初探混合高斯分布

周陈序（523120910191）

November 20, 2024

摘要

本文以混合高斯分布为研究对象，通过理论推导和数值实验，系统分析了其分布特性及参数对分布形态的影响。针对混合高斯分布，本文首先定义并绘制其频率分布直方图，深入探讨了各参数（如 μ_1、σ_1^2、μ_2、σ_2^2 及 p）对分布中“峰”特征的作用机制。随后，构造并分析了统计量 U，研究样本量 n 对其频率分布直方图的影响。实验结果显示，随着 n 的增加，分布形态逐渐趋于正态，验证了中心极限定理的作用。

本文还通过矩母函数推导和数值模拟，深入解释了混合高斯分布的多峰特性及其演化规律。本研究不仅加深了对混合高斯分布的理论理解，也进一步展示了概率论在复杂分布分析中的实际应用价值，为后续的学习和研究奠定了基础。

关键词：概率论，混合高斯分布，矩母函数，频率直方图，中心极限定理

Contents

1 引言

概率论作为数学的重要分支，为我们提供了一个严密而系统的理论框架，用于描述和分析随机现象的本质及其规律。在实际应用中，概率论已广泛渗透到数据科学、机器学习、经济学、工程和自然科学等诸多领域，为理解不确定性提供了坚实的基础。

本次大作业以混合高斯分布为核心研究对象，通过探索其频率分布直方图、参数影响以及统计量的行为特性，旨在加深对概率论基本概念和方法的理解，并将理论知识融会贯通于实际问题的解决中。混合高斯分布作为一种重要的概率模型，因其灵活性和强大的拟合能力，在数据聚类、模式识别和信号处理等领域具有广泛的应用价值。本作业不仅涵盖了混合高斯分布的基本性质，还对其在复杂环境下的行为进行系统化分析。

通过分析频率分布直方图以及参数对分布特征的影响，探讨了混合高斯分布在不同条件下的表现；同时，通过构造特定统计量 U，研究了样本量对其分布形态的影响，并利用矩母函数理论进行理论分析。这一过程充分体现了概率论理论与实践的结合，也进一步挖掘了混合高斯分布的统计特性和现实意义。

2 任务一

2.1 混合高斯分布的定义

随机变量 $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$，$Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, 则 $Z = X + \eta * Y$ 服从的分布称为混合高斯分布，其中 $P(\eta = 0) = p$，$P(\eta = 1) = 1 - p$，$p \in [0, 1]$。

2.2 不同参数下混合高斯分布频率分布直方图及参数对频率分布直方图中“峰”的影响

由上文定义可知，混合高斯分布的参数共有五个，分别为 μ_1，σ_1^2，μ_2，σ_2^2 以及 η，本文设定 $\mu_1 = 0$，$\sigma_1^2 = 1$，$\mu_2 = 5$，$\sigma_2^2 = 2$，$p = 0.4$ 绘制出如图1所示的混合高斯分布频率分布直方图，此处固定 X 服从标准正态分布，便于后续进一步探索不同参数对混合高斯分布“峰”的影响。

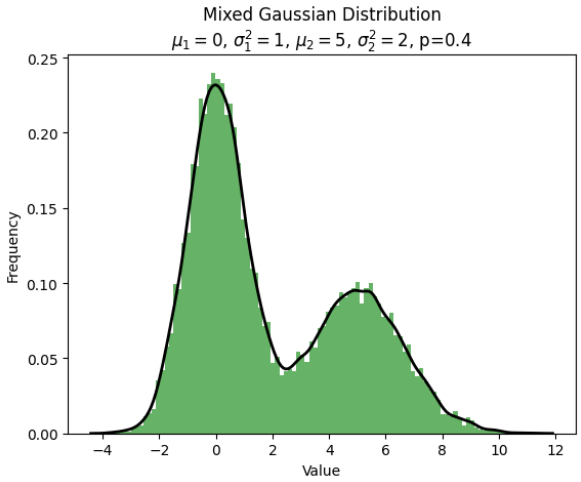

图 1: 混合高斯分布（$\mu_1 = 0$,$\sigma_1^2 = 1$,$\mu_2 = 5$,$\sigma_2^2 = 2$,$p = 0.4$）

2.2.1 μ_2 对“峰”的影响

下面本文将固定 μ_1，σ_1^2，σ_2^2，p，尝试设定不同的 μ_2 以分析 μ_2 对混合高斯分布“峰”的影响。为行文方便，下文每张图片对应的参数均已标注在图中，同时下文将用” 主峰” 代指由变量 X 生成的峰，用” 次峰” 代指由变量 Y 生成的峰。

首先，分别绘制 $\mu_2 = 5$ 与 $\mu_2 = -5$ 的频率分布直方图，如图2和图3所示。

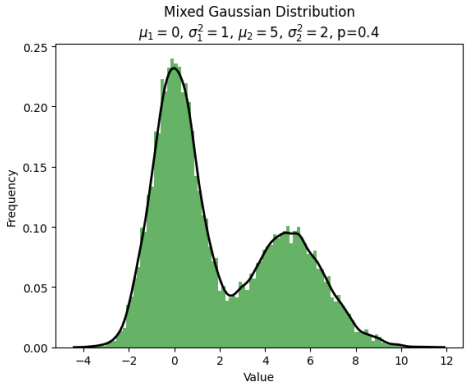

图 2
$\mu_2 = 5$

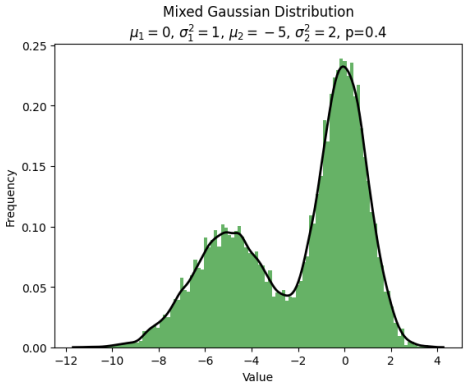

图 3
$\mu_2 = -5$

通过观察两张频率分布直方图，可以得出如下结论：当 μ_2 关于 μ_1 对称时，混合高斯分布的“峰”呈现对称性。通过改变其他参数并绘制相应的直方图（如图10所示，包含了三组不同参数下的频率分布直方图），发现当 $\mu_1 = 0$ 时，只要保持 μ_2 关于 μ_1 的对称性，得到的直方图均表现出对称的“峰”特征，上述结论在 $\mu_1 = 0$ 时的稳健性得到检验。然而，当 $\mu_1 \neq 0$ 时，尽管保持 μ_2 关于 μ_1 对称，直方图中的“峰”在相对位置上大致对称，但两峰之间会出现明显的“挤压”现象，导致图形的对称性破缺。随着 μ_1 的增大，这种对称性逐渐减弱。因此，可以推测，只有在特殊情况下，即 $\mu_1 = 0$ 时，混合高斯分布才具有显著的对称性。

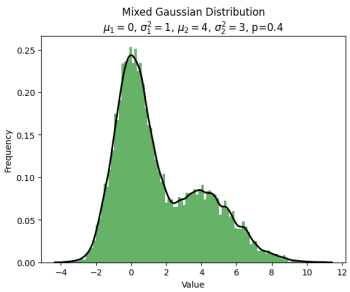

图 4: 对称 1-1

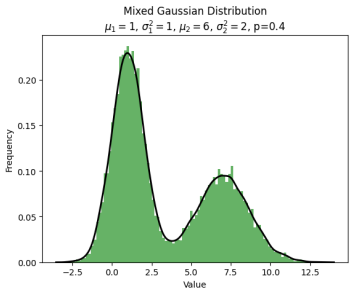

图 5: 对称 2-1

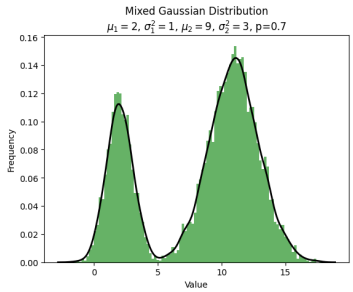

图 6: 对称 3-1

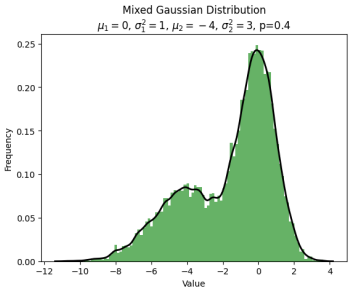

图 7: 对称 1-2

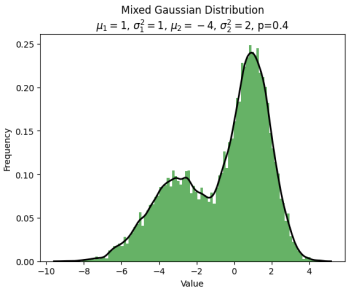

图 8: 对称 2-2

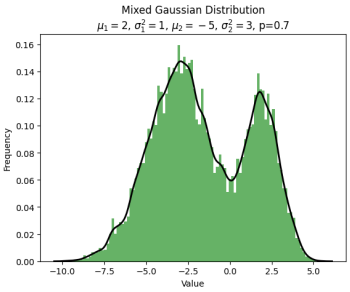

图 9: 对称 3-2

图 10: 对称性探索

本文从 $\mu_2 = -6$ 开始，逐步增大 μ_2 的值，绘制了对应的图像（图19）。通过观察“峰”的演变规律，可以得出以下结论：当 μ_2 接近 μ_1 时，图像中仅出现一个显著的主峰；然而，随着 μ_2 与 μ_1 之间的差距增大，次峰逐渐显现。进一步分析发现，主峰和次峰之间的距离随着 μ_2 与 μ_1 差距的增大而逐步扩大，推测该距离的变化主要由 μ_2 的取值所决定。

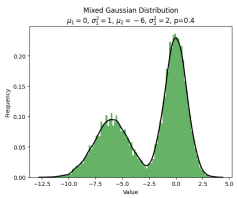

图 11: $\mu_2 = -6$

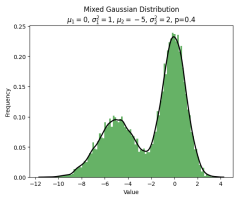

图 12: $\mu_2 = -5$

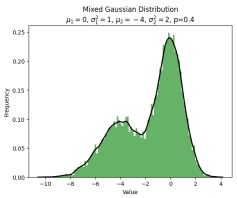

图 13: $\mu_2 = -4$

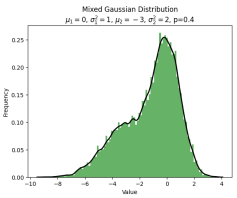

图 14: $\mu_2 = -3$

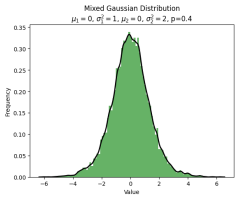

图 15: $\mu_2 = 0$

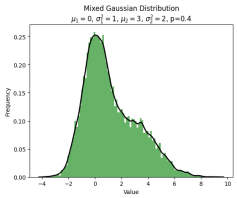

图 16: $\mu_2 = 3$

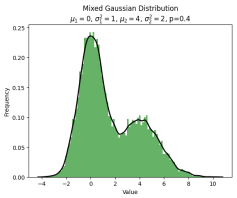

图 17: $\mu_2 = 4$

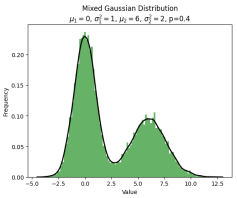

图 18: $\mu_2 = 6$

图 19: μ_2 对“峰”的影响

2.2.2 σ_2^2 对“峰”的影响

控制其它参数不变，改变 σ_2^2 的值，绘制出如下图26所示的一组图片。

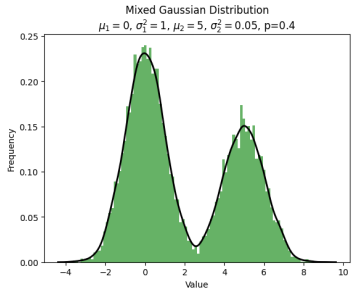

图 20: $\sigma_2^2 = 0.05$

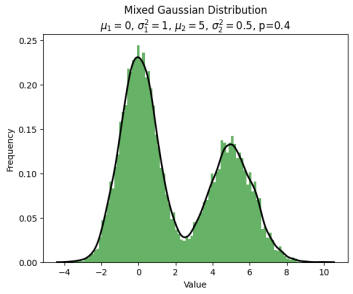

图 21: $\sigma_2^2 = 0.5$

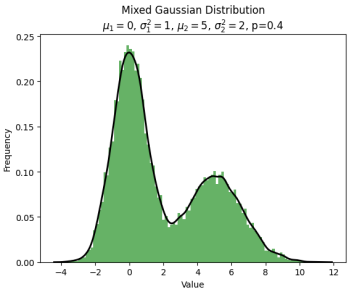

图 22: $\sigma_2^2 = 2$

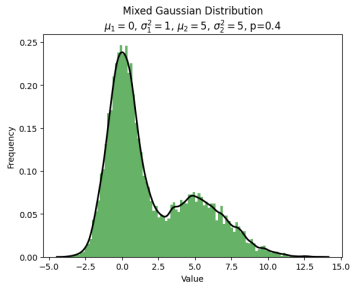

图 23: $\sigma_2^2 = 5$

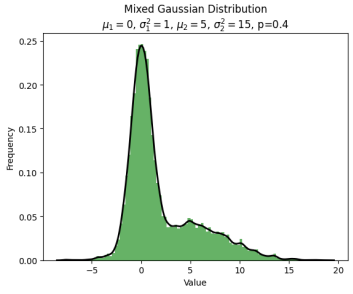

图 24: $\sigma_2^2 = 15$

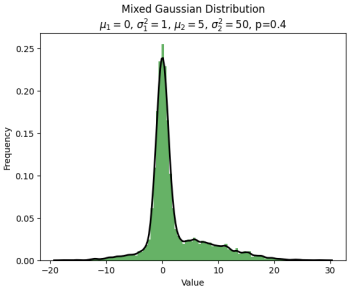

图 25: $\sigma_2^2 = 50$

图 26: σ_2^2 对“峰”的影响

通过对图中数据的仔细观察，我们可以发现 σ_2^2 对主峰的特征没有显著影响。然而，次峰的高度和陡峭程度明显依赖于 σ_2^2 的大小。随着 σ_2^2 的增加，次峰的高度逐渐降低，并且其周边变得更加平缓。当 σ_2^2 达到一定阈值后，次峰变得难以辨识。这些观察使我们可以合理推测，σ_2^2 对次峰的形态特征具有显著影响。

2.2.3 p 对“峰”的影响

控制其它参数不变，逐步改变 p 的值，绘制出如下图33所示的一组图片。

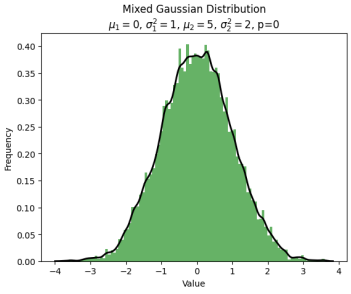

图 27: $p = 0$

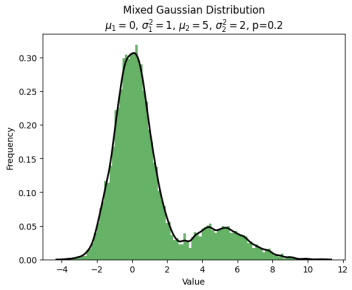

图 28: $p = 0.2$

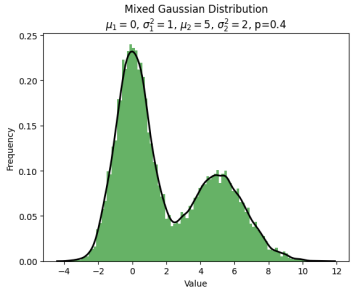

图 29: $p = 0.4$

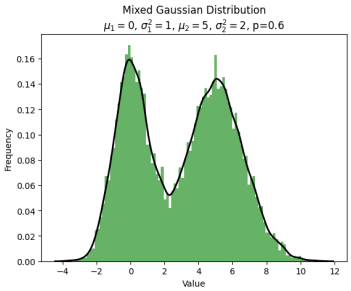

图 30: $p = 0.6$

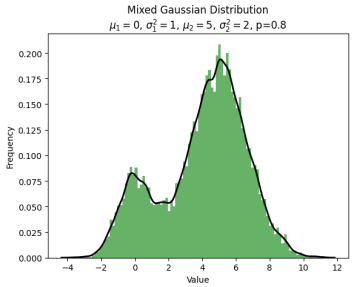

图 31: $p = 0.8$

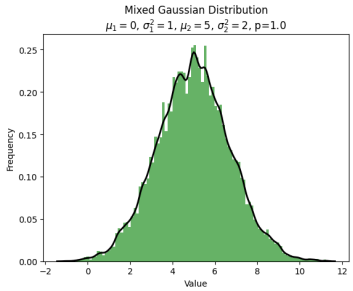

图 32: $p = 1.0$

图 33: p 对“峰”的影响

在上述图片中，本文通过逐渐调整第二高斯分量的权重参数 p 从 0 到 1.0，观察到分布形态的显著变化。最初，分布完全由第一个高斯分量控制，表现为一个集中在 μ_1 的单一峰，即一个标准正态分布的形状。随着 p 的增加，第二分量逐渐显示出其影响力，导致峰值从 μ_1 向 $\mu_2 = 5$ 移动，直至 $p = 1.0$ 时由 μ_2 和 μ_2 共同主导，表现为两个正态分布加和的图像。这一过程中，μ_1 附近的主峰逐渐降低，而 μ_2 附近的峰则逐步增高并变得更加显著，当 p 接近某一值时，会出现两峰高度相同的情况。

2.2.4 μ_1 对“峰”的影响

控制其它参数不变，仅改变 μ_1 的值，绘制出如图40一组图片。

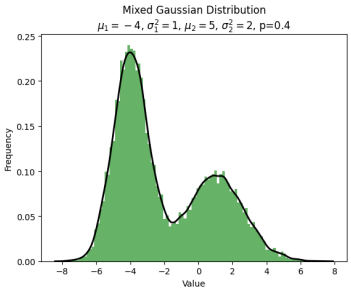

图 34: $\mu_1 = -4$

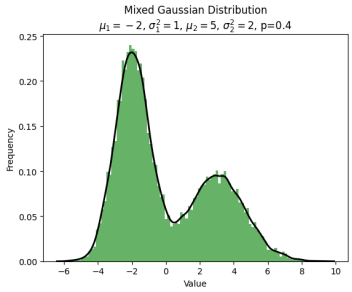

图 35: $\mu_1 = -2$

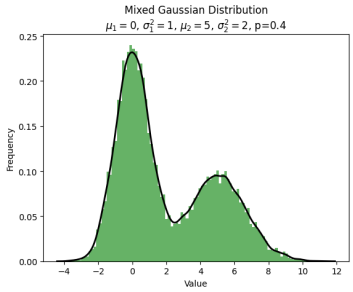

图 36: $\mu_1 = 0$

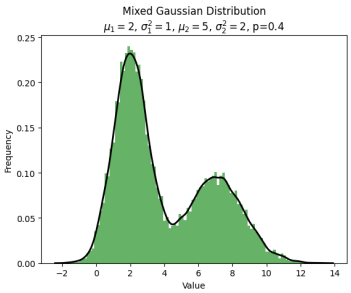

图 37: $\mu_1 = 2$

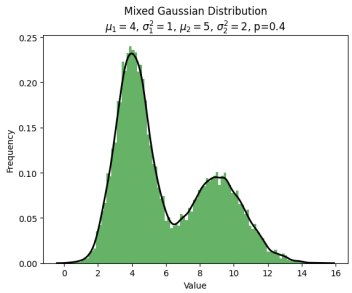

图 38: $\mu_1 = 4$

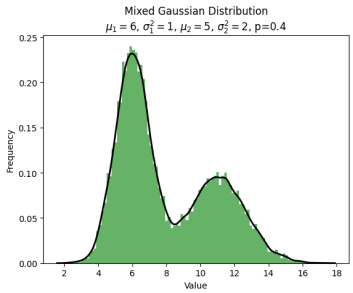

图 39: $\mu_1 = 6$

图 40: μ_1 对“峰”的影响

观察这组图片可以发现，混合高斯分布频率分布直方图的形状并没有发生改变，改变 μ_1 的结果表现为图像整体在水平方向上的平移，平移量即为 μ_1 的该变量。

2.2.5 σ_1^2 对“峰”的影响

控制其它参数不变，仅改变 σ_1^2 的值，绘制出如图47一组图片。

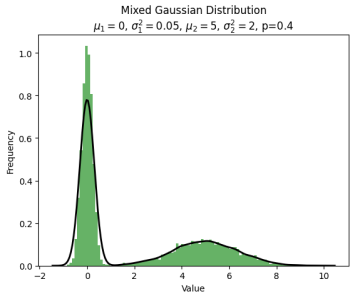

图 41: $\sigma_1^2 = 0.05$

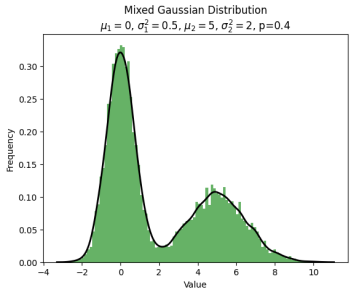

图 42: $\sigma_1^2 = 0.5$

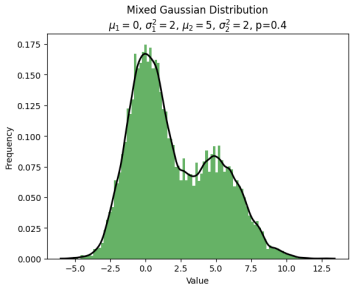

图 43: $\sigma_1^2 = 2$

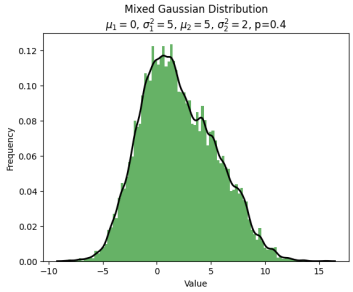

图 44: $\sigma_1^2 = 5$

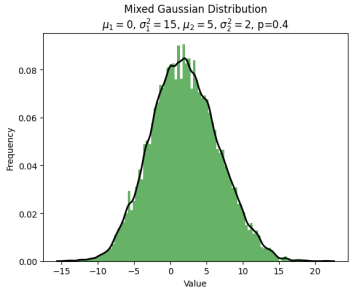

图 45: $\sigma_1^2 = 15$

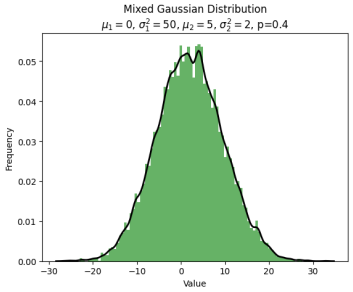

图 46: $\sigma_1^2 = 50$

图 47: σ_1^2 对“峰”的影响

通过观察这组图像，可以发现方差 σ_1^2 对主峰的影响与方差 σ_2^2 对次峰的影响具有一定的相似性。随着 σ_1^2 的增加，主峰的高度逐渐降低，峰形也变得更加平缓。当 σ_1^2 达到一定的值时，主峰变得难以辨识。为了与 σ_2^2 对次峰的影响进行比较，本实验中 σ_1^2 的变化范围与 σ_2^2 的变化范围保持一致。结果表明，σ_1^2 对主峰的影响比 σ_2^2 对次峰的影响显著更为剧烈。

2.3 利用混合高斯分布的密度函数解释上述现象

2.3.1 混合高斯分布密度函数推导

考虑随机变量 $X \sim \mathcal{N}(\mu_1, \sigma_1^2), Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$，以及 $Z = X + \eta Y$，其中 $P(\eta = 0) = p, P(\eta = 1) = 1 - p$，且 η 与 X、Y 相互独立。由此可以推得：

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

接下来我们推导混合高斯分布的累积分布函数（CDF）。首先，混合高斯分布的累积分布函数 $F_Z(z)$ 为：

$$F_Z(z) = P(Z \leq z) = P(X + \eta Y \leq z) \tag{1}$$

由于 η 取值为 0 或 1，我们可以展开：

$$F_Z(z) = P(X \leq z | \eta = 0)P(\eta = 0) + P(X + Y \leq z | \eta = 1)P(\eta = 1) \tag{2}$$

进一步代入 $P(\eta = 0) = 1 - p$ 和 $P(\eta = 1) = p$，我们得到：

$$F_Z(z) = (1 - p) \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx + p \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_2^2}} e^{-\frac{(x-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}} dx \tag{3}$$

对上式求导，即可得到密度函数 $f_Z(z)$：

$$f_Z(z) = (1 - p) \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(z-\mu_1)^2}{2\sigma_1^2}} + p \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_2^2}} e^{-\frac{(z-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}} \tag{4}$$

2.3.2 密度函数对前文推测的解释

通过分析混合高斯分布的密度函数4可以得出以下结论：

1. 增大 μ_2 会使两个峰之间的间距增大，这与2.2.1中的结果一致。而 μ_1 的变化则会使两个峰平移相同的距离，这与2.2.4中的结论相符。从密度函数的角度来看，混合高斯分布的对称性取决于 μ_1 和 μ_2 的相对位置。当 $\mu_1 = 0$ 且 μ_2 关于 μ_1 对称时，密度函数中的两个峰分别位于 0 和 μ_2 处。在这种情况下，若 μ_2 相对于 μ_1 对称，两个峰的位置和形状将保持一致，从而使混合高斯分布呈现出明显的对称性。然而，当 $\mu_1 \neq 0$ 时，即使 μ_2 相对于 μ_1 对称，密度函数的两个峰分别位于 μ_1 和 $\mu_1 + \mu_2$ 处。此时，因 μ_1 的偏移，两个峰之间的距离会减小，且出现“挤压”现象，导致分布的整体对称性减弱。这一现象验证了2.2.1中的观察结果。

2. σ_1^2 和 σ_2^2 的变化影响了峰的形状。当方差增大时，两个峰的高度会降低，且形状变得更加平缓。尤其是由于 σ_1^2 同时作用于密度函数的两项，这使得 σ_1^2 的变化对分布的影响更加显著，验证了2.2.5和2.2.2中的结论。

3. p 代表两个高斯分布的权重，通过调整 p 的值可以改变两个峰的相对高度。当 p 增大时，次峰相对于主峰的高度也随之增大。在密度函数中，权重较大的那一项对应的峰会更加明显，这一现象验证了2.2.3中的结论。

3 任务二

3.1 混合高斯分布的期望与方差推导

由 X、Y、η 独立，可由下面的式子得出混合高斯分布的期望与方差：

1. 期望 $E(Z)$：

$$E(Z) = E(X + \eta Y) = E(X) + E(\eta)E(Y) = \mu_1 + p\mu_2$$

2. 方差 $D(Z)$：

$$D(Z) = D(X + \eta Y) = D(X) + D(\eta Y)$$

又

$$D(\eta Y) = E(\eta^2 Y^2) - E^2(\eta Y)$$

$$E(\eta^2 Y^2) = E(\eta^2)E(Y^2), \quad E(\eta Y) = E(\eta)E(Y)$$

因此，

$$D(\eta Y) = E(\eta^2)E(Y^2) - (E(\eta)E(Y))^2$$

$$= (E^2(\eta) + D(\eta))(E^2(Y) + D(Y)) - (E(\eta)E(Y))^2$$

$$= E^2(\eta)D(Y) + E^2(Y)D(\eta) + D(\eta)D(Y)$$

从而，

$$D(Z) = \sigma_1^2 + p\sigma_2^2 + \mu_2^2 p(1-p)$$

3.2 随机变量 U 的频率分布直方图及样本量 n 对其峰的影响

3.2.1 随机变量 U 的定义及其含义

统计量 U 的定义为：

$$U_i = \frac{1}{\sqrt{n\mathrm{Var}(Z)}} \left(\sum_{j=1}^{n} Z_{i,j} - n\mathbb{E}[Z] \right) \tag{5}$$

其中 $Z_{i,j}$ 是从混合高斯分布中独立抽样得到的随机变量。

观察定义式5，可以得到 U_i 实际上为 n 个共同分布为混合高斯分布的独立同分布随机变量之和并标准化后的结果。

为使其频率分布直方图特性鲜明，此处采用 $\mu_1 = 0$、$\sigma_1^2 = 1$、$\mu_2 = 20$、$\sigma_2^2 = 0.1$、$p = 0.4$，对 $n = 2, 3, 4, 5, 10, 20, 50, 1000, 5000$ 分别采样 1000 组，并绘制如下频率分布直方图。

3.2.2 随机变量 U 的频率分布直方图

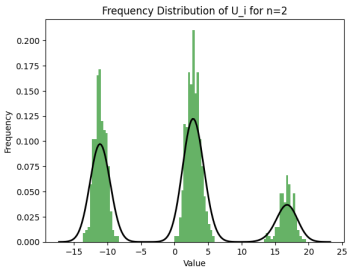

图 48: $n = 2$

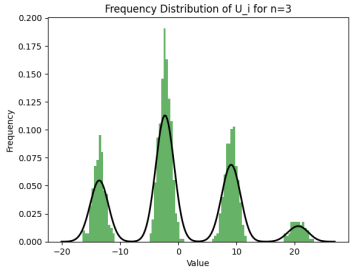

图 49: $n = 3$

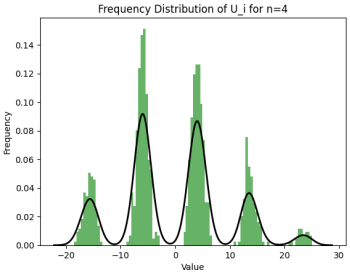

图 50: $n = 4$

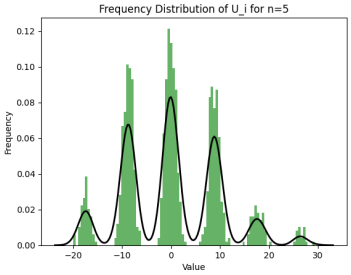

图 51: $n = 5$

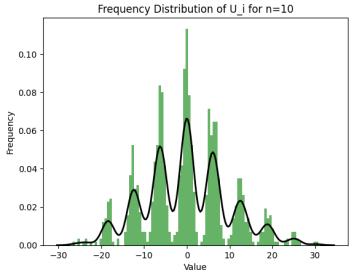

图 52: $n = 10$

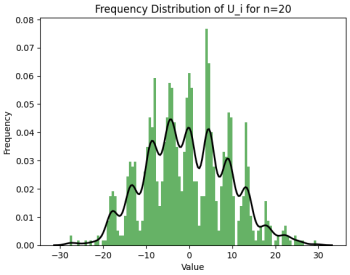

图 53: $n = 20$

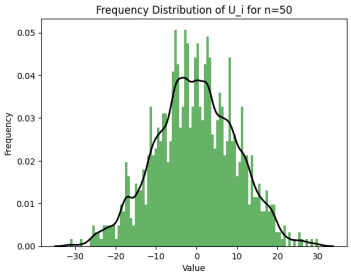

图 54: $n = 50$

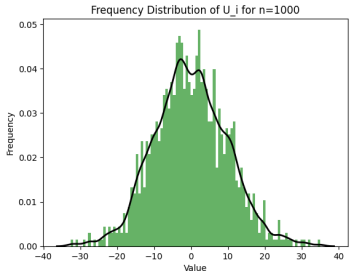

图 55: $n = 1000$

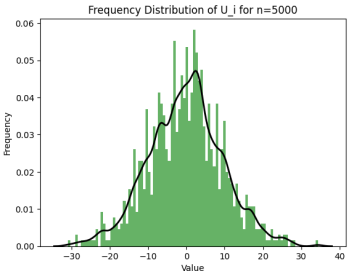

图 56: $n = 5000$

图 57: n 对“峰”的影响

观察这组图片可以发现，随着 n 的增大，“峰”的个数在逐渐增多，n 较小时，“峰“的个数约为 $n+1$ 个；但当 n 足够大时，“峰”的个数逐渐变为一个。

3.2.3 理论分析

首先我们推导混合高斯分布的矩母函数 $M_Z(t)$ ：

$$M_Z(t) = (1-p)\exp\left(t\mu_1 + \frac{t^2}{2}\sigma_1^2\right) + p\exp\left(t(\mu_1+\eta\mu_2) + \frac{t^2}{2}(\sigma_1^2+\eta^2\sigma_2^2)\right).$$

计算 $\mathbb{E}[e^{tU}]$，即 U_i 的矩母函数 $M_U(t)$：

$$M_U(t) = \mathbb{E}\left[\exp\left(t\cdot\frac{1}{\sqrt{n\mathrm{Var}(Z)}}\left(\sum_{j=1}^{n} Z_{i,j} - n\mathbb{E}[Z]\right)\right)\right].$$

这里 $\sum_{j=1}^{n} Z_{i,j}$ 是从混合高斯分布中抽取的 n 个样本的和，且由于这些样本是独立同分布的，我们

有：

$$\mathbb{E}\left[\exp\left(t\cdot\frac{1}{\sqrt{n\mathrm{Var}(Z)}}\left(\sum_{j=1}^{n}Z_{i,j}-n\mathbb{E}[Z]\right)\right)\right]=\prod_{j=1}^{n}\mathbb{E}\left[\exp\left(t\cdot\frac{1}{\sqrt{n\mathrm{Var}(Z)}}\left(Z_{i,j}-\mathbb{E}[Z]\right)\right)\right].$$

注意到 $Z_{i,j}-\mathbb{E}[Z]$ 是零均值的混合高斯分布随机变量。
对每个单独的随机变量 $Z_{i,j}$，我们有：

$$\mathbb{E}\left[\exp\left(t\cdot\frac{1}{\sqrt{n\mathrm{Var}(Z)}}\left(Z_{i,j}-\mathbb{E}[Z]\right)\right)\right]=\exp\left(-\frac{t^2}{2n\mathrm{Var}(Z)}\right).$$

因为 $Z_{i,j}$ 来自混合高斯分布，这个期望值将由两个高斯成分的加权平均组成：

$$M_{Z_{i,j}-\mathbb{E}[Z]}(t)=(1-p)\exp\left(-\frac{t^2}{2n\sigma_1^2}\right)+p\exp\left(-\frac{t^2}{2n(\sigma_1^2+\eta^2\sigma_2^2)}\right).$$

因此，U_i 的矩母函数为：

$$M_U(t)=\left[(1-p)\exp\left(-\frac{t^2}{2n\sigma_1^2}\right)+p\exp\left(-\frac{t^2}{2n(\sigma_1^2+\eta^2\sigma_2^2)}\right)\right]^n. \tag{6}$$

下面利用式6解释频率分布直方图中“峰”受 n 影响而产生的变化：

- 当 n 较小时，由于样本均值 $\sum_{j=1}^{n}Z_{i,j}$ 还没有完全收敛于理论期望 $\mathbb{E}[Z]$，统计量 U_i 受到混合分布的影响。因为混合分布有两个成分，U_i 的分布在有限样本时表现出多个峰值。
- 为什么会有多个峰？
  - 每个样本的值 $Z_{i,j}$ 可能来自两个不同的高斯成分。由于 $Z_{i,j}$ 的期望不同，样本均值可能集中在不同的区域，这些区域对应于混合分布中的不同高斯成分。
  - 由于统计量 U_i 是样本均值的标准化形式，它反映了这些区域的偏差。每个高斯成分的样本均值会贡献一个峰值。
  - 因此，在有限样本的情况下，矩母函数的结果是一个多峰分布，其峰的数量大致与混合分布成分的数量相关，即大约有 $n+1$ 个峰。
  - 当样本量足够大时，样本均值趋向于 $\mathbb{E}[Z]$，矩母函数逐渐趋近于标准正态分布，从而消除了多峰现象，分布服从独立同分布的中心极限定理，呈现单峰的正态分布。

4 感想

行文至此，概率论这门课的大作业终于圆满完成。尽管不至于历经艰险，却也经历了一些曲折。从课堂上初识混合高斯分布，到逐步理解并拆解任务，从敲下第一行代码，到完成所有绘图，从查阅资料、向学长求教，到最终完成理论分析，每一步都凝聚着努力与收获。完成这份大作业后，我不仅对高斯分布这一被称为“世界最重要分布”的理解更加深入，也对概率论中的公式、定理与分布有了更加具象的体悟，同时也极大地提升了自学能力。

关于这门课，我感触良多。学期初，或许是因为刚学完概统，再加上作为双学位学生，本学期的经济学课程也涉及了不少概率统计相关内容，我一度觉得概率论与概率统计的前半部分非常相似，只是在定义上借助测度论做了延拓，而实际应用部分却并无太大差异。然而，涉及测度论的部分对我来说相对陌生且艰深，起初让我对这门课的热情有所减退。但随着课程的推进，我逐渐意识到，事实并非如此。概率论不仅探讨了许多概率统计中浅尝辄止的知识点，还涵盖了自身独特的内容，例如各种收敛性以及更复杂、更贴近现实的分布。熊老师通过现实案例引入教学内容，让我切实感受到概率论在实际生活中的广泛应用。这些体验让我逐步领悟到，概率论的价值不仅仅体现在学科间的丰富应用，更在于它能够体系化地解释真实世界中的不确定性。

在此，我衷心感谢熊老师为我们提供了这样一个难得的机会，既加深了我们对概率论的理解，也锻炼了我们的多方面能力。同时，也感谢熊老师在整个学期中的辛勤付出和耐心教学，感谢助教们的悉心批改与答疑。特别感谢谭宇学长对我的指点，他的指导让我初步掌握了矩母函数的相关知识，并成功将其应用于任务二的分析中。希望这种有趣且富有实际意义的大作业能够继续传承，让更多同学受益。

再次致以诚挚的感谢！

A Python Code

以下是用于完成本次作业的 Python 代码：

```python
import numpy as np
import scipy
import matplotlib.pyplot as plt
import seaborn as sns

def generate_mixed_gaussian(mu1, sigma1, mu2, sigma2, p, size=10000,seed=None):
    if seed is not None:
        np.random.seed(seed)
    eta = np.random.choice([0, 1], size=size, p=[1-p, p])
    X = np.random.normal(mu1, np.sqrt(sigma1), size)
    Y = np.random.normal(mu2, np.sqrt(sigma2), size)
    Z = X + eta * Y
    return Z

def plot_histogram(data, title, filename):
    # 直方图绘制
    plt.hist(data, bins=100, density=True, alpha=0.6, color='g')

    # 使用Seaborn的kdeplot绘制核密度估计轮廓线
    sns.kdeplot(data, bw_adjust=0.5, color='k', linewidth=2)

    plt.title(title)
    plt.xlabel('Value')
    plt.ylabel('Frequency')
    plt.savefig(filename)
    plt.show()

# 参数
mu1, sigma1 = 0, 1
mu2, sigma2 = 5, 2
p = 0.4
seed = 52

# 生成混合高斯分布的随机数
data = generate_mixed_gaussian(mu1, sigma1, mu2, sigma2, p, seed=seed)

# 画出频率分布直方图
plot_histogram(data, f'Mixed␣Gaussian␣Distribution\nmu1={mu1},␣sigma1^2={sigma1},␣mu2={mu2},
    ␣sigma2^2={sigma2},␣p={p}', 'mixed_gaussian.png')
```

Listing 1: 混合高斯分布频率直方图生成代码

```python
import numpy as np
import scipy
import matplotlib.pyplot as plt
import seaborn as sns

def generate_mixed_gaussian(mu1, sigma1, mu2, sigma2, p, size=10000,seed=None):
    if seed is not None:
        np.random.seed(seed)
    eta = np.random.choice([0, 1], size=size, p=[1-p, p])
    X = np.random.normal(mu1, np.sqrt(sigma1), size)
    Y = np.random.normal(mu2, np.sqrt(sigma2), size)
    Z = X + eta * Y
    return Z

def calculate_U(n, EZ, DZ, mu1, sigma1, mu2, sigma2, p):
    U = []
    for i in range(1000):
        Z = generate_mixed_gaussian(mu1, sigma1, mu2, sigma2, p, size=n, seed=i)
        U_i = (1 / np.sqrt(n * DZ)) * (np.sum(Z) - n * EZ)
        U.append(U_i)
    return U

def plot_histogram(data, title, filename):
    # 绘制直方图
    plt.hist(data, bins=100, density=True, alpha=0.6, color='g')

    # 添加标题和标签
    plt.title(title)
    plt.xlabel('Value')
    plt.ylabel('Frequency')
    plt.show()

    # 使用Seaborn的kdeplot绘制核密度估计轮廓线
    sns.kdeplot(data, bw_adjust=0.5, color='k', linewidth=2)

    # 保存图片
    plt.savefig(filename)
    plt.show()

# 参数
mu1, sigma1 = 0, 1
mu2, sigma2 = 5, 2
p = 0.4
seed = 52

# 计算混合高斯分布的期望和方差
EZ = mu1 * (1 - p) + (mu1 + mu2) * p
DZ = sigma1 * (1 - p) + (sigma1 + sigma2) * p

# 不同的n值
n_values = [2, 3, 4, 5, 10, 20, 50, 1000, 5000]

for n in n_values:
    U = calculate_U(n, EZ, DZ, mu1, sigma1, mu2, sigma2, p)
    plot_histogram(U, f'Frequency␣Distribution␣of␣U_i␣for␣n={n}', filename=f'n={n}.png')
```

Listing 2: 随机变量 U 频率直方图生成代码