

WHERE DOES THE BELL CURVE COME FROM? - WEBINAR

The curve that works 'best' when our brain tries to comprehend uncertainty



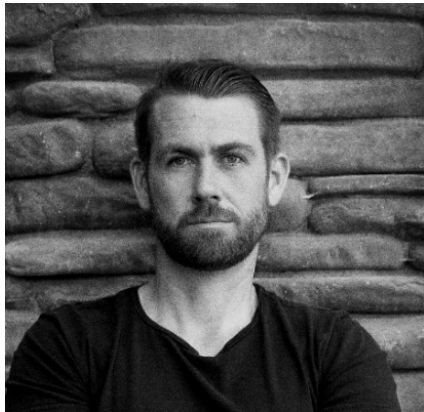
Workbook

Presented by

Dr Chris Jackson

Reliability engineer





Chris Jackson, PhD, CRE, CPEng

chris.jackson@acuitas.com

Reliability Engineer

Former Lieutenant Colonel in the Australian Army

Founder and director of Acuitas Reliability

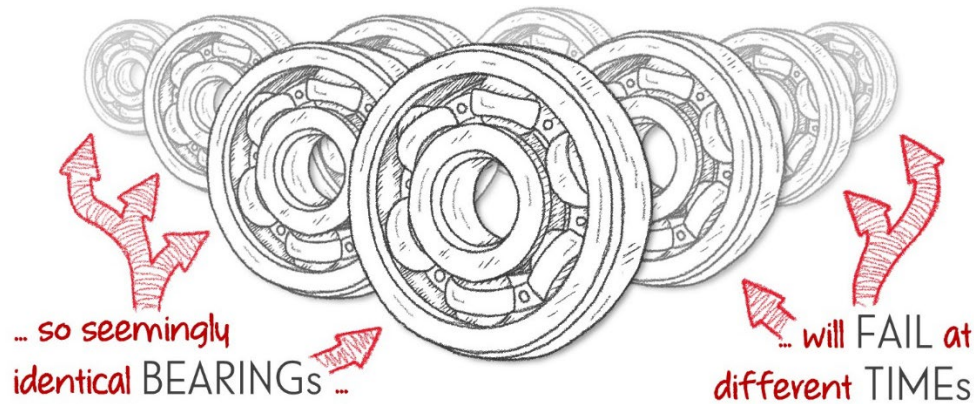
Co-founder of www.is4.com education

Author of books on how to do 'reliability stuff'

Still learning

FAILURE is a RANDOM PROCESS

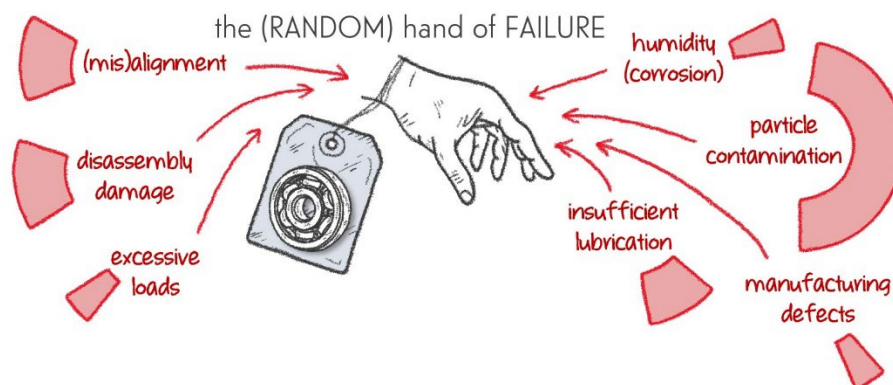
When it comes to understanding the **reliability** of a product, we need to embrace that there will always be some **uncertainty** (in our minds) about when it will **fail**.



From one perspective there is 'no RANDOMNESS' in product **failure**. If we were to hypothetically understand the location of every atom, every molecule, every force (along with its magnitude), and be able to precisely model every physical interaction ranging from user behaviour through to lubricant contamination ... we should be able to know precisely when a ball bearing will fail. The trouble is, we will never have this information ... nor will we be able to afford the supercomputer that we would then need to process this information.

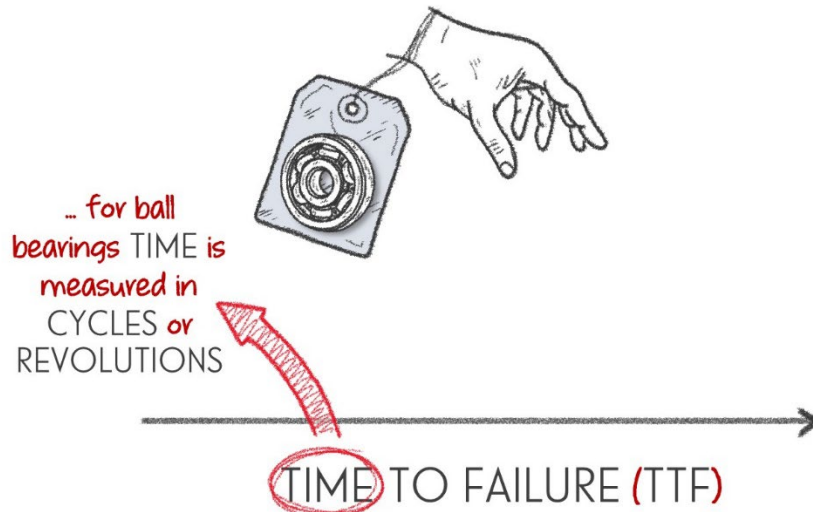
So even though we know each ball bearing is not identical, they are practically IDENTICAL to us because we can't tell which one is different from another.

But that doesn't mean we are completely clueless. We know through observation and study that there are several key factors that influence the random nature of product (ball bearing) **failure**. And these factors are represented with the **random hand of failure**. It is a simple image or illustration that represents all those factors that mean that different things will not **fail** at the same time.

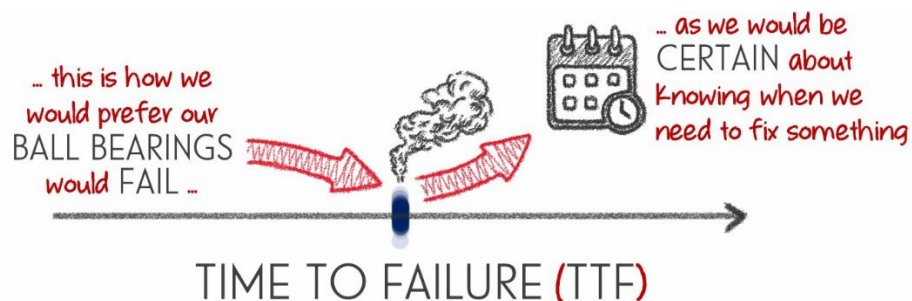


When it comes to 'time,' we need to also understand that this is measured differently for different products. There is usually never one 'perfect' measure of time. Some **failure mechanisms** are based on calendar hours. Others are based on operating hours. And yet others are based on throughput, distance, and cycles to name a few. This is another factor that drives uncertainty in terms of when our product will **fail**. And in many cases we need to pick the best one.

the (RANDOM) hand of FAILURE



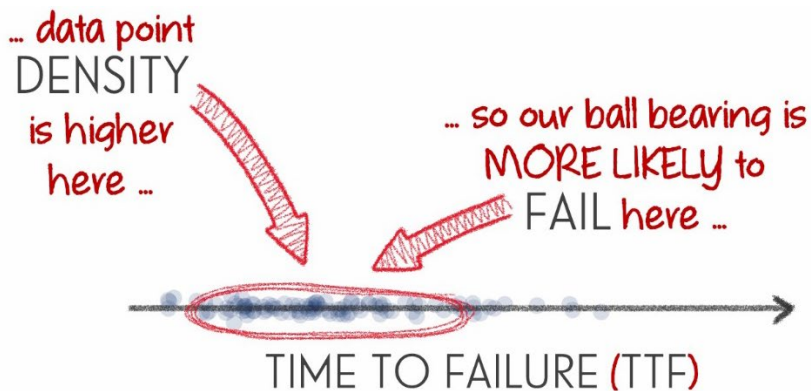
Of course, we would love this to not be the case. We would love all our (for example) engines to **fail** at precisely the same time or usage ... which means we can all put that **failure event** in our Outlook, Apple or Google calendar.



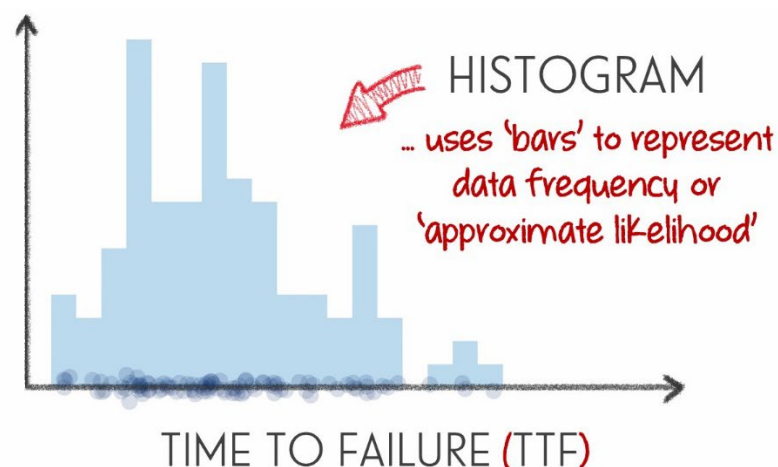
... but this is NOT POSSIBLE ...

just because something is RANDOM
... doesn't mean its UNPREDICTABLE

Let's say that we have 100 data points that represent the **Time to Failure (TTF)** of 100 ball bearings. They are each represented by a blue circle on the axis below. The first thing you can see is that even though they are **random variables**, there is at least one clear pattern or characteristic.



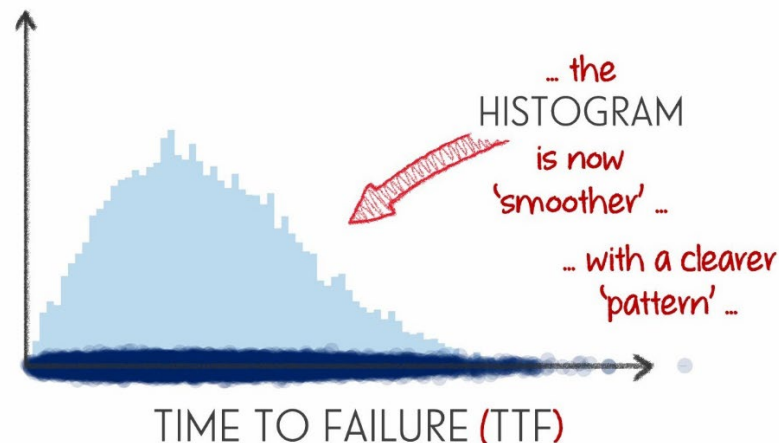
But it is not easy to see random failure characteristics from the line of data points above. There are other ways to visually represent this data. Using a **histogram** to get a little better idea of the **random nature** of ball bearing **failure**.



This **histogram** represents the '**TTF density**.' It measures the observed likelihood of **TTF** occurring at specific durations. Taller bars mean the ball bearing is more likely to **fail** within the time interval represented by width of that bars.

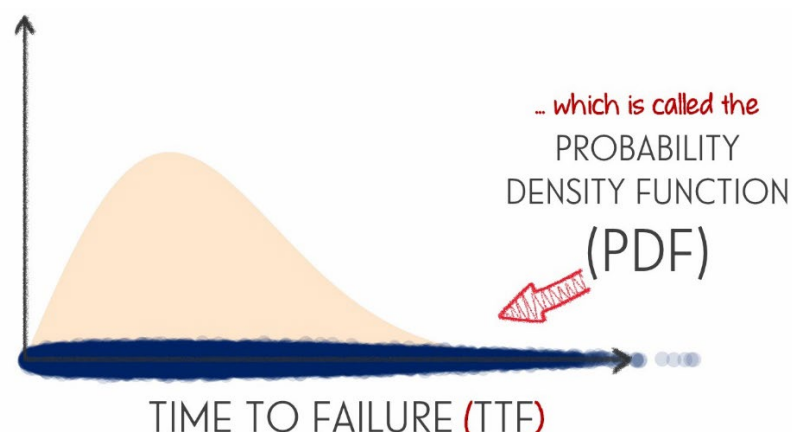
when will it fail? ... PROBABILITY DENSITY FUNCTION (PDF)

We can see that even with 100 data points (which might seem like a lot ... at least to some!) the **histogram** on the previous page is not perfectly smooth. The bars go up and down in a way that probably doesn't represent the true nature of the **random process**. One way to smooth out the **histogram** is to get a lot more **information** ... like 10 000 more bearing TTF data points.



The reason the **histogram** looks smoother is because each data point contains a little bit of **information**. So for every additional data point, the amount of **information** we have increases. And (hypothetically) we can get to a point where we have all the **information** we will ever get about our **random process**.

If we to get a million, billion, trillion data points (... a lot!) our **histogram** will start to look really, really smooth. This (hypothetically perfectly) smooth **histogram** (with lots of information) will create a curve, which represents what we call the **Probability Density Function (PDF)** of this **random process**. The **PDF** describes the random nature that our 'random hand of failure' places on each **random variable**.



the PROBABILITY DENSITY FUNCTION (PDF) gives the relative probability that a continuous random variable is equal to a specific value

Now because this is a **function**, we typically represent the **PDF** with the **notation** $f(t)$.

is it still working? ... RELIABILITY

We can use the **PDF** to advise us about all sorts of **random process** related metrics ... and importantly ... answer critical questions ...

... what should my WARRANTY PERIOD be if no more than 5 % of my bearings can FAIL?

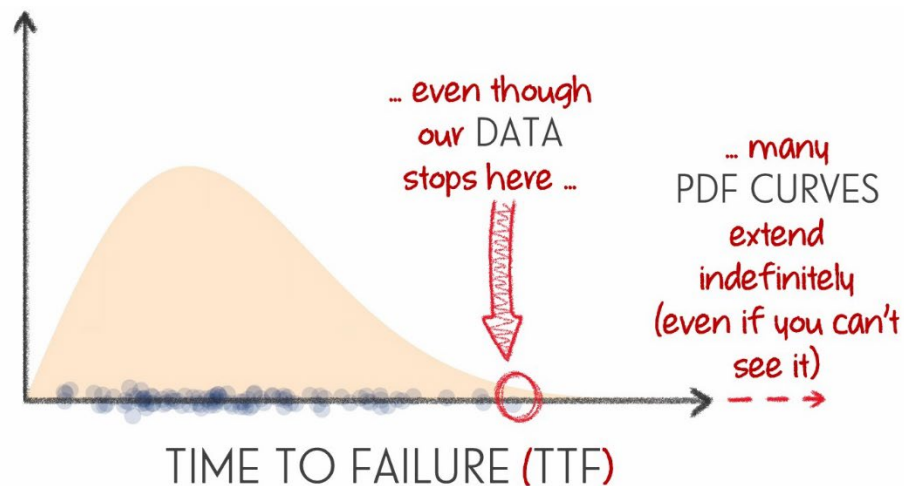
... what is my SERVICING INTERVAL if I want no more than 1 % of my bearings to FAIL?

... how many bearings will I expect to FAIL in one year if I don't SERVICE them?

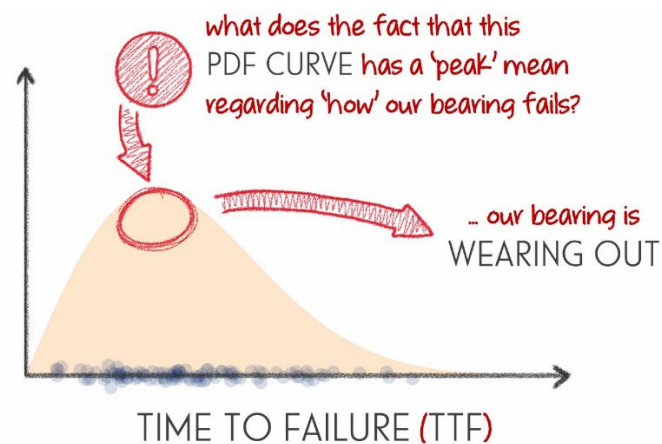
... is SERVICING going to benefit these bearings?

... and LOTS MORE questions for DIFFERENT PRODUCTS.

So let's look at **PDF curves** in greater detail to see what information we can extract from them. The first thing we need to understand about **PDF curves** (especially those that describe **TTF**) is that they can cover a lot of possible variable values.

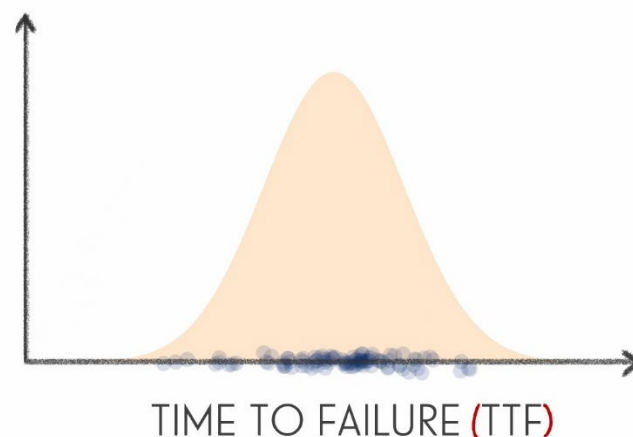


One thing we can see straight away is that our ball bearing is **wearing out**. Wear-out **failure mechanisms** means that **TTF** data 'clusters' around a central value, and the '**TTF density**' appears like a 'bell shape' or a 'hill.'



then there is the BELL CURVE

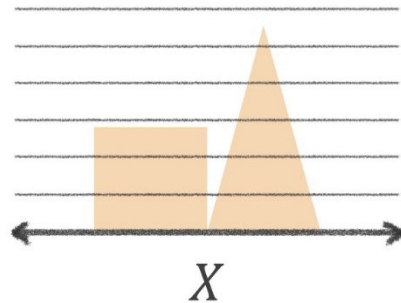
The **normal distribution PDF curve** is a 'perfectly symmetrical' bell shape. This shape isn't there because it looks 'pretty.' This shape is actually based on an underlying process that appears in lots of probability and statistics scenarios, random processes and the natural world.



So ... where does this nice shape come from?

... the CENTRAL LIMIT THEOREM

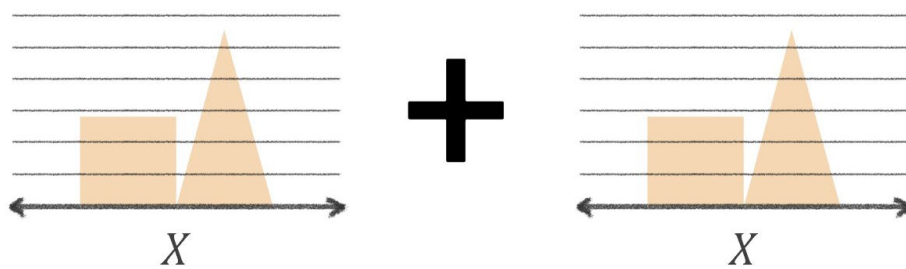
Let's make up a **random variable**, which has a clearly 'absurd' **PDF curve**. Let's make this absurd **PDF** look like it has a square on the left and a triangle on the right.



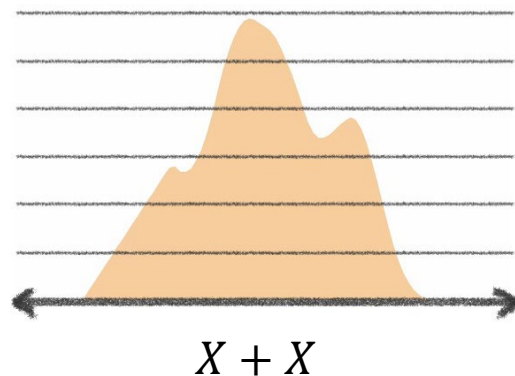
Hopefully you can agree that this **random variable** is not likely to exist! But now let's add this **random variable** to itself. That is ...

$$X + X$$

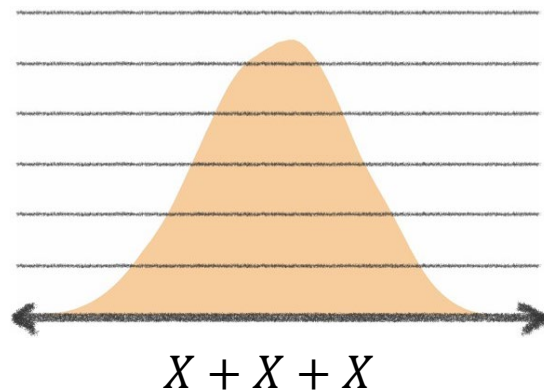
What does the **PDF** of the sum of these two **random variables** look like? So we are trying to ...



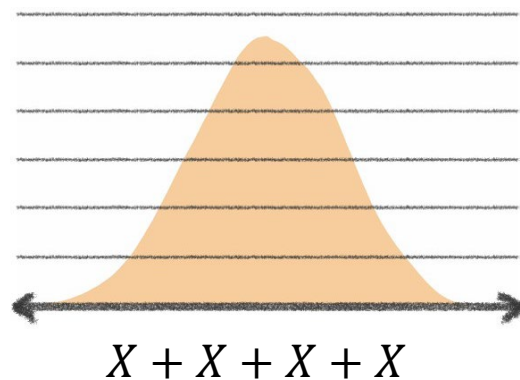
It turns out the PDF of this new **random variable** ($X + X$) looks like ...



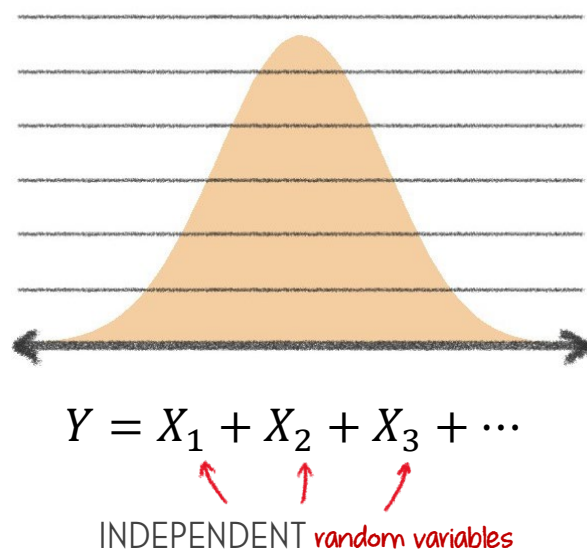
You can already see that the shape of this new **PDF curve** is substantially different. It is smoother, and perhaps more like a bell shape! But let's add another X and see what happens ...



And let's add another X ...



It should be really obvious that the **PDF** of this **random variable** that is the sum of lots of ' X 's' has a **PDF curve** that looks like a bell shape. And this is where the **normal distribution** comes from. If there a **random process** is based on the sum of lots of random processes, the output random variable will probably be described by a **normal distribution** with its bell-shaped **PDF curve**. In other words, **random variable 'Y'** is a **normally distributed random variable** if ...



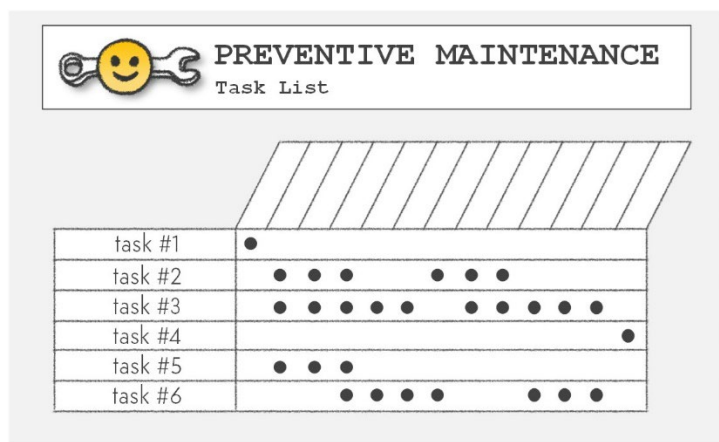
This phenomenon is known as the **Central Limit Theorem**.

the CENTRAL LIMIT THEOREM states that any sum of (independently individually distributed) random variables will be approximately NORMALLY DISTRIBUTED.

So ... the **normal distribution** IS the 'bell curve.'

... the normal distribution is great for PREVENTIVE MAINTENANCE (PM)

PM is **maintenance** we conduct at specific intervals. It is not in response to **failure** ... so we typically follow a list of **PM tasks**.



For example, a **PM action** may consist of ...

1. Inspection
2. Preparation
3. Change fluids
4. Clean filters
5. Replace seals
6. Functional Test

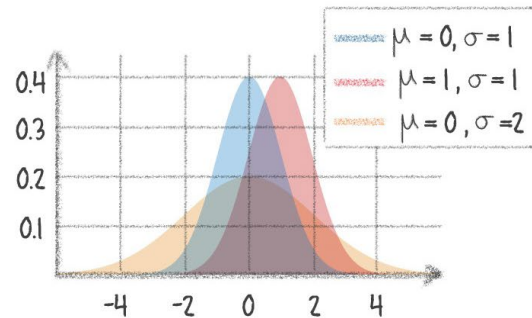
Total **PM time** is the sum of all the individual task times. Which themselves are **random variables**. So we know that **PM time** is always described by a **normal distribution**. This is an example of the **normal distribution** being relevant for applications beyond **reliability** and **TTF**. But remember ... when we use the **normal distribution** to describe **TTF**, we are *always* describing a product, system or service that wears out.

PROBABILITY DENSITY FUNCTION (PDF)

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



= norm.dist
([x], [μ], [σ], FALSE)



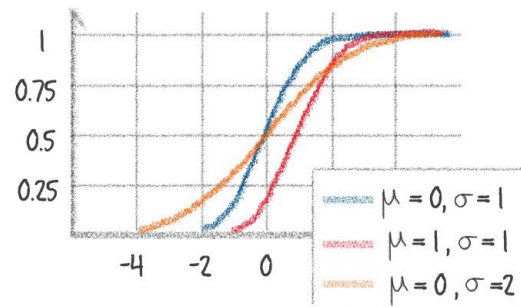
The **PDF curve** of all **normal distributions** are symmetrical bell shapes centered about the **mean** (μ , **MTTF** or **MTBF**). The two parameters you need to define a **normal distribution** are the **mean** (μ , **MTTF** or **MTBF**) and the **standard deviation** (σ).

CUMULATIVE DISTRIBUTION FUNCTION (CDF)

$$F(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^t e^{-\frac{(t'-\mu)^2}{2\sigma^2}} dt'$$



= norm.dist
([x], [μ], [σ], TRUE)

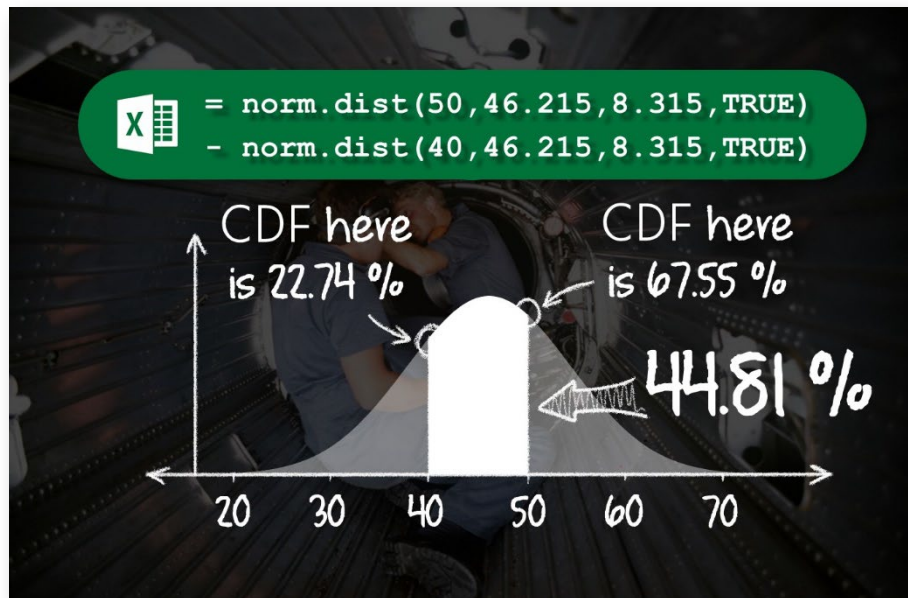


You can see here that the **normal distribution CDF** (like all other **CDFs**) starts at 0 and approaches 1 (or 100 %) as **TTF** increases.

EXERCISE

Let's go back to Preventive Maintenance (PM).

If Preventive Maintenance Time (PMT) is normally distributed with a MEAN (μ) of 46.215 minutes and STANDARD DEVIATION (σ) of 8.315 minutes, what is the probability that PMT will be between 40 and 50 minutes?

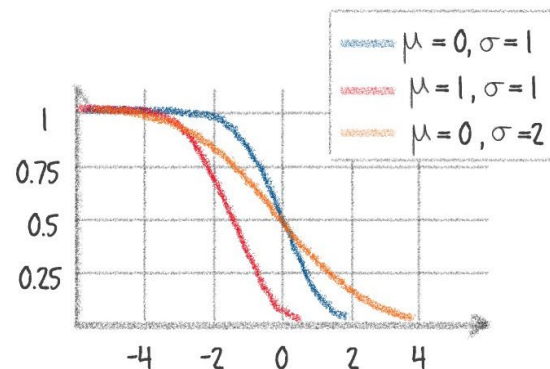


RELIABILITY FUNCTION

$$R(t) = 1 - \frac{1}{\sigma\sqrt{2\pi}} \int_0^t e^{-\frac{(t'-\mu)^2}{2\sigma^2}} dt'$$

Excel formula for the reliability function:


```
= 1 - norm.dist([x], [μ], [σ], TRUE)
```

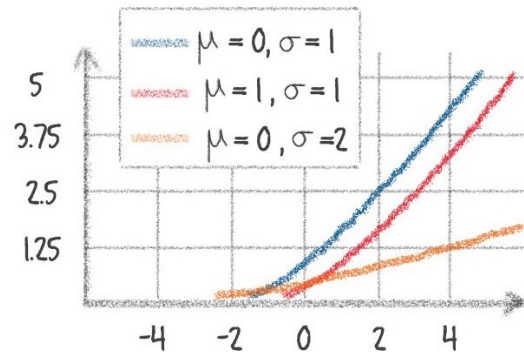


You can see here that the **normal distribution reliability function** (like all other **reliability functions**) starts at 1 (or 100 %) and approach 0 as **TTF** increases.

HAZARD RATE

$$\lambda(t) = \frac{e^{-\frac{(t-\mu)^2}{2\sigma^2}}}{\int_t^{\infty} e^{-\frac{(t'-\mu)^2}{2\sigma^2}} dt'}$$

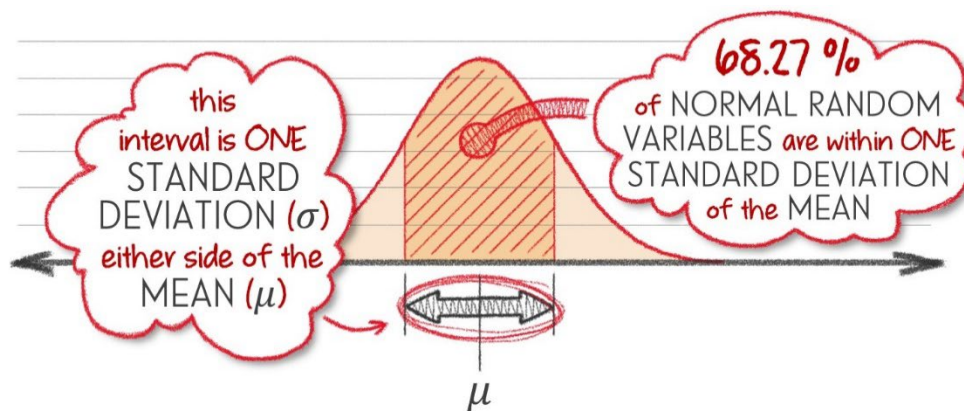
 = norm.dist([x],[μ],[σ], FALSE)/(1 - norm.dist([x],[μ],[σ], TRUE))



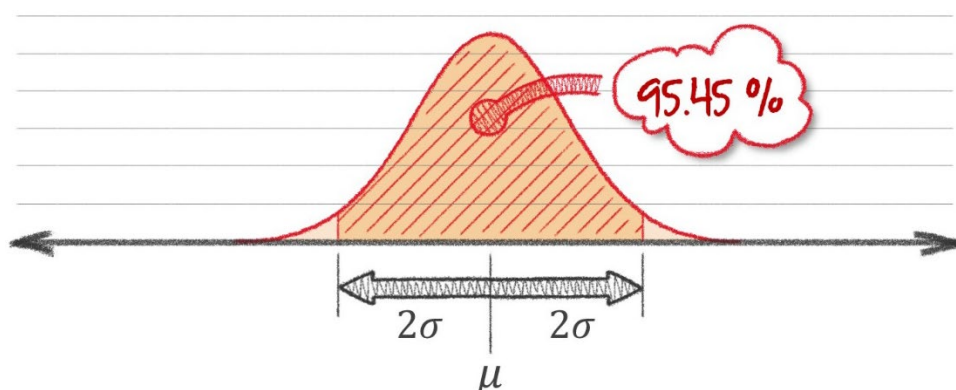
Every **normal distribution** has an increasing **hazard rate**. Which is why we know that 'bell shaped' **PDF curves** imply that the product, system or service is wearing out.

... other characteristics

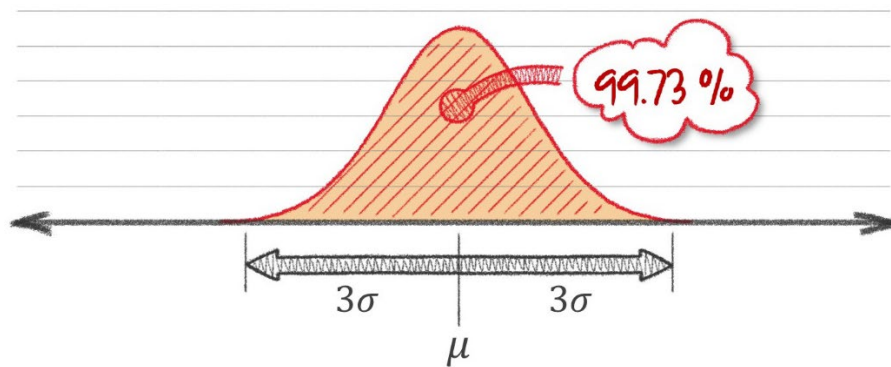
The **PDF curve** of every **normal distribution** is similar to all others. Which means they all have same shape. Which means that regardless of the **mean** (μ , **MTTF** or **MTBF**) or **standard deviation** (σ), they all have the following characteristics.



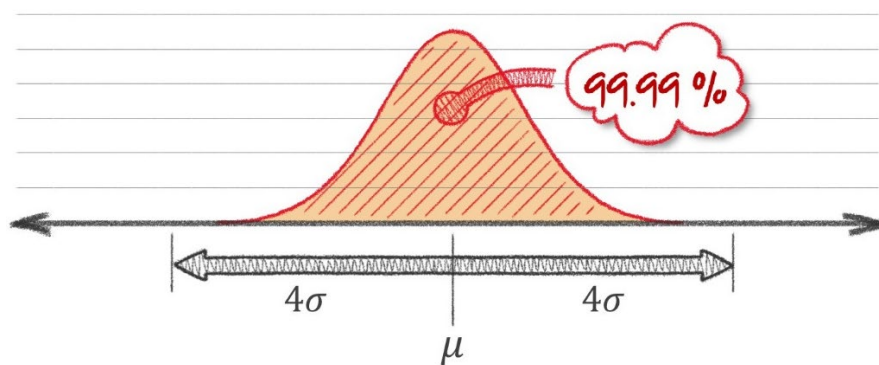
We can also say the following for all **normally distributed random variables** within two standard deviations (σ) of the **mean** (μ , **MTTF** or **MTBF**).



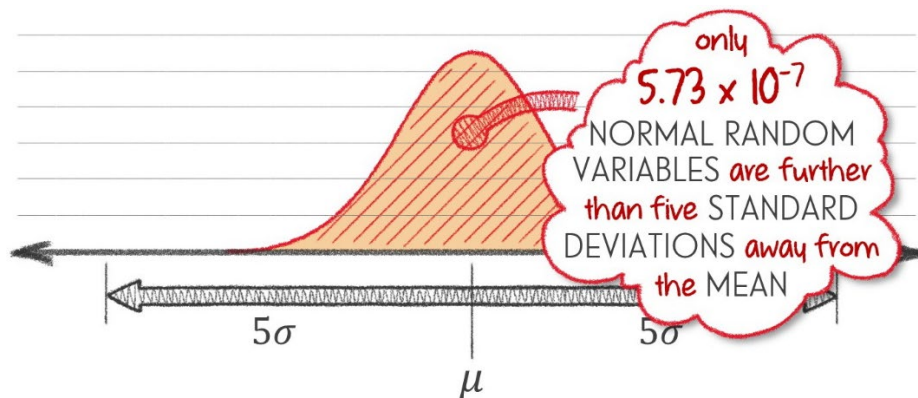
And for three **standard deviations** (σ) away from the **mean** (μ , **MTTF** or **MTBF**) ...



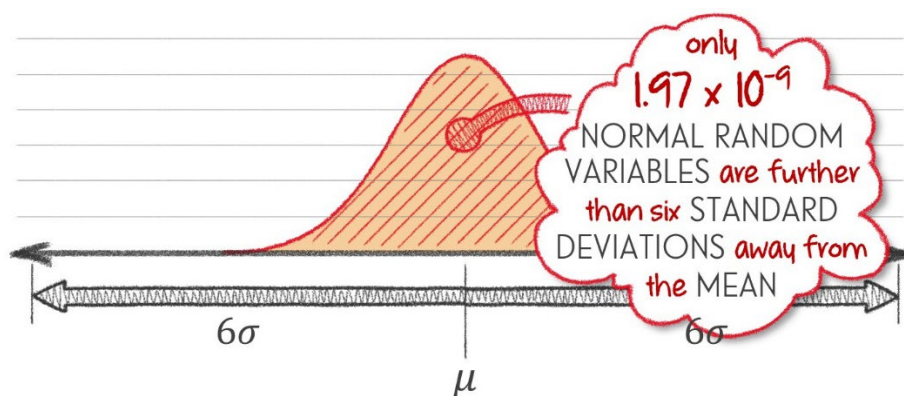
And for four **standard deviations** (σ) away from the **mean** (μ , **MTTF** or **MTBF**) ...



And for five **standard deviations** (σ) away from the **mean** (μ , **MTTF** or **MTBF**) ...

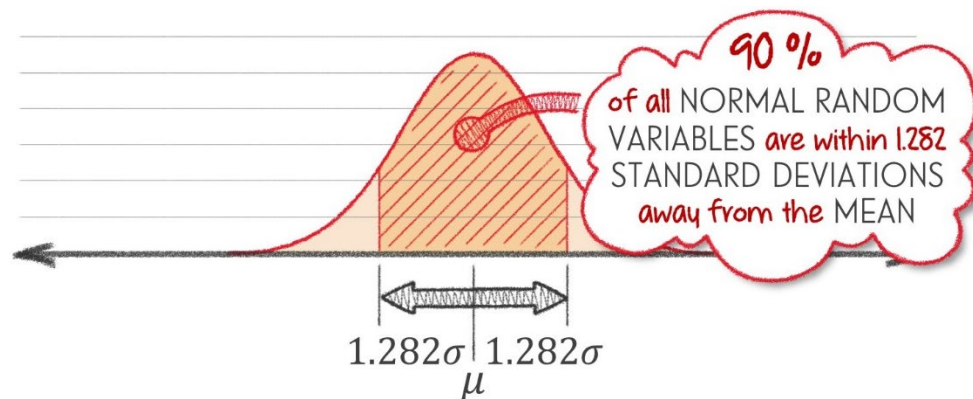


And for six **standard deviations** (σ) away from the **mean** (μ , **MTTF** or **MTBF**) ...

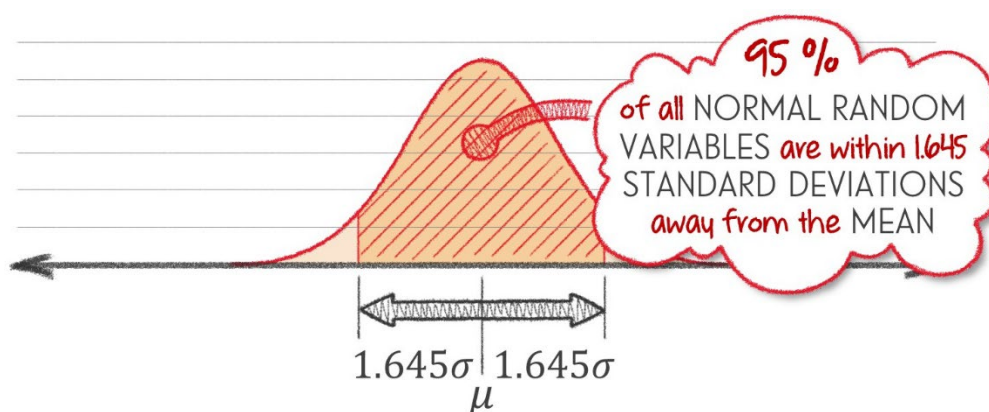


Another perspective that **reliability engineers** often need to adopt focuses on the fraction of **normally distributed random variables** that are within a certain 'deviation' from the **mean** (μ , **MTTF** or **MTBF**). This is particularly important for manufacturing (when we want to maximize the extent to which component dimensions fall within tolerances) and confidence bounds (where the **normal distribution** is often used to model our uncertainty in something).

If we are interested in the 'middle' 90 % of **normally distributed random variables** ...

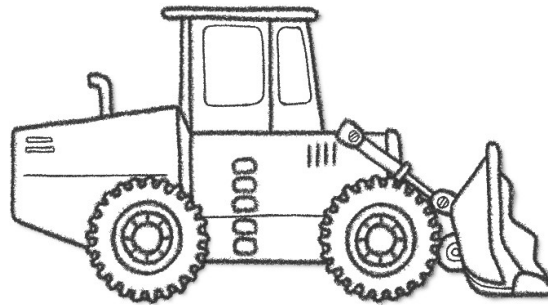


If we are interested in the 'middle' 95 % of **normally distributed random variables** ...



EXERCISE

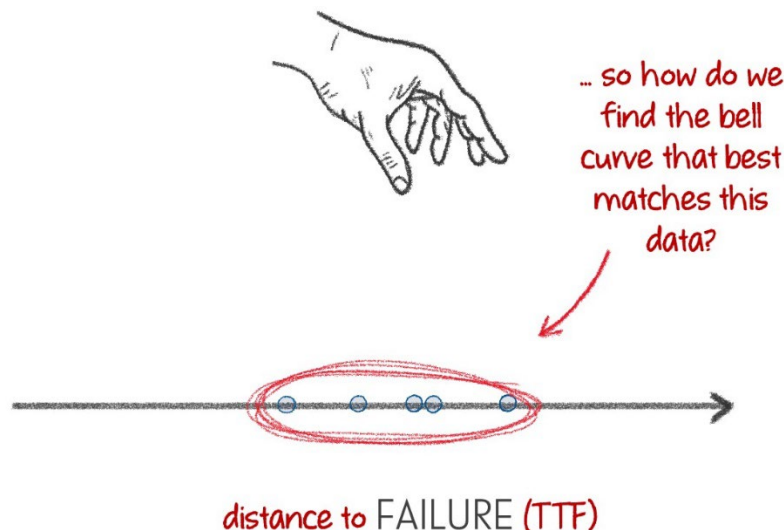
Let's look at the tires on a wheeled dozer we looked at



Let's now assume that our tires are no longer 'perfect' and will wear out.



We have historical data from five other tires that wore out. And because they wore out (meaning failure was driven by a wear out failure mechanism), and because we know how tires lose tread over time, we can reasonably assume that the tire's TTF can be described by a normal distribution.

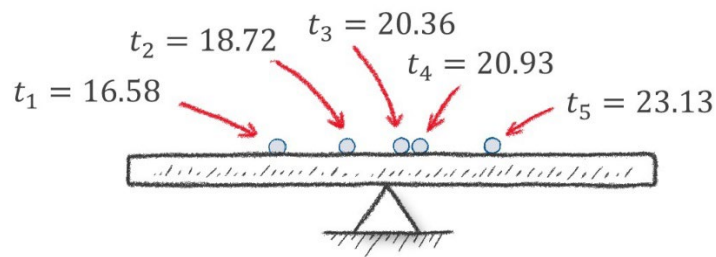


We know that the normal distribution needs the mean (μ , MTTF or MTBF) and standard deviation (σ) to be completely defined.

So how do we get these parameters from our five data points?

EXERCISE

These five data points are what we call a 'sample.' And there is a way that we can calculate the mean of this sample. It all comes down to finding the point at which those data points will be perfectly balanced.



The equation that gives us this perfect balance point (or sample mean) is ...

$$\bar{t} = \frac{\sum_{i=1}^n t_i}{n}$$

... where t_i is the value of each data point, and there is a total of n data points.

So what is the 'sample mean' of our 5 data points?

$$\bar{t} = \frac{\sum_{i=1}^5 t_i}{5}$$

$$= \frac{16.58 + 18.72 + 20.36 + 20.93 + 23.13}{5}$$

$$= 19.94$$

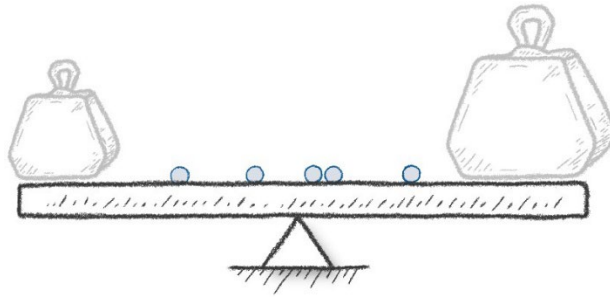
And you can use Microsoft Excel to help ...



= average
(16.58, 18.72, 20.36, 20.93, 23.13)

EXERCISE

So that gives us the 'sample mean.'
Now we need the 'sample standard deviation.'



The equation that gives us the sample standard deviation is ...

$$s = \sqrt{\frac{\sum_{i=1}^n (t_i - \bar{t})^2}{n - 1}}$$

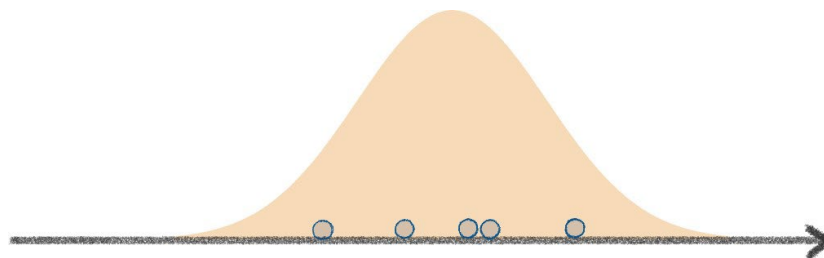
This is a little more challenging to evaluate manually!
So let's use Microsoft Excel.



= stdev.s
(16.58, 18.72, 20.36, 20.93, 23.13)

The reason we use the symbols ' \bar{t} ' and ' s ' to represent **SAMPLE** means and standard deviations instead of ' μ ' and ' σ ' is because they are not the same. **SAMPLE** parameters (\bar{t} and s) will be approximately equal to what we call the **POPULATION** parameters (μ and σ). This approximation will become more and more accurate as we get more and more data.

So when we approximate ' μ ' and ' σ ' with ' \bar{t} ' and ' s ' we get the following PDF curve. You can see that it does a 'good' job of aligning with the sample data points.



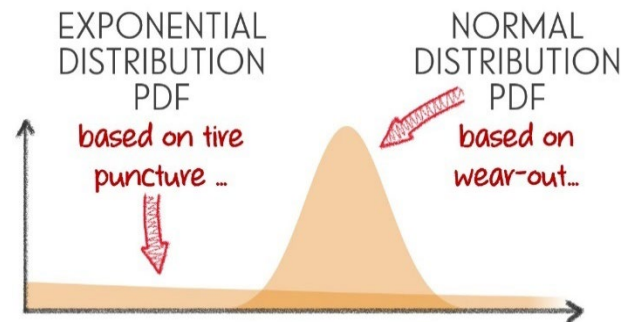
This data analysis technique is referred to as **MATCHING MOMENTS**. 'Moments' refers to parameters and characteristics of things like probability distributions. We essentially use **SAMPLE** characteristics to approximate the normal distribution parameters that define the PDF curve illustrated above.

EXERCISE

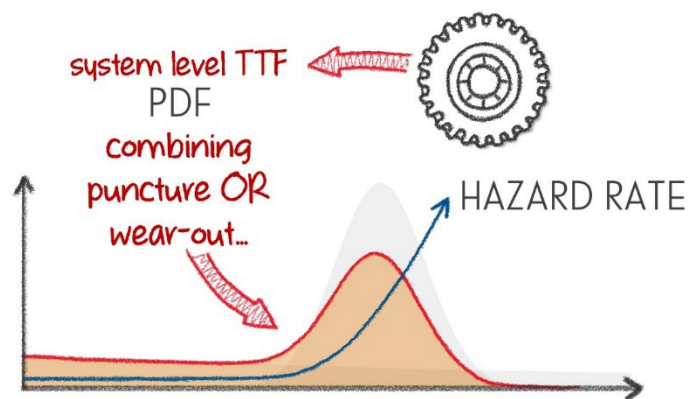
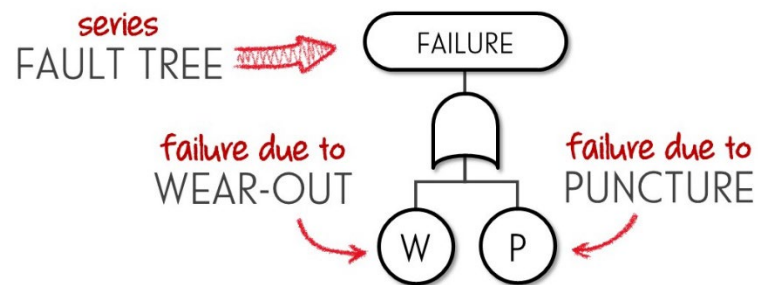
But we aren't done yet!

Let's say that we identified an exponential distribution that modelled failure based on 'sharp rock' puncture.

In this example, we identified a normal distribution that models failure based on wear out.



We know that our tire will fail due to either of these two failure modes. So how do we combine them? With a system reliability model!



... and the FINAL WORD

The **normal distribution** models the **TTF** of products, systems and services that wear out. But ... there are different ways things can wear out. The type of wear out that **normal distributions** model is those where the rate of degradation is constant. For example, we would expect a tire to lose the same amount of tread for every km or mile it is driven. This is not the case for fatigue, where the rate at which cracks grow increases over time.

Normal distributions are also good at modelling **PM** time and any other **random variable** that is the sum of lots of other **random variables**.

But there are other **probability distributions** we need to be aware of as **reliability engineers**.