Homework 5, DATA 556: Due Tuesday, 10/31/2018

Alexander Van Roijen

October 28, 2018

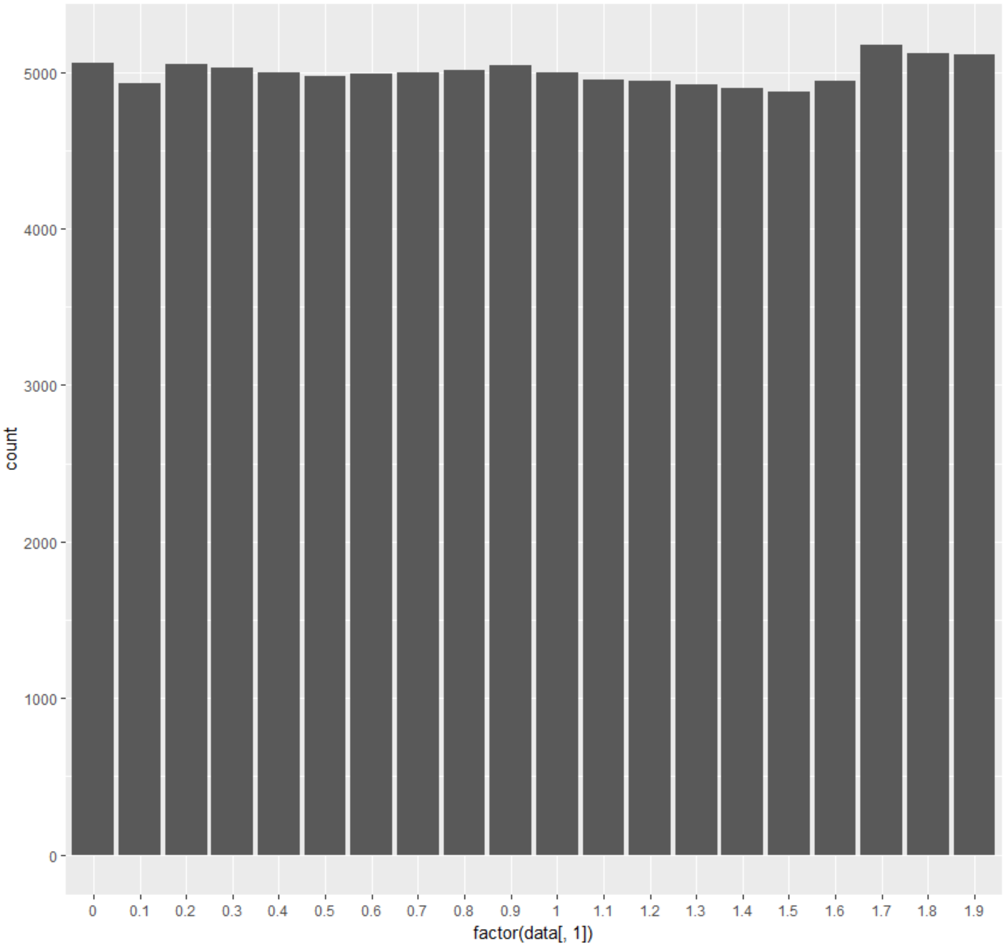
Please complete the following:

1. Problem 1 Let U ~Unif(a, b).

   (a) Use simulations in R (the statistical programming language) to numerically estimate the median and the mode of U for a = 0 and b = 2.

```
> binUnif= function(n,a,b)
+ {
+   resultsog = runif(n,a,b)
+   results = floor(resultsog*10)
+   data = data.frame(results)
+   vals = seq(a,b,.1)
+   p<-ggplot(data=data, aes(x=factor(data[,1]))) +geom_bar(stat="count") + scale_x_
+   p
+   #barplot(results,main="whatev",width=0.5)
+   print(median(resultsog))
+   getmode = table(resultsog)
+   print(which.max(getmode))
+   #print(forMode[c(1:10),])
+ }
> #this is 1a
> set.seed(123)
> n=1000000
> binUnif(n,0,2)
[1] 0.9983065
0.0349205480888486
17519
```

Figure 1: Mode roughly equivalent across all values of x



(b) Find the Median and Mode of $U \sim$ Unif(a,b)

$$\text{Take the result form above and take the derivative} \tag{1}$$

$$f(X|X>a) = F'(X|X>a) = \frac{dF(X|X>a)}{dx} = \frac{F'(x) - F'(a)}{1-F(a)} \text{ with} \frac{dF(a)}{dx} = 0 \tag{2}$$

$$=> f(X|X>a) = \frac{f(x)}{1-F(a)} \tag{3}$$

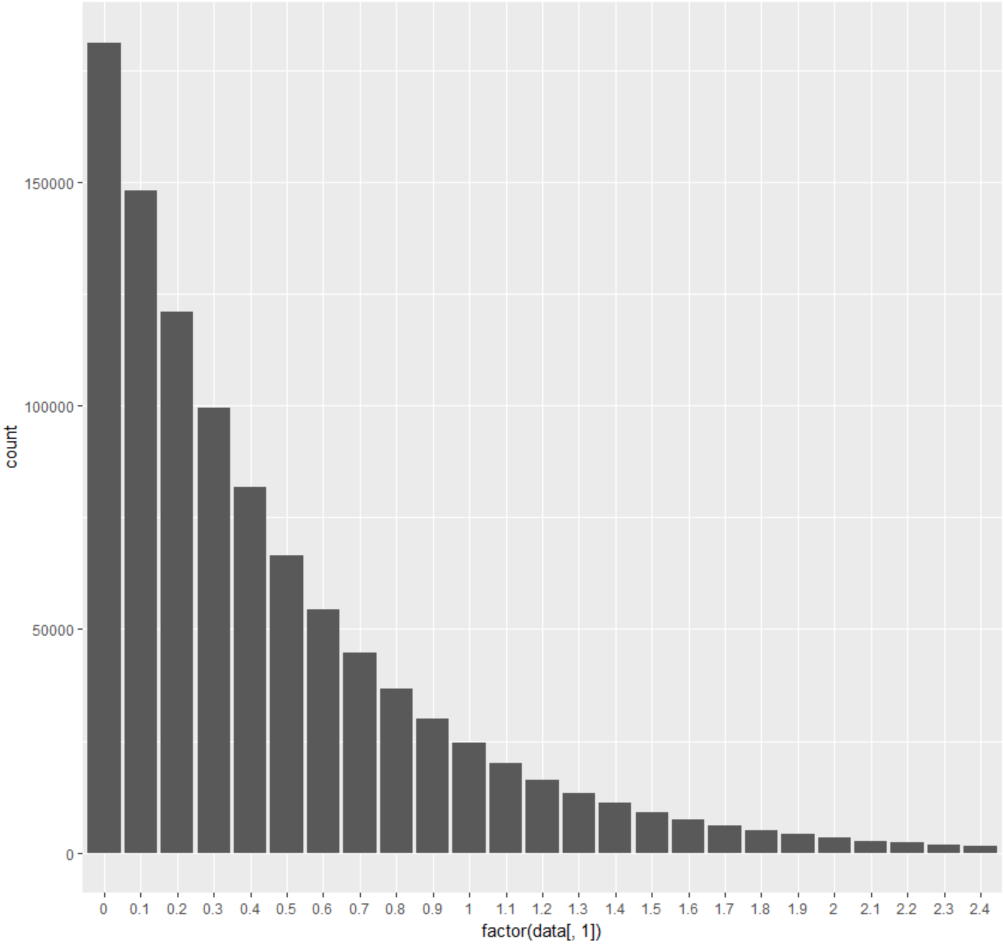
2. Let $X \sim$ Expo(λ)

(a) Use R to simulate median and mode of Expo(2)

```
expoMedMode = function(n,rate)
```

```
+ {
+   resultsog = rexp(n,rate=rate)
+   results = floor(resultsog*10)
+   data = data.frame(results)
+   data = data.frame(data[data[,1]<25,])
+   vals = seq(0,2.5,.1)
+   p<-ggplot(data=data, aes(x=factor(data[,1]))) +geom_bar(stat="count") + scale_x_
+   print(p)
+   #barplot(results,main="whatev",width=0.5)
+   print(median(resultsog))
+   getmode = table(resultsog)
+   print(which.max(getmode))
+ }
> #this is 2a
> set.seed(123)
> n=1000000
> expoMedMode(n,2)
[1] 0.3467215
0.00334986066445708
6662
```

Figure 2: Mode of an exponential with rate of 2 around 0 and median around .3467



This overall makes sense, as we will show below, 0 is the mode of the exponential distribution. This is confirmed as well by the barplot shown above.

(b) Find the Median and Mode of $X \sim \text{Expo}(\lambda)$

3. Let X be Discrete Uniform on 1,2,3,4,5...n .

(a) Use simulations in R to numerically estimate all medians and all modes of X for n = 1,2,3...10.

```
> #this is 3a
> set.seed(433)
> counter = 1
> n=10
> size = 1000
> while(counter <= n)
+ {
+   binUnif(size,1,counter,1,1)
+   counter = counter + 1
+ }
[1] "From 1 to 1"
[1] 1
1
1
[1] "From 1 to 2"
[1] 1.505739
1.0011387350969
1
[1] "From 1 to 3"
[1] 2.025223
1.00209179287776
1
[1] "From 1 to 4"
[1] 2.502714
1.00117637915537
1
[1] "From 1 to 5"
[1] 2.957827
1.00255306344479
1
[1] "From 1 to 6"
[1] 3.368198
1.00341863720678
1
[1] "From 1 to 7"
```

```
[1] 4.004081

1.0143808578141

1

[1] "From 1 to 8"

[1] 4.470458

1.00075664301403

1

[1] "From 1 to 9"

[1] 5.058819

1.0038467105478

1

[1] "From 1 to 10"

[1] 5.422321

1.01815693522803

1
```

Figure 3: Mode across demonstrated from Unif(1,1)

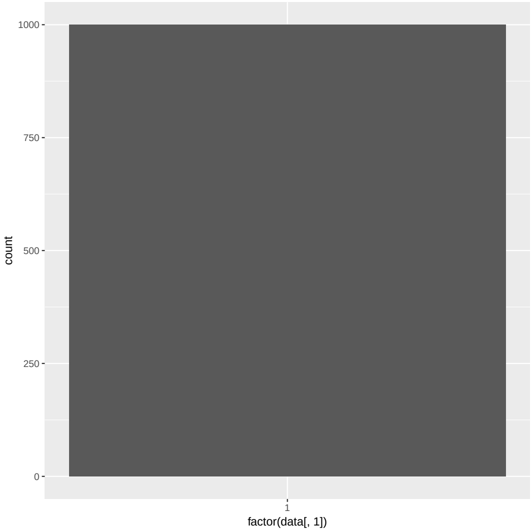


Figure 4: Mode across demonstrated from Unif(1,1)

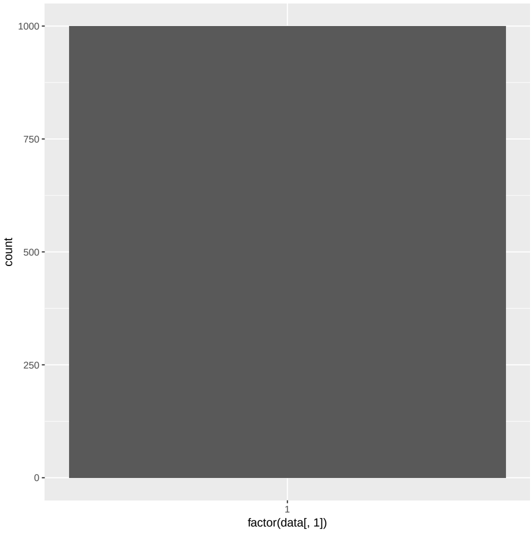


Figure 5: Mode across demonstrated from Unif(1,1)

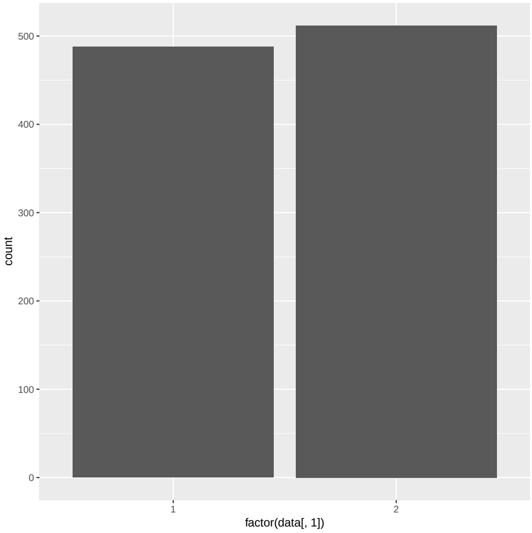


Figure 6: Mode across demonstrated from Unif(1,1)

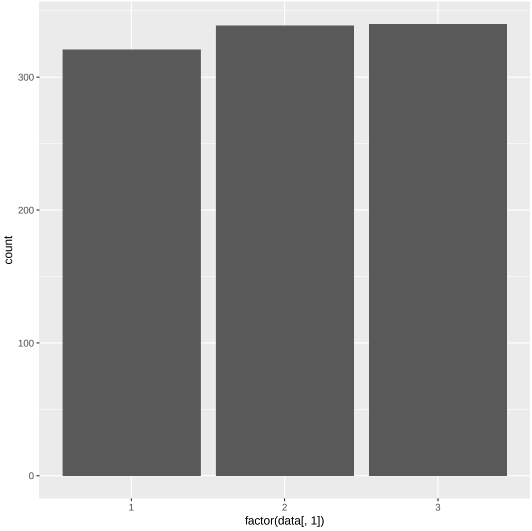


Figure 7: Mode across demonstrated from Unif(1,1)

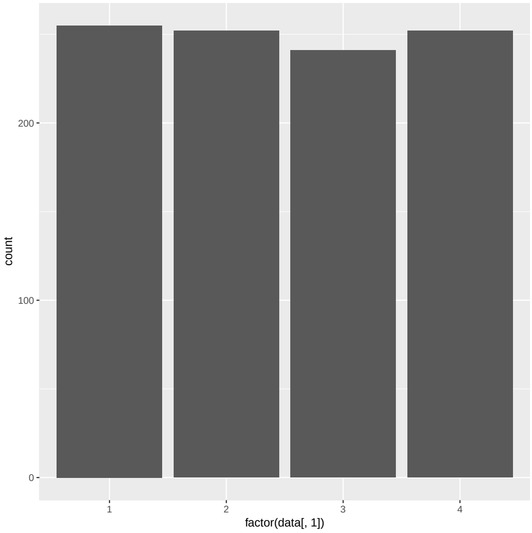


Figure 8: Mode across demonstrated from Unif(1,1)

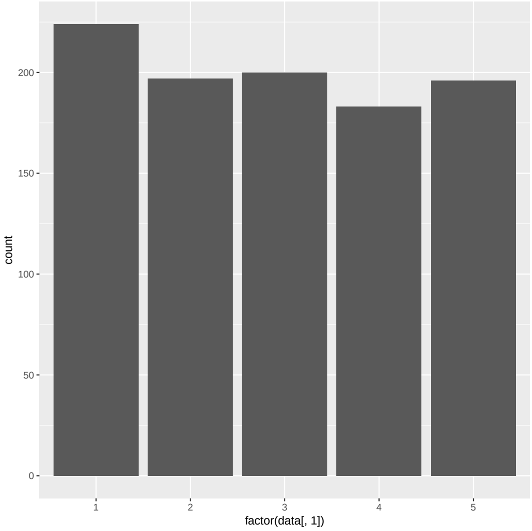


Figure 9: Mode across demonstrated from Unif(1,1)

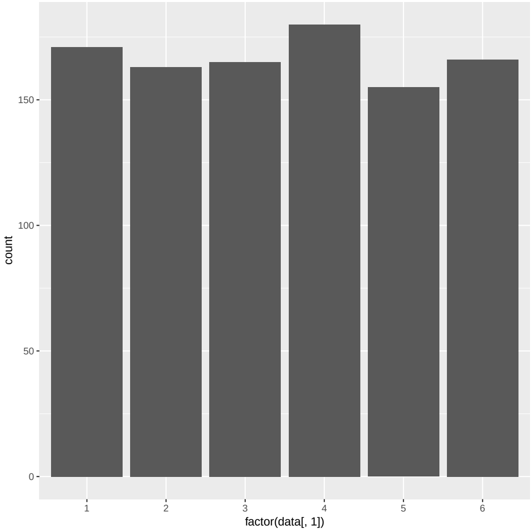


Figure 10: Mode across demonstrated from Unif(1,1)

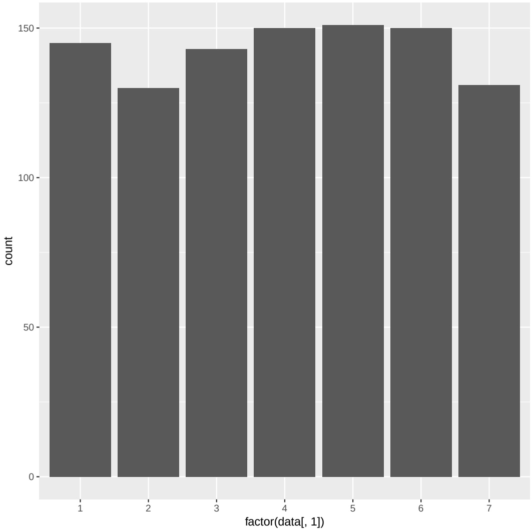


Figure 11: Mode across demonstrated from Unif(1,1)

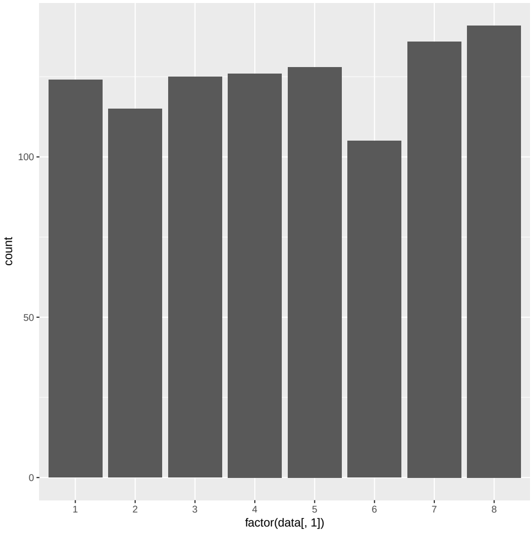
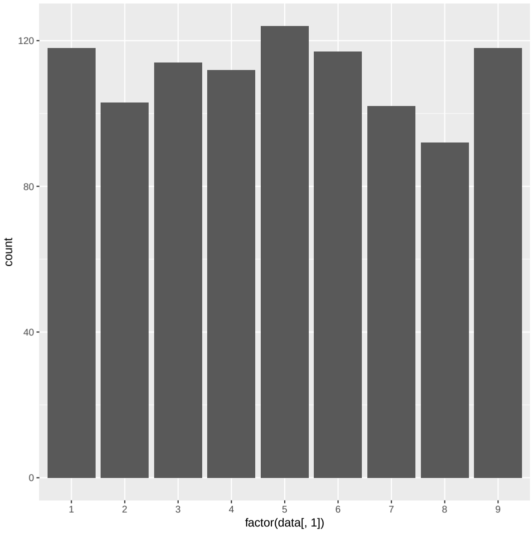


Figure 12: Mode across demonstrated from Unif(1,1)



(b) Find All medians and modes of X

Note the mode is trivial, as it is a similar case to 1.

$$\text{Want } P(X=c) \geq P(X=x) \forall x \in 1,2,...n \tag{4}$$

$$=> \frac{1}{n-1+1} \geq \frac{1}{n-1+1} \text{ by def of discrete Uniform} \tag{5}$$

$$=> c = x \forall x \in 1,2,3,...n \text{ which is almost the same as 1} \tag{6}$$

$$\text{Notable difference is that in the discrete case, c can only take discrete values} \tag{7}$$

$$\text{Now the median} \tag{8}$$

$$\text{Want } P(X \leq x) \geq \frac{1}{2} \& P(X \geq x) \geq \frac{1}{2} \tag{9}$$

$$P(X \leq x) = \sum_{i=1}^{x} \frac{1}{n} = \frac{x}{n} \geq \frac{1}{2} => x \geq \frac{n}{2} \tag{10}$$

$$\text{However, we can observe the patterns noted in the graphs below} \tag{11}$$

$$\text{If n is odd, the previous conclusion is the only solution as} \tag{12}$$

$$\text{if you go above or below } \frac{n}{2} \text{ you lose the probability @ } x = \frac{n}{2} \tag{13}$$

$$\text{This is true as we have a jump exactly at } \frac{n}{2} \tag{14}$$

$$\text{In the case n is even, you have some wiggle room} \tag{15}$$

$$\text{There is no jump at } \frac{n}{2} \tag{16}$$

$$\text{In fact, the next jump is at } \frac{n}{2} + 1 \tag{17}$$

$$\text{This means we have medians from } [\frac{n}{2}, \frac{n}{2} + 1] \tag{18}$$

$$\text{One last important factor to note, is that this result relies heavily on} \tag{19}$$

$$\text{how we define the median and our environment} \tag{20}$$

$$\text{In general we have} \tag{21}$$

$$Median(X) = \begin{cases} \frac{n}{2} & n = \text{odd} \\ [\frac{n}{2}, \frac{n}{2} + 1] & n = \text{even} \end{cases} \tag{22}$$

Figure 13: (n=5)Note the jumps taking place on the discrete values

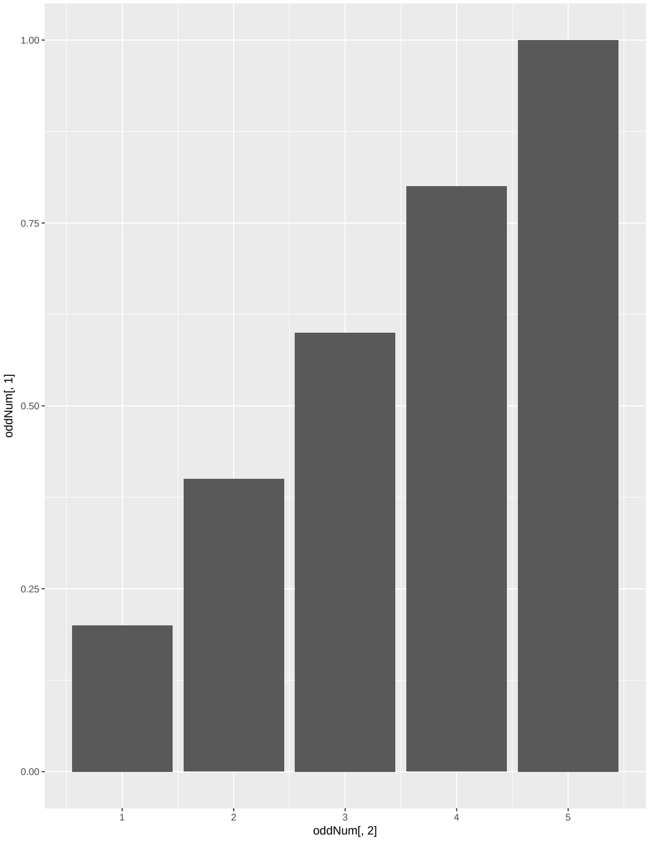
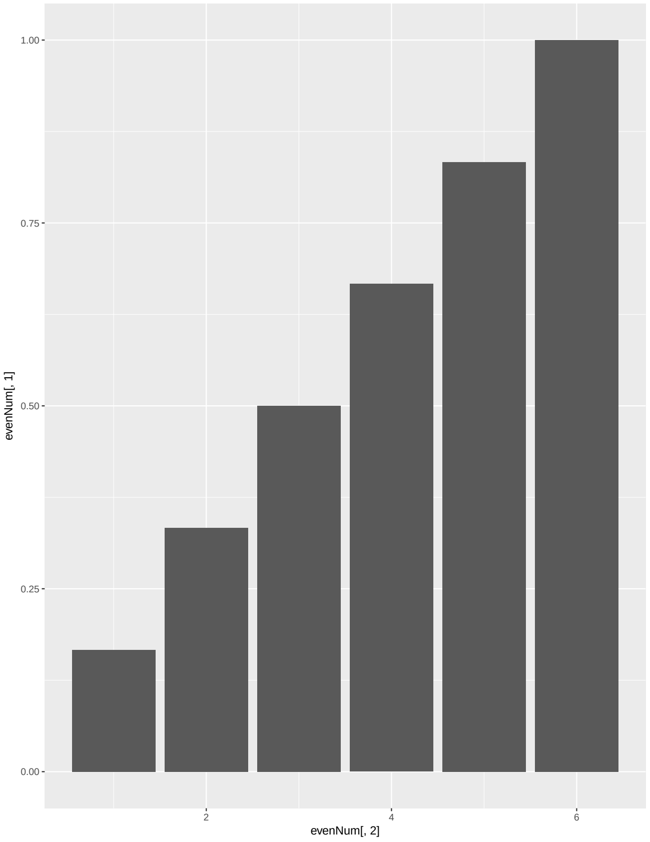


Figure 14: (n=6)These jump patterns hold for all discrete n



4. Let $U_1, ..., U_n$ be i.i.d. $Unif(0, 1)$, and $X = max(U_1, ..., U_n)$.

   (a) What is the PDF of X?

$$\text{CDF} = P(X \leq x) = P(U_1 \leq x, ..., U_n \leq x) = P(U_1 \leq x) * ... * P(U_n \leq x) \text{ since i.i.d} \tag{23}$$

$$=> \text{CDF} = x * x * x ... * x = x^n => \text{ PDF } = \frac{dF}{dx} = n * x^{n-1} \text{ with } 0 \leq x \leq 1 \tag{24}$$

   (b) what is the $E[X]$

$$\int_{-\infty}^{\infty} x * f(x) = \int_0^1 x * f(x) = \int_0^1 x * n * x^{n-1} = \int_0^1 n * x^n = \frac{n}{n+1} * (1^n) - 0 = \frac{n}{n+1} \tag{25}$$

   (c) R simulation results / approx

```
> #4c
> set.seed(123)

> n=10

> runplenty=10000

> og = runif(n,0,1)

> counter=1

> result= numeric(0)

> while(counter<=runplenty)

+ {

+   og = runif(n,0,1)

+   result[counter]=max(og)

+   counter= counter+1

+ }

> print(mean(result))

[1] 0.9091953

> print(n/(n+1))

[1] 0.9090909
```

5. (a) Find $P(X < Y)$ for X $\sim N(a, b), Y \sim N(c, d)$ with X and Y independent

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + Var(Y) = b^2 + d^2 \tag{26}$$

$$E[X - Y] = E[X] - E[Y] = a - c \tag{27}$$

$$=> P(X < Y) = P(X - Y < 0) \sim N(a - b, c^2 + d^2) \text{ determined via hint} \tag{28}$$

$$= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} \text{ Where } \sigma = \sqrt{c^2 + d^2} \text{ and } \mu = a - b \tag{29}$$

(b) R simulation

```
#5b

> set.seed(123)

> n= 1000000

> resultsx = rnorm(n,0,1)

> resultsy = rnorm(n,1,5)

> trueResults = rnorm(n,-1,sqrt(26))

> counter = 1
> numCorr=0
> numCorr2=0
> results=numeric(0)
> while(counter<=n)
+ {
+   if(resultsx[counter]<resultsy[counter])
+   {
+     numCorr = numCorr + 1
+   }
+   if(trueResults[counter]<0)
+   {
+     numCorr2 = numCorr2 + 1
+   }
+   counter= counter+1
+ }
> print(numCorr/n)
[1] 0.576969
> print(numCorr2/n)
[1] 0.577764
```

6. The heights of men in the United States are normally distributed with mean 69.1 inches and standard deviation 2.9 inches. The heights of women are normally distributed with mean 63.7 inches and standard deviation 2.7 inches. Let x be the average height of 100 randomly sampled men, and y be the average height of 100 randomly sampled women.

(a) What is the distribution of x - y?

Similar to 5, except now we need to calculate the expected value and variance over the 100 samples

(30)

$$E[x-y] = E[\frac{1}{n}\sum_{n=1}^{100} x_n - y_n] = \frac{1}{n}\sum_{n=1}^{100} E[X_n] - E[y_n] = \frac{1}{n}\sum_{n=1}^{100}(69.1-63.7) = \frac{1}{n}n*(69.1-63.7)$$

(31)

$$\text{we got this from the distributions of the random variables given} \tag{32}$$

$$=> E[x-y] = 5.4 \tag{33}$$

$$\text{Var}[x-y] = \text{Var}[\frac{1}{n}\sum_{n=1}^{100} x_n - y_n] = \frac{1}{n^2}\text{Var}[\sum_{n=1}^{100} x_n - y_n] \text{ from independence we get}$$

(34)

$$= \frac{1}{n^2}\sum_{n=1}^{100}(\text{Var}[x_n] + \text{Var}[y_n]) = \frac{1}{n^2}\sum_{n=1}^{100}(2.9^2+2.7^2) = \frac{1}{n^2}n(15.7) = \frac{1}{100}(15.7) \tag{35}$$

$$=> \text{Var}[x-y] = \sqrt{\frac{15.7}{100}}^2 \tag{36}$$

$$=> x-y \sim N(5.4, \sqrt{\frac{15.7}{100}}) \tag{37}$$

(b) R monte carlo simulations

```
> set.seed(123)

> n=100

> numTrials=100000

> counter=1

> diffRes = numeric(0)

> while(counter<numTrials)

+ {

+   resultsx = rnorm(n,69.1,2.9)

+   resultsy = rnorm(n,63.7,2.7)

+   diffRes[counter] = mean(resultsx)-mean(resultsy)

+   counter = counter + 1

+ }

> resultCalc = rnorm(n,5.4,sqrt(15.7/n))
> print(mean(diffRes))
[1] 5.400457
> print(mean(resultCalc))
[1] 5.381992
> print(var(diffRes))
[1] 0.1570096
> print(var(resultCalc))
[1] 0.1590042
```

Clearly, we can see that these make sense intuitively.

(c) What is the probability that a man is taller than a randomly sampled woman?

let X be the RV for a man sampled and Y be a RV for a woman sampled

$$\text{Note, that the } P(X-Y<0) = P(X<Y) => P(X>Y) = 1-P(X<Y) = 1-P(X-Y<0) \tag{38}$$

$$\text{we can assume this as the } P(X=Y)=0 \tag{39}$$

$$\text{We know } X-Y \sim N(5.4, \sqrt{15.7}) \tag{40}$$

$$=> 1-P(X-Y<0) = 1-0.08646693 = .9135331 \tag{41}$$

Note, the value above was derived using dbinom from R

7. Suppose we have a RV Y such that $Y \sim \text{Binom}(n=5, p=\theta)$

(a) Using Bayes Rule to determine $P(\theta|y)$ in terms of θ_i

$$P(\theta|y) = \frac{P(Y|\theta_i)*P(\theta_i)}{P(Y)} = \frac{\binom{5}{y}\theta_i^y(1-\theta_i)^{5-y}*\frac{1}{11}}{P(Y)} \tag{42}$$

$$\text{Now with the law of total probability, we get} \tag{43}$$

$$= \frac{\binom{5}{y}\theta_i^y(1-\theta_i)^{5-y}*\frac{1}{11}}{\sum_{i=0}^{11}\binom{5}{y}\theta_i^y(1-\theta_i)^{5-y}*\frac{1}{11}} = \frac{\frac{1}{11}\binom{5}{y}\theta_i^y(1-\theta_i)^{5-y}}{\binom{5}{y}\frac{1}{11}\sum_{i=0}^{11}\theta_i^y(1-\theta_i)^{5-y}} = \frac{\theta_i^y(1-\theta_i)^{5-y}}{\sum_{i=0}^{11}\theta_i^y(1-\theta_i)^{5-y}} \tag{44}$$

(b) & (c) The previous graphics make complete sense. In particular, notice that the end points are the edge cases of p = 0 and p = 1. In either case, you either need there to be all failures, or all successes, and is thus 0 probability or 1 depending on the y. Further, its interesting to note the symmetric nature of this distribution due to the binomial nature and θ and $1-\theta$ relationship

8. a) Starting from independent uniform random variables (U $\sim$ Unif(0, 1)), devise an algorithm to generate independent samples from a Logistic distribution, having density

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2} => F(x) = (1+e^{-x})^{-1} \tag{45}$$

$$=> \text{let } F(x) = u \text{ we want to solve for x in terms of you to find } F^{-1} \tag{46}$$

$$\frac{1}{1+e^{-x}} = u => \frac{1}{u} = 1+e^{-x} => \frac{1-u}{u} = e^{-x} => \log(\frac{1-u}{u}) = -x => F^{-1}(X) = \log(\frac{u}{1-u}) \tag{47}$$

b) R simulations

```
> set.seed(123)
> n=100000
> unifVals = runif(n,0,1)
> invert = log(unifVals/(1-unifVals))
> res = (invert<3 & invert>2)
> print(sum(res == TRUE))
[1] 7077
> print(sum(res==TRUE)/n)
[1] 0.07077
```