PMF visualize relationships and patterns well. when the number of values are few.

When the dispersion and the amount of data grows the plot becomes more generalized (smaller probabilities) and the proportion of noise to data grows.

Furthermore the crossection of one dataset on to others often becomes convoluted.

It becomes hard to distinguish between data sources and determine qualities of any distribution even by itself.

⟹ This can be mitigated by grouping data into non-overlapping buckets - a technique called "binning."

- however this can generalize real patterns & relationships, but when done appropriately remove noise.

- It's also just painstaking.

Cumulative Distribution Functions is a method which solves this problem without the friction of Binning.

Percentiles & Percentile Rank.

Percentile rank – the fraction of people who scored lower than or equal to a given score.

$$pr = \frac{size(S_L)}{size(S)},$$ where S is the set of all data points.

where S_L is the set of all data points $\leq$ to the variable 'L'.

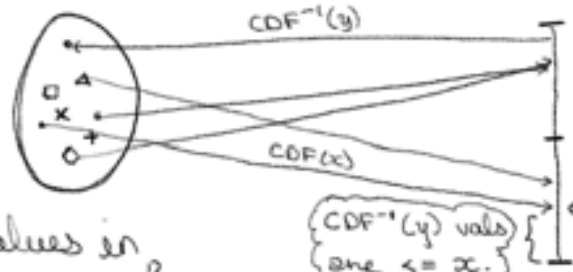
Percentile – the maximum percentile rank's corresponding score which does not exceed the definition defined by percentile rank.

* To get a percentile must sort data an access index based on percentile rank's called index.

Cumulative Distribution Function

Distrib. Vals

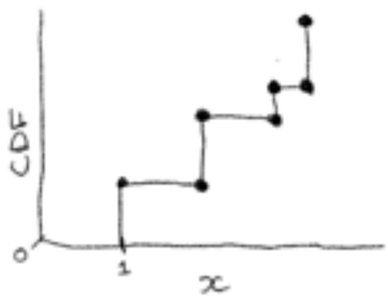
Percentile Rk.



« Values in the data »

The fraction (or probability) of values « in the distribution » less than or equal to some value.

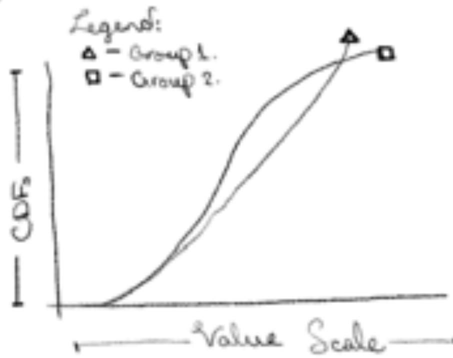
Graph of CDF



Graphed using a step function.

If x_z is greater than the longest x, $CDF(x_z) = 1$.

Comparing CDFs (Better than PMFs)



Interpreting a CDF, can use as a lookup; where the percentile rank indicates the values and fraction of the distribution which falls below those values maximum.

Vertical sections indicate density and common values in the distribution.

⇒ If there are many common values in a distribution this means significant fractions of the distrib. will be less than or equal to that value. Due to this significant ranges of the CDF (the vertical part) will be occupied by the preceding value.

CDFs are superior visualizations for comparing Distrib.

They present the change of the distrib. in a continuous and undisruptive way.

In addition they have the faculty to observe singular elements in relation to the distribution.

Percentile-based Summary Statistics

With CDFs in hand summary stats can be computed through use of the CDF to form percentile based descriptions.

Due to this distrib can be graphed alongside, then evaluated at an aggregate level. and then dissected at a granular level if need be.

50^{th} percentile – val. which divs. distrib. in half. called – median –.

inner quartile range (IQR) – the measure of a distrib. spread & dispersion. It's the diff. between the 25^{th} and 75^{th} percentile.

quantiles – often segmentation of a distrib into continuous ranges of percentiles, can be used to describe a distrib.

→ quintillion is percentile ranges of 20: 20^{th}, 40^{th}, 60^{th}, ..

RANDOM NUMBERS

* w replacement – without removing the value from the population.

Meaning:

10% of sample ≤ 10^{th} percentile

20% .... ≤ 20^{th} percentile

The random selection of samples

:

100% of samp ≤ 100th / 99th percentile.

Percentile Rk.

will uniformly span the original CDF, its percentiles.

Generating Random Numbers w- a Cumulative Distrib. Function.

1. Choose a percentile rank uniformly from [0, 100].

2. $CDF^{-1}(y)$ the percentile rank to find the percentile score.

Comparing Percentile Ranks.

Can compare across groups. via ranking.