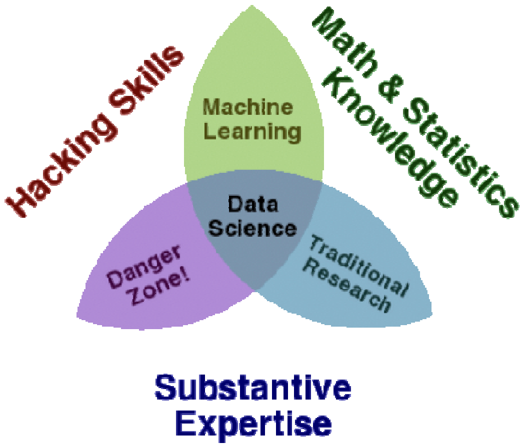


COMM054: Data Science Principles & Practices

Introduction Distributions Parameters, MLE & MAP

Dr. Manal Helal

7 November 2019

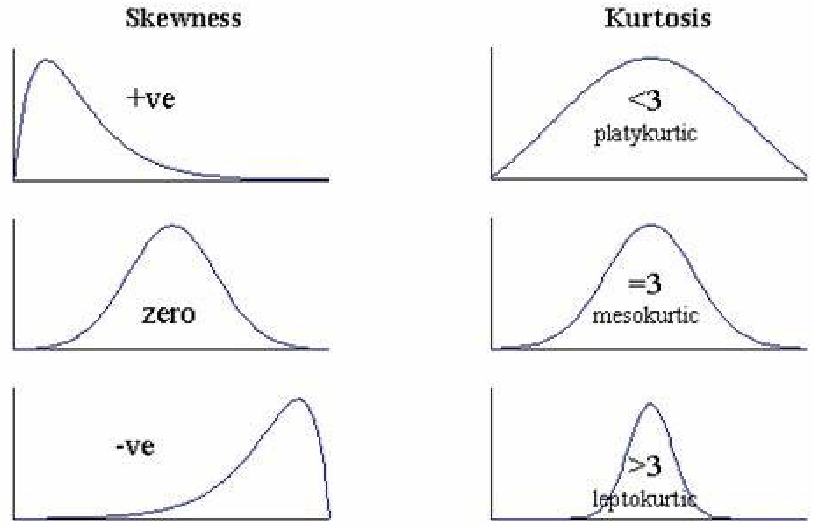
Outline

- Distributions Types
- Parameters
- Non-Parametric Analysis / KDE
- MLE
- MAP

Distribution Types

- The lab first python notebook introduces different types of distributions and their parameters:
  - Continuous: Normal /Gaussian, Log Normal, Beta, gamma, Uniform
  - Discrete: Uniform, Bernoulli, Binomial, Poisson, Geometric
- Kernel Distribution Estimation is applied when you do not know the distribution and want to estimate its parameters.

Distributions’ Parameters



- Measures of central tendency: Mode, Median, Mean, Geometric mean, Harmonic mean.
- Measures of statistical dispersion: Variance, Geometric variance and covariance, Mean absolute deviation around the mean, Mean absolute difference.
- Skewness: is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.
- Kurtosis: is a measure of the "tailedness" of the probability distribution of a real-valued random variable.

- These parameters help estimate the expected value of the random variable.

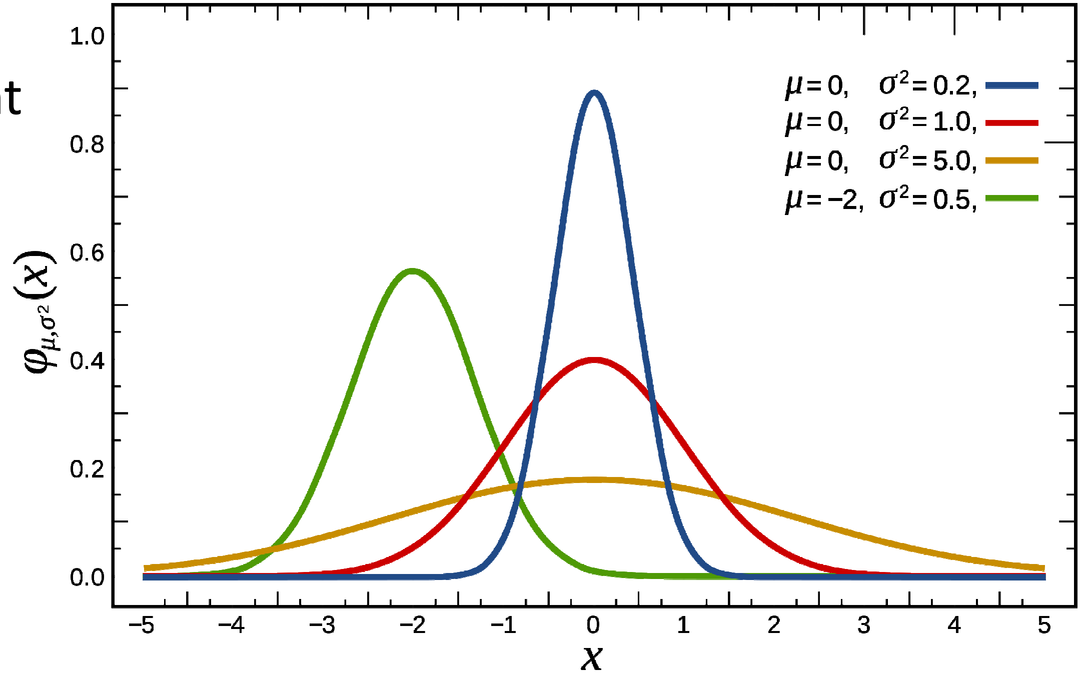
Normal /Gaussian Distribution

- It is a continuous probability distribution that is useful because of the central limit theorem.
- It has 2 parameters
  - μ is the mean or expectation of the distribution (and also its median and mode)
  - σ is the standard deviation and σ^2 is the variance

- The expectation of X conditioned on the event that X lies in an interval [a,b]:

$$\mathrm{E}[X \mid a < X < b] = \mu - \sigma^2 \frac{f(b) - f(a)}{F(b) - F(a)}$$

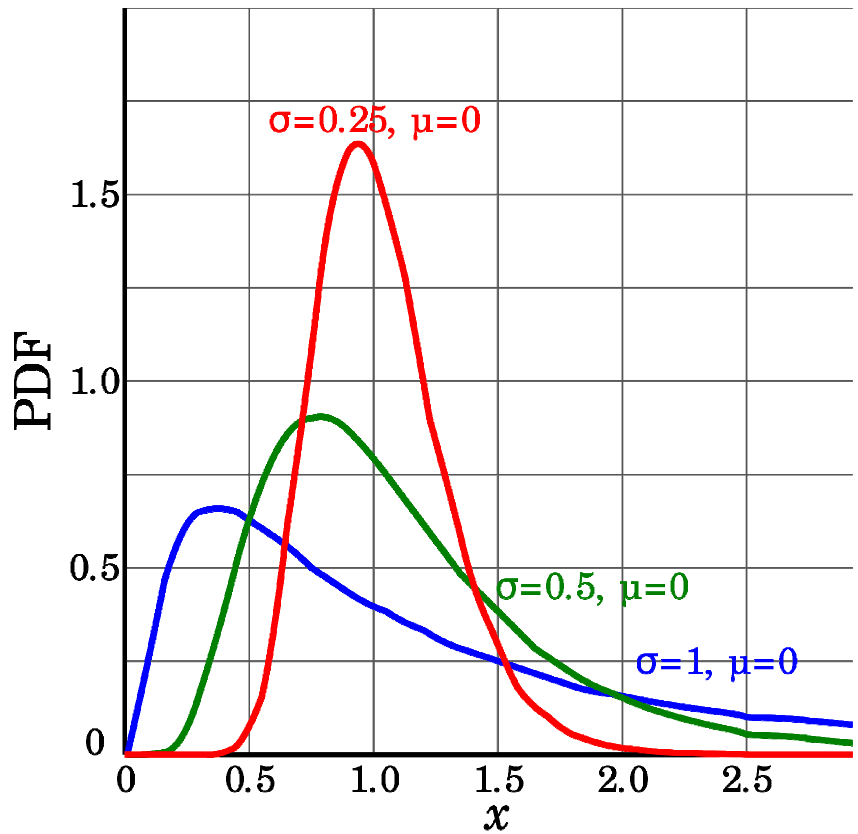
where f and F respectively are the density and the cumulative distribution function of X .



Log Normal Distribution

- It is a continuous probability distribution of a random variable whose logarithm is normally distributed. $Y = \ln(X)$.
- It has 2 parameters:
  - μ is the mean or expectation of the distribution (and also its median and mode)
  - σ is the standard deviation and σ^2 is the variance

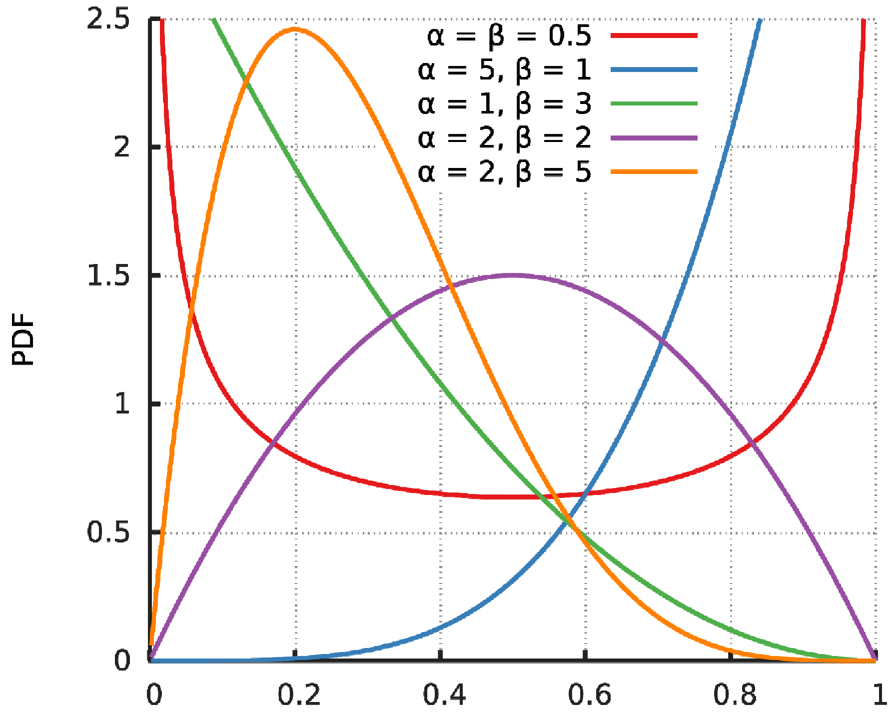
$$\mathrm{E}[X] = e^{\mu + \frac{1}{2}\sigma^2}$$



Beta Distribution

- It is a continuous probability distribution defined on the interval [0, 1].
- It has 2 positive shape parameters, denoted by α and β, that appear as exponents of the random variable and control the shape of the distribution. It is a special case of the Dirichlet distribution.

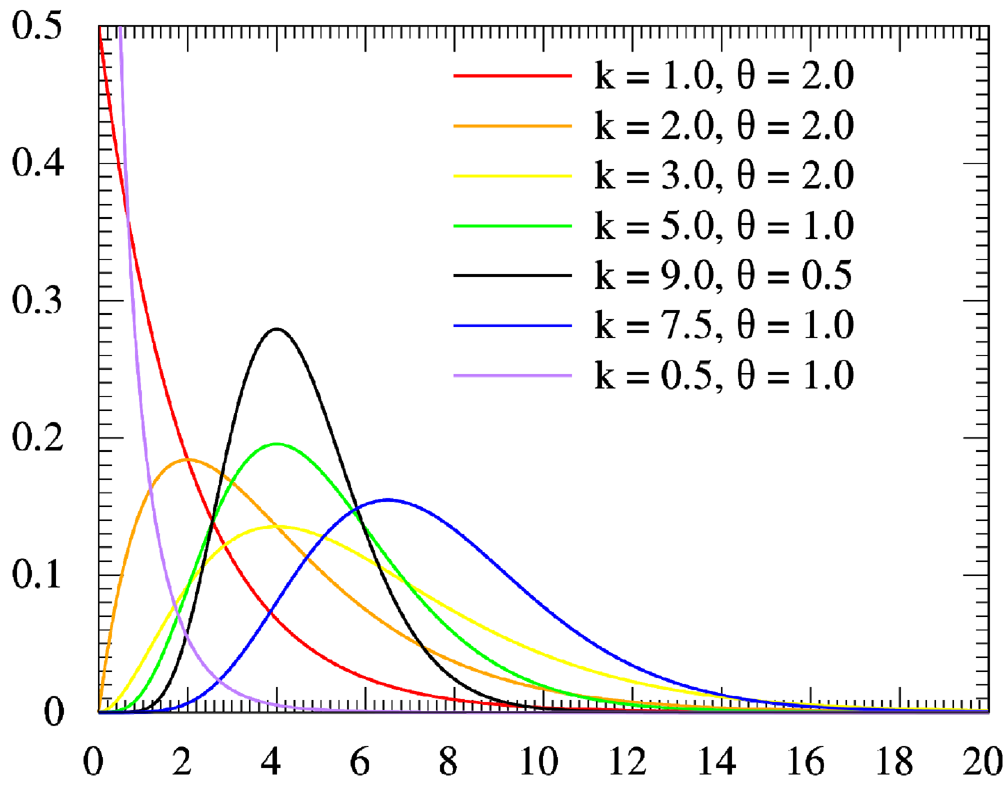
$$\begin{aligned}\mu = \operatorname{E}[X] &= \int_0^1 x f(x;\alpha,\beta)\,dx \\ &= \int_0^1 x\,\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\,dx \\ &= \frac{\alpha}{\alpha+\beta} \\ &= \frac{1}{1+\frac{\beta}{\alpha}}\end{aligned}$$



Gamma Distribution

- It is a two-parameter family of continuous probability distributions.
- The exponential distribution, Erlang distribution, and chi-squared distribution are special cases of the gamma distribution.
- There are three different parametrizations in common use:
  - With a shape parameter k and a scale parameter θ.
  - With a shape parameter α = k and an inverse scale parameter β = 1/θ, called a rate parameter.
  - With a shape parameter k and a mean parameter μ = kθ = α/β.

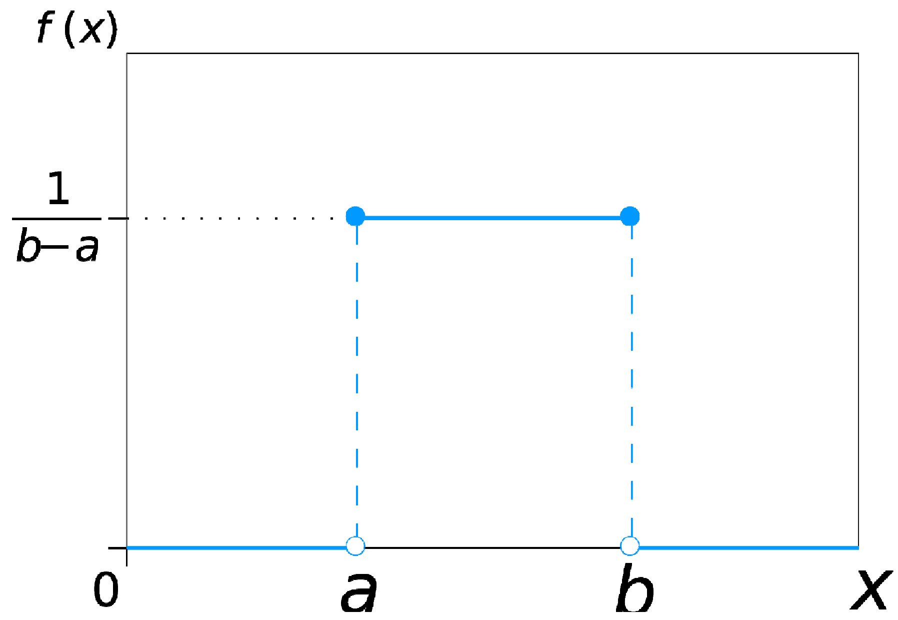
$$\mathbf{E}[X] = k\theta$$



Uniform Distribution

- It is a continuous probability distribution in which all intervals of the same length on the distribution's support are equally probable.
- The support is defined by the two parameters, *a* and *b*, which are its minimum and maximum values, or one parameter n = b - a.
- The expectation of X:

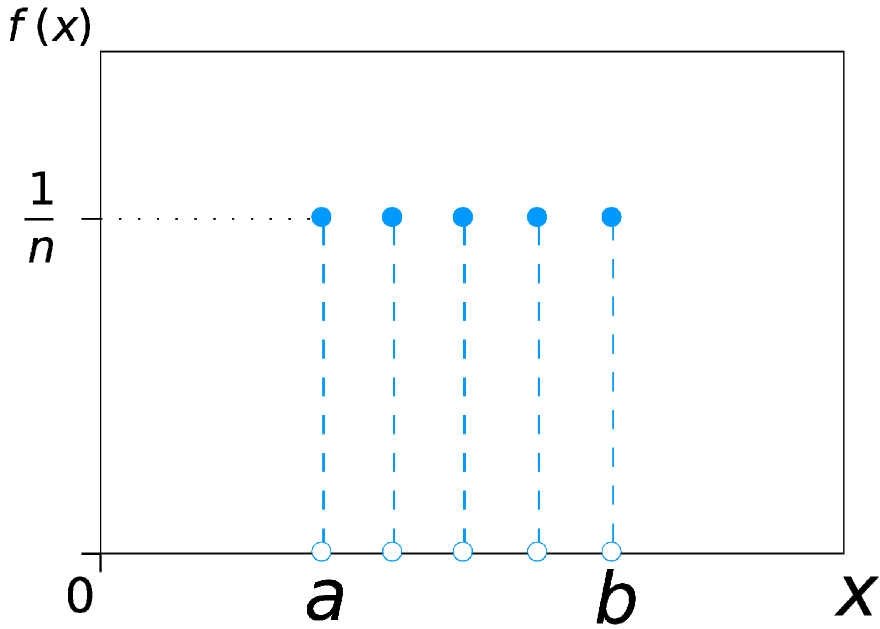
$$\mathbf{E}[X] = \frac{1}{2}(a+b)$$



Uniform Distribution (Discrete)

- It is a discrete symmetric probability distribution whereby a finite number of values are equally likely to be observed; every one of n values has equal probability 1/n.
- The discrete uniform distribution itself is inherently non-parametric.
- It is convenient, however, to represent its values generally by all integers in an interval [*a*,*b*], or interval [1,*n*] with the single parameter *n*
- The expectation of X:

$$\mathbf{E}[X] = \mu = P(X=1) = \frac{1}{n}$$



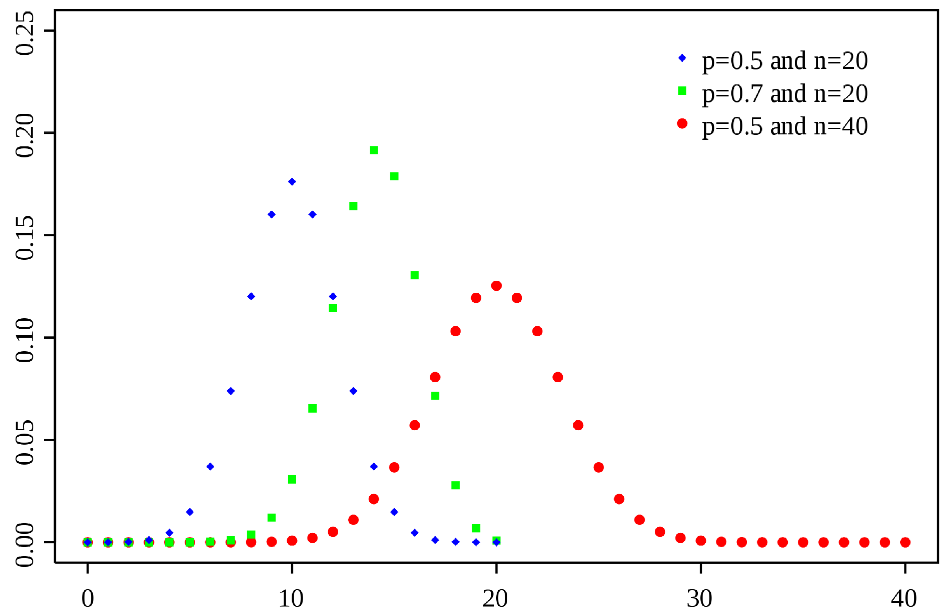
Bernoulli Distribution

- It is a discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$ (a yes–no question experiment)

$$\mathbf{E}[X] = p$$

Binomial Distribution

- It is a discrete probability distribution of the number of successes in a sequence of n independent experiments, each is Bernoulli trial or Bernoulli experiment with outcome as a random variable which takes the value 1 with probability p and the value 0 with probability q =1−p (a yes–no question experiment)
- If $X \sim B(n, p)$, that is, X is a binomially distributed random variable, n being the total number of experiments and p the probability of each experiment yielding a successful result:

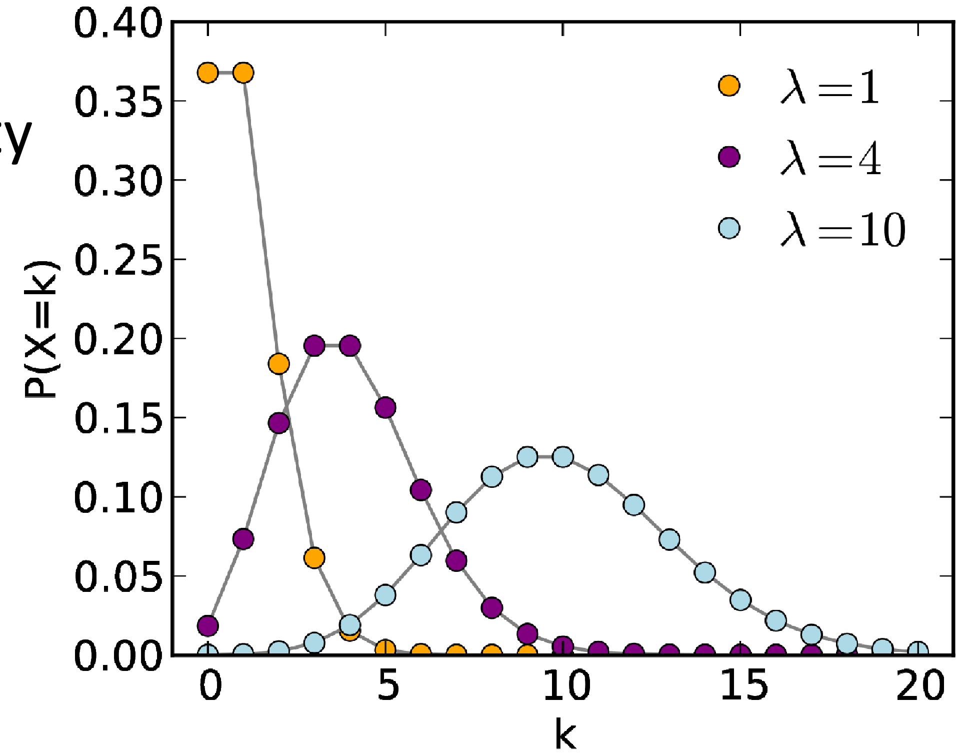


$\mathbf{E}[X] = np$

Poisson Distribution

- It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space (distance, area or volume) if these events occur with a known constant rate and independently of the time since the last event.

- Parameters: The vertical axis is the probability of *k* occurrences given λ *:* the expected number of occurrences, which need not be an integer.

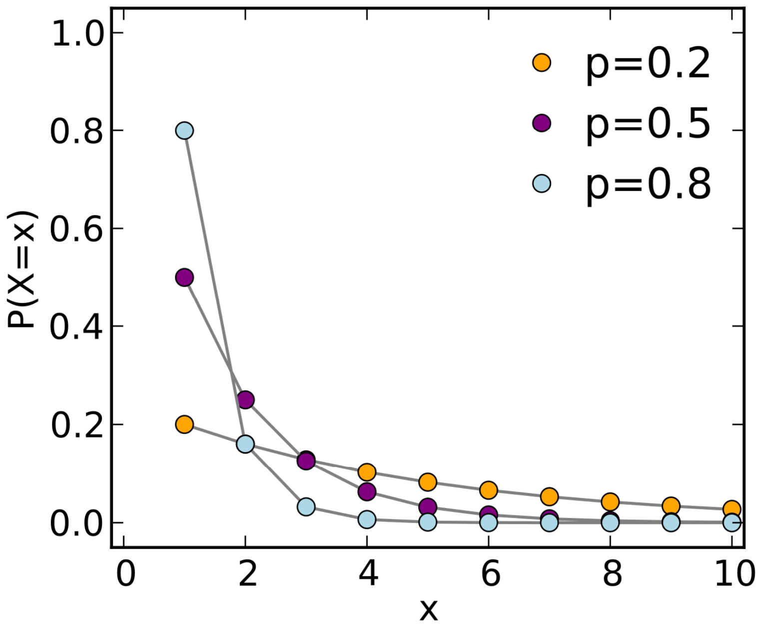


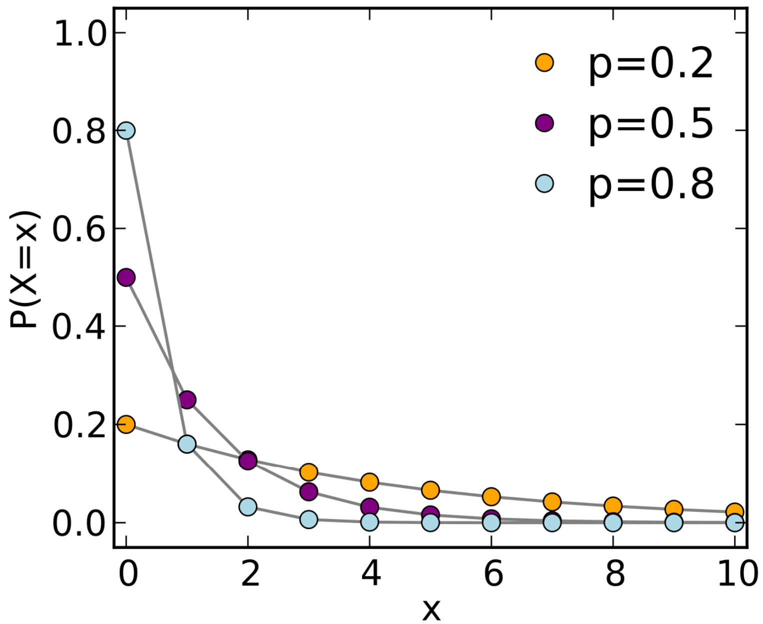
$\lambda = \mathbf{E}[X] = \mathrm{Var}(X)$

Geometric Distribution

- It is a discrete probability distribution with 1 parameter p that is either of two discrete probability distributions:
  - The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set { 1, 2, 3, ... }
  - The probability distribution of the number $Y = X - 1$ of failures before the first success, supported on the set { 0, 1, 2, 3, ... }

$\mathbf{E}[X] = \frac{1}{P}$





MLE & MAP

- Wikipedia:
  - In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the parameters. MLE can be seen as a special case of the maximum a posteriori estimation (MAP) that assumes a uniform prior distribution of the parameters, or as a variant of the MAP that ignores the prior and which therefore is unregularized.

MLE & MAP

Frequentists

- Wikipedia:
  - **In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the parameters. MLE can be seen as a special case of the maximum a posteriori estimation (MAP) that assumes a uniform prior distribution of the parameters, or as a variant of the MAP that ignores the prior and which therefore is unregularized.**

Bayesians

Bayes theorem

- **Bayes theorem** reverses the direction of the **dependencies**:

Likelihood: Probability of collecting this data when our hypothesis is true

Prior: The Probability of hypothesis being true before collecting the data

Posterior: The Probability of hypothesis being true given the data collected

$$\mathbf{P}\ (\boldsymbol{B}|\boldsymbol{A}) = \frac{\mathbf{P}\ (\mathbf{A}\mid\mathbf{B})\mathbf{P}(\boldsymbol{B})}{\boldsymbol{P}(\boldsymbol{A})}$$

Evidence: What is the probability of collecting this data under all possible hypothesis

For **Example 2** by Bayes theorem:

P(B|A) = (1·9/36)/(25/36) = 9/25, exactly what we got before.

https://gerardnico.com/data_mining/bayes

MLE vs MAP

- Both compute the best parameters/coefficients for a model, by computing a single estimate, instead of a full distribution.
- MLE is the probability of some specific coefficients, ignoring the prior and the evidence because of assuming they have uniform prior distribution — all coefficient values are equally likely.
- Bayesians MAP frame the question the exact opposite way, and same do MLE.
- *Probability attaches to possible results: mutually exclusive, exhaustive, and sums to 1*
- *likelihood attaches to hypotheses: neither mutually exclusive nor exhaustive*

https://towardsdatascience.com/a-gentle-introduction-to-maximum-likelihood-estimation-9fbff27ea12f

MLE is Frequentist, Bayesian motivated

- `posterior = likelihood x prior / evidence`
- `MLE ≈ likelihood, i.e.` $p(B|A) \approx L(A|B)$
- $p(b_1,b_2,\ldots,b_n|A) \approx L(A|b_1,b_2,\ldots,b_n)$
- $\because$ $p(A,B) = p(A)p(B)$
- $\therefore$ $L(A|b_1,b_2,\ldots,b_n) = p(b_1|A)p(b_2|A),\ldots,p(b_n|A) = \prod p(b_i|A)$
- $\max_{A}\{\prod_i p(b_i|A)\}$
- Remember logs from last week: turning product function into a sum function:

$$\max_{A}\{\ln\{\prod_i p(b_i|A)\}\}$$

- Further simplify:

$$L(A|B) = \max_{A}\{\sum_i \ln\{p(b_i|A)\}\}$$

MLE in Regression Problems

- MLE works great for classification problems with discrete outcomes, but we have to use different distribution functions, depending on how many classes we have.
- Regression has a continuous outcome, can be calculated by Ordinary Least squares (OLS).
- In (OLS) the residuals are normally distributed around mean zero, our fitted OLS model literally becomes the embodiment of a maximum expectation of y. And our probability distribution is Normal.
- In the lab you will see that MLE and OLS will estimate the same coefficients of a normal distributed variable.

MAP Estimation

- posterior = likelihood x prior / evidence
- We are ignoring the normalizing constant as we are strictly speaking about optimization here, so proportionality is sufficient.
- posterior ≈ likelihood x prior
- P $(B|A) \approx P\ (A \mid B)P(B)$
- $\because$ the likelihood is $P\ (A \mid B)$, which is MLE L(B/A) $= \underset{B}{max}\{\sum_i \ln\{p(a_i|B)\}\}$
- $\therefore$ MAP (B) = $\underset{B}{max}\{\sum_i \ln\{p(a_i|B)\}\}$ x $\ln P(B)$

- Comparing both MLE and MAP equation, the only thing differs is the inclusion of prior P(θ) in MAP. Using constant prior, reduces MAP to MLE

References

- https://github.com/ibab/python-mle
- https://zhiyzuo.github.io/MLE-vs-MAP/
- https://github.com/EgroegCai/MLE-vs-MAP

- Wikipedia and Module Textbooks