

# Lab 2: Probability Theory

W203: Statistics for Data Science

*Tako Hisada*

*10/8/2017*

## 1. Meanwhile, at the Unfair Coin Factory...

You are given a bucket that contains 100 coins. 99 of these are fair coins, but one of them is a trick coin that always comes up heads. You select one coin from this bucket at random. Let  $T$  be the event that you select the trick coin. This means that  $P(T) = 0.01$ .

- a. To see if the coin you have is the trick coin, you flip it  $k$  times. Let  $H_k$  be the event that the coin comes up heads all  $k$  times. If you see this occur, what is the conditional probability that you have the trick coin? In other words, what is  $P(T|H_k)$ .

$$P(T|H_k) = \frac{0.01}{(0.99 * 0.5^k + 0.01 * 1^k)}$$

- b. How many heads in a row would you need to observe in order for the conditional probability that you have the trick coin to be higher than 99%?

$$P(T|H_k) = 0.99 = \frac{0.01 * 1^k}{(0.99 * 0.5^k + 0.01 * 1^k)}$$

$$0.99(0.99 * 0.5^k + 0.01 * 1^k) = 0.01 * 1^k$$

Since  $0.01 * 1^k = 0.01$ , we can say

$$0.99(0.99 * 0.5^k + 0.01) = 0.01$$

$$0.99 * 0.99 * 0.5^k = 0.01 - 0.99 * .01 = 0.0001$$

$$0.5^k = \frac{0.0001}{0.99 * 0.99}$$

$$\log(0.5^k) = \log\left(\frac{0.0001}{0.99 * 0.99}\right)$$

$$\log(0.5^k) = \log(0.0001020304)$$

$$k * \log(0.5) = \log(0.0001020304)$$

$$k = \log(0.0001020304) / \log(0.5) = 13.25871$$

```
(q1b = log(0.0001020304)/log(0.5))
```

```
## [1] 13.25871
```

14 times

## 2. Wise Investments

You invest in two startup companies focused on data science. Thanks to your growing expertise in this area, each company will reach unicorn status (valued at \$1 billion) with probability  $3/4$ , independent of the other company. Let random variable  $X$  be the total number of companies that reach unicorn status.  $X$  can take on the values 0, 1, and 2. Note:  $X$  is what we call a binomial random variable with parameters  $n = 2$  and  $p = 3/4$ .

- a. Give a complete expression for the probability mass function of  $X$ .

$$P(X) = b(x; 2, 0.75) = \binom{2}{x} (.75)^x (.25)^{2-x}$$

- b. Give a complete expression for the cumulative probability function of  $X$ .

$$F(x) = P(X \leq x) = b(x; 2, 0.75) = \sum_0^x b(x; 2, 0.75) = \sum_0^x \binom{2}{x} (.75)^x (.25)^{2-x}$$

- c. Compute  $E(X)$ .

```
n <- 2
p <- 3/4
q2c <- n*p
```

$$E(X) = np = 2 * 0.75 = 1.5$$

- d. Compute  $var(X)$ .

```
n <- 2
p <- 3/4
q2d <- n*p*(1-p)
```

$$V(X) = np(1 - p) = 2 * 0.75(1 - 0.75) = 1.5 * 0.25 = 0.375$$

## 3. Relating Min and Max

Continuous random variables  $X$  and  $Y$  have a joint distribution with probability density function,

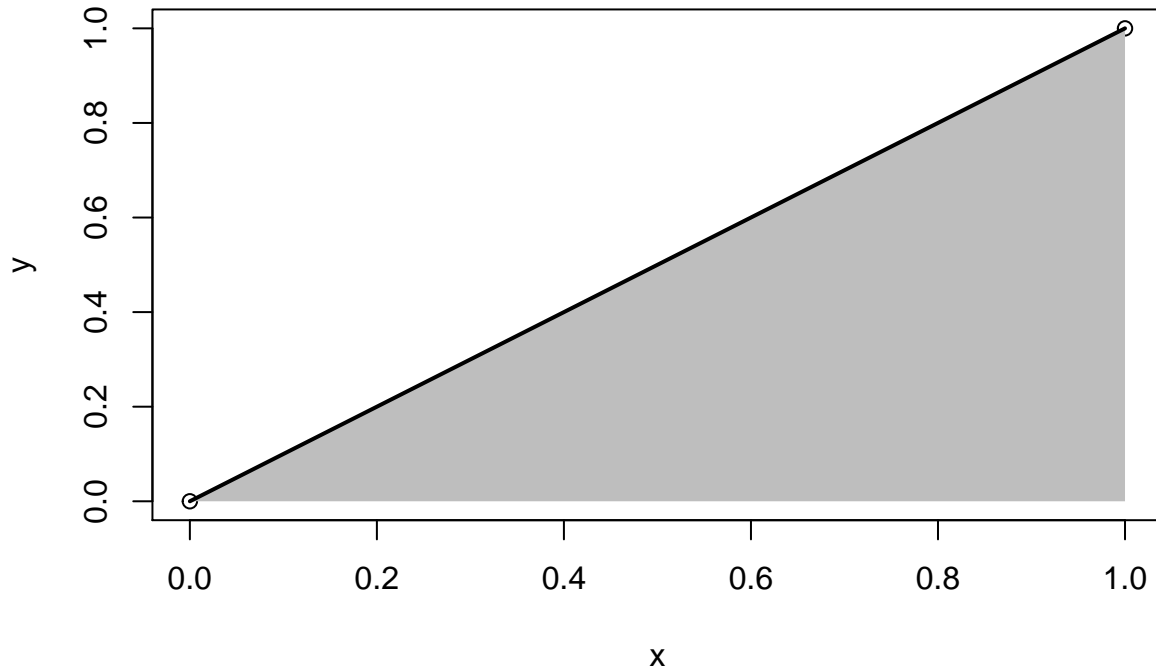
$$f(x, y) = \begin{cases} 2, & 0 < y < x < 1 \\ 0, & \text{otherwise.} \end{cases}$$

You may wonder where you would find such a distribution. In fact, if  $A_1$  and  $A_2$  are independent random variables uniformly distributed on  $[0, 1]$ , and you define  $X = \max(A_1, A_2)$ ,  $Y = \min(A_1, A_2)$ , then  $X$  and  $Y$  will have exactly the joint distribution defined above.

- a. Draw a graph of the region for which  $X$  and  $Y$  have positive probability density.

```
plot(c(0,1), c(0,1), xlab = "x", ylab = "y", main = "Relating Min and Max")
x <- c(0, 1, 1)
y <- c(0, 1, 0)
polygon(x, y, col = "grey", lty = 1, border = 0)
polygon(c(0, 1), c(0, 1), col = "black", lty = 1, lwd = 2)
```

## Relating Min and Max



- b. Derive the marginal probability density function of  $X$ ,  $f_X(x)$ .

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 2dy = 2y|_0^1 = 2$$

- c. Derive the unconditional expectation of  $X$ .

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 x2dx = x^2|_0^1 = 1$$

- d. Derive the conditional probability density function of  $Y$ , conditional on  $X$ ,  $f_{Y|X}(y|x)$

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

- e. Derive the conditional expectation of  $Y$ , conditional on  $X$ ,  $E(Y|X)$ .

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

- f. Derive  $E(XY)$ . Hint: if you take an expectation conditional on  $X$ ,  $X$  is just a constant inside the expectation. This means that  $E(XY|X) = XE(Y|X)$ .

$$E(XY) = E(XY|X) = XE(Y|X) = X * \frac{1}{2} = \frac{1}{2}$$

- g. Using the previous parts, derive  $cov(X, Y)$

$$Cov(X, Y) = E(XY) - \mu_X * \mu_Y = \frac{1}{2} - 1 * \frac{1}{2} = 0$$

## 4. Circles, Random Samples, and the Central Limit Theorem

Let  $X_1, X_2, \dots, X_n$  and  $Y_1, Y_2, \dots, Y_n$  be independent random samples from a uniform distribution on  $[-1, 1]$ . Let  $D_i$  be a random variable that indicates if  $(X_i, Y_i)$  falls within the unit circle centered at the origin. We can define  $D_i$  as follows:

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

Each  $D_i$  is a Bernoulli variable. Furthermore, all  $D_i$  are independent and identically distributed.

- Compute the expectation of each indicator variable,  $E(D_i)$ . Hint: your answer should involve a Greek letter.
- Compute the standard deviation of each  $D_i$ .
- Let  $\bar{D}$  be the sample average of the  $D_i$ . Compute the standard error of  $\bar{D}$ . This should be a function of sample size  $n$ .
- 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.
- Now let  $n = 100$ . Use R to simulate a draw for  $X_1, X_2, \dots, X_n$  and  $Y_1, Y_2, \dots, Y_n$ . Calculate the resulting values for  $D_1, D_2, \dots, D_n$ . 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.

```
plot(x,y, col=d+1, asp=1)
```

- What value do you get for the sample average,  $\bar{D}$ ? How does it compare to your answer for part a?
- 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.
- Compute the standard deviation of your sample averages to see if it's close to the value you expect from part c.
- 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.