Glossary,

- Probability Density Function - the density of a value's probability.
- Probability Density - seeks to express the relationship between an amount of probability per number of values.

  Similar to physics $D = m/v$.
  It is the derivative of the Cumulative Distribution function (CDF).

- Kernel Density Estimation (KDE) - an algorithm for generating a PDF given some sample data.
- Raw Moment - a summary stat. of data based on that data's deviation from zero, raised to a power
- Discretize - to reduce a continuous function to discrete segmentations; the opposite of smoothing - it is a discrete approximation of a continuous function.

- Central Moment - statistic based on deviation from mean; transformed by some power.

- Standardized Moment - a moment that accounts for the proportional contribution of other moments. Has no

- Skewness - the degree

to which a distribution   ᵈ units.
is asymmetric.

- <u>Sample Skewness</u> – computed from moment based statistics.

- <u>Pearson's Median Skewness Coefficient</u> – degree of a distribution's asymmetry based on mean, median and the standard deviation. Insulated from the effects of outliers due to it's substitution for moments.

- <u>Robust</u> – a resilience to outlier data.

PROBABILITY DENSITY FUNCTION (PDF):

Exponential Dist.

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

Normal Dist.

$$PDF_{normal}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

The derivative of a CDF is a Probability Density Function. To get probability mass, you have to integrate over x.

Kernel Density Estimation (KDE)

an algorithm that takes a sample and finds a smooth PDF. Does so non-parametrically; that is without parameters.

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

$$= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{n}\right),$$

K - is the kernel, a non-negative function

K_h - is the scaled kernel. defined $K_h(x) = \frac{1}{h}\cdot K\left(\frac{x}{h}\right)$

h - is a smoothing parameter called the bandwidth.

Kernel functions,

I. uniform
II. triangular
III. Epanechikov (*)
IV. Normal

" Create a smooth curve given a set of data "

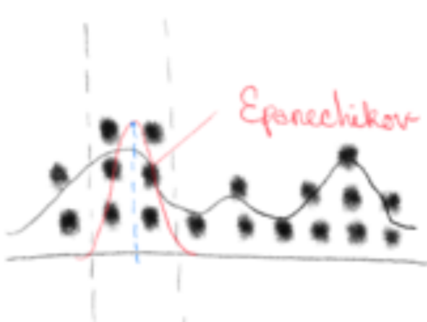
A contiguous replacement for the discrete histogram.

$K(x) = \phi(x)$, where ϕ is the standard normal distrib.

* Changing the bandwidth changes the shape of the kernel

lower bandwidth - only points close in position making the estimate squigly.

higher bandwidth - shallow kernel means distant points can contribute



• Bandwidth
• Amplitude

$$\hat{f}(x) = \sum_{\text{OBSERVATIONS}} K\left(\frac{x - \text{observ.}}{\text{bandwidth}}\right)$$

Estimating a density function with KDE is useful.

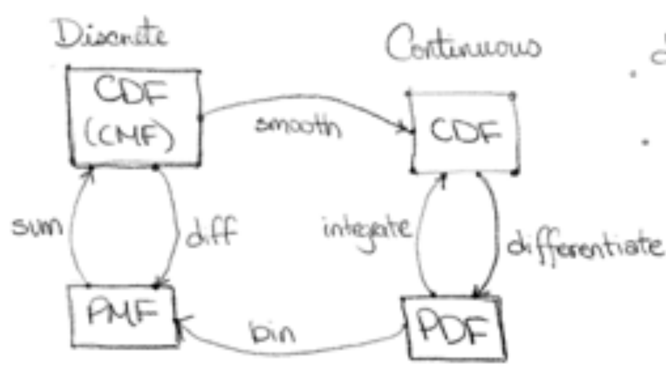
- <u>vizualization</u> – for exploration, CDFs are the usually the best vizualization. After a CDF, can decide if an estimated PDF is better to model a distribution. May be better choice for presenting distrib to an audience.
- <u>interpolation</u> – a way to use sample data/sparse data to model a distribution. Assuming the data's distribution is smooth.
- <u>simulation</u> – based on the distribution of a sample. The ability to generate outcomes; as opposed to replicating existing data.

A Distribution Framework

PMF – probabilities for a set of values.

CDF – cumulative probabilities

PDF – derivative of a CDF



- discretizing/quantizing
- cumulative mass function – not really used.

Moments

reducing a sample to a number is a statistic.

raw moment → $m'_k = \frac{1}{n} \sum (x_i)^k$

central moment → $m_k = \frac{1}{n} \sum (x_i - \bar{x})^k$

$\boxed{\text{Skewness}}$ 3rd-order central moment

* can be standardized.

Pearson's Median Skewness Coefficient

$$g_p = \frac{3(\bar{x} - m)}{S}$$

$\bar{x}$ is sample mean
m is sample median
S is sample deviation

- Sign is important – magnitude harder to decipher
- outliers make sample skewness unreliable
  ↳ making the sample moment less applicable to determining skewness of sample data seemingly it's obvious use.