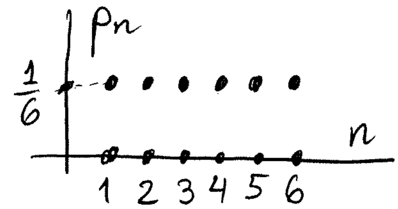
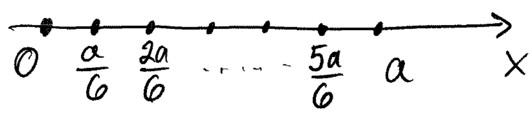
Lec. 7 – Probability Density Function

1 Discrete random variables

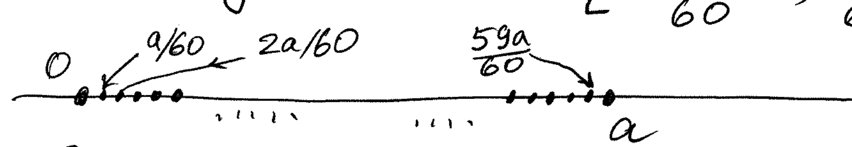
Suppose one rolls a fair die with numbers 1 through 6 on its faces. The probability that the die lands on any one number is: $p_n = 1/6$ for $n = 1, 2, \dots 6$.

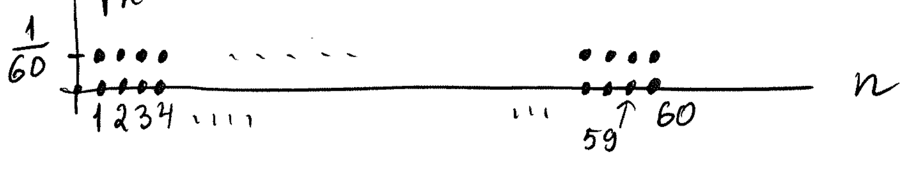


One can consider a different problem with a similar outcome. Suppose we take an interval $[0, a]$, divide it into 6 equal sections: $[0, \frac{a}{6}], [\frac{a}{6}, \frac{2a}{6}]$, etc., and cast the same die as before. If the die lands on face # n, we choose the n-th section of $[0, a]$ (i.e, section $[\frac{(n-1)a}{6}, \frac{na}{6}]$. The probability to choose any one section is the same number, $p_n = 1/6$.



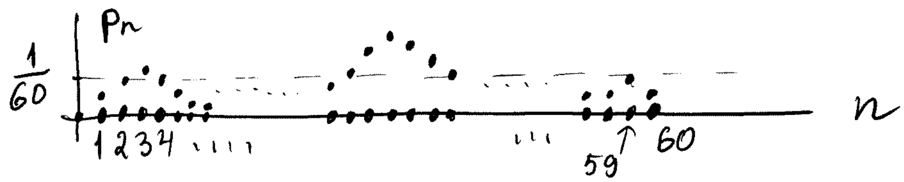
Now suppose that we divide that interval $[0, a]$ into 60 equal sections and cast a 60-face fair die (called **Pentakis dodecahedron**). The probability to choose any one section $[\frac{(n-1)a}{60}, \frac{na}{60}]$ is $p_n = \frac{1}{60}$.





<u>Questions:</u> What is the probability to choose (using the above 60-face die) <u>any one</u> of sections #27 ... #32? It is: $\underbrace{p_{27}+p_{28}+\dots+p_{32}}_{6} = 6\cdot\frac{1}{60} = \frac{1}{10}$.

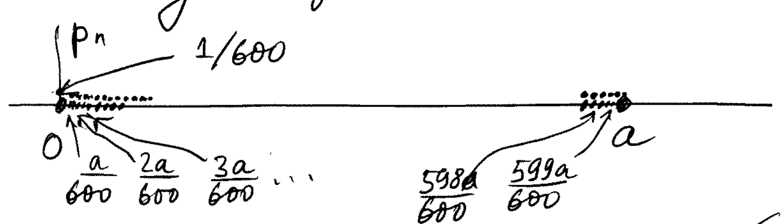
And what will the "probability diagram" be like if the die is <u>not</u> fair and so will land more often on some faces than on others? It will be non-uniform:



The probability to land in <u>any one</u> of sections 27...32 will still be: $p_{27\dots32} = p_{27}+p_{28}+\dots+p_{32}$.

② Passing from discrete to continuous random variables

If we increase the number of sections to, say, 600 (and use a 600-face die to choose a given section), the probability diagram will be (each $p_n = 1/600$):



<u>One notices two trends:</u>

1) The diagram becomes almost a continuous curve.

2) Its height becomes very small ($\frac{1}{600}$ in the above example).

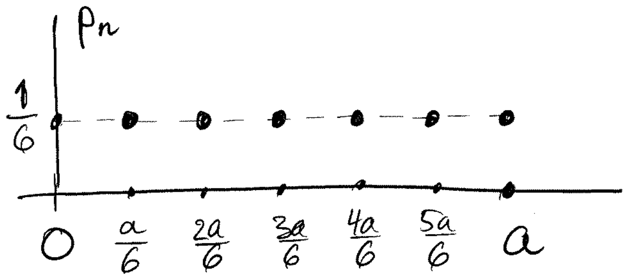
Trend 1) is <u>good</u>: we like continuous curves!

But Trend 2) is <u>bad</u>: the points get too close to 0 and become difficult to see them.

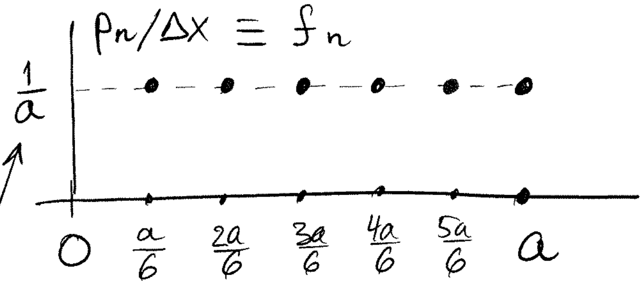
To counteract that bad trend, we can plot <u>not</u> p_n, but $p_n / \Delta x$, where Δx is the length of each section.

To see this, let's start with a <u>6-face die</u> (= cube).

<u>Before dividing by Δx</u>



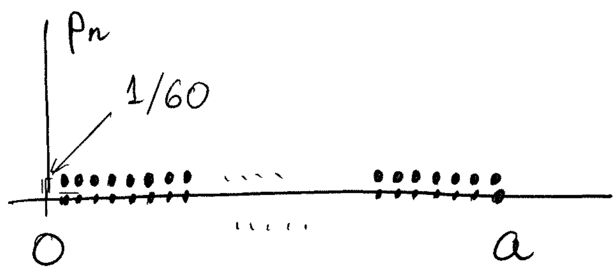
<u>After dividing by Δx</u>



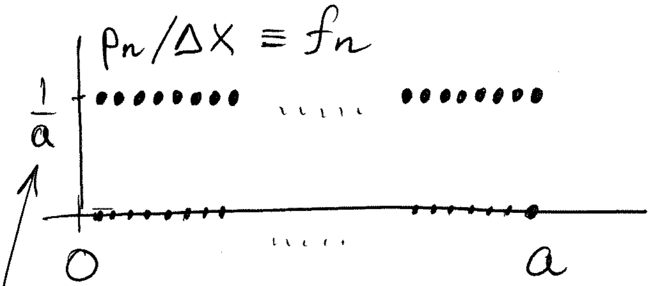
$$\frac{1/6}{\Delta x} = \frac{1/6}{a/6} = \frac{1}{a}$$

Repeat this for 60 sections:

<u>Before dividing by Δx</u>



<u>After dividing by Δx</u>



$$\frac{1/60}{\Delta x} = \frac{1/60}{a/60} = \frac{1}{a}$$

So we see that as we divide $[0, a]$ into finer and finer sections, the height of the curve does not change.

This is convenient!

Moreover, we see that:

$$(f_1+f_2+\dots+f_{60})\cdot\Delta X = \left(\frac{p_1}{\Delta X}+\frac{p_2}{\Delta X}+\dots+\frac{p_{60}}{\Delta X}\right)\cdot\Delta X$$

$$= \frac{p_1+p_2+\dots+p_{60}}{\cancel{\Delta X}}\cdot\cancel{\Delta X} = \underbrace{p_1+p_2+\dots+p_{60}}_{\text{probability to "get" in any one section of } [0,a].} = 1$$

We will use the fact that

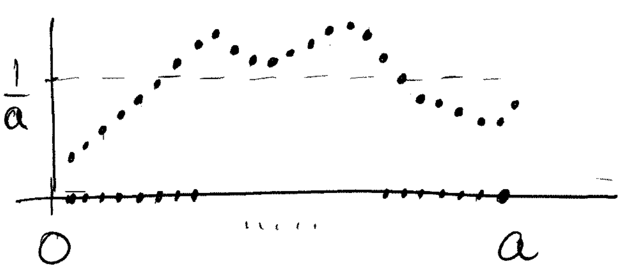
$$\boxed{(f_1+f_2+\dots+f_{60})\cdot\Delta X = 1}$$

later.

More Questions:

- What if the die is not fair (and we plot $p_n/\Delta x$)?

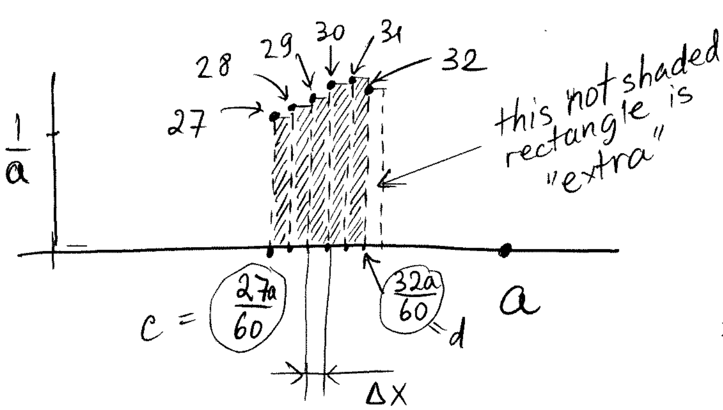
$p_n/\Delta X \equiv f_n$



The scaled probabilities $f_n = p_n/\Delta X$ still cluster around $1/a$.

- What is the probability that we "get" into sections 27 ... 32? It is still $(p_{27}+p_{28}+\dots+p_{32})$, but we can now write it differently:

$$p_{27}+p_{28}+\dots+p_{32} = \left(\frac{p_{27}}{\Delta X}+\frac{p_{28}}{\Delta X}+\dots+\frac{p_{32}}{\Delta X}\right)\Delta X = (f_{27}+f_{28}+\dots+f_{32})\Delta X$$



Area of rectangle n is $(f_n\cdot\Delta X)$.

$$(f_{27}+f_{28}+\dots+f_{31}+f_{32})\Delta X$$

$$= \underbrace{(f_{27}+\dots+f_{31})\Delta X}_{\text{term 1}} + \underbrace{f_{32}\cdot\Delta X}_{\text{term 2}}$$

term 1 $= (f_{27}+f_{28}+\dots+f_{31})\Delta X \approx \int_{c=\frac{27a}{60}}^{d=32a/60} f(x)\,dx$

if we consider the collection of f_n as a continuous curve.

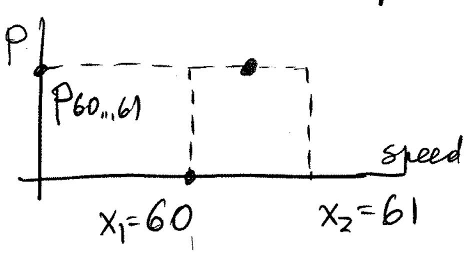
term 2 $= \underbrace{f_{32}}_{\text{number} \sim 1/a} \cdot \underbrace{\Delta X}_{\to 0} \approx 0$

Thus, the probability to "get" into sections 27,..,32 is:

$$(f_{27}+\dots+f_{31}+f_{32})\Delta X \approx (f_{27}+\dots+f_{31})\Delta X \approx \underbrace{\int_c^d f(x)\,dx}_{\substack{\text{area under curve} \\ y=f(x) \text{ over interval } [c,d].}}$$

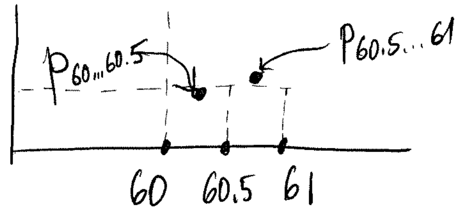
③ Continuous random variables

Ex. 1 Suppose X is the speed of a given car as it drives on a highway. We can ask: What is the probability that the car's speed is between $x_1=60$ and $x_2=61$ mph? The answer is some number $P_{60\dots61}$.



We can also ask: What is the probability to find the car's speed between 60 & 60.5 mph? Clearly,

$$P_{60\dots61} = P_{60\dots60.5} + P_{60.5\dots61},$$



and so $P_{60\dots60.5} \approx P_{60.5\dots61} \approx \frac{1}{2} P_{60\dots61}$.

We can keep asking about the probability of finding the car's speed within progressively narrow intervals, say, [60.3, 60.4] mph, and we'll get progressively smaller values, because:

$$p_{60\ldots61} = p_{60\ldots60.1} + p_{60.1\ldots60.2} + \ldots + p_{60.9\ldots61} \Rightarrow$$

$$p_{60\ldots60.1} \approx p_{60.1\ldots60.2} \approx \ldots p_{60.9\ldots61} \approx \frac{1}{10}\, p_{60\ldots61}.$$

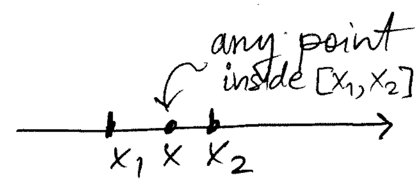
The Bottom Line is:

As the length Δx of interval $[x_1, x_2]$ gets smaller (think of $[x_1, x_2]$ shrinking from [60, 61] to [60.3, 60.4]), **the probability $P(x_1 \le \underline{X} \le x_2)$ that value of $\underline{X}$ is found within $[x_1, x_2]$ gets proportionally smaller.**

However, the probability density function (PDF)

$$f(x) = \frac{P(x_1 \le \underline{X} \le x_2)}{\underbrace{x_2 - x_1}_{\Delta x}}$$

(where x is any value inside $[x_1, x_2]$:



tends to a finite number as $\Delta x \to 0$.

Moreover,

$$\boxed{P(c \le \underline{X} \le d) = \int_c^d \underbrace{f(x)}_{\text{PDF of } \underline{X}}\, dx}$$

as we showed at the end of topic 2 (pp. 7-4, 5).

Difference between $\underline{X}$ and x:

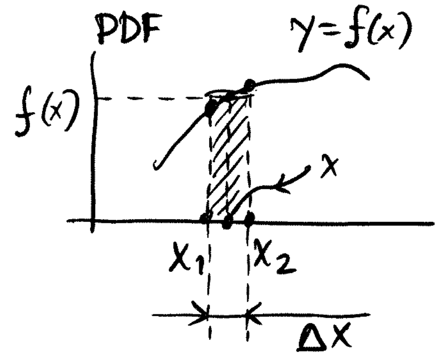
$\underline{X}$ is the random variable (like the speed of a car);

x is any value in the range of values that $\underline{X}$ can take.

(4) Properties of a PDF $f(x)$

1. $f(x) \geq 0$ for all x

2a. Probability that X is found in a small interval $[x_1, x_2]$ of length Δx is:



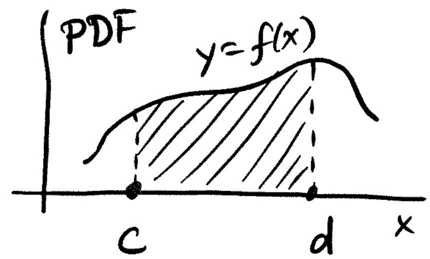
$$P(x_1 \leq X \leq x_2) \approx f(x)\,\Delta x$$

where x is any value in $[x_1, x_2]$.

(E.g., $P(60.1 \leq \text{speed} \leq 60.2) \approx f(\underbrace{60.15}_{\text{inside } [60.1,\,60.2]}) \cdot \overbrace{0.1}^{60.2-60.1}$.)

2b. Probability that X in any interval $[c,d]$ is:

$$\boxed{P(c \leq X \leq d) = \int_c^d f(x)\,dx}$$

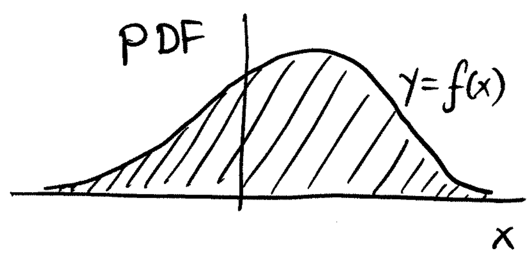


3. Probability to find X somewhere in $(-\infty, \infty)$ is 100%:

$$P(-\infty \leq X \leq \infty) = 1$$

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

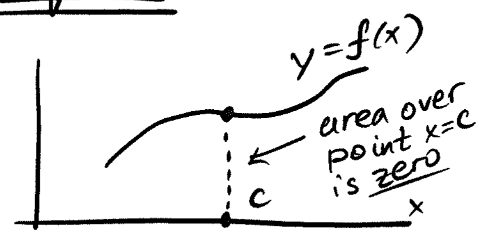
(see p. 7-4)



4. Probability that a continuous random variable X takes on precisely a value c is always 0.

$$P(X=c) \equiv P(c \leq X \leq c) = 0,$$

or $\quad \int_c^c f(x)\,dx = 0 .$



Ex. 2 (= Ex. 11. 2. {1+2} in textbook)

Given the PDF
$$f(x) = \begin{cases} 12x^2 - 12x^3, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$
perform these tasks:

(a) Graph $f(x)$ and verify that it satisfies Property 1 of a PDF : $f(x) \geq 0$

(b) Verify that it satisfies Property 3 of a PDF: $\int_{-\infty}^{\infty} f(x)dx = 1$.

(c) Find $P(0.4 \le X \le 0.7)$

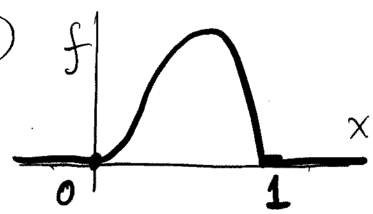
(d) Find $P(X = 0.4)$

(e) Find $P(0.4 < X \le 0.7)$

(f) Find $P(X \geq 0.6)$

(g) Find $P(X \le 0.5)$.

Sol'n: (a)



Use a graphic calculator.
Property 1 : $f(x) \geq 0$
yes.

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

1) Note that $\int_{-\infty}^{\infty} = \int_{-\infty}^{0} + \int_{0}^{1} + \int_{1}^{\infty}$

2) Since $f(x) = 0$ for $x \le 0$, $\int_{-\infty}^{0} f(x)\,dx = 0$

Since $f(x) = 0$ for $x \geq 1$, $\int_{1}^{\infty} f(x)dx = 0$

So $\int_{-\infty}^{\infty} f(x)dx = 0 + \int_{0}^{1} f(x)dx + 0 = \int_{0}^{1} f(x)dx$.

3) $\int_0^1 f(x)\,dx = \int_0^1 (12x^2 - 12x^3)\,dx = \left(\frac{12}{3}x^3 - \frac{12}{4}x^4\right)\Big|_0^1$

$= 4 \cdot (1^3 - 0^3) - 3(1^4 - 0^4) = 1.$ ✓

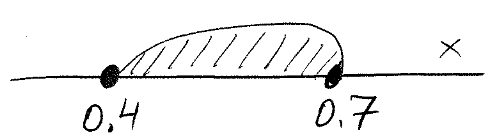
By combining steps 1)–3), we've verified Property 3.

(c) Find $\underline{P}(0.4 \le X \le 0.7) \overset{\text{Prop. 2b}}{=} \int_{0.4}^{0.7} (12x^2 - 12x^3)\,dx$

$= \left(\frac{12}{3}x^3 - \frac{12}{4}x^4\right)\Big|_{0.4}^{0.7} = 4(0.7^3 - 0.4^3) - 3\cdot(0.7^4 - 0.4^4) \approx 0.47$

(d) Find $\underline{P}(X = 0.4)$. This $= 0$ by Property 4.

(e) Find $\underline{P}(0.4 < X \le 0.7)$



Clearly:

$$\underbrace{\underline{P}(0.4 \le X \le 0.7)}_{\substack{\text{probability to land}\\ \text{on } [0.4,\,0.7],\text{ i.e.}\\ \text{INcluding the}\\ \text{end point } 0.4}} = \underbrace{\underline{P}(0.4 < X \le 0.7)}_{\substack{\text{probability to land}\\ \text{on } (0.4,\,0.7],\text{ i.e.}\\ \text{EXcluding the}\\ \text{end point } 0.4}} + \underbrace{\cancelto{0 \text{ by (d)}}{\underline{P}(X = 0.4)}}_{\substack{\text{probability to}\\ \text{land exactly}\\ \text{on } X = 0.4}}$$

$$\Rightarrow \quad \underline{P}(0.4 < X \le 0.7) = \underline{P}(0.4 \le X \le 0.7).$$

In general: $\boxed{\underline{P}(c \le X \le d) = \underline{P}(c < X \le d) = \underline{P}(c \le X < d) = \underline{P}(c < X < d)}$

for any c, d. I.e., excluding or including the end point(s) does __not__ change the probability to land on a given interval.

(f) $\underline{P}(X \ge 0.6) = \int_{0.6}^{\infty} f(x)\,dx = \int_{0.6}^{1} f(x)\,dx + \int_{1}^{\infty} \cancelto{0}{f(x)}\,dx$

$= \int_{0.6}^{1} (12x^2 - 12x^3)\,dx + 0 = \left(\frac{12}{3}x^3 - \frac{12}{4}x^4\right)\Big|_{0.6}^{1} \approx 0.52$

(g) $\underline{P}(X \le 0.5) = \int_{-\infty}^{0.5} f(x)\,dx = \int_{-\infty}^{0} \underbrace{f(x)}_{\to 0}\,dx + \int_{0}^{0.5} f(x)\,dx$

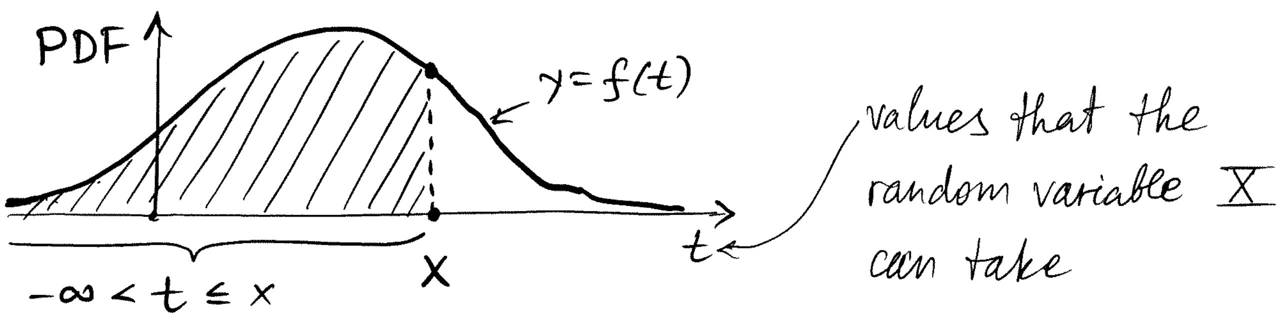
$= \left(\frac{12}{3}x^3 - \frac{12}{4}x^4\right)\Big|_{0}^{0.5} \approx 0.31$

⑤ Cumulative Distribution Function

[a] Motivation and Definition

Since each time that we compute the probability of a continuous random variable X we use $\int_c^d f(x)\,dx$, then it is convenient to give a name to some anti-derivative of $f(x)$. The convention is to name the particular anti-derivative $\int_{-\infty}^{x} f(t)\,dt$.

<u>Note:</u> Since x is now the limit of integration, we <u>cannot</u> <u>use it</u> <u>inside</u> the integral and therefore use another name, t, for the variable there.



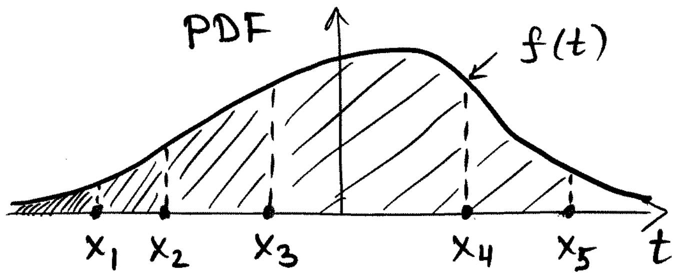
$\mathbb{P}(X \le x)$ ← probability that variable X (think of X as being the speed of a car) does not exceed a value x (think of x as being 60 mph)

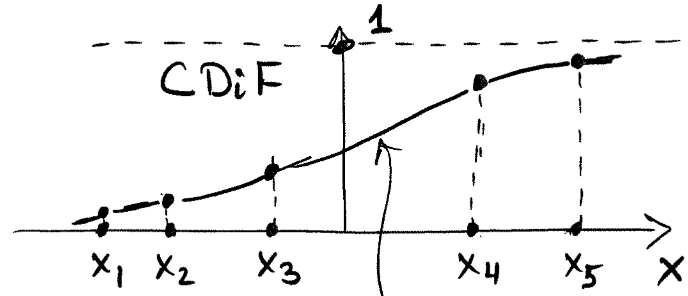
$\|$

$\int_{-\infty}^{x} f(t)\,dt \equiv F(x)$ ← CUMULATIVE Distribution FuncTION, <u>CDiF</u>.

$\boxed{b}$ Properties of the CDiF

Scenario 1: PDF $\neq 0$ for all finite values (t)

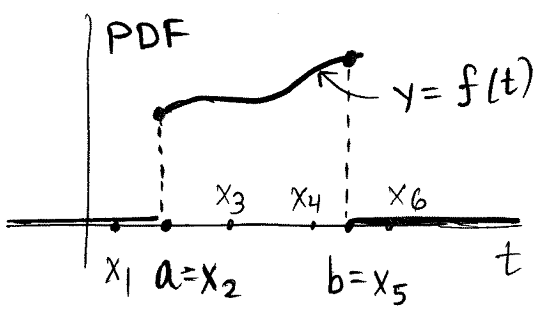


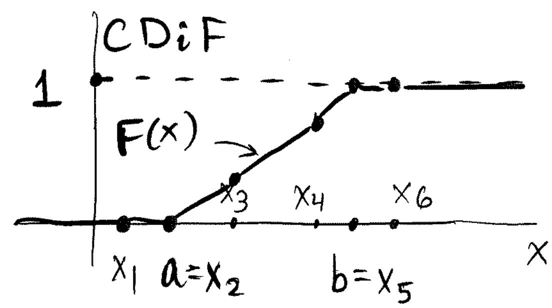


$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

ALWAYS: $\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1$

Scenario 2: PDF $\neq 0$ on a finite interval $[a,b]$





$$f(t) \neq 0, \quad a \le t \le b$$
$$= 0 \quad \text{otherwise}$$

$F(x_1) = 0$ since $\int_{-\infty}^{x_1} \underbrace{f(t)}_{\to 0\ (x_1 < a)}\,dt = 0$

$F(x_2) = 0 = \int_{a}^{x_2(=a)} f(t)\,dt = 0$

(Property 4 of PDF)

$F(x_3) = \int_{a}^{x_3} f(t)\,dt > 0$ (positive area under $y = f(t)$ over $[a, x_3]$)

$F(x_4) = \int_{a}^{x_4} f(t)\,dt = \left(\int_{a}^{x_3} + \underbrace{\int_{x_3}^{x_4}}_{>0 \text{ since } f(t) \ge 0}\right) f(t)\,dt > F(x_3)$

$F(x_5) = \int_{a}^{x_5 = b} f(t)\,dt = 1$ by Property 3 of PDF

$$F(x_6) = \int_a^{x_6} f(t)\,dt = \int_a^{b} f(t)\,dt + \int_b^{x_6} \underbrace{f(t)}_{=0 \text{ for } x_6 > b}\,dt$$

$$= 1 + 0 = 1.$$

Based on these examples, we formulate the following

Properties of CDiF:

If f is a PDF and $F(x) = \int_{-\infty}^{x} f(t)\,dt$ is the associated CDiF, then:

1. $F'(x) = f(x)$

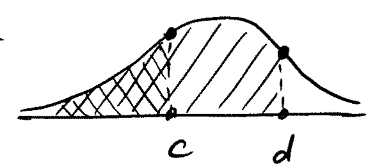
   (If $F(x)$ is an anti-derivative of $f(x)$, then the derivative of $F(x)$ must be $f(x)$.)

2. $0 \le F(x) \le 1$

3. $F(x)$ never decreases as a function of x

   (it can either increase or stay constant; see Scenario 2)

4. Since $\underbrace{\int_c^d f(t)\,dt}_{P(c \le X \le d)} = \underbrace{\int_{-\infty}^{d} f(t)\,dt}_{F(d)} - \underbrace{\int_{-\infty}^{c} f(t)\,dt}_{F(c)}$



then: $\boxed{P(c \le X \le d) = F(d) - F(c)}$

__Ex. 3__ ($\approx$ Ex. 11.2.4 in textbook)

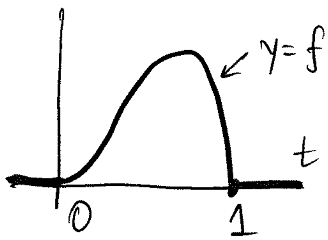
(a) Find the CDiF for the PDF in Ex. 2, i.e.

for $f(t) = \begin{cases} 12t^2 - 12t^3, & 0 \le t \le 1 \\ 0 & \text{otherwise.} \end{cases}$

(b) Sketch the PDF and CDiF.

(c) Using the CDiF, compute $P(0.4 \le X \le 0.7)$.

Sol'n: (a) It is actually a good idea to always sketch the PDF before computing the CDiF.



$\underline{x < 0}$

$$F(x) = \int_{-\infty}^{x} \underbrace{f(t)}_{=0 \text{ for } x \le 0} dt = 0.$$

$\underline{0 \le x \le 1}$

$$F(x) = \int_{-\infty}^{x} f(t)dt = \cancel{\int_{-\infty}^{0} f(t)dt}^{\,0} + \int_{0}^{x} f(t)\,dt$$

$$= \int_0^x (12t^2 - 12t^3)\,dt = \left(\frac{12}{3}t^3 - \frac{12}{4}t^4\right)\Big|_0^x = 4x^3 - 3x^2.$$

$\underline{x > 1}$

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{1} f(t)dt + \int_{1}^{x} \underbrace{f(t)}_{=0 \text{ for } x > 1} dt$$

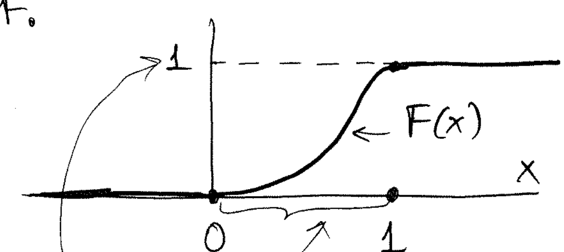
$$= \int_{-\infty}^{1} f(t)dt + 0 = F(1) + 0$$

$$= 4\cdot 1^3 - 3\cdot 1^4 + 0 = 1$$

(this agrees with Property 3 of PDF; see also Scenario 2 on pp. 7-11, 12.)

(b) We've already sketched the PDF.

Now sketch the CDiF.



Note: this particular PDF $\neq 0$ on $0 \le t \le 1$. On the other hand, any CDiF satisfies $F(x \to \infty) = 1$.

(c) $P(0.4 \le X \le 0.7) \overset{\text{Property 4 of CDiF}}{=} F(0.7) - F(0.4)$

$= (4\cdot 0{,}7^3 - 3\cdot 0{,}7^4) - (4\cdot 0{,}4^3 - 3\cdot 0{,}3^4) \approx 0.47$ (same as in Ex. 2(c))

Ex. 4 (= Ex. 11.2.{5 & 3} in textbook)

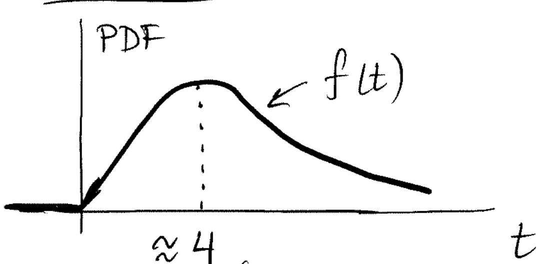
The shelf life of a drug is a continuous random variable X with a PDF
$$f(t)=\begin{cases}100t/(t^2+50)^2, & t\ge 0\\ 0, & t<0,\end{cases}$$
where t is the time in months. (This means that the probability that the drug goes bad between times $t=c$ and $t=d$ is: $\int_c^d f(t)\,dt$.

(a) Find the CDF.

(b) Sketch it

(c) Suppose that the pharmacist wants to be 95% certain that the drug is still good when it is sold. How long is it safe to keep the drug on the shelf before selling it (to satisfy the above condition)?

Sol'n: (a) As in Ex. 3, let us sketch the PDF first.



Incidentally, this PDF shows that the drug most likely (= max PDF) will go bad after about 4 months on the shelf.

$x<0$ $\quad F(x)=\int_{-\infty}^{x} f(t)\,dt = 0$ since $f(t)=0$ for $t\le x<0$.

$x\ge 0$ $\quad F(x)=\int_{-\infty}^{x} f(t)\,dt = \cancel{\int_{-\infty}^{0}}\to 0 + \int_0^x f(t)\,dt = \int_0^x \frac{100t\,dt}{(t^2+50)^2}$

$u=t^2+50$

$du=(t^2+50)'\,dt = 2t\,dt$

$t\,dt=\frac{1}{2}du$

$u(t=0)=0^2+50=50$

$u(t=x)=x^2+50$

$$=\int_{50}^{50+x^2}\frac{100\cdot\frac{1}{2}du}{u^2} =$$

$$= 50 \int_{50}^{50+x^2} u^{-2}\,du = 50 \cdot \frac{u^{-1}}{-1}\Bigg|_{50}^{50+x^2} =$$

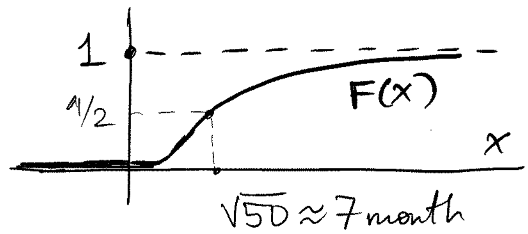
$$= -50\left(\frac{1}{50+x^2} - \frac{1}{50}\right) = \cancel{-50} \cdot \frac{\cancel{50} - (\cancel{50}+x^2)}{\cancel{50} \cdot (50+x^2)}$$

$$= \cancel{-} \frac{\cancel{-}x^2}{50+x^2} = \frac{x^2}{50+x^2} \leftarrow F(x),\ x \geq 0.$$

(and $F(x)=0$, $x<0$)

(b) Use a graphic calculator:

CDiF $\rightarrow$



It is a good idea to verify that

$\lim_{x\to\infty} F(x) = 1$ (since $\int_{-\infty}^{\infty} f(t)\,dt = 1$).

Try $x = 100$: $F(100) = \dfrac{100^2}{50+100^2} \approx 0.995$.

(c) Note that the PDF and the CDiF are of the event that the drug goes <u>bad</u>. So, the probability that by time $=x$ the drug goes <u>bad</u> is $F(x)$.

Then the probability that by time $=x$ the drug <u>is still good</u> must be $(1-F(x))$.

Since we want that by time x the drug be still good,

$1-F(x)=0.95 \leftarrow$ the certainty that the pharmacist wants.

$F(x) = 1-0.95 = 0.05$

$$\frac{x^2}{50+x^2} = 0.05 \Rightarrow x^2 = 50\cdot 0.05 + x^2 \cdot 0.05 \Rightarrow$$

$$x^2(1-0.05) = 50\cdot 0.05 \Rightarrow 0.95x^2 = 2.5 \Rightarrow x = \sqrt{\frac{2.5}{0.95}}$$

$x \approx 1.6$ month $\leftarrow$ **The drug must be sold within 1.6 month.**