3F3 Statistical Signal Processing

Sumeetpal S. Singh

Department of Engineering
University of Cambridge
E-mail: sss40@cam.ac.uk

Outline

Probability and Random variables:

- Sample space, events, probability measure, axioms.
- Conditional probability, probability chain rule, independence, Bayes rule.
- Random variables (discrete and continuous), probability mass function (pmf), probability density function (pdf), cumulative distribution function, transformation of random variables.
- Bivariate, conditional pmf, conditional pdf, expectation, conditional expectation.
- Multivariate: marginals, Gaussian (properties), characteristic function, change of variables (Jacobian)

Finally

✍ = Homework

Acknowledgement: These note were compiled by drawing on material from

[1] Wasserman, L. (2004), *All of statistics: a concise course in statistical inference*. Springer.

[2] Gubner, J.A. (2006), *Probability and random processes for electrical and computer engineers.* Cambridge University Press.

[3] Norris, J.R. (1997), *Markov chains*. Cambridge University Press.

Probability Space

Sample space, Events, Probability measure, Axioms.

The term *random experiment* is used to describe any situation which has a set of possible outcomes, each of which occurs with a particular probability.

- For example, the variation in temperature after subtracting seasonal effects, the outcome of a game of roulette, motor insurance claims, etc.
- How do we mathematically describe a random experiment?

To *mathematically* describe a random experiment we must specify:

1. The **sample space** Ω, which is the set of all possible outcomes of the random experiment. We call any subset of $A \subseteq \Omega$ an *event*.
2. A mapping/function P from events to a number in the interval $[0, 1]$. That is we must specify $\{P(A), A \subset \Omega\}$. We call P the **probability**.

We then call (Ω, P) the **probability space**.

Example

If we toss a coin twice then $\Omega = \{HH, HT, TH, TT\}$. Note Ω is a finite set and we can list it easily.

Example

On a particular day of the year (say in summer) the temperature is a random perturbation of the expected seasonal value. The sample space Ω in this case is the real line:
$\Omega = (-\infty, \infty)$.
We can answer all useful questions such as "what is the probability that the temperature "lies between -1 and +1."Although the temperature is clearly bounded, there is no harm in taking the sample space to be the whole real line as opposed to a subset of it.

Example

If we toss a coin infinite times, an elementary outcome is described by the sequence $\omega = (o_1, o_2, \ldots)$ where each $o_i \in \{H, T\}$ and the set of all possible outcomes is

$$\Omega = \{\omega = (o_1, o_2, \ldots) : o_i \in \{H, T\}\} .$$

The event E that the first head occurs on the third toss is

$$E = \{\omega = (T, T, H, o_4, o_5, \ldots) : o_i \in \{H, T\} \text{ for } i > 3\} .$$

The probability of this event is $P(E) = (1/2)^3$.

Remark

(A flexible framework.) We have complete freedom in defining Ω to desribe the real-world random experiment. We only considered two specific examples: when Ω is a general countable set or when Ω the real-line. We will not be introducing anymore.

Definition

(Axioms of probability.) A probability P assigns each *event* E, $E \subset \Omega$, a number in $[0, 1]$ and P must satisfy the following properties:

- $P(\Omega) = 1$.
- For events A,B such that $A \cap B = \emptyset$ (i.e. disjoint) then $P(A \cup B) = P(A) + P(B)$.
- If $A_1, A_2, \ldots$ are disjoint then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Example

Show

(i) if event $A \subset B$ then $P(A) \leq P(B)$.

(ii) $P(A^c) = 1 - P(A)$. ✍

Some set notation:

$$\boxed{B \backslash A = B \cap A^c}$$

Verify (i): Write $B = (B \cap A^c) \cup A = (B \backslash A) \cup A$. Then

$$P(B) = P(B \backslash A) + P(A) \geq P(A).$$

Verify (ii): Since $\Omega = A \cup A^c$,

$$P(\Omega) = P(A) + P(A^c) = 1.$$

Example (Total probability. ✍)

Let $A_1, A_2, \ldots, A_n$ be n mutually disjoint events and whose union is Ω. For an event B show that $P(B) = \sum_{i=1}^{n} P(BA_i)$.

Decompose B as the union of n disjoint sets

$$B = BA_1 \cup BA_2 \cup \cdots \cup BA_n$$

where (set notation)

$$\boxed{BA_i = B \cap A_i.}$$

Result follows from the additivity of P for disjoint sets.

The notion of a probability space formalizes our description of a random experiment. We always have to specify (Ω, P).
The space Ω will be apparent from the problem being studied, e.g. in the temperature example Ω was the real line.

- When Ω is a *discrete* set it is easy to construct a probability P as the next example shows.
- (Set notation.) Define the *indicator* function for a set or event E,

$$\mathbb{I}_E(t) = \begin{cases} 1 & \text{if } t \in E, \\ 0 & \text{if } t \notin E. \end{cases}$$

Example (Defining P.)

Ω is a finite discrete set, i.e. $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. Let $p_1, p_2, \ldots, p_n$ be non-negative numbers that add to 1. For any event A, set

$$P(A) = \sum_{i=1}^{n} \mathbb{I}_A(\omega_i) p_i.$$

Then P satisfies the axioms.
Let $p_i = 1/n$. Then

$$P(\{\omega_i\}) = p_i = 1/n,$$

i.e. each outcome is equally likely. This is the *uniform probability distribution*.

Example (Defining P. ✍)

Ω is an *infinite* discrete set, i.e. $\Omega = \{\omega_1, \omega_2, \ldots\}$. Given any non-negative sequence of numbers $p_1, p_2, \ldots$ that add to 1, let

$$P(A) = \sum_{i=1}^{\infty} \mathbb{I}_A(\omega_i) p_i.$$

Then P is a valid probability. (All axioms satisfied; inherited from the properties of the sum.)
This construction gives $P(\{\omega_j\}) = p_j$. If

$$p_j = e^{-\lambda} \frac{\lambda^{j-1}}{(j-1)!}$$

then P is the Poisson distribution.

When studying random experiments with Ω being the real line, P will be specified through a probability density function (pdf) $f(t)$.

A pdf is a non-negative function,

$$f(t) \geq 0 \quad \text{s.t} \quad \int_{-\infty}^{\infty} f(t)dt = 1,$$

i.e. total area is 1. For an event $E = [a, b]$ define

$$P(E) = \int_{a}^{b} f(t)\, dt.$$

Note that $P(E)$ is the area under f between a and b.

- This assignment for P gives a valid probability. That is the axioms of probability are satisfied.
- This definition implies

$$P(\{c\}) = 0 \quad \text{since} \quad \int_c^c f(t)\, dt = 0.$$

$$\text{Thus} \quad P([a, b]) = P((a, b]) = P((a, b)).$$

- For a more general event E, i.e. not just an interval $[a, b]$, we can calculate the probability using the *indicator* function $\mathbb{I}_E(t)$,

$$\boxed{P(E) = \int_{-\infty}^{\infty} \mathbb{I}_E(t) f(t)\, dt.}$$

Conditional Probability

Conditional probability, Probability chain rule, Independence, Bayes rule.

Definition

(Conditional probability.) Consider events A and B. The conditional probability of event A occuring given that event B has occured is defined to be

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0.$$

- This definition only applies when $P(B) > 0$ otherwise B *could not* have occured.
- Think of $P(A|B)$ as the fraction of times A occurs among those in which B occurs.
- Set notation: AB is shorthand for $A \cap B$.

Example (Showing $P(\cdot|G)$ is a probability. ✍)

For any fixed G such that $P(G) > 0$, show that $P(\cdot|G)$ is a probability.

Verification: firstly $P(\Omega|G) = P(\Omega \cap G)/P(G) = 1$. Secondly, for disjoint events A and B,

$$\begin{aligned} P(A \cup B|G) &= P(AG \cup BG)/P(G) \\ &= (P(AG) + P(BG))\,/P(G) \\ &= P(A|G) + P(B|G). \end{aligned}$$

(Probability chain rule.) For events $A_1, \ldots, A_n$, note that

$$\begin{aligned} P(A_1 \ldots A_{n-1}A_n) &= P(A_n|A_1 \ldots A_{n-1})P(A_1 \ldots A_{n-1}) \\ P(A_1 \ldots A_{n-1}) &= P(A_{n-1}|A_1 \ldots A_{n-2})P(A_1 \ldots A_{n-2}) \\ &\vdots \\ P(A_1A_2) &= P(A_2|A_1)P(A_1). \end{aligned}$$

$P(A_1 \ldots A_{n-1}A_n)$ can be written as

$$P(A_1)\left(\prod_{i=2}^{n} P(A_i|A_1 \ldots A_{i-1})\right).$$

The chain rule is often used in the study of jointly distributed random variables, e.g. a stochastic process.

Definition

(Independence.) Two events A and B are independent if

$$P(AB) = P(A \cap B) = P(A)P(B).$$

Independence will often arise in two ways. In the first case it is assumed, for example the random experiment corresponding to two independent tosses of a fair coin. In the second case we may be asked to verify two events are independent as in this example.

Example

✍

A fair die is thrown. Let $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$. Show A and B are independent, i.e. $P(AB) = P(A)P(B)$.

If A and B are independent then $P(A|B) = P(A)$. With or without knowledge that B has occured, the probability of event A occuring is the same, or put another way, we are none the wiser. This is another interpretation of the *mathematical definition* of independence.

Definition

(Bayes' Theorem) Its the relationship between $P(A|B)$ and $P(B|A)$ obtained via the definition of the conditional probability:

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

What is important here is the interpretation, which is explained using the following simple example.

Example

An incoming email is either spam or not.

- Let B be the event the email contains the word "free." From experience (or training data),

$$P(B|\text{spam}) = 0.8 \text{ and } P(B|\text{not spam}) = 0.1$$

  and spam emails are 25% of all my emails.

- I just received an email and it contains the words free. The probability the received email is spam is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.8 \times 0.25}{0.8 \times 0.25 + 0.1 \times 0.75} = 0.727$$

  which is significantly more than 25% or $P(A)$.

This is an example of an *expert* system.

Remarks:

- $P(B|\text{spam})$ and $P(B|\text{not spam})$ is not specific to an individual, e.g. think of these numbers being provided by the university mail server.
- $P(A)$ though is individual specific; depends on your internet browsing behaviour. So this expert system could be catered to individuals.
- In the language of statistical inference, $P(A)$ is known as the prior or *a priori* probability since it reflects prior knowledge about event A.
- $P(A|B)$ is known as the posterior or *a posteriori* probability since it gives the probability of A after having observed that event B occured. $P(B)$ is the (unconditional) probability of event B.

Named after Rev. Thomas Bayes, an 18th century British mathematician, who studied (statistical) inference.

Random Variables

Random variables (discrete and continuous), Probability mass function (pmf), Probability density function (pdf), Cumulative distribution function, Transformation of random variables.

Definition

(Random variable.) Given a probability space (Ω, P), a *random variable* is a function $X(\omega)$ which maps each element ω of the sample space Ω onto a point on the real line.

Example (Flipping a coin twice.)

The sample space is $\Omega = \{(T,T),(T,H),(H,T),(H,H)\}$.
Let $X(\omega)$ be the number of heads.

ω	$P(\{\omega\})$	$X(\omega)$
TT	$\frac{1}{4}$	0
TH	$\frac{1}{4}$	1
HT	$\frac{1}{4}$	1
HH	$\frac{1}{4}$	2

x	$\Pr(X = x)$
0	$\frac{1}{4}$
1	$\frac{1}{2}$
2	$\frac{1}{4}$

In the second table, x denotes a possible value of the rv X.

The second table does not mention the sample space.

> The range of X is listed along with the probability the random variable X takes those values. **This is the approach we will adopt when defining a rv.**

But keep in mind that

(i) there is a sample space lurking behind *every* definition of a rv.

(ii) The probability that $X = x$ is inherited from the defintion of (Ω, P) and the mapping $X(\omega)$.

For the table example, in fact for any set $A \subset (-\infty, \infty)$, we define

$$\Pr(X \in A) = P(\{\omega : X(\omega) \in A\}) \quad \text{(inherited from } P.)$$

- In the table example, the random variable's range was a finite set. Stating $\Pr(X \in A)$ for all $A \subset (-\infty, \infty)$ is unnessary when the rv is discrete as it is enough to state $\Pr(X = x)$ for the finite (or countable) set of values it can take.
- Random variables can also be mappings to *all* of the real line,

  $$X : \Omega \to (-\infty, \infty)$$

  and not just a finite subset of it. Such rvs are completely sepecified only when provided with $\Pr(X \in A)$ for all $A \subset (-\infty, \infty)$.

Definition

A random variable is called *discrete* if its range is a finite set, say $\{x_1, \ldots, x_i, \ldots, x_M\}$, or a countable set, say $\{x_1, x_2, \ldots\}$.

- A set E is countable if you can define a one-to-one mapping from E to the set of integers.
- This means the integers can exhaustively list the elements of E.
- Examples of countable sets are: all even numbers, all odd numbers, all rational numbers. The interval $[0, 1]$ is not countable.

To fully define a discrete random variable, we must define its *range* and its *probability mass function* (pmf.)

Definition

(Probability mass function.) For a discrete rv X with range $\{x_1, x_2 \ldots\}$, define the pmf of X to be the function $p_X : \{x_1, x_2 \ldots\} \rightarrow [0, 1]$ where

$$p_X(x_i) = \Pr(X = x_i).$$

Note that $\sum_{i=1}^{\infty} p_X(x_i) = 1$.
The pmf is a complete description: for any set A,

$$\Pr(X \in A) = \sum_{i=1}^{\infty} \mathbb{I}_A(x_i) p_X(x_i).$$

As an example of a pmf, let the range be $\{0, 1, \ldots\}$ and $p_X(k) = e^{-\lambda}\lambda^k/k!$. This is the Poisson random variable.

Continuous rvs are defined as having a *probability density function* (pdf.)

Definition

(Probability density function.) A random variable X is *continuous* if there exists a non-negative function $f_X(x) \geq 0$ such that $\int_{-\infty}^{\infty} f_X(x)dx = 1$ and for any set A

$$\Pr(X \in A) = \int_{-\infty}^{\infty} \mathbb{I}_A(x) f_X(x) dx.$$

For example, if $A = [a, b]$ then

$$\Pr(X \in A) = \Pr(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

Some important remarks about continuous random variables:

- We call the function f_X is the probability density function of X.
- *(Interpreting the pdf.)* The pdf answers questions of the form "what is the probability that X lies in subset $E = [a, b]$ of the real line?"
- Recall we showed that a pdf f_X assigns 0 probability to any particular point $x \in \mathbb{R}$. Thus $\Pr(X = x) = 0$ for all x. Also

$$\boxed{\Pr(X \in [a, b]) = \Pr(X \in (a, b]) = \Pr(X \in (a, b))}$$

- This means a continuous rv has no concentration of probability at particular points like a discrete rv does.

Definition

(Cumulative distribution function.) The cdf can describe both discrete or continuous random variables and is defined to be

$$F_X(x) = \Pr(X \leq x).$$

Note also that $\Pr(X > x) = 1 - F_X(x)$.
Where there is no ambiguity we will usually drop the subscript 'X' and refer to the cdf as $F(x)$.

The following properties follow directly from the axioms of probability:

1. $0 \leq F_X(x) \leq 1$.
2. $F_X(x)$ is non-decreasing as x increases.
3. $\Pr(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$.
4. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.
5. If X is a continuous r.v. then $F_X(x)$ is continuous.
6. If X is discrete then F_X is right-continuous: $F_X(x) = \lim_{t \downarrow x} F(t)$ for all x.

Example (Showing properties 5 and 6.)

For a continuous rv with pdf,

$$F_X(x) = \int_{-\infty}^{x} f(t)dt$$

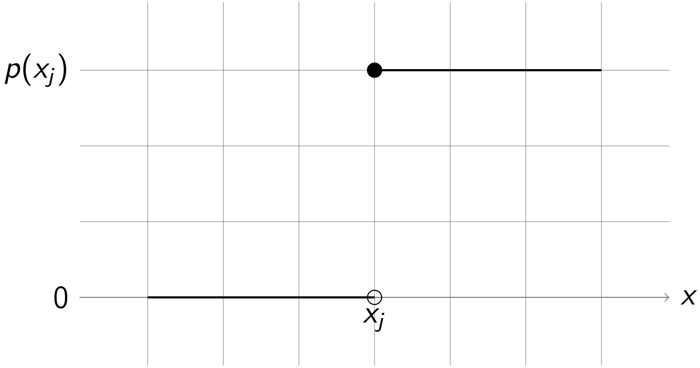
by definition. <u>The area under f is a continuous function of x.</u> For a discrete rv with range $\{x_1, \ldots, x_i, \ldots x_M\}$,

$$F_X(x) = \sum_{j=1}^{M} p(x_j)\mathbb{I}_{[x_j,\infty)}(x).$$

So $F_X(x)$ is a step function that jumps at each x_i. In particular, for $x_1 \leq \ldots \leq x_i \leq x < x_{i+1} \leq \ldots \leq x_M$

$$F_X(x) = \sum_{j=1}^{i} p(x_j).$$

Plot of $p(x_j)\mathbb{I}_{[x_j,\infty)}(x)$ as we vary x.



For a discrete rv

$$F_X(x) = \sum_{j=1}^{M} p(x_j)\mathbb{I}_{[x_j,\infty)}(x).$$

is a step-function.

For a continuous rv, the relationship between the cdf

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

and its pdf $f_X(t)$ is

$$f_X(t) = \frac{dF_X(t)}{dx}.$$

The relationship is a result from calculus (a result linking integration and differentiation.)

The cdf is useful when characterising the probability distribution of a transformation of a random variable. That is from X define a new rvs, eg. $Y = \exp(X)$ or $Y = X^2$.
Lets do a simple example first.

Example

Derive the density of $Y = X + a$.

First find the cdf of Y:

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= \Pr(X + a \leq y) \\ &= \Pr(X \leq y - a) \\ &= F_X(y - a). \end{aligned}$$

Next differentiate to get the pdf:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = f_X(y - a).$$

Fact

Let X have pdf f_X and cdf F_X. Define the new random variable $Y = r(X)$.
When r is strictly increasing or strictly decreasing we can derive a formula for f_Y. In this case r has an inverse, let $s = r^{-1}$. Then

$$f_Y(y) = f_X(s(y)) \left| \frac{ds(y)}{dy} \right| .$$

Example

Derive the density of $Y = X + a$.
Applying the previous result, $r(x) = x + a$, $s(y) = y - a$ and (since r is strictly increasing) $f_Y(y) = f_X(y - a)$.

We can verify the stated fact by following 3 simple steps. ✍
Step 1:

$$F_Y(y) = \Pr(Y \leq y) = \Pr(r(X) \leq y) \\ = \Pr(X \in (-\infty, s(y)))$$

if r is strictly increasing and $\Pr(X \in [s(y), \infty))$ if strictly decreasing.
Step 2:

$$F_Y(y) = F_X(s(y))$$

for r increasing and $F_Y(y) = 1 - F_X(s(y))$ for r decreasing.
Finally, step 3:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = f_X(s(y)) \left| \frac{ds(y)}{dy} \right|.$$

which holds for r increasing or decreasing.

Bivariates

Bivariate: Conditional pmf, Conditional pdf, Expectation, Conditional expectation.

Bivariates

A *bivariate* are two jointly distributed random variables.

Example

A bag contains balls numbered 1 to 5. You pick one at random, let X denote its number. Without replacement, you pick another one at random calling its number Y

Clearly if $X = i$ then $Y \neq i$ so knowing X narrows the range of Y.

> Jointly distributed random variables are not independent of each other.

Definition

Consider two discrete random variables X and Y where

$$X \in \{x_1, \ldots, x_m\}, \qquad Y \in \{y_1, \ldots, y_n\}.$$

Define the *joint* pmf to be

$$p_{X,Y}(x_i, y_j) = \Pr(X = x_i, Y = y_j)$$

which is the probability of the event $X = x_i$ and $Y = y_i$. The joint pmf is a complete description since it gives the probability of every possible outcome for the pair (X, Y).

Example

Returning to the prev. example, we can visualise $p_{X,Y}(x_i, y_j)$ as a table of numbers:

	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$	$Y = 5$
$X = 1$	0	1/20	1/20	1/20	1/20
$X = 2$	1/20	0	1/20	1/20	1/20
$X = 3$	1/20		0		
$X = 4$	1/20			0	
$X = 5$	1/20				0

Once the joint pmf $p_{X,Y}$ is given we can derive pmfs $p_X(x)$ and $p_Y(y)$. These are called the marginal pmfs. The marginal pmfs are

$$p_X(x_k) = \sum_{j=1}^{n} p_{X,Y}(x_k, y_j)$$

and

$$p_Y(y_k) = \sum_{i=1}^{m} p_{X,Y}(x_i, y_k).$$

Example

For previous example check $p_Y(i) = p_X(i) = 1/5$. ✍

Derive the result for X (as Y uses the same proof.)

$$
\begin{aligned}
&\Pr(X = x_k) \\
&= \Pr(X = x_k, Y \in \{y_1, \ldots, y_n\}) \\
&= \sum_{j=1}^{n} \Pr(X = x_k, Y = y_j) \\
&= \sum_{j=1}^{n} p_{X,Y}(x_k, y_j).
\end{aligned}
$$

Sometimes we are give the table of values $p_{X,Y}(x,y)$ and asked to verify dependence.

> Two discrete random variables X and Y are independent if
> $$p_{X,Y}(x,y) = p_X(x)p_Y(y) \quad \text{for all } (x,y).$$

Steps in verifying independence:

- Find $p_X(x)$ for all x values.
- Find $p_Y(y)$ for all y values.
- Show boxed relationship above.

(A similar checking result holds for jointly distributed continuos random variables.)

If X and Y are not independent then the conditional pmf is useful for statistical inference: when the value of one rv is observed, say $Y = y$, we wish to reason about the other unobserved one.

Definition

(Conditional probability mass function.) For the discrete random variables X and Y, the pmf of X given $Y = y$ is

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

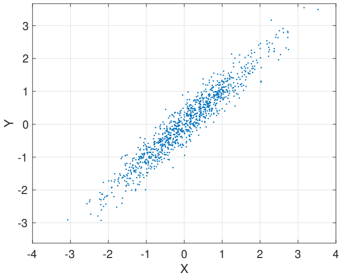
The cpmf $p_{X|Y}(\cdot|y)$ summarises all we know about X having observed $Y = y$.

Example

Show that $p_{X|Y}(\cdot|y)$ is itself a pmf. ✍
(A pmf must be non-negative and sum to 1.)

An example of jointly distributed continuous rvs.

X is $\mathcal{N}(0,1)$. Z is $\mathcal{N}(0,\epsilon)$ where $\epsilon < 1$ and let $Y = X + Z$. Here are many samples of (X, Y) for $\epsilon = 0.1$.



- Knowledge of X is useful in guessing the sign of Y, i.e. if $Y > 0$ or $Y < 0$.
- An alternative description of (X, Y) (which does not specify how they are generated as in the example) is to specify their joint probability density function.

Definition

For continuous random variables X and Y, we call a non-negative function $f(x, y)$ their joint probability density function if

- $\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x, y) dx \right) dy = 1$ and
- for any sets (events) $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$,

$$\Pr(X \in A, Y \in B) = \int_{-\infty}^{\infty} \mathbb{I}_B(y) \left(\int_{-\infty}^{\infty} \mathbb{I}_A(x) f(x, y) dx \right) dy.$$

(These double integrals can be evaluated in any order.)

Example

Let $f(x,y) = x + y$ for $x, y \in [0,1]$ and $f(x,y) = 0$ otherwise. Verify f is a pdf.

$$\int_0^1 \int_0^1 (x+y)dxdy = \int_0^1 \frac{1}{2}dy + \int_0^1 \frac{1}{2}dx = 1.$$

> The relationship between the joint pdf $f_{X,Y}(x,y)$ and the marginal pdfs $f_X(x)$ and $f_Y(x)$ is
>
> $$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \text{ and } f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx.$$

Example

Show $f_X(x) = x$ and $f_Y(y) = y$. ✍

Fact

(Verifying independence.) Two continuous rvs X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

Example

Recall previous example: $f_{X,Y}(x,y) = x + y$ and you showed $f_X(x) = x$ and $f_Y(y) = y$. Clearly $f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$ so X and Y are not independent of each other.

If X and Y are dependent then (as in the discrete case) the conditional pdf summarise all we know about the unobserved rv, say X, given the observed value y of the other rv Y.

Definition

(Conditional probability density function.) Let rvs X and Y have joint pdf $f_{X,Y}(x, y)$. The pdf of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)},$$

assuming $f_Y(y) > 0$. Moreover, for all sets A

$$\Pr(X \in A|Y = y) = \int_{-\infty}^{\infty} \mathbb{I}_A(x) f_{X|Y}(x|y) dx.$$

Example

(Sum of two independent random variables.) Let X_1 and X_2 be two independent rvs with pdfs $f_1(x_1)$ and $f_2(x_2)$ and let $Y = X_1 + X_2$. Find the pdf $f_{X_1,Y}$ and then f_Y.
Write the joint pdf using the conditional pdf formula

$$f_{X_1,Y}(x_1, y) = f_1(x_1)\, f_{Y|X_1}(y|x_1).$$

Since $Y = X_2 + x_1$, (from the transformation example) $f_{Y|X_1}(y|x_1) = f_2(y - x_1)$. Thus

$$\boxed{f_Y(y) = \int_{-\infty}^{\infty} f_2(y - x_1) f_1(x_1) dx_1}$$

which is the convolution of f_1 and f_2.

The example highlights a general result that says the pdf of the sum of two independent rvs is the convolution of their pdfs.

For a rv X (cts or discrete) or bivariate (X, Y), we would like to be able to compute the *average value* of $r(X)$ (or $r(X, Y)$ in the bivariate case) where r is a real valued function. Such computations are known as the *mathematical expectation*.

Definition

The *expected value* or *mean value* or *first moment* of X is

$$\mathbb{E}\{X\} = \begin{cases} \sum_x x p_X(x) & \text{(discrete)} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{(continuous.)} \end{cases}$$

Think of $\mathbb{E}\{X\}$ as the empirical average $n^{-1} \sum_{i=1}^{n} X_i$ where X_i are independent samples of X since the *law of large numbers (LLN) states that*

$$n^{-1} \sum_{i=1}^{n} X_i \to \mathbb{E}\{X\}.$$

Definition

For any function $r(\cdot)$ compute $\mathbb{E}\{r(X)\}$ by replacing x in the above formulae with $r(x)$. For example the *higher moments* are $\mathbb{E}(X^n)$ (for $n > 1$) so set $r(X) = X^n$.

Example (Calculating probabilities with the expectation operator. ✍)

For an event A

$$\mathbb{E}\{\mathbb{I}_A(X)\} = \begin{cases} \sum_x \mathbb{I}_A(x) p_X(x) & \text{(discrete)} \\ \int_{-\infty}^{\infty} \mathbb{I}_A(x) f_X(x) dx & \text{(continuous.)} \end{cases}$$

Then $\mathbb{E}\{\mathbb{I}_A(X)\} = \Pr\{X \in A\}$. The *frequency* interpretation of probability is formalised by the LLN:

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_A(X_i) \to \mathbb{E}\{\mathbb{I}_A(X)\}.$$

Example

Take a unit length stick and break it at random. Find the mean of the longer piece.

Call the longer piece Y and the break point X. Then X is a uniform rv in $[0,1]$, $Y = \max\{X, 1-X\}$ and

$$\begin{aligned}\mathbb{E}\{Y\} &= \mathbb{E}(\max\{X, 1-X\}) \\ &= \int_{-\infty}^{\infty} \max\{x, 1-x\} f_X(x) dx \\ &= \int_0^1 \max\{x, 1-x\} dx \\ &= \int_0^{0.5} (1-x) dx + \int_{0.5}^1 x dx = 0.75.\end{aligned}$$

Definition

The mean of a function $r(X, Y)$ of the bivariate (X, Y) is

$$\mathbb{E}\left\{r(X, Y)\right\} = \begin{cases} \sum_y \sum_x r(x, y) p_{X,Y}(x, y) & \text{(disc.)} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} r(x, y) f_{X,Y}(x, y) dx dy & \text{(cts.)} \end{cases}$$

The *conditional* expectation is

$$\mathbb{E}\left\{r(X, Y) | Y = y\right\} = \begin{cases} \sum_x r(x, y) p_{X|Y}(x|y) & \text{(disc.)} \\ \int_{-\infty}^{\infty} r(x, y) f_{X|Y}(x|y) dx & \text{(cts.)} \end{cases}$$

$\mathbb{E}\left\{r(X, Y) | X = x\right\}$ is computed similarly but this time fixing x and summing/integrating using either $p_{Y|X}(y|x)$ or $f_{Y|X}(y|x)$.

A new method to compute $\mathbb{E}\{r(X,Y)\}$ is obtained by exploiting the relationship $f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y)$.

$$\begin{aligned}
\mathbb{E}\{r(X,Y)\} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} r(x,y)f_{X,Y}(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} r(x,y)f_{X|Y}(x|y)dx\right) f_Y(y)dy \\
&= \int_{-\infty}^{\infty} \mathbb{E}\{r(X,Y)|Y=y\} f_Y(y)dy
\end{aligned}$$

where the last integrand can be interpreted as $\mathbb{E}\{g(Y)\}$ where $g(Y) = \mathbb{E}\{r(X,Y)|Y\}$.

Fact

(Rule of iterated expectation) For random variables X and Y,

$$\mathbb{E}\left\{r(X,Y)\right\} = \mathbb{E}\left(\mathbb{E}\left\{r(X,Y)|Y\right\}\right).$$

For continuous random variables

$$\mathbb{E}\left\{r(X,Y)|Y=y\right\} = \int_{-\infty}^{\infty} r(x,y) f_{X|Y}(x|y)dx$$

and

$$\mathbb{E}\left\{r(X,Y)\right\} = \int_{-\infty}^{\infty} \mathbb{E}\left\{r(X,Y)|Y=y\right\} f_Y(y)dy.$$

Example

If (X, Y) are jointly Gaussian, their joint pdf $f_{X,Y}(x, y)$ is

$$\frac{1}{2\pi |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}[x - m_1, y - m_2]\,\Sigma^{-1}\,[x - m_1, y - m_2]^T\right).$$

where $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix}$.

Show that the cpdf $f_{X|Y}(x|y)$ is a Gaussian pdf with ✍

$$\text{mean } m_1 + \frac{\rho}{\sigma_2^2}(y - m_2) \qquad \text{and variance } \sigma_1^2 - \frac{\rho^2}{\sigma_2^2}.$$

Calculate $\mathbb{E}\{XY\}$.

Use the iterated expectation rule: $\mathbb{E}\{XY\} = \mathbb{E}(\mathbb{E}\{XY|Y\})$ where

$$\mathbb{E}(\mathbb{E}\{XY|Y\}) = \int_{-\infty}^{\infty} \mathbb{E}\{XY|Y=y\} f_Y(y)dy.$$

$$\begin{aligned}\mathbb{E}\{XY|Y=y\} &= y\mathbb{E}\{X|Y=y\} \\ &= y\left(m_1 + \frac{\rho}{\sigma_2^2}(y-m_2)\right).\end{aligned}$$

$$\begin{aligned}&\int_{-\infty}^{\infty} \mathbb{E}\{XY|Y=y\} f_Y(y)dy \\ &= \int_{-\infty}^{\infty} y\left(m_1 - \frac{\rho}{\sigma_2^2}m_2\right) f_Y(y)dy + \int_{-\infty}^{\infty} \frac{\rho}{\sigma_2^2} y^2 f_Y(y)dy \\ &= m_2\left(m_1 - \frac{\rho}{\sigma_2^2}m_2\right) + \frac{\rho}{\sigma_2^2}(\sigma_2^2 + m_2^2) \\ &= \rho + m_1 m_2.\end{aligned}$$

✍

Multivariates

Random vectors: Joint pdf, Marginals, Gaussian (properties), Characteristic function, Change of variables (Jacobian.)

The study of two jointly distributed rvs is a special case of the following generalisation.

Definition

(Random vector.) Let $X_1, X_2, \ldots, X_n$ be n continuous (or n discrete) random variables. We call $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ a *continuous (or discrete) random vector*.

It is possible to define the joint pdf (pmf) and conditional pdf (pmf) as was done in the bivariate case.

Definition

Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a continuous random vector. Let $f(x_1, \ldots x_n)$ be a non-negative function that integrates to 1. Then f is called the pdf of the random vector X if

$$\begin{aligned} &\Pr\left(X_1 \in A_1, \ldots, X_n \in A_n\right) \\ &= \int_{-\infty}^{\infty} \mathbb{I}_{A_n}(x_n) \cdots \int_{-\infty}^{\infty} \mathbb{I}_{A_1}(x_1) f(x_1, \ldots, x_n) dx_1 \cdots dx_n \end{aligned}$$

for all events $A_1, \ldots, A_n$. (This integral should be read starting from the inner most integral, i.e. integrate wrt x_1, then wrt x_2 etc.)

Example

(Gaussian vector.) Let $X_1, X_2, \ldots, X_n$ be independent Gaussian random variables where X_i is $\mathcal{N}(\mu_i, \sigma_i^2)$ with pdf f_{X_i}.

$$\begin{aligned}
&\Pr(X_1 \in A_1, \ldots, X_n \in A_n) \\
&= \Pr(X_1 \in A_1) \ldots \Pr(X_n \in A_n) \quad \text{(independence)} \\
&= \left(\int_{-\infty}^{\infty} \mathbb{I}_{A_1}(x_1) f_{X_1}(x_1) dx_1\right) \ldots \left(\int_{-\infty}^{\infty} \mathbb{I}_{A_n}(x_n) f_{X_n}(x_n) dx_n\right) \\
&= \int_{-\infty}^{\infty} \mathbb{I}_{A_n}(x_n) \cdots \int_{-\infty}^{\infty} \mathbb{I}_{A_1}(x_1) f_{X_1}(x_1) \ldots f_{X_n}(x_n) dx_1 \cdots dx_n,
\end{aligned}$$

the last line writes the n integrals as one multi-integral. Thus the joint pdf $f(x, \ldots, x_n)$ is

$$f_{X_1}(x_1) f_{X_1}(x_1) \cdots f_{X_n}(x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right).$$

In general $f(x_1, \ldots, x_n)$ will not have such a product form.

Fact

The pdf of X_i is obtained by integrating $f(x_1, \ldots, x_n)$ over the full range of all its arguments except variable x_i:

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n.$$

We call the pdf of X_i the ith *marginal* of $f(x_1, \ldots, x_n)$.

We can show this result for X_n, and indeed any other X_i, as follows. Let $A_1 = \ldots = A_{n-1} = (-\infty, \infty)$ while A_n is arbitrary. Then

$$
\begin{aligned}
&\Pr(X_n \in A_n) \\
&= \Pr(X_1 \in A_1, \ldots, X_n \in A_n) \\
&= \int_{-\infty}^{\infty} \mathbb{I}_{A_n}(x_n) \cdots \int_{-\infty}^{\infty} \mathbb{I}_{A_1}(x_1) f(x_1, \ldots, x_n) dx_1 \cdots dx_n \\
&= \int_{-\infty}^{\infty} \mathbb{I}_{A_n}(x_n) \left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) dx_1 \cdots dx_{n-1} \right) dx_n \\
&= \int_{-\infty}^{\infty} \mathbb{I}_{A_n}(x_n) f_{X_n}(x_n) dx_n.
\end{aligned}
$$

Definition

The n random variables $X_1, \ldots, X_n$ are *independent* if and only if for every $A_1, \ldots, A_n$

$$\Pr\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \Pr\left(X_1 \in A_1\right) \cdots \Pr\left(X_n \in A_n\right).$$

Fact

Independence is equivalent to checking that the joint pdf reduces to the product of marginals:

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Example

The pdf $f(x_1, \ldots, x_n)$ of a Gaussian random vector $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is

$$\frac{1}{(2\pi)^{n/2}(\det C)^{1/2}} \exp\left\{-\frac{1}{2}(x-m)C^{-1}(x-m)^T\right\}$$

where $m = (m_1, \ldots m_n)$ is the (row vector) of means and C is the covariance matrix

$$m_i = \mathbb{E}\{X_i\} \qquad \text{and} \qquad [C]_{i,j} = \mathbb{E}\{(X_i - m_i)(X_j - m_j)\}.$$

Show that if $C_{i,j} = 0$ for $i \neq j$ then

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Call $C_{i,i} = \sigma_i^2$. Simplify the exponential term:

$$(x - m)C^{-1}(x - m)^T = \sum_{i=1}^{n} \frac{(x_i - m_i)^2}{\sigma_i^2}.$$

Hence $f(x_1, \ldots, x_n)$ is

$$\begin{aligned}
&\frac{1}{(2\pi)^{n/2}(\det C)^{1/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \frac{(x_i - m_i)^2}{\sigma_i^2}\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma_1 \cdots \sqrt{2\pi}\sigma_n} \prod_{i=1}^{n} \exp\left\{-\frac{1}{2}\frac{(x_i - m_i)^2}{\sigma_i^2}\right\} \\
&= f_{X_1}(x_1) \cdots f_{X_n}(x_n).
\end{aligned}$$

Now that we have introduced the concept of a random vector, we can state the following further properties of the expectation operator $\mathbb{E}(\cdot)$.

Fact

(Independence.) If $X_1, \ldots, X_n$ are independent random variables then $\mathbb{E}\left\{\prod_{i=1}^n X_i\right\} = \prod_{i=1}^n \mathbb{E}\left\{X_i\right\}$, that is the expectation of the product is the product of the expectation.

(Verification.) By independence, the joint pdf factorises $f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$.

$$
\begin{aligned}
&\mathbb{E}\left\{\prod_{i=1}^{n} X_i\right\} \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1 \cdots x_n f(x_1, \ldots, x_n) dx_1 \cdots dx_n \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1) \cdots x_n f_{X_n}(x_n) dx_1 \cdots dx_n \\
&= \left(\int_{-\infty}^{\infty} x_n f_{X_n}(x_n) dx_n\right) \cdots \left(\int_{-\infty}^{\infty} x_1 f_{X_1}(x_1) dx_1\right).
\end{aligned}
$$

Fact

(Linearity) If $X_1, \ldots, X_n$ are random variables and if $a_1, \ldots, a_n$ are real constants then $\mathbb{E}\left\{\sum_{i=1}^{n} a_i X_i\right\} = \sum_{i=1}^{n} a_i \mathbb{E}\left\{X_i\right\}$

This fact can be checked by executing the multi-integral for the integrand $x_1 a_1 + \ldots + x_n a_n$.

The change of variable formula can be applied to random vectors. Let

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} g_1(X_1, \ldots, X_n) \\ \vdots \\ g_n(X_1, \ldots, X_n) \end{bmatrix}$$

or

$$Y = G(X).$$

If G is invertible then $X = G^{-1}(Y)$. Let $H(Y) = G^{-1}(Y)$. So

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} h_1(Y_1, \ldots, Y_n) \\ \vdots \\ h_n(Y_1, \ldots, Y_n) \end{bmatrix}.$$

Form the matrix of partial derivatives of $H(y)$ (we need $H(y)$ to be continuous with continuous partial derivatives):

$$J(y) = \begin{bmatrix} \frac{\partial}{\partial y_1} h_1, \ldots, \frac{\partial}{\partial y_n} h_1 \\ \vdots \\ \frac{\partial}{\partial y_1} h_n, \ldots, \frac{\partial}{\partial y_n} h_n \end{bmatrix}$$

Then

$$\boxed{f_Y(y) = f_X\left(H(y)\right) \left|\det J(y)\right| .}$$

Remark

This is a result from calculus for performing a change of variable during integration and is not specific to the study of probability.

Example

Let $X_1, X_2, \ldots, X_n$ be independent Gaussian random variables where X_i is $\mathcal{N}(0,1)$. Let S be an invertible matrix and m a column vector. Let $Y = m + SX$ where $X = (X_1, \ldots, X_n)^T$. Show Y is also a Gaussian random vector.

Use the change of variable result:

$$H(Y) = S^{-1}(Y - m).$$

The matrix $J(y)$ is just

$$J(y) = S^{-1}.$$

Applying the change of variable formula gives

$$f_Y(y) = f_X\left(S^{-1}(y-m)\right)\left|\det S^{-1}\right|$$

where $f_X(x_1,\ldots,x_n) = \frac{1}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}x^Tx\right\}$. Thus

$$f_Y(y) = \frac{\left|\det S^{-1}\right|}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}(y-m)^T\left(S^{-1}\right)^T S^{-1}(y-m)\right\}$$

is the density of a Gaussian vector with mean m and covariance matrix SS^T. (Note that $\det S^{-1} = 1/\det S$, $\det(SS^T) = \det S \det S^T = (\det S)^2$.)

An affine transformation of a Gaussian vector is still a Gaussian vector. This gives a method for generating any Gaussian vector from iid Gaussian random variables.

To generate a $\mathcal{N}(m, \Sigma)$ vector:

- Decompose the symmetric matrix $\Sigma = S\ S^T$.
- Output $m + SX$ where $X = (X_1, \ldots, X_n)^T$

where $X_1, X_2, \ldots, X_n$ be independent $\mathcal{N}(0, 1)$ random variables.

But the transformation preserved the dimension of the vector. A more general result is that any affine transformation still yields a Gaussian (be it a variable or vector). We use the characteristic function to verify this.

Definition

(Characteristic function.) The characteristic function of a (discrete or continuous) random variable X is

$$\varphi_X(t) = \mathbb{E}\left\{\exp(itX)\right\}, \qquad t \in \mathbb{R}.$$

For a random vector $X = (X_1, X_2, \ldots, X_n)$, the characteristic function is

$$\varphi_X(t) = \mathbb{E}\left\{\exp(it^T X)\right\}, \qquad t \in \mathbb{R}^n.$$

Similarly to the Fourier transform, the characteristic function uniquely describes a pdf.

Example

(Gaussian.) Show $\varphi_X(t) = \exp(it\mu)\exp(-\frac{1}{2}\sigma^2 t^2)$ when X is a Gaussian random variable with mean μ and variance σ^2.

$$
\begin{aligned}
&\mathbb{E}\left\{\exp(itX)\right\} \\
&= \int_{-\infty}^{\infty} e^{itx} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx \\
&= e^{it\mu} \int_{-\infty}^{\infty} e^{its} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} s^2\right) ds, \quad \text{let } s = x - \mu \\
&= e^{it\mu} e^{-\frac{1}{2}\sigma^2 t^2} \qquad \text{(Fourier transform table.)}
\end{aligned}
$$

Example

Compute the characteristic function $\varphi_Y(t)$ of $Y = \sum_{i=1}^{n} X_i$ where X_i are independent random variables.

$$
\begin{aligned}
&\mathbb{E}\left\{\exp(itY)\right\} \\
&= \mathbb{E}\left\{\exp(itX_1)\exp(itX_2)\cdots\exp(itX_n)\right\} \\
&= \mathbb{E}\left\{\exp(itX_1)\right\}\mathbb{E}\left\{\exp(itX_2)\right\}\cdots\mathbb{E}\left\{\exp(itX_n)\right\} \\
&= \varphi_{X_1}(t)\cdots\varphi_{X_n}(t)
\end{aligned}
$$

where second equality used their independence.

The characteristic function of the sum of independent random variables is the product of their individual characteristic functions.

Example

(Moments.) Using $\varphi_X(t)$, compute $\mathbb{E}\left\{X^n\right\}$.

$$\frac{d^n}{dt^n}\varphi_X(t) = \mathbb{E}\left\{\frac{d^n}{dt^n}\exp(itX)\right\} = \mathbb{E}\left\{i^n X^n \exp(itX)\right\}.$$

Thus $i^n\mathbb{E}\left\{X^n\right\} = \frac{d^n}{dt^n}\varphi_X(t=0)$.

Fact

(Equality of characteristic functions.) Suppose that X and Y are random vectors with $\varphi_X(t) = \varphi_Y(t)$ for all $t \in \mathbb{R}^n$. Then X and Y have the same probability distribution.

Let $X_1, X_2, \ldots, X_n$ be independent Gaussian random variables where X_i is $\mathcal{N}(0,1)$. Then $Y = m + SX$, where $m \in \mathbb{R}^d$ with $d < n$, is multivariate Gaussian with mean m and covariance SS^T.

Verify the result using the characteristic function, that is let $t \in \mathbb{R}^d$ and compute $E\left\{\exp(it^T Y)\right\}$.

$$\begin{aligned} \exp(it^T Y) &= \exp(it^T m) \exp(it^T SX) \\ &= \exp(it^T m) \exp(ir_1 X_1) \cdots \exp(ir_n X_n) \end{aligned}$$

where vector $r = t^T S$.

$$
\begin{aligned}
&E\left\{\exp(it^TY)\right\} \\
&= \exp(it^Tm)E\left\{\exp(ir_1X_1)\right\}\cdots E\left\{\exp(ir_nX_n)\right\} \quad \text{(independence)} \\
&= \exp(it^Tm)\exp(-\frac{1}{2}r_1^2)\cdots\exp(-\frac{1}{2}r_n^2) \\
&= \exp(it^Tm)\exp(-\frac{1}{2}t^TSS^Tt)
\end{aligned}
$$

which is the characteristic function of a multivariate Gaussian with mean m and covariance SS^T.