18.05 Problem Set 4, Spring 2022 Solutions

Problem 1. (25: 5,5,10,5 pts.) **Time to failure**
Recall that an exponential random variable $X \sim exp(\lambda)$ has pdf given by $f(x) = \lambda e^{-\lambda x}$ on $x \geq 0$.

(a) *Compute $P(X \geq x)$.*

Solution:

$$P(X \geq x) = 1 - P(X < x) = 1 - \int_0^x \lambda \mathrm{e}^{-\lambda x} dx = 1 - (1 - \mathrm{e}^{-\lambda x}) = \mathrm{e}^{-\lambda x}.$$

(b) *Compute the mean and standard deviation of X. You need to set up the necessary integrals, but you can use Wolfram Alpha or another application to do the computation. (Of course, it will be good for you if you compute the integrals by hand!)*

Solution: First we compute the mean

$$E[X] = \int_0^\infty x f(x)\, dx = \int_0^\infty \lambda x \mathrm{e}^{-\lambda x}\, dx = -x\mathrm{e}^{-\lambda x} - \frac{\mathrm{e}^{-\lambda x}}{\lambda}\Big|_0^\infty = \boxed{\frac{1}{\lambda}.}$$

For the variance, we use the formula $\mathrm{Var}(X) = E[X^2] - E[X]^2$.

$$E[X^2] = \int_0^\infty x^2 f(x)\, dx = \int_0^\infty \lambda x^2 \mathrm{e}^{-\lambda x}\, dx = -x^2\mathrm{e}^{-\lambda x} - \frac{2x\mathrm{e}^{-\lambda x}}{\lambda} - \frac{2\mathrm{e}^{-\lambda x}}{\lambda^2}\Big|_0^\infty = \frac{2}{\lambda^2}$$

So, $\mathrm{Var}(X) = \dfrac{2}{\lambda^2} - \dfrac{1}{\lambda^2} = \dfrac{1}{\lambda^2}$. So, standard deviation $\boxed{\sigma = \dfrac{1}{\lambda}.}$

(c) *Suppose that X_1 and X_2 are independent $exp(\lambda)$ random variables. Let $T = min(X_1, X_2)$. Find the cdf and pdf of T. (Hint: first find a formula for $P(T \geq t)$?)*

Note: for independent continuous random variables X_1, X_2, you can assume the following formula:

$$P(X_1 \geq x_1,\, X_2 \geq x_2) = P(X_1 \geq x_1)P(X_2 \geq x_2).$$

Solution: For $t \geq 0$, we know that $T \geq t$ if and only if both $X_1 \geq t$ and $X_2 \geq t$. So $P(T \geq t) = P(X_1 \geq t, X_2 \geq t)$. Since X_1 and X_2 are independent, using part (a) we get,

$$P(X_1 \geq t, X_2 \geq t) = P(X_1 \geq t)P(X_2 \geq t) = e^{-2\lambda t}.$$

Thus, $F_T(t) = P(T \leq t) = 1 - e^{-2\lambda t}$. Differentiating with respect to t to get the pdf, we find

$$f_T(t) = F_T'(t) = 2\lambda e^{-2\lambda t}.$$

That is, T is an exponential random variable with mean $\frac{1}{2\lambda}$.

(d) *Suppose we are testing 3 different brands of light bulbs B_1, B_2, and B_3 whose lifetimes are exponential random variables with mean 1/2, 1/3, and 1/5 years, respectively. Assuming*

that all of the bulbs are independent, what is the expected time before one of the bulb fails. (Hint: part (c) was a warmup for this problem.)

Solution: Let X_1, X_2, and X_3 be the lifetimes of bulbs B_1, B_2 and B_3, respectively. Then we know $X_1 \sim \exp(2), X_2 \sim \exp(3), X_3 \sim \exp(5)$. Let $T = \min(X_1, X_2, X_3)$. Then T is the time to the first failure of a bulb. Following the same argument as in (c), we have

$$P(T \geq t) = P(X_1 \geq t)P(X_2 \geq t)P(X_3 \geq t) = \mathrm{e}^{-10t}.$$

Thus, the cdf of T is $F_T(t) = 1 - e^{-10t}$ and the pdf, $f_T(t)$ is given by

$$f_T(t) = F_T'(t) = 10e^{-10t}.$$

We found that $T \sim \exp(10)$. Therefore, $E[T] = \frac{1}{10}$.

Problem 2. (20: 10,10 pts.) **Elections**
To head the newly formed US Dept. of Statistics, suppose that 50% of the population supports Alessandre, 20% supports Sarah, and the rest are split between Gabriel, Sarah and So Hee. A poll asks 400 random people who they support.

(a) *Use the central limit theorem to estimate the probability that at least 52.5% of those polled prefer Alessandre?*

Solution: Let X_i be the result of polling person i:

$$X_i = \begin{cases} 1 & \text{if person } i \text{ supports Alessandre} \\ 0 & \text{if person } i \text{ does not support Alessandre} \end{cases}$$

Then $X_i \sim \text{Bern}(0.5)$ and the number of people who prefer Alessandre is

$$S = X_1 + \cdots + X_{400}.$$

We know $E[X_i] = 1/2$ and $\text{Var}(X_i) = 1/4$. This implies $E[S] = 200$ and $\text{Var}(S) = 100$. Thus, the central limit theorem tells us that

$$S \approx \text{N}(200, 100).$$

The problem asks for $P(S > 210)$:

$$P(S > 210) = P\left(\frac{S - 200}{10} > \frac{210 - 200}{10}\right) \approx P(Z > 1) \approx \boxed{0.16}.$$

(b) *Use the central limit theorem to estimate the probability that less than 31% of those polled prefer Gabriel, Sarah or So Hee?*

Solution: Now let $Y_i = 1$ if person i prefers one of Gabriel, Sarah or So Hee and 0 otherwise. We have $Y_1, \ldots, Y_{400}$ are independent Bern(0.3). So $E[Y_i] = \mu = 0.3$ and $\text{Var}(Y_i) = (0.3)(0.7) = 0.21$. If $\overline{Y} = \frac{1}{400}(Y_1 + \cdots + Y_{400})$, the Central Limit Theorem tells us

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{400}} = \frac{(\overline{Y} - 0.3)\sqrt{400}}{\sqrt{0.21}}$$

is approximately standard normal. Thus, using Z for a standard normal random variable,

$$P(\overline{Y} \leq 0.31) \approx P\left(Z < \frac{(0.31-0.3)\sqrt{400}}{\sqrt{0.21}}\right) = P(Z < 0.4364358) \approx 0.67.$$

Problem 3. (10 pts.) **A penny for your thoughts**
To save a mint, in 2012 Canada decided to do away with its pennies. The Chubby Chef in Equality, Illinois wants to be ready should the United States decide to pass a similar law. The Chubby Chef processes $n = 1000$ orders of assorted baked goods each day, and will round the price of each order to the nearest nickel (e.g., \$3.57 rounds to \$3.55 while \$3.58 rounds to \$3.60). Let p be the probability that the total rounding error over the course of a day is either greater than 100 or less than -100 cents, i.e. exceeds 100 in absolute value.

Estimate p using the central limit theorem.

Solution: Let S be the total rounding error for a day. The problems asks for

$$P(|S| > 100).$$

Let X_i be the rounding error (in cents) of the i^{th} order. Then X_i takes values $-2, -1, 0, 1, 2$, each with probability $\frac{1}{5}$. We compute

$$E[X_i] = \mu = 0, \quad \text{Var}(X_i) = \sigma^2 = 2.$$

The total rounding error $S = X_1 + \cdots + X_{1000}$. By the Central Limit Theorem, we know that $S \approx \text{N}(0, 2000)$.

$$P(|S| \geq 100) = P\left(\frac{|S-0|}{\sqrt{2000}} \geq \frac{100-0}{\sqrt{2000}}\right) \approx P\left(|Z| \geq \frac{100}{\sqrt{2000}}\right) = \boxed{0.0253}.$$

Extra credit 5 points *Simulate this in R with 10000 trials. (Each trial involves 1000 orders.) Print out or hand copy your code and include it. Give the result of running your code 3 times,*

Solution: Here's my code.

```
ntrials = 10000
n_orders = 1000
threshold = 100
cnt_above_threshold = 0
for (j in 1:ntrials) {
  # The rounding error is 0, -1, -2, 2, 1 depending on if the price modulo 5
is 0, 1, 2, 3, 4
  # Generate 1000 random orders, just keep the rounding error
  x = sample(c(0,-1,-2,2,1), n_orders, replace=TRUE)
  total_error = sum(x)
  if (abs(total_error) > threshold) {
    cnt_above_threshold = cnt_above_threshold + 1
  }
}
```

```
prob_above_thresh = cnt_above_threshold/ntrials

print(prob_above_thresh)
```

In three runs it gave 0.0245, 0.0263, 0.0216. This agrees nicely with the CLT estimate.

Problem 4. (30: 10,10,10 pts.) **Change of scale.**
In this problem we will look at scaling random variables. This is a simple, but common thing to do. As usual with transformations, if you don't approach it systematically, it is easy to make mistakes.

(a) *Suppose the random variable X has an exponential distribution with parameter 1, i.e. $X \sim \texttt{exp}(1)$. Give the range and pdf for the variables X and $Y = 3X$*

Sketch the graph of the density functions for each of these variables.

Solution: X: The range of X is $[0, \infty)$. The pdf is $f_X(x) = \mathrm{e}^{-x}$, for x inside the range and 0 elsewhere.

$Y = 3X$: The range of Y is $[0, \infty)$. To find f_Y, we work with the cumulative distributions.

$$\begin{aligned} F_Y(y) &= P(Y \le y) && \text{definition of CDF} \\ &= P(3X \le y) && \text{because } Y = 3X \\ &= P(X \le y/3) && \text{algebra} \\ &= F_X(y/3) && \text{definition of CDF} \end{aligned}$$

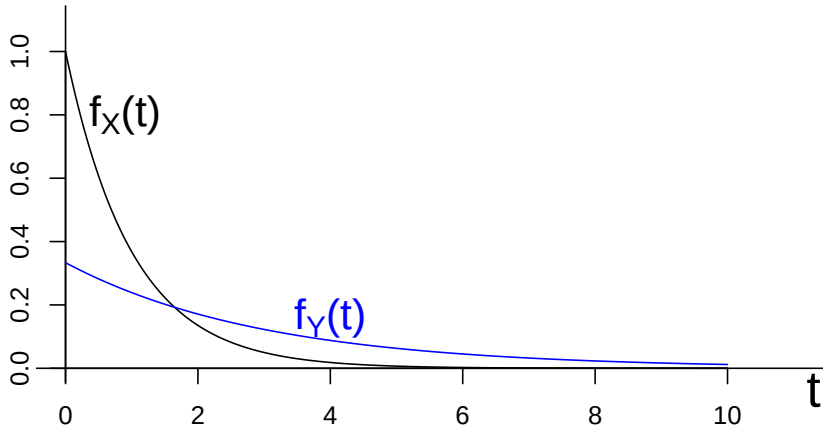
Now for the pdf:

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{1}{3}\frac{d}{dy}F_X(y/3) = \frac{1}{3}F_X'(y/3) = \frac{1}{3}f_X(y/3).$$

The second to last equality used the chain rule. The last equality is the fact that the pdf is the derivative of the cdf, i.e. $f_X = F_X'$.

Since we know $f_X(y/3) = \mathrm{e}^{-y/3}$, we have

$$\boxed{f_Y(y) = \frac{1}{3}\mathrm{e}^{-y/3}} \text{ for } y \text{ between 0 and } \infty.$$



(b) *For the random variable X from part (a), find the range and pdf of $W = aX + b$, where a and b are constants. Assume $a > 0$.*

Solution: We find the range and pdf by following the same pattern as in part (a). The range of W is $[b, \infty)$.

$$F_W(w) = P(W \le w) = P(aX + b \le w) = P\left(X \le \frac{w-b}{a}\right) = F_X\left(\frac{w-b}{a}\right).$$

Taking the derivative:

$$f_W(w) = \frac{d}{dw}F_W(w) = \frac{d}{dw}F_X\left(\frac{w-b}{a}\right) = \frac{1}{a}F_X'\left(\frac{w-b}{a}\right) = \boxed{\frac{1}{a}f_X\left(\frac{w-b}{a}\right).}$$

Since we know $f_X\left(\frac{w-b}{a}\right) = \mathrm{e}^{-(w-b)/a}$, we have

$$\boxed{f_W(w) = \frac{1}{a}\mathrm{e}^{-(w-b)/a}} \text{ for } w \text{ between } b \text{ and } \infty.$$

(c) *Let $V = X^3$. Find the range and pdf of V.*

Solution: We follow the same pattern as in the previous parts. The range of V is $[0, \infty)$.

$$F_V(v) = P(V \le v) = P(X^3 \le v) = P(X \le v^{1/3}) = F_X(v^{1/3}).$$

Taking the derivative:

$$f_V(v) = \frac{d}{dv}F_V(v) = \frac{d}{dv}F_X(v^{1/3}) = \frac{1}{3}v^{-2/3}F_X'(v^{1/3}) = \boxed{\frac{1}{3}v^{-2/3}f_X(v^{1/3}) = \frac{1}{3}v^{-2/3}\mathrm{e}^{-v^{1/3}}.}$$

Problem 5. (30: 10,10,10 pts.) *In this problem we will explore how the transformations in the previous problem affect the mean and median.*

(a) *For the variables X, Y, W in the previous problem, assume each of the variables are given in units of minutes. Find the expected value, variance and standard deviation of each variable. Be sure to include units in your answer.*

What are the units on a and b in the defnition of W?

Solution: Since $X \sim$ Exponential(1) we know

$$E[X] = 1 \text{ min.}, \ \mathrm{Var}(X) = 1 \text{ min.}^2, \ \sigma_X = 1 \text{ min.}$$

Since expected value is linear,

$$E[Y] = E[3X] = 3E[X] = 3 \text{ min.}, \quad E[W] = aE[X] + b = (a + b) \text{ min.}$$

Likewise, the variance is invariant under translation and scales by the square of the multiplier:

$$\mathrm{Var}(Y) = 9 \cdot \mathrm{Var}(X) = 9 \text{ min.}^2, \quad \sigma_Y = 3 \text{ min.},$$
$$\mathrm{Var}(W) = a^2 \cdot \mathrm{Var}(X) = a^2 \text{ min.}^2, \quad \sigma_W = a \text{ min.}.$$

Because both X and W are in units of minutes, a must be dimensionless and b has units of minutes.

(b) *For V from the previous problem, compute $E[V]$. As usual, you must set up the integral, but you can use a package like Wolfram Alpha to compute the integral.*

Solution: $E[V] = \displaystyle\int_0^\infty v f_V(v)\,dv = \frac{1}{3}\int_0^\infty v^{1/3}\mathrm{e}^{-v^{1/3}}\,dv$. This integral can be computed using the change of variable $u = v^{1/3}$, i.e. $u^3 = v$. The final answer is $E[V] = 6$. (Wolfram Alpha agrees!)

(c) *Compute the median value of both X and V.*

Solution: For X, the median value is the value $q_{0.5}$ such that $F_X(q_{0.5}) = 0.5$. Now,

$$F_X(q) = \int_0^q f_X(x)\,dx = \int_0^q \mathrm{e}^{-x}\,dx = -\mathrm{e}^{-x}\big|_0^q = 1 - \mathrm{e}^{-q}.$$

Solving $1 - \mathrm{e}^{-q} = 0.5$ gives the median of X is $\boxed{q_{0.5} = \ln(2)}$.

Since as we saw in the previous problem, $F_V(v) = F_X(v^{1/3})$, we have

$$F_V(v) = 0.5 \iff F_X(v^{1/3}) = 0.5 \iff v^{1/3} = q_{0.5}.$$

That is, the median of $V = X^3$ is just (the median of $X)^3$, i.e. $\ln(2)^3$.

Problem 6. (30: 5,5,10,10 pts.) **Fat tails**

This problem will explore the tails of two distributions. The tails are important when we want to think about probabilities of extreme events.

(a) *As an example, in the general population IQ has mean 100 and standard deviation of 15. IQ is normally distributed. Use the R function* `pnorm` *to give the probability that a randomly chosen person has IQ greater than 160, i.e. more than 4 standard deviations above the mean.*

Solution: We use the R code `p = 1 - pnorm(160, 100, 15)`. This gives the probability $p = 3.167124 \times 10^{-5}$.

(b) *Now, in order to be able to use R or Wolfram Alpha without a lot of distracting algebraic manipulation, we'll modify the definition of IQ. Suppose that Modified_IQ has mean 0 and standard deviation $\sqrt{3}$.*

Assuming Modified_IQ is normally distributed, find the probability that a randomly chosen person has Modified_IQ more than 4 standard deviations above the mean.

Solution: One of the important facts about normal distributions, is that, when measured in standard deviations above the mean they all give the same probabilities. That is, the answer is exactly the same as in part (a)

(c) *Now assume that Modified_IQ follows a t-distribution with 3 degrees of freedom. Later in the class we will work extensively with t-distributions. Here, it will be enough for us to know the following about this distribution.*

- *Range: $(-\infty, \infty)$*

- *PDF:* $f(x) = \dfrac{2}{3\pi}\left(1 + \dfrac{x^2}{3}\right)^{-2}$
- *Mean:* $\mu = 0$
- *Standard deviation:* $\sigma = \sqrt{3}$

(So this has the same mean and standard deviation as in part (b).)

For this problem, you can work with this pdf directly or you can look up how to use the R functions `dt` *and* `pt`*.*

Assuming it follows this t-distribution, find the probability that a randomly chosen person has Modified_IQ more than 4 standard deviations above the mean.

You can use R or another calculation package to do the calculation, but you must explicitly show the integral in terms of the probability density.

Compare this value with the probability in part (b)

Solution: If I is the Modified_IQ of a randomly chosen person, we want to compute $P(I > 4 * \sqrt{3})$. In terms of the pdf this is

$$P(I > 4\sqrt{3}) = \int_{4\sqrt{3}}^{\infty} f(x)\,dx = \int_{4\sqrt{3}}^{\infty} \frac{2}{3\pi}\left(1 + \frac{x^2}{3}\right)^{-2} dx$$

We computed this in R with the code `p = 1 - pt(4*sqrt(3), 3)`. This gives the probability $p = 0.003082687$.

This is a small probability, but it is about 100 times the probability in parts (a) and (b).

(d) *For this problem, compute probabilities using both the normal distribution in part (b) and the t-distribution in part (c). Do this for the following probabilities.*

(i) $P(\text{Modified\_IQ} > 20)$, *(ii)* $P(\text{Modified\_IQ} > 40)$, *(iii)* $P(\text{Modified\_IQ} > 200)$.

Why do we say that the t-distribution has a 'fat tail'?

Hence the moral of this problem: Knowing the mean and standard deviation of a quantity is often not enough for predicting the frequency of extreme events (high IQ, 100-year floods, etc.); you need to know the underlying distribution itself (which often requires finding out the underlying geophysics, geochemistry, and biology). In the solutions we will show you graphs of these distributions zoomed in around 4σ *above the mean. If you do that yourself, you will see that they look very different.*

Solution: For $x = 20, 40, 200$, we use the R code `p_t = 1 - pt(x, 3)`, `p_normal = 1 - pnorm(x, 0, sqrt(3))`. We get

(i) $x = 20$: `p_t = 0.00014,  p_normal = 0`

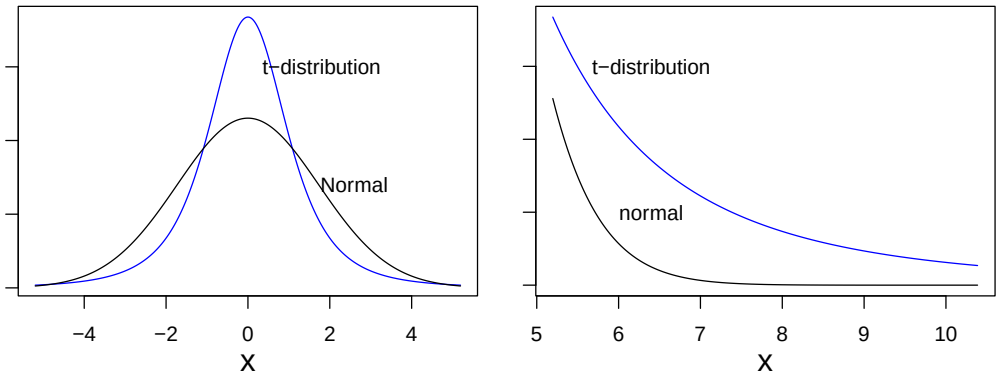
(ii) $x = 40$: `p_t = 1.7e-5,  p_normal = 0`

(iii) $x = 200$: `p_t = 1.4e-7,  p_normal = 0`

All three examples show that for x far from the mean, i.e. in the tail, the t-distribution probability is many orders of magnitude greater than the normal distribution probability

We say the t-distribution has a fat tail, because the its tail contains much more probability than the normal distribution. That is, extreme events are much more likely for the t-distribution.

Graphically, the following figures show the tail of the t-distribution is much greater, i.e. fatter, than that of the normal distribution.



Left: Center of distributions, Right: tails from 3σ to 6σ above mean

18.05 Introduction to Probability and Statistics
Spring 2022