IPDA1005 Introduction to Probability and Data Analysis

Worksheet 8

Solution

1. The article "Second Moment Reliability Evaluation vs. Monte Carlo Simulations for Weld Fatigue Strength" (*Quality and Reliability Engr. Intl.*, 2012: 887–896) considered the use of a uniform distribution with $A = 0.20$ and $B = 4.25$ for the diameter X (mm) of a certain type of weld.

   (a) Determine the PDF of X and plot it.

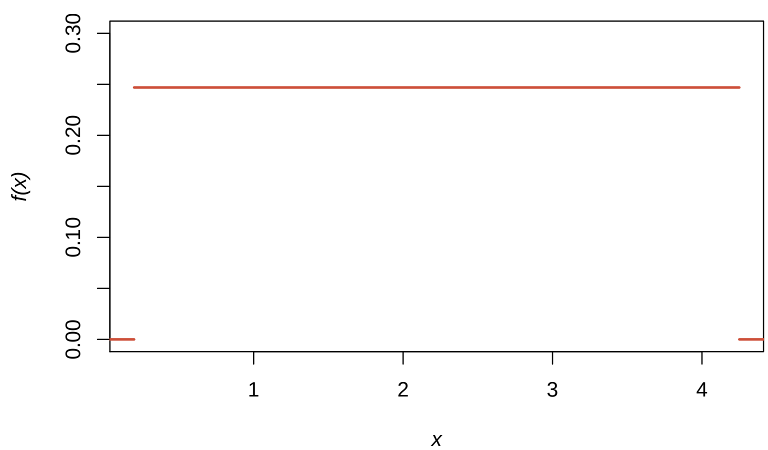
   (b) Identify the mean of X from the plot in (a).

   (c) What is the probability that the diameter X exceeds 3 mm?

   (d) For any value a satisfying $0.20 < a < a + 1 < 4.25$, what is $P(a < X < a + 1)$?

Solution:

(a) $X \sim \text{Unif}[0.20, 4.25]$ and therefore $f(x) = \frac{1}{4.25-0.20} = \frac{1}{4.05}$ for $0.20 \leq X \leq 4.25$.



(b) Because this is a symmetric distribution, the mean will be at the mid-way point, i.e., $E(X) = \frac{0.20+4.25}{2} = 2.225$.

(c) $P(X > 3) = \frac{4.25-3}{4.05} = 0.3086$.

(d) The interval between a and $a + 1$ is one, so for any value a that satisfies $0.20 < a < a + 1 < 4.25$, $P(a < X < a + 1) = \frac{1}{4.05}$.

2. Let X denote the vibratory stress (psi) on a wind turbine blade at a particular wind speed in a wind tunnel. The article "Blade Fatigue Life Assessment with Application to VAWTS" (*J. Solar Energy Engr.*, 1982: 107-–111) proposes the *Rayleigh* distribution, with PDF

$$f(x;\theta) = \frac{x}{\theta^2} \cdot e^{-x^2/(2\theta^2)}, \quad x > 0$$

as a model for X, where θ is a positive constant.

(a) Verify that $f(x;\theta)$ is a legitimate PDF.

(b) Fill in the blank: "θ is a .... of this distribution."

(c) Suppose that $\theta = 100$ (a value suggested by the article). What is the probability that X is between 100 and 200?

(d) What is the expression for the CDF of X?

Solution:

(a) First, it is clear that $f(x;\theta) \geq 0$ for $x > 0$ so the first condition is satisfied. Second, we need to show that $\int_{-\infty}^{\infty} f(x;\theta)\, dx = 1$. Thus,

$$\int_{-\infty}^{\infty} f(x;\theta)\, dx = \int_0^{\infty} \frac{x}{\theta^2} e^{-x^2/(2\theta^2)}\, dx = -e^{-x^2/(2\theta^2)}\Big|_0^{\infty} = 0 - (-1) = 1$$

(b) θ is a **parameter** of this distribution.

(c) $P(100 \leq X \leq 200) = \int_{100}^{200} f(x;100)\, dx = -e^{-x^2/20000}\Big|_{100}^{200} = 0.4712$

(d) For $x \leq 0$, F(x)=0. For $x > 0$,

$$F(x) = P(X \leq x) = \int_0^x \frac{y}{\theta^2} e^{-y^2/(2\theta^2)}\, dy = -e^{-y^2/(2\theta^2)}\Big|_0^x = 1 - e^{-x^2/(2\theta^2)}$$

3. Let X be a continuous random variable with CDF given by

$$F(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x}{4}\left[1 + \ln\left(\frac{4}{x}\right)\right] & 0 < x \leq 4 \\ 1 & x > 4 \end{cases}$$

This type of CDF is suggested in the article "Variability in Measured Bedload-Transport Rates" (*Water Resources Bull.*, 1985: 39-–48) as a model for a hydrologic variable.] What is

(a) $P(X \leq 1)$?

(b) $P(1 \leq X \leq 3)$?

(c) the PDF of X?

Solution:

(a) $P(X \leq 1) = F(1) = 0.25[1 + \ln(4)] = 0.5966$

(b) $P(1 \leq X \leq 3) = F(3) - F(1) = 0.9658 - 0.5966$

(c) Recall that $f(x) = F'(x)$, so we need to differentiate $F(x)$ for $0 < x < 4$. Thus,

$$f(x) = F'(x) = \frac{1}{4} + \frac{\ln 4}{4} - \frac{1}{4}\ln x - \frac{1}{4}x\frac{1}{x} = \frac{1}{4}(\ln 4 - \ln x)$$

for $0 < x < 4$ and zero at all other points.

4. In countries where Imperial and SI (système internationale, or metric) co-existed, scientists and engineers needed to know how to convert between one and the other, for example, between °C and °F. Let X be the temperature in °C at which a chemical reaction takes place, and let Y be the temperature in °F, and recall that $Y = 1.8X + 32$.

   (a) If the median of the distribution of X is η_X, show that the median of Y is $1.8\eta_X + 32$.

   (b) Based on your result in (a), conjecture how is the 90th percentile of Y related to the 90th percentile of X. Verify your conjecture.

Solution:

(a) Let η_Y denote the median of Y, and by definition, $P(Y \leq \eta_Y) = 0.5$. Hence,

$$\begin{aligned} P(Y \leq \eta_Y) &= 0.5 \\ P(1.8X + 32 \leq \eta_Y) &= 0.5 \\ P\left(X \leq \frac{\eta_Y - 32}{1.8}\right) &= 0.5 \Rightarrow \eta_X = \frac{\eta_Y - 32}{1.8} \end{aligned}$$

and solving for the median of Y yields $\eta_Y = 1.8\eta_X + 32$.

(b) We might conjecture that $\eta_{Y,0.9} = 1.8\eta_{X,0.9} + 32$, and indeed we find that this is so if we substitute 0.9 for 0.5 in the derivation above.

5. If $X \sim U(A, B)$, then it's easy to see that $\mu = E(X) = (A + B)/2$. It's not so obvious, however, that that $\text{Var}(X) = (B - A)^2/12$. Show that it is so, by beginning with the definition

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)\, dx$$

and then carrying out the integration by substitution.

Solution: : The variance of X is given by

$$\text{Var}(X) = \int_A^B (x-\mu)^2 \cdot \frac{1}{B-A}\,dx = \frac{1}{B-A}\int_A^B \left(x - \frac{A+B}{2}\right)^2\,dx$$

If we let $u = x - \frac{A+B}{2}$, then we can write

$$\begin{aligned}
\text{Var}(X) &= \frac{1}{B-A}\int_{-(B-A)/2}^{(B-A)/2} u^2\,du \\
&= \frac{2}{B-A}\int_0^{(B-A)/2} u^2\,du \qquad \text{because of symmetry} \\
&= \frac{2}{B-A}\frac{u^3}{3}\Big|_0^{(B-A)/2} = \frac{2}{B-A}\frac{(B-A)^3}{2^3\cdot 3} = \frac{(B-A)^2}{12}
\end{aligned}$$

6. The weekly demand for propane gas (in thousands of gallons) from a particular facility is a random variable X with PDF

$$f(x) = \begin{cases} 2\left(1-\frac{1}{x^2}\right) & 1 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Compute the CDF of X.

(b) Obtain an expression for the ($100p$)th percentile. What is the value of the median?

(c) Calculate $E(X)$. Compare the value of the mean and median and comment on the shape of the distribution.

(d) Compute $\text{Var}(X)$ and then the standard deviation of X.

Solution:

(a) For $x \le 1$, F(x)=0 and for $x \ge 2$, F(x)=1. For $1 \le x \le 2$,

$$F(x) = \int_1^x 2\left(1-\frac{1}{y^2}\right)\,dy = 2\left(y+\frac{1}{y}\right)\Bigg|_1^x = 2\left(x+\frac{1}{x}\right) - 4$$

(b) To obtain an expression for the ($100p$)th percentile, we set $F(x) = p$ and then solve for x. Thus,

$$2\left(x+\frac{1}{x}\right) - 4 = p \Rightarrow 2x^2 - (p+4)x + 2 = 0 \Rightarrow \eta_p = x = \frac{p+4+\sqrt{p^2+8p}}{4}$$

(We ignore the other root of the quadratic equation because a little bit of algebra will show that this will result in an x-value that is less than 1.)
To calculate the median, we set $p = 0.5$ and obtain $\eta_{0.5} = 1.640$, and hence the median weekly demand for propane gas is 1640 gallons.

(c) $E(X) = \int_1^2 x \cdot 2\left(1 - \frac{1}{x^2}\right)\,dx = 2\int_1^2 \left(x - \frac{1}{x}\right)\,dx = 2\left(\frac{x^2}{2} - \ln x\right)\Big|_1^2 = 1.614$, or 1614 gallons of propane gas. The mean is slightly lower than the median, because the distribution is slightly left-skewed.

(d) $E(X^2) = 2\int_1^2 (x^2 - 1)\,dx = 2\left(\frac{x^3}{3} - x\right)\Big|_1^2 = \frac{8}{3} = 2.667$ and hence $\mathrm{Var}(X) = 2.667 - (1.614)^2 = 0.0626$, which implies that the standard deviation is $\sqrt{0.0626} \approx 0.2502$, or about 250 gallons.

7. Marks on an exam were not very good. When graphed, their distribution had a shape similar to the PDF

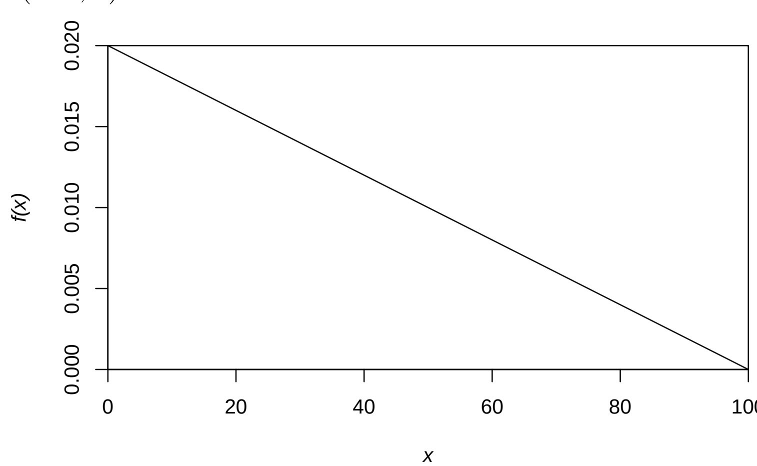
$$f(x) = \frac{1}{5000}(100 - x), \qquad 0 \le x \le 100$$

(a) Sketch the distribution above. What was the average mark?

(b) In an effort to 'curve' the distribution and increase the average mark, the lecturer decides to assign a new grade to everyone: if an individual's mark was X, it will be replaced by $10\sqrt{X}$. Is that strategy successful in raising the class average above 60?

Solution:

(a) The PDF is simply the equation of a straight line that begins at $(0, 100/5000)$ and ends at $(100, 0)$:



The average mark can be obtained by calculating the definite integral

$$
\begin{aligned}
E(X) &= \int_0^{100} x \cdot f(x)\,dx \\
&= \frac{1}{5000}\int_0^{100} x(100-x)\,dx \\
&= \frac{1}{5000}\int_0^{100} (100x - x^2)\,dx \\
&= \frac{1}{5000}\left[50x^2 - \frac{1}{3}x^3\right]_0^{100} \\
&\vdots \\
&= 33.3
\end{aligned}
$$

(b) Recall that the expected value of a function of a random variable $h(X)$ is $E[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f(x)\,dx$. Hence, the average of the re-scaled marks will be

$$
\begin{aligned}
E(10\sqrt{X}) &= \int_0^{100} h(x) \cdot f(x)\,dx \\
&= \frac{1}{5000}\int_0^{100} 10x^{1/2}(100-x)\,dx \\
&= \frac{10}{5000}\int_0^{100} (100x^{1/2} - x^{3/2})\,dx \\
&= \frac{1}{500}\left[\frac{200}{3}x^{3/2} - \frac{2}{5}x^{5/2}\right]_0^{100} \\
&\vdots \\
&= 53.3
\end{aligned}
$$

The strategy was not successful in raising the class average above 60.

8. If X is a continuous random variable with PDF $f(x)$, mean μ, and standard deviation σ, show that for any constants a and b, $\mathrm{Var}(aX+b) = a^2\sigma^2$.

Solution: First, it's straightforward to show that that $E(aX+b) = a\mu + b$. Then, recall that $\mathrm{Var}(aX+b) = E[((aX+b) - E(aX+b))^2]$, and hence

$$\begin{aligned}\mathrm{Var}(aX+b) &= E[(aX+b-a\mu-b)^2] \\ &= E[(aX-a\mu)^2] \\ &= E[a^2(X-\mu)^2] \\ &= a^2E[(X-\mu)^2] \\ &= a^2\mathrm{Var}(X) \\ &= a^2\sigma^2\end{aligned}$$

9. Based on extensive data from an urban freeway to the west of Toronto, we can assume that "free speeds can be best represented by a Normal distribution" ("Impact of Driver Compliance on the Safety and Operational Impacts of Freeway Variable Speed Limit Systems", *Journal of Transportation Engineering*, 2011: 260–268). In the article, the mean and standard deviation were reported to be 119 km/h and 13.1 km/h, respectively.
   (a) What is the probability that the speed of a randomly selected vehicle is between 100 and 120 km/h?
   (b) What speed characterizes the fastest 10% of all speeds?
   (c) What would the mean and standard deviation of the distribution of speeds be if they were to be expressed in miles/h?

Solution: Using R functions such as `pnorm` and `qnorm`, it is trivial to calculate the required quantities in (a) and (b), but obtaining them 'by hand' requires us to calculate Z-scores and then refer to tables of the standard Normal distribution to obtain probabilities or quantiles. Recall that $Z \sim N(0,1)$ and that $P(Z \leq z) = \Phi(z)$, and from the problem statement, $X \sim N(119, 13.1^2)$.

(a) $P(100 \leq X \leq 120) = \Phi\left(\frac{120-119}{13.1}\right) - \Phi\left(\frac{100-119}{13.1}\right) = \Phi(0.0763) - \Phi(-1.4504) = 0.5304 - 0.0735 = 0.4569$.

(b) We require k such that $P(X > k) = 0.1$, or equivalently, $P(X \leq k) = 1 = 0.1 = 0.9$. Hence,

$$0.9 = \Phi\left(\frac{k-119}{13.1}\right) \Rightarrow \Phi^{-1}(0.9) = 1.28 = \frac{k-119}{13.1} \Rightarrow k = 135.8 \text{ km/h}$$

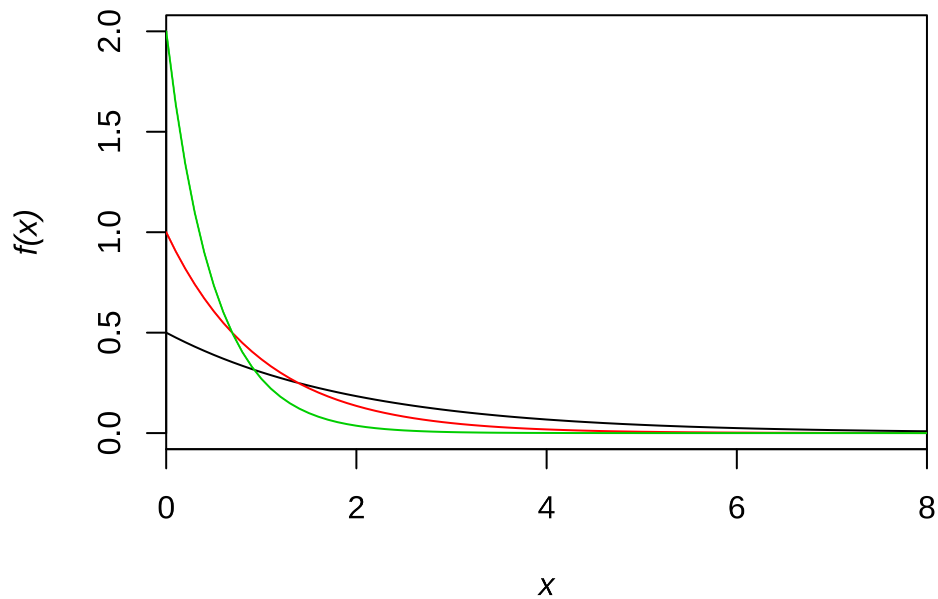
(c) There are 1.60934 km in a mile, so the mean and standard deviation expressed in miles would be $119/1.60934 = 73.9$ mi and $13.1/1.60934 = 8.1$ mi, respectively.

10. The exponential distribution is widely used in engineering, science, and finance. The random variable X is said to have an exponential distribution with parameter λ if the

PDF of X is

$$f(x;\lambda) = \lambda e^{-\lambda x}, \qquad x > 0$$

(a) For $\lambda = 0.5, 1, 2$, the plot below shows graphs of several exponential PDFs. Which is which?



(b) The standard deviation of X is $1/\lambda$, but what is $E(X)$? [Hint: obtaining this expected value requires integration by parts.]

(c) Obtain the CDF.

(d) Data collected at an international airport suggests that an exponential distribution with mean value 2.725 h is a good model for rainfall duration (*Urban Stormwater Management Planning with Analytical Probabilistic Models*, 2000, p. 69). What is the probability that the duration of a particular rainfall event at this location is at least 2 h? Between 2 and 3 h?

Solution:

(a) Clearly, $f(0;\lambda) = \lambda$, so we can immediately idenfity the black, red, and green curves as corresponding to $\lambda = 0.5, 1, 2$, respectively.

(b) Using the definition of $E(X)$, we write that

$$\begin{aligned} E(X) &= \int_0^\infty x \cdot f(x)\, dx \\ &= \lambda \int_0^\infty x e^{-\lambda x}\, dx \end{aligned}$$

For definite integrals the expression for integration by parts is (you would have learned this in high-school or first-year calculus)

$$\int_a^b u\,dv = uv\Big|_a^b - \int_a^b v\,du$$

If we let $u = x$ and $dv = e^{-\lambda x}\,dx$, then $du = dx$ and $v = -\lambda^{-1}e^{-\lambda x}$, and we can write

$$\begin{aligned} E(X) &= \lambda\left[-\frac{x}{\lambda}e^{-\lambda x}\Big|_0^\infty + \frac{1}{\lambda}\int_0^\infty e^{-\lambda x}\,dx\right] \\ &= \lambda\left[-\frac{1}{\lambda^2}e^{-\lambda x}\right]_0^\infty \\ &= \frac{1}{\lambda} \end{aligned}$$

(c) The CDF is defined as $F(x) = P(X \leq x) = \int_{-\infty}^x f(y)\,dy$. For the exponential distribution, therefore,

$$\begin{aligned} F(x) &= \lambda\int_0^x e^{-\lambda y}\,dy \\ &= \lambda\left[-\frac{1}{\lambda}e^{-\lambda y}\right]_0^x \\ &= 1 - e^{-\lambda x} \end{aligned}$$

(d) A mean value of 2.725 implies that $\lambda = 1/2.725$, and we can use the expression we derived for $F(x)$ above to calculate the required probabilities. Thus,

$$P(X \geq 2) = 1 - P(X < 2) = 1 - P(X \leq 2) = 1 - F(2) = e^{-2/2.725} = 0.48.$$

Furthermore,

$$P(2 \leq X \leq 3) = F(3) - F(2) = e^{-2/2.725} - e^{-3/2.725} = 0.48 - 0.3326 = 0.1474$$

(*Sources: All questions adapted from Devore and Berk (2012) and Carlton and Devore (2017).*)

Bibliography

1. Carlton, M.A. and Devore, J.L. (2017) *Probability with Applications in Engineering, Science, and Technology*, 2nd ed. Springer: New York.

2. Devore, J.L. and Berk, K.N. (2012) *Modern Mathematical Statistics with Applications.* Springer: New York.