Lab 2: Probability Theory
w203: statistics for Data Science

Adam Yang, Section 4.

1a. T = event you select a trick coin

H_K = event where all K flips are heads.

We know:

$P(T) = 0.01 = \frac{1}{100}$: We have 100 coins, 1 is a trick coin

$P(H_K \mid T) = 1$: Trick coin can only come up with heads

$P(H_K \mid T') = (0.5)^K$: Given that we selected a fair coin, probability of getting only heads is $(0.5)^K$

Solve $P(T \mid H_K) = \dfrac{P(T \wedge H_K)}{P(H_K)}$

$P(T \wedge H_K) = P(H_K \wedge T) = P(H_K \mid T)\,P(T) = (1)(0.01) = 0.01$

$$\begin{aligned} P(H_K) &= P(H_K \wedge T) + P(H_K \wedge T') \\ &= P(H_K \wedge T) + P(H_K \mid T')\,P(T') \\ &= 0.01 + (0.99)(0.5)^K \end{aligned}$$

$$\therefore P(T \mid H_K) = \frac{0.01}{0.01 + (0.99)(0.5)^K}$$

1b) How many flips to get $P(T|H_K) > 0.99$?

$$\frac{0.01}{0.01 + (0.99)(0.5)^K} > 0.99$$

$$\frac{0.01}{0.99} > 0.01 + (0.99)(0.5)^K$$

$$\frac{\left(\frac{0.01}{0.99} - 0.01\right)}{0.99} > 0.5^K$$

$$\ln\left(\frac{\frac{0.01}{0.99} - 0.01}{0.99}\right) > K \ln(0.5)$$

$$\frac{\ln\left(\frac{\frac{0.01}{0.99} - 0.01}{0.99}\right)}{\ln(0.5)} \overset{K}{<} K$$

$\therefore \ln(0.5)$ is negative, flip signs when dividing by a negative number

$$\therefore K > \frac{\ln\left(\frac{\frac{0.01}{0.99} - 0.01}{0.99}\right)}{\ln(0.5)} = 13.2587$$

$K > 13.2587$

You need at least 14 heads in a row

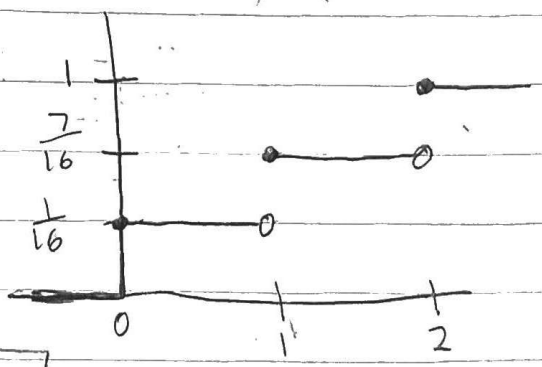
2a. Let S_1 = event where company 1 becomes a unicorn
Let S_2 = event where company 2 becomes a unicorn,
$P(S_1) = P(S_2) = \frac{3}{4}$ S_1 and S_2 are independent.
Let X = total number of companies that become unicorns.

Possible outcomes, $X = 0$ $(S_1' \wedge S_2')$
$X = 1$ $(S_1 \wedge S_2') \cup (S_1' \wedge S_2)$
$X = 2$ $(S_1 \wedge S_2)$

for $X=0$, $\left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{1}{16}$
$X=1$, $\left(\frac{3}{4}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{6}{16}$
$X=2$, $\left(\frac{3}{4}\right)\left(\frac{3}{4}\right) = \frac{9}{16}$

$$f(x) = \begin{cases} \frac{1}{16}, & x=0 \\ \frac{6}{16}, & x=1 \\ \frac{9}{16}, & x=2 \\ 0, & \text{otherwise} \end{cases}$$

2b. 0, $x<0$
$\frac{1}{16}$, $0 \le x < 1$
$\frac{1}{16} + \frac{6}{16}$, $1 \le x < 2$
$\frac{1}{16} + \frac{6}{16} + \frac{9}{16}$, $x \ge 2$



$$F(x) = \begin{cases} 0, & x<0 \\ \frac{1}{16}, & 0 \le x < 1 \\ \frac{7}{16}, & 1 \le x < 2 \\ 1, & x \ge 2 \end{cases}$$

2c) $E(x) = \sum_{i=1}^{2} x_i f(x_i) = \frac{1}{16}(0) + \left(\frac{6}{16}\right)(1) + \left(\frac{9}{16}\right)(2)$

$E(x) = 1.5$

2d) $Var(x) = E(x^2) - [E(x)]^2$

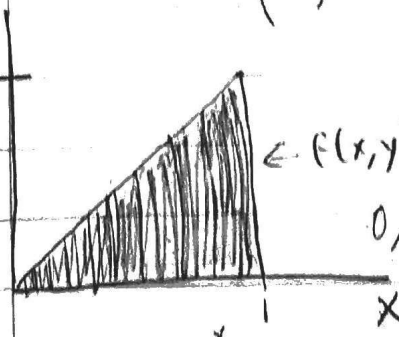
$E(x^2) = \sum_{i=1}^{2} x_i^2 f(x_i) = (0)^2\left(\frac{1}{16}\right) + (1)^2\left(\frac{6}{16}\right) + (2)^2\left(\frac{9}{16}\right)$

$= \frac{6}{16} + \frac{9(4)}{16} = \frac{42}{16} = 2.625$

$E(x)^2 = (1.5)^2 = \frac{9}{4} = 2.25$

$Var(x) = 2.625 - 2.25 = \underline{0.375}$

3a) $f(x,y) = \begin{cases} 2, & 0 < y < x \leq 1 \\ 0, & \text{otherwise.} \end{cases}$



3b) $f_x(x) = \int_{y=0}^{x} f_{x,y}(x,y)\, dy = \int_{0}^{x} 2\, dy = \Big[2y\Big]_{0}^{x} = 2x$

$$f_x(x) = \begin{cases} 2x, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

3c) $E(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_0^1 2x^2\,dx = \left[\frac{2}{3}x^3\right]_0^1 = \frac{2}{3}$

$\boxed{E(X) = \frac{2}{3}}$

3d) $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{2}{2x} = \frac{1}{x}$

$$\boxed{f_{Y|X}(y|x) = \begin{cases} \frac{1}{x}, & 0<x<1 \\ 0, & \text{otherwise} \end{cases}}$$

3e) $E(Y|X) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\,dy = \int_0^X \frac{y}{x}\,dy = \left[\frac{y^2}{2x}\right]_0^X = \frac{x^2}{2x}$

$\boxed{E(Y|X=x) = \frac{x}{2} \quad \text{for} \quad 0<x<1}$

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

3f) $E(XY) = E\big(E(YX|X)\big) = E\big(xE(Y|X)\big) = E\left(\frac{X^2}{2}\right)$

$E\left(\frac{X^2}{2}\right) = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx = \int_0^1 \frac{x^2}{2}\cdot 2x\,dx = \int x^3\,dx$

$= \left[\frac{1}{4}x^4\right]_0^1$ $\qquad \boxed{E(XY) = \frac{1}{4}}$

3g) $Cov(X,Y) = E(XY) - E(X)E(Y)$

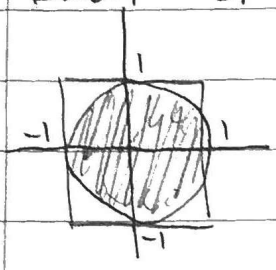
$E(Y) = E\big(E(Y|X)\big) = E\left(\frac{X}{2}\right) = \frac{1}{2}E(X) = \frac{1}{2}\cdot\frac{2}{3}$

$E(Y) = \frac{1}{3}$

$Cov(X,Y) = \frac{1}{4} - \left(\frac{2}{3}\right)\left(\frac{1}{3}\right) = \frac{1}{36}$ $\qquad \boxed{Cov(X,Y) = \frac{1}{36}}$

4a) $D_i = \begin{cases} 1, & X_i^2 + Y_i^2 < 1 \\ 0, & \text{otherwise} \end{cases}$

Each D_i is a Bernoulli variable and are i.d.d.



← ratio of these areas is the probability that (X_i, Y_i) falls in the circle.

$$= \frac{\pi r^2}{2 \cdot 2} = \frac{\pi}{4}$$

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

$$E(D_i) = \sum_{i=0}^{1} d_i f(d_i)$$

$$= \left(\tfrac{\pi}{4}\right)(1) + \left(1 - \tfrac{\pi}{4}\right)(0) = \frac{\pi}{4}$$

$$\boxed{E(D_i) = \frac{\pi}{4}}$$

4b) $\mathrm{Var}(D_i) = E(D_i^2) - [E(D_i)]^2$

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

$$\boxed{\begin{aligned} \mathrm{Var}(D_i) &= \tfrac{\pi}{4} - \left(\tfrac{\pi}{4}\right)^2 \approx 0.1685 \\ \sigma &= \sqrt{\tfrac{\pi}{4} - \left(\tfrac{\pi}{4}\right)^2} \approx 0.4105 \end{aligned}}$$

4c) $E(D_i)$ and $\mathrm{Var}(D_i)$ are know,

Central limit theorem says that is n is large,

$$\boxed{\sigma_{\bar{D}} = \frac{\sigma}{\sqrt{n}} = \frac{0.4105}{\sqrt{n}}}$$

w203 Lab2: Probability Theory

Adam Yang

6/21/2018

4) Circles, Random Samples, and the Central Limit Theorem

Parts a, b, and c are handwritten

d. Now let n = 100. Using the Central Limit Theorem, compute the probability that $\bar{D}$ is larger than 3/4. Make sure you explain how the Central Limit Theorem helps you get your answer.

Answer:

Let X_1, X_2, ..., X_n be a random sample from a distribution with mean, μ and variance, σ^2. The Central Limit Theorem states that if sample size, n is large enough, the distribution of sample means, $\bar{X}$, has approximately a normal distribution with $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Therefore, by applying the Central Limit Theorem, the distribution of $\bar{D}$ has a mean of $\mu_{\bar{D}} = \mu = \frac{\pi}{4}$ and a standard deviation of $\sigma_{\bar{D}} = \frac{\sigma}{\sqrt{n}} \approx \frac{0.4105}{\sqrt{n}}$.

We can subtract the μ from the sample mean and then divide the result by $\frac{\sigma}{\sqrt{n}}$ to get the standard normal variable, otherwise known as the Z-value:

$$Z = \frac{\bar{D}-\mu}{\sigma/\sqrt{n}}$$

$$Z = \frac{\frac{3}{4}-\frac{\pi}{4}}{0.4105/\sqrt{100}}$$

By finding pnorm(Z), we will know the probability that $\bar{D}$ is less than 3/4. By subtracting that number from 1, we will know the probability that $\bar{D}$ is larger than 3/4.

```
# list variables
miu <- pi/4
sigma <- sqrt(pi/4-(pi/4)^2)
D_bar <- 3/4
n = 100
# solve for z-value
z <- (D_bar-miu)/(sigma/sqrt(n))
# solve for probability that D_bar > 3/4
1 - pnorm(z)
```

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

Therefore, the probability that $\bar{D}$ is larger than 3/4 is 0.806.

I should mention that the Central Limit Theorem's rule of thumb for sufficient sample size is anything larger than 30. However, if there is a large skew to the data, the sample size would have to be much larger to be sufficient to approximate a normal distribution of sample means.

e. Now let n = 100. Use R to simulate a draw for X_1, X_2, ..., X_n and Y_1, Y_2, ..., Y_n. Calculate the resulting values for D_1, D_2, ..., D_n. Create a plot to visualize your draws, with X on one axis and Y on the other. We suggest using a command like the following to assign a different color to each point, based on whether it falls inside the unit circle or outside it. Note that we pass $d + 1$ instead of d into the color argument because 0 corresponds to the color white.

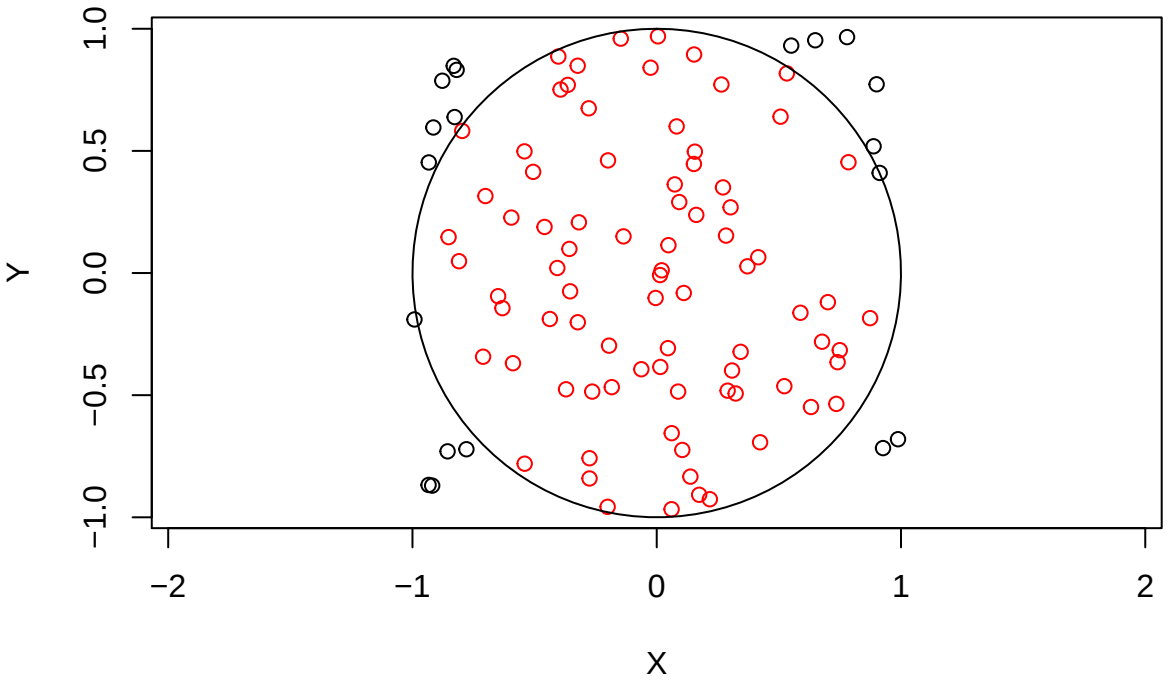
Answer:

```
n = 100
# Get samples for X and Y
X <- runif(n, min = -1, max = 1)
Y <- runif(n, min = -1, max = 1)

# Compute the samples for D based on X and Y.
D = c()
for(i in 1:n) {
  if(X[i]^2 + Y[i]^2 < 1) {
    D = c(D,1)
  } else {
    D = c(D,0)
  }
}


# plot x and y, highlight the values where D=1
plot(X,Y, xlim=c(-1,1), col = D+1, asp =1)

# plot a unit circle
theta <- seq(0,2*pi,length=100)
circle <- t(rbind(sin(theta), cos(theta)))
lines(circle)
```



f. What value do you get for the sample average, $\bar{D}$? How does it compare to your answer in part a?

Answer:

```
mean(D)
```

```
## [1] 0.81
```

The sample average I got is shown above. In part a, I calculated the expectation to be $\frac{\pi}{4} \approx 0.785$. The sample average I calculated above, is close to 0.785, but not as much as I would have liked. I think although we

approximated a normal distribution, the distribution is still fairly wide. If we increased our sample size, we would yield a better sample average. This speaks a little bit towards the skew of our Bernoulli distribution. Even though 100 samples is much more than 30, it seems like there is still room for improvement by increasing the sample size even more.

g. Now use R to replicate the previous experiment 10,000 times, generating a sample average of the D_i each time. Plot a histogram of the sample averages.

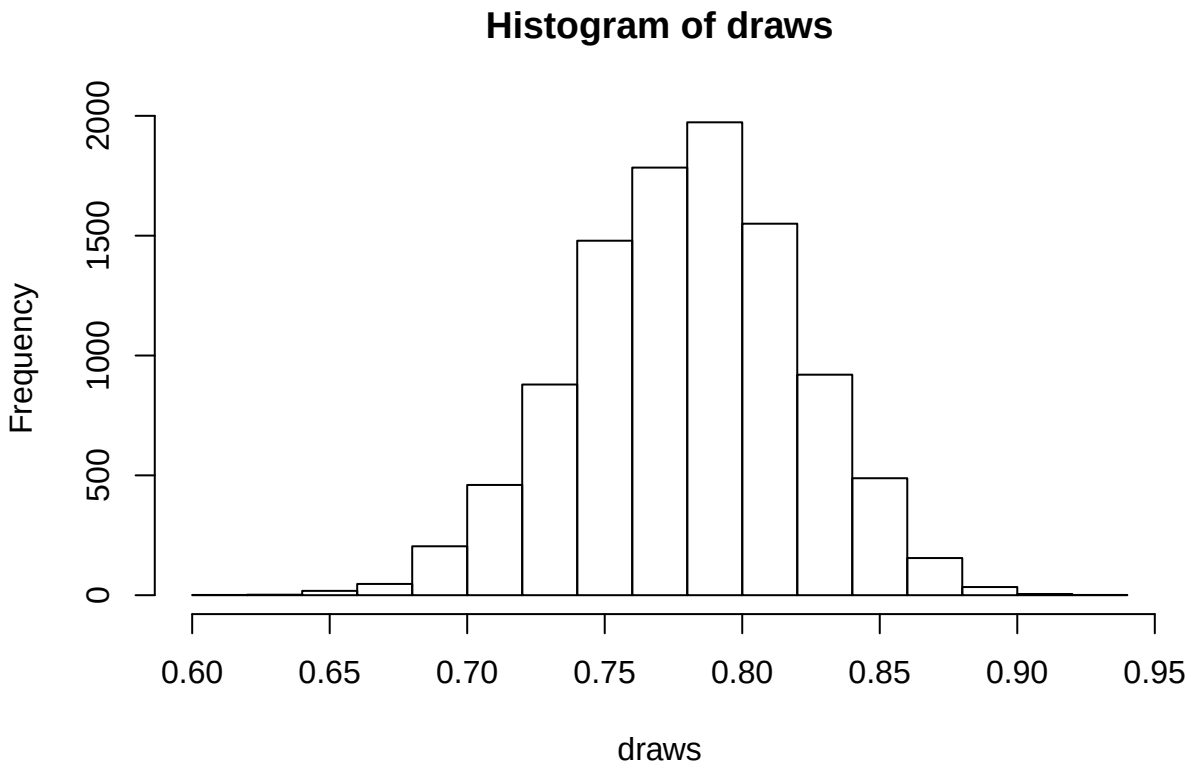
Answer:

```
# Function to get the sample mean of D
sample_D <- function(n) {
  # Get samples for X and Y
  X <- runif(n, min = -1, max = 1)
  Y <- runif(n, min = -1, max = 1)

  # Compute the samples for D based on X and Y.
  D = c()
  for(i in 1:n) {
    if(X[i]^2 + Y[i]^2 < 1) {
      D = c(D,1)
    } else {
      D = c(D,0)
    }
  }
  return(mean(D))
}

draws <- replicate(10000, sample_D(100))

hist(draws)
```



h. Compute the standard deviation of your sample averages to see if it's close to the value you expect from part c.

Answer:

```
sd(draws)
```

```
## [1] 0.04052066
```

In part c, we calculated $\sigma_{\bar{D}} = \frac{\sigma}{\sqrt{n}} = \frac{0.4105}{\sqrt{100}} \approx 0.04105$. The standard deviation of our sample averages is shown above. It is within the ballpark of 0.04105, although, like the sample mean, it is not as close as I would have liked. Again, I think the skew of the Bernoulli distribution requires us to use a larger sample size for the Central Limit Theorem to work better.

i. Compute the fraction of your sample averages that are larger than 3/4 to see if it's close to the value you expect from part d.

```
# Calulate the fraction of my sample averages that are larger than 3/4
length(draws[draws > 3/4])/10000
```

```
## [1] 0.7712
```

In part d, I got the probability that $\bar{D}$ is larger than 3/4 to be 0.8057173. From my sample averages, however, the probability that $\bar{D}$ is larger than 3/4 is quite a bit lower than 0.8057173. I believe the reason for this is that a Bernoulli distribution should ideally have a mean of 0.5. However, our calculated expectation is around 0.785 which is quite a bit larger. This suggests there is a skew to the distribution. It is possible that our sample size of 100 is not quite sufficiently large enough to approximate a normal distribution for the sample averages. I believe if we increase the sample size, our sample probability would be closer to our calculated probability.