Applied Statistic HW1

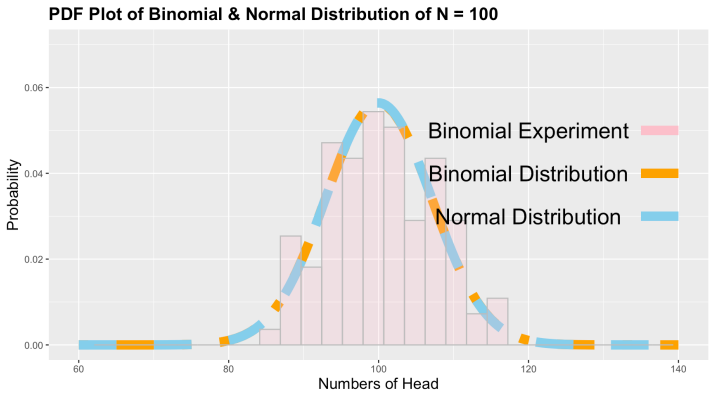
2020270026 王姿文、2020211316 周斯萤、2020211314 徐颢轩

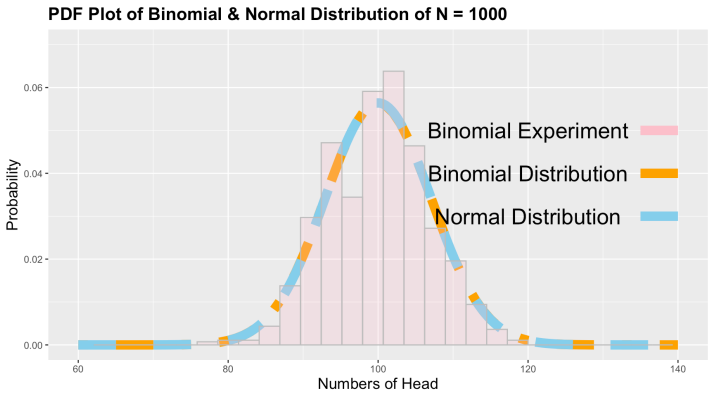
2020/10/02

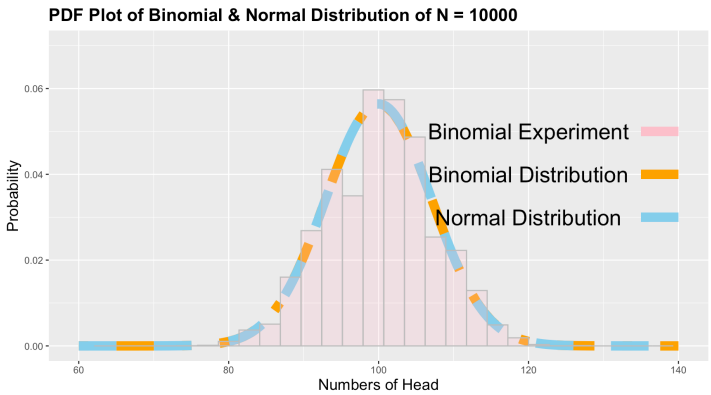
Question 1 模拟将一个公平的硬币独立地抛掷200次的过程，正面记1，反面记0。

1 给出正面个数的近似分布和经验分布，比较两个分布的差别:

数据来自连续200个{0,1}随机抽样，重复n次，计算得到n次实验每次的正面(Head)个数。这是一个重复的独立Bernoulli实验，因此该数据服从$Binomial(200, 0.5)$的二项分布。 此外，根据中央极限定理，在大样本下，若随机变数$\{X_i\}|_{i=0}^{m}$ i.i.d于同样的分布，且$E(X) = \mu$、$Var(X) = \sigma^2 < \infty$，则$\sum_{i=0}^{m} X_i \sim N(m\mu, \sqrt{mVar(\sigma^2)})$。由于 $X_i \sim^{i.i.d} Bernoulli(0.5)$ ，因此 $Binomial(200, 0.5)$ 近似分布 $N(100, 50)$ 。 以下分别取n=100,1000,10000，比较真实分布（粉色频率直方图）、$Binomial(200, 0.5)$的分布（橘色曲线）、$N(100, 50)$的分布（蓝色曲线）。可以明显看出，随着n的数字增大，真实分布（粉色频率直方图）越符合它的近似分佈$N(100, 50)$（蓝色曲线）。

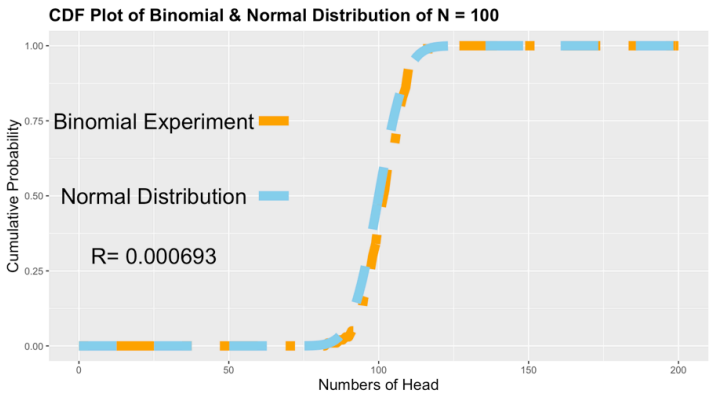


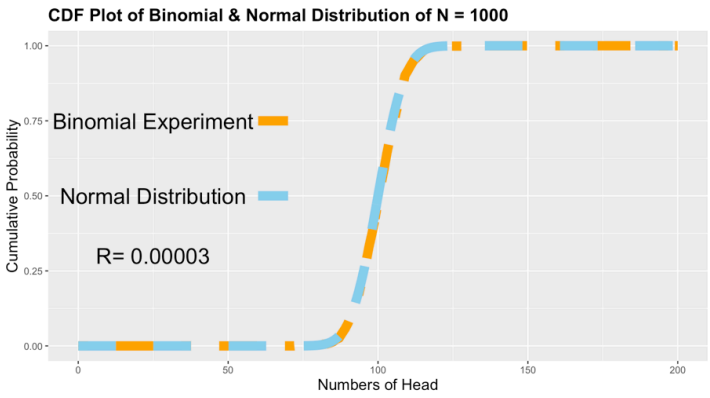


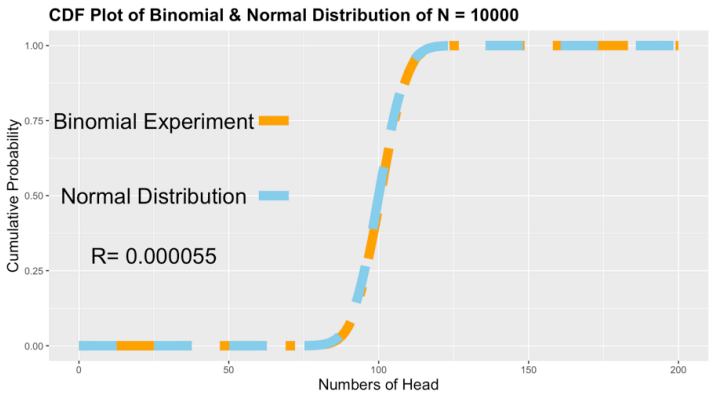


以下分别取n=100,1000,10000，比较真实经验分布（橘色曲线）和近似分布$N(100, 50)$的经验分布（蓝色曲线）。可以明显看出，随着n的数字增大，真实分布（橘色曲线）越符合它的近似分布$N(100, 50)$（蓝色曲线）。

除了统计图的比较，也计算误差R，以数字比较两个分布的拟和度。能够发现随着n增大，误差渐小，可以在图中看出R的值逐渐变小。





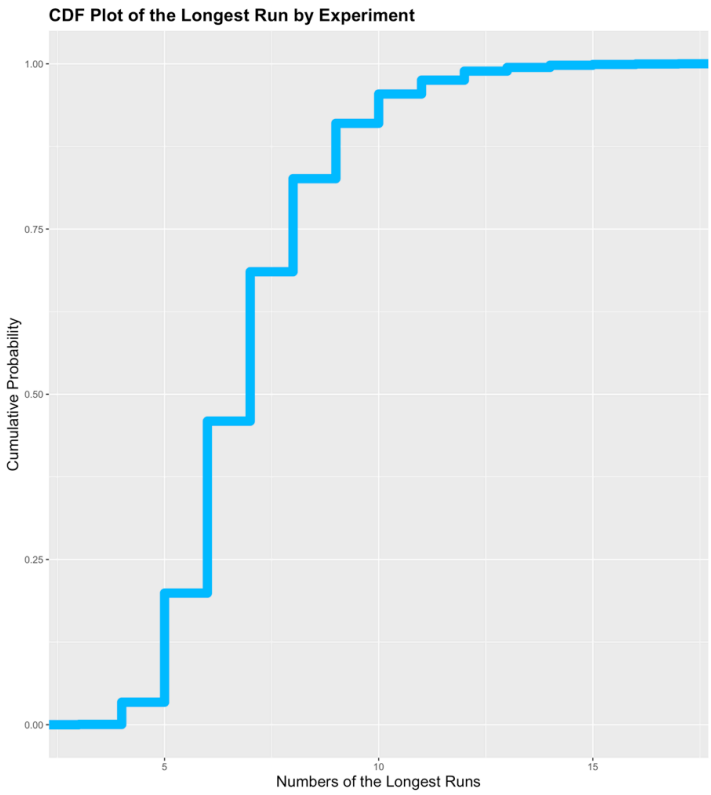


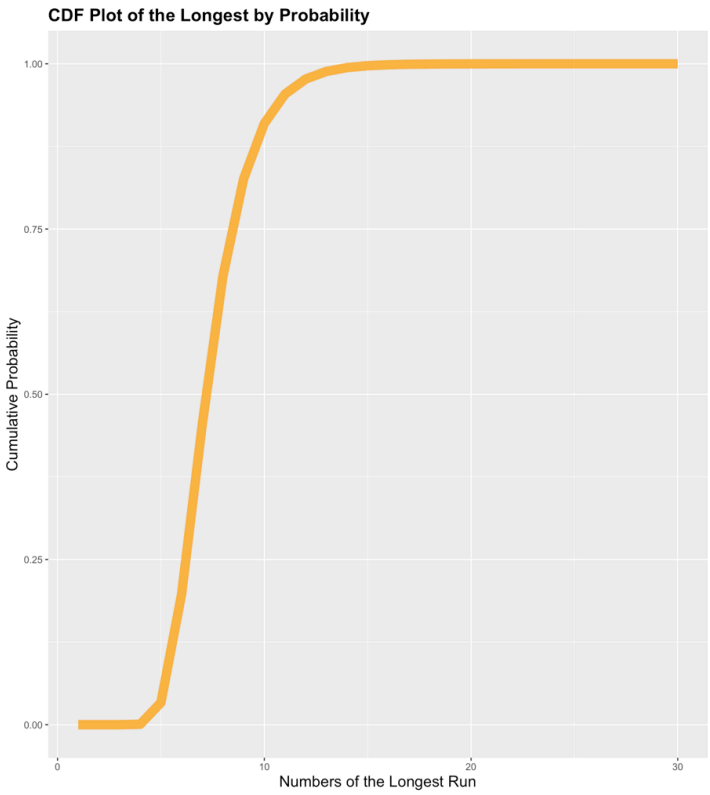
2 给出最长0或1串的长度的经验分布，并尽可能解释所得结果的准确性

我们令 $A_n(x)$ 代表使得最长1串的长度不超过x时总计投掷硬币的次数。显然，所求经验函数 $F_n(x) = 2^{-n}A_n(x)$ 。 我们首先考虑 $x = 3$ 时候的场景。当 $n \leq 3$ 时，由于此时最长1串的长度一定不大于3，因此 $A_n(3) = 2^n$ 。当 $n > 3$ 时，任何一个使得 $x = 3$ 的投掷硬币出来的结果序列一定以0、10、110或者1110开头，然后连接上一个满足最长1串的长度不大于3的序列，因此我们可以得到递推关系式：$A_n(3) = A_{n-1}(3) + A_{n-2}(3) + A_{n-3}(3) + A_{n-4}(3), n \geq 3$ 。同理，对于 x 取任意非负整数值时，

$$f(x) = \begin{cases} \sum_{j=0}^{x} A_{n-1-j}(x) & ,for \quad n > x \\ 2^n & ,for \quad n \leq x \end{cases}$$

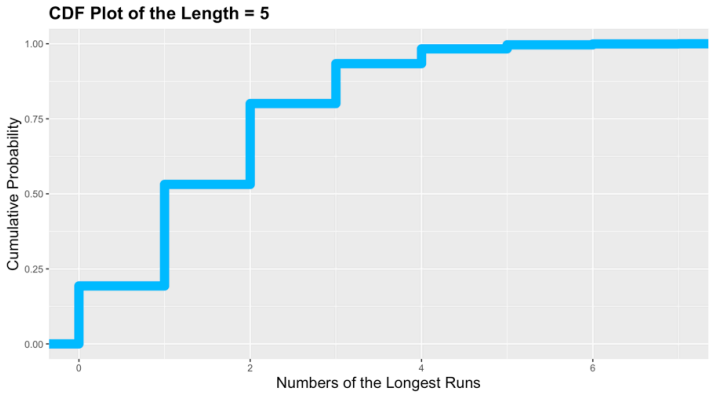
下面左图是我们实验结果的经验分布（取n=10000），右图则是以 $A_n(x)$ 画出的经验分布，可以看出实验结果和理论分布走向一致。

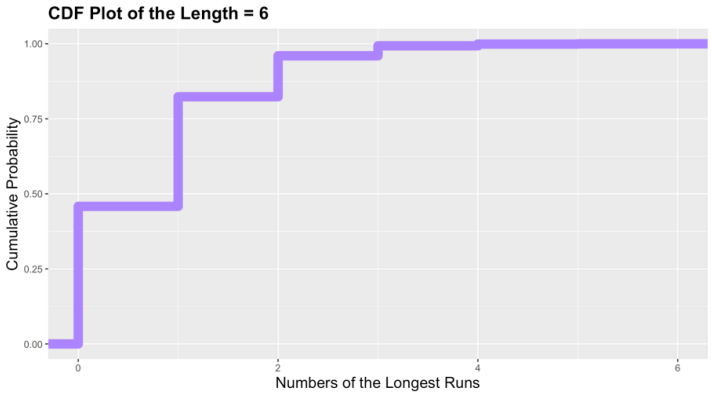


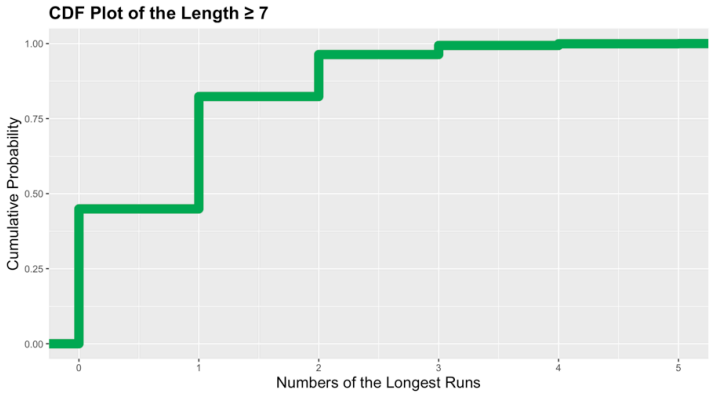


3 分别给出长度为5、6和不小于7的0或1串的个数的经验分布

根据题1.2，可以得出最长0或1串的长度的经验分布，下图因而得出长度分别为5、6和不小于7的0或1串的经验分布。





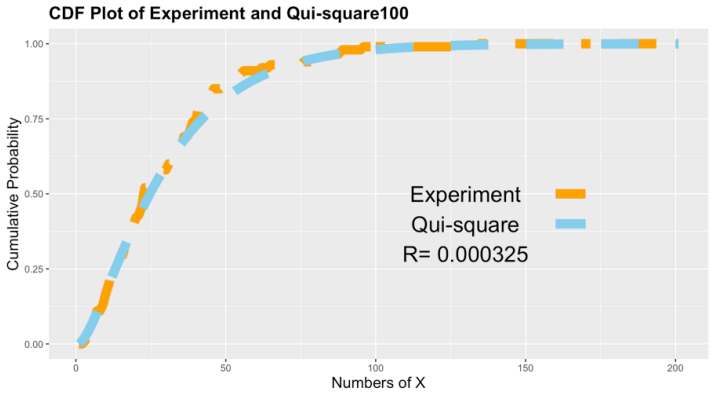


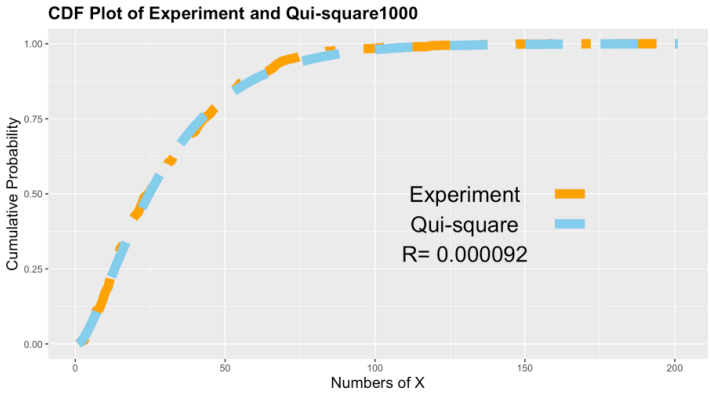
4 连续2个数作为一组，将长度为200的0-1串分割为100个2位二进制数，记这100个数中0-3的个数分别为n_0, n_1, n_2, n_3，$X = \sum_{i=0}^{3} \frac{(n-25)^2}{25}$，试比较$X$的经验分布与自由度为3的$\chi^2$分布的近似程度。

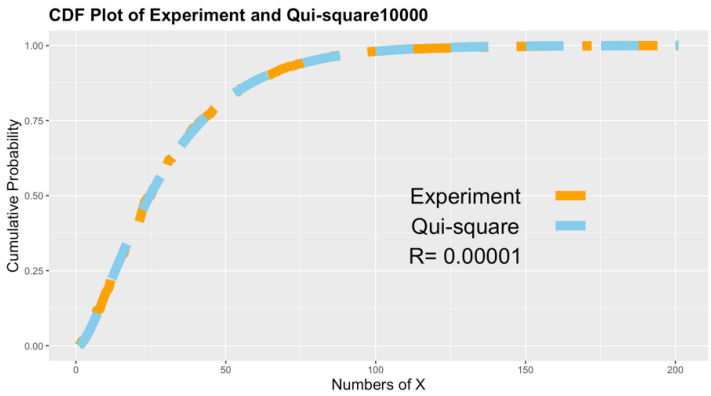
根據統計知識，$X = \sum_{i=0}^{3} \frac{(n-25)^2}{25} \sim \chi^2(3)$，因此分別取n=100,1000,10000，來探討X的经验分布与自由度为3的χ^2分布的近似程度。

下圖為$X = \sum_{i=0}^{3} \frac{(n-25)^2}{25}$的經驗分布（橘色曲线）與它的近似分佈$\chi^2(3)$（蓝色曲线）。可以明显看出，随着n的数字增大，$X = \sum_{i=0}^{3} \frac{(n-25)^2}{25}$的經驗分布（橘色曲线）越符合它的近似分佈$\chi^2(3)$（蓝色曲线）。

除了统计图的比较，也计算误差R，以数字比较两个分布的拟和度。能够发现随着n增大，误差渐小，可以在图中看出R的值逐渐变小。

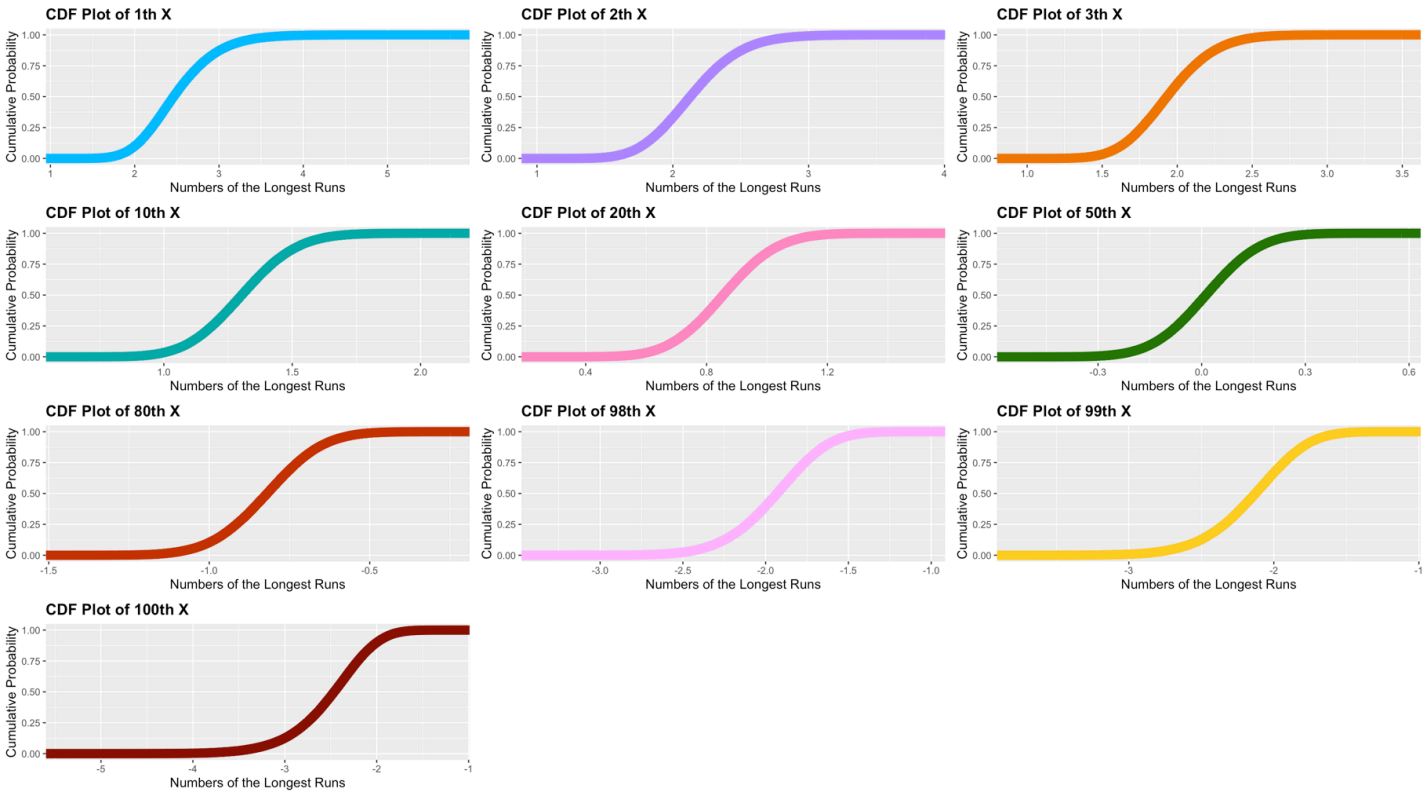






Question 2 $X_1, X_2, \ldots, X_{100}$为一个标准正态分布总体的样本，$X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(100)}$，分别给出$X_{(1)}, X_{(2)}, X_{(3)}, X_{(10)}, X_{(20)}, X_{(50)}, X_{(80)}, X_{(98)}, X_{(99)}, X_{(100)}$等的经验分布，并分别估计它们的期望和方差。

下圖為n=100000的Order Ststistic的經驗分佈圖，可以由圖看出不同$X_{(i)}$的分佈。



下表則為期望值和方差結果。

OrderStatistic	Expectation	Variation
X1	2.5050	0.1845
X2	2.1470	0.0953
X3	1.9456	0.0684
X10	1.3062	0.0297
X20	0.8573	0.0204
X50	0.0124	0.0155
X80	-0.8219	0.0201
X98	-1.9478	0.0686
X99	-2.1502	0.0961
X100	-2.5095	0.1849