混合高斯分布相关问题的讨论

518021910677 朱展达

Dec. 10th, 2019

1 问题说明

混合高斯分布：$X \sim N(\mu_1, \sigma_1^2)$，$Y \sim N(\mu_2, \sigma_2^2)$，变量 η 满足二项分布。则称 $Z = X + \eta Y$ 服从的分布为混合高斯分布。其中，η 服从的二项分布如下表：

表 1: η 分布列

η	0	1
P	$1-p$	p

问题 1：

- 自己设定参数，用计算机生成 10000 个混合高斯分布的随机数；
- 画出其频率分布直方图；
- 讨论不同参数对其分布“峰”的影响。

问题 2：自己设定参数，用计算机生成 1000 组，每组 n 个混合高斯分布的随机数。第 i 组随机数记为:$Z_{i,1}, Z_{i,2}, ..., Z_{i,n}$，$i = 1, 2, ..., 1000$。定义

$$U_i = \frac{\sum_{j=1}^{n} Z_{i,j} - nEZ}{\sqrt{nDZ}} \tag{1}$$

- 画出 $U_1, U_2, ..., U_{1000}$ 的频率分布直方图；
- 讨论不同 $n = 10, 20, 50, 100, 1000$ 对频率直方图“峰”的影响；
- 你能从中得到什么结论？

2 问题分析、求解思路与代码

2.1 问题分析

问题 1 主要是探索混合高斯分布，根据其形式，其应为两个正态分布的加权平均，需要考虑 $\mu_1, \sigma_1, \mu_2, \sigma_2, p$ 对峰的影响；问题 2 主要是利用混合高斯分布来探索验证 Lindeberg-Lévy 中心极限定理。

2.2 问题 1 求解思路与代码

通过 matlab 生成相应混合高斯分布，利用 hist 和 bar 函数生成相应的频率分布直方图，通过固定其中四个参数，多次改变另一个参数，来分析其对峰的影响。代码如下：

```
clear;
n = 10000;
mu1 = Input_mu1, sigma1 = Input_sigma1;
mu2 = Input_mu2, sigma2 = Input_sigma2;
p = Input_p;
x1 = normrnd(mu1,sigma1,[n,1]);
x2 = normrnd(mu2,sigma2,[n,1]);
x3 = unifrnd(0,1,[n,1]);
eta = zeros(n,1);
[counts_x1, centers_x1] = hist(x1, 100);
[counts_x2, centers_x2] = hist(x2, 100);
% bar(centers_x1, counts_x1 / sum(counts_x1));
% bar(centers_x2, counts_x1 / sum(counts_x1));
eta(x3 ≤ p) = 1;
Z = x1 + eta.* x2;
[counts_Z, centers_Z] = hist(Z, 80);
bar(centers_Z, counts_Z / sum(counts_Z))
```

2.3 问题 2 求解思路与代码

由于式（1）中 EZ 和 DZ 为理论值，先计算混合高斯分布 Z 的均值和方差：

$$EZ = E(X+\eta Y) = EX + E\eta \cdot EY = \mu_1 + p\mu_2 \tag{2}$$

$$\begin{aligned} EZ^2 &= E((X+\eta Y)^2) = E(X^2 + 2\eta XY + \eta^2 Y^2) \\ &= EX^2 + 2(EX)(EY)(E\eta) + E(\eta^2)E(Y^2) \\ &= \mu_1^2 + \sigma_1^2 + 2\mu_1\mu_2 p + (\mu_2^2 + \sigma_2^2)p \end{aligned} \tag{3}$$

$$DZ = EZ^2 - (EZ)^2 = \sigma_1^2 + p\sigma_2^2 + p(1-p)\mu_2^2 \tag{4}$$

确定合适的参数，生成 1000 组混合高斯分布的随机数，分别计算 U_i，$i = 1, 2, ..., 1000$。做 $n = 10, 20, 50, 100, 1000$ 的图进行比较讨论。

由于 EZ 和 DZ 在 $(\mu_1, \sigma_1, \mu_2, \sigma_2, p)$ 确定时是常量，由 Lindeberg-Lévy 中心极限定理知，当 n 足够大时，U_i 服从标准正态分布。为了使图像随着 n 的改变变化明显，要尽量破坏原分布的正态分布性，因为对于 Z 分布得到的频率直方图的两个“峰”，当 σ 较小时，取值分布在“峰”两边。因此令 $|\mu_2|$ 很大时，x 取值在两峰之间的概率很小，可以破坏原有的正态分布性。

代码如下：

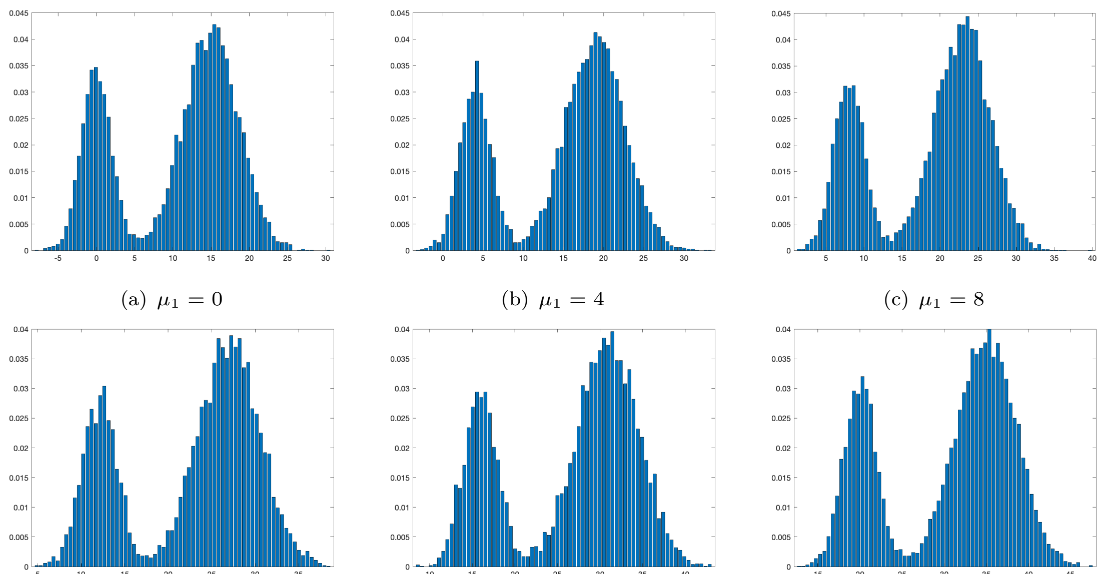
```
clear;
n = 1000;
mu1 = 0, sigma1 = 2, mu2 = 15, sigma2 = 3;
p = 0.7;
x1 = normrnd(mu1,sigma1,[1000,n]);
x2 = normrnd(mu2,sigma2,[1000,n]);
x3 = unifrnd(0,1,[1000,n]);
```

```
eta = zeros(1000,n);
eta(x3 ≤ p) = 1;
Z = x1 + eta.∗ x2;

EZ = mu1 + p ∗ mu2;
DZ = sigma1^2 + p ∗ sigma2^2 + p ∗ (1-p) ∗ mu2^2;

for i = 1 : 1000
    temp = 0;
    for j = 1 : n
        temp = Z(i,j) + temp;
    end
    U(i) = (temp - n ∗ EZ) / sqrt(n ∗ DZ);
end

[counts_U, centers_U] = hist(U, 80);
bar(centers_U, counts_U / sum(counts_U))
```

3 问题 1 求解

3.1 讨论 μ_1 对分布“峰”的影响

固定参数 $\sigma_1 = 2, \mu_2 = 15, \sigma_2 = 3, p = 0.7$，改变 μ 取值，令 μ_1 分别为 $0, 4, 8$ $12, 16, 20$, 观察得到的频率分布直方图。



(a) $\mu_1 = 0$　(b) $\mu_1 = 4$　(c) $\mu_1 = 8$

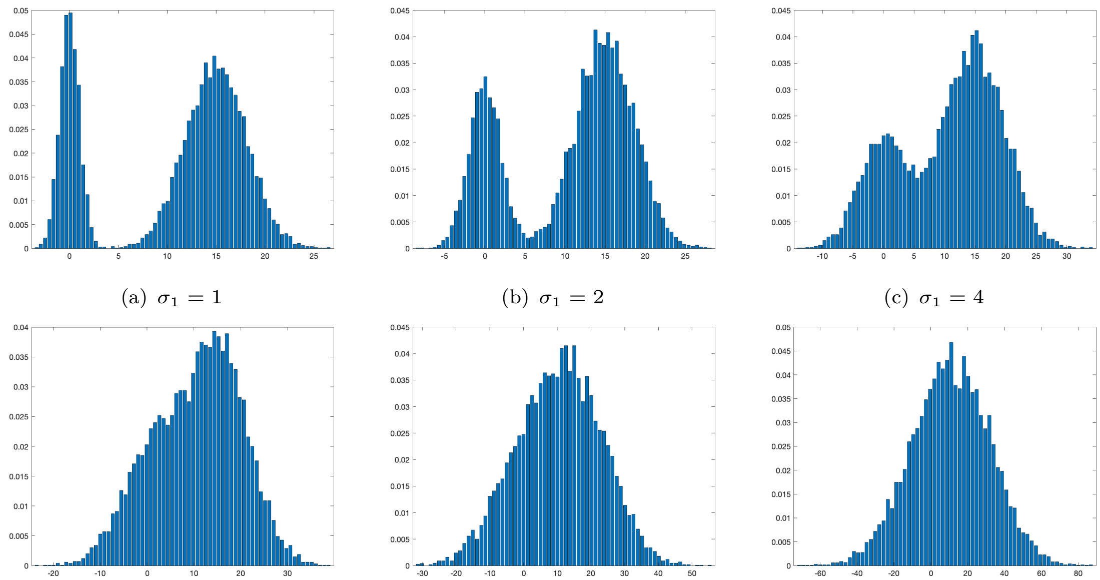
(d) $\mu_1 = 12$　(e) $\mu_1 = 16$　(f) $\mu_1 = 20$

图 1: $\sigma_1 = 2, \mu_2 = 15, \sigma_2 = 3, p = 0.7$ 下，μ_1 变化时生成的随机数的频率分布直方图

结论：从上述六图中我们可以发现，在 μ_1 改变时，随机数的频率分布直方图的形态基本保持不变，整体图像仅随 μ_1 的变化而平移。其中，第一个峰对应的 x 坐标为 μ_1，第二个峰对应的 x 坐标为 $\mu_1 + \mu_2$。

3.2 讨论 σ_1 对分布“峰”的影响

固定参数 $\mu_1 = 0, \mu_2 = 15, \sigma_2 = 3, p = 0.7$，改变 σ_1 取值，令 σ_1 分别为 $1, 2, 4, 6, 10, 20$，观察得到的频率分布直方图。



(a) $\sigma_1 = 1$ (b) $\sigma_1 = 2$ (c) $\sigma_1 = 4$

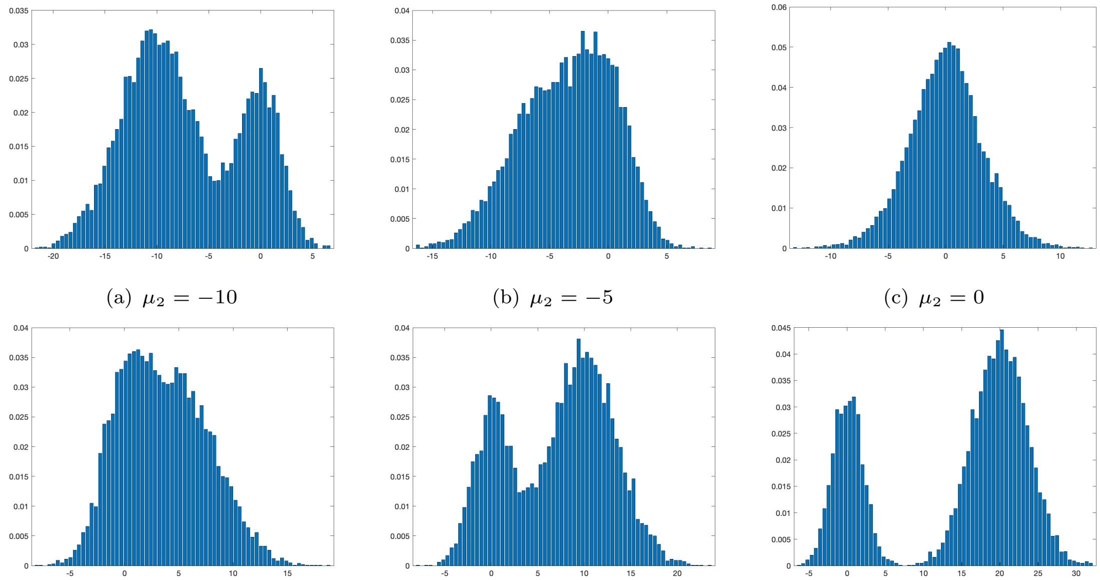
(d) $\sigma_1 = 6$ (e) $\sigma_1 = 10$ (f) $\sigma_1 = 20$

图 2: $\mu_1 = 0, \mu_2 = 15, \sigma_2 = 3, p = 0.7$ 下，σ_1 变化时生成的随机数的频率分布直方图

结论：从上述六图中我们可以发现，在 σ_1 变化时，“峰”的分布、数量、高度和峰两侧对应的斜率都会产生变化。当 μ_2 不为 0 时，σ_1 较小时会出两个“峰”，σ_1 较大时两个“峰”会合并成一个“峰”。随着 σ_1 的增大，μ_1 对应的“峰”高度下降并且变得逐渐趋于平缓，$\mu_1 + \mu_2$ 对应的“峰”高度有所升高，最终会产生两峰合并，只留下 $\mu_1 + \mu_2$ 对应的峰的情况。

3.3 讨论 μ_2 对分布“峰”的影响

固定参数 $\mu_1 = 0, \sigma_1 = 2, \sigma_2 = 3, p = 0.7$，改变 μ_2 取值，令 μ_2 分别为 $-10, -5, 0, 5, 10, 20$，观察得到的频率分布直方图。



(a) $\mu_2 = -10$ (b) $\mu_2 = -5$ (c) $\mu_2 = 0$

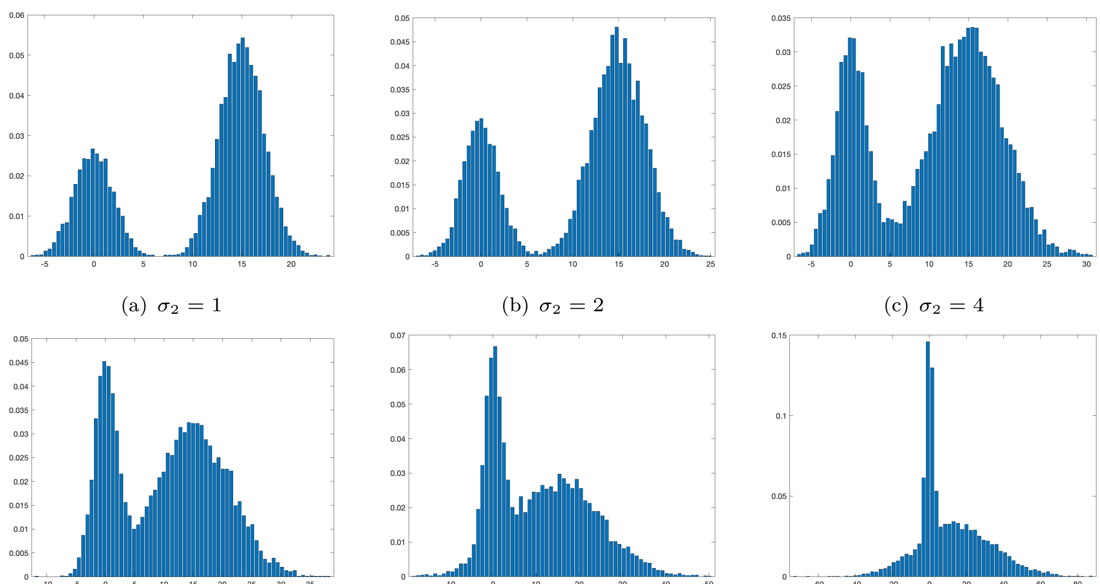
(d) $\mu_2 = 5$ (e) $\mu_2 = 10$ (f) $\mu_2 = 20$

图 3: $\mu_1 = 0, \sigma_1 = 2, \sigma_2 = 3, p = 0.7$ 下，μ_2 变化时生成的随机数的频率分布直方图

结论：从上述六图中我们可以发现，在 μ_2 变化时，“峰”的数量和分布会发生变化。其中一个“峰”对应的 x 坐标一直为 μ_1，另一个“峰”对应的 x 坐标 $\mu_1 + \mu_2$ 会发生改变。当 $|\mu_2|$ 较大时，有明显的两个“峰”，当 $|\mu_2|$ 较小时，两座峰逐渐融合，当 $\mu_2 = 0$ 时，完全只剩下一个“峰”。同时显然的，μ_2 的正负会影响两座峰的左右排布。

3.4 讨论 σ_2 对分布“峰”的影响

固定参数 $\mu_1 = 0, \sigma_1 = 2, \mu_2 = 15, p = 0.7$，改变 σ_2 取值，令 σ_1 分别为 $1, 2, 4, 6, 10, 20$，观察得到的频率分布直方图。



(a) $\sigma_2 = 1$ (b) $\sigma_2 = 2$ (c) $\sigma_2 = 4$

(d) $\sigma_2 = 6$ (e) $\sigma_2 = 10$ (f) $\sigma_2 = 20$

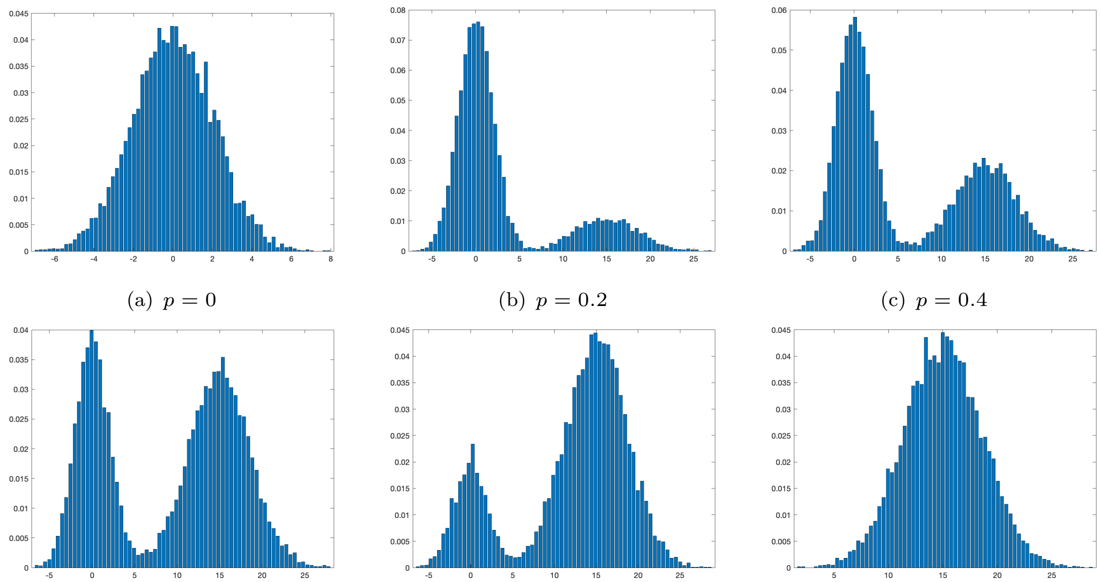
图 4: $\mu_1 = 0, \sigma_1 = 2, \mu_2 = 15, p = 0.7$ 下，σ_2 变化时生成的随机数的频率分布直方图

结论：从上述六图中我们可以发现，在 σ_2 变化时，“峰”的相对高度会发生变化。随着 σ_2 的

增大，μ_1 对应的峰的高度升高，$\mu_1+\mu_2$ 对应的峰的高度下降，μ_1 对应的峰与 $\mu_1+\mu_2$ 对应的峰的相对高度的代数值增大。且特别地，当 σ_2 特别大时，$\mu_1+\mu_2$ 所对应的峰会趋于消失，只剩下 μ_1 所对应的峰。

3.5 讨论 p 对分布“峰”的影响

固定参数 $\mu_1=0,\sigma_1=2,\mu_2=15,\sigma_2=3$，改变 p 取值，令 p 分别为 $0,0.2,0.4,0.6,0.8,1.0$，观察得到的频率分布直方图。



(a) $p=0$　(b) $p=0.2$　(c) $p=0.4$

(d) $p=0.6$　(e) $p=0.8$　(f) $p=1.0$

图 5: $\mu_1=0,\sigma_1=2,\mu_2=15,\sigma_2=3$ 下，p 变化时生成的随机数的频率分布直方图

结论： 从上述六图中我们可以发现，在 p 变化时，“峰”的数量、分布和高度会发生变化。$p=0$ 时，只有一个对应 x 坐标为 μ_1 的“峰”；随着 p 的增大，μ_1 对应的“峰”的高度下降，$\mu_1+\mu_2$ 对应的“峰”的高度升高，μ_1 对应的“峰”与 $\mu_1+\mu_2$ 对应的“峰”的相对高度的代数值降低；当 $p=1$ 时，只剩下一个对应 x 坐标为 $\mu_1+\mu_2$ 的峰。

3.6 问题一结论总结与概括

- 混合高斯分布是两个正态分布的加权平均，每个正态分布单独存在时的频率分布直方图存在一个“峰”。故混合高斯分布得到的频率分布直方图存在两个“峰”，他们对应的 x 坐标分别为 μ_1，$\mu_1+\mu_2$，但这两个“峰”的位置、高度、平缓程度会受五个参数 $(\mu_1,\sigma_1,\mu_2,\sigma_2,p)$ 的影响，有时两个峰变成只有一个峰。

- 1. μ_1 会影响“峰”分布的 x 坐标，“峰”随 μ_1 的变化而发生平移。
  2. σ_1 会影响“峰”的数量、高度、平缓程度。随着 σ_1 的增大，μ_1 对应的“峰”高度下降并且变得逐渐趋于平缓，$\mu_1+\mu_2$ 对应的“峰”的高度升高；当 σ_1 极大时，只剩下 $\mu_1+\mu_2$ 对应的“峰”。
  3. μ_2 会影响“峰”的数量和分布。$|\mu_2|$ 较大时，有明显的两个“峰”，$|\mu_2|$ 逐渐变小时，两座峰逐渐融合，当 $\mu_2=0$ 时，只剩下一个“峰”。

4. σ_2 会影响“峰”的数量、高度、平缓程度。随着 σ_2 的增大，μ_1 对应的“峰”高度升高，$\mu_1+\mu_2$ 对应的“峰”高度下降并逐渐趋于平缓；当 σ_2 极大时，只剩下 μ_1 对应的“峰”。
5. p 会影响“峰”的数量、分布和高度。$p=0$ 时，只有一个对应 x 坐标为 μ_1 的“峰”；随着 p 的增大，μ_1 对应的“峰”的高度下降，$\mu_1+\mu_2$ 对应的“峰”的高度升高，μ_1 对应的“峰”与 $\mu_1+\mu_2$ 对应的“峰”的相对高度的代数值降低；当 $p=1$ 时，只剩下一个对应 x 坐标为 $\mu_1+\mu_2$ 的峰。

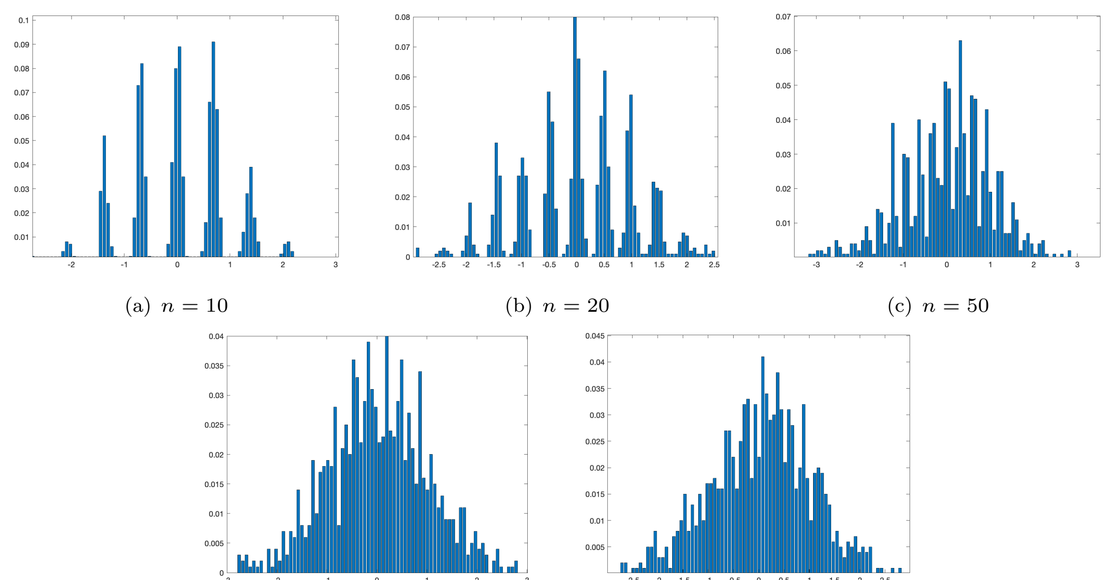
4 问题 2 求解

4.1 参数选择

根据 **2.3** 的分析，需要让 $|\mu_2|$ 值较大，取 $\mu_1=0$，$\sigma_1=2$，$\mu_2=100$，$\sigma_2=3$，$p=0.7$。

4.2 频率直方图

分别选择 $n=10,20,50,100,1000$，得到相应的关系图。



(a) $n=10$ (b) $n=20$ (c) $n=50$

(d) $n=100$ (e) $n=1000$

图 6: $\mu_1=0,\sigma_1=2,\mu_2=100,\sigma_2=3,p=0.7$ 下，n 变化时生成的 U_i 的频率分布直方图

4.3 结论和讨论

结论：当 n 比较小时，频率直方图的“峰”数量较多，且“峰”与“峰”之间间隔较大，随着 n 的增大，“峰”与“峰”之间的间距逐渐变小。同时随着 n 的增大，U_i 的分布逐渐趋向于标准正态分布。

讨论：当 $|\mu_2|$ 较小时，频率直方图的峰之间距离较小，使得原本便近似于正态分布，因此改变 n 的取值并不能得到很好的效果。但是当 $|\mu_2|$ 很大的情况下时，若 n 很小，频率直方图峰间距较大，使其与正态分布存在偏差，若 n 很大，根据 Lindeberg-Lévy 中心极限定理，其分布应近似于标准正态分布。而结果恰恰与标准正态分布相符合。成功验证了 Lindeberg-Lévy 中心极限定理。

5 总结与体悟

通过这次对混合高斯分布相关问题的讨论和实践，我对混合高斯分布中各参数的地位、作用和影响有了更深层次的理解，对其图像和性质有了较好的把我；通过 matlab 生成对应的分布分布，让我具体理解了 Monte-Carlo method 和计算机仿真模拟在概率统计中的重要作用；通过对混合高斯分布的讨论，还进而验证了 Lindeberg-Lévy 中心极限定理，让我明白了其实际体现。总之，此次探索和讨论让我对概率统计问题的研究方法有所涉猎，更加理解概率密度函数在现实中的表现，还提高了我的计算机编程能力和论文撰写能力。

6 致谢

感谢熊德文老师的认真授课和点拨启发。

感谢助教对本次大作业付出的辛劳。

感谢方泓杰同学在 Latex 版式上提供的帮助。

参考文献

[1] 上海交通大学数学系.《概率论与数理统计》. 上海交通大学出版社.2011