2018年秋季统计201A（高级概率概论）-所有课堂讲稿

阿迪蒂亚·冈图博伊纳

2018年10月10日

# 目录

0.1样本空间、事件、概率。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …四

0.2条件概率和独立性。. . . . . . . . . . . . . . . . . . . . . . . . . . …五

0.3随机变量。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .6条

1. 随机变量、期望和方差7
   1. 随机变量的期望. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …八
   2. 方差。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .9条
2. 随机变量的独立性10
3. 共同分配10
   1. Ber（p）分布. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …十
   2. Bin（n，p）分布。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .10条
   3. 泊松分布. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …11个
4. 协方差、相关和回归13
5. 相关与回归14
6. 回到共同分配15
   1. 几何分布。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .15条
   2. 负二项分布. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …16个
7. 连续分布16
   1. 正态或高斯分布. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .16条
   2. 均匀分布。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …16个
   3. 指数密度。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .17条
   4. 伽马密度. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …17年
8. 变量转换18
9. 分布函数与分位数变换19
10. 接缝密度21
11. 转换下的接头密度22
    1. 绕行到卷积。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …23个
12. 转换下的接头密度24
    1. 线性变换。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …24个
    2. 可逆线性变换. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .25点
    3. 一般可逆变换. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …26年
13. 不可逆变换下的联合密度28
14. 更多有序统计：固定i 29的X（i）密度
    1. 方法一。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .29条
    2. 方法二。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …29年
    3. 方法三。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …30个
    4. 统一订单统计。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …30个
    5. 独立制服的最大值。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .31条
    6. 独立指数的最小值。. . . . . . . . . . . . . . . . . . . . . . . . . . . . …31年
15. 中心极限定理32
16. 分布收敛33
17. 弱大数定律35
18. 力矩生成函数36
19. 使用MGFs 37证明CLT
20. 关于CLT 39的两点评论
21. 再论弱律与概率39的收敛性
22. Slutsky定理、连续映射定理及其应用42
23. 三角法46
24. Delta方法在方差稳定变换中的应用47
    1. 激励方差稳定转换。. . . . . . . . . . . . . . . . . . . . . . …47个
    2. 方差稳定变换的构造。. . . . . . . . . . . . . . . . . . …48个
    3. 回到伯努利的例子. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …48个
    4. 回到泊松例子。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .49点
    5. 卡方检验。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …50个
    6. 几何示例. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …50个
25. g0（θ）=051时的Delta方法
26. 条件52
    1. 基础知识. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …52年
    2. 离散随机的条件分布、全概率律和贝叶斯规则

变量。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …52年

1. 离散随机变量的条件55
2. 连续随机变量的条件密度55
3. 条件密度与接头密度成正比57
4. 条件密度与独立性58
5. 连续随机变量的全概率律和贝叶斯规则58
6. 连续随机变量的全概率定律和贝叶斯规则59
7. 一般随机变量的LTP和Bayes规则61
   1. X和Θ都是离散的. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .61点
   2. X和Θ都是连续的。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .61点
   3. X是离散的，而Θ是连续的. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …61年
   4. X是连续的，而Θ是离散的。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …62年
8. 条件接缝密度63
9. 条件关节密度64
   1. 正态先验正态数据模型的应用。. . . . . . . . . . . . . . . . . . . . …66年
10. 条件期望67
    1. 迭代/总期望定律. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …68年
    2. 迭代/总期望定律。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …69个
    3. 总期望定律在统计风险最小化中的应用. . . . . . …69个
11. 条件方差71
12. 随机向量72
13. 随机向量73
14. 对称矩阵的绕道谱定理74
    1. 正交基。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .74点
    2. 谱定理。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …75个
    3. 谱定理的三个应用. . . . . . . . . . . . . . . . . . . . . . . . . . .76点
       1. 每个对称半正定矩阵都是协方差矩阵。. . . . . . …76年
       2. 变白。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …77个
       3. 随机向量的第一主成分。. . . . . . . . . . . . . . . . . . …77个
15. 最佳线性预测77
16. 最佳线性预测79
17. 剩余81
18. 绕道：舒尔补充81
19. 偏相关82
20. 偏相关和逆协方差83
21. 偏相关和最佳线性预测85
22. 当Y是随机向量时的BLP 87
23. 随机向量的矩母函数88
24. 多元正态分布89
    1. 多元正态的矩母函数。. . . . . . . . . . . . . . . . . . …89年
    2. 与i.i.d N（0,1）随机变量的连接。. . . . . . . . . . . . . . . . . . . . . . . . . . .90分
25. 多元正态分布的联合密度90
26. 多元正态随机变量90的性质
27. 多元正态随机变量的性质92
28. 幂等矩阵与卡方分布92
29. 多元正态分布95
30. 关于多元正态分布和卡方分布的一些注记97
31. 多元正态分布98

我们将通过对大学生概率材料的复习来开始这门课程。复习会很快，大约六节课就可以完成。

## 0.1样本空间、事件、概率

概率论在偶然或随机实验的情况下被调用在这样一个随机实验中，样本空间是所有可能结果的集合，我们用Ω表示例如，假设我们掷硬币两次那么一个合理的样本空间是{hh，ht，th，tt}。

样本空间的子集称为事件例如，在上面的例子中，{hh，ht}是一个事件，它表示两个抛出中的第一个导致头部的事件。类似地，两次投掷中至少有一次造成头部的事件用{hh，ht，th}表示。

给定一组事件A1，A2，。。。，

1. 表示A1不发生的事件。我们说这是事件A1的补充。
2. ∪i≥1Ai表示A1，A2发生。
3. ∩i≥1Ai表示所有A1，A2发生了。

概率定义为将事件映射（或关联）到介于0和1之间的实数并满足某些自然一致性属性的函数具体来说，P是一个概率，前提是：

1. 每次事件A的0≤P（A）≤1。
2. 对于空子集Φ（=不可能事件），P（Φ）=0
3. 对于整个样本空间（=特定事件），P（Ω）=1。
4. 如果事件A是事件序列A1、A2的不相交并，。。。（这意味着A中的每个点恰好属于集合A1，A2，…）中的一个），则P（A）=Pi≥1 P（Ai）。

示例0.1（非转换骰子）考虑以下一组骰子：

1. 模具A有2，2，4，4，9，9面。
2. 模具B有边1、1、6、6、8、8。三。模具C有3、3、5、5、7、7面。

A滚动的次数比B高的概率是多少？B滚得比C高的概率是多少？C滚得比A高的概率是多少假设，在任何骰子中，所有的结果都是一样的。

## 0.2条件概率和独立性

考虑概率P和事件B，其中P（B）>0然后我们可以定义每个事件A的P（A | B）为

（一）

P（A | B）称为给定B的条件概率。

这里有两个非常有趣的问题，来自Mosteller的愉快的书（题为50个挑战性的概率问题），说明了条件概率的使用。

例0.2（摘自Mosteller的书（问题13；囚犯困境））。三名囚犯，A、B和C，显然有着同样良好的记录，已经申请假释假释委员会决定释放三人中的两人，囚犯们知道这一点，但不知道是哪两人囚犯A的看守朋友知道谁会被释放。囚犯A意识到，询问典狱长A是否要被释放是不道德的，但想到要问一个囚犯的名字，而不是他自己要被释放。他认为在他提出要求之前，他被释放的机会是2/3他认为如果典狱长说“B会被释放”，他自己的机会现在已经下降到1/2，因为A和B或者B和C都会被释放所以A决定不通过询问来减少他的机会。然而，A的计算是错误的解释一下。

例0.3（摘自Mosteller的书（问题20：三角决斗））甲、乙、丙三方将进行一场三角手枪决斗大家都知道A命中目标的几率是0.3，C是0.5，B从不失误。他们将按顺序A、B、C依次射击他们选择的目标（但一名击中者失去了进一步的转弯，不再被击中），直到只有一名男子未被击中为止。A的策略应该是什么？

我们说两个事件A和B是独立的（在概率P下），如果

P（A | B）=P（A）。

等效地，独立性由P（A∩B）=P（A）P（B）或P（B | A）=P（B）给出。

涉及条件概率的一个非常重要的公式是Bayes规则。上面写着

（二）

你能从条件概率的定义（1）中得出这个结论吗？

这里有一个非常有趣的应用贝叶斯规则。

例0.4考虑一个癌症的临床测试，它可以产生阳性或阴性的结果假设一个真正患有癌症的病人有1%的机会在未被发现的情况下通过测试另一方面，假设一个没有癌症的病人有5%的几率得到阳性的检测结果假设2%的人口患有癌症。假设一个接受过检测的病人得到了阳性检测结果，他们患癌症的概率是多少？

假设C和Cc分别是患者有癌和无癌的事件。还假设+和-分别是测试产生正结果和负结果的事件。根据提供的信息，我们已经

P（－| C）=0.01 P（＋| Cc）=0.05 P（C）=0.02。

我们需要计算P（C |+）。根据贝氏法则，我们有

.

因此，这名患者患癌症的概率（假设测试结果为阳性）约为29%这尤其意味着，即使检测结果呈阳性，他们仍然不太可能患上癌症（尽管癌症的概率从2%增加到29%）。上述计算的另一个有趣的方面是

P（+=P（+C）P（C）+P（+C）P（Cc）=0.99\*0.02+0.05\*0.98=0.0688。

这意味着该检测将在大约7%的时间内产生阳性结果（注意，只有2%的人患有癌症）。

现在假设P（C）=0.001（与P（C）=0.02相反），并假设P（－| C）和P（＋| Cc）与之前一样保持在0.01和0.05。那么

P（+=P（+C）P（C）+P（+C）P（Cc）=0.99\*0.001+0.05\*0.999=0.0509。

在这里，真正的癌症发病率为0.001，表观发病率为0.05（增加了50倍）。在美国步枪协会（National Rifle Association）正在进行一项调查的背景下，不妨考虑一下这一点，调查对象是在过去一年中是否使用枪支自卫的公民。取C为真用法，+为报告用法。如果真的只有千分之一的人用枪自卫，那么二十分之一的人会用枪自卫这些例子摘自丹尼斯·林德利（Dennis Lindley）的一本名为《理解不确定性》（Understanding Uncertainty）的书（我觉得概率论和统计学的每个学生都应该读这本书）。

## 0.3随机变量

随机变量是一个函数，它将一个数字附加到样本空间的每个元素上。换句话说，它是一个将样本空间映射为实数的函数。

例如，在掷硬币50次的机会实验中，人头的数目是一个随机变量另一个随机变量是第一条尾巴前面的头数另一个随机变量是模式hththt出现的次数。

许多现实生活中的数据，例如（a）伯克利明天的平均气温，（b）这个房间里随机挑选的学生的身高，（c）我明天将接到的电话数量，（d）赫斯特大街9月份发生的事故数量，等等，都可以作为随机变量来处理。

对于每个事件A（回想一下，事件是样本空间的子集），可以关联一个随机变量，如果A发生，则取1，如果A不发生，则取0这称为对应于事件A的指示符随机变量，并由I（A）表示。

随机变量的分布，非正式地说，是对随机变量所取的一组值及其概率的描述。

如果一个随机变量X取一个有限或可数的可能值的不确定集（在这种情况下，我们说X是一个离散随机变量），它的分布由值a1，a2，。。。加上概率的具体说明：

P{X=ai}对于i=1,2，。。。。

将ai映射到P{X=ai}的函数称为离散随机变量X的概率质量函数（pmf）。

如果一个随机变量X取一组连续的值，它的分布通常由一个称为概率密度函数（pdf）的函数来描述。pdf是R上的一个函数f，它满足f（x）≥0对每个x∈R和

.

随机变量的pdf f可用于计算每个集合a通孔的P{X∈a}

Z轴

P{X∈A}=f（X）dx。

一个

注意，如果X有密度f，那么对于每个y∈R，

.

重要的是要记住，随机变量的密度函数f（x）并不代表概率（特别是，f（x）取的值远大于1是很常见的）相反，f（x）的值可以被认为是一个比例常数。这是因为通常（只要f在x处是连续的）：

.

# 随机变量、期望和方差

随机变量是一个函数，它将一个数字附加到样本空间的每个元素上。换句话说，它是一个将样本空间映射为实数的函数。

例如，在掷硬币50次的机会实验中，人头的数目是一个随机变量另一个随机变量是第一条尾巴前面的头数另一个随机变量是模式hththt出现的次数。

许多现实生活中的数据，例如（a）伯克利明天的平均气温，（b）这个房间里随机挑选的学生的身高，（c）我明天将接到的电话数量，（d）赫斯特大街9月份发生的事故数量，等等，都可以作为随机变量来处理。

对于每个事件A（回想一下，事件是样本空间的子集），可以关联一个随机变量，如果A发生，则取1，如果A不发生，则取0这称为对应于事件A的指示符随机变量，并由I（A）表示。

随机变量的分布，非正式地说，是对随机变量所取的一组值及其概率的描述。

如果一个随机变量X取一个有限或可数的可能值的不确定集（在这种情况下，我们说X是一个离散随机变量），它的分布由值a1，a2，。。。它需要

|  |  |
| --- | --- |
| 连同概率说明： |  |
| P{X=ai} | 对于i=1,2，。。。。 |

将ai映射到P{X=ai}的函数称为离散随机变量X的概率质量函数（pmf）。

如果一个随机变量X取一组连续的值，它的分布通常由一个称为概率密度函数（pdf）的函数来描述pdf是R上的一个函数f，它满足f（x）≥0对每个x∈R和

.

随机变量的pdf f可用于计算每个集合a通孔的P{X∈a}

Z轴

P{X∈A}=f（X）dx。

一个

注意，如果X有密度f，那么对于每个y∈R，

.

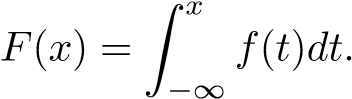
重要的是要记住，随机变量的密度函数f（x）并不代表概率（特别是，f（x）取的值远大于1是很常见的）相反，f（x）的值可以被认为是一个比例常数这是因为通常（只要f在x处是连续的）：

.

随机变量X的累积分布函数（cdf）定义为

F（x）：=P{x≤x}，对于－∞<x＜∞。

这是为所有离散或连续的随机变量定义的如果随机变量X具有密度，则其cdf由



cdf具有以下性质：（a）它是非递减的，（b）右连续的，（c）limx ~－－∞F（x）=0和limx ~＋∞F（x）=1。

## 随机变量的期望

设X为离散随机变量，g为X范围内的实值函数，则g（X）具有有限期望

十

|g（x）| P{x=x}<∞

十

其中总和在x的所有可能值x上。注意g（x）上的绝对值。

当g（X）有有限期望时，我们定义Eg（X）为

（三）

在这里，求和是在所有可能的x的x上。

类似地，如果X是具有密度（pdf）f的连续随机变量，那么我们说g（X）具有有限期望

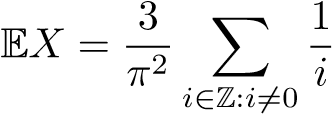
.

当g（X）有有限期望时，我们定义Eg（X）为

（四）

为什么在定义期望值之前，我们需要确保绝对和（或积分）的有限性？因为否则（3）中的和（或（4）中的积分）可能定义不清例如，当X是离散随机变量时，其值为…，-3，-2，-1,1,2,3有概率的

对于i∈Z，i 6=0然后g（x）=x的（3）和变成



这是毫无意义的这里很容易看出P x | x | P{x=x}=∞。

如果A是一个事件，那么回想一下I（A）表示对应的指示符随机变量，当A保持时等于1，当A不保持时等于0。很容易注意到I（A）的期望值正好等于P（A）。

需要记住的一点是期望是一个线性算子，即。，

E（aX+bY）=aE（X）+bE（Y）

对于任意两个期望值有限的随机变量X和Y以及实数a和b。

## 方差

如果X2有有限期望，则随机变量X的方差是有限的（你知道当X2有有限期望时，X也会有有限期望吗你将如何证明这一点？）。在这种情况下，X2的方差定义为

V ar（X）：=E（X––X）2=E（X2）–––2X，其中μX：=E（X）。

从定义中可以清楚地看出，随机变量X的方差衡量X在其均值E（X）周围所取值的平均变异性。

假设X是一个离散随机变量，具有有限多个概率值x1，…，xn。那么X的方差是多少？

X方差的平方根称为X的标准差，通常用SD（X）表示。

随机变量X的期望具有以下变分性质：是a的值使所有实数a上的数量E（X-a）2最小化。你知道如何证明这一点吗？

如果随机变量X的方差很小，则X不能偏离其平均值（=E（X）=μ）。这可以通过Chebyshev不等式来精确说明，它说明了以下几点。

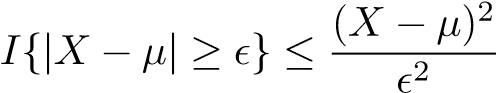
Chebyshev不等式：设X为方差有限且均值为μ的随机变量然后，对于每一个>0，以下不等式成立：

.

换言之，X偏离其平均值以上的概率由

.

切比雪夫不等式的证明



同时考虑两边的期望值（在左边，我们有一个指标随机变量，它取值1 when，否则取0）。

# 随机变量的独立性

我们说两个随机变量X和Y是独立的，如果对任何涉及Y的事件的条件作用不改变任何涉及X的事件的概率，即。，

P{X∈A | Y∈B}=P{X∈A}

对每个A和B。

同样地，X和Y的独立性与

P{X∈A，Y∈B}=P{X∈A}P{Y∈B}

对每个A和B。

以下是独立的后果。如果X和Y是独立的，那么

1. 对于每对函数g和h，g（X）和h（Y）是独立的。
2. E（g（X）h（Y））=每对函数g和h的Eg（X）Eh（Y）。

更一般地，我们说随机变量X1，…，Xk是（相互）独立的，如果对于每1≤i≤k，任何涉及Xj的事件的条件作用，j 6=i不改变任何涉及Xi的事件的概率。

从这里可以很容易地导出独立性的性质，例如

P{X1∈A1，…，Xk yn Ak}=P{X1 yn A1}P{X2 yn A2}…P{Xk yn Ak}

对于事件A1，…，Ak的所有可能选择。

# 共同分配

## Ber（p）分布

如果随机变量X取p{X=1}=p的两个值0和1，则称其具有Ber（p）（带参数p的Bernoulli）分布。

然后注意EX=p和V ar（X）=p（1-p）。p的哪个值是X的最大变量？最小变量？

## 仓（n，p）分布

随机变量X具有参数n和p的二项分布（n是正整数，p∈[0,1]），如果它取0,1，…，n的值，pmf由

每k=0,1，…，n。

这是二项式系数：

.

一个Bin（n，p）随机变量的主要例子是一枚硬币在n次独立投掷中获得的人头数，人头的概率等于p。

这里有一个关于Mosteller书中二项分布的有趣问题（你可以很容易地在R中计算出这些概率）。

例3.1（摘自Mosteller的书（问题19：Issac Newton帮助Samuel Pepys））。佩皮斯写信给牛顿，问三个事件中哪一个更有可能：一个人（a）掷6个骰子时至少得到16个，（b）掷12个骰子时至少得到2个6，或（c）掷18个骰子时至少得到3个6，答案是什么？

设X表示一枚硬币n次独立抛掷中的头部数，头部概率为p，那么我们知道X∼Bin（n，p）。如果，现在，Xi表示二元随机变量，如果第i次投掷是头部，则取1，如果第i次投掷是尾部，则取0，那么应该清楚

X=X1+····+Xn。

注意，每个Xi是Ber（p）随机变量，X1，…，Xn是独立的。因此，Bin（n，p）随机变量可以看作n个独立Ber（p）随机变量的和中心极限定理（我们将在后面的课程中详细研究）意味着大量i.i.d的和（i.i.d是什么？）随机变量近似正常。这意味着当n大且p固定时，Bin（n，p）分布看起来像正态分布。我们稍后会把这一点说清楚的。特别地，这意味着二项式概率可以通过N个大P和P固定的正常概率近似计算。从这个角度来看，当k较大时，从6k卷的模具中得到k个或更多个六角的概率是多少？

Bin（n，p）分布的平均值是多少？根据一枚硬币的n次独立投掷，对头部概率的无偏估计是什么？Bin（n，p）的方差是多少？

## 泊松分布

如果X取0,1,2，…，则随机变量X具有参数λ>0的泊松分布（用Poi（λ）表示）pmf由

对于k=0,1,2，。。。。

泊松分布的主要效用来自以下事实：

事实：二项式分布Bin（n，p）很好地近似于泊松分布POI（NP），只要数量NP2小。

要直观地看到为什么这是真的，只要看看

P{Bin（n，P）=0}=（1P）n=exp（nlog（1P））。

现在注意，np2很小意味着p很小（注意p可以写成pnp2/n≤pnp2，所以小np2必然意味着p很小）当p为小时，我们可以将log（1～p）近似为p p2／2，从而得到

.

现在因为NP2小，所以我们可以忽略上面的第二个术语，得到p{bin（n，p）＝0 }近似由Ep（NP）精确地等于p{POI（NP）＝0 }。对于每一个固定k＝0，1，2，……，同样可以用p{POI（NP）＝k}近似地逼近p{ Bin（n，p）＝k}。

有一个形式定理（称为Le-Cam定理）严格证明了当np2很小时Bin（n，p）≈Poi（np）以下无需证明（其证明不在本类范围内）。

定理3.2（勒卡姆定理）假设X1，…，X n是独立的随机变量，因此对于某些pi∈[0,1]对于i=1，…，n，Xi∼Ber（pi），设X=X1+····+Xn和λ=p1+…pn那么

.

在特殊情况下，当p1=···=pn=p时，上述定理表明

∞

X | P{Bin（n，P）=k}-P{Poi（np）=k}|<2np2

k=0

因此，当np2很小时，对于每个k=0,1，…，概率P{Bin（n，P）=k}接近于P{Poi（np）=k}这个事实的一个含义是，对于每个固定的λ>0，我们有

当n很大时。

这是因为当p=λ/n时，我们有np2=λ2/n，当n较大时，它会很小。

泊松分布的这种近似性质是泊松分布用于模拟稀有事件计数的原因。例如，一个电话接线员一天接到的电话号码、一天中某条街道上发生的事故数目、一本书中发现的打字错误数目、一场足球比赛中进球的数目，都可以被模拟成Poi（λ）的某个λ>0。你能解释为什么这些现实生活中的随机量可以用泊松分布来建模吗？

下面的例子给出了泊松分布提供良好近似的另一种情况。

例3.3。假设n个字母编号为1，…，n和n个信封编号为1，…，n。字母i的右信封是信封i。假设我取随机排列σ1，…，σn为1，…，n，然后将字母σi放入信封i中。让X表示在其右信封中的字母数X的分布是什么？

设Xi为随机变量，当第i个字母在第i个信封中时取1，否则取0那么很明显X=X1+····+Xn。注意

对于每个i=1，…，n。

这是因为第i封信同样可能在n个信封中的任何一个信封中。因此，这意味着

当i=1，…，n时，Xi∼Ber（1/n）。

如果X i也是独立的，那么X=X1+····+X n将是Bin（n，1/n），这对于大n非常接近Poi（1）。但是Xi在这里不是独立的，因为对于i 6=j，

.

但依赖性相对较弱，X的分布与Poi（1）非常接近。我们将通过证明p{x＝0 }近似等于p{POI（1）＝0 }＝E＝1来证明这一点。我将作为一个练习来展示，对于每个固定的k，P{X=k}≈P{Poi（1）=k}。要计算P{X=0}，我们可以编写

.

现在请注意，对于每个i1

.

这给了

.

很容易证明Poi（λ）随机变量的期望和方差都等于λ。这也很有意义，因为两者之间的联系：

Poi（λ）≈Bin（n，λ/n）

作为

E（Bin（n，λ/n））=λ和。

当通过泊松分布对计数数据建模时，可以通过经验检验方差等于平均值的假设。如果经验方差似乎远高于平均值，那么有人说存在过度分散，在这种情况下，泊松可能不是一个好的数据模型。

# 协方差、相关与回归

给定两个随机变量X和Y（使得X2和Y 2具有有限期望），X和Y之间的协方差用Cov（X，Y）表示，并定义为

Cov（X，Y）：=E[（X-μX）（Y-μY）]（5）

式中，μX：=E（X）和μY：=E（Y）。换句话说，Cov（X，Y）被定义为随机变量（X-？X）（Y-？Y）的期望值（但是这个随机变量有有限的期望值吗？你能证明这是假设X2和Y 2有有限期望的结果吗？）.

重要的是要注意协方差是双线性算子，即。，

Cov（XaiXi，XbjYj）=XXaibjCov（Xi，Yj）。（6）

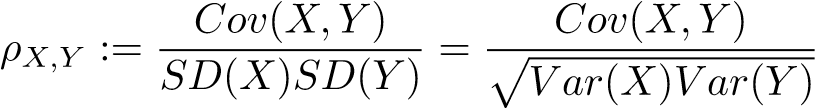
我会的

你能证明这是协方差定义（5）和期望算子线性的结果吗？

当X=Y时，很容易看出Cov（X，X）只是X的方差。利用协方差与方差和（6）之间的关系，可以推导出方差的下列标准性质：

1. V ar（aX+b）=a2V ar（X）。
2. V ar（Pi aiXi）=Pi a2iV ar（Xi）+Pi6=j aiajCov（Xi，Xj）。

两个随机变量X和Y（使得X2和Y 2具有有限方差）之间的相关性定义为：



如果ρX，Y=0，我们说X和Y是不相关的。

提议4.1关于相关性的两个事实：

1. 相关系数ρX，Y总是介于-1和1之间。
2. 对于每一个a，b，c，d∈（－∞，∞）。换句话说，在线性变换下，相关性是不变的（直至符号翻转）。

证明写

.

使用标准不等式：

（七）

用a=（X-μX）/pV ar（X）和b=（Y-μY）/pV ar（Y）获得

.

这证明了ρX，Y≤1为了证明ρX，Y<-1，用

（八）

关于相关性和线性函数的事实作为练习留在这里。

Cauchy-Schwartz不等式：相关性ρX，Y介于-1和1之间的事实有时通过Cauchy-Schwartz不等式得到证明，该不等式指出：对于每对随机变量Z1和Z2，我们有

)（九）

根据上述不等式，取Z1=X-μX和Z2=Y-μY，即可推导出|ρX，Y |≤1的事实。

你能用（7）和（8）证明柯西-施瓦兹不等式（9）吗？

不相关性与独立性：下面总结了不相关性与独立性的关系：

1. 两个独立的随机变量X和Y是不相关的。
2. 存在许多不相关的随机变量X和Y的非独立的例子。你能想出几个吗？
3. 当且仅当g（X）和h（Y）对于每对函数g和h不相关时，两个随机变量X和Y是独立的。

# 相关与回归

ρX，Y的一个重要性质是它测量X和Y之间的线性关联强度。本节对此进行了说明。考虑由x的线性函数β0＋β1x逼近随机变量y的问题。对于给定数β0和β1，让我们用均方误差测量β0 +β1x近似y的精度：

L（β0，β1）：=E（Y-β0-β1X）2。

如果β0＋β1x是一个很好的近似y，则L（β0，β1）应该是低的。相反，如果β0＋β1X是Y的差近似，则L（β0，β1）应该是高的。当β0和β1在所有实数上变化时，L（β0，β1）的最小可能值是多少。

可以看出

（十）

你知道如何证明以上的情况吗？

事实（10）精确地捕获了相关测量Y和X之间的线性关联强度的解释。这是因为Minβ0、β1 L（β0、β1）代表了Y和X（10）的线性组合近似Y的最小可能均方误差，这表明它与Y和X.之间的相关性直接相关。

你能明确地写下使L（β0，β1）最小的β0和β1的值吗？

以上有没有让你想起线性回归？以什么方式？

# 回到共同分配

在上一节课中，我们研究了二项分布和泊松分布。这两种分布都是离散分布，可由抛硬币引起（Bin（n，p）是硬币在n次独立抛硬币中的头部数分布，头部概率p和Poi（λ）≈Bin（n，λ/n））我们现在将修正另外两个由抛硬币产生的离散分布：几何分布和负二项分布。

## 几何分布

如果X取1,2，…，则X具有参数p∈[0,1]的几何分布（写成X∼Geo（p））概率：

P{X=k}=（1-P）k-1p对于k=1,2，。。。。

很容易看出，获得第一个头部所需的独立投掷次数（硬币的头部概率为p）具有Geo（p）分布。

Geo（p）分布具有无记忆性的有趣性质，即如果X∼Geo（p），那么

P{X>m+n | X>n}=P{X>m}。（11）

当P{X>m}=（1-P）m时，这很容易检查。有趣的是，几何分布是{1,2，…}上唯一满足无记忆特性（11）的分布为此，假设X是一个满足（11）的随机变量，它接受{1,2，…}中的值设G（m）：=P{X>m}，m=1,2，。。。。那么（11）与

G（m+n）=G（m）G（n）。

这清楚地给出了G（m）=（G（1））m，每m=1,2现在G（1）=P{X>1}=1-P{X=1}如果p=1-p{X=1}，则

G（m）=（1–（p）m

这意味着P{X=k}=P{X>k-1}-P{X>k}=P（1-P）k-1对于每k≥1意味着X是

地理位置（p）。

## 负二项分布

设X表示获得第k个头所需的投掷次数（硬币的头部概率为p）。下面给出X的分布。X取k，k+1，。。。和

这称为负二项分布，参数为k和p（用NB（k，p）表示）如果G1，…，Gk是独立的Geo（p）随机变量，那么G1+···+Gk∼NB（k，p）（你能证明吗？）.

# 连续分布

## 正态或高斯分布

随机变量X具有正态分布，平均值μ，方差σ2>0，如果它具有以下pdf：

.

我们写X∼N（μ，σ2）。当μ=0和σ2=1时，我们认为X具有标准正态分布，标准正态pdf简单地用φ（·）表示：

.

2√

你知道为什么φ（·）是一个有效的密度，也就是为什么R e-x/2dx=2π吗？

标准正态cdf用Φ（x）表示：

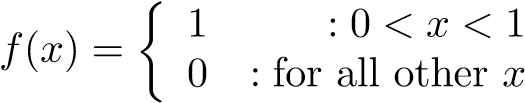
Z xΦ（x）：=φ（t）dt。

−∞

如果X∼N（μ，σ2），则E（X）=μ，V ar（X）=σ2有关正态分布的许多属性的列表，请参见相应的维基百科页面中心极限定理是正态分布是统计学中最突出的分布的主要原因。

## 均匀分布

如果随机变量U具有以下pdf，则称其在（0,1）上具有均匀分布：



我们写U∼U[0,1]你是什么意思你的方差是多少？在统计学中，统一分布在哪里空分布下的p值通常根据U[0,1]分布进行分布（稍后详细介绍）。

更一般地，给定一个区间（a，b），我们说随机变量U在（a，b）上具有均匀分布，如果它具有以下pdf：

：a<x<b

0:对于所有其他x

我们把它写成U∼U（a，b）。

## 指数密度

速率参数λ>0的指数密度（用Exp（λ）表示）由下式给出

f（x）=λe--λxI{x>0}。

它可以说是建模被限制为非负的随机量的最简单密度。它用来模拟话务员从现在开始接到的第一个电话的时间一般来说，它是随着Poisson过程中到达间隔时间的分布而产生的（稍后我们研究Poisson过程时将对此进行详细说明）。Exp（λ）的cdf很容易被认为是

对于x>0。

指数密度具有无记忆性（就像几何分布一样）。的确，

.

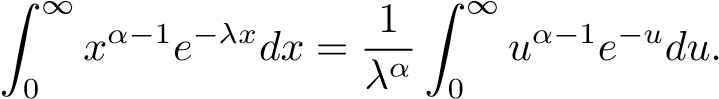
事实上，指数密度是（0，~）上唯一具有无记忆性的密度（证明留作练习）。在这个意义上，指数分布x可以看作是几何分布的连续模拟。

## 伽马密度

通常在指数密度之后讨论伽马密度形状参数α>0、速率参数λ>0的伽马密度由下式给出

f（x）107xα-1e-λxI{x>0}（12）

为了找到上面的比例常数，我们需要评估



现在函数

对于α>0

在数学上叫做伽玛函数所以（12）中的比例常数由



所以伽马密度的公式是：

.

我们称之为伽马（α，λ）密度。

注意，当α=1时，伽马（α，λ）密度减小到Exp（λ）密度因此，伽马密度可以看作是指数密度的推广。事实上，伽马密度可以看作是负二项分布的连续模拟，因为如果X1，…，Xk是独立的Exp（λ）随机变量，那么X1+···+Xn∼伽马（k，λ）（因此伽马分布作为i.i.d指数之和出现，正如负二项分布作为i.i.d几何随机变量之和出现一样）。

下面是伽马函数的一些基本性质，对我们以后的工作有帮助对于任意α>0，Gamma函数没有闭合表达式但是当α是正整数时，

|  |  |  |
| --- | --- | --- |
| 可以看出 |  |  |
| Γ（n）=（n-1）！  上述不平等是该性质的结果 | n≥1时。 | （十三） |
| Γ（α+1）=αΓ（α） | 对于α>0 | （十四） |

以及Γ（1）=1这一微不足道的事实你可以很容易地通过部分集成来验证（14）。

√

关于Gamma函数的另一个简单事实是Γ（1/2）=π（这是

2√

R e-x/2dx=2π）。

# 变量转换

通常采用随机变量的函数或变换。考虑一个随机变量X，对X应用一个函数u（·）将X转换成另一个随机变量Y=u（X）。如何从X的分布中找到Y=u（X）的分布？

如果X是离散随机变量，那么Y=u（X）也将是离散的，那么Y的pmf可以直接用X的pmf来表示：

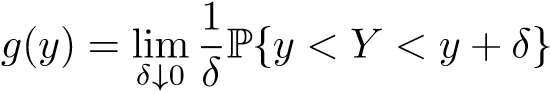
P{Y=Y}=P{u（X）=Y}=X P{X=X}。

x:u（x）=y

如果X是一个密度为f的连续随机变量，u（·）是一个光滑函数，那么用f来记下Y=u（X）的密度是相当简单的。有一些通用的公式可以这样做，但最好从第一性原理开始学习。我将用以下两个例子来说明这个总体思路。

例8.1。假设X∼U（π/2，π/2）。Y=tan（X）的密度是多少这是从第一原理开始的方法注意tan（x）与x的距离在（－∏/2，π/2）上的范围是R，所以固定y∈R，我们会发现y在y的密度g以下。

g（y）的公式是



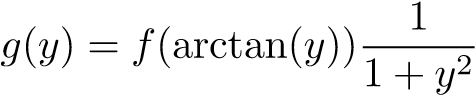
以便

当δ较小时，P{y<y<y+δ}≈g（y）δ。（15）

对于小δ，



其中f是X的密度，与（15）相比，我们可以得出结论

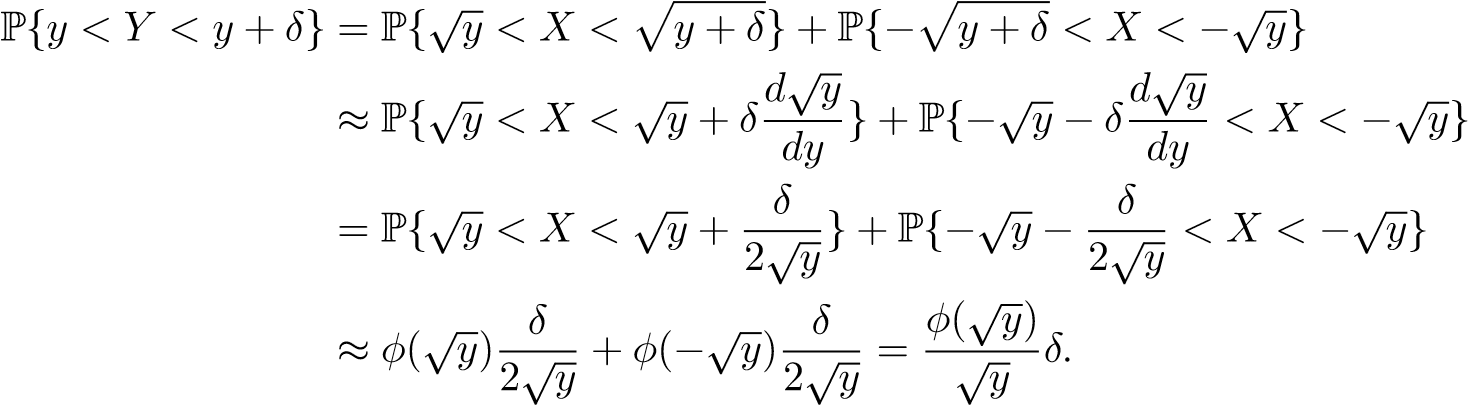


利用X∼U（－∏/2，π/2）的密度，我们推断

对于y∈R。

这是柯西密度。你能把上述论点严格化吗如果教你计算变换密度的公式，这个公式和这里的答案匹配吗？

例8.2。假设X∼N（0,1），使X具有标准法向密度φ（·）。Y=X2的密度是多少？下面的方法是从第一原则开始的。X2的范围是X的范围在（－～～，～）以上，所以我们把y定为0。我们将在Y处找到Y的密度g



这给了

当y>0时。

这是具有1个自由度的卡方随机变量（或形状参数α=1/2和比例参数β=1/2的伽马随机变量）的密度。

# 分布函数与分位数变换

随机变量X的分布函数（cdf）定义为

F（x）：=P{x≤x}，对于－∞<x＜∞。

cdf具有以下属性：

1. 它是非递减的，即当x≤y时，F（x）≤F（y）。
2. 它接受介于0和1之间的值。三。每一个x，我们都有

还有林。

这特别意味着F是右连续的，F的连续性相当于每X的P{X=X}=0。

四。函数F（x）接近0作为x······，接近1作为x·········。

上述性质刻画了cdf，即满足这四个性质的（－～～，∞）上的每个函数F等于某个随机变量的cdf证明这一点的一种方法是通过分位数变换。

给定满足上述四个性质的函数F，相应的分位数变换（或分位数函数）qF是（0,1）上的实值函数，定义为

qF（u）：=inf{x∈R:F（x）≥u}。（16）

分位数变换可视为cdf F的某种逆变换。事实上，当cdf F连续且严格增加时，分位数函数qF正好等于F−1。

分位数变换的基本性质如下对于x∈R和0<u<1：

当且仅当x≥qF（u）时F（x）≥u（17）

这是（17）的证据。当F（x）≥u时，从qF的定义可以明显看出x≥qF（u）。

另一方面，根据qF（u）的定义和F不递减的事实，我们得到了每大于0。设0，利用F的右连续性，我们推导出F（qF（u））≥u。这意味着当x≥qF（u）时，我们有F（x）≥F（qF（u））≥u。这证明了（17）。

分位数变换的一个重要事实是：

P{X≤qF（u）}≥u和P{X<qF（u）}≤u。（18）

因此，qF（u）是X的分位数（当u=0.5时，则qF（0.5）是X的中位数）因此命名为分位数变换。上面的第一个不等式是

P{X≤qF（u）}=F（qF（u））≥u。

要证明第二个不等式，请注意，根据qF（u）的定义，

为每一个人。

设0，我们得到了证明（18）的P{X<qF（u）}≤u。

下面的结果是分位数变换的一个重要应用。

提议9.1以下两种说法是正确的。

1. 假设U是根据（0,1）上的均匀分布而分布的随机变量然后qF（U）有cdf。
2. 假设X是具有连续cdf F的随机变量，则F（X）在（0,1）上具有均匀分布。

证明对于第一部分，首先注意均匀随机变量U的cdf FU满足FU（U）=U

0≤u≤1因此，X=qF（U）的cdf FX由

F x（x）=P{x≤x}=P{qF（U）≤x}=P{U≤F（x）}=F（x）

我们最常用的（17）这证明X有cdf。

对于第二部分，假设F是连续的。注意，对于每一个>0，qF的定义意味着。设0，利用F的连续性，我们推导出F（qF（u））≤u，并与（17）结合，得到F（qF（u））=u，因此对于每0<u<1，我们有

P{F（X）≥u}=P{X≥qF（u）}=1-P{X<qF（u）}=1-P{X≤qF（u）}=1-F（qF（u））=1-u

其中我们使用了F的连续性（这意味着P{X=X}对于每个X）每0<u<1，P{F（X）≥u}=1-u这一事实意味着F（X）∼u（0,1）。

例9.2（在零假设下，与具有连续分布的检验统计量相对应的p值具有均匀分布）。统计假设检验问题通常是根据数据计算相关的检验统计量而形成的。假设Tobs是根据数据计算的统计值的观测值。与检验相对应的p值被定义为在零假设下观察统计值的概率，该统计值比Tobs更极端。通常计算为

p=1-F0（托布斯）

其中F0是零假设下检验统计量的cdf如果F0是一个连续的cdf，那么当Tobs∼F0时，应该清楚p是根据U（0,1）分布的换句话说，在零分布（即Tobs∼F0）下，p值具有标准均匀分布。

# 接缝密度

节点密度用于描述有限组连续随机变量的分布我们关注双变量关节密度（即当存在两个连续变量X和Y时）。对于两个以上的变量，这两种观点是相同的。

以下是关于关节密度要记住的要点：1f（·，·）称为接头密度，如果

f（x，y）≥0，对于所有x，y和。

1. 我们说两个随机变量X和Y有联合密度f（·，·）如果

Z Z Z Z

P{（X，Y）∈B}=f（X，Y）dxdy=I{（X，Y）∈B}f（X，Y）dxdy。

乙

对于R2的每一个子集B我们通常用fX，Y来表示（X，Y）的节理密度。

1. 如果∏是R2中围绕一个点（x0，yo）的一个小区域，我们有（在某些规律性条件下，fX，Y在（x0，y0）处的行为）

P{（X，Y）∈∏}≈（面积∏）fX，Y（x0，y0）。

更正式地说，

P{（X，Y）∈{X

其中，极限值取∏收缩至（x0，y0）。

1. 如果（X，Y）具有关节密度fX，Y，则X的密度由fX给出，Y的密度为fY，其中

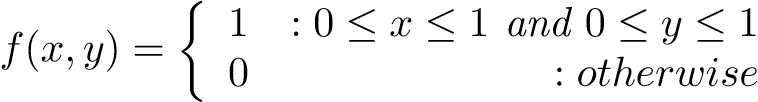
Z Z轴

fX（x）=fX，Y（x，Y）dy和fY（Y）=fX，Y（x，Y）dx。

密度fX和fY分别称为X和Y的边缘密度。

1. 独立性和接头密度：以下陈述是等效的：
   1. 随机变量X和Y是独立的。
   2. 关节密度fX，Y（x）分解为一个单独依赖于x的函数和一个单独依赖于Y的函数的乘积。
   3. fX，Y（x，Y）=所有x，Y的fX（x）fY（Y）。

例10.1考虑功能



检查这确实是一个密度函数这个密度取单位平方上的值1如果随机变量X和Y有这个密度f，那么我们说它们在单位平方上是均匀分布的。使用指示符函数，我们还可以将此密度写为：

f（x，y）=I{0≤x≤1，0≤y≤1}=I{0≤x≤1}I{0≤y≤1}

上面的因式分解立即表明，如果f=f X，Y，那么X和Y是独立的。X和Y的边缘密度是[0,1]上的均匀密度。

问题：如果X，Y有这个密度f，计算P{X2+Y 2≤1}（Ans:π/4）。

例10.2假设X，Y有关节密度

.

表明X的边缘密度由

.

X和Y是独立的吗（回答：不，为什么？）

# 转换下的节理密度

我们处理以下一般性问题假设X和Y具有关节密度fX，Y。假设现在我们考虑两个新的随机变量

（U，V）：=T（X，Y）

其中T:R2→R2是可微可逆函数。就fX，Y而言，U，V的关节密度fU，V是多少？

下面的简单示例将很好地激发一般的想法。

例11.1假设X，Y有关节密度fX，Y当U=X和V=X+Y时，U和V的连接密度是多少？

我们看到（U，V）=T（X，Y），其中T（X，Y）=（X，X+Y）。这个变换T显然是可逆的，其逆由S（u，v）=T-1（u，v）=（u，v-u）给出。为了确定（U，V）在某一点（U，V）的接缝密度，让我们考虑

（十九）

让R表示连接点和的矩形那么上述概率与

P{（U，V）∈R}=P{（X，Y）∈S（R）}。

什么是S（R）区？很容易看出这是连接点（u，v-u），（u）的平行四边形+

还有当δ和很小时，S（R）显然是一个很小的区域

在（u，v-u）附近，它允许我们写

P{（U，V）∈R}=P{（X，Y）∈S（R）}≈fX，Y（U，VU）（S（R）的面积）。

平行四边形S（R）的面积可以计算为δ（使用平行四边形面积等于基部乘以高度的公式），以便



与（19）相比，我们得到了fU，V（u，V）=fX，Y（u，V-u）。

这给出了用（X，Y）的节理密度计算（U，V）的节理密度的公式。

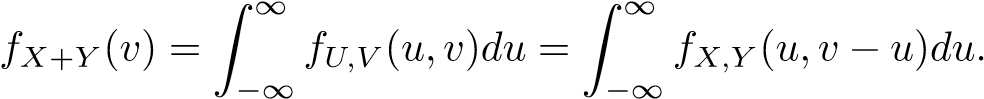
## 绕道卷积

在简单的卷积之后，我们将回到求变换密度的一般问题。

我们在上面的例子中证明了U=X和V=X+Y的连接密度由

fU，V（u，V）=fX，Y（u，V-u）

其中fX，Y是（X，Y）的关节密度因此，我们看到V=X+Y的密度由



假设现在X和Y是独立的。那么fX，Y（x，Y）=fX（x）fY（Y），因此

（二十）

其中最后一个等式是变量v-u=w简单变化的结果。

定义11.2（卷积）给定两个密度f1和f2，我们定义它们的卷积，f1？f2为密度：

Z Z轴

（f1f2）（v）：=f1（u）f2（v-u）du=f1（v-u）f2（u）du。

因此，方程（20）表示，X+Y的密度，其中X∼fX和Y∼fY是独立的，等于fX和fY的卷积。

例11.3。假设X和Y是独立的随机变量，它们与速率参数λ成指数分布。X+Y的分布是什么？

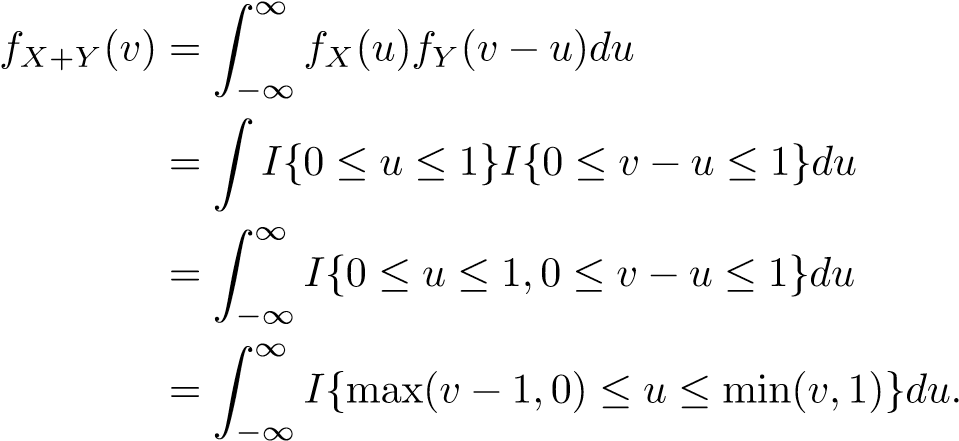
根据卷积公式，

.

这表明X+Y具有形状参数2和速率参数λ的γ分布。

例11.4假设X和Y是在[0,1]上均匀分布的独立随机变量。X+Y的密度是多少？

根据卷积公式，



只有当max（v−1,0）≤min（v，1）时，积分才是非零的，这很容易被视为等于0≤v≤2当0≤v≤2时，fX+Y（v）=最小（v，1）－最大（v－1，0）

可以简化为

：0≤v≤1

：1≤v≤2

0:否则

这叫做三角形密度。

# 转换下的节理密度

在上一个类中，我们根据（X，Y）的关节密度计算（X，X+Y）的关节密度在这堂课中，我们首先计算一对随机变量的线性可逆变换的联合密度，从而推广了这一计算背后的思想。我们还讨论了非线性可逆变换的情形。

在下一小节中，我们将回顾线性变换的一些标准性质。

## 线性变换

通过线性变换L:R2→R2，我们指的是

（21个）

其中M是2×2矩阵，c是2×1向量。上面右手边的第一项涉及2×2矩阵M与2×1向量（分量x和y）的乘法。

我们将2×2矩阵M称为对应于线性变换L的矩阵，并且经常为矩阵M写ML。

当且仅当矩阵M可逆时，（21）中的线性变换L是可逆的。在后半部分中，我们只讨论可逆线性变换下面是线性变换的两个标准属性，您需要熟悉它们才能完成后续工作。

1. 如果P是R2中的平行四边形，那么L（P）也是R2中的平行四边形换句话说，线性变换将平行四边形映射到平行四边形。
2. 对于每个平行四边形P，以下恒等式成立：

L（P）面积

=|检测（ML）|P面积

换言之，L（P）与P的面积之比由矩阵ML的行列式的绝对值给出。

## 可逆线性变换

假设X，Y具有线性可逆变换T:R2→R2的联合密度fX，Y和let（U，V）=T（X，Y）让T的逆变换用S表示。在上一节课的例子中，我们得到T（x，y）=（x，x+y）和S（u，v）=（u，v-u）。假设T是线性的和可逆的，这意味着S也是线性的和可逆的。

为了计算某一点（u，V）的fU，V，我们考虑



对于小δ和。让R表示连接点（）和(

那么上述概率与

P{（U，V）∈R}=P{（X，Y）∈S（R）}。

什么是S（R）区？很明显，现在S（R）是围绕S（u，v）点的一个小区域（δ和很小），所以

P{（U，V）∈R}=P{（X，Y）∈S（R）}≈fX，Y（S（U，V））（S（R）的面积）。

根据上一小节提到的事实，我们现在注意到S（R）是一个平行四边形，其面积等于| det（MS）|乘以R的面积（注意R的面积等于δ）因此我们有



所以我们可以推断

fU，V（u，V）=fX，Y（S（u，V））| detMS |（22）

这里有助于记住MS是对应于线性变换S的2×2矩阵。

例12.1假设X和Y是独立的标准正态随机变量。找到U=X+Y和V=X-Y的接头密度。

我们可以使用T（x，y）=（x+y，x-y）的公式（22），其逆变换为S（u，v）=，并且清楚地给出了对应于S的矩阵。公式（22）给出

.

因为X和Y是独立的标准法线，我们有

以便

这意味着U和V是独立的N（0,2）随机变量。

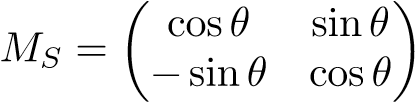
例12.2。假设X和Y是独立的标准正态随机变量。那么（U，V）=T（X，Y）的分布是什么

T（x，y）=（x cosθ-y sinθ，xsinθ+y cosθ）。

在几何上，变换T对应于逆时针方向将点（x，y）旋转一个角θ。T的逆变换S：=T−1由下式给出

S（u，v）=（u cosθ+v sinθ，——usinθ+v cosθ）

这相当于将点（u，v）顺时针旋转一个角度θ。对应于S的矩阵是



公式（22）给出

.

这意味着U和V是独立的随机变量，每个都具有标准正态分布。

接下来，我们将研究在不一定是线性的可微和可逆变换下获得联合密度的问题。

## 一般可逆变换

设（X，Y）为关节密度fX，Y。我们通过（U，V）=T（X，Y）将（X，Y）变换为两个新的随机变量（U，V）关节密度fU，V是多少假设T是可逆的（具有逆S=T-1）并且是可微的。注意S和T不一定是线性变换。

为了计算某一点（u，V）的fU，V，我们考虑



对于小δ和。让R表示连接点（）和(

那么上述概率与

P{（U，V）∈R}=P{（X，Y）∈S（R）}。

什么是S（R）区？如果S是线性的，那么S（R）（如我们前面看到的）将是一个平行四边形对于一般的S，主要的想法是，只要δ和小，区域S（R）可以用平行四边形来近似。这是因为S本身可以通过在区域R上的线性变换来近似，因此，让我们将函数S（a，b）写成（S1（a，b），S2（a，b）），其中S1和S2映射在R2到r中。假设S1和S2是可微的，我们可以近似（a，b）接近（u，v）。

.

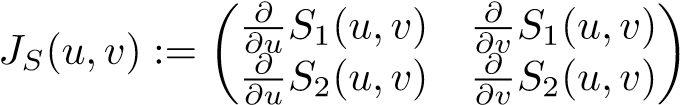
类似地，我们可以近似（a，b）为（a，b）靠近（u，v）。

.

把上述两个方程组合起来，我们得到，对于（a，b）接近（u，v），

.

因此S可以通过线性变换来逼近，矩阵由



对于（a，b）接近（u，v）注意，特别地，当δ和小时，S的线性映射在区域R上是有效的。矩阵JS（U，V）被称为S（U，V）=（S1（u，v），S2（u，v））在点（u，v）处的雅可比矩阵。

因为上面的线性近似，我们可以写。

P{（X，Y）∈S（R）}≈fX，Y（S（u，v））| det（JS（u，v））|（R面积）

这给了我们一个重要的公式

fU，V（u，V）=fX，Y（S（u，V））| detJS（u，V）|（23）

例12.3假设X和Y有关节密度fX，Y。U=X/Y和

V=Y？

我们需要计算（U，V）=T（X，Y）的关节密度，其中T（X，Y）=（X/Y，Y）。这个变换的逆是S（u，v）=（uv，v）那么公式（23）给出

.

因此，U=X/Y的边际密度由

Z Z轴

fU（u）=fU，V（u，V）dv=fX，Y（uv，V）| V | dv。

在特殊情况下，当X和Y是独立的标准正态随机变量时，U=X/Y的密度由

.

这是标准的柯西密度。

√

例12.4。假设X和Y是独立的标准正态随机变量。设r＝x2+y 2，并让x表示向量（x，y）与平面中的正x轴所形成的角。（R，Θ）的接缝密度是多少？

显然（R，Θ）=T（X，Y），其中T的倒数由S（R，θ）=（rcosθ，rsinθ）给出f的密度

（R，Θ）at（R，θ）为零，除非R>0和0<θ<2π公式（23）给出

.

从这里很容易看出，Θ在（0，2π）上均匀分布，R具有密度

fR（r）=re-r2/2I{r>0}。

而且R和Θ是独立的R的密度称为瑞利密度。

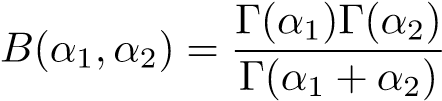
例12.5这里有一个关于伽马分布的重要事实：假设X∼伽马（α1，λ）和Y∼伽马（α2，λ）是独立的，那么X+Y∼伽马（α1+α2，λ）。这可以用独立随机变量和密度的卷积公式来证明。另一个公式使用Jacobian公式导出U和V的联合密度，其中V=X/（X+Y）。这里的相关逆变换S（u，v）=（uv，u-uv），因此雅可比公式给出：

fU，V（u，V）=fX，Y（uv，u（1-V））u=fX（uv）fY（u（1-V））u。

插入fX和fY的相关伽马密度，我们可以推断

.

这意味着U∼γ（α1+α2，λ）这也意味着V∼β（α1，α2），U和V是独立的，以及



在上面的右边，我们有Beta函数。注意，因为Γ（n）=（n-1）因为当n是整数时，上面的公式给出了当α1和α2是正整数时计算β函数B（α1，α2）的方法。

# 不可逆变换下的关节密度

在上一类中，我们研究了根据原始随机变量的联合密度计算连续随机变量变换集的联合密度的雅可比公式。这个公式假定变换是可逆的换言之，如果转换是不可逆的，则公式不起作用然而，基于第一原理的一般方法（我们用来推导雅可比公式）工作良好下面的例子说明了这一点。

例13.1（订单统计）假设X和Y有关节密度fX，Y。U=min（X，Y）和V=max（X，Y）的接头密度是多少？

让我们找出（U，V）在（U，V）处的关节密度。由于U<V，当U≥V时，密度fU，V（U，V）将为零。所以让U<V。对于δ和小的，让我们考虑

.

如果δ和比v-u小得多，则上述概率等于



进一步近似等于



我们已经证明了

)：u<v

0:否则

我们可以把它推广到两个以上随机变量的情况。假设X1，…，Xn是具有关节密度fX1，…，Xn（X1，…，Xn）的随机变量设X（1）≤·························································然后是X（1），…，X（n）的联合分布上述两个变量的计算可以很容易地推广到

<·····<联合国

0:否则

其中和在所有置换π上（即，一个置换和映射{1，…，n}到{1，…，n}的函数上）。

当变量X1，…，Xn是i.i.d（独立且相同分布）时，则从上面可以看出

<·····<联合国

0:否则

# 更多有序统计：固定i的X（i）密度

假设X1，…，X n是具有公共密度f的i.i.d随机变量。在上一节中，我们导出了阶统计量X（1），…，X（n）的联合密度在这里，我们主要讨论确定固定i的X（i）密度的问题

（二十四）

有三种标准的方法可以得出这个结论。

## 方法一

第一种方法将关节密度fX（1），…，X（n）（u1，…，u i-1，u，ui+1，….un）集成到u1，…，ui-1，ui+1，…，un上，以获得fX（i）（u）。更准确地说，

Z Z fX（i）（u）=··n！f（u1）….f（u I-1）f（u）f（ui+1）….f（un）I{u1·····································

首先对u1（在范围（－∞，u2））进行积分，然后对u2（在范围（－∞，u3））进行积分，直至对u1进行积分。然后就联合国，然后就联合国-1，一直到ui+1这将导致（24）。

## 方法二

这种方法使用多项式概率假设我们重复实验n次，n次重复的结果是独立的。假设每个单独的实验都有k个结果，我们用O1，…，Ok表示，并让这些结果的概率由p1，…，pk给出（注意，这些是总和为1的非负数）。

现在让Ni表示结果Oi出现的次数（超过n次重复）（注意N1，…，Nk是和为n的非负整数）。（N1，…，Nk）的联合分布称为参数n和p1，…，pk的多项式分布。这是一个证明

（25个）

当n1，…，nk是和为n的非负整数。

现在让我们回到求X（i）密度的问题考虑概率



对于固定的u和小的δ如果δ是小的，那么这个概率可以用E定义如下的事件E的概率来近似。E是X1，…，Xn中的（i-1）观测值严格小于u，X1，…，Xn中的一个观测值位于[u，u+δ]，X1，…，Xn中的n-i观测值严格大于u+δ的事件。后一种概率是多项式概率公式的特例

（25）当δ很小时，我们得到这个概率等于



式中，F是对应于F的cdf。然后紧跟公式（24）。

## 方法三

在这里，我们首先计算X（i）的cdf，然后对其进行区分以得到pdf。注意



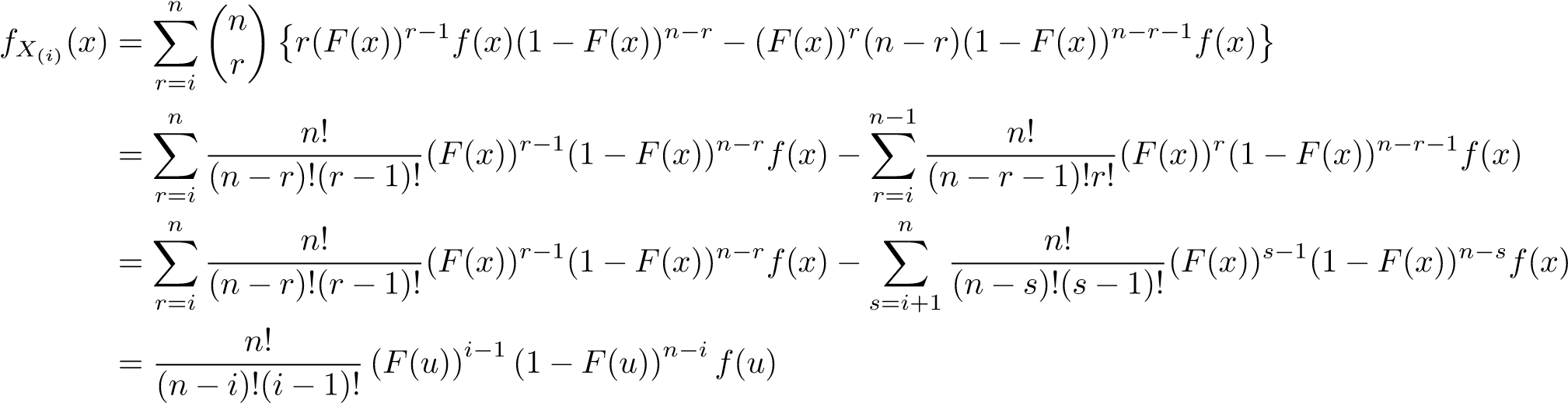
=P{X1的至少i，…，Xn≤x}

n个

=XP{正好是X1的r，…，Xn≤x}

.

为了计算密度，我们必须区分F x（i）与x的关系。这给出了（注意，F的导数是F）



因此我们再次得到公式（24）。

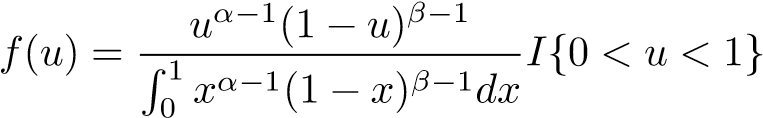
接下来，我们看一些关于单阶统计量密度的公式（24）的特殊例子。

## 统一顺序统计

假设X1，…，Xn是i.i.d，密度为（0,1）。然后公式（24）（插入f（u）=1，f（u）=u表示0<u<1）给出X（i）的以下密度：

对于0<u<1.（26）

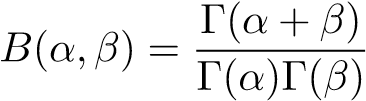
这是参数i和n-i+1的β密度。通常，参数α>0和β>0的β密度由



上面分母中的积分称为β函数：

对于α>0，β>0。

Beta函数通常没有封闭形式的表达式，但我们可以通过公式用Gamma函数来表示



我们在上节课上看到的当α和β是整数时，这个公式允许我们用闭合形式写B（α，β）（注意，Γ（n）=（n−1）！），例如，

.

## 独立制服的最大值

假设X1，…，Xn是独立随机变量，对于某些θ>0，在区间（0，θ）上具有均匀分布。结果表明，最大阶统计量x（n）是最大似然估计。X（n）的密度很容易被认为是（作为（24）的结果）：

.

那么E（X（n））是什么？使用上面的密度公式，

.

因此，这意味着X（n）作为θ的估计量具有-θ/（n+1）的轻微负偏差，并且（（n+1）/n）X（n）是θ的无偏估计量。

## 独立指数的最小值

速率参数λ>0的指数密度（用Exp（λ）表示）由下式给出

f（x）=λe--λxI{x>0}。

它可以说是建模被限制为非负的随机量的最简单密度。它产生于Poisson过程中的到达时间分布（稍后我们研究Poisson过程时将对此进行更多讨论）。

Exp（λ）的cdf很容易被认为是

对于x>0。

假设X1，…，Xn是Exp（λ）的i.i.d观测值。X（1）的密度是多少根据公式（24）：

当u>0时，fX（1）（u）=n（1–（F（u））n–（1f（u）=（nλ）e–（nλ）u。

因此X（1）具有速率参数nλ的指数密度。

# 中心极限定理

我用埃里希·莱曼的《大样本理论的要素》一书的第二章作为我们处理CLT的参考。

中心极限定理（CLT）不是一个单一的定理，而是包含与大量随机变量之和有关的各种结果，这些随机变量经过适当的规范化，具有正态极限分布下面是CLT的最简单版本，这是我们在这个类中主要要处理的版本。

定理15.1（中心极限定理）。假设Xi，i=1,2i.i.d，E（Xi）=μ，var（Xi）=σ2<∞那么

在分布上收敛到N（0,1），其中。

我们将讨论关于CLT的以下几点：

1. “分布趋同”是什么意思？
2. 如何证明CLT？
3. 后果和应用。

非正式地，CLT说，对于I.Id观测值X1，…，Xn，具有有限平均方差和方差的2，N（X，N，N）/α近似（或渐近）N（0，1）。非正式地说，CLT还意味着

1. n（x，n，y）约为n（0，2）。
2. xn近似为n（α，2）。
3. Sn= x1+···+xn近似为n（n，n，2）。
4. Sn n n约为n（0，n＝2）。

√

1. （Sn，n）/（n次）近似为n（0，1）。

注意到这一点可能会有所帮助

E（X′n）=μ，var（X′n）=σ2/n

还有

E（Sn）=nμ，var（Sn）=nσ2。

CLT最显著的特点是它不考虑Xi的分布（只要它们是i.i.d，来自具有有限均值和方差的分布F）因此，在这个意义上，CLT是免费的这使得有可能使用CLT导出在没有特定分布假设的情况下渐近有效的统计过程。为了说明Xi的分布可以是任意的，让我们考虑下面的例子。

1. Burnulixi＝p和：Supposevar（Xi）＝Xi是Ip伯努利随机变量，其概率为p（1～p），使得Ct意味着n（x n n p）/pp（1π）近似为p p。那么

N（0,1）。这实际上被称为De Moivre定理，它在1733年一般CLT之前被证明过1810年，拉普拉斯证明了上述一般CLT。

CLT也意味着Sn大约为N（NP，NP（1～p））。我们知道Sn是按Bin（n，p）分布精确分布的因此，当P是固定的，N是大的时，二项式分布Bin（n，p）与平均NP和方差NP（1～p）的正态分布基本相同。

1. 泊松：假设Xi是i.i.d Poi（λ）随机变量。然后EXI＝LaM= VaR（XI），使得CLT称Sn= x1+Fo.+xn是平均正态的，平均n阶和方差n为α。这里不难证明Sn是作为Poi（nλ）随机变量精确分布的（证明这个！）因此，我们推导出当N大且固定不变时，Poi（n个）与平均n阶和方差n阶正态分布基本相同。
2. Gamma：假设Xi是具有Gamma（α，λ）分布的i.i.d随机变量然后检查EXi=α/λ和var（Xi）=α/λ2。然后从CLT推导出Sn= x1+Fo.+xn近似平均分布，平均nα/α，方差nα/ 2。在最后一类中，我们得到了Sn精确分布为Gamma（nα，λ）因此，当N大且α和α保持固定时，Gamma（Nα，La）与CLT（nα/α，nα/ 2）分布密切相关。
3. 卡方检验假设Xi是具有1个自由度的i.i.d卡方随机变量，即，对于i.i.d标准正态随机变量Z1，Z2，…，Xi=Zi2，。。。。很容易检查Xi是否是Gamma（1/2,1/2）随机变量。这说明X1+···+Xn是伽马（n/2,1/2）。X1+····+Xn的精确分布也称为n自由度的卡方分布（用x2n表示）因此，Ct 2n的分布与n（n，2n）密切相关。
4. 柯西假设Xi是i.i.d标准Cauchy随机变量那么Xi没有有限的均值和方差。因此，CLT不适用于这里事实上，这里可以证明（X1+····+Xn）/n对每个n都有柯西分布。

# 分布收敛

为了理解CLT的确切含义，我们需要理解分布收敛的概念。

定义16.1（分布趋同）假设Y1，Y2，。。。是随机变量，F是cdf。

我们说Yn在分布上收敛于F（或Yn在定律上收敛于F），如

P{y n≤y}→F（y）as n······

对于cdf为连续的y我们用Yn→L F来表示。

换句话说，如果Fn表示Yn的cdf，那么Yn→L F if和only if

Fn（y）→F（y）作为n→∞

对于每一个y，这是F的连续点。

在讨论分布的收敛性时，我们将使用以下约定。

1. 如果F是标准分布的cdf，如N（0,1），那么我们取

Yn→Ln（0,1）

意味着Yn在分布上收敛到N（0，1）的cdf。

1. 对于随机变量Y，我们取

是的→是的

意味着Yn在分布上收敛到Y的cdf。

注意，分布的收敛性是根据cdfs定义的，这使得讨论离散随机变量序列收敛到连续分布成为可能例如，如果Yn在有限集{1/n，2/n，…，1}上具有离散均匀分布，则根据上述定义Yn→L Unif（0,1）但请注意，Yn是离散的，而U（0,1）是连续分布的。

注意，Yn→L Y的定义只要求Fn（Y）在F的连续点Y处收敛到F（Y）（这里Fn和F分别是Yn和Y的cdf）。如果F是一个连续的cdf（如正规或均匀的cdf），那么每个点都是一个连续点，那么Yn→lf等于

每y的P{Yn≤y}→F（y）。

但当F是离散的cdf时，对于Yn→lf，我们不坚持P{Yn≤y}在F是不连续的y点收敛到F（y）。这在如下情况下是有利的。假设每n≥1且Y=0，Yn=1/n。很容易看出

当y 6=0时，P{Yn≤y}→P{y≤y}。

然而，当P{y≤0}=1时，上述收敛性对于y=0不成立，因为P{y n≤0}=0对于每n因此，如果坚持让P{y n≤y}在所有点y（相对于仅连续点）收敛到P{y≤y}，那么Yn=1/n将不会在分布上收敛到y=0（这将是非常不自然的）。这是在分布收敛性定义中引入F的连续点限制的一个理由。

现在让我们分离以下两个特殊的Yn→L Y情况。

1. Y有一个连续的cdf F：在这种情况下，如果P{Yn≤Y}收敛到P{Y≤Y}，则Y→L Y。

这实际上意味着

每y的P{Yn<y}→P{y≤y}

以及

P{a≤Yn≤b}→P{a≤Y≤b}对于每个a和b。

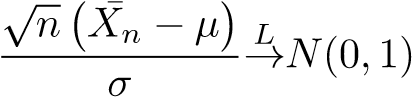
1. Y等于常数。假设极限随机变量Y等于一个常数c，那么Y的cdf（Y）很容易被看作是Y<c时等于0，Y>c时等于1，>L的定义则意味着，当且仅当P{Yn≤Y}收敛到Y<c时为0，Y>c时收敛到1，那么很容易被看作等价于：

0作为n→∞（27）

每大于0在这种情况下，当Y是一个常数时，我们实际上写下并说Yn收敛于c的概率，或者，你可以取（27）作为

今天的主要目标是引入矩生成函数并用它们来证明中心极限定理。让我们首先回顾一下CLT的声明。

定理16.2（中心极限定理）。假设Xi，i=1,2i.i.d，E（Xi）=μ，var（Xi）=σ2<∞那么



式中，X∏n=（X1+····+Xn）/n。

为了理解上述定理的表述，我们首先需要知道符号→L的意思。这是分布收敛的概念，定义如下我们说一系列随机变量Y1，Y2如果P{y n≤y}在函数F连续的每一y收敛为n············································此外，我们还说，如果P{Y n≤Y}在Y的cdf连续的每一Y收敛到P{Y≤Y}，则Yn→ly。

Yn→lY的陈述可能表明，对于大n，Yn接近于Y。这实际上不是真的。Y→L Y只表示Y的分布与Y的分布接近。实际上更适合写Yn→L F，其中F是Y的cdf例如，假设Y∼Unif（0,1），对于n的奇数值，Yn等于Y，对于n的偶数值，Yn等于（1-Y）。然后，显然每个Yn∼Unif（0,1），使得Yn→L Y和Yn→L 1-Y都是真的。但很明显，对于偶数n，Yn不接近Y，对于奇数n，Yn不接近1-Y。

当F是一个连续的cdf（当F是N（0,1）的cdf时，语句

Yn→L F等于

每y的P{Yn≤y}→F（y）。

在这种情况下（即，当F是连续的时），它也遵循

每Y的P{Y<Y}→F（Y）

而且

P{a≤Yn≤b}→P{a≤Y≤b}对于每个a和b。

因此，CLT的确切含义是

)作为n→∞

对于每个y∈R，这里Φ（·）是N（0，1）的cdf同样地，CLT也意味着

)作为n→∞

每a≤b，这与

)如n·····。

现在假设zα/2>0是实线上的点，使得Φ（zα/2）=1-α/2表示0<α<1然后取a=-zα/2和b=zα/2，我们推断

如n·····。

这意味着

是μ的渐近100（1α）%置信区间（假设σ已知）。CLT的应用确保了对X1，X2是此结果所必需的。

# 弱大数定律

√ √

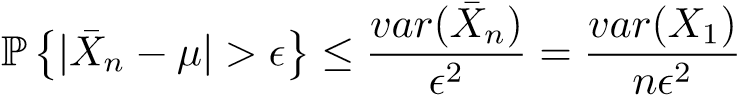
Ct表示n（xn n－ω）/（α）分布在n（0，1）上收敛，这意味着n（xn n ^）/Ω约为n（0，1）。这意味着xn n近似为n（α，2）。由于N（μ，σ2/N）变得越来越集中在单点μ上，因此可以推测X′N收敛到单点μ为N·····这在下面的结果中是精确的，这就是所谓的大数弱定律。

定理17.1（弱大数定律）假设X1，X2，。。。是独立且相同分布的随机变量假设E | Xi |<∞使得EXi定义良好设EXi=μ。那么

如n·····。

回想上一个类，Yn→P c（这里c是常数）意味着当n······等价地，如果F是常数随机变量的cdf，它总是等于c，那么Yn→pc与Yn→lf是相同的。

如上所述的弱大数定律是不可证明的然而，如果我们进一步假设Xi的方差是有限的，则可以给出一个简单的证明在这种情况下，我们可以简单地使用切比雪夫不等式。实际上，切比雪夫不等式给出了



它收敛到零作为n→∞（因为分母中的n）注意，当var（X1）=∞时，这个证明不起作用。

# 矩生成函数

下一步我们将试图证明CLT。我们证明的主要工具是现在介绍的矩母函数。

随机变量X的矩生成函数（MGF）定义为：



对于E（etX）<∞的所有t∈R。注意MX（0）=1。存在随机变量（例如根据标准CAOCY分布分布的），对于每个T 6＝0，Mx（t）是无限的。

例18.1（标准高斯函数的MGF）。如果X∼N（0,1），则其MGF可以简单地计算如下：

.

因此，MX（t）=et2/2对于所有t∈R。

MGFs的基本性质总结如下。

1. 独立随机变量和的因式分解：假设X1，…，Xn是独立的-

那么，丹特

MX1+····+Xn（t）=MX1（t）MX2（t）…MXn（t）。

这是因为

,

最后的平等是独立的结果。

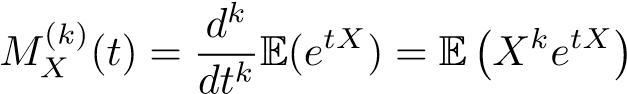
1. 标度：Ma+bX（t）=eatMX（bt），所有t（a和b在此为常数）这很容易证明。
2. mgf决定分布：如果两个随机变量的mgf在包含0的开放区间内是有限且相等的，则它们具有相同的分布（即，到处都有相同的cdf）。这意味着N（0，1）是相同的分布，所有t的MGF等于et2/2。
3. MGFs提供力矩信息：对于k≥1，数字E（Xk）称为第k个力矩

对随机变量X的了解使我们能够很容易地读出X的矩。实际上，MGF的幂级数展开是：

.

因此，X的第k个力矩就是MX（t）的幂级数中的tk系数乘以k！.

或者，我们可以导出力矩E（Xk）作为MGF在0的导数。事实上，很容易看出



以便

.

换句话说，E（Xk）等于MX在0的第k阶导数。因此

MX0（0）=E（X）和MX00（0）=E（X2）

等等。

作为一个例子，我们可以从标准正态分布的

MGF等于et2/2。的确，因为

,

紧接着，当k为奇数且等于时，N（0,1）的第k个矩等于0

当k=2j时。

MGF的最后一个重要特性如下。

5）MGFs与分布收敛性的关系：假设Y，Y1，Y2，。。。正在运行-

在包含零的开放间隔中具有有限mgf的dom变量。假设MY n（t）收敛到MY（t）为n························然后是你→我是。

# 用MGFs证明CLT

让我们回忆一下基本设置。我们有i.i.d随机变量X1，X2，。。。具有平均μ和有限方差σ2。

√

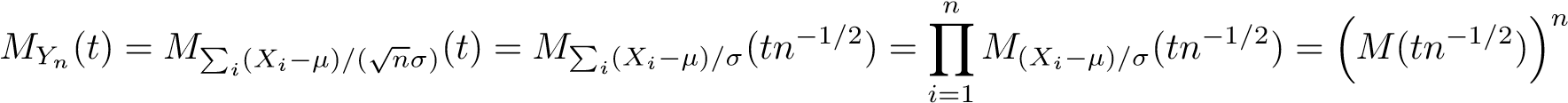
设Yn：=n（X’-n–）/σ我们需要证明1）。从上一节对MGFs的讨论中可以清楚地看出

MYn（t）→et2/2，对于每一个t∈（－∞，∞）。

注意

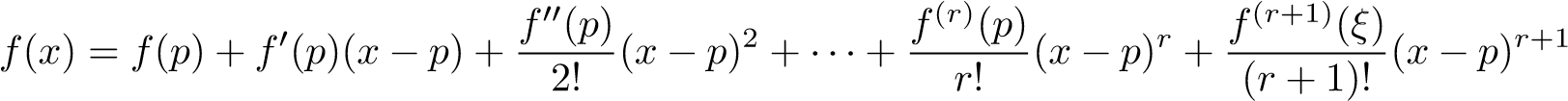
.

因此，



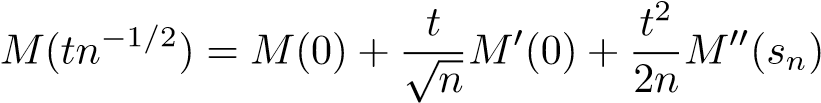
式中，M（·）是（X1∏μ）/σ的MGF我们现在使用泰勒定理将M（tn-1/2）展开到一个约为0的二次多项式。

首先让我们快速回顾一下泰勒定理。这说明对于函数f和f域中的两点x和p，我们可以



式中，ζ是位于x和p之间的某点。此公式要求f在包含p和x的开放区间内具有（r+1）导数。

利用泰勒定理，在r=1，x=tn-1/2和p=0时，我们得到



对于一些介于0和tn-1/2之间的sn。因此，这意味着sn→0作为n·····现在请注意，M（0）=0，M0（0）=E（（X1-μ）/σ）=0。因此我们推断

.

现在请注意

=1作为n→∞。

因此，我们援引以下事实：

n个

（28个）

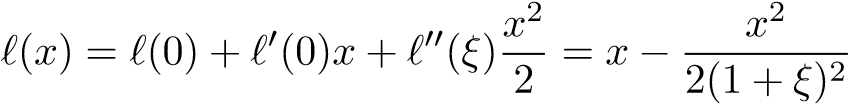
推断出

.

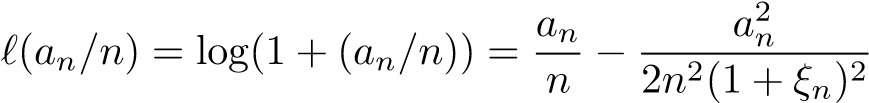
这就完成了CLT假设事实的证明（28）还有待证明。有很多证据证明这一点。这是一个。写

.

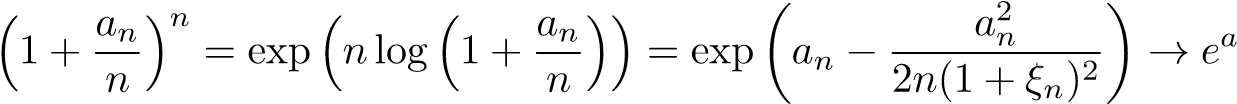
设`（x）：=对数（1+x）关于r=2和p=0的泰勒定理给出了



对于介于0和x之间的某个ζ，取x=an/n，我们得到



对于介于0和an/n之间的某个ζn（因此ζn→0作为n·····）。因此，



如n·····这证明了（28）。

这就完成了CLT的证明请注意，我们默认地假设所有的T都存在X1，…，Xn的矩生成函数，这比XI的方差的存在强得多。如果MGF不是有限的，这个证明就不起作用存在更先进的证明（例如，它工作的特征函数，而不是MGFS），仅在有限方差假设下工作。这些都超出了这个类的范围。

# 关于CLT的两点评论

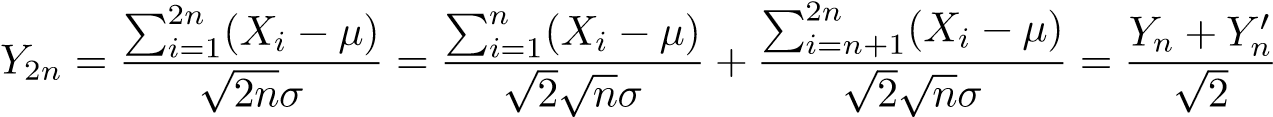
关于CLT的一个自然问题是：为什么N（0,1）（或N（0，σ2））作为独立随机变量（而不是其他分布）和的极限出现？

这可以用多种方式来解释我将在下面提到两个常见的解释。

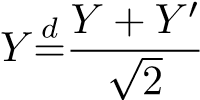
1. CLT计算

.

现在考虑一下：



其中是Yn的独立副本（独立副本意味着与Yn是独立的并且具有相同的分布）。因此，如果对一个随机变量Y来说Yn→ly，那么它必须保持



d 0）/√2具有相同的分布。

其中=表示“分配中的平等”表示Y和（Y+Y

√

很容易看出，如果Y∼N（0，τ2），则Y和（Y+Y 0）/2具有相同的分布相反，值得注意的是，N（0，τ2）分布是唯一具有这种性质的分布（难以证明）。这以及所有n的var（Yn）=1的事实，意味着n（0,1）是Yn唯一可能的极限分布。

1. 对CLT的另一个有趣的解释和解释来自信息论的考虑。注意，对于每个n，随机变量Yn的方差等于1。但是，随着n的增加，Yn的公式中涉及更多的变量Xi因此可以说，当方差保持在1时，Yn的“熵”随n增加。现在有一种形式化的熵的概念，并且有可能表明N（0，1）是最大化熵的分布，其方差等于1。因此，YN的熵随着n的增加而增加（在计算YN时涉及更多的变量Xi），并且最终作为n＝-ω，得到最大熵分布，n（0，1），作为极限。有一种方法可以使这些精确。

# 再论弱律与概率收敛

在最后几类中，我们研究了弱大数定律和中心极限定理弱大数定律如下：

定理21.1（弱大数定律）。假设X1，X2，。。。是独立且相同分布的随机变量假设E | Xi |<∞使得EXi定义良好设EXi=μ。那么

如n·····。

回想一下，从上节课开始，P的定义如下：当n············下面的结果给出了一个直观的关于概率收敛的简单事实然而，这个结果有点难以证明（欢迎您尝试证明这个；结果本身对我们有用，但不是证明）。

引理21.2。如果X1，X2，。。。和Y1，Y2对于两个常数c和d，两个随机变量序列满足Xn→pc和Yn→pc

1. Xn+Yn→P c+d

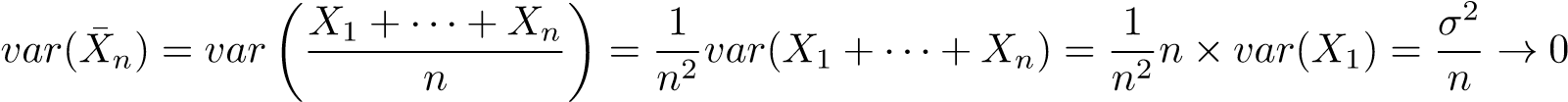
第页

1. Xn-Yn→c-d
2. XnYn→P cd
3. Xn/Yn→P c/d，前提是d 6=0。

现在让我们回到弱大数定律注意（21.1）适用于随机变量X1，X2（只有独立的假设和相同的分布和期望的存在就足够了）。在随机变量具有有限方差的附加假设下，弱定律易于证明。我们在上节课中已经看到的这个证明，是基于切比雪夫不等式的

（二十九）

因为



如n·····结果，从（29）开始，我们得到（29）的左手边收敛到0，这意味着X′n→Pμ。

一般来说，如果Y1，Y2，。。。是一系列随机变量，其中EYn收敛到某个参数θ，var（Yn）收敛到零，然后Yn→Pθ。这在下面的结果中给出。

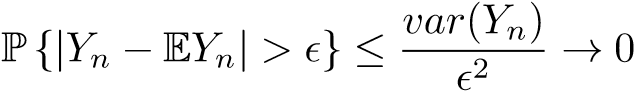
引理21.3假设Y1，Y2，。。。是一系列随机变量

1. EYn→θas n→∞
2. 变量（Yn）→0作为n→∞。

第页

然后Yn→θ为n·····。

证明写Yn=EYn+（Yn-EYn）。切比雪夫不等式（以及var（Yn）·····）给出



每大于0，则Yn-EYn→P 0。这个和EYn→θ意味着（通过引理21.2的第一个断言）

第页

Yn=EYn+（Yn−EYn）→θ。

在数理统计中，当Yn→Pθ时，我们说Yn是θ的一致估计，或者简单地说Yn是θ的一致估计弱大数定律简单地说X′n与E（X1）是一致的更一般地，引理21.3指出，当E（Yn）>0和var（Yn）>0时，Yn与θ是一致的下面的例子展示了另外两种保持一致性的情况。

例21.4假设X1，X2，。。。i.i.d在（0，θ）上的均匀分布，对于固定θ>0。

然后，最大阶统计量x（n）＝max（x1，…，Xn）是一个一致的估计，即x，（n）~p p。我们可以从两个方面来看待这个问题第一种方法是利用上述结果（引理21.3）计算X（n）的均值和方差X（n）/θ是来自Unif（0,1）中n大小的i.i.d样本的最大阶统计量，正如我们在上一类中看到的，X（n）/θ具有β（n，1）分布。因此，使用Beta分布的均值和方差公式（这些公式见wikipedia），我们有

和

.

它给予

和

.

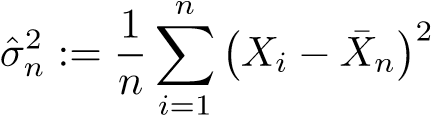
显然，EX（n）收敛到θ，var（X（n））分别收敛到0，这意味着（通过引理21.3）X（n）收敛到θ的概率。

有第二种（更直接的）方法可以看到X（n）→Pθ这包括写作

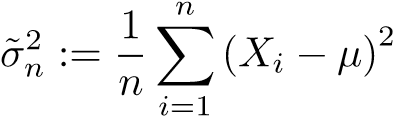
为所有人

当n··································根据概率收敛的定义，这表明X（n）>Pθ。

例21.5。假设X1，X2，。。。是具有平均μ和有限方差σ2的i.i.d观测值那么



是σ2的一致估计量。看到第一个音符



利用弱大数定律，以n····················。这是因为，对于i=1，…，n，σ-n2是i.i.d随机变量Yi=（Xi-μ）2的平均值。因此，弱定律表明，σ-n2在概率上收敛于EY1=E（X1-μ）2=σ2。

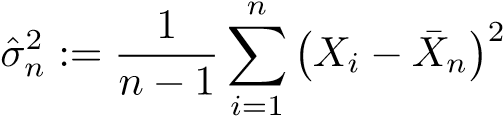
现在，为了证明这一点，我们的想法仅仅是将σˆn2与σ倬n2联系起来。具体操作如下：



上面右边的第一项通过弱大数定律收敛到σ2（注意，σ2=E（X1–）2）。第二项收敛到零是因为（引理21.2）。我们用引理

21.2再次得出σˆn2→Pσ2的结论注意，我们没有假设任何关于X1，X2，…，Xn的分布假设（唯一的要求是它们有均值零和方差σ2）。

顺便说一下，我们还可以通过



系数为1/（n−1）而不是1/n。这也会在概率上收敛到σ2，因为

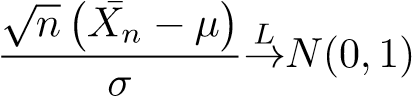
.

由于上面的第一项在概率上收敛于σ2，而第二项在概率上收敛于σ1，因此乘积在概率上收敛于σ2（引理21.2）。

# Slutsky定理、连续映射定理及其应用

我们还研究了中心极限定理的陈述和证明，如下所示。

定理22.1（中心极限定理）。假设Xi，i=1,2i.i.d，E（Xi）=μ，var（Xi）=σ2<∞那么



式中，X∏n=（X1+····+Xn）/n。

这里，分布的收敛性（→L）定义如下：序列Y1，Y2，。。。当P{Yn≤y}收敛于F（y）的连续点的每一个y的F（y）时，称为随机变量的分布收敛于F。虽然分布收敛是用cdfs定义的，但由于独立随机变量和的cdfs不易处理，故用矩母函数证明了CLT。

从统计学角度来看，CLT的一个重要结果是，它给出了μ的渐近有效置信区间事实上，由于CLT，我们

)作为n→∞

每a≤b，这与

)如n·····。

现在假设zα/2>0是实线上的点，使得Φ（zα/2）=1-α/2表示0<α<1然后取a=-zα/2和b=zα/2，我们推断

如n·····。

这意味着

（三十）

是μ的渐近100（1α）%置信区间（假设σ已知）。CLT的应用确保了对X1，X2是此结果所必需的。

区间（30）的问题在于它取决于在统计设置中未知的σ（唯一可用的数据是X1，…，Xn）自然的想法是用自然估计代替σ，如例21.5中定义的ˆσn：

（31个）

这将导致间隔：

（32）Slutsky定理下一步将暗示

1）（33）

这意味着（32）也是μ的渐近100（1α）%置信区间。

定理22.2（Slutsky定理）和Bn→pB，然后



.

我们经常使用的另一个有用的结果是连续映射定理：

定理22.3（连续映射定理）一。假设f是在Y值范围内连续的函数，那么f（Yn）→lf（Y）。

2。假设f在c处是连续的，那么f（Yn）→pf（c）。

这两个结果的一个直接应用是（33），如下所示。

例22.4。设X1，…，Xn为i.i.d观测值，平均值μ，方差σ2。我们需要考虑以下限制性分布：

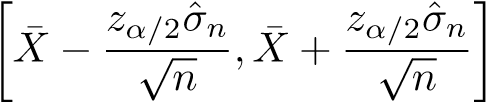
（三十四）

式中，σˆn如（31）中所定义。注意

.

上面右手边的第一项通过通常的CLT在概率上收敛到N（0,1）对于第二项，请注意σn2→Pσ2（如例21.5所证明），因此应用连续映射定理

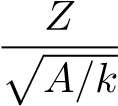
f（x）=pσ2/x意味着。这使得上面的第二项收敛于1的概率因此，我们可以用Slutsky定理来观察，由于上面的第一项在分布上收敛到N（0,1），而第二项在概率上收敛到1，所以随机变量Tn在分布上收敛到N（0,1）。因此，



仍然是100（1-α）%的渐近有效C.I（μ）。注意，我们还没有对X1，…，Xn假设任何分布假设特别是，数据可以是非高斯的。

（34）中的随机变量Tn称为样本t统计量。这个名字来源于t分布或t密度对于给定的整数k≥1，具有k个自由度的t密度是

变量



其中Z∼N（0,1），A具有k度的卡方密度（即A∼x2k），Z和A是独立随机变量。

现在，当X1，…，Xn是i.i.d N（μ，σ2）时，可以显示（稍后我们将看到如何这样做）

和

而且这两个随机变量是独立的因此，当X1，…，Xn为i.i.d n（μ，σ2）时，t统计量Tn具有n−1自由度的tdi分布。

因此

1. 当X1，…，Xn为i.i.d N（μ，σ2）时，样本t统计量Tn具有N-1自由度的t分布。
2. 当X1，…，Xn为i.i.d，平均μ和有限方差σ2（无分布假设）时，t统计量，t N在分布上收敛到N（0,1）。

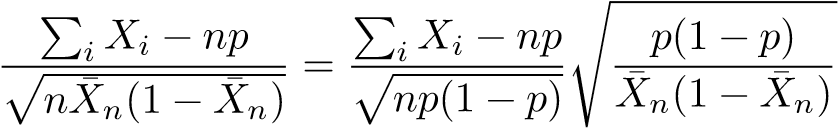
有必要注意到，k自由度的t分布本身在分布上收敛到N（0,1）作为k→∞。

例22.5（伯努利参数估计）假设X1，X2，…，Xn是具有Ber（p）分布的i.i.d然后CLT给出

它给予

如n·····。

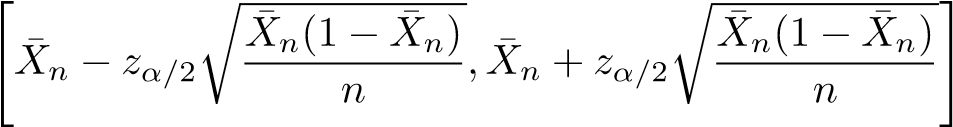
这不会直接导致p的C.I。要做到这一点，用X'代替分母中的p是很自然的。这样做是因为



根据Slutsky定理，上述随机变量在分布上收敛到N（0,1）为了给出更多的细节，我们使用了这样一个事实：上面的第一个随机变量通过CLT在分布上收敛到N（0,1），第二个随机变量在概率上收敛到1（基本上然后使用连续映射定理）。这让我们可以推断

如n·····。

以便



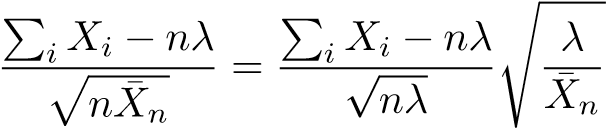
是p的渐近有效100（1α）%C.I。

例22.6（泊松平均估计）假设X1，X2，…，Xn是具有Poi（λ）分布的i.i.d。然后CLT给出

它给予

如n·····。

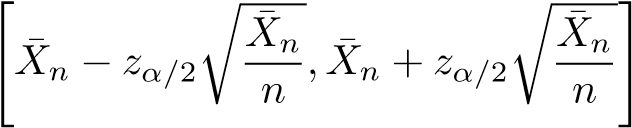
很难将其转换为λ的C.I如果我们能用X′代替分母中的λ，这将变得简单得多。这样做是因为



根据Slutsky定理，上述随机变量在分布上收敛到N（0,1）（这里我们使用的是弱大数定律的结果）这让我们可以推断

如n·····。

以便



是λ的渐近有效100（1-α）%C.I。

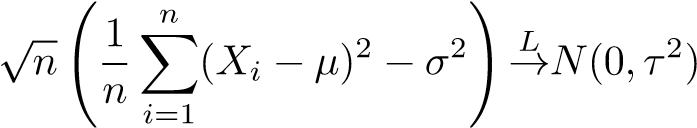
例22.7（样本方差的渐近分布）。设X1，X2为i.i.d，平均值μ，有限方差σ2让

.

我们知道σˆn2→Pσ2。我们也能找到极限分布吗？要做到这一点，写下

.

现在到了中央电视台，



式中，τ2=var（（X1～）2）（当然，我们假设τ2＜∞），根据Slutsky定理，



因此，通过Slutsky定理，我们得到

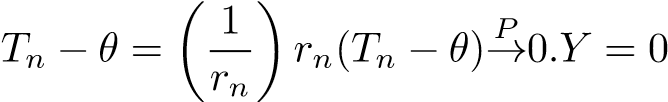
.

Slutsky定理的另一个简单结果是：。

L√P

事实：如果对于某个速率rn······然后Tn→θ。

这直接从Slutsky定理出发，因为



如1/rn→0和rn（Tn−θ）→L Y。

这里是CLT和连续映射定理的一个简单结果假设X1，X2，。。。是

i.i.d平均μ和有限方差σ2的随机变量然后CLT说

.

然后，连续映射定理给出

.

# 三角法

Delta方法是关于分布收敛性的另一个一般性声明，与CLT结合使用时有着有趣的应用。

定理23.1（Delta方法）。如果√n（Tn-θ）L 2），则→n（0，τ

√-g（θ））→Ln（0，τ2（g0（θ））2）N（g（Tn）

G0（Th）存在且非零。

Delta方法非正式地指出，如果Tn具有极限正态分布，则g（Tn）也具有极限正态分布，并给出了g（Tn）渐近方差的显式公式这是令人惊讶的，因为g可以是线性的，也可以是非线性的一般来说，正态随机变量的非线性函数不具有正态分布。但是Delta方法是有效的，因为在假设下

当n（Tn＝Th）＝L n（0，Ta1 2）时，Tn＝pπ，使得Tn至少在大的N附近接近于Th，在一个邻域中，非线性函数G可以用线性函数逼近，这意味着G的行为类似于一个线性函数。事实上，delta方法是近似的结果：

g（Tn）－g（θ）≈g0（θ）（Tn－θ）。

下面是Delta方法的一个应用。

例23.2。假设0≤p≤1是一个固定参数，假设我们要估计p2。假设我们有两种估算p2的选择：

1. 我们可以通过X/n来估计p2，其中X是n个二项试验的成功次数，成功概率为p2。
2. 我们可以通过（Y/n）2估计p2，其中Y是n个二项试验中成功的次数，成功的概率为p。

以上哪一个是p2更好的估计量？为什么Delta方法为这个问题提供了一个简单的答案。注意，到CLT，我们已经

还有那个

Delta方法现在可用于将上述限制语句转换为（Y/n）2的精度语句，如下所示：

.

因此，我们推断（X/n）比（Y/n）2是p2的更好的估计量

p2（1-p2）<4p（1-p）p2

相当于p>1/3。因此，当p>1/3时，X/n比（Y/n）2是p2的更好的估计量，当p<1/3时，（Y/n）2是更好的估计量。

# Delta方法在方差稳定变换中的应用

## 激励方差稳定转换

Delta方法可以应用于方差稳定变换例如，考虑这样一个例子，我们观察到数据X1、X2、…、Xn是具有Ber（p）分布的i.i.d。然后CLT说

（三十五）

不方便的是p也出现在方差项中。这在寻找p的置信区间时是一个麻烦。解决这个问题的一个方法是通过Slutsky定理来观察，

.

这是上节课做的。虽然这个方法是可以的，但是人们可能仍然怀疑是否有可能获得一个具有



当方差c2不依赖于p时，这样的函数f称为方差稳定变换。

对于另一个例子，考虑我们观察到的数据X1，…，Xn是i.i.d，具有Poi（λ）分布的情况然后CLT说

（三十六）

在寻找λ的置信区间时，在上述方差项中出现λ这一事实令人烦恼。正如上节课所做的，我们可以通过观察（通过Slutsky定理）来解决这个问题

.

虽然这个方法是可以的，但是人们可能仍然怀疑是否有可能获得一个具有



其中方差c2不依赖于λ。如果真的能找到这样一个f，则称之为方差稳定变换。

## 方差稳定变换的构造

更一般而言，考虑到结果：

√−θ）→Ln（0，τ2（θ））（37）N（Tn

当方差τ2（θ）依赖于θ时，是否有可能找到一个变换f

√–-f（θ））→L N（0，c2）（38）N（f（Tn）

其中方差c2不依赖于θ然后我们可以说函数f是一个方差稳定变换。

这可以通过Delta方法来实现。实际上，Delta方法表明

√-f（θ））→Ln（0，（f0（θ））2τ2（θ））N（f（Tn）

所以，为了保证（38），我们只需要选择f

（39个）

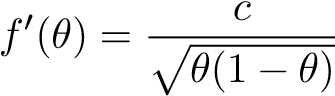
也就是说（不定积分）。

## 回到伯努利的例子

这里我们有X1，…，Xn，它是i.i.d，具有Ber（p）分布，因此通过CLT

√∏p）→L N（0，p（1−p））。n（Xn

因此（37）适用于Tn=X′n，θ=p和τ2（θ）=θ（1−θ）因此公式（38）表明我们选择f作为



√

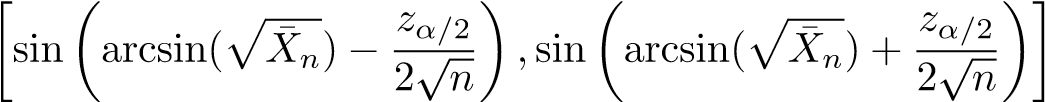
也就是说f（θ）=2carcin（θ）。然后Delta方法保证

.

这意味着

作为n→∞

以便

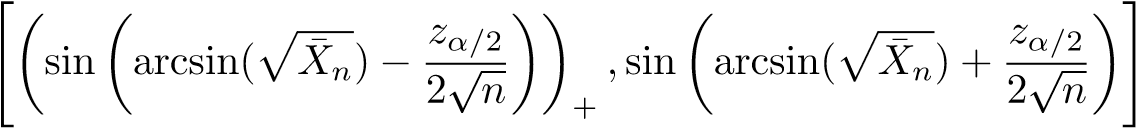


√

约为100（1α）%C. I为p。上述区间的下端可以为负值（注：

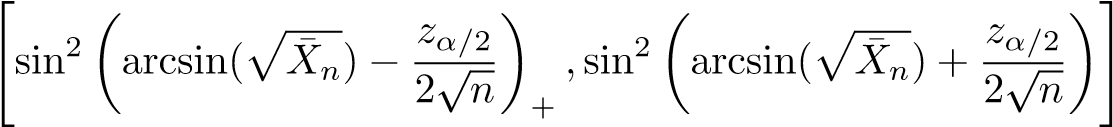
√p’）取0和π/2之间的值，但arcin（pX∏n）–zα/2/（2√n）可以为负），而arcin（Xn

p总是正的。所以如果结果是负的，我们可以用0来代替下端点使用符号x+=max（x，0），我们看到



√

是一个近似100（1α）%C.I。对于P，得到P的置信区间，我们可以简单地对上述区间的两个端点进行平方。这让我们可以推断

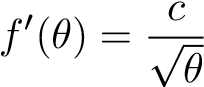


是一个近似100（1α）%C。

## 回到泊松例子

现在让我们回到泊松分布，这里有X1，…，Xn是i.i.d Poi（λ），CLT给出

（36）。因此，Tn=X′n，θ=λ，τ2（θ）=θ方程（39）表明我们选择f作为



√

其中表示f（θ）=2cθ。然后Delta方法保证

（四十）

因此，应用于X′n的平方根变换确保（pX′n的）结果方差不依赖于λ（在极限意义上）。

事实（40）将导致近似的置信区间。确实，（40）立即意味着

作为n→∞

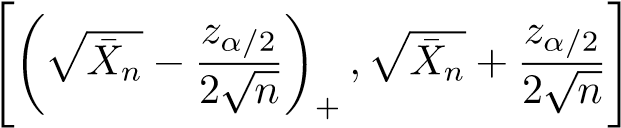
以便



√

是一个近似的100（1α）%C。注意，当λ始终为正时，上述区间的下端点可以为负。所以如果结果是负的，我们可以用0来代替下端点。

再次使用符号x+：=max（x，0），我们看到



√

是一个近似的100（1α）%C。为了得到λ的置信区间，我们可以简单地平方上述区间的两个端点。这让我们可以推断

（41个）

是一个近似的100（1α）%C。

这个区间可以与上节课用Slutsky定理得到的区间相比较。间隔时间是

（42）

间隔（41）和（42）看起来可能不同，但实际上它们对于大n来说彼此非常接近。要看到这一点，请注意这两个间隔的上界之间的差异最多为zα/22/（4n），当n较大时，这一差异非常小（下界也是如此）。

## 卡方检验

现在让我们看看另一个例子，其中的方差稳定转换是对数函数。

假设X1，X2，。。。i.i.d使得Xi/σ2具有一个自由度的卡方分布换句话说，

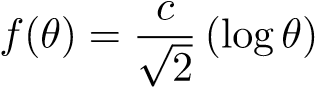
.

因为E（X1）=σ2和var（X1）=2σ4，CLT说

（43个）

我们现在能不能找到一个函数f，使得f（X′n）具有独立于σ2的极限方差因为（43）的形式是（37），其中√Tn=X∏n，θ=σ2，τ2（θ）=2θ2，所以我们可以使用（39）来表示取f so

即f0（θ）=c/τ（θ）=c/（2θ）这给了



让我们得出结论

.

平方根和对数是常见的转换，当存在变化的方差时应用于数据（参见，例如。

## 几何示例

假设X1，X2，。。。i.i.d是否具有参数p的几何分布。回想一下，如果X取1,2，…，X具有Geo（p）分布，。。。有可能

P{X=k}=（1-P）k-1p对于k=1,2，。。。。

获得第一个硬币头部所需的独立投掷次数（硬币头部概率为p）具有Geo（p）分布。

我离开是为了验证X∼Geo（p）

还有。

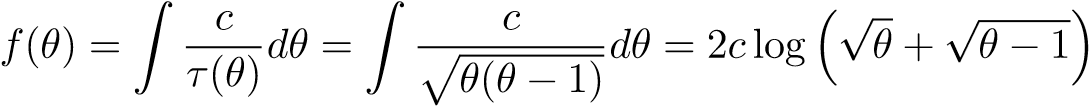
因此，CLT指出，对于i.i.d观测值X1，X2，。。。在地理分布上，我们有

.

什么是X′n的方差稳定变换，即f（X′n）具有恒定渐近方差的变换f是什么要回答这个问题，请注意，上面显示的方程与（37）的形式相同，Tn=X′n，θ=1/p，τ2（θ）=（1−p）/p2。然后我们用θas来写τ（θ）（注意p=1/θ）

.

因此，方差稳定变换由



√ √

前f（θ）=2clog（θ+θ-1）是这里的方差稳定变换

.

# 当g0（θ）=0时的Delta方法

假设√n（Tn−θ）→ln（0，τ2）我们现在对g（Tn）的渐近分布感兴趣Delta方法指出当g0（θ）6=0时

√-g（θ））→Ln（0，τ2（g0（θ））2）。（44）N（g（Tn）

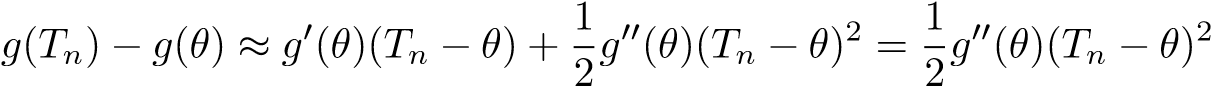
这实质上是泰勒近似的结果：G（Tn）G（Th）G0（Th）（Tn×Th）。如果g0（θ）=0会发生什么在这种情况下，如果将右手侧解释为常数0，即当g0（θ）=0时，则语句（44）仍然正确，如下所示：

打√P

n（g（Tn）–g（θ））→0。

然而，这仅说明g（Tn）－g（θ）与n－1/2相比具有较小的阶数，但当按正确的阶数缩放时，并不能精确地说明确切的阶数和极限分布。为了解决这些问题，我们需要考虑g（Tn）在θ附近的泰勒展开式中的高阶项。在续集中，假设g0（θ）=0，g0（θ）6=0。

在这种情况下，我们做两项泰勒近似：



因此，我们

.

现在根据连续映射定理：



因此我们有

（四十五）

因此，当g0（θ）=0和g0（θ）6=0时，右标度因子为n，极限分布为的标度倍数（注意，极限不是正态分布）。

例25.1假设X1，X2，…，Xn是i.i.d Ber（p）随机变量假设我们要估计p（1-p）自然估计器是X′n（1−X′n）。这个估计器的极限行为是什么？

这可以用Delta方法通过取g（θ）=θ（1-θ）来回答。首先请注意，按照通常的CLT，

.

对于g（θ）=θ（1-θ），注意g0（θ）=1-2θ

所以当p 6=1/2时g0（p）6=0。因此，当p 6=1/2时，Delta方法给出

.

但是当p=1/2时，我们必须使用（45）而不是（44），这就给出了（注意g00（p）=-2）

.

# 条件作用

我们的下一个主题是条件作用这是统计类的一个非常重要的主题。

## 基础

首先让我们看看条件概率的定义。给定P（A）>0的两个事件A和B，我们将B的条件概率定义为

（46）

参见Jim Pitman 2016年201A笔记第10课第1.1节，以获得条件概率定义的一些直观理由。

使用这个条件概率的定义，我们可以看到

P（B）=P（B∩A）+P（B∩Ac）=P（B | A）P（A）+P（B | Ac）P（Ac）

注意这里A和Ac是不相交的事件，其联合是结果的整个空间。一般来说，如果A1，A2是不相交的事件，其联合是Ω，我们有

P（B）=XP（B | Ai）P（Ai）。（47）

i≥1

这被称为全概率定律。

现在让我们来看看贝叶斯规则

.

## 离散随机变量的条件分布、全概率律和贝叶斯规则

考虑两个随机变量X和Θ。假设两者都是离散随机变量然后，可以通过定义条件概率来定义给定的X的条件分布Θ=θ：

（48个）

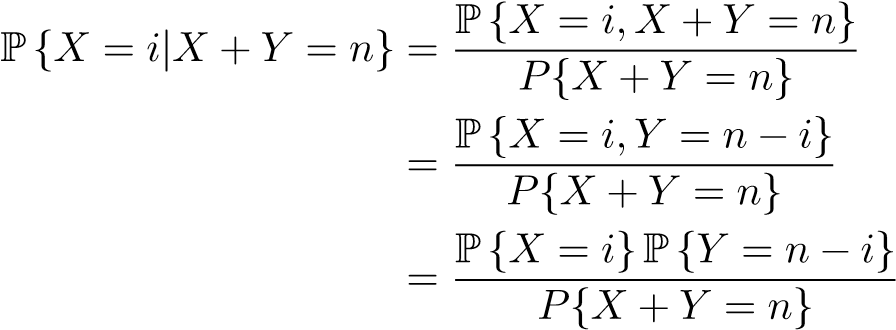
假设P{Θ=θ}>0如果P{X=X |=θ}=0，我们就不会试图定义P{X=X |=θ}。

当x在随机变量x所取的所有值上变化时，概率（51）确定给定的x的条件分布Θ=θ。注意，条件概率P{X=X |=θ}总是在0和1之间，我们得到Px P{X=X |=θ}=1。

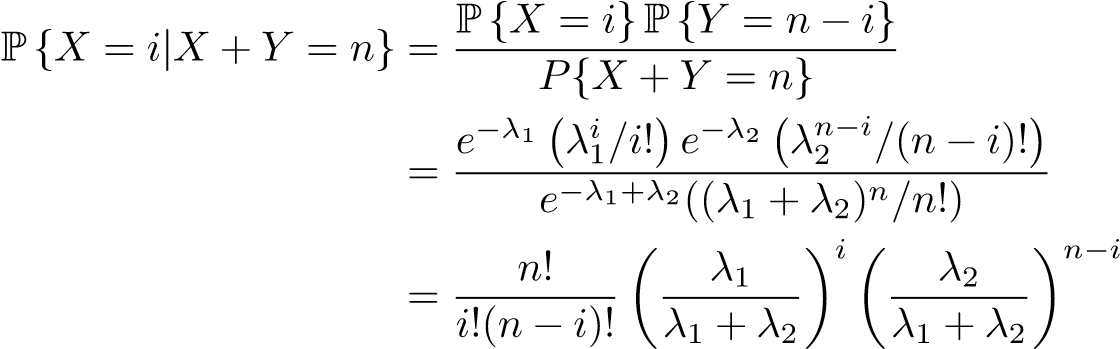
例26.1假设X和Y是分别具有Poi（λ1）和Poi（λ2）分布的独立随机变量对于n≥0，给定X+Y=n时X的条件分布是什么？

我们需要计算P{X=i | X+Y=n}

对于i的各种值，显然只有当i是介于0和n之间的整数时，上述概率才是非零的。因此，我们假设i是介于0和n之间的整数。根据定义



当X和Y分别独立分布为Poi（λ1）和Poi（λ2）时，可以直接计算上述分子。对于分母，我们使用X+Y是Poi（λ1+λ2）的事实（这个事实的证明留作练习）因此我们有



这意味着给定X+Y=n的X的条件分布是参数n和p=λ1/（λ1+λ2）的二项分布。

现在让我们来看看离散随机变量X和Θ的全概率定律和贝叶斯规则由于（47），我们有

P{X=X}=XP{X=X |=θ}P{X=θ}（49）

θ

其中，求和是随机变量Θ所取θ的所有值这个公式允许人们使用P{X=X |=θ}和P{X=θ}的知识来计算P{X=X}我们将（52）称为离散随机变量的总概率定律。

贝耶斯法则是

（五十）

Bayes规则允许计算给定X的条件概率，使用给定X的条件概率和边缘概率的知识。我们将（53）称为离散随机变量的Bayes规则。

例26.2假设N是具有Poi（λ）分布的随机变量同样假设，在N=N的条件下，随机变量X具有Bin（N，p）分布。此设置称为二项式的泊松化找到X的边际分布，也就是给定X=i的N的条件分布是什么？

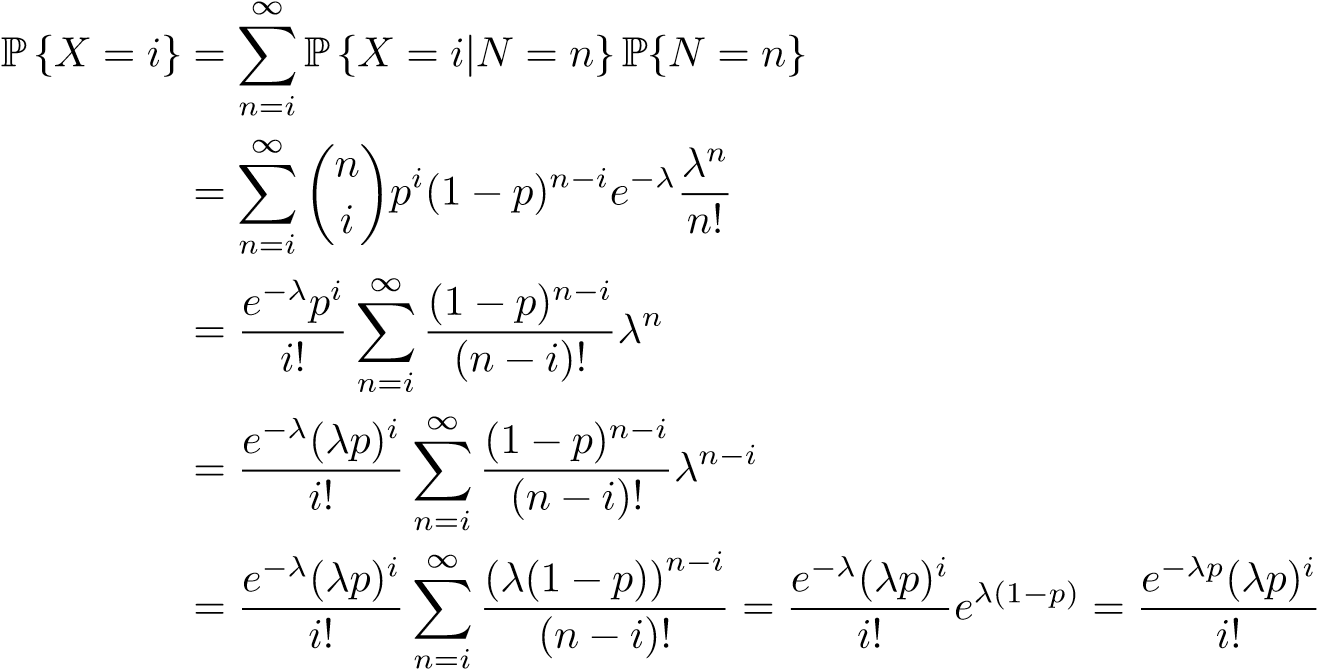
为了找到X的边际分布，我们需要找到每一个i≥0的整数的P{X=i}为此，我们使用总概率定律，它指出

∞

P{X=i}=XP{X=i | N=N}P{N=N}。

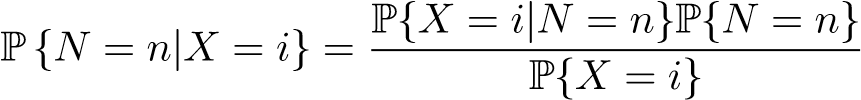
n=0

因为X | N=N是Bin（N，p），只有当0≤i≤N时，概率p{X=i | N=N}才是非零的，因此只有当N≥i时，上述和中的项才是非零的，我们得到



这意味着X具有Poi（λp）分布。

为了找到给定X=i的条件分布，我们需要使用Bayes规则，它声明

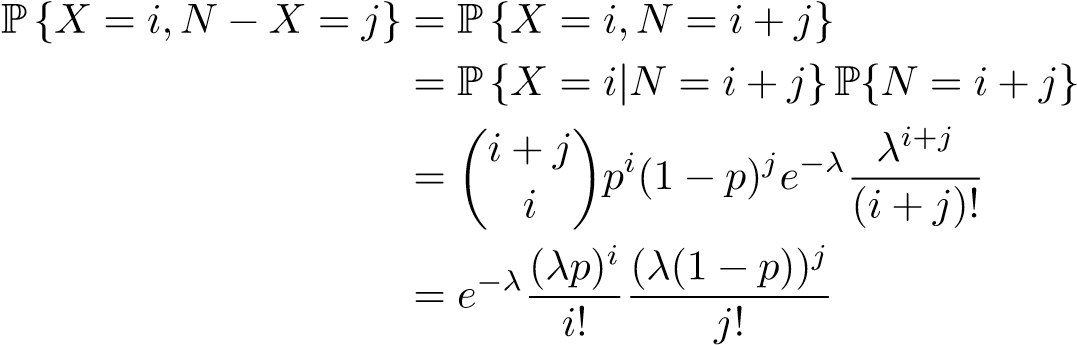


当n≥i时，这仅为非零（否则P{X=i | n=n}将为零）。当n≥i时，我们有

n≥i时。

这意味着，在X=i的条件下，随机变量N以i+Poi（λ（1-p））的形式分布。

在这个例子中，X和N-X的联合分布是什么要计算这个，请注意



注意，这将分解成只涉及i的项和只涉及j的项。因此，这意味着X和N-X是独立的同样从上面的表达式，很容易推断X的边缘分布是Poi（λp）（我们已经通过总概率定律导出了该分布），N-X是Poi（λ（1-p））。

这个例子的背景是当一个人掷硬币的概率为p时，硬币的头部独立于Poi（λ）的次数然后，N表示投掷的总数，X表示头部的数量，N-X表示尾部的数量。因此，我们证明了X和N-X是独立的，并且分别根据Poi（λp）和Poi（λ（1-p））分布。这里X和N-X的独立性特别有趣当一枚硬币被投掷固定的n次时，正面和反面的数量显然是不独立的（因为它们必须和n相加）。但当抛掷次数本身是随机的且具有泊松分布时，则头部和尾部的数目成为独立的随机变量。

# 离散随机变量的条件

在上一课中，我们研究了离散随机变量的条件作用给定两个离散随机变量X和Θ，然后可以通过定义条件概率来定义X的条件分布给定Θ=θ：

（51个）

假设P{Θ=θ}>0如果P{X=X |=θ}=0，我们就不会试图定义P{X=X |=θ}。

当x在随机变量x所取的所有值上变化时，概率（51）确定给定的x的条件分布Θ=θ。注意，条件概率P{X=X |=θ}总是在0和1之间，我们得到Px P{X=X |=θ}=1。

我们还研究了全概率定律和贝叶斯规则总概率定律是

P{X=X}=XP{X=X |=θ}P{X=θ}（52）

θ

其中，求和是随机变量Θ所取θ的所有值这个公式允许人们使用P{X=X |=θ}和P{X=θ}的知识来计算P{X=X}贝耶斯法则是

（五十三）

Bayes规则允许计算给定X的条件概率，使用给定X的条件概率和边缘概率的知识。

本课程的目标是将以上所有内容扩展到X和Θ是连续随机变量的情况。

# 连续随机变量的条件密度

现在考虑两个连续的随机变量X和Θ有一个联合密度fX，Θ（X，θ）回想一下，对于所有x，θ和RR f（x，θ）dxdθ=1，fX，Θ（x，θ）≥0。还记得X和Θ的边缘密度是由

和

我们现在定义X的条件密度，对于θ的固定值，给定Θ=θ。为了在x点定义条件密度，我们需要考虑

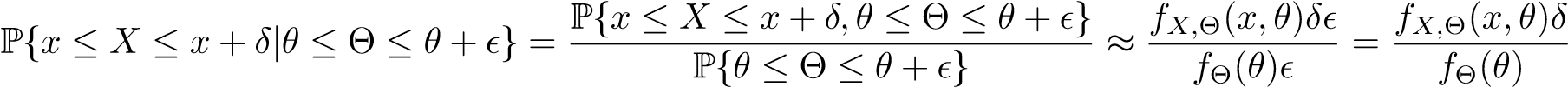
P{x≤x≤x+δ|=θ}（54）

对于较小的δ>0。因为P{Θ=θ}=0（注意，Θ是一个连续的随机变量），我们不能用定义P（B | a）：=P（B∩a）/P（a）来定义这个条件概率但是，直觉上，条件作用

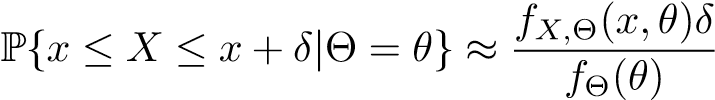
Θ=θ应等于对小的进行调节所以我们可以写



小的。对于上面右边的概率，我们可以使用P（B | A）：=P（B∩A）/P（A）来获得



我们就这样得到了



对于小δ这表明了

（55个）

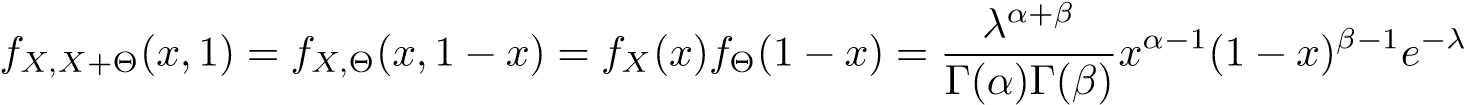
对于给定的X的条件密度Θ=θ。只要fΘ（θ）>0，这个定义就有意义如果fΘ（θ）=0，我们不试图定义fX|920;=θ。

例28.1假设X和Θ分别是具有Gamma（α，λ）和Gamma（β，λ）分布的独立随机变量那么给定X+Θ=1，X的条件密度是多少。

定义（55）给出

.

通过计算转换随机变量密度的雅可比公式，可以检验



对于0<x<1我们以前也看到X+Θ分布为Γ（α+β，λ）因此

.

因此

对于0<x<1。

因此，这意味着

X |（X+Θ=1）∼β（α，β）。

例28.2假设X和Y是独立的Unif（0,1）随机变量。什么是fU | V=V，其中U=min（X，Y）和V=max（X，Y）和0<V<1？

首先要注意

.

当0<u<v<1时，我们知道

fU，V（u，V）=fX，Y（u，V）+fX，Y（V，u）=2。

同时V=最大值（X，Y）∼β（2，1），以便

fV（v）=2vI{0<v<1}。

因此我们有

对于0<u<v。

换句话说，U | V=V在区间（0，V）上均匀分布。

# 条件密度与接头密度成正比

条件密度

（56个）

具有以下重要特性。作为x的函数（并保持θ不变），fX |Θ=θ（x）是有效密度，即。，

Z∞

fX |Θ=θ（x）≥0，fX |Θ=θ（x）dx=1。

−∞

上面的积分等于1，因为

.

因为f x |Θ=θ（x）作为x的函数积分到一，并且因为定义（56）中的分母fΘ（θ）不依赖于x，所以通常写入

fX |Θ=θ（x）ΘfX，Θ（x，θ）。（57）

这里的符号表示“与”成比例，上面的说法是，fX |Θ=θ（x），作为x的函数，与fX，Θ（x，θ）成比例。比例常数必须是fΘ（θ），因为这等于f x的积分值，当x的范围超过（－～～，∞）时，x（x，θ）的积分值。

比例陈述（57）通常使得涉及条件密度的计算简单得多。为了说明这一点，让我们分别回顾示例（28.1）和（28.2）中的计算。

例29.1（例28.1回顾）。假设X和Θ分别是具有Gamma（α，λ）和Gamma（β，λ）分布的独立随机变量那么X的条件密度是多少

X+Θ=1？在（57）之前，

fX | X+Θ=1（X）∮fX，X+Θ（X，1）

=fX，Θ（x，1-x）

=fX（x）fΘ（1-x）

{e∏λx xα-1I{x>0}e∏λ（1-x）（1-x）β-1I{1-x>0}xα-1（1-x）β-1I{0<x<1}

这立即意味着X | X+Θ=1具有参数α和β的β分布。

例29.2（例28.2回顾）。假设X和Y是独立的Unif（0,1）随机变量。什么是fU | V=V，其中U=min（X，Y）和V=max（X，Y）和0<V<1？

写

fU | V=V（u）| fU，V（u，V）

=2fU（u）fV（v）I{u<v}

{fU（u）I{u<v}

=I{0<u<1}I{u<v}=I{0<u<min（v，1）}

因此，对于v<1，给定v=v的条件密度是[0，v]上的均匀密度。对于v>1，给定v=1时U的条件密度不定义为v>1时v的密度等于0。

# 条件密度与独立性

1. 且Θ是独立的，当且仅当fX |Θ=θ=fX对于θ的每个值后一种说法正好等于f x，Θ（x，θ）=fX（x）fΘ（θ）。通过交换X和Θ的角色，也可以得出X和Θ是独立的，当且仅当fΘ| X=X=fΘ对于每个X。

当且仅当给定的X的条件密度对fΘ（θ）>0的所有θ值相同时，也不难看出X和Θ是独立的。

示例30.1（回到Gamma示例）我们之前已经看到，当X∼γ（α，λ）和

1. ∼伽马（β，λ），然后

X |（X+Θ=1）∼β（α，β）。

这也可以直接看到（通过观察X/（X+920）分布为β（α，β），X/（X+920）独立于X+920），如下所示：

.

注意，在上面的最后一步中，我们去掉了X+Θ=1上的条件作用，因为X/（X+Θ）与X+Θ无关。

# 连续随机变量的全概率律和贝叶斯规则

现在让我们回到条件密度的性质。

1. 根据fX |Θ=θ（x）的定义，它直接如下

f x，Θ（x，θ）=fX |Θ=θ（x）fΘ（θ）。

这就告诉我们如何利用已知的边值和条件密度来计算X和Θ的联合密度。

1. 连续随机变量的总概率定律：回顾（52）中离散随机变量的总概率定律。连续随机变量的类似语句是

Z f x（x）=fX |Θ=θ（x）fΘ（θ）dθ。

这直接来自fX |Θ=θ（x）的定义这个公式允许我们利用给定的X的条件密度和X的边际密度的知识来推导X的边际密度。

1. 连续随机变量的Bayes规则：回顾（53）中离散随机变量的Bayes规则连续随机变量的类似语句是

.

这允许使用给定X的条件密度和边缘密度的知识来推导给定X的条件密度。

# 连续随机变量的全概率律和贝叶斯规则

假设X和Θ是具有联合密度fX的连续随机变量。我们已经看到了

.

注意上面右边的分母不包含x，所以我们可以写

fX |Θ=θ（x）ΘfX，Θ（x，θ）。

注意我们也有

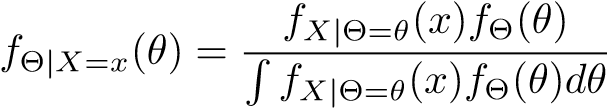
fΘ| X=X（θ）ΘfX，Θ（X，θ）。

以下是条件密度定义的直接结果。

一我们有f x，Θ（x，θ）=fX |Θ=θ（x）fΘ（θ）。2。总概率定律指出

Z f x（x）=fX |Θ=θ（x）fΘ（θ）dθ

三。Bayes规则指出



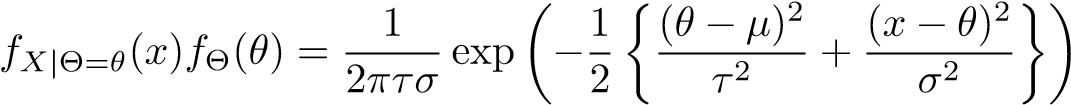
这里有两个例子来说明这些。

例32.1。假设Θ∼N（μ，τ2）和X |Θ=θ∼N（θ，σ2）那么X的边际密度和X=X的条件密度是多少？

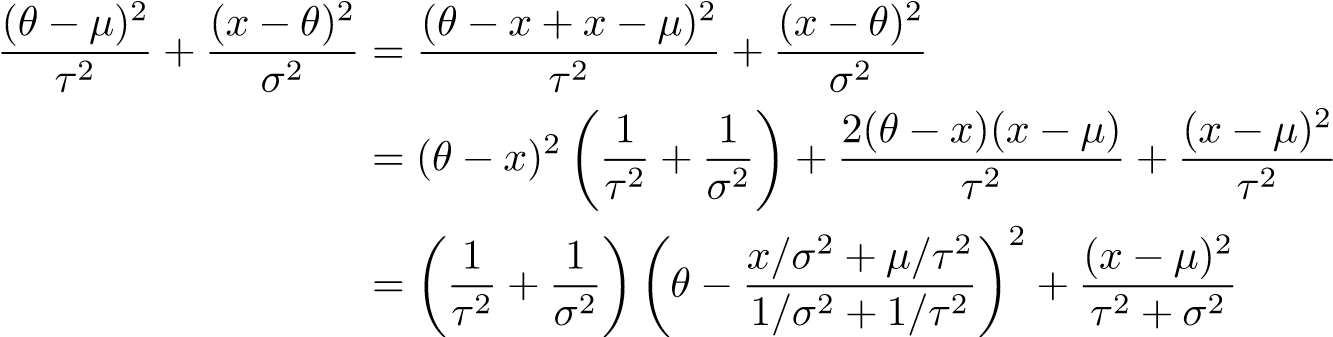
为了得到X的边缘密度，我们用LTP表示

Z f x（x）=fX |Θ=θ（x）fΘ（θ）dθ

现在

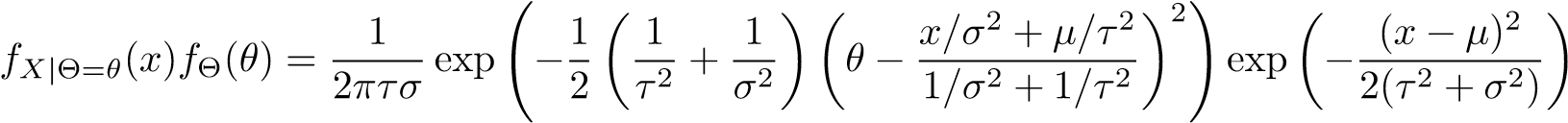


指数中的项可以简化为

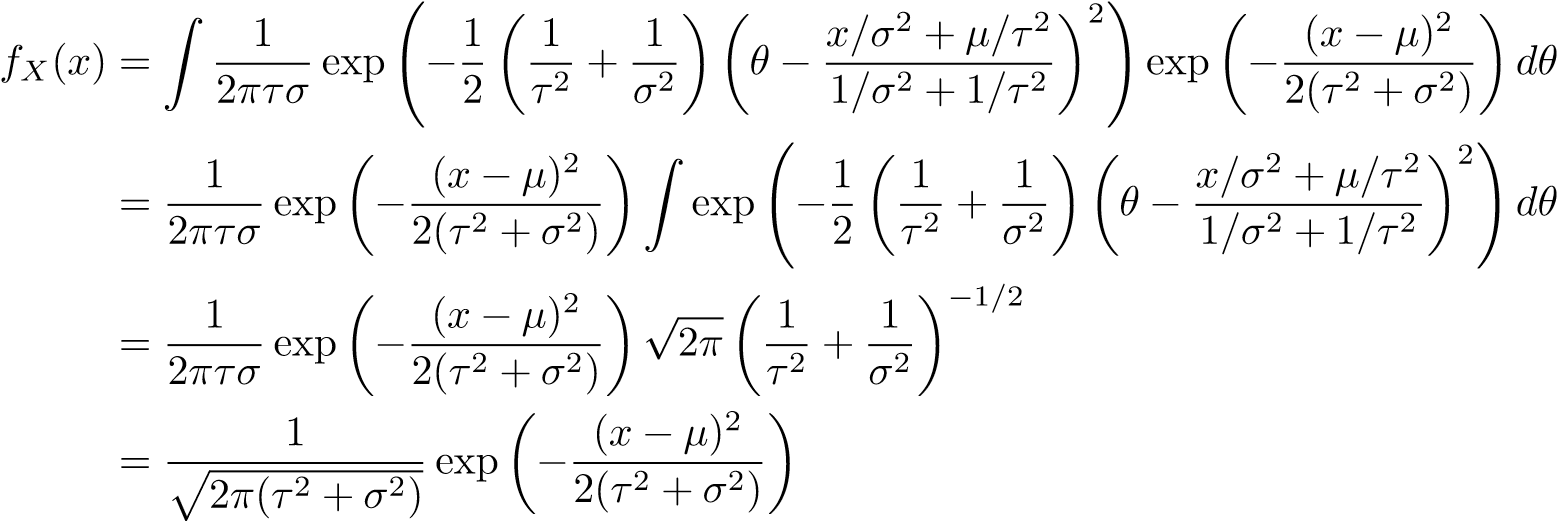


在这里，我跳过了几个步骤来获得最后的等式（完成正方形并简化得到的表达式）。

因此



因此，



它给予

X∼N（0，τ2+σ2）。

为了得到fΘ| X=X（θ），我们使用Bayes规则：

!

也就是说

.

对于均值m和方差v2的正态密度，方差1/v2的逆称为精度。因此，上述计算表明，给定X的条件分布的精度分别等于X的分布和X的分布的精度之和。

在统计术语中，通常称为：

1. 作为未知参数θ的先验分布的边缘分布。
2. X |Θ=θ的条件分布，作为以真参数值为条件的数据分布。
3. 在给定数据的情况下，X=X的条件分布作为X的后验分布。

在这个特定的例子中，后验分布的平均值是先验平均值以及权重与精度成正比的数据的加权线性组合而且，后验精度等于先验精度和数据精度之和，这在非正式意义上特别意味着后验精度比先验精度更高。

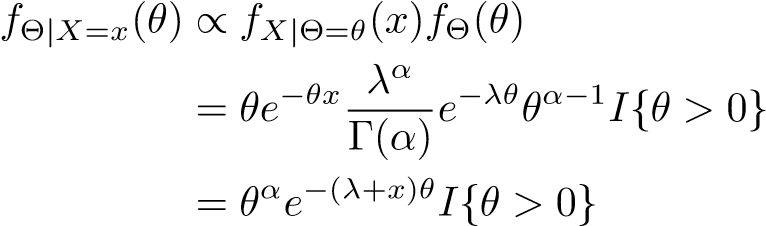
例32.2。设Θ∼γ（α，λ）和X |Θ=θ∼Exp（θ）。X的边际密度是多少，给定X=X的条件密度是多少？

对于X的边缘，使用LTP：

.

这称为Lomax分布，形状参数α>0，比例/速率参数λ>0（请参见。

对于给定X=X的条件分布，通过比例论证



也就是说

Θ| X=X∼伽马（α+1，λ+X）。

# 一般随机变量的LTP和Bayes规则

LTP描述了如何根据给定的X的条件分布和条件分布来计算X的分布。Bayes规则描述了如何在已知X的条件分布和X的条件分布的基础上，计算给定X=X的条件分布。到目前为止，我们研究了当X和Θ都是离散的或它们都是连续的时的LTP和Bayes规则。现在我们还要考虑一个是离散的，另一个是连续的情况。

## X和Θ都是离散的

在这种情况下，我们已经看到LTP是

P{X=X}=XP{X=X |=θ}P{X=θ}

θ

贝耶斯法则是

.

## X和Θ都是连续的

这里是LTP

贝耶斯法则是

.

## X是离散的，而Θ是连续的

LTP为Z

P{X=X}=P{X=X |=θ}fΘ（θ）dθ

贝耶斯法则是

.

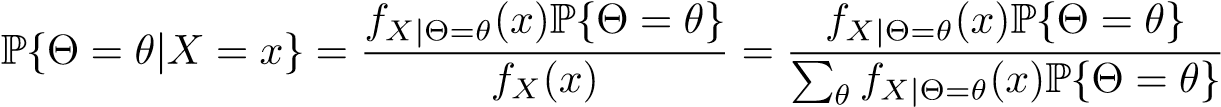
## X是连续的，而Θ是离散的

LTP是

fX（x）=XfX |Θ=θ（x）P{Θ=θ}

θ

贝耶斯法则是

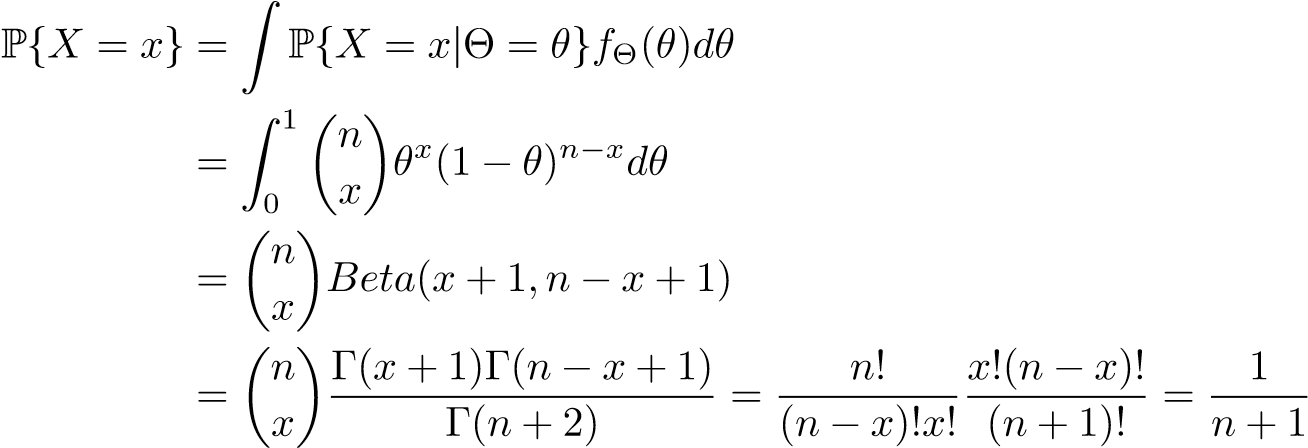


当X给定的条件分布Θ=θ以及Θ的边际分布易于确定（或作为模型规范的一部分给出）并且目标是确定X的边际分布以及X给定的条件分布Θ=X时，这些公式是有用的。

我们现在来看看当X和Θ中的一个是离散的，而另一个是连续的时，LTP和Bayes规则的两个应用。

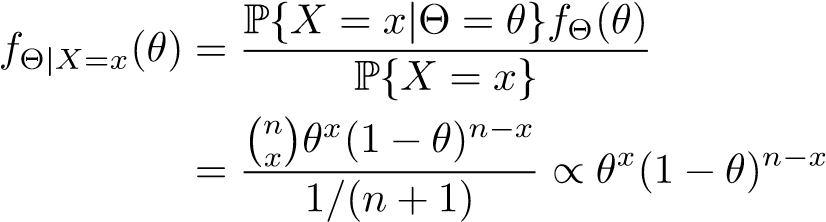
例33.1。假设Θ是（0,1）上的均匀分布，且设X|=θ具有参数n和θ的二项式分布（即，在Θ=θ的条件下，随机变量X以成功概率θ的硬币n次独立投掷中的成功次数分布）那么X的边际分布和给定X=X的条件分布是什么？

注意，在这种情况下，X是离散的（取0,1，…，n中的值），而Θ是连续的（取区间（0,1）中的值）为了计算X的边际分布，我们使用适当的LTP来写（对于X=0,1，…，n）



这意味着X在有限集{0,1，…，n}上均匀分布。

现在让我们计算给定X=X的后验分布。使用Bayes规则，我们得到



对于0<θ<1。从这里开始，紧接着

Θ| X=X∼β（X+1，n−X+1）。

β（α，β）分布的平均值为α/（α+β）。因此，给定X=X的条件分布的平均值（也称为后验平均值）等于

.

因为前面的平均值等于1/2，我们可以写

,

因此，后验平均值介于前验平均值和x/n之间。当n变大时，后验平均值接近x/n。

例33.2（统计分类）在统计分类问题中，随机变量Θ是离散的，X通常是连续的。最简单的情况是当Θ是二进制时让我们这么说吧

P{Θ=1}=P和P{Θ=0}=1−P。

还假设X给定的条件密度为f0，X给定的条件密度为f1，即。，

X | 920;=0∼f0和X | 920;=1∼f1。

使用LTP，我们看到X的边缘密度等于

fX=（1−p）f0+pf1。

换言之，FX是F0和F1的混合，其混合权重等于β的边际概率。

根据Bayes规则，给定X=X的条件分布由



和

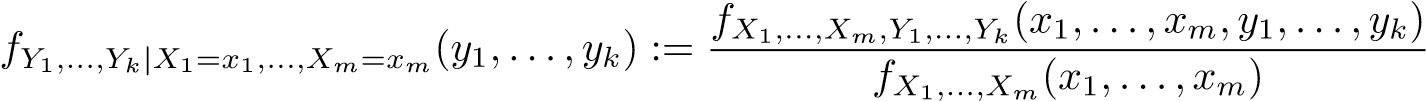
.

这些也被称为给定X=X的后验概率。

# 条件节理密度

给定连续随机变量X1，…，Xm，Y1，…，Yk，条件联合密度Y1，…，Yk

X1=X1，X2=X2，…，Xm=Xm定义为



假设x1，…，xm是这样的：fX1，…，xm（x1，…，xm）>0。

下面是条件连接密度的一些简单但重要的性质。

1. 对于每个x1，…，xm，y1，…，yk，我们都有

fX1，…，Xm，Y1，…，Yk（x1，…，Xm，Y1，…，Yk）=fY1，…，Yk | x1=x1，…，Xm=Xm（Y1，…，Yk）fX1，…，Xm（x1，…，Xm）。

1. 每组随机变量Y1，…，Yn的联合密度满足以下条件：fY1，…，Yn（Y1，…，Yn）=fYn | Y1=Y1，…，Yn−1=Yn−1（Yn）fYn−1 | Y1=Y1，…，Yn−1=Yn−2（Yn−1）…fY2 | Y1=Y1（y2）fY1（Y1）。
2. 这是先前事实的概括。条件关节密度

fY1，…，Yn | X1=X1，…，Xm=Xm（y1，…，Yn）

Y1，…，Yn给定X1=X1，…，Xm=Xm等于乘积

n个

是的

仅供参考：Y1=Y1，…，Yi-1=Yi-1，X1=X1，…，Xm=Xm（Yi）i=1

1. 这可以看作是总条件概率定律：对于随机变量Y1，…，Yk，X1，…，Xm和Θ，我们有

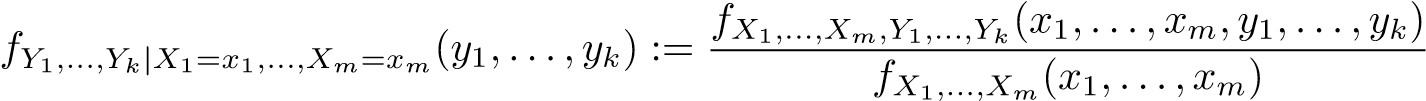
Z轴

fY1，…，Yk | X1=X1，…，Xm=Xm（y1，…，Yk）=fY1，…，Yk，Θ| X1=X1，…，Xm=Xm（y1，…，Yk，θ）fΘX1=X1，…，Xm=Xm（θ）dθ。

我们将在下一节课上讨论上述事实的一些应用。

# 条件节理密度

给定连续随机变量X1，…，Xm，Y1，…，Yk，条件联合密度Y1，…，Yk给定X1=X1，X2=X2，…，Xm=Xm定义为



假设x1，…，xm是这样的：fX1，…，xm（x1，…，xm）>0。

例35.1假设U1，…，Un是在（0,1）上具有均匀密度的独立观测值。当U（n）=U时，U（1），…，U（n-1）的条件连接密度是多少根据定义，

.

根据我们之前计算出的订单统计量的联合分布，首先，只有当0<u1<····················

.

对于上面的分母，我们使用了U（n）∼Beta（n，1）这个事实我们已经证明了

对于0<u1<····<un-1<u<1。

注意，上面的右手边是从区间（0，u）上的均匀分布得出的（n-1）i.i.d观测值的阶统计量的联合密度。因此，我们证明了，在U（n）=U的条件下，U（1），…，U（n-1）的联合密度与（0，U）上均匀分布的（n-1）i.i.d观测的阶统计量的联合密度相同。

下面是条件连接密度的一些简单但重要的性质。

1. 对于每个x1，…，xm，y1，…，yk，我们都有

fX1，…，Xm，Y1，…，Yk（x1，…，Xm，Y1，…，Yk）=fY1，…，Yk | x1=x1，…，Xm=Xm（Y1，…，Yk）fX1，…，Xm（x1，…，Xm）。

1. 每组随机变量Y1，…，Yn的联合密度满足以下条件：fY1，…，Yn（Y1，…，Yn）=fYn | Y1=Y1，…，Yn−1=Yn−1（Yn）fYn−1 | Y1=Y1，…，Yn−1=Yn−2（Yn−1）…fY2 | Y1=Y1（y2）fY1（Y1）。
2. 这是先前事实的概括。条件关节密度

fY1，…，Yn | X1=X1，…，Xm=Xm（y1，…，Yn）

Y1，…，Yn给定X1=X1，…，Xm=Xm等于乘积

n个

是的

仅供参考：Y1=Y1，…，Yi-1=Yi-1，X1=X1，…，Xm=Xm（Yi）i=1

1. 这可以看作是总条件概率定律：对于随机变量Y1，…，Yk，X1，…，Xm和Θ，我们有

Z轴

fY1，…，Yk | X1=X1，…，Xm=Xm（y1，…，Yk）=fY1，…，Yk，Θ| X1=X1，…，Xm=Xm（y1，…，Yk，θ）fΘX1=X1，…，Xm=Xm（θ）dθ。

以下是上述事实的一些应用。

例35.2（自回归过程的联合密度）假设X1，Z2，…，Zn是独立随机变量，Z2，…，Zn分布为N（0，σ2）。定义新的随机变量X2，…，Xn via

Xi=φXi-1+Zi，对于i=2，…，n

其中φ是一个实数进程X1，…，Xn被称为顺序为1的自回归进程给定X1=X1，X2，…，Xn的条件连接密度是多少？X1，…，Xn的连接密度是多少让我们首先计算条件连接密度X2，…，Xn，给定X1=X1。为此，写下

n个

Fx2，…，Xn x1＝x1（x2，…，xn）＝yfxi x1= x1，…，Xi＝1＝Xi＝1（Xi）（58）

i=2

现在对于每个i=2，…，n，请注意

Xi x1= x1，…，Xi＝1＝Xi＝1＝（DωXi＝1＋Zi）x1＝x1，…，Xi＝1＝Xi＝1。

丁

=（πXi＝1＋Zi）x1= x1，…，Xi＝1＝Xi＝1

第二天

=ω- 1＋Zi~（n）（πXi，1，Sig.）。

我们可以去除X1= x1，…，Xi＝Xi＝1以上的条件，因为X1，…，Xi＝1只依赖于X1，Z2，…，Zi，1，因此独立于ZI。

从上面的断言链，我们推断

对于i=2，…，n。

结合（58），我们得到

.

要获得X1，…，Xn，write的连接密度

.

在统计设置中，利用该联合密度通过最大似然估计来估计参数ω和2。然而，对于该模型，使用条件密度X2，…，Xn（给定X1=X1）而不是全关节密度X1，…，Xn更容易。

## 正态先验正态数据模型的应用

现在让我们来看看条件密度公式在正态先验正态数据模型中的应用这里我们首先得到一个具有N（μ，τ2）分布的随机变量Θ。我们也有随机变量X1，…，Xn+1这样

X1，…，Xn+1 |Θ=θ∼i.i.d N（θ，σ2）。

换句话说，以Θ=θ为条件，随机变量X1，…，Xn+1为i.i.d N（θ，σ2）。

让我们首先找到给定X1=X1，…，Xn=Xn的条件分布答案是

（59个）

式中，xn：=（x1+···+xn）/n。让我们看看下面为什么是这样首先请注意，我们在上一节课中已经解决了n=1的这个问题，我们证明了：

.

一般n≥1的结果（59）实际上可以从n=1的上述结果中导出。有两种方式可以看到这一点。

方法一：对n≥1采用数学归纳法。我们已经知道（59）对于n=1是正确的。假设n是真的，我们将试图证明n+1是真的。关键是要注意

（Θ| X1=X1，…，Xn+1=Xn+1）=dΘ| Y=Xn+1（60）

哪里

Y |Θ=∼θ∼N（θ，σ2）和Θ∼∼X1=X1，…，Xn=Xn。

也就是说，（60）表示在观测（n+1）观测值X1=X1，…，Xn+1=Xn+1后的后验与在先前的观测值X1=X1，…，Xn=Xn下观测一个观测值Y=Xn+1后的后验相同。

要正式了解为什么（60）是真的，请注意

fΘX1=X1，…，Xn=Xn，Xn+1=Xn+1（θ）≤fXn+1 | X1=θ，X1=X1，…，Xn=Xn（Xn+1）fΘX1=X1，…，Xn=Xn（θ）

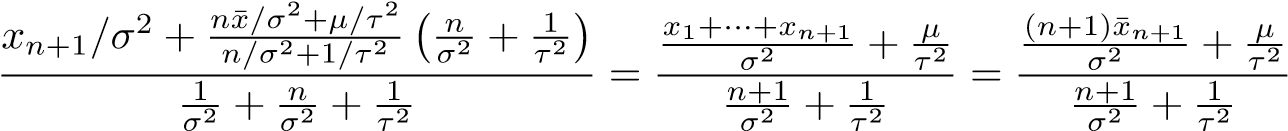
=f xn+1 | X1=X1，…，xn=xn（θ）。

第一等式是条件密度性质的结果。上面的第二个等式是Xn+1独立于X1，…，Xn条件为Θ的结果。

语句（60）允许我们使用n=1的结果和（59）适用于n的归纳假设

和

并且x=xn+1，我们推断出Θ| X1=X1，…，xn+1=xn+1是具有平均值的正态分布

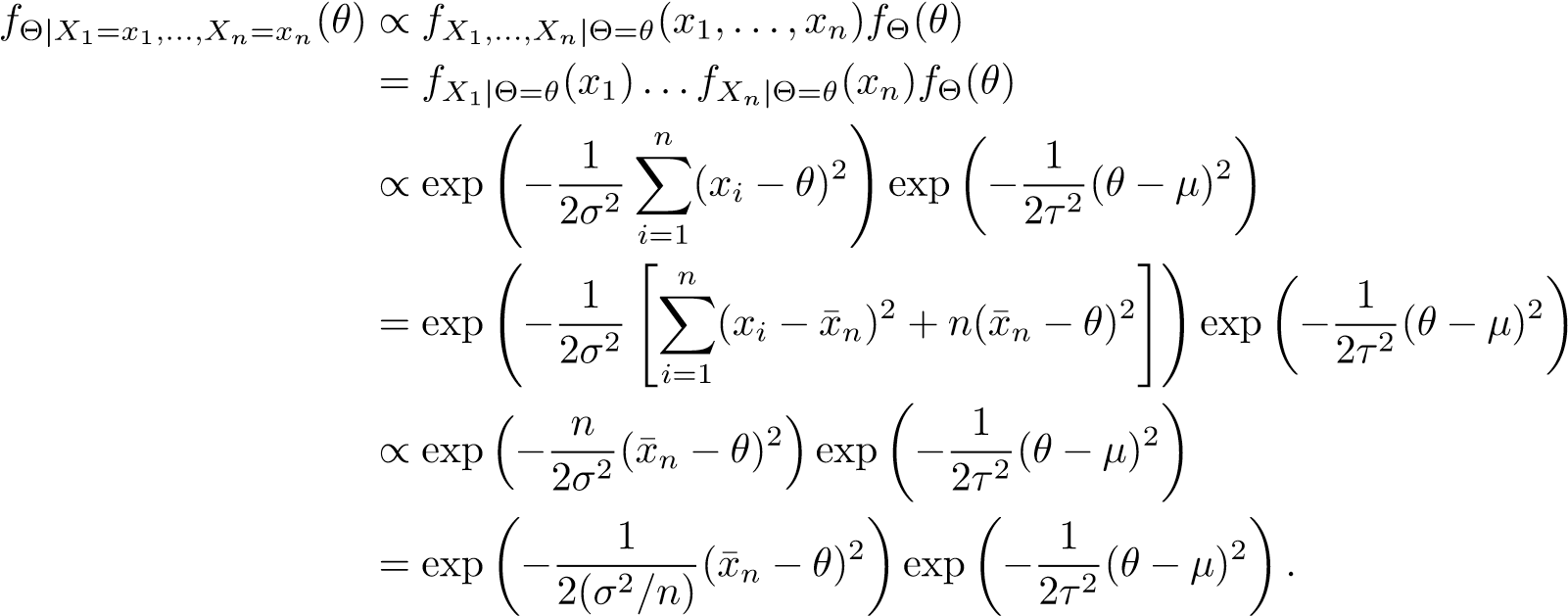


和变化

.

这证明了n+1的（59）（59）的证明是通过归纳法完成的。

方法二证明（59）的第二种方法通过书写更直接地进行：



这与我们之前对n=1的计算类似。唯一的区别是x现在被∏x n代替，而σ2被σ2/n代替。因此，应用于x→x∏n和σ2→σ2/n的n=1结果应产生（59）这证明了（59）。

现在让我们计算Xn+1的条件密度，给定X1=X1，…，Xn=Xn为此，我们可以用总条件概率定律来写

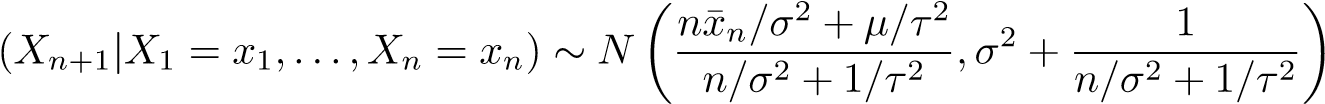
Z轴

f Xn+1 | X1=X1，…，Xn=Xn（x）=fXn+1 | X1=θ，X1=X1，…，Xn=Xn（x）fΘX1=X1，…，Xn=Xn（θ）dθ

Z轴

=f Xn+1 |Θ=θ（x）fΘX1=X1，…，Xn=Xn（θ）dθ

这又类似于n=1问题中X的边缘密度的计算（其中答案是X∼n（μ，τ2+σ2））。唯一的区别是先前的N（μ，τ2）现在被（59）给出的后验密度所代替因此，我们得到



# 条件期望

给定两个随机变量X和Y，给定X=X的Y的条件期望（或条件平均值）表示为

E（Y | X=X）

定义为给定X=X时Y的条件分布的期望值。

我们可以写

Y是连续的

：如果Y是离散的

更普遍地说

R g（y）fY | X=X（y）dy:如果y是连续的

E Y X：如果Y是离散的

还有

Y是连续的

：如果Y是离散的

关于条件期望，最重要的事实是迭代期望定律（也称为总期望定律）。我们下次再看。

## 迭代/总期望定律

总期望定律指出

X是连续的

（）=P X E（Y | X=X）P{X=X}：如果X是离散的

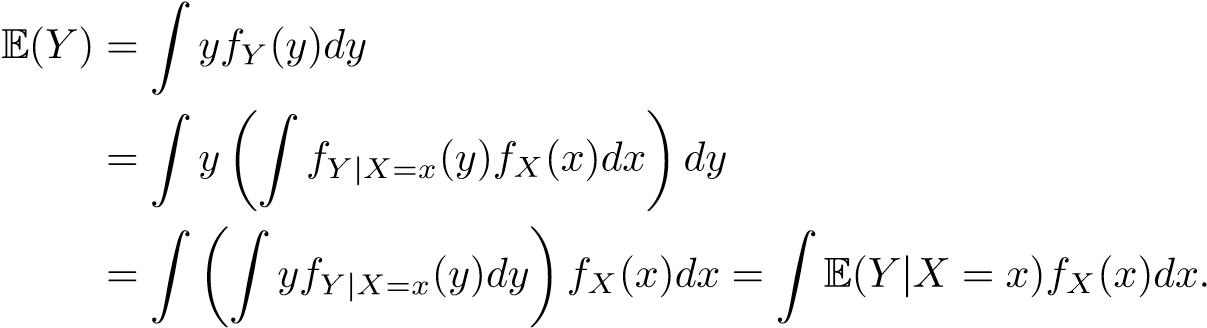
基本上，总期望定律告诉我们如何利用给定X=X的条件期望的知识来计算E（Y）的期望。注意与总概率定律的相似性，总概率定律规定了如何利用给定Y的条件分布的知识来计算Y的边际分布

X=X。

总期望定律可以作为总概率定律的结果来证明。下面给出Y和X是连续的证明在其他情况下（当Y和X中的一个或两个是离散的）的证明是相似的，作为一个练习。

总期望定律的证明：假设Y和X都是连续的那么

Z E（Y）=yfY（Y）dy。根据全概率定律，我们有



证明了总期望定律。

有一种更简洁的形式来表述总期望定律，这使得我们有理由称之为迭代期望定律我们下次再看。注意E（Y | X=X）依赖于X。换句话说，E（Y | X=X）是X的函数。让我们用h（·）表示这个函数：

h（x）：=E（Y | x=x）。

如果我们现在把这个函数应用到随机变量X，我们得到一个新的随机变量h（X）。这个随机变量用简单的E（Y | X）表示，即。，

E（Y | X）：=h（X）。

注意，当X是离散的时，这个随机变量E（Y | X）的期望变成

.

当X是连续的，E（Y | X）的期望值是

Z Z轴

E（E（Y | X））=E（h（X））=h（X）fX（X）dx=E（Y | X=X）fX（X）dx。

观察到，这些期望中的右手边正是总期望定律中右手边的术语。因此，总期望定律可以重新表述为

E（Y）=E（E（Y | X））。

因为在右手边有两个期望，所以总期望定律也被称为迭代期望定律。

迭代期望定律有许多应用下面给出几个简单的例子，我们将探讨风险最小化的应用。

例36.1。考虑一根长的棍子。在一个随机点X处断开它，该点被均匀地选在杆的长度上。然后在任意点Y再次折断木棍，该点也均匀地选择在木棍长度上。最后一块的预期长度是多少？

根据问题的描述，

Y | X=X∼Unif（0，X）和X∼Unif（0，`）

我们需要计算E（Y）首先注意E（Y | X=X）=X/2，这意味着

E（Y | X）=X/2。因此根据迭代期望定律，

E（Y）=E（E（Y | X））=E（X/2）=`/4。

## 迭代/总期望定律

在上一节课中，我们研究了迭代期望定律（或总期望定律），它指出

E（Y）=E（E（Y | X））。

在上面的右侧，E（Y | X）是通过将函数h（X）：=E（Y | X=X）应用于随机变量X（即E（Y | X）=h（X））而获得的随机变量。

迭代期望定律有许多应用下面给出几个简单的例子，我们将探讨风险最小化的应用。

例36.2。假设X，Y，Z是i.i.d Unif（0,1）随机变量找到P{X≤Y Z}的值根据迭代期望定律，

P{X≤Y Z}=E（I{X≤Y Z}）=E[E（I{X≤Y Z}Y Z]=E（Y Z）=E（Y）E（Z）=1/4。

例36.3（i.i.d随机变量的随机数之和）假设X1，X2，。。。是i.i.d随机变量，E（Xi）=μ还假设N是一个离散随机变量，它接受{1,2，…，}中的值，并且它独立于X1，X2定义

S:=X1+X2+····+XN。

换句话说，S是随机变量Xi的随机数（N）的和迭代期望定律可用于计算S的期望值，如下所示：

E（S）=E（E（S | N））=E（Nμ）=（μ）（EN）=（EN）（EX1）。

这个事实实际上是一个称为Wald的身份的一般结果的特例。

## 总期望定律在统计风险最小化中的应用

迭代期望律在统计风险最小化问题中有着重要的应用。其中最简单的问题如下。

问题1：给定两个随机变量X和Y，X的函数g∗（X）是什么使

R（g）：=E（g（X）－Y）2

所有的功能由此产生的随机变量g∗（X）可以被称为Y的最佳预测因子，作为X的函数，用期望平方误差表示。

为了找到g\*，我们使用迭代期望定律来编写

R（g）=E（g（X）－Y）2=EnEh（g（X）－Y）2 | Xio

使内部期望最小化的值g∗（x）：

EH（Y·G（x））2×x= Xi

简单地说就是g∗（x）=E（Y | x=x）。

这是因为当c在c∗=E（Z）时c在R上变化时，E（Z-c）2最小化。因此，我们证明了在所有函数g上最小化R（g）的函数g∗（X）是由

g∗（X）=E（Y | X。

因此，就期望平方误差而言，X的函数最接近Y，由条件平均E（Y | X）给出。

现在让我们考虑一个不同的风险最小化问题。

问题2：给定两个随机变量X和Y，X的函数g∗（X）是什么使

R（g）：=E | g（X）–Y|

所有的功能由此产生的随机变量g∗（X）可以被称为Y的最佳预测因子，作为X的函数，用期望绝对误差表示。

为了找到g\*

R（g）=E | g（X）–Y |=E{E[| g（X）–Y | X]}

使内部期望最小化的值g∗（x）：

E[| Y−g（x）| x=x]

简单地由给定X=X的Y的任何条件中值给出。这是因为当c在Z的任何中值处变化超过R时，E | Z−c |最小化。为此，假设Z具有密度f并写入

Z轴

E | Z−c |=| Z−c | f（Z）dz

Z c Z∞

=（c-z）f（z）dz+（z-c）f（z）dz

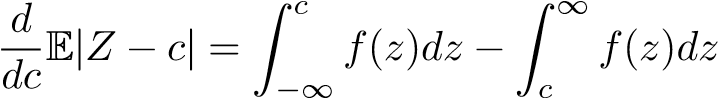
-∞c

Z c Z c Z∞Z∞

=c f（z）dz-zf（z）dz+zf（z）dz-cf（z）dz。

－～－～∞控制中心

关于c的区别，我们得到



因此，当c是中值时，E | Z−c |的导数将等于零这表明，当c是Z的中值时，c 7→E | Z−c |最小化。

因此，我们证明了在所有函数g上最小化R（g）的函数g∗（x）是由给定x=x的任何条件平均值给出的。因此，给定x=x的条件平均值是就预期绝对误差而言最接近Y的x的函数。

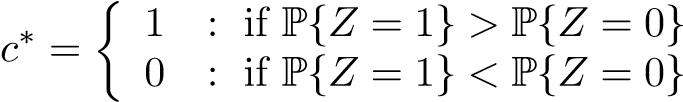
问题3：假设Y是一个二进制随机变量，取值0和1，X是任意随机变量。X的函数g∗（X）是什么使

R（g）：=P{Y 6=g（X）}

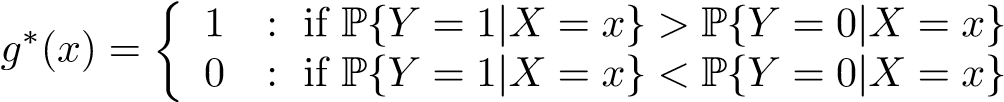
所有的功能为了解决这个问题，再次使用迭代期望定律来编写

R（g）=P{Y 6=g（X）}=E（P{y6=g（X）| X}）。

在上面的内部期望中，我们可以把X当作一个常数，使得这个问题类似于二元随机变量Z在c∈R上最小化P{Z 6=c}，很容易看出P{z6=c}在c\*处最小化，其中



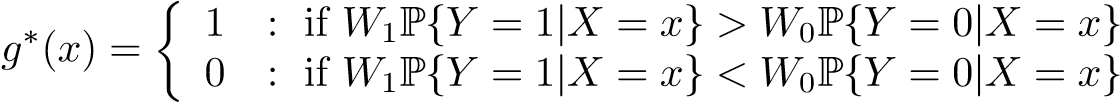
如果P{Z=1}=P{Z=0}，我们可以将c\*取为0或1从这里可以（通过迭代期望定律）推断出最小化P{Y 6=g（X）}的函数g∗（X）由



问题4：再次假设Y是二进制的，取值为0和1，X是任意随机变量X的函数g∗（X）是什么使

R（g）：=W0P{Y 6=g（X），Y=0}+W1P{y6=g（X），Y=1}。

使用类似于前面问题的参数，推断以下函数使R（g）最小化：



在上述四个问题中使用的论点（通过迭代期望定律）可以总结如下。使

R（g）：=EL（Y，g（X））

所有函数g由

g∗（x）=c∈R上E（L（Y，c）| x=x）的最小值。

# 条件方差

给定两个随机变量Y和X，Y给定X=X的条件方差被定义为Y给定X=X的条件分布的方差，

.

与条件期望一样，条件方差V ar（Y | X=X）也是X的函数，我们可以把这个函数应用到随机变量X上，得到一个新的随机变量，我们用V ar（Y | X）表示。注意

V ar（Y | X）=E（Y 2 | X）－（E（Y | X））2。（61）

与总期望定律类似，也有总方差定律这个公式说明

V ar（Y）=E（V ar（Y | X））+V ar（E（Y | X））。

为了证明这个公式，把右手边展开为

E（V ar（Y | X））+V ar（E（Y | X））=EnE（Y 2 | X）－（E（Y | X））2o+E（E（Y | X））2－（E（E（Y | X））2

=E（E（Y 2 | X））-E（E（Y | X））2+E（E（Y | X））2-（E（Y））2=E（Y 2）-（EY）2=V ar（Y）。

例37.1我们以前见过

X |Θ=θ∼N（θ，σ2）和ΘN（μ，τ2）=X∼N（μ，σ2+τ2）。

当然，这意味着

E（X）=μ，V ar（X）=σ2+τ2。

利用总期望和总方差定律，可以直接证明如下。

E（X）=E（E（X | 920;））=E（Θ）=μ

和

V ar（X）=E（V ar（X |Θ））+V ar（E（X |Θ））=E（σ2）+V ar（Θ）=σ2+τ2。

例37.2（i.i.d随机变量的随机数之和）。假设X1，X2，。。。是i.i.d随机变量，E（Xi）=μ，V ar（Xi）=σ2<∞还假设N是一个离散随机变量，它接受{1,2，…，}中的值，并且它独立于X1，X2定义

S:=X1+X2+····+XN。

我们以前见过

E（S）=E（E（S | N））=E（Nμ）=（μ）（EN）=（EN）（EX1）。

利用总方差定律，我们可以计算V-ar（X）如下。

V ar（S）=E（V ar（S | N））+V ar（E（S | N））=E（Nσ2）+V ar（Nμ）=σ2（EN）+μ2V ar（N）。

.

# 随机向量

在这一节中，我们将有限个随机变量看作一个随机向量，并讨论了随机向量均值和协方差的一些基本公式。

随机向量是一个向量，其条目是随机变量。设Y=（Y1，…，Yn）T为随机向量它的期望EY被定义为一个向量，其第i项是Yi的期望，即EY=（EY1，EY2，…，EYn）T。

Y的协方差矩阵，用Cov（Y）表示，是一个n×n矩阵，其（i，j）项是Yi和Yj之间的协方差。关于Cov（Y）的两个重要但简单的事实是：

1. Cov（Y）的对角线项是Y1，…，Yn的方差更具体地说，矩阵Cov（Y）的第（i，i）项等于var（Yi）。
2. Cov（Y）是一个对称矩阵，即Cov（Y）的第（i，j）项等于（j，i）项这是因为Cov（Yi，Yj）=Cov（Yj，Yi）。

以下公式很简单，但非常有用。他们的证据留作练习。

1. 对于每个确定性矩阵A和每个确定性向量c，E（A Y+c）=AE（Y）+c。
2. Cov（A Y+c）=每个确定性矩阵A和每个确定性向量c的ACov（Y）。

这里有一个例子说明了这些公式的使用。

例38.1对于长度为n（或n×1阶）的向量a和n×1随机向量Y，我们有

V ar（aTY）=Cov（aTY）=aTCov（Y）a.（62）

这是写公式的简洁方法

.

注意（62）是公式Cov（a Y+c）=ACov（Y）AT的直接结果。

# 随机向量

在上一节课中，我们开始研究随机向量一个p×1随机向量Y由p个随机变量Y1，…，Yp即Y=（Y1，…，Yp）T组成。

平均向量Y由EY=（EY1，…，EY p）T给出，Y的协方差矩阵由p×p矩阵给出，p×p矩阵的（i，j）项等于Yi和Yj之间的协方差注意Cov（Y）的对角线项是Y1，…，Yp的方差还要注意Cov（Y）是一个对称矩阵。

上节课我们还学习了以下两个重要公式：

1. 对于每个确定性矩阵A和每个确定性向量c，E（A Y+c）=AE（Y）+c。
2. Cov（A Y+c）=每个确定性矩阵A和每个确定性向量c的ACov（Y）。

例39.1（白噪声）如果随机变量Z1，…，Zp的均值为零，方差为1且不相关，则称其为白噪声。设Z为分量为Z1，…，Zp的随机向量当且仅当EZ=0且Cov（Z）=Ip（这里Ip是p×p恒等式矩阵）时，Z的分量是白噪声。

作为上面第二个公式的结果，我们看到

var（aTY）=aTCov（Y）a，对于每个p×1向量a。

因为方差总是非负的，这意味着Cov（Y）满足以下性质：

aTCov（Y）a=var（aTY）≥0，对于每个p×1向量a。（63）

现在回想一下线性代数的以下定义：

定义39.2。设∑表示p×p对称矩阵。

1. 当aT∑a≥0时∑为正半正定。
2. 如果在∑a>0时，对于a∈Rp，a为6=0，则∑为正定。

从这个定义和事实（63）可以看出，每个随机向量Y的协方差矩阵Cov（Y）是对称的和半正定的。

然而Cov（Y）不一定是正定的。要了解这一点，只需对随机变量Y1取p=2和Y=（Y1，-Y1）T当a=（1,1）时，很容易看出aTCov（Y）a=vr（aTY）=vr（Y1+Y2）=0。

但是如果Cov（y）不是正定的，则存在6＝0，使得V AR（ATY）＝ATCOV（y）a＝0。这必然意味着在（Y-μ）=0时，其中μ=E（Y）。换句话说，随机变量Y1，…，Yn必须满足一个线性方程。因此我们可以说：Cov（Y）是正定的，当且仅当随机变量Y1，…，Yn不满足线性方程。

# 对称矩阵的绕道谱定理

由于Cov（Y）是一个对称的半正定矩阵，因此在处理协方差矩阵时，关于这类矩阵的一些标准事实是有用的特别地，我们将利用对称矩阵的谱定理。在研究谱定理之前，我们需要回顾正交基的概念。

## 正交基

定义40.1（正交基）。Rp中的正交基是一组p向量u1，…，up in Rp具有以下特性：

1. u1，…，up是正交的，即hui，uj i：=uTi uj=0，对于i 6=j。
2. 每个ui都有单位长度，即每个i的kuik=1。

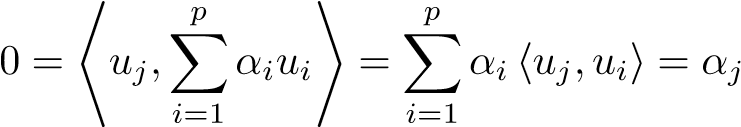
正交基的最简单例子是e1，…，en，其中ei是第i个位置的向量1和所有其他位置的向量0。

每个正交基u1，…，up满足以下性质：

1. u1，…，up是线性独立的，因此构成Rp的基础（这解释了在正交基的定义中存在“基础”一词）想知道为什么这是真的，假设

α1u1+···+αpup=0，对于某些α1，…，αp.（64）

取上述等式两边与uj的点积（对于固定j），我们得到



因为只有当i=j和huj，uji=1时，uii才是非零的。因此（64）意味着，对于每个j=1，…，p和u1，…，up，αj=0是线性独立的，因此形成Rp的基础。

1. 以下公式适用于每个向量x∈Rp：

第页

x=Xhx，uiiui.（65）

i=1

要了解为什么这是真的，首先注意前面的属性意味着u1，…，up构成Rp的基，这样每个x∈Rp都可以写成一个线性组合

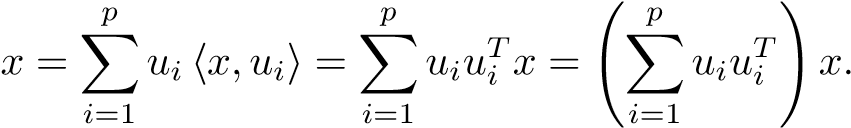
x=β1u1+···+βpup

u1的，…，向上现在取两边有uj的点积证明βj=hx，uji。

1. 公式

（66个）

保持其中Ip是p×p恒等式矩阵要了解为什么这是真的，请注意（65）可以重写为



由于上述恒等式的两边对于每个x都是相等的，所以我们必须有（66）。

1. 假设U是p×p矩阵，其列是向量u1，…，up。那么

UTU=UUT=Ip。

要了解为什么这是真的，请注意UTU的第（i，j）项等于（根据正交基的定义）当i 6=j时为零，否则为1。另一方面，UUT=Ip语句与（66）相同。

1. 对于每个向量x∈Rp，公式

第页

kxk2=Xhx，uii2

i=1

持有。想知道为什么这是真的，就写

磷

kxk2=xTx=xtutx=kUTxk2=X（uTi X）2=Xhx，uii2。

i=1 i=1

## 谱定理

定理40.2（谱定理）。假设∑是一个p×p对称矩阵然后存在一个正交基U1，…，UP和实数，1，…，，

（六十七）

谱定理通常也用下面的另一种形式来写。假设U是p×p矩阵，其列是向量u1，…，up还假设∧是p×p对角矩阵（对角矩阵是非对角项都为零的矩阵），其对角项为λ1，…，λp，则（67）等价于

σ=U∧UT和UT∑U=λ。

下面是谱定理的一些直接结果：

1. 对于每1≤j≤p，我们有

σuj=λjuj和。

这些直接来自（67）。上面的第一个恒等式意味着每个λj是具有特征向量uj的∑的特征值。

1. 在表示式（67）中，特征值λ1，…，λp是唯一的，而特征向量u1，…，up不一定是唯一的（对于每个uj，可以用-uj替换，如果λ1=λ2，则u1和u2可以由u1和u2跨度内的任意一对正交单位范数向量u1，u∮2替换）。
2. σ的秩正好等于非零的λ0j的个数。
3. 如果所有的λ1，…，λp都是非零的，则∑具有全秩，因此是可逆的而且，我们可以写

.

1. 如果∑是半正定的，那么（67）中的每个λj都是非负的（这是

0）。

1. 半正定矩阵的平方根：如果∑是半正定的，那么我们可以定义一个新的矩阵

.

很容易看出∑1/2是对称的，半正定的，满足∑1/2（∑1/2）=∑。我们将∑1/2称为∑的平方根。

1. 如果∑是正定的，那么（67）中的每个λj都是严格正的（这是

0）。

## 谱定理的三个应用

40.3.1每个对称半正定矩阵都是协方差矩阵

我们已经看到每个随机向量Y的协方差矩阵Cov（Y）是对称的和半正定的。结果表明，这句话的逆也成立，即对于某个随机向量Y，每一个对称半正定矩阵等于Cov（Y）。为了解释这一点，假设∑是任意p×p对称半正定矩阵。回想一下，通过谱定理，我们定义了∑1/2（∑的平方根），这是一个对称的非负定矩阵，因此∑1/2∑1/2=是∑。

现在假设Z1，…Zp是不相关的随机变量，都有单位方差，让Z=（Z1，…，Zp）T是相应的随机向量因为Z1，…，Zp是不相关的并且具有单位方差，所以很容易看出Cov（Z）=Ip。假设Y=1/2Z，那么

Cov（Y）=Cov（1/2Z）=∑1/2Cov（Z）（1/2）T=1/2（In）∑1/2=σ。

因此，我们从一个任意的半正定矩阵∑开始，证明了它等于某个随机向量Y的Cov（Y）。

因此，我们可以总结协方差矩阵的以下性质。

1. 每个随机向量的协方差矩阵都是半正定的。
2. 每一个半正定矩阵等于某个随机向量的协方差矩阵。
3. 除非随机变量Y1，…，Yn满足一个精确的线性方程，否则它们的协方差矩阵是正定的。

40.3.2美白

给定一个p×1随机向量Y，我们如何将其转换为p×1白噪声向量Z（回想一下，Z是白噪声意味着EZ=0，Cov（Z）=Ip）这种转变被称为美白。如果Cov（Y）为正定，则可以进行美白。确实，假设∑：=Cov（Y）是正定的，具有谱表示：

.

是正定的这一事实意味着，对于每个i=1，…，p，λi>0。在这种情况下，很容易看出

1/2是可逆的

.

此外，很容易检查∑-1/2∑∑-1/2=Ip。使用这个，很容易检查Z=

是白噪声。实际上EZ=0和

.

因此，谱定理被用来定义矩阵（Cov（Y））-1/2，该矩阵可用于白化给定的随机向量Y。

40.3.3随机向量的第一主成分

设Y为p×1向量我们认为单位向量a∈Rp（单位向量是范数等于1的向量）是Y-if的第一主成分

每单位向量b的var（aTY）≥var（bTY）。

换言之，单位向量A在所有单位向量B上最大化BTY的方差。

假设∑：=Cov（Y）具有谱表示（67）假设在不丧失一般性的情况下，（67）中出现的特征值λ1，…，λp按降序排列，即。，

λ1≥λ2≥····≥λp≥0。

然后证明向量u1是Y的第一主分量。要看到这一点，只需注意



对于每个单位向量b，

.

因此u1是Y的第一主成分注意，第一主成分不是唯一的确实，√

-u1也是第一主成分，如果λ1=λ2，则u2和（u1+u2）/2也是第一主成分。

# 最佳线性预测

考虑具有有限方差的随机变量Y，X1，…，Xp。我们想在X1，…，Xp的基础上预测Y。给定基于X1，…，Xp的Y的预测因子g（X1，…，Xp），我们通过

R（g）：=E（Y-g（X1，…，Xp））2。

我们在上一节课中看到，最佳预测因子（即使R（g）最小化的函数g\*由条件期望给出：

g∗（x1，…，xp）：=E（Y | x1=x1，…，xp=xp）。

这种有条件的期望往往是一个相当复杂的量。例如，在实际估计中，需要大量关于变量X1，…，Xp，Y的数据。

我们现在考虑一个相关的问题，仅基于X1，…，Xp的线性函数来预测Y。具体而言，我们考虑β0+β1X1+····+βp X p=β0+βT X形式的预测（其中β：=（β1，…，βp）T和X=（X1，…，Xp）T）就X1，…，Xp而言，Y的最佳线性预测（BLP）是线性函数

具有

哪里最小化

L（β0，…，βp）=E（Y-β0-β1X1-····-βpXp）2

超过β0，β1，…，βp。

通过微积分直接使L最小，可以得到和β\*的显式公式取关于β0，β1，…，βp的偏导数，并将其设为零，我们得到以下方程：

)=0（68）

和

=0，i=1，…，p.（69）

上面的第一个方程意味着这是一个均值为零的随机变量。利用这个，我们可以将第二个方程重写为

)=0表示i=1，…，p

这和

)=0，i=1，…，p.（70）

重新排列上面的内容，我们得到

第页

X∗

βi Cov（Xi，Xj）=i=1，…，p时的Cov（Y，Xi）。

j=1

在矩阵表示法中，我们可以将其重写为

Cov（X）β∗=Cov（X，Y）与。

这里Cov（X，Y）是p×1向量，有项Cov（X1，Y），…，Cov（Xp，Y）。上面的公式给出了

β∗=（Cov（X））-1 Cov（X，Y）

假设Cov（X）是可逆的这个方程决定了我们可以用（73）来写

.

注意，上面出现的术语Cov（Y，X）是Cov（X，Y）的转置更一般地，给定两个随机向量W=（W1，…，W p）和Z=（Z1，…，Z q），我们将Cov（W，Z）定义为p×q矩阵，其（i，j）项是Wi和Zj之间的协方差。

根据X1，…，Xp，Y的最佳线性预测（BLP）等于

β0∗+β1∗X1+…βp∗Xp=β0∗+（β∗）TX

=E（Y）–Cov（Y，X）（Cov（X））–1E（X）+Cov（Y，X）（Cov（X））–1X

=E（Y）+Cov（Y，X）（Cov（X））-1（X−E（X））（71）

如前所述，就X1，….Xp而言，Y的最佳预测因子（BP）是X1，….Xp的函数g∗（X1，…，Xp），它使

L（g）：=E（Y-g（X1，…，Xp））2

在所有函数g中我们都看到了

g∗（X1，…，Xp）=E（Y | X1，…，Xp）。

换句话说，最好的预测因素是条件期望。一般来说，BP和BLP是不同的，与BLP相比，BP对Y的预测更准确。注意，BLP只依赖于随机变量Y，X1，…，Xp之间的方差和协方差，而BP则潜在地依赖于整个联合分布。因此，与BP相比，BLP通常更容易根据数据进行估计。

一般来说，我们将把任何涉及Y，X1，…，Xp分布的量作为二阶性质，它只依赖于Y，X1，…，Xp的均值、方差和协方差。请注意，BLP是第二个订单数量，而BP不是。

# 最佳线性预测

在上一堂课中，我们讨论了基于其他随机变量X1，…，Xp的线性函数预测随机变量Y的问题。具体而言，我们考虑β0+β1X1+····+βp X p=β0+βT X形式的预测（其中β：=（β1，…，βp）T和X=（X1，…，Xp）T）。就X1，…，Xp而言，Y的最佳线性预测（BLP）是线性函数

具有

哪里最小化

L（β0，…，βp）=E（Y-β0-β1X1-····-βpXp）2（72）

超过β0，β1，…，βp。

设置（72）相对于β0，…，βp的导数，并将其设置为零，我们观察到满足以下方程：

)=0（73）

和

)=0，i=1，…，p.（74）

（74）中的方程式可以用矩阵表示法简洁地写成：

Cov（X）β∗=Cov（X，Y）与。

这里Cov（X，Y）是p×1向量，有项Cov（X1，Y），…，Cov（Xp，Y）。上面的公式给出了

β∗=（Cov（X））–1个Cov（X，Y）这就决定了我们可以用（73）来写

.

注意，上面出现的术语Cov（Y，X）是Cov（X，Y）的转置更一般地，给定两个随机向量W=（W1，…，W p）和Z=（Z1，…，Z q），我们将Cov（W，Z）定义为p×q矩阵，其（i，j）项是Wi和Zj之间的协方差。

Y的最佳线性预测（BLP）X1，…，Xp等于β0∗+β1∗X1+…βp∗Xp=β0∗+（β∗）TX

=E（Y）–Cov（Y，X）（Cov（X））–1E（X）+Cov（Y，X）（Cov（X））–1X

=E（Y）+Cov（Y，X）（Cov（X））-1（X−E（X））（75）

以下是BLP的一些重要特性：

1. BLP求解方程（73）和（74）。这些方程称为正规方程。
2. 如果Cov（X）是可逆的（等价地，正定的），那么BLP由（75）唯一给出。
3. Y-BLP的平均值为零（因为（73）），Y-BLP与每个Xi不相关，i=1，…，p（因为（74））事实上，这个属性是BLP的特征（见下一步）。
4. 如果Cov（X）是可逆的，那么从正态方程的形式可以清楚地看出，BLP是X1，…，Xp的唯一线性组合，使得Y-BLP具有平均零，并且与X1，…，Xp不相关。

例42.1（情况p=1）当p=1时，随机向量X只有元素X1，因此Cov（X）刚好等于V ar（X1）在这种情况下，Y关于X1的BLP由

.

换句话说，当p=1时，

.

在进一步的特殊情况下，当V ar（Y）=V ar（X1）和E（Y）=E（X1）=0时，我们有



所以BLP是由ρY，X1X1给出的。

例42.2。假设X1，X2，Z3，…，Zn，Zn+1是不相关的随机变量和均值为零的随机变量将随机变量X3，…，Xn+1定义为

对于t=3，…，n+1，Xt=φ1Xt-1+φ2Xt-2+Zt。

当n≥2时，Xn+1在X1，…，Xn的BLP是多少？

根据定义，

Xn+1=φ1Xn+φ2Xn−1+Zn+1

也就是说Xn+1-φ1Xn-φ2Xn-1=Zn+1现在很容易看出，对于t≥3，每个Xt仅依赖于X1，X2，Z3，…，Zt，这意味着Zn+1与所有X1，…，Xn不相关。

因此，φ1Xn+φ2Xn-1是X1，…，Xn的线性组合，使得Xn+1-φ1Xn-φ2Xn-1与X1，…，Xn中的每一个都不相关（它也具有平均零）。因此，我们推断Xn+1的BLP在X1，…，Xn方面等于φ1Xn+φ2Xn-1。

如前所述，就X1，….Xp而言，Y的最佳预测因子（BP）是X1，….Xp的函数g∗（X1，…，Xp），它使

L（g）：=E（Y-g（X1，…，Xp））2

在所有函数g中我们都看到了

g∗（X1，…，Xp）=E（Y | X1，…，Xp）。

换句话说，最好的预测因素是条件期望。一般来说，BP和BLP是不同的，与BLP相比，BP对Y的预测更准确。注意，BLP只依赖于随机变量Y，X1，…，Xp之间的方差和协方差，而BP则潜在地依赖于整个联合分布。因此，与BP相比，BLP通常更容易根据数据进行估计。

一般来说，我们将把任何涉及Y，X1，…，Xp分布的量作为二阶性质，它只依赖于Y，X1，…，Xp的均值、方差和协方差。请注意，BLP是第二个订单数量，而BP不是。

# 残差

随机变量Y在X1，…，Xp方面的余数将用rY | X1，…，Xp表示，并定义为Y与Y的BLP在X1，…，Xp方面的差异换句话说，

rY | X1，…，Xp=Y−BLP。

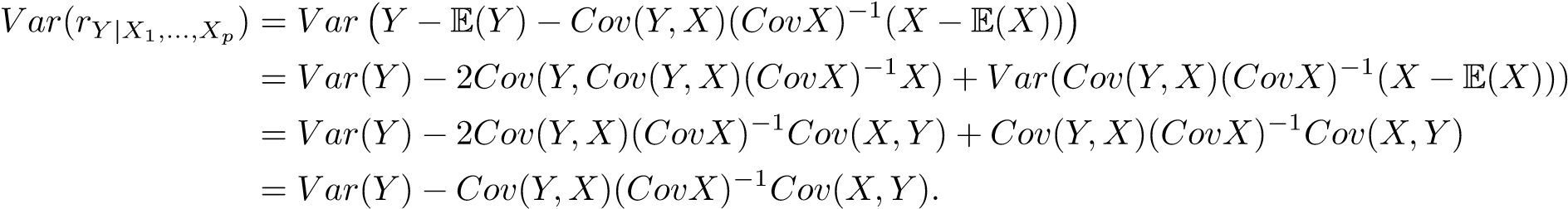
使用BLP的公式，我们可以为残差写下以下公式：

rY | X1，…，Xp=Y−E（Y）–Cov（Y，X）（CovX）–1（X−E（X））（76）

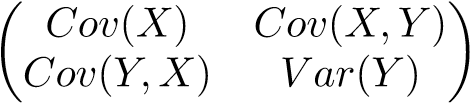
其中X是p×1随机向量，分量为X1，…，Xp。

残差的均值为零，与X1，…，Xp中的每一个都不相关。这可以直接从公式（76）或从BLP的性质来证明。

残差的方差可由公式（76）计算如下：



换言之，V ar（rY | X1，…，Xp）等于协方差矩阵中V ar（Y）的Schur补码（下一次调用）：

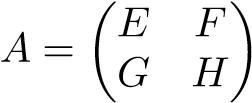


（n+1）×1随机向量（X1，…，Xp，Y）T。

注意，剩余量也是第二个订单数量。

# 绕道：Schur补语

假设一个n×n矩阵A被分成四个块



其中E为p×p，F为p×q，G为q×p，H为q×q（p和q为p+q=n）。

我们定义

ES:=E-FH-1G和HS:=H-GE-1F

假设H 1和E 1存在。我们将把ES和HS分别称为E和H的Schur补语（警告：这不是标准术语；更常见的说法是把ES称为H的Schur补语，把HS称为E的Schur补语。我觉得把ES看作E的Schur补语和HS看作H的Schur补语更自然）。

注意E和ES都是p×p，而H和HS都是q×q。

Schur补语有许多有趣的性质，例如：

1. det（A）=det（E）det（HS）=det（H）det（ES）。
2. 如果A是正定的，那么E，ES，H，HS都是正定的。

还有很多其他人请随意阅读Diane Ouellette的专著《Schur Completions and Statistics》，以证明和阐述这些事实（这对本课程并不一定）。

但Schur补的一个非常重要的性质是它们自然地出现在分块矩阵的逆矩阵中。分块矩阵的逆矩阵的标准公式（例如，参见

（77个）

必须从这个公式中注意到，A-1的第一个（或（1,1）第块等于A的第一个块的Schur补的逆。同样，A-1的最后一个（或（2,2）第块等于A的最后一个块的Schur补的逆。

我们将使用表达式（77）来求分块矩阵A的逆，但我们不知道如何证明（77）。你可以在其他地方找到这一事实的许多证据（只是谷歌类似于“分区矩阵的逆”）。

# 偏相关

给定随机变量Y1，Y2和X1，…，Xp，给定X1，…，Xp的Y1和Y2之间的偏相关用ρY1，Y2 | X1，…，Xp表示，定义为

.

换句话说，ρY1，Y2 | X1，…，Xp被定义为给定X1，…，Xp的Y1的余数与给定X1，…，Xp的Y2的余数之间的相关性。ρY1，Y2 | X1，…，Xp也被称为Y1和Y2在控制X1，…，Xp后的偏相关由于残差是二阶量，因此偏相关也是二阶量我们现在来看看如何用Y1，Y2和X的协方差来显式地写出偏相关。

如rY1 | X1，…，Xp=Y1−E（Y1）–Cov（Y1，X）（CovX）–1（X−E（X））

和rY2 | X1，…，Xp=Y2−E（Y2）–Cov（Y2，X）（CovX）–1（X−E（X）），

可以检查（作为练习留下）

Cov（rY1 | X1，…，Xp，rY2 | X1，…，Xp）=Cov（Y1，Y2）–Cov（Y1，X）（CovX）–1Cov（X，Y2）。

这与前一小节的残差方差公式一起，给出了偏相关ρY1，Y2 | X1，…，Xp的以下公式：

.

当p=1，使X等于标量随机变量X1时，上述公式简化为（勾选此项）：

.

把残差rY1 | X1，…，Xp和ry2 | X1，…，Xp的方差及其协方差放在矩阵中是有指导意义的。首先回顾一下：

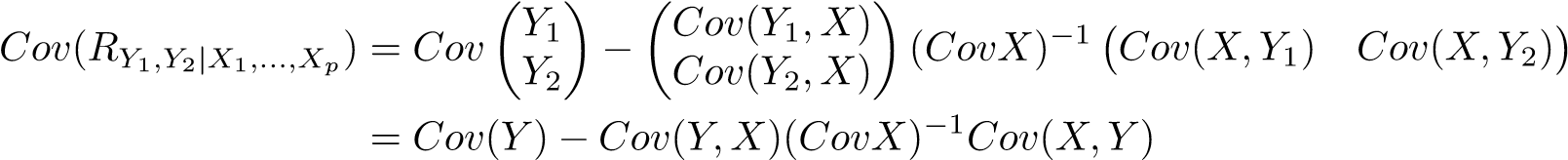
V ar（rY1 | X1，…，Xp）=V ar（Y1）–Cov（Y1，X）（CovX）–1Cov（X，Y1），

V ar（rY2 | X1，…，Xp）=V ar（Y2）–Cov（Y2，X）（CovX）–1Cov（X，Y2）

和

Cov（rY1 | X1，…，Xp，rY2 | X1，…，Xp）=Cov（Y1，Y2）–Cov（Y1，X）（CovX）–1Cov（X，Y2）。

设RY1，Y2 | X1，…，Xp表示由残差RY1 | X1，…，Xp和rY2 | X1，…，Xp组成的2×1随机向量。残差的方差和协方差公式允许我们将RY1，Y2 | X1，…，Xp的2×2协方差矩阵写成



哪里

X1

X2

 · 



和X=·。

 · 

经验值

Cov公式的右边（RY1，Y2 | X1，…，Xp）正好等于矩阵中Cov（Y）的Schur补

.

因此，如果∑表示（p+2）×1随机向量（X1，…，Xp，Y1，Y2）T的协方差矩阵，那么Cov（RY1，Y2 | X1，…，Xp）正好等于∑中Cov（Y）的Schur补我们将在下一节课中回到这个事实，并用它来描述一个涉及∑-1的偏相关ρY1，Y2 | X1，…，Xp的表达式。

# 偏相关与反协方差

上节课我们定义了偏相关给定随机变量Y1，Y2和X1，…，Xp，给定X1，…，Xp的Y1和Y2之间的偏相关用ρY1，Y2 | X1，…，Xp表示，定义为

.

换句话说，ρY1，Y2 | X1，…，Xp被定义为给定X1，…，Xp的Y1的余数与给定X1，…，Xp的Y2的余数之间的相关性。

记住，残差rY1 | X1，…，Xp和rY2 | X1，…，Xp有以下表达式：

rY1 | X1，…，Xp=Y1-E（Y1）-Cov（Y1，X）（CovX）-1（X-E（X））

和rY2 | X1，…，Xp=Y2−E（Y2）–Cov（Y2，X）（CovX）–1（X−E（X）），

在上一个类中，我们计算了rY1 | X1，…，Xp和rY2 | X1，…，Xp的方差以及它们之间的协方差这给了我们公式：

V ar（rY1 | X1，…，Xp）=V ar（Y1）–Cov（Y1，X）（CovX）–1Cov（X，Y1），

V ar（rY2 | X1，…，Xp）=V ar（Y2）–Cov（Y2，X）（CovX）–1Cov（X，Y2）

和

Cov（rY1 | X1，…，Xp，rY2 | X1，…，Xp）=Cov（Y1，Y2）–Cov（Y1，X）（CovX）–1Cov（X，Y2）。

我们可以把这些表达式放在一起，得到给定X1，…，Xp的Y1和Y2之间的偏相关的下列公式：

.

现在我们将描述部分相关和协方差矩阵的逆之间的关系。

设RY1，Y2 | X1，…，Xp表示由残差RY1 | X1，…，Xp和rY2 | X1，…，Xp组成的2×1随机向量残差的方差和协方差公式允许我们将RY1，Y2 | X1，…，Xp的2×2协方差矩阵写成



哪里

X1

X2



而X=·····································。

 · 

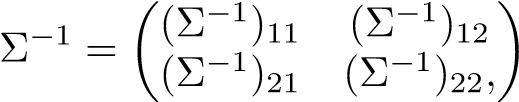
经验值

Cov公式的右边（RY1，Y2 | X1，…，Xp）正好等于矩阵中Cov（Y）的Schur补

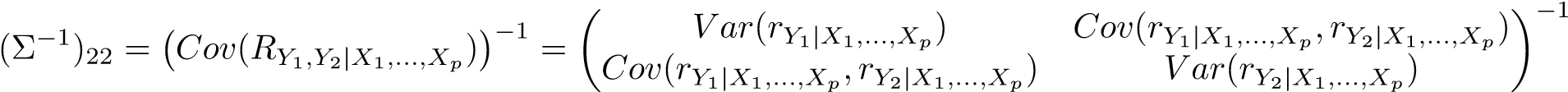
.

因此，如果∑表示（p+2）×1随机向量（X1，…，Xp，Y1，Y2）T的协方差矩阵，那么Cov（RY1，Y2 | X1，…，Xp）正好等于∑中Cov（Y）的Schur补。

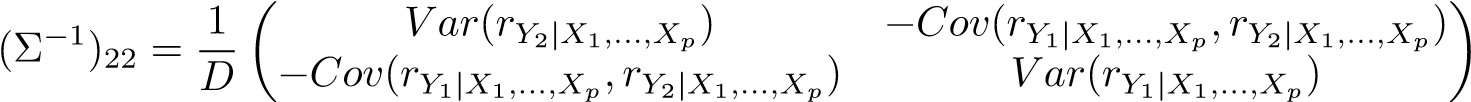
但是我们知道如果我们反转∑，那么∑-1的最后一个对角线块（或第（2,2）个块）等于∑的第（2,2）个块的Schur补的逆Schur补和RY1，Y2 | X1，…，Xp的协方差之间的上述联系允许我们推断



然后



然后给出了2×2矩阵逆的一般公式



其中D是Cov的行列式（RY1，Y2 | X1，…，Xp）。

由此可知，偏相关ρY1，Y2 | X1，…，Xp有另一个表达式：

.

这表明了部分相关矩阵和逆协方差矩阵之间的联系。

更一般地，如果Y1，…，Yn是随机变量（这里不需要分布假设），协方差矩阵由∑给出。那么给定Yk，k 6=i，k6=j，Yi和Yj之间的偏相关可以用∑-1写成

.

这尤其意味着

（？-1）（i，j）=0⇐•？ρYi，Yj | Yk，k6=i，k6=j=0

因此，当Yk，k 6=i，k6=j为0时，（？-1）（i，j）=0等于Yi和Yj之间的偏相关。

阿尔索

（σ-1）（i，j）≤0⇐ρYi，Yj | Yk，k6=i，k6=j≥0和（σ-1）（i，j）≥0⇐ρYi，Yj | Yk，k6=i，k6=j≤0

换言之，当Yk，k 6=i，k6=j为非负时，∑-1（i，j）为非正态，相当于Yi与Yj之间的偏相关。类似地，当Yk，k 6=i，k6=j为非正时，∑-1（i，j）为非负相当于Yi和Yj之间的偏相关。

# 偏相关与最佳线性预测

考虑随机变量Y和X1，…，Xp用X1，…，Xp表示Y的BLP。

我们之前已经看到，如果p=1，那么X等于标量随机变量X1，BLP则具有表达式：

.

换句话说，当p=1时，BLP的斜率系数由

（78）

当p≥1时，我们将得到p“斜率”系数X1，…，Xp。在这种情况下，可以写出类似于（78）的公式，如下所示：

（79个）

也就是说，根据rXi | Xk，k6=i等于rY | Xk的BLP的斜率系数。

我们现在要证明这一事实。我们可以假设在不丧失一般性的情况下i=p。对其他i的证明可以通过重新排列X1，…，Xp来完成，这样Xi出现在最后一个位置。β∗=（β1，…，βp）∗的公式是

.

让我们写信

式中，X-p：=（X1，…，Xp-1）T由除Xp以外的所有X组成我们可以将Cov（X）划分为

.

然后β\*的公式变成



为了从这个表达式导出的显式公式，我们需要计算（CovX）－1的最后一行分块矩阵的逆矩阵的标准公式表明

一些东西

式中，HS:=H-GE-1F是A中H的Schur补码。我们将此公式应用于

E=Cov（X−p），F=Cov（X−p，X p），G=Cov（Xp，X−p），H=V ar（Xp）

所以A等于Cov（X）。在这种情况下，

HS=V ar（X p）－Cov（Xp，X－p）（Cov（X－p））－1Cov（X－p，Xp）＝V ar（rXp | Xk，k6＝p）

以便

.

因此我们得到

.

这证明了i=p的结果，一个可以证明另一个i的结果，只要重新排列X1，…，Xp，使Xi显示为最后一个变量。

（79）的一个重要结果是：

=0（80）

换句话说，当且仅当Y与Xi之间的偏相关系数Xk，k 6=i等于0时，基于X1，…，Xp的Y的BLP中的Xi系数等于零。

例47.1。假设n≥2的X1，X2，Z3，…，Zn，Zn+1是均值为零的不相关随机变量将随机变量X3，…，Xn+1定义为

对于t=3，…，n+1，Xt=φ1Xt-1+φ2Xt-2+Zt。

在上一个类中，我们已经看到Xn+1的BLP，就X1，…，Xn等于φ1Xn+φ2Xn-1这意味着Xn+1的BLP中Xi的系数X1，…，Xn等于0，对于i=1，…，n-2作为（80）的结果，我们推断

ρXn+1，Xi | Xk，k6=i，i=1，…，n-2时，1≤k≤n=0。

利用偏相关和逆协方差的关系，我们可以进一步推导出如果∑表示X1，…，Xn+1的（n+1）×（n+1）协方差矩阵，则

当i=1，…，n-2时，∑-1（i，n+1）=0。

# 当Y是随机向量时的BLP

让我们先快速回顾一下BLP。给定随机变量Y和X1，…，Xp，Y的线性预测因子X1，…，Xp是β0+β1X1+···+βpXp形式的随机变量。BLP由where minimize给出：

L（β0，β1，…，βp）：=E（Y−β0−β1X1——···−βpXp）2

在β0，…，βp上，我们已经看到可以用微积分计算出来，这给出了公式：

BLP=EY+Cov（Y，X）（CovX）–1（X–EX）

其中X代表p×1随机向量，分量X1，…，Xp。剩余的rY | X1，…，Xp简单地等于Y-BLP，我们已经看到rY | X1，…，Xp的方差等于：

var（rY | X1，…，Xp）=var（Y）–Cov（Y，X）（CovX）–1Cov（X，Y）。

注意，这是

现在假设我们有两个随机变量Y1和Y2，Y表示2×1随机向量，分量Y1和Y2。考虑根据X求Y的BLP的问题（其中X，如前所述，是p×1随机向量，分量X1，…，Xp）为了形式化，我们首先需要定义什么是线性预测（注意Y是2×1随机向量，而不是标量随机变量）。根据X给出Y的线性预测

斧头+c

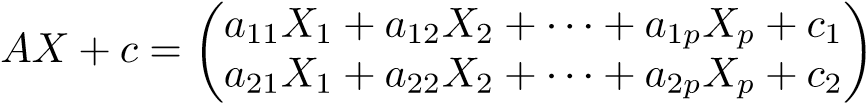
其中A是2×p矩阵，c是2×1向量。预测Y的线性预测器的精度可以通过

L（A，c）：=EkY–-AX–-ck2。

然后由A\*Y+c\*给出BLP，其中A\*和c\*将A和c上的L（A，c）最小化。为了解决这个最小化问题，让我们先将A和c写成

和

以便



和

L（A，c）=EkY–-AX–-ck2=E（Y1–-a11X1–-a12X2–··–-a1pXp–-c1）2+E（Y2–-a21X1–-a22X2–··–-a2pXp–-c2）2

从这里可以清楚地看出，为了使上述关于A和c的最小，我们可以将a11，a12，…，a1p，c1上的第一项最小化，然后将a21，a22，…，a2p，c2上的第二项最小化从这里很容易看到

以X表示的是

就Y2的BLP而言

因此，即使Y是2×1随机向量，相同的公式EY+Cov（Y，X）（CovX）－1（X－EX）也会根据X给出Y的BLP。现在很容易看出，当Y是k×1的随机向量时，对于每个k≥1（而不仅仅是k=1或k=2），这是成立的可以用X1，…，Xp定义Y的余数

Y | X：=Y−EY−Cov（Y，X）（CovX）–1（X−EX）

这正是向量，它的第i个分量是Yi在X1，…，Xp方面的余数。检查RY | X的协方差矩阵是否由

Cov（RY | X）=Cov（Y）–Cov（Y，X）（CovX）–1Cov（X，Y）

也就是矩阵中Cov（Y）的Schur补

# 随机向量的矩母函数

接下来我们将讨论这门课的最后一个主题：多元正态分布。因此，了解随机向量的矩母函数是很有帮助的。

n×1随机向量Y的矩母函数定义为

阿蒂

我的（a）：=Ee

对于期望存在的每一个RN。注意，当a=（0，…，0）T是零向量时，很容易看到MY（a）=1。

就像在单变量情况下，矩生成函数在A＝0的邻域中确定分布。

矩生成函数在独立的情况下表现得非常好。假设Y（1）和Y（2）是两个随机向量，设为将Y（1）和Y（2）放在同一列向量中得到的向量。那么Y（1）和Y（2）是独立的当且仅当

我的（a）=我的（1）（a（1））我的（2）（a（2））对于每个。

因此在独立性下，MGF分解，反之，当MGF分解时，我们有独立性。

# 多元正态分布

多元正态分布的定义如下。

定义50.1。随机向量Y=（Y1，…，Yn）T是多元正态分布，如果Y的每一个线性函数都是单变量正态分布。

备注50.1。必须强调的是，要使Y=（Y1，…，Yn）T为多元正态函数，Y=a1Y1+…anYn的每个线性函数都需要为单变量正态函数举个例子来说，仅仅让每个Yi成为单变量正态是不够的很容易构造这样的例子，其中每个Yi都是单变量正规，但是对于许多向量（a1，…，a N）T，a1 Y1+·····+anYn不是单变量正规。例如，假设Y1∼N（0,1）和Y2=ζY1，其中ζ是离散随机变量，取概率为1/2的两个值1和−1，且ζ独立于Y1。很容易看出

Y2 |ζ=1∼N（0,1）和Y2 |ζ=-1∼N（0,1）。

因此，这意味着Y2∼N（0,1）（并且Y2独立于ζ）。但是请注意，Y1+Y2不是正常的

.

因此，在这个例子中，即使Y1和Y2都是N（0,1），向量也不是多元正态的。

例50.2。我们在前面的课中已经看到，如果Z1，…，Zn是独立的单变量正态分布，那么a1Z1+…anZn是每个a1，…，an的正态分布。因此，由独立正态随机变量组成的随机向量Z=（Z1，…，Zn）T具有多元正态分布。事实上，我们将在下面证明，如果Y有一个多元正态分布，那么它必然是

Y是由独立的单变量正态随机变量构成的随机向量Z的线性函数。

## 多元正规的矩母函数

假设Y=（Y1，…，Yn）T是多元正态分布。设μ=E（Y）和∑=Cov（Y）分别为Y的均值向量和协方差矩阵然后，作为多元正态性定义的直接结果，Y的MGF等于

（81）

要了解为什么这是真的，请注意，根据多元正态性的定义，aTY是单变量正态的现在aTY的均值和方差由

E（aT Y）=在μ和V ar（aTY）=aTCov（Y）a=在∑a

以便

aTY∼N（aTμ，aT∑a）表示每一个a∈Rn。

然后（81）直接从一元正规的MGF公式得出。

注意Y的MGF（由（81）给出）仅取决于Y的平均向量μ和协方差矩阵∑。因此，每个多元正态向量Y的分布特征是平均向量μ和协方差∑。因此，我们使用符号Nn（μ，σ）表示平均μ和协方差∑的多元正态分布。

## 与i.i.d N（0,1）随机变量的连接

假设Y的协方差矩阵∑是正定的，所以∑-1/2是定义良好的设Z：=∑-1/2（Y-μ）。公式（81）允许计算Z的MGF，如下所示：

.

上面的右手边显然是一个随机向量的MGF，这个随机向量具有n i d标准正态随机变量因此，由于MGFs唯一地决定分布，我们得出结论Z=（Z1，…，Zn）T具有独立的标准正态随机变量。因此，我们证明了：如果Y∼Nn（μ，∑）和∑是p.d，那么Z=的Z1，….Zn是独立的标准正态随机变量。

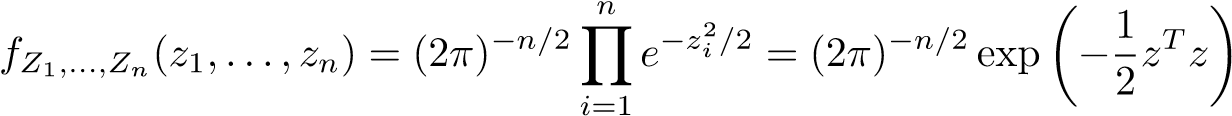
# 多元正态分布的联合密度

假设Y=（Y1，…，Yn）T是具有多元正态分布的随机向量。那么Y1，…，Yn的节理密度是多少？

设μ=E（Y）和∑=Cov（Y）分别为Y的均值向量和协方差矩阵为了使Y有一个连接密度，我们需要假设∑是正定的在上一节中，我们看到Z的Z1，…，Zn是独立的标准正态随机变量，其中

Z=∑-1/2（Y–）。

因为Z1，…，Zn是独立的标准法线，它们的关节密度等于



式中z=（z1，…，zn）T。

利用上述公式和Y=μ+σ1/2Z的事实，我们可以通过雅可比公式计算Y1，…，Yn的节点密度这给了



式中y=（y1，…，yn）T。

# 多元正态随机变量的性质

假设Y=（Y1，…，Yn）T∼Nn（μ，σ）注意，μ是Y的平均向量E（Y），而∑是协方差矩阵Cov（Y）。以下特性非常重要。

1. Y的线性函数也是多元正态的：如果A是m×n确定性矩阵，c是m×1确定性向量，则Y+c∼Nm（Aμ+c，A∑AT）。

原因：a Y+c的每一个线性函数显然也是Y的一个线性函数，因此，这个事实遵循了多元正态分布的定义。

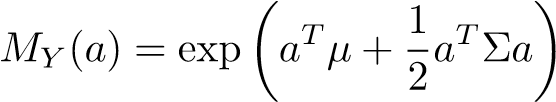
1. 如果Y是多元正态的，那么取Y的一个子集所形成的每个随机向量也是多元正态的。

理由：从以前的事实来看。

1. 独立性与不相关性相同：如果Y（1）和Y（2）是两个随机向量，那么Y=（Y（1）T，Y（2）T）T是多元正态的。当且仅当Cov（Y（1），Y（2））=0时，Y（1）和Y（2）是独立的。

原因：独立性意味着Cov（Y（1），Y（2））=0这一事实是显而易见的，不需要任何正态性。关键是另一个含义，即零协方差意味着独立性。因此，可以证明Y的MGF等于Y（1）和Y（2）的MGF的乘积Y的MGF

等于



其中∑=Cov（Y）。

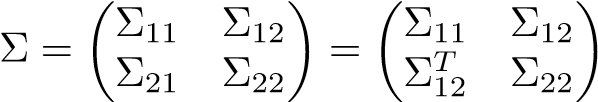
注意Y（1）和Y（2）也是多元正态的，所以

当i=1,2时

哪里

μ（i）：=E（Y（i））和∑ii:=Cov（Y（i））。

如果∑12：=Cov（Y（1），Y（2））和∑，那么观察



因此，如果a=（a（1），a（2））T，则

.

在∑12=0的假设下，我们可以写出



从中可以看出

MY（a）=MY（1）（a（1））MY（2）（a（2））。

由于Y=（Y（1），Y（2））T的MGF分解为Y（1）的MGF和Y（2）的MGF的乘积，因此Y（1）和Y（2）是独立的。因此，在（Y（1），Y（2））T的多元正态性假设下，不相关等同于独立性。

1. 假设Y=（Y1，…，Yn）T是多元正态随机向量那么两个分量Yi和Yj是独立的当且仅当∑ij=0其中∑Cov（Y）。

理由：直接从前面三个事实出发。

1. 线性函数的独立性可以通过矩阵相乘来检验：假设Y是多元正态函数。那么当且仅当A∑BT=0时，AY和BY是独立的。原因：首先要注意

是多元正态的。因此，当且仅当Cov（AY，BY）=0时，AY和BY是独立的。然后，根据Cov（AY，BY）=A∑BT的观察结果，提出了该断言。

# 多元正态随机变量的性质

假设Y=（Y1，…，Yn）T∼Nn（μ，σ）注意，μ是Y的平均向量E（Y），而∑是协方差矩阵Cov（Y）。在上一节课中，我们研究了以下属性。

1. Y的线性函数也是多元正态的：如果A是m×n确定性矩阵，c是m×1确定性向量，则Y+c∼Nm（Aμ+c，A∑AT）。
2. 由Y的一个子集构成的随机向量也是多元正态的。
3. 独立性与不相关性相同：如果Y（1）和Y（2）是从Y获得的两个子向量，那么Y（1）和Y（2）是独立的，当且仅当Cov（Y（1），Y（2））=0。
4. 两个分量Yi和Yj是独立的当且仅当∑ij=0其中∑Cov（Y）。
5. 线性函数的独立性可以通过矩阵相乘来检验：当且仅当A∑BT=0时，y和by是独立的。

# 幂等矩阵与卡方分布

接下来，我们将证明具有恒等协方差的多元正态随机变量的二次型具有卡方分布，前提是定义二次型的对称矩阵是等幂的。如果A2=A，则称平方矩阵A为幂等矩阵。关于幂等矩阵的一个重要事实如下。

事实：如果A是秩r的n×n对称且幂等矩阵当且仅当

（82个）

对于r正交和单位长度向量u1，…，ur。

为了证明这个事实，首先要注意如果A是对称的，那么通过谱定理



对于正交基u1，…，un和实数λ1，…，λn，A的秩正好等于非零的λi的个数。如果r是A的秩，那么我们可以这样写（假设在不丧失一般性的情况下，λ1，…，λr是非零的，且λr+1=·····=λn=0）

.

接下来就是

.

因此，如果A是等幂的，那么A2=A



这意味着λ2i=λi，它给出了λi=1（注意，我们假设λi6=0）这证明了（82）。

下面的结果表明，如果下垫矩阵是等幂的，则具有恒等协方差的多元正态随机向量的二次型是卡方的。

定理54.1。假设Y∼n n（μ，In），设A是秩为r的n×n对称幂等矩阵，则

（Y-μ）TA（Y-μ）∼x2R。

证明因为A是对称的和幂等的，并且有秩r，我们可以把A写成



对于一些正交和单位范数向量u1，…，ur。那么

.

式中，Vi：=uTi（Y–）。现在请注意，每个Vi∼N（0,1）和V1，…，Vr是不相关的，因此是独立的（因为正态性）这证明了。

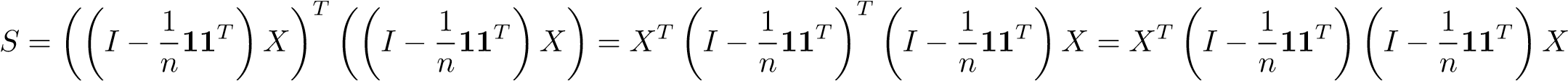
例54.2假设X1，…，Xn是i.i.d N（0,1）然后X∏∼N（0,1/N）和

.

此外，X′和S是独立的。

X∏N（0,1/N）很容易为了证明S∼x2n-1和S与X∼是独立的，我们将给出两种方法。

方法一：要证明S∼x2n-1，关键是要注意



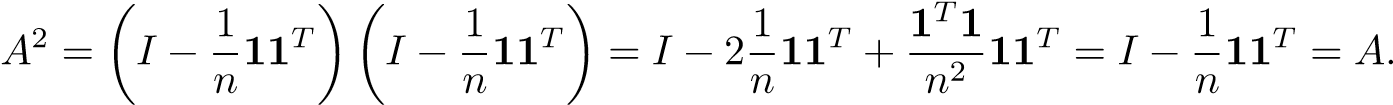
其中X=（X1，…，Xn）T和1=（1，…，1）T。在上面的最后一步中，我们使用了对称的事实第一步，我们使用了

.

现在如果

,

那么很明显A是对称的和等幂的



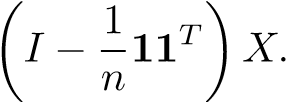
A的秩也等于n-1因此通过定理54.1（注意X=（X1，…，Xn）T∼Nn（0，In）），我们得到

S=XTAX∼x2n-1。

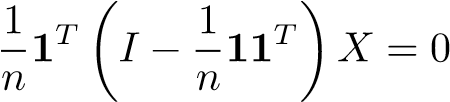
为了证明S和X是独立的，我们只需要观察

和（83）

是独立的，因为S是



（83）中随机变量的独立性如下：



Rn，u1=1/√n（检查u1是否有单位方法二：设u1，…，un为范数的正交基）设U是列u1，…，un的矩阵，即。，

U=[u1:···：联合国]。

注意，UUT=UTU=In（根据正交基的性质）。现在让Y=UTX那么Y是X（和X∼Nn（0，In））的线性函数，因此

Y∼Nn（UT（0），utiu）=Nn（0，UTU）=Nn（0，In）。

进一步注意

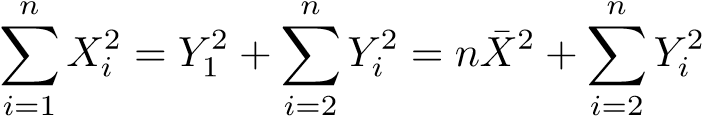
北

X 2吨T T T X 2

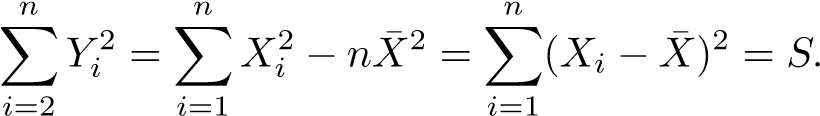
Yi=Y Y=X UU X=X X=Xi（84）

i=1 i=1

还有那个因此，（84）给出



以便



这和Y∼N N（0，In）（也就是说Y1，…，Yn是i.i.dn（0，1））的事实意味着

. 还要注意，S仅依赖于√Y2，…，Yn，因此它独立于Y1，因此S和X

独立的（注意X‘=Y1/n）。

例54.3假设X∼n n（0，∑），其中∑是一个n×n矩阵，对角线上有1，非对角线上有ρσ也可以表示为

σ=（1−ρ）In+ρ11T，其中1：=（1，…，1）T。

也就是说，X1，…，Xn是多元正态的，具有均值零，单位方差，且每对之间的相关性等于ρ。找出X′的分布，并证明它们是独立的。

X是X的线性函数，所以它是正常的我们只需要找到它的均值和方差。显然，EX’=0（因为每个EXi=0）和

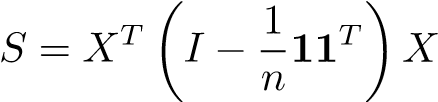
.

因此

.

观察到这意味着1+（n−1）ρ≥0或ρ≥-1/（n−1）换句话说，如果ρ<-1/（n-1），那么∑将不是半正定的。

为了找到S的分布，我们可以像前面的例子一样



但是我们不能不幸地使用定理54.1，因为X没有恒等协方差（定理54.1）仅适用于具有恒等协方差的多元正态随机向量。在这里，第二种方法（在前面的例子中描述）在这里工作并给出了S的分布，这在下面解释。

n其中u1=1/√n，设U为带列的矩阵，设u1，…，un为ru1，…，un的正交基，这样UTU=UUT=In设Y=UTX并注意（如前一个示例中所示）

Y1=√nX‘，S=Xn（X−X‘）2=Y22+····+Yn2。

我

i=1

Y的分布现在由Y∼Nn（0，UT∑U）和

.

要计算1TU，请注意

T T T√

1 U=（1 u1，1 u2，…，1 un）=（n，0，…，0）

√Tu1=1T1/√n=√n，事实上1与u2正交，…，un（这是因为

我们用的地方1

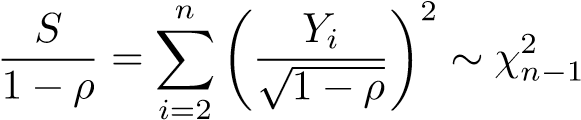
h1，uii=nhu1，i>1时uii=0）我们就这样得到了

UT∑U=（1−ρ）In+ρ（√n，0，…，0）T（√n，0，…，0）。

这意味着UT∑U是一个对角线矩阵，有对角线项（1--ρ+nρ），1--ρ，1--ρ，…，1--ρ因此Y∼Nn（0，UT∑U）表示Y1，…，Yn与

当i>1时，Y1∼N（0,1+（N-1）ρ）和Yi∼N（0,1-ρ）。

因此

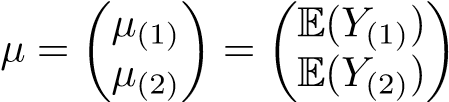


或S∼（1-ρ）x2n-1另外，因为X′只依赖于Y1，S只依赖于Y2，…，Yn，我们认为S和X′是独立的。

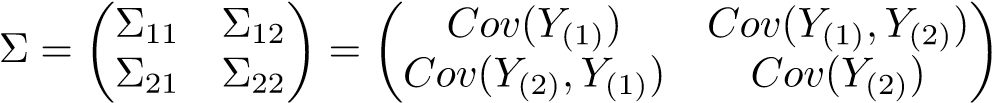
# 多元正态分布的条件分布

假设Y∼Nn（μ，σ）让我们把Y分成两部分Y（1）和Y（2），其中Y（1）=（Y1，…，Y p）T由Y的前p个分量组成，Y（2）=（Yp+1，…，Y n）由Y的最后q:=n−p个分量组成。

然后我们可以类似地划分平均向量μ



协方差矩阵∑as



我们现在要解决的问题是：给定Y（1）=y1，Y（2）的条件分布是什么？答案如下。

事实：假设Y∼Nn（μ，σ），我们有

（85）

也就是说，给定Y（1）=y1的Y（2）的条件分布也是多元正态分布，平均向量由下式给出：



协方差矩阵由

.

我们将把下面（91）的证明翻一遍。在此之前，让我们就有条件分配的形式作几点简短的评论：

1. 条件分布也是多元正态分布。
2. 条件期望E（Y（2）| Y（1）=y1是y1的线性函数。
3. 就Y（1）而言，E（Y（2）| Y（1））正好等于Y（2）的BLP因此BP和BLP是一致的。
4. 条件协方差矩阵Cov（Y（2）| Y（1）=y1）不依赖于y1（这可视为某种同态性）。
5. 条件协方差矩阵Cov（Y（2）| Y（1）=y1）等于∑S22（∑中∑22的Schur补）。换句话说，条件协方差矩阵Cov（Y（2）| Y（1）=y1精确地等于给定Y（1）的Y（2）的残差的协方差矩阵。

事实证明（91）：很容易看出（91）相当于：

.

进一步等同于

（86）

注意，上面右边的分布不依赖于y1因此（86）相当于

和

其中⊥⊥表示独立因为Y是多元正态函数，我们知道Y的线性函数也是多元正态函数，并且Y的线性函数是独立的，只要它们是不相关的。因此，上述断言相当于以下三个等式：

一

2。。

三。

换句话说，我们注意到证明上述三个方程等同于证明（91）。我们现在通过证明上述三个方程来完成（91）的证明。我们已经证明了这三个事实。第一个事实简单地说，给定Y（1）的Y（2）的余数具有零期望。第二个事实是残差的协方差矩阵等于Schur补。第三个事实是给定Y（1）的Y（2）的残差与Y（1）不相关。为了完整起见，让我们按如下所示快速地重新定义它们。

，（87）

（88个）

最后

=0（89）

这就完成了（91）的证明。

让我们重申，所有这三种计算（87），（88）和（89）不需要对Y（1）和Y（2）进行任何分配假设。它们适用于所有随机变量Y（1）和Y（2）多元正态性假设允许我们从这些二阶（即均值和协方差）计算中推断（91）。

# 多元正态分布与卡方分布的若干注记

在上节课上，我们看到了下面的结果。

定理56.1。如果Z∼Nn（0，In）和A是对称的幂等矩阵，则ZTAZ∼x2r，其中r是A的秩。

事实证明，这个结果的倒数也是真的，我们有

定理56.2。假设Z∼Nn（0，In）和A是对称矩阵。则ZTAZ∼x2r当且仅当A是秩r的幂等矩阵。

换句话说，ZTAZ具有卡方分布的唯一方法是当A是幂等的。因此，对于ZTAZ作为卡方分布来说，具有幂等性既是必要的，也是充分的。暗示A的幂等性的事实可以通过矩生成函数来证明，但我们将跳过这个论点。

当协方差不是恒等矩阵时，定理56.1需要修改如下所示。现在假设Y∼Nn（0，∑），我们有兴趣看看Q：=Y-TAY是幂等的

（这里A是对称矩阵）。我们知道Z：=是∑-1/2Y∼Nn（0，In），所以我们可以写（如Y=是∑1/2Z）

Q=Y TAY=ZT∑1/2A∑1/2Z。

因此Q是卡方分布的，当且仅当∑1/2A∑1/2是等幂的，它等价于

∑1/2A∑1/2∑1/2A∑1/2=1/2A∑1/2⇐⇐A∑A=A。

因此我们有

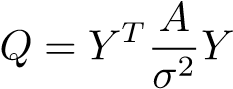
定理56.3假设Y∼Nn（0，∑）和A是对称矩阵。那么Y-TAY∼x2r当且仅当A∑A=A且r=秩（∑1/2A∑1/2）=秩（A）。

让我们看看这个结果的一些例子。

例56.4。假设Y∼n n（0，σ2In），设A为秩为r的n×n对称幂等矩阵，结果表明

.

这可以作为定理56.3的结果来证明，因为



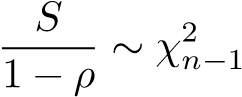
和

.

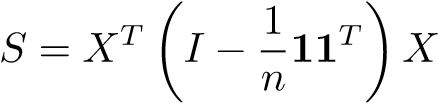
例56.5。假设X∼n n（0，∑），其中∑是一个n×n矩阵，对角线上有1，非对角线上有ρ。σ也可以表示为

Σ = (1 − *ρ*)*In* + *ρ***11***T where* **1** := (1*,...,*1)*T.*

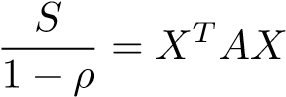
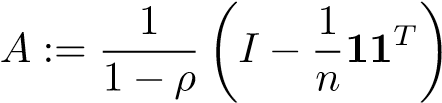
*In the last class, we showed that* *satisfies*

*.* (90)

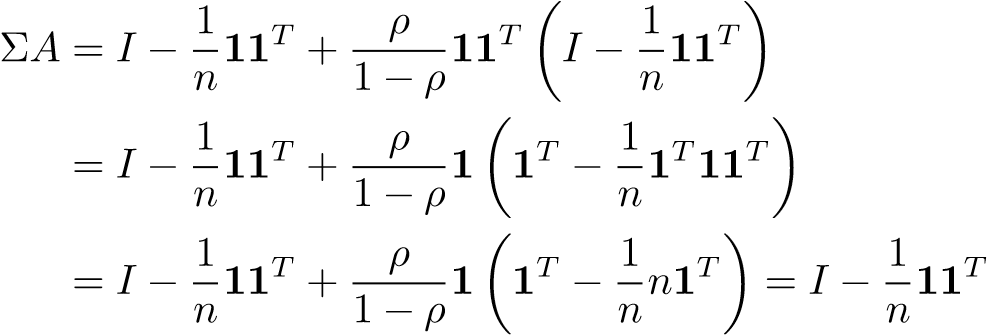
*We shall show this here using Theorem 56.3. Note first that*



*so that*

 *where* 

*Thus to show* (90)*, we only need to prove that A*Σ*A* = *A. For this, see that*



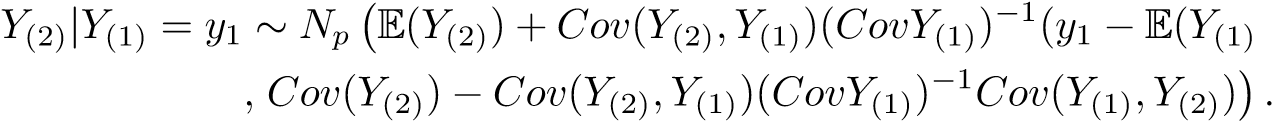
*so that A*Σ*A* = *A. Thus Theorem 56.3 immediately gives* (90)*.*

**Example 56.6.** *Suppose Y* ∼ *Nn*(0*,*Σ)*. Then**. This follows directly from Theorem 56.3 by taking A* = Σ−1*.*

Finally let us mention that when *Z* ∼ *Nn*(*µ,In*) and *A* is idempotent, then *ZTAZ* will be a non-central chi-squared distribution. We will not study these in this class.

# Conditional Distributions of Multivariate Normals

Suppose that *Y*(1) and *Y*(2) are two random vectors (with possibly different lengths) such that the vector  is multivariate normal. Then, as we saw in the last class, we have

 )) (91)

(92)

In words, the conditional distribution of *Y*(2) given *Y*(1) = *y*1 is also multivariate normal with mean vector given by:



and covariance matrix given by

*Cov*(*Y*(2)|*Y*(1) = *y*1) = *Cov*(*Y*(2)) − *Cov*(*Y*(2)*,Y*(1))(*CovY*(1))−1*Cov*(*Y*(1)*,Y*(2))

It is important to note that

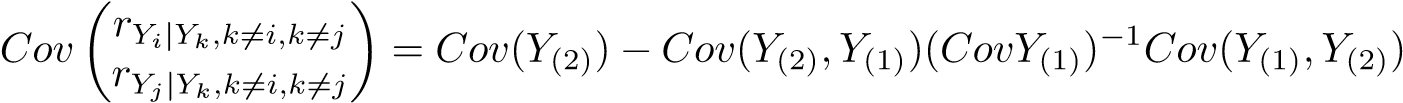
1. E(*Y*(2)|*Y*(1)) is exactly equal to the BLP of *Y*(2) in terms of *Y*(1). Thus the BP and BLP coincide.
2. The conditional covariance matrix *Cov*(*Y*(2)|*Y*(1) = *y*1) does not depend on *y*1 (this can be viewed as some kind of *homoscedasticity*).
3. The conditional covariance matrix *Cov*(*Y*(2)|*Y*(1) = *y*1) is precisely equal to the Covariance Matrix of the Residual of *Y*(2) given *Y*(1).

We can look at the following special case of this result. Fix two components *Yi* and *Yj* of *Y* . Let

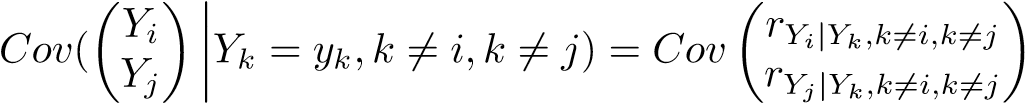
*Y*(2) := (*Yi,Yj*)*T* and let *Y*(1) denote the vector obtained from all the other components *Yk,k* 6= *i,k* 6= *j*. Then

*Cov*(*Y*(2)|*Y*(1) = *y*1) = *Cov*(*Y*(2)) − *Cov*(*Y*(2)*,Y*(1))(*CovY*(1))−1*Cov*(*Y*(1)*,Y*(2))

Now let *rYi*|*Yk,k*6=*i,k*6=*j* and *rYj*|*Yk,k*6=*i,k*6=*j* denote the residuals of *Yi* in terms of *Yk,k* 6= *i,k* 6= *j* and *Yj* in terms of *Yk,k* 6= *i,k* 6= *j*, then we have seen previously that

*.*

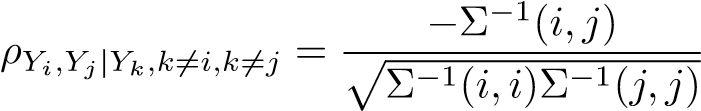
We thus have

 for every *yk,k* 6= *i,k* 6= *j.*

This means, in particular, that the conditional correlation between *Yi* and *Yj* given *Yk* = *yk,k* 6= *i,k* 6= *j*

is precisely equal to the partial correlation *ρYi,Yj*|*Yk,k*6=*i,k*6=*j* (recall that *ρYi,Yj*|*Yk,k*6=*i,k*6=*j* is the correlation between *rYi*|*Yk,k*6=*i,k*6=*j* and *rYj*|*Yk,k*6=*i,k*6=*j*).

Now recall the following connection between partial correlation and entries of Σ−1 that we have seen earlier:

*.*

Putting the above observations together, we can deduce that the following are **equivalent** when *Y* is multivariate normal with covariance matrix Σ:

1. Σ−1(*i,j*) = 0
2. *ρYi,Yj*|*Yk,k*6=*i,k*6=*j* = 0
3. The conditional correlation between *Yi* and *Yj* given *Yk* = *yk,k* 6= *i,k* =6 *j* equals 0 for every choice of *yk,k* 6= *i,k* 6= *j*.
4. *Yi* and *Yj* are conditionally independent given *Yk* = *yk,k* 6= *i,k* 6= *j* equals 0 for every choice of *yk,k* 6= *i,k* 6= *j*.