



# 计算社会科学导论

## —— 数据分析基础

金耀辉、许岩岩

2023年3月2日

CS1126



上海交通大学  
SHANGHAI JIAO TONG UNIVERSITY

# Kenya



肯尼亚面临哪些社会问题?



肯尼亚是一个发展中国家，面临着许多社会问题，以下是其中的一些：

1. 贫困和不平等：肯尼亚的贫困率很高，大约一半的人口生活在贫困线以下。贫困与不平等在城市和农村地区都存在，并且在不同族群和性别之间也存在差异。
2. 教育：肯尼亚的教育系统面临着很多挑战，其中包括缺乏基础设施、师资力量不足以及贫穷家庭难以支付学费等问题。这导致了教育水平的低下，特别是在农村地区和贫困家庭中。
3. 卫生：肯尼亚的卫生系统也面临着很多挑战，其中包括医疗设施不足、医生和护士的短缺、药品供应不足等。肯尼亚也面临着疟疾、艾滋病等传染病的高发率。
4. 犯罪和安全：肯尼亚的犯罪率很高，特别是在城市地区。盗窃、抢劫和贩毒是常见的犯罪行为，而政治暴力和恐怖主义也时有发生。
5. 环境污染：肯尼亚的环境问题包括土地退化、森林砍伐、土地开垦、水资源枯竭等。这些问题对当地社区的生计和经济产生了严重的影响，并且导致了生态系统的破坏。
6. 少数民族和种族关系：肯尼亚是一个多民族国家，不同的族群之间存在着紧张关系。政治家往往利用种族差异来获得选票，并且民间的种族主义和仇恨犯罪也时有发生。

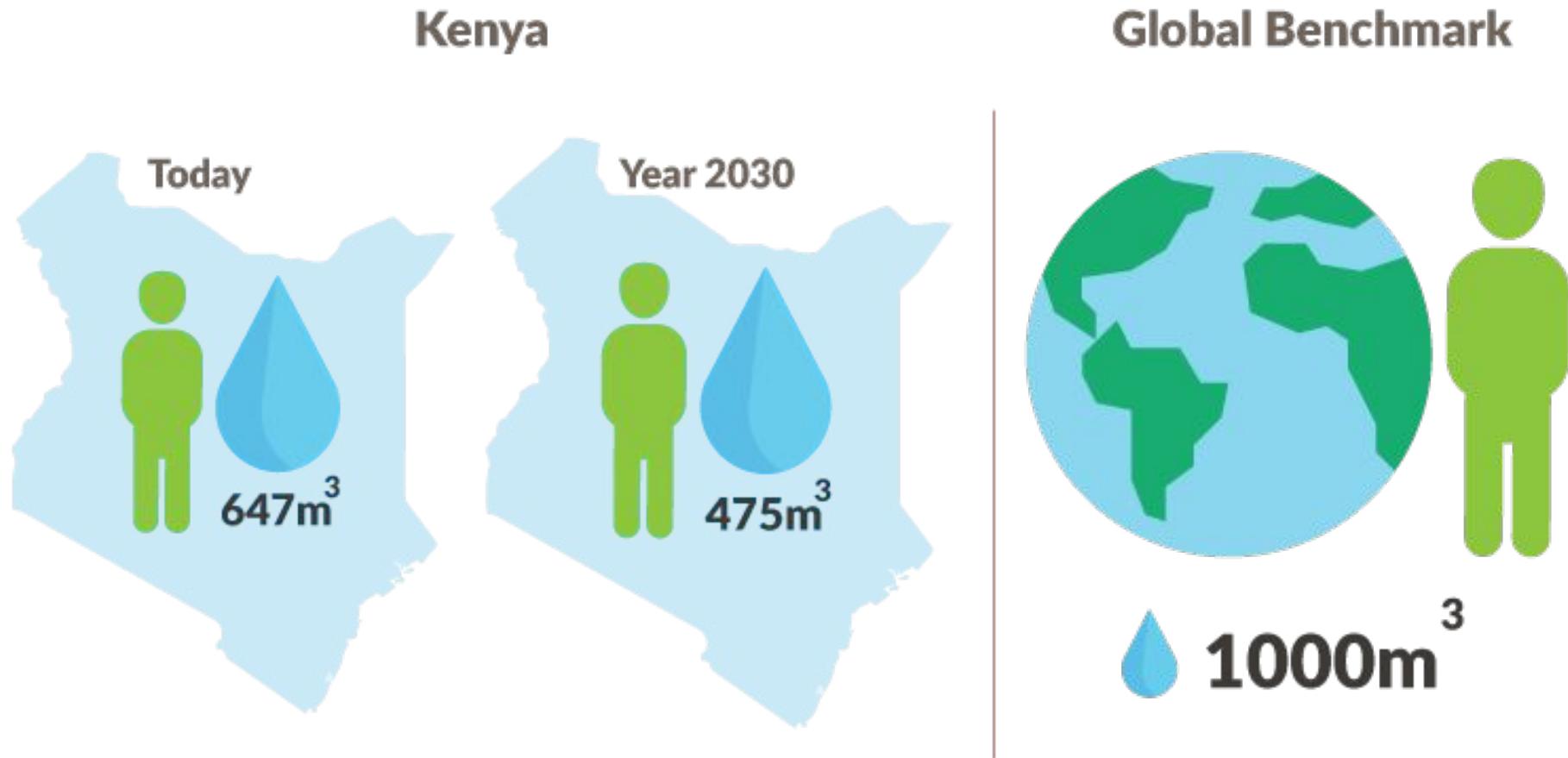


# Kenya

Water Pollution in Kenya  
Water Crisis



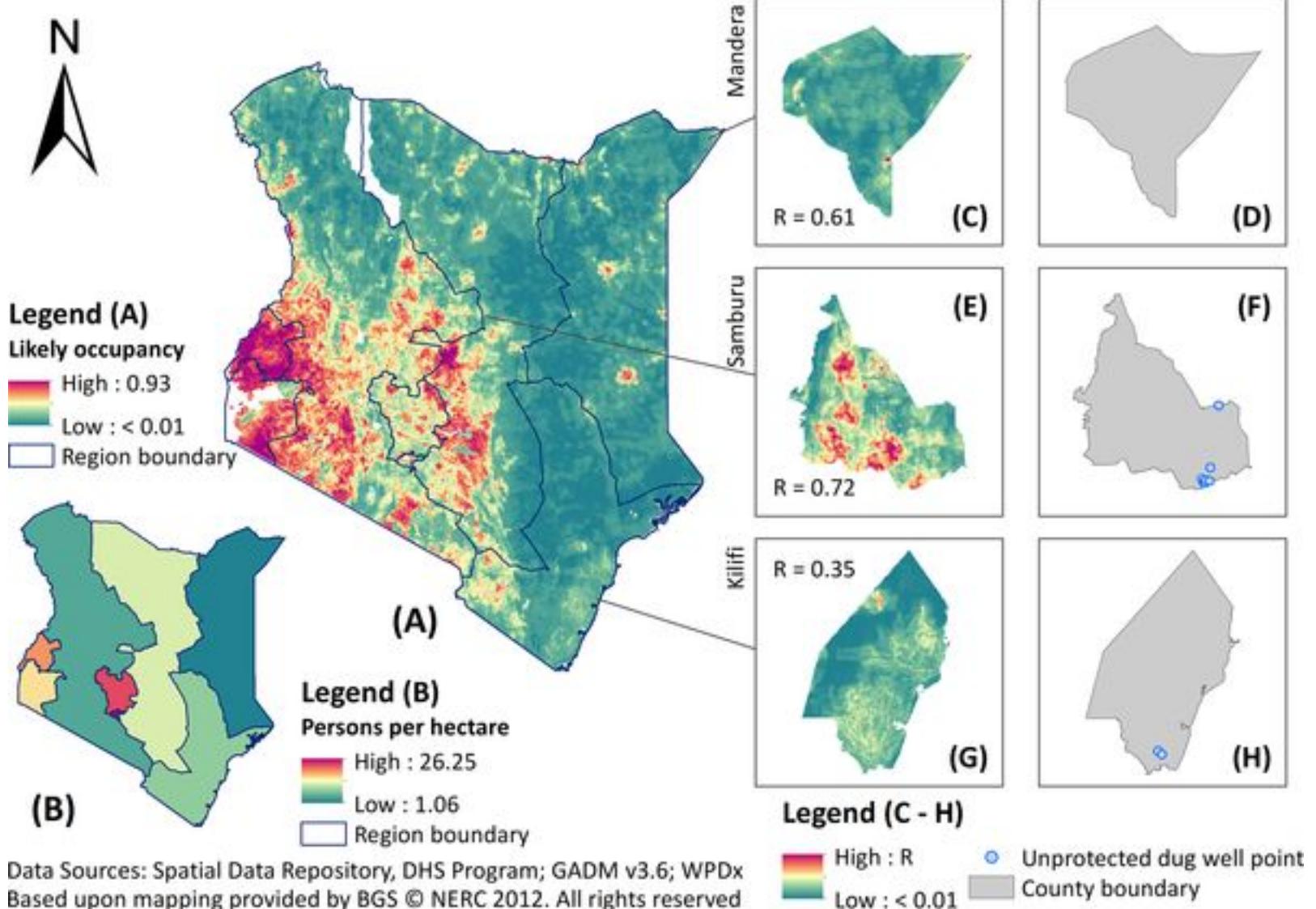
# Kenya



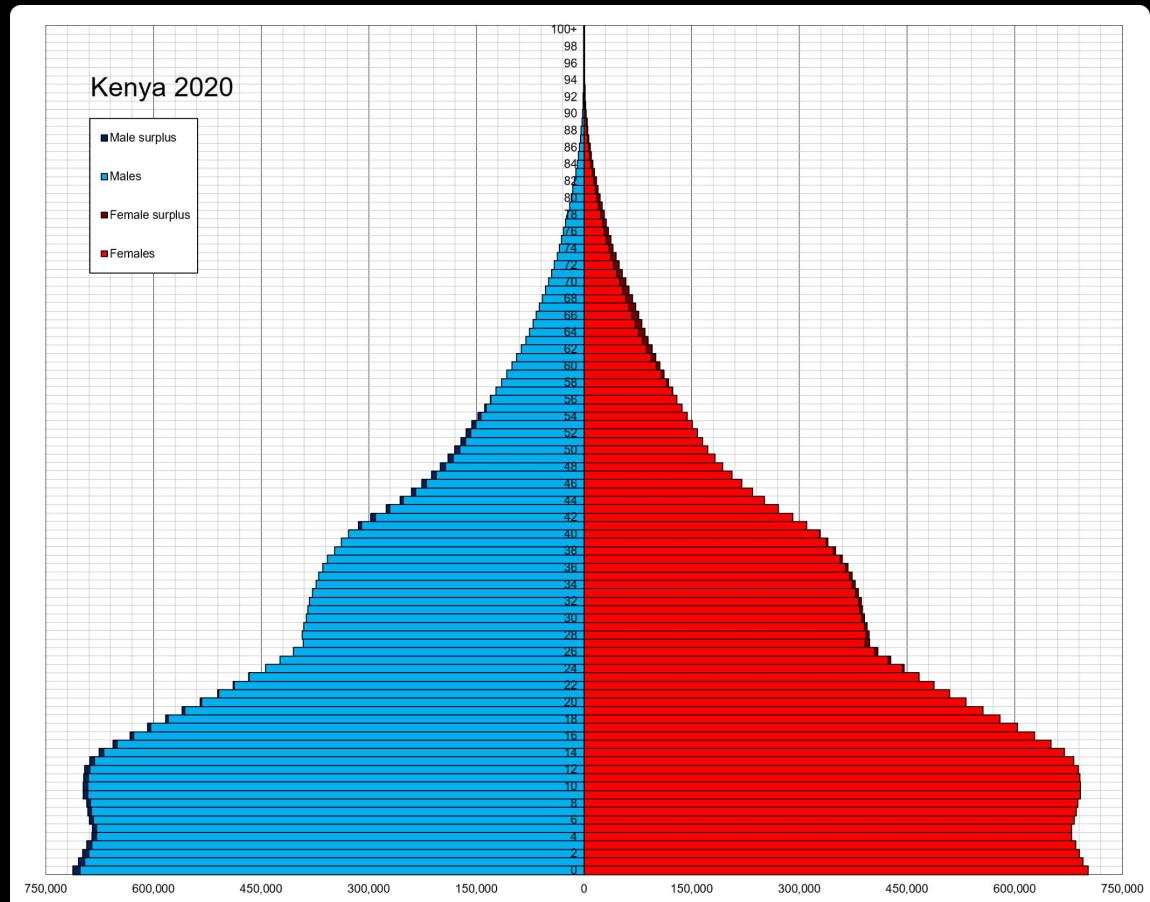
# Kenya

Predicted surface water occupancy across Kenya with inset maps highlighting Mandera, Samburu, and Kilifi counties.

Yu, Weiyu, et al. "Mapping access to domestic water supplies from incomplete data in developing countries: An illustrative assessment for Kenya." PloS one 14.5 (2019): e0216923.



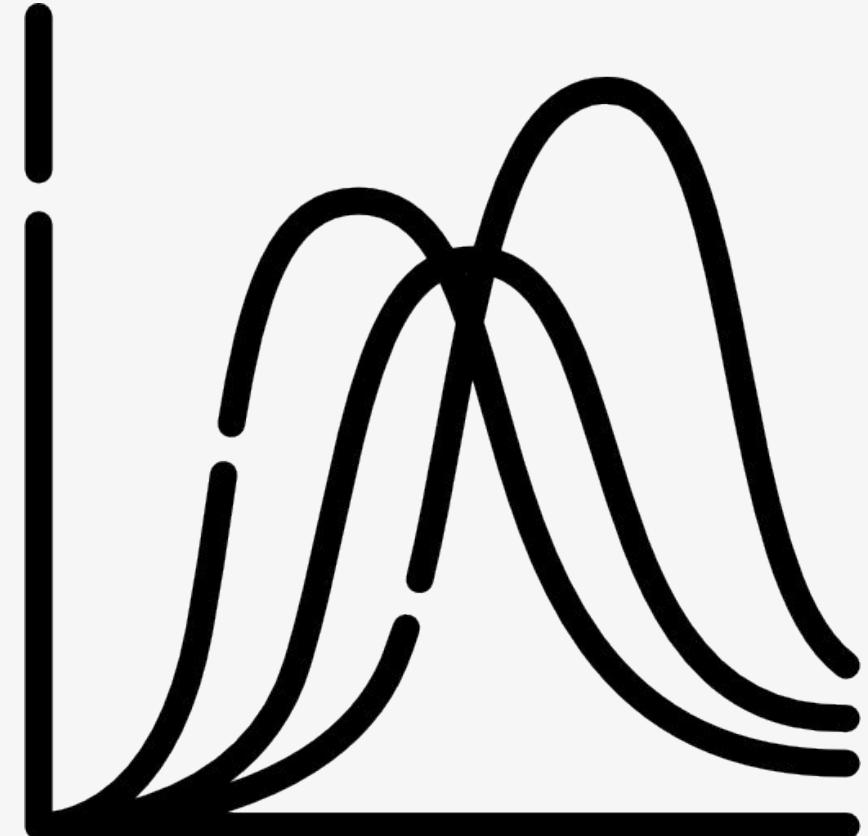
# Kenya



- 调查问卷: 调查本科毕业人群的收入分布
- 问题: 给定收入水平, 推测本科毕业的概率

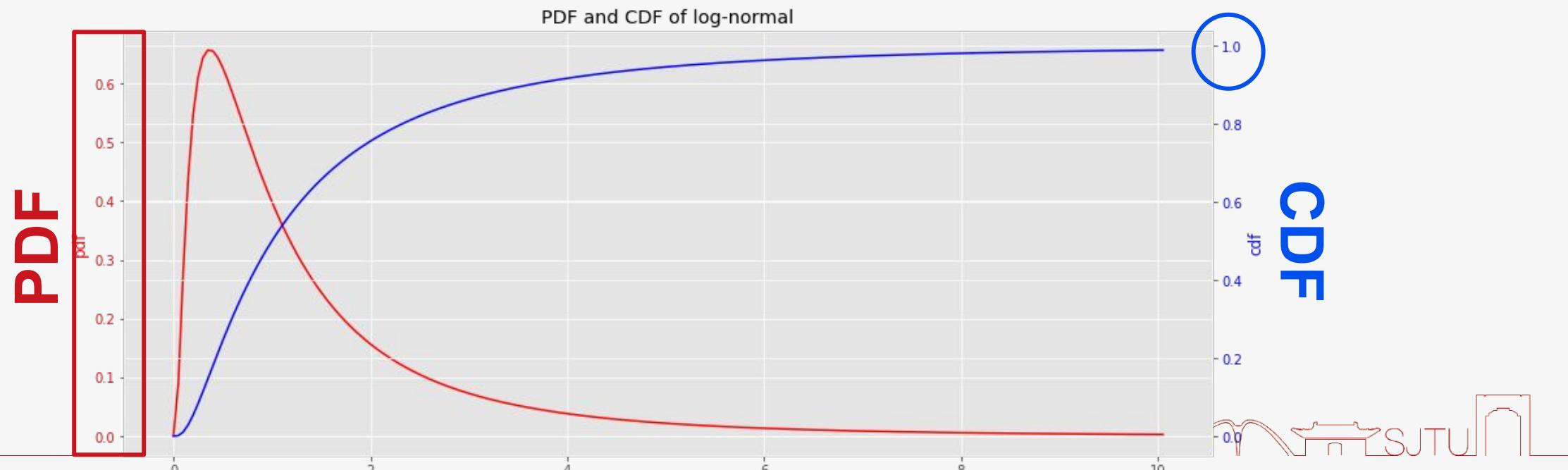


- 基本概率分布
- 量化数据分布
- 贝叶斯公式
- 函数拟合
- 数据相关性



PDF: 概率密度函数, probability density function, 描述连续型随机变量的输出值在某个取值点附近的可能性大小的函数。

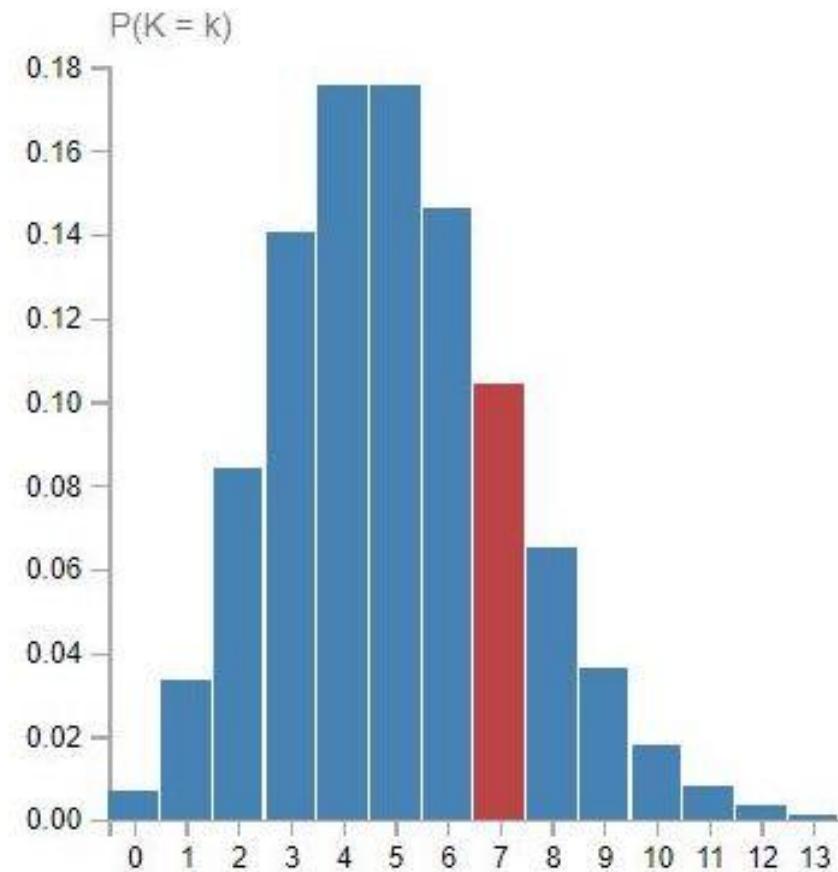
CDF: 累计分布函数, cumulative distribution function, 是概率密度函数从负无穷到某个位置 $x$ 的积分(曲线下的面积), 函数在 $x$ 的取值表示变量小于等于 $x$ 的概率。



# PDF vs. CDF

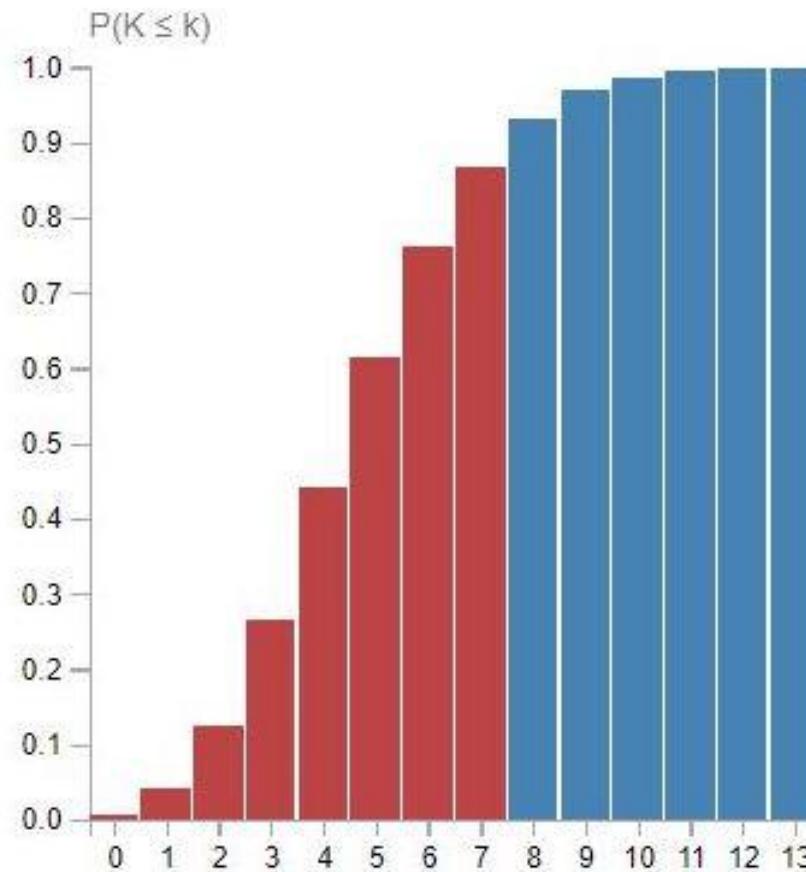
Probability mass function

$$P(K = 7) = 0.10444$$



Cumulative distribution function

$$P(K \leq 7) = 0.86663$$





Quiz:

$x$ 服从一个0-1的均匀分布

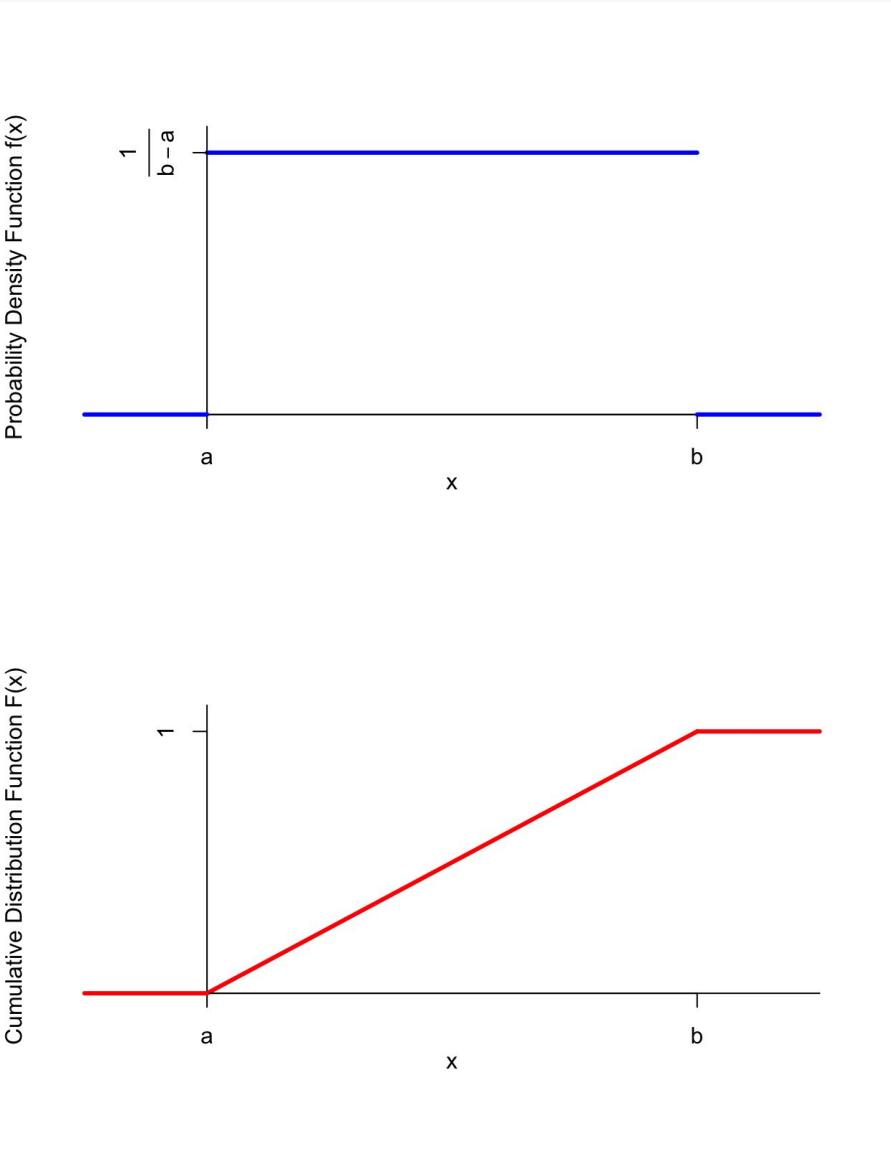
$x$ 的PDF为？

$x$ 的CDF为？

$P(x=0.5)=?$

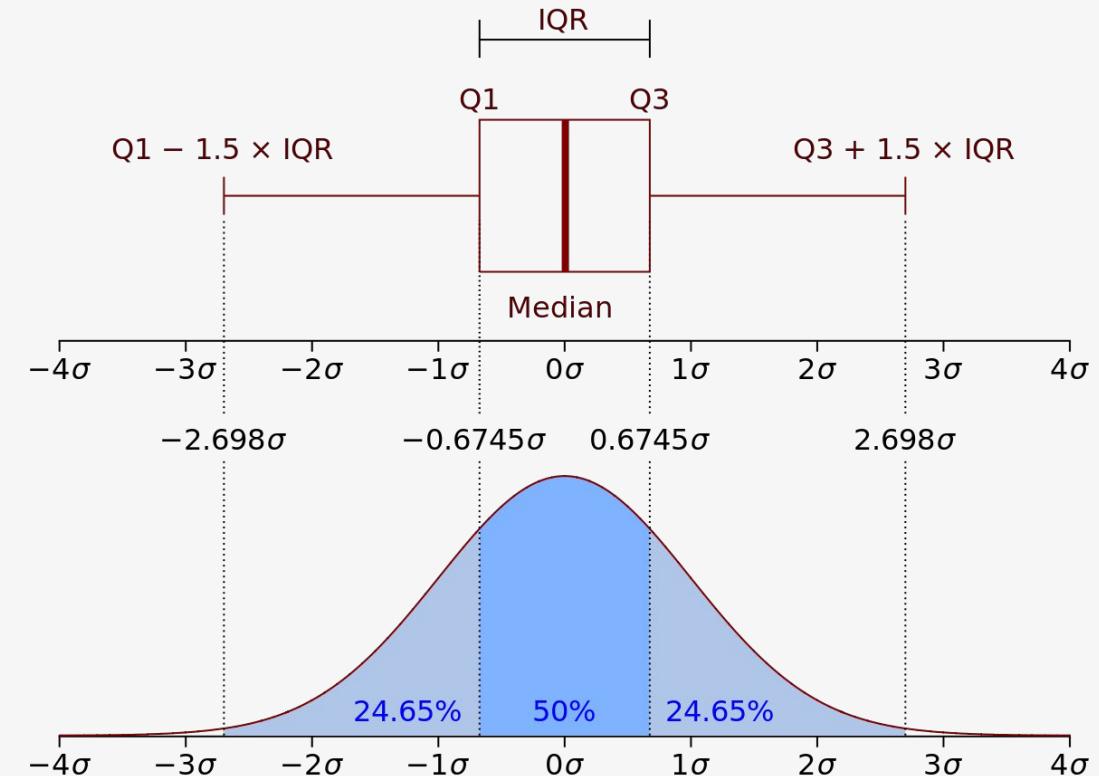
$P(x<0.5)=?$

$P(x \leq 0.5)= ?$



# 箱线图

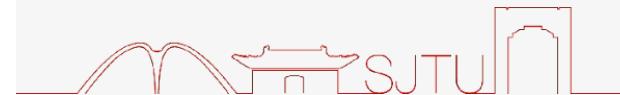
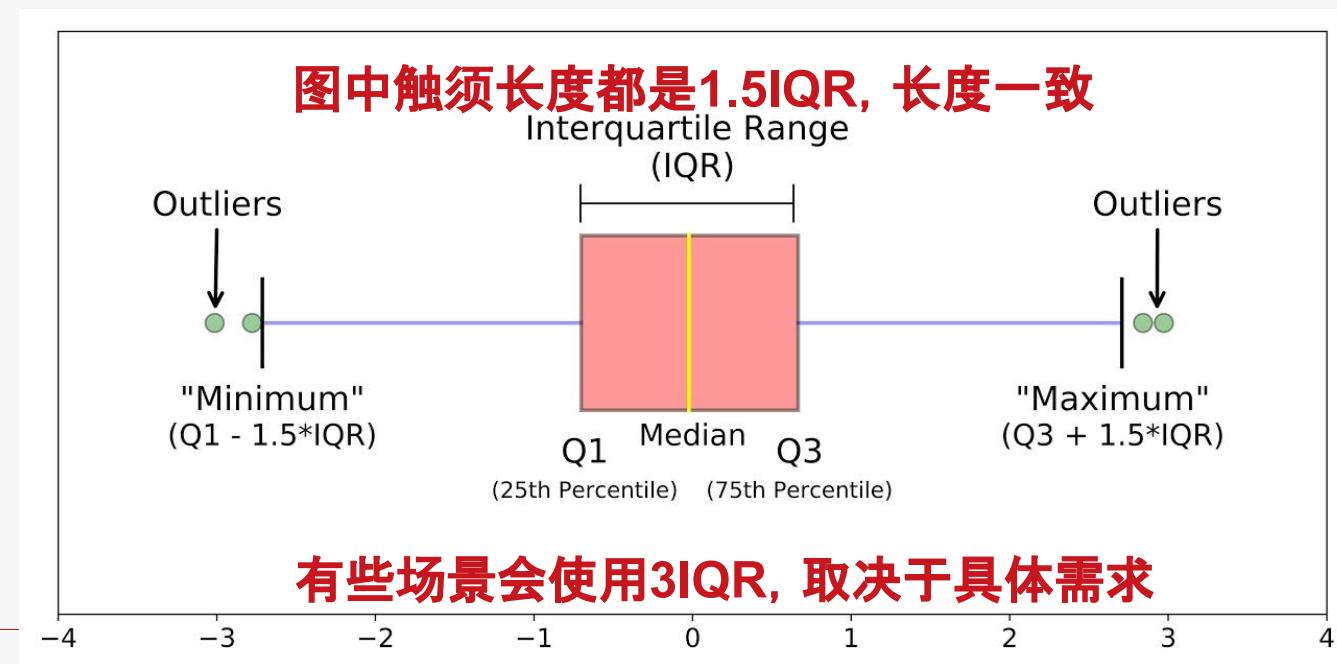
- 分位数Quantile
  - p分位数为t: 有比例p的数据小于t。例如50%分位数即为中位数
- 25%、50%、75%分位数
  - 一般也记为Q1、Q2、Q3
  - 最小值和最大值记为Q0、Q4
  - 正态分布的Q1在 $\mu - 0.6745\sigma$ 处
- IQR (Interquartile range, 四分位距)
  - $Q3 - Q1$
  - 一般认为小于 $Q1 - 1.5 \times IQR$ , 或是大于 $Q3 + 1.5 \times IQR$ 的为离群点(outlier)
  - 也有使用 $3 \times IQR$ 的情况



# 箱线图

- 箱线图boxplot

- 箱:由Q1、Q2、Q3画出;
- 线:右图中给出了 $Q1 - 1.5\text{IQR}$ 和 $Q3 + 1.5\text{IQR}$ 的触须whiskers;
- 离群点:数据中超出触须范围的, 需要标注出来作为离群点;
- !当 $Q3 + 1.5\text{IQR}$ 大于数据最大值时, 右侧whisker定义为数据的最大值, 此时右侧无outlier。
- 实际上右whisker应定义为 $\min(Q4, Q3 + 1.5\text{IQR})$ , 左侧同理。



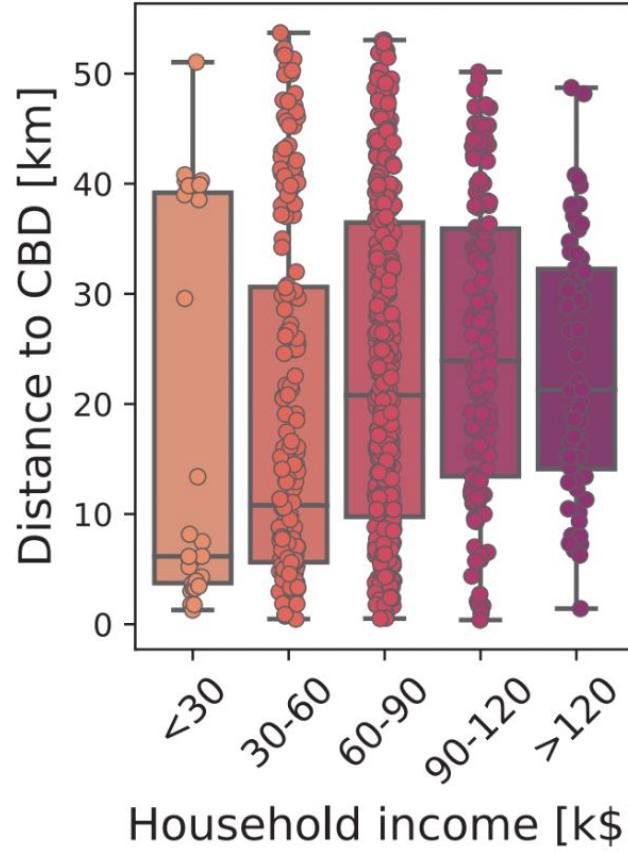
# 箱线图



上海交通大学  
SHANGHAI JIAO TONG UNIVERSITY

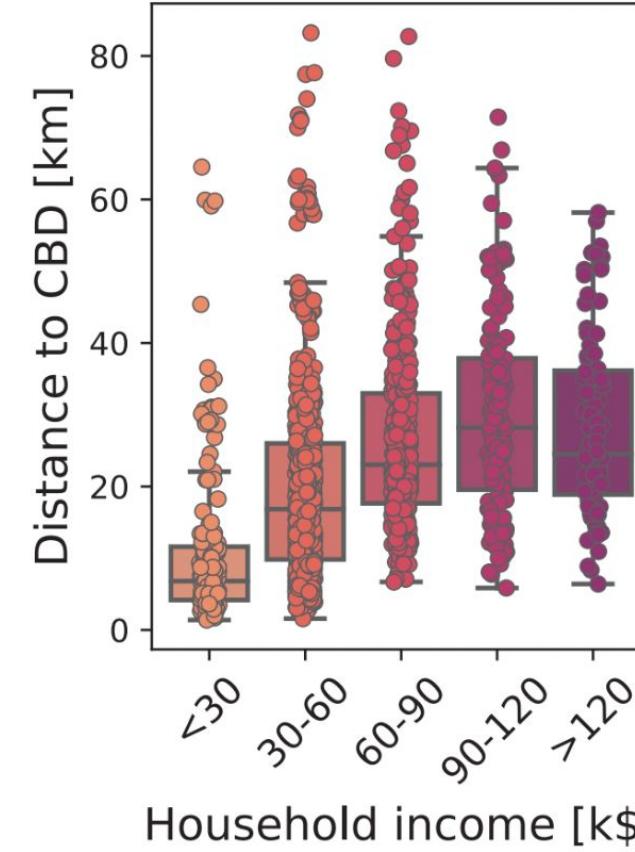
A

Boston



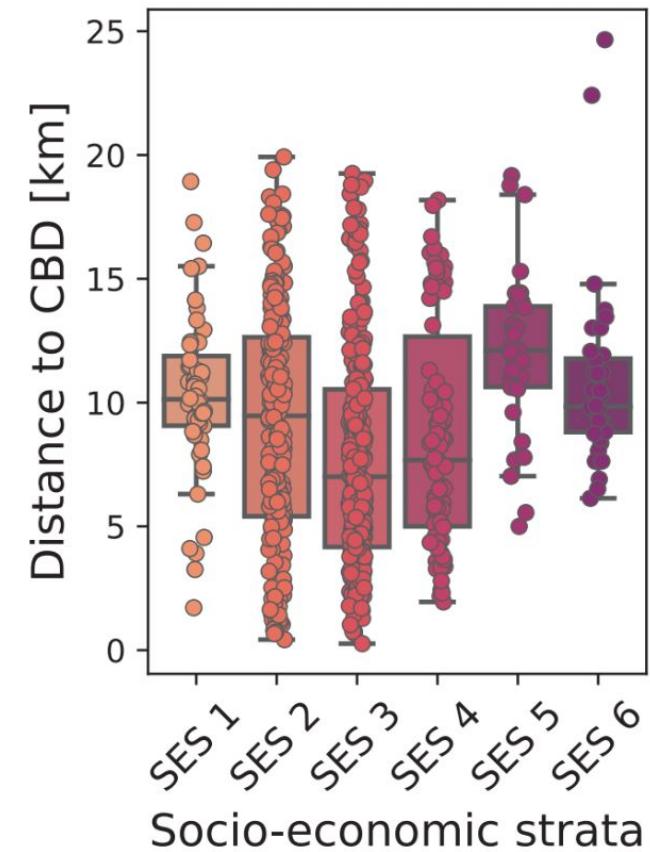
B

LA



C

Bogota



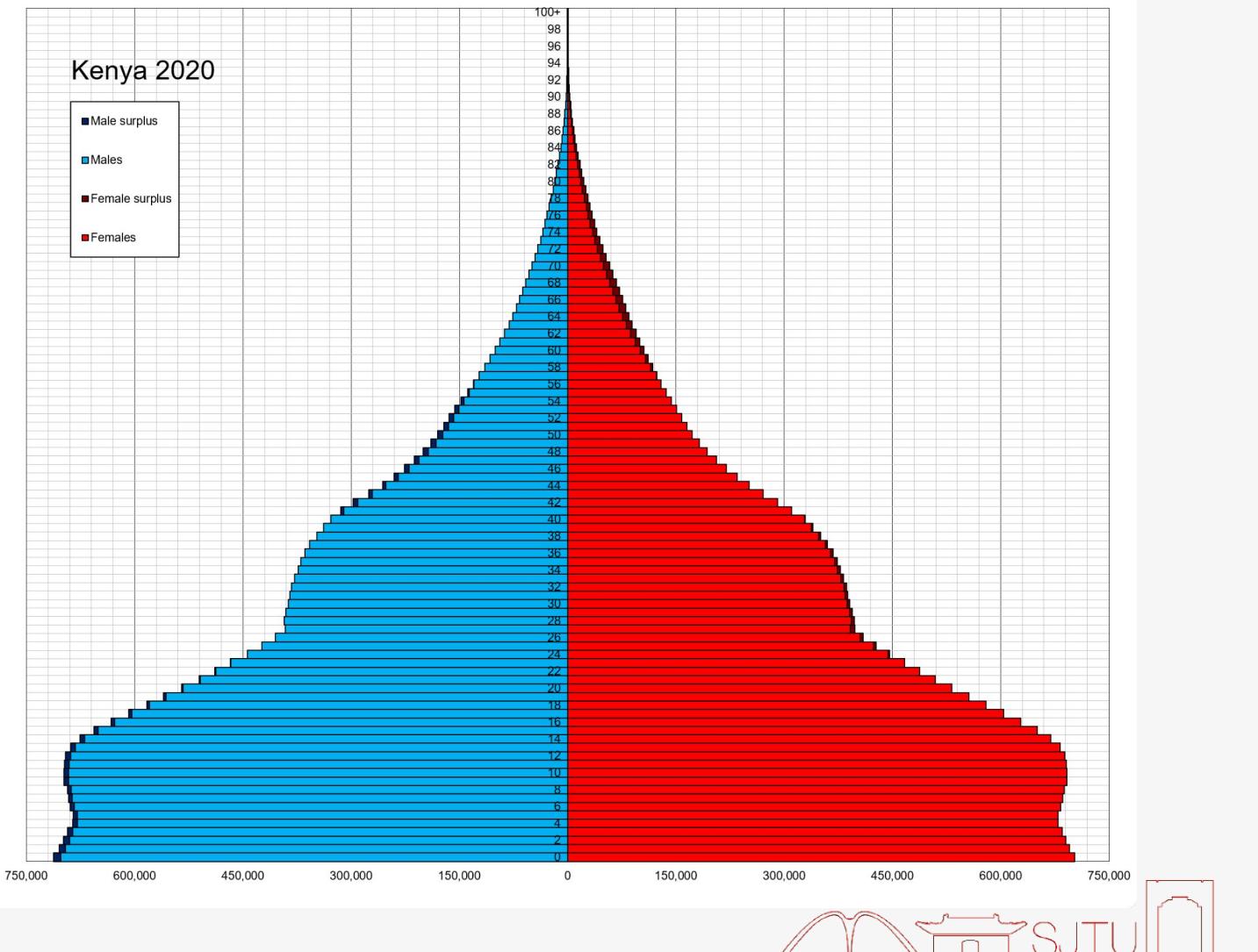
利用箱线图量化居住地与市中心距离 vs. 居民家庭收入之间的关系



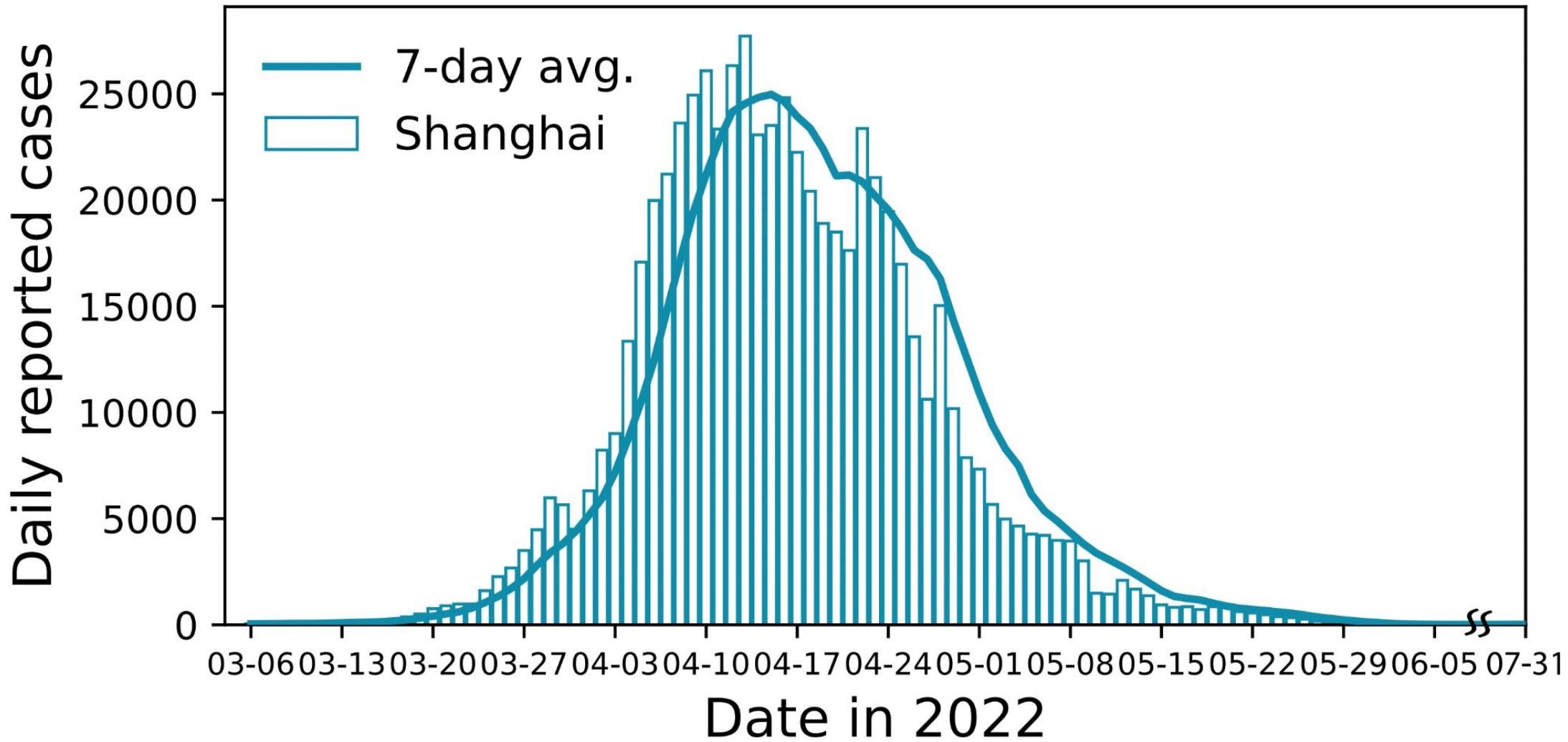
# 直方图Histogram和核密度估计KDE

Histogram和KDE都是密度估计  
(Density Estimation)的一种方  
法。

直方图(Histogram)是一种统计  
报告图, 统计每个区间内的频数

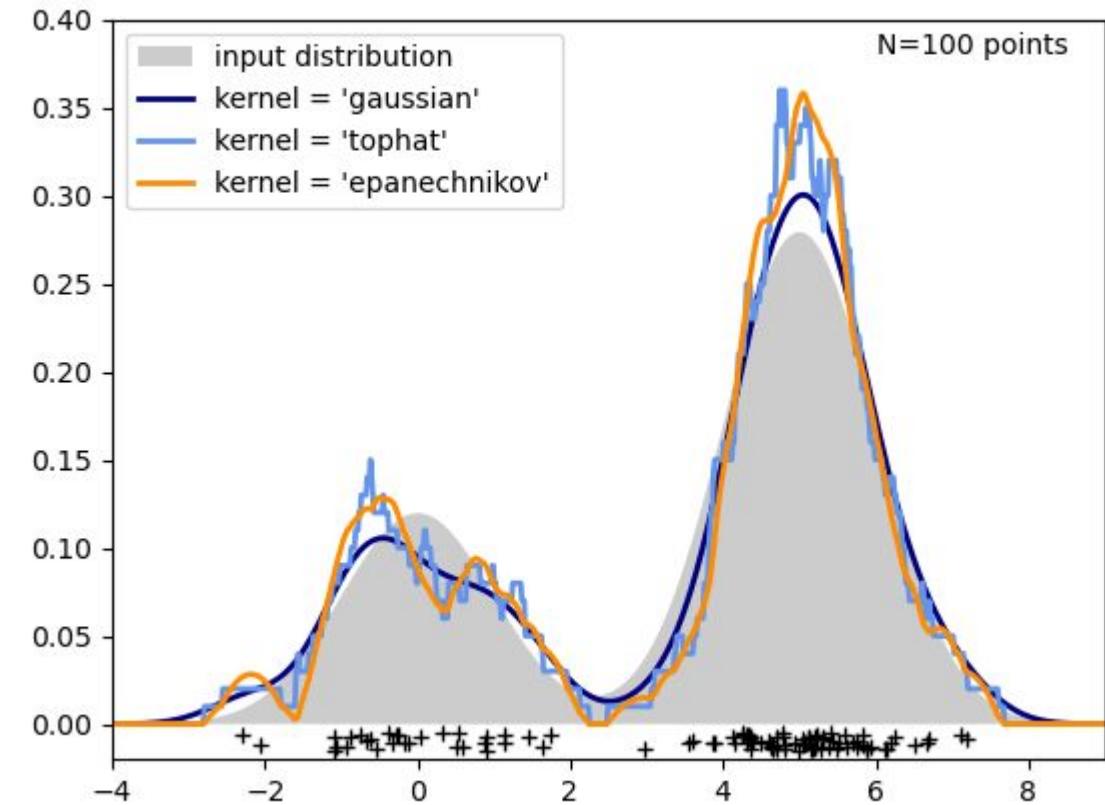
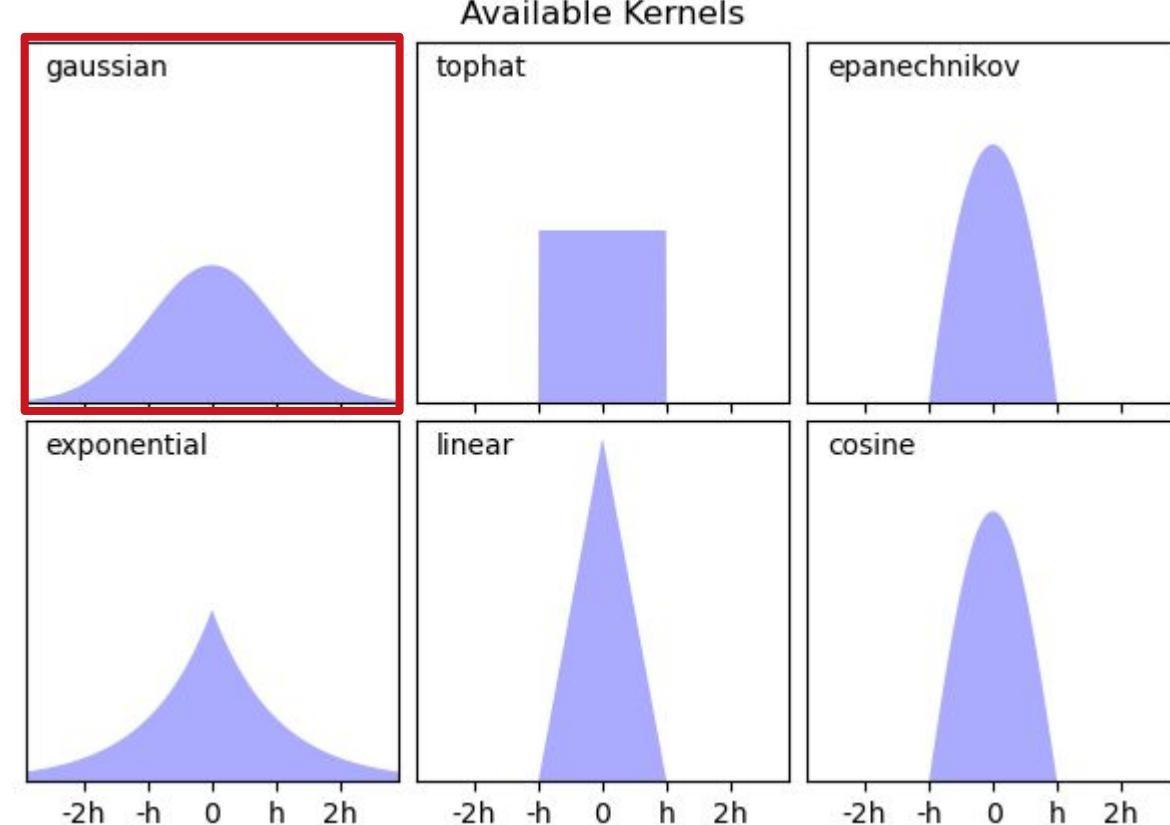


# 上海市COVID-19每日新增确诊数量



# 直方图Histogram和核密度估计KDE

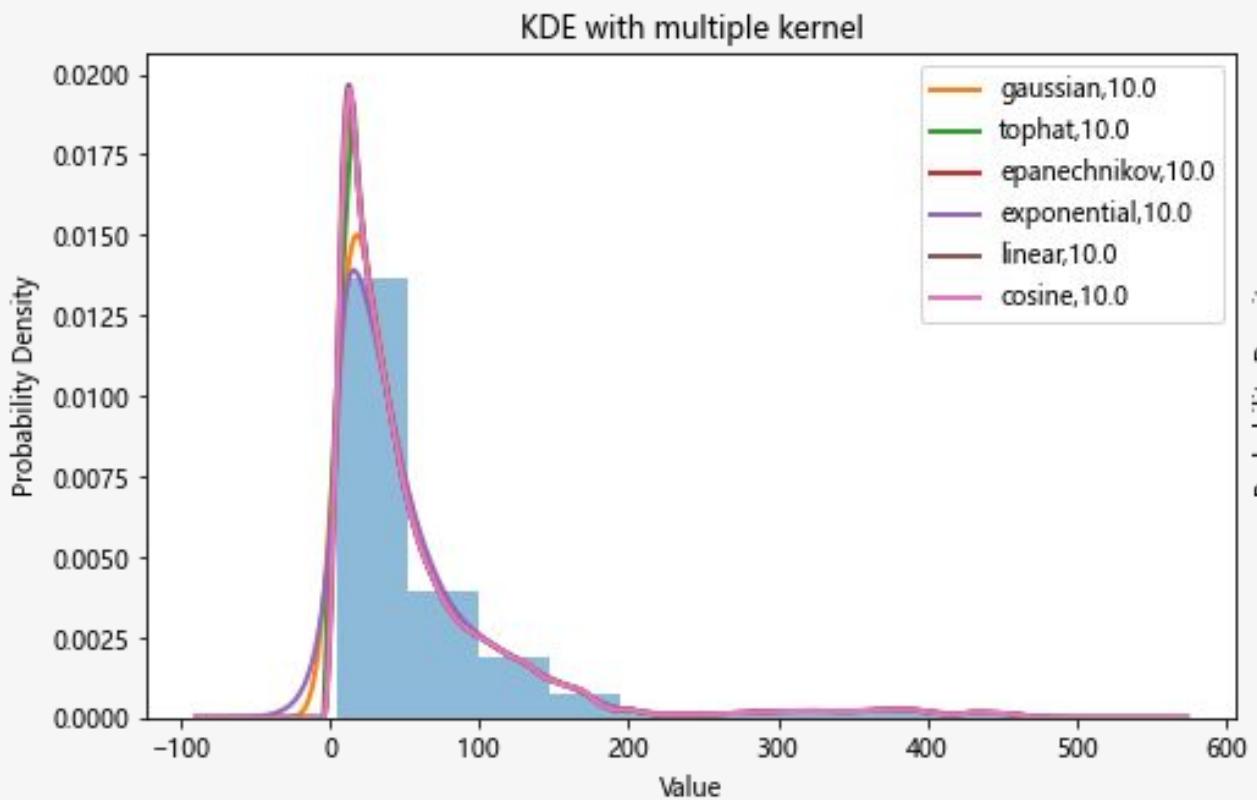
KDE采用不同的Kernel估计分布密度函数, 图示为一维情况



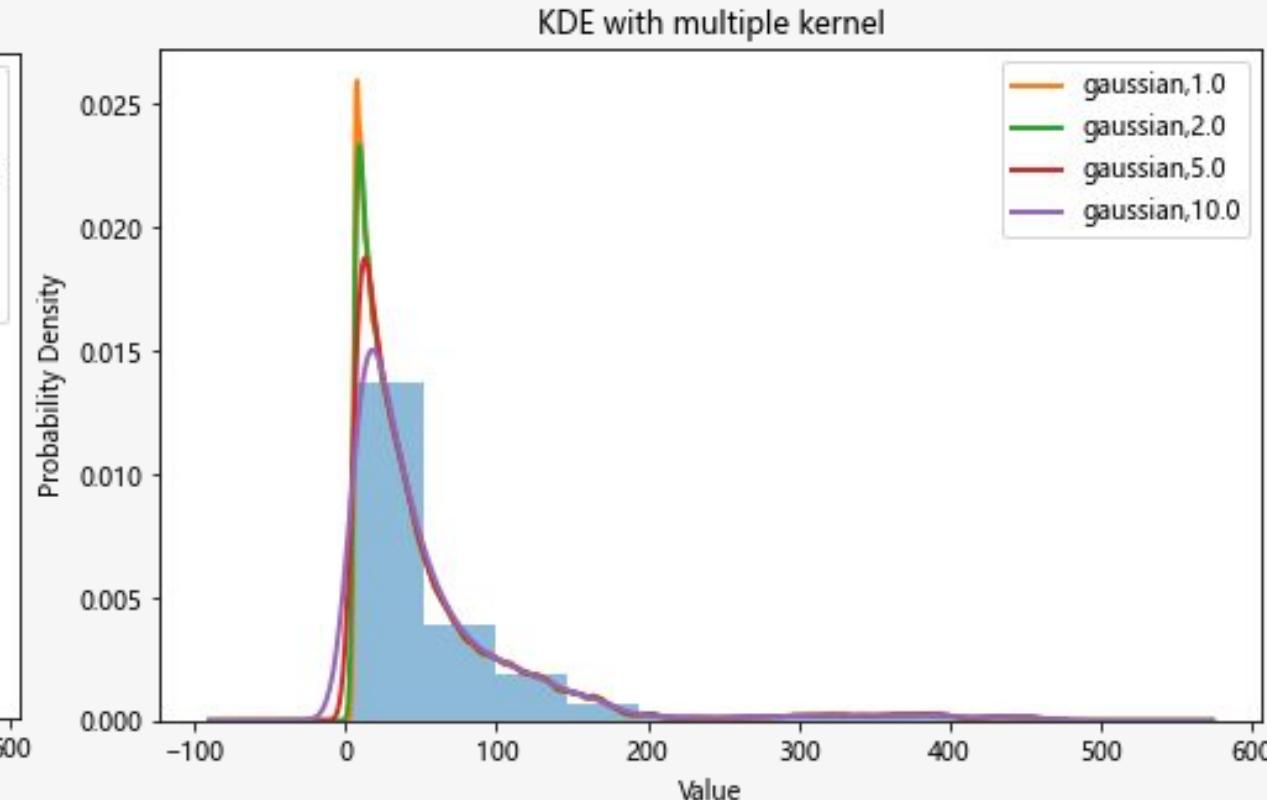
[Kernel Density Estimation \(mathisonian.github.io\)](https://mathisonian.github.io)链接详细解释了不同kernel的含义和作用效果, 简单来说, 估计某个中心点 $x$ 的密度函数时,  $x$ 周围点的**重要性**如何。

# 对同一样本集采用不同kernel和带宽进行密度估计

具体代码见课程Notebook, 数据为上海市百米网格人口密度



不同kernel拟合出的概率密度函数

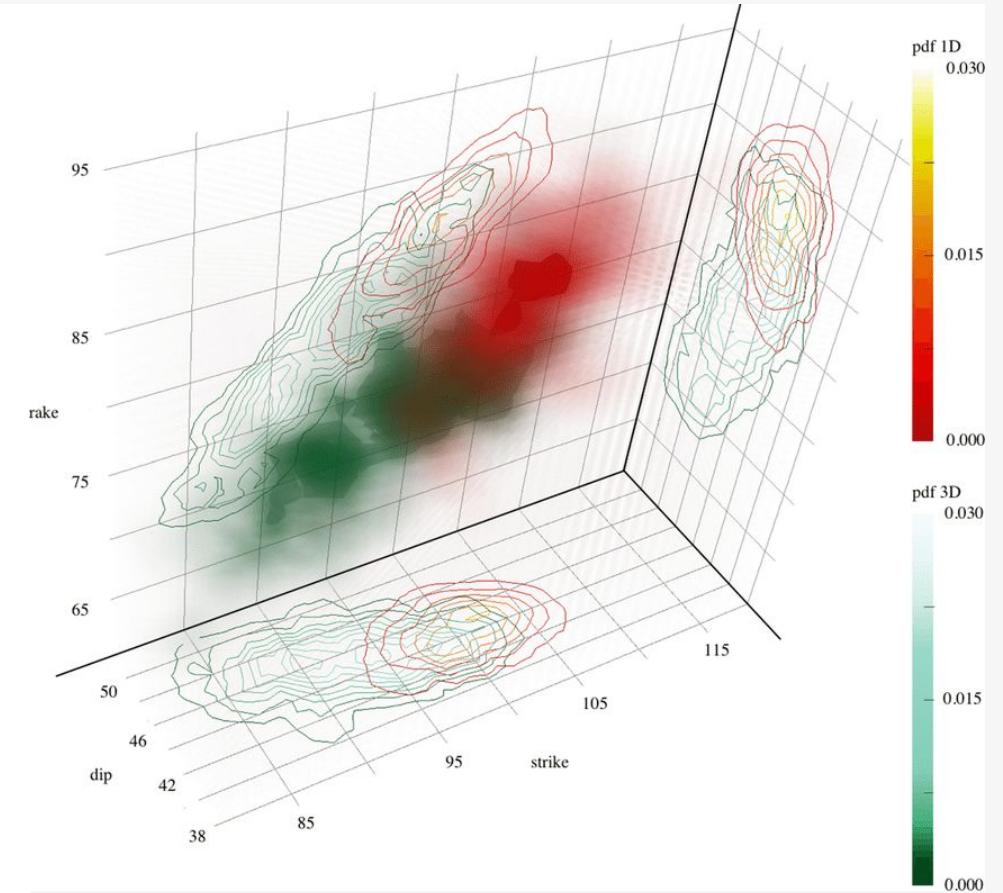
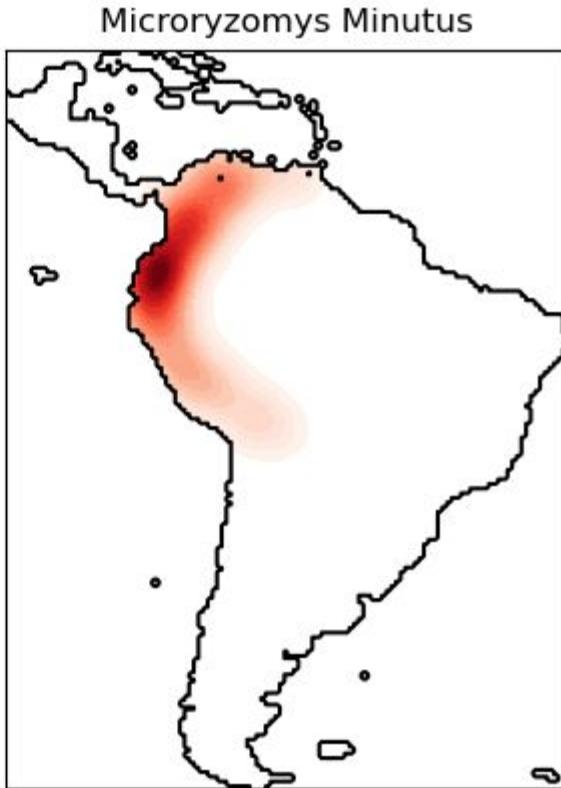
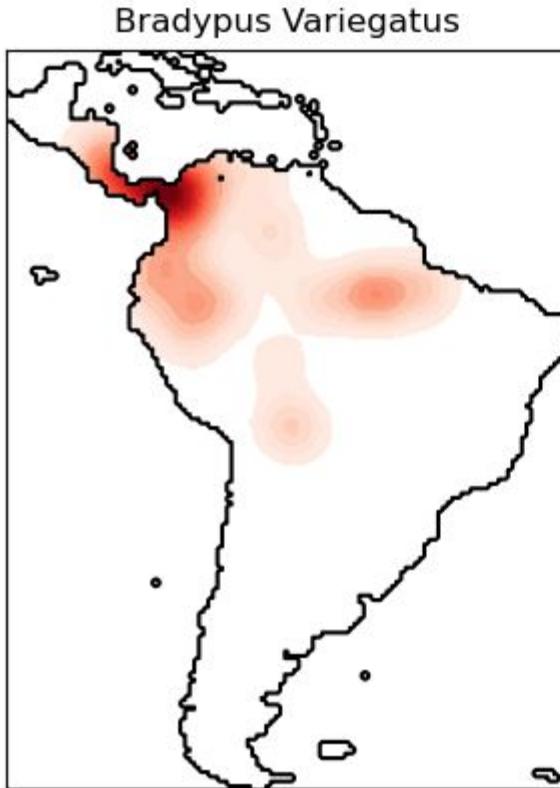


不同宽度的高斯函数拟合出的概率密度函数



# Histogram和KDE

## 高维数据分布的KDE估计



# 伯努利分布



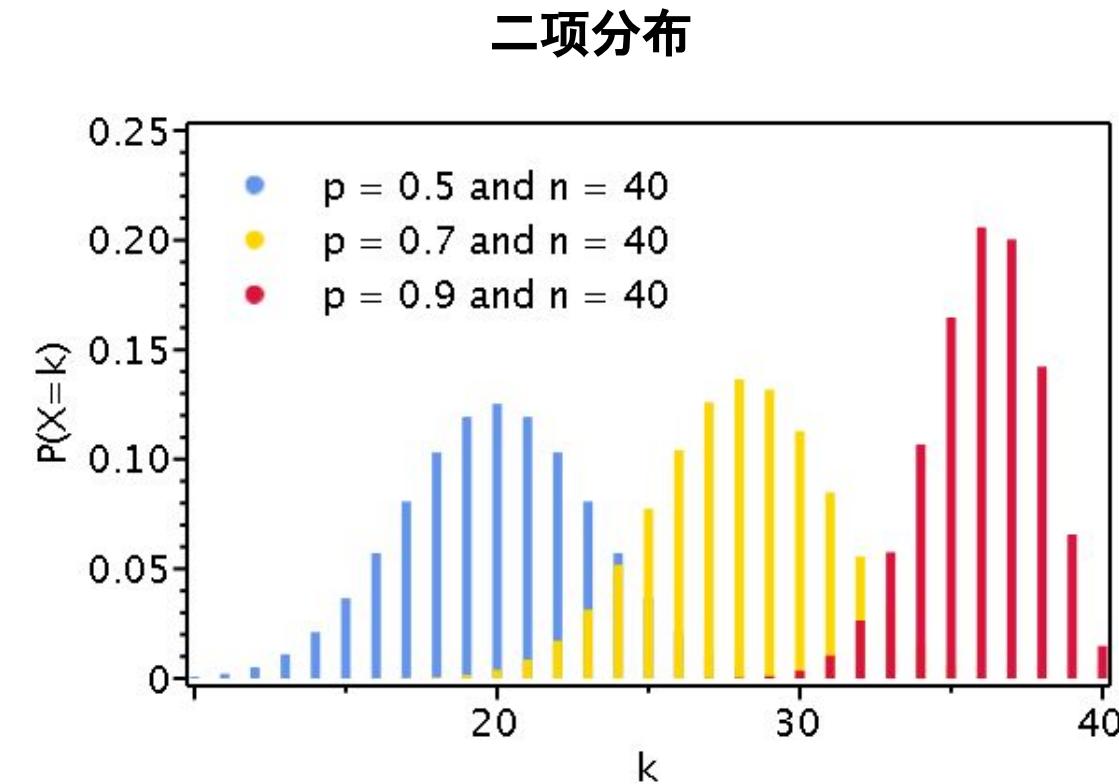
离散分布：

伯努利分布：

X	0	1
P	$1 - p$	$p$

n次伯努利实验的成功次数-二项分布：

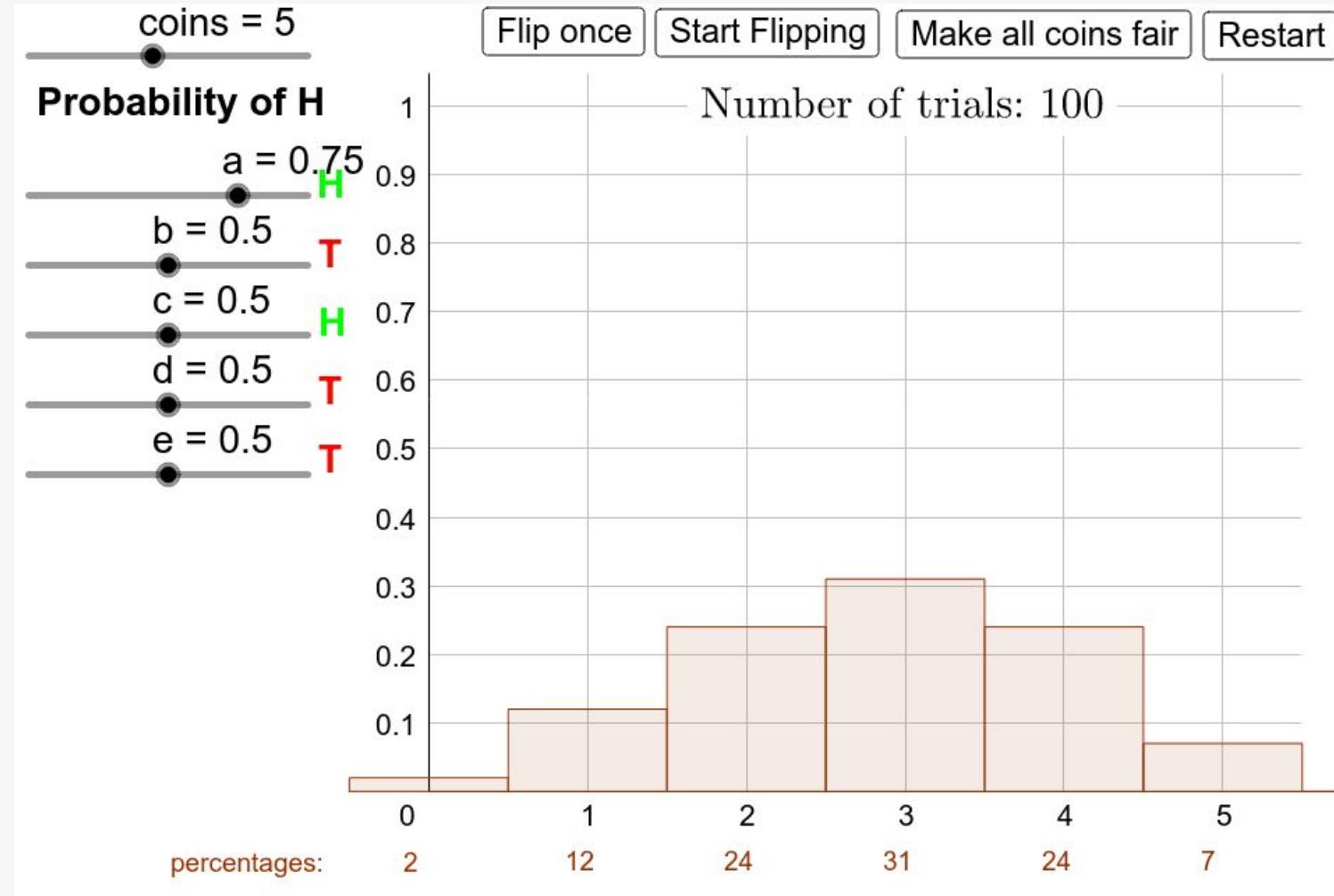
$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n.$$



# 抛硬币实验



<https://www.geogebra.org/m/m6zkdqtw>



# 泊松分布

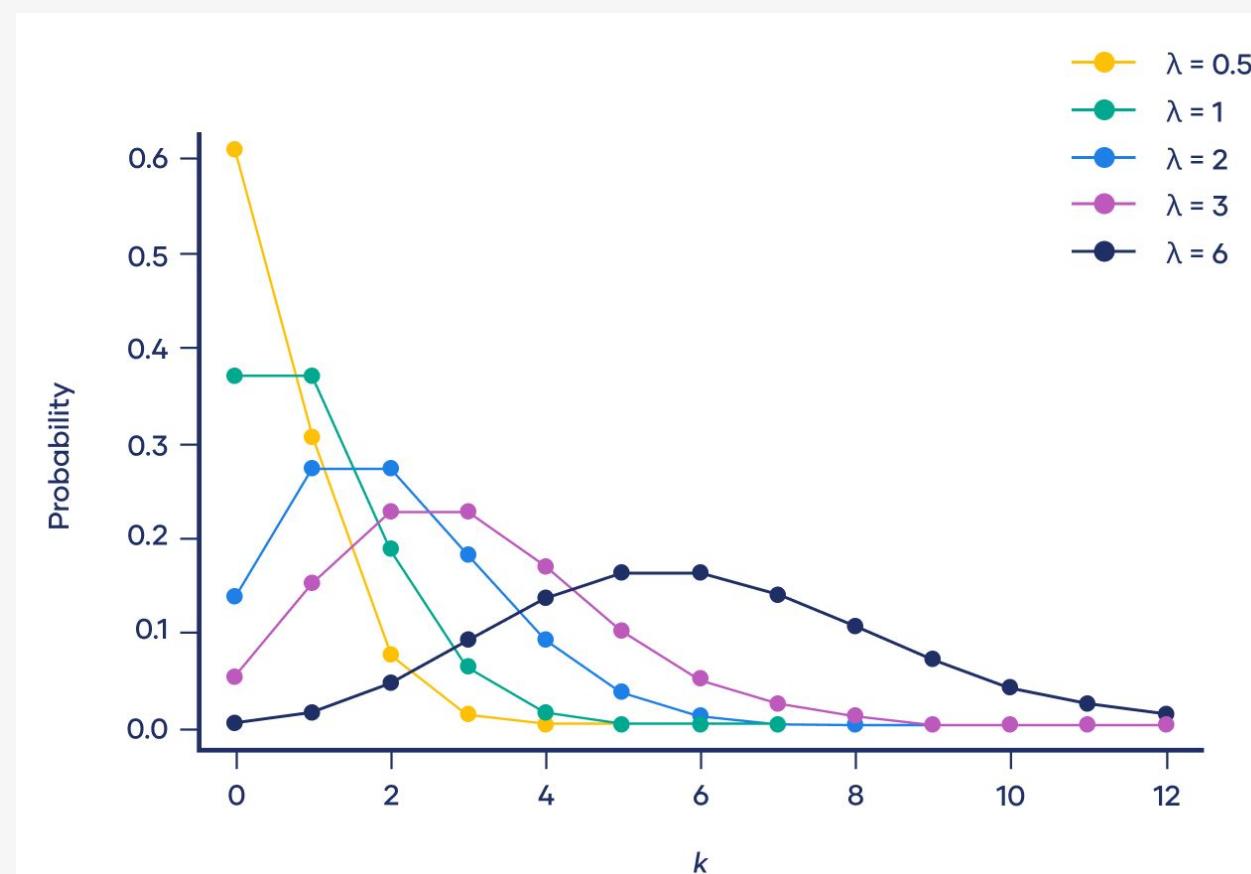
在伯努利实验中n极大, p极小, np有限, 一段时间内实验成功的次数符合泊松分布。

泊松分布：

参数 $\lambda=np$ , 期望 $=\lambda$

例如小明的店铺平均一天有100人光顾, 他建模的时候就可以假设每天的顾客数量符合 $\lambda=100$ 的泊松分布

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

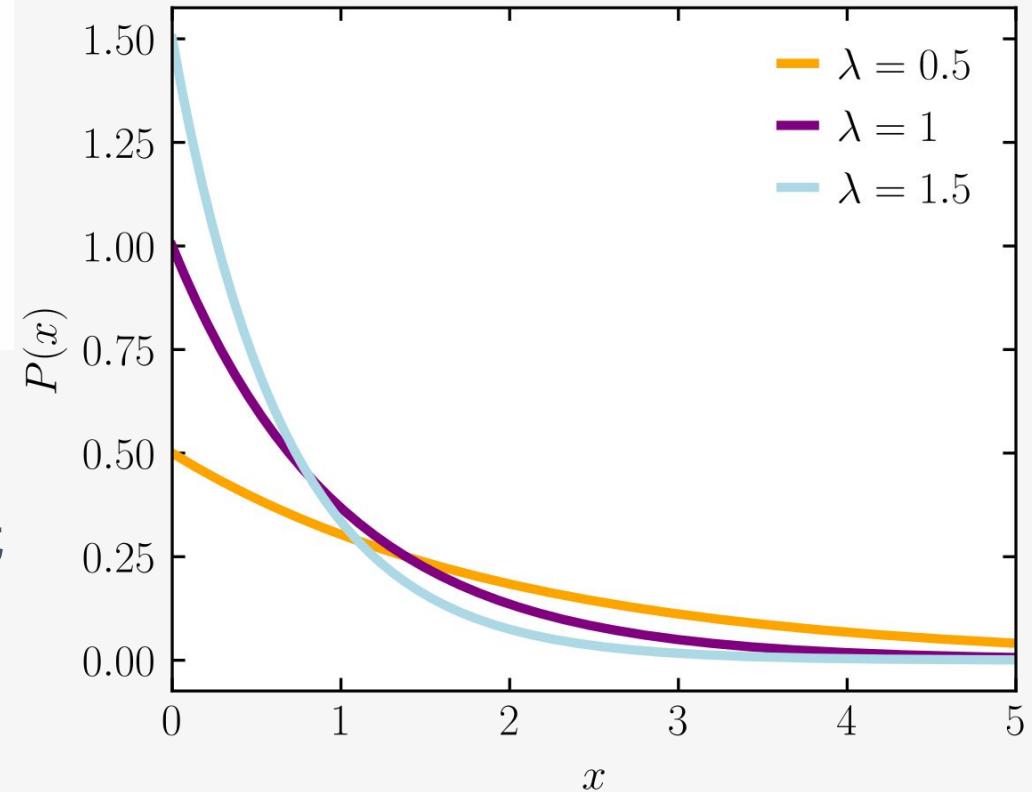


# 指数分布

指数分布的PDF

$$f(x) = \begin{cases} \lambda e^{-\lambda * x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

指数分布常被用作各种“寿命”分布，譬如电子元器件的寿命、电话的通话时间、随机服务系统中的服务时间等都可假定服从指数分布。



# 指数分布

- 指数分布的PDF
- 通过积分计算出指数分布的CDF表达式

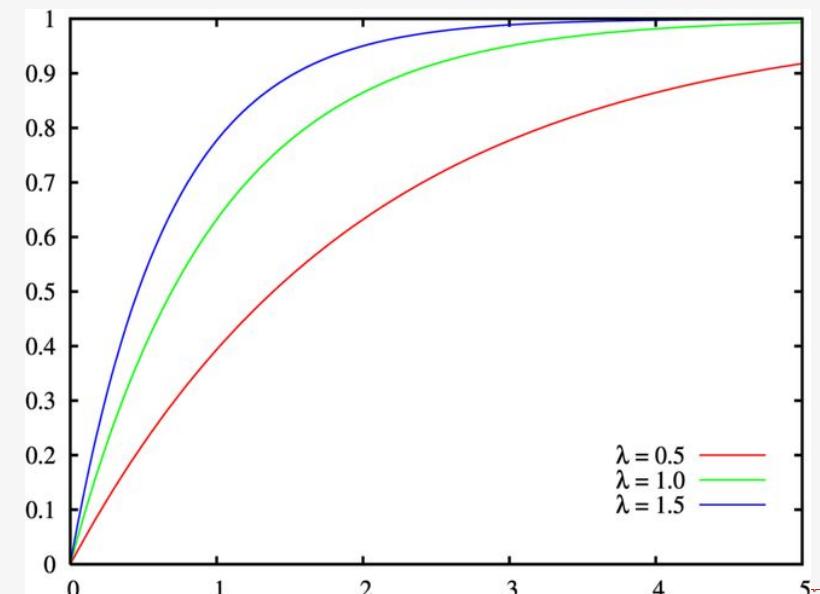
$$F(x) = 1 - e^{-\lambda x}$$

- 指数分布的期望为 $1/\lambda$ , 方差为 $1/\lambda^2$

PDF

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

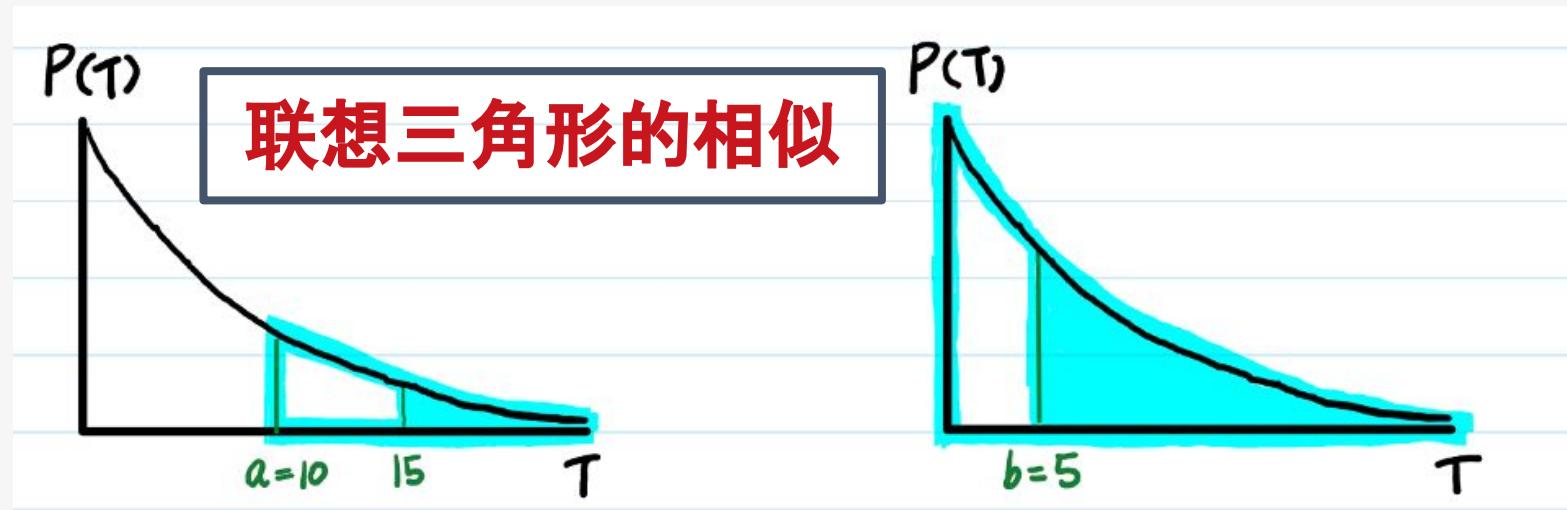
CDF图像



# 指数分布



特点：无记忆性，简单理解为从现在开始 $t$ 秒灯泡没坏掉的概率 $P(X>t)$ 和已知过了 $s$ 秒后灯泡没坏，再过 $t$ 秒后灯泡还没坏掉的概率 $P(X>s+t|X>s)$ 相等。（条件概率）



$$P(T>a+b|T>a) = \frac{P(T>a+b)}{P(T>a)} = \frac{P(T>b)}{1}$$

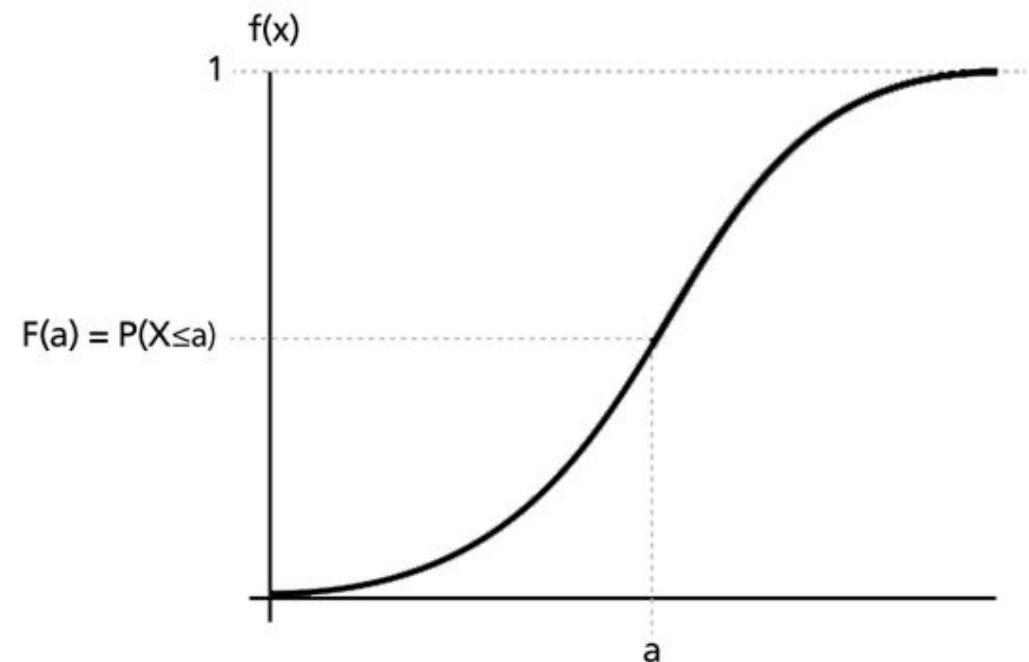
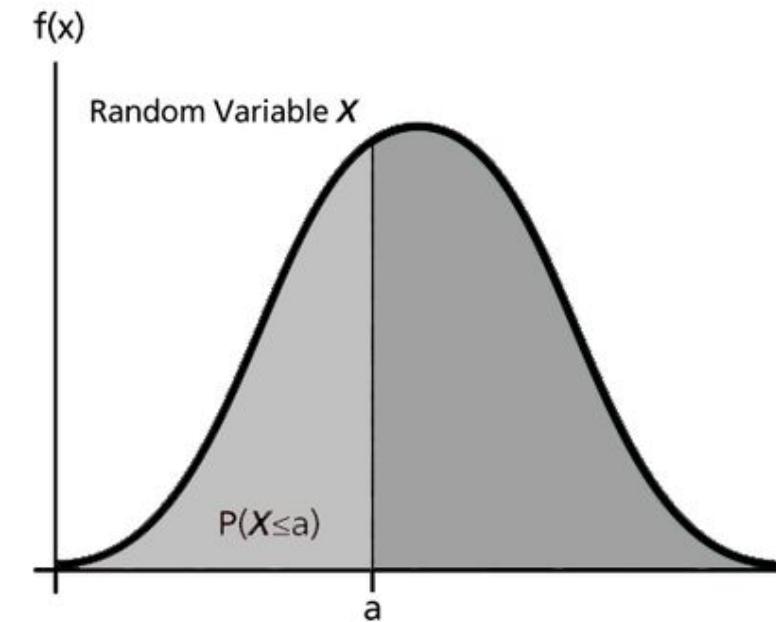
# 正态分布

## 一维正态分布的PDF

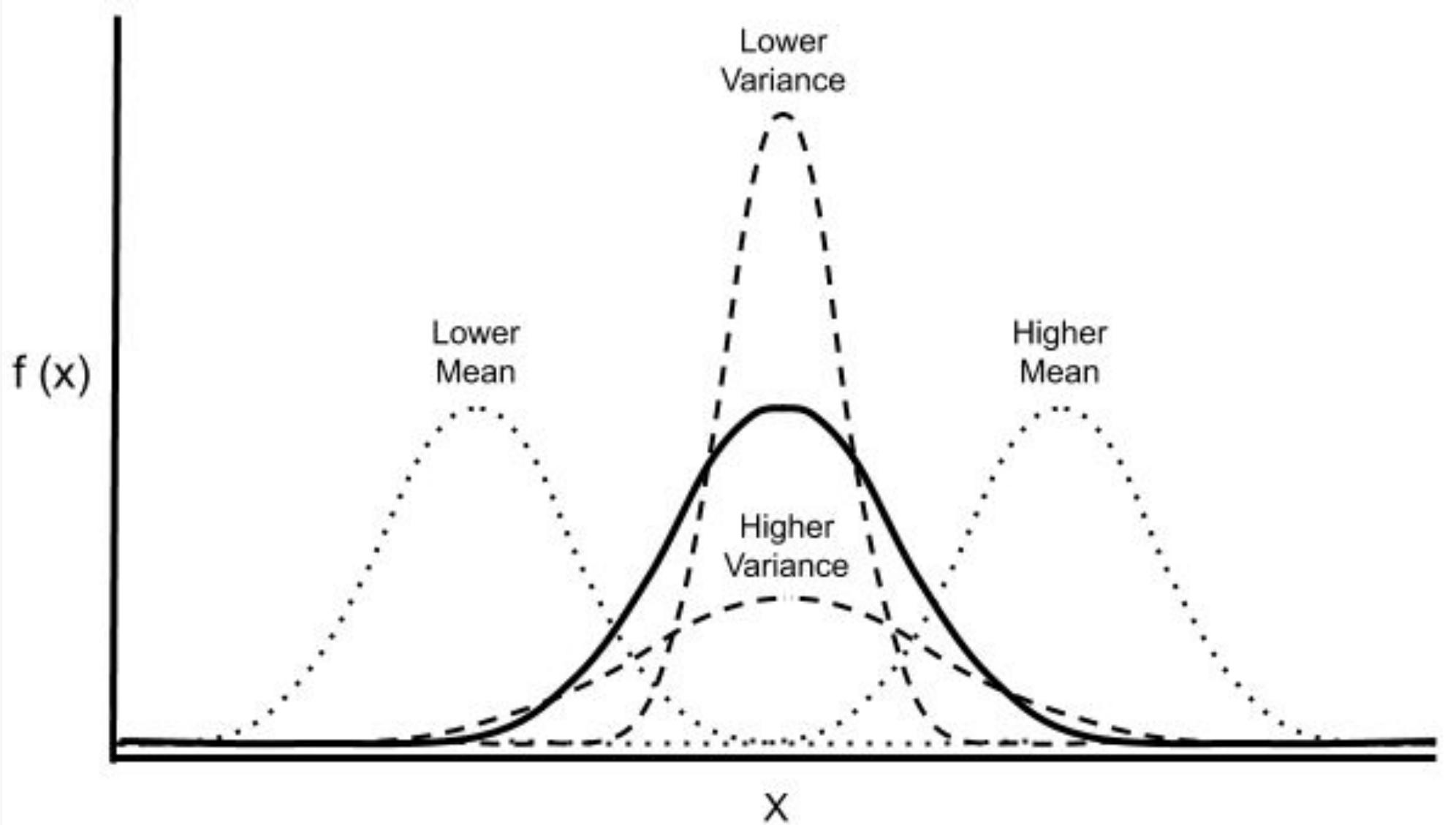
由两个参数 $\mu$ 和 $\sigma$ 来描述

对于正态分布, 期望 $=\mu$ , 标准差 $=\sigma$

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$



# 期望 方差

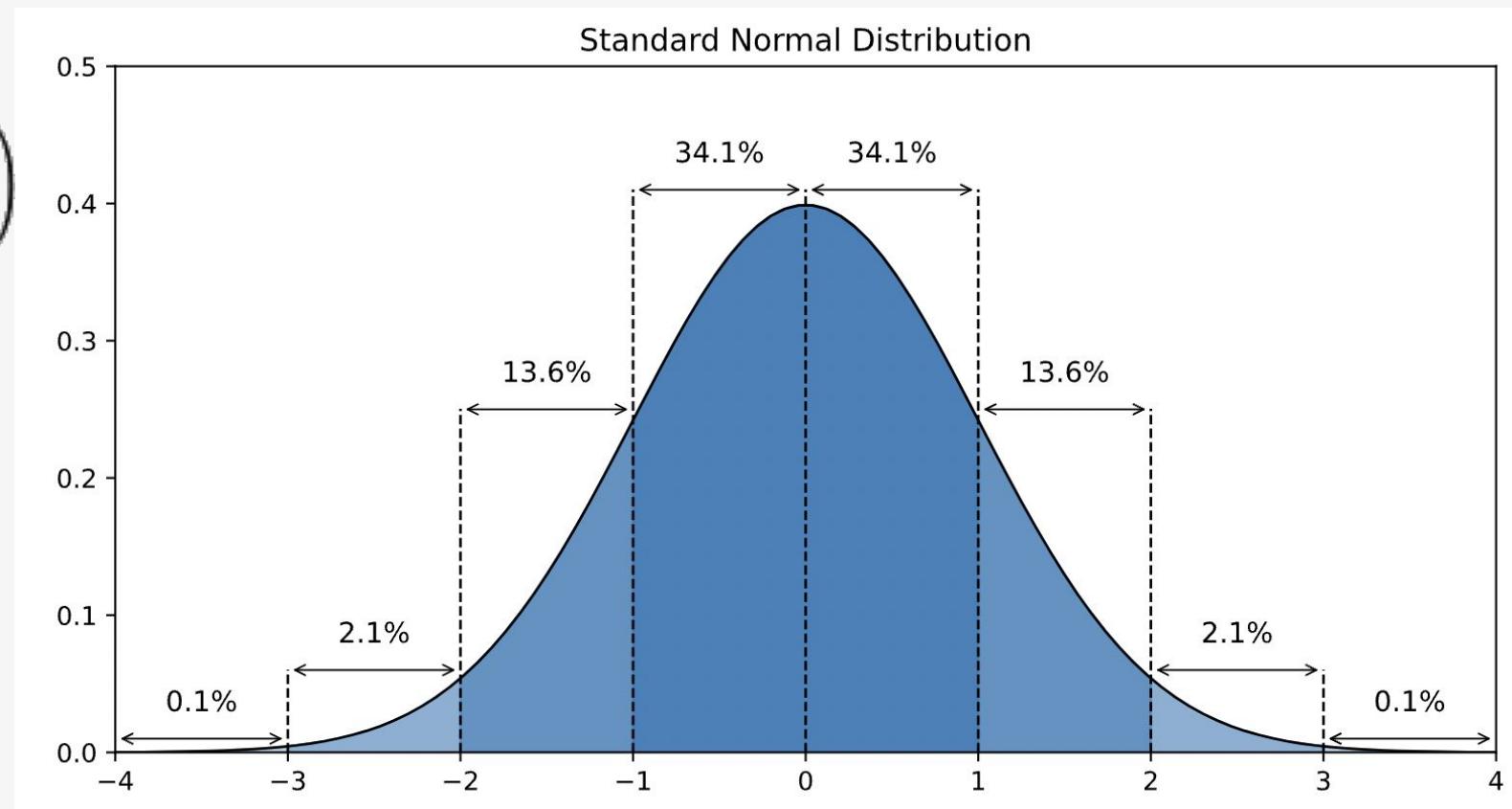


# 正态分布

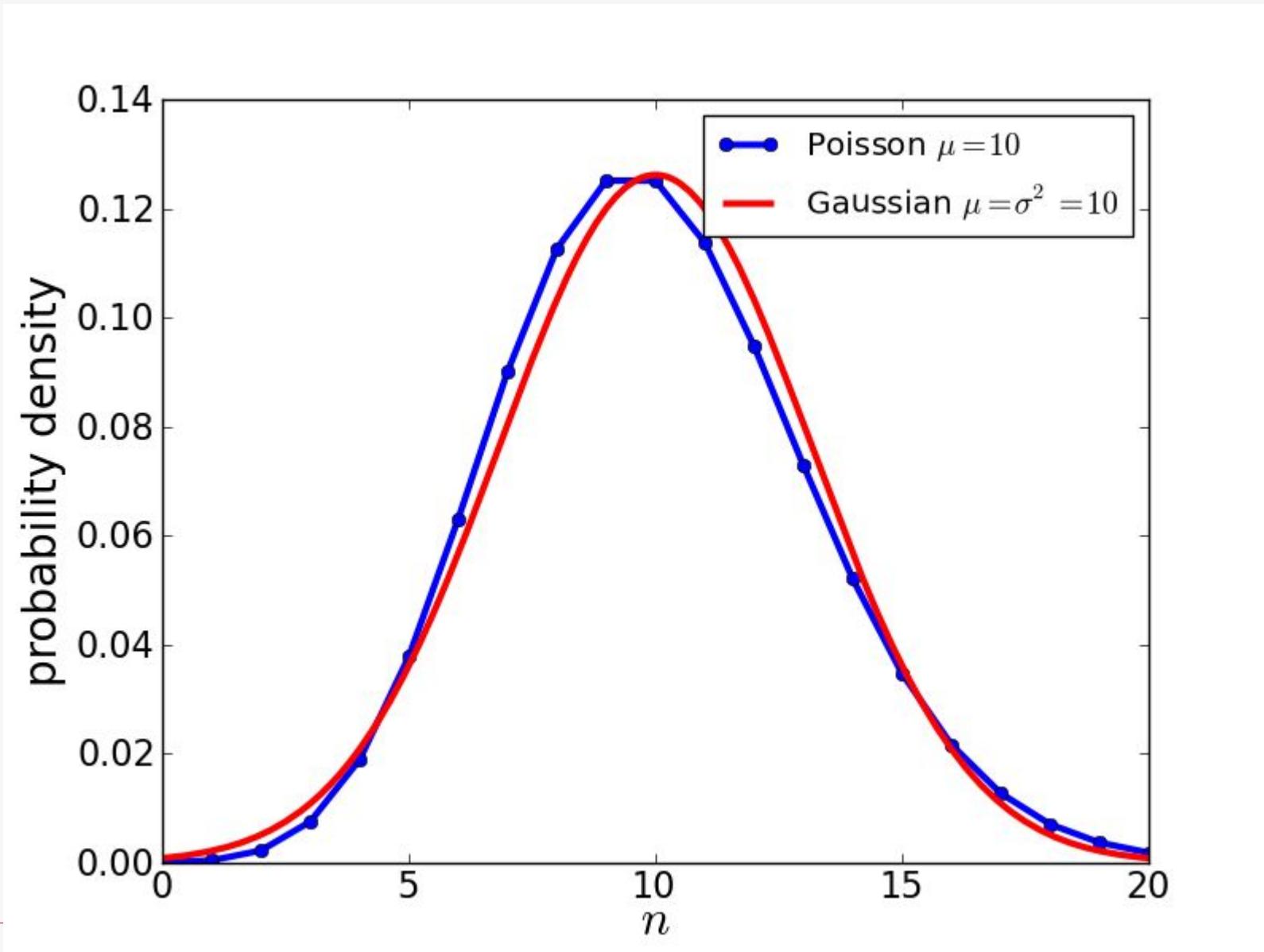
标准正态分布是均值为0, 方差为1的正态分布

任何正态分布减去均值, 除以方差后都可以变为标准正态分布

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

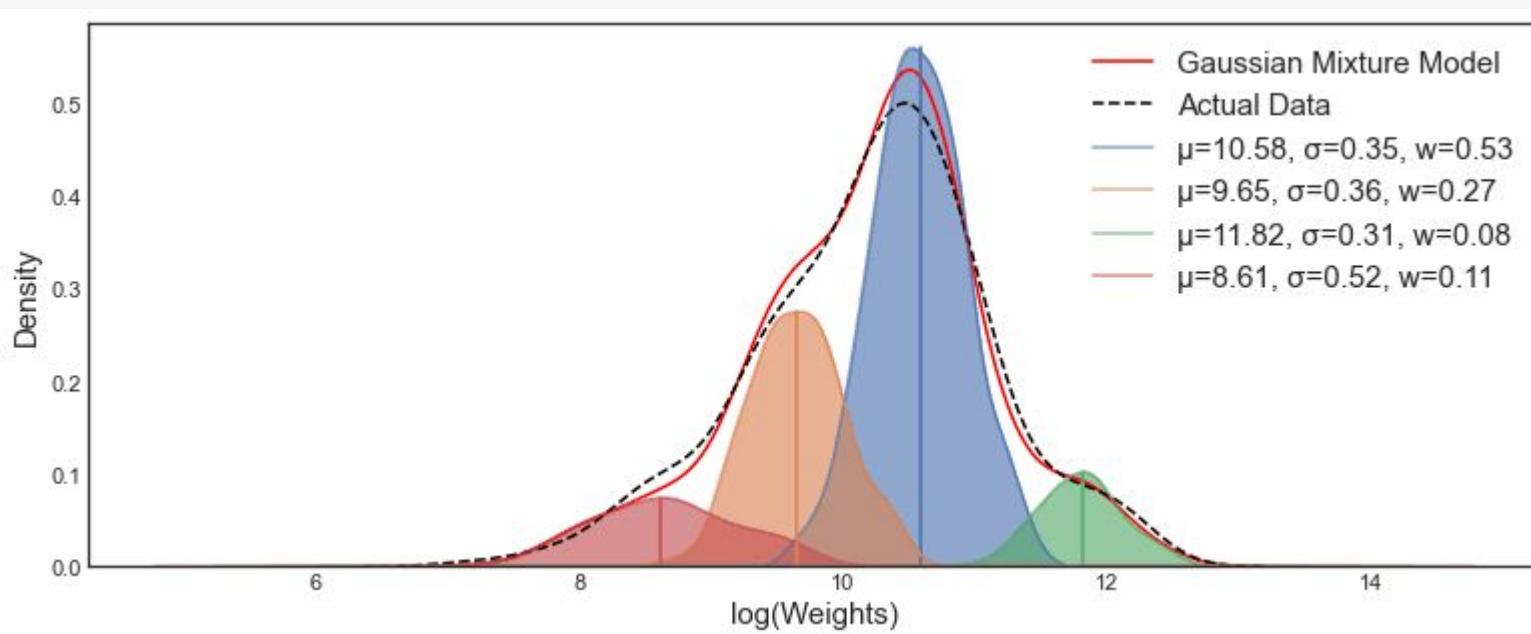


# 正态分布 vs 泊松分布



# 高斯混合模型(GMM)

- 高斯混合模型可以视为 $K$ 个高斯分布的加权和。
- 从生成数据的角度来看，也可以认为是从 $K$ 个高斯模型中按照概率 $a_k$ 先从 $K$ 个高斯分布中选出一个，再根据这个高斯分布生成一个样本。

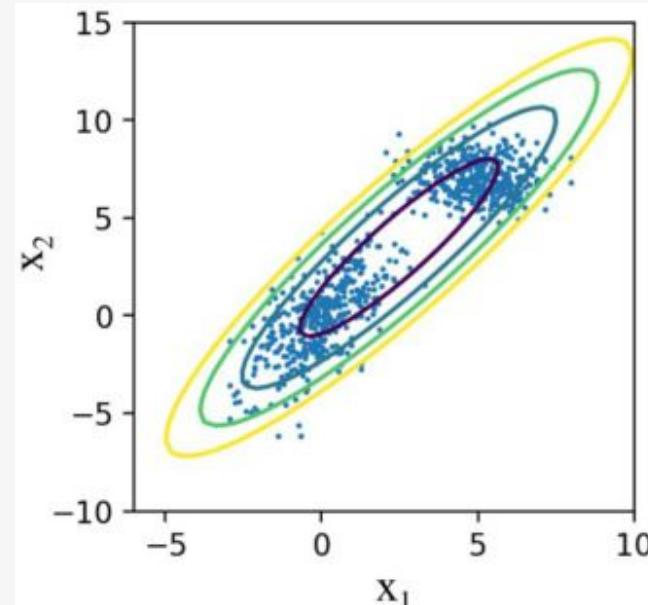
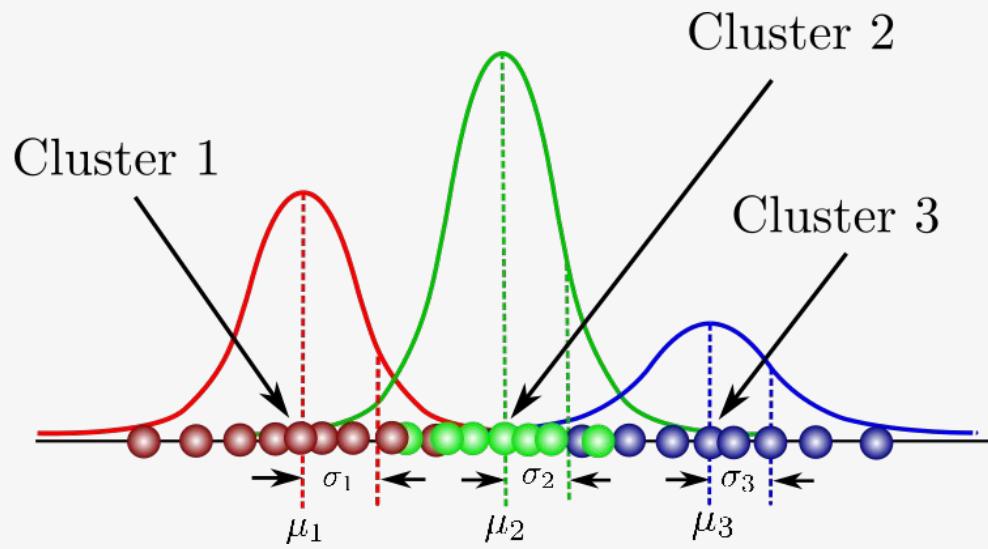


$$\sum_{k=1}^K a_k \mathcal{N}(\mu_k, \sigma_k)$$

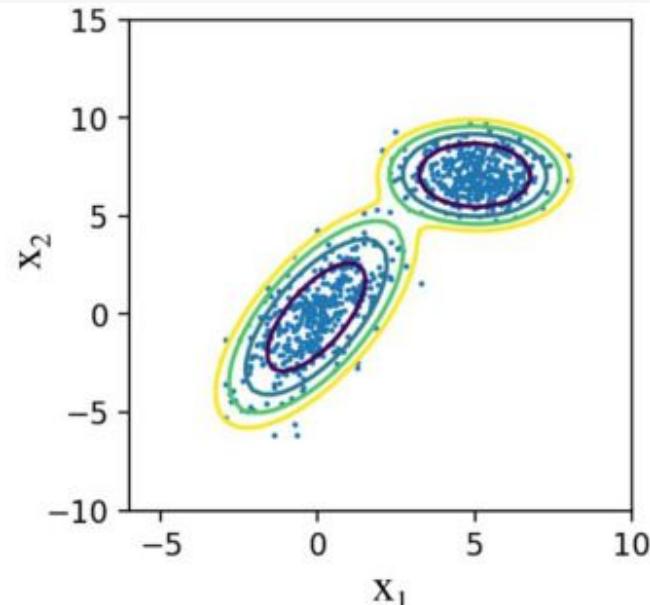
$$\sum_{k=1}^K a_k = 1$$

# 高斯混合模型(GMM)

- 高斯混合模型可以用于聚类任务，本质上是估计出生成所给样本的K个高斯分布，以及参数 $a_k$ 。高斯混合模型可以给出每个样本属于每个类的概率，也被称为“软聚类”。



(a) Single Gaussian



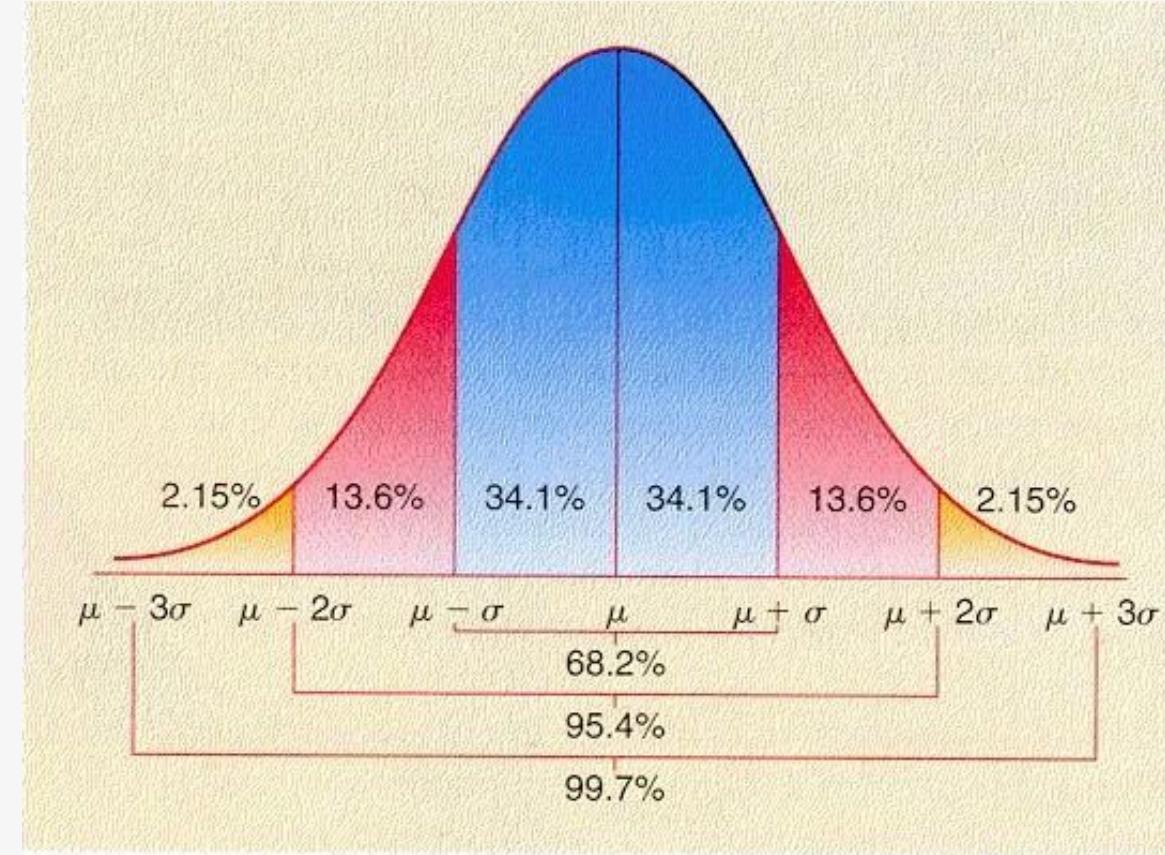
(b) Two Gaussians

# 置信区间

置信区间是什么？我虽然不能具体地得到某个变量 $X$ 是多少，但我可以给出某个区间 $[a,b]$ ，自信地说出：

我在某种程度上确定， $X$ 会落在 $[a,b]$ 之间！例如“ $X$ 的95置信区间为 $[a,b]$ ”表示 $X$ 有95%的概率在区间 $[a,b]$ 之间。

经过实验求得： $X$ 的期望为0.3，标准差为0.1，假设 $X$ 是一个正态分布，那么 $X$ 的95置信区间就是 $[0.1, 0.5]$ 。



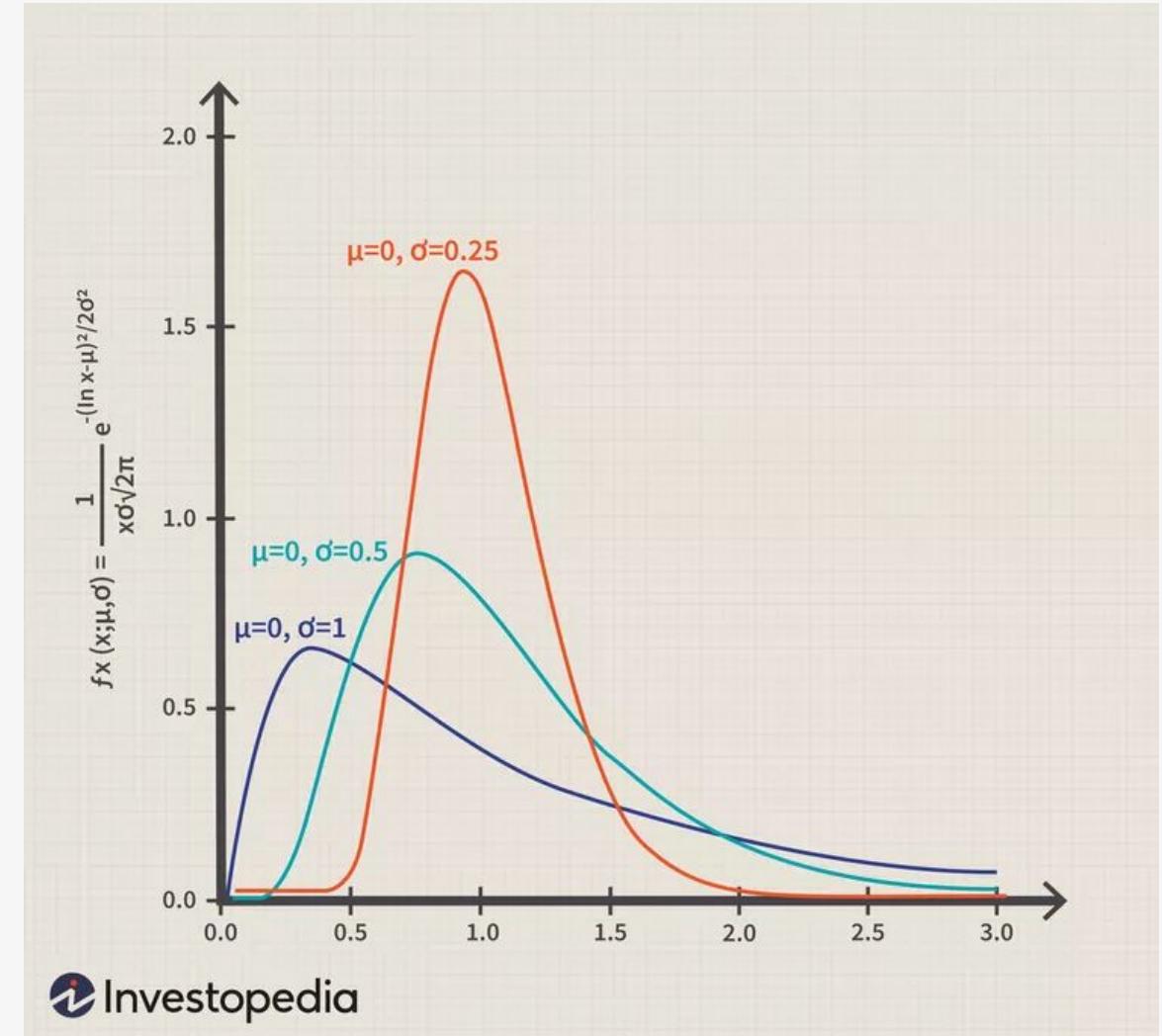
均值 $\pm 2\sigma$ 下的面积达到95%，均值 $\pm 3\sigma$ 下的面积达到99%，可以作为常识记住。还记得PDF曲线下面积的意义吗？

# Log-normal分布

常用于假设非对称的连续分布

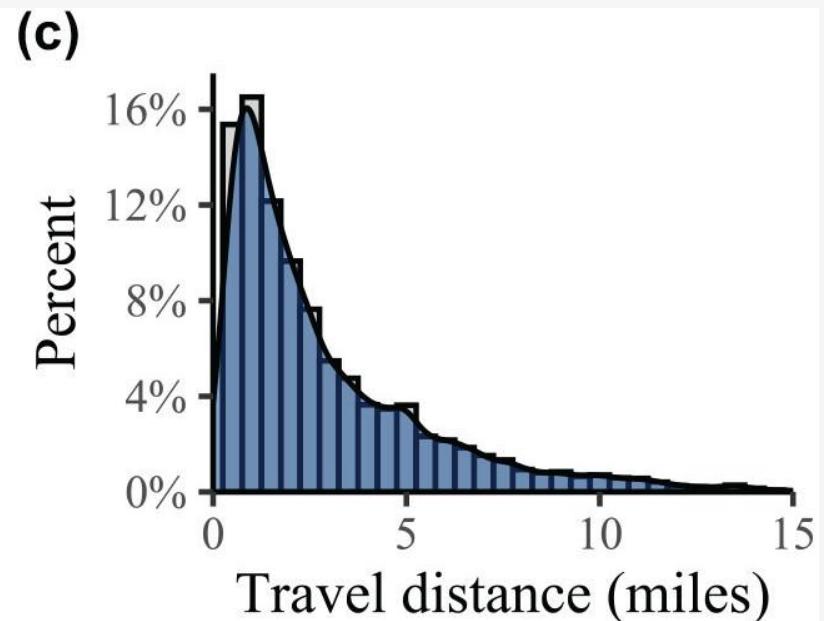
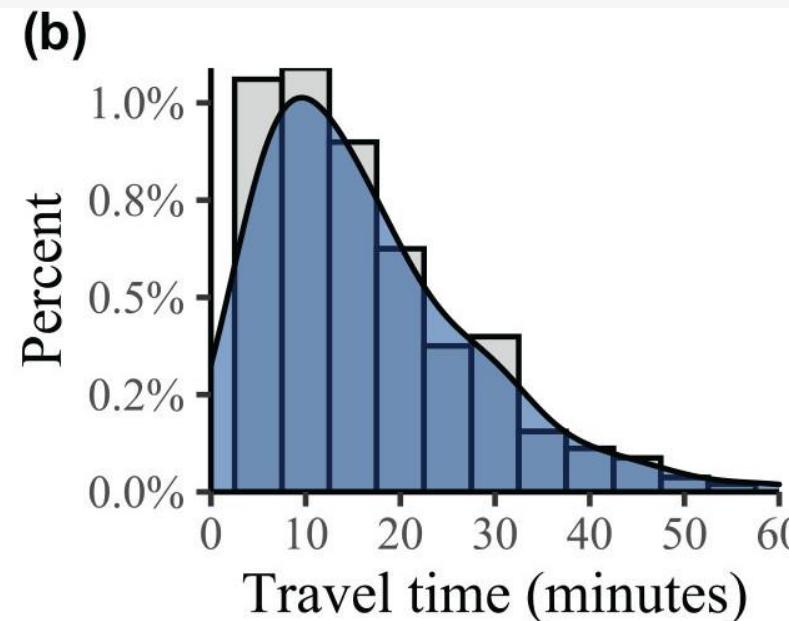
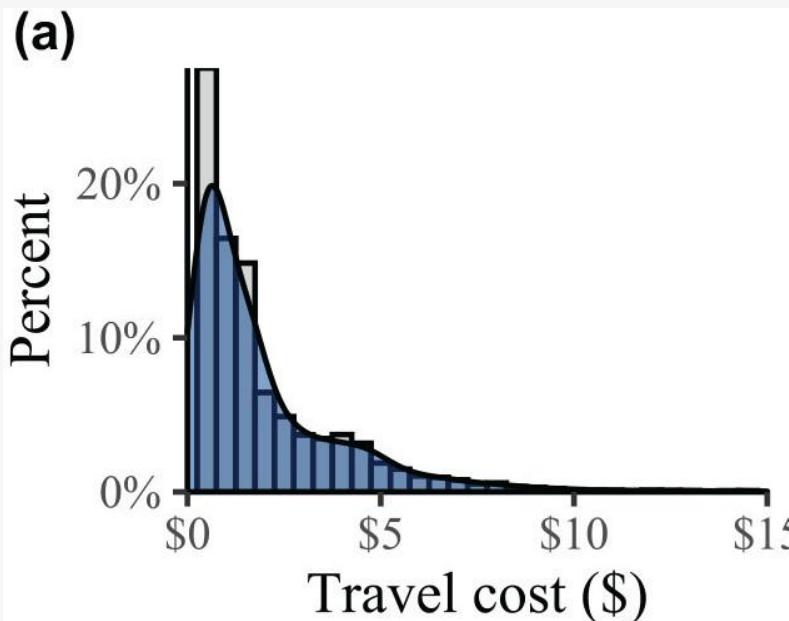
$$f_X(x) = \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}},$$

两个参数分别控制什么pattern？



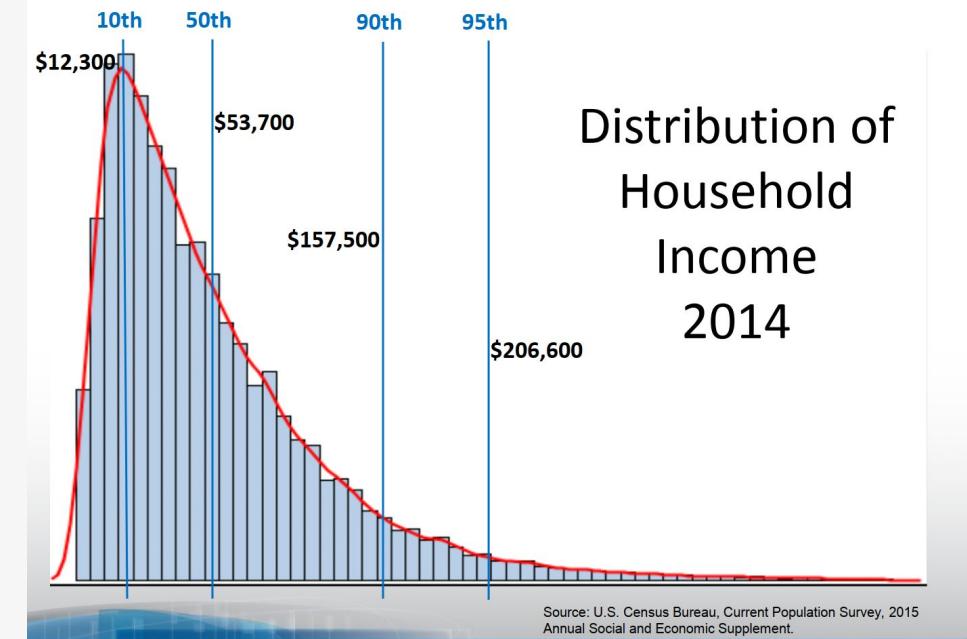
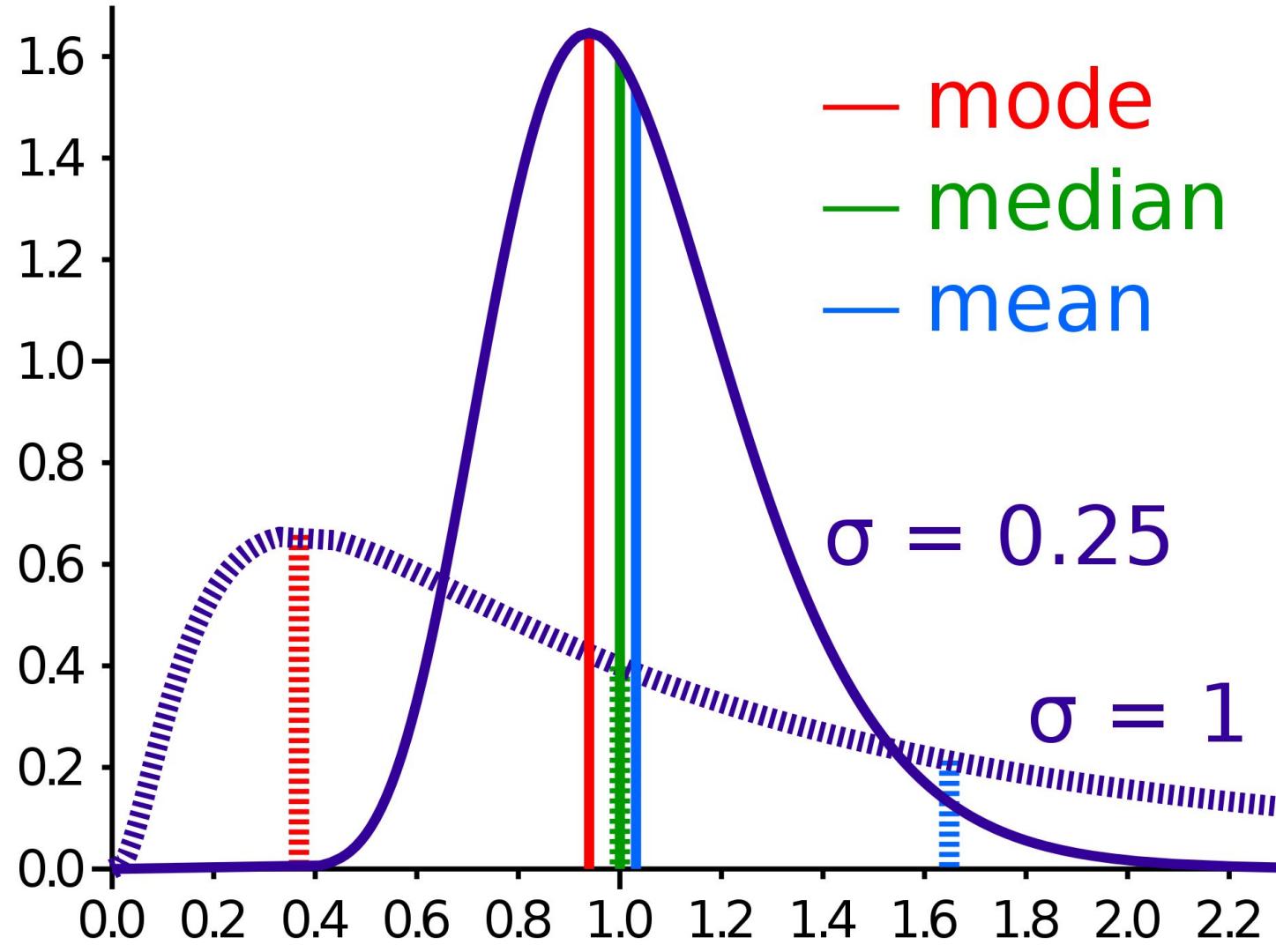
# Log-normal分布

Distributions of continuous variables in Massachusetts Travel Survey dataset.



Fournier, Nicholas, and Eleni Christofa. "On the impact of income, age, and travel distance on the value of time." *Transportation research record* 2675.3 (2021): 122-135.

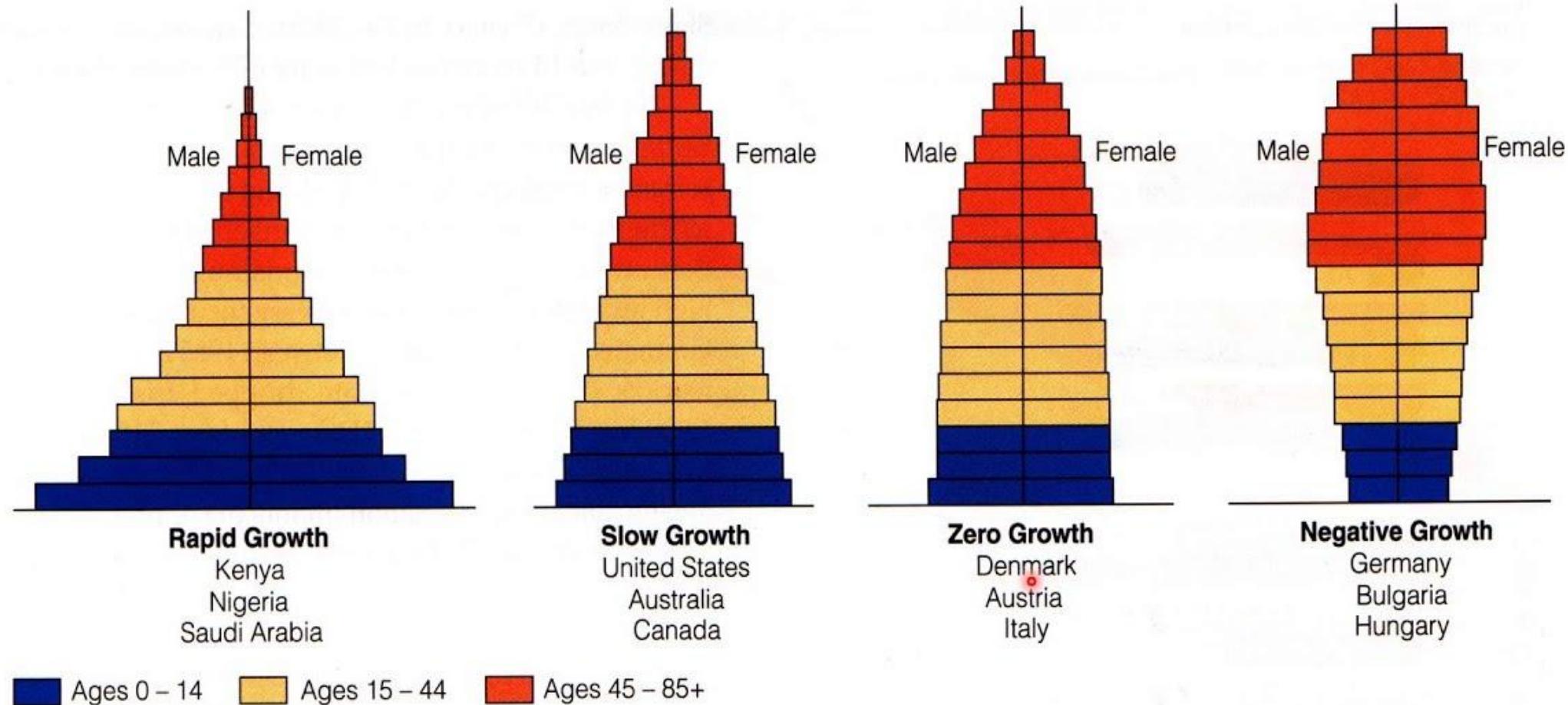
# 收入分布



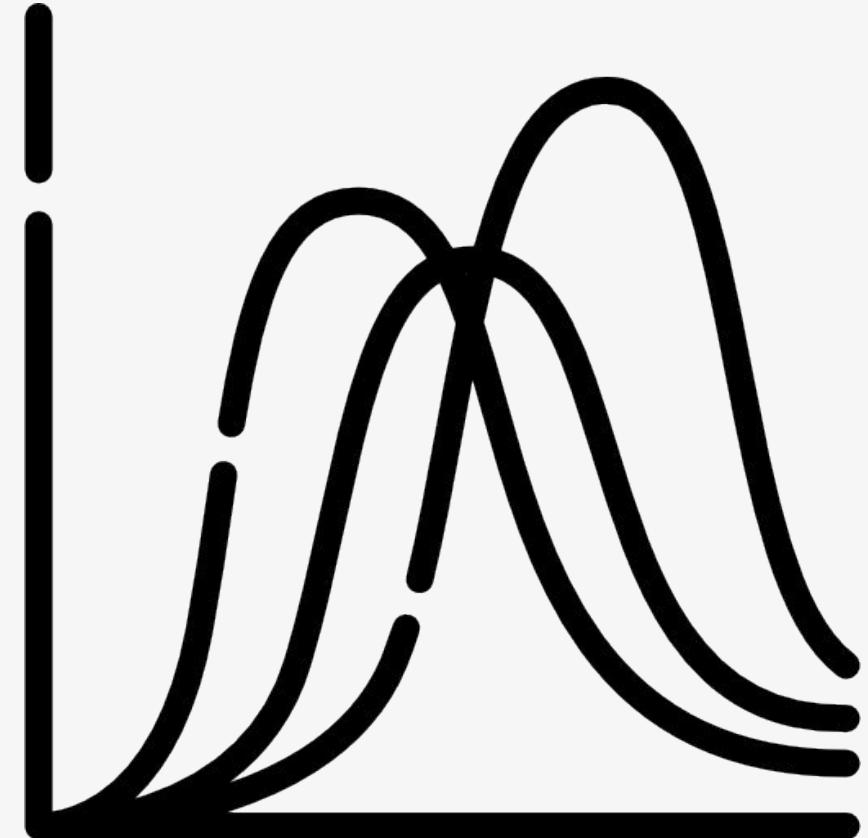
延伸阅读：  
What the difference between mean and median tells us about income inequality  
<https://blog.datawrapper.de/weekly-charts-income/>

# 如何量化年龄结构分布差异？

## Types Of Population Pyramids



- 基本概率分布
- 量化数据分布
- 贝叶斯公式
- 函数拟合
- 数据相关性

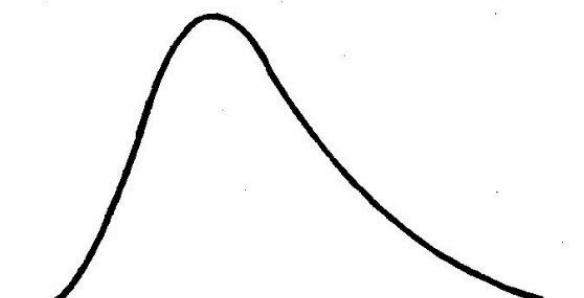


# 偏度skewness

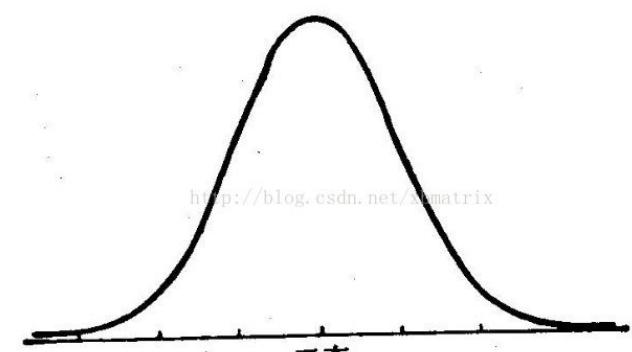
偏度(skewness)，是统计数据分布偏斜方向和程度的度量，是统计数据分布非对称程度的数字特征。正态分布的偏度为0；<0负偏态，左偏；>0正偏态，右偏；

$$Skew(X) = E \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right]$$

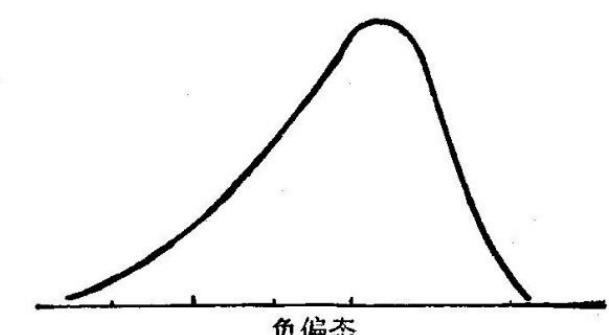
收入分布是如何偏斜的？  
均值和中位数的大小关系如何？



正偏态

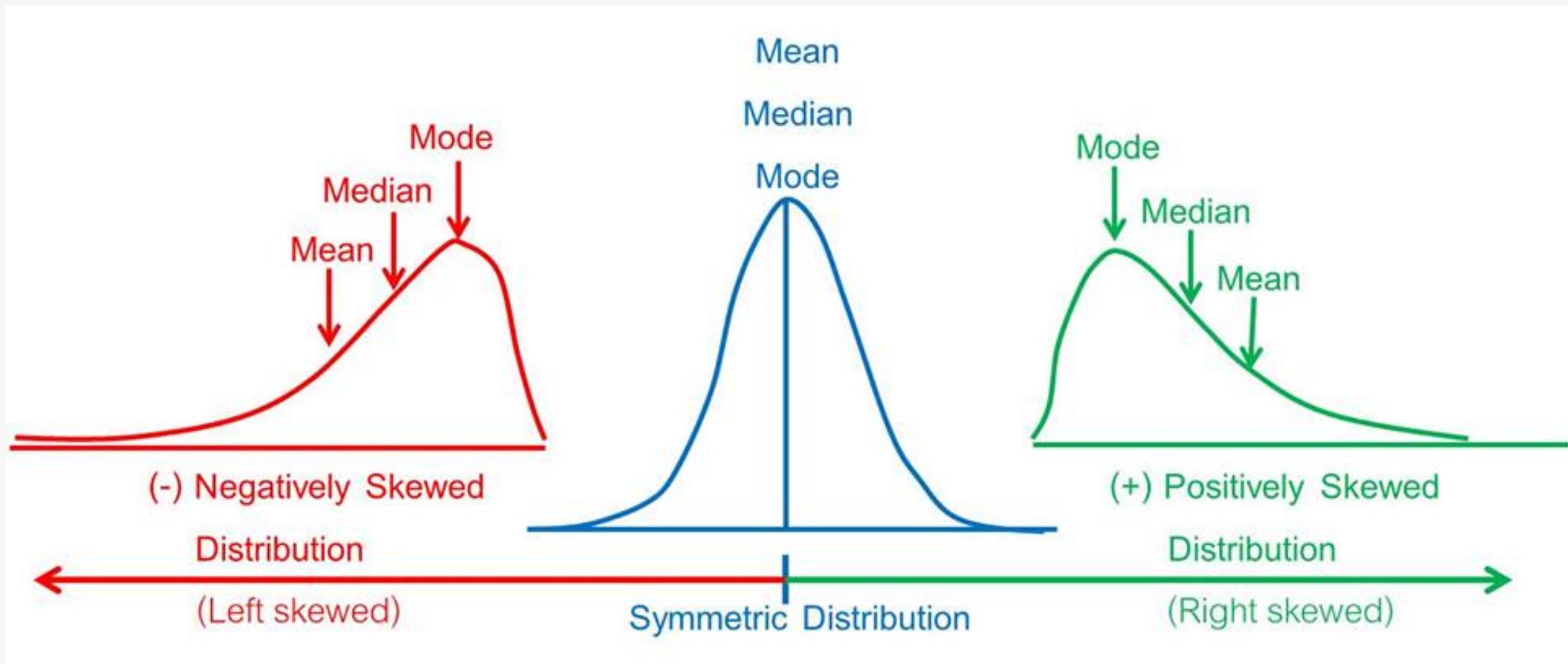


正态



负偏态

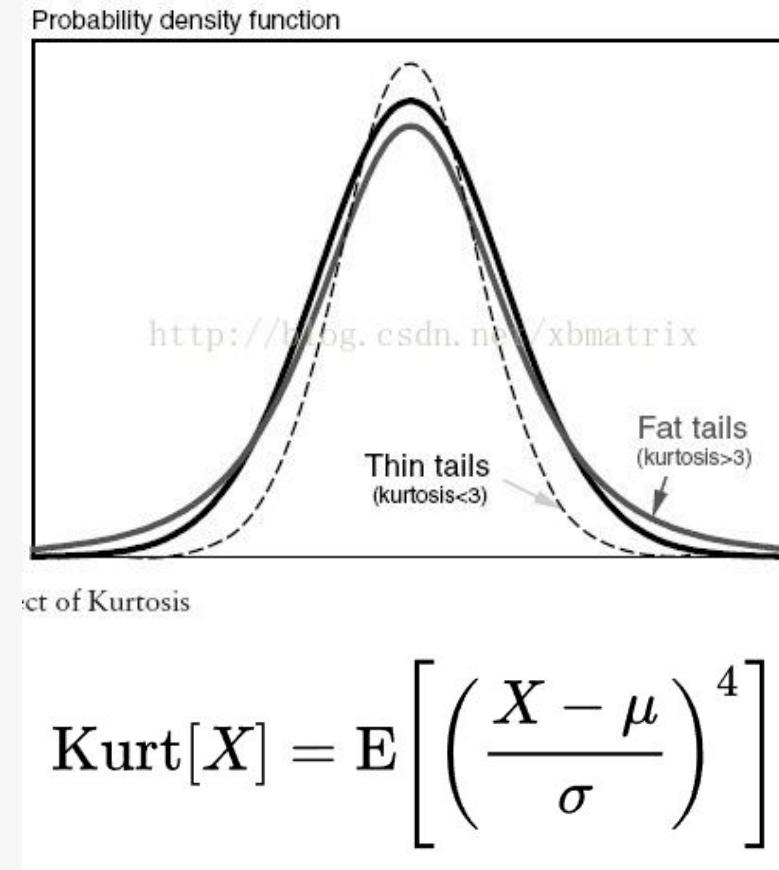
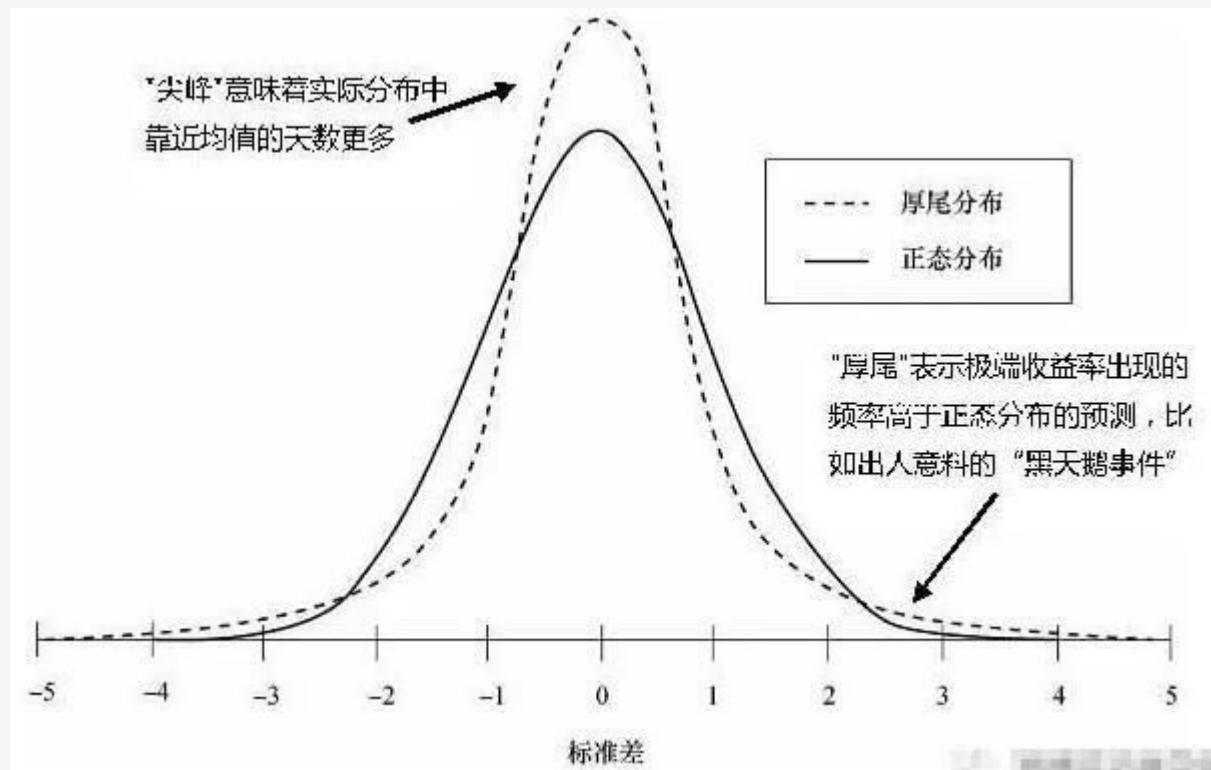
# 偏度skewness



# 峰度kurtosis



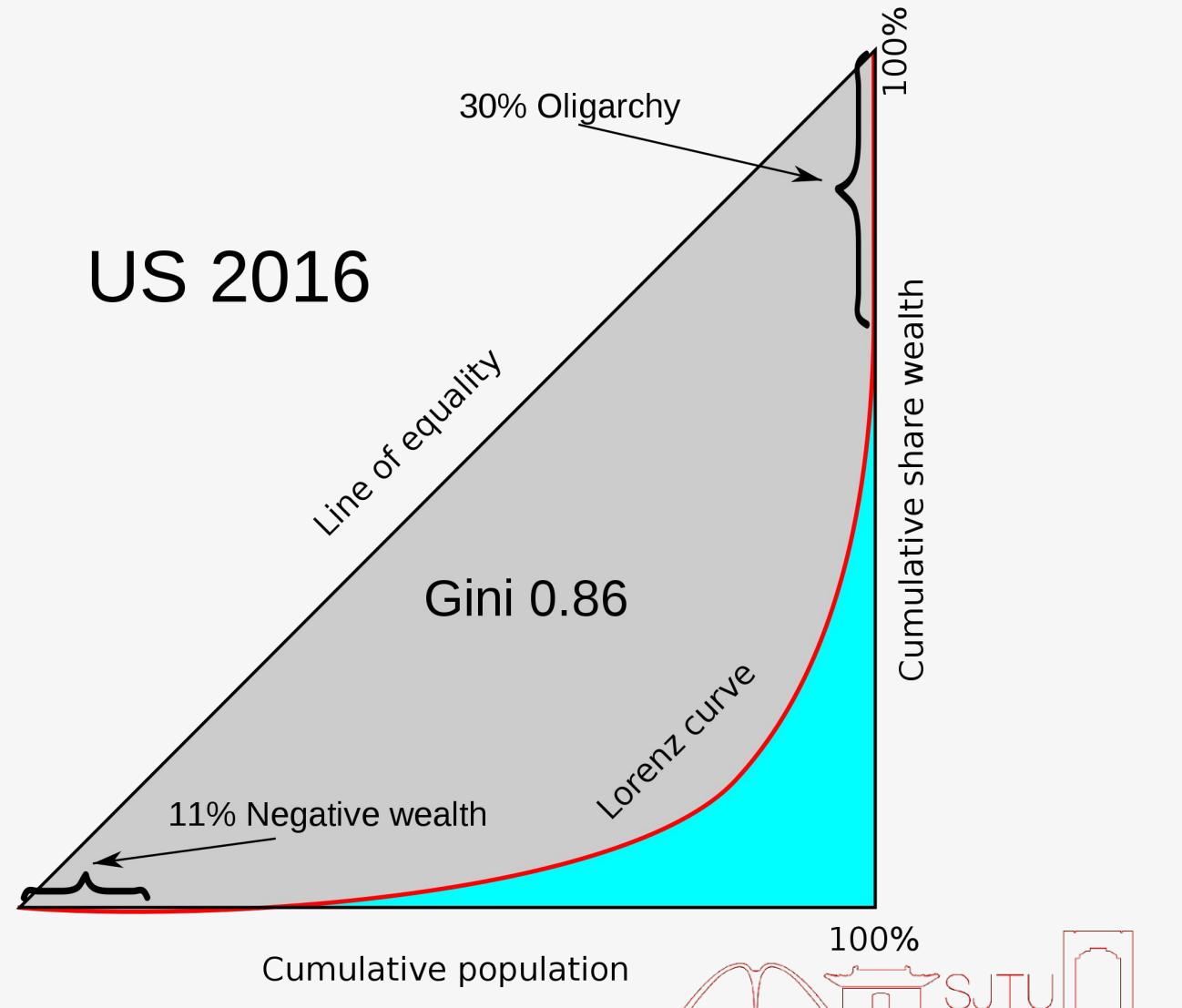
峰度(kurtosis)，表征概率密度分布曲线在平均值处峰值高低的特征数。直观看来，峰度反映了峰部的尖度。正态分布峰度为3，小于3瘦尾，大于3厚尾。



<https://towardsdatascience.com/skewness-kurtosis-simplified-1338e094fc85>

# 基尼系数 Gini Coefficient

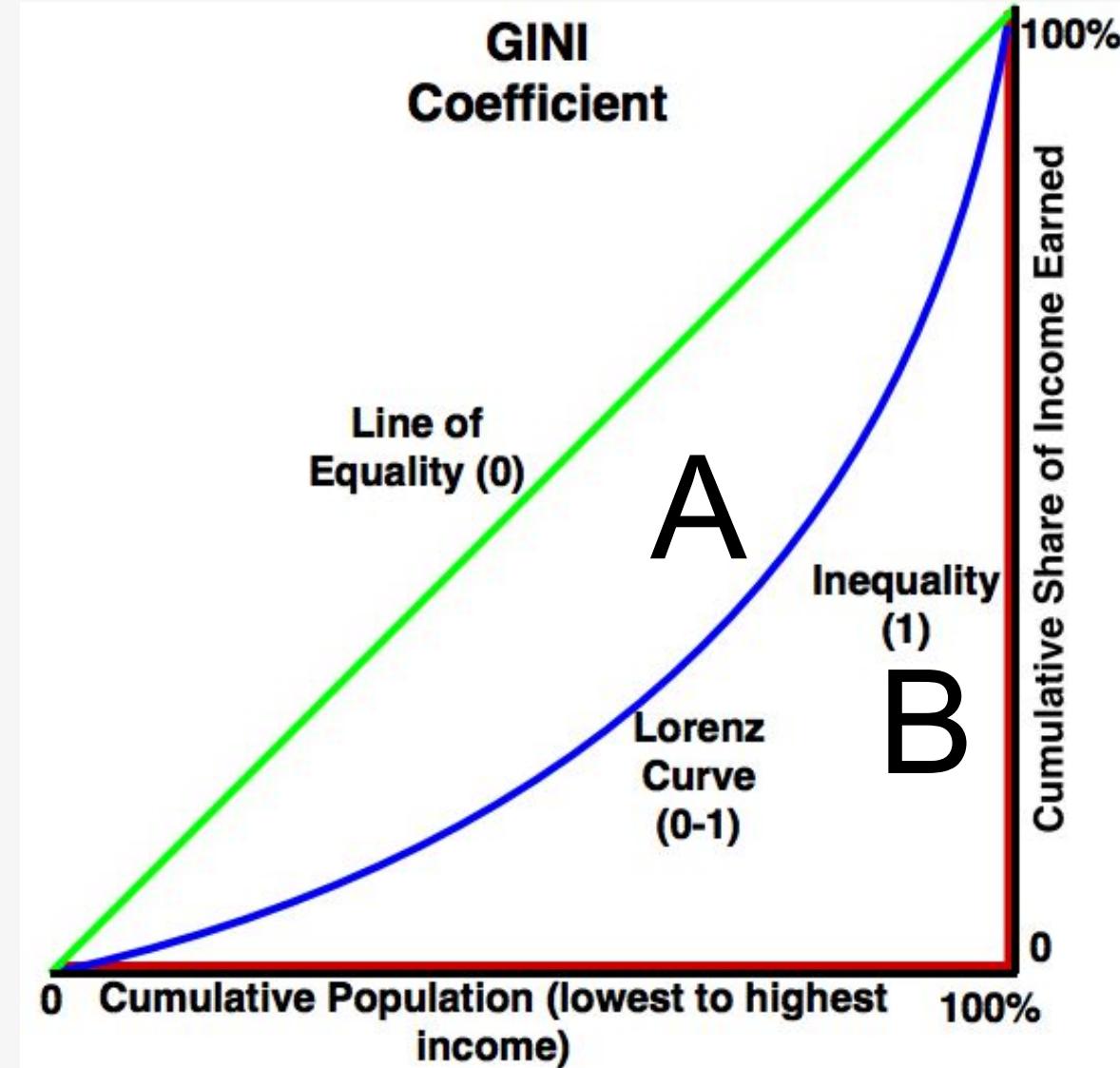
- **Lorenz曲线**: 类似于CDF的定义, 表示一定比例的最低收入家庭(x轴)掌握了社会财富的占比(y轴);
- **Lorenz曲线越靠近绝对公平线**, A区域的面积和基尼系数越小, 越接近绝对公平分布。



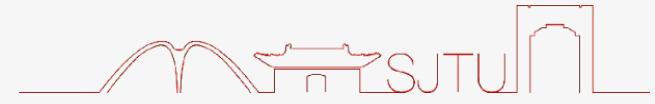
# 基尼系数 Gini Coefficient



- 一个用于判断收入分配公平程度的指标，范围为0(完全平等)-1(完全不平等)；
- 基尼系数定义为A区域的面积除以A和B的面积和；
- 基尼系数同样可以评价**分布的均匀性**



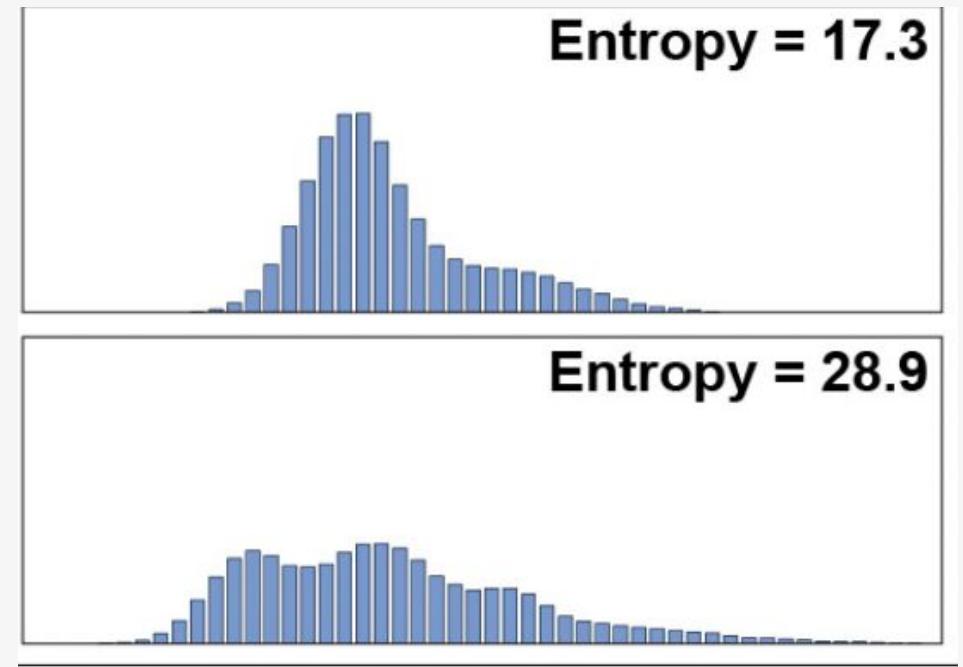
# 数据分布的熵



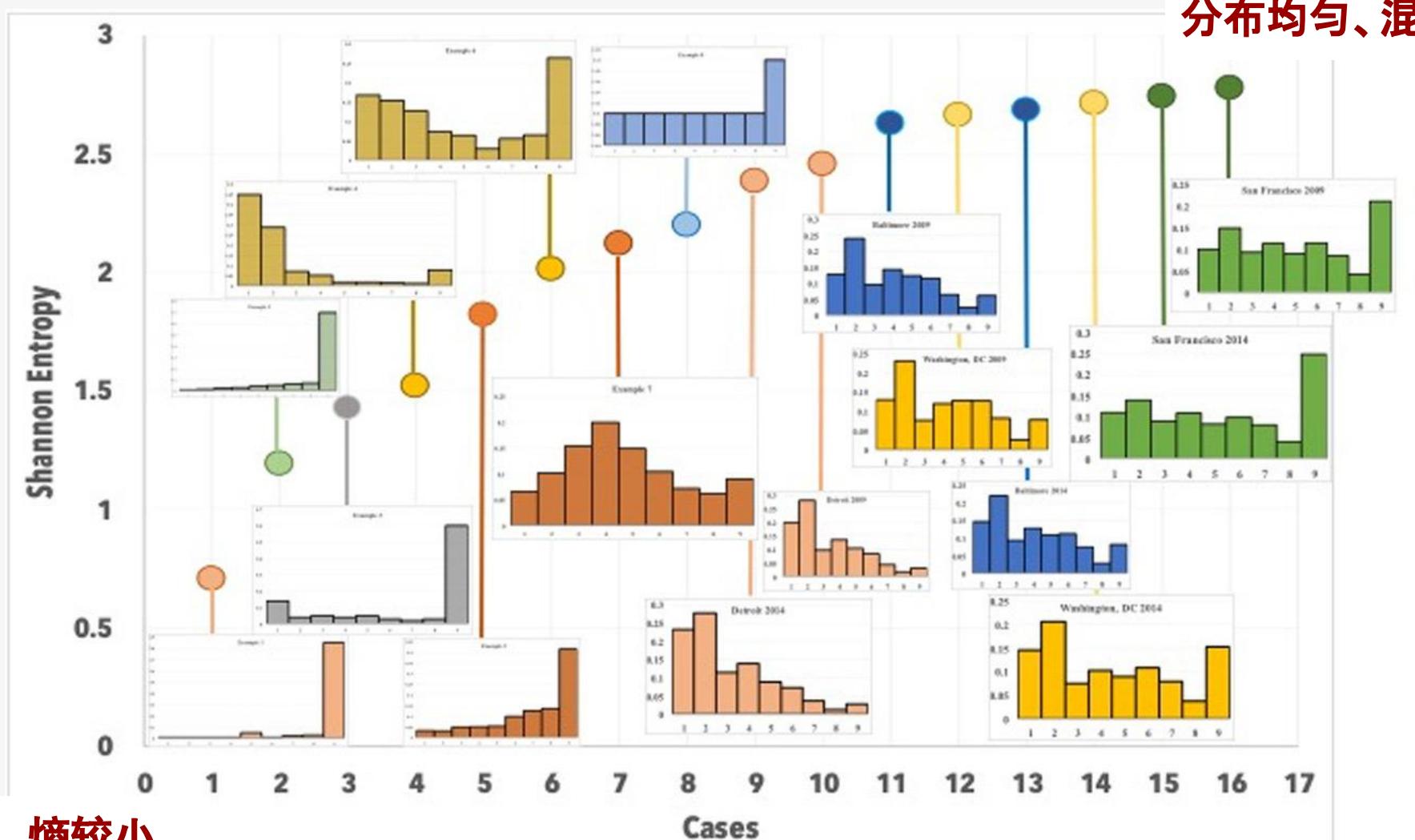
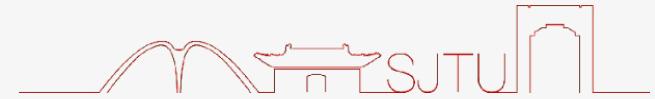
- **信息熵**是C.E.Shannon从热力学中借用的概念，定量描述了信息的不确定程度和信息量的多少。
- 信息不确定性越高，信息量越大，熵越大

$$H(x) = - \sum_{i=1}^n P(x_i) \log_2 P(x_i)$$

信息熵中 $\log$ 底数常用2，是由于计算机中采用0-1作为最小信息单元。实际上根据换底公式，不同底数的熵可以简单地进行转换计算。

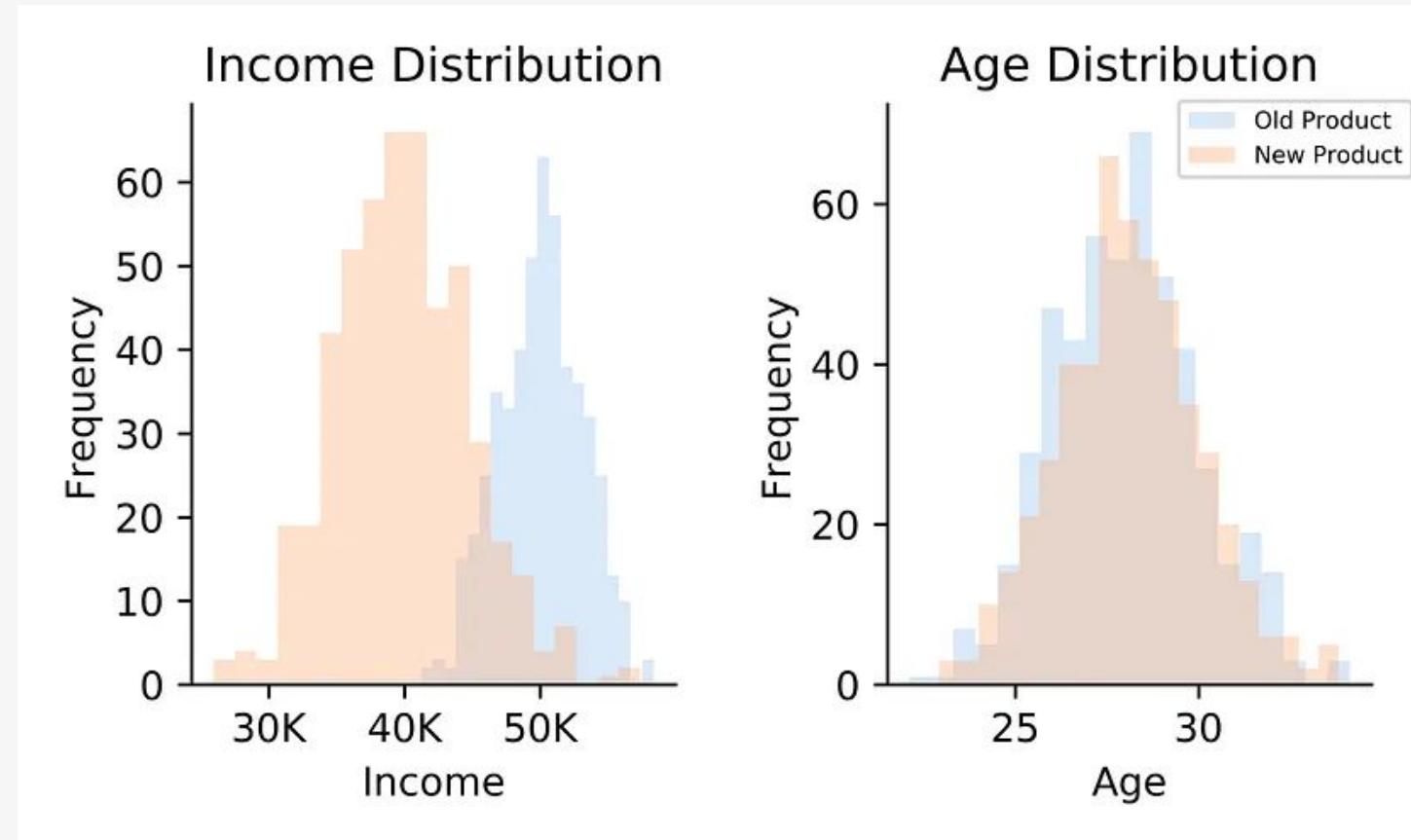


# 数据分布的熵



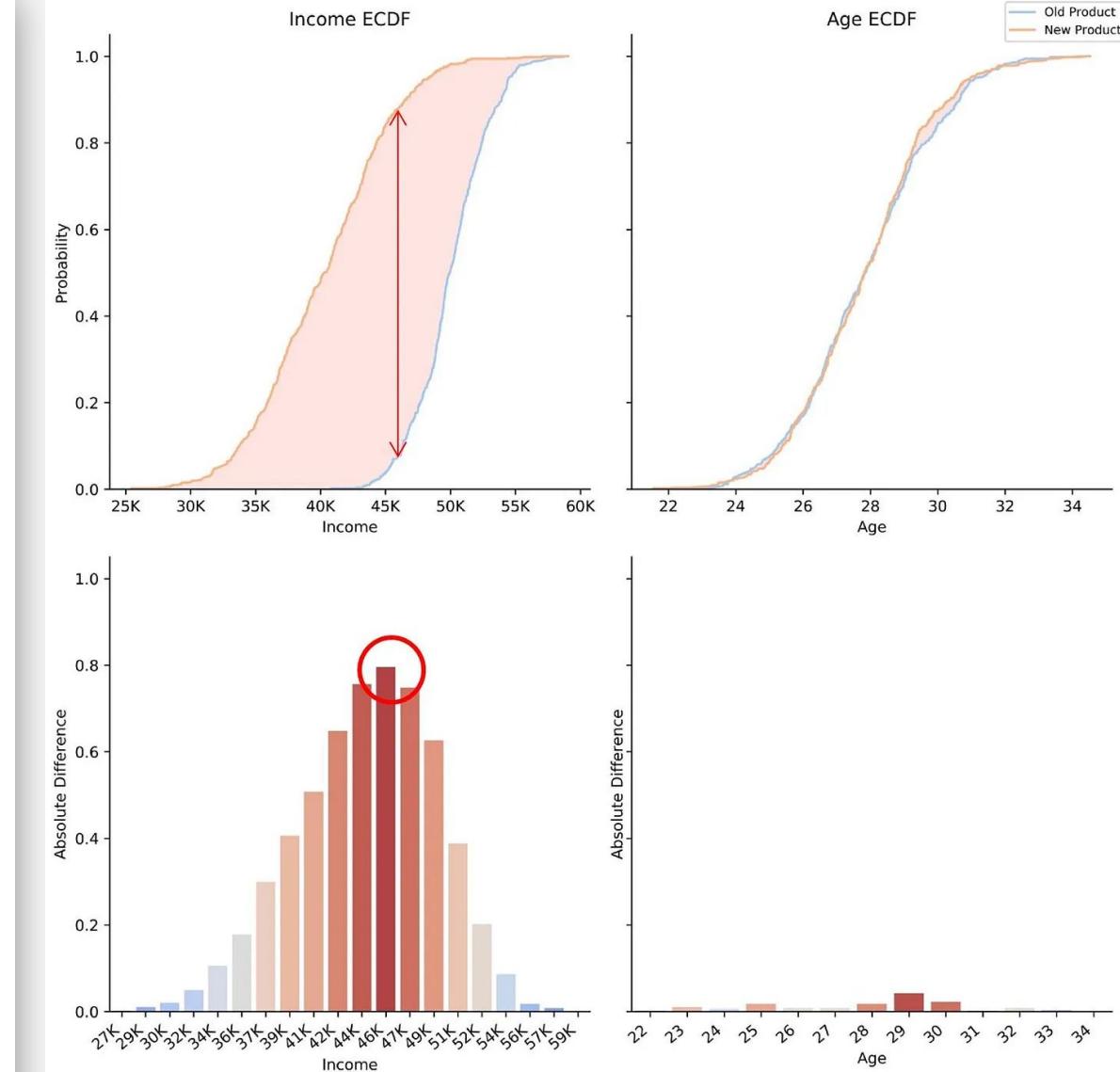
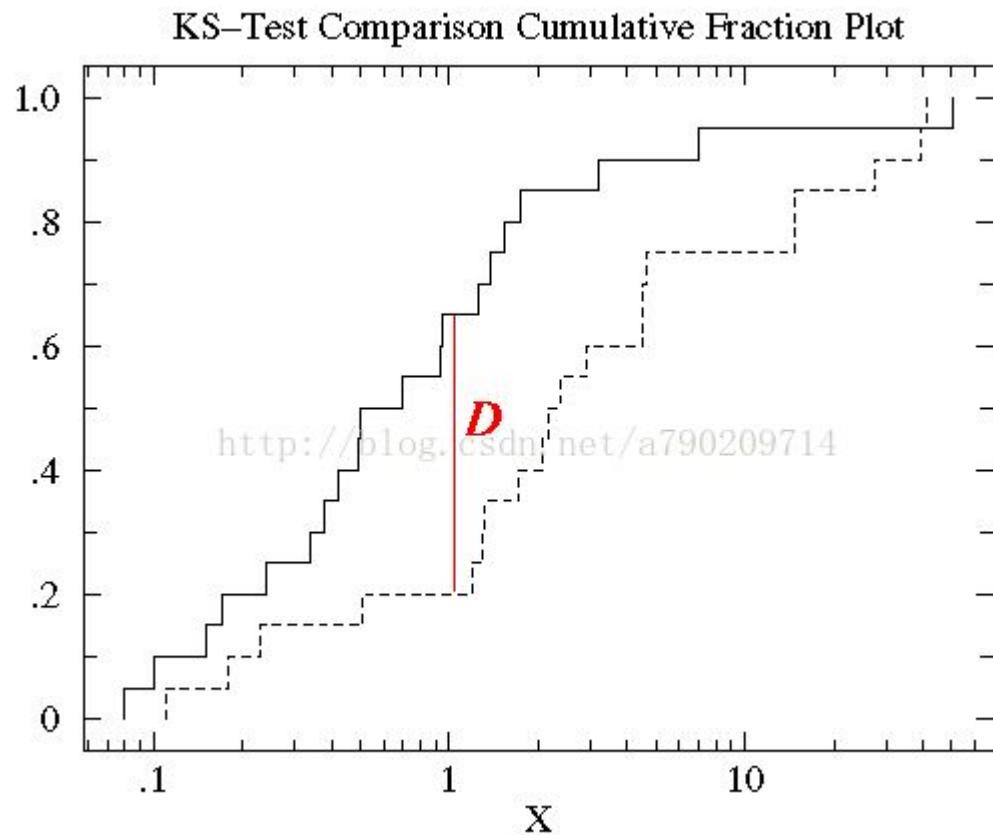
# 分布的差异

- 右图中给出了新旧两个产品的用户特征分布差异
- 直观来看，用户的收入水平分布存在很大差异，而年龄分布几乎没有差异
- 怎么量化？
  - 比较均值、方差？
  - 非高斯分布只用均值和方差进行描述会损失信息



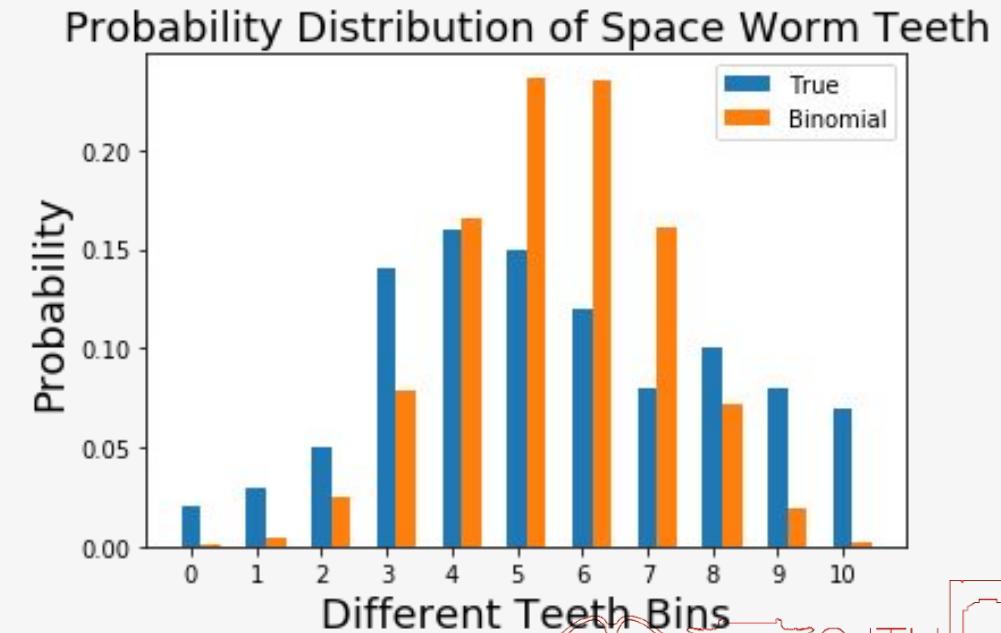
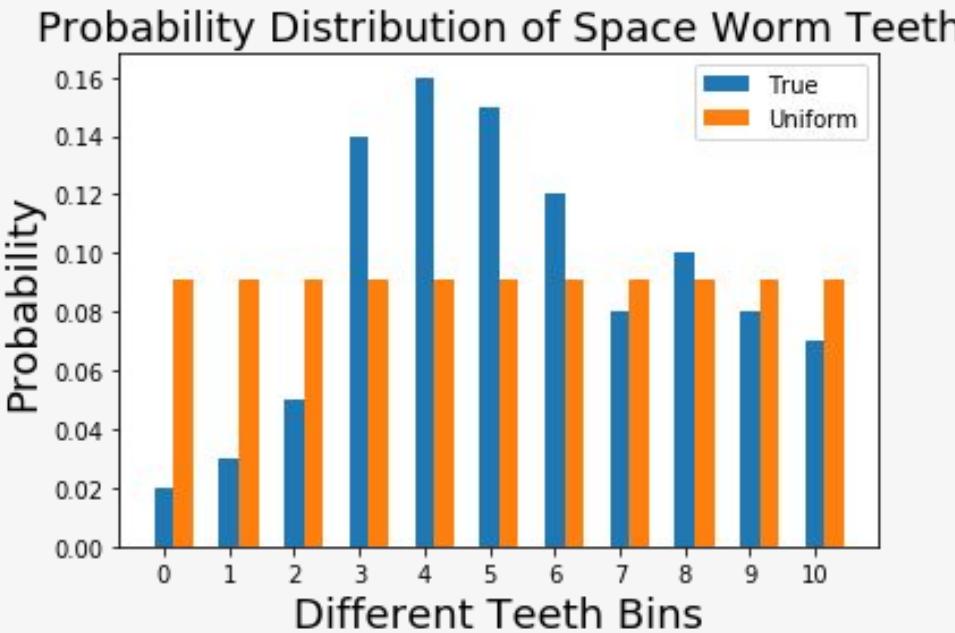
# 度量分布的相似性

KS distance基于累计分布函数, 用以检验两个经验分布是否不同或一个经验分布与另一个理想分布是否不同, 定义为两个CDF间的最大垂直距离



# 分布的差异

- 左右两图分别表示用均匀分布和二项分布拟合蓝色的真实分布，哪个分布的拟合效果更好(更接近蓝色的真实分布)？
- “看起来”好像是二项分布的变化趋势和真实分布更像，如何量化比较？
- 不同的二项分布(单次实验成功率)拟合效果有什么不同？



# 度量分布的相似性

KL(True || Uniform): 0.13667971094966938  
KL(True || Binomial): 0.42734972485619327



上海交通大学  
SHANGHAI JIAO TONG UNIVERSITY

KL divergence, 在机器学习中常用来比较两个分布的差异

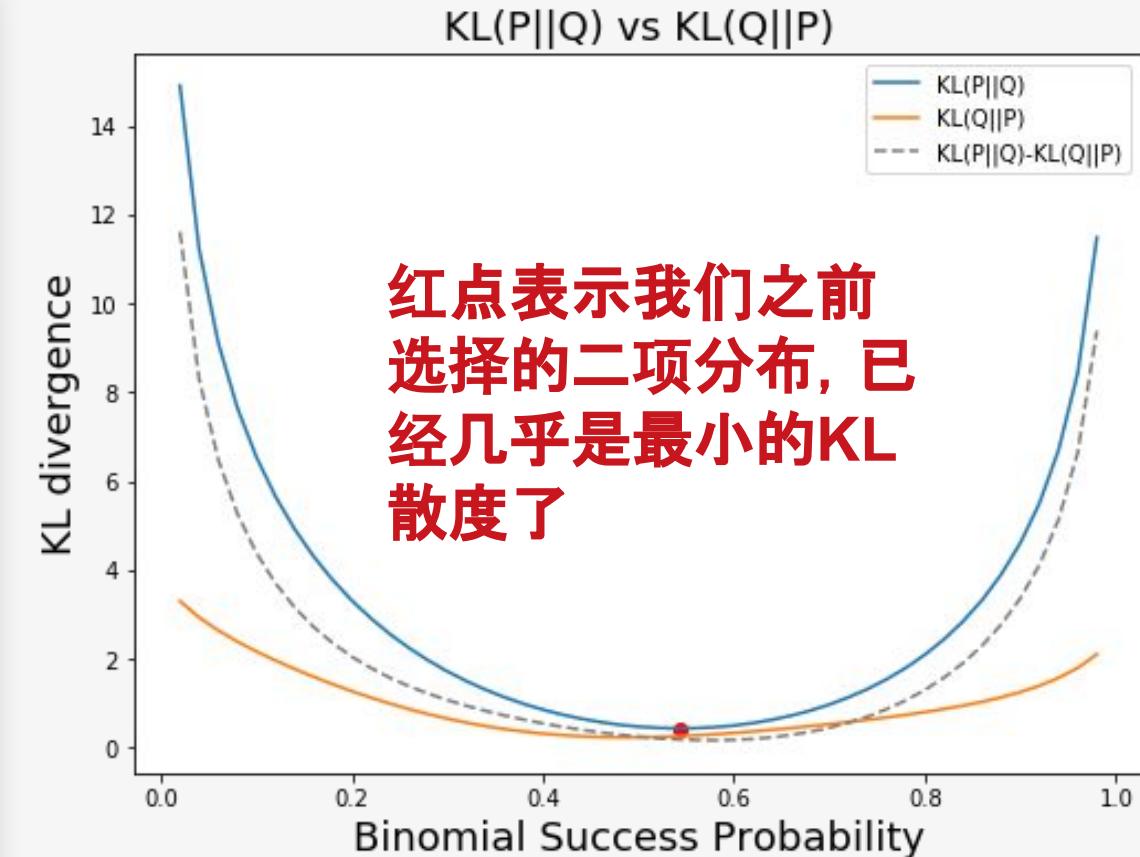
离散和连续形式：

$$KL(P||Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$

$$KL(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

通常非对称，不是距离，定义来自信息论中的熵，视为信息差

$KL(P||Q) \neq KL(Q||P)$



红点表示我们之前选择的二项分布，已经几乎是最小的KL散度了

横轴是二项分布的单次实验成功概率

思考 🤔 :  $P(i)=0$  或  $Q(i)=0$  如何处理？

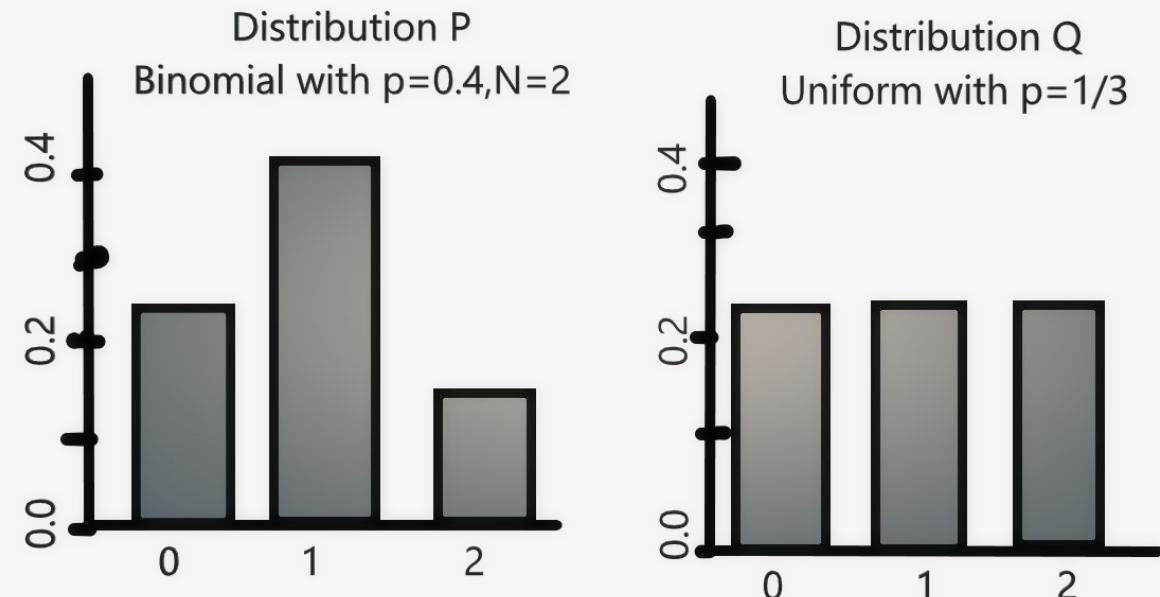
# 度量分布的相似性

计算右表中P和Q的KL散度

$$KL(P||Q)$$

$$KL(Q||P)$$

$$KL(P||Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$



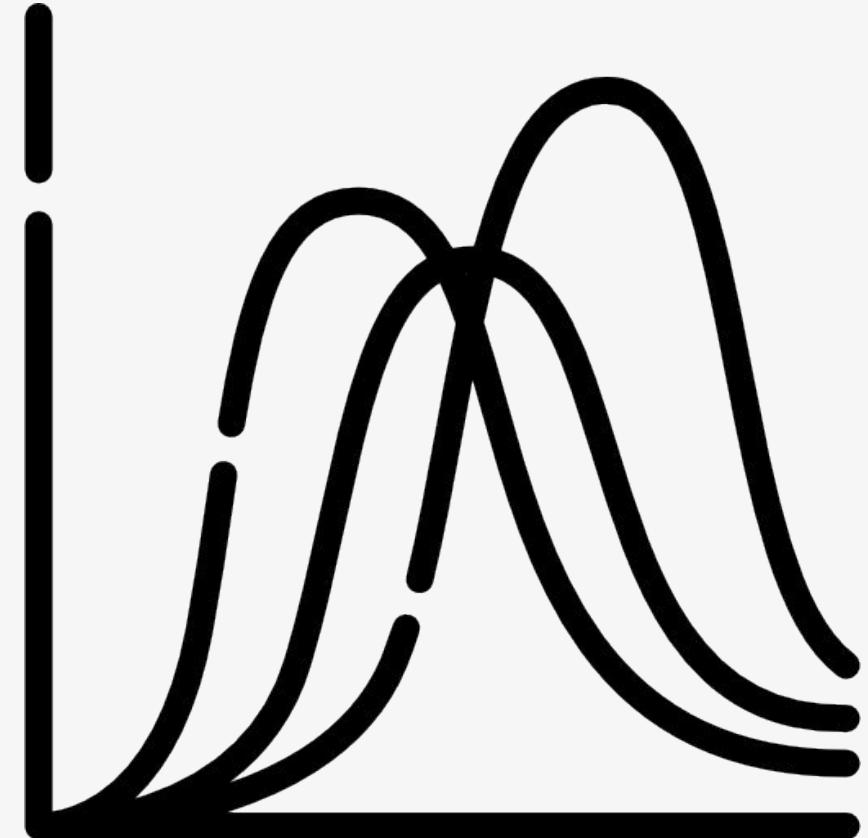
Canvas Quiz

0.085

0.097

X	0	1	2
Distribution P(X)	0.36	0.48	0.16
Distribution Q(X)	0.333	0.333	0.333

- 基本概率分布
- 量化数据分布
- 贝叶斯公式
- 函数拟合
- 数据相关性



- 调查问卷: 调查本科毕业人群的收入分布
- 问题: 给定收入水平, 推测本科毕业的概率



# 贝叶斯公式-条件概率



- $P(A, B)$ 称为联合分布, 表示A和B同时发生的概率。

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

- 条件概率: 给定事件B的前提下(条件), 发生事件A的概率, 记为 $P(A|B)$

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

条件概率公式

- \*在A和B独立时,  $P(A, B)=P(A)*P(B)$ , 此时根据条件概率公式,  $P(A|B)=P(A)$ , 即是否发生B不影响发生A的概率。



# 贝叶斯公式-全概率公式



- 全概率公式为:  $P(A) = P(A,B) + P(A, \neg B)$

- B发生和不发生, A发生的联合概率之和

- 如果A和B是离散变量

$$P(A = a) = \sum_b P(A = a, B = b)$$

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

条件概率公式

可以记为

$$P(a) = \sum_b P(a,b)$$

全概率公式

结合条件概率公式, 可以得到

$$P(a) = \sum_b P(a|b)P(b)$$



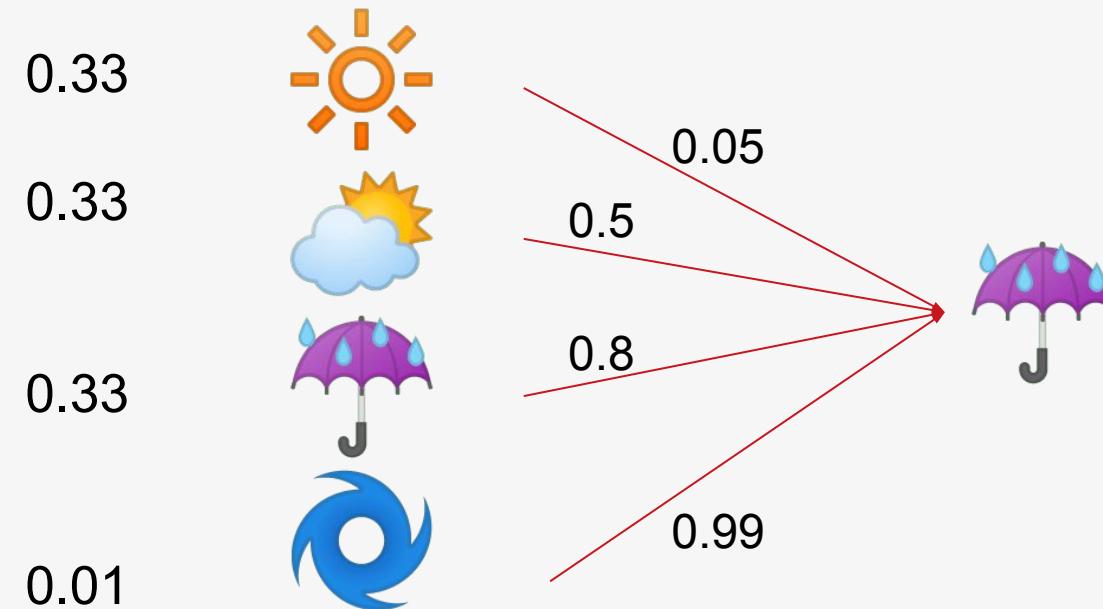
# 贝叶斯公式-全概率公式

- 全概率公式为:  $P(A) = P(A,B) + P(A,\neg B)$
- 明天下午下雨的概率有多大?

$$P(a) = \sum_b P(a|b)P(b)$$

明天早上

明天下午



Canvas Quiz

# 贝叶斯公式

- 对条件概率公式中联合概率项采用条件概率变形, 得到(离散变量形式)

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

- 再对分母采用全概率公式, 得到

$$P(a|b) = \frac{P(b|a)P(a)}{\sum_{a^*} P(b|a^*)P(a^*)}$$

$$P(a|b) = \frac{P(a,b)}{P(b)}$$

条件概率公式

$$P(b) = \sum_{a^*} P(b|a^*)P(a^*)$$

全概率公式

- 调查问卷: 调查本科毕业人群的收入分布
- 问题: 给定收入水平, 推测本科毕业的概率



# 贝叶斯公式的社会科学应用

- A:本科毕业
- B:代表收入水平
- $P(B|A)$ :本科毕业人群的收入水平分布
  - 调查本科毕业人群获得
- $P(A|B)$ :给定收入水平, 推测本科毕业的概率
- $P(A)$ :全国本科毕业率
- $P(B)$ :全国收入水平分布

似然概率

LIKELIHOOD

The probability of "B" being True, given "A" is True

$P(A|B)$

POSTERIOR

The probability of "A" being True, given "B" is True

后验概率

先验概率

PRIOR

The probability "A" being True. This is the knowledge.



$P(B|A).P(A)$

$P(B)$



MARGINALIZATION

The probability "B" being True.

边缘概率



# 贝叶斯公式的社会科学应用



- A:本科毕业
- B:收入水平
- $P(B|A)$ :本科毕业人群的收入水平分布
  - 调查本科毕业人群获得
- $P(A|B)$ :给定收入水平, 推测本科毕业的概率
- $P(A)$ :全国本科毕业率17%
- $P(B)$ :全国收入水平分布

收入水平B (月薪)	0-5000	5000-15000	15000+
本科毕业 $P(B A)$	0.4	0.55	0.05
全国 $P(B=b)$	0.69	0.29	0.02
$P(A B=b)$			

求  $P(A|B=b) =$



# 贝叶斯公式的社会科学应用



**目标:** 依据居民收入与里程信息, 推测是否使用新能源汽车(EV)

**已有数据:** EV车主调查数据(收入、里程) California Plug-in Electric Vehicle Driver Survey (2013)、全市居民收入分布、全市居民里程分布。

$$P(EV | I_u, D_u) = \frac{P(I_u, D_u | EV)P(EV)}{P(I_u, D_u)}$$

$$= \frac{P(I_u | EV)P(D_u | EV)P(EV)}{P(I_u)P(D_u)}$$

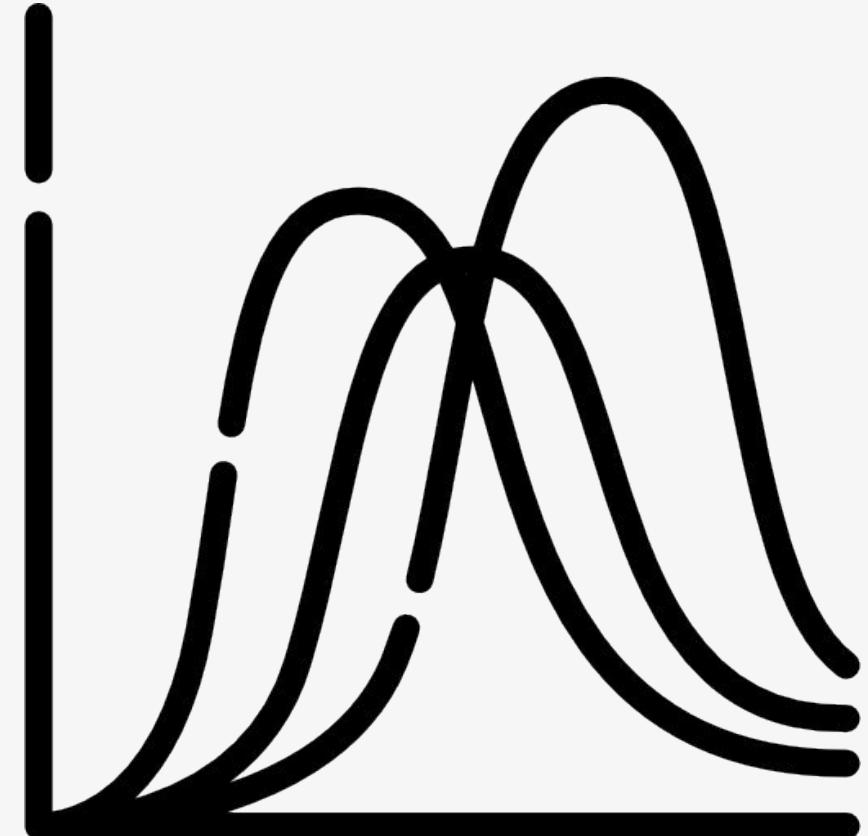
$I_u$ , 居民收入分布

$D_u$ , 居民里程分布

**P(EV) = 0.62%** 调查数据发生时的EV市场占有率为 (share of EVs within all cars in the Bay Area in the end of 2013)



- 基本概率分布
- 量化数据分布
- 贝叶斯公式
- 函数拟合
- 数据相关性



# 常用的拟合函数

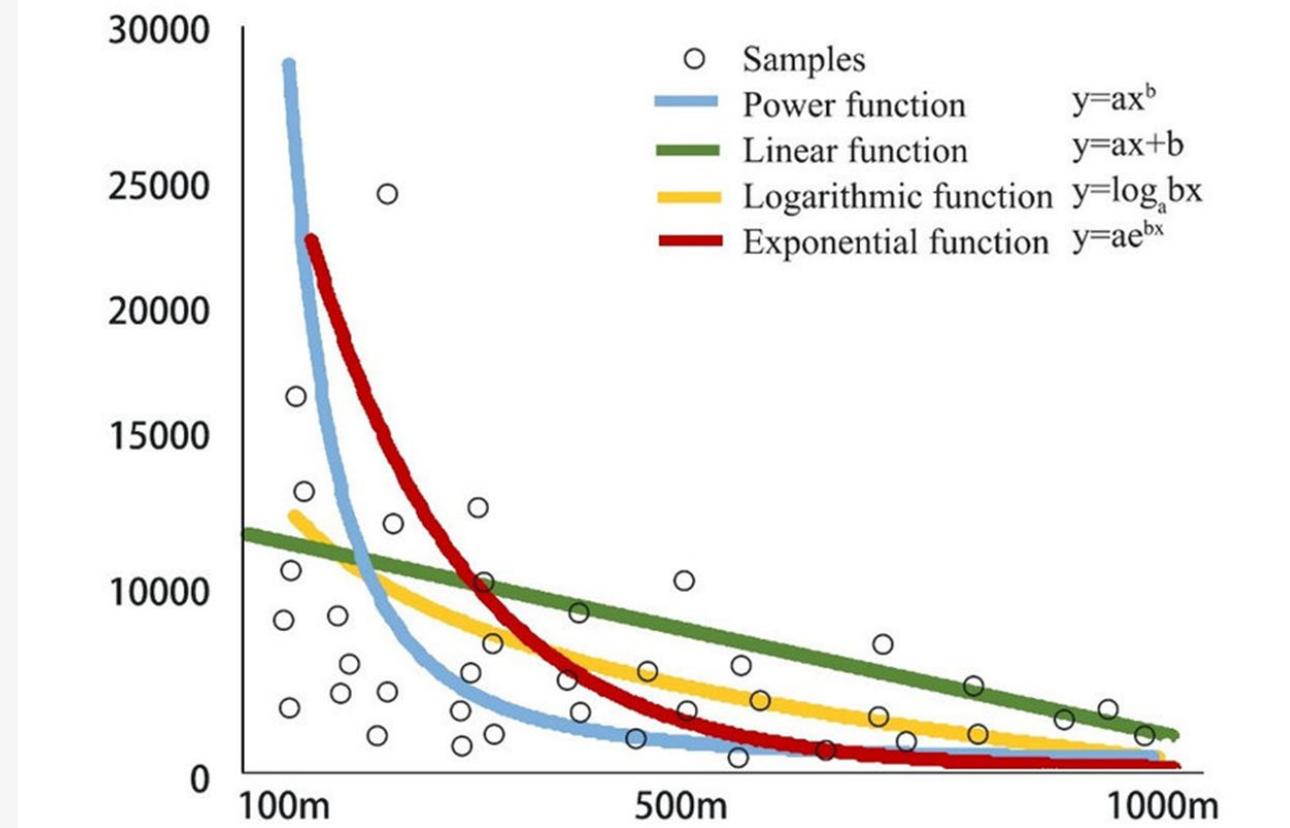
线性函数

二次/三次函数

对数函数  $y = a \ln(x) + b$

指数函数  $y = ae^{bx}$

幂函数  $y = ax^b$



# Log-scale和power-law

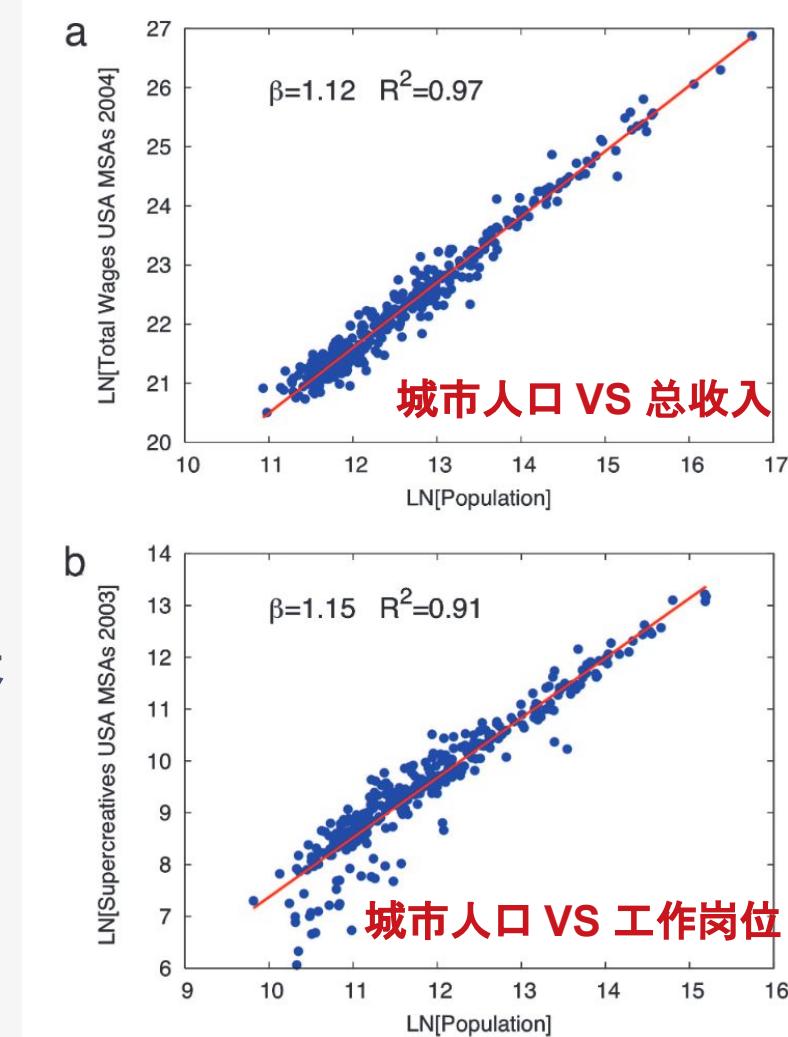
注意右图中，横纵坐标都做了取对数的变换，称为双对数坐标，并分别拟合出了斜率为1.12和1.15的线性模型。

在双对数坐标下的线性模型是什么含义？

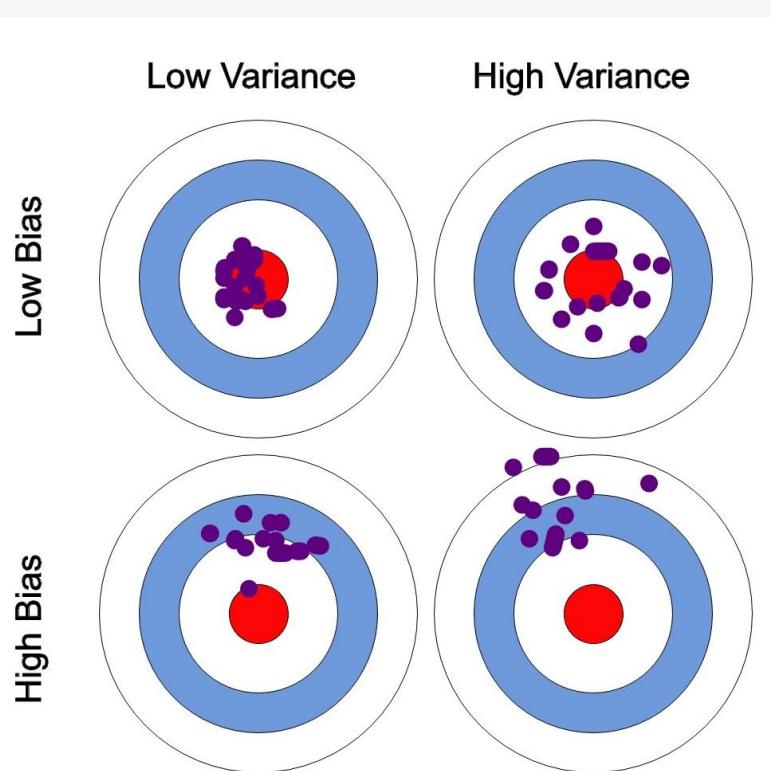
$$y_{ln} = \beta x_{ln} + \alpha$$

$$y = x^\beta e^\alpha$$

也就是当x变为两倍时，y变为原来的  $2^\beta$  倍，说明y和x的关系是超线性关系。



简单来说，拟合好的模型预测和真实数据之间的Error由什么构成？一部分是模型的偏差(Bias)，另一部分是模型本身方差(Variance)。



- Basic Model:  $Y = f(X) + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$
- The expected prediction error of a regression fit  $\hat{f}(X)$

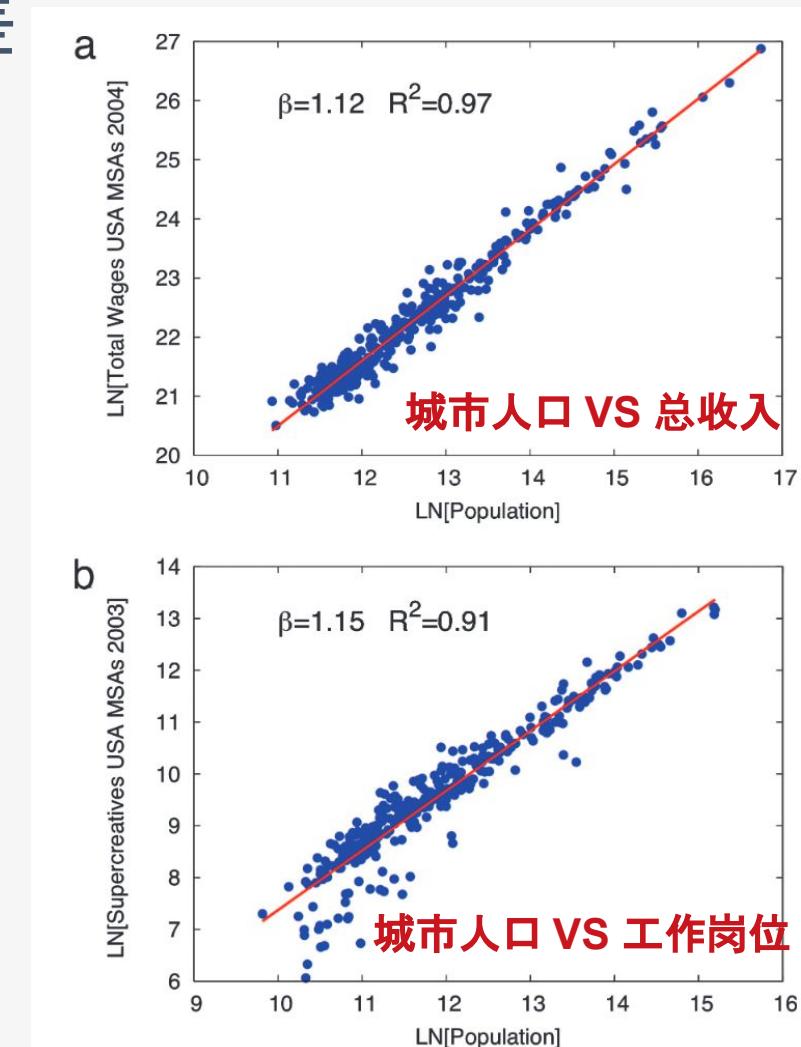
$$\begin{aligned} Err(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\ &= \sigma_\varepsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 \\ &= \sigma_\varepsilon^2 + Bias(\hat{f}(x_0))^2 + Var(\hat{f}(x_0)) \\ &= \text{Irreducible Error} + Bias^2 + Variance \end{aligned}$$

# R-squared

- $R^2$  表示经过模型拟合，残差的方差比变量的方差减少了多少。

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- 例如拟合前变量的方差为10，经过拟合后，残差的方差只有2，那么  $R^2$  为0.8；



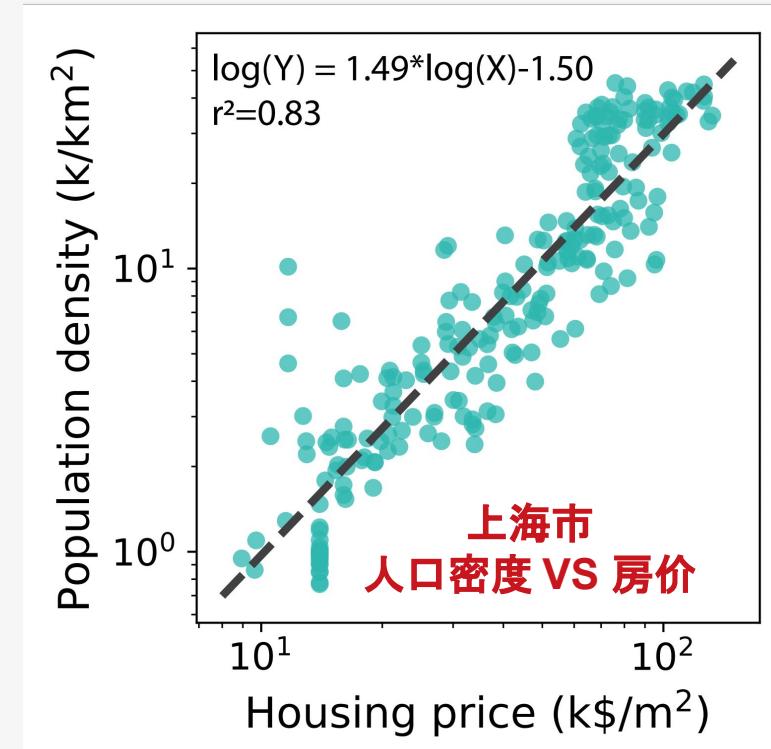
# R-squared

- 什么是 $R^2=0$ ? 预测房价的时候, 把上海市所有房子的房价全部加起来取平均, 粗暴地认为上海房价都是这个平均值, 此时这个预测模型的 $R^2$ 为0。因为它这个预测根本没有减少残差!

- 什么是 $R^2=1$ ? 做出完美预测?

- **⚠**  $R^2$ 并不是某个数的平方, 因此 $R^2$ 可以小于0, 说明这个预测比均值预测还要差。

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$



# 模型复杂度和误差的关系

- 模型复杂度可以简单用参数数量来衡量
- 线性模型有两个独立参数，二次曲线有三个参数...
- 一般来讲，模型复杂度越高，能在训练集上拟合得更好；但复杂度太高的时候，换一组数据，效果就不好了，甚至可能很差！

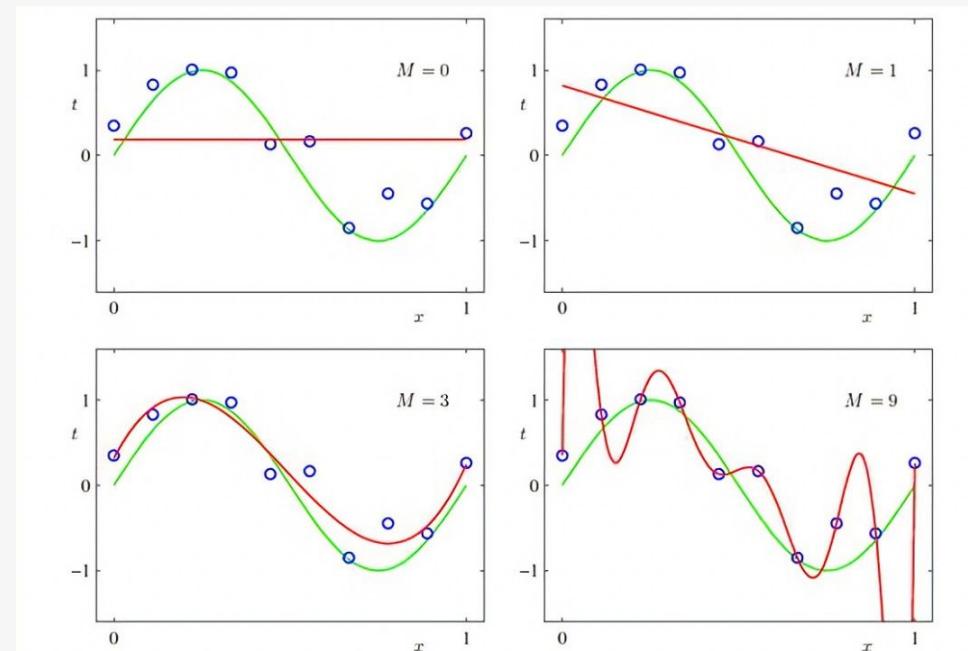
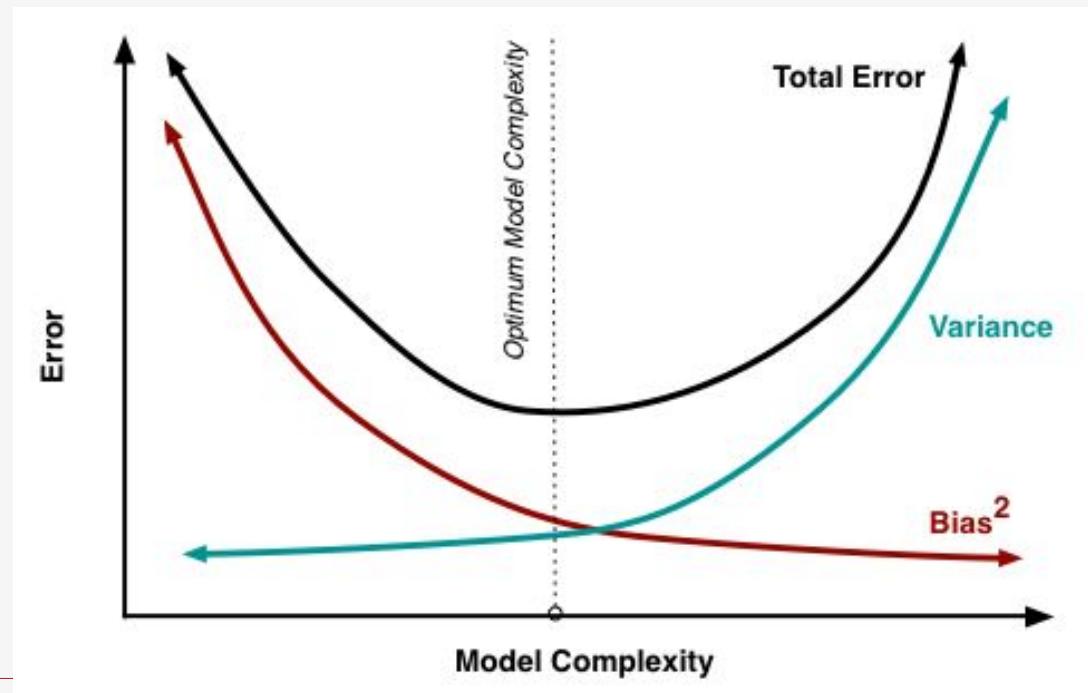
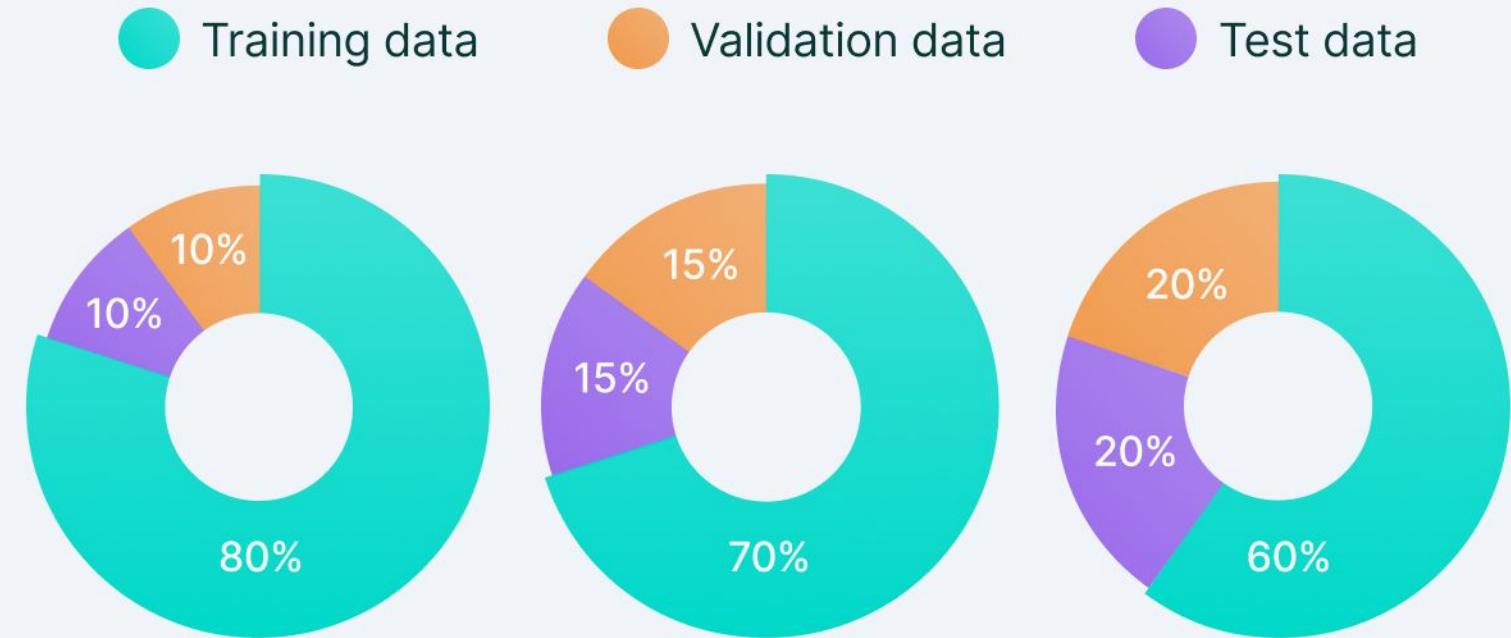


Figure 1.4 Plots of polynomials having various orders  $M$ , shown as red curves, fitted to the data set shown in Figure 1.2.

# 模型性能的估计

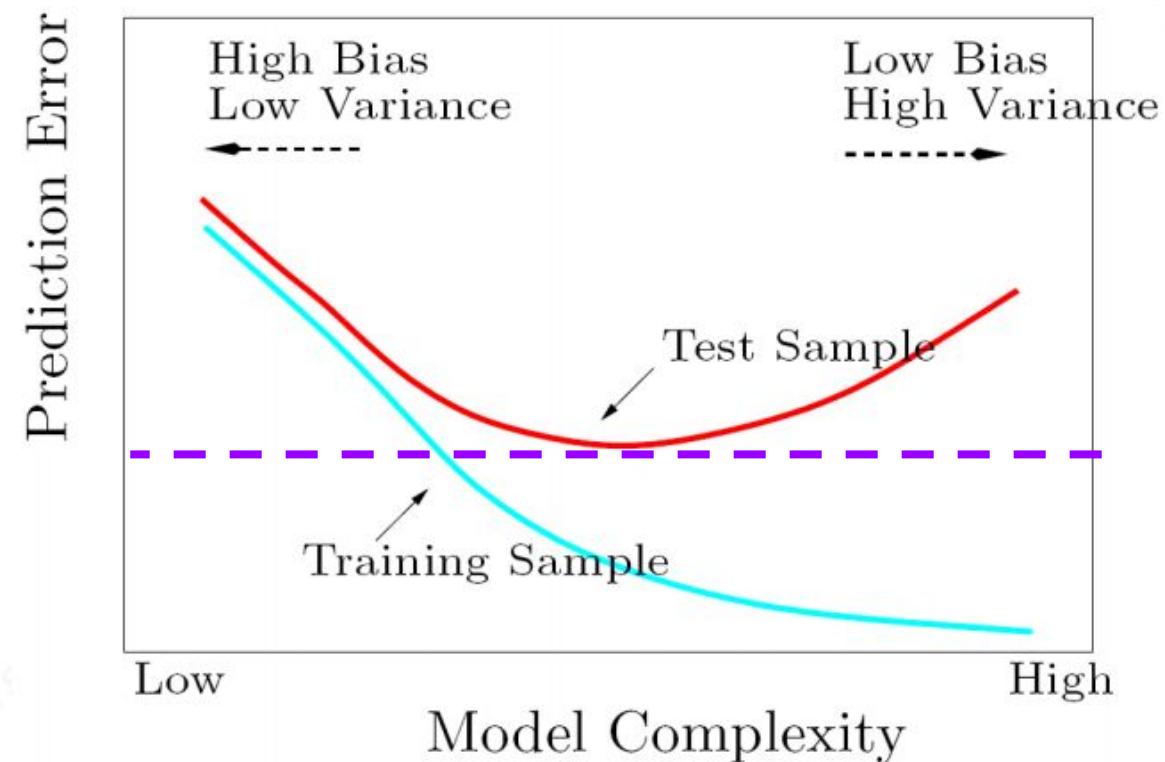
- **泛化能力**指模型在非训练数据集上的性能
- 实际项目中，通常将原始数据集划分为**训练集**、**验证集**和**测试集**；
- 其中验证集会在模型训练、模型和超参数选择的过程中用于评估，而测试集则是为拟合后的模型进行泛化性能评估。



V7 Labs

# 模型性能的估计

- 越复杂的模型，在训练集上可能可以获得更低的误差，但在测试集上却可能出现过大的误差，此时模型的方差过大，也称为**过拟合**；
- 与之相对应地，模型复杂度越低，在训练集和测试集上的误差都很大，称为**欠拟合**；



# 过拟合vs欠拟合

- 欠拟合:模型参数不足, 偏差大;
- 过拟合:模型参数过多, 方差大;
- 右图中, 3阶曲线已经足够拟合, 9阶曲线在训练数据(蓝点)以外严重偏离真实数据(绿线);

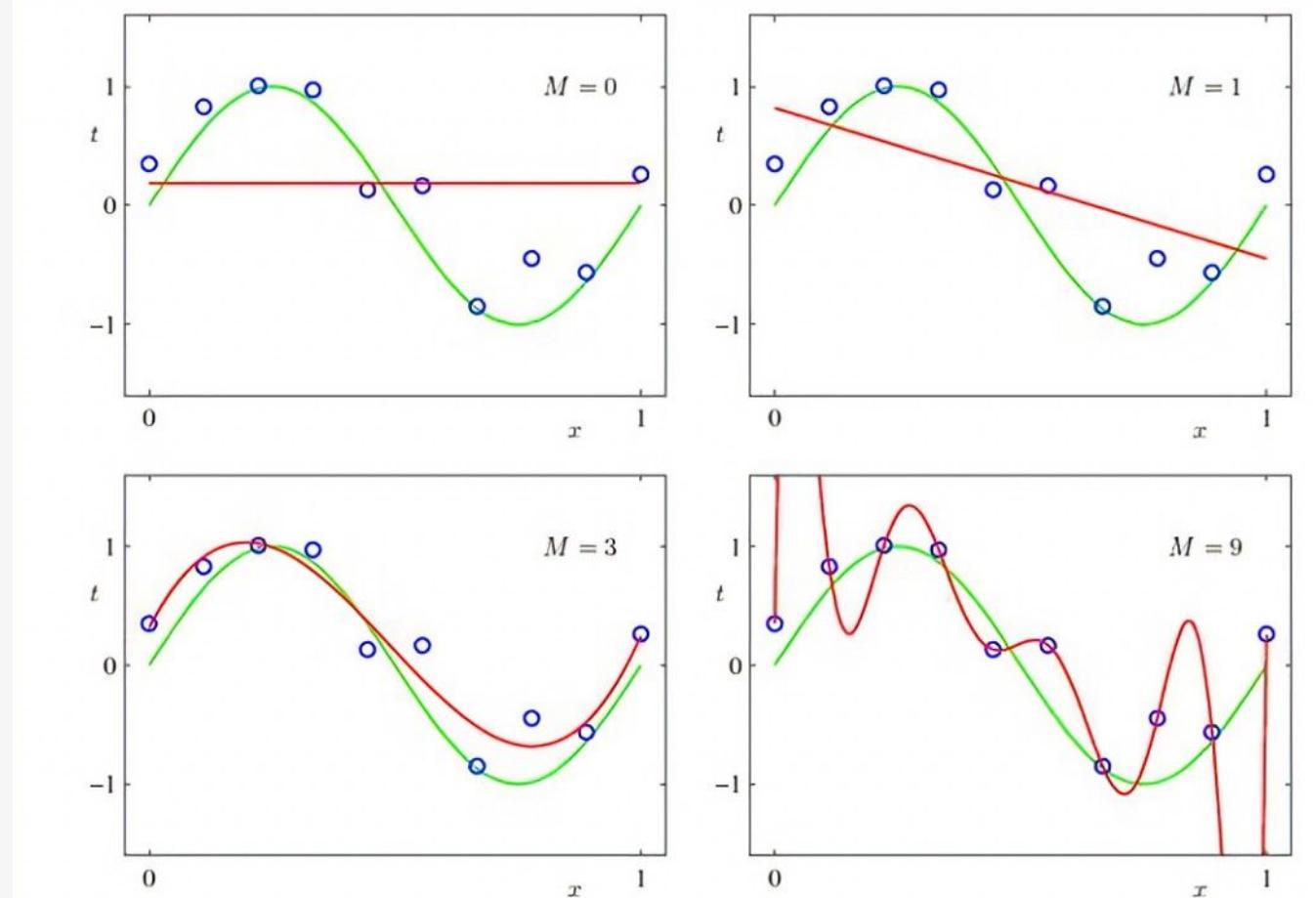
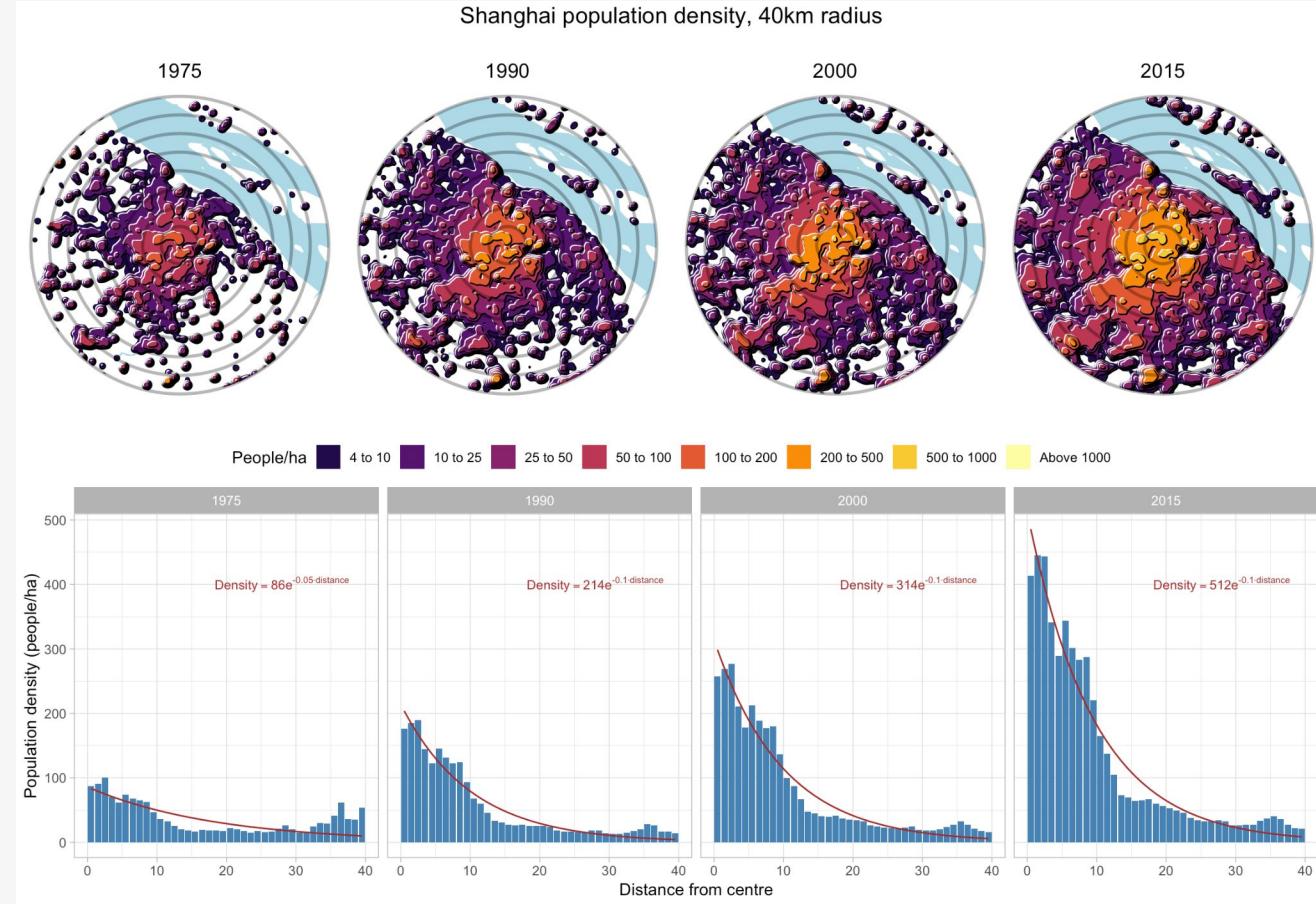
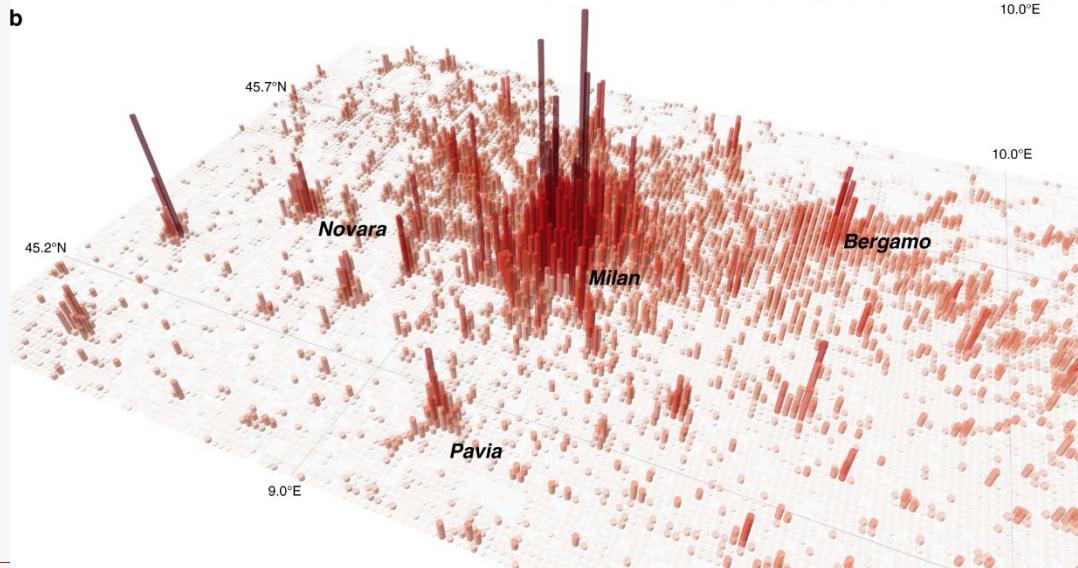
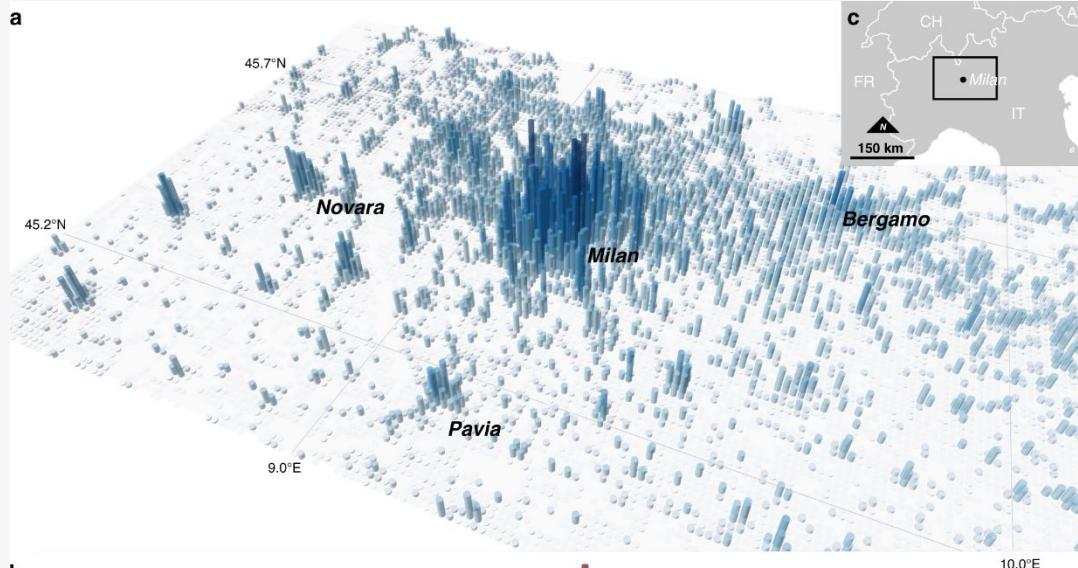


Figure 1.4 Plots of polynomials having various orders  $M$ , shown as red curves, fitted to the data set shown in Figure 1.2.

# 理解城市人口分布



上海交通大学  
SHANGHAI JIAO TONG UNIVERSITY



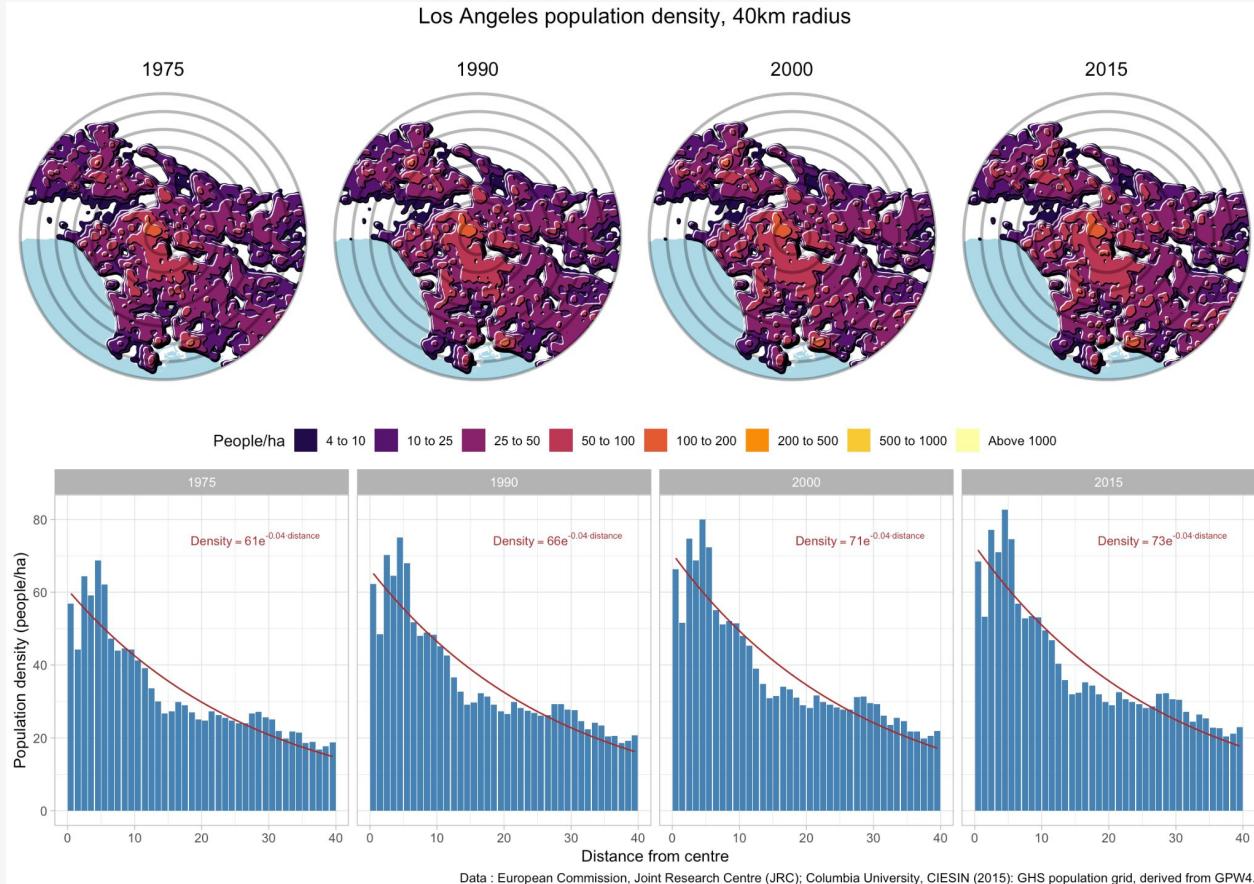
Data : European Commission, Joint Research Centre (JRC); Columbia University, CIESIN (2015): GHS population grid, derived from GPW4.



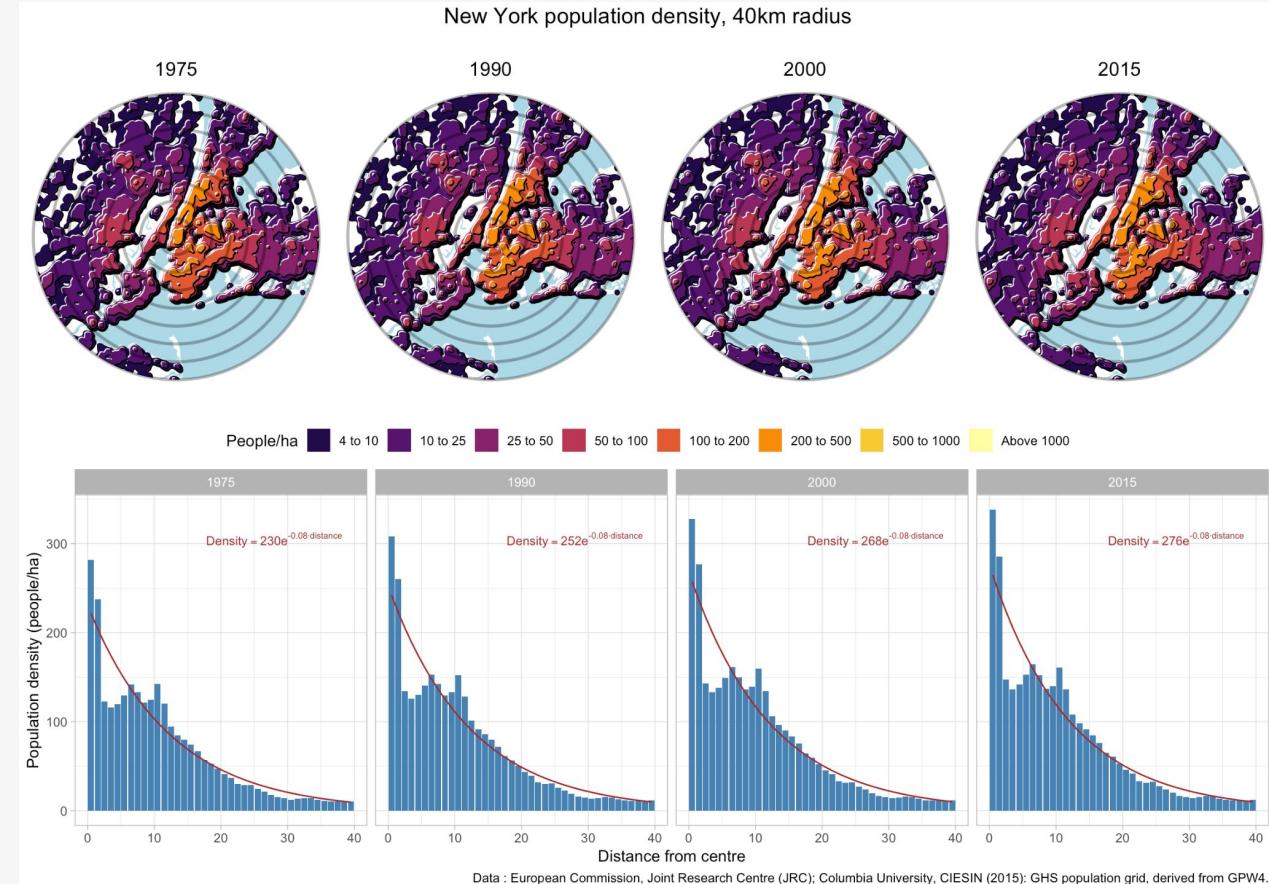
# 理解城市人口分布



上海交通大学  
SHANGHAI JIAO TONG UNIVERSITY



洛杉矶



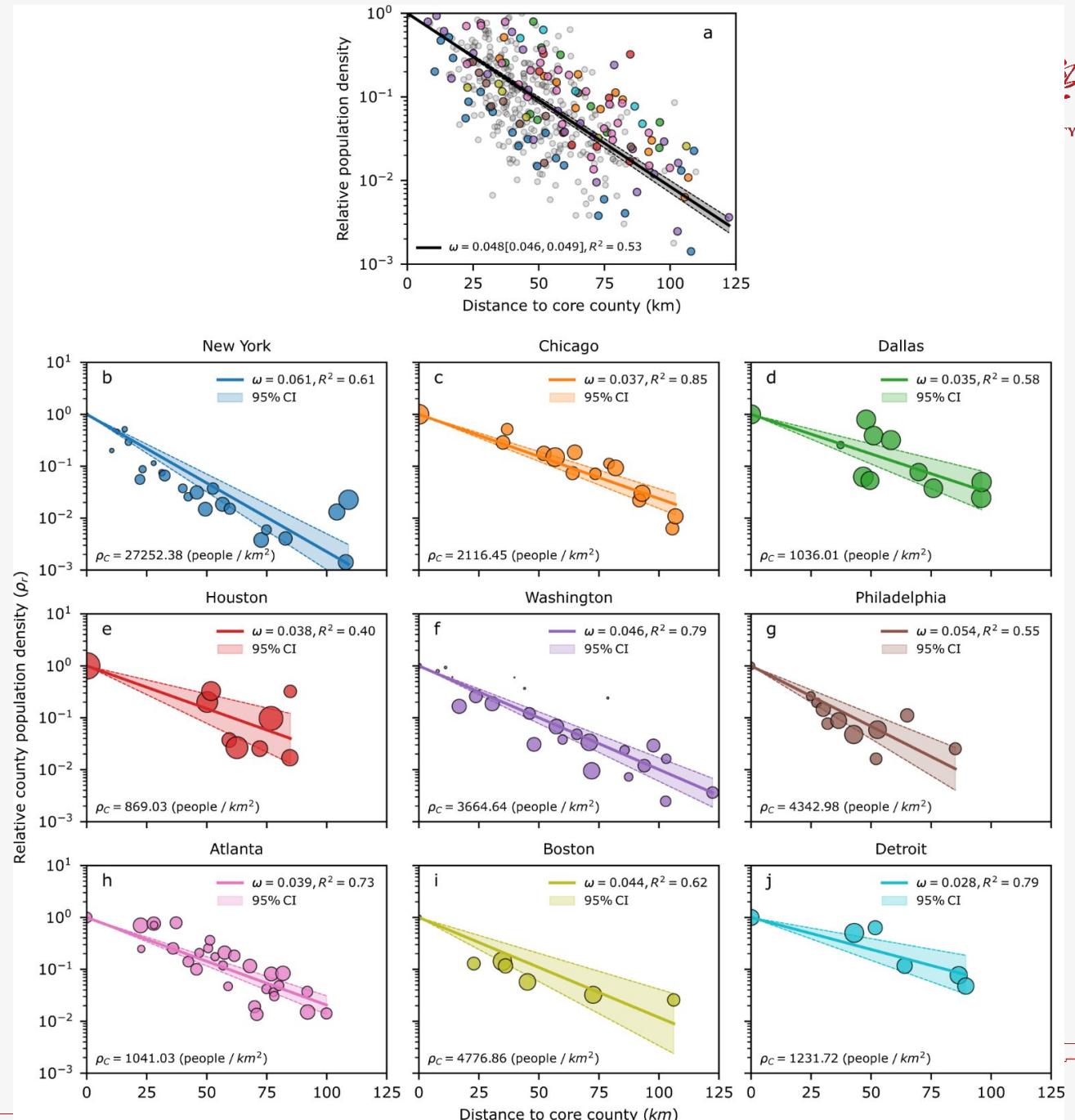
纽约



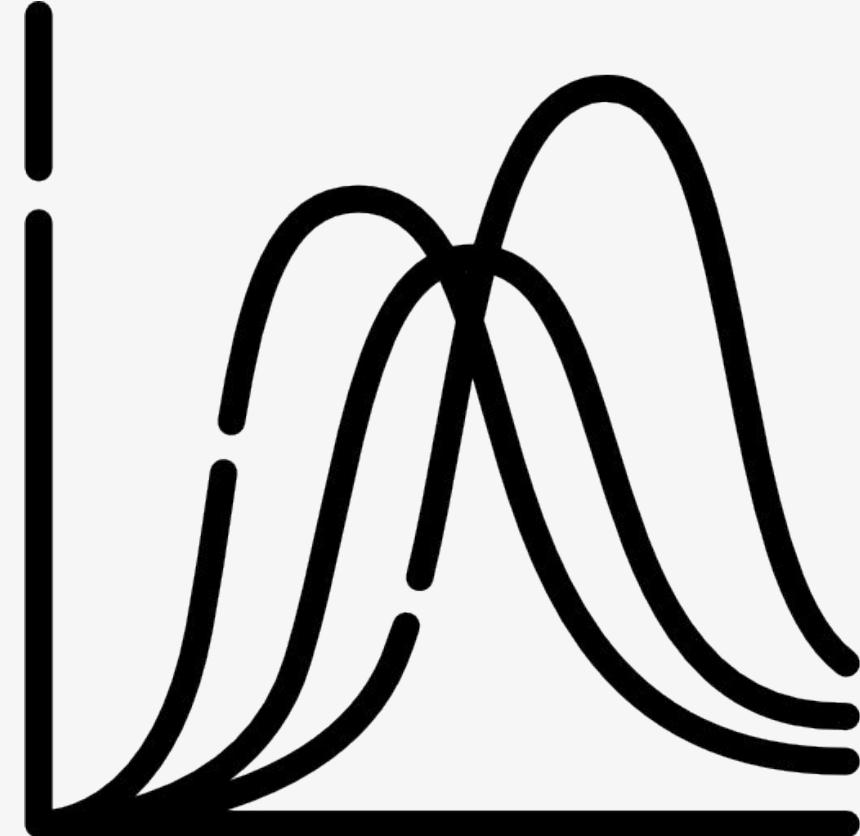
# 理解城市人口分布

- 右图表示每个county到城市中心的距离和county人口密度的关系；
- 注意这里y轴为对数坐标，x轴为线性坐标，因此此时拟合出的直线表示人口密度随着远离市中心而指数式下降。

<https://www.nature.com/articles/s42949-022-00075-9>



- 基本概率分布
- 量化数据分布
- 贝叶斯公式
- 函数拟合
- 数据相关性



# 独立、线性相关、相关系数

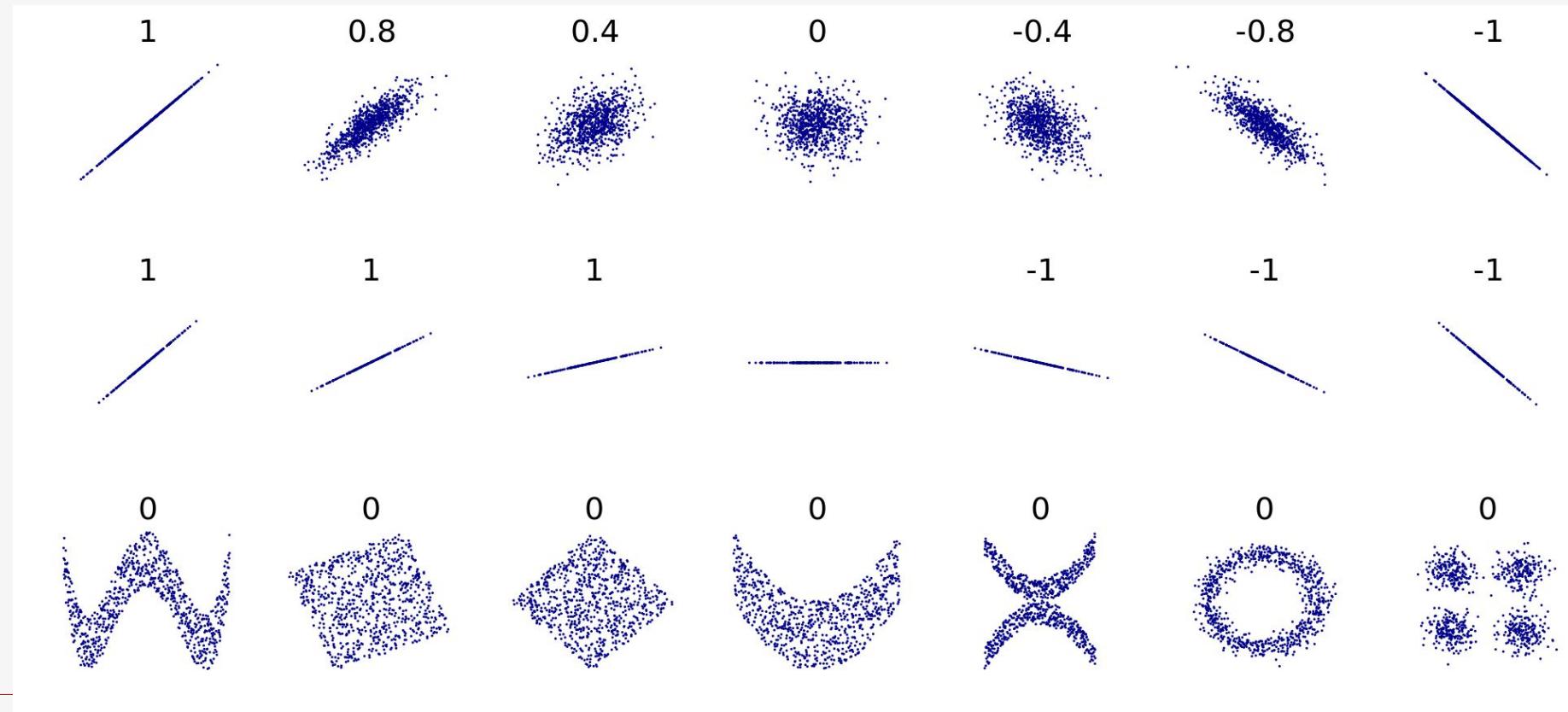


- 从概率的角度看, AB两个事件**独立**定义为 $P(A\text{发生}) * P(B\text{发生}) = P(AB\text{都发生})$
- 统计学中, 相关一般指线性相关, 可以采用相关系数来评估两个变量之间相关性的大小。对于连续变量, 一般采用Pearson相关系数。

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

# 独立、线性相关、相关系数

- 相关系数为正数, 说明两个变量的变化趋势相同; 相关系数为负则相反; 相关系数=0, 说明线性不相关。
- 独立则线性不相关, 但线性不相关不能推出独立。(第三行)

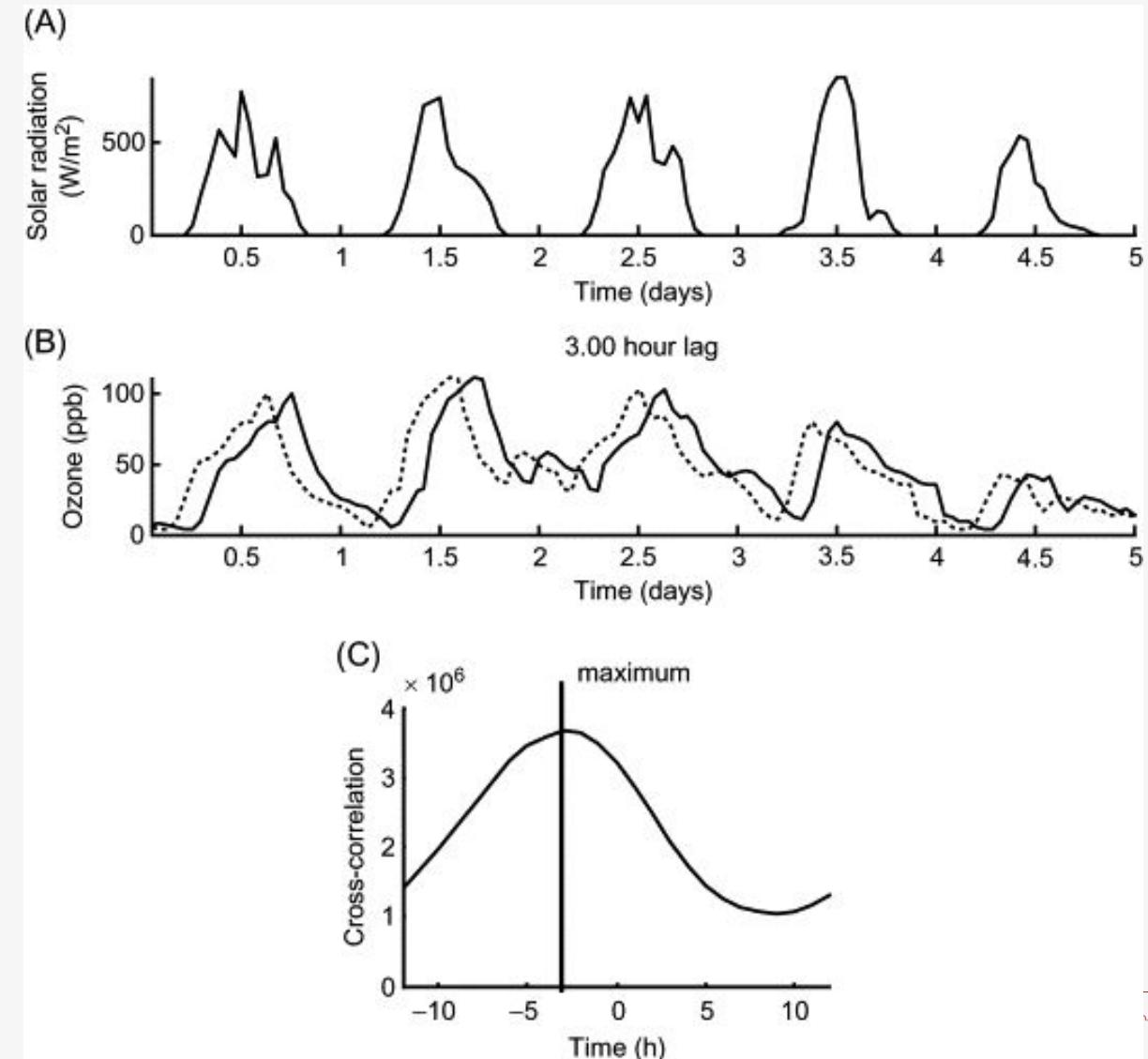


# time-lag correlation

- 观察右图中A图和B图的实线，相似但存在水平方向的错位。
- 将B图的实线向左偏移3个时间单位，形成虚线( $\tau=-3$ )
- 虚线和A图的相关系数很高
- 如何找到这个最佳的偏移量？这个偏移量说明了什么？

$$R_\tau = \text{corr}(X_t, Y_{t-\tau})$$

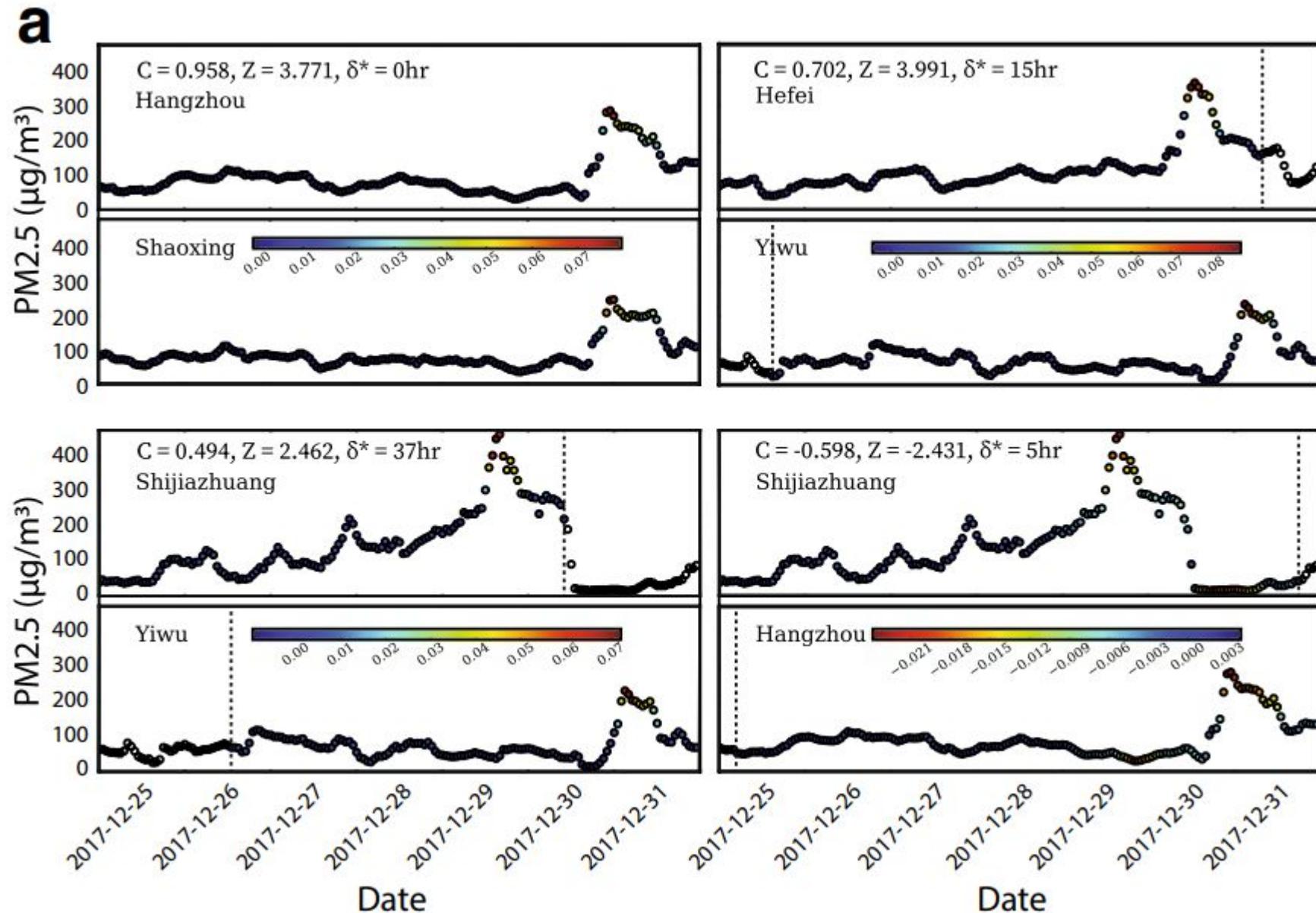
评价经过一段时间后产生的影响



# time-lag correlation



- 两地空气污染存在相似的曲线，但时间上存在延迟
- 通过时延相关系数，可以知道污染传播的速度。
- 延时相关是否表示因果？



# 数字也可能欺骗你

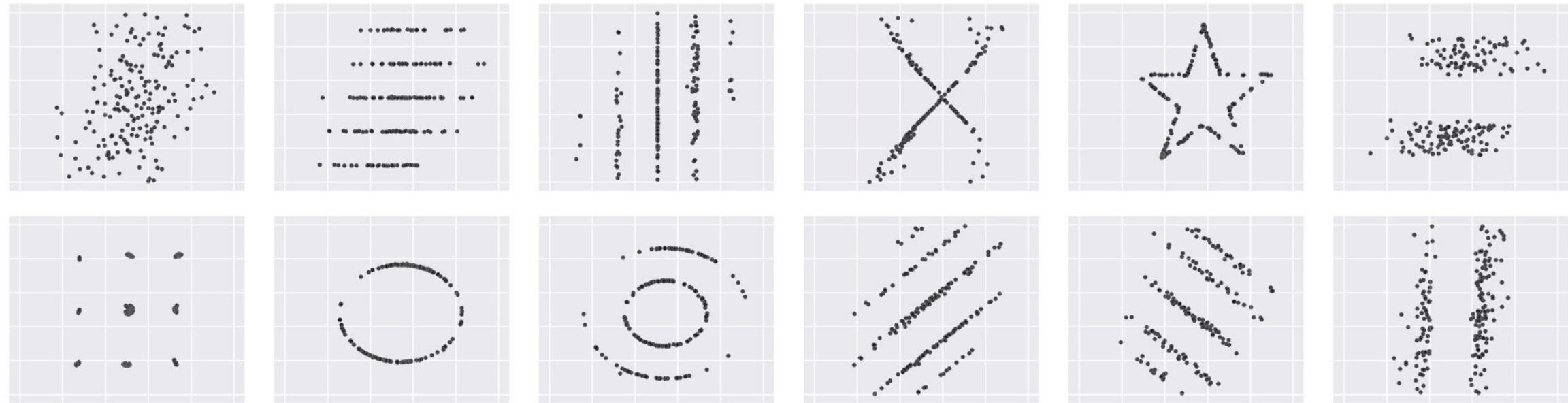


Figure 1. A collection of data sets produced by our technique. While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places. ( $\bar{x} = 54.02$ ,  $\bar{y} = 48.09$ ,  $sdx = 14.52$ ,  $sdy = 24.79$ , Pearson's  $r = +0.32$ )

以上数据集尽管看起来各不相同，他们都拥有相同的统计量(平均数(一阶统计量)、标准差(二阶统计量)、皮尔森相关系数)

Matejka J, Fitzmaurice G. Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing[C]//Proceedings of the 2017 CHI conference on human factors in computing systems. 2017: 1290-1294.

# 辛普森悖论

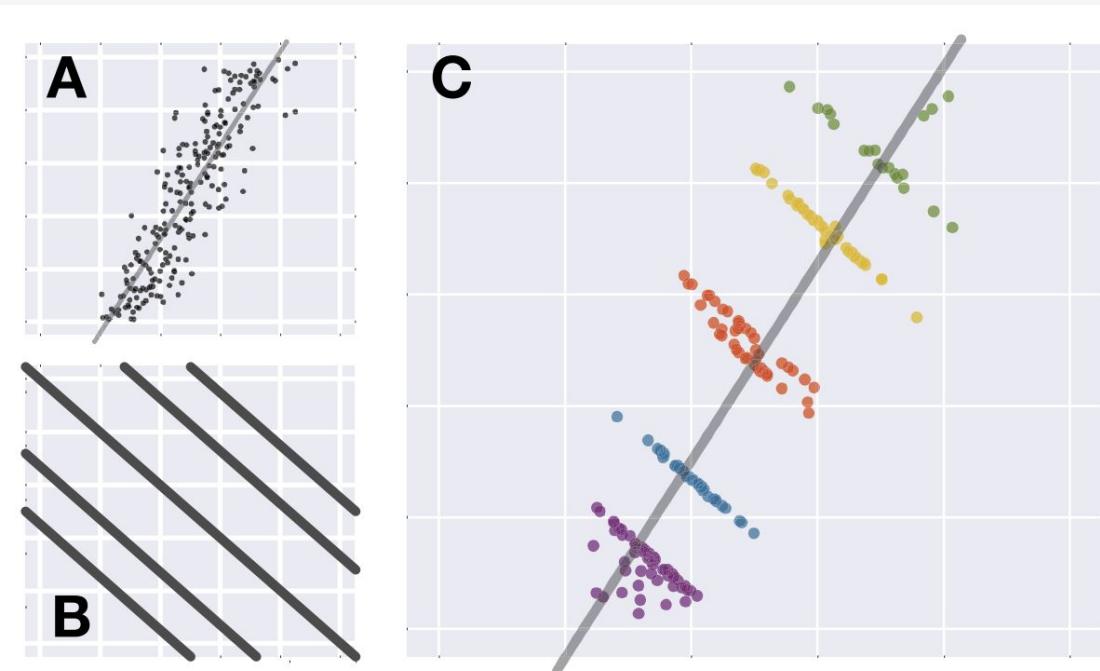


Figure 7. Demonstration of Simpson's Paradox. Both datasets (A and C) have the same overall Pearson's correlation of +0.81, however after coercing the data towards the pattern of sloping lines (B), each subset of data in (C) has an individually negative correlation.

A和C都有很高的正相关系数(+0.81)，但如果C图的数据引入了一个新的类别变量进行分类(按照图B)，每一类内横轴和纵轴居然形成了非常高的负相关！

医院	病情	入院总人数	死亡人数	存活人数	存活率
医院 A					
	合计	1000	100	900	90%
医院 B					
	合计	1000	200	800	80%

你会选择哪个医院？

# 作业



## 阅读论文：

Um, Jaegon, Seung-Woo Son, Sung-Ik Lee, Hawoong Jeong, and Beom Jun Kim. "Scaling laws between population and facility densities." *Proceedings of the National Academy of Sciences* 106, no. 34 (2009): 14236-14240.

## 分析数据：

在美国county(郡县)level上探讨人口分布与充电设施分布之间的关系。

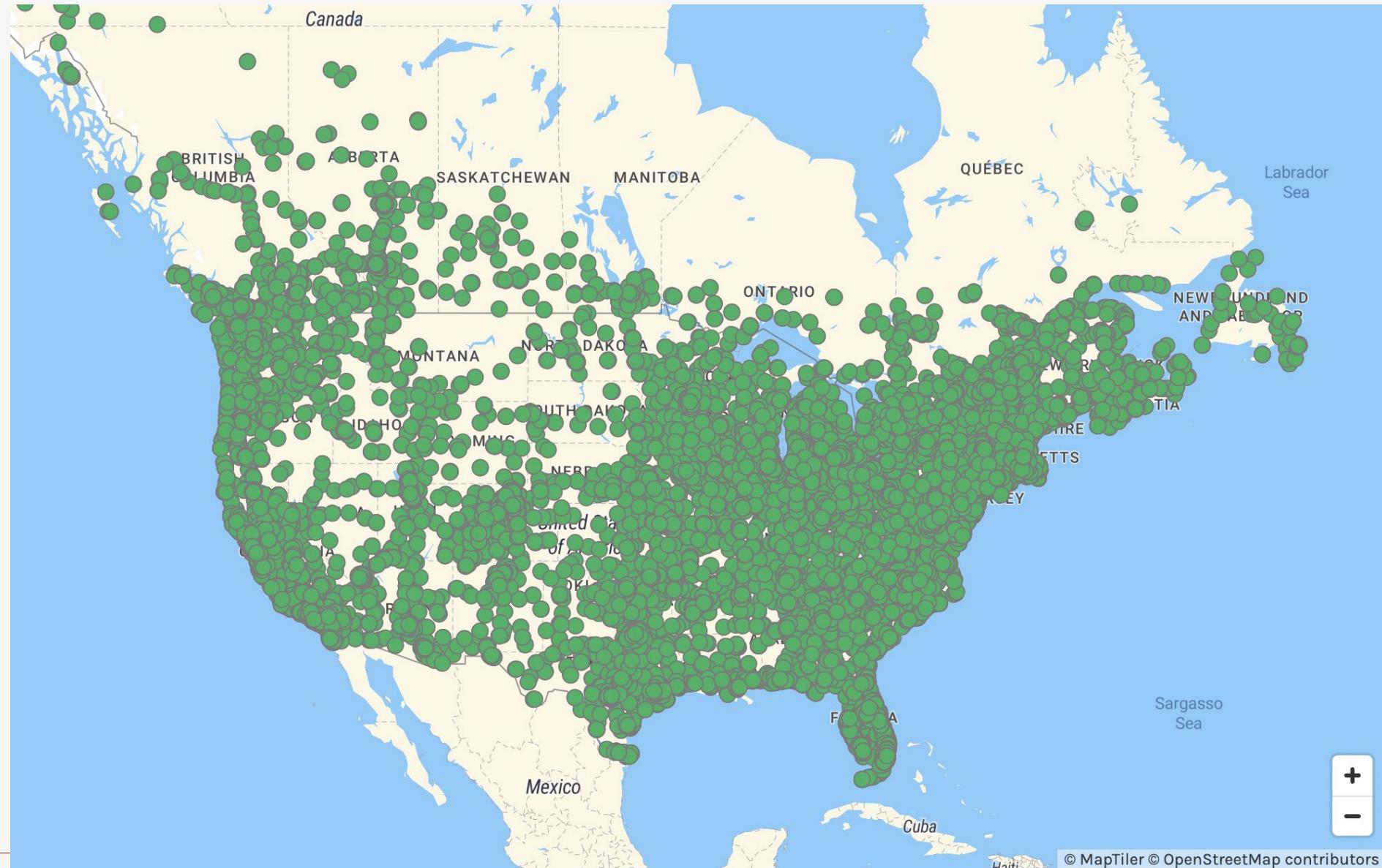
**提示:**是否存在统一规律？是否存在离群点？哪些因素影响了充电设施分布？充电设施分布是否存在不公平性？etc.



# 作业



上海交通大学  
SHANGHAI JIAO TONG UNIVERSITY



# 作业



## 已有数据：

- (1) County data: county ID, population, area, GDP, household income, education
- (2) Charging station: Lon, Lat, county ID. 来源

:[https://afdc.energy.gov/fuels/electricity\\_locations.html#/find/nearest?fuel=ELEC](https://afdc.energy.gov/fuels/electricity_locations.html#/find/nearest?fuel=ELEC)

## 其他数据资源举例：

<https://www.statsamerica.org/downloads/default.aspx>

<https://simplemaps.com/data/us-cities>



# 作业



## 提交形式：

Jupyter Notebook, 需里包含文字(中英文不限)、图片、代码等。

评分标准：以上提示全部探讨，得满分。有额外思考与数据分析，得奖励分(加入期末成绩)。

## 数据及作业提交入口：

<http://10.123.4.32:8887/tree>

在workspaces目录下，建立自己姓名(拼音)为名字的文件夹。





---

上海交通大学

SHANGHAI JIAO TONG UNIVERSITY