高中统计学·第叁课

矩的定义与计算

何濯羽 2024年2月12日

目录

CONTENT

前言

在第贰课中，我们已经学习三种可以“完美”描述随机变量的工具：

- 概率质量/密度函数
- 累积分布函数
- 分位数函数

可是，有时候我们并不想获得关于某个随机变量的全部信息，而只对它的部分特征感兴趣。这种情况下，我们会依赖于一些带有总结性质的统计工具（例如小学、初中学习的平均数、方差等）。我们把这类工具统称为“**矩**”（moment）。一种矩只能反映随机变量的一个特征（trait）。

注：在本节课中，我们只学习离散型随机变量的矩的定义和计算。这是因为连续型随机变量的矩的定义与计算涉及积分（integral calculus），而积分已经超出高中数学的范畴。

01

原始矩

(Raw Moment)

原始矩的定义

随机变量 X 的 m 阶原始矩（m-th raw moment，简称 m 阶矩）是

$$E(X^m)$$

其中，$E(\cdot)$ 是期望算子（expectation operator），代表针对 X^m 的概率分布的特定算法，而非 X^m 的函数。

当 $E(|X^m|)=\infty$ 时，我们称“随机变量 X 的 m 阶原始矩不存在”。

在很多统计学教材和论文中，我们都可以看见 $E(|X^m|)<\infty$ 形式的假设或条件，这一假设或条件通常就是为了确保随机变量 m 阶矩的存在。

数学期望

离散型随机变量 X 的**数学期望**（expectation 或 expected value，简称“**期望**”）是 X 的**一阶原始矩** μ_X，即

$$\boldsymbol{\mu_X = E(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x)}$$

其中，$f_X(\cdot)$ 是 X 的概率质量函数（PMF），$\mathcal{X}$ 是 X 的支撑集。

例： 随机变量 Y 的概率分布列如下所示。Y 的数学期望是多少？

Y 的取值	-1	0	1
概率	0.5	0.2	0.3

$$E(Y) = \sum_{y \in \mathcal{Y}} y \cdot f_Y(y) = -1 \times 0.5 + 0 \times 0.2 + 1 \times 0.3 = -0.2$$

数学期望与平均数

离散型随机变量的数学期望是试验中所有可能的结果乘以其相应的概率的总和。

注意：数学期望不等于我们平时常说常用的平均数!

我们日常所说的平均数在统计学中被称为“简单算术平均数”（ordinary arithmetic mean）：

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

其中，X_i 代表我们观测到的数值，n 代表观测的次数。显然，这和数学期望的定义式不同!

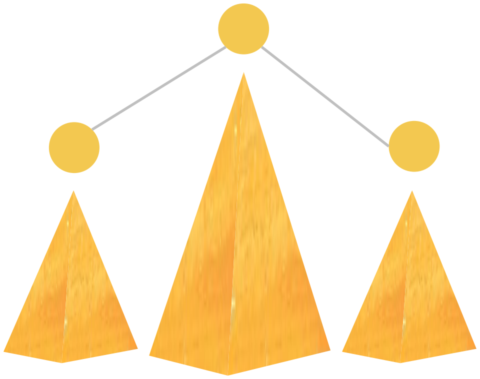
实际上，数学期望可以理解为随机变量 X 所有可能取值的加权算数平均数（weighted arithmetic mean），权重就是各个可能取值相应的概率。

$$\bar{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \Rightarrow E(X) = \frac{\sum_{x\in\mathcal{X}} x \cdot \Pr(X = x)}{\sum_{x\in\mathcal{X}} \Pr(X = x)} = \sum_{x\in\mathcal{X}} x \cdot \Pr(X = x) = \sum_{x\in\mathcal{X}} x \cdot f_X(x)$$

数学期望的用途

数学期望反映的是一个随机变量取值的中心趋势（central tendency）——它回答了“随机变量的取值集中趋向于何值？”

数学期望并非唯一可以反映中心趋势的统计工具；中位数（median）、众数（mode）等都可以反映中心趋势。



数学期望的局限

Expectation is not everything!（数学期望不等于一切！）

数学期望，如同接下来我们将要学习的二阶矩、三阶矩等，只反映了随机变量所服从的概率分布的一个特征。数学期望并不能提供给我们关于一个概率分布的全部信息。

例一

X 的取值	-1	0	1
概率	0.3	0.4	0.3

Y 的取值	-2	0	2
概率	0.3	0.4	0.3

显然，X 和 Y 服从的是两个不同的概率分布，但是它们的数学期望是相等的。

$$E(X) = 0 = E(Y)$$

数学期望的局限

Expectation is not everything!（数学期望不等于一切！）

数学期望，如同接下来我们将要学习的二阶矩、三阶矩等，只反映了随机变量所服从的概率分布的一个特征。数学期望并不能提供给我们关于一个概率分布的全部信息。

例二

X 的取值	0	1	10000
概率	0.4	0.59	0.01

当 X 的可能取值中存在极端值时，即使它出现的概率较低，X 的数学期望仍会收到巨大影响：

$$E(X) = 100.59$$

数学期望的重要性质

- $E(c) = c$
- $E(aX + b) = a \cdot E(X) + b$
- $E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$
- $E(c_1X_1 + c_2X_2 + \cdots + c_nX_n) = c_1 \cdot E(X_1) + c_2 \cdot E(X_2) + \cdots + c_n \cdot E(X_n)$

其中，小写字母代表常数（即“定值”），大写字母代表随机变量。

练习： 请证明这些性质对于离散型随机变量恒成立（设 $E(X) < \infty, \forall X$）。

以上性质对于离散型和连续型随机变量均成立。

02

中心矩

(Central Moment)

中心矩和方差的定义

随机变量 X 的 m 阶中心矩（m-th central moment）是

$$E\{[X-E(X)]^m\}$$

离散型随机变量 X 的方差（variance）是 X 的二阶中心矩 σ_X，即

$$\boldsymbol{\sigma_X^2=D(X)=\sum_{x\in\mathcal{X}}(x-\mu_X)^2\cdot f_X(x)}$$

中国大陆高中、本科教材一般使用 $D(X)$ 代表方差（来源于英文单词 deviation，意为“离差”），而西方教材一般使用 $Var(X)$ 或 $var(X)$。

例题：方差的计算

X 的概率分布列如下所示。它的方差是多少？

X 的取值	0	1	2
概率	0.3	0.2	0.5

X 的数学期望为 $E(X) = 0 \times 0.3 + 1 \times 0.2 + 2 \times 0.5 = 1.2$。接下来，我们可以计算 X 各个取值与数学期望的差，然后再计算各个差的平方数。

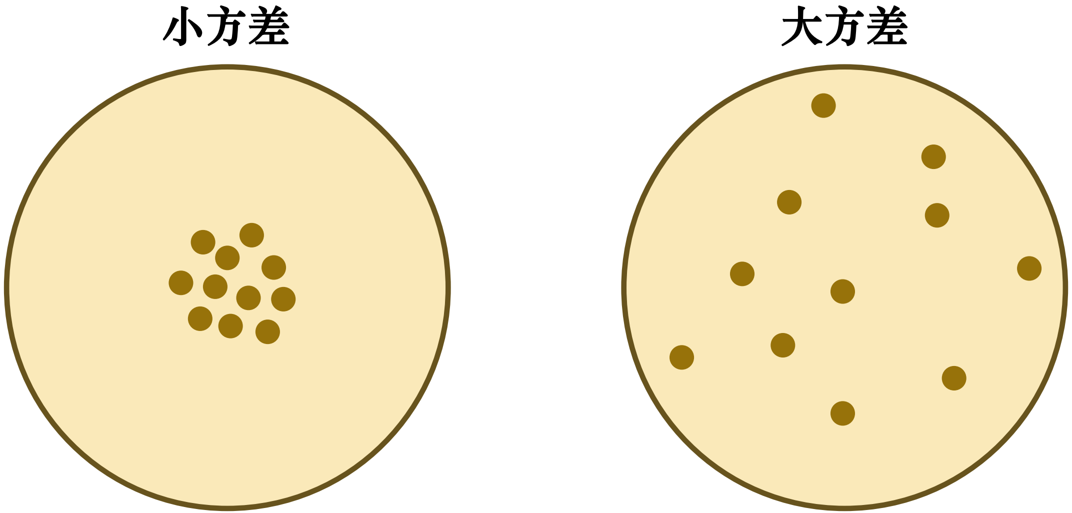
X 的取值	0	1	2
$E(X)$		1.2	
做差	-1.2	-0.2	0.8
平方	1.44	0.04	0.64
概率	0.3	0.2	0.5

最后，将这些平方差乘以各自相应的概率，加总，我们就得到了 X 的方差：

$$D(X) = 1.44 \times 0.3 + 0.04 \times 0.2 + 0.64 \times 0.5 = 0.76$$

方差的用途

方差反映的是一个随机变量取值的离散程度（dispersion）——它回答了“随机变量的取值有多么分散？”



方差为什么是“方”差?

让我们来仔细观察一下离散型变量的方差定义式。

$$\sigma_X^2 = D(X) = E\{[X - E(X)]^2\} = \sum_{x \in \mathcal{X}} (x - \mu_X)^2 \cdot f_X(x)$$

方差的“差”来自于 $x - \mu_X$，相当于 X 的支撑集内的一个点与 X 的数学期望之间的“距离”，这也被称为“离差”（deviation）。

为了衡量离散程度，我们可以将这些“距离”乘以相应的概率，然后加总:

$$\sum_{x \in \mathcal{X}} (x - \mu_X) \cdot f_X(x)$$

这其实是“一阶中心矩”。

请问，你认为用这个矩来衡量离散程度的缺点/问题是什么?

方差为什么是“方”差?

使用一阶中心矩来衡量离散程度的问题在于：等概率的“正距离”和“负距离”会相互抵消。

例三

X 的取值	-1000	1000
概率	0.5	0.5

X 的数学期望为 $E(X) = 0$，一阶中心矩也为 0，尽管 X 的离散程度看起来非常大。

为了解决这个正负相互抵消的问题，我们可以在一阶中心矩中加入“绝对值”：

$$\sum_{x \in \mathcal{X}} |x - \mu_X| \cdot f_X(x)$$

这个统计工具被称为“平均绝对离差”（mean absolute deviation），记为 $MAD(X)$。

方差为什么是“方”差？

使用平均绝对离差来衡量离散程度并没有问题。但是，统计学家们更喜欢使用“平方”而非“绝对值”。

$$\sum_{x \in \mathcal{X}} (x - \mu_X)^2 \cdot f_X(x)$$

这不是为了符合二阶中心矩的定义，而是因为考虑到计算机语言的特点：

现实中，大多数统计研究或应用都需要统计软件的帮助（笔算再快也快不过 Excel）。在计算机语言中，平方运算是一个如同加减乘除般简单的任务。然而，绝对值运算是一个条件语句（conditional statement）：当输入值大于或等于 0 时，返回原数值；当输入值小于 0 时，返回原数值的相反数。编写过软件代码的同学应该知道，条件语句的代码行数往往会超过普通运算代码的行数，这也意味着在计算机眼中，绝对值运算是一个比平方运算更加繁琐的任务。

方差为什么是“方”差?

$$\sum_{x\in\mathcal{X}}(x-\mu_X)^2\cdot f_X(x)$$

使用“平方”的另一个优点是：“平方”惩罚了极端值。

当随机变量的取值极大或者极小时，这些极端值与数学期望的距离会在平方之后被放大，使得它们对方差的增大有着更显著的作用。

方差的重要性质

- $D(c) = 0$
- $D(a + X) = D(X)$
- $D(bX) = b^2 \cdot D(X)$
- $D(a + bX) = b^2 \cdot D(X)$
- $D(X) = E(X^2) - [E(X)]^2$
- 一般情况下，$D(X + Y) \neq D(X) + D(Y)$

其中，小写字母代表常数（即“定值”），大写字母代表随机变量。

练习：请证明这些性质对于离散型随机变量恒成立（设 $E(|X^2|) < \infty, \forall X$ ）。

注：$D(X + Y)$ 在一般情况下的公式将会在下节课中介绍。

以上性质对于离散型和连续型随机变量均成立。

标准差的定义

“方差”的单位是古怪的。例如，随机变量 X 代表一所学校内男生的身高，单位为米（m），那么 X 的方差 $D(X)$ 的单位就是平方米（m^2）。这个方差的值到底有什么现实含义呢?

为了解决这个问题，统计学家们选择对方差进行开方，并将这种统计工具称为“**标准差**”（standard deviation）。对于离散型随机变量 X，它的标准差为

$$\sigma_X = SD(X) = \sqrt{D(X)} = \sqrt{\sum_{x \in \mathcal{X}} (x - \mu_X)^2 \cdot f_X(x)}$$

显然，标准差的单位和随机变量的单位是保持一致的。

非高考内容!

03

标准矩

(Standardized Moment)

标准矩的定义

随机变量 X 的 m 阶标准矩（m-th standardized moment 或 normalized m-th central moment）是

$$\frac{E[(X-\mu_X)^m]}{\sigma_X^m}=\frac{E\{[X-E(X)]^m\}}{[D(X)]^m}$$

简言之，m 阶标准矩就是经过标准化（standardization）后的 m 阶中心距。

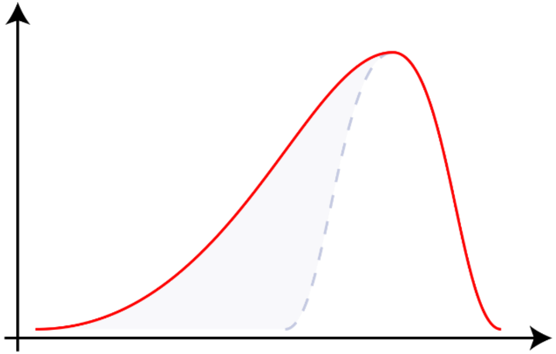
偏度的定义和用途

随机变量 X 的**偏度**（skewness）是 X 的**三阶标准矩**：

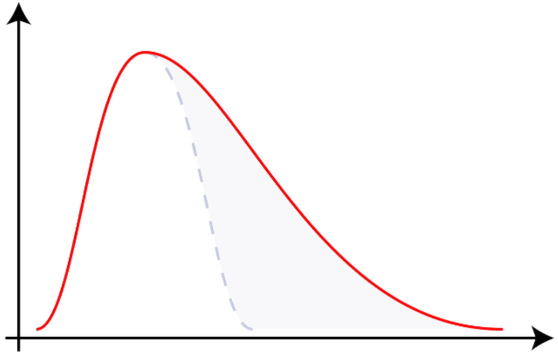
$$Skew(X) = \frac{E[(X-\mu_X)^3]}{\sigma_X^3} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^3\right]$$

偏度的用途是衡量随机变量所服从的概率分布是否关于其数学期望呈现对称性。

偏度 <0 表示概率分布的左尾较长。



偏度 >0 表示概率分布的右尾较长。



峰度的定义和用途

随机变量 X 的峰度（Pearson’s kurtosis）是 X 的四阶标准矩：

$$Kurt(X) = \frac{E[(X-\mu_X)^4]}{\sigma_X^4} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^4\right]$$

峰度的用途是衡量随机变量所服从的概率分布“尾部”的厚度——即衡量极端值出现的概率较高还是较低。

注意：不要被“峰度”这个名称误导—— $Kurt(X)$ 真正衡量的是概率分布的“尾部”而非“峰部”。

非高考内容!

04

矩量母函数

(MGF)

矩量母函数的定义和用途

随机变量 X 的**矩量母函数**（moment generating function，简写为 MGF）是

$$M_X(t) = E(e^{tX})$$

其中，$t \geq 0$，$E(\cdot) < \infty$。

这个以非负数 t 为自变量的函数被称为矩量**母**函数是因为我们可以通过它快速地算出任何阶数的原始矩：

$$\frac{d^m}{dt^m} M_X(t)\bigg|_{t=0} = E(X^m)$$

换言之，随机变量 X 的矩量母函数在 $t = 0$ 上的 m 阶导数就是 X 的 m 阶原始矩。

矩量母函数的描述功能

如果我们已知一个随机变量的所有阶矩（从一阶到 ∞ 阶），那么这个随机变量所服从的概率分布可被唯一确定。

换言之，如果两个随机变量的所有阶矩（从一阶到 ∞ 阶）均相等，那么这两个随机变量必定服从同一个概率分布。

我们已知矩量母函数包含了所有原始矩的信息，因此，如果两个随机变量的矩量母函数完全相同，那么这两个随机变量也必定服从同一个概率分布。

矩量母函数，与概率质量（或密度）函数、累积分布函数、分位数函数一样，可以完美地描述一个随机变量（或概率分布）。

高中统计学

第叁课

谢谢聆听