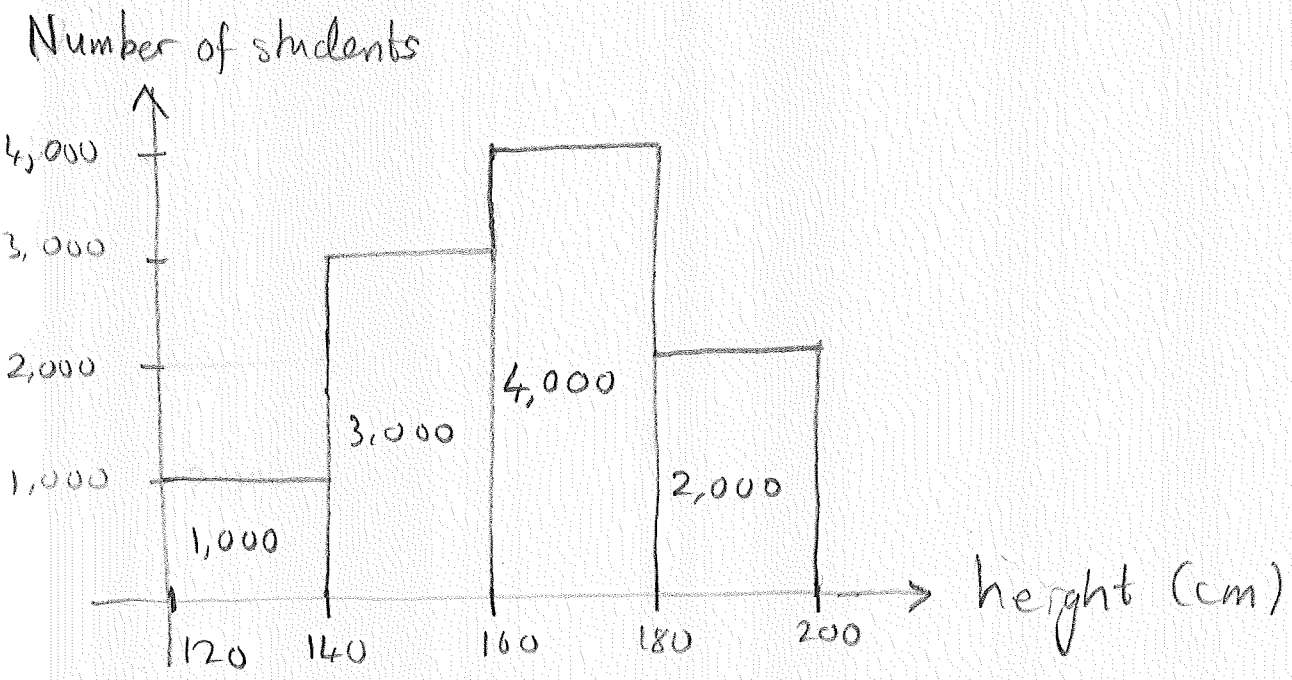
§ 8.7, 8.8 Distribution Functions, Densities, Probability, Mean and Median

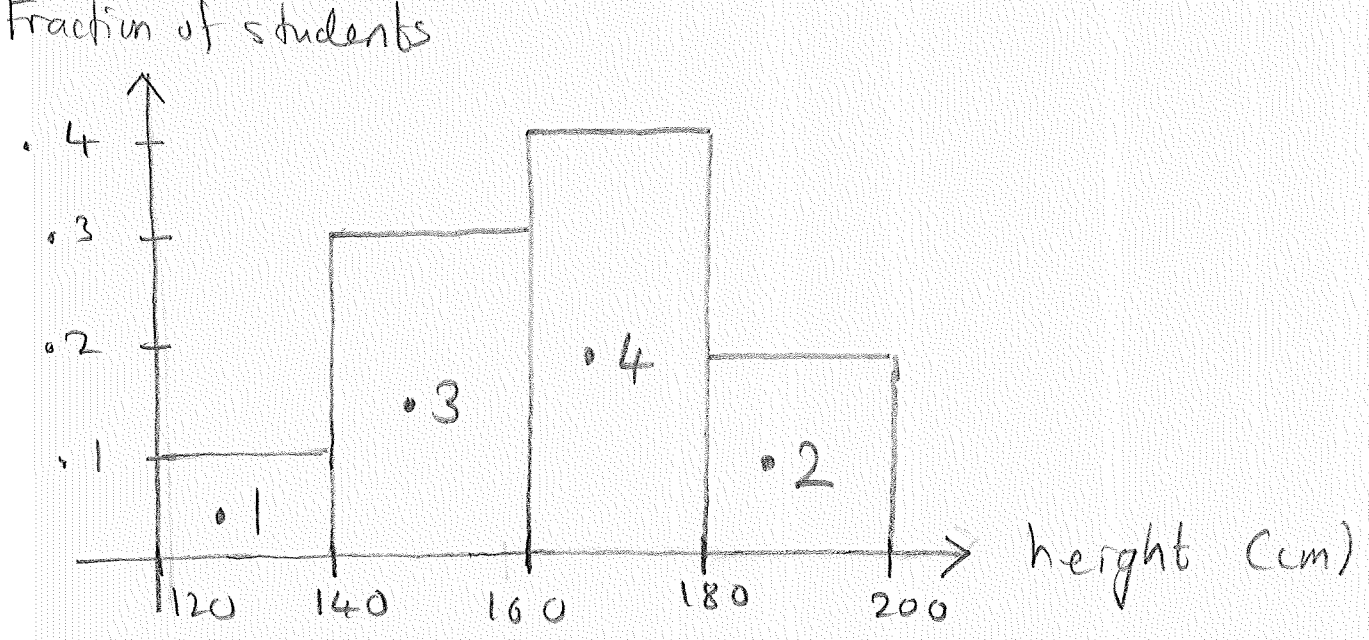
Consider the following graph which represents the heights of URI students (for the sake of simplicity, we'll assume URI has exactly 10,000 students).



From this it is easy to see that, for example 4,000 = 40% of students are between 160cm and 180cm tall while 3,000 + 4,000 = 70% of students are between 140cm and 180cm tall.

So if we picked a student at random (all students being equally likely to be chosen), we would see that the chance the student is between 160cm and 180cm tall is 40% while the chance the student is between 140cm and 180cm tall is 70%.

A better way to do this is to redo the graph with the fraction of the total number of students on the vertical axis instead of the actual number.



So the probability (chance) a randomly chosen student is between 160cm and 180cm tall is .4 while the chance the student is between 120cm and 180cm tall is .3 + .4 = .7

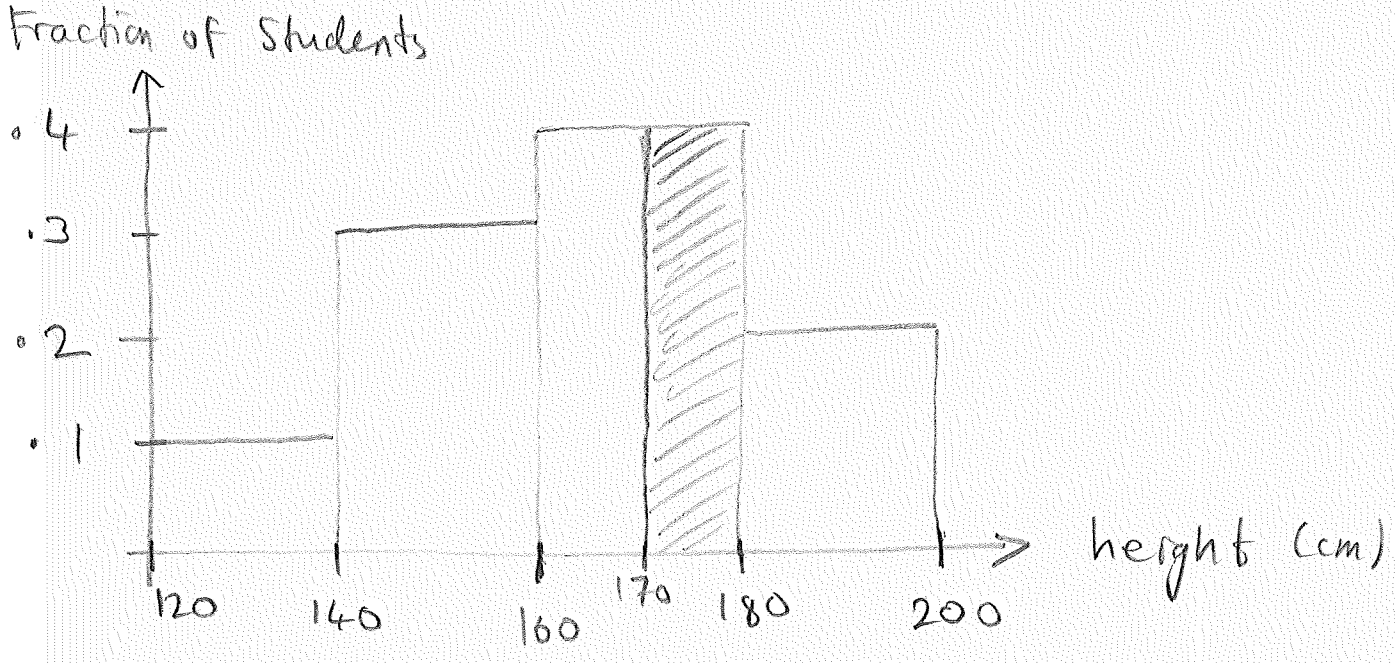
Note that the total probability is

$$.1 + .3 + .4 + .2 = 1 \quad (100\%)$$

as we'd expect.

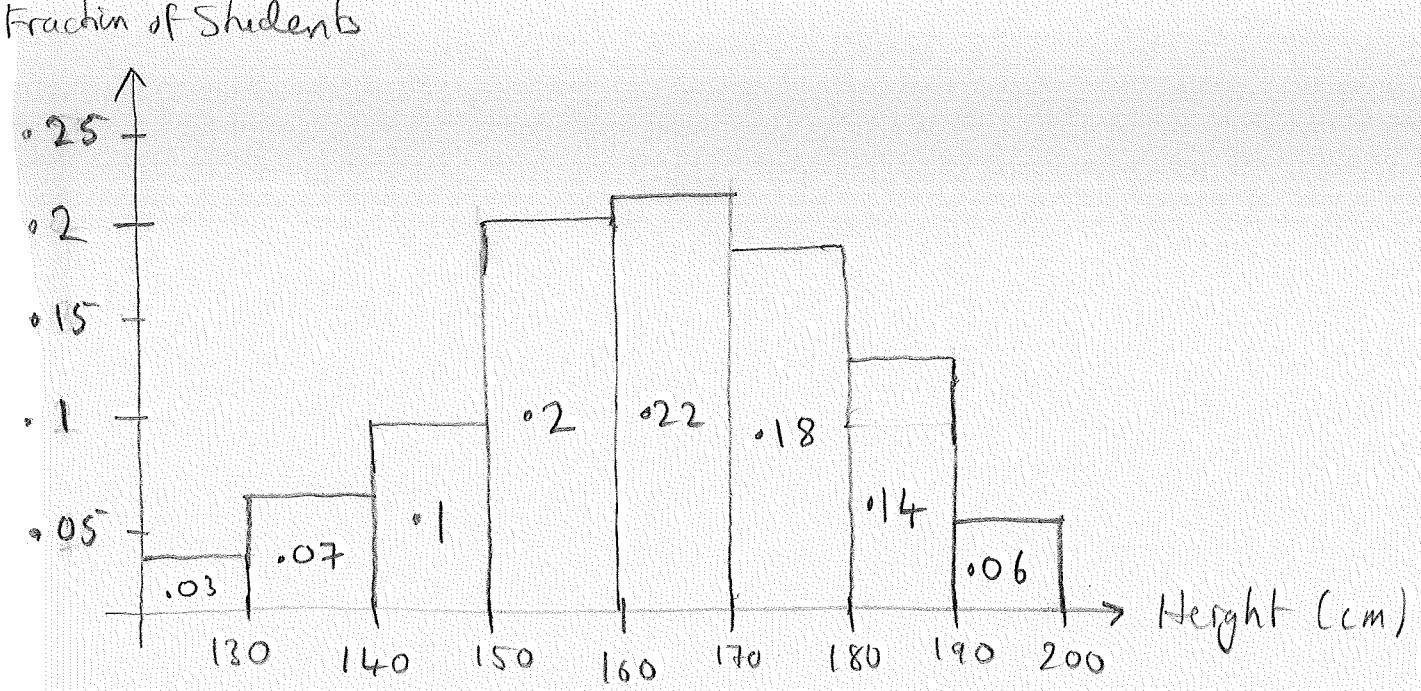
What about the probability the student is between 170 cm and 180 cm tall?

We don't have enough information to say exactly. However, since .4 of the students are between 160 cm and 180 cm, a reasonable approximation would be $\frac{1}{2} \times .4 = .2$



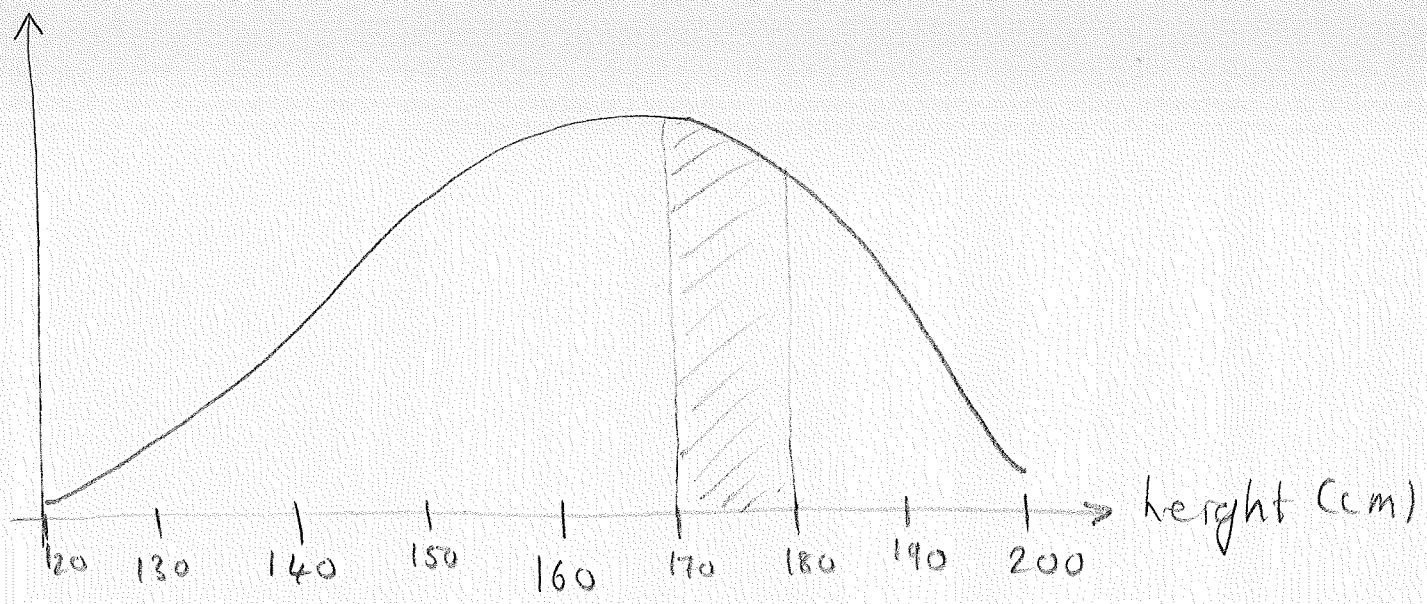
Notice that what we're really doing is finding the area under a graph.

Of course, what we really need is a graph with more detailed data



Notice that here the probability a student is between 160cm and 180cm = .22 + .18 which is still .4. However the probability the student is between 170cm and 180cm is .18.

Of course, we could go on dividing up our intervals again and again. We would end up with a curve which looked something like



The total area under the curve would be 1 and the probability a student has a height in a given range would be found by integrating the curve over this range to find the area.

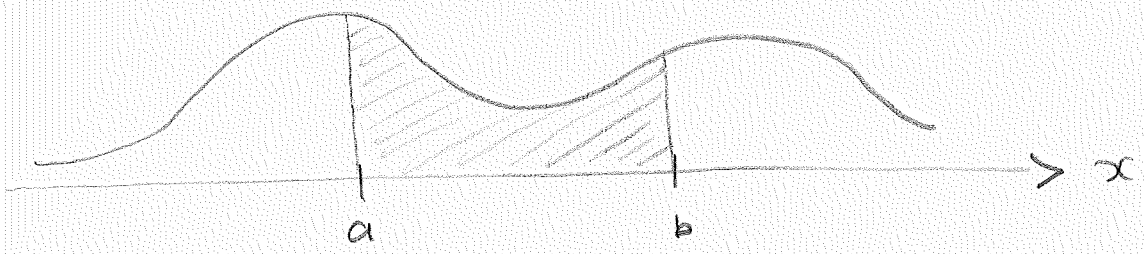
This motivates the following definition

A function $p(x)$ is a <u>probability density function</u> (pdf) for some quantity x associated with a population if

$$p(x) \geq 0 \quad \forall x \in \mathbb{R}, \quad \int_{-\infty}^{\infty} p(x)\,dx = 1$$

and

Fraction of pop. for which x is between a and b = Area under graph of p between a and b = $\int_a^b p(x)\,dx$.



IMPORTANT NOTE!

The value of $p(x)$ at some value a does NOT measure the probability x has value a.

In fact, for any density p and any a, the prob. that x has value a is always 0 (e.g. what is the chance a student is exactly 180 cm tall).

$p(x)$ is never used just on its own to calculate probabilities. Instead it is always integrated

e.g. $\int_a^b p(x)\,dx$.

Ex. The outcome of a certain scientific experiment is random and governed by the density

$$p(x) = \frac{1}{\pi(1+x^2)} \qquad \text{(Cauchy distribution)}$$

Find the prob. that the outcome is between -1 and 1

A.

$$\int_{-1}^{1} p(x)\,dx = \frac{1}{\pi}\int_{-1}^{1} \frac{1}{1+x^2}\,dx$$

$$= \frac{1}{\pi}\left[\arctan x\right]_{-1}^{1}$$

$$= \frac{1}{\pi}\left(\frac{\pi}{4} - \left(-\frac{\pi}{4}\right)\right)$$

$$= \frac{\frac{\pi}{2}}{\pi}$$

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

Ex. I'm waiting outside my house for the RIPTA bus.

a) For which value of k could the fn.

$$p(t) = \begin{cases} 0, & t < 0 \\ k e^{-3t}, & t \geq 0 \end{cases} \qquad (t \text{ in hours})$$

serve as a density which models the (random) amount of time I'm waiting at the bus stop?

b) What is the prob. I must wait at least 20 minutes?

Soln. Since we need $p(t) \geq 0 \ \forall t \in \mathbb{R}$, we certainly require that $k \geq 0$. (Note that the fact that $p(t) = 0$ for $t < 0$ reflects the fact that I cannot wait for a negative amount of time!)

The other requirement is that

$$\int_{-\infty}^{\infty} p(t)\,dt = 1.$$

Here, since $p = 0$ for $t < 0$, this means that

$$\int_0^\infty k e^{-3t} dt = 1$$

ie

$$\lim_{b \to \infty} \int_0^b k e^{-3t} dt = 1$$

$$\lim_{b \to \infty} \left[-\frac{k}{3} e^{-3t} \right]_0^b = 1.$$

$$\lim_{b \to \infty} \left(-\frac{k}{3} e^{-3b} + \frac{k}{3} \cdot 1 \right) = 1$$

$$\frac{k}{3} = 1$$

So $k = 3$.

b) 20 min $= \frac{1}{3}$ hr, so the desired prob. is given by

$$\int_{\frac{1}{3}}^\infty 3e^{-3t} dt = \lim_{b \to \infty} \int_{\frac{1}{3}}^b 3e^{-3t} dt$$

$$= \lim_{b \to \infty} \left[-e^{-3t} \right]_{\frac{1}{3}}^{b}$$

$$= \lim_{b \to \infty} \left(-e^{-3b} + e^{-1} \right)$$

$$= e^{-1} = \frac{1}{e} \approx .3679$$

Cumulative Distribution Functions

Often we are interested in the fraction of the population for which x is less than (or equal to) some value t. Using the density $p(x)$, this gives us a fn $P(t)$ defined by

$$P(t) = \int_{-\infty}^{t} p(x)\, dx.$$

$P(t)$ is called the cumulative distribution function (CDF) of the density p.

So $P(t)$ is got from $p(x)$ by integrating.

Note that in view of the Second Fundamental Theorem of Calculus (§ 6.4, Thm 6.2, P. 299), P is an antiderivative of p, ie.

$$P'(t) = \frac{d}{dt}\left(\int_{-\infty}^{t} p(x)\,dx\right) = p(t).$$

Since $p(x) \geq 0$ everywhere, it follows that $P(t)$ is non-decreasing, ie.

if $a < b$, then $P(a) \leq P(b)$.

Also easy to see is that

$$\lim_{t \to -\infty} P(t) = 0, \quad \lim_{t \to \infty} P(t) = 1$$

e.g. with the students - nobody's shorter than $-\infty$ cm, but everybody's shorter than $+\infty$ cm.

Finally, we have that

$$\text{Fraction of pop. having values of } x \text{ between } a \text{ and } b = \int_a^b p(x)\,dx$$

$$= P(b) - P(a)$$

Finally, using the First Fundamental Theorem of Calculus

$$\text{Fraction of pop. having values of } x \text{ between } a \text{ and } b = \int_a^b p(x)\,dx$$

$$= P(b) - P(a) \quad \text{as } P \text{ is an antid. of } p.$$

Ex. For the experiment whose outcome had density

$$p(x) = \frac{1}{\pi(1+x^2)},$$

the general antid of $p(x)$ is (of course!)

$$F(t) = \frac{1}{\pi} \arctan t + C.$$

To find the CDF, we need the (unique) antid P which satisfies

$$\lim_{t \to -\infty} P(t) = 0, \quad \lim_{t \to \infty} P(t) = 1.$$

Now

$$\lim_{t \to -\infty} \frac{1}{\pi} \arctan t = \frac{1}{\pi} \cdot \frac{-\pi}{2} = -\frac{1}{2}$$

while

$$\lim_{t \to \infty} \frac{1}{\pi} \arctan t = \frac{1}{\pi} \cdot \frac{\pi}{2} = +\frac{1}{2}.$$

Thus if we let $C = \frac{1}{2}$ and set

$$P(t) = \frac{1}{\pi} \arctan t + \frac{1}{2}$$

we have the CDF of $\rho(x)$.

Ex. Find the CDF for the waiting for the bus problem and use it to calculate the prob. I have to wait for between 20 minutes and 30 minutes.

The density was given by

$$p(t) = \begin{cases} 0, & t < 0 \\ 3e^{-3t}, & t \geq 0 \end{cases}$$

An antid. of this is given for $t \geq 0$ by

$$P(t) = \int_0^t p(s)\,ds = \int_0^t 3e^{-3s}\,ds$$

$$= \left[-e^{-3s}\right]_0^t$$

$$= -e^{-3t} - (-1)$$

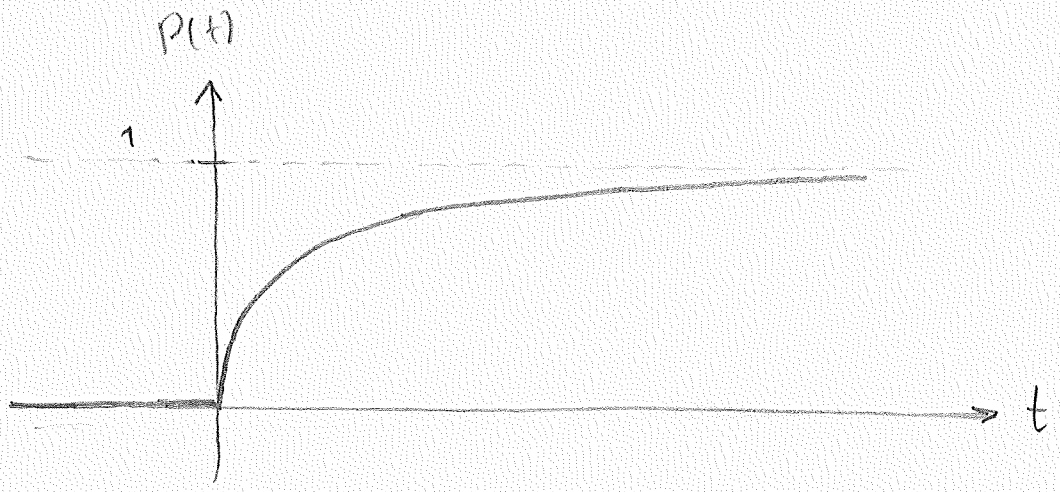
$$= 1 - e^{-3t}$$

If we then set

$$P(t) = \begin{cases} 0, & t < 0 \\ 1 - e^{-3t}, & t \geq 0 \end{cases}$$

then clearly $\lim_{t \to -\infty} P(t) = 0$ and $\lim_{t \to \infty} P(t) = 1$

so that we have the CDF.



The prob. I'm waiting between 20 min and 30 min is then given by

$$P(\tfrac{1}{2}) - P(\tfrac{1}{3}) = 1 - e^{-3/2} - (1 - e^{-\frac{3}{3}})$$

$$= e^{-1} - e^{-1.5} \approx .1447$$

The Median and the Mean

Both of these measure the 'average' value of a distribution, but in different ways.

The Median

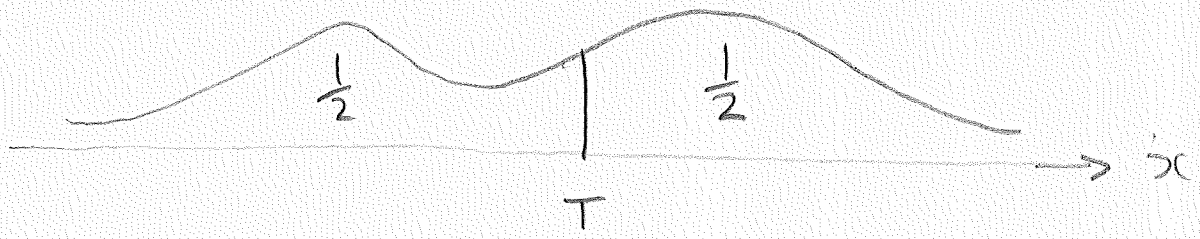
The median of a quantity x distributed through a population is a value T s.t. half the pop. has values of x less than (or equal to) T and half the pop. has values greater than (or equal to) T. Thus, the median satisfies

$$\int_{-\infty}^{T} p(x)\,dx = 0.5$$

where p is the pdf. In terms of the CDF, P, this can be restated as

$$P(T) = .5.$$

Physically, one can think of the median as the place to cut the density so that half the total area lies on each side of the cut.



Ex. For the URI students, in the first version the median height is somewhere between 160cm and 180cm.

In the second version with more data, the median height is somewhere between 160 cm and 170 cm.

Ex For the bus example, find the median waiting time.

Recall that the CDF was given by

$$P(t) = \begin{cases} 0, & t < 0 \\ 1 - e^{-3t}, & t \geq 0 \end{cases}$$

We want the value T where $P(T) = .5$.

$$1 - e^{-3T} = .5$$

$$-e^{-3T} = -.5$$

$$e^{-3T} = .5$$

Take ln of both sides

$$\ln(e^{-3T}) = \ln(.5)$$

$$-3T = \ln(.5)$$

$$T = \frac{\ln(.5)}{-3} = \frac{\ln 2}{3} \approx 0.23 \text{ hours}$$

$$\approx 13.86 \text{ min}.$$

The Mean

This is a more common way of finding the average value of some quantity. For example, to calculate the average height of a UR1 student from our graph, we could add up all the (approximate) heights of the students and then divide by the total number of students.

One way of doing this (using a left-hand sum) would give

$$\frac{(120)(1,000) + (140)(3,000) + (160)(4,000) + (180)(2,000)}{10,000}$$

$$= 154\text{cm}.$$

Another way of doing the same calculation is to divide by 10,000 first before adding to get

$$120(.1) + 140(.3) + 160(.4) + 180(.2) = 154 \text{ cm}.$$

Note that this is the same answer we would have got if we'd used the version of the graph with probabilities, or the fraction of the total number of students on the vertical axis.

If instead we use the version with more data where we had 10cm intervals, we'd get an average height of

$$120(.03) + 130(.07) + 140(.1) + 150(.2) + 160(.22) + 170(.18) + 180(.14) + 190(.06)$$

$$= 159.1 \text{ cm}.$$

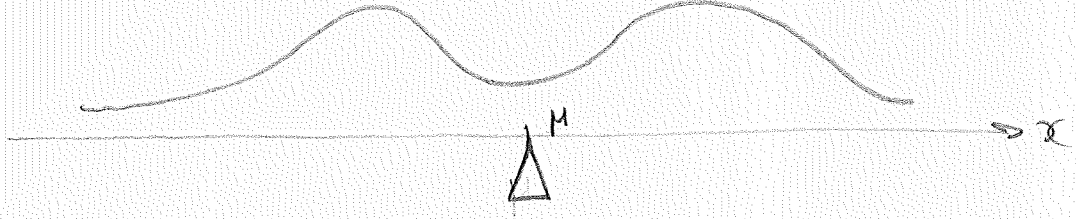
Clearly the second approximation ought to be more accurate than the first. In the cts. case, these Riemann sums will become integrals and we get the following defn.

The <u>mean</u> μ of a density ρ is given by

$$\mu = \int_{-\infty}^{\infty} x\,\rho(x)\,dx$$

provided this integral converges.

Note the strong resemblance to the formula for centre of mass in § 8.4. In fact, one can think of μ as the place to put a fulcrum so that the distribution balances



On the other hand, the centre of mass of a body can be thought of as the 'mean position' of a body.

Ex. The distribution $p(x) = \frac{1}{\pi(1+x^2)}$ of the experiment example doesn't have a mean at all. The reason for this is that both

$$\int_{-\infty}^{0} \frac{x}{\pi(1+x^2)}\,dx \quad \text{and} \quad \int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,dx$$

diverge and so therefore does

$$\int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\,dx.$$

Ex Find the mean for the waiting for the bus example.

Recall that $p(t) = \begin{cases} 0, & t < 0 \\ 3e^{-3t}, & t \geq 0 \end{cases}$

$$\mu = \int_{-\infty}^{\infty} t\, p(t)\, dt = \int_{0}^{\infty} 3t e^{-3t}\, dt$$

$$= \lim_{b \to \infty} \int_{0}^{b} 3t\, e^{-3t}\, dt.$$

To find this integral, we do integration by parts with

$$u = 3t, \quad dv = e^{-3t}\, dt$$
$$du = 3\, dt, \quad v = -\tfrac{1}{3} e^{3t}$$

so that

$$\int_{0}^{b} 3t e^{-3t}\, dt = \left[(3t)(-\tfrac{1}{3} e^{3t})\right]_{0}^{b} - \int_{0}^{b} -\tfrac{1}{3} e^{3t} \cdot 3\, dt$$

$$= \left[-te^{-3t}\right]_0^b + \int_0^b e^{-3t}\,dt$$

$$= \left[-te^{-3t}\right]_0^b + \left[-\frac{1}{3}e^{-3t}\right]_0^b$$

$$= -be^{-3b} - (0) + \left(-\frac{1}{3}e^{-3b} - \left(-\frac{1}{3}\right)\right)$$

$$= -be^{-3b} - \frac{1}{3}e^{-3b} + \frac{1}{3}.$$

Now $\lim_{b\to\infty} -be^{-3b} = \lim_{b\to\infty} \frac{-b}{e^{3b}} = \lim_{b\to\infty} \frac{-1}{3e^{3b}} = 0$

by l'Hôpital's rule.

Thus

$$\mu = \lim_{b\to\infty} \int_0^b 3te^{-3t}\,dt = \lim_{b\to\infty} \left(-be^{-3b} - \frac{1}{3}e^{-3b} + \frac{1}{3}\right)$$

$$= 0 + 0 + \frac{1}{3}$$

$$= \frac{1}{3} \text{ hours } (= 20 \text{ min})$$

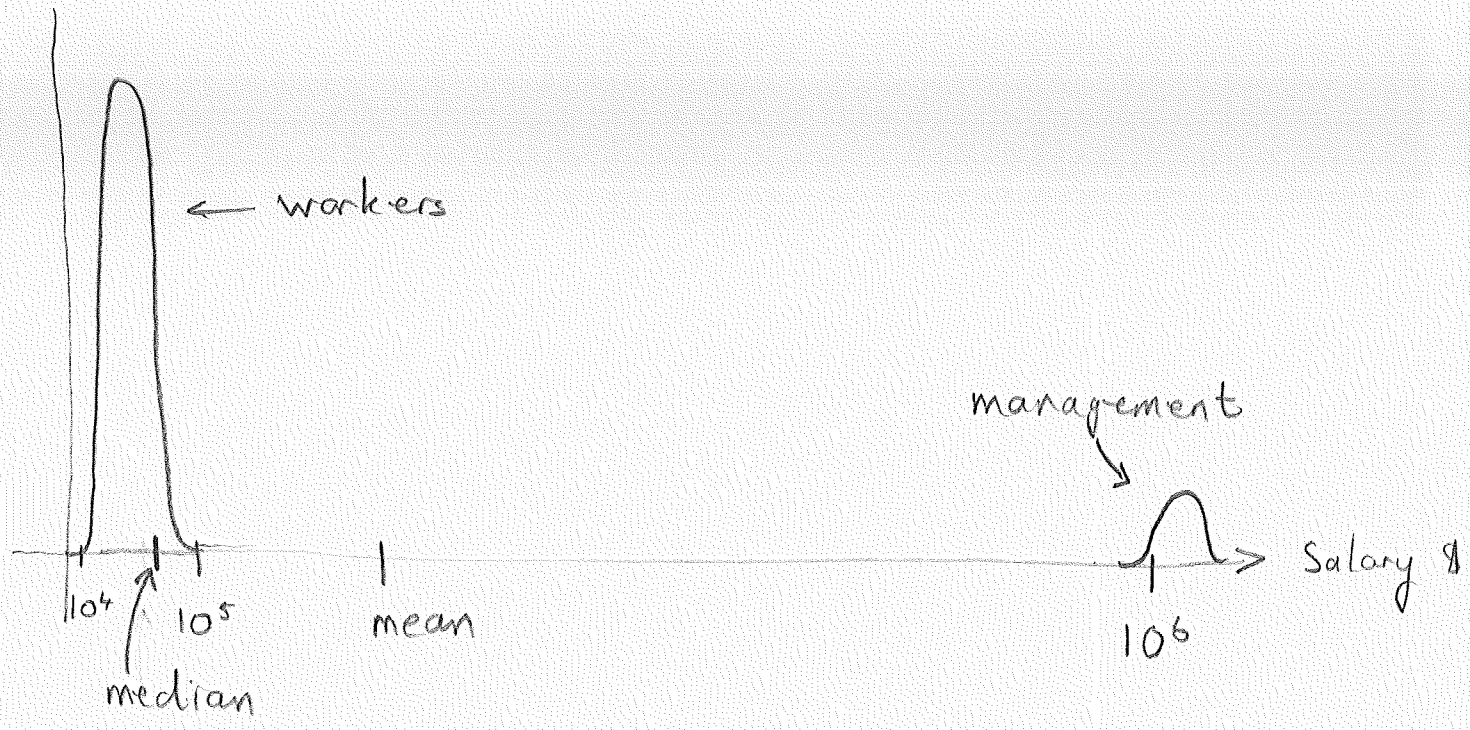
Distinguishing between Median and Mean

Consider the example of a large company where the management is paid a great deal more than the workers.

If we use the mean to calculate the average salary at the company, there is a good chance this will work out to be a fairly healthy amount. Even if the workers are poorly paid, the management will make a large contribution to the mean since, although few in numbers, they make so much.

On the other hand, the median salary is likely to be much lower as there are much more workers than management so most of the area of the distribution will be concentrated around them.

The picture of the income distribution at such a company might look something like

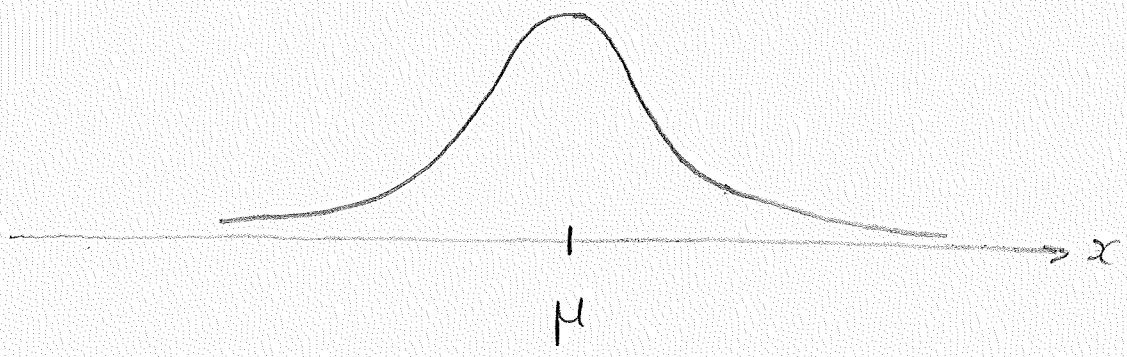


This is the reason why statistics concerning the incomes of groups of people are generally compiled using the median!

Normal Distributions

Many quantities are distributed according to the normal distribution

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$



Here μ is the mean of the distribution and σ is the standard deviation (a measure of how 'spread out' the distribution is), where $\sigma > 0$.

Ex. The yearly rainfall (in inches - includes snow) in Anchorage, AK is normally distributed with mean $\mu = 15$ (i.e. average yearly rainfall is 15 in) and standard deviation $\sigma = 1$.

Find the fraction of the years with rainfall between

a) 14 in and 16 in, b) 13 in and 17 in

c) 12 in and 18 in.

Here $p(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-15)^2/2}$ and as there is no elementary antid. for this fn, we must proceed numerically.

a) Fraction of years with rainfall between 14 in & 16 in $= \int_{14}^{16} \frac{1}{\sqrt{2\pi}} e^{-(x-15)^2/2}\, dx \approx .68$

b) ——— 13 in & 17 in $= \int_{13}^{17} \frac{1}{\sqrt{2\pi}} e^{-(x-15)^2/2}\, dx \approx .95$

c) ——— 12 in & 18 in $= \int_{12}^{18} \frac{1}{\sqrt{2\pi}} e^{-(x-15)^2/2}\, dx \approx .997$

To put this in context, the chance of having <13 in or >17 in is about 5% or 1 in 20. Since $\sigma = 1$, this is referred to in statistics as a 2σ event.

The chance of having <12 in or >18 in is even less at about .3% or 1 in 300. This is called a 3σ event.