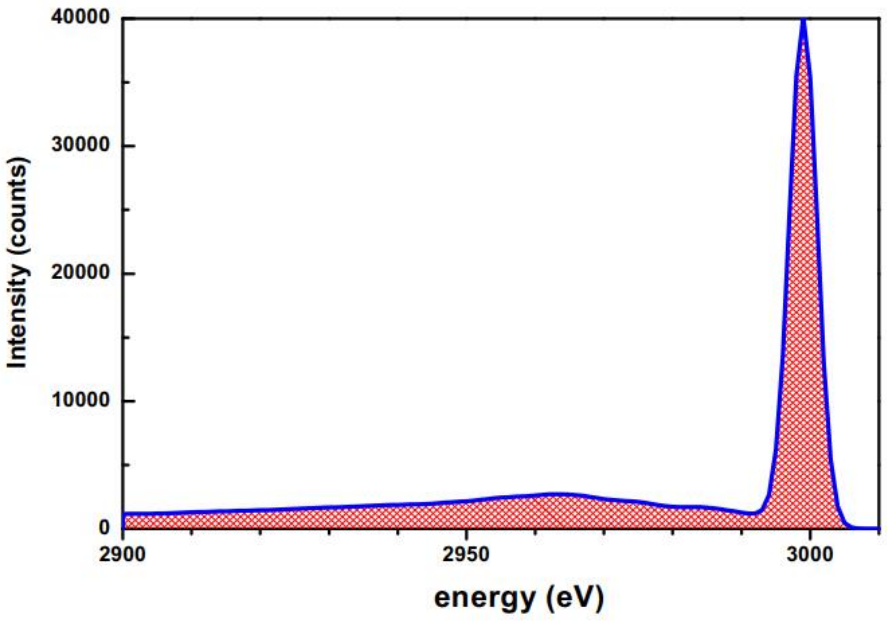
10:直接抽样法和舍选抽样法

许传奇 PB16021546

1 题目

对一个实验谱函数$p(x)$，自设$F(x)$，分别用直接抽样法和舍选法对$p(x)$抽样。比较原曲线和抽样得到的曲线并验证。讨论抽样效率。



2 原理与算法

2.1 原理

2.1.1 离散型变量的直接抽样法

设变量x是离散性的，取值为$x_1, x_2, \ldots$，相应值出现的几率为$p_1, p_2, \ldots$。对于一个物理量常常给出的不是归一化的几率值p，而是其截面值σ，则可将其归一化为几率。

$$o_i = \frac{\sigma_i}{\sum_{i=1}^{n} \sigma_i} \tag{1}$$

如果从$[0,1]$区间中均匀抽样得到的随机数ξ满足：

$$\sum_{i=1}^{n-1} p_i < \xi \leq \sum_{i=1}^{n} p_i \tag{2}$$

则物理量x取值为x_n。

这种方法就是离散型变量的直接抽样法。

2.1.2 舍选抽样法

在某些情况下，我们无法采用直接抽样法和变换抽样法来进行抽样，这时我们可以采取舍选抽样法。

舍选抽样法的思想是：对于一个难以用常规方法抽样的分布（设其概率密度函数为$f(x)$），我们选取另一个容易抽样的函数（设其概率密度函数为$g(x)$）进行抽样，得到$g(x)$的随机抽样。然后再对这个分布内的点进行舍去和选择，就能得到$f(x)$的抽样。

具体步骤如下：

1. 对$g(x)$进行抽样，得到ξ_x（如在[0,1]上抽取ξ，再根据累计函数的反函数求得ξ_x，就得到了$g(x)$的抽样ξ_x）；
2. 抽取在[0, g(ξ_x)]均匀分布的ξ_y，比较$\xi_y < f(\xi_x)$是否成立；
3. 如果不成立，则返回到（1）；如果成立，则ξ_y就是f(x)的抽样。

$g(x)$选取的原则是：容易求得其抽样；在抽样的区间内满足：$f(x) \leq g(x)$

2.2 算法

读取数据后，画出直方图，如下图所示：

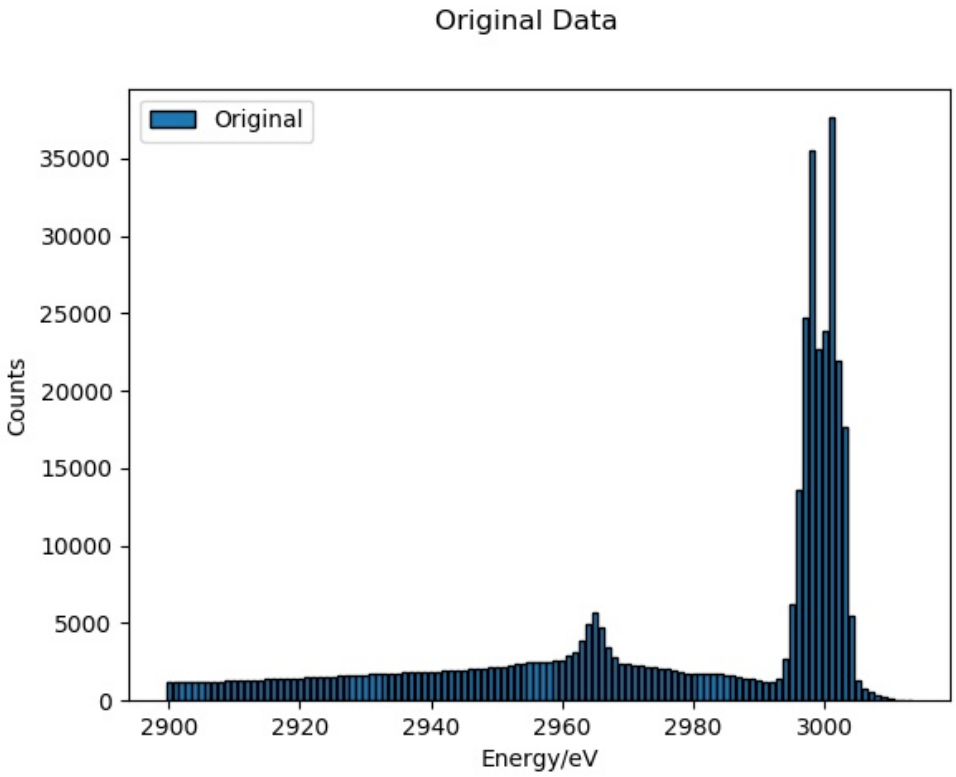


图 1: 读取数据对应直方图

2.2.1 直接抽样法

由于data.TXT文件中的能量与计数是离散的，因此我们使用离散分布的直接抽样法。
记能量为$x_k = 2900 + k,\ k = 0, 1, ..., 113$，与$x_k$对应的计数为$y_k$。
先抽取在$[0, 1]$上分布的ξ，当其满足：

$$\frac{\sum_{k=0}^{m-1} y_k}{\sum_{k=0}^{113} y_k} < \xi \leq \frac{\sum_{k=0}^{m} y_k}{\sum_{k=0}^{113} y_k} y_k \tag{3}$$

与ξ对应的x为x_m。
因此，直接抽样法的步骤如下：

1. 在$[0, 1]$上抽取均匀分布的ξ;
2. 先遍历随机数ξ;
3. 再从小到大遍历m，找到第一个使该ξ满足上述关系的m;
4. 将对应的x_m打印到文件中，退出当前遍历m的循环。

2.2.2 舍选法

由数据对应的直方图可知，当能量在[2995, 3007]时，计数非常大，而其他的地方计数非常小。为了方便，我们舍选法的$f(x)$选择分段的常函数，具体形式如下：

$$f(x) = \begin{cases} 0.015 & 2900 \leq x < 2994 \\ 0.100 & 2994 \leq x \leq 3013 \end{cases} \tag{4}$$

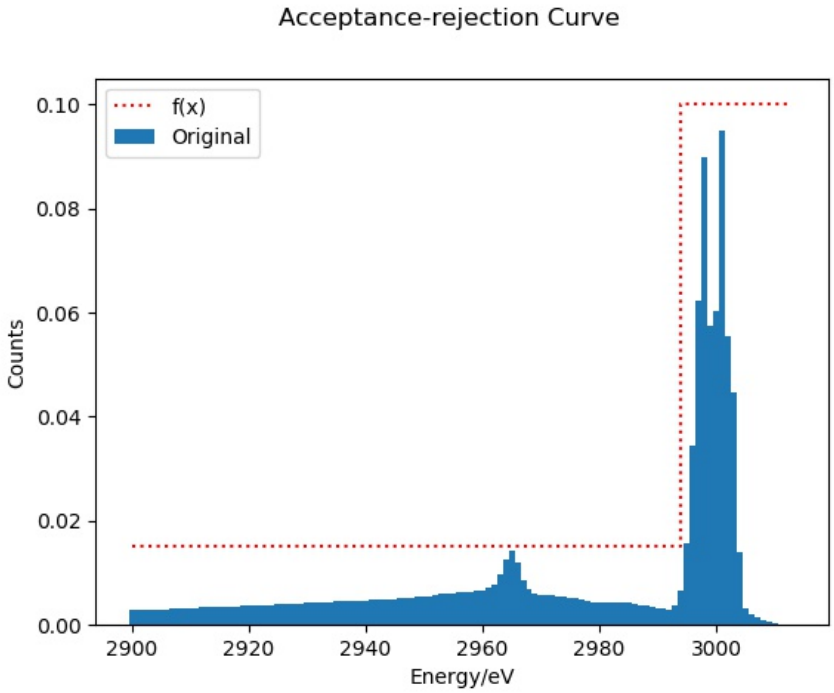


图 2: 舍选法函数选择

如图所示，函数能够包住待抽样的区域，因此可以选为舍选法的函数。

因此，离散的累积函数为：

$$F(x)=\begin{cases}0.015(x-2899) & 2900\le x<2994\\ 0.100(x-2994)+1.41 & 2994\le x\le 3013\end{cases} \tag{5}$$

总的$F(x)$为：

$$F(3013)=\sum_{x=2900}^{3013}f(x)=3.41 \tag{6}$$

舍选法的步骤如下：

1. 在$[0,1]$上抽取均匀分布的ξ_x；
2. 判断ξ_x的大小：
   (a) 当$\xi_x\le\frac{1.41}{3.41}=\frac{141}{341}$时，$x=int(\frac{3.41}{0.015}\xi_x)+2900=int(\frac{682}{3}\xi_x)+2900$；
   (b) 当$\xi_x>\frac{1.41}{3.41}=\frac{141}{341}$时，$x=int(\frac{3.41\xi_x-1.41}{0.100})+2994=int(34.1\xi_x-14.1)+2994$。
3. 在$[0,1]$随机抽取均匀分布的ξ_y，判断是否满足：
$$f(x)\cdot\xi_y\le\frac{y_{x-2900}}{\sum_{k=0}^{113}y_k} \tag{7}$$
其中的y_k是在直接抽样法中叙述的data.TXT中能量为$x=2900+k$处的计数。
4. 若ξ_y满足以上关系，则ξ_x为舍选法抽取的样本；否则，重新抽取ξ_x。

3 源文件使用说明

编译并运行“10Direct_and_A-R.cpp”，将弹出命令行，要求输入总的抽样个数num。

输入总的抽样个数后，程序运行，将直接抽样法数据和舍选法数据输出到文件“direct_num=输入的num.txt”和“ra_num=输入的nun.txt”中。

编译并运行“plot.py”即可绘制出直方图。

4 计算结果及具体分析

4.1 直接抽样法

用直接抽样法抽取随机数个数分别为1000、100000、10000000的样本，如下所示：

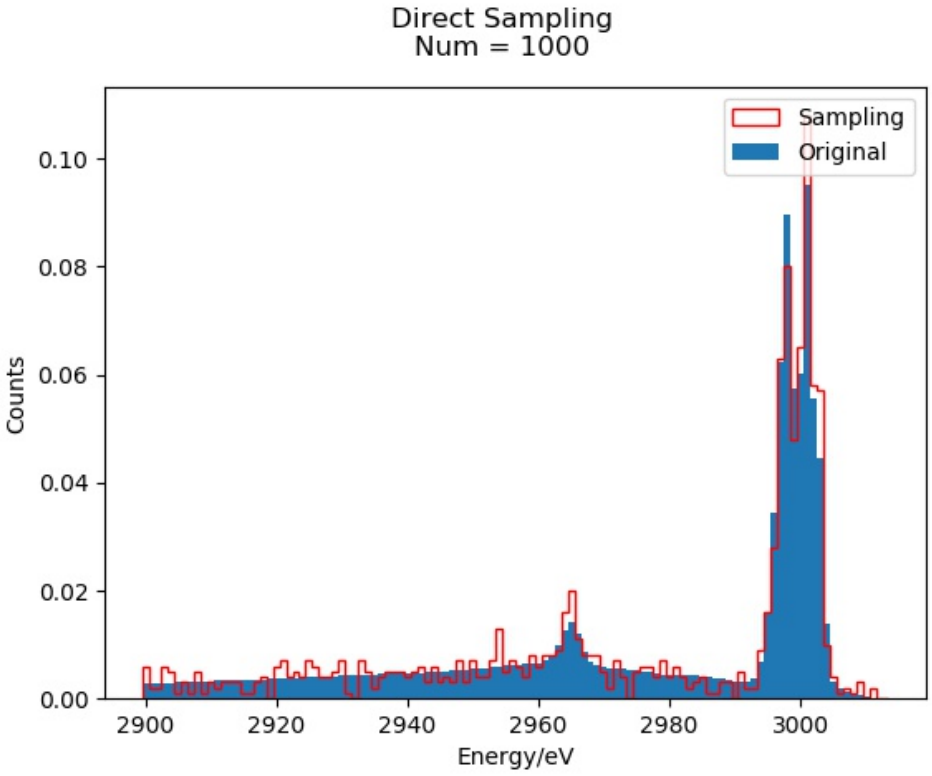


图 3: 直接抽样法Num=1000直方图

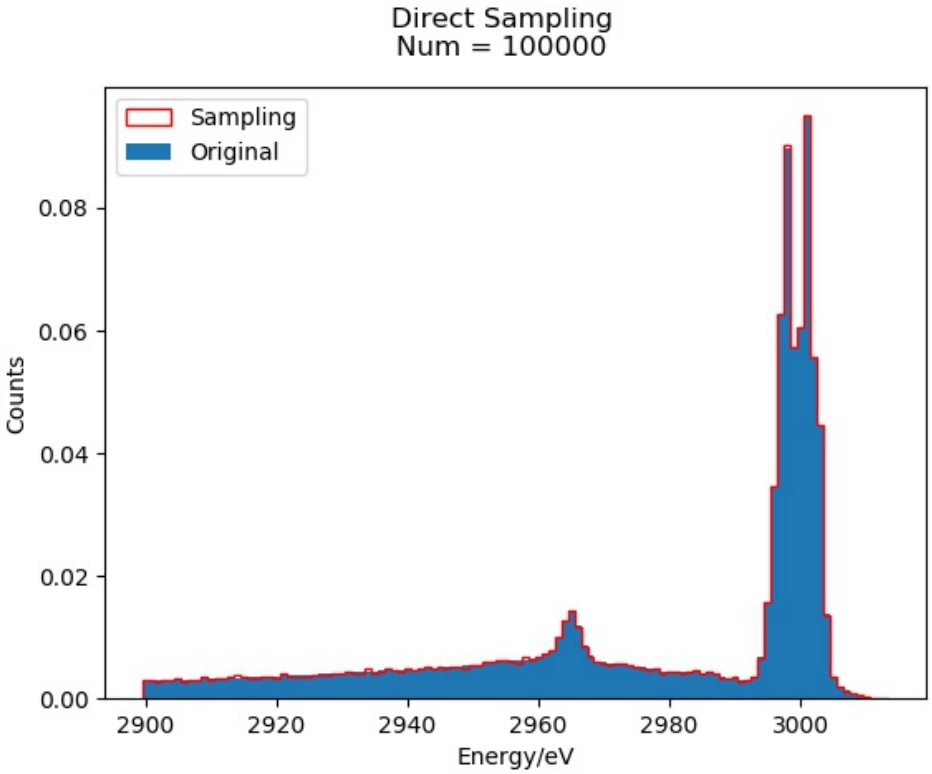


图 4: 直接抽样法Num=100000直方图

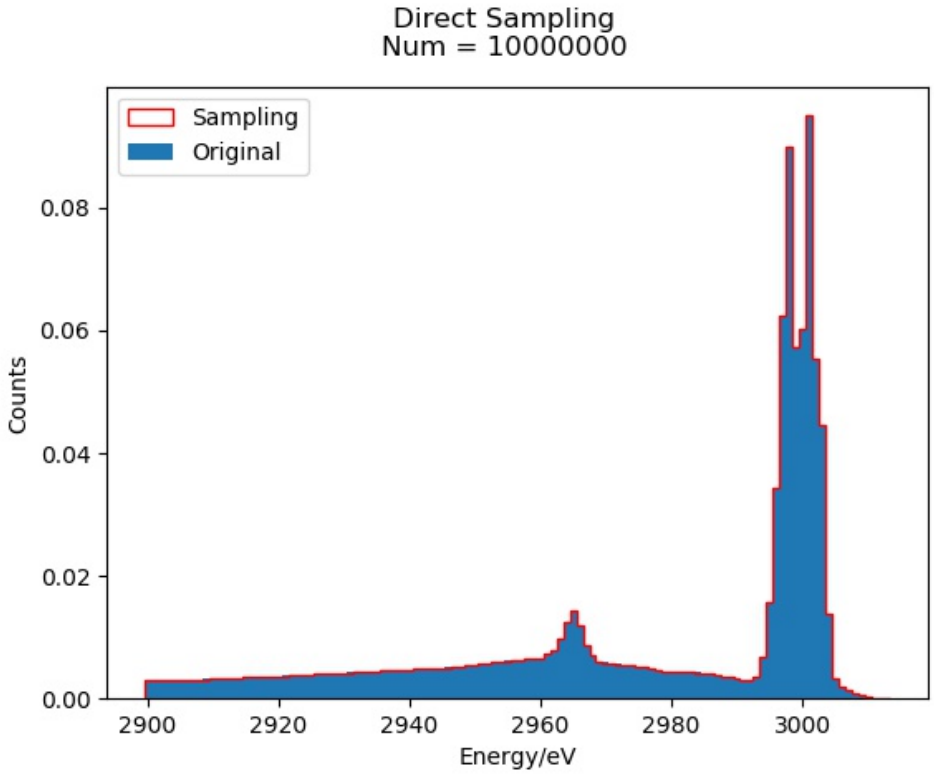


图 5: 直接抽样法Num=10000000直方图

4.2 舍选法

用舍选法抽取随机数个数分别为1000、100000、10000000的样本，如下所示：

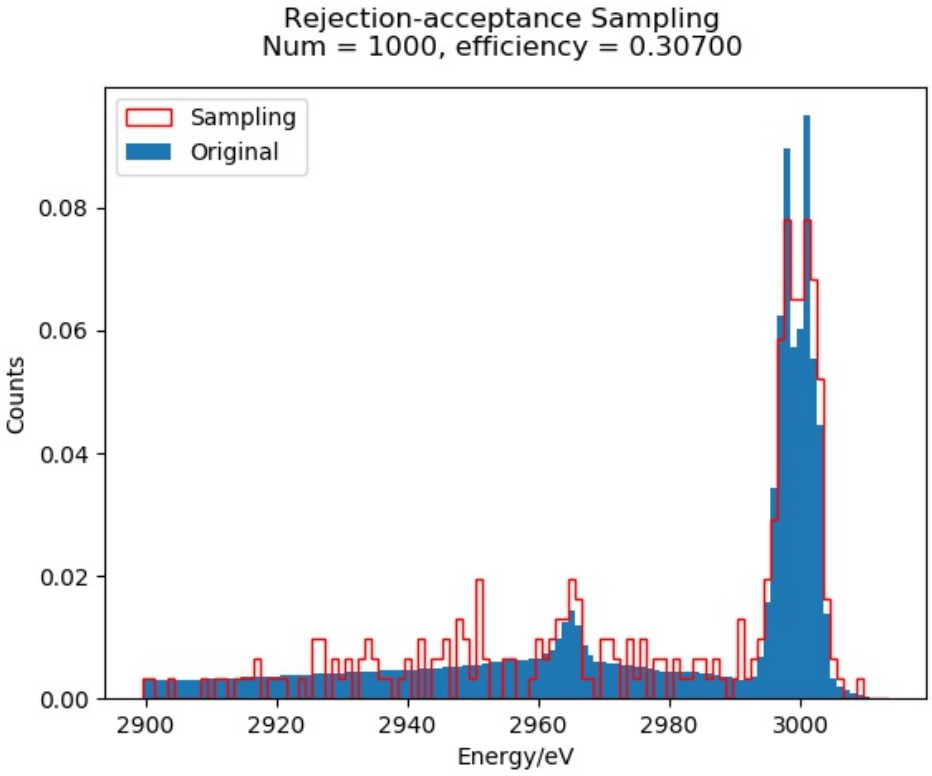


图 6: 舍选法Num=1000直方图

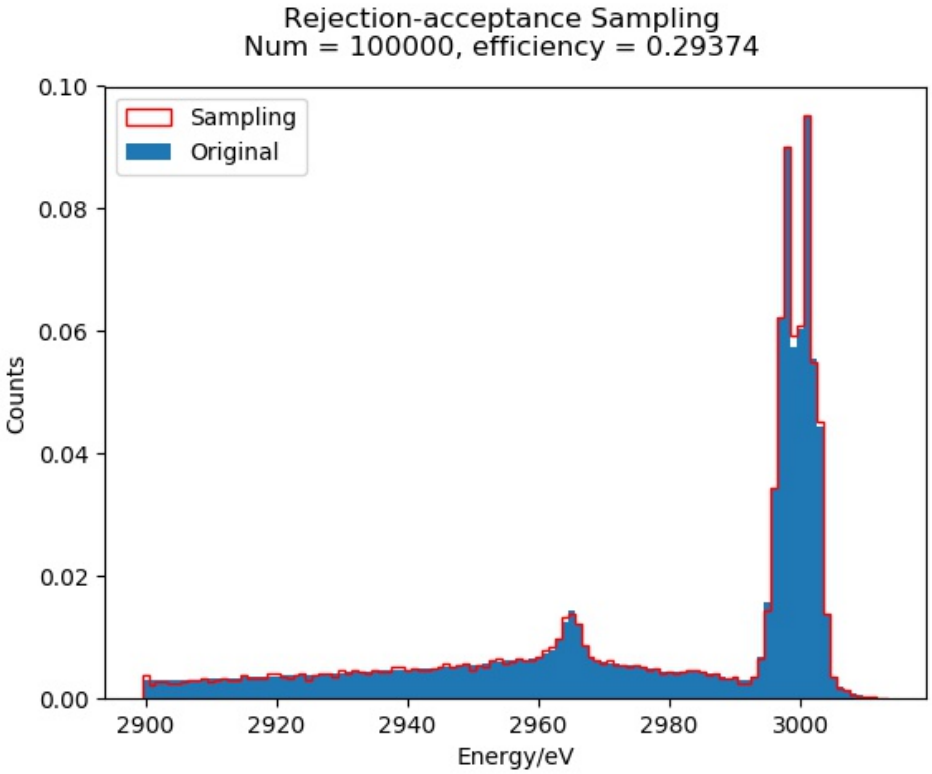


图 7: 舍选法Num=100000直方图

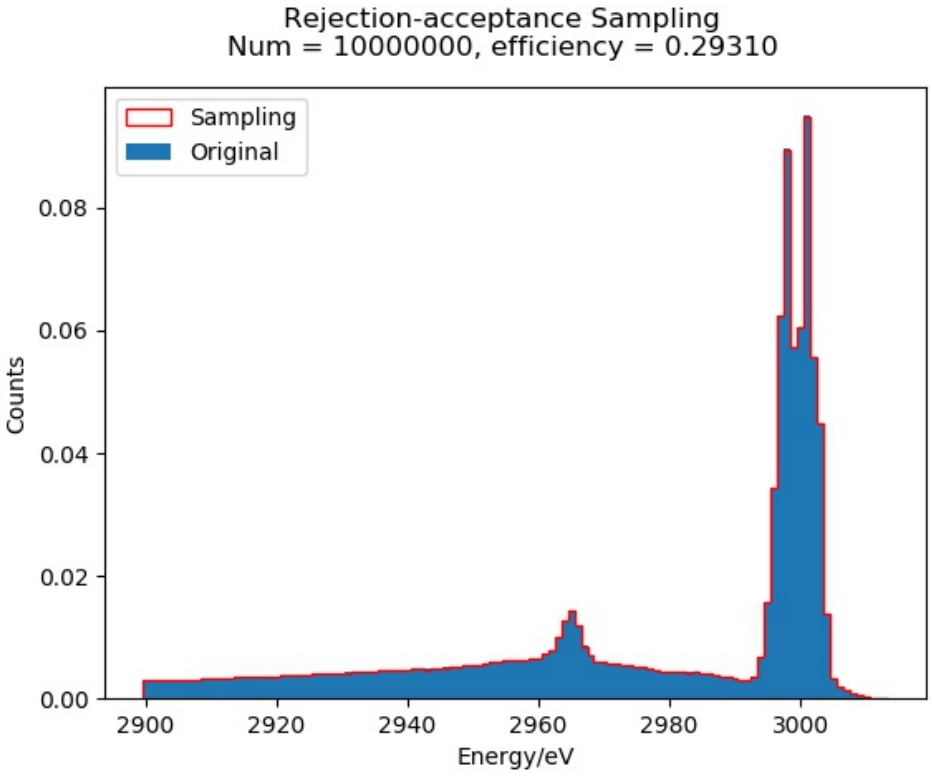


图 8: 舍选法Num=10000000直方图

5 讨论

5.1 原理的讨论

本题原理较简单，直接按照原理部分的叙述编程即可。

本题的关键是在于对原始数据的处理。由于原始数据是离散分布的计数，不需要对其进行函数插值或者其他的方法进行拟合，各种抽样方法也会因为是离散分布而变得简单。

比如直接抽样法时，不需要考虑累积函数，更不需要考虑累积函数的反函数，直接按照计数的比重进行比较就可以得到抽样。但缺点是不能直接根据累积函数的反函数直接求出抽样点，而是用循环进行遍历，这会使时间复杂度的指数加一，当数据点较多时，会显著增加时间复杂度。

对于舍选法，则需要先把原始数据化成归一化的直方图，选择既简单抽样效率又高的函数进行抽样。由于是离散分布，舍选法的抽样也不难，关键还是在进行舍选的函数的选择上。根据得到的结果显示，抽样效率并不高，因此本次作业选择的舍选函数并不是十分好。

5.2 算法的改进

本题中为了简单，直接选择的是两段的分段函数进行舍选法抽样，这样选取$F(x)$得到的抽样效率适中。

如果直接选择不分段的函数进行舍选法抽样，从原始数据的直方图中可以看出，数据最高点对应的比重要大于0.10，因此抽样效率要低于10%。这样的抽样效率太低了。

从抽样效率的结果来看，抽样效率约等于29%，对于舍选法来说实在不高。这也主要是因为用来进行舍选的函数选取的不好。可以分段更多，使其与原始数据拟合的更好，即与原始数据之间的差距减小，能够显著提高抽样效率。但考虑舍选法的目的就是为了把一个难以抽样的函数换成另一个容易抽样的函数进行抽样，如果用来舍选的函数过于复杂，就违背了使用舍选法的初衷。

5.3 结果的讨论

从结果中可以看出，随着随机数个数的增加，用直接发和舍选发得到的抽样与时间读取的抽样吻合程度逐渐增加。这是因为随着抽样点个数的增加，在中心极限定理的要求下，它们的样本方差逐渐减小，因此拟合程度逐渐增加。

另外，计算舍选法的理论误差为：

$$efficiency = \frac{1}{F(3013)} = \frac{1}{3.41} \approx 0.29326 \tag{8}$$

从舍选法的结果来看，随着抽样点个数的增加，也与这个抽样效率越来越符合。

同时可以证明程序编写无误，用直接抽样法和舍选法得到的分布都是题目中给的数据的分布。