"The world is a bell curve."

Simon Sinek

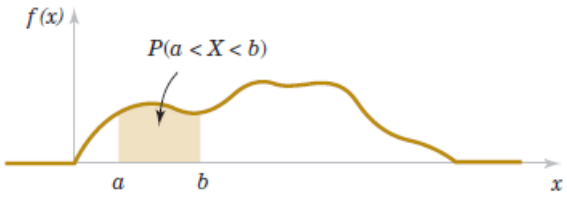
8 Continuous random variables

General facts

A continuous random variable can assume any value in a given interval, for example: the random variable X that measures the necessary time to accomplish something. For a discrete random variable the verb was "to count", now it becomes "to measure".

• X is a continuous random variable when there exists a function $f(x)$, called probability density function, such that for all $-\infty \leq a \leq b \leq \infty$

$$P(a < X < b) = \int_a^b f(x)dx$$



• a probability density function (PDF) has the defining properties

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad \text{and} \quad f(x) \geq 0$$

• the cummulative distribution function defined by $F(x) := P(X \leq x)$ can substitute the role of the function f in computing probabilities

$$P(a < X \leq b) = F(b) - F(a)$$

• for a continuous random variable X, when F is continuous, one has the identities

$$P(a \leq X < b) = P(a \leq X \leq b) = P(a < X < b) = P(a < X \leq b) = F(b) - F(a)$$

In general the following properties hold

• $F'(x) = f(x)$ (if the derivative exists)

• $F(x) = \int_{-\infty}^{x} f(t)dt$

• $F(x_1) \leq F(x_2)$ daca $x_1 < x_2$

• $\lim_{x\to\infty} F(x) = 1$ si $\lim_{x\to-\infty} F(x) = 0$

• it is worth mentioning that $P(X = c) = 0$, for every value $c \in \mathbb{R}$, this contrasts with the discrete case

• the mean value $E(X)$ and the variance $var(X)$ are defined in terms of the corresponding PDF

$$E(X) = \int_{-\infty}^{\infty} xf(x)\ dx,$$

$$var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\ dx.$$

• the moments of order k denoted by M_k are

$$M_k(X) = \int_{-\infty}^{\infty} x^k f(x)\ dx$$

and the central moments of order k

$$m_k(X) = \int_{-\infty}^{\infty} (x - E(X))^k f(x)\ dx$$

• the covariance $cov(X, Y)$ and the correlation coefficient $\rho_{X,Y}$ are defined similarly to the discrete case

Functions of random variables

• for a discrete random variable the PDF of $Y = g(X)$ one usually finds first the CDF using

$$F_Y(y) = P(g(X) \leq y) = P\left(X \leq g^{-1}(y)\right)$$

• the expected value is

$$E[Y] = \int_{-\infty}^{\infty} g(x) f_X(x)\ dx$$

Conditional random variables

• given an event B, with $P(B) > 0$ the conditional probability density function of X is

$$f_{X|B}(x) = \begin{cases} \frac{f_X(x)}{P(B)} & \text{, if } x \in B \\ 0 & \text{, otherwise} \end{cases}$$

• the conditional expected value

$$E[X|B] = \int_{-\infty}^{\infty} x f_{X|B}(x)\, dx$$

• the conditional variance is

$$var(X|B) = E[X^2|B] - E^2[X|B]$$

• a random variable X resulting from an experiment with event space $B_1, B_2, \ldots B_m$ has the PMF

$$f_X(x) = \sum_{i=1}^{m} f_{X|B_i}(x) \cdot P(B_i)$$

Classical continuous random variables

Continuous uniform random variables

• if X has the probability density function

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{for } \; x \in [a,b] \\ 0, & \text{otherwise} \end{cases}$$

we call X uniformly distributed and write $X \sim U(a,b)$.

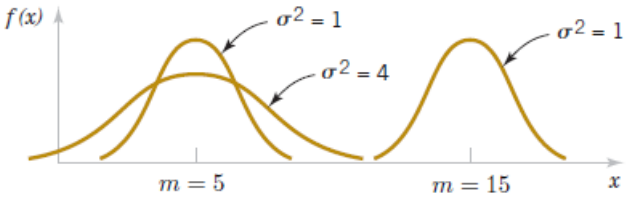
• the formulae $M(X) = \frac{a+b}{2}$ and $D^2(X) = \frac{(b-a)^2}{12}$ hold

Normal random variables

• if X has the probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

we call X normally distributed and write $X \sim N(m, \sigma^2)$.



• for such a random variable one has $M(X) = m$ si $D^2(X) = \sigma^2$.

Standard normal random variables

• a variable with a standard normal distribution Z is a normally distributed variable that corresponds to $m = 0$ and $\sigma = 1$, $Z \sim N(0, 1)$.

• its cummulative distribution function is denoted by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

and it has the important values in a z-table.

• when working with normally distributed random variables we frequently use the following standardization argument

• for a random variable $X \sim N(m, \sigma^2)$ we compute probabilites in this way

$$P(x_1 \leq X \leq x_2) = P\left(\frac{x_1 - m}{\sigma} \leq Z \leq \frac{x_2 - m}{\sigma}\right) = \Phi\left(\frac{x_2 - m}{\sigma}\right) - \Phi\left(\frac{x_1 - m}{\sigma}\right)$$

where $Z := \dfrac{X - m}{\sigma}$ is a standard normal random variable and the corresponding values $\Phi\left(\dfrac{x_2 - m}{\sigma}\right)$, $\Phi\left(\dfrac{x_1 - m}{\sigma}\right)$ can be found in the z-table.

• the above identities establish a connection between the CDF of the random variable $X \sim N(m, \sigma^2)$ and the CDF of $Z \sim \mathcal{N}(0, 1)$

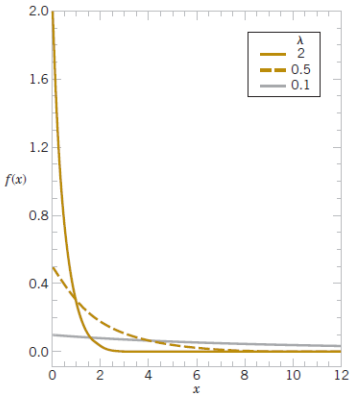
$$F_X(x) = \Phi\left(\frac{x - m}{\sigma}\right)$$

Exponential random variables

• if X has the probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{for } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

we call X exponentially distributed and write $X \sim Exp(\lambda)$.



- the following formulae hold $M(X) = \frac{1}{\lambda}$ and $D^2(X) = \frac{1}{\lambda^2}$

Normal approximations of some random variables

- the central limit theorem plays an important role in probabilities and statistics, it helps us to simplify problems by allowing us to work with a distribution that is approximately normal
- we present here two practical situations:

1 when X is a binomial random variable $X \sim Bin(n, p)$ and n is large enough, then X can be approximated by $Y \sim \mathcal{N}(np, np(1-p))$ and the continuity corrections hold

$$P(X = k) \approx P\left(k - \frac{1}{2} < Y < k + \frac{1}{2}\right)$$

since here X is discrete, it makes sense computing the probability of the event $X = k$, moreover

$$P(X \leq k) = P(X < k + 1) \approx P\left(Y < k + \frac{1}{2}\right)$$

$$P(X \geq k) = P(X > k - 1) \approx P\left(Y > k - \frac{1}{2}\right)$$

$$P(k_1 \leq X \leq k_2) \approx P\left(k_1 - \frac{1}{2} < Y < k_2 + \frac{1}{2}\right)$$

2 when X is a Poisson random variable with parameter λ and λ is large enough, then X can be approximated by a normally distributed random variable $Y \sim N(\lambda, \lambda)$

- one can use the same continuity corrections

Solved problems

Problem 1. *The random variable X has the probability density function*

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

a) Find the cummulative distribution function
b) Find the PDFs corresponding to $Y=e^X$ and $Z=2X^2+1$.

Solutie: a) The function f is a probability density function because

$$\int_{-\infty}^{\infty} f(x)\,dx=\int_{-1}^{1}\frac{1}{2}dx=\frac{x}{2}\,|_{-1}^{1}=1$$

and it is nonnegative.

By its very definition the cummulative distribution function is

$$F_X(x)=\int_{-\infty}^{x} f(t)\,dt=\begin{cases}0, & x<-1\\ \frac{x+1}{2}, & -1\le x<1\\ 1, & 1\le x\end{cases},$$

since

$$x<-1\Rightarrow F(x)=\int_{-\infty}^{x}0\;dt=0,$$

$$x\in[-1,1)\Rightarrow F(x)=\int_{-\infty}^{-1}0\;dt+\int_{-1}^{x}\frac{1}{2}\;dt=\frac{x}{2}+\frac{1}{2}=\frac{x+1}{2},$$

$$1\le x\Rightarrow F(x)=\int_{-\infty}^{-1}0\;dt+\int_{-1}^{1}\frac{1}{2}\;dt+\int_{1}^{\infty}0\;dt=1.$$

b) We start by finding the cummulative distribution function $G(x)$ of the random variable Y. Since $Y>0$ for every $x\le 0$ one gets $G(x)=P(Y\le x)=0$. If $x>0$ then

$$G(x)=P(Y\le x)=P\left(e^X\le x\right)=P(X\le\ln x)=F(\ln x)$$

Putting it together one gets

$$G(x)=\begin{cases}0, & \ln x<-1 \text{ and } x\le 0\\ \frac{1+\ln x}{2}, & -1\le\ln x<1\\ 1, & 1\le\ln x\end{cases}=\begin{cases}0, & x\in\left(-\infty,\frac{1}{e}\right)\\ \frac{1+\ln x}{2}, & x\in\left[\frac{1}{e},e\right)\\ 1, & x\in[e,\infty)\end{cases}$$

The corresponding PDF will be

$$g(x) = G'(x) = \begin{cases} \frac{1}{2x}, & x \in \left(\frac{1}{e}, e\right) \\ 0, & \text{otherwise} \end{cases}$$

Since X is nonzero only on $(-1,1)$, $Z = 2X^2 + 1$ will be nonzero on $(1,3)$. For $x \in (1,3)$, the cummulative distribution function $H(x)$ of Z will be

$$\begin{aligned} H(x) &= P(Z \leq x) = P\left(2X^2 + 1 \leq x\right) = P\left(X^2 \leq \frac{x-1}{2}\right) = \\ &= P\left[-\sqrt{\frac{x-1}{2}} \leq X \leq \sqrt{\frac{x-1}{2}}\right] = F\left[\sqrt{\frac{x-1}{2}}\right] - F\left[-\sqrt{\frac{x-1}{2}}\right] \\ &= \frac{1}{2}\left[1 + \sqrt{\frac{x-1}{2}}\right] - \frac{1}{2}\left[1 - \sqrt{\frac{x-1}{2}}\right] = \sqrt{\frac{x-1}{2}}. \end{aligned}$$

The corresponding PDF, obtained by $h(x) = H'(x)$, is

$$h(x) = \begin{cases} \frac{1}{2\sqrt{2x-2}}, & x \in (1,3) \\ 0, & \text{otherwise} \end{cases}$$

Problem 2. *The probability density function of a continuous random variable X is given by*

$$f(x) = \begin{cases} \frac{1}{2}\cos x, & x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \\ 0, & \textit{otherwise} \end{cases}$$

a) Compute the mean value and the variance of X.
b) Find the CDF and calculate the probability $P\left(\frac{\pi}{4} < X < \frac{\pi}{3}\right)$.

Solutie: a) The mean value

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx = \frac{1}{2} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \underbrace{x \cos x}_{\text{f is odd}}\, dx \overset{\substack{\text{symmetric} \\ \text{interval}}}{=} 0,$$

and the variance

$$\begin{aligned} var(X) &= \int_{-\infty}^{+\infty} [x - E(X)]^2 f(x)\, dx = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} (x-0)^2 f(x)\, dx = \\ &= \frac{1}{2} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \underbrace{x^2 \cos x}_{\text{f is even}}\, dx \overset{\substack{\text{symmetric} \\ \text{interval}}}{=} 2 \cdot \frac{1}{2} \int_{0}^{\frac{\pi}{2}} x^2 \cos x dx, \end{aligned}$$

hence

$$var\left(X\right)=\frac{\pi^2}{4}-2.$$

b) The CDF is defined as

$$F\left(x\right)=\int\limits_{-\infty}^{x}f\left(t\right)dt.$$

Thus, for $x<-\frac{\pi}{2}\implies F\left(x\right)=\int\limits_{-\infty}^{x}0\ dt=0.$

For $x\in\left[-\frac{\pi}{2},\frac{\pi}{2}\right)$ one gets

$$F\left(x\right)=\int\limits_{-\infty}^{x}f\left(t\right)dt=\int\limits_{-\infty}^{-\frac{\pi}{2}}0\ dt+\int\limits_{-\frac{\pi}{2}}^{x}\frac{1}{2}\cos t\ dt=\frac{1}{2}+\frac{1}{2}\sin x.$$

and if $x\geq\frac{\pi}{2}$ we have

$$F\left(x\right)=\int\limits_{-\infty}^{x}f\left(t\right)dt=\int\limits_{-\infty}^{-\frac{\pi}{2}}0\ dt+\int\limits_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\frac{1}{2}\cos t\ dt+\int\limits_{\frac{\pi}{2}}^{x}0\ dt=1.$$

hence

$$F\left(x\right)=\begin{cases}0, & x\leq-\frac{\pi}{2}\\ \frac{1}{2}+\frac{1}{2}\sin x, & x\in\left(-\frac{\pi}{2},\frac{\pi}{2}\right)\\ 1, & x\geq\frac{\pi}{2}\end{cases}.$$

For a continuous random variable we can use the PDF or the CDF, in oder to compute probabilities

$$P\left(\frac{\pi}{4}<X<\frac{\pi}{3}\right)=\int\limits_{\frac{\pi}{4}}^{\frac{\pi}{3}}f\left(x\right)dx=\int\limits_{\frac{\pi}{4}}^{\frac{\pi}{3}}\frac{1}{2}\cos x\ dx=F\left(\frac{\pi}{3}\right)-F\left(\frac{\pi}{4}\right)=\frac{\sqrt{3}-\sqrt{2}}{4}.$$

Problem 3. *Overbooking of passengers on intercontinental flights is a common practice among airlines, see the United Airlines case. Aircrafts which are capable of carrying* 300 *passengers are booked to carry* 320 *passengers. If* 10% *of passengers who have a booking fail to turn up for their flights, what is the probability that at least one passenger who has a booking, will end up without a seat on a particular flight? What is the probability that between* 25 *and* 45 *passengers will fail to turn up for their flights?*

Solutie: First of all, one has to recognize a binomial experiment. There are $n=320$ trials in this experiment. In every trial one passenger, with a reserved

seat, tries to board his plane. We call "success" the situation when the passenger fail to turn up for his flight. The probability of a success is $p = 0.10$

We denote by X the discrete random variable that counts the passengers who fail to turn up for their flight. The variable X has a binomial distribution $X \sim Bin(320, 0.10)$ and we have to compute $P(X \leq 19)$ and $P(25 \leq X \leq 45)$.

We can do that using the binomial distribution but it leads to tedious computations, for example

$$P(X \leq 19) = \sum_{k=1}^{19} C_{320}^{k} (0.10)^k (0.90)^{320-k}$$

A better idea is to approximate X by a continuous random variable Y which is normally distributed

$$Y \sim N(np, np(1-p)) = N(32, 28.8)$$

In order to improve the estimates we'll use the continuity corrections:

$$P(k_1 \leq X \leq k_2) \approx P\left(k_1 - \frac{1}{2} < Y < k_2 + \frac{1}{2}\right)$$

and

$$P(X \leq k) \approx P\left(Y < k + \frac{1}{2}\right)$$

Thus

$$P(25 \leq X \leq 45) \approx P\left(25 - \frac{1}{2} < Y < 45 + \frac{1}{2}\right)$$

and

$$P(X \leq 19) \approx P\left(Y < 19 + \frac{1}{2}\right)$$

Now the last probability is related to a normally distributed random variable with means $m = 32$ and variance $\sigma^2 = 28.8$. We can not use the PDF of a normally distributed r.v. since it involves the function e^{-x^2} which can not be integrated using elementary methods.

The idea is to use a standardization algorithm and the z-scores. Hence the random variable $Z = \frac{Y-m}{\sigma}$ will be standard normally distributed $Z \sim N(0, 1)$ since

$$E(Z) = Z\left(\frac{Y-m}{\sigma}\right) = \frac{1}{\sigma}(E(Y) - m) = 0$$

and

$$var(Z) = \frac{1}{\sigma^2}(var(E) - 0) = \frac{\sigma^2}{\sigma} = 1$$

It is worth mentioning the identities

$$P(x_1 \leq X \leq x_2) = P\left(\frac{x_1 - m}{\sigma} \leq Z \leq \frac{x_2 - m}{\sigma}\right) = \Phi\left(\frac{x_2 - m}{\sigma}\right) - \Phi\left(\frac{x_1 - m}{\sigma}\right)$$

where Φ is the CDF of a standard normally distributed random variable. It's values are given in a z-table.

Now we are able to approximate the requested probabilities

$$P(25 \leq X \leq 45) \approx P\left(24.5 < Y < 45.5\right) = P\left(\frac{24.5-32}{5.36} \leq Z \leq \frac{45.5-32}{5.36}\right)$$
$$= \Phi\left(2.51\right) - \Phi\left(-1.39\right) = 0.9940 - 0.0823 = 0.92 = 92\%$$

and

$$P(X \leq 19) \approx P\left(Y < 19 + \frac{1}{2}\right) = P\left(Z \leq \frac{19.5-32}{5.36}\right)$$
$$= \Phi\left(-2.33\right) = 0.0102 = 1\%$$

Above we read the z-scores from a z-table.

Lipsa memoriei unei variabile exponential distribuite
Fie X timpul scurs intre detectarea particulelor cu un contor Geiger si sa presupunem ca X are o distributie exponentiala cu $M(X) = 1.4$ minute. Aflati probabilitatea de a detecta o particula in primele 30 de secunde de la pornirea contorului. Sa presupunem ca am asteptat 3 minute fara sa fi detectat o particula. Care este probabilitatea sa detectam apoi o particula in urmatoarele inca 30 de secunde ?

Solutie: Pentru o variabila cu distributia exponentiala $X \sim Exp(\lambda)$ stim ca $M(X) = \frac{1}{\lambda}$. Prin urmare $\lambda = \frac{1}{1.4}$ si apoi probabilitatea de a detecta particula in primele 30 de secunde va fi estimata prin

$$P(X < 0.5) = \int_{-\infty}^{0.5} f(x)\ dx = \int_{0}^{0.5} \lambda e^{-\lambda x}\ dx = 1 - e^{-\frac{0.5}{1.4}} \approx 30\%$$

unde am folosit minutul ca unitate de masura si formula densitatii de probabilitate pentru variabilele exponential distribuite. Vom folosi pentru comparare valoarea exacta $1 - e^{-\frac{0.5}{1.4}}$ si nu cea aproximativa, afectata de erorile de aproximare.

Daca nu vom detecta nicio particula timp de trei minute, senzatia generala este ca probabilitatea de detectare ar trebui sa fie mai mare in cele 30 de secunde scurse dupa aceste trei minute. Insa vom demonstra matematica contrariu. Probabilitatea ceruta se exprima matematic prin $P(X < 3.5 | X > 3) =?$ Adica timpul scurs sa fie mai mic decat 3min 30 sec daca stim ca e sigur mai mare decat 3min. Conform formulei probabilitatilor conditionate

$$P(X < 3.5 \mid X > 3) = \frac{P(3 < X < 3.5)}{P(X > 3)}$$

caci consideram cele doua evenimente $X > 3$ si $X < 3.5$ iar intersectia lor se exprima prin evenimentul $3 < X < 3.5$. Folosind densitatea de probabilitate a distributiei exponentiale gasim

$$P(3 < X < 3.5) = \int_{3}^{3.5} \frac{1}{1.4} e^{-\frac{x}{1.4}}\ dx = -e^{-\frac{3.5}{1.4}} + e^{-\frac{3}{1.4}}$$

si

$$P(X > 3) = \int_{3}^{3.5} \frac{1}{1.4} e^{-\frac{x}{1.4}}\ dx = e^{-\frac{3}{1.4}}$$

In consecinta

$$P(X < 3.5 \mid X > 3) = \frac{-e^{-\frac{3.5}{1.4}} + e^{-\frac{3}{1.4}}}{e^{-\frac{3}{1.4}}} = 1 - e^{-\frac{0.5}{1.4}} = P(X < 0.5)$$

Aceasta lipsa de memorie reprezinta o proprietate specifica variabilelor exponential distribuite, fiind singurele variabile aleatoare continue cu aceasta proprietate, si poate fi exprimata general prin relatia

$$P(X < t_1 + t_2 \mid X > t_1) = P(X < t_2).$$

Proposed problems

Problem 1. *The probability density for the rolling amplitudes of a ship has the following form, according to Rayleigh's law:*

$$f(x) = \frac{x}{a^2} e^{-\frac{x^2}{2a^2}}, \qquad x \geq 0$$

Determine the expectation $E(X)$, the variance $var(X)$, the standard deviation $\sigma(X) = \sqrt{var(X)}$, the central moments of third and fourth order m_3 and m_4.

Problem 2. *A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of* 90 *km/hr and a standard deviation of* 10 *km/hr. What is the probability that a car picked at random is travelling at more than* 100 *km/hr?*

Problem 3. *Entry to Politehnica University is determined by a test. The scores on this test are normally distributed with a mean of 500 and a standard deviation of 100. John wants to be admitted to this university and he knows that he must score better than at least* 70% *of the students who took the test. John takes the test and scores 585. Will he be admitted to this university?*

Problem 4. *A person tosses 1000 coins. Find the chance (approximately) that the number of heads is between* 475 *and* 525, *inclusive.*

Problem 5. *X is a continuous uniform $(-5, 5)$ random variable*

i) What is the PDF $f_X(x)$?

ii) What is the CDF $F_X(x)$?

iii) What is $E[X^5]$?

iv) What is $E[e^X]$ *?*

Problem 6. *Consider the function:*

$$f(x) = \begin{cases} \frac{a}{x^2+4}, & \text{if } x \in (-2\sqrt{3}, 2\sqrt{3}) \\ 0, & \text{otherwise} \end{cases}$$

i) Find a *such that* f *is the probability density function of a continuous random variable* X

ii) Compute the mean $E(X)$ *and the variance* $var(X)$ *of this random variable*

iii) Find the distribution function $F(x)$ *and compute* $P(-2 < X < 2)$

Problem 7. *A football club provides a bus service for his fans. A bus arrives in a certain bus stop every* 15 *minutes between* 4 *and* 12 *pm during the matchday. Fans arrive at the bus stop at random times. The time that a fan waits is uniformly distributed from* 0 *to* 15 *minutes. How long will a fan tipically have to wait for a bus ? (i.e. the mean waiting time ?) What is the probability a fan will wait more than* 10 *minutes ? What is the probability a fan will wait between* 5 *and* 10 *minutes ?*

Problem 8. *The peak temperature* T*, as measured in degrees Fahrenheit, on a July day in New Jersey is a normal random variable* $\mathcal{N}(85, 100)$*. What is* $P(T > 100)$*,* $P(T < 60)$ *and* $P(70 \leq T \leq 100)$ *?*

Hint: use a standardization argument

Problem 9. *The tickets for the "Untold Festival" are sold online according to a Poisson distribution with a mean of* 25 *per day. What are the probabilities that:*

a) more than 20 *tickets are sold in one day?*

b) between 20 *and* 30 *tickets are sold in one day ?*

Hint: use a normal approximation

Problem 10. *The random variable* Y *has an exponential distribution with parameter* $\lambda = 0.2$*. Given the event* $A = \{Y < 2\}$*.*

i) What is the conditional $f_{Y|A}(y)$ *?*

ii) Find the conditional expected value $E[Y|A]$*.*

Bibliography

[1] R. Yates and D. Goodman. *Probability and Stochastic processes*, Wiley&Sons, 2005.

[2] D. Montgomery and G. Runger. *Applied Statistics and Probability for Engineers*, Wiley, 2014.

[3] C. Ariesanu. *Lecture Notes on Special Mathematics*, 2020.