Homework 4, DATA 556: Due Tuesday, 10/23/2018

Alexander Van Roijen

October 25, 2018

Please complete the following:

1. Problem 1 Let X be a continuous random variable with CDF F and PDF f.

   (a) Find the conditional CDF of X given $X > a$ (where a is a constant with $P(X > a) \neq 0$). That is, find $P(X \leq x | X > a)$ for all a, in terms of F.

$$P(X \leq x | X > a) = \frac{P(X \leq x, X > a)}{P(X > a)} = \frac{P(X \leq x, X > a)}{1 - P(X < a)} = \frac{P(X \leq x, X > a)}{1 - F(a)} \tag{1}$$

$$P(X \leq x, X > a) = P(X < x) - P(X < a) = F(x) - F(a) \tag{2}$$

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

   (b) Find the conditional PDF of X given $X > a$.

$$\text{Take the result form above and take the derivative} \tag{4}$$

$$f(X|X > a) = F'(X|X > a) = \frac{dF(X|X > a)}{dx} = \frac{F'(x) - F'(a)}{1 - F(a)} \text{ with} \frac{dF(a)}{dx} = 0 \tag{5}$$

$$=> f(X|X > a) = \frac{f(x)}{1 - F(a)} \tag{6}$$

   (c) Check that the conditional PDF from (b) is a valid PDF, by showing directly that it

is non negative and integrates to 1.

$$f(X|X > a) \geq 0 \forall x \text{ as } f(X|X > a) = \frac{f(x)}{1 - F(a)} \tag{7}$$

$$\text{and we know that } f(x) \geq 0 \forall x \text{ as well as that } 1 - F(a) \text{ is a const} \tag{8}$$

$$\text{Thusly, } f(X|X > a) \geq 0 \forall x \text{ according to these facts} \tag{9}$$

$$\tag{10}$$

$$\text{Now we show } \int_{-\infty}^{\infty} f(X|X > a)dx = 1 \text{ using the fact that } \int_{-\infty}^{\infty} f(x)dx == 1 \tag{11}$$

$$\int_{-\infty}^{\infty} f(X|X > a)dx = \int_{-\infty}^{a} f(X|X > a)dx + \int_{a}^{\infty} f(X|X > a)dx = \int_{a}^{\infty} f(X|X > a)dx \tag{12}$$

$$\text{the above is true as we know } f(X|X > a) \text{ has no density for x below a} \tag{13}$$

$$= \int_{a}^{\infty} \frac{f(x)}{1 - F(a)}dx = \frac{1}{1 - F(a)} \int_{a}^{\infty} f(x)dx \text{ and since we know F is a valid cdf we get} \tag{14}$$

$$= \frac{1}{1 - F(a)}[1 - F(a)] = 1\square \tag{15}$$

2. Problem 2: A circle with a random radius $R \sim Unif(0, 1)$ is generated. Let A be its area.

   (a) Use R to simulate mean and variance of A

```
> set.seed(123)
> #this is 2a
> n=1000000
> results = runif(n,0,1)
> counter = 1
> areas=numeric(0)
> while(counter<n)
+ {
+   areas[counter] = results[counter] * results[counter] * pi
+   counter= counter + 1
+ }
```

```
> print(mean(areas))
[1] 1.045967
> print(var(areas))
[1] 0.8775327
```

(b) Find the theoretical mean and the variance of A, without first finding the CDF or PDF of A. Compare with your numerical results from (a).

$$E[A] = \pi E[r^2] \tag{16}$$

$$\text{we know r folows Unif(0,1)} => E[r^2] = Var(r) + E[r]^2 = \frac{1}{12} + \frac{1}{2}^2 = \frac{1}{3} \tag{17}$$

$$=> E[A] = \frac{\pi}{3} = 1.047198 \tag{18}$$

$$\text{Var}[A] = \text{Var}[\pi * r^2] = \pi^2 \text{Var}[r^2] = \pi^2 * (E[r^4] - E[r^2]^2) = \pi^2(\int_0^1 r^4 dr - (\frac{1}{3})^2) \tag{19}$$

$$= \pi^2(\frac{1}{5} - \frac{1}{9}) = 0.87729816898 \tag{20}$$

(c) Find the CDF and PDF of A.

$$F(x) = P(A \leq x) = P(\pi R^2 \leq x) = P(R \leq \frac{\sqrt{x}}{\sqrt{\pi}}) = F(x) = \frac{\sqrt{x}}{\sqrt{\pi}} \tag{21}$$

$$f(x) = F'(x) = \frac{1}{\pi} * \frac{d\sqrt{x}}{\sqrt{dx}} = \frac{1}{2\sqrt{x\pi}} \tag{22}$$

$$\text{with } 0 \leq x \leq \pi \tag{23}$$

3. A stick of length 1 is broken at a uniformly random point, yielding two pieces. Let X and Y be the lengths of the shorter and longer pieces, respectively, and let $R = \frac{X}{Y}$ be the ratio of the lengths of X and Y.

(a) Use simulations in R (the statistical programming language) to gain some understanding about the distribution of the random variable R. Numerically estimate the expected value of R and $1/R$.

```
> set.seed(123)
> n= 1000
> results = runif(n,0,1)
> counter = 1
```

```
> xs = numeric(0)
> ys = numeric(0)
> while(counter <= n)
+ {
+   if(results[counter]>=0.5)
+   {
+     ys[counter]=results[counter]
+     xs[counter] = 1-ys[counter]
+   }
+   else
+   {
+     xs[counter] = results[counter]
+     ys[counter] = 1 -xs[counter]
+   }
+   counter = counter + 1
+ }
> rs = xs/ys
> print(mean(rs))
[1] 0.3877773
> print(mean(1/rs))
[1] 15.82287
```

(b) Find the CDF and PDF of R.

We note that since x is exclusively smaller than y of a unit length stick, then $0 \leq X \leq 0.5$ and $0.5 \leq Y \leq 1$ and $X = 1 - Y$ and $Y = 1 - X$

and that X = min(U,1-U) and Y = max(U,1-U)

$$P(R \leq r) = P(\frac{X}{Y} \leq r) = P(\frac{X}{(1-X)} \leq r) = P(X \leq r * (1-X)) = P(X + rX \leq r) \tag{24}$$

$$= P(X(1+r) \leq r) = P(X \leq \frac{r}{1+r}) = P(\min(U \text{ or } 1-U) \leq \frac{r}{1+r}) \tag{25}$$

$$= P(U \leq \frac{r}{1+r}) + P((1-U) \leq \frac{r}{1+r}) + P(U \cap (1-U) \leq \frac{r}{1+r}) \text{ by def of OR/union} \tag{26}$$

$$= P(U \leq \frac{r}{1+r}) + P((1-U) \leq \frac{r}{1+r}) + 0 \tag{27}$$

the above is true as in the intersect case, both U values cant be realized simultaneously

(28)

(except for when U = .5, but this is a single point in a continuous distribution)

(29)

$$= P(U \leq \frac{r}{1+r}) + P(U \geq 1 - \frac{r}{1+r}) = P(U \leq \frac{r}{1+r}) + 1 - P(U \leq 1 - \frac{r}{1+r}) \tag{30}$$

$$= \frac{r}{1+r} + 1 - (1 - \frac{r}{1+r}) = \frac{r}{1+r} + 1 - \frac{1}{1+r} \tag{31}$$

$$=> F(r) = \frac{2r}{1+r} \tag{32}$$

$$=> f(r) = (1+r)^{-1}\frac{d}{dr}(2r) + 2r\frac{d}{dr}(1+r)^{-1} = (1+r)^{-1}(2) + 2r(-1)(1+r)^2 \tag{33}$$

$$= \frac{2+2r-2r}{(1+r)^2} = f(r) = \frac{2}{(1+r)^2} \tag{34}$$

(c) Find the expected value of R (if it exists).

$$E[R] = \int_{-\infty}^{\infty} r * f(r)dr = \int_{-\infty}^{\infty} \frac{2r}{(1+r)^2}dr = \int_0^1 \frac{2r}{(1+r)^2}dr = 2(\frac{1}{1+r} + \log(1+r))\Big|_0^1 \tag{35}$$

$$= 2((\frac{1}{2} + \log(2)) - (\frac{1}{1} + \log(1))) = 2((\frac{1}{2} + \log(2)) - 1) = 2\log(2) - 1 = 0.3862944 \tag{36}$$

(d) Find the expected value of 1/R if it exits.

$$E[\frac{1}{R}] = \int_{-\infty}^{\infty} \frac{1}{r} * f(r)dr = \int_{-\infty}^{\infty} \frac{2}{r(1+r)^2}dr = \int_{0}^{1} \frac{2}{r(1+r)^2}dr = 2 * (\frac{1}{r+1} + \log(r) - log(1+r)\Big|_0^1) \tag{37}$$

However, this can not be evaluated as the log of zero is undefined (38)

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?

$$\text{CDF} = P(X \leq x) = P(U_1 \leq x, ..., U_n \leq x) = P(U_1 \leq x) * ... * P(U_n \leq x) \text{ since i.i.d} \tag{39}$$

$$=> \text{CDF} = x * x * x ... * x = x^n => \text{ PDF } = \frac{dF}{dx} = n * x^{n-1} \text{ with } 0 \leq x \leq 1 \tag{40}$$

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

$$\int_{-\infty}^{\infty} x * f(x) = \int_{0}^{1} x * f(x) = \int_{0}^{1} x * n * x^{n-1} = \int_{0}^{1} n * x^n = \frac{n}{n+1} * (1^n) - 0 = \frac{n}{n+1} \tag{41}$$

(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

$$\text{Var}(X - Y) = \text{Var}(X) + Var(Y) = b^2 + d^2 \tag{42}$$

$$E[X - Y] = E[X] - E[Y] = a - c \tag{43}$$

$$=> P(X < Y) = P(X - Y < 0) \sim N(a - b, c^2 + d^2) \text{ determined via hint} \tag{44}$$

$$= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} \text{ Where } \sigma = \sqrt{c^2 + d^2} \text{ and } \mu = a - b \tag{45}$$

(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

   (46)

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

   (47)

   we got this from the distributions of the random variables given (48)

$$=> E[x-y] = 5.4 \quad (49)$$

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

   (50)

$$= \frac{1}{n^2}\sum_{n=1}^{100}(\text{Var}[x_n] + \text{Var}[y_n]) = \frac{1}{n^2}\sum_{n=1}^{100}(2.9^2 + 2.7^2) = \frac{1}{n^2}n(15.7) = \frac{1}{100}(15.7) \quad (51)$$

$$=> \text{Var}[x-y] = \sqrt{\frac{15.7}{100}}^2 \quad (52)$$

$$=> x - y \sim N(5.4, \sqrt{\frac{15.7}{100}}) \quad (53)$$

   (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

Note, that the $P(X - Y < 0) = P(X < Y) => P(X > Y) = 1 - P(X < Y) = 1 - P(X - Y < 0)$ (54)

we can assume this as the $P(X = Y) = 0$ (55)

We know $X - Y \sim N(5.4, \sqrt{15.7})$ (56)

$=> 1 - P(X - Y < 0) = 1 - 0.08646693 = .9135331$ (57)

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

$$P(\theta|y) = \frac{P(Y|\theta_i) * P(\theta_i)}{P(Y)} = \frac{\binom{5}{y}\theta_i^y(1-\theta_i)^{5-y} * \frac{1}{11}}{P(Y)} \tag{58}$$

Now with the law of total probability, we get (59)

$$= \frac{\binom{5}{y}\theta_i^y(1-\theta_i)^{5-y} * \frac{1}{11}}{\sum_{i=0}^{11}\binom{5}{y}\theta_i^y(1-\theta_i)^{5-y} * \frac{1}{11}} = \frac{\frac{1}{11}\binom{5}{y}\theta_i^y(1-\theta_i)^{5-y}}{\binom{5}{y}\frac{1}{11}\sum_{i=0}^{11}\theta_i^y(1-\theta_i)^{5-y}} = \frac{\theta_i^y(1-\theta_i)^{5-y}}{\sum_{i=0}^{11}\theta_i^y(1-\theta_i)^{5-y}} \tag{60}$$

   (b) & (c)

Figure 1: Notice the high probability for $\theta = 0.0$

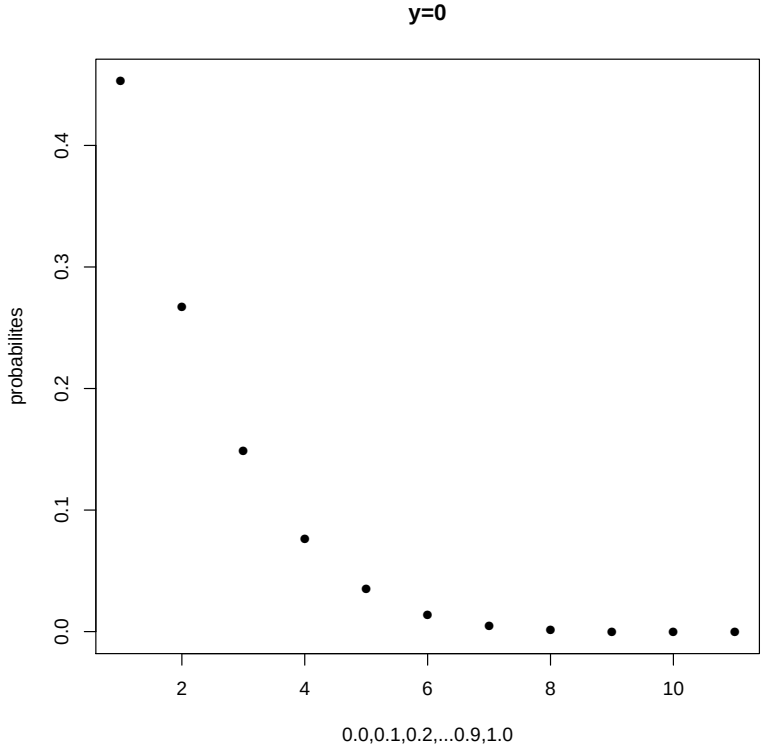


Figure 2: Here, we see a severe left skew due to the low success rate

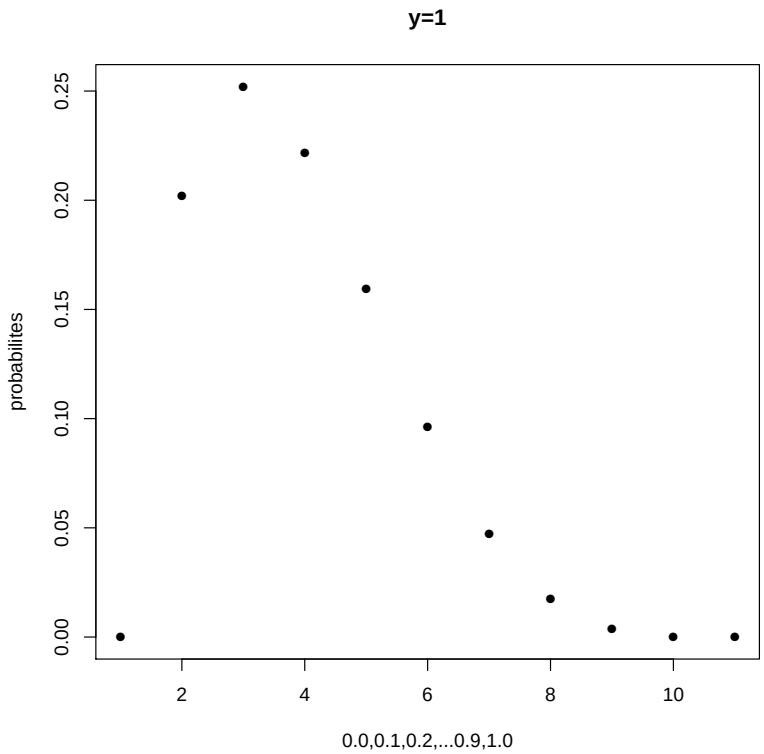


Figure 3: The skew becomes less severe as we have more successes

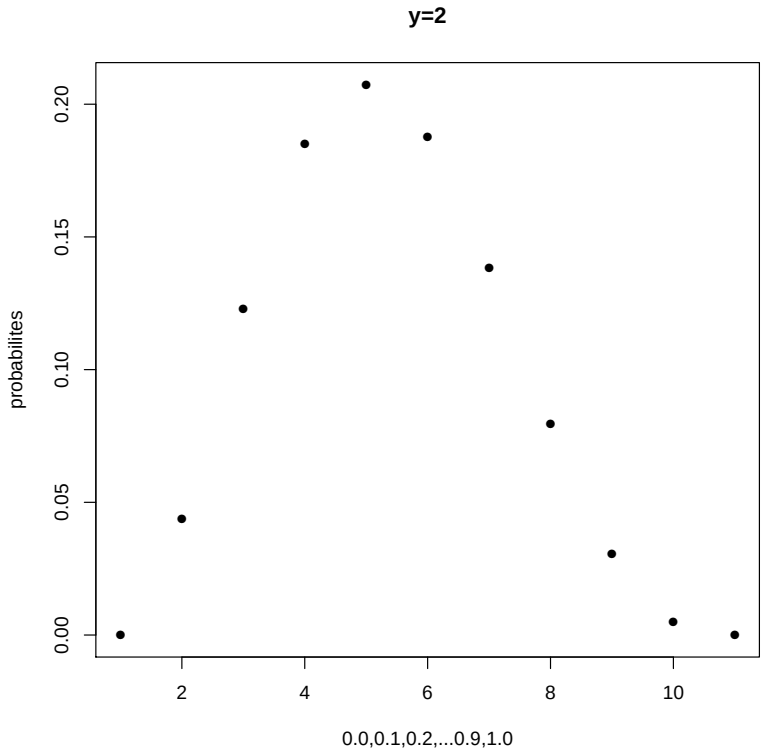


Figure 4: The skew has shifted to the right side as we approach y = 5

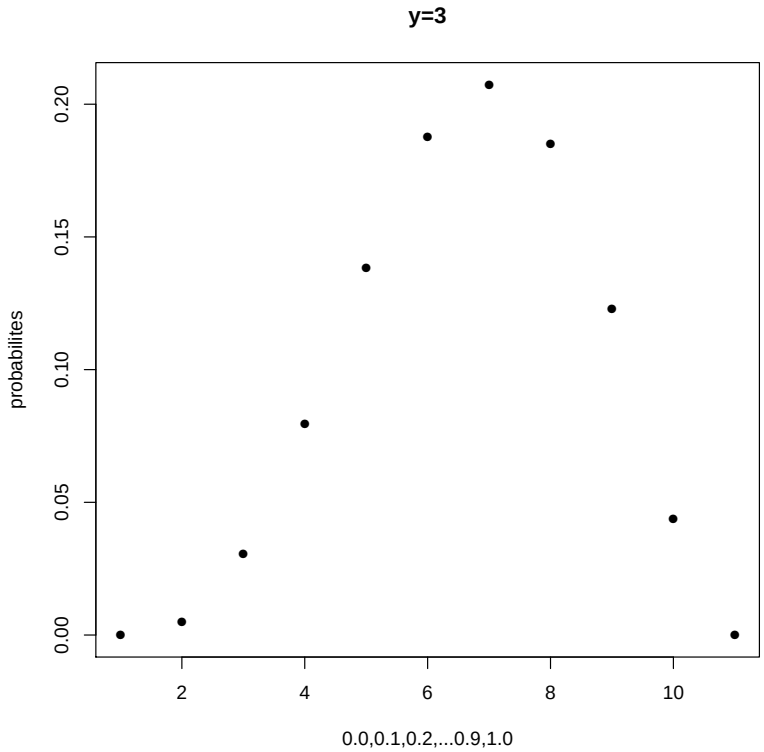


Figure 5: The skew is now extremely towards higher θ as the success rate increases

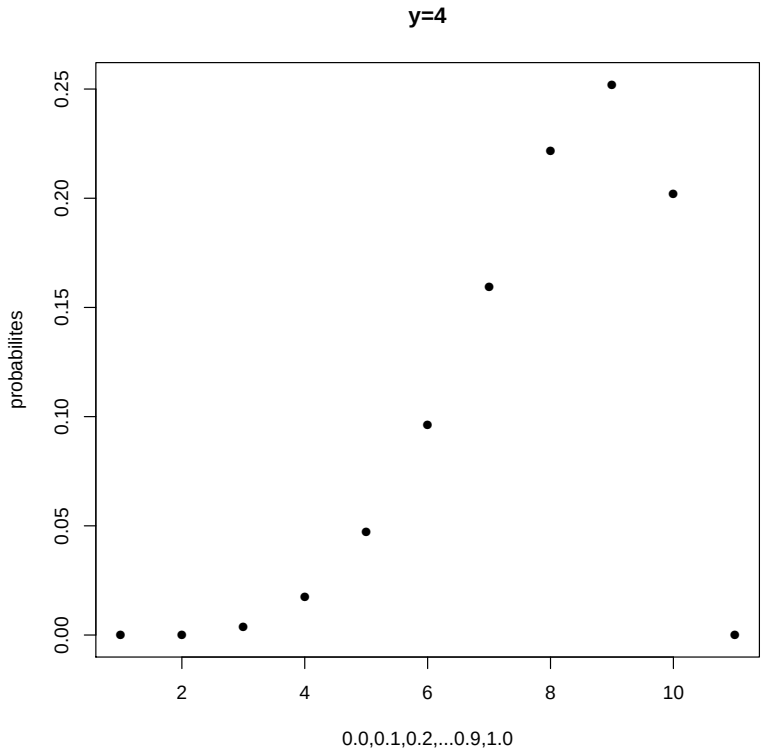
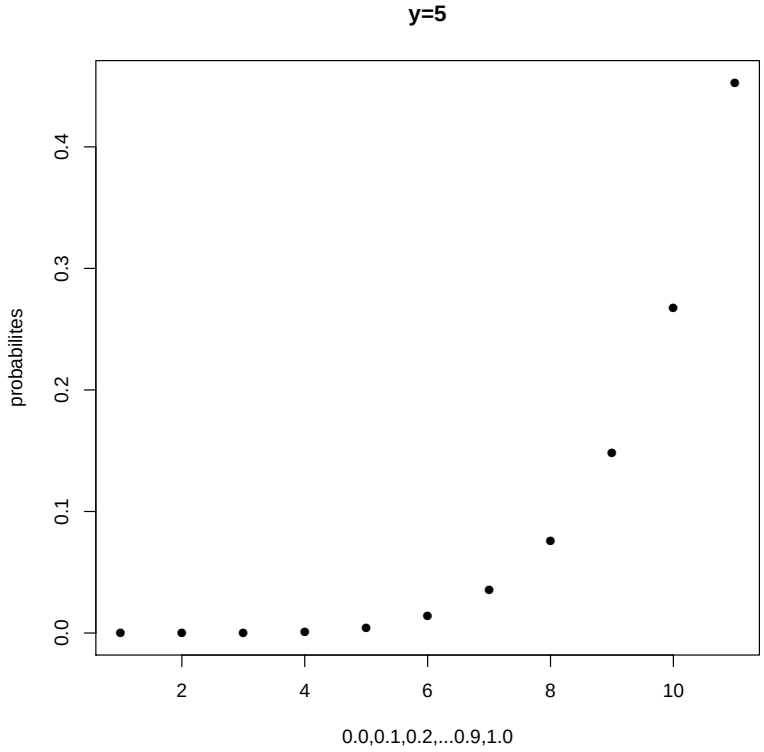


Figure 6: Now since the success is all 5 cases, $\theta = 1$ has the highest likelihood



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

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2} => F(x) = (1+e^{-x})^{-1} \tag{61}$$

$$=> \text{let } F(x) = u \text{ we want to solve for x in terms of you to find } F^{-1} \tag{62}$$

$$\frac{1}{1+e^{-x}} = u => \frac{1}{u} = 1 + e^{-x} => \frac{1-u}{u} = e^{-x} => \log(\frac{1-u}{u}) = -x => F^{-1}(X) = \log(\frac{u}{1-u}) \tag{63}$$

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
```