Applied Statistic HW4

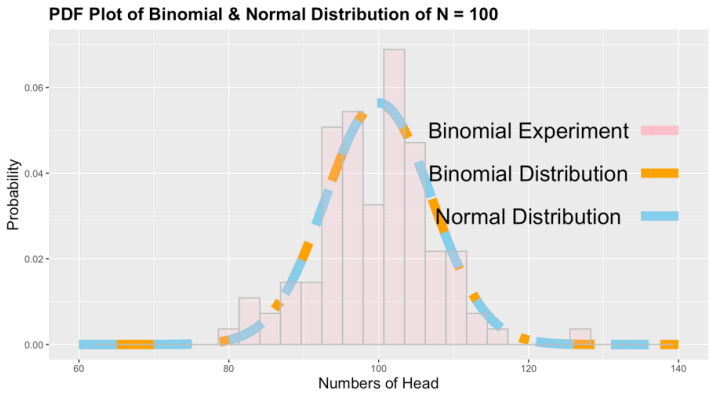
2020270026 王姿文、2020211316 周斯萤、2020211314 徐颢轩

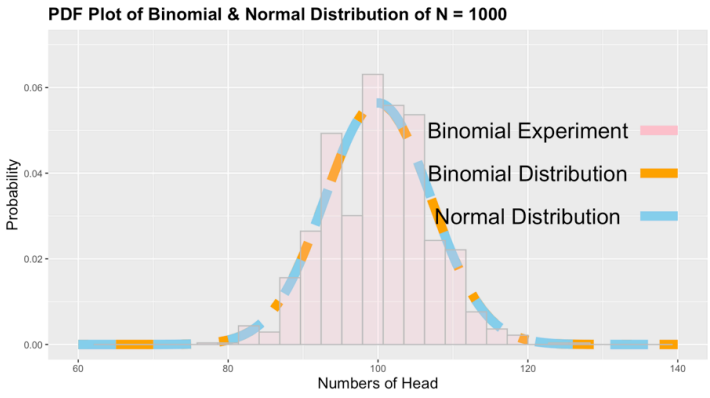
2020/11/24

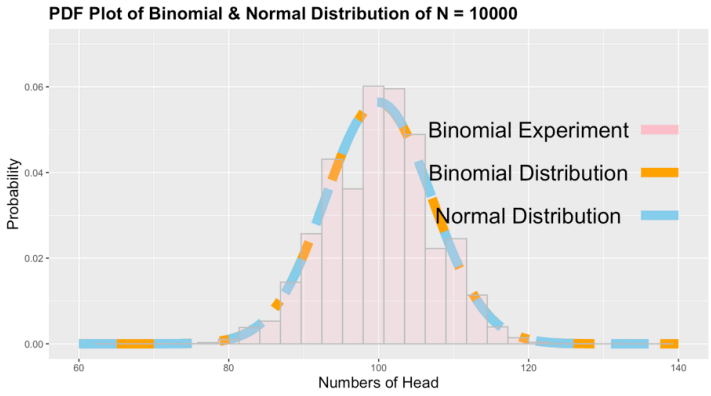
Question 1 定义关于200位0-1序列的统计量；找到公平抛掷(即b(1,0.5))假设下统计量的分布，统计量的分布可能可以解析地算出，也可用直方图等近似。

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）。

解析来看，是直接代入 $Binomial(200, 0.5)$ 的公式，而直方图近似来看则是接近 $N(100, 50)$ 分布







Question 2 提供几个对这个问题不一定是很好的统计量供大家参考

1 正面的个数

可以使用卡方拟合优度检验来判断，令 H_o：分布服从$B(200, 0.5) v.s. H_a$：分布不服从$B(200, 0.5)$， 统计检定量为 $ Y=_{i=0}^{1} $ ，且$Y \sim \mathcal{X}^2(1)$。 A的1个数有93个，0个数有107个；B的1个数有109个，0个数有91个。

$Y_A = \frac{(93-100)^2+(107-100)^2}{200*0.5} = 0.98$ $Y_B = \frac{(91-100)^2+(109-100)^2}{200*0.5} = 1.62$

下为 $\mathcal{X}^2(1)$ 的检定值，可以看出在$\alpha = 0.05$时，两者均不拒绝H_0，均为随机分布，但若$\alpha > 0.1$，则但B拒绝H_0，B为伪随机分布，A为真实随机分布。

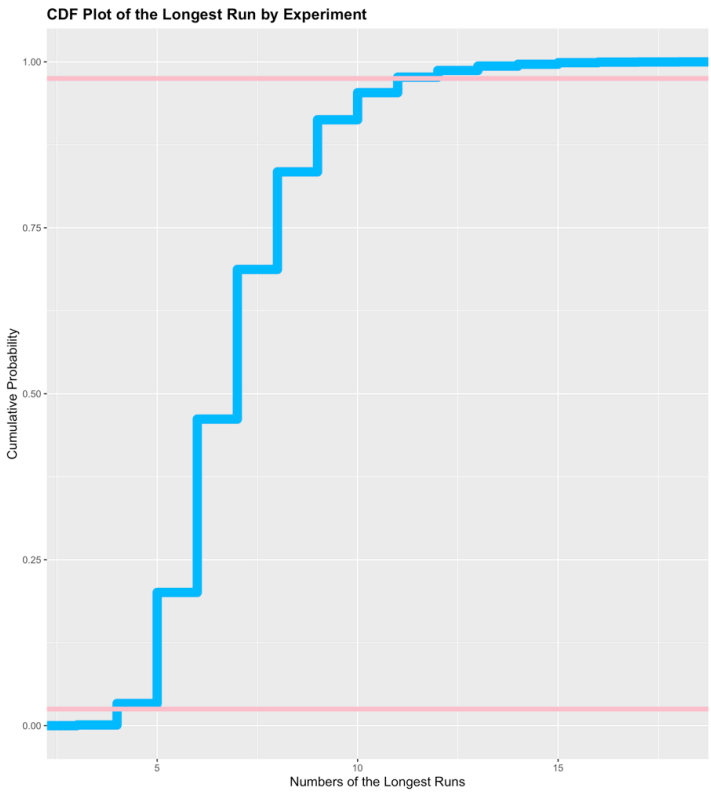
Alpha	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
value	•	•	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879

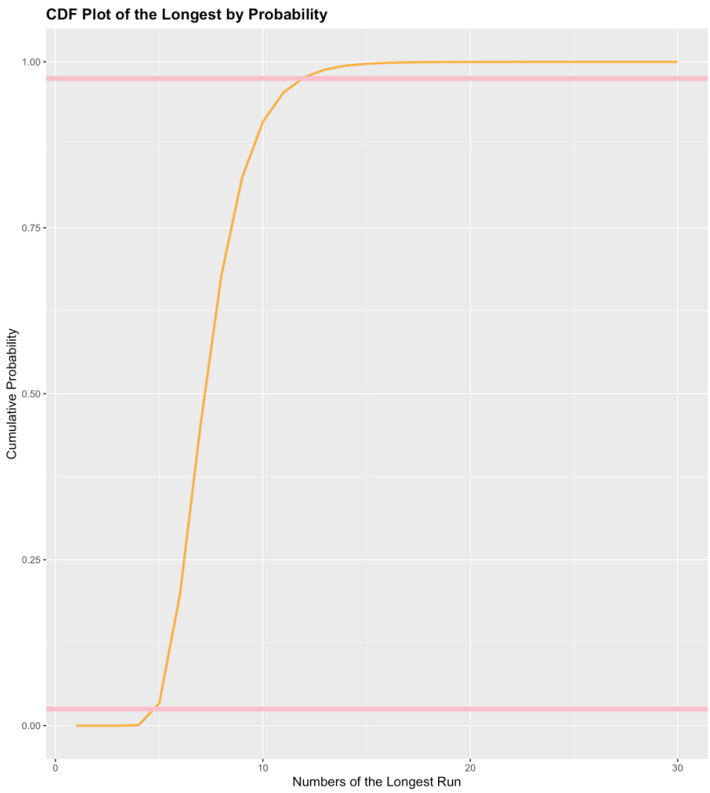
2 最长0或1串的长度

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，而无论是实验结果或理论分布，都能看到在p=0.5时，其对应的最长0或1串的长度约在9左右，而粉色水平线为$\alpha = 0.05$的接受域和拒绝域界限（p=0.025 or p=0.975）。





下表為最長串的次數：

Type	Test_Statistics
理論分佈	8
A	7

下表为理论上最长0或1串的累积分布，可以看出在$\alpha = 0.05$时，接受域为$\mathcal{A} = (5, 12)$，因此A和B都是真实的随机变数，但若在$\alpha = 0.1$时，接受域为$\mathcal{A} = (6, 10)$，则B拒绝H_0，A为真实的随机数，B为伪的随机数。

Max.Length	Prob
1	0.0000
2	0.0000
3	0.0000
4	0.0007
5	0.0335
6	0.1974
7	0.4545
8	0.6789
9	0.8260
10	0.9096
11	0.9541
12	0.9769
13	0.9885
14	0.9942
15	0.9971
16	0.9986
17	0.9993
18	0.9996
19	0.9998
20	0.9999
21	1.0000
22	1.0000

23	1.0000
24	1.0000
25	1.0000
26	1.0000
27	1.0000
28	1.0000
29	1.0000
30	1.0000

3 0-1变化次数，比如01001的切换次数为4， 0-1-00-1

令H_o：分布服从$B(200, 0.5) v.s. H_a$：分布不服从$B(200, 0.5)$，使用**游程檢驗**的方法，構造检验统计量 $\frac{R-\frac{2n_1}{1+c}}{\sqrt{\frac{4cn_1}{(1+c)^2}}} \sim N(0,1)$，R是切換次數，$n_1$ 是0的個數，n_2 是1的個數。 根據下表求得的結果，可以看出，在 $\alpha = 0.05$時，A與B均不落入於拒絕域，因此A和B都是真實的隨機數，但還是能看出B已快落入拒絕域，因此若改變α，例如令$\alpha = 0.17$，則此時B落入拒絕域，A是真的隨机分布，B是伪的随机分布。

Type	理論分佈	A	B
Alpha.0.05的檢定統計量	98	96	110
Alpha.0.05的接受域	(85.0984,113.4816)	(85.71862,114.30138)	(85.44297,113.93703)
Alpha.0.17的檢定統計量	98	96	110
Alpha.0.17的接受域	(90.22792,110.27208)	(90.01083,110.00917)	(89.72138,109.65862)

Question 3 争取提出更多的检验办法进行判断，并给出分析。并且可以将自己的方法用来对计算机模拟生成的0-1串，真正的抛掷结果，以及自己尽力伪造的0-1随机串进行判断和对比。

使用**序列檢验**的方法，对于
$0 < i \leq n,$
$(x_{2i-1}, x_{2i}) = (0,0), (0,1), (1,0), (1,1)$的概率应均为
$\frac{1}{4}$，使用卡方拟合优度检验进行检验。构造检验统计量
$Y = \sum_{i=1}^{4} \frac{((n_i - 25)^2)}{25} \sim \mathcal{X}^2(3)$，
H_o：分布服从$B(200, 0.5) v.s. H_a$：分布不服从$B(200, 0.5)$，在
$\alpha = 0.05$时，拒绝域为
$C = Y \geq 7.8147$，因此A是真的随机分布，B是伪的随机分布。

Type	Test_Statistics

理論分佈	3.36
A	3.60
B	8.56

我们使用Matlab来实现线性同余生成器(LCG)，以此生成随机数。

```
#m=2^31-1;a=65539;
#x=23;n=200;
#for k=1:n
#     x(k+1)=mod(x(k)*a,m);
#     y(k)=x(k+1)/m;
#end;
#z=round(y(1:1:n))
#0,0,0,0,0,1,0,1,1,0,1,1,0,0,1,1,0,0,1,1,1,0,0,0,1,0,1,1,1,0,0,0,1,0,0,1,0,1,1,0,0
,0,0,0,1,0,0,1,0,0,0,1,1,1,0,1,1,1,0 ,0,1,1,0,0,0,1,0,1,1,0,0,0,1,1,1,1,1,0,1,0,0,
1,0,1,1,0,0,0,0,0,1,0,0,0,0,1,1,0,1,0,1,0,1,0,0,1,0,0,1,0,0,0, 0 ,0, 0 ,0 ,0 ,0 ,0
,0, 0, 1, 1,0,0,0,1,1,0,1,0,0,0,0,1,0,1,0,0,1,0,0,1,1,1,0,0,1,1,0,0,0,1,1,1,1,1,0,
0,0,0,0,1,0,1,0,0,1,0,1,0,0,1,1,1,0,1,0,1,0,0,0,1,0,1,1,0,1,0,0,1,1,0,1,0,1,0,0,0,
0
```

由下面三种检验方法，可以得知生成的随机数检验结果为真实随机数，虽然LCG在超平面上存在缺点，但在生成的随机数量少的条件下，LCG的表现还是不错的。

方法	序列检验法	0-1变化次数	最长0或1串的长度
檢定統計量	6.8	100	7
Alpha.0.05的接受域	<=7.8147	(83.65,111.57)	(5,12)