Empirical Distributions

Distributions based on empirical observations, of a finite sample.

Analytic Distribution

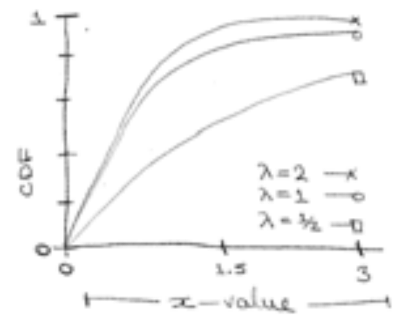
Generated from a mathematical function — the function forms the CDF.

Can be used to mimick or model empirical data in a generalized way.

Exponential Distribution

$$CDF(x) = 1 - e^{-\lambda x}$$

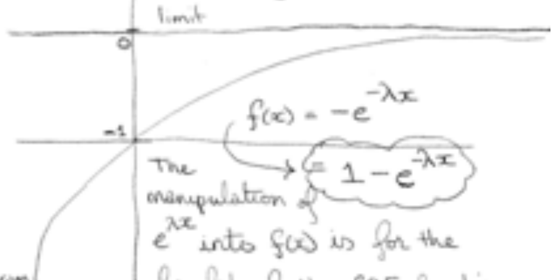
* λ – determines the shape of a distribution. A parameter.



When events are equally likely to occur at any time, and there is observation of a sequence of these events — the measurement of time occurring between these events is called interarrival times.

The distribution of interarrival times is similar to an exponential distribution.

There is brilliance in this function; the e-base is arbitrary the exponential is congruent w- compounding probabilities.



$$f(x) = -e^{-\lambda x}$$

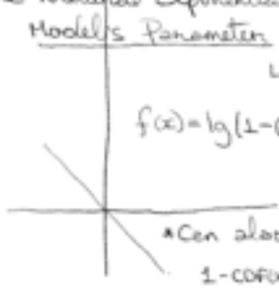
$$= 1 - e^{-\lambda x}$$

The manipulation of $e^{\lambda x}$ into $f(x)$ is for the benefit of the CDF function which must begin at zero and complete at one. Exponentials are equal to (1) when their powers are zero; furthermore positive based exponents are always positive —

The Linear Descent of the Natural Exponential Model's Parameter

Linear Param

$$f(x) = \lg(1 - (1 - e^{-\lambda x}))$$



* Can also just map $1 - CDF(x) \rightarrow 1 - (1 - e^{\lambda x})$ on a log-y scale.

IC.

$$y \approx e^{-\lambda x}$$
$$\Leftrightarrow \lg(y) \approx -\lambda x.$$
} Therefor linear w/ $\frac{d}{dx}(\log(y)) = -\lambda$

CCDF.

Complementary CDF is $1 - CDF(x)$.

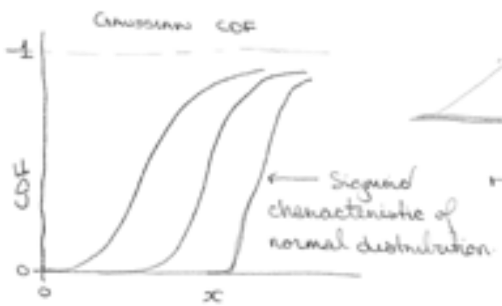
The mean value is λ^{-1}. Since the distribution begins at zero and approaches (1) when $\lambda = 1$. The parameter just scales the distribution in a proportionally inverse manner.

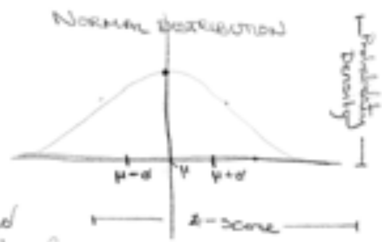
approaching but never falling below zero. By manipulating the exponential the $\langle 0,1 \rangle$ (i.e. $\langle x,y \rangle$) end can correspond to the beginning of the CDF starting at absolute zero. The approach to zero can represent the approach to $CDF(y) = 1$.

λ is the only param; it is the rate of events which occur in a unit of time.

NORMAL DISTRIBUTION
(GAUSSIAN)

- describes many phenomena
- standard normal distrib
  - ↳ $\mu = 0$, $\sigma = 1$ } two params
  - → avg + std dev.





TESTING THE FIT OF ANALYTICAL MODEL

1. Histograms
2. CDFs
3. Stem and Leaf Plots
4. Box Plots (Box + Whisker Plots)
5. P-P Plots
6. Q-Q Plots (Normal Probability Plot)

There is an easy and hard way to create Normal Probability Plots

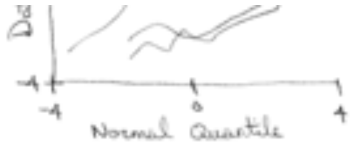
Easy:

1. Sort values in the sample
2. From a standard normal distribution ($\mu = 0$ and $\sigma = 1$), generate a random sample with the same size as the sample, and sort it.
3. Plot the sorted values from the sample versus the random values.

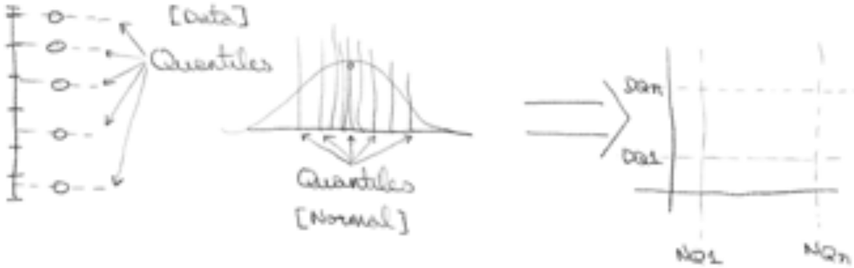
Normal Probability Plot
Q-Q Plot

StatQuest Method

1. split data into quantiles
2. Choose model
3. split data into quantiles
4. Split model into same number of quantiles
   ↳ where there is equal probability of observing values within a group



* μ – intercept
* σ – slope



If the Q-to-Q mapping is congruent mean the quantiles between distributions are similarly distributed then the data can be said to follow a distribution to a higher degree.

Where NQ_x is the normal quantile of index x.
Where DQ_x is the data quantile of index x.

* Can also compare data distrib to other data distrib using quantile-to-quantile.

— THE LOGNORMAL DISTRIBUTION —

When the logarithm of a dataset is intrinsically of a normal distribution.

① Sub x for $\log(x)$. ② Take $CDF(\log(x))$.

* Can just plot CDF on a log-x scale

◦ $CDF_{lognormal}(x) = CDF(\log(x))$

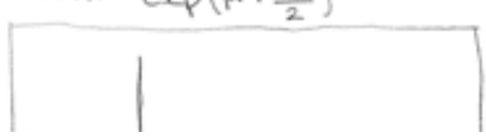
Parameters:

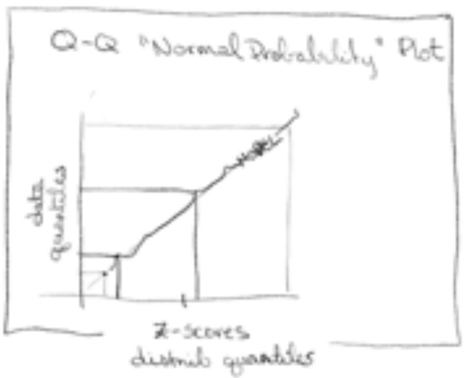
μ – not the mean

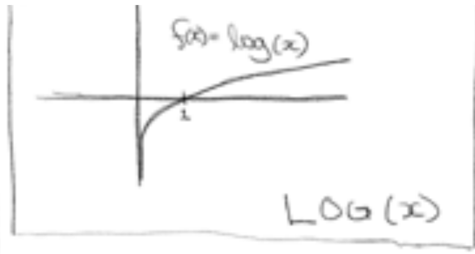
σ – not the std. dev.

Summary Statistics:

mean $= \exp\left(\mu + \frac{\sigma^2}{2}\right)$







The Pareto Distribution

$$CDF(x) = 1 - \left(\frac{x}{x_m}\right)^{-\alpha}$$

The parameters;

x_m — the minimum possible value

α — degree of compound

The Cumulative Distribution of a Pareto Model can be re-expressed as a Complementary CDF; allowing for rework into quantiles.

$$\begin{aligned} CCDF &= y \\ &= 1 - \left[1 - \left(\frac{x}{x_m}\right)^{-\alpha}\right] \\ &= \left(\frac{x}{x_m}\right)^{-\alpha} \\ &= 1/x^{\alpha} \cdot x_m^{\alpha} \\ &= (x_m/x)^{\alpha} \end{aligned}$$

$\longrightarrow$

The logarithm.

$$\begin{aligned} \log y &= \log\left[(x_m/x)^{\alpha}\right] \\ &= \alpha \log x_m/x \text{, by log power rule} \\ &= \alpha(\log x_m - \log x) \text{, log quotient rule.} \end{aligned}$$

$\Rightarrow$ Since the Pareto distribution develops exponentially, the appropriate logarithm of the y will be a reflection of the linear contribution of x. By taking the same logarithm of the x-axis the logarithm of the y-axis should be congruent, by a factor of $-\alpha$.

$$\text{Key: } \log y = \alpha(\log x_m - \log x)$$

<u>Generating Random Numbers with a Distribution</u>

A CDF computes the probability of a distribution containing values of a lesser or equal value.

The inverse would when provided a probability determine a value which is greater than or equal to the probability's corresponding fraction of the distribution.

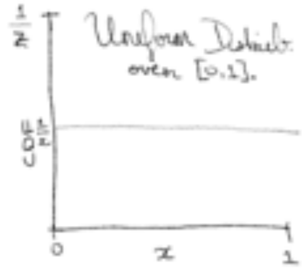
Through equitable selection over a uniform distribution [0, 1] one can apply the Inverse CDF (ICDF) and ...

[illegible] generate samples from the distribution; with values which would model the very same distribution.

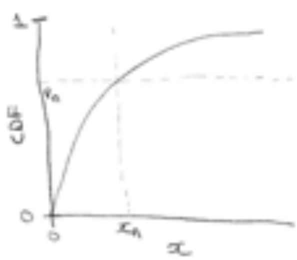
Example,

$$p = 1 - e^{-\lambda x} \quad \} \text{ exponential distribution}$$

$$\Leftrightarrow x = -\log(1-p)/\lambda$$



bounded by $x \in [0,1]$.



where p_n is a probability,
x_n is a value.

WHY MODEL

- compression