18.05 Problem Set 3, Spring 2022 Solutions

Problem 1. (25: 10,5,10 pts.) **Independence**
Three events A, B, and C are pairwise independent *if each pair is independent. They are* mutually independent *if they are pairwise independent and if, in addition,*

$$P(A \cap B \cap C) = P(A)P(B)P(C). \qquad (1)$$

(a) *Suppose we roll two 6-sided die. Consider the events:*

$$A = \text{'odd on die 1'} \qquad B = \text{'odd on die 2'} \qquad C = \text{'odd sum'}$$

Are A, B, and C pairwise independent? Are they mutually independent?

Solution: We have $P(A) = P(B) = P(C) = 1/2$. Writing the outcome of die 1 first, we can easily list all outcomes in the following intersections.

$$A \cap B = \{(1,1),(1,3),(1,5),(3,1),(3,3),(3,5),(5,1),(5,3),(5,5)\}$$
$$A \cap C = \{(1,2),(1,4),(1,6),(3,2),(3,4),(3,6),(5,2),(5,4),(5,6)\}$$
$$B \cap C = \{(2,1),(4,1),(6,1),(2,3),(4,3),(6,3),(2,5),(4,5),(6,5)\}$$

By counting we see

$$P(A \cap B) = \frac{1}{4} = P(A)P(B).$$

Likewise,

$$P(A \cap C) = \frac{1}{4} = P(A)P(C) \quad \text{and} \quad P(B \cap C) = \frac{1}{4} = P(B)P(C).$$

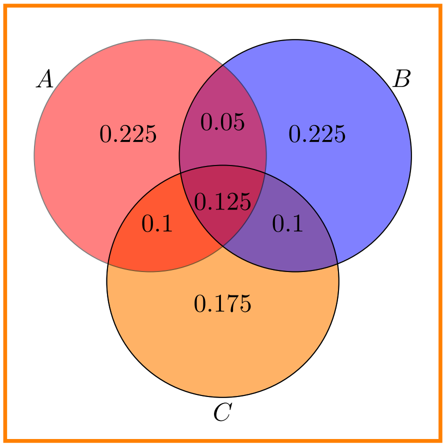
So, we see that $A, B,$ and C are pairwise independent.

However, $A \cap B \cap C = \emptyset$, since if we roll an odd on die 1 and an odd on die 2, then the sum of the two will be even. So, in this case,

$$P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C),$$

and we conclude that A, B and C are not mutually independent.

(b) *Consider the Venn diagram below. A, B and C are the overlapping circles and the probabilities of each region are as marked. Does equation (1) hold. Are the events A, B, C mutually independent?*



Solution: We start by totaling the probabilities in the regions shown to get the following probabilities.

$$\begin{aligned}
&P(A) = 0.225 + 0.05 + 0.1 + 0.125 = 0.5, && P(B) = 0.225 + 0.05 + 0.1 + 0.125 = 0.5\\
&P(C) = 0.175 + 0.1 + 0.1 + 0.125 = 0.5, && P(A \cap B) = 0.05 + 0.125 = 0.175\\
&P(A \cap C) = 0.1 + 0.125 = 0.225, && P(B \cap C) = 0.1 + 0.125 = 0.225\\
&P(A \cap B \cap C) = 0.125
\end{aligned}$$

We see that $P(A)P(B)P(C) = 0.5 \cdot 0.5 \cdot 0.5 = 0.125 = P(A \cap B \cap C)$. So, yes, equation (1) holds.

But, $P(A)P(B) = 0.5{\cdot}0.5 = 0.25 \neq P(A{\cap}B)$. Since, mutual independence requires pairwise independence as well as the multiplication formula for all three events, we see that the three events are not independent. (Likewise $P(A)P(C) \neq P(A{\cap}C)$ and $P(B)P(C) \neq P(B{\cap}C)$.)

(c) *Consider a litter of n puppies. What value(s) of n makes the events 'the litter has puppies of both sexes' and 'there is at most one female' independent.*

Solution: Let A be the event "the litter has puppies of both sexes" and B be the event "there is at most one female." In order for A to ever be true, we first assume that $n > 1$.

Now, if we let X be the number of female puppies then we have

$$P(A) = P(1 \leq X \leq n-1), \qquad P(B) = P(X \leq 1), \qquad P(A \cap B) = P(X = 1)$$

Since $X \sim \text{binomial}(n, 1/2)$ we have: $P(X = 0) = \dfrac{1}{2^n}$ and $P(X = 1) = \dbinom{n}{1}\dfrac{1}{2^n} = \dfrac{n}{2^n}$. So,

$$\begin{aligned}
P(A) &= 1 - P(X = 0) - P(X = n) = 1 - \frac{2}{2^n}\\
P(B) &= P(X = 0) + P(X = 1) = \frac{n+1}{2^n}\\
P(A \cap B) &= P(X = 1) = \frac{n}{2^n}.
\end{aligned}$$

Since we are told that A and B are independent, we must have $P(A)P(B) = P(A \cap B)$. This implies

$$\left(\frac{n+1}{2^n}\right)\left(1 - \frac{2}{2^n}\right) = \frac{n}{2^n}$$

Some algebra yields

$$(n+1)\left(1 - \frac{2}{2^n}\right) = n \Leftrightarrow n + 1 - \frac{n+1}{2^{n-1}} = n \Leftrightarrow 2^{n-1} = n + 1$$

Plugging small values of n into the above equation, we find that the two events are independent when $\boxed{n = 3.}$

If $n = 1$ then $P(A) = 0$, so A and B are (vacuously) independent, i.e. $n = 1$ is technically also a solution.

Problem 2. (25: 5,5,10,5 pts.) **What does the data say?**
Suppose there is an experimental medical treatment for a cancer that, if untreated, is nearly always fatal within 12-15 months. The doctors enroll 5000 patients in a study in which

each patient is given the treatment and followed for 5 years. Let X be the length of time a random patient given the treatment survives. (If a patient is still alive at the end of the study, then $X = 5$ for this patient.)

As the statistician it is your job to analyze the data.

To load the data into R you should do the following:

1. Download the data `mit18_05_s22_ps3prob2-data.r`*. You can find this on our course website on the page with R code.*
2. Put this file in your 18.05 R directory.
3. In R studio make sure the working directory is set to your 18.05 R directory.
4. Run the commands:

```
> source('mit18_05_s22_ps3prob2-data.r')
> x = get_prob2_data()
```

The variable `x` *should now hold an array of the 5000 data points.*

(a) *Use R to compute the mean, variance and standard deviation of the data.*

(b) *Use the* `hist` *command to get R to plot a frequency histogram of the data. Set the histogram so each bin has width 0.1 years. Print the histogram and turn it in with the pset. The* `hist()` *command was introduced in Studio 3. There is also a short tutorial on using R to plot histograms on our class R page.*

(c) *Using your answers in (a) and (b), write a short paragraph summarizing the data in a useful way.*

(d) *Based on the (c), what are your conclusions about the effectiveness of the treatment? What recommendations would you make for avenues of further research?*

Solution: Here is the R code for both (a) and (b).

```
# 2a: Compute statistics
source('mit18_05_s22_ps3prob2-data.r')
years = get_prob2_data()
m2a = mean(years)
v2a = var(years)
sd2a = sqrt(var(years))

cat('2(a) mean =', m2a, ', var = ', v2a, ', std. dev =', sd2a, '\n' )

# 2(b): Make histogram. (Export image from R Studio to save)
hist(years, freq=T, breaks=seq(0, 5, 0.1), main= "Years of Survival", col="yellow")
```

(a) The results are:

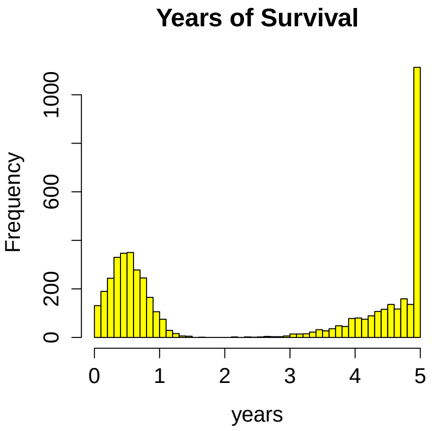
$$\text{Mean} \approx 2.554528, \quad \text{variance} \approx 4.3018, \quad \text{std. dev} \approx 2.074.$$

Note: we used the R function `var()` to compute the variance. This uses the following formula for the variance of n points of data:

$$\frac{\sum_{j=1}^{5000}(x_j - \mu)^2}{4999}.$$

We will learn the reason for dividing by $4999 = 5000 - 1$ instead of 5000 when we do the statistics portion of the class. In this case, there is very little difference between dividing by 5000 or 4999.

(b) Here is the histogram produced by the abouve code



(c) Looking at the distribution we see it is bimodal with a spike at 5 years. About half the patients die in the first year but about half live more than 2.5 years with over 20% still alive after 5 years. The spike is because everyone who survives to 5 years is lumped into that category. The average of 2.5 years is not that meaningful because there seem to be two categories of patients. This is reflected in the large standard deviation.

(d) The fact that the disease is almost always fatal in 12-15 months gives us an implicit control group. So, the treatment appears to be effective for about half the patients. More research would be needed to understand what characteristics of the disease or patients predict the treatment will be effective.

Problem 3. (30: 10,10,10 pts.) **Dice**
Let X be the result of rolling a fair 4-sided die. Let Y be the result of rolling a fair 6-sided die. Let Z be the average of X and Y.

(a) *Find the standard deviation of X, of Y, and of Z.*

Solution: We compute $\text{Var}(X) = E[X^2] - E[X]^2$ etc. from the tables.

X	1	2	3	4
$p(x)$	1/4	1/4	1/4	1/4
X^2	1	4	9	16

Y	1	2	3	4	5	6
$p(y)$	1/6	1/6	1/6	1/6	1/6	1/6
Y^2	1	4	9	16	25	36

So, $E[X] = \frac{1}{4}(1+2+3+4) = \frac{5}{2}$, $E[X^2] = \frac{1}{4}(1+4+9+16) = \frac{15}{2}$. Thus,

$$\text{Var}(X) = E[X^2] - E[X]^2 = 5/4.$$

Similarly, $E[Y] = \frac{7}{2}$, $E[Y^2] = \frac{91}{6}$. So, $\text{Var}(Y) = \frac{35}{12}$.

Since X and Y are independent,

$$\text{Var}(Z) = \text{Var}\left(\frac{X+Y}{2}\right) = \frac{1}{4}\left(\text{Var}(X) + \text{Var}(Y)\right) = \frac{25}{24}.$$

Taking the square root of the variances we get:

$$\sigma_X = \sqrt{5}/2 \approx 1.118 \qquad \sigma_Y = \sqrt{35/12} \approx 1.708 \qquad \sigma_Z = \sqrt{25/24} \approx 1.021$$

(b) *Carefully graph the pmf and cdf of Z.*

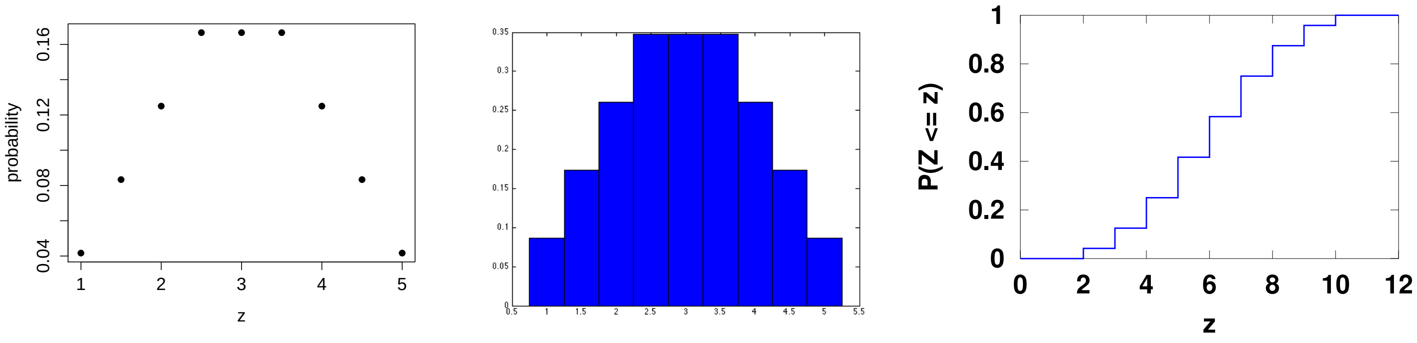
Solution: To compute the pmf of Z we need all the ways X and Y can be averaged.

$$\begin{aligned}
Z &= 1 && \longleftrightarrow (X,Y) = \{(1,1)\} \\
Z &= 3/2 && \longleftrightarrow (X,Y) = \{(1,2),\ (2,1)\} \\
Z &= 2 && \longleftrightarrow (X,Y) = \{(1,3),\ (2,2),\ (3,1)\} \\
Z &= 5/2 && \longleftrightarrow (X,Y) = \{(1,4),\ (2,3),\ (3,2),\ (4,1)\} \\
Z &= 3 && \longleftrightarrow (X,Y) = \{(1,5),\ (2,4),\ (3,3),\ (4,2)\} \\
Z &= 7/2 && \longleftrightarrow (X,Y) = \{(1,6),\ (2,5),\ (3,4),\ (4,3)\} \\
Z &= 4 && \longleftrightarrow (X,Y) = \{(2,6),\ (3,5),\ (4,4)\} \\
Z &= 9/2 && \longleftrightarrow (X,Y) = \{(3,6),\ (4,5)\} \\
Z &= 5 && \longleftrightarrow (X,Y) = \{(4,6)\}
\end{aligned}$$

Since each (X,Y) pair has probability 1/24, the pmf and cdf of Z are

Z	1	3/2	2	5/2	3	7/2	4	9/2	5
$p(z)$	1/24	2/24	3/24	4/24	4/24	4/24	3/24	2/24	1/24
$F_Z(z)$	1/24	3/24	6/24	10/24	14/24	18/24	21/24	23/24	24/24

We graph the pmf of Z as point plot and then as a density histogram. The cdf is a staircase graph.



(c) *Here is a gambling game: You win $2X$ dollars if $X > Y$ and lose 1 dollar otherwise. After playing this game 60 times, what is your expected total gain (positive) or loss (negative)?*

Solution: We see that the only pairs of (X,Y) which satisfy $X > Y$ are $\{(2,1),(3,1),(3,2),(4,1),(4,2),(4,3)\}$. So $P(X > Y) = \frac{6}{24}$. Let W be the amount won in one gaime. Again, since each (X,Y) pair has probability 1/24, W has the probability table

W	4	6	8	-1
$p(w)$	1/24	2/24	3/24	18/24

So $E[W] = 4 \cdot \dfrac{1}{24} + 6 \cdot \dfrac{2}{24} + 8 \cdot \dfrac{3}{24} - 1 \cdot \dfrac{18}{24} = \dfrac{22}{24} = \dfrac{11}{12}$

Now if you played the game 60 times, and received winnings $W_1, \ldots, W_{60}$, (with $E[W_i] = \frac{11}{12}$), your expected total gain is

$$E[W_1 + \cdots + W_{60}] = E[W_1] + \cdots + E[W_{60}] = 60 \cdot \frac{11}{12} = 55.$$

Problem 4. (20: 5,5,5,5 pts.) **Two scoops**
Boxes of Raisin Bran cereal are 30cm tall. Due to settling, boxes have a higher density of raisins at the bottom ($h = 0$) than at the top ($h = 30$). Suppose the density (in raisins per cm of height) is given by $f(h) = 40 - h$.

(a) *How many raisins are in a box?*

Solution: The number of raisins is

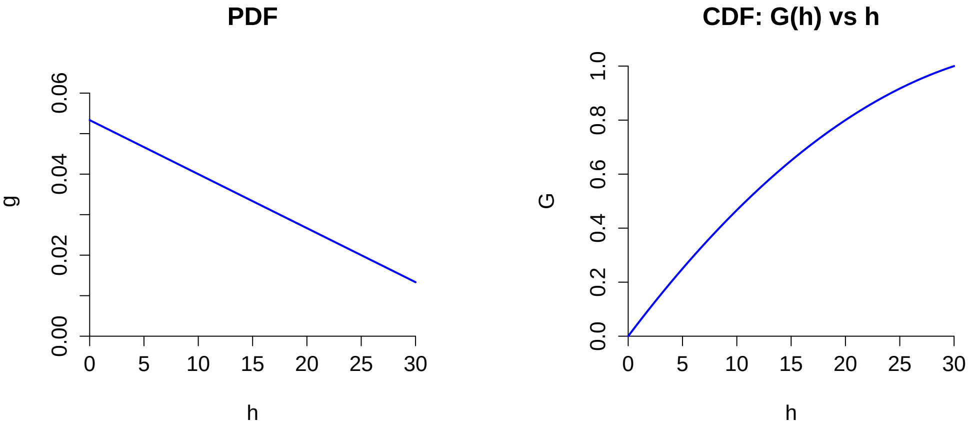
$$\int_0^{30} f(h)dh = \int_0^{30} (40 - h)dh = 750$$

(b) *Let H be the height of a random raisin. Find and graph the pdf $g(h)$ of H.*

Solution: The probability density is just the actual density divided by the total number of raisins. $g(h) = \frac{1}{750}(40 - h)$.

(c) *Find and graph the cdf $G(h)$ of H.*

Solution: For $0 \leq h \leq 30$ we have $G(h) = \displaystyle\int_0^h g(x)\,dx = \frac{40h}{750} - \frac{h^2}{1500}$.



The R code for these plots is posted in mit18_05_s22_ps3_sol.r

(d) *What is the probability that a random raisin is in the bottom third of the box?*

Solution: Since the height is 30 we need to find $P(H \leq 10)$.

$$P(H \leq 10) = \int_0^{10} g(h)dh = \frac{1}{750}\int_0^{10} (40 - h)dh = \frac{7}{15}. \quad \text{Or, } P(H \leq 10) = G(10) = \frac{400}{750} - \frac{100}{1500} = \frac{7}{15}.$$

Problem 5. (20: 5,5,10 pts.) **Gallery of continuous random variables.**
The `pnorm()` function on R gives the cdf of the normal distribution, e,g, if $X \sim N(\mu, \sigma^2)$ then $\texttt{pnorm}(x, \mu, \sigma) = P(X \leq x) = F_X(x)$.

(a) *Suppose Z is a standard normal random variable. Use R to compute*

(i) $P(Z \leq 0)$, (ii) $P(Z > 1.5)$ (iii) $P(|Z| < 1.5)$.

Solution: (i) You don't need R for this: $P(Z \leq 0) = 0.5$.

(ii) $P(Z > 1.5) =$ `1 - pnorm(1.5)` $= 0.0668072$.

(iii) $P(|Z| \leq 1.5) =$ `pnorm(1.5) - pnorm(-1.5)` $= 0.8663856$.

(b) *Let $X \sim N(\mu, \sigma^2)$ where $\mu = 2$ and $\sigma = 3$. Use R to compute*

(i) $P(X \leq \mu)$, (ii) $P(X - \mu > 1.5\sigma)$ (iii) $P(|X - \mu| < 1.5\sigma)$.

Solution: The trick you were supposed to notice is that these answers are identical to the ones in part (a). This is because $Z = (X - \mu)/\sigma$ is standard normal. In the R code below `mu = 2, sigma = 3`

(i) $P(X \leq \mu) =$ `pnorm(mu, mu, sigma)` $= 0.5$.

(ii) $P(X - \mu > 1.5 * \sigma) = P(X > \mu + 1.5 * \sigma) =$ `1 - pnorm(mu + 1.5*sigma, mu, sigma)` $= 0.0668072$.

(iii) $P(|X - \mu| < 1.5 * \sigma) = P(X < \mu + 1.5 * \sigma) - P(X < \mu - 1.5\sigma) =$ `pnorm(mu + 1.5*sigma, mu, sigma) - pnorm(mu - 1.5*sigma, mu, sigma)` $= 0.8663856$.

(c) *Let $Y \sim exp(\lambda)$. Compute the cdf of Y by integrating the pdf. What is the probability $Y \leq 1/\lambda$? (You need to do an integration, but you can check your work numerically using the `pexp()` function in R.)*

Solution: The pdf is $f_Y(y) = \lambda \mathrm{e}^{-\lambda y}$. The range of Y starts at 0. We compute the cdf by integration

$$F_Y(a) = P(Y \leq a) = \int_0^a \lambda \mathrm{e}^{-\lambda y}\, dy = 1 - \mathrm{e}^{-\lambda a}.$$

We use this to compute the probability asked for:

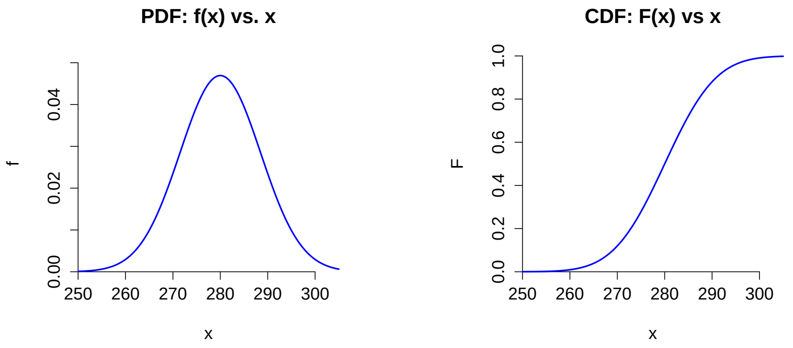
$$P(Y \leq 1/\lambda) = F_Y(1/\lambda) = 1 - \mathrm{e}^{-1} \approx 0.6321206.$$

Problem 6. (20: 5,5,5,5 pts.) **Birth day**

The length of human gestation is well-approximated by a normal distribution with mean $\mu = 280$ days and standard deviation $\sigma = 8.5$ days.

(a) *Graph the corresponding pdf and cdf. You should do this using the dnorm, pnorm and plot commands in R. Print the results and turn them in with the pset.*

Solution: Suppose $Y \sim N(280, 8.5)$. The pdf, $f(y)$ and cdf $F(y)$ are plotted below.



(b) *Suppose your final exam is scheduled for May 18 and your pregnant professor has a due date of May 25. Find the probability she will give birth on or before the day of the final.*

Solution: There is some ambiguity here depending on the exact time of day of the due date. In order to get whole numbers, we'll assume the due date is at the same hour on the 25th as the final is on the 18th. So the final is exactly 7 days before the due date. (We'll accept any number between 6 and 8.)

Let X be the number of days before or after May 25 that the baby is born. We want the probability $X \leq -7$ We know $X \sim N(0, 8.5)$.

$$P(X \leq -7) = \texttt{pnorm(-7, 0, 8.5)} = 0.205$$

(Or we could have computed $P(X \leq -7) = P(Z \leq -\frac{7}{8.5}) =$ `pnorm(-7/8.5, 0, 1)` $\approx$ 0.205., where Z is standard normal.)

(c) *Find the probability she will give birth in May sometime after the exam. (Assume this means from the start of May 19 to the end of May 31.)*

Solution: We want the probability that the baby is born between May 19 ($X = -6$) and May 31 ($X = 6$). We compute

$$P(-6 \leq X \leq 6) = P\left(-\frac{6}{8.5} \leq Z \leq \frac{6}{8.5}\right) \approx 0.520$$

Again there is some ambiguity about the range. We'll accept any reasonable choice.

(d) *The professor decides to move up the exam date so there will be a 95% probability that she will give birth afterward. What date should she pick?*

Solution: We want to find x such that $P(X \geq x) = 0.95$. That is, we want

$$P(Z \geq \frac{x}{8.5}) = 0.95.$$

This says that $x/8.5$ must be the 0.05 quantile for Z.

Using R: `x = 8.5*qnorm(0.05)`, we find $x \approx -14$ (May 11).

Note: we could have skipped standardizing and computed this with `qnorm(0.05, 0, 8.5)`