Probability Review

Dipanwita Guhathakurta

May 16, 2021

Approaches for understanding probability

Classical approach: If we perform an experiment in which there are n equally likely outcomes, and the event A consists of exactly m of these outcomes, we say that the probability of A, i.e. P(A) is $\frac{m}{n}$.
Empirical Approach: Suppose we toss a coin 1000 times. Let's say number of heads is n_H. Total number of tosses is n.
P(H) $= n_H/n$
To get a better estimate we can repeat our experiment a large number of times, so as $n \to \infty$ this ratio n_H/n converges to $\frac{1}{2}$
Axiomatic approach: A more general mathematical approach - probability is any function from events to $\mathbb{R}$ satisfying three axioms.

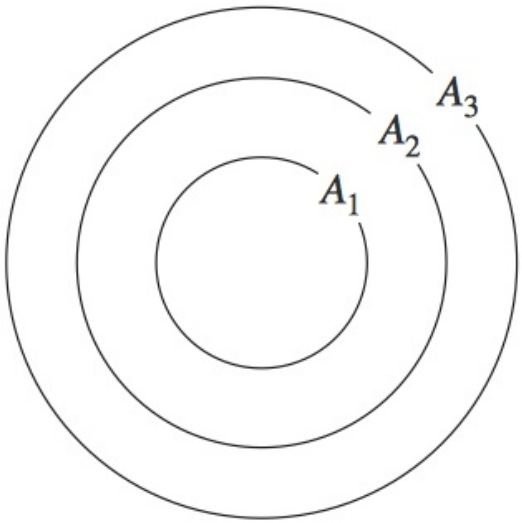
Probability Space

Probability space: Probability space $(\Omega, \mathcal{F}, P)$ is a mathematical construct that provides a formal model of a random process or "experiment". It consists of the following:

- **Sample space:** the set Ω of all possible outcomes
- **Event space:** a set of events $\mathcal{F}$, an event being a set of outcomes in the sample space.
- **Probability function:** a function P assigned to each event in the event space that satisfies:
  1. For every event A, $P(A) \geq 0$
  2. $P(\Omega) = 1$
  3. If E_1, E_2, ..., such that $E_i \cap E_j = \phi$ (all the events are mutually disjoint), then $P(\cup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$

Continuity of Probability

- Let A_1, A_2, ... be an increasing sequence of events: i.e.
  $A_1 \subseteq A_2 \subseteq ... \subseteq A_n \subseteq A_{n+1} \subseteq ...$
  Then $P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n)$
  Note: because $A_1 \subseteq A_2 \subseteq ...$, we have: $\lim_{n\to\infty} A_n = \cup_{n=1}^{\infty} A_n$
- Let B_1, B_2, ... be a decreasing sequence of events: i.e.
  $B_1 \supseteq B_2 \supseteq ... \supseteq B_n \supseteq B_{n+1} \supseteq ...$
  Then $P(\lim_{n\to\infty} B_n) = \lim_{n\to\infty} P(B_n)$
  Note: because $B_1 \subseteq B_2 \subseteq ...$, we have: $\lim_{n\to\infty} B_n = \cap_{n=1}^{\infty} B_n$



Conditional Probability

Conditional probability is the probability of one event occurring with some relationship to one or more other events. For example, given two events A and B, the conditional probability of A given B is defined to be:

$$P(A) = \frac{P(A \cap B)}{P(B)}$$

where $P(A \cap B)$ is the probability that both events A and B occur.

- If $P(A \mid B) = P(A)$, then events A and B are said to be independent.i.e, $P(A \cap B) = P(A)P(B)$
- If $P(A \mid B) = 0$ then events A and B are mutually exclusive.i.e, $P(A \cap B) = 0$

Examples: Mutually exclusive events mean that if one event occurs, the other cannot occur. However, when we say that two events are independent, it means that the occurrence and outcome of one event will not have any impact on that of the other.

- Outcomes of rolling of a die are mutually exclusive events,i.e it is not possible to get a 5 and a 6 at the same time.
- Outcomes of rolling two different die are independent events,i.e the number we get on the first die is not dependent on the number we get on the second die.

Law of Total Probability

Partition:

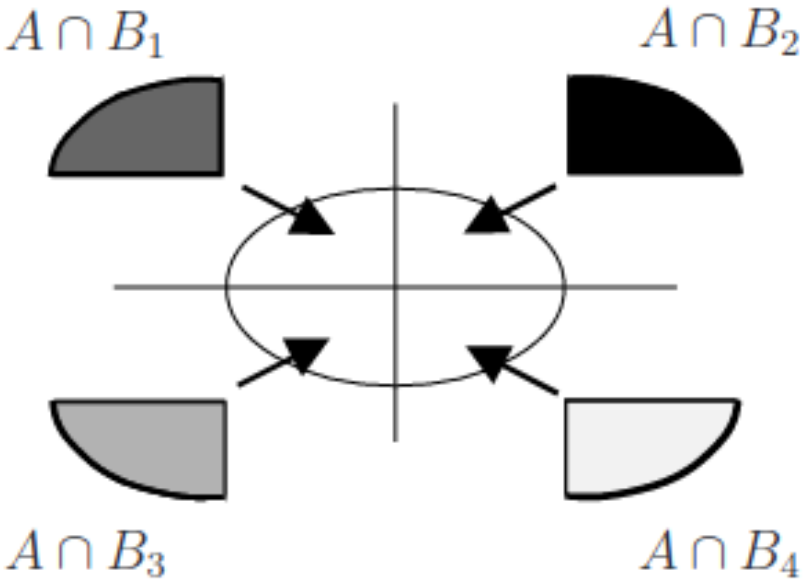
A partition of A is a collection of mutually exclusive events whose union is A .That is, sets B_1, B_2, . . . , B_k form a partition of A if:

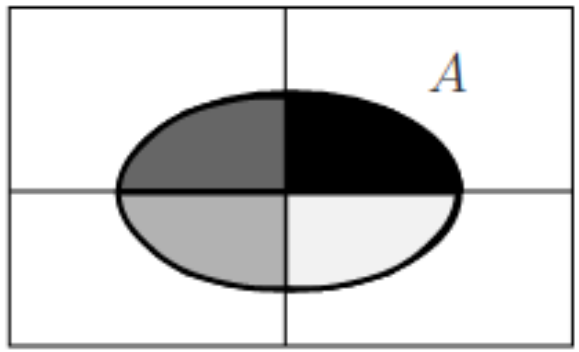
$B_i \cap B_j = \phi$ for all i, j with $i \neq j$,and $\cup_{i=1}^{k} B_i = B_1 \cup B_2 \cup .. \cup B_k = A$

The Partition Theorem/ Total Probability Theorem:

Let B_1, . . . , B_m form a partition of Ω . Then for any event A,

$$P(A) = \sum_{i=1}^{m} P(A \cap B_i) = \sum_{i=1}^{m} P(A \mid B_i) P(B_i)$$





Bayes' Theorem

Bayes' Theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. It is stated mathematically as the following equation:

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

where A and B are events and $P(B) \neq 0$.

- $P(A \mid B)$ is a conditional probability: the probability of event A occurring given that B is true. It is also called the posterior probability of A given B.
- $P(B \mid A)$ is also a conditional probability: the probability of event B occurring given that A is true.
- P(A) and P(B) are the probabilities of observing A and B respectively without any given conditions; they are known as the prior probabilities.

Random Variables

Informally, a random variable assigns a real number to every possible outcome of a random experiment. More formally, a random variable is a measurable function from X: $\Omega \to \mathbb{R}$ such that the pre-image $X^{-1}(-\infty, x] \in \mathcal{F}$ i.e, the event space.
For example: Suppose $\Omega = \{a,b,c\}$, $\mathcal{F} = \{\phi, \{a\}, \{b, c\}, \Omega\}$ and we define

$$X(\omega) = \begin{cases} 0, \omega = b \\ 1, \omega = a, c \end{cases}$$

Then, $X^{-1}(-\infty, x]$ includes $\{b\} \notin \mathcal{F}$, so X is not a random variable.
However, had we defined:

$$X(\omega) = \begin{cases} 0, \omega = a \\ 1, \omega = b, c \end{cases}$$

$X^{-1}(-\infty, x] \in \mathcal{F}$ as we vary x, so X is a random variable.

Distribution functions

We associate probability that X takes on a value in a measurable set $S \subseteq E$ as:

$$P(X \in S) = P(\{\omega \in \Omega \mid X(\omega) \in S\})$$

So, "How likely is it that the value of X is equal to 2?" is the same as the probability of the event $\omega : X(\omega) = 2$ which is often written as P(X=2) or $p_X(2)$ for short.

Cumulative Distribution Function (CDF) - The cumulative distribution function of a random variable X is given by:

$$F_X(x) = P(X \leq x) \tag{1}$$

Probability Density Function (PDF) - The probability distribution function of a random variable X is given by: $f_X(x) = \frac{d}{dx} F_X(x)$

So conversely,

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du \tag{2}$$

Properties of CDF and PDF

Properties of the cumulative distribution function:

- $\lim_{x\to-\infty} F_X(x) = 0, \quad \lim_{x\to+\infty} F_X(x) = 1$
- $F_X(x)$ is a non-decreasing function of x
- $F_X(x)$ is right-continuous.

Properties of the probability density function:

- $f_X(x) \geq 0 \quad \forall x$
- $\int_{-\infty}^{\infty} f_X(x)dx =$ area under the entire graph of f(x) $= 1$

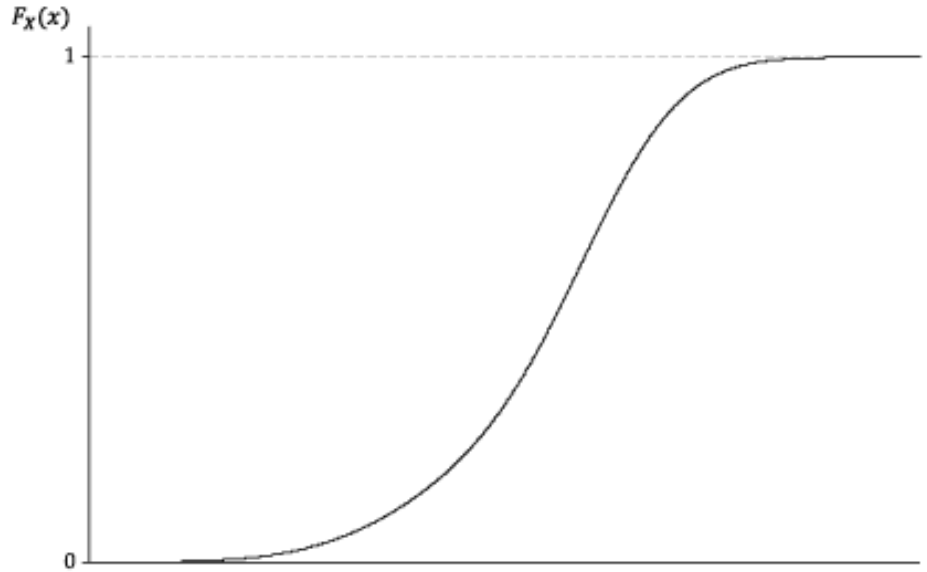
Continuous Random Variables

A continuous random variable is a random variable whose cumulative distribution function is continuous everywhere.

- Continuous random variables almost never take an exact prescribed value $\forall c \in \mathbb{R}: \ \Pr(X = c) = 0$
- Continuous random variables are usually characterized by their probability density functions (PDF).

A few example distributions:

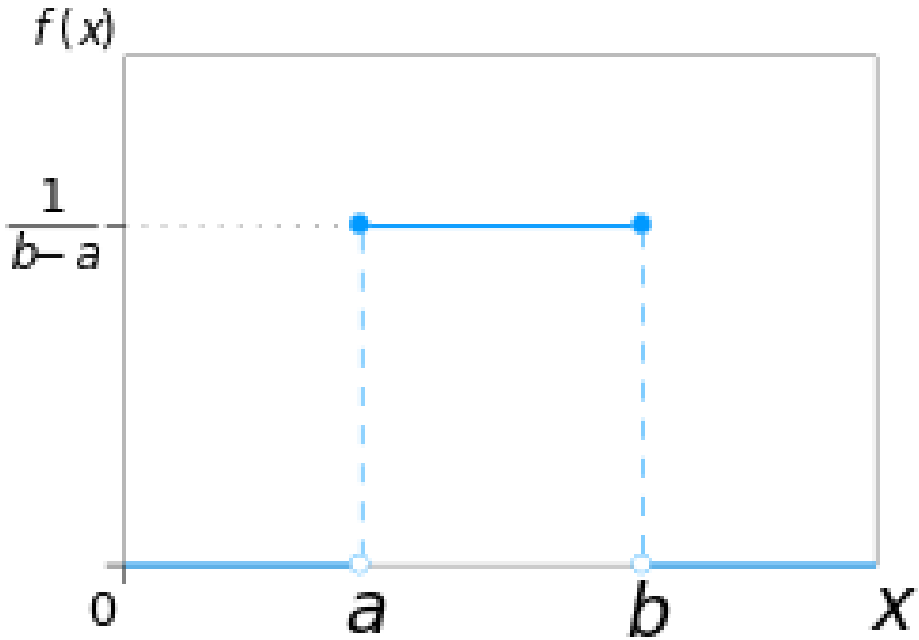
- Uniform Distribution
- Gaussian Distribution
- Exponential Distribution



Continuous RV - Common Distributions

- **Uniform distribution:**

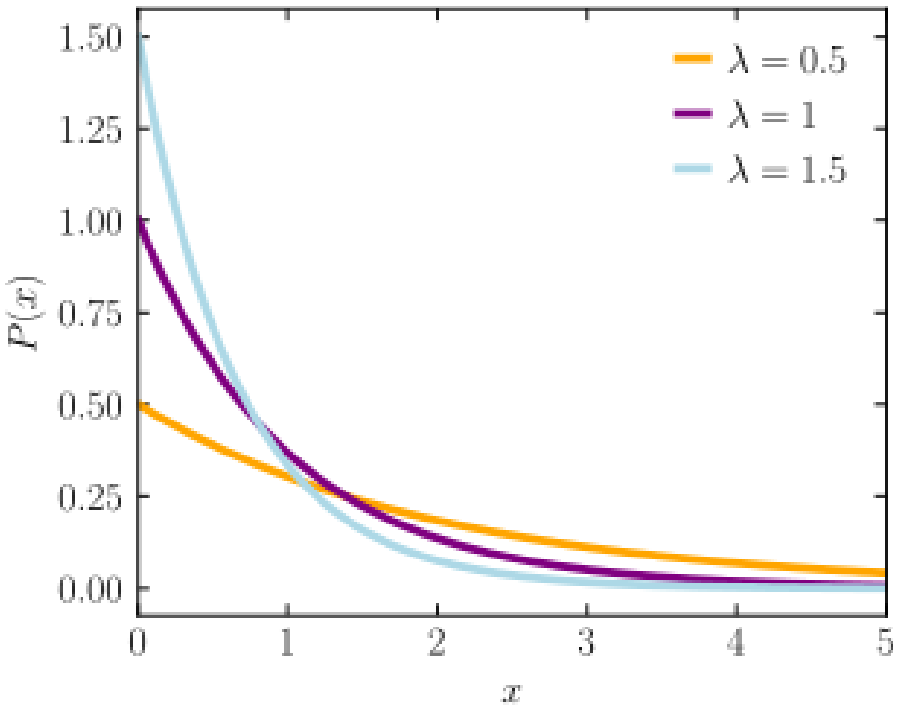
PDF: $\begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$



CDF: $\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases}$

Continuous RV - Common Distributions

- **Exponential distribution:**
  PDF: $\lambda e^{-\lambda x} \quad \forall \lambda \geq 0, x \in [0, \infty)$

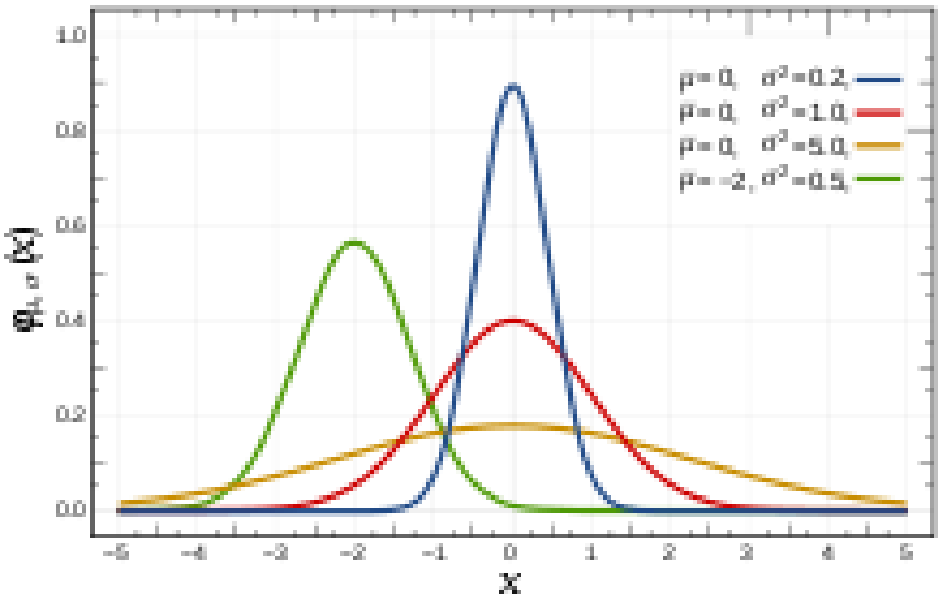


CDF: $\quad 1 - e^{-\lambda x} \quad \forall \lambda \geq 0, x \in [0, \infty)$

Continuous RV - Common Distributions

- **Gaussian distribution:** $\mathcal{N}(\mu, \sigma^2)$

  PDF: $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $\quad x \in \mathbb{R}$



Standard Normal Distribution:

This is a special case when $\mu = 0, \sigma = 1$, and is described by this probability density function:

$$\varphi(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

Discrete Random Variables

Discrete random variables can take on either a finite or at most a countably infinite set of discrete values, i.e. its range is countable.

- Their probability distribution is given by a probability mass function which directly maps each value of the random variable to a probability value.
- **Probability Mass Function:** A function that gives the relative probability that a discrete random variable is exactly equal to some value.
  $P_X(x_k) = P(X = x_k)$ for k=1,2,3,...
- **Cumulative Distribution Function:** A function evaluating the probability that X will take a value less than or equal to x for a discrete random variable.
- **Example Distributions:**
  - Binomial Distribution
  - Poisson Distribution

Discrete RV - Common Distributions

- **Binomial Distribution:**
  PMF:

$$f(k, n, p) = \Pr(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

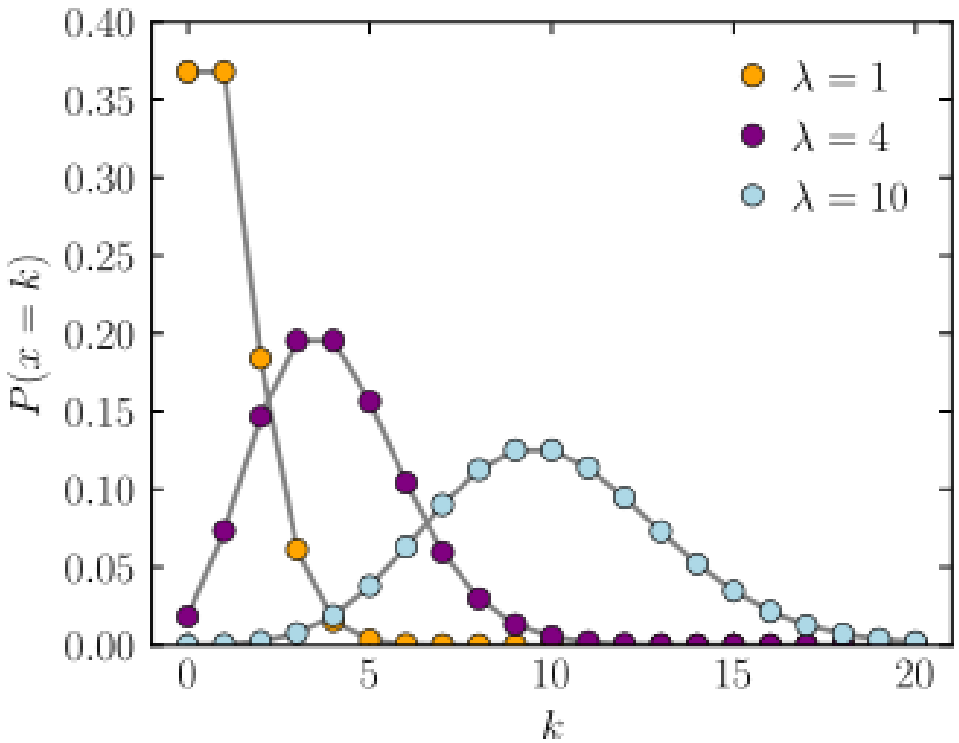
for k = 0, 1, 2, ..., n, where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient
This can be thought of as getting exactly k successes in n independent Bernoulli trials where probability of success in each trial = p $\in [0, 1]$. k successes can occur anywhere among the n trials, and there are $\binom{n}{k}$ different ways of distributing k successes in a sequence of n trials.

Discrete RV - Common Distributions

- **Poisson Distribution**
  PMF:

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

for k = 0, 1, 2, ..., n



Random Variables - Independence

Independence:
We say that random variables X and Y are independent if
$P(X = x, Y = y) = P(X = x)P(Y = y) \quad \forall$ x,y.
In general, if two random variables are independent, then we can write
$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for all sets A and B.
We can also write:
$P(Y = y \mid X = x) = P(Y = y) \quad \forall x, y.$

Joint Distributions - Discrete

Joint PMF: The joint probability mass function of two discrete random variables X and Y is defined as

$$P_{XY}(x, y) = P(X = x, Y = y)$$

where $\sum_{(x_i, y_j)} P_{XY}(x_i, y_j) = 1$

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 0$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{8}$
$X = 1$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$

Marginal PMF:

$$P_X(x) = P(X = x) = \sum_{y_j} P(X = x, Y = y_j) = \sum_{y_j} P_{XY}(x, y_j)$$

Question: Find $P(Y = 1 \mid X = 0)$ from the above table.

Joint Distributions - Continuous

Joint CDF: The joint cumulative distribution function of two random variables X and Y is defined as

$$F_{XY}(x, y) = P(X \leq x, Y \leq y).$$

Joint PDF: Two random variables X and Y are jointly continuous if there exists a nonnegative PDF function $f_{XY} : R^2 \to R$, such that, for any set $A \in R^2$, we have

$$P((X, Y) \in A) = \int_A \int f_{XY}(x, y) dx dy$$

Marginal PDF: Given the Joint PDF, we can find marginal PDF of X and Y as follows:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy, \quad \forall x$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx, \quad \forall y$$

Covariance of 2 random variables:

$$Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Mean,Variance,Standard Deviation

- **Discrete case:** The mean or expectation of a discrete random variable, X, is its weighted average.

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot P(X = x)$$

The variance of a discrete random variable X is given by:

$$\sigma^2[X] = \sum_{x \in \mathcal{X}} (x - \mu_x)^2 \cdot P(X = x)$$

- **Continuous case:** The mean is given by:

$$\mathbb{E}[X] = \int_{x \in \mathcal{X}} x \cdot P(X = x)$$

The variance of a discrete random variable X is given by:

$$\sigma^2[X] = \int_{x \in \mathcal{X}} (x - \mu_x)^2 \cdot P(X = x)$$

The standard deviation σ_x is the square root of the variance.

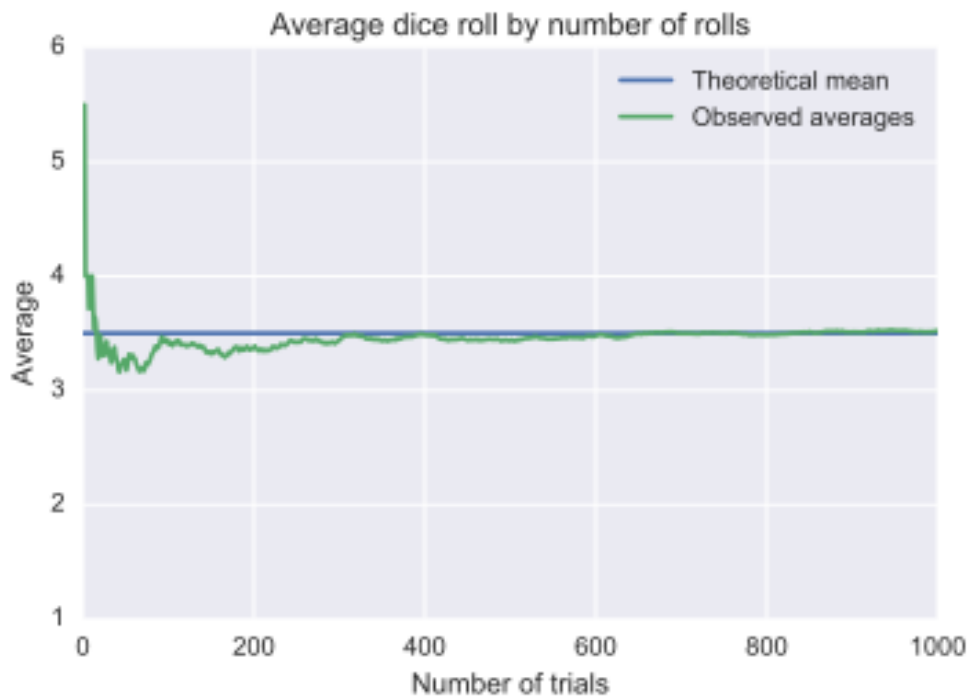
Common Distributions and their parameters

Distributions and Parameters		
Distribution	Mean	Variance
Bernoulli(p)	p	p(1-p)
Binomial(n,p)	np	np(1-p)
Poisson(λ)	λ	λ
Uniform(a,b)	$\frac{(a+b)}{2}$	$\frac{(b-a)^2}{12}$
Exponential(λ)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gaussian(μ, σ^2)	μ	σ^2

Law of Large Numbers

The law of large numbers states that the observed random mean from an increasingly large number of observations of a random variable will always approach the distribution mean. That is, as the number of observations increases, the mean of these observations will become closer and closer to the true mean of the random variable.

$$\overline{X}_n \to \mu \quad \text{as } n \to \infty$$



Functions of Random Variables

Suppose X is a continuous random variable and Y = g(X) is a function of X, then Y itself is a random variable. Thus, we should be able to find the CDF and PDF of Y.

Example: Let X be a Uniform(0,1) random variable, and let $Y = e^X$. We already know the CDF and PDF of X. In particular, CDF[X] $= \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \in [0,1] \\ 1 & \text{for } x > 1 \end{cases}$

Range of Y?

Since e^x is an increasing function of x and $R_X = [0,1]$, we conclude that $R_Y = [1, e]$. So we immediately know that

$F_Y(y) = P(Y \leq y) = 0$, for y< 1,

$F_Y(y) = P(Y \leq y) = 1$, for y$\geq e$

for $y \in [1, e]$, we can write

$F_Y(y) = P(Y \leq y)$

$=P(e^X \leq y)$

$=P(X \leq lny)$

$=F_X(lny) = lny$

Functions of Random Variables-PDF

Suppose that X is a continuous random variable and $g : R \to R$ is a strictly monotonic differentiable function. Let Y=g(X). Then the PDF of Y is given by:

$$f_Y(y) = \begin{cases} \frac{f_X(x_1)}{|g'(x_1)|} = f_X(x_1) \cdot \mid \frac{dx_1}{dy} \mid \text{where } g(x_1) = y \\ 0 \quad \text{if g(x)=y does not have a solution} \end{cases}$$

Question: Let X be a continuous random variable with PDF:

$$f_X(x) = \begin{cases} 4x^3 & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

Let $Y = \frac{1}{X}$. Find $f_Y(y)$

Mean and Variance of functions of Random Variables

- **Linear Combinations:** Suppose Y = aX+b where X is a random variable. Then the mean and variance are affected as follows:

$$\mathbb{E}[Y] = a\mathbb{E}[X] + b$$
$$\sigma^2[Y] = a^2\sigma^2[X]$$

- **Sum of two random variables:** Suppose Z = X+Y where X and Y are random variables. The mean and variance of Z are:

$$\mathbb{E}[Z] = \mathbb{E}[X] + \mathbb{E}[Y]$$
$$\sigma^2[Z] = \sigma^2[X] + \sigma^2[Y]$$

Gaussian Random Variables

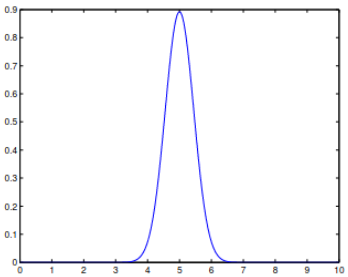
Recall that the density function of a univariate normal (or Gaussian) distribution is given by:

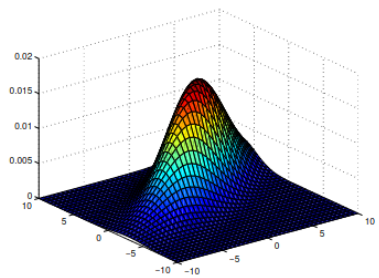
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad x \in \mathbb{R}$$

In the case of the multivariate Gaussian density, the argument of the exponential function, $-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$ is a quadratic form in the random vector x = vstack$(x_1, x_2 ...)$ and Σ is positive definite matrix called Covariance Matrix given by:

$$\mu = \frac{x_1 + x_2 + ...x_k}{k}$$

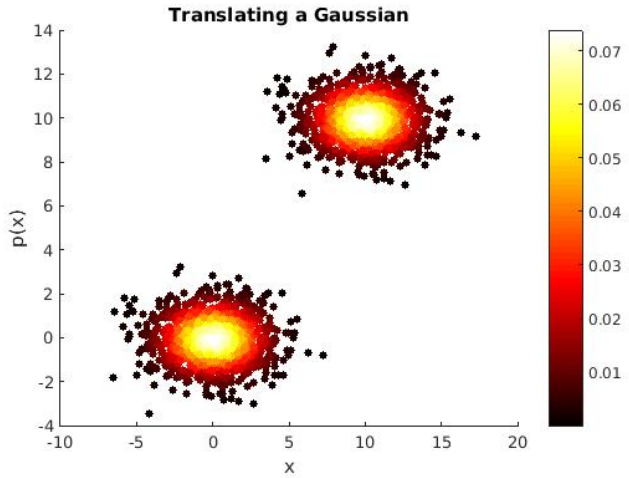
$$\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^T] = \mathbb{E}[XX^T] - \mu\mu^T$$



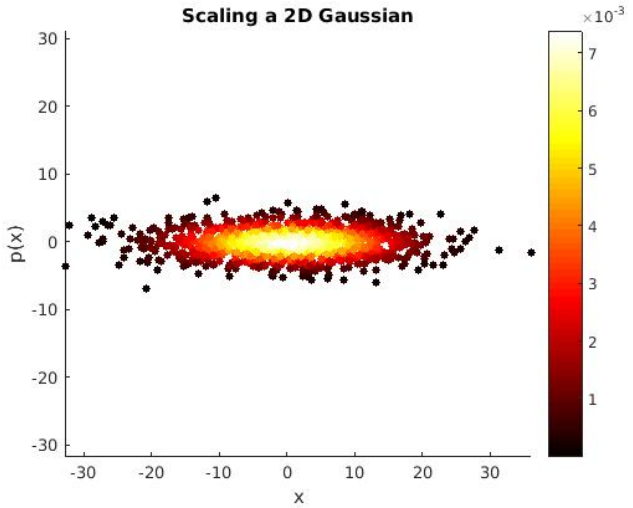


2D Gaussians-Transforms

- Translation - only change in mean:

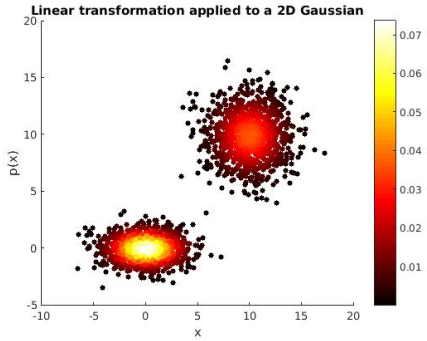


- Scaling:



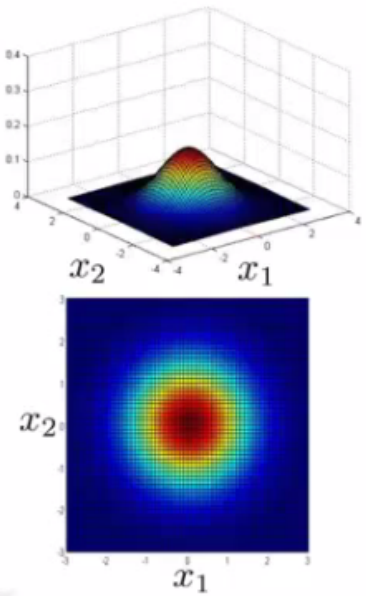
2D Gaussians-Transforms

- Linear transforms:

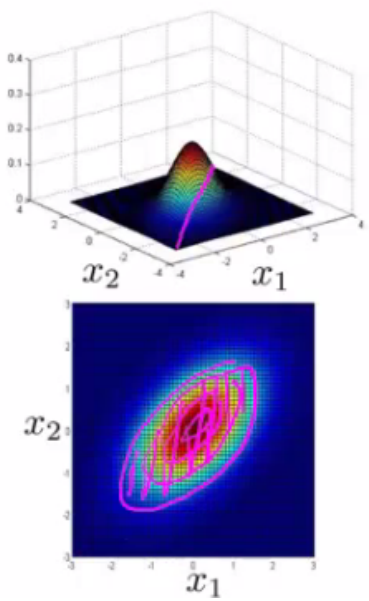


- Impact of covariance:

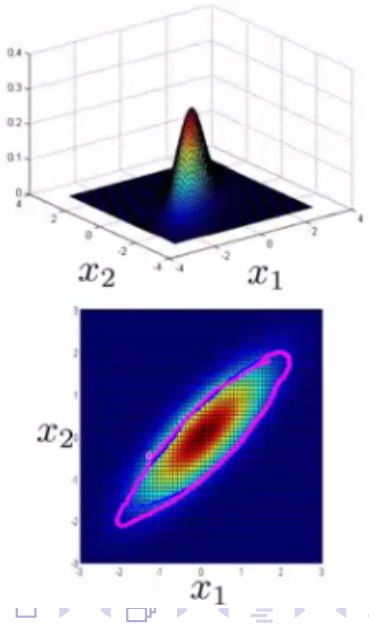
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



Maximum Likelihood Estimate

When to use: ML rule is used for parameter estimation, which involves maximizing the probability of observing the data from the joint probability distribution given a specific probability distribution and its parameters θ.
Mathematically:
Let us denote a function $f(x \mid \theta)$ as the "likelihood" function, basically the probability of x when the underlying population parameter is θ.
Log-likelihood: Consider $\log f(x \mid \theta)$
The estimate:

$$\hat{\theta}_{\mathrm{MLE}}(x) = \arg\max_{\theta} f(x \mid \theta)$$

is called the maximum likelihood estimate of θ.
In terms of log likelihood this will become,

$$\hat{\theta}_{\mathrm{MLE}}(x) = \arg\max_{\theta} \log f(x \mid \theta)$$

Task: Estimate the parameters of a Gaussian distribution using MLE.

Parameters of Gaussian Distribution

Log likelihood $\mathcal{LL} = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2$. We need to solve:

$\underset{\mu}{\text{argmax}}\mathcal{LL}(X|\mu, \sigma^2) := \frac{\partial \mathcal{LL}}{\partial \mu} = 0$

$$\frac{\partial \mathcal{LL}}{\partial \mu} = \frac{\partial}{\partial \mu}\Big(-\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big)$$

$$= \frac{\partial}{\partial \mu}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big) = \sum_{n=1}^{N}\frac{\partial}{\partial \mu}\Big(-\frac{1}{2\sigma^2}(x_n - \mu)^2\Big)$$

$$= -\frac{1}{2\sigma^2}\sum_{n=1}^{N} 2(x_n - \mu)\cdot -1$$

$$= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)$$

$$\text{We have } \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) = 0 \implies \mu = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Try: Derive variance using ML rule for Gaussian distributions.

MAP rule

Problem with ML estimate? Does not incorporate any prior knowledge about the experiments performed.
The method of Maximum A-Posteriori estimation utilizes prior distribution $g(\theta)$ as follows:
MAP estimate: $\hat{\theta}_{\mathrm{MAP}}(x)$

$$
\begin{aligned}
&= \arg\max_{\theta} \; f(\theta \mid x) \\
&= \arg\max_{\theta} \; f(\theta \mid x) \\
&= \arg\max_{\theta} \; \frac{f(x|\theta)\, g(\theta)}{\int_{\Theta} f(x \mid \theta)\, g(\theta)\, d\theta} \\
&= \arg\max_{\theta} \; f(x \mid \theta)\, g(\theta).
\end{aligned}
$$

The denominator of the posterior distribution (so-called marginal likelihood) doesn't depend on θ and plays no role in optimization.
Question: When is MAP equivalent to ML estimate?

References:

- https://www.probabilitycourse.com
- https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/ppt/22-MAP.pdf
- http://cs229.stanford.edu/section/gaussians.pdf
- https://web.stanford.edu/class/cs109/