Lab 2

Alex Dauenhauer

June 11, 2018

Problem 1

(a) find $P(T|H_k)$

The first thing to do here is to set up the problem. We are given that T is the event that we have selected the trick coin. Let F be the event that the coin is fair. Therefore,

$$P(T) = 0.01, \qquad P(F) = 1 - P(T) = 0.99$$

We are given that H_k is the event that we get heads for all k coin tosses and we are trying to find $P(T|H_k)$ which is the probability that we have selected the trick coin, given that we have gotten heads on each of k tosses. To find $P(T|H_k)$ we can apply Bayes Theorem which says,

$$P(T|H_k) = \frac{P(H_k|T) \cdot P(T)}{P(H_k)}$$

We know that the probability of getting heads in all k tosses given that we have selected the trick coin is 1 since that is a property of the trick coin. Additionally we are given the probability of selecting the trick coin in the problem statement as 0.01. So we only need to find $P(H_k)$ to solve this equation. We know from the law of total probability that, if events, $A_1, A_2, ..., A_n$ are mutually exclusive, then for any other event B,

$$P(B) = \sum_n P(A_n \cap B)$$

Therefore, since there are only two events (either you have selected the trick coin or you haven't) and they are mutually exclusive,

$$P(H_k) = P(F \cap H_k) + P(T \cap H_k)$$

Using the multiplication rule we can rewrite this as,

$$P(H_k) = P(H_k|F) \cdot P(F) + P(H_k|T) \cdot P(T)$$

As stated above, the probability of getting heads k times if you have the trick coin is simply 1, the probability of selecting the trick coin is 0.01, and the probability of selecting a fair coin is 0.99. So we only need to find the probability of getting heads k times if you have selected a fair coin $P(H_k|F)$. If you have a fair coin, the probability of getting heads on a single flip is 0.5. Since there are only two outcomes, heads or tails, this can be represented by a Bernoulli distribution with probability $p = 0.5$. Then the probability of getting a heads k times given that you have a fair coin is,

$$P(H_k|F) = \left(0.5^1(1-0.5)^{1-1}\right)^k = 0.5^k$$

Therefore,

$$\begin{aligned} P(H_k) &= P(H_k|F) \cdot P(F) + P(H_k|T) \cdot P(T) \\ &= 0.5^k \cdot 0.99 + 1 \cdot 0.01 \end{aligned}$$

Now we have all the pieces we need to find $P(T|H_k)$, Therefore,

$$\begin{aligned} P(T|H_k) &= \frac{P(H_k|T) \cdot P(T)}{P(H_k)} \\ &= \frac{0.01}{0.5^k \cdot 0.99 + 0.01} \end{aligned}$$

(b) How many heads would need to be observed for the conditional probability that I have the trick coin to be 99%?

To solve this we use the solution to part (a) and set the LHS of the equation to 0.99 and solve for k. Therefore,

$$\begin{aligned}
0.99 &= P(T|H_k) \\
0.99 &= \frac{0.01}{0.5^k \cdot 0.99 + 0.01} \\
0.5^k \cdot 0.99 + 0.01 &= \frac{.01}{0.99} \\
0.5^k \cdot 0.99 &= \frac{.01}{0.99} - 0.01 \\
0.5^k &= \frac{\frac{.01}{0.99} - 0.01}{0.99} = 0.0001020304 \\
k \cdot \log(0.5) &= \log(0.0001020304) \\
k &= \frac{\log(0.0001020304)}{\log(0.5)} = 13.25871 \\
\Rightarrow k &= 14
\end{aligned}$$

Therefore, if we flip heads 14 times in a row, the probability that we have selected the trick coin will be greater than 99%

Problem 2

(a) Write the complete probability density function for random variable X, the total number of companies that reach unicorn status

As stated in the problem description, X is a binomial random variable with parameters $n = 2$ and $p = 3/4$. Therefore,

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2; \quad n = 2; \quad p = \frac{3}{4}$$

$$\begin{aligned}
\Rightarrow f(x) &= \begin{cases} \binom{2}{0}\left(\frac{3}{4}\right)^0 (1-\frac{3}{4})^{2-0}, & x = 0 \\ \binom{2}{1}\left(\frac{3}{4}\right)^1 (1-\frac{3}{4})^{2-1}, & x = 1 \\ \binom{2}{2}\left(\frac{3}{4}\right)^2 (1-\frac{3}{4})^{2-2}, & x = 2 \end{cases} \\
&= \begin{cases} 0.0625, & x = 0 \\ 0.3750, & x = 1 \\ 0.5625, & x = 2 \end{cases}
\end{aligned}$$

Checking this result in R gives,

```
n = 2
p = 3/4
x = c(0,1,2)
(P = dbinom(x, size=2, prob=p))
```

```
## [1] 0.0625 0.3750 0.5625
```

(b) give a complete expression for the cumulative probability of X

The cumulative distribution function of X is probability that $X \leq x$. This is calculated by taking the sum of the probabilities of events where $X \leq x$. This is written as,

$$F(X) = P(X \leq x) = \sum_{y:y \leq x} f(y)$$

So using our probability mass function above,

$$\begin{aligned} F(X) &= \begin{cases} 0.0625, & 0 \leq x < 1 \\ 0.0625 + 0.3750, & 1 \leq x < 2 \\ 0.0625 + 0.3750 + 0.5625, & 2 \leq x \end{cases} \\ &= \begin{cases} 0.0625, & x = 0 \\ 0.4375, & x = 1 \\ 1, & x = 2 \end{cases} \end{aligned}$$

Checking with R,

```
(F = pbinom(x, size=2, prob=p))
```

```
## [1] 0.0625 0.4375 1.0000
```

(c) Compute $E(X)$

The expectation value of X is given by taking the sum of the product of the value of x and the probability of getting that value of x. This is written as,

$$\begin{aligned} E(X) &= \sum_X x \cdot f(x) \\ &= (0.0625 \times 0) + (0.375 \times 1) + (0.5625 \times 2) \\ &= 1.5 \end{aligned}$$

Checking with R,

```
(E = sum(x * P))
```

```
## [1] 1.5
```

(d) Compute $var(X)$

The variance is written as,

$$\begin{aligned} V(X) &= E((X - \mu_x)^2) \\ &= E((X - \mu_x) \cdot (X - \mu_x)) \\ &= E(X^2 - 2X\mu_x + \mu_x^2) \\ &= E(X^2) - E(2X\mu_x) + E(\mu_x^2) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

We first need to find $E(X^2)$,

$$\begin{aligned} E(X^2) &= \sum_X x^2 \cdot f(x) \\ &= (0.0625 \times 0^2) + (0.375 \times 1^2) + (0.5625 \times 2^2) \\ &= 2.625 \end{aligned}$$

Therefore,

$$V(X) = E(X^2) - E(X)^2 = 2.625 - 1.5^2 = 0.375$$

Checking with R,

```
(V = sum(x^2 * P) - E^2)
```

```
## [1] 0.375
```

Problem 3

(a) Draw a graph of the region for which X and Y have positive probability density

I will use R to graph this. First step is to set up the variables. The problem statement says this can be modeled by letting A_1 and A_2 be two independent random variables which are uniformly distributed on $[0, 1]$ and letting $X = \max(A_1, A_2)$ and $Y = \min(A_1, A_2)$. So the first step is to generate A_1 and A_2 variables

```
a1 = seq(0,1,.0001)
a2 = seq(0,1,.0001)
```

The next step is to define a way to randomly select an a_1 and an a_2 from these two distributions and then define X as the max and Y as the min. To do this I will write a function that takes a single sample from the A_1 and A_2 distributions and sets $X = max(A_1, A_2)$ and $Y = min(A_1, A_2)$. I append the result to an existing dataframe so that I can repeat the experiment many times and store the results in a single variable.

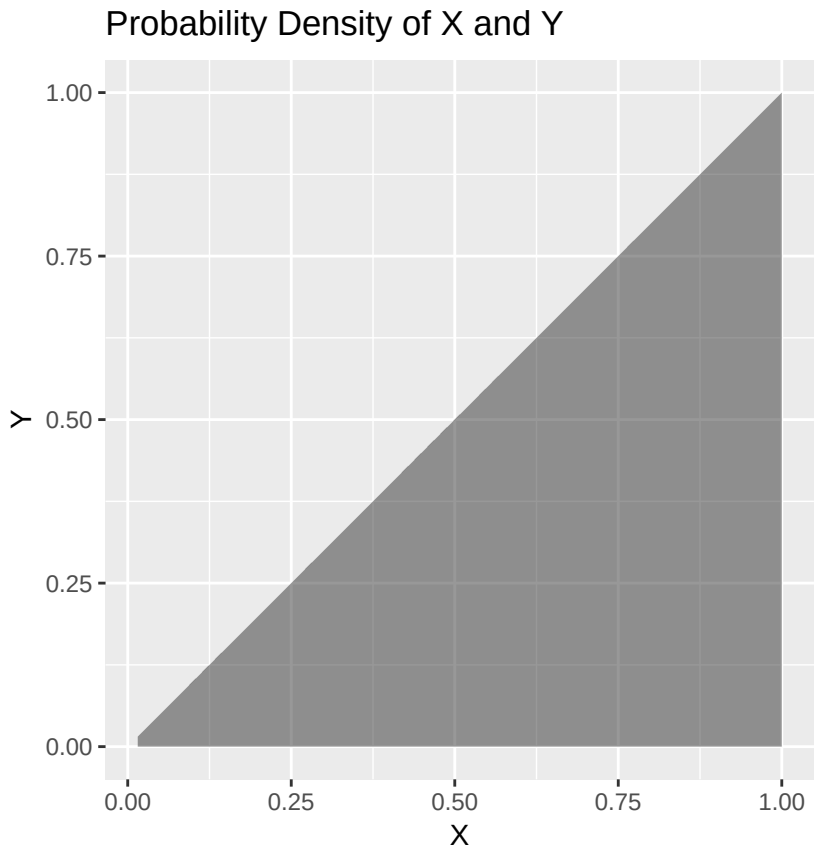
```
single_sample = function(df,a1,a2){
  S = c(sample(a1, 1, replace=T), sample(a2, 1, replace=T))
  x = max(S)
  y = min(S)
  return(rbind(df, data.frame('x'=x,'y'=y)))
}
```

Next, I want to repeat this experiment many times in order to view the distribution.

```
df = data.frame()
for (i in 1:10000){
  df = single_sample(df, a1, a2)
}
```

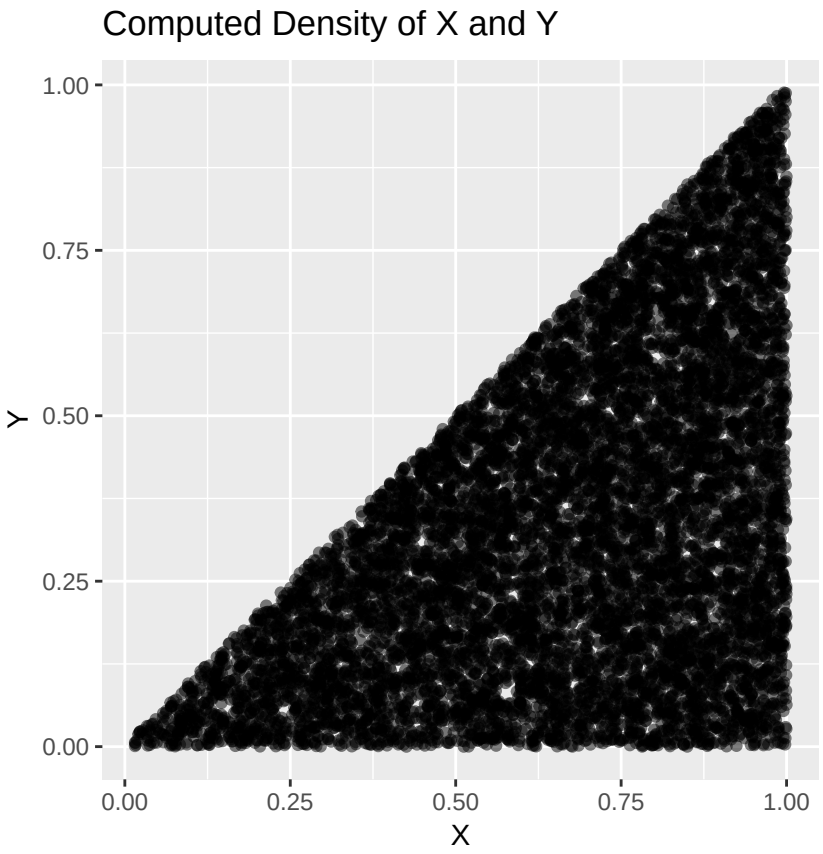
Now that I have collected my sample, I can plot the probability distribution. In our distribution, X exists from $[0, 1]$ and Y exists from $[0, X]$. So the area where both X and Y have positive probability density is the area under the line $y = x$. This can be plotted in the following way,

```
f_xy = function(x){x}
ggplot(df, aes(x,y)) +
  stat_function(fun=f_xy, geom='area', alpha=0.5) +
  labs(title='Probability Density of X and Y',
       x='X',
       y='Y'
  ) +
  coord_fixed()
```



To confirm this is correct, we can plot the simulated X and Y values from our sample

```
ggplot() +
  geom_point(data=df, mapping=aes(x,y), alpha=0.5) +
  labs(title = 'Computed Density of X and Y',
       x = 'X',
       y = 'Y'
  ) +
  coord_fixed()
```



(b) Derive the marginal probability density function of X

The marginal pdf of X is calculated by integrating over all Y the joint pdf $f(x, y)$. Therefore,

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy = \int_0^x 2dy = 2y\Big|_0^x = 2x$$

(c) Derive the unconditional expectation of X

For a joint distribution of two continuous variables, the unconditional expectation of one variable is calculated by integrating the product of the single variable in question with it's marginal distribution, over all possible values of that variable. Therefore,

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)dx = \int_0^1 2x^2 dx = \frac{2x^3}{3}\Big|_0^1 = \frac{2}{3}$$

I can use the sample collected above to check whether this result makes sense

```
(Ex = mean(df$x))
```

```
## [1] 0.6669748
```

```
2/3
```

```
## [1] 0.6666667
```

(d) Derive the conditional probability density function of Y, conditional on X

To get the conditional pdf of Y conditional on X, we can use the multiplication rule since we know the joint pdf and the marginal pdf of x

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{2}{2x} = \frac{1}{x}$$

(e) Derive the conditional expectation of Y, conditional on X

The conditional expectation of Y conditional on X can be calculated by integrating the product of the variable Y and the conditional pdf of Y conditional on X, over all possible values of Y. Therefore,

$$E(Y|X) = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y|x)dy = \int_0^x y\Big(\frac{1}{x}\Big)dy = \frac{y^2}{2x}\Big|_0^x = \frac{x}{2}$$

This result makes sense, since $f_{X,Y}(x,y)$ is only defined in the region of $0 < y < x < 1$, and X and Y represent the max and min of two random variables uniform on $[0,1]$, then given a particular X, we know Y will be uniform on the interval $0 < y < x$ and so the conditional expectation of Y given X is simply the midpoint of this interval.

(f) Derive $E(XY)$

$E(XY)$ can be solved for using the Law of Iterated Expectations. As noted in the prompt, if we take an expectation conditional on X, then X is just a constant inside the expectation. Therefore,

$$\begin{aligned}
E(XY) &= E(E(XY|X)) \\
&= E(XE(Y|X)) \\
&= E(x^2/2) \\
&= \frac{1}{2}\int_{-\infty}^{\infty} x^2 f_X(x)dx \\
&= \frac{1}{2}\int_0^1 2x^3 dx \\
&= \frac{x^4}{4}\Big|_0^1 \\
&= \frac{1}{4}
\end{aligned}$$

(g) Derive cov(X,Y)

The $cov(X,Y)$ is given by

$$\begin{aligned}
cov(X,Y) &= E((X-\mu_X)(Y-\mu_Y)) \\
&= E(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y) \\
&= E(XY) - E(X\mu_Y) - E(Y\mu_X) + E(\mu_X\mu_Y) \\
&= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\
&= E(XY) - E(X)E(Y)
\end{aligned}$$

We know $E(XY)$ and $E(X)$ so we just need to find $E(Y)$ and we can directly find $cov(X,Y)$. We can find $E(Y)$ using the law of iterated expectations, since we know $E(Y|X)$. Therefore,

$$E(Y) = E(E(Y|X)) = E\Big(\frac{X}{2}\Big) = \frac{E(X)}{2} = \frac{1}{3}$$

I can again use the sample collected above to check that this result makes sense

```
(Ey = mean(df$y))
```

```
## [1] 0.3375355
```

```
1/3
```

```
## [1] 0.3333333
```

Therefore, the $cov(X, Y)$ is just

$$cov(X, Y) = E(XY) - E(X)E(Y) = \frac{1}{4} - \left(\frac{2}{3} \cdot \frac{1}{3}\right) = \frac{1}{36}$$

Using the sample collected above to check this,

```
cov(df$x, df$y)
```

```
## [1] 0.0281874
```

```
1/36
```

```
## [1] 0.02777778
```

Problem 4

(a) Compute the expectation of each indicator variable, $E(D_i)$

In order to compute the expectation of D_i, we first need to compute the pdf of D_i, $f(D_i)$. We are given that both X and Y are uniformly distributed on $[-1, 1]$ and that the indicator D_i is a success if $X_i^2 + Y_i^2 < 1$. So the probability that D_i will be a success is the area of success divided by the total area of the distribution. The total area is given by,

$$A_{total} = (1 - (-1)) \times (1 - (-1)) = 4$$

And the area of success is given by

$$A_{success} = \pi r^2 = \pi(1)^2 = \pi$$

Therefore,

$$f(D_i) = \begin{cases} \pi/4, & D_i = 1 \\ 1 - \pi/4 & D_i = 0 \end{cases}$$

Now we can calculate the expectation value of D_i,

$$E(D_i) = \sum_D d \cdot f(d) = 1 \cdot \left(\frac{\pi}{4}\right) + 0 \cdot \left(1 - \frac{\pi}{4}\right) = \frac{\pi}{4}$$

(b) Compute the standard deviation of each D_i

Since the standard deviation is simply the square root of the variance, I will start by calculating the variance

$$V(D_i) = E(D_i^2) - E(D_i)^2$$

I know $E(D_i)$ from above. I can calculate $E(D_i^2)$ in the same way

$$E(D_i^2) = \sum_D d^2 \cdot f(d) = 1^2 \cdot \left(\frac{\pi}{4}\right) + 0^2 \cdot \left(1 - \frac{\pi}{4}\right) = \frac{\pi}{4}$$

Therefore,

$$V(D_i) = E(D_i^2) - E(D_i)^2 = \frac{\pi}{4} - \left(\frac{\pi}{4}\right)^2 = \frac{\pi}{4} - \frac{\pi^2}{16}$$

So the standard deviation σ_{D_i} is

$$\sigma_{D_i} = \sqrt{V(D_i)} = \sqrt{\frac{\pi}{4} - \frac{\pi^2}{16}}$$

(c) Compute the standard error of $\bar{D}$

The standard error is related to the standard deviation, by dividing by the standard deviation by the square root of the sample size n. Therefore,

$$\sigma_{\bar{D}} = \frac{\sigma_{D_i}}{\sqrt{n}} = \sqrt{\frac{\frac{\pi}{4} - \frac{\pi^2}{16}}{n}}$$

(d) Compute the probability that $\bar{D}$ is greater than 3/4

To do this we need to use the Central Limit Theorem. The Central Limit Theorem states that, if the number of samples n is sufficiently large, the distribution of the sample mean will be a normal distribution with mean equal to the population mean, μ, and sample standard deviation equal to the population standard deviation, σ, divided by the square root of the sample size, n, $\sigma/\sqrt{n}$. This is extremely useful because now we can easily determine the probability that a specific range of values will contain the mean of our random variable. The probability that a normally distributed random variable will be less than or equal to a critical value can be computed by calculating a z-score, and looking up the z-value in a z-score table. The z-score calculates the value we should plug into the CDF of our distribution, denoted as $\Phi(z)$, and since we want the probability that our random variable is greater than a specific value, we need to subtract the $\Phi(z)$ from 1 to get the area of the curve greater than our specific value. Therefore,

$$\begin{aligned} z &= \frac{X - \mu}{\sigma/\sqrt{n}} \\ &= \frac{\frac{3}{4} - \frac{\pi}{4}}{\sqrt{\frac{\frac{\pi}{4} - \frac{\pi^2}{16}}{n}}} \\ &= -0.8622219 \end{aligned}$$

Looking this value for z up in a z-score table gives a value for for the area under the curve of $\Phi(z) = 0.1949$. Therefore,

$$P(3/4 \leq \bar{D}) = 1 - \Phi(z) = 1 - 0.1949 = 0.8051$$

We can compare our result to the output of the `pnorm` function which calculates the CDF of the input normal distribution

```
n = 100
(mu = pi/4)
```

```
## [1] 0.7853982
```

```
sigma = sqrt((pi/4) - (pi^2/16))
(sigma_bar = sigma/sqrt(n))
```

```
## [1] 0.04105458
```

```
x = 3/4
(Z = (x - mu)/(sigma_bar))
```

```
## [1] -0.8622219
```

```
phi = .1949
(1 - phi)
```

```
## [1] 0.8051
```

```
(P = 1 - pnorm(x, mean=mu, sd=sigma_bar))
```

```
## [1] 0.8057173
```

Which gives the same result. So $P\left(\frac{3}{4} < \bar{D}\right) \approx 0.80$

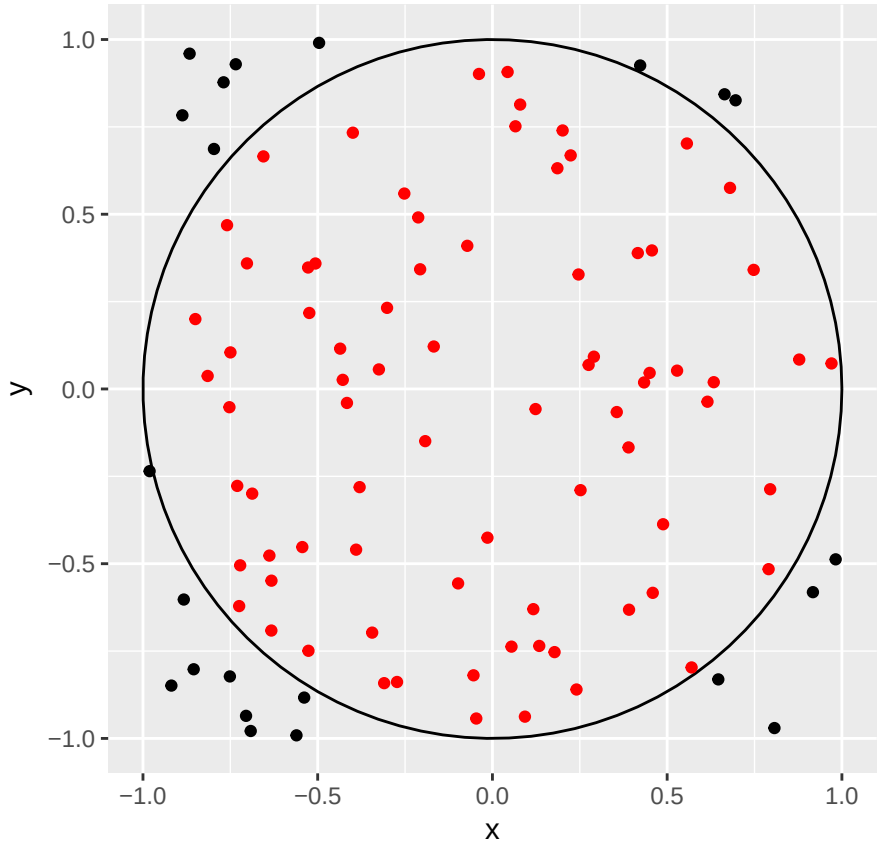
(e) Visualize drawing 100 observations for X, Y and D

To do this I created a function called `get_sample()`, which takes the number of draws n as an input and outputs a dataframe with each row representing a single draw and columns X, Y, and D. I then use `geom_point()` to plot the resulting $[x, y]$ points and color them based on whether they are a success or not.

```
set.seed(25)
get_sample = function(n){
  x = sample(seq(-1,1,.0001),n,replace=T)
  y = sample(seq(-1,1,.0001),n,replace=T)
  d = as.numeric(x^2 + y^2 < 1)
  df = data.frame('x'=x, 'y'=y, 'd'=d)
  return(df)
}

circle = function(center=c(0,0), radius=1, n=NULL){
    theta = seq(0, 2*pi, length.out=n)
    x = center[1] + radius * cos(theta)
    y = center[2] + radius * sin(theta)
    return(data.frame(x=x, y=y))
}

n = 100
df = get_sample(n)
c = circle(n=n)
ggplot() +
  geom_path(data=c, mapping=aes(x,y)) +
  geom_point(data=df, mapping=aes(x,y), col=df$d+1) +
  coord_fixed()
```



(f)

```
mean(df$d)
```

```
## [1] 0.78
```

```
mu
```

```
## [1] 0.7853982
```

The expected value is $\pi/4$ which is approximately 0.78, so the measured average for this particular sample was very close to that.

(g) Plot a histogram of 10000 sample averages

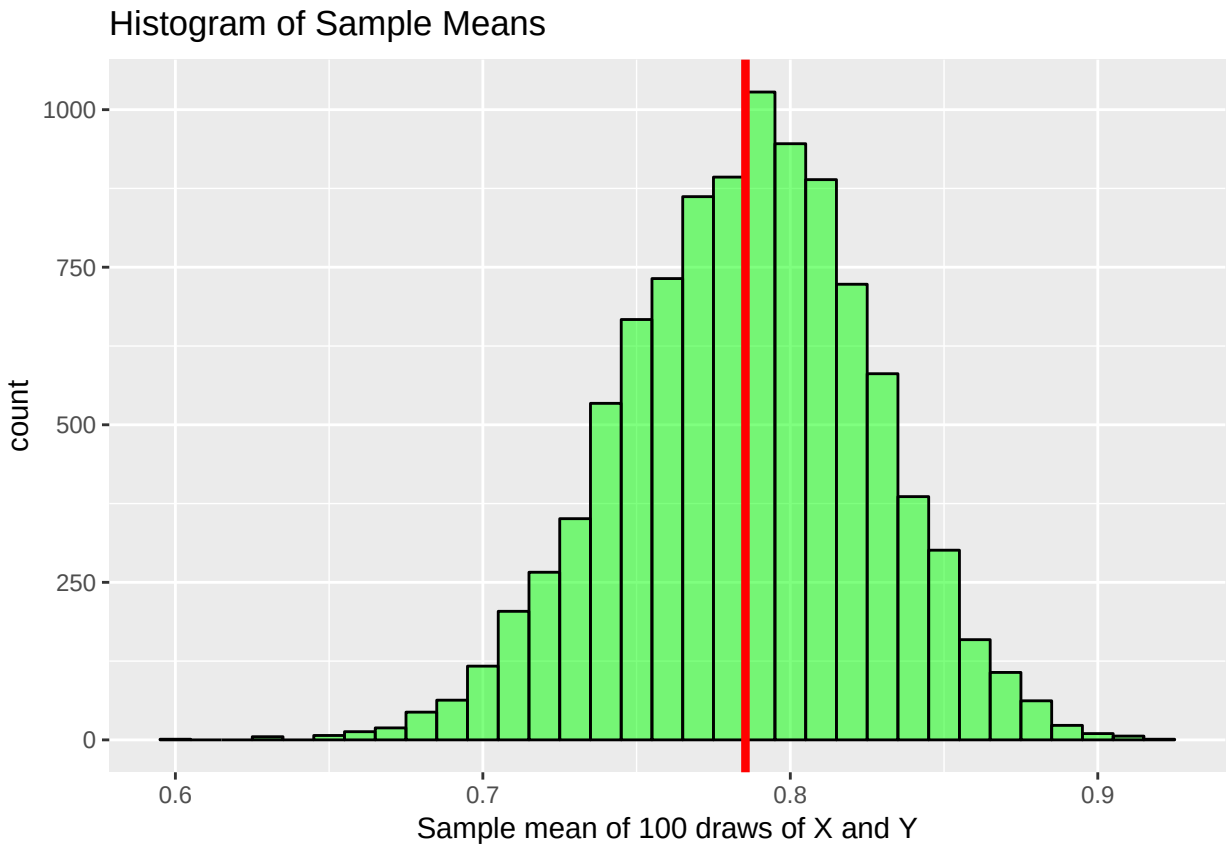
I want to use R's built in replicate function to improve speed rather than using a for loop. To do this I need to create a function that runs my `get_sample()` function and then returns the mean of D from the sample.

```
sample_mean = function(n){
  df = get_sample(n)
  return(mean(df$d))
}

iter = 10000
n = 100
S = replicate(iter, sample_mean(n))
```

Now that I have my distribution of samples, I can plot the histogram. I have overlaid the population mean, μ as a vertical red line

```
dS = data.frame(S)
ggplot() +
  geom_histogram(mapping=aes(S), data=dS, binwidth = 0.01,
                 fill='green', alpha=0.5, col='black') +
  geom_vline(xintercept=mu, col='red', lwd=1.5) +
  labs(title='Histogram of Sample Means',
       x='Sample mean of 100 draws of X and Y')
```



(h) Compute the standard deviation of your sample and compare to the expected population standard deviation

```
(sigma2 = sd(dS$S))
```

```
## [1] 0.04069559
```

```
sigma_bar
```

```
## [1] 0.04105458
```

```
(percent_diff = abs(((sigma_bar - sigma2)/sigma_bar) * 100))
```

```
## [1] 0.8744279
```

The distribution standard deviation is around 1% different from the expected standard deviation of the population so this is very close and confirms that the expected population standard deviation was calculated

correctly

(i) Compute the fraction of your sample averages that are larger than 3/4 and compare to the answer from part d

```
length(dS$S[dS$S > 3/4]) / length(dS$S)
```

```
## [1] 0.7709
```

```
P
```

```
## [1] 0.8057173
```

The result is very close. The true probability that $D_i < 3/4$ is roughly 80% and we calculated 77% in our distribution. This agreement is very good and implies that we have performed the calculations correctly. Increasing n improves the comparison between measured and expected probability of $D_i < 3/4$.

```
n = 1000
iter = 10000
S = replicate(iter, sample_mean(n))
length(S[S > 3/4]) / length(S)
```

```
## [1] 0.9962
```

```
(P2 = 1 - pnorm(3/4, mean=pi/4, sd=sigma/sqrt(n)))
```

```
## [1] 0.9968003
```