STAT2001/STAT3101: Probability and Inference

Lecturer: Yvo Pokern,
Room 144 Dept. of Statistical Science
Lectures:

- **Medical Sciences 131 A V Hill LT on Thursdays 1pm-2pm**
- **Medical Sciences 131 A V Hill LT on Fridays 2pm-4pm**

Office Hour: Fridays 4:15pm-5:15pm
Email: `y.pokern@ucl.ac.uk`

Lecturecast

Lecturecast recording is active capturing both video of the front of the lecture theatre and audio using the lecturer's lapel microphone.

Lecture Topics

1. **Joint (or multivariate) distributions: describe the joint behaviour of more than one random variable.**
2. **Transformation of variables: e.g. taking logarithms to get closer to normality, the Body Mass Index**

Lecture Topics

1. Joint (or multivariate) distributions: describe the joint behaviour of more than one random variable.
2. Transformation of variables: e.g. taking logarithms to get closer to normality, the Body Mass Index
3. Generating functions: For simplifying calculations for probability distributions, e.g. for sums of independent random variables leading to the Central Limit Theorem
4. Distributions of functions of normally distributed variables: chi-squared, t, F etc.
5. Statistical estimation: how to derive estimators and obtain their properties

How to use the lecture notes

The lecture notes contain all relevant material for the course, *i.e.* definitions and results as well as the methods required to derive these results.
However, we will work through the examples in detail during the lecture and additional explanations not contained in this booklet will also be given in the lectures. It is therefore essential to attend the lectures and supplement the lecture notes with your own notes. The lecture notes would be woefully incomplete without the weekly exercise sheets that will be handed out separately and discussed in tutorials.

Exercise Sheets

Your solutions to the B-section of exercise sheets 1-8 should be handed in by the deadline stated on the exercise sheet. One randomly selected B-section question will be marked (the same question for everybody) and the 7 best marks out of 8 will constitute the 10% in-course assessed component of this course.

> ***UG Student Handbook:***
>
> ***Plagiarism means attempting to pass off someone else's work as your own, while collusion means passing off joint work as your own unaided effort. Both are unacceptable, particularly in material submitted for examination purposes including exercises done in your own time for in-course assessment.***

Tutorials

- For your tutorial slot: check your online timetable and email, if in serious doubt, contact Karen Leport (Statistical Science, General Office, Room 120)
- Attendance at Tutorials is compulsory
- Next week's tutorial: section A of exercise sheet 1
- Subsequent weeks: section B of previous week's exercise sheet

Content on Moodle at `https://moodle.ucl.ac.uk/`

Enrolment key: Rao-Blackwell

- **full answers to section A questions on the exercise sheets, very succinct answers to section B questions normally published Thursday mornings at 9:15am**
- **very succinct answers to section C questions (in time for exam preparation)**
- **messages, such as rearranged lectures**
- **general discussion forum moderated by the lecturer**
- **polls & quizzes to check your understanding and inform lecture progress**
- **Links to lecturecast material (recorded lectures and additional short explanatory videos for basic maths)**

Learning outcomes

At the end of each chapter and of important sections you will find a list of *Learning Outcomes.* These summarize key aspects, and point out what you are expected to be able to do once you have 'learned' the material. You can use them to monitor your own progress and to check whether you are well prepared for in–course assessments or the exam. The learning outcomes will be reflected in the examples and exercises given throughout the course.

Revision of Basic Probability

The fundamental idea of probability is that chance can be measured on a scale which runs from zero, which represents *impossibility*, to one, which represents *certainty*.

Sample space, Ω: the set of all outcomes of an experiment (real or hypothetical).

Event, *A*: a subset of Ω. The elements $\omega \in \Omega$ are called elementary events or outcomes.

Event Space, $\mathcal{A}$: The family of events whose probability we might be interested in.

Probability measure

P: a *mapping* from the Event Space $\mathcal{A}$ to $[0, 1]$ such that

1. $\mathbf{P}(A) \geq 0$
2. $\mathbf{P}(\Omega) = 1$
3. *Countable additivity:* **If $A_1, A_2, \ldots$ is a sequence of mutually disjoint sets (*i.e.* $A_i \cap A_j = \emptyset$, for all $i \neq j$) then**

$$\mathbf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbf{P}(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} \mathbf{P}(A_i).$$

These conditions are Kolmogorov's axioms of probability.

If Ω is countable (*i.e.* $\Omega = \{\omega_1, \omega_2, \dots\}$) then for P to be a probability measure, the axioms will generally hold for *all* subsets *A* and mutually disjoint A_i in Ω.

If Ω is uncountable, (like any interval of real numbers) we have to define a 'suitable' class of subsets, $\mathcal{A}$, the event space, for which the axioms hold — in practice this can always be constructed to include all events of interest.

For *any* two events *A* and *B* we have the addition rule

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$$

Independence of events

Events A and B are said to be independent if $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$.
Events $A_1, A_2, \ldots, A_n$ are independent if

$$\mathrm{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathrm{P}(A_{i_1}) \ldots \mathrm{P}(A_{i_k})$$

for all possible choices of k and
$1 \leq i_1 < i_2 < \cdots < i_k \leq n$.
That is, the product rule must hold for every subclass of the events $A_1, \ldots, A_n$.

Example 1.1

Consider two independent tosses of a fair coin and the events $A =$ ‘first toss is head’, $B =$ ‘second toss is head’, $C =$ ‘different results on two tosses’.
Find the sample space, the probability of an elementary event and the individual probabilities of A, B, and C.
Show that A, B, and C are not independent. □

Conditional Probability

Suppose that $P(B) > 0$. Then the conditional probability of A given B, $P(A|B)$ is defined as

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

***i.e.* the relative weight attached to event A within the restricted sample space B. The conditional probability is undefined if $P(B) = 0$. Note that $P(\cdot|B)$ is a probability measure on B.**

Further note that if A and B are independent events then $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

Multiplication Rule

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

Implies

$$\begin{aligned} \mathrm{P}(A \cap B) &= \mathrm{P}(A|B)\mathrm{P}(B) \\ &= \mathrm{P}(B|A)\mathrm{P}(A) \end{aligned}$$

Note that if $\mathrm{P}(B|A) = \mathrm{P}(B)$ then we recover the multiplication rule for independent events.

Conditional Independence

Two events A and B are conditionally independent given *a third event C if*

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

Conditional independence means that once we know that C is true A carries no information on B. Note that conditional independence does not imply marginal independence, nor vice versa.
Example 1.2(1.1 ctd.)

Show that A and B are not conditionally independent given $\overline{C}$. □

Total Probability

The law of total probability, or partition law follows from the additivity axiom and the definition of conditional probability: suppose that $B_1, \dots, B_k$ are mutually exclusive and exhaustive events (*i.e.* $B_i \cap B_j = \emptyset$ for all $i \neq j$ and $\cup_i B_i = \Omega$) and let A be any event. Then

$$\boxed{\mathrm{P}(A) = \sum_{j=1}^{k} \mathrm{P}(A \cap B_j) = \sum_{j=1}^{k} \mathrm{P}(A|B_j)\mathrm{P}(B_j)}$$

Example 1.3

A child gets to throw a fair die. If the die comes up 5 or 6, she gets to sample a sweet from box A which contains 10 chocolate sweets and 20 caramel sweets. If the die comes up 1,2,3 or 4 then she gets to sample a sweet from box B which contains 5 chocolate sweets and 15 caramel sweets. What is the conditional probability she will get a chocolate sweet if the die comes up 5 or 6? What is the conditional probability she will get a chocolate sweet if the die comes up 1,2,3 or 4? What is her probability of getting a chocolate sweet?

Bayes Theorem – Bayesian Statistics

Bayes theorem follows from the law of total probability and the multiplication rule.
Again, let $B_1, \dots, B_k$ be mutually exclusive and exhaustive events and let A be any event with $\mathrm{P}(A) > 0$. Then Bayes theorem states that

$$\boxed{\mathrm{P}(B_i|A) = \frac{\mathrm{P}(A|B_i)\mathrm{P}(B_i)}{\sum_{j=1}^{k} \mathrm{P}(A|B_j)\mathrm{P}(B_j)}}$$

1.2 Revision of random variables (univariate case)

A random variable, *X*, assigns a real number *x* to each element ω of the sample space Ω. The probability measure P on Ω then gives rise to a probability distribution for *X*. More formally, any function $X : \Omega \to \mathbb{R}$ is called a random variable.

The random variable *X* may be discrete or continuous.

1.2 Revision of random variables (univariate case)

A random variable, X, assigns a real number x to each element ω of the sample space Ω. The probability measure P on Ω then gives rise to a probability distribution for X. More formally, any function $X : \Omega \to \mathbb{R}$ is called a random variable.
The random variable X may be discrete or continuous.

The probability measure P on Ω induces a probability distribution for X. In particular, X has (cumulative) distribution function (cdf) $F_X(x) = \mathrm{P}(\{\omega : X(\omega) \leq x\})$, which is usually abbreviated to $\mathrm{P}(X \leq x)$.
It follows that $F_X(-\infty) = 0$, $F_X(\infty) = 1$, F_X is monotone increasing, and $\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a)$.

ANY function?

A random variable, X, assigns a real number x to each element ω of the sample space Ω. The probability measure P on Ω then gives rise to a probability distribution for X. More formally, any function $X : \Omega \to \mathbb{R}$ is called a random variable.

The random variable X may be discrete or continuous.

The probability measure P on Ω induces a probability distribution for X. In particular, X has (cumulative) distribution function (cdf) $F_X(x) = \mathrm{P}(\{\omega : X(\omega) \leq x\})$, which is usually abbreviated to $\mathrm{P}(X \leq x)$.
It follows that $F_X(-\infty) = 0$, $F_X(\infty) = 1$, F_X is monotone increasing, and $\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a)$.

ANY function? **ALL** of Gallia?

ANY function? **ALL** of Gallia? No!

The function must be measurable but it is *very hard* to come up with a non-measurable function...

Example 1.4

Give an example of a random variable whose cdf is right-continuous (it has to be) but not continuous.

Discrete random variables.

X takes only a finite or countably infinite set of values $\{x_1, x_2, \ldots\}$. F_X is a step-function, with steps at the x_i of sizes $p_x(x_i) = \mathrm{P}(X = x_i)$, and $p_x(\cdot)$ is the *probability mass function (pmf)* of X. (*E.g.* X = place of horse in race, grade of egg.)

Example 1.5 (1.1 ctd. II)

Consider, for example the random variable $X =$ number of heads obtained on the two tosses. Obtain the pmf and cdf of X. □

Example 1.6

Consider the random variable $X \sim \mathrm{Geo}(p)$ with $\mathrm{P}(X = k) = (1 - p)^{k-1} p$ where $k \in \mathbb{N}$. Compute the cdf and sketch it. Is X a discrete or a continuous random variable?

Continuous random variables.

If F_X can be expressed as

$$F_X(x) = \int_{-\infty}^{x} f_x(u)\, du.$$

with f_x non-negative and integrating to one, then X is a continuous random variable.
In this case, X takes values in a non-countable set and

$$\mathrm{P}(x < X \leq x + dx) \simeq f_x(x)\, dx.$$

Thus $f_x(x)\, dx$ is the probability that X lies in the infinitesimal interval $(x, x + dx)$. Note that the probability that X is *exactly* equal to x is zero for all x (*i.e.* $\mathrm{P}(X = x) = 0$).

Example 1.7

Suppose $f_X(x) = k(2 - x^2)$ on $(-1, 1)$. Calculate k and sketch the pdf. Calculate and sketch the cdf. Is the cdf differentiable? Calculate P$(|X| > 1/2)$.

Example 1.8

The lecturer L recruits a student volunteer V and obtains his/her height in feet and inches, converts to cm denoting the result as $h \in \mathbb{R}$.

1in=2.54cm
1ft=12in
=30.12cm

Example 1.8

The lecturer L recruits a student volunteer V and obtains his/her height in feet and inches, converts to cm denoting the result as $h \in \mathbb{R}$. Then ask for the maximal wager V would be willing to make on each of the following bets involving V's true height H:

1. **L pays V £1 if $H \in (h - 15cm, h + 15cm)$**
2. **L pays V £1 if $H \in (h - 5cm, h + 5cm)$**
3. **L pays V £1 if $H \in (h - 1cm, h + 1cm)$**
4. **L pays V £1 if $H \in (h - 0.5cm, h + 0.5cm)$**
5. **L pays V £1 if $H \in (h - 0.1cm, h + 0.1cm)$**

1in=2.54cm
1ft=12in
=30.12cm

Example 1.8

The lecturer L recruits a student volunteer V and obtains his/her height in feet and inches, converts to cm denoting the result as $h \in \mathbb{R}$. Then ask for the maximal wager V would be willing to make on each of the following bets involving V's true height H:

1in=2.54cm
1ft=12in
=30.12cm

1. L pays V £1 if $H \in (h - 15cm, h + 15cm)$
2. L pays V £1 if $H \in (h - 5cm, h + 5cm)$
3. L pays V £1 if $H \in (h - 1cm, h + 1cm)$
4. L pays V £1 if $H \in (h - 0.5cm, h + 0.5cm)$
5. L pays V £1 if $H \in (h - 0.1cm, h + 0.1cm)$

Assuming that fair bets have been offered, calculate the volunteer's subjective probability for each of the intervals.

Expectation.

A distribution has several characteristics that could be of interest such as its shape or skewness. Another one is its expectation, which can be regarded as a summary of the 'average' value of a random variable.

Discrete case:

$$\mathbb{E}(X) = \sum_i x_i \, p_X(x_i) = \sum_\omega X(\omega)\mathrm{P}(\{\omega\})\,.$$

That is, the averaging can be taken over the (distinct) values of X with weights given by the probability distribution p_X, *or* **over the sample space Ω with weights $\mathrm{P}(\{\omega\})$.**

Continuous case:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \, f_X(x) \, dx.$$

Example 1.9

The discrete random variable X has pmf $p_X(k) = \frac{1}{e^\mu - 1} \frac{\mu^k}{k!}$ for $k \in \mathbb{N}$. Compute its expectation.

Functions of a random variable

Let $\phi : \mathbb{R} \to \mathbb{R}$.
Then the random variable $Y = \phi(X)$ is defined by

$$Y(\omega) \equiv \phi(X)(\omega) = \phi(X(\omega))$$

Since $X : \Omega \to \mathbb{R}$, it follows that $\phi(X) : \Omega \to \mathbb{R}$. Thus $Y = \phi(X)$ is also a random variable. Its expectation is

$$\begin{aligned} \mathbb{E}\{\phi(X)\} &= \sum_i \phi(x_i) p_X(x_i). \\ &= \sum_\omega \phi(X(\omega)) \mathrm{P}(\{\omega\}) \end{aligned}$$

The first expression on the right-hand side averages the values of $\phi(x)$ over the distribution of X, whereas the second expression averages the values of $\phi(X(\omega))$ over the probabilities of $\omega \in \Omega$. A third method would be to compute the distribution of Y and average the values of y over the distribution of Y.

Example 1.10 (1.1 ctd III)

Let X be the random variable indicating the number of heads on two tosses. Consider the transformation ϕ with $\phi(0) = \phi(2) = 0$ and $\phi(1) = 1$.
Find $\mathbb{E}(X)$ and $\mathbb{E}\{\phi(X)\}$. □

Variance

The variance of X is

$$\sigma^2 = \mathrm{Var}(X) = \mathbb{E}\{X - \mathbb{E}(X)\}^2.$$

Equivalently $\sigma^2 = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2$ (*exercise: prove*). The square root, σ, of σ^2 is called the standard deviation.

Variance

The variance of X is

$$\sigma^2 = \text{Var}(X) = \mathbb{E}\{X - \mathbb{E}(X)\}^2.$$

Equivalently $\sigma^2 = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2$ (*exercise: prove*). The square root, σ, of σ^2 is called the standard deviation.

Example 1.11 (1.1 ctd. IV)

Find Var(X) and Var$\{\phi(X)\}$. □

Linear functions of X.

The following properties of expectation and variance are easily proved (*exercise/previous notes*):

$$\boxed{\mathbb{E}(a+bX)=a+b\mathbb{E}(X),\ \ \mathrm{Var}(a+bX)=b^2\mathrm{Var}(X)}$$

Linear functions of X.

The following properties of expectation and variance are easily proved (*exercise/previous notes*):

$$\boxed{\mathbb{E}(a+bX) = a + b\mathbb{E}(X), \;\; \text{Var}(a+bX) = b^2\text{Var}(X)}$$

Example 1.12 (1.1 ctd. V)

Let Y be the excess of heads over tails obtained on the two tosses of the coin. Write down $\mathbb{E}(Y)$ and Var(Y). □

Linear functions of X.

The following properties of expectation and variance are easily proved (*exercise/previous notes*):

$$\boxed{\mathbb{E}(a + bX) = a + b\mathbb{E}(X), \;\; \mathrm{Var}(a + bX) = b^2\mathrm{Var}(X)}$$

Example 1.12 (1.1 ctd. V)

Let Y be the excess of heads over tails obtained on the two tosses of the coin. Write down $\mathbb{E}(Y)$ and $\mathrm{Var}(Y)$. $\square$

Standard distributions: For ease of reference, Appendices 1 and 2 provide definitions of standard discrete and continuous distributions given in earlier courses.

1.3 Joint distributions

Let us first consider the bivariate case. Suppose that the two random variables *X* and *Y* share the same sample space Ω (*e.g.* the height and the weight of an individual). Then we can consider the event

$$\{\omega : X(\omega) \le x, Y(\omega) \le y\}$$

and define its probability, regarded as a function of the two variables *x* and *y*, to be the joint (cumulative) distribution function of *X* and *Y*, denoted by

$$\begin{aligned} F_{X,Y}(x,y) &= \mathrm{P}(\{\omega : X(\omega) \le x, Y(\omega) \le y\}) \\ &= \mathrm{P}(X \le x, Y \le y). \end{aligned}$$

1.3.1 Joint CDF

The joint cumulative distribution function (cdf) has similar *properties* to the univariate cdf. The joint cdf $F_{X,Y}(x, y)$ has the properties

1. **$F_{X,Y}(-\infty, y) = F_{X,Y}(x, -\infty) = 0$ and $F_{X,Y}(\infty, \infty) = 1$ and**
2. **$F_{X,Y}$ is a nondecreasing function of each of its arguments**
3. **$F_{X,Y}$ is *right-continuous*. That is, $F_{X,Y}(x + h, y + k) \to F_{X,Y}(x, y)$ as $h, k \downarrow 0$ for all x, y.**

Note that $F_{X,Y}$ is not necessarily left-continuous as it will have 'jumps' if X and/or Y is discrete.

1.3.1 Joint CDF

The joint cumulative distribution function (cdf) has similar *properties* to the univariate cdf. The joint cdf $F_{X,Y}(x,y)$ has the properties

1. $F_{X,Y}(-\infty, y) = F_{X,Y}(x, -\infty) = 0$ and $F_{X,Y}(\infty, \infty) = 1$ and
2. $F_{X,Y}$ is a nondecreasing function of each of its arguments
3. $F_{X,Y}$ is *right-continuous*. That is, $F_{X,Y}(x+h, y+k) \to F_{X,Y}(x,y)$ as $h, k \downarrow 0$ for all x, y.

Note that $F_{X,Y}$ is not necessarily left-continuous as it will have ‘jumps’ if X and/or Y is discrete.

The marginal cdfs of X and Y can be found from

$$F_X(x) = \mathrm{P}(X \leq x, Y < \infty) = F_{X,Y}(x, \infty)$$

and

$$F_Y(y) = \mathrm{P}(X < \infty, Y \leq y) = F_{X,Y}(\infty, y)$$

We already know in the univariate case that $\mathrm{P}(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$. Similarly, we find in the bivariate case that

$$\mathrm{P}(x_1 < X \leq x_2, \quad y_1 < Y \leq y_2) =$$

$$F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_2) - F_{X,Y}(x_2, y_1) + F_{X,Y}(x_1, y_1).$$

We already know in the univariate case that $P(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$. Similarly, we find in the bivariate case that

$$P(x_1 < X \leq x_2, \quad y_1 < Y \leq y_2) =$$

$$F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_2) - F_{X,Y}(x_2, y_1) + F_{X,Y}(x_1, y_1).$$

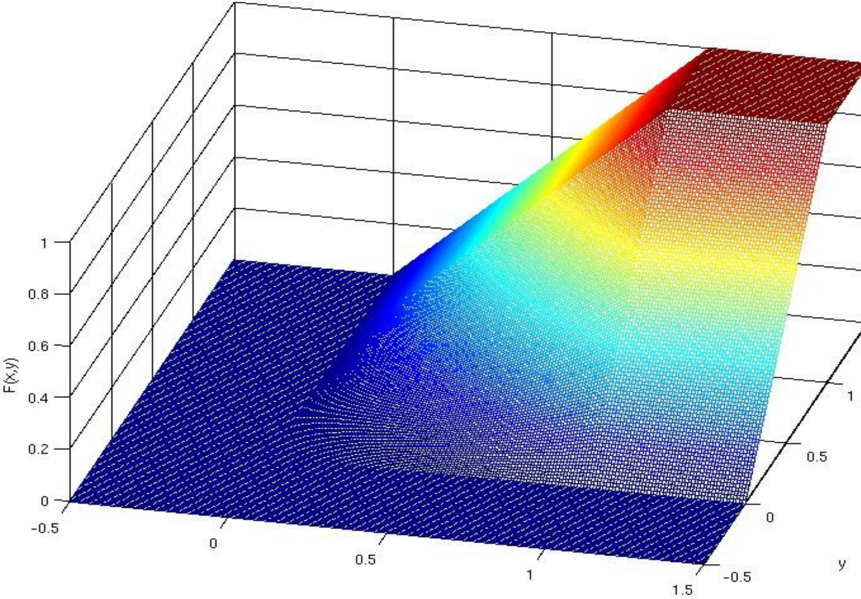
Example 1.13

Consider the function

$$F_{X,Y}(x, y) = x^2y + y^2x - x^2y^2\,, \quad 0 \leq x \leq 1,\, 0 \leq y \leq 1\,.$$

Show that $F_{X,Y}$ has the above properties of CDFs. Find the marginal cdfs of X and Y. Also find $P(0 \leq X \leq \frac{1}{2}, 0 \leq Y \leq \frac{1}{2})$. □

$$F_{X,Y}(x, y) = x^2y + y^2x - x^2y^2$$



1.3.2 The discrete case

In many cases of interest, X and Y take only values in a discrete set. Then $F_{X,Y}$ is a step function in each variable separately and we consider the joint probability mass function

$$p_{X,Y}(x_i, y_j) = \mathrm{P}(X = x_i, Y = y_j).$$

It is often convenient to represent a discrete bivariate distribution – a joint distribution of two variables – by a *2-way table.* In general, the entries in the table are the *joint probabilities* **$p_{X,Y}(x, y)$, while the row and column totals give the** *marginal* **probabilities $p_X(x)$ and $p_Y(y)$. As always, the total probability is 1.**

Example 1.14

Consider three independent tosses of a fair coin. Let $X =$ ‘number of heads in first and second toss’ and $Y =$ ‘number of heads in second and third toss’. Give the probabilities for any combination of possible outcomes of X and Y in a two-way table and obtain the marginal pmfs of X and Y. □

In general, from the joint distribution we can use the law of total probability to obtain the marginal pmf of Y as

$$\begin{aligned} p_Y(y_j) = \mathrm{P}(Y = y_j) &= \sum_{x_i} \mathrm{P}(X = x_i,\ Y = y_j) \\ &= \sum_{x_i} p_{X,Y}(x_i, y_j)\,. \end{aligned}$$

Similarly, the marginal pmf of X is given by

$$p_X(x_i) = \sum_{y_j} p_{X,Y}(x_i, y_j)\,.$$

The marginal distribution is thus the distribution of just one of the variables.

The joint cdf can be written as

$$F_{X,Y}(x, y) \quad = \quad \sum_{x_i \le x} \sum_{y_j \le y} p_{X,Y}(x_i, y_j)\,.$$

Note that there will be jumps in $F_{X,Y}$ at each of the x_i and y_j values.

Random Question to be Marked

A set of English scrabble letters are sorted into three bags: all letters with 1 point are sorted into bag (i), all letters with 0,2 or 3 points are sorted into bag (ii) and all other letters go to bag (iii). A die is thrown: if it comes up 1, bag (i) is sampled from, if it comes up 2 or 3, bag (ii) is sampled from, otherwise bag (iii) is sampled from. Exactly one letter is sampled from the selected bag.

The three events

- A_1 **One of the letters B,C,M,N,Q,R,T,Z is sampled.**
- A_2 **One of the letters A,D,E,F,H,I,G,V,W**
- A_3 **Neither A_1 nor A_2 happens.**

are assigned to the exercise questions B1,B2,B3 by audience vote.

What assignment would you like?

Independence

The random variables X and Y, defined on the sample space Ω with probability measure P, are independent if the events

$$\{X = x_i\} \text{ and } \{Y = y_j\}$$

are *independent events*, for all possible values x_i and y_j. Thus X and Y are independent if,

$$p_{X,Y}(x_i, y_j) = \mathrm{P}(X = x_i, Y = y_j) = p_X(x_i)p_Y(y_j)$$

for all x_i, y_j.

This implies that $\mathrm{P}(X \in A, Y \in B) = \mathrm{P}(X \in A)\mathrm{P}(Y \in B)$ for all sets A and B, so that the two events $\{\omega : X(\omega) \in A\}, \{\omega : Y(\omega) \in B\}$ are independent. (*Exercise:* prove this.)

NB: If x is such that $p_X(x) = 0$, then $p_{X,Y}(x, y_j) = 0$ for all y_j and the above factorisation holds automatically. Thus it does not matter whether we require the factorisation for all *possible* **x_i, y_j *i.e.* those with positive probability, or all** *real* **x, y. (That is, $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ for all x, y would be an equivalent definition of independence.)**

If X, Y are independent then the entries in the two way table are the products of the marginal probabilities. In Example 1.14 we see that X and Y are *not* **independent.**

Conditional probability distributions

These are defined for random variables by analogy with conditional probabilities of events.

$$\mathrm{P}(X = x_i \mid Y = y_j) = \frac{\mathrm{P}(X = x_i,\ Y = y_j)}{\mathrm{P}(Y = y_j)} = \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)}$$

as a function of x_i, for fixed y_j. Then this is a pmf – it is non-negative and

$$\sum_{x_i} \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)} = \frac{1}{p_Y(y_j)} \underbrace{\sum_{x_i} p_{X,Y}(x_i, y_j)}_{p_Y(y_j)} = 1,$$

and it gives the probabilities for observing $X = x_i$ given that we already know $Y = y_j$.

We therefore *define* the conditional probability distribution of X given $Y = y_j$ as

$$\boxed{p_{X|Y}(x_i|y_j) = \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)}}$$

Conditioning on $Y = y_j$ can be compared to selecting a subset of the population, i.e. only those individuals where $Y = y_j$. The conditional distribution $p_{X|Y}$ of X given $Y = y_j$ then describes the distribution of X within this subgroup.

From the above definition we immediately obtain the multiplication rule for pmfs:

$$\boxed{p_{X,Y}(x_i, y_j) = p_{X|Y}(x_i|y_j)p_Y(y_j)}$$

which can be used to find a bivariate pmf when we known one marginal distribution and one conditional distribution. Note that if X and Y are *independent* then $p_{X,Y}(x_i, y_j) = p_X(x_i)p_Y(y_i)$ so that $p_{X|Y}(x_i|y_j) = p_X(x_i)$ *i.e.* the *conditional* distribution is the same as the *marginal* distribution.

In general, X and Y are independent *if and only if* the conditional distribution of X given $Y = y_j$ is the same as the marginal distribution of X *for all* y_j. (This condition is equivalent to $p_{X,Y}(x_i, y_j) = p_X(x_i)p_Y(y_j)$ for all x_i, y_i, above). The conditional distribution of Y given $X = x_i$ is defined similarly.

Example 1.15 (1.14 ctd.)

Obtain the conditional pmf of X given $Y = y$. Use this conditional distribution to verify that X and Y are not independent. □

Example 1.16

Suppose that R and N have a joint distribution in which $R|N$ is Bin(N, π) and N is Poi(λ). Show that R is Poi$(\pi\lambda)$.

□

Conditional expectation

Since $p_{X|Y}(x_i|y_j)$ is a probability distribution, it has a mean or expected value:

$$\mathbb{E}(X \mid Y = y_j) = \sum_{x_i} x_i \, p_{X|Y}(x_i|y_j)$$

which represents the average value of X among outcomes ω for which $Y(\omega) = y_j$. This may also be written $\mathbb{E}_{X|Y}(X \mid Y = y_j)$. We can also regard the conditional expectation $\mathbb{E}(X \mid Y = y_j)$ as the mean value of X in the subgroup characterised by $Y = y_j$.

Example 1.17 (1.14 ctd. II)

Find the conditional expectations $\mathbb{E}(X \mid Y = y)$ for $y = 0, 1, 2$. Plot the graph of the function $\phi(y) = \mathbb{E}[X|Y = y]$. What do these values tell us about the relationship between X and Y? □

In general, what is the relationship between the *unconditional* expectation $\mathbb{E}(X)$ and the *conditional* expectation $\mathbb{E}(X \mid Y = y_j)$?

Example 1.18

Collect the joint distribution of X: gender ($x_1 = M, x_2 = F$) and Y: number of cups of tea drunk today ($y_1 = 0, y_2 = 1, y_3 = 2, y_4 = 3 \text{ or more}$). Are X and Y independent? What is the expectation of Y? What is the conditional expectation $Y|X = M$ and $Y|X = F$? □

Consider the conditional expectation $\phi(y) = \mathbb{E}_{X|Y}(X|Y = y)$ as a function of y. We may compute its expectation $\mathbb{E}\{\phi(Y)\} = \mathbb{E}_Y\{\mathbb{E}_{X|Y}(X|Y)\}$. But, from the definition of the expectation of a function of Y, we have E$\{\phi(Y)\} = \sum_{y_j} \phi(y_j)p_Y(y_j)$, so that

$$\mathbb{E}_Y\{\mathbb{E}_{X|Y}(X \mid Y)\} = \sum_{y_j} \underbrace{\mathbb{E}(X \mid Y = y_j)}_{\text{function of } y_j} p_Y(y_j)$$

Consider the conditional expectation $\phi(y) = \mathbb{E}_{X|Y}(X|Y = y)$ as a function of y. We may compute its expectation $\mathbb{E}\{\phi(Y)\} = \mathbb{E}_Y\{\mathbb{E}_{X|Y}(X|Y)\}$. But, from the definition of the expectation of a function of Y, we have E$\{\phi(Y)\} = \sum_{y_j} \phi(y_j)p_{Y}(y_j)$, so that

$$\mathbb{E}_Y\{\mathbb{E}_{X|Y}(X \mid Y)\} = \sum_{y_j} \underbrace{\mathbb{E}(X \mid Y = y_j)}_{\text{function of } y_j} p_{Y}(y_j)$$

We show in the lectures that this gives the marginal expectation of $\mathbb{E}(X)$. That is,

$$\boxed{\mathbb{E}(X) = \mathbb{E}_Y\{\mathbb{E}_{X|Y}(X|Y)\}}$$

which is known as the iterated conditional expectation.

The iterated conditional expectation formula is most useful when the conditional distribution of X given $Y = y$ is known and easier to handle than the joint distribution (requiring integration to find the marginal of X if it is not known).

Example 1.19 (1.14 ctd. II)

Verify that $\mathbb{E}(Y) = \mathbb{E}_X\{\mathbb{E}_{Y|X}(Y|X)\}$ in this example. □

Example 1.20 (1.16 ctd.)

Find the mean of R using the iterated conditional expectation formula. □

Expectation of Functions of two variables

The definition of expectation generalises immediately to functions of two variables, *i.e.* we can compute it from the joint pmf:

$$
\begin{aligned}
\mathbb{E}\left\{\phi(X, Y)\right\} &= \sum_{\omega} \phi(X(\omega), Y(\omega))\mathrm{P}\left(\{\omega\}\right) \\
&= \sum_{x_i} \sum_{y_j} \phi(x_i, y_j)\mathrm{P}\left(\{\omega : X(\omega) = x_i, Y(\omega) = y_j\}\right) \\
&= \sum_{x_i} \sum_{y_j} \phi(x_i, y_j)p_{X,Y}(x_i, y_j)
\end{aligned}
$$

Iterated Conditional Expectation Formula:

$$
\begin{aligned}
\mathbb{E}\left\{\phi(X, Y)\right\} &= \sum_{x_i}\sum_{y_j}\phi(x_i, y_j)p_{X|Y}(x_i \mid y_j)p_Y(y_j) \\
&= \sum_{y_j} p_Y(y_j)\underbrace{\sum_{x_i}\phi(x_i, y_j)p_{X|Y}(x_i \mid y_j)}_{\mathbb{E}_{X|Y}(\phi(X,y_j)|Y=y_j)} \\
&= \mathbb{E}_Y\{\mathbb{E}_{X|Y}(\phi(X, Y) \mid Y)\}\,.
\end{aligned}
$$

TOK

Taking out what is known (TOK)

$$\mathbb{E}_{X|Y}[\phi(Y)\psi(X, Y)|Y] = \phi(Y)\mathbb{E}_{X|Y}[\psi(X, Y)|Y]$$

This will be shown in lectures for discrete random variables only. It also holds for continuous and mixed random variables, however.

Example 1.21

Consider two random variables X and Y, where the marginal probabilities of Y are $\mathrm{P}(Y = 0) = 3/4$, $\mathrm{P}(Y = 1) = 1/4$ and the conditional probabilities of X are $\mathrm{P}(X = 1|Y = 0) = \mathrm{P}(X = 2|Y = 0) = 1/2$ and $\mathrm{P}(X = 0|Y = 1) = \mathrm{P}(X = 1|Y = 1) = \mathrm{P}(X = 2|Y = 1) = 1/3$. Use the iterated conditional expectation formula to find $\mathbb{E}(XY)$.

$\square$

1.3.3 The continuous case

Now, both X and Y take values in a continuous range and their joint cdf $F_{X,Y}(x,y)$ is differentiable with respect to both x and y. Then $F_{X,Y}$ can be expressed as

$$\boxed{F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)dvdu}$$

where $f_{X,Y}(x,y)$ is the joint probability density function of X and Y. Letting $y \to \infty$ we get

$$F_X(x) = F_{X,Y}(x,\infty) \; = \; \int_{-\infty}^{x}\Big(\int_{-\infty}^{\infty} f_{X,Y}(u,v)dv\Big)du$$

But from §1.2 we also know that $F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$. It follows that the marginal density function of X is

$$\boxed{f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,v)dv}$$

Similarly, Y has marginal density

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(u, y) du$$

As for the univariate case, we have

$$\begin{aligned} &\mathrm{P}(x < X \leq x + dx, y < Y \leq y + dy) \\ &\qquad = \int_x^{x+dx} \int_y^{y+dy} f_{X,Y}(u, v) dvdu \simeq \quad f_{X,Y}(x, y) dxdy \,. \end{aligned}$$

That is, $f_{X,Y}(x, y)dxdy$ is the probability that (X, Y) lies in the infinitesimal rectangle $(x, x + dx) \times (y, y + dy)$. As in the univariate case, $\mathrm{P}(X = x,\ Y = y) = 0$ for all x, y.

Example 1.22

Consider two continuous random variables X and Y with joint density

$$f_{X,Y}(x,y) = \begin{cases} 8xy & 0 \leq x \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Sketch the area where $f_{X,Y}$ is positive. Derive the marginal pdfs of X and Y.

Independence

By analogy with the discrete case, two random variables X and Y are said to be independent if their joint density factorises, ie if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \text{ for all } x,y.$$

Slogan:

Independence means Factorising

IF X and Y are independent
THEN $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x, y.

Equivalent characterisation of independence

Two continuous random variables are independent if and only if there exist functions $g(\cdot)$ and $h(\cdot)$ such that for all (x, y) the joint density factorises as $f_{X,Y}(x, y) = g(x)h(y)$, where g is a function of x only and h is a function of y only.

Proof. If X and Y are independent then simply take $g(x) = f_X(x)$ and $h(y) = f_Y(y)$. For the converse, suppose that $f_{X,Y}(x, y) = g(x)h(y)$ and define

$$G = \int_{-\infty}^{\infty} g(x)dx, \qquad H = \int_{-\infty}^{\infty} h(y)dy.$$

Note that both G and H are finite (why?).

Then the marginal densities are $f_X(x) = g(x)H$, $f_Y(y) = Gh(y)$ and either of these equations implies that $GH = 1$. It follows that

$$f_{X,Y}(x,y) = g(x)h(y) = \frac{f_X(x)}{H}\frac{f_Y(y)}{G} = f_X(x)f_Y(y)$$

and so X and Y are independent. □

The advantage of knowing that under independence $f_{X,Y}(x,y) = g(x)h(y)$ is that we don't need to find the marginal densities $f_X(x)$ and $f_Y(y)$ (which would typically involve some integration) to verify independence.

Example 1.23 (1.22 ctd.)

Reminder: Example 1.22

Consider two continuous random variables X and Y with joint density

$$f_{X,Y}(x,y) = \begin{cases} 8xy & 0 \le x \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Sketch the area where $f_{X,Y}$ is positive. Derive the marginal pdfs of X and Y.

Example 1.23

Are X and Y independent?

Conditional distributions

For the conditional distribution, we cannot condition on $Y = y$ in the usual way, as for any arbitrary set A, $\mathrm{P}(X \in A \text{ and } Y = y) = \mathrm{P}(Y = y) = 0$ when Y is continuous, so that

$$\mathrm{P}(X \in A \mid Y = y) = \frac{\mathrm{P}(X \in A, Y = y)}{\mathrm{P}(Y = y)}$$

is not defined (0/0). However, we can consider

$$\frac{\mathrm{P}(x < X \leq x + dx, y < Y \leq y + dy)}{\mathrm{P}(y < Y \leq y + dy)} \quad \simeq \quad \frac{f_{X,Y}(x, y)dxdy}{f_Y(y)dy}$$

and interpret $f_{X,Y}(x, y)/f_Y(y)$ as the conditional density of X given $Y = y$ written as $f_{X|Y}(x \mid y)$.

Note that this *is* a probability density function – it is non-negative and

$$\int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)}{f_Y(y)} dx = \frac{1}{f_Y(y)} \underbrace{\int_{-\infty}^{\infty} f_{X,Y}(x,y) dx}_{f_Y(y)} = 1.$$

If X and Y are independent then, as before, the conditional density of X given $Y = y$ is just the marginal density of X.

Example 1.24 (1.22 ctd. II)

Give the conditional densities of X given $Y = y$ and of Y given $X = x$ indicating clearly the area where they are positive. Also, find $\mathbb{E}(X|Y = y)$ and $\mathbb{E}(X)$, using the law of iterated conditional expectation for the latter. Compare this with the direct calculation of $\mathbb{E}(X)$. □

Experiment: Question to be marked

Spaghetto Breaking

Measure the length L of a randomly chosen Spaghetto. Break it in two pieces at a random point. Measure the length L_1 of one of the pieces. Convert this length to the question to be marked using the following table:

Experiment: Question to be marked

Spaghetto Breaking

Measure the length L of a randomly chosen Spaghetto. Break it in two pieces at a random point. Measure the length L_1 of one of the pieces. Convert this length to the question to be marked using the following table:

$L_1 \leq 10\mathrm{cm}$	B2
$10\mathrm{cm} \leq L_1 \leq 20\mathrm{cm}$	B3
$20\mathrm{cm} < L_1$	B1

Experiment: Question to be marked

Spaghetto Breaking

Measure the length L of a randomly chosen Spaghetto. Break it in two pieces at a random point. Measure the length L_1 of one of the pieces. Convert this length to the question to be marked using the following table:

$L_1 \leq 10\text{cm}$	B2
$10\text{cm} \leq L_1 \leq 20\text{cm}$	B3
$20\text{cm} < L_1$	B1

Want to know what the odds were? Do exercise sheet 3, question B3!

1.4 Further results on expectations

Expectation of a sum.
Consider the sum $\phi(X) + \psi(Y)$ when X, Y have joint probability mass function $p_{X,Y}(x, y)$. (The continuous case follows similarly, replacing probabilities by probability densities and summations by integrals.)

1.4 Further results on expectations

Expectation of a sum.
Consider the sum $\phi(X) + \psi(Y)$ when X, Y have joint probability mass function $p_{X,Y}(x, y)$. (The continuous case follows similarly, replacing probabilities by probability densities and summations by integrals.) Then

$$\mathbb{E}_{X,Y}\left[\phi(X) + \psi(Y)\right] \;=\; \sum_{x_i}\sum_{y_j}\{\phi(x_i) + \psi(y_j)\}\, p_{X,Y}(x_i, y_j)$$

1.4 Further results on expectations

Expectation of a sum.
Consider the sum $\phi(X) + \psi(Y)$ when X, Y have joint probability mass function $p_{X,Y}(x,y)$. (The continuous case follows similarly, replacing probabilities by probability densities and summations by integrals.) Then

$$\begin{aligned}
\mathbb{E}_{X,Y}[\phi(X)+\psi(Y)] &= \sum_{x_i}\sum_{y_j}\{\phi(x_i)+\psi(y_j)\}\, p_{X,Y}(x_i,y_j) \\
&= \sum_{x_i}\phi(x_i)\underbrace{\sum_{y_j} p_{X,Y}(x_i,y_j)}_{p_X(x_i)} \\
&\quad + \sum_{y_j}\psi(y_j)\underbrace{\sum_{x_i} p_{X,Y}(x_i,y_j)}_{p_Y(y_j)}
\end{aligned}$$

1.4 Further results on expectations

Expectation of a sum.

Consider the sum $\phi(X) + \psi(Y)$ when X, Y have joint probability mass function $p_{X,Y}(x, y)$. (The continuous case follows similarly, replacing probabilities by probability densities and summations by integrals.) Then

$$
\begin{aligned}
\mathbb{E}_{X,Y}\left[\phi(X) + \psi(Y)\right] &= \sum_{x_i}\sum_{y_j}\{\phi(x_i) + \psi(y_j)\}\, p_{X,Y}(x_i, y_j) \\
&= \sum_{x_i}\phi(x_i)\underbrace{\sum_{y_j} p_{X,Y}(x_i, y_j)}_{p_X(x_i)} \\
&\quad + \sum_{y_j}\psi(y_j)\underbrace{\sum_{x_i} p_{X,Y}(x_i, y_j)}_{p_Y(y_j)} \\
&= \mathbb{E}_X\left[\phi(X)\right] + \mathbb{E}_Y\left[\psi(Y)\right] .
\end{aligned}
$$

$$\mathbb{E}_{X,Y}\left[\phi(X)+\psi(Y)\right] \;=\; \mathbb{E}_{X}\left[\phi(X)\right]+\mathbb{E}_{Y}\left[\psi(Y)\right]$$

Note that the subscripts on the E's are unnecessary as there is no possible ambiguity in this equation, and also that this holds regardless of whether or not X and Y are independent.

In particular we have $\mathbb{E}(X+Y)=\mathbb{E}(X)+\mathbb{E}(Y)$. Note the power of this result: there is no need to calculate the probability distribution of $X+Y$ (which may be hard!) if all we need is the mean of $X+Y$.

Expectation of a product.

Now consider $\phi(X)\psi(Y)$. Then

$$\begin{aligned}\mathbb{E}_{X,Y}\left[\phi(X)\psi(Y)\right] &= \sum_{x_i}\sum_{y_j}\{\phi(x_i)\psi(y_j)\}p_{X,Y}(x_i,y_j)\\ &= ?\end{aligned}$$

Expectation of a product.

Now consider $\phi(X)\psi(Y)$. Then

$$\begin{aligned}\mathbb{E}_{X,Y}\left[\phi(X)\psi(Y)\right] &= \sum_{x_i}\sum_{y_j}\{\phi(x_i)\psi(y_j)\}p_{X,Y}(x_i,y_j)\\ &= ?\end{aligned}$$

If **X and Y are** *independent*, **then $p_{X,Y}(x_i,y_j) = p_X(x_i)\ p_Y(y_j)$:**

$$\mathbb{E}_{X,Y}\left[\phi(X)\psi(Y)\right] = \underbrace{\sum_{x_i}\phi(x_i)p_X(x_i)}_{\mathbb{E}_X[\phi(X)]}\ \underbrace{\sum_{y_j}\psi(y_j)p_Y(y_j).}_{\mathbb{E}_Y[\psi(Y)]}$$

Thus, *except* **for the case where X and Y are** *independent*, **we typically have that**

E(product) $\neq$ product of expectations

But it is *always* **true that**

E(sum) = sum of expectations.

Slogan:

Independence means Factorising

IF *X* and *Y* are independent
THEN $\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left[\mathbf{XY}\right] = \mathbf{E}_{\mathbf{X}}\left[\mathbf{X}\right]\mathbf{E}_{\mathbf{Y}}\left[\mathbf{Y}\right]$

Covariance

A particular function of interest is the covariance between X and Y. As we will see, this is a measure for the strength of the linear relationship between X and Y. The covariance is defined as

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

An alternative formula for the covariance follows on expanding the bracket, giving

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}\ [XY - X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X)\ \mathbb{E}(Y)] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\ \mathbb{E}(Y) - \mathbb{E}(X)\ \mathbb{E}(Y) + \mathbb{E}(X)\ \mathbb{E}(Y) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\ \mathbb{E}(Y)\end{aligned}$$

Note that $\text{cov}(X, X) = \text{var}(X)$, giving the familiar formula $\text{var}(X) = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2$.

If *X* and *Y* are *independent* then, from above,

$$\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$$

and it follows that

$$\mathrm{Cov}(X, Y) = 0\,.$$

However **in general cov(X, Y) $= 0 \not\Rightarrow$ *X* and *Y* are independent! (Example below)**

Also if $Z = aX + b$, then E$(Z) = a\mathbb{E}(X) + b$ and $Z - \mathbb{E}(Z) = a\{X - \mathbb{E}(X)\}$, so that

$$\begin{aligned} \text{Cov}(\mathrm{Z}, Y) &= \mathbb{E}\{a(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\} \\ &= a\text{Cov}(X, Y). \end{aligned}$$

Using a similar argument we get

$$\text{Cov}(X + Y, W) = \text{Cov}(X, W) + \text{Cov}(Y, W).$$

Exercise: **Using the fact that Var$(X + Y) =$ Cov$(X + Y, X + Y)$, derive the general formula Var$(X + Y) =$ Var$(X) +$ Var$(Y) + 2$Cov(X, Y).**

Correlation:

From above, we see that the covariance varies with the scale of measurement of the variables (lbs/kilos etc), making it difficult to interpret its numerical value. The correlation is a standardised form of the covariance, which is *scale-invariant* **and therefore its values are easier to interpret.**

The correlation between X and Y is defined by

$$\text{Corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Suppose that $a > 0$. Then $\text{cov}(aX, Y) = a\,\text{Cov}(X, Y)$ and $\text{Var}(aX) = a^2\,\text{Var}(X)$, so it follows that $\text{corr}(aX, Y) = \text{Corr}(X, Y)$, and thus the correlation is *scale-invariant*.

A key result is that

$$\boxed{-1 \leq \text{Corr}(X, Y) \leq +1}$$

for all random variables X and Y. The proof is STAT3101 only and (hopefully) made available via moodle/lecturecast.

Example 1.21 ctd.

Example 1.21

Consider two random variables X and Y, where the marginal probabilities of Y are $\mathrm{P}(Y = 0) = 3/4$, $\mathrm{P}(Y = 1) = 1/4$ and the conditional probabilities of X are $\mathrm{P}(X = 1|Y = 0) = \mathrm{P}(X = 2|Y = 0) = 1/2$ and $\mathrm{P}(X = 0|Y = 1) = \mathrm{P}(X = 1|Y = 1) = \mathrm{P}(X = 2|Y = 1) = 1/3$. Use the iterated conditional expectation formula to find $\mathbb{E}(XY)$.

Example 1.24

Find the covariance and correlation of X and Y.

Experiment: Question to be marked

Spaghetti Breaking continued

Measure the length L of a Spaghetto. Break it in half at a randomly point. Measure the length L_1 of the first piece and L_2 of the second piece and compute the realisations u_1, u_2 of the random variables $U_1 = \frac{L_1}{L}$ and $U_2 = \frac{L_2}{L}$.

Question:

- **What is the distribution of U_1? Discrete? Continuous? Mixed?**
- **What was the probability of each question (B1,B2,B3)?**
- **What is the correlation of U_1 and U_2?**

Example 1.26

Compute the correlation of $X \sim \mathrm{U}(-1, 1)$ and $Y = X^2$. Sketch a typical scatter plot of X and Y, e.g. for a sample of size 20. Are X and Y independent?

The conditional variance

Consider random variables X and Y, and the conditional probability distribution of X given $Y = y$. This conditional distribution has a mean, denoted E$(X|Y = y)$, and a variance, var$(X|Y = y)$. We have already shown that the marginal (unconditional) mean E(X) is related to the conditional mean via the formula

$$\mathbb{E}(X) \quad = \quad \mathbb{E}_Y\{\mathbb{E}_{X|Y}(X|Y)\}\,.$$

In the lectures we will obtain a similar result for the relation between the marginal and conditional variances. The result is that

$$\boxed{\mathrm{Var}(X) = \mathbb{E}_Y\{\mathrm{Var}(X|Y)\} + \mathrm{Var}_Y\{\mathbb{E}(X|Y)\}}$$

Example 1.21

Consider two random variables X and Y, where the marginal probabilities of Y are $\mathrm{P}(Y = 0) = 3/4$, $\mathrm{P}(Y = 1) = 1/4$ and the conditional probabilities of X are $\mathrm{P}(X = 1|Y = 0) = \mathrm{P}(X = 2|Y = 0) = 1/2$ and $\mathrm{P}(X = 0|Y = 1) = \mathrm{P}(X = 1|Y = 1) = \mathrm{P}(X = 2|Y = 1) = 1/3$. Use the iterated conditional expectation formula to find $\mathbb{E}(XY)$.

Example 1.27 (1.21 ctd.)

Find the conditional variances of X given $Y = 0, 1$. Compute the marginal variance of X by using the above result.

Example 1.16

Suppose that R and N have a joint distribution in which $R|N$ is $\mathrm{Bin}(N, \pi)$ and N is $\mathrm{Poi}(\lambda)$. Show that R is $\mathrm{Poi}(\pi\lambda)$.

Example 1.28 (1.16 ctd.)

Find the variance of R using the iterated conditional variance formula.

1.5 Standard multivariate distributions

The idea of joint probability distributions extends immediately to more variables, giving general multivariate distributions, i.e. the variables $X_1, \ldots, X_n$ have a joint cumulative distribution function

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathrm{P}(X_1 \le x_1, \ldots, X_n \le x_n)$$

and may have a joint probability mass function

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathrm{P}(X_i = x_i;\ i = 1, \ldots, n)$$

or joint probability density function

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n),$$

so that a function $\phi(X_1, \ldots, X_n)$ has an expectation with respect to this joint distribution etc.

Conditional distributions of a subset of variables given the rest then follow as before; for example, for discrete random variables X_1, X_2, X_3,

$$p_{X_1,X_2|X_3}(x_1, x_2 \mid x_3) = \frac{p_{X_1,X_2,X_3}(x_1, x_2, x_3)}{p_{X3}(x_3)}$$

is the conditional pmf of (X_1, X_2) given $X_3 = x_3$. Similarly, the discrete random variables $X_1, \ldots, X_n$ are mutually independent if and only if

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$$

for all $x_1, x_2, \ldots, x_n$. Mutual independence of $X_1, \ldots, X_n$ implies mutual independence of the events $\{X_1 \in A_1\}, \ldots, \{X_n \in A_n\}$ (*exercise: prove*).

Finally, we say that X_1 and X_2 are *conditionally independent given* X_3 if

$$p_{X_1,X_2|X_3}(x_1, x_2 \mid x_3) = p_{X_1|X_3}(x_1 \mid x_3)\, p_{X_2|X_3}(x_2 \mid x_3)$$

for all x_1, x_2, x_3.
These definitions hold for continuous distributions by replacing the pmf by the pdf.

1.5.1 The multinomial distribution

The multinomial distribution is a generalization of the binomial distribution. Suppose that a sample of size n is drawn (*with replacement*) from a population whose members fall into one of $m+1$ categories. Assume that, for each individual sampled, independently of the rest

$$\text{P(individual is of type } i) = p_i, \qquad i = 1, \ldots, m+1$$

where $\sum_{i=1}^{m+1} p_i = 1$. Let N_i be the number of type i individuals in the sample. Note that, since $N_{m+1} = n - \sum_{i=1}^{m} N_i$, N_{m+1} is determined by $N_1, \ldots, N_m$. We therefore only need to consider the joint distribution of the m random variables $N_1, \ldots, N_m$.

The joint pmf of $N_1, \ldots, N_m$ is given by

$$\mathrm{P}(N_1 = n_1, \ldots, N_m = n_m) = \begin{cases} \frac{n!}{n_1! \ldots n_{m+1}!} p_1^{n_1} \ldots p_{m+1}^{n_{m+1}}, & n_1 + \ldots + n_m \leq n \\ 0 & \text{otherwise.} \end{cases}$$

where $n_{m+1} = n - \sum_{i=1}^{m} n_i$.
This is the multinomial distribution with index n and parameters $p_1, \ldots, p_m$, where $p_{m+1} = 1 - \sum_{i=1}^{m} p_i$ (so p_{m+1} is not a 'free' parameter).

$$\mathrm{P}(N_1 = n_1, \ldots, N_m = n_m) = \begin{cases} \frac{n!}{n_1! \ldots n_{m+1}!} p_1^{n_1} \ldots p_{m+1}^{n_{m+1}}, & n_1 + \ldots + n_m \leq n \\ 0 & \text{otherwise.} \end{cases}$$

To justify the above joint pmf note that we want the probability that the n trials result in exactly n_1 outcomes of the first category, n_2 of the second, . . . , n_{m+1} in the last category. Any specific ordering of these n outcomes has probability $p_1^{n_1} \ldots p_{m+1}^{n_{m+1}}$ by the assumption of independent trials, and there are $\frac{n!}{n_1! \ldots n_{m+1}!}$ such orderings.

$$\mathrm{P}(N_1 = n_1, \ldots, N_m = n_m) = \begin{cases} \frac{n!}{n_1! \ldots n_{m+1}!} p_1^{n_1} \ldots p_{m+1}^{n_{m+1}}, & n_1 + \ldots + n_m \leq n \\ 0 & \text{otherwise.} \end{cases}$$

If $m = 1$ the multinomial distribution is just the binomial distribution, i.e. $N_1 \sim \mathrm{Bin}(n, p_1)$, which has mean np_1 and variance $np_1(1 - p_1)$.

$$\mathrm{P}(N_1 = n_1, \ldots, N_m = n_m) =$$
$$\begin{cases} \frac{n!}{n_1! \ldots n_{m+1}!} p_1^{n_1} \ldots p_{m+1}^{n_{m+1}}, & n_1 + \ldots + n_m \leq n \\ 0 & \text{otherwise.} \end{cases}$$

Example 1.29

Suppose that a bag contains five red, five black and five yellow balls and that three balls are drawn at random with replacement. What is the probability that there is one of each colour? $\square$

Marginal distribution of N_i.

Clearly N_i can be regarded as the number of successes in n independent Bernoulli trials if we define *success* **to be** *individual is of type* i**. Thus N_i has a binomial distribution, $N_i \sim \text{Bin}(n, p_i)$, with mean np_i and variance $np_i(1 - p_i)$.**

Example 1.30

Let N_A, N_B and N_F be the numbers of A grades, B grades and fails respectively amongst a class of 100 students. Suppose that generally 5% of students achieve grade A, 30% grade B and that 5% fail. Write down the joint distribution of N_A, N_B and N_F and find the marginal distribution of N_A. □

Joint distribution of N_i and N_j.

Again we can regard individuals as being one of three types, i, j and *k={not i or j}*. This is the trinomial distribution with probabilities

$$\boxed{\mathrm{P}(N_i = n_i, N_j = n_j) = \begin{cases} \frac{n!}{n_i!n_j!n_k!} p_i^{n_i} p_j^{n_j} p_k^{n_k}, & n_i + n_j \leq n \\ 0 & \text{otherwise} \end{cases}}$$

where $n_k = n - n_i - n_j$ and $p_k = 1 - p_i - p_j$. It is intuitively clear that N_i and N_j are dependent and negatively correlated, since a relatively large value of N_i implies a relatively small value of N_j and conversely.
We show this as follows. First,

$$\mathbb{E}(N_i N_j) = \sum_{\{n_i, n_j \geq 0, n_i + n_j \leq n\}} n_i n_j \frac{n!}{n_i! n_j! n_k!} p_i^{n_i} p_j^{n_j} p_k^{n_k}$$

- **goal: create a pmf that we know sums to one.**

$$\mathbb{E}(N_iN_j)$$

$$= \sum_{\{n_i,n_j\geq 0, n_i+n_j\leq n\}} n_in_j\frac{n!}{n_i!n_j!n_k!}p_i^{n_i}p_j^{n_j}p_k^{n_k}$$

$$= n(n-1)p_ip_j \cdot$$

$$\sum_{\{n_i-1,n_j-1\geq 0, n_i+n_j-2\leq n-2\}} \frac{(n-2)!}{(n_i-1)!(n_j-1)!n_k!}p_i^{n_i-1}p_j^{n_j-1}p_k^{n_k}$$

- goal: create a pmf that we know sums to one.
- Note that we may take $n_i, n_j \geq 1$ in the sum, since if either n_i or n_j is zero then the corresponding term in the sum is zero.

$$
\begin{aligned}
&\mathbb{E}(N_i N_j)\\
&= \sum_{\{n_i, n_j \geq 0, n_i + n_j \leq n\}} n_i n_j \frac{n!}{n_i! n_j! n_k!} p_i^{n_i} p_j^{n_j} p_k^{n_k}\\
&= n(n-1) p_i p_j \cdot\\
&\qquad \sum_{\{n_i - 1, n_j - 1 \geq 0, n_i + n_j - 2 \leq n - 2\}} \frac{(n-2)!}{(n_i - 1)!(n_j - 1)! n_k!} p_i^{n_i - 1} p_j^{n_j - 1} p_k^{n_k}\\
&= n(n-1) p_i p_j \cdot 1\\
&= n(n-1) p_i p_j
\end{aligned}
$$

- goal: create a pmf that we know sums to one.
- Note that we may take $n_i, n_j \geq 1$ in the sum, since if either n_i or n_j is zero then the corresponding term in the sum is zero.
- summation through using known multinomial pmf

Finally

$$\begin{aligned}\mathrm{Cov}(N_i, N_j) &= \mathbb{E}(N_i N_j) - \mathbb{E}(N_i)\mathbb{E}(N_j)\\ &= n(n-1)p_i p_j - (np_i)(np_j) = -np_i p_j\end{aligned}$$

and so

$$\begin{aligned}\mathrm{Corr}(N_i, N_j) &= \frac{-np_i p_j}{\sqrt{np_i(1-p_i)np_j(1-p_j)}}\\ &= -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}}.\end{aligned}$$

Note that corr(N_i, N_j) is negative, as anticipated, and also that it does not depend on n.

Random Question to be Marked

A box of sweets contains 4 sweets with black wrapper, 3 with light blue wrapper and 4 with dark blue wrapper. Three sweets are sampled from the box (with replacement). The three events

- A **Three sweets of three different colours.**
- B **Three sweets of the same colour.**
- C **Two sweets black, one of a different colour.**

are assigned to the exercise questions B1,B2,B3 by audience vote. If none of the three events occurs, the experiment is repeated.

What assignment would you like?

Conditional distribution of N_i given $N_j = n_j$.

Given $N_j = n_j$, there are $n - n_j$ remaining independent Bernoulli trials, each with probability of being type i given by

$$\mathrm{P}(\text{type } i|\text{not type } j) = \frac{\mathrm{P}(\text{type } i)}{\mathrm{P}(\text{not type } j)} = \frac{p_i}{1 - p_j}.$$

Thus, given $N_j = n_j$, N_i has a binomial distribution with index $n - n_j$ and probability $\frac{p_i}{1-p_j}$.

Exercise: Verify this result by using the definition of conditional probability together with the joint distribution of N_i and N_j and the marginal distribution of N_j.

Example 1.31 (1.30 ctd.)

Find the conditional distribution of N_A given $N_F = 10$ and calculate $\mathrm{Corr}(N_A,\ N_F)$.

Remark:

The multinomial distribution can also be used as a model for *contingency tables*. Let X and Y be discrete variable with a number

of I and J different outcomes, respectively. Then, in a trial of size n, N_{ij} will count the number of outcomes where we observe $X = i$ and $Y = j$. The counts N_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, are typically arranged in a contingency table, and from the above considerations we know that their joint distribution is multinomial with parameters n and $p_{ij} = \mathrm{P}(X = i, Y = j)$, $i = 1, \ldots, I$, $j = 1, \ldots, J$.

This leads to the analysis of *categorical data*, for which a question of interest is often 'are the categories independent?', *i.e.* is $p_{ij} = p_i p_j$ for all i, j? Exact significance tests of this hypothesis can be constructed from the multinomial distribution of the entries in the contingency table.

Announcement

Lecturecast video of solution

A video of the fully worked solution of one of the exercise questions is available for you to watch on moodle. You are expected to watch the video before the tutorial.

Linear algebra and quadratic equation revision videos have also been made available on lecturecast (links on Moodle) - please use these, especially if you do not know that $\forall \alpha \in \mathbb{R}\, \forall \boldsymbol{A} \in \mathbb{R}^{n \times n} : \operatorname{det}(\alpha \boldsymbol{A}) = \alpha^{n} \operatorname{det}(\boldsymbol{A})$.

1.5.3 The multivariate normal distribution

The continuous random variables X and Y are said to have a bivariate normal distribution if they have joint probability density function

$$f_{X,Y}(x,y) =$$

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}\right]$$

for $-\infty < x, y < \infty$, where
$-\infty < \mu_X, \mu_Y < \infty; \sigma_X, \sigma_Y > 0; \rho^2 < 1$.
The parameters of this distribution are μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ. As we will see below, these turn out to be the marginal means, variances, and the correlation of X and Y.

The bivariate normal is widely used as a model for many observed phenomena where dependence is expected, *e.g.* height and weight of an individual, length and width of a petal, income and investment returns. Sometimes the data need to be transformed (*e.g.* by taking logs) before using the bivariate normal.

Marginal distributions. In order to simplify the integrations required to find the marginal densities of X and Y, we set

$$\frac{x-\mu_X}{\sigma_X}=u, \qquad \frac{y-\mu_Y}{\sigma_Y}=v.$$

Then, integrating with respect to y, the density of X can be found as

Marginal of a Bivariate Normal

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

- **Marginal Distribution is integral of joint distribution.**

Marginal of a Bivariate Normal

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\
&\qquad \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left\{u^2 - 2\rho uv + v^2\right\}\right]\sigma_Y dv
\end{aligned}
$$

- **Marginal Distribution is integral of joint distribution.**
- **Insert joint bivariate normal density; express in *u*, *v*.**

Marginal of a Bivariate Normal

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\ &\qquad \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left\{u^2 - 2\rho u v + v^2\right\}\right]\sigma_Y dv \end{aligned}$$

- Marginal Distribution is integral of joint distribution.
- Insert joint bivariate normal density; express in *u*, *v*.
- Take the quadratic and linear term in *v* ...

Marginal of a Bivariate Normal

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\
&\qquad \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left\{u^2 - 2\rho u v + v^2\right\}\right]\sigma_Y dv \\
&= \frac{1}{\sigma_X\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \\
&\qquad \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left\{(v-\rho u)^2 + u^2(1-\rho^2)\right\}\right] dv
\end{aligned}$$

- **Marginal Distribution is integral of joint distribution.**
- **Insert joint bivariate normal density; express in *u*, *v*.**
- **Take the quadratic and linear term in *v* ...**
- **... and complete the square.**

Marginal of a Bivariate Normal

$$\begin{aligned} f_X(x) &= \frac{1}{\sigma_X\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \\ &\qquad \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left\{(v-\rho u)^2 + u^2(1-\rho^2)\right\}\right] dv \\ &= \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \\ &\qquad \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left\{(v-\rho u)^2\right\}\right] dv \end{aligned}$$

- **Marginal Distribution is integral of joint distribution.**
- **Insert joint bivariate normal density; express in *u*, *v*.**
- **Take the quadratic and linear term in *v* ...**
- **... and complete the square.**
- **Take term not involving *v* outside integral.**

Marginal of a Bivariate Normal

$$
\begin{aligned}
f_X(x) &= \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \\
&\qquad \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left\{(v-\rho u)^2\right\}\right] dv \\
&= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right\}
\end{aligned}
$$

- **Marginal Distribution is integral of joint distribution.**
- **Insert joint bivariate normal density; express in u, v.**
- **Take the quadratic and linear term in v ...**
- **... and complete the square.**
- **Take term not involving v outside integral.**
- **Density $N(\rho u, 1-\rho^2)$ integrates to 1; re-express in x.**

Marginal of a Bivariate Normal

We have shown

Marginal of a Bivariate Normal

For (X, Y) bivariate normal with parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$, the marginal distribution of X is normal, with mean μ_X and variance σ_X^2.

- **By symmetry in X and Y we get that $Y \sim N(\mu_Y, \sigma_Y^2)$ is the marginal distribution of Y.**
- **It will be shown later that the fifth parameter, ρ also has a simple interpretation, for $\rho = \text{Corr}(X, Y)$.**

Conditional distributions

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

- Conditional density is quotient of joint and marginal.

Conditional distributions

$$
\begin{aligned}
f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\
&= \frac{\sqrt{2\pi}\sigma_Y}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\
&\quad \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \right.\right. \\
&\qquad\qquad \left.\left. + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - (1-\rho^2)\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}\right]
\end{aligned}
$$

- Conditional density is quotient of joint and marginal.
- Insert joint and marginal densities.

Conditional distributions

$$\begin{aligned}
f_{X|Y}(x|y) &= \frac{\sqrt{2\pi}\sigma_Y}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\
&\quad\cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \right.\right. \\
&\qquad\left.\left. + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - (1-\rho^2)\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}\right] \\
&= \frac{\sqrt{2\pi}\sigma_Y}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\
&\quad\cdot \exp\left[-\frac{1}{2\sigma_X^2(1-\rho^2)}\left\{(x-\mu_X)^2 - 2\rho\frac{\sigma_X}{\sigma_Y}(x-\mu_X)(y-\mu_Y)\right.\right. \\
&\qquad\left.\left. + \rho^2\frac{\sigma_X^2}{\sigma_Y^2}(y-\mu_Y)^2\right\}\right]
\end{aligned}$$

- **Conditional density is quotient of joint and marginal.**
- **Insert joint and marginal densities.**
- **Rewrite Expression in $[\cdot]$, factorise σ_X^{-2}.**

Conditional distributions

$$
\begin{aligned}
f_{X|Y}(x|y) &= \frac{\sqrt{2\pi}\sigma_Y}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\
&\quad \cdot \exp\Bigg[-\frac{1}{2\sigma_X^2(1-\rho^2)} \Bigg\{ (x-\mu_X)^2 - 2\rho\frac{\sigma_X}{\sigma_Y}(x-\mu_X)(y-\mu_Y) \\
&\qquad\qquad + \rho^2\frac{\sigma_X^2}{\sigma_Y^2}(y-\mu_Y)^2 \Bigg\} \Bigg] \\
&= \frac{\sqrt{2\pi}\sigma_Y}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\
&\quad \cdot \exp\Bigg[-\frac{1}{2\sigma_X^2(1-\rho^2)} \Bigg\{ (x-\mu_X) - \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y) \Bigg\}^2 \Bigg]
\end{aligned}
$$

- Conditional density is quotient of joint and marginal.
- Insert joint and marginal densities.
- Rewrite Expression in $[\cdot]$, factorise σ_X^{-2}.
- Summarise into one complete square.

Conditional distributions

$$f_{X|Y}(x|y) = \frac{\sqrt{2\pi}\sigma_Y}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2\sigma_X^2(1-\rho^2)}\left\{(x-\mu_X) - \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right\}^2\right]$$

- Conditional density is quotient of joint and marginal.
- Insert joint and marginal densities.
- Rewrite Expression in $[\cdot]$, factorise σ_X^{-2}.
- Summarise into one complete square.
- So $X|Y = y \sim N(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y), \sigma_X^2(1-\rho^2))$.

The role of ρ

Note that knowledge of $Y = y$ reduces the variability of X by a factor $(1 - \rho^2)$. The closer the correlation between X and Y, the smaller the conditional variance becomes. Note also that the conditional mean of X is a *linear function* of y. If y is relatively large then the conditional mean of X is also relatively large if $\rho = \text{Corr}(X, Y) > 0$, or is relatively small if $\rho < 0$.

The role of ρ

Note that knowledge of $Y = y$ reduces the variability of X by a factor $(1 - \rho^2)$. The closer the correlation between X and Y, the smaller the conditional variance becomes. Note also that the conditional mean of X is a *linear function* of y. If y is relatively large then the conditional mean of X is also relatively large if $\rho = \text{Corr}(X, Y) > 0$, or is relatively small if $\rho < 0$.
Suppose $\rho = 0$. Then

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left[-\frac{1}{2}\left\{\left(\frac{x - \mu_X}{\sigma_X}\right)^2 + \left(\frac{y - \mu_Y}{\sigma_Y}\right)^2\right\}\right] \\ &= f_X(x)f_Y(y) \end{aligned}$$

showing that uncorrelated normal variables are independent (remember that this is not true in the general case).

Example 1.32

Let X be the one-year yield of portfolio A and Y be the one-year yield of portfolio B. From past data, the marginal distribution of X is modelled as $N(7, 1)$, whereas the marginal distribution of Y is $N(8, 4)$ (being a more risky portfolio but having a higher average yield). Furthermore, the correlation between X and Y is 0.5. Assuming that X, Y have a bivariate normal distribution, find the conditional distribution of X given that $Y = 9$ and compare this with the marginal distribution of X. Calculate the probability $\mathrm{P}(X > 8|Y = 9)$. □

Linear Algebra Reminder 1

Scalars:

$$\alpha, \beta, \gamma \in \mathbb{R}$$

Vectors:

$$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^{\boldsymbol{n}}, \qquad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = (x_1, x_2, \ldots, x_n)^T$$

Matrices:

$$\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}, \qquad \boldsymbol{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ \vdots & & \ddots & \vdots \\ a_{m,1} & \ldots & \ldots & a_{m,n} \end{pmatrix}$$

Linear Algebra Reminder 2

The usual rules of linear algebra hold:

$$(AB)C = A(BC) \quad A(B+C) = AB + AC \quad (A+B)C = AC + BC$$

$$(\alpha + \beta)x = \alpha x + \beta x \quad \alpha(x+y) = \alpha x + \alpha y$$

Multiplying in component form:

$$(Ax)_i = \sum_{j=1}^{n} a_{i,j} x_j \quad (AB)_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}$$

Transposition

- A^T is the transpose of A with entries $(A^T)_{i,j} = a_{j,i}$, $A^T \in \mathbb{R}^{n\times m}$ if $A \in \mathbb{R}^{m\times n}$.
- Transpose to denote the inner product,

  $$x^T y = \sum_{i=1}^{n} x_i y_i,$$

  works well in matrix expressions, e.g.

  $$x^T A y = (A^T x)^T y$$

  since $(A^T)^T = A$.
- transposition and taking inverses:

  $$(A^T)^{-1} = (A^{-1})^T =: A^{-T}$$

Symmetric Matrices

A matrix is symmetric if $A^T = A$.
Lemma: Covariance matrices are symmetric.
Proof:

Symmetric Matrices

A matrix is symmetric if $A^T = A$.
Lemma: Covariance matrices are symmetric.
Proof:

$$\begin{aligned}\boldsymbol{\Sigma}_{i,j} = \mathrm{Cov}(X_i, X_j) &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] = \mathbb{E}[X_j X_i] - \mathbb{E}[X_j]\mathbb{E}[X_i] \\ &= \mathrm{Cov}(X_j, X_i) = \boldsymbol{\Sigma}_{j,i}\end{aligned}$$

Determinants

$$\det : \mathbb{R}^{n\times n} \to \mathbb{R}$$

- $\det(A) = 0$ if and only if A is non-invertible (i.e. singular).
- $\det(A^{-1}) = \frac{1}{\det(A)}$ if A is invertible
- $\det(A) = \det(A^T)$
- $\det(AB) = \det(A)\det(B)$ for $A, B \in \mathbb{R}^{n\times n}$
- The identity matrix has determinant one:

$$\det \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix} = 1$$

Matrix notation:

Define

$$\boldsymbol{X} = \begin{pmatrix} X \\ Y \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

Here $\boldsymbol{X}$ is a random vector, $\mu = \mathbb{E}(\boldsymbol{X})$ is its mean vector and $\Sigma = \mathrm{Cov}(\boldsymbol{X})$ is the covariance matrix, or dispersion matrix of $\boldsymbol{X}$.

Matrix notation:

Define

$$\boldsymbol{X} = \begin{pmatrix} X \\ Y \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

Here $\boldsymbol{X}$ is a random vector, $\mu = \mathbb{E}(\boldsymbol{X})$ is its mean vector and $\Sigma = \mathrm{Cov}(\boldsymbol{X})$ is the covariance matrix, or dispersion matrix of $\boldsymbol{X}$.
Then

$$|\Sigma| = \sigma_X^2\sigma_Y^2(1-\rho^2)\,, \quad \Sigma^{-1} = \frac{1}{|\Sigma|}\begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix}$$

Also, writing $x = \binom{x}{y}$,

$$\begin{aligned}
(x-\mu)^T\Sigma^{-1}(x-\mu) &= (x-\mu_X, y-\mu_Y)\frac{1}{|\Sigma|}\begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix}\begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix} \\
&= \frac{1}{\sigma_X^2\sigma_Y^2(1-\rho^2)}\{(x-\mu_X)^2\sigma_Y^2 - 2(x-\mu_X)(y-\mu_Y)\rho\sigma_X\sigma_Y \\
&\qquad\qquad + (y-\mu_Y)^2\sigma_X^2\} \\
&= \frac{1}{1-\rho^2}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}.
\end{aligned}$$

It follows that the joint density $f_{\boldsymbol{X}}(x)$ of X, Y given at the beginning of this section can be written as

$$f_X(x) \;=\; \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

on noting that $|2\pi\Sigma|^{1/2} = 2\pi|\Sigma|^{1/2}$. The quantity in $\{\cdot\}$ is a quadratic form in $x - \mu$. Note that the above way of writing the joint density resembles much more the univariate density than the explicit formula given at the beginning of the section.

The usefulness of this matrix representation is that the bivariate normal distribution now extends immediately to a general multivariate form, with joint density as given above with

$$\boldsymbol{X} = (X_1, \ldots, X_k)^T, \qquad \boldsymbol{x} = (x_1, \ldots, x_k)^T, \qquad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^T$$

and

$$(\boldsymbol{\Sigma})_{ij} = \mathrm{cov}(X_i, X_j) = \rho_{ij}\sigma_i\sigma_j$$

Further note that, since $\boldsymbol{\Sigma}$ is $k \times k$, we can write $|2\pi\boldsymbol{\Sigma}|^{1/2} = (2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}$.

The usefulness of this matrix representation is that the bivariate normal distribution now extends immediately to a general multivariate form, with joint density as given above with

$$X = (X_1, \dots, X_k)^T, \qquad x = (x_1, \dots, x_k)^T, \qquad \mu = (\mu_1, \dots, \mu_k)^T$$

and

$$(\mathbf{\Sigma})_{ij} = \mathrm{cov}(X_i, X_j) = \rho_{ij}\sigma_i\sigma_j$$

Further note that, since $\mathbf{\Sigma}$ is $k \times k$, we can write $|2\pi\mathbf{\Sigma}|^{1/2} = (2\pi)^{k/2}|\mathbf{\Sigma}|^{1/2}$.

For this k–dimensional joint distribution, denoted by $MN(mu, Sigma)$ or $N_k(mu, Sigma)$, $\rho_{ij} = \mathrm{Corr}(X_j, X_j)$ and $\mathrm{var}(X_i) = \sigma_i^2$.

It can then be shown that X_i has marginal distribution $N(\mu_i, \sigma_i^2)$, that any two of these variables have a bivariate normal distribution as above, and therefore that the conditional distribution of one variable given the other is also normal.

Example 1.33

Let X_1, X_2, X_3 have a trivariate normal distribution with mean vector (μ_1, μ_2, μ_3) and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}.$$

Show that $f_{X_1,X_2,X_3} = f_{X_1} f_{X_2} f_{X_3}$ and give the marginal distributions of X_1, X_2, and X_3. □

Chapter 2: Transformation of Variables

In this section we will see how to derive the distribution of transformed random variables. This is useful because many statistics applied to data analysis (*e.g.* test statistics) are transformations of the sample variables.

2.1 Univariate case
Suppose that we have a sample space Ω, a probability function P on Ω, a random variable $X : \Omega \to \mathbb{R}$, and a function $\phi : \mathbb{R} \to \mathbb{R}$.
Recall from §1.2 : $Y = \phi(X) : \Omega \to \mathbb{R}$ is defined by $Y(\omega) = \phi(X)(\omega) = \phi(X(\omega))$. Since $Y = \phi(X)$ is a random variable it also has a probability distribution, which can be determined either directly from P or via the distribution of X.

Discrete case:

$$\begin{aligned}
\mathrm{P}(Y = y) &= \mathrm{P}(\{\omega : \phi(X(\omega)) = y\}) = \sum_{\{\omega : \phi(X(\omega)) = y\}} \mathrm{P}(\{\omega\}) \\
&= \sum_{\{x : \phi(x) = y\}} \mathrm{P}(\{\omega : X(\omega) = x\}) \\
&= \sum_{\{x : \phi(x) = y\}} p_X(x).
\end{aligned}$$

Discrete case:

$$\begin{aligned}
\mathrm{P}(Y=y) &= \mathrm{P}(\{\omega : \phi(X(\omega))=y\}) = \sum_{\{\omega:\phi(X(\omega))=y\}} \mathrm{P}(\{\omega\}) \\
&= \sum_{\{x:\phi(x)=y\}} \mathrm{P}(\{\omega : X(\omega)=x\}) \\
&= \sum_{\{x:\phi(x)=y\}} p_X(x).
\end{aligned}$$

So, for example

$$\begin{aligned}
\mathrm{E}\{Y\} &= \sum_{\omega} \phi(X(\omega))\mathrm{P}(\{\omega\}) && \textbf{with respect to P on } \Omega \\
&= \sum_{x} \phi(x)p_X(x) && \textbf{with respect to distribution of } X \\
&= \sum_{y} y p_{\phi(X)}(y) && \textbf{with respect to distribution of } \phi(X)
\end{aligned}$$

Examples

Example 2.1
Consider two independent throws of a fair die. Let X be the sum of the numbers that show up. Give the distribution of X.
Now consider the transformation $Y = (X - 7)^2$. Derive the distribution of Y. $\square$

transformation of univariate continuous random variable: ϕ increasing

In general, suppose that $Y = \phi(X)$ where ϕ is a strictly increasing and differentiable function. Then,

$$F_Y(y) = \mathrm{P}(\phi(X) \le y) = \mathrm{P}(X \le \phi^{-1}(y)) = F_X(\phi^{-1}(y)).$$

transformation of univariate continuous random variable: ϕ increasing

In general, suppose that $Y = \phi(X)$ where ϕ is a strictly increasing and differentiable function. Then,

$$F_Y(y) = \mathrm{P}(\phi(X) \leq y) = \mathrm{P}(X \leq \phi^{-1}(y)) = F_X(\phi^{-1}(y)).$$

Then, differentiating, Y has density

$$f_Y(y) = f_X(\phi^{-1}(y))\, \frac{d}{dy}\phi^{-1}(y) = f_X(x)\frac{dx}{dy}\bigg|_{x=\phi^{-1}(y)},$$

where the index $x = \phi^{-1}(y)$ means that any x in the formula has to be replaced by the inverse $\phi^{-1}(y)$ because $f_Y(y)$ is a function of y.

transformation of univariate continuous random variable: ϕ decreasing

Similarly, if ϕ is decreasing then

$$F_Y(y) = \mathrm{P}(\phi(X) \le y) = \mathrm{P}(X \ge \phi^{-1}(y)) = 1 - F_X(\phi^{-1}(y))$$

so that

$$f_Y(y) = -f_X(x)\frac{dx}{dy}\bigg|_{x=\phi^{-1}(y)}.$$

transformation of univariate continuous random variable: ϕ decreasing

Similarly, if ϕ is decreasing then

$$F_Y(y) = \mathrm{P}(\phi(X) \leq y) = \mathrm{P}(X \geq \phi^{-1}(y)) = 1 - F_X(\phi^{-1}(y))$$

so that

$$f_Y(y) = -f_X(x) \frac{dx}{dy}\bigg|_{x=\phi^{-1}(y)} .$$

In the first case $dy/dx = d\phi(x)/dx$ is positive (since ϕ is increasing), in the second it is negative (since ϕ is decreasing) so either way the transformation formula is

$$\boxed{f_Y(y) = f_X(x) \left|\frac{dx}{dy}\right|_{x=\phi^{-1}(y)}}$$

We can check that the right-hand side of the above formula is a valid pdf as follows. Recall that $\int_{-\infty}^{\infty} f_X(x)dx = 1$. Changing variable to $y = \phi(x)$ we have, for ϕ increasing,

$$1 = \int_{-\infty}^{\infty} \left\{ f_X(x)\frac{dx}{dy} \right\}_{x=\phi^{-1}(y)} dy$$

so that $f_X(x)\left|\frac{dx}{dy}\right|$ is a valid pdf. Similarly for ϕ decreasing.

We can check that the right-hand side of the above formula is a valid pdf as follows. Recall that $\int_{-\infty}^{\infty} f_X(x)dx = 1$. Changing variable to $y = \phi(x)$ we have, for ϕ increasing,

$$1 = \int_{-\infty}^{\infty} \left\{ f_X(x) \frac{dx}{dy} \right\}_{x=\phi^{-1}(y)} dy$$

so that $f_X(x) \left| \frac{dx}{dy} \right|$ is a valid pdf. Similarly for ϕ decreasing.

Example 2.2

Consider X $\sim$Uniform$[-\frac{\pi}{2}, \frac{\pi}{2}]$, i.e.

$$f_X(x) = \begin{cases} \frac{1}{\pi} & -\frac{\pi}{2} \leq x \leq \frac{\pi}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Derive the density of $Y = \tan(X)$. □

Many-to-one functions

When ϕ is a *many-to-one* function we use the generalised formula $f_Y(y) = \sum f_X(x)\left|\frac{dx}{dy}\right|$, where the summation is over the set $\{x : h(x) = y\}$. That is, we add up the contributions to the density at y from all x values which map to y.

Example 2.3

Suppose that $f_X(x) = 2x$ on $(0, 1)$ and let $Y = (X - \frac{1}{2})^2$. Obtain the pdf of Y. □

2.2 Bivariate case

For the bivariate case we consider two random variables X, Y with joint density $f_{X,Y}(x, y)$. What is the joint density of transformations $U = u(X, Y)$, $V = v(X, Y)$ where $u(\cdot, \cdot)$ and $v(\cdot, \cdot)$ are functions from $\mathbb{R}^2$ to $\mathbb{R}$, such as the ratio X/Y or the sum $X + Y$?

In order to use the following generalisation of the method of §2.1, we need to assume that u, v are such that each pair (x, y) defines a unique (u, v) and conversely, so that $u = u(x, y)$ and $v = v(x, y)$ are differentiable and invertible. The formula that gives the joint density of U, V is similar to the univariate case but the derivative, as we used it above, now has to be replaced by the *Jacobian* $J(u, v)$ of this transformation.

The result is that $U = u(X, Y)$, $V = v(X, Y)$ have joint density

$$\boxed{f_{U,V}(u, v) = f_{X,Y}(x(u, v), y(u, v))|J(x, y)|_{\substack{x=x(u,v)\\ y=y(u,v)}}}$$

Again, the index $\substack{x=x(u,v)\\ y=y(u,v)}$ means that the x, y have to be replaced by the suitable transformations involving u, v only.
But how do we get the Jacobian $J(x, y)$? It is actually the determinant of the *matrix of partial derivatives* :

$$J(x, y) = \det\left(\frac{\partial(x, y)}{\partial(u, v)}\right) = \det\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}$$

We finally take its absolute value, $|J(x, y)|$.

There are two ways of computing this:
(1) Obtain the inverse transformation $x = x(u, v), y = y(u, v)$, compute the matrix of partial derivatives $\partial(x, y)/\partial(u, v)$ and then its determinant and absolute value.
(2) Alternatively find the determinant $J(u, v)$ from the matrix of partial derivatives of (u, v) with respect to (x, y) and then its absolute value and invert this.

There are two ways of computing this:
(1) Obtain the inverse transformation $x = x(u, v), y = y(u, v)$, compute the matrix of partial derivatives $\partial(x, y)/\partial(u, v)$ and then its determinant and absolute value.
(2) Alternatively find the determinant $J(u, v)$ from the matrix of partial derivatives of (u, v) with respect to (x, y) and then its absolute value and invert this.
The two methods are equivalent since

$$\frac{\partial(x, y)}{\partial(u, v)} = \left\{ \frac{\partial(u, v)}{\partial(x, y)} \right\}^{-1}$$

Which way to choose in a specific case will depend on which functions are easier to derive. But note that the inverse transformations $x = x(u, v)$ and $y = y(u, v)$ are required anyway so that the first approach is often preferable.

Example 2.4

Let X and Y be two independent exponential variables with $X \sim$ Exp(λ) and $Y \sim$ Exp(μ). Find the distribution of $U = X/Y$ □

Example 2.5

Consider two independent and identically distributed random variables X and Y having a uniform distribution on $[0, 2]$. Derive the joint density of $Z = X/Y$ and $W = Y$, stating the area where this density is positive. Are Z and W independent?
Obtain the marginal density of $Z = X/Y$. □

Sums of random variables

The distribution of a sum $Z = X + Y$ of two (not necessarily independent) random variables X and Y can be derived directly as follows.

In the discrete case note that the marginal distribution of Z is

$$\mathrm{P}(Z = z) = \sum_x \mathrm{P}(X = x, Z = z) = \sum_x \mathrm{P}(X = x, Y = z - x)$$

Sums of random variables

The distribution of a sum $Z = X + Y$ of two (not necessarily independent) random variables X and Y can be derived directly as follows.

In the discrete case note that the marginal distribution of Z is

$$\mathrm{P}(Z = z) = \sum_x \mathrm{P}(X = x, Z = z) = \sum_x \mathrm{P}(X = x, Y = z - x)$$

That is,

$$\boxed{p_Z(z) = \sum_x p_{X,Y}(x, z - x)}$$

Analogously, in the continuous case we get

$$\boxed{f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z - x)\, dx}$$

Example 2.6

Let X and Y be two positive random variables with joint pdf

$$f_{X,Y}(x,y) = xye^{-(x+y)}\,, \;\; x, y > 0\,.$$

Derive and name the distribution of their sum $Z = X + Y$. $\square$

2.3 Multivariate case

The ideas of §2.2 extend in a straightforward way to the case of more than two variables. The general problem is to find the distribution of $Y = \phi(X)$, where Y is $s \times 1$ and X is $r \times 1$, from the known distribution of X. Here X is the random *vector*

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_r \end{pmatrix} .$$

Discrete case.

$p_Y(y) = \sum p_X(x)$, where the summation is over the set $\{x : \phi(x) = y\}$. That is, we just add up the probabilities of all x-values that give $\phi(x) = y$.

Discrete case.

$p_Y(y) = \sum p_X(x)$, where the summation is over the set $\{x : \phi(x) = y\}$. That is, we just add up the probabilities of all x-values that give $\phi(x) = y$.

Continuous case.

Case (i): ϕ is a one-to-one transformation (so that $s = r$). Then the rule is

$$\boxed{f_Y(y) = f_X(x(y))\,|J(x)|_{x=x(y)}}$$

where $J(x) = \det\left(\frac{dx}{dy}\right)$ is the Jacobian of transformation.

Here $\frac{dx}{dy}$ is the matrix of partial derivatives $\left(\frac{dx}{dy}\right)_{ij} = \frac{\partial x_i}{\partial y_j}$.

Case (ii): $s < r$. First transform the s-vector Y to the r-vector Y', where $Y'_i = Y_i$, $i = 1, \ldots, s$, and the other $r - s$ random variables Y'_i , $i = s + 1, \ldots, r$, are chosen for convenience. Now find the density of Y' as in case (i) and then integrate out $Y'_{s+1}, \ldots, Y'_r$ to obtain the marginal density of Y, as required. (*c.f.* Examples 2.6 & 2.7 in the bivariate case.)

Case (iii): $s = r$ but $\phi(\cdot)$ is not monotonic. Then there will generally be more than one value of x corresponding to a given y and we need to add the probability contributions from all relevant xs.

Multivariate Transformation: Example

Example 2.7 (linear transformation)

Suppose that $Y = AX$, where A is an $r \times r$ nonsingular matrix. Then $f_Y(y) = f_X(A^{-1}y)|\det(A)|^{-1}$. □

2.4 Approximation of moments

Sometimes we may not need the complete probability distribution of $\phi(X)$, but just the first two moments. Recall that $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$, so the relation $\mathbb{E}\{\phi(X)\} = \phi(\mathbb{E}(X))$ is true whenever ϕ is a linear function. However, in general if $Y = \phi(X)$ it will not be true that $\mathrm{E}(Y)$ = $\mathrm{E}\{\phi(X)\}$ = $\int \phi(x)f_x(x)dx$ is the same as $\phi(\mathrm{E}(X))$ = $\phi(\int xf_x(x)dx)$, (or equivalent summations if X is discrete). (*c.f.* Example 1.1 ctd. in §1.2.)

To find moments of Y we can use the distribution of X, as above. However, the sums or integrals involved may be analytically intractable. In practice an approximate answer may be sufficient.

Intuitively, if X has mean μ_X and X is not very variable, then we would expect $\mathrm{E}(Y)$ to be quite close to $\phi(\mu_X)$. How good is this approximation?

Suppose that $\phi(x)$ is a continuous function of x for which the following Taylor expansion about μ_X exists (which requires the existence of the derivatives of ϕ):

$$\phi(x) = \phi(\mu_X) + (x - \mu_X)\phi'(\mu_X) + \frac{1}{2}(x - \mu_X)^2\phi''(\mu_X) + \ldots$$

Suppose that $\phi(x)$ is a continuous function of x for which the following Taylor expansion about μ_X exists (which requires the existence of the derivatives of ϕ):

$$\phi(x) = \phi(\mu_X) + (x - \mu_X)\phi'(\mu_X) + \frac{1}{2}\,(x - \mu_X)^2\phi''(\mu_X) + \ldots$$

Replacing x by X and taking expectations (or, equivalently, multiplying both sides of the above equation by $f_x(x)$ and integrating over x) term by term, we get

$$\begin{aligned}\mathbb{E}\{\phi(X)\} = \phi(\mu_X) + \phi'(\mu_X)\ &\underline{\mathrm{E}(X - \mu_X)}_{=0} + \\ &\frac{1}{2}\,\phi''(\mu_X)\underline{\mathrm{E}\ \{(X - \mu_X)^2\}}_{\sigma_X^2} + \ldots\end{aligned}$$

so that

$$\boxed{\mathrm{E}(Y) \simeq \phi(\mu_X) + \frac{1}{2}\,\phi''(\mu_X)\sigma_X^2}$$

A usually sufficiently good approximate formula for the variance is based on a first order approximation yielding

$$\begin{aligned}\textbf{Var}\Big\{\phi(X)\Big\} &= \mathbb{E}\{\phi(X) - \mathrm{E}(\phi(X))\}^2 \\ &\approx \mathbb{E}\{\phi(X) - \phi(\mu_X)\}^2 \approx \mathbb{E}\{(X - \mu_X)^2(\phi'(\mu_X))^2\} \\ &= \{\phi'(\mu_X)\}^2\mathbb{E}\{(X - \mu_X)^2\}\,,\end{aligned}$$

where we have used the approximation $\mathrm{E}(\phi(X)) \simeq \phi(\mu_X)$. Therefore

$$\boxed{\mathrm{var}(Y) \simeq \{\phi'(\mu_X)\}^2\sigma_X^2}$$

A usually sufficiently good approximate formula for the variance is based on a first order approximation yielding

$$\begin{aligned}
\textbf{Var}\big\{\phi(X)\big\} &= \mathbb{E}\{\phi(X) - \mathrm{E}(\phi(X))\}^2 \\
&\approx \mathbb{E}\{\phi(X) - \phi(\mu_X)\}^2 \approx \mathbb{E}\{(X - \mu_X)^2(\phi'(\mu_X))^2\} \\
&= \{\phi'(\mu_X)\}^2\mathbb{E}\{(X - \mu_X)^2\}\,,
\end{aligned}$$

where we have used the approximation $\mathrm{E}(\phi(X)) \simeq \phi(\mu_X)$. Therefore

$$\boxed{\mathrm{var}(Y) \simeq \{\phi'(\mu_X)\}^2\sigma_X^2}$$

Example 2.8

Consider a Poisson variable X $\sim$Poi(μ). Find approximations to the expectation and variance of $Y = \sqrt{X}$. □

2.5 Order Statistics

Order statistics are a special kind of transformation of the sample variables. Their joint and marginal distributions can be derived by combinatorial considerations. Suppose that $X_1, \ldots, X_n$ are independent with common density f_X. Denote the ordered values by $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. What is the distribution F_r of $X_{(r)}$? In particular, $X_{(n)} = \max\,(X_1, \ldots, X_n)$ is the sample maximum and $X_{(1)} = \min\,(X_1, \ldots, X_n)$ is the sample minimum.

To find the distribution of $X_{(n)}$, note that $\{X_{(n)} \leq x\}$ and $\{\text{all } X_i \leq x\}$ are the same event – and so have the same probability!

Therefore the distribution function of $X_{(n)}$ is

$$\begin{aligned} F_n(x) = \mathrm{P}(X_{(n)} \leq x) &= \mathrm{P}(\text{all } X_i \leq x) \\ &= \mathrm{P}(X_1 \leq x, X_2 \leq x, \ldots, X \leq x_n) \\ &= \{F_X(x)\}^n \end{aligned}$$

since the X_i are independent with the same distribution function F_X. Thus

$$\boxed{F_n(x) = \{F_X(x)\}^n}$$

To find the distribution of $X_{(n)}$, note that $\{X_{(n)} \le x\}$ and $\{\text{all } X_i \le x\}$ are the same event – and so have the same probability!

Therefore the distribution function of $X_{(n)}$ is

$$\begin{aligned} F_n(x) = \mathrm{P}(X_{(n)} \le x) &= \mathrm{P}(\text{all } X_i \le x) \\ &= \mathrm{P}(X_1 \le x, X_2 \le x, \ldots, X \le x_n) \\ &= \{F_X(x)\}^n \end{aligned}$$

since the X_i are independent with the same distribution function F_X. Thus

$$\boxed{F_n(x) = \{F_X(x)\}^n}$$

Furthermore, differentiating this expression we see that the density f_n of $X_{(n)}$ is

$$\boxed{f_n(x) = n\{F_X(x)\}^{n-1} f_X(x)}$$

Using a similar argument for $X_{(1)} = \min(X_1, \ldots, X_n)$ we see that

$$\begin{aligned} F_1(x) &= \mathrm{P}(X_{(1)} \leq x) = \mathrm{P}(\text{at least one } X_i \leq x) \\ &= 1 - \mathrm{P}(\text{all } X_i > x) = 1 - \{1 - F_X(x)\}^n \end{aligned}$$

so the distribution function of $X_{(1)}$ is

$$\boxed{F_1(x) = 1 - \{1 - F_X(x)\}^n}$$

and, differentiating, the pdf f_1 of $X_{(1)}$ is

$$\boxed{f_1(x) = n\{1 - F_X(x)\}^{n-1} f_X(x)}$$

Consider next the situation for general $1 \leq r \leq n$. For dx sufficiently small we have

$$\mathrm{P}(x < X_{(r)} \leq x + dx) = \mathrm{P}\left(\begin{array}{l} r-1 \text{ values } X_i \text{ such that } X_i \leq x, \text{ and} \\ \text{one value in } (x, x+dx]\,, \text{ and} \\ n-r \text{ values such that } \; X_i > x + dx \end{array}\right)$$

Consider next the situation for general $1 \leq r \leq n$. For dx sufficiently small we have

$$\mathrm{P}(x < X_{(r)} \leq x + dx) = \mathrm{P}\left(\begin{array}{l} r-1 \text{ values } X_i \text{ such that } X_i \leq x, \text{ and} \\ \text{one value in } (x, x+dx]\text{, and} \\ n-r \text{ values such that } X_i > x + dx \end{array}\right)$$

$$= \underbrace{\frac{n!}{(r-1)!(n-r)!}}_{\text{no. of ways of ordering the } r-1, 1 \text{ and } n-r \text{ values}} \{F_X(x)\}^{r-1} f_X(x) dx \{1 - F_X(x+dx)\}^{n-r}$$

Recalling that $f_r(x) = \lim_{dx \to 0} \mathrm{P}(x < X_{(r)} \leq x + dx)/dx$, dividing both sides of the above expression by dx and letting $dx \to 0$ we obtain the density function of the rth order statistic $X_{(r)}$ as

$$\boxed{f_r(x) = \frac{n!}{(r-1)!(n-r)!}\{F_X(x)\}^{r-1}\{1 - F_X(x)\}^{n-r} f_X(x)}$$

Exercise: Show that this formula gives the previous densities when $r = n$ and $r = 1$.

A hint of extreme value theory

Example 2.9

A village is protected from a river by a dike of height *h*. The maximum water levels X_i reached by the river in subsequent years $i = 1, 2, 3, \ldots$ are modelled as independent following an exponential distribution with mean $\lambda^{-1} = 10$. What is the probability that the village will be flooded (in statistical language this would be called a "threshold exceedance") at least once in the next 100 years? How high does the dike need to be to make this probability smaller than 0.5?