MATH 320: Probability

Lecture 8: Distribution Functions

Chapters 2 and 3: Distributions (2.1 and 3.1)

Probability functions

Probabilities for discrete random variables

- Definition: The **probability mass function (pmf)** of a discrete random variable X is given by
$$f_X(x) = P(X = x), \quad \text{for all } x$$
- For a discrete random variable with a small number of outcomes, the pmf can be given in a table. When there is a very large or infinite number of possible outcomes, $f(x)$ can be given in a formula.
- Example: The number of injury claims per month is modeled by a random variable N with
$$P(N = n) = \frac{1}{(n+1)(n+2)}, \quad \text{where } n \geq 0.$$

n	0	1	2	...
$P(N=n)$	$1/2$	$1/6$	$1/12$	...

a) Determine the probability of two claims.

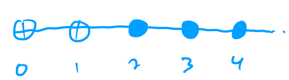
$$P(N=2) = \frac{1}{(2+1)(2+2)} = \frac{1}{3(4)} = 1/12$$

b) Determine the probability of at most two claims.

$$P(N \le 2) = P(N=0,1,2) = P(N=0) + P(N=1) + P(N=2) = 1/2 + 1/6 + 1/12 = 9/12$$

c) Determine the probability of at least two claims.

$$P(N \ge 2) = 1 - P(N < 2) = 1 - P(N \le 1) = 1 - [P(N=0) + P(N=1)] = 1 - [1/2 + 1/6] = 2/6$$

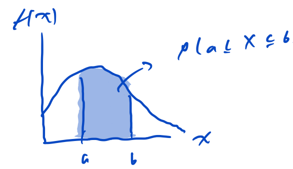


Probabilities for continuous random variables

- For continuous random variables, does the pmf exist? No!

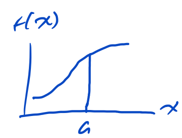
  These intervals are not countable, so we cannot use a pmf to assign probabilities.
- Instead, to find the general probability $P(a \le X \le b)$, we find the area bounded by $f(x)$ and the x-axis between $x = a$ and $x = b$ using integration.
- More specifically, we use a density function and find areas under the density function curve.
- Definition: A **probability density function (pdf)** is a continuous random variable X is a real-valued function that can be used to find probabilities using

$$P(a \le X \le b) = \int_a^b f(x)\,\mathrm{d}x$$



- Notes:
  - There is no probability associated with a single point:

    For $a \in \mathcal{X}$, $\quad P(X = a) = \displaystyle\int_a^a f(x)\,\mathrm{d}x = 0$



  - As a result: For any interval (a, b), it doesn't matter if we include or exclude the endpoints in the continuous case (unlike with discrete).

    For $(a, b) \in \mathcal{X}$, $\quad P(a < X < b) = P(a \le X \le b) = \displaystyle\int_a^b f(x)\,\mathrm{d}x$

    $a < X \le b$

    $a \le X < b$ (implied $a < b$)

- Example: Let $f(x) = \frac{1}{12}(x^2 + 1)$, $\quad$ for $0 \le x \le 3$.

  Find: (a) $P(X \le 1)$ (b) $P(X \ge 1)$ (c) $P(0.5 \le X \le 1.5)$.

(a)

$$\int_0^1 \frac{1}{12}(x^2+1)\,dx$$
$$= \frac{1}{12}\left[\frac{1}{3}x^3 + x\right]\Big|_0^1 \quad F(1) - F(0)$$
$$= \frac{1}{12}\left[\frac{1}{3} + 1\right]$$
$$= 1/9$$

(b)

$$\int_1^3 \frac{1}{10}(x^2+1)\,dx$$
$$= 1 - P(X \le 1)$$
$$= 1 - 1/9$$
$$= 8/9$$

(c)

$$\int_{1/2}^{3/2} \frac{1}{12}(x^2+1)\,dx$$
$$= \frac{1}{12}\left[\frac{1}{3}x^3 + x\right]\Big|_{1/2}^{3/2} \quad F(3/2) - F(1/2)$$
$$= \frac{1}{12}\left\{\frac{1}{3}\left(\frac{3}{2}\right)^3 + \frac{3}{2}\right\} - \frac{1}{12}\left[\frac{1}{3}\left(\frac{1}{2}\right)^3 + \frac{1}{2}\right]$$
$$\approx 0.174$$

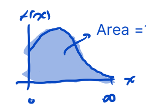
← Show work at least up to this step on Integrals, then can calc

Valid pmfs and pdfs

- There are rules that these new probability functions must follow, similar to the axioms our original probability assignments needed to satisfy.

- Theorem: A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only if
  (a) $f_X(x) \geq 0$ for all x. Non -negative
  (b) $\sum_x f_X(x) = 1$ (pmf) or $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ (pdf). Total probability

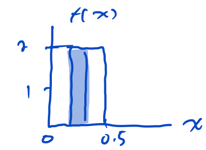
ex) if $x = [0,\infty)$



- Important note about $f(x)$ values.
  - For the discrete case, $f(x)$ values were actually probabilities.
    e.g. if $f(5) = 0.23$, there is a 23% probability of $x = 5$.
  - For the continuous case, the values of $f(x)$ are NOT probabilities themselves; they define areas which give probabilities.

    The values $f(x)$ must be positive, but they can be greater than 1.

    Example: Let $f(x) = 2, \quad 0 \leq x \leq 0.5$.



$P(.25 \leq x \leq .27) = 0.04$

- Note these are why continuous variables use density functions and not mass functions.
  Cannot assign probability to every single point $\iff$ Can't satisfy $\sum_{x \in \mathcal{X}} f(x) = 1$.

infinite range on $\mathbb{R}$ $x : [2, 5]$ 2 5

- Examples:
  1. Verify the pmf for X is valid.

x	1	2	3	4
$f_X(x)$	0.43	0.12	0.3	0.15

① all $f(x) \geq 0$ ✓

② $\sum f(x) = 1$ ✓

  2. Verify $f_X(x) = 2x^{-3}, \quad 1 \leq x < \infty$ is a valid pdf.

① $f(x) \geq 0$ ✓ over $[1, \infty)$

② $\int_1^\infty 2x^{-3}\,dx = -x^{-2}\Big|_1^\infty \; F(\infty) - F(1)$

$= \frac{1}{-\infty^2} + \frac{1}{1^2}$

$= 1$ ✓

  3. Let X have the following pdf:

$$f_X(x) = \begin{cases} cx^2 & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Find the value of c for which $f_X(x)$ is a valid pdf.

$\rightarrow 1 = \int_0^2 cx^2 = \frac{c}{3}x^3\Big|_0^2$

$= \frac{c}{3}[2^3 - 0^3]$

$1 = \frac{8}{3}c$

$\Rightarrow c = \frac{3}{8}$

$\rightarrow \int_0^2 \frac{3}{8}x^2 \overset{\checkmark}{=} 1$ + all $f(x) \geq 0$ over $[0, 2]$

The cumulative distribution function

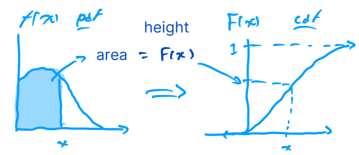
Concept of a cdf

- Examples: Discrete case

x	1	2	3	4
$f_X(x)$	0.43	0.12	0.3	0.15

Continuous case

$f_X(x) = \frac{3}{8}x^2 \quad 0 \le x \le 2$



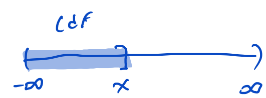
- In these examples, probabilities were obtained by cumulatively adding successive probabilities in the table above or by accumulating more probability as we increased the upper bound.

  If we do this throughout the entire range of either case, we obtain the cumulative distribution function $F(x)$. Every random variable X has an associated cumulative distribution function that can be defined as follows.

Defining a cdf

- Definition: The **cumulative distribution function** or **cdf** of a random variable X, denoted $F_X(x)$, is defined by

$$F_X(x) = P_X(X \le x), \quad \text{for all } x.$$



- Notes about $F(x)$: ALWAYS
  - The cdf is defined for $-\infty < x < \infty$ always.
  - The range of every cdf is $0 \le F(x) \le 1 \iff$ Limits: $\lim_{x\to-\infty} F(x) = 0$ + $\lim_{x\to\infty} F(x) = 1$
  - $F_X(x)$ is a non-decreasing function.

because if $\underset{x}{A} \subset \underset{x+\varepsilon,\ \varepsilon>0}{B}$, $P(A) \le P(B) \to F(x) \le F(x+\varepsilon)$

- DISCRETE case

  Example:

  (a) Using the pmf table for X below, find the cdf $F_X(x)$ as a table.

x	1	2	3	4
$f_X(x)$	0.43	0.12	0.3	0.15
$F_X(x)$	0.43	0.55	0.85	1

  (b) Write $F_X(x)$ as a piecewise function:

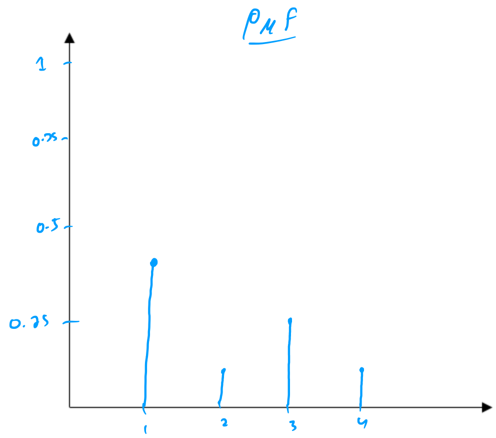
$$F_X(x) = \begin{cases} 0 & x < 1 \\ 0.43 & 1 \le x < 2 \\ 0.55 & 2 \le x < 3 \\ 0.85 & 3 \le x < 4 \\ 1 & x \ge 4 \end{cases}$$

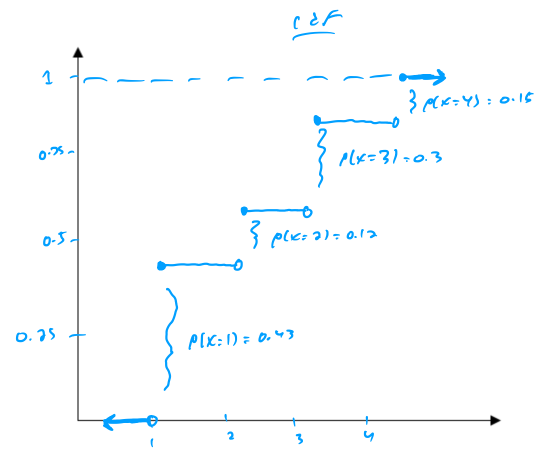
No probability until 12

Spike at 1, constant until 2

⋮

(c) Plot the pmf and cdf.





Observations of pmf — Properties of cdf

1) Sum of the heights = 1 $\Longrightarrow$ left limit is 0 & right limit is 1

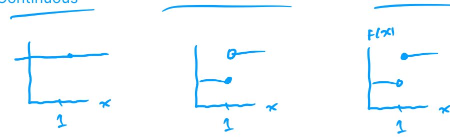
2) Positive values (probabilities) only at $x = 1, 2, 3, 4$ $\Longrightarrow$ Jump (discontinuity) at $x = 1, 2, 3, 4$

3) No probability between x values $\Longrightarrow$ Constant (horizontal) between x values (with value at next smallest x)

Properties

- The last entry in a table for $F(x)$ of a finite discrete random variable will always be 1.
- Even though $f(x)$ is only defined for certain values of x, we can define $F(x)$ for any real number.
- $F_X(x)$ is a right-continuous step-function.

Continuous — left continuous — Right Continuous



- CONTINUOUS case

Example: Let [redacted]

$$f(x) = \begin{cases} \frac{3}{8}x^2 & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

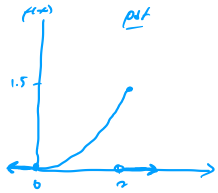
(a) Find the cdf $F(x)$.

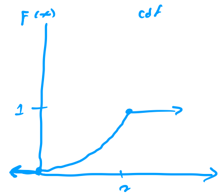
$$F_X(x) = \int_0^x \frac{3}{8} t^2 \, dt = \frac{1}{8} t^3 \Big|_0^x \quad F(x) - F(0)$$

$$= \frac{1}{8}x^3, \quad 0 \le x \le 2$$

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{8}x^3 & 0 \le x \le 2 \\ 1 & x \ge 2 \end{cases} \longrightarrow \text{check} \quad F_X(0) = \frac{1}{8} 0^3 = 0 \checkmark, \quad F_X(2) = \frac{1}{8} 2^3 = 1 \checkmark$$

(b) Plot the pdf and cdf.





Observations of cdf | Properties of cdf

1) Plot starts at ___0___ and ends at ___1___ $\Longrightarrow$ $0 \le F(x) \le 1$

2) ___Continuous___ at change points $\Longrightarrow$ continuous over $\mathbb{R}$

(not necessarily differentiable)

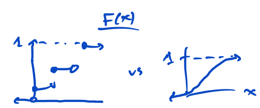
Properties

check

– F_X(lower limit) = 0 and F_X(upper limit) = 1

– $F_X(x)$ is always a continuous function (even though the pdf $f_X(x)$ does not necessarily have to be continuous over $\mathbb{R}$).

- Types of random variables (another way to define).



  – A random variable X is **discrete** $\Longleftrightarrow$ $F_X(x)$ is a step function of x.
  – A random variable X is **continuous** $\Longleftrightarrow$ $F_X(x)$ is a continuous function of x.

Relationship between the cdf and pdf

- Since $F(x)$ is defined by integrating $f(x)$, it is clear that the derivative of $F(x)$ is $f(x)$. This simple relationship is very important when the derivative $F'(x)$ exists.

$$\boxed{F'(x) = f(x)}$$

- Said another way: We can define the **pdf** of a continuous random variable X as the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t \qquad \text{for all } x.$$

Then using the Fundamental Theorem of Calculus, if $f_X(x)$ is continuous,

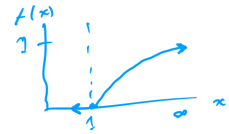
$$\frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = f_X(x)$$

- This relationship means that the pdf (or pmf) contains the same information as the cdf. So knowing one of these about a random variable in very important and allows researchers to analyze it many, many ways.

know pdf (or pmf) $\xLeftrightarrow[\text{same info}]{}$ know Cdf

- Example: Let X have the following cdf:

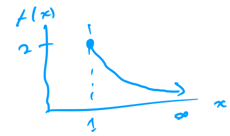
$$F(x) = P(X \le x) = 1 - \frac{1}{x^2}, \qquad 1 \le x < \infty$$



Find the pdf $f(x)$.

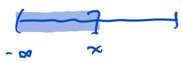
$f(x) = F'(x) = \frac{d}{dx}\left(1 - x^{-2}\right)$

$= 2x^{-3}, \quad x \ge 1$



Finding probabilities using the cdf

- ALWAYS
  - Cdf gives a <u>LEFT</u> probability.

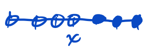


- DISCRETE case
  - Cdf adds all probabilities of points less than or equal to x. Formula version of this for a particular value $x = a$:

$$F(a) = P(X \le a) = \sum_{x \le a} f(x)$$

  - The complement of this is:

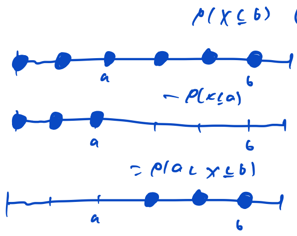
$$1 - F(x) = 1 - P(X \le x) = P(X > x)$$



which represents the probability X is greater than x, <u>exclusive</u>.

  - Interval probabilities: $P(a < X \le b) = P(X \le b) - P(X \le a) = F(b) - F(a)$

$P(a \le X \le b) = P(X \le b) - P(X \le a-1) = F(b) - F(a-1)$



- CONTINUOUS case
  - Cdf accumulates all of the probability less than or equal to x, which means we are finding the probability up to x. So this x the upper bound of the integral.

    To not confuse our letters, we change the letter of the variable in the function being integrated (and in the differential dt).

$$\boxed{F_X(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t}$$

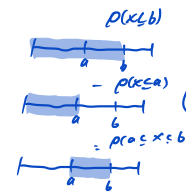
- For a specific value of $x = a$, we find probability with

$$F(a) = \int_{-\infty}^{a} f(x)\,\mathrm{d}x$$

- The complement of this is:

$$1 - F(a) = 1 - P(X \leq a) = P(x \geq a)$$

- Interval probabilities: $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$



$\leq, <, \geq, >$ doesn't matter for continuous

- Examples

1. Let X have the cdf table below.

x	1	2	3	4	5
$F_X(x)$	0.16	0.63	0.67	0.78	1.00

Find (a) $P(X \leq 2)$ (b) $P(X > 3)$ (c) $P(X \geq 3)$ (d) $P(X < 4)$

(a) $= F(2) = 0.63$

(b) $1 - P(x \leq 3)$
$= 1 - F(3)$
$= 1 - 0.67$
$= 0.33$

(c) $1 - P(x < 3)$
$= 1 - F(2)$
$= 1 - 0.63$
$= 0.37$

(d) $P(x \leq 3)$
$= F(3)$
$= 0.67$

(e) $P(2 \leq X \leq 4)$ (f) $P(1 < X \leq 5)$

(e) $F(4) - F(1)$
$= 0.78 - 0.16$
$= 0.62$

(f) $P(2 \leq x \leq 5)$
$= F(5) - F(1)$
$= 1 - 0.16$
$= 0.84$

2. Let $F_X(x) = \frac{x^3}{8}$, $\quad 0 \leq x \leq 2$.

Find (a) $P(X \leq 1)$ (b) $P(X < 1.5)$ (c) $P(X \geq 1.25)$ (d) $P(X > 0.75)$

(a) $F(1) = \frac{1}{8}$

(b) $F(1.5) = \frac{1.5^3}{8}$

(c) $1 - F(1.25)$
$= 1 - \frac{1.25^3}{8}$

(d) $1 - F(0.75)$
$= 1 - \frac{0.75^3}{8}$

(e) $P(0.25 \leq X \leq 1)$ (f) $P(1 < X \leq 1.5)$

(e) $F(1) - F(0.25)$
$= \frac{1^3}{8} - \frac{0.25^3}{8}$

(f) $F(1.5) - F(1)$
$= \frac{1.5^3}{8} - \frac{1^3}{8}$

Examples

- Discrete

1. The cdf for the years until patients are asymptomatic for a certain disease is shown below:

Number of years (x)	1	2	3	4	5
$F(x)$	0.53	0.78	0.9	0.97	1.00

a) Find $P(X < 4)$ and $P(X = 3)$.

↳ $= F(3) = 0.9$

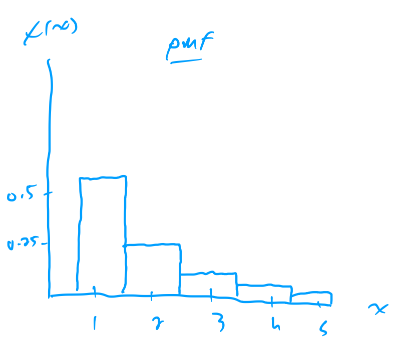
↳ $F(3) - F(2)$
$= 0.9 - 0.78$
$= 0.12$

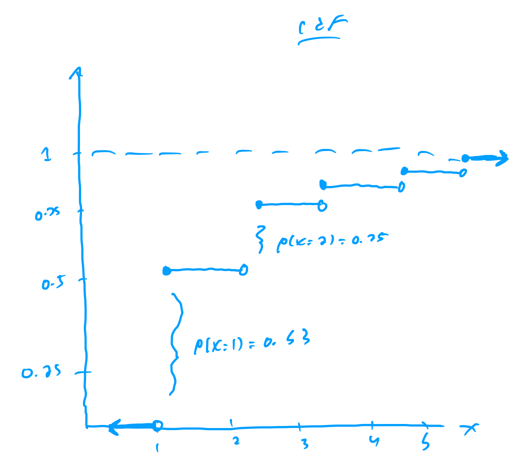
b) Write the piecewise functions of the pmf and cdf.

$$f(x) = \begin{cases} 0.53 & x = 1 \\ 0.25 & x = 2 \\ 0.12 & x = 3 \\ 0.07 & x = 4 \\ 0.03 & x = 5 \end{cases}$$

$$F_X(x) = \begin{cases} 0 & x < 1 \\ 0.53 & 1 \le x < 2 \\ 0.78 & 2 \le x < 3 \\ 0.90 & 3 \le x \le 4 \\ 0.97 & 4 \le x \le 5 \\ 1 & x \ge 5 \end{cases}$$

c) Plot a histogram of the pmf and a line graph of the cdf.



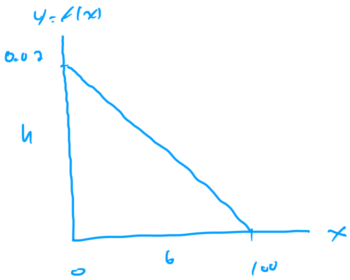


- Continuous

  Straight-line densities

  2. Suppose we are offering a warranty insurance policy which pays for repairs on a new appliance. We know from experience that repair costs X on a single policy will be on the interval $[0, 100]$, with probabilities highest for the lowest cost (\$0) and decreasing in a straight line fashion until x reaches \$100.

     (a) Find an appropriate density function.



$1 = \text{Area}$

$= \frac{1}{2} bh$

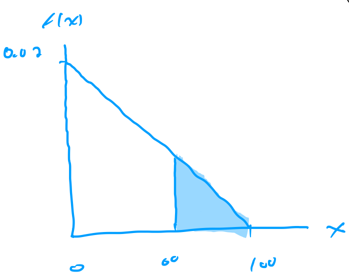
$= \frac{1}{2} 100 f(0)$

$f(0) = 0.02$

equation of line

$f(x) = 0.02 - 0.0002x, \quad 0 \le x \le 100$

     (b) Calculate $P(X > 60)$.



$P(X > 60) = \text{Area } \triangle$

$= \frac{1}{2}(100-60) f(60)$

$= \frac{1}{2} 40 (0.008)$

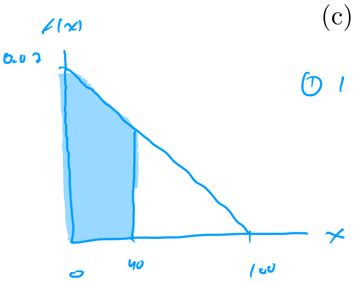
$= 0.16$

$\int_{60}^{100} f(x)\,dx = \int_{60}^{100} (0.02 - 0.0002x)\,dx$

$= 0.02x - 0.0001x^2 \Big|_{60}^{100}$

$= 1 - 0.84 = 0.16$

     (c) Calculate $P(X \le 40)$.



① $1 - \triangle \text{ area} = 1 - P(X \ge 40)$

$= 1 - \frac{1}{2} 60 \underbrace{f(40)}_{0.012}$

$= 1 - 0.36$

$= 0.64$

② Trapezoid area $= \frac{1}{2}(b_1 + b_2)h$

$= \frac{1}{2}(f(0) + f(40))\, 40$

$= \frac{1}{2}(0.02 + 0.012)\, 40$

$= 0.64$

③ Integral

$\int_0^{40} f(x)\,dx = 0.64$

     – Conclusion: For straight-line densities, it is usually easier to find probabilities as areas of trapezoids or triangles.

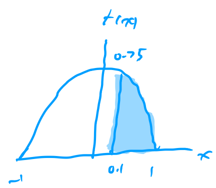
Symmetric densities

3. A risky investment has widely varying possible return percentages for next year. Best case: Return on investment (ROI) 100% (doubles money by getting money invested back plus 100% of the amount invested; Worst case: -100% (loses all money invested).

   The percentage return is a random variable X with that can be anywhere between the worst and best case, depending on the state of the economy. The pdf is:

$$f(x) = \begin{cases} 0.75(1-x^2) & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   (a) Find the probability the investor has a ROI greater than 10%.



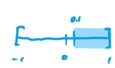
$$P(X > 0.1) = \int_{0.1}^{1} 0.75(1-x^2)\,dx$$
$$= 0.75x - 0.25x^3 \Big|_{0.1}^{1}$$
$$= (0.75 - 0.25) - (0.75(0.1) - 0.25(0.1)^3)$$
$$= 0.42525$$

   (b) Find the cdf $F(x)$; then find $P(X \geq 0.1)$.

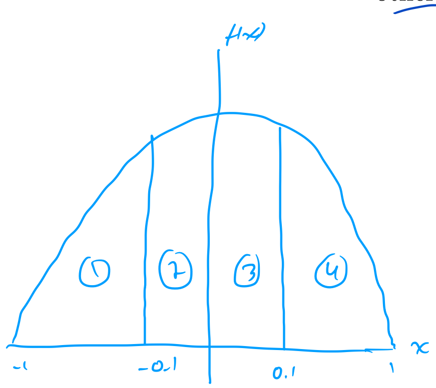
$$F_X(x) = \int_{-1}^{x} 0.75(1-t^2)\,dt = 0.75t - 0.25t^3 \Big|_{-1}^{x} \quad F(x) - F(-1)$$
$$= 0.75x - 0.25x^3 - (0.75(-1) - 0.25(-1)^3)$$
$$= 0.25(-x^3 + 3x + 2), \quad -1 \leq x \leq 1$$

check: $F(-1) = 0$ ✓, $F(1) = 1$ ✓

$$P(X \geq 0.1) = 1 - F(0.1)$$
$$= 1 - 0.25(-(0.1)^3 + 3(0.1) + 2)$$
$$= 1 - 0.57475$$
$$= 0.42525 \checkmark$$



– This is actually a symmetric density, which means we have information about other probabilities as well just from finding the probability in part (b).



$P(X \geq 0.1) =$ (4) $= 0.42525 =$ (1)

$P(0 < X < 0.1) =$ (3) $= 0.5 -$ (4) $= 0.07475 =$ (2)

$P(-0.1 < X < 0.1) =$ $P(|X| < 0.1) =$ (2) + (3) $= 2(0.07475) = 0.1495$

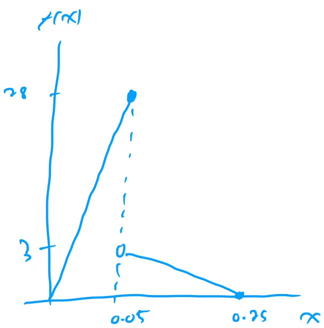
$P(\{X < -0.1\} \cup \{X > 0.1\}) =$ $P(|X| > 0.1) =$ (1) + (4) $= 1 - P(|X| < 0.1) = 1 - 0.1495 = 0.8505$

(2)+(3)

Piecewise densities

4. The density function for a continuous random variable can be defined piecewise and fail to be continuous at some points. Let X have the following pdf:

$$f_X(x) = \begin{cases} 0 & x < 0 \\ 560x & 0 \le x \le 0.05 \\ -15x + 3.75 & 0.05 < x \le 0.25 \\ 0 & x > 0.25 \end{cases}$$



(a) Show that the total probability is 1.

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

$$= \int_{0}^{0.05} 560x\,dx + \int_{0.05}^{0.25} (-15x + 3.75)\,dx$$

$$= \quad 28 \quad .05 \quad + \quad 3 \quad 0.2$$

$$= \tfrac{1}{2}(.05)(28) + \tfrac{1}{2}(3)(0.2)$$

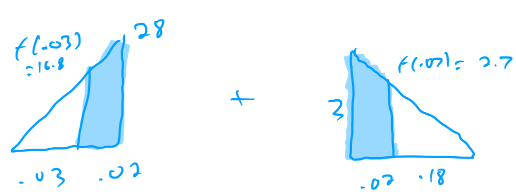
$$\checkmark = 0.7 + 0.3$$

Need separate integral for each function

straight line densities ⇒ use areas of shapes

(b) Find $P(0.03 \le X \le 0.07)$.

$$= \int_{0.03}^{0.05} 560x\,dx + \int_{0.05}^{0.07} (-15x + 3.75)\,dx$$



or $(0.7 - \quad)$ + or $(0.3 - \quad)$

$$(0.7 - \tfrac{1}{2}(0.03)(16.8)) + (0.3 - \tfrac{1}{2}(0.18)(2.7))$$

$$= 0.448 + 0.057$$

$$= 0.505$$

(c) Find the cdf $F_X(x)$.

STRATEGY: Find the cdf in cases.

Range   function

Case 1: $0 \le x \le 0.05 \Rightarrow f(x) = 560x$

$$\rightarrow P(X \le x) = \int_0^x 560t\,dt$$
$$= 280t^2\Big|_0^x$$
$$= 280x^2$$

Case 2: $0.05 < x \le 0.25 \Rightarrow f(x) = -15x + 3.75$

$$\rightarrow P(X \le x) = F(0.05) + \int_{0.05}^{x} (-15t + 3.75)\,dt$$
$$= 0.7 + \left[-7.5t^2 + 3.75t\Big|_{0.05}^{x}\right]$$
$$= \left[-7.5x^2 + 3.75x - \underbrace{\left(-7.5(.05)^2 + 3.75(.05)\right)}_{=0.16875}\right]$$
$$= -7.5x^2 + 3.75x + 0.53125$$

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 280x^2 & 0 \le x < 0.05 \\ -7.5x^2 + 3.75x + 0.53125 & 0.05 < x \le 0.25 \\ 1 & 0.25 < x \end{cases}$$