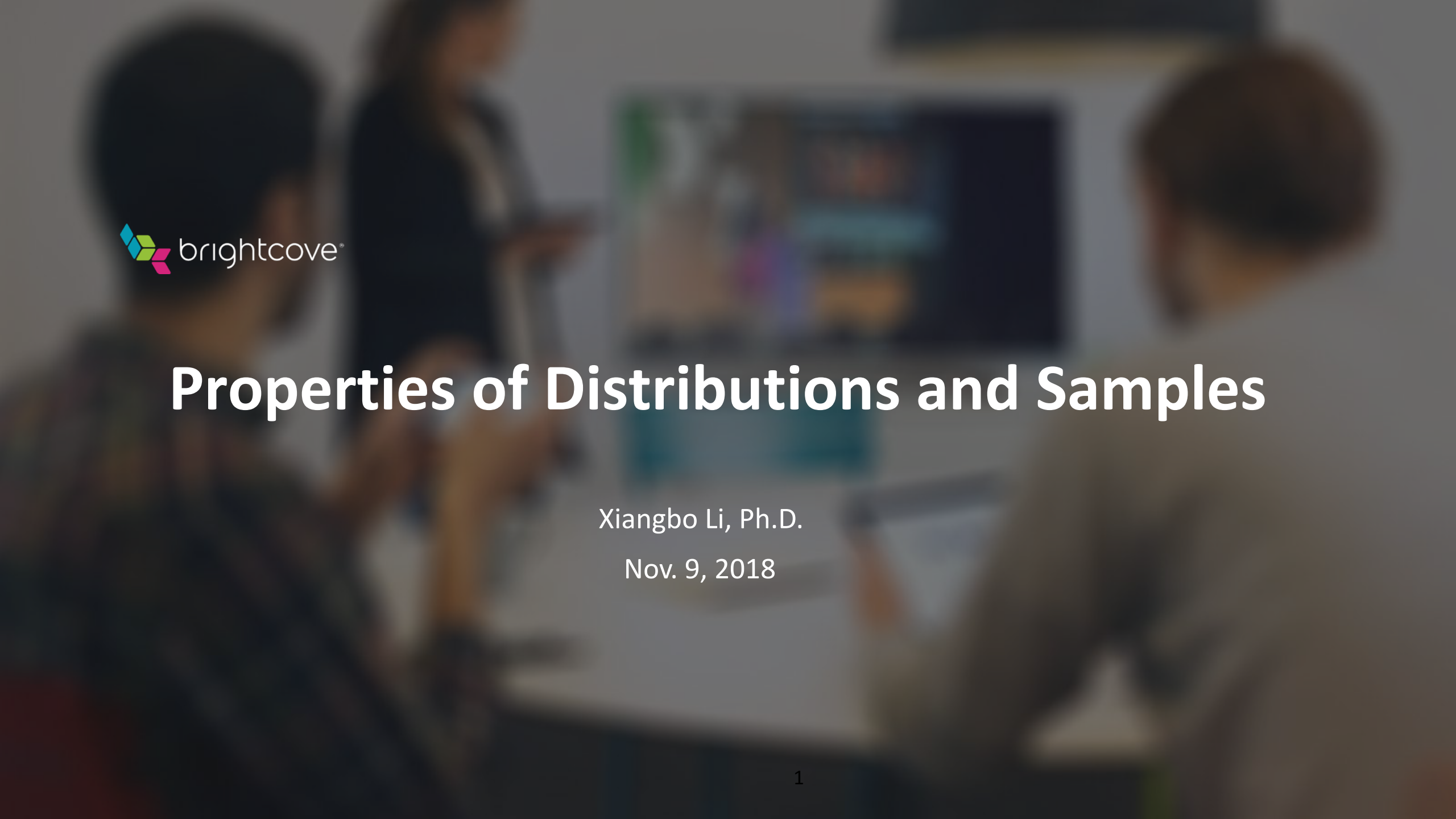


Properties of Distributions and Samples

Xiangbo Li, Ph.D.

Nov. 9, 2018

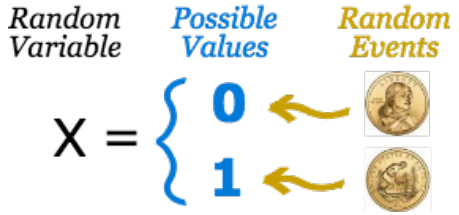
Outline

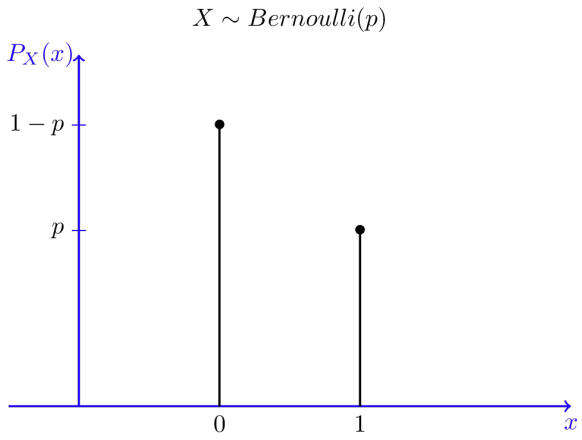
- Random variables
- Distribution (discrete and conditional cases)
  - examples: Bernoulli, Laplace, Gaussian, etc.
- Analytical properties of distributions
  - moments, quantiles, etc.
- Samples
- Sample -based statistics
  - sample mean, sample variance, etc
- Connection between analytical and sample-based statistics
- Sample based Sufficient statistics
- What it all means practically?

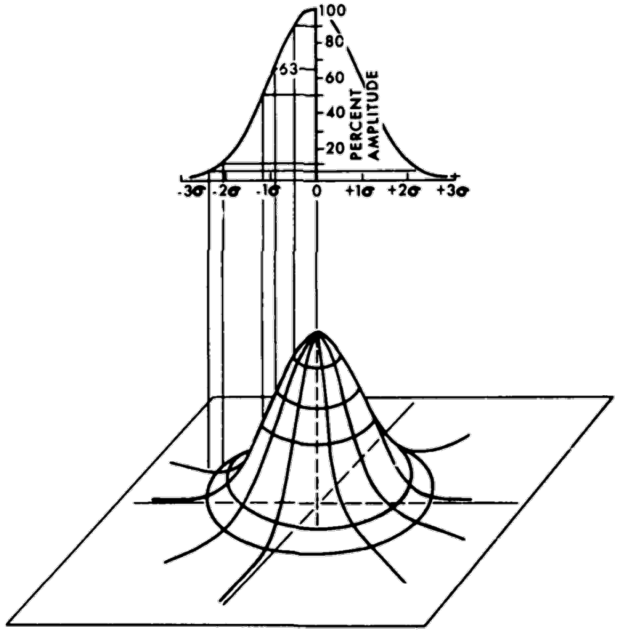
Random Variables

A random variable is a variable whose possible values are numerical outcomes of a random process. There are two types of random variables:

- Discrete random variables are distinct or separate values, the space of random variable X is countable.
- Continuous random variables are any values in an interval.







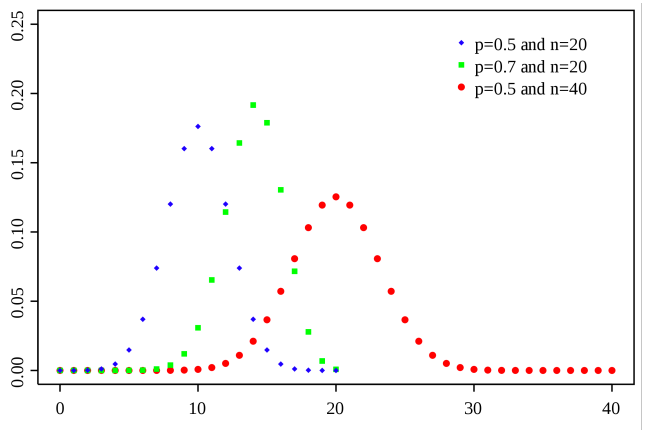
Properties of Distributions

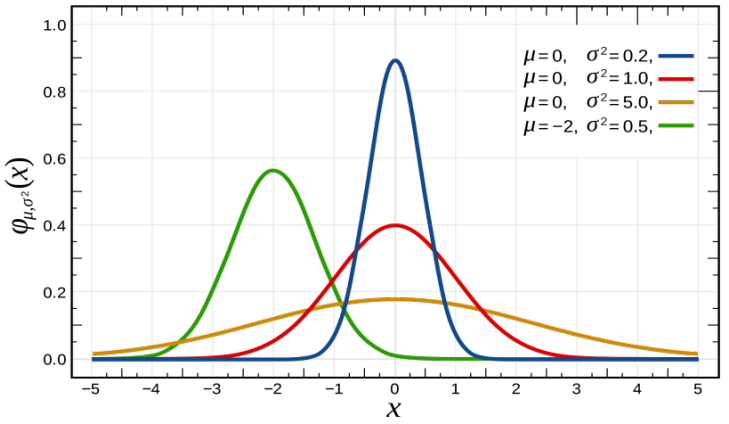
Distribution Functions

- Probability Mass / Density Function

$$f(x) = \Pr(X = x)$$

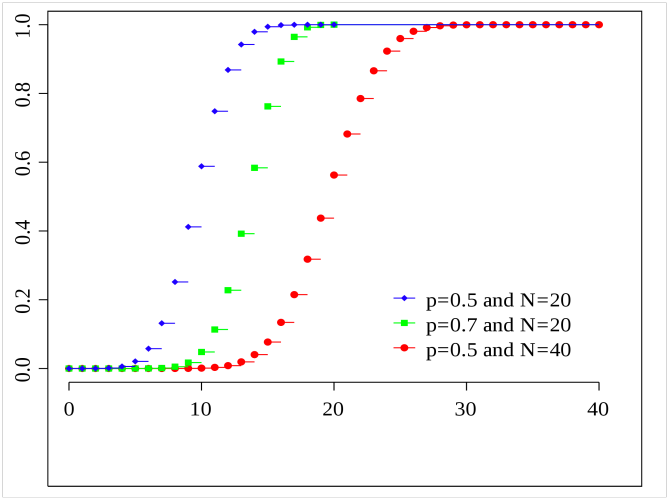
  - $f(x) \geq 0$
  - $\sum_{x \in R_x} f(x) = 1$ or $\int_{-\infty}^{\infty} f(x) = 1$

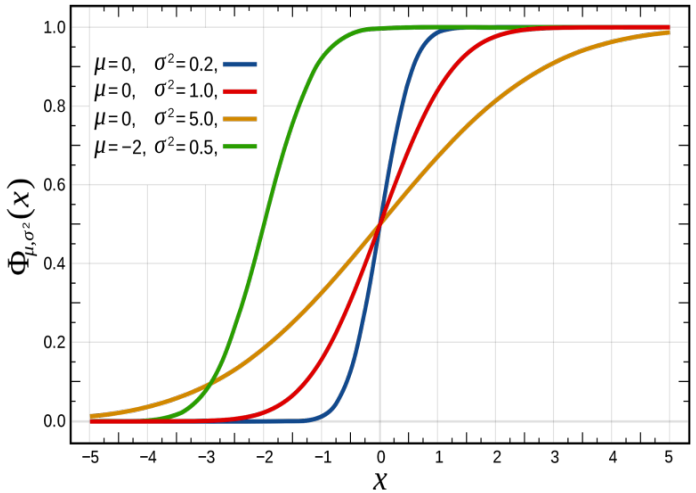




- Cumulative Probability Function

$$F(x) = \sum_{t<x} f(t) \quad \text{or} \quad F(x) = \int_{-\infty}^{x} f(x)dx$$





Expected Value

The expected value of a random variable, intuitively, is the long-run **average** value of repetitions of the same experiment it represents

- Discrete random variables:

$$E[X] = \sum_{i=1}^{n} x_i p(x_i)$$

- Continuous random variables:

$$E[X] = \int_a^b x f(x) dx$$

1. If X and Y are random variables on a sample space Ω then

$$E(X + Y) = E(X) + E(Y).$$

2. If a and b are constants then

$$E(aX + b) = aE(X) + b.$$

Moments

A moment is a specific quantitative measure of the shape of a function. If the function is a probability distribution, then :

- Raw moment:

$$\mu'_n = E[X^n] = \sum_{x \in R_x} x^n f(x) dx \qquad discrete$$

$$\mu'_n = E[X^n] = \int_{-\infty}^{\infty} x^n f(x) dx \qquad continuous$$

- Central moment:

$$\mu = E[X] = \sum_{x \in R_x} x f(x) dx \qquad discrete$$

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx \qquad continuous$$

- Standardized moment: $\tilde{\mu}_n = \frac{E[(X-\mu)^n]}{\sigma^2}$

- The 1st moment is the mean

$$E[X] = \mu'_n$$

- The 2nd central moment is the variance

$$Var[X] = \mu_2 = E\left[X - E[X]\right]^2$$

- The 3rd standardized moment is the skewness

$$Skew[X] = \tilde{\mu}_3 = \frac{E[(X - E[X])]^3}{\sigma^3}$$

- The 4th standardized moment is the kurtosis

$$Kurt[X] = \tilde{\mu}_4 = \frac{E[(X - E[X])]^4}{\sigma^4}$$

Moment Generation Function (MGF)

Moment-generating function of a real-valued random variable is an alternative specification of its probability distribution. Thus, it provides the basis of an alternative route to analytical results compared with working directly with probability density functions or cumulative distribution functions.

$$M_X(t) = E[e^{tX}] = \sum_{x \in R_x} e^{tX} x f(x) dx \qquad discrete$$

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tX} x f(x) dx \qquad continuous$$

$$M_X(t) = E[e^{tX}] = 1 + tE[X] + \frac{t^2 E[X^2]}{2!} + \frac{t^3 E[X^3]}{3!} + \cdots + \frac{t^n E[X^n]}{n!} + \cdots$$
$$= 1 + t\mu_1 + \frac{t^2 \mu_2}{2!} + \frac{t^3 \mu_3}{3!} + \cdots + \frac{t^n \mu_n}{n!} + \cdots$$

where μ_n is the n^{th}moment. Differentiating $M_X(t)$ i times with respect to t and setting $t = 0$, we obtain the i^{th} moment about the origin u_i

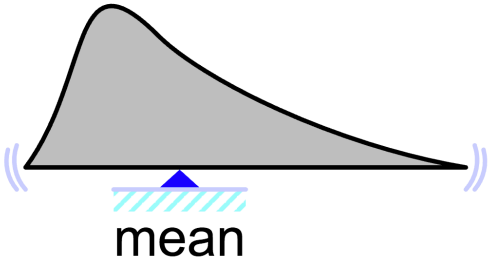
Measures of Central Tendency

➢ Mean / Expected Value

The mean of a probability distribution is the long-run arithmetic average value of a random variable having that distribution

$$\mu = E[X] = \sum_{x \in R_x} xf(x)dx \qquad discrete$$

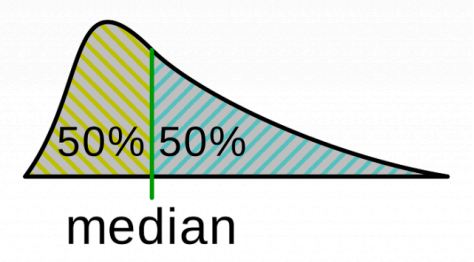
$$\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx \qquad continuous$$



Good for symmetric distribution

➢ Median

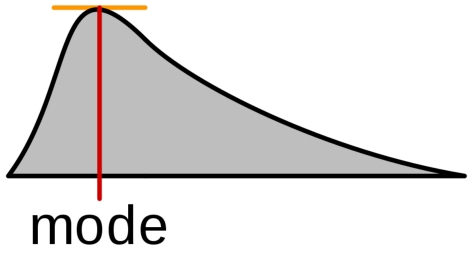
The value separating the higher half from the lower half of a probability distribution.



Good for skewed and outliers data

➢ Mode

The value appears most often in a data set. The value x at which a probability distribution takes its locally maximum value.



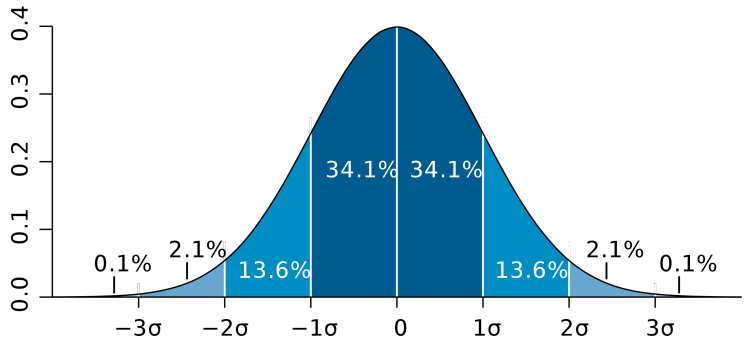
Good for categorical data

Measures of Spread / Disparity / Scale

➢ Standard Deviation

population standard deviation is found by taking the square root of the average of the squared deviations of the values subtracted from their average value

$$\sigma = \sqrt{E\left[X - E[X]\right]^2}$$



- Pair with mean, suitable for symmetric distribution
- standard deviation squares its differences, it tends to give more weight to larger differences and less weight to smaller differences compared to the mean absolute difference

➢ Mean/Average Absolute Deviation

average absolute deviation (or mean absolute deviation) of a data set is the average of the absolute deviations from a central point

$$m(X) = E[|X - E[X]|] = \int_{-\infty}^{\infty} |x - E[X]| f(x) dx$$

➢ Quantile

Quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way

$$\Pr[X < x] \leq \frac{k}{q}$$

0 quartile = 0 quantile = 0 percentile

1 quartile = 0.25 quantile = 25 percentile

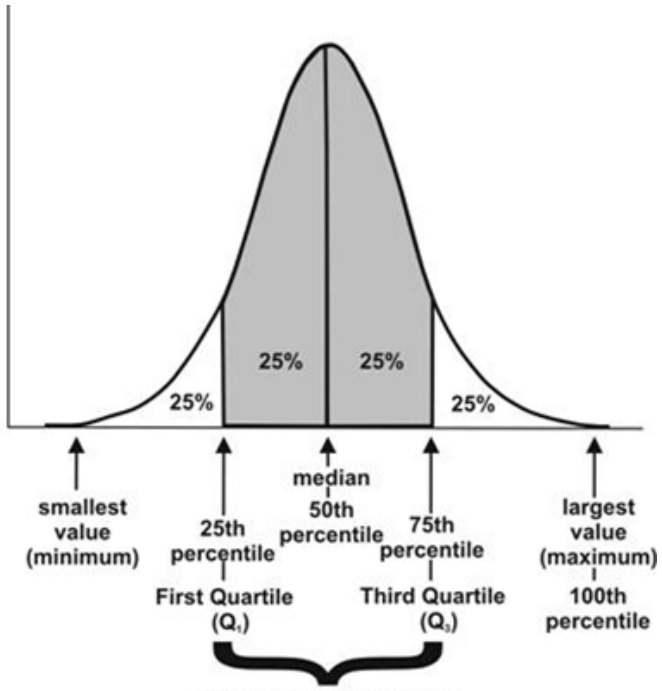
2 quartile = .5 quantile = 50 percentile (median)

3 quartile = .75 quantile = 75 percentile

4 quartile = 1 quantile = 100 percentile

➢ Quartile

A quartile is a type of quantile. The first quartile (Q1) is defined as the middle number between the smallest number and the median of the data set. The second quartile (Q2) is the median of the data. The third quartile (Q3) is the middle value between the median and the highest value of the data set.



➢ Percentile

A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls

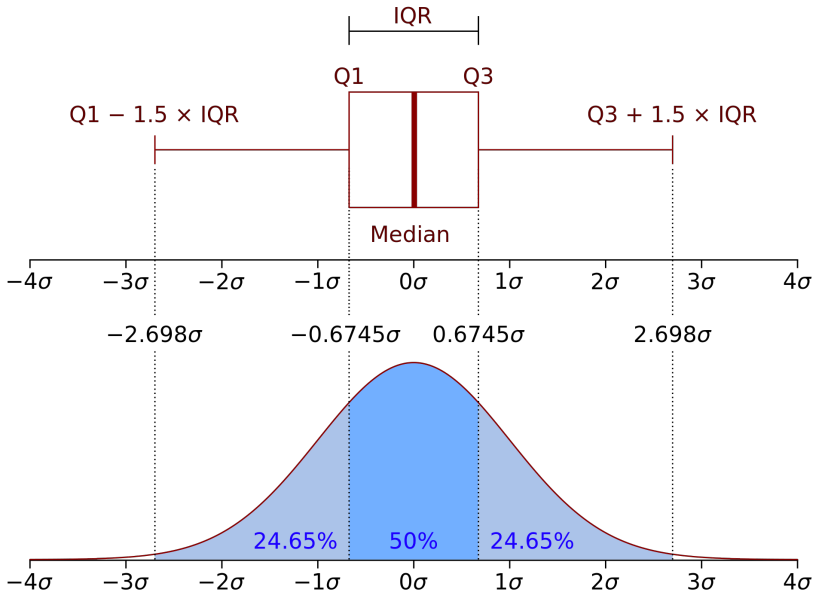
➢ Interquartile Range (IQR)

a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, $IQR = Q_3 - Q_1$

$$Q_1 = Q(0.25) = CDF^{-1}(0.25)$$

$$Q_3 = Q(0.75) = CDF^{-1}(0.75)$$

$$IQR = Q_3 - Q_1$$



➢ Median Absolute Deviation (MAD)

the median of the absolute values of the differences between the data values and the overall median of the data set

$$MAD(X) = E[|X - M_d|] = \int_{-\infty}^{\infty} |x - M_d| f(x) dx$$

- Pair with median, suitable for skewed distribution

Scale Factor: $k = \frac{\sigma}{MAD}\ or \frac{\sigma}{IQR}$

Measures of Shape

Skewness

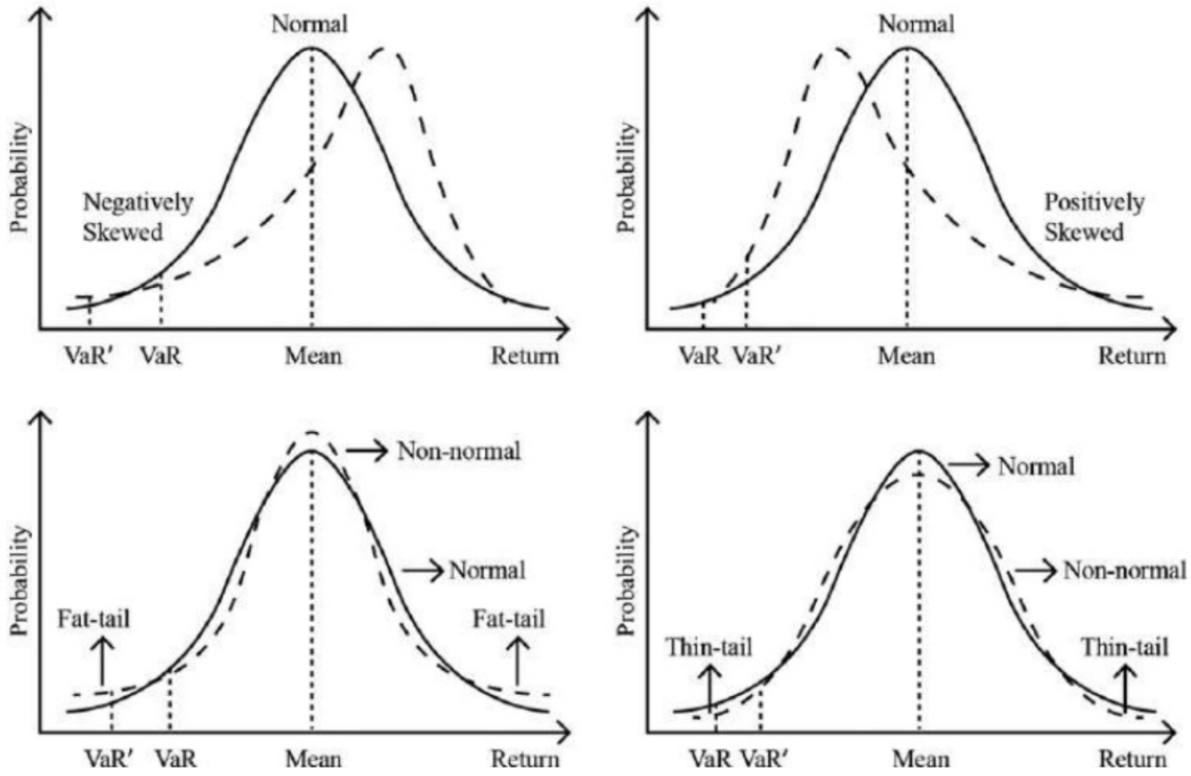
The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left. Positive values for the skewness indicate data that are skewed right.

$$Skew[X] = \tilde{\mu}_3 = \frac{E[(X - E[X])]^3}{\sigma^3}$$

Kurtosis

Standard normal distribution has a kurtosis of zero. Positive kurtosis indicates a "heavy-tailed" distribution and negative kurtosis indicates a "light tailed" distribution.

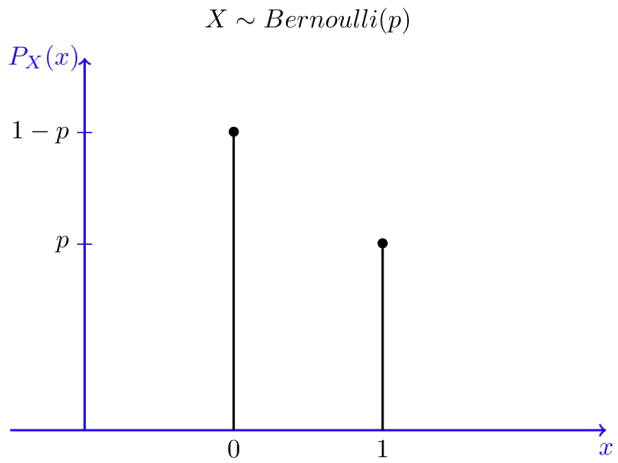
$$Kurt[X] = \tilde{\mu}_4 = \frac{E[(X - E[X])]^4}{\sigma^4}$$



Discrete Distributions

Bernoulli Distribution

$$f(x) = \begin{cases} p, & head \\ 1-p, & tail \end{cases}$$



Binomial Distribution

Flip the coin n times, what's the probability of k heads.


$n = 40, k = 14$


$n = 40, k = 23$

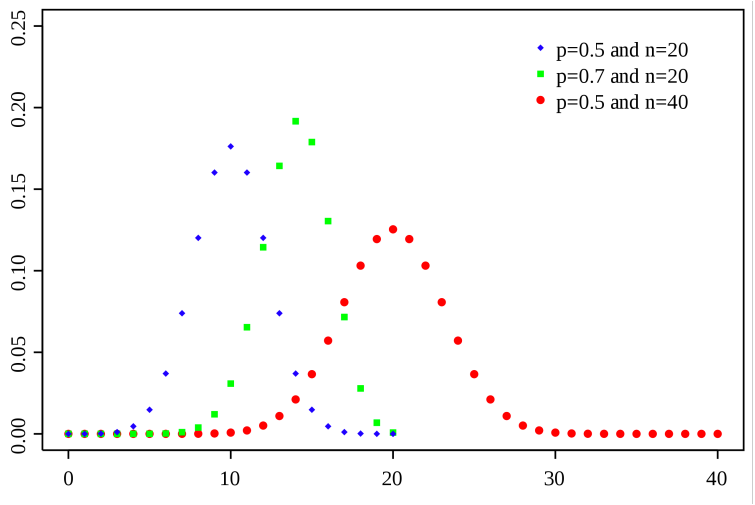

$n = 40, k = 20$


$n = 40, k = 18$


$n = 40, k = 8$

.....

$$f(x) = \binom{n}{k} p^k (1-p)^{n-k}$$



Geometric Distribution

Flip the coin k times, 1^{st} time get head.


$k = 1$


$k = 3$

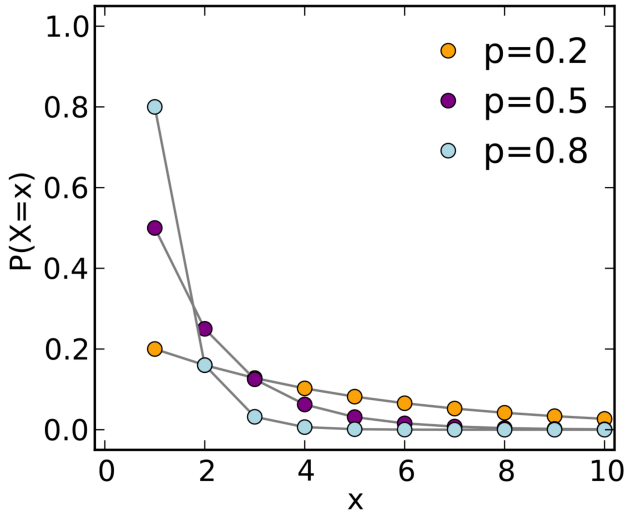

$k = 2$


$k = 8$


$k = 1$

.....

$$f(x) = (1-p)^{k-1}p$$

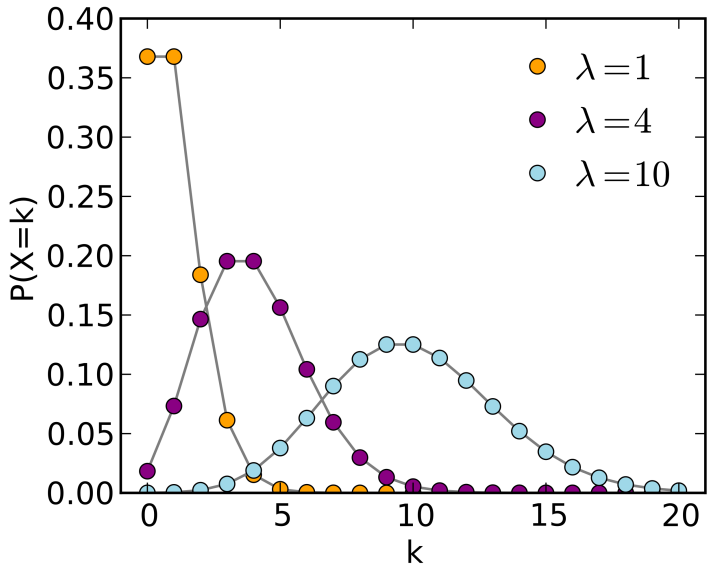


Poisson Distribution

The Poisson distribution is popular for modeling the number of times an event occurs in an interval of time or space:

- The number of meteorites greater than 1 meter diameter that strike Earth in a year
- The number of patients arriving in an emergency room between 10 and 11 pm
- The number of photons hitting a detector in a particular time interval

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



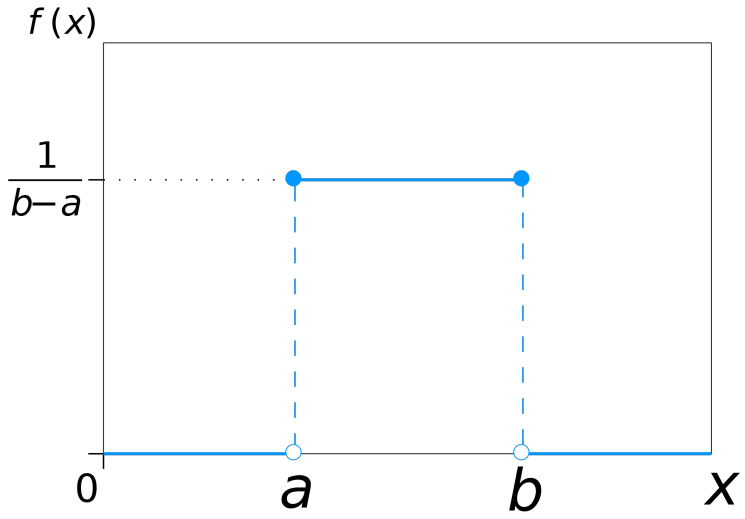
PMF CDF

Continuous Distributions

Uniform Distribution

Consider an experiment which consists of choosing a point from the interval [a, b] such that “all points are equally likely to be chosen”

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a, b] \\ 0, & otherwise \end{cases}$$



PDF

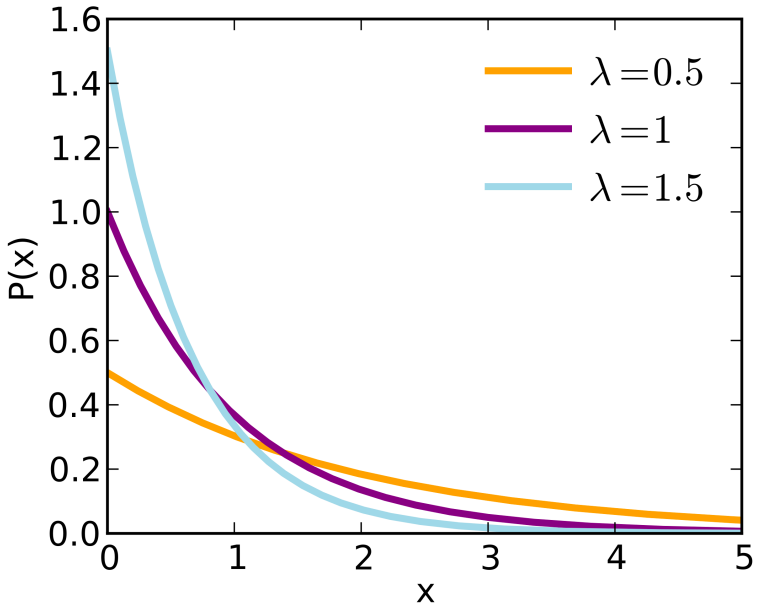
CDF

Exponential Distribution

The exponential distribution is often concerned with the amount of time until some specific event occurs, there are fewer large values and more small values:

- the amount of time (beginning now) until an earthquake occurs has an exponential distribution
- the amount of money customers spend in one trip to the supermarket

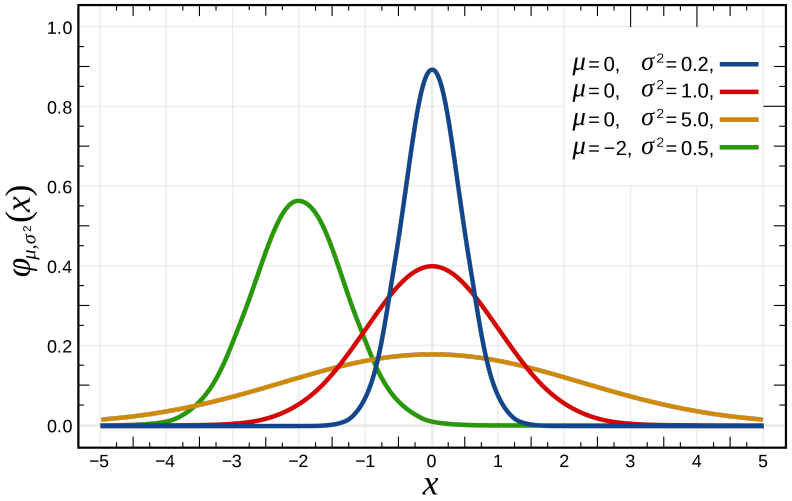
$$f(x) = \lambda e^{-\lambda x}$$



Normal / Gaussian Distribution

The sum or average of independent events always follows normal distribution, no matter what's the original distribution is.

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

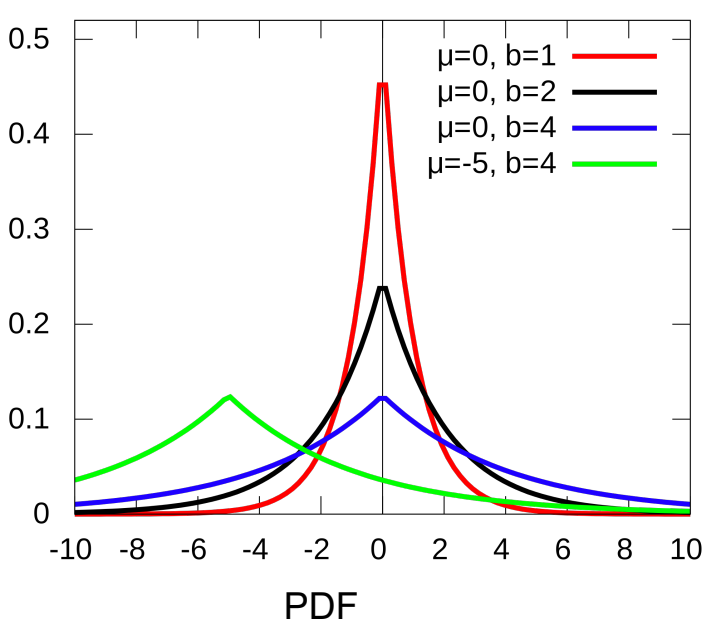


PDF

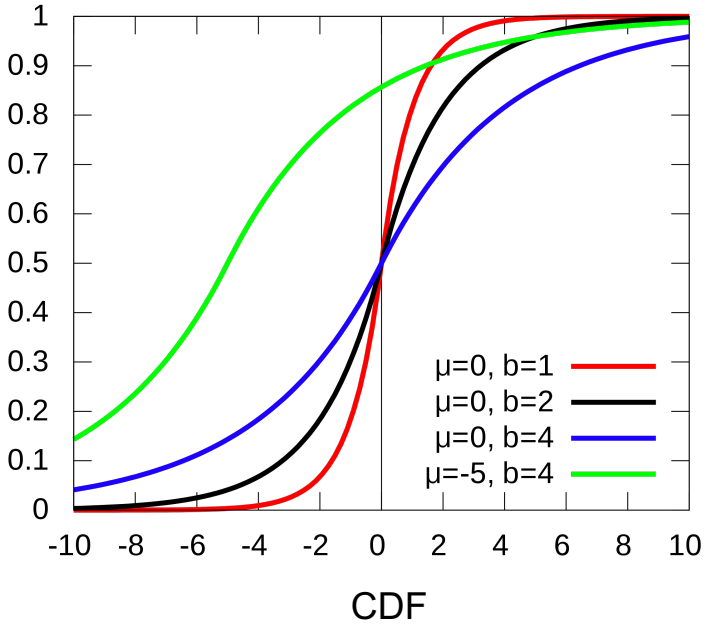
CDF

Laplace Distribution

Consider an experiment which consists of choosing a point from the interval [a, b] such that “all points are equally likely to be chosen”



PDF

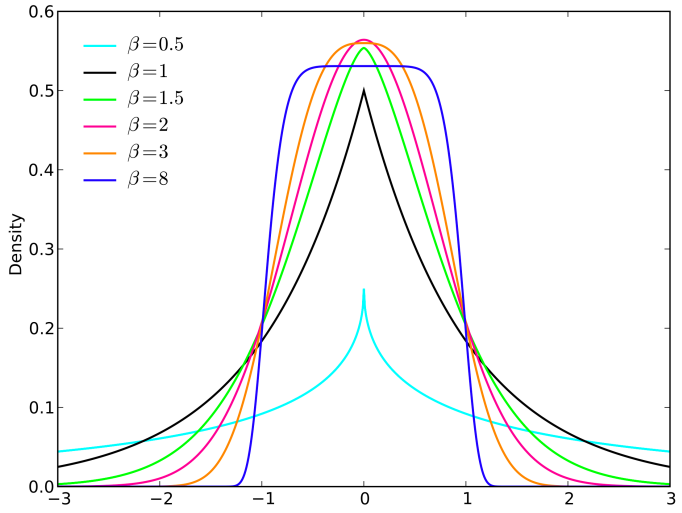


CDF

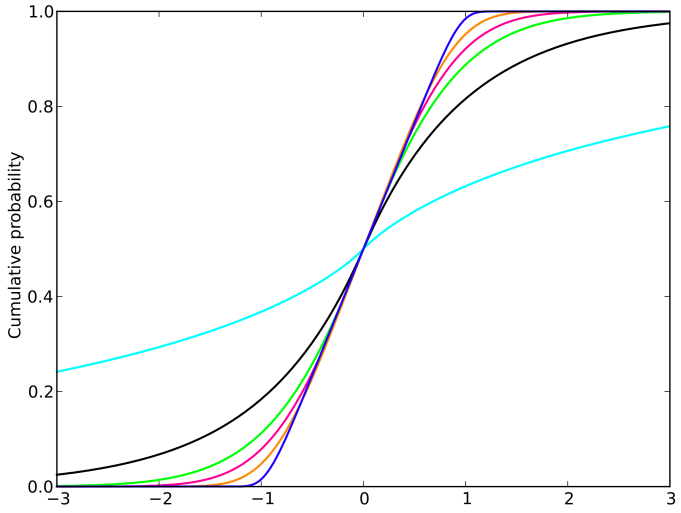
Parameters	μ location (real) $b > 0$ scale (real)
Support	$\mathbb{R}$
PDF	$\frac{1}{2b}\exp\left(-\frac{\lvert x-\mu\rvert}{b}\right)$
CDF	$\begin{cases} \frac{1}{2}\exp\left(\frac{x-\mu}{b}\right) & \text{if } x \le \mu \\ 1-\frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \ge \mu \end{cases}$
Quantile	$\begin{cases} \mu + b\ln(2F) & \text{if } F \le \frac{1}{2} \\ \mu - b\ln(2-2F) & \text{if } F \ge \frac{1}{2} \end{cases}$
Mean	μ
Median	μ
Mode	μ
Variance	$2b^2$
Skewness	0
Ex. kurtosis	3
Entropy	$\log(2be)$
MGF	$\frac{\exp(\mu t)}{1-b^2t^2}$ for $\lvert t\rvert < 1/b$

Generalized Gaussian Distribution

- Binomial distribution, n->infinity
- Central Limit Theorem



PDF

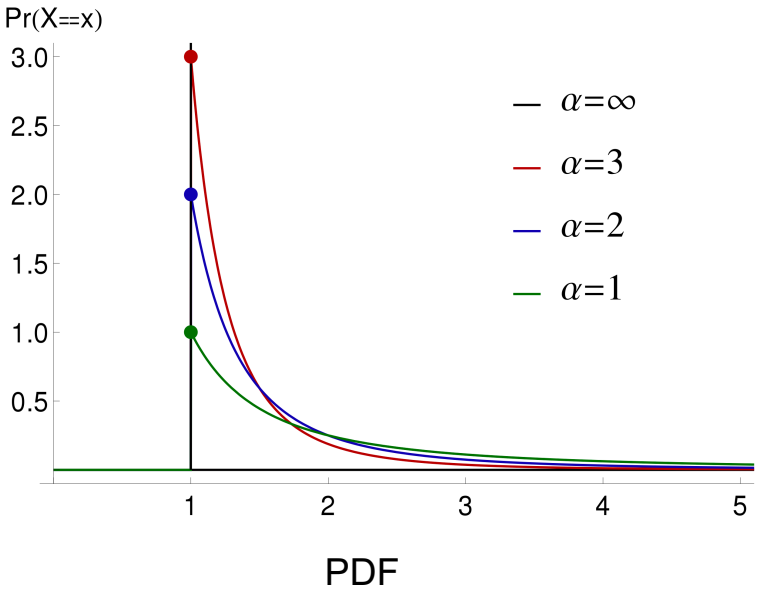


CDF

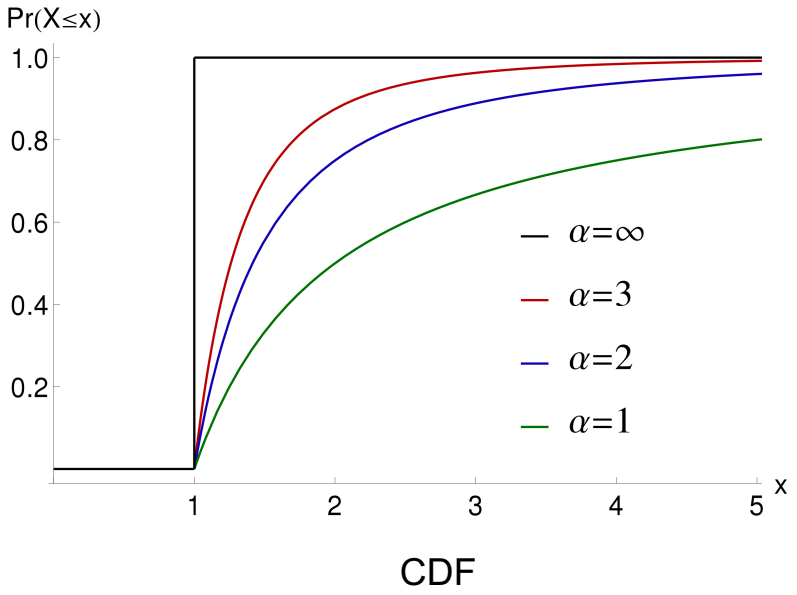
Parameters	μ location (real) α scale (positive, real) β shape (positive, real)
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{\beta}{2\alpha\Gamma(1/\beta)}\, e^{-(\lvert x-\mu\rvert/\alpha)^\beta}$ Γ denotes the gamma function
CDF	$\frac{1}{2} + \operatorname{sgn}(x-\mu)\frac{\gamma\left[1/\beta, \left(\frac{\lvert x-\mu\rvert}{\alpha}\right)^\beta\right]}{2\Gamma(1/\beta)}$ γ denotes the lower incomplete gamma function
Mean	μ
Median	μ
Mode	μ
Variance	$\frac{\alpha^2\Gamma(3/\beta)}{\Gamma(1/\beta)}$
Skewness	0
Ex. kurtosis	$\frac{\Gamma(5/\beta)\Gamma(1/\beta)}{\Gamma(3/\beta)^2} - 3$
Entropy	$\frac{1}{\beta} - \log\left[\frac{\beta}{2\alpha\Gamma(1/\beta)}\right]$ [1]

Pareto Distribution

Consider an experiment which consists of choosing a point from the interval [a, b] such that “all points are equally likely to be chosen”



PDF

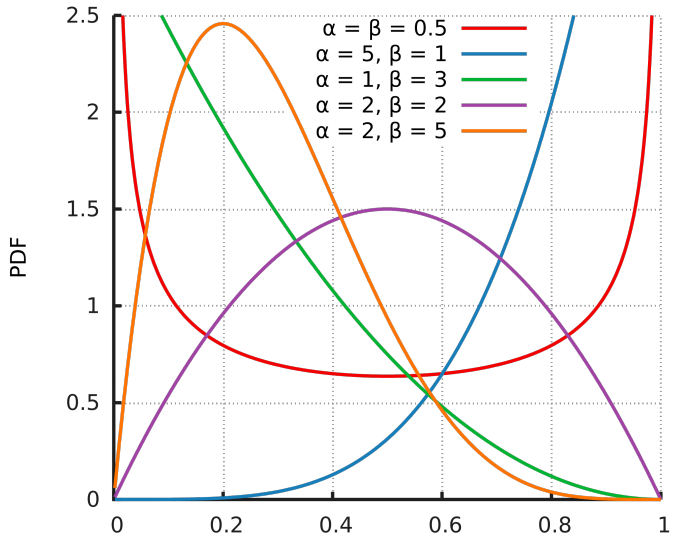


CDF

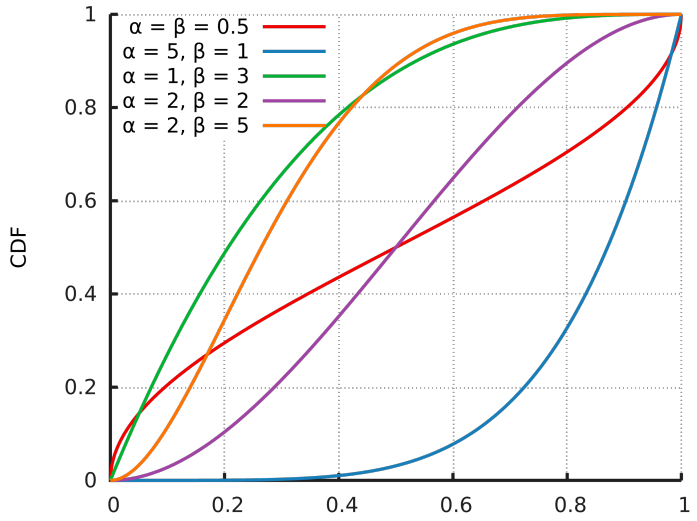
Parameters	$x_\mathrm{m} > 0$ scale (real) $\alpha > 0$ shape (real)
Support	$x \in [x_\mathrm{m}, \infty)$
PDF	$\frac{\alpha x_\mathrm{m}^\alpha}{x^{\alpha+1}}$
CDF	$1 - \left(\frac{x_\mathrm{m}}{x}\right)^\alpha$
Mean	$\begin{cases} \infty & \text{for } \alpha \le 1 \\ \frac{\alpha x_\mathrm{m}}{\alpha - 1} & \text{for } \alpha > 1 \end{cases}$
Median	$x_\mathrm{m} \sqrt[\alpha]{2}$
Mode	x_m
Variance	$\begin{cases} \infty & \text{for } \alpha \le 2 \\ \frac{x_\mathrm{m}^2 \alpha}{(\alpha-1)^2(\alpha-2)} & \text{for } \alpha > 2 \end{cases}$
Skewness	$\frac{2(1+\alpha)}{\alpha-3}\sqrt{\frac{\alpha-2}{\alpha}}$ for $\alpha > 3$
Ex. kurtosis	$\frac{6(\alpha^3+\alpha^2-6\alpha-2)}{\alpha(\alpha-3)(\alpha-4)}$ for $\alpha > 4$
Entropy	$\log\left(\left(\frac{x_\mathrm{m}}{\alpha}\right) e^{1+\frac{1}{\alpha}}\right)$
MGF	$\alpha(-x_\mathrm{m}t)^\alpha \Gamma(-\alpha, -x_\mathrm{m}t)$ for $t < 0$

Beta Distribution

Consider an experiment which consists of choosing a point from the interval [a, b] such that “all points are equally likely to be chosen”



PDF

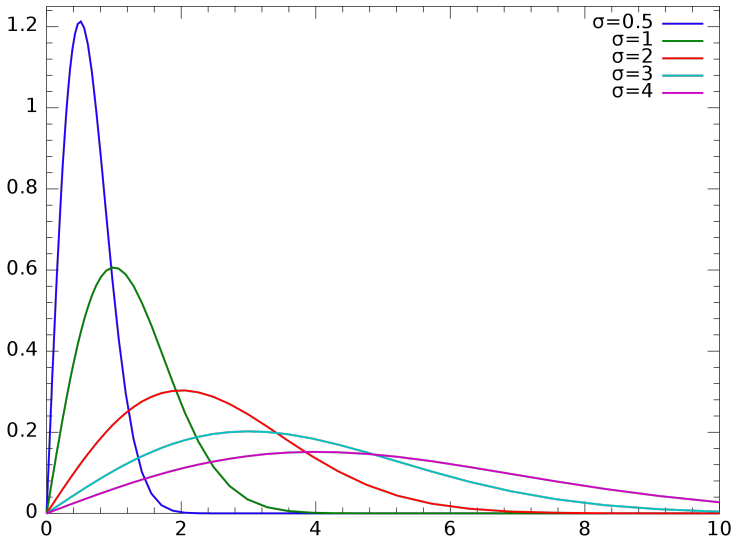


CDF

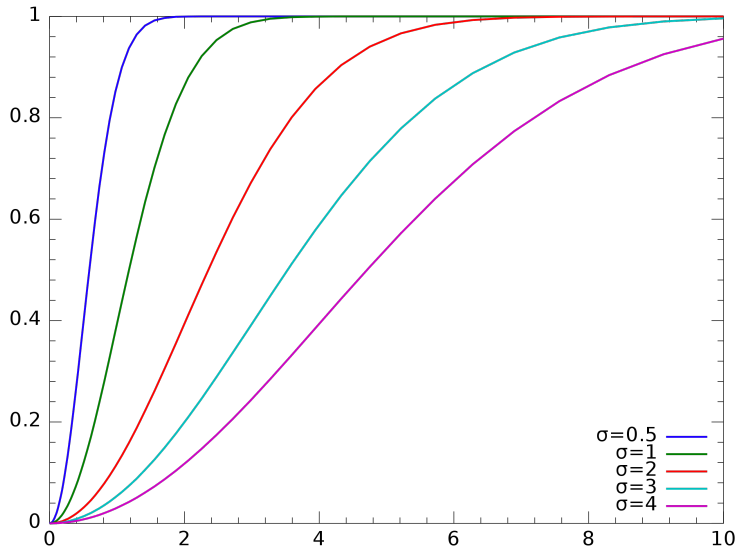
Parameters	$x_\mathrm{m} > 0$ scale (real) $\alpha > 0$ shape (real)
Support	$x \in [x_\mathrm{m}, \infty)$
PDF	$\frac{\alpha x_\mathrm{m}^\alpha}{x^{\alpha+1}}$
CDF	$1 - \left(\frac{x_\mathrm{m}}{x}\right)^\alpha$
Mean	$\begin{cases} \infty & \text{for } \alpha \le 1 \\ \frac{\alpha x_\mathrm{m}}{\alpha - 1} & \text{for } \alpha > 1 \end{cases}$
Median	$x_\mathrm{m} \sqrt[\alpha]{2}$
Mode	x_m
Variance	$\begin{cases} \infty & \text{for } \alpha \le 2 \\ \frac{x_\mathrm{m}^2 \alpha}{(\alpha-1)^2(\alpha-2)} & \text{for } \alpha > 2 \end{cases}$
Skewness	$\frac{2(1+\alpha)}{\alpha-3}\sqrt{\frac{\alpha-2}{\alpha}}$ for $\alpha > 3$
Ex. kurtosis	$\frac{6(\alpha^3+\alpha^2-6\alpha-2)}{\alpha(\alpha-3)(\alpha-4)}$ for $\alpha > 4$
Entropy	$\log\left(\left(\frac{x_\mathrm{m}}{\alpha}\right) e^{1+\frac{1}{\alpha}}\right)$
MGF	$\alpha(-x_\mathrm{m}t)^\alpha \Gamma(-\alpha, -x_\mathrm{m}t)$ for $t < 0$

Rayleigh Distribution

Consider an experiment which consists of choosing a point from the interval [a, b] such that “all points are equally likely to be chosen”



PDF



CDF

Parameters	scale: $\sigma > 0$
Support	$x \in [0, \infty)$
PDF	$\frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}$
CDF	$1 - e^{-x^2/(2\sigma^2)}$
Quantile	$Q(F;\sigma) = \sigma\sqrt{-2\ln(1-F)}$
Mean	$\sigma\sqrt{\frac{\pi}{2}}$
Median	$\sigma\sqrt{2\ln(2)}$
Mode	σ
Variance	$\frac{4-\pi}{2}\sigma^2$
Skewness	$\frac{2\sqrt{\pi}(\pi-3)}{(4-\pi)^{3/2}}$
Ex. kurtosis	$-\frac{6\pi^2 - 24\pi + 16}{(4-\pi)^2}$
Entropy	$1 + \ln\left(\frac{\sigma}{\sqrt{2}}\right) + \frac{\gamma}{2}$
MGF	$1 + \sigma t e^{\sigma^2 t^2/2}\sqrt{\frac{\pi}{2}}\left(\operatorname{erf}\left(\frac{\sigma t}{\sqrt{2}}\right) + 1\right)$

Sample Statistics

Sample Statistics

Pythagorean Mean

- Arithmetic Mean

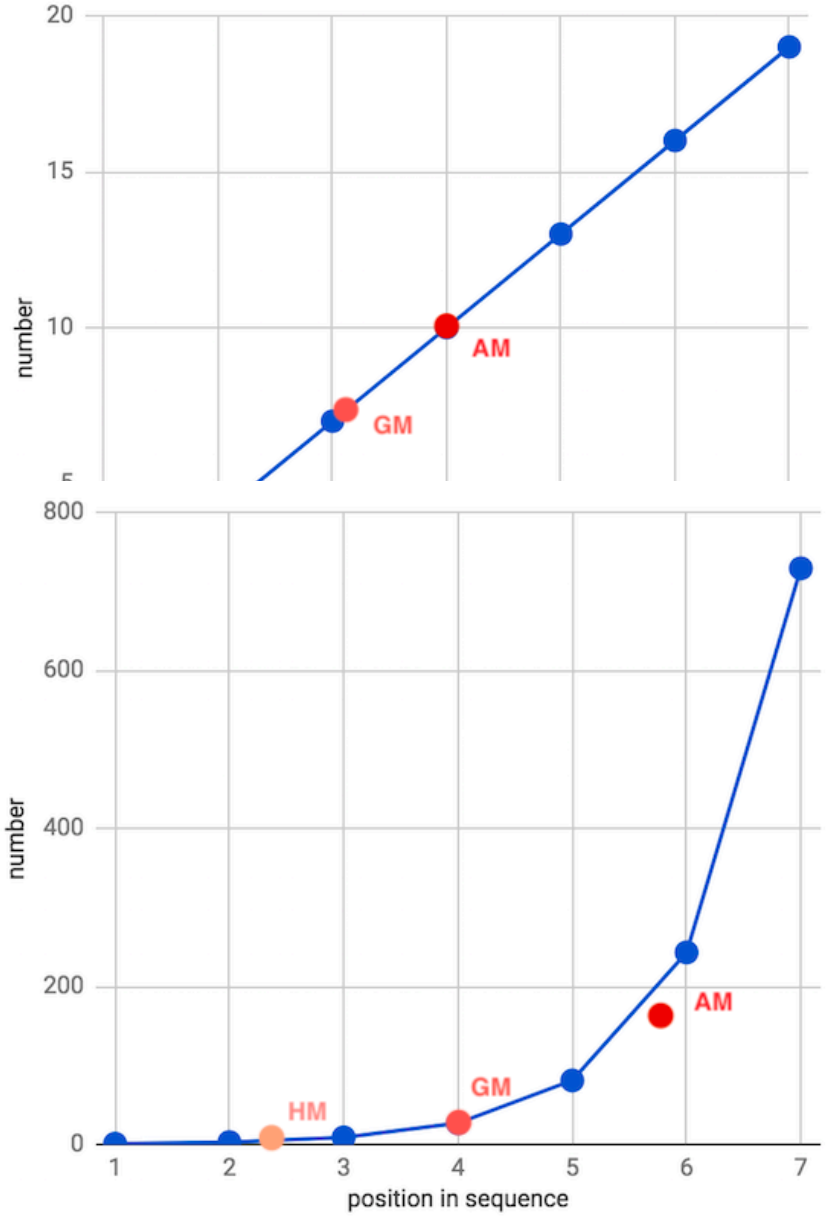
$$AM(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \cdots + x_n)$$

- Geometric Mean

$$GM(x_1, \ldots, x_n) = \sqrt[n]{|x_1 \times \cdots \times x_n|}$$

- Harmonic Mean

$$HM(x_1, \ldots, x_n) = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$$



Q & A

Thank You!