第2章 概率论回顾

中山大学人工智能学院
毛旭东

Email：maoxd3@mail.sysu.edu.cn

4.1 什么是概率（probability）？

- 随机事件发生的***可能性（possibilities）***是通过***概率（probability）***来衡量***不确定性（uncertainty）***的。
- 随机事件所有可能出现的结果，被称为样本空间（sample space）。

- 比如，随机事件X表示抛硬币，样本空间为{正面，反面}，$P(X = \text{正面}) = 0.5$表示结果为正面的概率是0.5。

- $P(X = x)$记为$p_X(x)$或$p(x)$。

参数表示

- 给定一组抛硬币的结果，我们想要推理的是这个硬币“正面的偏向性”和“反面的偏向性”。
  - 由于$p(\text{正面}) + p(\text{反面}) = 1$，因此只需推理其中一个。
- 很自然地，我们用参数θ来表示“正面的偏向性”。
  - θ本身就表示$p(\text{正面})$，即$p(\text{正面}) = \theta$。
  - 比如，$\theta = 0.5$表示：正面的偏向性是0.5。
  - $p(\theta = 0.5) = 0.9$表示：正面的偏向性是0.5的概率是0.9。
- 我们需要推理的就是θ所有的可能取值的概率值，也就是θ的概率分布。
  - θ的样本空间是[0,1]。

抛硬币

- 我们后面会经常用抛硬币这个例子。

- 抛硬币代表的是一类随机事件（样本空间为2种）：
  - 做完心脏手术后，病人存活超过一年和未超过一年的概率；
  - 服用某种新药后，引发头痛和不引发头痛的概率；
  - 1场2个候选人的选举，候选人A获胜和候选人B获胜的概率；
  - 研究大脑偏侧性，左偏侧性和右偏侧性的概率；
  - 。。。

4.3 概率分布（probability distribution）

- 概率分布是随机事件的所有结果和相应的概率值的集合。

- 比如，抛硬币的概率分布是：正面或者反面，以及他们对应的概率值θ和$1-\theta$。
  - 离散分布

- 随机事件的结果也有可能是连续值，比如身高、降雨量等。
  - 连续分布

4.3.1 离散分布（discrete distribution）

- 当样本空间的可能结果是离散值时，我们称其为离散分布。
  - 离散分布的取值可能是无穷多个，比如泊松分布。

- 离散分布可能结果的概率值，称为***概率质量（probability mass）***。

- 离散分布所有可能结果的概率质量的总和是1：

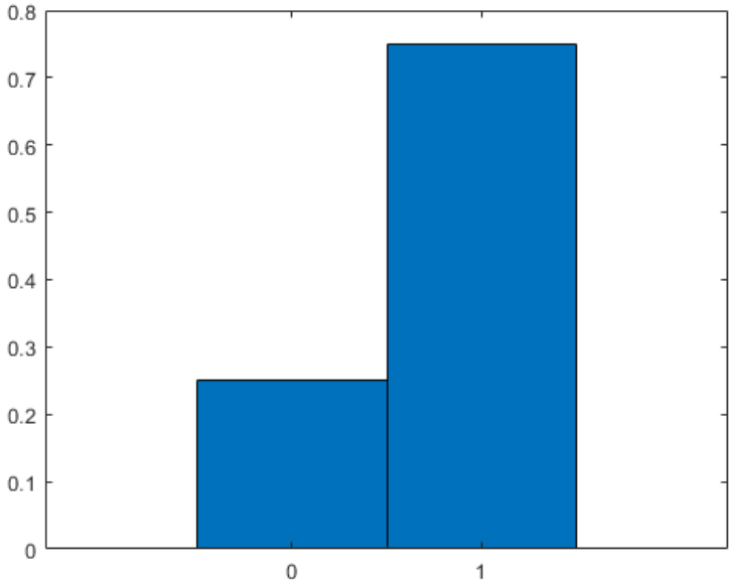
$$\sum_x p(x) = 1$$

伯努利分布（Bernoulli Distribution）

- 随机事件Y的的样本空间为$\{Y=1, Y=0\}$，用参数θ表示$Y=1$的概率，即$P(Y=1)=\theta$，$P(Y=0)=1-\theta$。
- 可以将上式合并：

$$p(y|\theta)=\theta^{y}(1-\theta)^{1-y}$$

- 该分布称为**伯努利分布**。



连续分布（continuous distribution）

- 当样本空间的可能结果是连续值时，我们称其为连续分布。

- 对于连续分布的1个点，我们用***概率密度（probability density）***来表示，相当于是概率质量和区间长度的比值。

连续分布

- 任意1个点的概率质量为0。
- 将连续值分成互斥且完备的区间，区间的概率质量和为1：

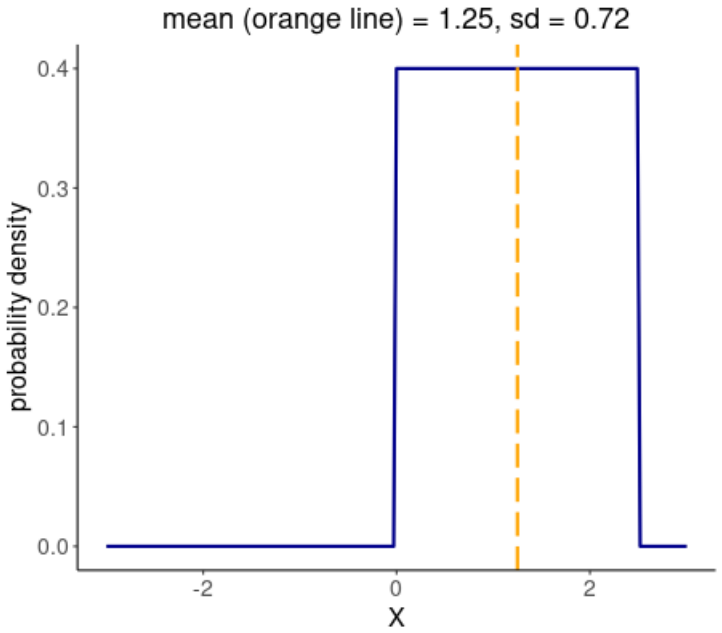
$$\int p(x)dx = 1$$

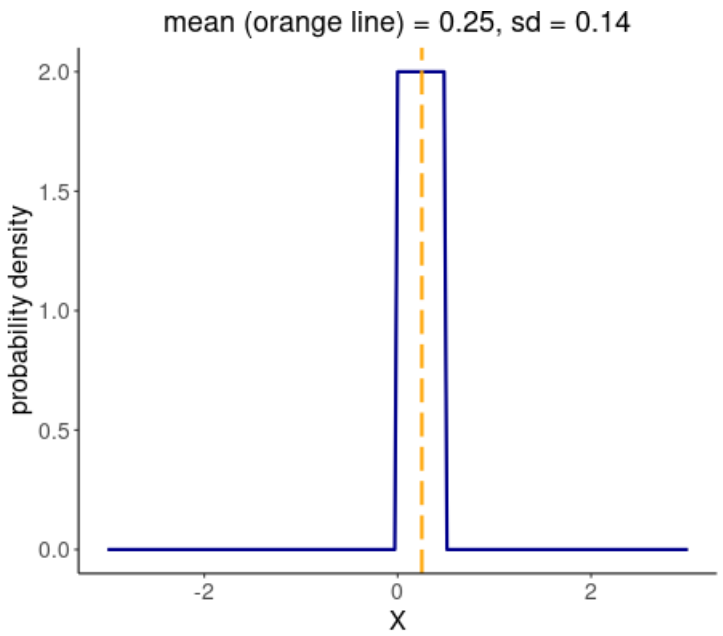
- 区间[a,b]之间的概率质量：

$$P(a \leq X \leq b) = \int_a^b p(x)dx$$

连续分布

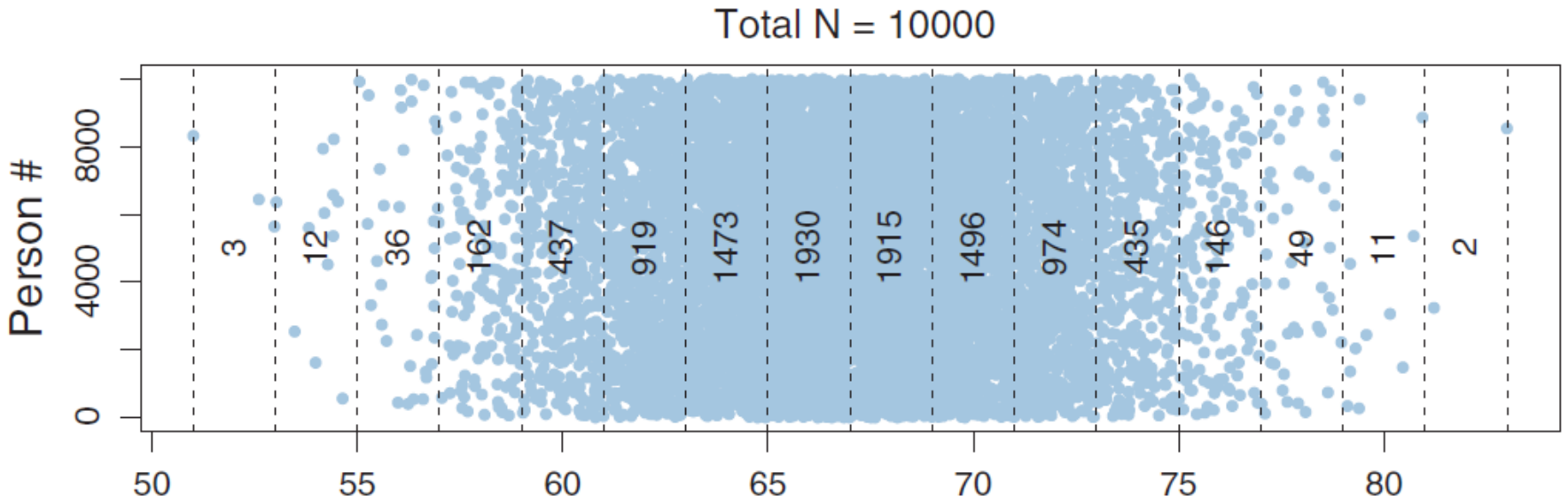
- 概率密度值可能大于1。





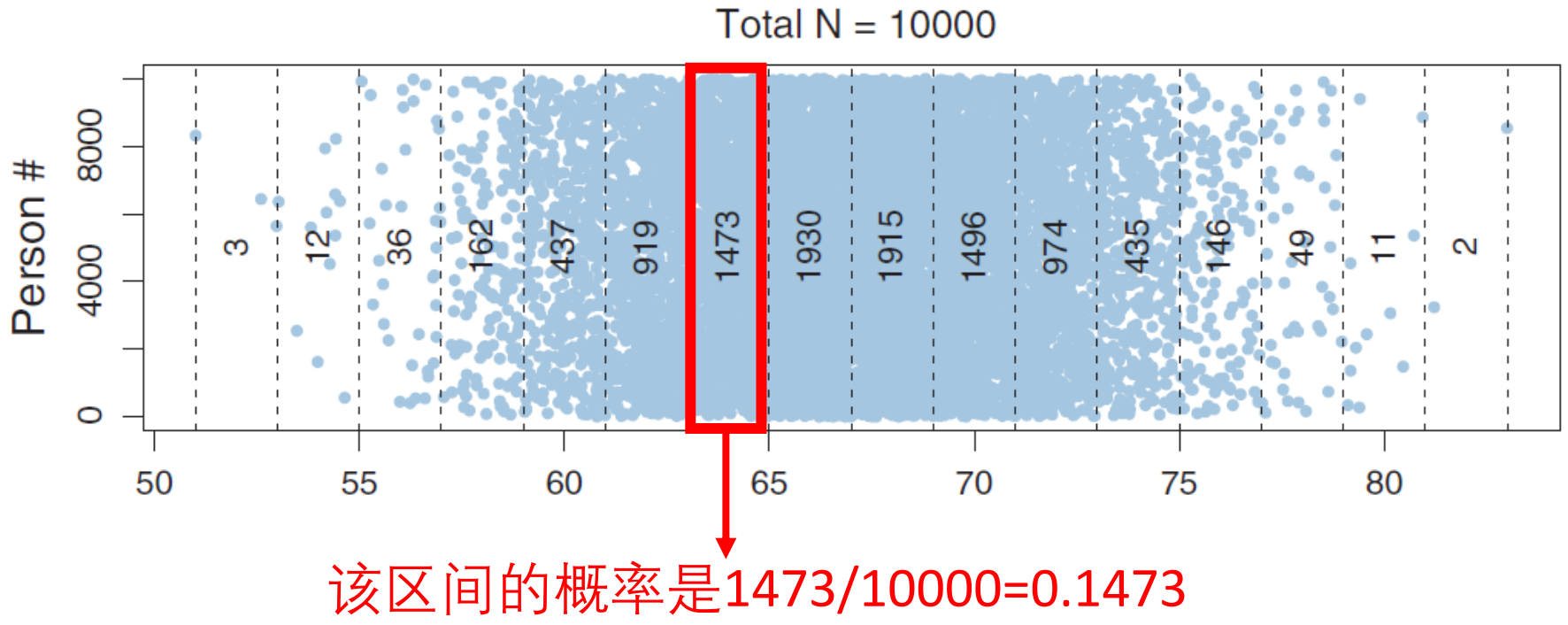
离散化

- 对于可能结果是连续值的，我们也可以对齐进行离散化（discretize），将其转换为离散分布。
- 比如，身高是连续值，我们可以将身高的取值切分成几个区间：



离散化

- 比如，身高是连续值，我们可以将身高的取值切分成几个区间：

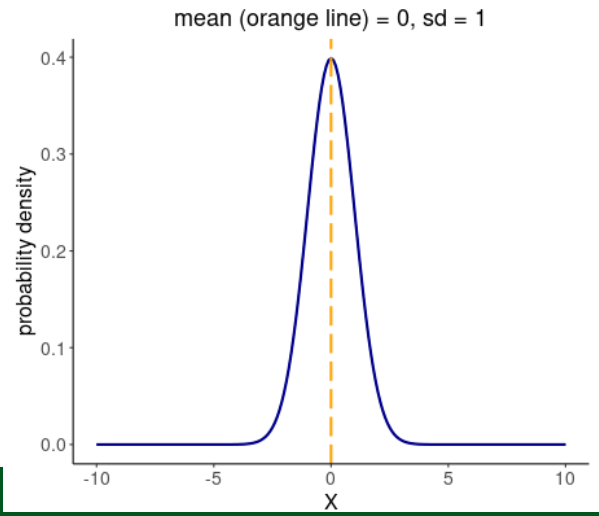


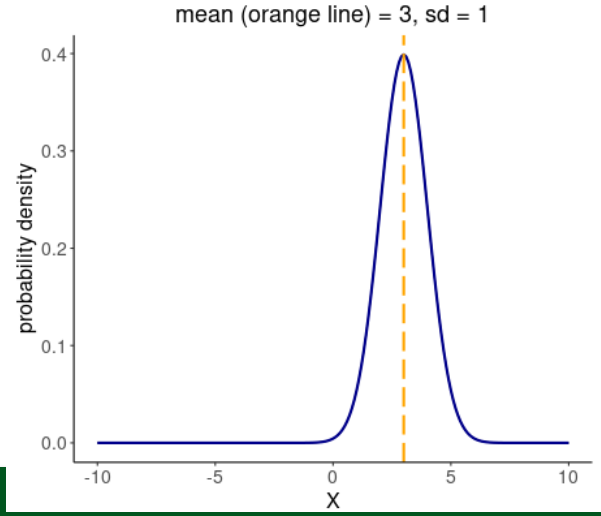
4.3.2.2 正态分布（Normal Distribution）

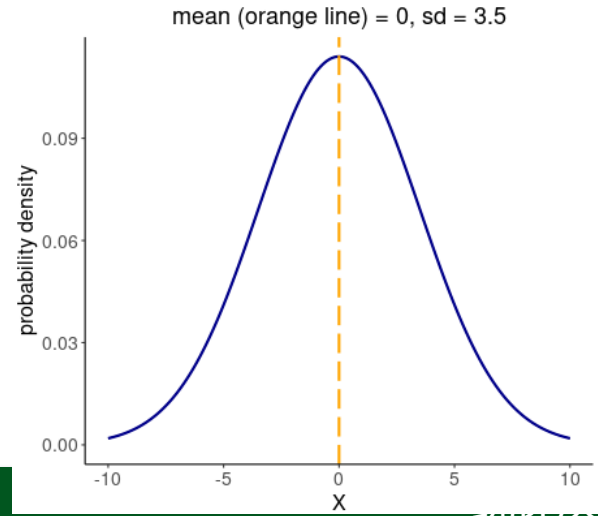
- 概率密度函数为：

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

- 参数μ是均值，被称为位置参数。
- 参数σ是标准差，被称为尺度参数。







4.3.3 均值和方差

- ***均值（mean）***，又被称为***期望值（expected value）***，定义如下：

$$E[x] = \sum_x p(x)x$$

- 比如1个6面的骰子，每一个面出现的概率相同，则均值为：$\left(\frac{1}{6}\right) * 1 + \left(\frac{1}{6}\right) * 2 + \left(\frac{1}{6}\right) * 3 + \left(\frac{1}{6}\right) * 4 + \left(\frac{1}{6}\right) * 5 + \left(\frac{1}{6}\right) * 6 = 3.5$。

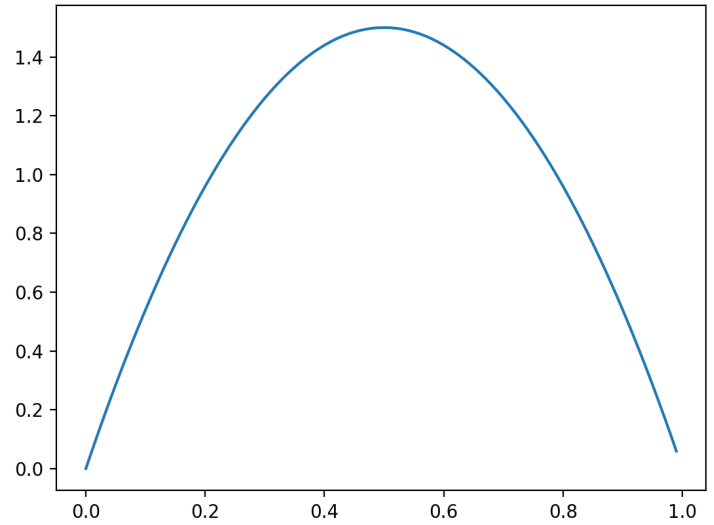
- 连续值的均值：

$$E[x] = \int p(x)x \; dx$$

均值的例子

- 概率密度函数为$p(x) = 6x(1-x), x \in [0,1]$，求均值。

$$
\begin{aligned}
E[x] &= \int \mathrm{d}x\, p(x)\, x \\
&= \int_0^1 \mathrm{d}x\ 6x(1-x)\, x \\
&= 6\int_0^1 \mathrm{d}x\ \left(x^2 - x^3\right) \\
&= 6\left[\frac{1}{3}x^3 - \frac{1}{4}x^4\right]_0^1 \\
&= 6\left[\left(\frac{1}{3}1^3 - \frac{1}{4}1^4\right) - \left(\frac{1}{3}0^3 - \frac{1}{4}0^4\right)\right] \\
&= 0.5
\end{aligned}
$$



方差

- *方差（variance）*的定义如下：

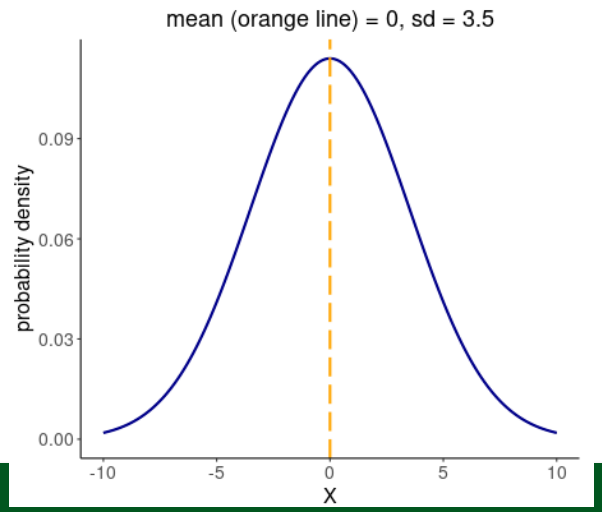
$$\mathrm{var}_x = \int p(x)(x - E[x])^2 dx$$

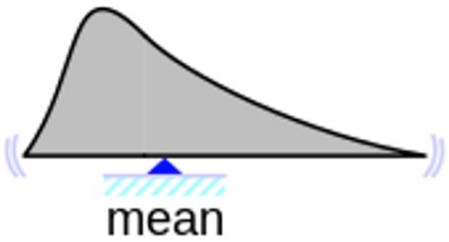
- 方差就是x和均值之间的平方误差的均值，即均方误差（mean squared deviation，MSD）。
- 方差衡量的是概率分布对其均值的离散度。
  - 方差越大，越分散。方差越小，越集中。
- 方差的平方根是概率分布的标准差。

集中趋势（central tendency）：均值

- 集中趋势表示概率分布的中心位置或者代表值，包括**均值、中位数、众数**等。
- 均值（mean）：

$$E[x] = \int p(x)x \, dx$$





集中趋势：众数（mode）

- 离散分布：

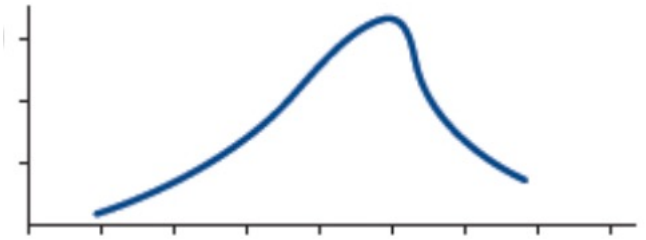
$$\underset{M}{\operatorname{argmax}}\, p(x = M)$$

- 连续分布：

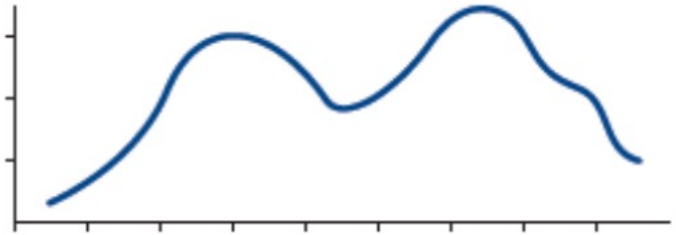
$$p'(M) = 0 \quad and \quad p''(M) < 0$$

在M处达到局部最大值。

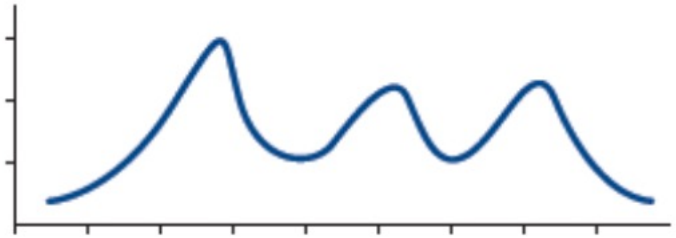
- mode有时称为峰值，但不严谨。



“单峰”分布（unimodal）



“双峰”分布（bimodal）

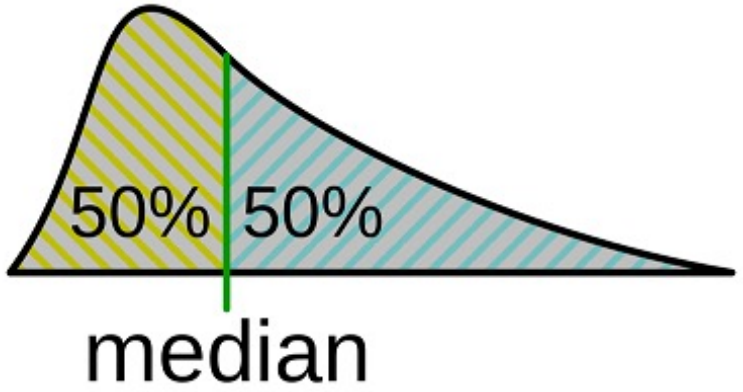


“多峰”分布（multimodal）

集中趋势：中位数（median）

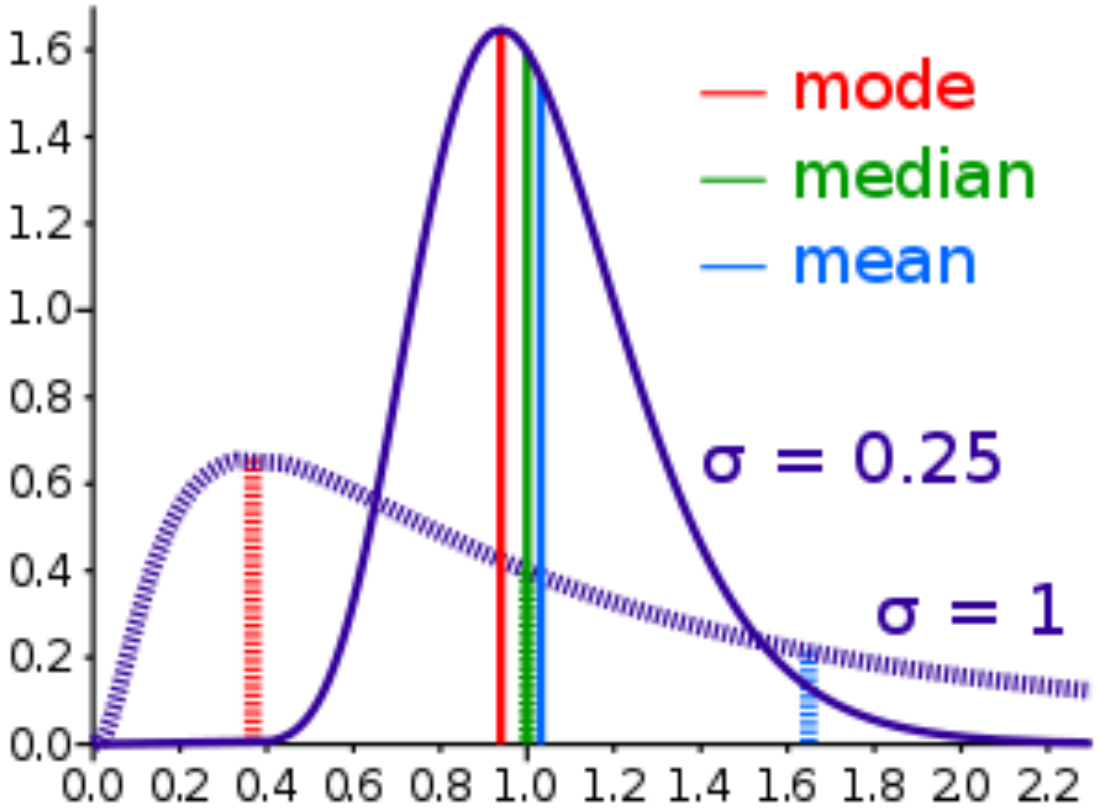
- 中位数（median）：

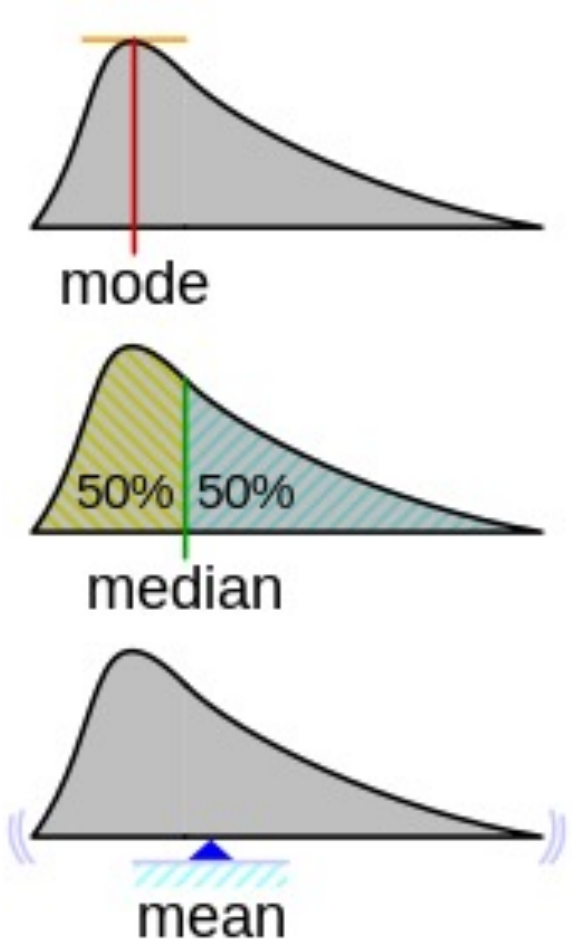
$$P(x \leq M) = P(x \geq M) = \frac{1}{2}$$



- 当概率分布是对称的，中位数=均值。
- 当概率分布是对称且单峰的，中位数=均值=众数。

集中趋势（central tendency）

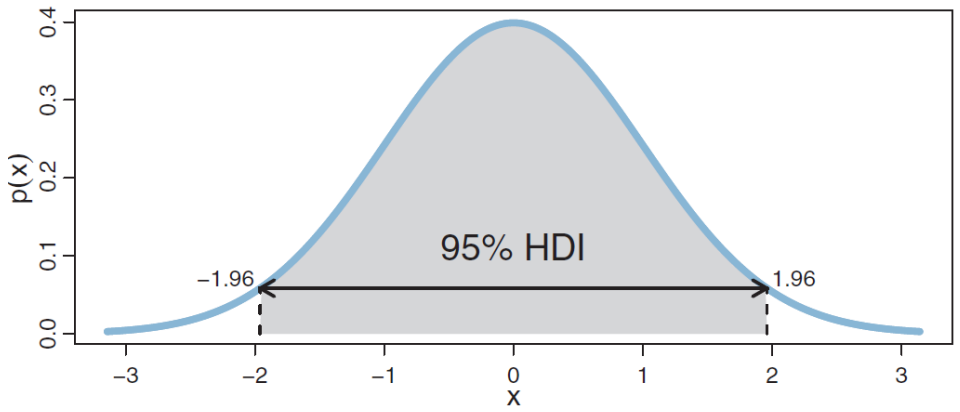


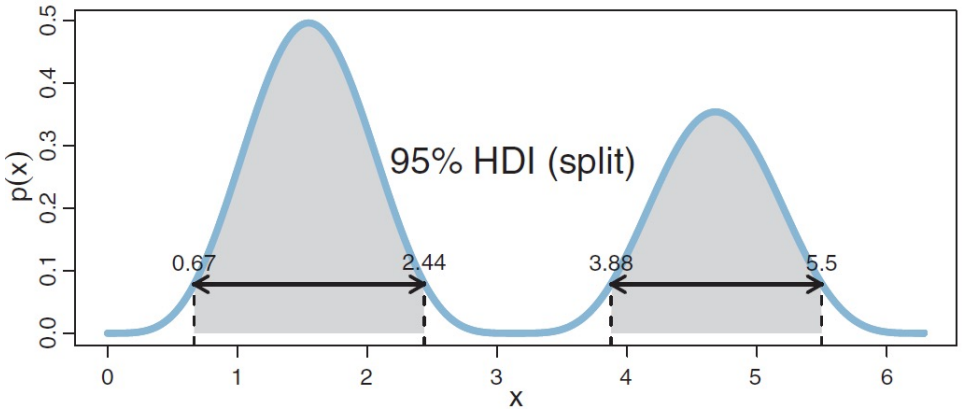


4.3.4 最高密度区间（Highest Density Interval，HDI）

- 最高密度区间指的是占据95%概率质量值的区间，且该区间内的点的概率密度大于其他点的概率密度。

$$\int_{x:p(x)>W} p(x)dx = 0.95$$





240

Discussion [D]What are some "important" problems in machine learning/AI? (self.MachineLearning)

Netero1999 於 17 天前 發表

I am not talking about "hot stuff" like self driving cars or anything, but topics important to the field(like maybe interpretability of machine learning?) which is fundamental to the advancement of the field.

146 留言 分享 儲存 隱藏 give award 檢舉 crosspost

[–] **lorepieri** 242 指標 17 天前

Estimating the uncertainty and confidence interval in AI/ML predictions.

永久連結 embed 儲存 檢舉 give award 回覆

[–] **quadprog** 111 指標 17 天前

Commenting because this answer is short so people might overlook it. This is a huge, important problem. Simpler ML models were good at providing uncertainty estimates. Now deep networks do not have analytical solutions and the obvious technique (MCMC) is extremely computationally expensive. Anyone who figures out a reliable and inexpensive way to get uncertainty estimates from deep NNs will be a hero of ML.

永久連結 embed 儲存 上層留言 檢舉 give award 回覆

4.4 双因素分布（two-way distribution）

- 两个随机事件相结合的概率分布。

- 例子：

Table 4.1 Proportions of combinations of hair color and eye color

	Hair color			
Eye color	Black	Brunette	Red	Blond
Brown	0.11	0.20	0.04	0.01
Blue	0.03	0.14	0.03	0.16
Hazel	0.03	0.09	0.02	0.02
Green	0.01	0.05	0.02	0.03

联合概率（joint probability）

Table 4.1 Proportions of combinations of hair color and eye color

	Hair color			
Eye color	Black	Brunette	Red	Blond
Brown	0.11	0.20	0.04	0.01
Blue	0.03	0.14	0.03	0.16
Hazel	0.03	0.09	0.02	0.02
Green	0.01	0.05	0.02	0.03

- 两个随机事件同时发生的概率。比如眼睛颜色是e，头发颜色是h的概率是：

$$p(e,h) \text{ 或者 } p(h,e)$$

- $\sum_e \sum_h p(e,h) = 1$

边缘概率（marginal probability）

Table 4.1 Proportions of combinations of hair color and eye color

	Hair color			
Eye color	Black	Brunette	Red	Blond
Brown	0.11	0.20	0.04	0.01
Blue	0.03	0.14	0.03	0.16
Hazel	0.03	0.09	0.02	0.02
Green	0.01	0.05	0.02	0.03

- 绿眼睛的人的概率是多少？

$$0.01 + 0.05 + 0.02 + 0.03 = 0.11$$

边缘概率（marginal probability）

Table 4.1 Proportions of combinations of hair color and eye color

Eye color	Hair color				Marginal (eye color)
	Black	Brunette	Red	Blond	
Brown	0.11	0.20	0.04	0.01	0.37
Blue	0.03	0.14	0.03	0.16	0.36
Hazel	0.03	0.09	0.02	0.02	0.16
Green	0.01	0.05	0.02	0.03	0.11
Marginal (hair color)	0.18	0.48	0.12	0.21	1.0

- 从一个联合分布，得到随机事件子集的概率分布，称为边缘概率分布。

$$p(e) = \sum_h p(e,h) \text{ 或者 } \int p(e,h)dh$$

条件概率（conditional probability）

Table 4.1 Proportions of combinations of hair color and eye color

Eye color	Hair color				Marginal (eye color)
	Black	Brunette	Red	Blond	
Brown	0.11	0.20	0.04	0.01	0.37
Blue	0.03	0.14	0.03	0.16	0.36
Hazel	0.03	0.09	0.02	0.02	0.16
Green	0.01	0.05	0.02	0.03	0.11
Marginal (hair color)	0.18	0.48	0.12	0.21	1.0

- 蓝眼睛的人里面，金色头发的概率是多少？

$$\frac{0.16}{0.36} = 0.45$$

条件概率（conditional probability）

- 一个随机事件已经发生的情况下（$Y = y$），另一个随机事件发生的概率（$X = x$），称为条件概率：

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

$$= \frac{p(x,y)}{\sum_{x^*} p(x^*, y)} \text{ 或者 } \frac{p(x,y)}{\int p(x,y)dx}$$

- 另一种理解：在$Y = y$的情况下，$X = x$所占的比例。
- 可得**链式法则**：$p(x,y) = p(x|y)p(y)$
- $p(x,y,z) = p(x|y,z)p(y|z)p(z)$

独立性（Independence）

- 当对任意x和y，满足 $p(x|y) = p(x)$时， 我们称x和y是独立的，记为$x \perp y$。
  - y的取值，不影响x的概率。

- 由$p(x, y) = p(x|y)p(y)$，可得：

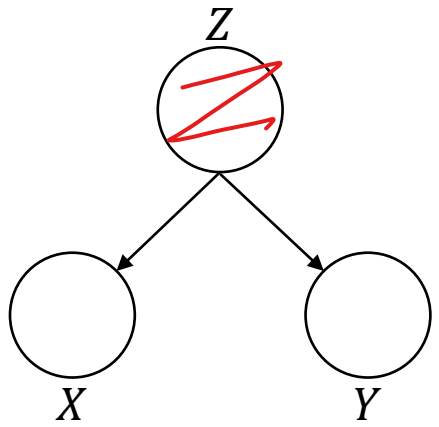
$$p(x, y) = p(x)p(y)$$

条件独立性（conditional independence）

- 给定z，当对任意x和y，满足 $p(x|y,z) = p(x|z)$时， 我们称x和y是条件独立的，记为$x \perp y|z$。
- 等价于$p(x,y|z) = p(x|z)p(y|z)$
- 例子：事件X和Y分别是张三和李四是否准时回家吃晚饭，事件Z是是否有暴雨。
  - 是否有暴雨未知的情况下，“张三没有准时回家吃饭” 可能是由于“有暴雨”，因此，“李四不能准时回家吃饭”的概率增加了。=> 事件X和Y不独立。
  - 已知有暴雨的情况下，“张三没有准时回家吃饭”并不能帮助判断“李四是否准时回家吃饭”，因此$X \perp Y|Z$。

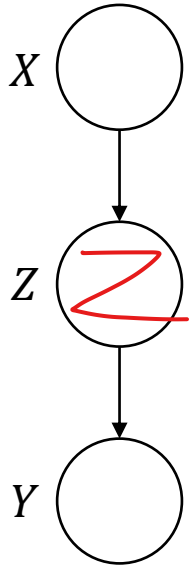
条件独立性

- 右图是常见的一种条件独立。
  - 暴雨是Z。
  - 张三是否准时回家吃饭是X。
  - 李四是否准时回家吃饭是Y。
- 当X和Y都依赖于Z，X和Y本身是不独立的，X和Y会通过Z互相影响。
- 当Z给定时，X和Y是条件独立的，此时不能通过Z互相影响。



条件独立性

- 右图是常见的另一种条件独立。
- 给定Z时， X和Y是条件独立的。



- 比如X：是否有乌云， Z：是否下暴雨，Y：是否准时回家。

5.1 贝叶斯法则（Bayes’ Rule）

- 条件概率的定义：

$$p(x|y) = \frac{p(x,y)}{p(y)} \qquad p(y|x) = \frac{p(x,y)}{p(x)}$$

- 由此可得：

$$p(x|y)p(y) = p(y|x)p(x)$$

- 可得贝叶斯法则：

$$\boldsymbol{p(x|y) = \frac{p(y|x)p(x)}{p(y)}}$$

贝叶斯法则

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

贝叶斯法则

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

$$= \frac{p(y|x)p(x)}{\sum_{x^*} p(y, x^*)} \text{或者} \frac{p(y|x)p(x)}{\int p(y, x)dx}$$

贝叶斯法则

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

$$= \frac{p(y|x)p(x)}{\sum_{x^*} p(y, x^*)} \text{或者} \frac{p(y|x)p(x)}{\int p(y, x)dx}$$

$$= \frac{p(y|x)p(x)}{\sum_{x^*} p(y|x^*)p(x^*)} \text{或者} \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}$$

- 当x是参数θ，y是数据D时：

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

例子：抗原检测

- 国家卫生健康委公布的抗原检测试剂的准确率在75%-98%，假阳性在1%-5%之间。
- 我们假设准确率在87%，假阳性在3%，先验患新冠的概率是0.1%，某人的抗原检测是阳性，计算患病的概率。

- 用$\theta = 1$表示某人患病，$\theta = 0$表示某人不患病。
- 用$T = 1$表示抗原是阳性，$T = 0$表示抗原是阴性。
- 求$p(\theta = 1|T = 1)$。

(数据来源：https://finance.sina.com.cn/jjxw/2022-03-21/doc-imcwiwss7188854.shtml)

例子：抗原检测

$$p(\theta = 1|T = 1)$$

$$= \frac{p(T = 1|\theta = 1)p(\theta = 1)}{p(T = 1|\theta = 1)p(\theta = 1) + p(T = 1|\theta = 0)p(\theta = 0)}$$

$$= \frac{0.87 * 0.001}{0.87 * 0.001 + 0.03 * 0.999}$$

$$= 0.028$$

- 即使抗原阳性，患病的概率很低。
- 解决办法：做多次抗原检测。

例子：抗原检测

- 如果某人住的楼栋有了确诊，我们假设该楼栋患新冠的先验概率是0.05。

$$\frac{p(T=1|\theta=1)p(\theta=1)}{p(T=1|\theta=1)p(\theta=1)+p(T=1|\theta=0)p(\theta=0)}$$

$$=\frac{0.87*0.05}{0.87*0.05+0.03*0.95}$$

$$=0.6$$

- 此时，抗原阳性，患病的概率很高。

贝叶斯法则用于模型参数和数据

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- $p(\theta)$：先验（prior）
- $p(\theta|D)$：后验（posterior）
- $p(D|\theta)$：似然（likelihood）
- $p(D)$：证据（evidence），也称为边缘似然（marginal likelihood）
  - $\sum_{\theta^*} p(D|\theta^*)\, p(\theta^*)$ 或者 $\int p(D|\theta)\, p(\theta)\, d\theta$