### Contents

1. **Events and Probabilities**  
   1.1 Probability Spaces  
   1.2 A Few Simple Theorems  
   1.3 Discrete Probability Spaces  
   1.4 Interpretation of Probability  
   1.5 Conditional Probability and Independence  
   1.6 Two Important Theorems  
   1.7 The 3-door (or Monty Hall) problem  
   1.8 The Birthday Problem

2. **Discrete Random Variables**  
   2.1 Some Common Discrete Random Variables  
   2.2 Expectation  
   2.3 Expectation of Functions of $X$  
   2.4 Conditional Expectation  
   2.5 Properties of Expectation  
   2.6 Variance

3. **Continuous Random Variables**  
   3.1 Properties of the pdf  
   3.2 Examples of Continuous Random Variables  
   3.3 Expectation and Variance  
   3.4 Properties of Expectations  
   3.5 Functions of a Random Variable

4. **Multivariate Random Variables**  
   4.1 Multivariate Discrete Distributions  
   4.2 Expectation  
   4.3 Independence  
   4.4 Generalising to $n$ Random Variables  
   4.5 Multivariate Distribution Functions  
   4.6 Continuous Random Variables  
   4.7 Marginal Density Functions and Independence  
   4.8 Conditional Density Functions  
   4.9 Expectations of Continuous Random Variables  
   4.10 Variance and Covariance  
   4.11 Sums of Random Variables  
   4.12 The Multivariate Gaussian Distribution

5. **Moment and Characteristic Functions**  
   5.1 Moment Generating Functions  
   5.2 Characteristic Functions

6. **Two Probability Theorems**  
   6.1 The Law of Averages  
   6.2 The Central Limit Theorem

© Simon Wilson, 2012
1 Events and Probabilities

HANDOUT: outline

Probability was developed by gamblers in the 17th century. However, its rigorous development had to wait for the start of the 20th century.

Probability starts with the idea of an experiment or trial:

**Definition:** An experiment is any course of action whose consequence is not predetermined.

When we have an experiment, there is more than one possible outcome, and we cannot determine which will happen.

**Definition:** The set of all possible outcomes of an experiment is called the sample space. It is usually denoted \( \Omega \).

Example: experiment = “throw a die”. Then \( \Omega = \{1, 2, 3, 4, 5, 6\} \).

**Definition:** An event is just a subset of the sample space that interests us. The members of \( \Omega \) are also events, called elementary events. The set of events is the event space and is denoted \( \mathcal{F} \).

For simple things such as a die, \( \mathcal{F} \) is usually the power set. However, in more complicated situations, it may be considerably less than the set of all subsets (this is to do with \( \sigma \)-algebra, which we do not go into here). However, again for reasons that we do not go into, \( \mathcal{F} \) must satisfy:

- \( \mathcal{F} \) is non-empty;
- \( A \in \mathcal{F} \Rightarrow \Omega - A \in \mathcal{F} \);
- \( A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_i A_i \in \mathcal{F} \)

Usually, we think that some outcomes of an experiment are more likely to occur than others. Probability is a way to quantify the likelihood of each event in \( \mathcal{F} \) occurring. It is a number between 0 and 1, and so is a mapping from \( \mathcal{F} \) to \([0, 1]\). We give it certain properties:

**Definition:** A probability measure on \((\Omega, \mathcal{F})\) is a mapping \( P : \mathcal{F} \to [0, 1] \) that satisfies:

- \( P(\Omega) = 1 \) and \( P(\emptyset) = 0 \);
- If \( A_1, A_2, \ldots \) are disjoint events then:

\[
P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)
\]

Example: throwing a die: \( A_1 = \) get a 1 or 2, \( A_2 = \) get a 5; \( P(A_1 \cup A_2) = P(A_1) + P(A_2) = P(1) + P(2) + P(5) = 3/6 = 0.5 \).

1.1 Probability Spaces

For any experiment, there is a mathematical object called a probability space that describes it.

**Definition:** A probability space is a triple \((\Omega, \mathcal{F}, P)\) where

- \( \Omega \) is a set;
- \( \mathcal{F} \) is an event space of subsets of \( \Omega \);
- \( P \) is a probability measure on \((\Omega, \mathcal{F})\)
1.2 A Few Simple Theorems

Theorem: $A, B \in \mathcal{F} \Rightarrow A - B \in \mathcal{F}$.

Proof: $\Omega - (A - B) = (\Omega - A) \cup B$, which is an event by properties of $\mathcal{F}$.

Theorem: $A_1, A_2, \ldots, \in \mathcal{F} \Rightarrow \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$

Proof: This is just de Morgan’s law. $\Omega - \bigcap_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} (\Omega - A_i)$. $A_i \in \mathcal{F} \Rightarrow \Omega - A_i \in \mathcal{F} \Rightarrow \Omega - \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$ (complement of $\Omega - \bigcap_{i=1}^{\infty} A_i$).

Theorem: For $A \in \mathcal{F}$, $P(\Omega - A) = 1 - P(A)$ (i.e. $P(\text{not } A) = 1 - P(A)$).

Proof: $\Omega - A \in \mathcal{F}$, and disjoint with $A$, thus $P(A) + P(\Omega - A) = P(\Omega) = 1$.

Theorem: For $A, B \in \mathcal{F}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: Draw a Venn diagram. $A - B$ and $A \cap B$ are disjoint thus $P(A) = P((A - B) \cup (A \cap B)) = P(A - B) + P(A \cap B)$. Similarly $P(B) = P((B - A) \cup (B \cap A)) = P(B - A) + P(B \cap A)$; adding, we get $P(A) + P(B) = P(A - B) + P(B - A) + 2P(A \cap B) = \left[P((A - B) \cup (A \cap B)) \cup (B - A)\right] + P(A \cap B) = P(A) + P(B) - P(A \cap B)$. Useful hint: it is usually best when trying to prove things like this to draw a Venn diagram and put things in terms of disjoint sets, then use the rule about unions of disjoint sets.

1.3 Discrete Probability Spaces

When $\Omega$ is countable, we say that the probability space is discrete. In this case, $\mathcal{F}$ is almost always the set of all subsets of $\Omega$. We always assume that is the case.

If that is so, then the elementary events $\omega \in \Omega$ are in $\mathcal{F}$ so we can define $P(\{\omega\}) = P(\omega)$. Then, for any event $A$, we have $P(A) = P(\bigcup_{\omega \in A} \{\omega\}) = \sum_{\omega \in A} P(\omega)$.

So, for discrete spaces, $P(A)$ is always the sum of probabilities of the outcomes in $A$.

Example: Let $\Omega = \{1, \ldots, N\}$ and suppose $P(i) = 1/N$ i.e. all outcomes are equally likely. Then $P(A) = \frac{|A|}{N}$. This is very common: (lottery, picking a card from a pack, etc.) We look at an example of this in the tutorials.

1.4 Interpretation of Probability

1.5 Conditional Probability and Independence

I take a pack of cards and you pick one: $P(\heartsuit) = 1/52$.

Suppose I pick a card and tell you “it is a heart”. Now what is $P(\heartsuit)$?

The extra information “it is a heart” has affected the prob. In fact I have told you that an event “it is a heart” has occurred, and it has affected the probability of the event “it is $\spadesuit$”.

In general, if $A, B \in \mathcal{F}$, and we want $P(A)$ and are given that $B$ has occurred, then $A$ will only occur if $A \cap B$ occurs and $P(A)$ changes to something we call the probability of $A$ given $B$ or the conditional probability.

Definition: If $A, B \in \mathcal{F}$ and $P(B) > 0$ then the conditional probability of $A$ given $B$ is denoted $P(A \mid B)$ and is defined to be $P(A \mid B) = P(A \cap B) / P(B)$.

Example: Consider two urns. Urn 1 has 3 white and 2 black balls, urn 2 has 1 white and 6 black balls. An experiment consists of tossing a fair coin. If it lands H then a ball is picked from urn 1, else a ball is picked from urn 2:

1. What is $\Omega$?

2. What is $P(\text{W ball} \mid \text{H thrown})$?

3. What is $P(\text{W ball} \mid \text{T thrown})$?
**Definition:** Two events $A$, $B$ are independent if the occurrence of one does not affect the prob. of the other; or if $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.

From the definition of $P(A \cap B)$, we see that $A$ and $B$ are independent if $P(A \cap B) = P(A)P(B)$.

More generally, a family of events $\{A_i : i \in I\}$ are independent if for all finite subsets $J \subseteq I$:

$$P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i).$$

The family is pairwise independent if the above holds for $|J| = 2$.

**Example:** Take 3 events $A$, $B$ and $C$.

1. What must hold for these 3 events to be independent?

2. Consider throwing a 4 sided die, where each outcome is equally likely. Let $A = \{1, 2\}$, $B = \{1, 3\}$ and $C = \{1, 4\}$. Show that these are pairwise independent but not independent.

**1.6 Two Important Theorems**

Two very useful laws of probability are the partition law and Bayes law:

**Theorem:** (Partition law) Let $B_1, B_2, \ldots, \in \mathcal{F}$ be a partition of $\Omega$. Then for any $A \in \mathcal{F}$:

$$P(A) = \sum_i P(A \mid B_i) P(B_i).$$

**Proof:** $P(A) = P(A \cap (\bigcup_i B_i)) = P(\bigcup_i (A \cap B_i)) = \sum_i P(A \cap B_i) = \sum_i P(A \mid B_i) P(B_i)$.

Note: a special case is if the partition sets are just each outcome, so you have $P(A) = \sum_{x \in \Omega} P(A \mid x)P(x)$.

**Example:** Tomorrow’s weather is either rain or snow but not both, with $P(\text{rain})=0.4$ and $P(\text{snow})=0.6$. If it rains, then I am late for the class with prob. 0.2; if it snows with prob. 0.6. What is $P(\text{l}

**Theorem:** (Bayes’ law). Let $B_1, B_2, \ldots, \in \mathcal{F}$ be a partition of $\Omega$. Then for any $A \in \mathcal{F}$:

$$P(B_i \mid A) = \frac{P(A \mid B_i) P(B_i)}{\sum_j P(A \mid B_j) P(B_j)}.$$

**Proof:** $P(B_i \mid A) = P(B_i \cap A) / P(A) = P(A \mid B_i)P(B_i) / P(A) =$result (by partition law).

Again, a special case is for any $x \in \Omega$, $P(x \mid A) = P(A \mid x) P(x) / \sum_{y \in \Omega} P(A \mid y) P(y)$

**1.7 The 3-door (or Monty Hall) problem**

There are 3 identical doors. Behind 1 door is a nice prize (say a car). Behind the other two is something not very desirable (nothing, or a donkey). You play the following game:

- You will select a door and of course want to win the nice prize. Your host is Monty Hall who knows what lies behind the doors but will not tell you!

- Once you have picked a door then Monty will open one of the other doors that does not have the prize behind it. Note that there will always be at least one such door;

- You now have a choice. You can:

  1. Stay with your original choice of door;
2. Switch to the other unopened door.

The question is: does it make any difference to your chances of winning the nice prize by switching or not?

Let the doors be labelled \( W, L_1, L_2 \). The probabilities you pick each is 1/3. If you initially pick \( W \) then Monty has a choice of opening \( L_1 \) or \( L_2 \). Assume wlog that he opens \( L_1 \) in this case. However if you pick \( L_1/L_2 \) then Monty can only open \( L_2/L_1 \) (no choice). If you decide to switch then there’s only one remaining door to switch to.

The outcomes of this experiment are a pair \((x, y)\) where \( x \) is your initial choice and \( y \) is the door that you finally choose after switching or not. Consider the two options:

**Switch**: \( \Omega_{\text{switch}} = \{(W, L_2), (L_1, W), (L_2, W)\} \). These occur with probabilities 1/3 each. The event “win”=\{(L_1, W), (L_2, W)\} and \( P(\text{win}) = 1/3 + 1/3 = 2/3 \);

**Stay**: \( \Omega_{\text{stay}} = \{(W, W), (L_1, L_1), (L_2, L_2)\} \). These occur with probabilities 1/3 each. The event “win”=\{(W, W)\} and \( P(\text{win}) = 1/3 \);

So it is better to switch!

Another (and simpler) way to see what is happening is to draw a table of the outcomes of the game according to which door you intially pick: so when switching, you win iff you initially pick \( L_1 \) or \( L_2 \)

<table>
<thead>
<tr>
<th>Pick first</th>
<th>Switch</th>
<th>Stay</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W )</td>
<td>( L_2 )</td>
<td>( W )</td>
</tr>
<tr>
<td>( L_1 )</td>
<td>( W )</td>
<td>( L_1 )</td>
</tr>
<tr>
<td>( L_2 )</td>
<td>( W )</td>
<td>( L_2 )</td>
</tr>
</tbody>
</table>

which has probability 2/3. When staying, you win iff you initially pick \( W \) with probability 1/3.

### 1.8 The Birthday Problem

There are \( N \) people in a room. What is the chance that none of them share a birthday? What do you think for this class (ignoring 29th February for now)? Is it more than 25%? More than 50%? More than 75%?

We can figure out the probability as follows. Assume 365 days and that each date is equally likely to be a birthday (not quite true but we’ll ignore that for now). Let \( p_N = P(\text{N people do not have a birthday in common}) \).

Label the people in the room 1, 2, \ldots, \( N \). We’ll use an induction argument.

- For \( N = 2, p_2 = P(\text{person 2 does not have the same birthday as person 1}) = 364/365 \);

- We have:

  \[
  p_N = P(\text{person } N \text{ does not have the same birthday as persons } 1, \ldots, N - 1 \text{ AND persons } 1, \ldots, N - 1 \text{ do not share a birthday})
  
  = P(\text{person } N \text{ does not have the same birthday as persons } 1, \ldots, N - 1)
  
  \times P(\text{persons } 1, \ldots, N - 1 \text{ do not share a birthday})
  
  = \frac{365 - (N - 1)}{365} p_{N-1}.
  
\]

Hence

\[
 p_N = \frac{365 - (N - 1)}{365} p_{N-1} = \frac{365 - (N - 1)}{365} \left( \frac{365 - (N - 2)}{365} \right) p_{N-2} = \cdots = \prod_{i=2}^{N} \frac{366 - i}{365}.
\]
The handout shows a plot of these probabilities for different $N$. Are you surprised?

HANDOUT: BIRTHDAY PROBLEM
HANDOUT: BAYES LAW - EXPERT SYSTEMS
HANDOUT: TUTORIAL 1
2 Discrete Random Variables

Random variables are a very important idea in probability theory.

**Simplistic Definition:** a *discrete random variable* \( X \) is the sample space of a discrete probability space that is a subset of \( \mathbb{R} \).

i.e. throws of a die. \( \Omega = \{1, 2, 3, 4, 5, 6\} \) is a random variable; throwing a coin \( \Omega = \{H, T\} \) is not a random variable.

The big advantage of random variables is that \( P \) can be represented by a function of a real variable \( p: \Omega \to [0, 1] \) for the probability of each outcome. We would like to be able to do this for non-real valued \( \Omega \) as well, which we can do if we have a mapping from \( \Omega \) to \( \mathbb{R} \) to represent \( \Omega \). To make it compatible with the definition of a probability space, it should have certain properties as well. This gives us:

**Formal Definition:** A *discrete random variable* \( X \) is a mapping \( X: \Omega \to \mathbb{R} \) on the space \((\Omega, F, P)\), where \( \Omega \) is discrete and

- \( X(\Omega) \) is a countable subset of \( \mathbb{R} \);
- \( \{\omega \in \Omega \mid X(\omega) = x\} \in F, \forall x \in \mathbb{R} \)

Note: when \( \Omega \subset \mathbb{R} \), \( X \) is almost always the identity mapping, which gives us the simplistic definition.

Given a random variable, we can propose:

**Definition:** If \( X \) is a discrete random variable on \((\Omega, F, P)\) then the *probability mass function* (pmf) \( p_X \) of \( X \) is the function such that

\[
p_X(x) = P(\{\omega \in \Omega \mid X(\omega) = x\}).
\]

We often say \( P(X = x) \) for \( p_X(x) \).

\( p_X(x) \) is a probability, and so must satisfy \( 0 \leq p_X(x) \leq 1 \) and \( \sum_{x \in \mathbb{R}} p_X(x) = 1 \).

**Example:** Toss a coin. \( \Omega = \{H, T\} \). An obvious random variable to associate with tossing coin is \( X(T) = 0 \) and \( X(H) = 1 \). If \( P(H) = p \) and \( P(T) = 1 - p \) then the pmf can be written \( p^x(1-p)^{1-x} \).

**Example:** Throw a die. \( \Omega = \{1, 2, 3, 4, 5, 6\} \). Suppose you play a game where you: win £2 if a 5 is thrown, lose £1 if a 1 or a 3 is thrown, and lose nothing otherwise. Then your winnings are a discrete random variable \( X: \Omega \to \mathbb{R} \) with \( X(1) = X(3) = -1 \), \( X(2) = X(4) = X(6) = 0 \) and \( X(5) = 2 \).

2.1 Some Common Discrete Random Variables

Certain random variables occur very frequently, and we list 6 of them here:

1. **Bernoulli.** This is the coin toss. \( X \) takes value 1 with probability \( p \) and 0 with probability \( 1 - p \); no other value is possible. Often 1 is associated with success an 0 with failure, and so \( p \) is called ”success probability”. The pmf is

\[
P(X = x) = p^x(1-p)^{1-x}, \quad x = 0, 1.
\]

2. **Binomial.** Suppose we throw a coin \( n \) times, where \( P(H) = p \), and each throw is independent. What is the chance that we get \( x \) heads? In general, we have \( P(kH) = p^k(1-p)^{n-k} \times \text{no. of ways of getting } kH \). The number of ways of getting \( x \) H from \( n \) is \( n!/(x!(n-x)!)) \). This gives us the binomial distribution. It is the number of ”successes” of \( n \) independent Bernoulli trials, with prob. of success \( p \). The pmf is:

\[
P(X = x) = \binom{n}{x} p^x(1-p)^{n-x}, \quad x = 0, 1, \ldots, n.
\]
Example: throw a die 4 times, what is the probability of getting 0, 1, 2, etc sixes? (Binomial, \( n = 6, p = 1/6 \))

3. **Geometric.** If we keep tossing the above coin, how many tosses until we get a H? For the first to be on the \( n \)th throw, we have to have \( n-1 \) throws T then a H, with prob. \((1-p)^{n-1}p\). This is the geometric. The pmf is:

\[
P(X = x) = (1 - p)^{x-1}p, \ x = 1, 2, \ldots ;
\]

Example: what is the probability of seeing the first 6 on the \( x \)th throw of a die? For \( x = 3 \)?

4. **Negative binomial.** How many coin tosses until we get \( k \) H? This is the negative binomial. Its pmf is:

\[
P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \ x = k, k+1, \ldots ,
\]

When \( k = 1 \) we get the geometric.

5. **Poisson.** The Poisson is a random variable on \( \{0, 1, \ldots \} \) with pmf:

\[
P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \ x = 0, 1, 2, \ldots .
\]

The Poisson describes many physical systems, such as number of customers entering a bank per minute, radioactive decays per second, etc. Example: the number of server failures in a data centre per day is Poisson with \( \lambda = 3 \). What is the probability of 0 failures, 4 failures, 2 or more failures? The Poisson is justified by thinking of a binomial with large \( n \) and small \( p \) and letting \( n \to \infty \) and \( p \to 0 \) such that \( np = \lambda \) is a constant.

\[
P(X = k) = \lim_{n \to \infty} \frac{n!}{k!(n-k)!} \left( \frac{\lambda}{n} \right)^k \left( 1 - \frac{\lambda}{n} \right)^{n-k}
\]

\[
= \lim_{n \to \infty} \frac{n(n-1) \cdots (n-k+1)}{n^k} \left( \frac{\lambda^k}{k!} \right) \left( 1 - \frac{\lambda}{n} \right)^n \left( 1 - \frac{\lambda}{n} \right)^{-k}
\]

\[
= \frac{\lambda^k}{k!} \lim_{n \to \infty} \frac{n(n-1) \cdots (n-k+1)}{n^k} \lim_{n \to \infty} \left( 1 - \frac{\lambda}{n} \right)^n \lim_{n \to \infty} \left( 1 - \frac{\lambda}{n} \right)^{-k}
\]

\[
= \frac{\lambda^k}{k!} e^{-\lambda}.
\]

6. **Hypergeometric.** This is the “lottery probability”. There are \( N \) balls in a bag, of which \( K \) are white and \( N - K \) are black; in a lottery, the white ones correspond to the numbers that you picked. \( n \) of these are picked without replacement - these correspond to the numbers chosen to be the winning ones. The probability that \( x \) of the \( n \) are white is:

\[
P(X = x) = \binom{K}{x} \binom{N-K}{n-x} \binom{N}{n}, \ x = 0, 1, \ldots , \min(n, K).
\]

For example, a lottery where you choose 6 from 45 numbers (so \( N = 45, K = 6, n = 6 \)) and 6 numbers are picked as the winning ones without replacement (so \( n = 6 \)), then:

\[
P(x \text{ numbers picked}) = \frac{6}{x} \binom{39}{6-x} \binom{45}{6}, \ x = 0, 1, \ldots , 6.
\]
Thus, \( P(0) = 0.401 \), \( P(1) = 0.424 \), \( P(2) = 0.151 \), \( P(3) = 0.02247 \), \( P(4) = 0.0013605 \), \( P(5) = 0.00002875 \), \( P(6) = 0.000001235 \) (1 in 8.1 million).

### 2.2 Expectation

When we throw a die, we do not know of course which outcome will occur. However, there are properties of random experiments that are not uncertain. For example, suppose that we repeatedly toss the die. Let \( X_i \) be the \( i \)th outcome and look at the average number thrown \( \sum_1^n X_i/n \) as a function of \( n \).

**Handout: Die Throws**

Such a property holds for many random variables. It is called the expectation.

**Definition:** The expectation of a discrete random variable \( X \) with pmf \( p_X(x) \) is

\[
E(X) = \sum_{\forall x} x p_X(x).
\]

It is quite easy to show that the running average \( \sum_1^n X_i/n \) does indeed “converge” to \( E(X) \) as \( n \to \infty \) (The law of large numbers).

The expectation is also called the mean of \( X \). It is also the ‘centre of gravity” of a probability distribution.

**Example:** Let \( X \) be Bernoulli(\( p \)). Then \( E(X) = \sum_{\forall x} x p_X(x) = 0 \times (1-p) + 1 \times p = p \).

**Example:** Let \( X \) be geometric(\( p \)). Then

\[
E(X) = \sum_{\forall x} x p_X(x) = \sum_{x=1}^\infty x(1-p)^{x-1}p = p \sum_{x=1}^\infty \frac{d}{dp} \left(-(1-p)^x\right) = -p \frac{d}{dp} \sum_{x=1}^\infty (1-p)^x = -pd/dp \frac{1}{p} = 1/p.
\]

**Example:** Let \( X \) be Poisson(\( \lambda \)). Then

\[
E(X) = \sum_{x=0}^\infty \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=1}^\infty \frac{\lambda^{x-1}}{(x-1)!} = e^{-\lambda} \sum_{x=0}^\infty \frac{\lambda^x}{x!} = \lambda e^{-\lambda} e^\lambda = \lambda.
\]

Other mean values are: for the binomial(\( n, p \)), \( E(X) = np \), for the negative binomial(\( n, p \)), \( E(X) = n/p \).

**Handout: Probability distribution examples**

### 2.3 Expectation of Functions of \( X \)

For any function \( g : \mathbb{R} \to \mathbb{R} \), we define the expectation of \( g(X) \) to be

\[
E(g(X)) = \sum_{\forall x} g(x) p_X(x).
\]

**Example:** \( X \) is Bernoulli(\( p \)). What is \( E(e^X) \)? \( E(e^X) = \sum_{x=0,1} e^x p^x (1-p)^{1-x} = (1-p) + pe = 1 + (e-1)p \).

### 2.4 Conditional Expectation

We define the conditional expectation of \( X \) given \( A \) to be:

\[
E(X \mid A) = \sum_{\forall x} x P(X = x \mid A).
\]

There is an equivalent of the Partition theorem for conditional expectations, which goes as follows:
Theorem: Let $B_1, B_2, \ldots$ be a partition of $\Omega$ and let $X$ be any rv on $\Omega$. Then

$$E(X) = \sum_i E(X \mid B_i) P(B_i),$$

whenever this sum converges absolutely.

Proof: using definition of conditional probability:

$$\sum_i E(X \mid B_i) P(B_i) = \sum_i \sum_x x P(\{X = x\} \cap B_i) = \sum_i x P(\{X = x\} \cap \bigcup_i B_i) = \sum_i x P(X = x) = E(X).$$

2.5 Properties of Expectation

Lemma: The following properties hold:

1. For a constant $c$, $E(c) = c$.
2. For any rv $X$ and $c \in \mathbb{R}$, $E(cX) = cE(X)$;
3. For any 2 rvs $X$ and $Y$, $E(X + Y) = E(X) + E(Y)$.
4. If $X$ and $Y$ are independent then $E(XY) = E(X)E(Y)$.
5. In fact, $X$ and $Y$ are independent iff $E(g(X)h(Y)) = E(g(X))E(h(Y))$ for all functions $g$ and $h$ for which the expectations exist.

Proof: trivial except for the last two, which we will prove when we come to do multivariate distributions in Section 4.

2.6 Variance

Here are two probability distributions on $\{0, 1, 2, 3, 4, 5\}$: distribution 1 is $(0.05, 0.05, 0.4, 0.4, 0.05, 0.05)$, distribution 2 is $(0.2, 0.1, 0.2, 0.2, 0.1, 0.2)$. PLOT THEM. We see that the expectation in both cases is 2.5 but that no. 1 is much less “spread out” or variable than no. 2.

The variance of a random variable measures the amount of spread. It is defined to be

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

the mean squared distance between $X$ and its mean. Note that:

$$E((X - E(X))^2) = E(X^2 - 2E(X)X + E(X)^2) = E(X^2) - 2E(X)^2 + E(X)^2 = E(X^2) - E(X)^2;$$

where $E(X^2) = \sum_{x=0}^\infty x^2 p_X(x)$. We usually use this formula to compute the variance.

Example: $X$ is Bernoulli($p$). We know that $E(X) = p$, and $E(X^2) = \lambda = 0^2(1-p) + 1^2 p = p$, thus $\text{Var}(X) = p - \lambda^2 = p(1-p)$.

Example: $X$ is Poisson($\lambda$). We know that $E(X) = \lambda$ and:

$$E(X^2) = \sum_{x=0}^\infty x^2 p_X(x) = \sum_{x=0}^\infty x^2 \lambda^x e^{-\lambda} / x! = e^{-\lambda} \sum_{x=1}^\infty x \lambda^x / (x-1)!$$

$$= e^{-\lambda} \sum_{x=1}^\infty (x-1) \lambda^x / (x-1)! + \sum_{x=1}^\infty \lambda^x / (x-1)!$$

$$= e^{-\lambda} \lambda^2 \sum_{x=1}^\infty (x-1) \lambda^{x-2} / (x-1)! + \lambda \sum_{x=1}^\infty \lambda^{x-1} / (x-1)!$$

$$= e^{-\lambda} \lambda^2 \sum_{x=0}^\infty x \lambda^x / x! + \lambda \sum_{x=0}^\infty \lambda^x / x! = e^{-\lambda} (\lambda^2 e^\lambda + \lambda e^\lambda) = \lambda^2 + \lambda$$
thus \( \text{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda \). So for the Poisson distribution, mean = variance.

Other variances are: the binomial, \( \text{Var}(X) = np(1 - p) \), geometric, \( \text{Var}(X) = (1 - p)/p^2 \)
3 Continuous Random Variables

Discrete rvs take only countably many values, which is very restrictive; continuous random quantities (lifelength of a light bulb, height of a randomly picked individual etc.) are also needed. To model continuous quantities, we need to redefine the rv.

**Definition** A random variable $X$ on the space $(\Omega, \mathcal{F}, P)$ is a mapping $X : \Omega \to \mathcal{F}$ such that \{\{ω ∈ Ω | X(ω) ≤ x\} ∈ \mathcal{F} for all x ∈ \mathbb{R}.

This subsumes the definition of a discrete rv.

Random variables in general are studied by their distribution function $F_X : \mathbb{R} → [0, 1]$, which is defined to be:

$$F_X(x) = P(\{\omega ∈ Ω : X(\omega) ≤ x\}),$$

or more simply $P(X ≤ x)$.

**Note:** For discrete random variables with a mass function $p_X(x)$, $P(X ≤ x) = \sum_{k≤x} p_X(k)$, so looks like an increasing step function.

**Properties of $F_X$:** Intuitively $F_X(x) → 1$ as $x → \infty$, and $F_X(x) → 0$ as $x → -\infty$. Also, $F$ is non-decreasing, since for $y > x$:

$$F_X(y) = P(X ≤ y) = P(\{\omega ∈ Ω : X(\omega) ≤ y\}) = P(\{\omega ∈ Ω : X(\omega) ≤ x\} \bigcup \{\omega ∈ Ω : x ≤ X(\omega) ≤ y\})$$

$$= F_X(x) + P(\{\omega ∈ Ω : x ≤ X(\omega) ≤ y\}) ≥ F_X(x).$$

Finally, for a bounded interval $(a, b]$, we have:

$$P(a < X ≤ b) = P(\{X ≤ b\} - \{X ≤ a\}) = P(X ≤ b) - P(X ≤ a) = F_X(b) - F_X(a),$$

**Definition:** A continuous random variable $X$ is one where $F_X(x)$ can be written

$$F_X(x) = \int_{-\infty}^{x} f_X(s) \, ds,$$

for some function $f_X(s)$. $f_X$ is called the probability density function (pdf).

### 3.1 Properties of the pdf

Any function $f_X : \mathbb{R} → \mathbb{R}$ satisfying:

1. Since $F_X$ is non-decreasing, we must have $f_X(x) ≥ 0$.
2. Since $F_X(x) → 1$, we must have that $\int_{-\infty}^{\infty} f_X(x) \, dx = 1$.

can be a pdf. In addition, we observe that:

1. $f_X(x) = dF_X(x)/dx$, where such a differential exists.
2. For $a < b$, $P(a ≤ X ≤ B) = P(X ≤ b) - P(X ≤ a) = \int_{-\infty}^{b} f_X(x) \, dx - \int_{-\infty}^{a} f_X(x) \, dx = f_X^b f_X(x) \, dx$.
3. $f_X(x)$ is NOT the probability that $X = x$. For small $\delta x$,

$$P(x ≤ X ≤ x + \delta x) = \int_{x}^{x+\delta x} f_X(u) \, du \approx f_X(x)\{\delta x, $$

so $f_X(x)\delta x$ is a probability.
4. In fact, for a continuous rv, \( P(X = x) = 0 \)! This seems counter-intuitive but isn’t really. It’s a result of measure theory (\( \{x\} \) has measure 0). Practically, we can never measure a continuous quantity exactly, there is always some uncertainty, and so we always will really be talking about \( X \) in an interval.

5. \( f_X \) does not have to be in \([0, 1]\), continuous or even bounded (as long as it is integrable to 1).

### 3.2 Examples of Continuous Random Variables

- **Uniform** Let \( a < b \), and define

\[
F_X(x) = \begin{cases} 
0, & \text{if } x < a \\
(x - a)/(b - a), & \text{if } a \leq x \leq b \\
1, & \text{if } x > b.
\end{cases}
\]

\( F \) satisfies the properties of distribution function. The pdf is \( f_X(x) = 1/(b - a) \), for \( a \leq b \), 0 otherwise.

- **Exponential** Let \( \lambda > 0 \) and define

\[
F_X(x) = \begin{cases} 
0, & \text{if } x \leq 0 \\
1 - e^{-\lambda x}, & \text{if } x > 0
\end{cases}
\]

The pdf is \( f_X(x) = \lambda e^{-\lambda x} \), for \( x > 0 \).

- **Normal, or Gaussian**. A very important distribution. The pdf has parameters \( \mu \in \mathbb{R} \) and \( \sigma^2 > 0 \):

\[
f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, x \in \mathbb{R}.
\]

\( F_X(x) \) is not in close form.

We’ll introduce others as we need to.

**HANDOUT: GAUSSIAN, TABLES OF THE NORMAL DISTRIBUTION**

### 3.3 Expectation and Variance

The expected value of a continuous rv \( X \) is:

\[
E(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx,
\]

the expected value of a function of \( X \), \( g(X) \) is:

\[
E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx.
\]

The variance of \( X \) is still \( E((X - E(X))^2) \), which can be calculated as \( E(X^2) - E(X)^2 \).

**Example:** \( f_X(x) = 3x^2, 0 \leq x \leq 1, \) 0 otherwise. Is this a legitimate pdf? If so, what are \( E(X) \) and \( \text{Var}(X) \)?

Since \( f_X(x) \geq 0 \) and \( \int_{-\infty}^{\infty} f_X(x) \, dx = x^3|_0^1 = 1 \), then it is a legitimate pdf.

\[
E(X) = \int_{0}^{1} x \, 3x^2 \, dx = 0.75x^4|_0^1 = 0.75.
\]
\[ E(X^2) = \int_0^1 3x^4 \, dx = 0.6; \ Var(X) = 0.6 - 0.75^2 = 0.0375. \]

**Example** Uniform distribution.

\[ E(X) = \int_a^b x/(b-a) \, dx = x^2/(2(b-a)) \bigg|_a^b = (b^2 - a^2)/(2(b-a)) = (a+b)/2. \]

\[ E(X^2) = x^3/(3(b-a)) \bigg|_a^b = (b^3 - a^3)/(3(b-a)) = (a^2 + ab + b^2)/3. \]

thus

\[ Var(X) = (a^2 + ab + b^2)/3 - (a + b)^2/4 = (b - a)^2/12. \]

**Example** Exponential distribution.

\[ E(X) = \int_0^\infty x \lambda e^{-\lambda x} \, dx = (\text{by parts}) = x \int_0^\infty \lambda e^{-\lambda x} - \ldots = 1/\lambda \]

It turns out that \( E(X^2) = 2/\lambda^2 \), thus \( Var(X) = 1/\lambda^2 \).

**Example**: Gaussian, we have \( E(X) = \mu \) and \( Var(X) = \sigma^2 \).

**HANDOUT: RETURN TO PROBABILITY DISTRIBUTION EXAMPLES**

### 3.4 Properties of Expectations

...are exactly those of expectation of discrete random variables.

### 3.5 Functions of a Random Variable

Given a random variable \( X \) with pdf \( f_X(x) \), can we say anything about the pdf of any function of \( X \), say \( g(X) \)? Yes, we can, as long as we make some restrictions on the class of \( g \):

**Theorem:** Let \( X \) be a continuous rv with pdf \( f_X(x) \) and let \( g \) be a strictly increasing and differentiable function from \( \mathbb{R} \) to \( \mathbb{R} \). Then \( Y = g(X) \) has pdf:

\[ f_Y(y) = f_X(g^{-1}(y)) \frac{d}{dy}[g^{-1}(y)]. \]

**Proof:** if we look at the distribution function, then since \( g \) is strictly increasing:

\[ P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)); \]

so \( F_Y(y) = F_X(g^{-1}(y)). \) Differentiate both sides wrt \( y \) to get the result.

Note that if \( g \) is strictly decreasing then by the same proof we have

\[ Y(y) = -f_X(g^{-1}(y)) \frac{d}{dy}[g^{-1}(y)]. \]

**Example:** If \( X \) has density \( f_X(x) \) then \( Y = X^3 \) has distribution function:

\[ P(Y \leq y) = P(X^3 \leq y) = P(X \leq \sqrt[3]{y}) \]

and density (by theorem) \( f_X(\sqrt[3]{y}) y^{-2/3}/3 \)

The density of \( g(X) \) exists for many cases outside that of the theorem, but there are no general results. The general idea is to work out the distribution function first and then differentiate to get the pdf.
Example: Let $X$ have pdf $f_X(x)$ and let $Y = X^2$. The distribution function of $Y$ is then:

$$P(Y \leq y) = P(X^2 \leq y) = \begin{cases} 0, & \text{if } y < 0 \\ P(-\sqrt{y} \leq X \leq \sqrt{y}), & \text{if } y \geq 0 \end{cases}$$

Differentiating we see that $f_Y(y) = 0$ for $y < 0$ and for $y > 0$ we have:

$$f_Y(y) = \frac{d}{dy} P(Y \leq y) = \frac{d}{dy} P(-\sqrt{y} \leq X \leq \sqrt{y}) = \frac{d}{dy} (F_X(\sqrt{y}) - F_X(-\sqrt{y})) = \frac{1}{2\sqrt{y}} (f_X(\sqrt{y}) + f_X(-\sqrt{y})).$$

HANDOUT: Tutorial 3.
4 Multivariate Random Variables

Often we want to make probability statements about 2 or more random variables at the same time. We do this using the methods of multivariate probability.

4.1 Multivariate Discrete Distributions

Definition: Let \( X \) and \( Y \) be discrete random variables on \((\Omega, \mathcal{F}, P)\). The joint probability mass function of \( X \) and \( Y \) is the function \( p_{X,Y} : \mathbb{R}^2 \to [0, 1] \) defined by:

\[
p_{X,Y}(x, y) = P(\{\omega \in \Omega | X(\omega) = x \text{ and } Y(\omega) = y\}),
\]

often written \( P(X = x, Y = y) \)

Lemma: The individual pmf of \( X \) and \( Y \) are

\[
p_X(x) = \sum_y P(X = x, Y = y), \quad p_Y(y) = \sum_x P(X = x, Y = y),
\]

these are also called the marginal distributions of \( X \) and \( Y \).

Proof: by Partition theorem: \( P(X = x) = \sum_y P(X = x | Y = y) P(Y = y) = \sum_y P(X = x, Y = y) \).

Similarly, for a vector of rvs \((X_1, X_2, \ldots, X_n)\), the joint pmf is \( P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \).

Example: \( X \) and \( Y \) take values in \( \{0, 1, 2\} \). The joint pmf can be written in a table as follows:

<table>
<thead>
<tr>
<th></th>
<th>( X )</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 )</td>
<td>0.15</td>
<td>0.1</td>
</tr>
<tr>
<td>( 1 )</td>
<td>0.15</td>
<td>0.2</td>
</tr>
<tr>
<td>( 2 )</td>
<td>0.25</td>
<td>0.05</td>
</tr>
</tbody>
</table>

We see that \( P(X = 1) = P(X = 1, Y = 0) + P(X = 1, Y = 1) + P(X = 1, Y = 2) = 0.35. \)

4.2 Expectation

For any function \( g : \mathbb{R}^2 \to \mathbb{R} \), the expectation of \( g(X, Y) \) is

\[
E(g(X,Y)) = \sum_{x} \sum_{y} g(x, y) p_{X,Y}(x,y).
\]

Theorem: The expectation operator is linear, that is for \( a, b \in \mathbb{R} \) and discrete rvs \( X \) and \( Y \),

\[
E(aX + bY) = aE(X) + bE(Y).
\]

Proof:

\[
E(aX + bY) = \sum_{x} \sum_{y} (ax + by)p_{X,Y}(x,y) = a \sum_{x} \sum_{y} p_{X,Y}(x,y) + b \sum_{y} \sum_{x} p_{X,Y}(x,y)
\]

\[
= a \sum_{x} xp_{X}(x) + b \sum_{y} yp_{Y}(y) = aE(X) + bE(Y).
\]

4.3 Independence

Definition: Random variables \( X \) and \( Y \) are independent if \( p_{X,Y}(x, y) = p_X(x)p_Y(y), \forall x, y. \)

This could also be written \( p_{X,Y}(x, y) = (\sum_y p_{X,Y}(x,y))((\sum_x p_{X,Y}(x,y)). \) In fact:
Theorem: Discrete rvs $X$ and $Y$ are independent iff there exist functions $g$, $h$ such that $p_{X,Y}(x,y) = g(x)h(y)$, \(\forall x,y\).

Proof: \(\Leftarrow\). Choose $g(x) = p_X(x)$ and $h(y) = p_Y(y)$.

\(\Rightarrow\) If $g$ and $h$ exist then $p_X(x) = \sum_y p_{X,Y}(x,y) = g(x) \sum_y h(y)$ and $p_Y(y) = h(y) \sum_x g(x)$. We also know that $1 = \sum_x \sum_y p_{X,Y}(x,y) = \sum_x \sum_y g(x)h(y) = \sum_x g(x) \sum_y h(y)$. Thus $p_{X,Y}(x,y) = g(x)h(y) = g(x)h(y) \sum_y g(x) \sum_y h(y) = p_X(x)p_Y(y)$.

Theorem: If $X$ and $Y$ are independent rvs then $E(XY) = E(X)E(Y)$.

Proof: $E(XY) = \sum_{x,y} xy p_{X,Y}(x,y) = \sum_{x,y} xyp_X(x)p_Y(y) = \sum_x x p_X(x) \sum_y y p_Y(y) = E(X)E(Y)$. The converse is not true — there are dependent rvs for which $E(XY) = E(X)E(Y)$.

Example: Let $P(X = -1) = P(X = 0) = P(X = 1) = 1/3$, and let $Y = 0$ if $X = 0$ and $Y = 1$ if $X \neq 0$. Then: $X$ and $Y$ are dependent (since $P(X = 0, Y = 1) = 0$ but $P(X = 0)P(Y = 1) = 2/9$) but $E(XY) = 0$ and $E(X) = 0$, $E(Y) = 2/3$ $\Rightarrow$ $E(X)E(Y) = 0$ also.

However, the following is true:

Theorem: Discrete rvs $X$ and $Y$ are independent iff $E(g(X)h(Y)) = E(g(X))E(h(Y))$ for all functions $g, h : \mathbb{R} \to \mathbb{R}$ for which these expectations exist.

Proof: \(\Leftarrow\). Choose

$$g(x) = \begin{cases} 1, & \text{if } x = a \\ 0, & \text{if } x \neq a \end{cases}, \quad h(y) = \begin{cases} 1, & \text{if } y = b \\ 0, & \text{if } y \neq b \end{cases}$$

Then $E(g(X)h(Y)) = p_{X,Y}(a,b)$ and $E(g(X))E(h(Y)) = p_X(a)p_Y(b)$. So $p_{X,Y}(a,b) = p_X(a)p_Y(b)$.

Since we can repeat this \(\forall a, b\) we have that $p_{X,Y}(a,b) = p_X(a)p_Y(b) \forall a, b$ so $X$ and $Y$ are independent.

\(\Leftarrow\) is just the previous proof.

### 4.4 Generalising to $n$ Random Variables

Everything we have done generalises to the case where we have a vector of rvs $X(X_1, X_2, \ldots, X_n)$.

The joint pmf is given by

$$p_X(x) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

The marginal pmfs are given by

$$p_X(x_i) = \sum_{\forall x_j \neq i} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

If the $X_i$ are independent then $p_X(x) = \prod_i p_X(x_i)$ and $E(X_1X_2\ldots X_n) = E(X_1)E(X_2)\ldots E(X_n)$.

### 4.5 Multivariate Distribution Functions

- For any two rvs $X$ and $Y$, we define the joint distribution function to be:
  $$F_{X,Y}(x,y) = P(X \leq x, Y \leq y).$$

- We see that $\lim_{x,y \to -\infty} F_{X,Y}(x,y) = 0$ and that $\lim_{x,y \to \infty} F_{X,Y}(x,y) = 1$. Further $F$ is non-decreasing in the sense that $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$ whenever $x_1 < x_2$ and $y_1 < y_2$.

- We can also compute the marginal distribution functions of $X$ and $Y$ by noting that $F_X(x) = P(X \leq x) = P(X \leq x, Y \leq \infty) = F_{X,Y}(x, \infty)$; similarly $F_Y(y) = F_{X,Y}(\infty, y)$.

- If $X$ and $Y$ are independent iff $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$, which is to say $F_{X,Y}(x,y) = F_X(x)F_Y(y)$.

- Finally, for $n$ rvs $X_1, X_2, \ldots, X_n$, the joint distribution is $F_{X}(x) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$. The variables are independent if $F_{X}(x) = F_{X_1}(x_1)F_{X_2}(x_2)\ldots F_{X_n}(x_n)$.
Example: Let \( F_{X,Y}(x,y) = 1 - e^{-x} - e^{-y} + e^{-(x+y)} \) for \( x, y \geq 0 \). This is a distribution function, and we have \( F_X(x) = F_{X,Y}(x,\infty) = 1 - e^{-x} \) and \( F_Y(y) = 1 - e^{-y} \) (exponential rvs). Further, we see that \( F_{X,Y}(x,y) = F_X(x)F_Y(y) \) so \( X \) and \( Y \) are independent.

Example: Let \( F_{X,Y}(x,y) = 1 - 0.5e^{-x} - 0.5e^{-y} \) for \( x, y \geq 0 \). Then \( F_X(x) = 1 - 0.5e^{-x} \) and \( F_Y(y) = 1 - 0.5e^{-y} \) but now \( F_{X,Y}(x,y) \neq F_X(x)F_Y(y) \) and so these two rvs are not variable.

4.6 Continuous Random Variables

\( X \) and \( Y \) are called continuous if \( F_{X,Y}(x,y) \) can be written

\[
F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v) \, du \, dv,
\]

for a function \( f_{X,Y} : \mathbb{R^2} \rightarrow [0,\infty) \) called the joint probability density function. 

\( f \) has many of the properties of the pdf. We have that \( f_{X,Y}(x,y) \geq 0 \) and

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, dx \, dy = 1.
\]

Any function \( f \) with these properties is a pdf. Also, \( P(x < X, x + \delta x < Y < y + \delta y) \approx f_{X,Y}(x,y) \delta x \delta y \). Also, for any \( A \subset \mathbb{R^2} \), we have

\[
P((X,Y) \in A) = \int_{A} f_{X,Y}(x,y) \, dx \, dy.
\]

Everything generalises to \((X_1, X_2, \ldots, X_n)\). For continuous rvs there is a pdf \( f(x_1,\ldots,x_n) \) such that:

\[
P((X_1, X_2, \ldots, X_n) \in A) = \int_{A} f(x_1,\ldots,x_n) \, dx,
\]

for \( A \subset \mathbb{R^n} \), and further

\[
f(x_1,\ldots,x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} P(X_1 \leq x_1, \ldots, X_n \leq x_n).
\]

4.7 Marginal Density Functions and Independence

This follows closely the ideas of joint discrete rvs. The marginal pdfs of \( X \) and \( Y \) are:

\[
f_X(x) = \frac{d}{dx} P(X \leq x) = \frac{d}{dx} \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u,v) \, du \, dv = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, dy,
\]

and \( f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, dx \).

If \( X \) and \( Y \) are independent then \( F_{X,Y}(x,y) = F_X(x)F_Y(y) \), which implies \( f_{X,Y}(x,y) = f_X(x)f_Y(y) \).

Example: \( X \) and \( Y \) have the joint pdf:

\[
f(x,y) = \begin{cases} \quad cx, & \text{if } 0 < y < x < 1, \\ 0, & \text{otherwise} \end{cases}
\]

What is \( c \)? What are the marginal densities of \( X \) and \( Y \). Are \( X \) and \( Y \) independent? What is \( P(X > 2Y) \)?

Answer: Since \( f \) integrates to 1, we have

\[
1 = \int_{0}^{1} \int_{0}^{x} cx \, dy \, dx = \int_{0}^{1} cx^2 = c/3,
\]

\[\text{Simon Wilson, 2012}\]
so $c = 3$. Marginal of $X$ is:

$$f_X(x) = \int_0^x 3x \, dy = 3x^2,$$

and of $Y$ is

$$f_Y(y) = \int_y^1 3x \, dx = 3(1 - y^2)/2.$$

Since $f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$, so $X$ and $Y$ are dependent. Finally:

$$P(X > 2Y) = \int_0^{1/2} \int_{2y}^1 3x \, dx \, dy = \int_0^{1/2} (3/2) - 6y^2 \, dy = 1/4.$$

### 4.8 Conditional Density Functions

If $X$ and $Y$ have joint pdf $f_{X,Y}(x,y)$ then we can talk about the pdf of $X$ conditional on $Y = y$. The pdf that describes this is the conditional density function and is defined as:

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

similarly $f_{Y|X}(y | x) = f_{X,Y}(x,y)/f_X(x)$.

**Example:** Let $f_{X,Y}(x,y) = e^{-x}$, for $0 < x < y < \infty$, and 0 otherwise. What are $f_{X|Y}(x | y)$ and $f_{Y|X}(y | x)$?

$f_X(x) = \int_x^\infty e^{-y} \, dy = e^{-x}$ and $f_Y(y) = \int_0^y e^{-y} \, dx = ye^{-y}$. Thus $f_{Y|X}(y | x) = e^{x-y}$, $y > x$ and $f_{X|Y}(x | y) = 1/y$, $0 < x < y$.

Conditional expectation works in the intuitive way:

$$E(X \mid Y = y) = \int_{\forall x} x f_{X|Y}(x | y) \, dx.$$

The equivalent of the partition theorem for continuous expectation is:

**Theorem:** For jointly continuous rvs $X$ and $Y$, $E(Y) = \int E(Y \mid X = x) \, f_X(x) \, dx$.

### 4.9 Expectations of Continuous Random Variables

The properties of expectations of discrete random quantities also hold for continuous. For $X$ and $Y$ jointly continuous and a function $g : \mathbb{R}^2 \to \mathbb{R}$, $E(g(X,Y))$ is defined as

$$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) \, f_{X,Y}(x,y) \, dx \, dy.$$

Expectation is linear ($E(aX + bY) = aE(X) + bE(Y)$). If $X$ and $Y$ are independent then $E(XY) = E(X)E(Y)$. $X$ and $Y$ are independent iff $E(g(X)h(Y)) = E(g(X))E(h(Y))$ for all functions $g$ and $h$ for which these expectations exist.

The proofs of these results are very similar to those for the discrete case, but with the pdf replacing the pmf and integrals replacing sums.

**Example:** Let $(X,Y)$ be the co-ordinates of a point uniformly distributed on the unit circle, so that

$$f_{X,Y}(x,y) = \pi^{-1}, \text{ for } x^2 + y^2 \leq 1,$$
0, otherwise. What is the expected distance of the point from the origin? This distance is \( \sqrt{X^2 + Y^2} \).

Its expected value is:

\[
E(\sqrt{X^2 + Y^2}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sqrt{x^2 + y^2} f_{X,Y}(x, y) \, dx \, dy = \int_{-1}^{1} \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \pi^{-1} \sqrt{x^2 + y^2} \, dx \, dy
\]

= (letting \( x = r \cos(\theta) \), \( y = r \sin(\theta) \)) = 0.5

### 4.10 Variance and Covariance

Remember that \( \text{Var}(X) = E((X - E(X))^2) \) is supposed to measure the amount of dispersion in \( X \). Recall that we can also write \( \text{Var}(X) = E(X^2) - (E(X))^2 \). Variance has the following properties:

**Lemma**: \( \text{Var}(X) = 0 \) iff \( P(X = E(X)) = 1 \).

**Proof**: Let \( \mu = E(X) \). \( \iff \ E(X^2) = \mu^2 \), hence \( \text{Var}(X) = \mu^2 - (E(X))^2 = \mu^2 - \mu^2 = 0 \).

\( \Rightarrow \) If \( \text{Var}(X) = 0 \) then \( E(X^2) = \mu^2 \). Define \( Y = X - \mu \). Then \( E(Y^2) = E((X - \mu)^2) = 0 \). Since \( E(Y^2) = \sum_y y^2 P(Y = y) \), the only way this can happen is of \( P(Y = 0) = 1 \), hence \( P(X = \mu) = 1 \).

**Lemma**: \( \text{Var}(aX + b) = a^2 \text{Var}(X) \), for \( a, b \in \mathbb{R} \).

**Proof**:

\[
\text{Var}(aX + b) = E(((aX + b) - E(aX + b))^2) = E((aX + b - aE(X) - b)^2)
\]

= \( a^2(E(X - E(X))^2) = a^2 \text{Var}(X) \).

Often, statisticians work with the standard deviation, which is \( \sqrt{\text{Var}(X)} \). This has the nice property that \( \text{SD}(aX) = a \text{SD}(X) \).

**Definition**: The covariance of \( X \) and \( Y \) is \( \text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) \).

We can expand the definition of covariance to get that \( \text{Cov}(X, Y) = E(XY) - E(X)E(Y) \).

The covariance contains info about the joint behaviour of \( X \) and \( Y \) are. If \( \text{Cov}(X, Y) > 0 \) then it is more likely that \( X - E(X) \) and \( Y - E(Y) \) have the same sign.

**Lemma**: If \( X \) and \( Y \) are independent then \( \text{Cov}(X, Y) = 0 \).

**Proof**: \( \text{Cov}(X, Y) = E(XY) - E(X)E(Y) \) so it’s sufficient to show that \( E(XY) = E(X)E(Y) \) when \( X \) and \( Y \) are independent. This is easy: The converse is not true in general; see:

<table>
<thead>
<tr>
<th>( Y=-1 )</th>
<th>( X=-1 )</th>
<th>( X=0 )</th>
<th>( X=1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y=1 )</td>
<td>1/6</td>
<td>1/3</td>
<td>1/6</td>
</tr>
<tr>
<td>( Y=1 )</td>
<td>1/6</td>
<td>0</td>
<td>1/6</td>
</tr>
</tbody>
</table>

You can check that \( X \) and \( Y \) are dependent e.g. \( p_{X,Y}(0,1) = 0 \neq 1/3 \times 1/3 = p_X(0)p_Y(1) \) but \( E(XY) = E(X) = E(Y) = 0 \) so \( \text{Cov}(X, Y) = 0 \).

Often we normalise the covariance to:

**Definition**: The correlation between \( X \) and \( Y \) is

\[
\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}.
\]

because:

**Theorem**: \(-1 \leq \rho(X,Y) \leq 1 \).

**Proof**: is a direct results of the Cauchy-Schwartz inequality by setting \( U = X - E(X) \) and \( V = Y - E(Y) \).

**Theorem**: (Cauchy-Schwartz Inequality) If \( U \) and \( V \) are random variables then \( E(UV)^2 \leq E(U^2)E(V^2) \).

**Proof**: Let \( W = sU + V \) for \( s \in \mathbb{R} \). Since \( W^2 \geq 0 \), we have that \( 0 \leq E(W^2) = E((sU + V)^2) = E(U^2)s^2 + 2E(UV)s + E(V^2) \). Clearly \( E(U^2) \geq 0 \). If \( E(U^2) = 0 \) then \( P(U = 0) = 1 \).
and result holds. If $E(U^2) > 0$ then the inequality implies that, as a quadratic in $s$, $E(U^2)s^2 + 2E(UV)s + E(V^2)$ intersects the origin at most once. Thus the discriminant of the quadratic $(2E(UV))^2 - 4E(U^2)E(V^2) \leq 0$ and result follows.

$\rho$ has the following property:

**Theorem:** $\rho(X, Y) = 1$ iff $Y = a + bX$ for $a > 0, b \in \mathbb{R}$. $\rho(X, Y) = -1$ iff $Y = a + bX$ for $a < 0, b \in \mathbb{R}$.

**Proof:** $\Leftarrow$ is easy to see since $\text{Var}(Y) = b^2 \text{Var}(X)$ and so $\sqrt{\text{Var}(X) \text{Var}(Y)} = b \text{Var}(X)$ and

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = E(X(a + bX)) - E(X)E(a + bX) = aE(X) + bE(X^2) - aE(X) - bE(X)^2 = b(E(X^2) - E(X)^2) = b \text{Var}(X),$$

thus $\rho(X, Y) = 1$.

$\Rightarrow$ Let $\rho(X, Y) = 1$. Let $a = \text{Var}(X), b = 2\text{Cov}(X, Y) > 0$ (since $\rho > 0$) and $c = \text{Var}(Y)$. Then

$$b^2 - 4ac = 4\text{Var}(X)\text{Var}(Y)[\rho(X, Y)^2 - 1] = 0.$$

Thus the quadratic $as^2 + bs + c = 0$ has 2 equal roots, say $\alpha = -b/2a$, which must be negative since $b > 0$.

Now let $W = \alpha(X - E(X)) + (Y - E(Y))$. Then

$$E(W^2) = a\alpha^2 + b\alpha + c = 0,$$

which can only happen if $P(W = 0) = 1$, thus $P(Y = -\alpha X + \beta) = 1$ for $\beta = \alpha E(X) + E(Y)$, and since $\alpha < 0$ the relationship is increasing linear.

We can repeat this for $\rho = -1$, getting that $\alpha > 0$ and thus the relationship is decreasing linear.

### 4.11 Sums of Random Variables

**Theorem:** Let $X$ and $Y$ be discrete random variables with joint probability mass function $p_{X,Y}(x, y)$. Then the pmf of $Z = X + Y$ is given by:

$$p_Z(z) = \sum_{x=-\infty}^{\infty} p_{X,Y}(x, z-x).$$

**Proof:**

$$p_Z(z) = P(Z = z) = P(X + Y = z) = \sum_{z_X} P(X = x \text{ and } Y = z-x) = \sum_{x=-\infty}^{\infty} p_{X,Y}(x, z-x) \text{ as required.}$$

Note that if $X, Y \geq 0$ then this reduces to:

$$p_Z(z) = \sum_{x=0}^{z} p_{X,Y}(x, z-x).$$

**Example:** For the joint pmf at the start of the Section, what is the pmf of $Z = X + Y$?

<table>
<thead>
<tr>
<th></th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>$X$</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Example: Let $X$ and $Y$ be independent Poisson random variables with mean $\lambda_x$ and $\lambda_y$ respectively. What is the distribution of $Z = X + Y$?

The equivalent result for continuous random variables replaces (as usual) the sum by an integral:

**Theorem:** Let $X$ and $Y$ be continuous random variables with joint probability density function $f_{X,Y}(x,y)$. Then the pmf of $Z = X + Y$ is given by:

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

If $X, Y \geq 0$ then

$$f_Z(z) = \int_{0}^{z} f_{X,Y}(x, z-x) \, dx.$$ 

Example: What is the pdf of $Z = X + Y$ for the example at the start of Section Example: Let $X$ and $Y$ be independent exponential random variables with the same mean $\lambda$. What is the distribution of $Z = X + Y$?

Can we say anything about the mean and variance of the sum of random variables?

**Theorem:** Let $X$ and $Y$ be random variables. Then

$$E(aX + bY) = aE(X) + bE(Y)$$

and

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y).$$

Note that in vector notation we can write

$$\text{Var}(a,b)(X,Y)^T = (a,b) \begin{pmatrix} \text{Var}(X) & \text{Cov}(X,Y) \\ \text{Cov}(X,Y) & \text{Var}(Y) \end{pmatrix} (a,b)^T.$$ 

This is useful later!!

**Lemma:** If $X$ and $Y$ are independent then $\text{Cov}(X,Y) = 0$ and so $\text{Var}(X+Y) = \text{Var}(X)+\text{Var}(Y)$.

### 4.12 The Multivariate Gaussian Distribution

This is the most important distribution in statistics; it forms the basis of many statistical methods where data are vectors rather than single numbers. Here we will concentrate on its definition, properties and examples. But you will meet it in many places if you study statistical methods further in sophisticated postgraduate courses.

Try [http://www.youtube.com/watch?v=TC0ZAX3DA88](http://www.youtube.com/watch?v=TC0ZAX3DA88) for a 15 minute video on it!!

**Definition:** The *multivariate normal (also called the multivariate Gaussian)* distribution is a continuous probability distribution defined for a set of random variables $X = (X_1, \ldots, X_k)$. Its density has the form:

$$f_X(x) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right), \ x \in \mathbb{R}^k,$$

where $\mu = (\mu_1, \ldots, \mu_k)^T$ is called the *mean vector*, $\Sigma$ is a $k \times k$ matrix, called the *variance-covariance matrix* and $|\Sigma|$ is the determinant of $\Sigma$. It must be a positive definite matrix for this density to be
well defined e.g. \( v^T \Sigma v > 0 \) for all \( v \in \mathbb{R}^k - \{0\} \). The entries in this matrix are the variances (on the main diagonal) and covariances (off the main diagonal) e.g.
\[
\Sigma_{ii} = \text{Var}(X_i) = \sigma_i^2,
\Sigma_{ij} = \text{Cov}(X_i, X_j) = \rho_{ij} \sigma_i \sigma_j,
\]
where \( \rho_{ij} \) is the correlation between \( X_i \) and \( X_j \).

Handout: bivariate normal examples

There are many properties of this distribution; here we simply state them.

1. All marginal distributions of \( X \) are also Gaussian with the mean and variance given in \( \mu \) and \( \Sigma \). So \( X_i \) is normal with mean \( \mu_i \) and variance \( \Sigma_{ii} \).

2. Further, any joint marginal of a subset of \( \{X_1, \ldots, X_k\} \) is also multivariate Gaussian. Its mean vector is the vector of the means of the components of the subset, and the variance-covariance matrix is just the rows and columns of \( \Sigma \) corresponding to the components. To put this in terms of an equation, partition \( X \) into two sub-vectors \( X_1 = (X_1, \ldots, X_j) \) and \( X_2 = (X_{j+1}, \ldots, X_k) \). Similarly partition the mean vector into \( \mu_1 = (\mu_1, \ldots, \mu_j) \) and \( \mu_2 = (\mu_{j+1}, \ldots, \mu_k) \) and the variance-covariance matrix:
\[
\Sigma = \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix},
\]
where
\[
\Sigma_{11} = \begin{pmatrix}
\sigma_1^2 & \rho_{12} \sigma_1 \sigma_2 & \cdots & \rho_{1k} \sigma_1 \sigma_k \\
\rho_{21} \sigma_2 \sigma_1 & \sigma_2^2 & \cdots & \rho_{2k} \sigma_2 \sigma_k \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{k1} \sigma_k \sigma_1 & \rho_{k2} \sigma_k \sigma_2 & \cdots & \sigma_k^2
\end{pmatrix}
\]
is the top left \( i \times i \) sub-matrix of \( \Sigma \), etc. Then \( X_1 \) is multivariate normal with mean \( \mu_1 \) and variance-covariance matrix \( \Sigma_{11} \), and \( X_2 \) is multivariate normal with mean \( \mu_2 \) and variance-covariance matrix \( \Sigma_{22} \).

3. Conditionals. All conditional distributions of any subset of \( \{X_1, \ldots, X_k\} \) conditional on any distinct other subset are also Gaussian. Defining \( X_1 = (X_1, \ldots, X_i) \) and \( X_2 = (X_{i+1}, \ldots, X_k) \) as before, the conditional distribution of \( X_1 \) given \( X_2 = x_2 \) is Gaussian with mean
\[
\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2)
\]
and variance-covariance matrix
\[
\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}.
\]

4. Linear combinations. Let \( Y = b_1 X_1 + b_2 X_2 + \cdots + b_k X_k \). Letting \( b = (b_1, \ldots, b_k)^T \), we can write \( Y = b^T X \). Then \( Y \) is also Gaussian and its mean is \( b^T \mu = b_1 \mu_1 + \cdots + b_k \mu_k \) and variance is
\[
\sum_{i=1}^{k} \sum_{j=1}^{k} b_i b_j \Sigma_{ij} = b^T \Sigma b.
\]

Note that the variance can be rewritten
\[
\sum_{i=1}^{k} b_i^2 \text{Var}(X_i) + \sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} b_i b_j \text{Cov}(X_i, X_j) = \sum_{i=1}^{k} b_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^{k} \sum_{j=1}^{i-1} b_i b_j \text{Cov}(X_i, X_j).
\]

The mean and variance of \( b^T X \) are \( b^T \mu \) and \( b^T \Sigma b \) for any vector of random variables \( X \). The special property when \( X \) is Gaussian is that the distribution of \( b^T X \) is also Gaussian.
5. Linear transformations. More generally, let $B$ be a $m \times k$ matrix and let $Y = BX \in \mathbb{R}^m$. Then $Y$ is Gaussian with mean $B\mu$ and variance-covariance matrix $B^T\Sigma B$.

Example: let $X = (X_1, X_2, X_3, X_4)^T$ be multivariate normal distributed with mean $(1, 0, 3, 2)^T$ and variance-covariance matrix

\[
\begin{pmatrix}
4 & 2.4 & 1 & -1 \\
2.4 & 9 & -0.3 & 2.4 \\
1 & -0.3 & 1 & 0 \\
-1 & 2.4 & 0 & 1
\end{pmatrix}.
\]

• What is $E(X_2)$? What is $\text{Var}(X_2)$? What is $\text{Cov}(X_1,X_3)$? What is $\text{Corr}(X_2,X_3)$?

• What is the distribution of $X_4$? What is the distribution of $(X_1, X_2)^T$? What is the distribution of $(X_2, X_4)^T$?

• What is the distribution of $X_1|X_2 = x_2$? What is the distribution of $(X_1, X_2)^T|(X_3, X_4)^T$? What is the distribution $(X_1, X_4)^T|X_2$?

• What is the distribution of $Y = X_1 + 3X_2 - X_3 + 2X_4$? What is the distribution of

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_2 - X_4 \\ 3X_1 + X_2 + 5X_3 - 2X_4 \end{pmatrix}?$$

HANDOUT: Tutorial 4
5 MOMENT AND CHARACTERISTIC FUNCTIONS

5.1 Moment Generating Functions

**Definition:** For a random variable $X$, the *moment generating function* (mgf) of $X$ is denoted $M_X(t)$ and defined as:

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum_x e^{tx} p_X(x), & \text{if } X \text{ is discrete}, \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) \, dx, & \text{if } X \text{ is continuous}. \end{cases}$$

wherever this sum or integral converges absolutely.

**Example** Normal distribution: For a random variable $X$ with normal distribution $N(\mu, \sigma^2)$,

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi} \sigma} e^{-(x-\mu)^2/2 \sigma^2} \, dx = e^{t\mu + \frac{t^2 \sigma^2}{2}}.$$
5.2 Characteristic Functions

The problem that the mgf is not necessarily defined for all \( t \) can be solved if we look at \( E(e^{itX}) \), for \( i = \sqrt{-1} \). This exists always for all \( t \).

**Definition:** The *characteristic function* of a random variable \( X \) is a function \( \phi_X(t) \) defined by:

\[
\phi_X(t) = E(e^{itX}) = \left\{ \begin{array}{ll}
\sum_{x} e^{itx} p_X(x), & \text{if } X \text{ is discrete,} \\
\int_{-\infty}^{\infty} e^{itx} f_X(x) \, dx, & \text{if } X \text{ is continuous.}
\end{array} \right.
\]

**Note:** Since \( e^{itx} = \cos(tx) + i\sin(tx) \), we can also write \( \phi_X(t) = E(\cos(tX)) + iE(\sin(tX)) \). We see that, since \( e^{itX} \) lies on the unit circle in \( \mathbb{C} \), that we have \( |\phi_X(t)| \leq 1 \). We can also write \( \phi_X(t) = M_X(it) \) if the mgf exists in an open neighbourhood of the origin.

**Example:** \( X \sim N(\mu, \sigma^2) \). Since \( M_X(t) = e^{\mu t} + \frac{\sigma^2 t^2}{2} \) which exists \( \forall t \), we can write \( \phi_X(t) = M_X(it) = e^{\mu t} + \frac{\sigma^2 t^2}{2} \).

**Example:** If \( X \) is exponential(\( \lambda \)). Then

\[
\phi_X(t) = \int_{0}^{\infty} e^{itx} \lambda e^{-\lambda x} \, dx = \lambda \int_{0}^{\infty} e^{(it-\lambda)x} \, dx = \lambda/(\lambda - it).
\]

(this may be found by splitting the integral into real and complex parts, or using residues).

Characteristic functions have the following properties of the mgf (proofs similar to the mgf case):

1. If \( X \) has cf \( \phi_X(t) \) then \( Y = aX + b \) has cf \( \phi_Y(t) = e^{ibt} \phi_X(at) \).
2. If \( X \) and \( Y \) are independent then \( \phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) \).
3. As long as the expectations exist,

\[
i^k E(X^k) = \left. \frac{d^k}{dt^k} \phi_X(t) \right|_{t=0}.
\]

Observant ones among you will have noticed that the characteristic function looks like a Fourier transform of \( f_X(x) \). Therefore we have:

**Theorem:** (Uniqueness) Let \( X \) and \( Y \) have cfs \( \phi_X \) and \( \phi_Y \) respectively. Then \( X \) and \( Y \) have the same distributions iff \( \phi_X(t) = \phi_Y(t) \) for all \( t \in \mathbb{R} \).

**Theorem:** (Inversion) Let \( X \) be continuous with pdf \( f_X(x) \) and cf \( \phi_X(t) \). Then

\[
f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_X(t) \, dt
\]

at every point \( x \) at which \( f \) is differentiable.
6 Two Probability Theorems

6.1 The Law of Averages

We motivated expectation by looking at the running average \( \bar{X}_n = (X_1 + \cdots + X_n)/n \) of iid rvs. As \( n \to \infty \), it was stated that \( \bar{X}_n \to E(X) \). The law of averages makes this statement explicit.

**Definition:** A sequence of random variables \( Y_1, Y_2, \ldots \) converges in mean square to the rv \( Y \) if

\[
E((Y_n - Y)^2) \to 0 \text{ as } n \to \infty.
\]

**Example:** Let \( Y_n \) be a Bernoulli rv with success probability \( 1/n \). Let \( Y \) be the constant rv 0. Then

\[
E((Y_n - Y)^2) = E(Y_n^2) = 0^2(1 - 1/n) + 1^2(1/n) = 1/n \to 0,
\]

so \( Y_n \to Y \) in mean square.

**Example:** Let \( Y_n \) take the value \( n \) with probability \( 1/n \) and 0 with probability \( 1 - 1/n \) and \( Y \) as before. Then

\[
E((Y_n - Y)^2) = E(Y_n^2) = 0^2(1 - 1/n) + n^2 \times 1/n = n \to \infty,
\]

so \( Y_n \) does not tend to \( Y \) in mean square (even though \( P(Y_n \neq 0) \to 0) \)

**Theorem:** (Law of Large Numbers). Let \( X_1, X_2, \ldots \) be a sequence of independent random variables each having mean \( \mu \) and variance \( \sigma^2 \). Then

\[
\frac{1}{n}(X_1 + \cdots + X_n) \to \mu
\]

in mean square.

**Proof:** Let \( S_n = \sum_1^n X_i \). So we are interested in what happens to \( S_n/n \). We have

\[
E(S_n/n) = \frac{1}{n}E(S_n) = \frac{1}{n}(E(X_1 + \cdots + X_n) = \frac{1}{n}(E(X_1) + E(X_2) + \cdots E(X_n)) = \mu.
\]

and so

\[
E \left( \left( \frac{1}{n}S_n - \mu \right)^2 \right) = \text{Var}(S_n/n) = \frac{1}{n^2}\text{Var}(X_1 + X_2 + \cdots X_n)
\]

\[
= \frac{1}{n^2}(\text{Var}(X_1) + \cdots + \text{Var}(X_n)) = \frac{1}{n^2}n\sigma^2 = \sigma^2/n \to 0,
\]

thus \( S_n/n \) converges in mean square to \( \mu \) as reqd.

6.2 The Central Limit Theorem

We know that \( E(S_n) = n\mu \). We know that in the limit \( S_n \) is converging to \( n\mu \) in the sense that \( S_n/n \to \mu \) in expected mean square. The next problem is to ask how much variation there is about this mean for \( S_n \) as \( n \to \infty \). The central limit theorem tells us this for iid rvs.

To start, we define

\[
Z_n = \frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}}
\]

to be the standardised version of \( S_n \). Note that \( Z_n = a_nS_n + b_n \) for constants \( a_n = 1/\sqrt{\text{Var}(S_n)} \) and \( b_n = -a_nE(S_n) \), chosen so that \( E(Z_n) = 1 \) and \( \text{Var}(Z_n) = 1 \).

Next, since \( E(S_n) = n\mu \) and \( \text{Var}(S_n) = n\sigma^2 \) (see proof of LLN), we can write

\[
Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.
\]
Theorem: (Central Limit Theorem). Let $X_1, X_2, \ldots$ be iid rvs having mean $\mu$ and variance $\sigma^2$. The standardised sum

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}},$$

where $S_n = X_1 + \cdots + X_n$, satisfies, as $n \to \infty$,

$$P(Z_n \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du,$$

for $z \in \mathbb{R}$. In other words, the distribution of $Z_n$ is converging point-wise (for each $z \in \mathbb{R}$) to the standard normal distribution.

Note: This is a remarkable result! Not only does the distribution of $Z_n$ converge to something, it ALWAYS converges to the standard normal, NO MATTER WHAT the distribution of the $X_i$. The only requirement is that $X_i$ have a mean and variance.

Note: Since $Z_n$ is approximately $N(0, 1)$, by the properties of the normal distribution, we have that $S_n = n\mu + \sigma\sqrt{n}Z_n$ is approximately $N(n\mu, n\sigma^2)$.

Proof: Let $U_i = X_i - \mu$, so that $E(U_i) = 0$ and $\text{Var}(U_i) = \sigma^2$. The mgf of the $U_i$ are $M_U(t) = e^{-\mu M_X(t)}$.

Now $Z_n = \sum_{i=1}^{n} U_i/(\sigma\sqrt{n})$, so the mgf of $Z_n$ is

$$M_{Z_n}(t) = E(e^{tZ_n}) = E \left( \exp \left( \frac{t}{\sigma\sqrt{n}} \sum_{i=1}^{n} U_i \right) \right) = \prod_{i=1}^{n} E \left( \exp \left( \frac{t}{\sigma\sqrt{n}} U_i \right) \right) = M_U \left( \frac{t}{\sigma\sqrt{n}} \right)^n,$$

by properties of the mgf.

Now recall the theorem that defines the mgf as a power series expansion about 0:

$$M_U(x) = 1 + xE(U_1) + 0.5x^2E(U_1^2) + o(x^2) = 1 + 0.5\sigma^2x^2 + o(x^2),$$

since $E(U_1) = 0$ and $E(U_1^2) = E(U_1^2) - E(U_1)^2 = \text{Var}(U_1) = \sigma^2$.

Thus

$$M_{Z_n}(t) = M_U \left( \frac{t}{\sigma\sqrt{n}} \right)^n = \left[ 1 + 0.5\sigma^2 \left( \frac{t}{\sigma\sqrt{n}} \right)^2 + o \left( \left( \frac{t}{\sigma\sqrt{n}} \right)^2 \right) \right]^n = \left[ 1 + 0.5 \frac{t^2}{n} + o(1/n) \right]^n,$$

As $n \to \infty$, the above $\to \exp(t^2/2)$ (Recall that $\lim_{n \to \infty} (1 + x/n)^n = e^x$). This is the mgf of the standard normal. By uniqueness theorem for mgfs, this means that the distribution must be the standard normal.

Example: A fair coin is thrown 20000 times. Let $S$ be the number of heads. Find an approximation to $P(9900 < S < 10010)$.

$S$ is in fact binomial(20000,0.5). But, we can write $S = \sum_{i=1}^{20000} X_i$, where $X_i = 1$ is Bernoulli(0.5), taking the value 1 for a H and 0 for T. The mean and variance of $X_i$ exists: $\mu = E(X_i) = 0.5$ and $\sigma^2 = \text{Var}(X_i) = 0.25$. We can apply the CLT to say that for large $n$, $S$ is approximately $N(n\mu, n\sigma^2) = N(10000, 5000)$. Thus

$$P(9900 < S < 10010) = P((9900-10000)/\sqrt{5000} < (S-10000)/\sqrt{5000} < (10010-10000)/\sqrt{5000})$$

$$= P(-1.41 < Z < 0.141) = P(Z < 0.141) - P(Z < -1.41) = 0.5557 - (1 - P(Z < 1.41))$$

$$= 0.5557 - 1 + 0.9207 = 0.4764.$$

Note that the actual probability using the true binomial distribution is 0.4793.

HANDOUT: Tutorial 5

HANDOUT: Sample Exam Questions