ENGRD 2700: Basic Engineering Probability and Statistics
Fall 2019

Homework 5

Due **Friday November 1** at 11:59pm. Submit to Gradescope by clicking the name of the assignment. See `https://people.orie.cornell.edu/yudong.chen/engrd2700_2019fa.html#homework` for detailed submission instructions.

The same stipulations from Homework 1 (e.g., independent work, computer code, etc.) still apply.

1. The joint density function of two random variables is given by

$$f_{X,Y}(x,y) = \begin{cases} k(2y+xy) & \text{if } 0 \le x \le 1 \text{ and } x \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

   (a) What is the constant k?

   We know that for a density function, the integral of it should be one. Therefore, we have

$$\begin{aligned} 1 = \iint f_{X,Y}(x,y)dydx &= \int_0^1 \int_x^1 k(2y+xy)dydx \\ &= k\int_0^1 (1+\frac{x}{2} - x^2 - \frac{1}{2}x^3)dx \\ &= k \cdot \frac{19}{24}, \end{aligned}$$

   from which we know that $k = \frac{24}{19}$.

   (b) What is the marginal pdf of X?

   To obtain the marginal pdf of X, we just have to integrate the joint density function with respect to Y,

$$\begin{aligned} f_X(x) &= \int_x^1 f_{X,Y}(x,y)dy \\ &= \int_x^1 \frac{24}{19}(2y+xy)dy \\ &= \frac{24}{19}(1+\frac{x}{2} - x^2 - \frac{1}{2}x^3), \end{aligned}$$

   Therefore, the marginal distribution of X is:

$$f_X(x) = \begin{cases} \frac{24}{19}(1+\frac{x}{2} - x^2 - \frac{1}{2}x^3) & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

   (c) What is the conditional density of X, given that $Y = 1/4$?

   We know that $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$, thus first of all, we should figure out $f_Y(y)$,

$$f_Y(y) = \int_0^y f_{X,Y}(x,y)dx = \frac{24}{19}y^2(2+\frac{1}{2}y)$$

   So,

$$f_{X|Y}(x|y) = \frac{\frac{24}{19}(2y+xy)}{\frac{24}{19}y^2(2+\frac{1}{2}y)} = \frac{2+x}{y(2+\frac{1}{2}y)}$$

which will hold for $0 \le x \le y$. Plug in $y = \frac{1}{4}$, we get

$$f_{X|Y}(x \,|\, \frac{1}{4}) = \frac{32(2+x)}{17}$$

Therefore, the condition density of X given $Y = \frac{1}{4}$ is :

$$f_{X|Y}(x \,|\, \frac{1}{4}) = \begin{cases} \frac{32(2+x)}{17} & \text{if } 0 \le x \le \frac{1}{4} \\ 0 & \text{otherwise} \end{cases}$$

2. The lifetimes X and Y of two batteries (in years) are distributed according to the following joint PDF:

$$f_{X,Y}(x,y) = \begin{cases} 8e^{-4x}e^{-2y} & \text{if } x \ge 0 \text{ and } y \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) Compute the probability that both batteries last at least 2 years.

$$P(X \ge 2, Y \ge 2) = \int_2^\infty \int_2^\infty 8e^{-4x}e^{-2y}\,dy\,dx = e^{-8} \cdot e^{-4} = 6.14 \times 10^{-6}$$

(b) Find the marginal PDFs $f_X(x)$ and $f_Y(y)$.
To find $f_X(x)$, we integrate out y

$$f_X(x) = \begin{cases} \int_0^\infty 8e^{-4x}e^{-2y}\,dy = 4e^{-4x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

Similarly,

$$f_Y(y) = \begin{cases} \int_0^\infty 8e^{-4x}e^{-2y}\,dx = 2e^{-2y} & y \ge 0 \\ 0 & y < 0 \end{cases}$$

(c) Compute $E[XY]$.
Since $X \sim Exp(4)$ and $Y \sim Exp(2)$, then $E[X] = \frac{1}{4}$, $E[Y] = \frac{1}{2}$.

$$\begin{aligned} E[XY] &= \int_{-\infty}^\infty \int_{-\infty}^\infty xyf(x,y)\,dx\,dy \\ &= \int_0^\infty \int_0^\infty 8xye^{-4x}e^{-2y}\,dx\,dy \\ &= \int_0^\infty 4xe^{-4x}\,dx \int_0^\infty 2ye^{-2y}\,dy \\ &= E[X]E[Y] = \frac{1}{4} \cdot \frac{1}{2} \\ &= \frac{1}{8} \end{aligned}$$

(d) What is $P(X \le \frac{Y}{2})$, the probability that the first battery lasts at most half as long as the second?

$$
\begin{aligned}
P(X \leq \frac{Y}{2}) &= \int_0^{\infty} \int_0^{y/2} 8e^{-4x} e^{-2y}\, dx\, dy \\
&= \int_0^{\infty} 2e^{-2y} \left(e^{-4x} \Big|_{x=y/2}^{x=0} \right) dy \\
&= \int_0^{\infty} 2e^{-2y} (1 - e^{-2y})\, dy \\
&= \int_0^{\infty} \left(2e^{-2y} - 2e^{-4y} \right) dy \\
&= e^{-2y} \Big|_{x=\infty}^{x=0} - \frac{1}{2} e^{-4y} \Big|_{x=\infty}^{x=0} \\
&= 1 - \frac{1}{2} \\
&= \frac{1}{2}
\end{aligned}
$$

(e) Are X and Y independent? Why or why not?

Yes.
We check whether or not $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for all pairs of (x,y).
If $x < 0$ or $y < 0$, then $f_{X,Y}(x,y) = 0 = f_X(x) f_Y(y)$.
If $x \geq 0$ and $y \geq 0$, then $f_{X,Y}(x,y) = 8e^{-4x} e^{-2y} = (4e^{-4x})(2e^{-2y}) = f_X(x) f_Y(y)$.
Therefore, X and Y are independent.

3. For two discrete random variables X and Y, recall the definition of $f_{X|Y}(x|y)$, the *conditional PMF* of X given that $Y = y$. Also recall that conditional PMFs are still PMFs, so they must sum to one; that is, $\sum_x f_{X|Y}(x|y) = 1$.

A bakery sells two types of cupcakes: red velvet and salted caramel. Let R and S denote the number of each type of cupcake an individual customer buys. Suppose that R and S are distributed according to the following joint PMF:

		s		
$f_{R,S}(r,s)$		0	1	2
	0	0	0.25	0.15
r	1	0.20	0.10	0.10
	2	0.10	0.05	0.05

(a) Compute $f_{R|S}(1|1)$, the conditional probability that a customer buys 1 red velvet cupcake, given that (s)he buys 1 salted caramel cupcake. How does this probability compare to $f_R(1)$?

$$f_{R|S}(1|1) = \frac{f_{R,S}(1,1)}{f_S(1)} = \frac{P(R=1, S=1)}{P(S=1)} = \frac{0.1}{0.25 + 0.1 + 0.05} = 0.25$$

We see $f_{R|S}(1|1) = 0.25 < f_R(1) = 0.2 + 0.1 + 0.1 = 0.4$, i.e., given that a customer buys one salted caramel cupcake, the chance of him/her buying one red velvet cupcake decreases.

(b) Find $f_{R|S}(r|2)$, the conditional PMF of R, given that the customer buys 2 salted caramel cupcakes. (This involves computing $f_{R|S}(0|2)$, $f_{R|S}(1|2)$, and $f_{R|S}(2|2)$.)

$$f_{R|S}(r|2) = \begin{cases} \dfrac{f_{R,S}(0,2)}{f_S(2)} = \frac{1}{2} & r = 0 \\ \dfrac{f_{R,S}(1,2)}{f_S(2)} = \frac{1}{3} & r = 1 \\ \dfrac{f_{R,S}(2,2)}{f_S(2)} = \frac{1}{6} & r = 2 \end{cases}$$

(c) The *conditional expectation* of X, given that $Y = y$, is defined as

$$E[X \,|\, Y = y] = \sum_x x\, f_{X|Y}(x|y).$$

Find $E[R \mid S = 2]$, the expected number of red velvet cupcakes a customer buys, given that (s)he buys 2 salted caramel cupcakes.

$$E[R \mid S = 2] = \sum_{r=0}^{2} r\, f_{R|S}(r|2) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{6} = \frac{2}{3}$$

4. Given two continuous random variables X and Y, recall the definition of $f_{X|Y}(x|y)$, the *conditional PDF* of X given that $Y = y$. Conditional PDFs are still PDFs, so they must *integrate* to one: $\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx = 1$.

   Suppose X and Y are distributed as in the following joint PDF:

$$f_{X,Y}(x,\, y) = \begin{cases} 2ye^{-y(2+x)} & x,\, y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

   For $y > 0$, compute $f_{X|Y}(x|y)$.

   We first find the marginal with respect to Y. Given $y > 0$,

$$\begin{aligned} f_Y(y) &= \int_0^{\infty} 2ye^{-y(2+x)}\, dx \\ &= 2e^{-2y} \int_0^{\infty} ye^{-yx}\, dx \\ &= 2e^{-2y} \cdot \left(-e^{-yx} \Big|_{x=0}^{x=\infty} \right) \\ &= 2e^{-2y}. \end{aligned}$$

   Since $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$, we have that

$$f_{X|Y}(x|y) = \begin{cases} ye^{-yx} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

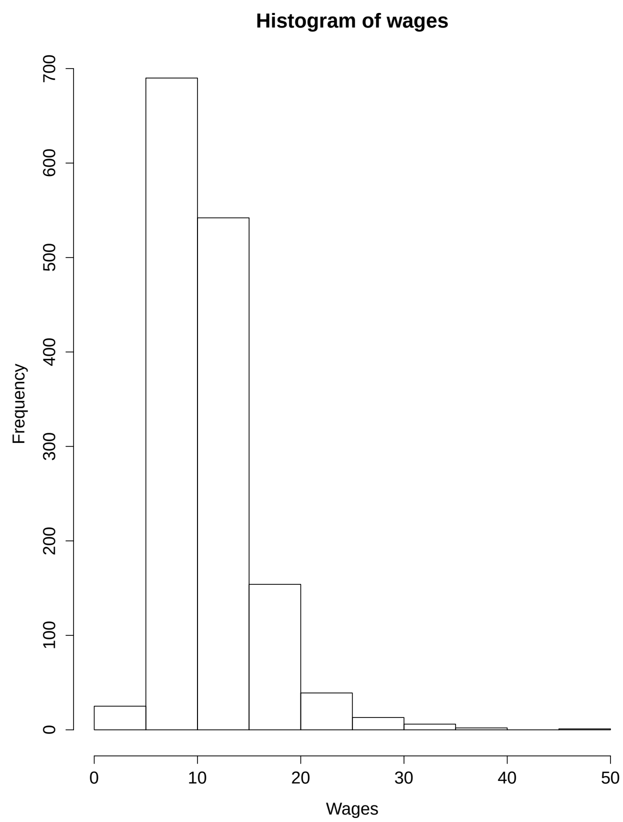
5. The file `Bwages.csv` contains hourly wages of 1473 randomly selected individuals living in Belgium in 1994. For the questions that follow, attach your plots, as well as any code used to generate them.

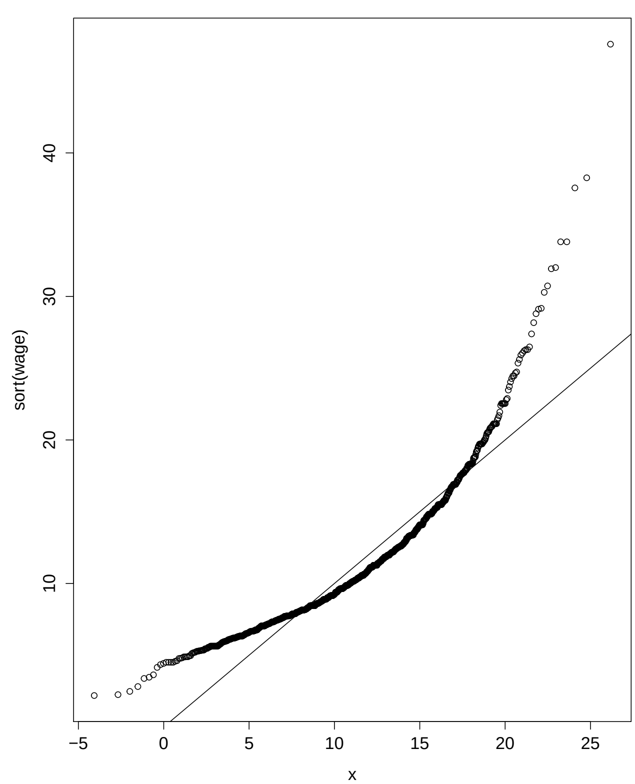
   (a) Import this dataset into RStudio, and generate a histogram of the data.

   The histogram is shown as above.
   Code:
```
Bwages = read.csv("Bwages.csv")
wage = Bwages[,1]
hist(wage,xlab = "Wages",main="Histogram of wages")
```

   (b) Construct a normal Q–Q plot, by hypothesizing that the data originate from a Normal($\bar{x}$, s^2) distribution, where $\bar{x}$ and s^2 are the sample mean and sample variance, respectively. Overlay the line $y = x$ onto your plot. (See Recitation 6 for an example of how to do this.)
   The Q-Q plot is shown in Figure. 2.
   Code:
```
n = length(wage)
quantiles = (1:n - 0.5)/n
mu = mean(wage)
sd = sd(wage)
x = qnorm(quantiles,mu,sd)
plot(x, sort(wage))
abline(0, 1)
```

   (c) Does the normal distribution appear to be a reasonable fit for the data? Why or why not?
   No. The fit is especially poor on the left because normal distribution can take on negative values (but wages can't), and that the fit is poor on the right because the data are much more likely to take on large values than the normal distribution would suggest.



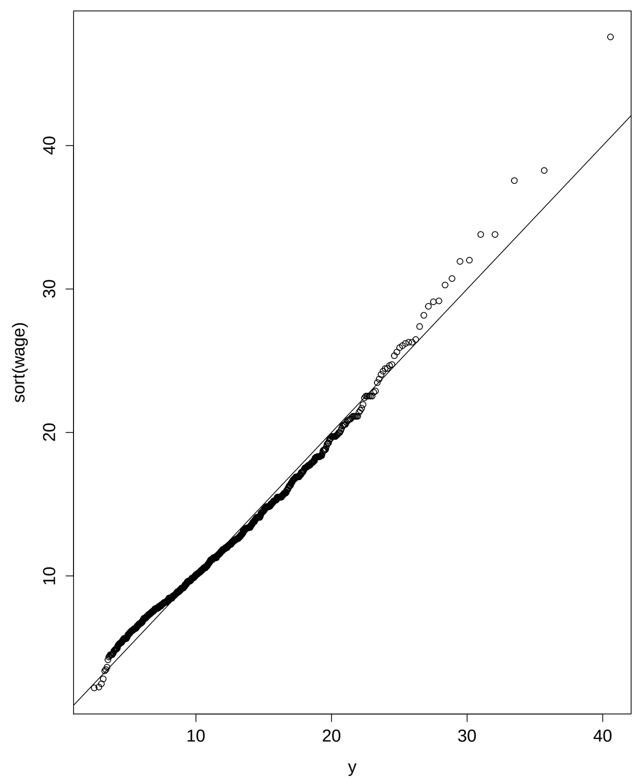


(d) We'll now attempt to fit a *lognormal distribution* to the data. If $X \sim \text{Lognormal}(\mu, \sigma^2)$, it has

PDF

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln(x)-\mu)^2/(2\sigma^2)} \qquad x > 0$$

The lognormal random variable gets its name because $\ln(X)$ is normally distributed (with mean μ and variance σ^2). Construct a lognormal Q–Q plot, by hypothesizing that the data originate from the Lognormal(2.31, 0.41) distribution. Use R's `qlnorm` function to compute theoretical quantiles, and set `meanlog` to 2.31 and `sdlog` to 0.41. Overlay the line $y = x$ onto your plot.



The Q-Q plot is shown in Figure. 3.

Code:
```
meanlog = 2.31
sdlog = 0.41
y = qlnorm(quantiles,meanlog,sdlog)
plot(y, sort(wage))
abline(0, 1)
```

(e) Does the lognormal distribution appear to be a better fit for the data? Where could the fit be improved? Comment on what you see.

Yes, the lognormal distribution provides a better fit for the data. This is mainly because lognormal distribution cannot take negative values and has a heavier tail than normal distribution. However, the fit isn't very good for larger values in the dataset since the tail of lognormal distribution is not heavy enough.