Introduction to Probability

Ioannis Patras

(with acknowledgements to
Mark Plumbley, Mike Davies, Kevin Murphy and Sam Roweis)

Introduction

Probability Theory deals with phenomena that have some degree of randomness/uncertainty/chance.

Example

The tossing of a coin.

Loose concept of probability:

$$\text{Prob(Heads)} = \frac{\text{No. of Heads}}{\text{No. of Trials}} = 140/250 = 0.56$$

Clearly Coin tossing is not predictable, however, we can expect that given another 100 spins we will see *approximately*

$$100 \times \text{Prob(Heads)} = 56$$

heads

Axioms of probability

An *event* is represented by a set A of possible outcomes.

Example: Die Roll

Some possible events:

$\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$, (shorthand: '1', '2', '3', etc.)
'odd' $= \{1, 3, 5\}$,
'even' $= \{2, 4, 6\}$,
'less than 3' $= \{1, 2\}$

Axioms of probability

Let A and B be possible *events*
Let Ω be the sample space (the set of all possible outcomes).

The axioms of probability can then be given as:

1. $P(A) \geq 0$ (Events have nonnegative probability)
2. $P(\Omega) = 1$ (The certain event)
3. $P(A \cup B) = P(A) + P(B)$ if A and B are *mutually exclusive*

For '$P(A \cup B)$' read 'the probability that A occurs or B occurs'.
Mutually exclusive means that events A and B cannot both occur at once, i.e. $A \cap B = \emptyset$.

These axioms are fairly intuitive, and are sufficient to define probability theory.

Mathematical Framework for Probability

Sample Space

We define the sample space to be, Ω, the space of all possible outcomes from an experiment.

Examples of Sample Spaces

Coin toss

$$\Omega = \{H, T\}$$

Die roll

$$\Omega = \{1, 2, \ldots, 6\}$$

Spinning pointer

$$\Omega = \{\theta : 0 \leq \theta < 2\pi\}$$

note this last space is not discrete.

Probability measure

We can now define a measure of probability for an element of $\mathcal{A}$ (i.e. a subset of Ω)

P is just a function from $\mathcal{A}$ to the interval $[0, 1]$ that satisfies:

1. $P(\Omega) = 1$;
2. if $E_1 \cap E_2 = \emptyset$ then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

That is we have just reformulated the axioms of probability.

Construction of Probability Models

A complete probability model is defined by $\{\Omega, \mathcal{A}, P\}$.

Getting a probability model from a real system is more difficult. We will look at this later in the course.

Exercise

Having defined probability, we can draw some corollaries:

1. $P(\emptyset) = 0$
   proof:

2. $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$
   proof:

Conditional probability

The *conditional probability* of A given B is defined to be:

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}$$

Example:

- Probability of being Greek (A) given being blond (B)

Caution: There is no time order between the events.

Conditional probability

Corollaries

- B becomes the new *certain event*, $P(B|B) = 1$.
- If $A \cap B = \emptyset$ then $P(A|B) = 0$.
  i.e. if A and B are mutually exclusive events,
  and B has occurred, then A cannot have occurred.
- If $B \subset A$ then $P(A|B) = 1$
  i.e. if B is a subset of A,
  and B occurred, then A must have occurred.

Important fact

$P(\cdot\,|B)$ defines a *new* probability model on $(\Omega, \mathcal{A})$.

Proof

The quantity $P(A|B)$ satisfies all three probability axioms:

1. $P(A|B) \geq 0$
2. $P(\Omega|B) = 1$
3. if $A \cap C = \emptyset$, $P(A \cup C|B) = P(A|B) + P(C|B)$

Total Probability

Now let B be some other event. Then

$$\begin{aligned} P(B) &= \sum_{k=1}^{n} P(B \cap A_k) \\ &= \sum_{k=1}^{n} P(B|A_k)P(A_k) \end{aligned}$$

where $A_1, \ldots, A_n$ be mutually exclusive events that *cover* Ω, that is

1. $A_i \cap A_j = \emptyset, i \neq j$ (Mutually exclusive)
2. $\bigcup_i A_i = \Omega$ (Cover)

This is called *the Total Probability Rule*.

Useful when we know the conditionals but not the marginal.

Bayes Rule

From the definition of Conditional Probability we have

$$P(A|B)P(B) = P(B \cap A) = P(B|A)P(A)$$

Eliminating $P(B \cap A) = P(A \cap B)$ we get

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

This is *Bayes Rule*.

Bayesian Inference

Bayes Rule is used a lot for inference.
The various elements are given special terms:

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

$$\text{'Posterior Probability'} = \frac{\text{'Likelihood'} \times \text{'Prior Probability for A'}}{\text{'Prior Probability for B'}}$$

- $P(A)$ is *prior* probability of A before observation of B
- $P(A|B)$ is *posterior* probability of A after B observed

Often B will constitute some observed data,
and A is a class to assign, or decision to be made based on B.

Difference between $P(A|B)$ and $P(B|A)$ can be very dramatic.

Exercise

An international airport has a problem with terrorism
(0.1% of its users are terrorists).
Therefore it installs a face recognition system to help identify them.

Given a terrorist,
the system will correctly identify them 99% of the time.

Similarly, Given a non-terrorist,
the system will correctly identify them as such 99% of the time.

What is the probability that a person is a terrorist, given that the system has identified him as one?

Independence

A key concept in statistics is *Independence*. That is when one event has no influence on the probability of another.

Definition: Events A and B are independent if:

$$P(A|B) = P(A)$$

or $P(A \cap B) = P(A)P(B)$ (alternative definition)

If A_1, A_2, A_3 are all mutually independent then we have:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$$

Pair-wise independence is necessary but *not sufficient* for mutual independence.

Properties of independent events

If A and B are independent we have:

- A^c and B^c are mutually independent
- A^c and B or A and B^c are independent

If A, B and C are mutually independent then:

- A and $B \cap C$ are independent
- A and $B \cup C$ are independent

etc.

Example

Tossing two fair coins (elements of sample space are *ordered pairs*):

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$$

assigning 'classical' probabilities:

$$P(\{(H, H)\}) = \frac{1}{4}, \qquad P(\{(H, T)\}) = \frac{1}{4}, \qquad \text{etc.}$$

With these probabilities, coin 1 / coin 2 events are independent:

$$P(\text{coin 1} = \text{H} | \text{coin 2} = \text{H}) = \frac{P(\{(H, H)\})}{P(\text{coin 2} = \text{H})} = \frac{1}{2} = P(\text{coin 1} = \text{H})$$

(The same holds for all other combinations of H and T.)
Hence the *experiments* are independent.

In general relative frequency view of repeated experiments only holds if the experiments are independent.

Random Variables

A random Variable (RV) X represent outcomes or states of the world. Instantiations usually in lower case: x.

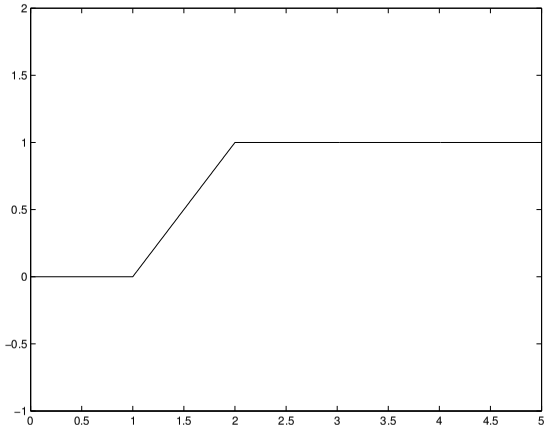
For a discrete Random Variable X, $P(X = x)$ is the probability that the RV X takes the value x.

For a real Random Variable X, $P(X \leq x)$ is the probability that the RV X takes a value smaller than x. This is the Cummulative Distribution function (cdf)

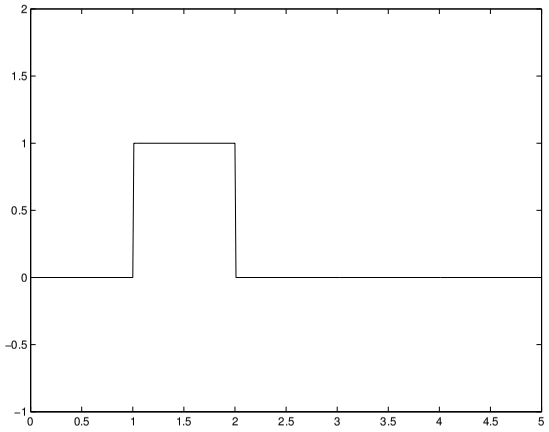
Probability mass (density) function (pdf) $p(x) \geq 0$
Discrete case: $p(x) = p(X = x)$
Continuous case: $p(x) = \frac{d}{dx} P(X \leq x) \approx \frac{P(x \leq X < x + \Delta x)}{\Delta x}$

Example pdf: The Uniform Distribution

$P(X \leq x)$ (cdf)



$p(x)$ (pdf)



Interpretation of $p(x)$

Probability of 'small interval' event $\{x \leq X < x + \Delta x\}$:

$$p(x) \approx \frac{P(x \leq X < x + \Delta x)}{\Delta x}$$

This is exact in the limit $\Delta x \to 0$.

Properties of pdf

1. $f(x) \geq 0$ (from definition of $F(x)$)
2. $\int_{-\infty}^{\infty} f(x)dx = 1$ (probability of the whole space)

Remember: "Densities are for integrating."

Ensemble Averages and Expectations

Often we are only interested in a partial characterization of the RV.

Example

When gambling the expected rate of return is our main statistical interest.

(Here we ignore important practical considerations, like available cash reserve.)

Hence we define 'average' properties.

There are two complementary definitions:

- in terms of 'data' (sample properties)
- in terms of the model (expectations)

Sample Mean and Expected Value

Sample Mean: Given N samples x_i drawn (independently) from an experiment, we define the *mean* to be:

$$\hat{m}_N = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

This measures the 'average' of the samples.

Expected Value: Given a RV X with pdf $p(x)$ we define:

$$E\{X\} = \int_{-\infty}^{\infty} x \cdot p(x) dx \qquad \text{if } X \text{ is continuous}$$

$$E\{X\} = \sum_{x} x \cdot P(X = x) \qquad \text{if } X \text{ is discrete (no pdf)}$$

We sometimes write $m_X = E\{X\}$ or $\mu_X = E\{X\}$.

Intuitively they are related (e.g. relative frequency) but how?

Law of Large Numbers

Roughly (and in many different ways!) as we have a very large number of data points ($N \to \infty$), we find that

$$\hat{m}_N \to E\{x\}$$

This general idea is known as the 'Law of Large Numbers'. (More on this later...)

Example

Roll many (large N) independent fair dice,
each die roll giving a number between 1 and 6.
(Remember: 'fair' means 'equal probability of each outcome'.)

Assuming equal probabilities, we would expect:

$$\hat{m}_N = \frac{1}{N}\sum_i x_i \approx E\{x\} = \sum_{k=1}^{6} \frac{k}{6} = 3.5$$

Functions of an RV - Means and expectations

Suppose that $g(x)$ is a function,
mapping real values $x \in \mathbb{R}$ to real values $y = g(x) \in \mathbb{R}$.
We can define means and expectations of functions of a RV X:

$$E\{Y\} = E\{g(X)\} = \int_{-\infty}^{\infty} g(x)p(x)dx$$

(note Y is a random variable itself and hence we are again considering the mean of an RV)

Example: $p(x)$ the pdf of the age of the population, $g(x)$ the medical costs associated with age x

Again there is an analogous definition for *sample* mean of Y:

$$\hat{m}_N = \frac{y_1 + y_2 + \ldots + y_N}{N}$$

Variance

Another important statistic is one that measures the 'spread' of the outcomes. One such statistic is the *variance*:

$$\text{Var}\{X\} = \sigma_X^2 = E\{(X - E\{X\})^2\}$$

The quantity $\sigma_X = \sqrt{\text{Var}\{X\}}$ is called the *standard deviation*.

$$\begin{aligned}
\text{Var}\{X\} &= \int_{-\infty}^{\infty} (x - E\{X\})^2 p_X(x) dx \\
&= \int_{-\infty}^{\infty} (x^2 - 2xE\{X\} + E\{X\}^2) p_X(x) dx \\
&= \int_{-\infty}^{\infty} x^2 p_X(x) dx - 2E\{X\} \int_{-\infty}^{\infty} x p_X(x) dx + E\{X\}^2 \int_{-\infty}^{\infty} p_X(x) dx \\
&= E\{X^2\} - 2E\{X\}^2 + E\{X\}^2 \\
&= E\{X^2\} - E\{X\}^2
\end{aligned}$$

Higher Order Statistics

Experiments often only need 2*nd order* statistics, i.e. only mean $m_X = E\{X\}$ and variance $\sigma_X^2 = E\{X^2\} - E\{X\}^2$.

There are also *Higher Order Statistics* (HOS), often based upon polynomial expectations, $(E\{X^3\}, E\{X^4\}, \ldots)$, e.g.:

$$\text{Skewness} = E\left\{\left(\frac{X - m_X}{\sigma_X}\right)^3\right\}$$

(this measures the lopsidedness of a distribution) and

$$\text{Kurtosis} = E\left\{\left(\frac{X - m_X}{\sigma_X}\right)^4\right\} - 3$$

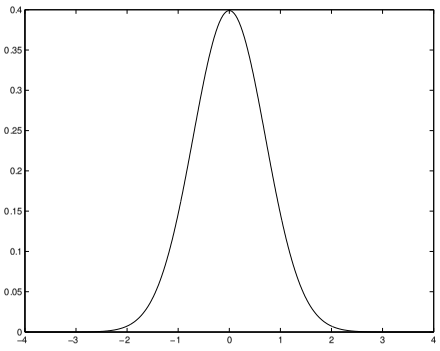
(this measures how much of the probability mass lies within the tails).
Positive kurtosis is also termed 'heavy-tailed'.
('Heavy' with respect to a Gaussian: hence the -3).

The Gaussian Distribution

aka the 'normal distribution', with probability density function (pdf):

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-(x-\mu)^2/2\sigma^2},$$

written in terms of the mean, μ, and variance, σ^2.



Multiple Random Variables

Extend the idea of RVs to two or more dimensions,
e.g. RVs X (age) and Y (wealth)

RVs X and Y have respective distributions $p_X(x)$ and $p_Y(y)$ but this does not tell us about the relation between X and Y.

Joint Density

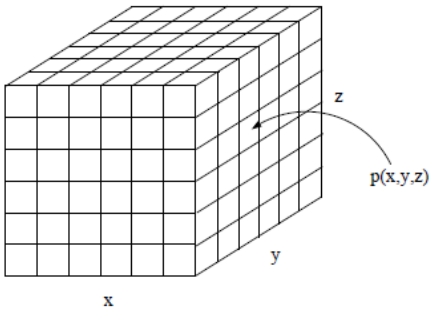
$$p_{XY}(x,y) \approx \frac{P(\{x < X \leq x + \Delta x\}, \{y < Y \leq y + \Delta y\})}{\Delta x \Delta y}$$

The probability of any joint area $D \subset \mathbb{R}^2$ (e.g. $P(\{x_1 < X \leq x_2\}, \{y_1 < Y \leq y_2\})$ can be evaluated from the 2-dimensional integral:

$$P((X,Y) \in D) = \int_D p_{XY}(x,y)dx\,dy$$

Joint Density

$p(x, y) = \text{prob}(X = x \text{ and } Y = y)$



Joint probability of three random variables x, y, and z .

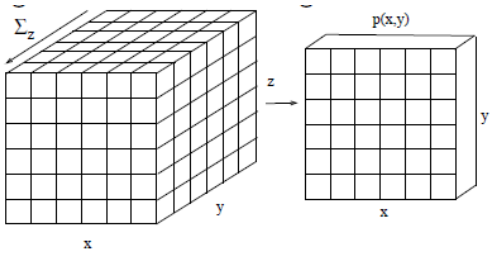
Joint and Marginal densities

Given a joint distribution $p_{XY}(x, y)$ we define the *marginal density*:

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, y)dy \tag{1}$$

So: Area of slice through $p_{XY}(x, y)$ at $X = x$ is equal to $p_X(x)$.

Relationship between Joint and Marginal densities

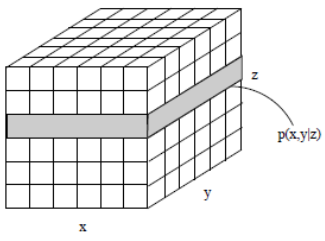


Marginal probability of x and y, when z is marginalised.

Conditional Densities

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{f_X(x)}$$

$$\boxed{p(x|y) = p(x,y)/p(y)}$$

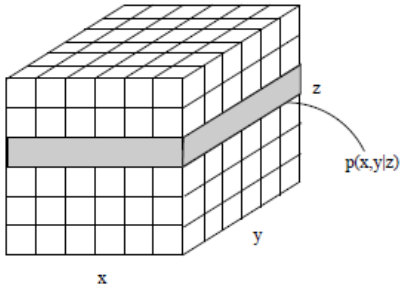


Conditional probability of x and y, given z are (normalised) slices in the $p(x, y, z)$ cube.

(NB. Remember that $\int \int p(x, y|z)dxdy = 1$)

Conditional densities

$$\boxed{p(x|y) = p(x,y)/p(y)}$$



Conditional probability of x and y, given z are (normalised) slices in the $p(x, y, z)$ cube.

(NB. Remember that $\int \int p(x, y|z)dxdy = 1$)

Total Probability and Bayes Rule

As before we now have Total Probability Rule:

$$\begin{aligned} p_Y(y) &= \int_{-\infty}^{\infty} p(x,y)dx \\ &= \int_{-\infty}^{\infty} p(y|x)p(x)dx \end{aligned}$$

We also have a density version of Bayes rule:

$$p(y|x)p(x) = p(x|y)p(y) = p(x,y)$$

or

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Independent RVs

Recall: Events A and B are independent if $P(A, B) = P(A)P(B)$.

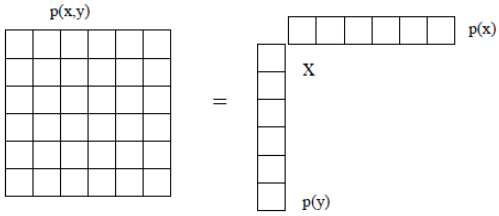
We say the RVs X and Y are *independent RVs* if:

$$f_{XY}(x, y) = f_X(x)\, f_Y(y) \tag{2}$$

So joint density has a factorial (or product) form.

Independent RVs

$$p(x, y) = p(x)p(y)$$



So joint distribution & density has a factorial (or product) form.

Example: Sum of two independent RVs

Let $Z = X + Y$ and X and Y be independent.

Given $p_X(x)$ and $p_Y(y)$ how do we calculate $p_Z(z)$?

We note:

$$p_Z(z) = \int_X p_{XZ}(x, z)\, dx = \int_X p_{Z|X}(z|x) p_X(x)\, dx$$

For fixed $X = x$, we get $Z = x + Y$ or $Y = Z - x$.
So Z is simply a translation of Y:

$$p_{Z|X}(z|x) = p_Y(z - x)$$

which gives:

$$p_Z(z) = \int_X p_Y(z - x)\, p_X(x)\, dx$$

Example: Expectation of sum of two independent RVs

To prove (which we used in the weak law of large numbers):

$$E\{X + Y\} = E\{X\} + E\{Y\}$$

Proof:

$$E\{X + Y\} = \int_{\{X,Y\}} (x + y) p_{XY}(x, y) dx\, dy$$

Given independence:

$$\begin{aligned} E\{X + Y\} &= \int_X \int_Y (x + y) p_X(x)\, p_Y(y) dx\, dy \\ &= \int_X \int_Y x p_X(x) p_Y(y) dx\, dy + \int_X \int_Y y p_X(x)\, p_Y(y) dx\, dy \\ &= \int_X x p_X(x) dx + \int_Y y p_Y(y) dy \\ &= E\{X\} + E\{Y\} \qquad \text{QED} \end{aligned}$$

Exercise

Show that the same holds for the variance of the sum of independent RVs, i.e.

$$\mathrm{Var}\{X + Y\} = \mathrm{Var}\{X\} + \mathrm{Var}\{Y\}$$

Covariance and correlations

The *covariance* of two RVs X and Y is:

$$\mathrm{Cov}(X, Y) \equiv \sigma_{XY} = E\{(X - m_X)(Y - m_Y)\} = E\{XY\} - m_X m_Y$$

(Pearson) correlation coefficient (sometimes just *correlation*) is:

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

which measures how linearly related the two variables are.

Fact: $-1 \leq \rho_{XY} \leq 1$. (Can you show this?)
If $\rho_{XY} = 0$, the RVs are said to be *uncorrelated*.
If $\rho_{XY} = \pm 1$ then Y is simply a linear rescaling of X.

NB: In e.g. signal processing, can also talk about *correlation*, $\mathrm{Cor}(X, Y) = E\{XY\}$, so that $\mathrm{Cov}(X, Y) = \mathrm{Cor}(X, Y) - m_X m_Y$.

Beware possible confusion in use of word "correlation"!

Independent vs Uncorrelated RVs

If X and Y are independent scalar RVs, we have:

$$p_{XY}(x, y) = p_X(x)\, p_Y(y)$$

Hence $E\{XY\} = E\{X\}E\{Y\} = m_X m_Y$. (Exercise: Show this).
Hence $\mathrm{Cov}(X, Y) = E\{XY\} - m_X m_Y = 0$ and so also $\rho_{XY} = 0$

Therefore: Independent $\Rightarrow$ Uncorrelated.

However: Uncorrelated $\not\Rightarrow$ Independent

Covariance $\mathrm{Cov}(X, Y)$ only measures 2nd order statistics.
Independence is *much* stronger: all nonlinear statistics must be unrelated.

Nevertheless, "uncorrelated" is much easier to calculate (and enforce!) than "independent", so is of great practical interest.

N-dimensional Random Variables

Distributions, densities, etc. generalize to N dimensions.

Let $\vec{X} = [X_1, \ldots, X_N]$ then:

$$p_{\vec{X}}(\vec{x}) = p([x_1, \ldots, x_k] | [x_{k+1}, \ldots, x_N]) f([x_{k+1}, \ldots, x_N])$$

etc.

Multi-Dimensional Expectations

Just as we defined $E\{x\}$ for scalar RVs, we have the same for vector RVs:

$$\vec{m}_x = E\{\vec{X}\} = \int_{x_1} \int_{x_2} \cdots \int_{x_N} \vec{x}\, p_{\vec{X}}(\vec{x}) dx_1 \ldots dx_N$$

where $\vec{\cdot}$ represents a vector value.

Multi-Dimensional Variance: Covariance Matrix

The equivalent of multi-dimensional variance is more complex.
There are several ways of generalising $E(X^2)$ to N-dimensions.

To capture the full generality we define the *covariance matrix*:

$$\text{Cov}(\vec{X}) = E\{(\vec{X} - \vec{m}_{\vec{X}})(\vec{X} - \vec{m}_{\vec{X}})^T\}$$

Note that $\text{Cov}(\vec{X})$ is an $N \times N$ matrix.

For two N-d RVs $\vec{X}$ and $\vec{Y}$ we can define the *cross-covariance*:

$$\text{Cov}(\vec{X}, \vec{Y}) = E\{(\vec{X} - \vec{m}_{\vec{X}})(\vec{Y} - \vec{m}_{\vec{Y}})^T\}$$

Exercise

Let X and Y be uncorrelated scalar RVs, and let $\vec{Z} = [X, Y]$.

Calculate the form of the covariance matrix $\text{Cov}(\vec{Z})$.

N-dimensional Gaussian RVs

The Gaussian distribution has a natural N-dimensional form:

$$p_X(x) = \frac{det(\text{Cov}(\vec{X}))^{-1/2}}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}(x - \vec{m}_x)^T \text{Cov}(\vec{X})^{-1}(x - \vec{m}_x)\right)$$

As for the scalar case, it is completely specified by
(1) the mean $\vec{m}_x$, and (2) the covariance $\text{Cov}(\vec{X})$.

The N-d Gaussian also has the following interesting properties:

1. Let $Z = aX + bY$.
   If X and Y are jointly Gaussian,
   then Z is also Gaussian with $\vec{m}_z = a\vec{m}_x + b\vec{m}_y$.
2. If X and Y are jointly Gaussian and uncorrelated,
   then they are also *Independent*.
   So correlation measures dependence for Gaussian RVs.