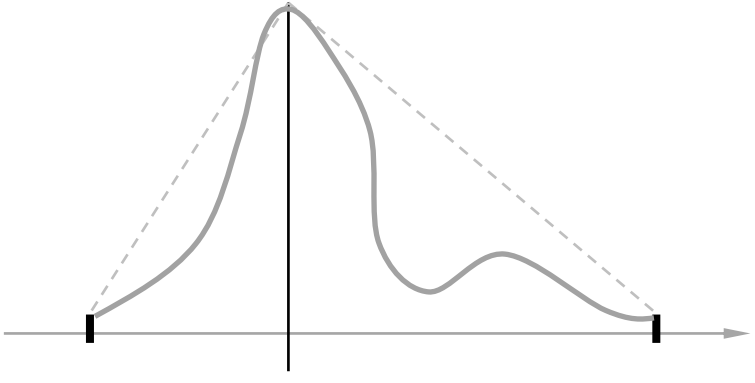
数值变量：统计量

- 统计量 (statistic)：从样本数据中计算得到的一个数值，用来描述某个特定的统计特征。
- 描述**集中**趋势

  均值、中位数、众数
- 描述**离散**趋势

  方差、标准差、最值、极差、变异系数
- 描述**分布**特征

  分位数、偏度、峰度



描述性统计方法

数值变量：集中趋势统计量

1, 1, 2, 3, 4, 5, ~~6~~
1000

- 均值 (mean)
  - 定义：$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
  - 考虑了所有数据点，但易受极端值的影响

mean = ~~3.14~~
145.14

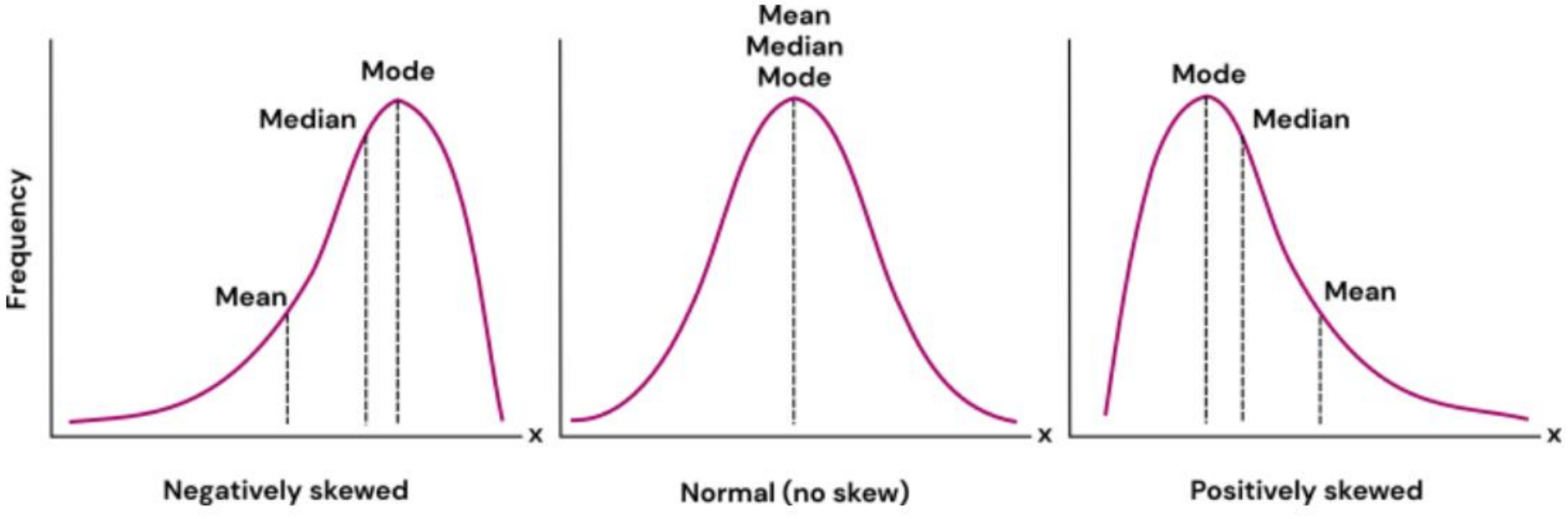
- 中位数 (median)
  - 定义：将数据从低到高排列后，位于中间的数据点
  - 只与排序后的中间值相关，不易受极端值影响，相对稳健

median = ~~3~~
3

- 众数 (mode)
  - 定义：出现频率最高的数据点

mode = ~~1~~
1

数值变量：集中趋势统计量



Negatively skewed

Normal (no skew)

Positively skewed

负偏/左偏/左长尾

正态/对称

正偏/右偏/右长尾

大部分数据集中在右侧高值区
少量低值分布在左侧较远的位置

大部分数据集中在左侧低值区
少量高值分布在右侧较远的位置

描述性统计方法

数值变量：离散趋势统计量

- 极差 (range)

  最大值 – 最小值

- 方差 (variance)

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- 标准差 (standard deviation, std)

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{\bar{X}})^2}{n-1}}$$

自由度 (degrees of freedom, df)

用于估计的独立信息的数量，等于独立数据点的数量减去估计过程中间步骤的参数数量

为什么用平方?

因为方差其实是中心矩的一种特例

为什么不是除以n?

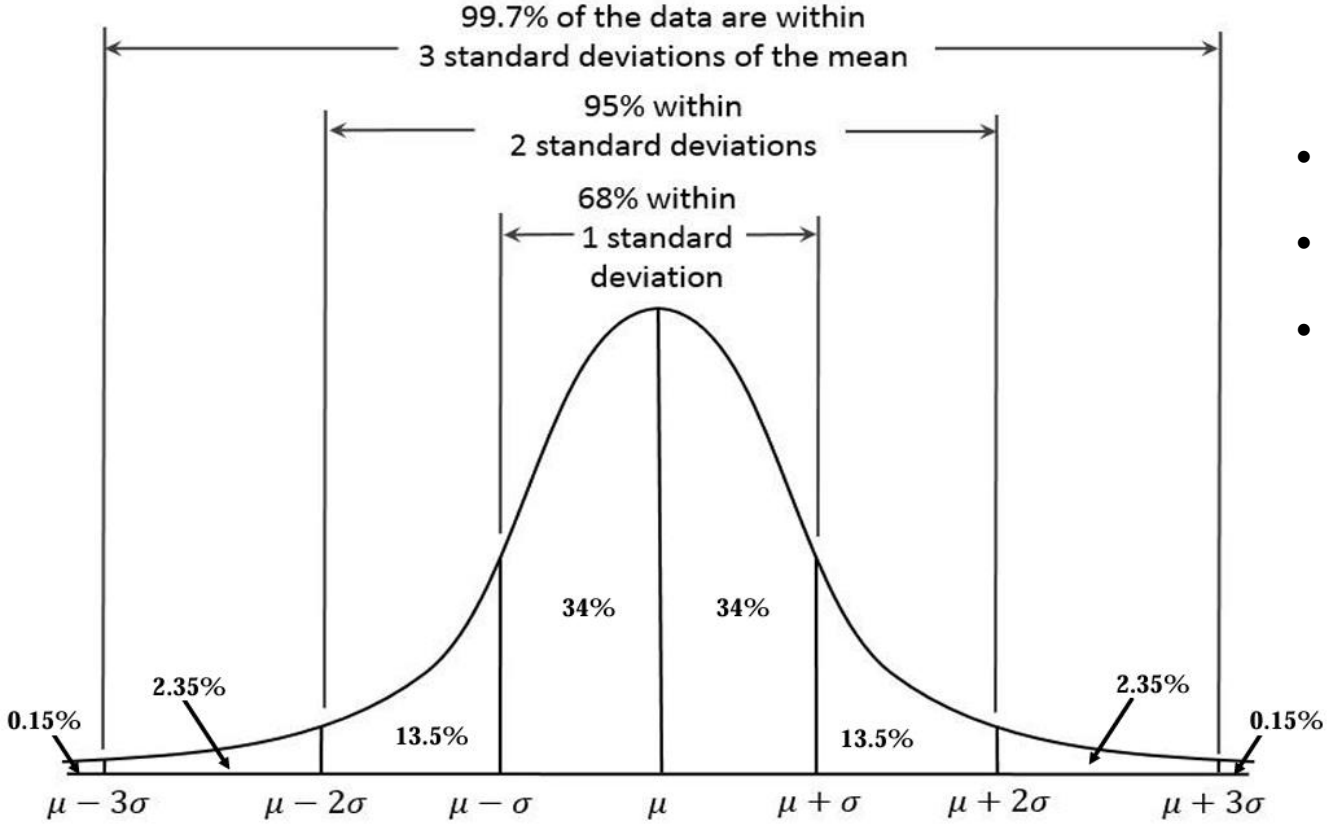
因为计算均值已经用了一个自由度，需调整

为什么不直接用方差?

因为标准差与原始数据具有相同的单位和尺度，更加直观

描述性统计方法

经验法则（empirical rule）/ 三西格玛法则（3sr）



基于正态分布

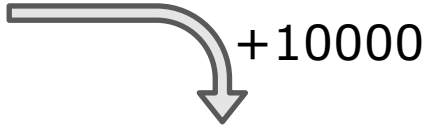
- 1倍标准差：68%
- 2倍标准差：95%
- 3倍标准差：99.7%

数值变量：离散趋势统计量

- 变异系数 (coefficient of variation, CV)
  - 定义：标准差除以均值 $CV = \frac{s}{\bar{X}} \times 100\%$
  - 相对离散程度：CV越大，说明数据的离散程度相对于均值越大。

1, 2, 3, 4, 5, 6, 7

$\bar{X} = 4$， $s = 2$ ， $CV = 50\%$



10001, 10002, 10003, 10004, 10005, 10006, 10007

$\bar{X} = 10004$， $s = 2$ ， $CV = 0.02\%$

  - 无量纲，可以在不同变量之间进行比较。

数值变量：分布特征统计量

- 分位数，百分位数 (quantile, percentile)
  - 将数据从小到大排列后，根据指定的比例位置找到对应的数值。
  - p%分位数（第p百分位数）：数据集中有p%的样本小于或等于该值。

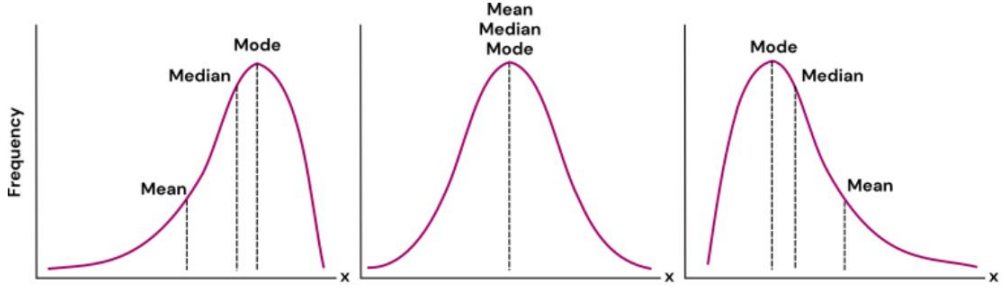
> 对于收入在10%分位数的家庭，“宽带中国”使其收入增加了29.7%；
> 对于收入在50%分位数的家庭，“宽带中国”使其收入增加了10.8%。

  - 中位数就是50%分位数。
  - 四分位数：25%，50%，75%。
  - 极端值：1%，5%，95%，99%。

数值变量：分布特征统计量

- 偏度 (skewness)

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s}\right)^3$$



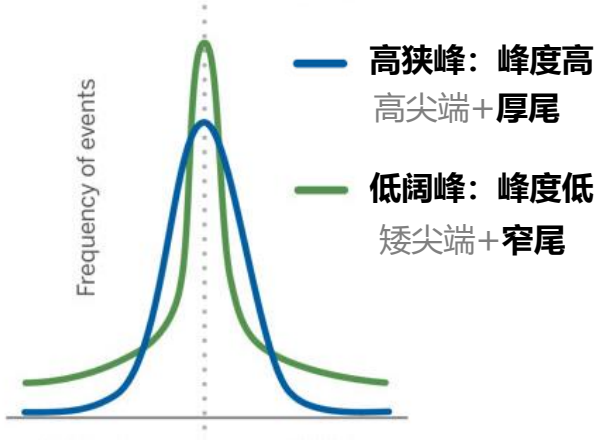
负偏/左偏/左长尾　　**正态/对称**　　**正偏/右偏/右长尾**

偏度 < 0　　**偏度 = 0**　　**偏度 > 0**

- 峰度 (kurtosis)

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s}\right)^4-\frac{3(n-1)^2}{(n-2)(n-3)}$$

  - 正态分布峰度=3，但被大多软件置为0。
  - 正、负峰度：与正态分布相比的高、低峰度。



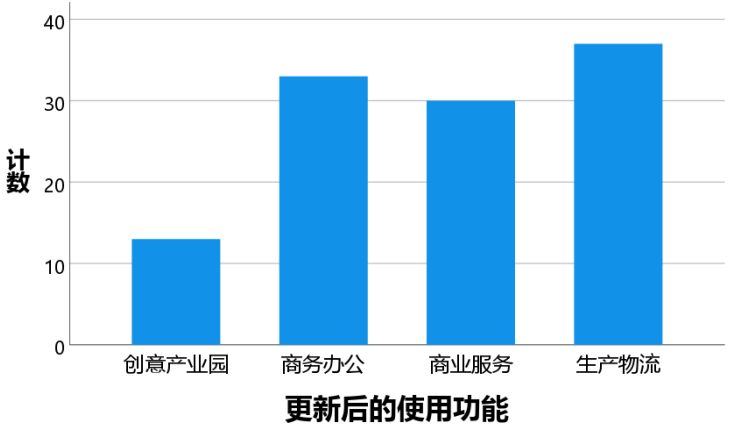
描述性统计方法

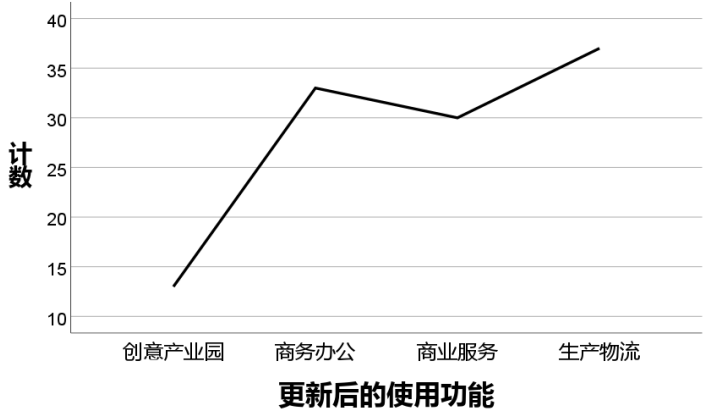
数值变量

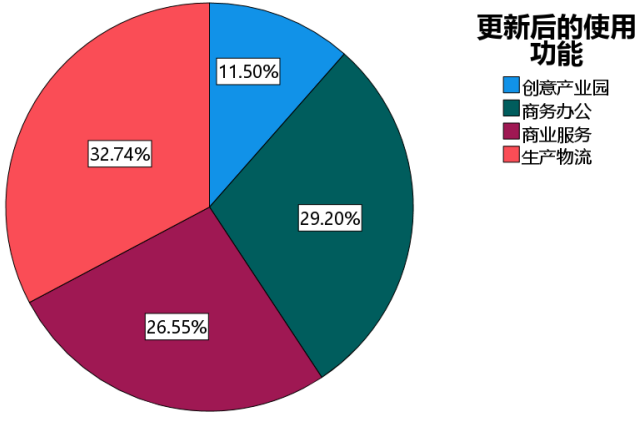
统计量		用地面积 (ha)	到市中心距离 (km)
均值		0.5133	1.1509
中位数		0.3600	0.8869
众数		0.66	0.10
方差		0.318	0.638
标准差		0.56385	0.79891
变异系数		119.85%	69.42%
偏度		2.759	1.016
峰度		9.194	0.011
最小值		0.03	0.10
最大值		3.34	3.30
极差		3.31	3.20
分位数	5%	0.0696	0.2091
	25%	0.1583	0.6119
	75%	0.6475	1.5027
	95%	1.8570	2.7610

统计图

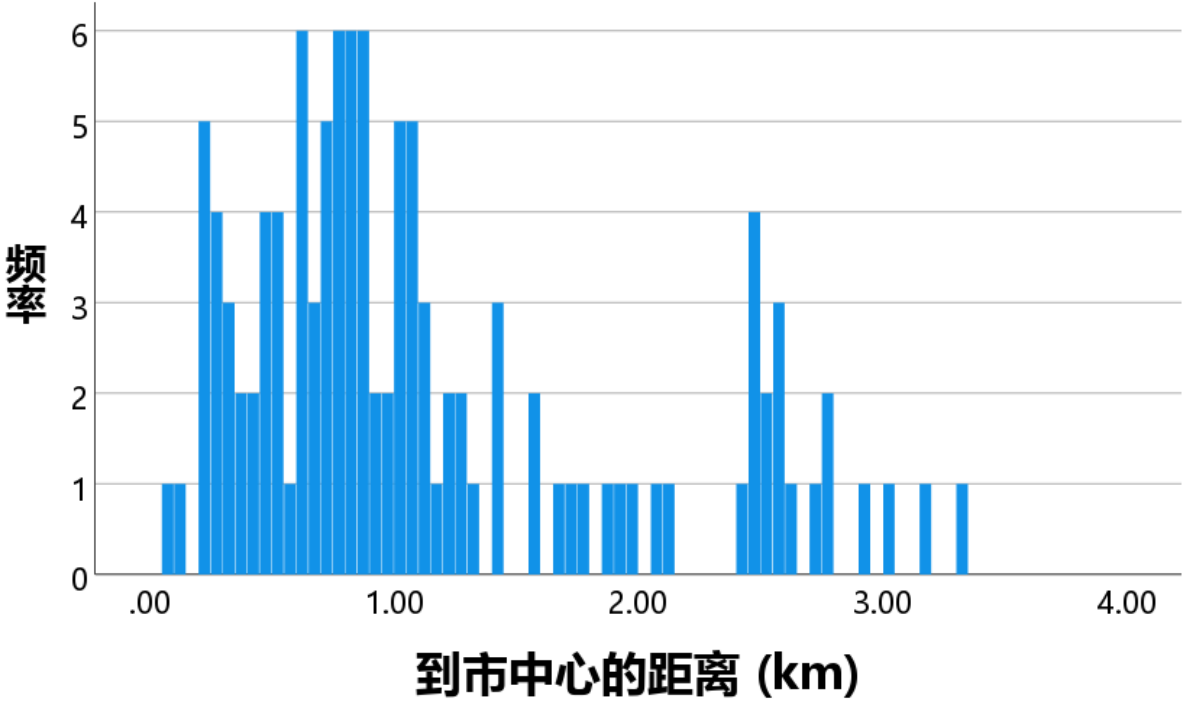
分类变量：柱状图、饼状图、折线图





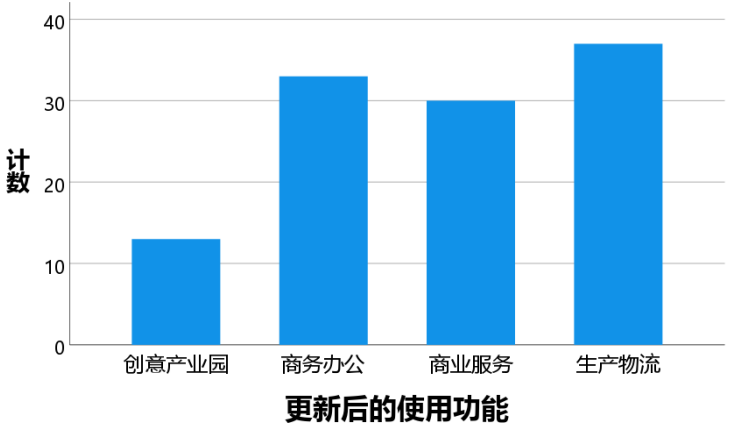


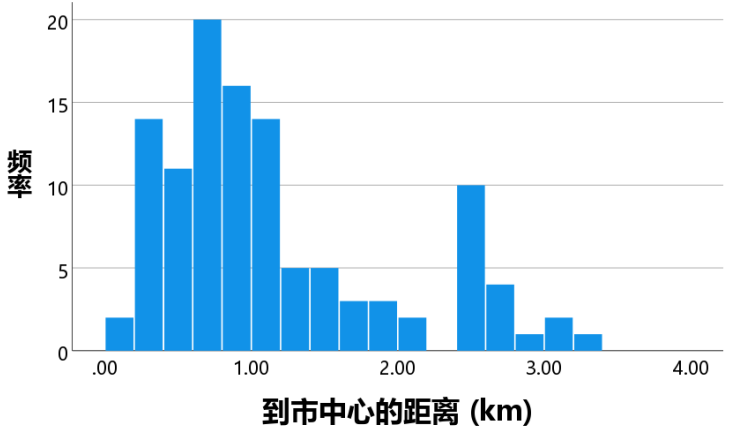
数值变量：直方图 (histogram)



- 把数据划分为一系列等宽的区间 (bin)；
- 每个柱子的高度代表了该区间的数据量，即频数或频率。
- 可根据需要调整bin的宽度。

直方图 vs. 柱状图





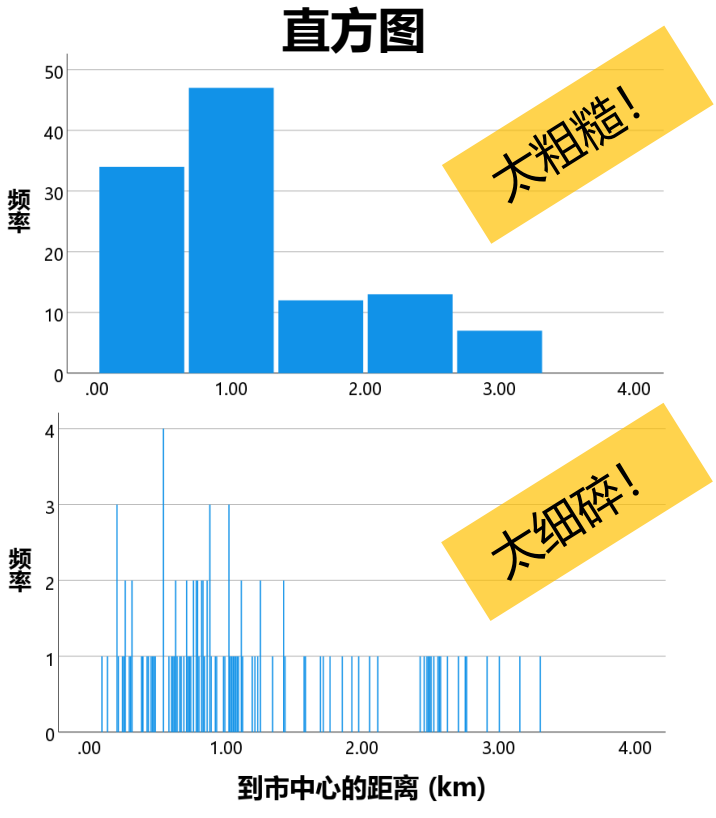
柱状图

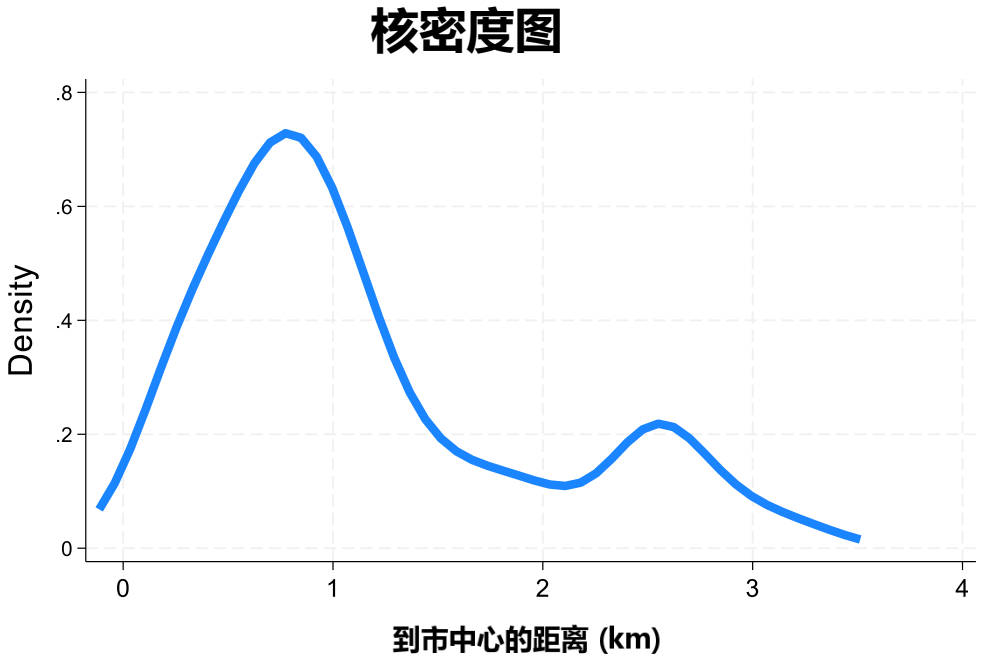
- 适用于分类变量
- 柱子之间有间隔，顺序大多不重要
- 类别是给定的

直方图

- 适用于连续的数值变量
- 柱子之间无间隔，顺序不可调
- 类别是自定的

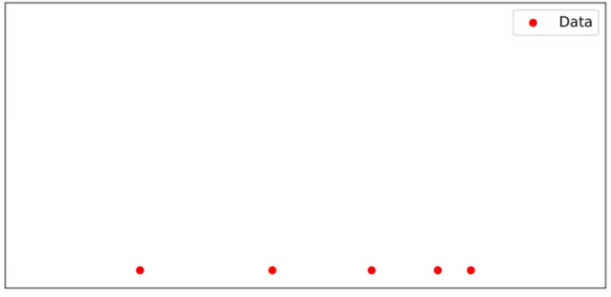
数值变量：核密度图 (kernel density)

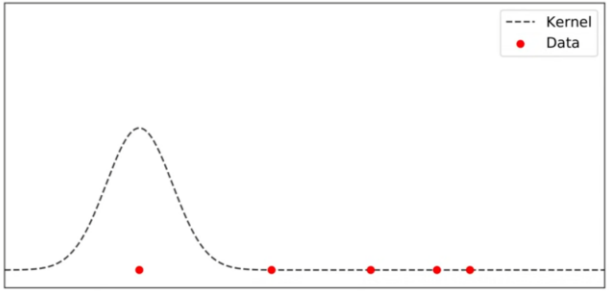


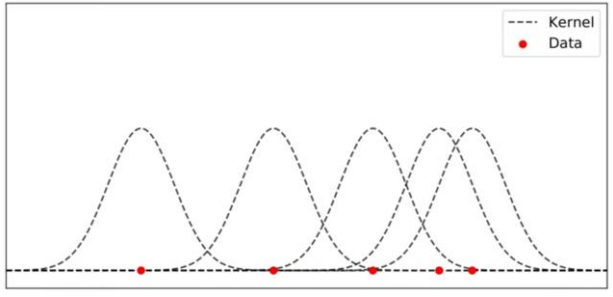


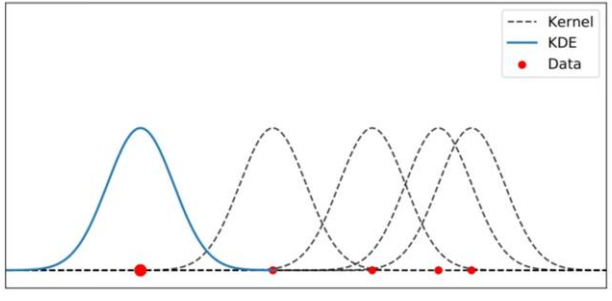
统计图

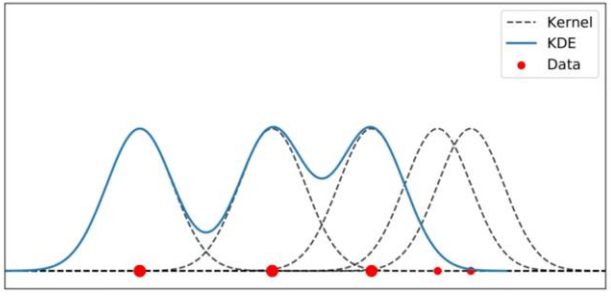
数值变量：核密度图 (kernel density)

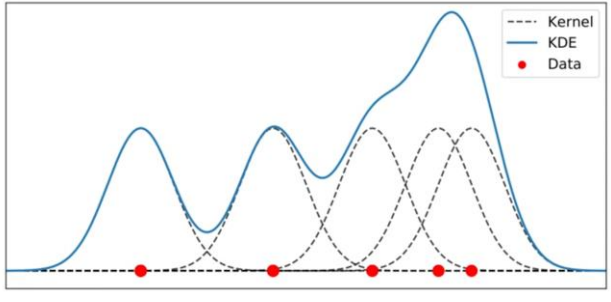












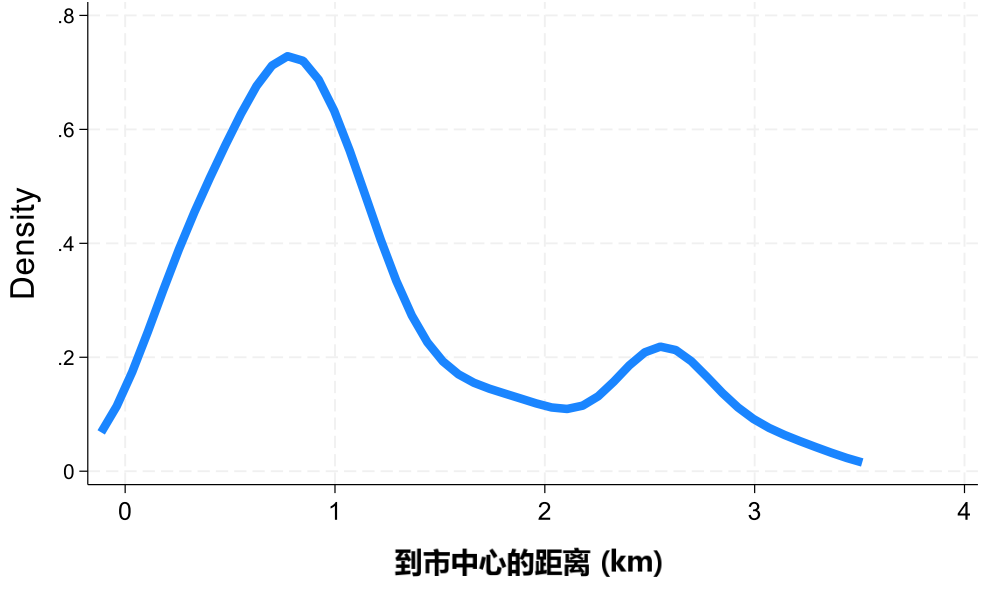
数值变量：核密度图（kernel density）

- 核密度图将数据分布进行平滑处理，生成一条连续的曲线来表示数据的分布情况。

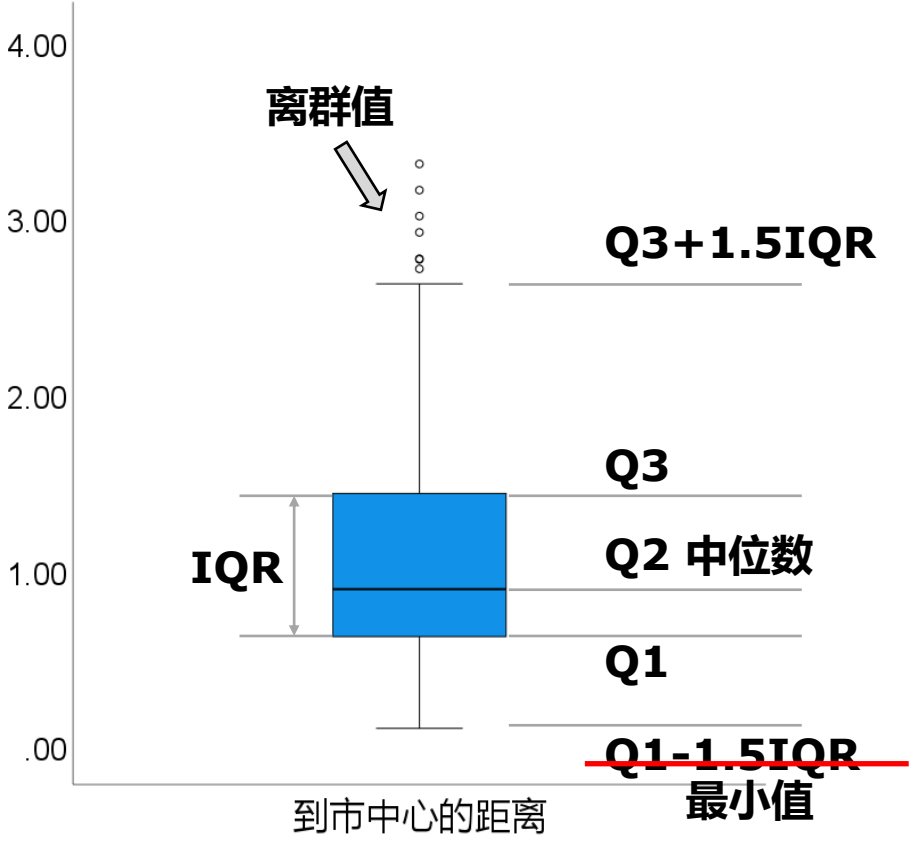
$$f(x) = \frac{\sum_{i=1}^{n} K(x - X_i)}{n}$$

$K(\cdot)$：核函数

- 适用于连续的数值变量。
- 比直方图更细致。

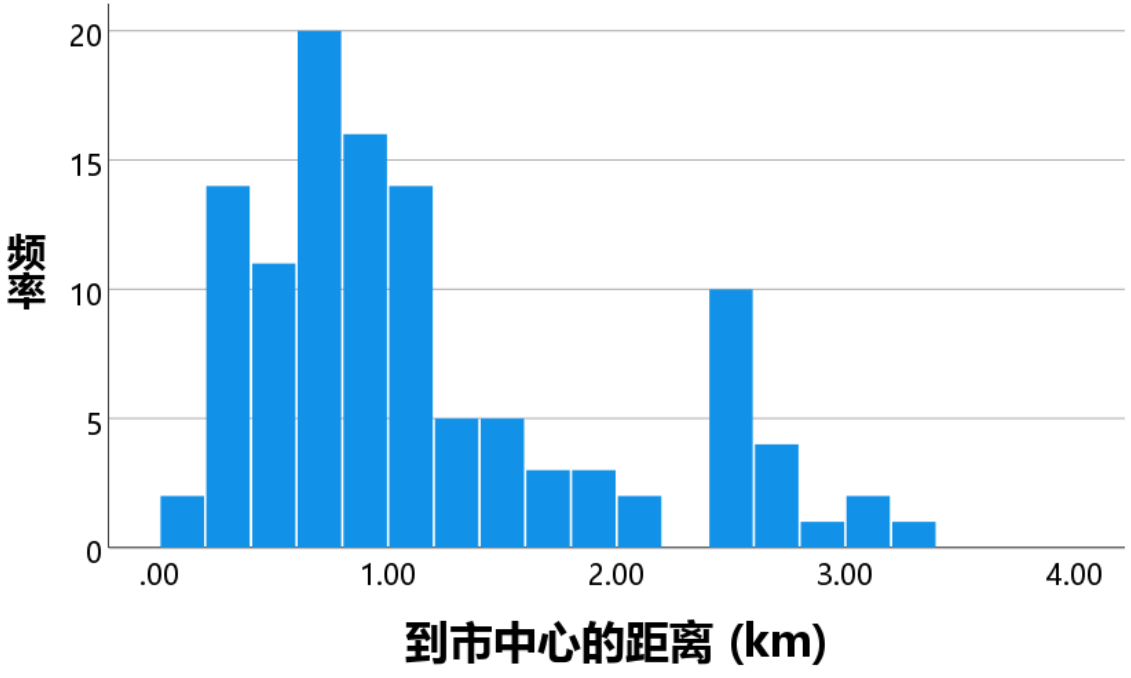


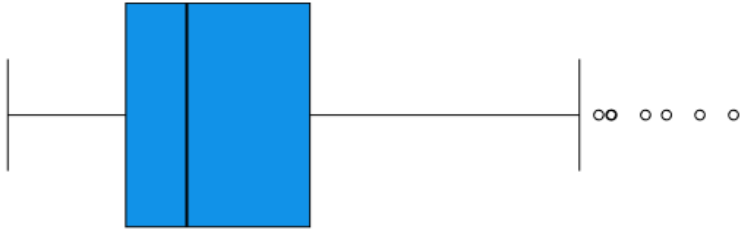
数值变量：箱形图 (box plot)



- 中位数：箱体中的粗线
- 四分位数：
  - 下四分位数$Q3$ （25%分位数）
  - 上四分位数$Q1$ （75%分位数）
  - 四分位距：$IQR = Q3 - Q1$
- 须：
  - 上界：$Q3 + 1.5IQR$
  - 下界：$Q1 - 1.5IQR$
  - 不得超过最大/最小值
- 离群值 (outliers)：
  - 超出上下“须”，单独标记

统计图



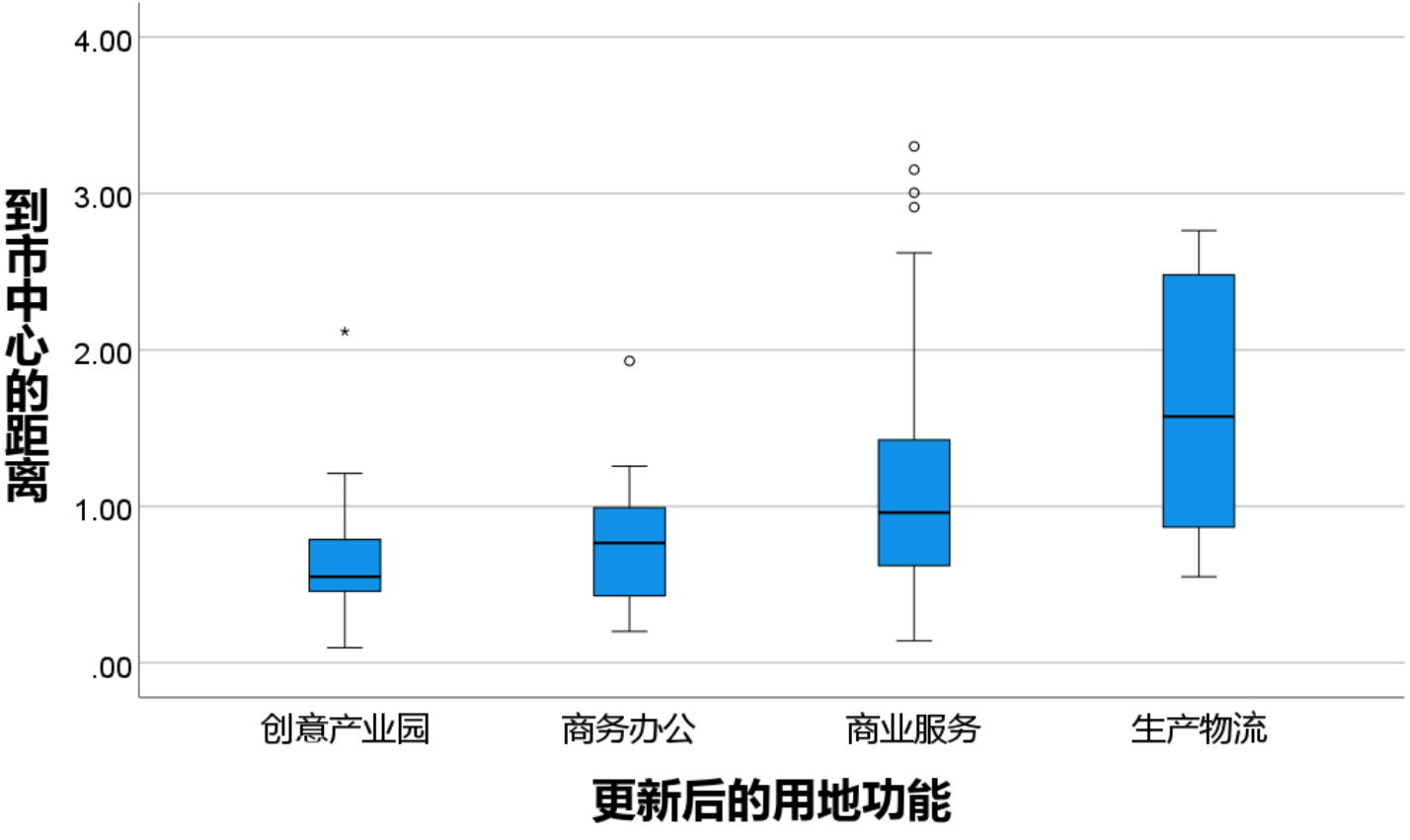


数值变量：箱形图

与直方图相比，箱形图……

- 概括性强，可以显示多个关键统计量；
- 更容识识别离群值和异常值；
- 特别适合多组比较。

数值变量：箱形图 (box plot)



总结

- 分布反映了随机变量每种取值发生的概率。常见的分布包括聚焦于均值的泊松分布（离散）、正态分布（连续），均质的均匀分布，不平衡的幂律分布等。
- 统计学的任务是利用观察数据建立概率模型，包括描述性统计和推断性统计。
- 数据中的变量类型，以及它们常用的描述性统计方法和统计图：
  - 分类变量：
    - ✓ 频数分析
    - ✓ 统计图：柱状图，折线图，饼图
  - 数值变量：
    - ✓ 统计量：集中趋势、离散趋势、分布特征
    - ✓ 统计图：直方图、核密度图、箱形图