# MATH 323: Lecture Notes

Ryan Ordille

April 13, 2012

These course notes are for McGill University's MATH 323: Probability course, offered in the Winter 2012 semester, taught by David Wolfson. These notes are simply what the instructor was writing on the board, and may contain errors. The notes are written by me, Ryan Ordille, and no copyright infringement is intended.

The notes up until 28 February (first class after the break and midterm) are concise – without review or obvious proofs. Most pre-midterm examples (especially the "nuts and bolts" examples) will be left out since, when these notes were typed up towards the end of the course, the solutions are obvious.

# 1 17 January

## 1.1 Axioms and Theorems

1. $P(E) \geq 0$
2. $P(S) = 1$
3. $\forall E_i \cap E_J = \emptyset$:
$$P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$$
4. **Theorem 1:** $P(\emptyset) = 0$
5. **Theorem 2:** $P(E^c) = 1 - P(E)$
6. **Theorem 3:** $P(E \cap F^c) = P(E) - P(E \cap F)$

7. **Theorem 4:** If $E \subset F$, then $P(E) \leq P(F)$.
8. **Theorem 5:** If $E$ and $F$ are *any* two events, then $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

Word problem hints:

- "either/or" corresponds to a union of two events
- "at least one" – union
- "not" – complement
- "and" – intersection
- "proportion" – "a probability statement about an individual"

# 2 19 January

## 2.1 Using counting methods to compute probabilities

**Theorem 6:** Let $S$ be a sample space (set of all possible outcomes) with finitely many outcomes $N$. If all these outcomes are equally likely, then, if $E$ is any event,

$$P(E) = \frac{\text{number of ways that } E \text{ can occur}}{\text{total number of possible outcomes} = N}$$

Counting the number of ways that $E$ can occur (and $N$) can be difficult without the following counting rules.

### 2.1.1 Factorial

By definition,

$$n! = n \times (n-1) \times \cdots \times 2 \times 1 \text{ where } 0! = 1$$

### 2.1.2 Multiplication rule for counting

Suppose you have $k$ sets of distinct objects of size $n_1, n_2, \ldots, n_k$ respectively. Then, the number of ways to choose one object from each set is $n_1, n_2, \ldots, n_k$.

### 2.1.3 $n$ choose $k$

Suppose that a set contains $n$ distinct objects. Then, the number of ways to select $k$ objects from these $n$ objects, if we sample *without replacement*, is denoted by $\binom{n}{k}$ ("$n$ choose $k$"). It turns out that:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Note that here the order of selections is not important, e.g. the selection $(A, B)$ is equivalent to the selection $(B, A)$.

### 2.1.4 Permutations

If the order is important in the previous example, this is denoted by $P_k^n$, and

$$P_k^n = \frac{n!}{(n-k)!}$$

# 3 24 January

This lecture just contained examples of the previous counting rules.

# 4 26 January

## 4.1 Conditional Probabilities

Very often, if we know that some event $A$ has occurred, then this will affect the probability that the event $B$ will occur.

**Definition:** for two events $A$ and $B$, with $P(A) \neq 0$, we define the probability of "$B$ given $A$", denoted by $P(B \mid A)$, as

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

If $P(A) = 0$, then $P(B \mid A)$ can be defined arbitrarily, so we consider it to be undefined.

Notice that the right-hand side of the definition is given in terms of already-defined quantities – straightforward, unconditional probabilities.

### 4.2 Notes

**(1)** Conditional probabilities satisfies the three axioms, just like unconditional probabilities.

**(2) (the multiplication rule for conditional probabilities)** Let $A$ and $B$ be two events. Then,

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

**(3)** When solving conditional probability problems, use the *conditioning technique* to break a long intersection into a series of conditional and unconditional probabilities.

## 5 31 January

Note that it's not true in general that $P(A \mid (B_1 \cup B_2)) = P(A \mid B_1) + P(A \mid B_2)$ even if $B_1 \cap B_2 = \emptyset$. However, it is true that $P(B_1 \cup B_2 | A) = P(B_1|A) + P(B_2|A)$.

### 5.1 The Law of Total Probability

Let $A$ be any event and let $B_1, B_2, \ldots, B_m$ be a collection of $m$ events satisfying:

1. $B_i \cap B_j = \emptyset \; \forall i \neq j$
2. $\cup_{i=1}^{m} B_i = S$ (i.e. $B_1, B_2, \ldots B_m$ form a partition of $S$)

Then,

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

Notice here that the left-hand side ($P(A)$) may be difficult to find directly, while the components of the right-hand side might be known or easy to find.

### 5.2 Bayes' Theorem

Let $A$ be any event. Let $B_1, B_2, \ldots, B_m$ form a partition of $S$. Then, for every $k = 1, 2, \ldots, m$,

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

## 6 02 February

### 6.1 Independence

Sometimes, knowing that an event $A$ has occurred will not affect the probability that $B$ will occur. In such a situation, we say that $A$ and $B$ are *independent*. More formally, two events $A, B$ are said to be independent ($A \perp B$) if and only if

$$P(B \mid A) = P(B) \text{ or } P(A \mid B) = P(A)$$

**Theorem:** if $A$ and $B$ are disjoint, then $A$ and $B$ can only be independent if either $P(A) = 0$ or $P(B) = 0$.

There's another important (albeit less intuitive) definition of independence: $A \perp B$ if and only if $P(A \cap B) = P(A)P(B)$.

More generally, events $A_1, A_2, \ldots, A_n$ are *mutually independent* if and only if, for every subset $A_{i1}, A_{i2}, \ldots, A_{ik}$ of $A_1, A_2, \ldots, A_n$,

$$P(\cap_{j=1}^{k} A_{ij}) = \prod_{j=1}^{k} P(A_{ij})$$

It follows that, if $A \perp B$, then $P(A \cup B) = P(A) + P(B) - P(A)P(B)$.

In general, sampling without replacement assumes dependence, while sampling with replacement assumes independence. Disjointness is entirely a set property, and should not be confused with independence.

# 7 07 February

## 7.1 Random variables

Often, we are not so interested in the outcomes of an experiment themselves, but rather in numerical values that can be associated with these outcomes.

**Definition:** a function $X$ that maps the sample space $S$ to the real line in such a way that, for every $\omega \in S$, $X(\omega)$ is a real number, is called a *random variable* (or r.v. for short).

Note that two or more distinct $\omega$'s can give the same value of $X(\omega)$. However, one value of $\omega$ is not allowed to give two different values of $X(\omega)$, as this would not make $X$ a function.

The term "random variable" comes about because the outcomes of the experiment are random or uncertain before we perform our experiment, and hence the value of $X$ will also be uncertain before the experiment.

We denote random variables with capitals, and the values of random variables after experiments with lowercase letters.

# 8 09 February

## 8.1 Random variables continued

### 8.1.1 The cumulative distribution function

We define $P(x \in B)$ to be $P(w \in S : x \in B : P(X^{-1}(B))$ (i.e. we refer back to the events of $S$ to find the probabilities of events on the real line).

**Definition:** $P(X \in (-\infty, x]) = P(X \leq x)$ is a function of $X$ called the *cumulative distribution function* (or c.d.f.) of the random variable $X$, and

$$F_X(x) = P(X \leq x) \forall x \in (-\infty, \infty).$$

### 8.1.2 Properties of the c.d.f.

**(1)** The c.d.f. is a real-valued function of $x$.

**(2)** In order to specify $F_X$, we need to specify $F_X(x)\ \forall x \in (-\infty, \infty)$.

**(3)** All c.d.f.s are non-decreasing and right-continuous.

**(4)** $F_X(-\infty) = \lim_{x\to-\infty} F_X(x) = 0$ and $F_X(\infty) = \lim_{x\to\infty} F_X(x) = 1$.

### 8.1.3 Continuous vs discrete random variables

**Definition:** We call a random variable *continuous* if its c.d.f. is a real-valued function of $x$ (i.e. it has no jumps). If a random variable is not continuous, the random variable is said to be *discrete*, and can assume at most a countable number of distinct values.

**Definition:** For a discrete random variable $X$, the real-valued function of $x$ specified by $P_X(x) = P(X = x)\ \forall x$ that $X$ can assume is called the *probability function* of $X$.

**Theorem:** If we know $P_X(x)\ \forall x$ that $X$ can assume, then we can find $F_X(x)\ \forall x \in (-\infty, \infty)$. Conversely, if we're given the c.d.f., we can find the probability function.

# 9 14 February

## 9.1 Named probability distributions

### 9.1.1 Discrete uniform distribution

**Definition:** $X$ has a *discrete uniform distribution* on the set of $N$ real numbers $a_1 < a_2 < \ldots, a_N$ if $P(X = a_i) = \frac{1}{N}\ \forall i = 1, 2, \ldots, N$.

The total probability (mass) is equally or uniformly spread out on each of the numbers $a_i$.

### 9.1.2 Bernoulli distribution

**Definition:** a random variable $X$ has a *Bernoulli distribution* with parameter $p$ if $P(X = 1) = p$ and $P(X = 0) = 1 - p = q$.

This is used as a "building block" for more complicated random variables.

**Starting from after the midterm and break:**

# 10 28 February

## 10.1 Random variable distributions

Cumulative distribution function (c.d.f.): $F_X(x) = P(X \leq x)$

For discrete random variables: $P(X = x) = P_X(x)$

### 10.1.1 Binomial distribution

The random variable $X$ has a *binomial distribution* with parameters $n$ and $p$ if

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

for $x = 0, 1, \ldots, n$, where $n$ is a non-negative integer and $0 \leq p \leq 1$.

Note that we should check if:

1. $P_X(x) \geq 0$ (obvious)
2. $\sum_{\text{all x in range}} P_X(x) = 1$ (easy to check)

We write $X \sim Bin(n, p)$ to mean "$X$ has the binomial distribution".

**How the binomial distribution arises:** (the binomial setup)

Firstly, we have a sequence of $n$ independent trials – that is, the outcomes of these trials are mutually independent.

Secondly, each trial can result in exactly one of two possible outcomes: a "success" (S) or a "failure" (F). We call such trials *Bernoulli trials*.

Thirdly, the probability of success at trial $i$ is constant and equal to $p$ for every $i = 1, 2, \ldots, n$. For example, in a coin toss, we cannot change the probability of heads halfway through, so the probability of success is constant.

**Theorem:** Let $X$ be the number of successes observed in these $n$ trials. Then,

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, \ldots, n.$$

**Proof:** (same idea for the genetic mutation example) First, note that the probability of any particular configuration in which there are $x$ successes (and $n-x$ failures) is just $p^x(1-p)^{n-x}$.

E.g:

$$P(S_1 \cap S_2 \cap \ldots \cap S_x \cap F_{x+1} \cap F_{x+2} \cap \ldots \cap F_n) = P(S_1)P(S_2)\ldots P(S_x)P(F_{x+1})\ldots P(F_n)$$

Assuming these are all independent. This is equal to

$$p \times p \times \ldots \times p \times (1-p) \times \ldots \times (1-p) = p^x(1-p)^{n-x}$$

But,

$$\{X = x\} = \cup_{\text{all possible configurations}}\{\text{configuration } i \text{ with } x \text{ successes}\}$$

Note that the configurations are disjoint, so therefore, by axiom 3:

$$P(X = x) = \sum_{\text{all configurations}} P(\text{configuration } i) = \sum p^x(1-p)^{n-x}$$

Since each configuration is defined by a choice of $x$ objects from $n$ total objects, then for every $x = 0, 1, \ldots, n$:

$$P(X = x) = \binom{n}{x} p^x(1-p)^{n-x}$$

### 10.1.2 Binomial distribution example

Suppose that the five year survival probability for lung cancer is .10. If thirty people with lung cancer are sampled, what is the probability that at least three will survive five years or longer?

**Solution:** Let $Y =$ the number out of thirty who will survive five or more years. We shall reasonably assume the binomial setup, with $n = 30$ and $P(S_i) = .10$, where $S_i$ is the event where the $i$th subject survives five or more years. Therefore,

$$P(Y \geq 3) = \sum_{y=3}^{30} \binom{30}{y} (.10)^y (1-.10)^{30-y} = 1 - \sum_{y=0}^{2} \binom{30}{y} (.10)^y (1-.10)^{30-y}$$

*Important note*: do not immediately have the "burning desire" to use the binomial distribution as soon as you see a bunch of trials, each of which can result in exactly one of two outcomes. You must check if the trials are independent – they will not be if you are sampling without replacement!

Observe that the Bernoulli distribution is a special case of the binomial distribution, where $n = 1$. So, $P_X(x) = p^x(1-p)^{1-x}$ for $x = 0, 1$.

### 10.1.3 Poisson distribution

The random variable $X$ is said to have a *Poisson distribution* with parameter $\lambda > 0$ if

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x = 0, 1, 2, \ldots$$

(Note the infinite range of $x$)

Check that:

1. $P_X(x) \geq 0$ – obvious
2. $\sum_0^\infty P_X(x) = 1$ – Taylor series expansion for $e^\lambda$

The Poisson distribution arises as an approximation to the binomial for "large $n$ and small $p$".

**Theorem:** Let $X \sim Bin(n, p)$. The limit of $P(X = x)$ as $n$ tends to infinity and $p$ tends to 0, in such a way that $n \times p$ is constant $(= \lambda)$, is

$$\frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x = 0, 1, 2, \ldots$$

(e.g. $p = \frac{6}{n} : np = n(\frac{6}{n}) = 6$)

**Proof:** We shall use the following result in proving our theorem.

$$\lim n \to \infty (1 + \frac{a}{n})^n = e^a$$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, \ldots, n$$

$$(*) = \frac{n!}{x!(n-x)!} p^x (1-p)^n (1-p)^{-x}$$

Now since $\lambda = np$, we have $p = \frac{\lambda}{n}$. Then, $(*)$ becomes:

$$\frac{n!}{x!(n-x)!}\frac{\lambda^x}{n^x}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-x}$$

After some algebraic manipulation, this becomes:

$$\frac{1}{x!}\frac{n(n-1)(n-2)\dots(n-x+1)}{nn\dots n}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-x}\lambda^x$$

Notice that the second term has $x$ terms on both the top and the bottom. Now, let $n$ go to infinity to get the limit:

$$\frac{\lambda^x}{x!}e^{-\lambda} \blacksquare$$

# 11 01 March

## 11.1 Poisson distribution continued

### 11.1.1 Poisson distribution example

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ where } x = 0, 1, 2, \dots$$

Suppose that in a book of 1000 pages, on any particular page, there can be either zero errors or one error. Suppose further that the probability of an error on any particular page is .002. what is the approximate probability that there will be at most three errors in the book?

**Exact solution:** There is a binomial setup – errors are likely to occur independently amongst the $n = 1000$ trials. Each trial can result in either a success (an error is found) or a failure (an error is not found). There is also a constant probability of success (.002) for every trial. Let $X$ be the number of successes in 1000 pages. Then $X \sim Bin(1000, .002)$.

$$P(X \leq 3) = \sum_{x=0}^{3}\binom{1000}{x}(.002)^x(1-.002)^{1000-x}$$

**Approximate solution:** Since $n$ is large and $p$ is small, we can use the Poisson approximation with $\lambda = n \times p = 1000 \times \frac{2}{1000} = 2$. Therefore:

$$P(X \leq 3) = \sum_{x=0}^{3} P(X = x) = \sum_{x=0}^{3} \frac{2^x e^{-2}}{x!}$$

If $X$ has a Poisson distribution with parameter $\lambda$, we write $X \sim Po(\lambda)$.

## 11.2 The Hypergeometric Distribution

The random variable X has a hypergeometric distribution with parameters $N$, $a$, and $n$ if:

$$P(X = x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}$$

(for $x \leq a$ and $n - x \leq N - a$).

The hypergeometric distribution discusses the probability of $a$ successes in $n$ draws of $N$ objects.

## 11.3 The Geometric Distribution

$X$ is said to have a geometric distribution with parameter $p$ if

$$P(X = x) = P_x(x) = (1-p)^{x-1}p \text{ for } x = 1, 2, \ldots.$$

$$p\sum_{x=1}^{\infty}(1-p)^{x-1} = \frac{p}{1-(p-1)} = 1$$

The geometric random variable is used to describe or model the trial number at which the first success occurs in a sequence of independent Bernoulli trials, each with the probability of success $p$.

## 11.4 Expected values

The probability distribution of a random variable provides the complete story about the random variable. There is no information about a random variable once we

know its probability distribution. However, we often wish to summarize a probability distribution. The two most common summaries are:

1. a parameter that discusses the "centre of distribution" and
2. a parameter that discusses how spread out the values are from the centre.

To this end, we give the following definition:

**Expected value:** let $Y$ be a discrete random variable with the probability function $P_Y(y)$. Then, we define the expected value of $y$ denoted by $E(Y)$ to be

$$E(Y) = \sum_{\text{all } y} yP_Y(y) = \sum_{\text{all } y} yP(Y = y)$$

# 12 06 March

## 12.1 Review

Recall: given a random variable $Y$, we define the expected value or expectation of $Y$ denoted by $E(Y)$ to be:

$$E(Y) = \sum_{\text{all } y} yP(Y = y)$$

provided that the sum is finite.

## 12.2 Theory

Notes:

1. $E(Y)$ is often denoted by $\mu_Y$, and also is termed the *mean* of $Y$ (also called the "population mean").
2. **Interpretation:** $E(Y)$ is a weighted average or mean of the possible values of the random variable $Y$, where the weights are the probabilities of these values. For example, in the special case where $Y$ has a discrete uniform distribution

$a_1, a_2, \ldots, a_N$, then

$$E(Y) = \sum_{i=1}^{N} a_i P(Y = a_i) = \frac{1}{N} \sum_{i=1}^{N} a_i$$

So think of $\mu$ as the "average value" of $Y$.

3. $\mu$ is a constant, a parameter specific to a given probability distribution. You need the distribution to compute $\mu$.
4. $E(cY) = cE(Y)$ where $c$ is a constant, and $E(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} E(Y_i)$ (the proof will come later)
5. Note, however, that in general, $E(XY) \neq E(X)E(Y)$.
6. Let $g$ be some real-valued function of a random variable $Y$. Then, if $X = g(Y)$, we have
$$E(X) = E(g(Y)) = \sum_{\text{all } y} g(y) P(Y = y)$$
The point of this: by definition, $E(g(Y)) = E(X) = \sum_{\text{all } x} xP(X = x)$. Thus, in order to find $E(g(Y))$, it would appear that we first need to find the probability distribution of $X = g(Y)$. We'll see such transformations later – they can be difficult. You do not have to first find the distribution of $X$ – you can use the distribution of the original $Y$ and sum $g(y)P(Y = y)$.

In particular, if $g(Y) = Y^k$, we can call $E(g(Y)) = E(Y^k)$ is called the $k$th moment of $Y$. $\mu$ is called the first moment of $Y$.

$$E(Y^k) = \sum_{\text{all } y} y^k P(Y = y)$$

Of particular importance is a special function of $Y$:

$$g(Y) = (Y - \mu_Y)^2$$

In this case, $E(g(Y)) = E((Y - \mu_Y)^2)$ is called the *variance* of $Y$, and is denoted by $Var(Y)$, also $\sigma_Y^2$. This gives the average squared distance between the values of $Y$ and its mean. It is a measure of spread or variation of $Y$ (i.e. its distribution). We call $\sqrt{\sigma_Y^2}$ the *standard deviation* of $Y$. This is more convenient than $Var(Y)$, since it is in the same units as $Y$, unlike $Var(Y)$.

**Result:**

$$Var(Y) = E((Y - \mu_Y)^2) = E(Y^2) - \mu_Y^2$$

Proof:

$$\begin{aligned} E((Y - \mu_Y)^2) &= E(Y^2 - 2\mu_Y Y + \mu_Y^2) \\ &= E(Y^2) - E(2\mu_Y Y) + E(\mu_Y^2) \\ &= E(Y^2) - 2\mu_Y E(Y) + \mu_Y^2 \\ &= E(Y^2) - 2\mu_Y^2 + \mu_Y^2 \\ &= E(Y^2) - \mu_Y^2 \ \blacksquare \end{aligned}$$

*Final note:* $Var(cY) \neq c \times Var(Y)$, but $Var(cY) = c^2 \times Var(Y)$.

## 12.3 Examples

### 12.3.1 Example 1

**The calculation of insurance premiums:** An insurance company will insure your computer against theft for \$1000. It is known with a probability .05 that your computer will be stolen. What premium should the insurance company charge so that its expected gain is 0?

**Solution:** Let $c$ be the required premium. Let $Y$ be the gain of the company in a given year. We need to find the value of $c$ such that $E(Y) = 0$.

First, we need $P_Y(y)$ for all $y$. We have $P(Y = c) = .95$ (where the computer was not stolen) and $P(Y = (c - 1000)) = .05$ (where the computer was stolen). Therefore,

$$E(Y) = c \times .95 + (c - 100) \times .05$$

Setting $E(Y) = 0$ and solving for $c$, we get $c = 50$. Thus, if the company charges \$50 for the policy, on average, over a large number of clients, they would neither lose nor gain money.

### 12.3.2 Example 2

**"Nuts and Bolts" Example:** suppose that the random variable $x$ has the probability distribution:

$$\begin{aligned} P(X = -1.2) &= .32 \\ P(X = 2.6) &= .40 \\ P(X = 0) &= .28 \end{aligned}$$

Find $E(X)$ and $Var(X)$.

$$E(X) = -1.2 \times .32 + 0 \times .28 + 2.6 \times .40 = .66 = \mu_X$$

$$\begin{aligned} Var(X) &= E(X^2) - \mu_X^2 = \sum x^2 P(X = x) \\ E(X^2) &= (-1.2)^2 \times .32 + 0^2 \times .28 + (2.6)^2 \times .40 \\ &= 3.1648 \\ \therefore Var(X) &= 3.1648 - (.66)^2 \\ &= 2.7292 \\ \text{Also, } \sigma &= \sqrt{2.7292} = 1.652 \end{aligned}$$

# 13 08 March

## 13.1 Summary

Centre – $E(X) = \mu$

Spread – $Var(x) = \sigma^2$

Standard Deviation – $\sqrt{\sigma^2} = \sigma$

$$\sum (x - \mu)^2 P(X = x) = \sum x^2 P(X = x) - \mu^2$$

## 13.2 The Mean and Variance of Some Named Distributions

### 13.2.1 Binomial

$$\begin{aligned}
E(X) = \mu &= \sum_{x=1}^{n} x\binom{n}{x}p^x(1-p)^{n-x} \\
&= \sum \frac{n(n-1)!}{(x-1)!(n-1-(x-1))!}pp^{x-1}(1-p)^{n-1-(x-1)} \text{ as } (n-x)! = (n-1-(x-1))! \\
&= np\sum\binom{n-1}{x-1}p^{x-1}(1-p)^{n-1-(x-1)} \\
&= np\sum_{y=0}^{n-1}\binom{n-1}{y}p^y(1-p)^{n-1-y} \\
&= np
\end{aligned}$$

The sum above is equal to 1 since it is just a sum of $n-1$ $Bin(n-1,p)$ probabilities.

To find $Var(X)$, we first have to find $E(X^2)$. By definition:

$$E(X^2) = \sum_{x=1}^{n} x^2\binom{n}{x}p^x(1-p)^{n-x}$$

Note that $x^2$ will not cancel with the leading terms of $x!$ as before, so we have to use a trick. We first calculate $E(X(X-1))$, which is easy. Then, notice that

$$E(X(X-1)) = E(X^2) - E(X) = E(X^2) - \mu$$

So, $E(X^2) = E(X(X-1)) + \mu$. Finally, $Var(X) = E(X(X-1)) + \mu - \mu^2$.

$$\begin{aligned}
E(X(X-1)) &= \sum_{x=2}^{n} x(x-1)\binom{n}{x}p^x(1-p)^{n-x} \\
&= n(n-1)p^2\sum\frac{(n-2)!}{(x-2)!(n-2-(x-2))!}p^{x-2}(1-p)^{n-2-(x-2)} \\
&= n(n-1)p^2\sum\binom{n-2}{x-2}p^{x-2}(1-p)^{n-2-(x-2)} \\
&= n(n-1)p^2
\end{aligned}$$

$$Var(X) = n(n-1)p^2 + np - n^2p^2 = np(1-p)$$

### 13.2.2 Bernoulli

In particular, the mean and variance of a Bernoulli random variable are $p$ and $p(1-p)$ respectively (because $n = 1$).

### 13.2.3 Poisson

$$\begin{aligned} E(X) &= \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda}\lambda \sum_{x-1=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda \end{aligned}$$

So, the mean of a Poisson random variable is just its parameter $\lambda$.

For the variance, first find $E(X(X-1))$. We find $Var(X) = \lambda$, the same as the mean.

### 13.2.4 Geometric

$$\begin{aligned}
E(X) &= \sum_{x=1}^{\infty} xp(1-p)^{x-1} \\
\text{using our trick...} \quad &= p\sum_{x=1}^{\infty} x(1-p)^{x-1} \\
&= p\sum_{x=1}^{\infty} -\frac{d}{dp}(1-p)^{x} \\
&= -p\frac{d}{dp}\sum_{x=1}^{\infty}(1-p)^{x} \text{ we can interchange the derivative and the sum} \\
&= -p\frac{d}{dp}\frac{1-p}{1-(1-p)} \quad \text{where } (1-p) = r \\
&= -p\frac{d}{dp}(\frac{1}{p} - 1) \\
&= \frac{1}{p}
\end{aligned}$$

For the variance, first find $E(X(X-1))$ (2 derivatives), and then you can find the variance.

$$Var(X) = \frac{1-p}{p^2}$$

## 13.3 Continuous probability distributions

**Definition:** a random variable $X$ with c.d.f. $F_X$ is said to be continuous if $F_X$ is continuous for all $-\infty < x < \infty$.

The continuous c.d.f.s split into two types:

1. the so-called "absolutely continuous" c.d.f.s (essentially, they are differentiable) and
2. the so-called "singular" c.d.f.s (without derivatives).

From now on in this course, we'll assume for continuos c.d.f.s $F_X$, that they are differentiable for all $-\infty < x < \infty$.

It follows that if $X$ is continuous, then $P(X = x) = 0$ for every $x$. Consider the "area under a curve" analogy.

Remember (as $F_X(x) = P(X \leq x)$):

$$P(X = x) = F_X(x) - P(X < x)$$

Hence, we cannot specify a continuous random variable by specifying the values $P(X = x)$ for all $x$ that $X$ can assume, as we did in the discrete case. Instead, we introduce an analogue of the probability function known as the *probability density function* (p.d.f.). It will turn out that the p.d.f. also uniquely determines the probability distribution.

**Definition:** A real-valued function $f_x$ is said to be the probability density function of a random variable $X$ if:

1. $f_x(x) \geq 0$ for all $-\infty < x < \infty$ and
2. $P(X \in A) = \int_A f_x(x)dx$ (i.e. $f_x$ has the property that when you integrate it over a set, you get the probability of the set).

# 14 13 March

## 14.1 The probability density function

**Definition:** the probability density function (p.d.f.) of a random variable $X$ is any function $f_X(x)$ such that:

1. $f_X(x) \geq 0 \; \forall -\infty < x < \infty$ and
2. $P(X \in A) = \int_A f_X(s)dx \; \forall$ events $A \in [R]$

A p.d.f. gives $P(X \in A)$ by integrating over $A$.

## 14.2 Notes on the p.d.f.

**(1)** In particular, if $A$ is of the form $(-\infty, x]$ then

$$P(X \in A) = P(X \leq x) = \int_{-\infty}^{x} f_X(y)dy$$

In short, $F_X(x) = \int_{-\infty}^{x} f_X(y)dy$.

**(2)** Conversely, we can recover the p.d.f. from the c.d.f. by the Fundamental Theorem of Calculus, since:

$$\frac{d}{dx}F_X(x) = F_X'(x) = f_X(x) \forall x$$

Thus, in particular, if $f_X$ is a p.d.f., then:

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

**(3)** Interpretation of a p.d.f: We have, for small $\Delta x$, that:

$$\frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} \approx f_X(x)$$

So it follows that:

$$F_X(x + \Delta x) - F_X(x) \approx \Delta x f_X(x)$$

But the left hand side is just $P(x < X \leq x + \Delta x)$. Finally, we have that $f_X(x)\Delta x$ is approximately the probability that $X$ lies in $(x, x + \Delta x]$.

Note that, because the p.d.f. does not represent a probability on its own, it can be greater than 1 or less than 0. When multiplied by a small $\Delta x$, it gives an *approximate* probability.

Any function whose total area equals 1 can qualify as a p.d.f., even if some points are larger than 1.

**(4)** For continuous random variables, we define:

$$E(g(X)) = \sum_{\text{all}x} g(x)P_X(x) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

As before, when $g(X) = X^k$, then we refer to the $E(g(X)) = E(X^k)$ as the $k$th moment. Of particular importance are the first moment ($k = 1$) and the second moment ($k = 2$). Again, as before, we call the first moment the *mean* or the *expected value* of $X$.

So, by definition:

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

and

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_x(x) dx)$$

Finally, as before:

$$Var(X) = E((X - \mu_X)^2) = \int_{\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx = E(X^2) - \mu_X^2$$

and

$$Var(X) = \int_{-\infty}^{\infty} x^2 f_X(x) dx - (\int_{-\infty}^{\infty} x f_X(x) dx)^2$$

Note: watch out for a p.d.f. that may change its form in different ranges of $(-\infty, \infty)$ when carrying out an integration.

## 14.3 Examples

### 14.3.1 Example 1

Let $f_X(x) = c(x^2 + 1)$ for $0 < x < 1$ and $f_X(x) = 0$ elsewhere ($c$ is a constant).

1. Find $c$.
2. Find $P(.25 < X \leq .50)$.
3. Find $P(.25 < X < .50)$.
4. Find $F_X$.
5. Find $E(X)$ and $\sigma_X$.

**(1)** Since $\int_{-\infty}^{\infty} f_X(x) dx = 1$, we must have:

$$\int_{-\infty}^{0} 0 dx + \int_{0}^{1} c(x^2 + 1) dx + \int_{1}^{\infty} 0 dx = 1$$

This gives $c = .75$.

**(2)**

$$P(.25 < X \leq .50) = \int_{.25}^{.50} (.75)(x^2+1)dx = \frac{55}{256}$$

**(3)** For a continuous p.d.f., these two values are the same (by rules of integration).

$$P(.25 < X < .50) = P(.25 < X \leq .50) = \frac{55}{256}$$

**(4)**

$$\begin{aligned}
F_X(x) &= 0 \ \forall x \leq 0 \\
F_X(x) &= \int_{-\infty}^{x} f_X(y)dy \ \forall 0 < x < 1 \\
&= \int_{-\infty}^{0} 0dy + \int_{0}^{x} (.75)(y^2+1)dy \\
&= (.75)(\frac{x^3}{3} + x) \\
F_X(x) &= 1 \ \forall x \geq 1
\end{aligned}$$

**(5)**

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f_X(x)dx \\
&= \int_{0}^{1} x(.75)(x^2+1)dx \\
&= (.75)(\frac{x^4}{4} + \frac{x^2}{2})\Big|_0^1 \\
&= (.75)(.25 + .50) \\
&= \frac{9}{16}
\end{aligned}$$

To find $\sigma_X$, first find:

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{0} 0dx + \int_{0}^{1} x^2(.75)(x^2+1)dx + \int_{1}^{\infty} 0dx \\ &= (.75)(\frac{x^5}{5} + \frac{x^3}{3})\Big|_0^1 \\ \sigma^2 &= Var(X) \\ &= E(X^2) - (\frac{9}{16})^2 \\ \sigma &= \sqrt{\sigma^2} \end{aligned}$$

# 15 15 March

## 15.1 Named continuous probability distributions

### 15.1.1 The uniform distribution

**Definition:** The random variable $X$ is said to be uniformly distributed on the interval $[a, b]$ if

$$f_X(x) = \int_a^x \frac{1}{b-a}\, dx \text{ for } a \leq x \leq b$$

and $f_X(x) = 0$ elsewhere.

**Notes:**

1. The p.d.f. is constant on $[a, b]$. On the graph, the height is constantly $\frac{1}{b-a}$, so the area under the curve is 1. The probability is uniformly spread out in the interval.
2. The uniform distribution is often used to model situations in which we believe outcomes occur completely at random.
3. An important special case is the uniform distribution $[0, 1]$.
4. Notation – we write $X \sim U(a, b)$ to mean that $X$ has a uniform distribution on the interval $[a, b]$.

5. The c.d.f. of $X$ is:

$$\begin{aligned} F_X(x) &= 0 \text{ for} x < a \\ &= \frac{x-a}{b-a} \text{ for} a \leq x \leq b \\ &= 1 \text{ for} b < x \end{aligned}$$

   The graph of the c.d.f. is 0 up to $a$, then grows linearly up to $b$, then is constantly 1 after.

6. $$\begin{aligned} \mu = E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= 0 + \int_a^b x \frac{1}{b-a} dx + 0 \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{a+b}{2} \end{aligned}$$

   It is easy to see that the variance of a uniform distribution is $Var(X) = \frac{(b-a)^2}{12}$.

### 15.1.2 The exponential distribution

**Definition:** $X$ has an exponential distribution with parameter $\beta > 0$ if:

$$f_X(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}} \text{ for } x \geq 0$$

The distribution is equal to 0 elsewhere.

**Notes:**

1. we write $X \sim Exp(\beta)$ to describe this distribution.
2. If $X$ is exponential, then $X$ is a non-negative random variable, meaning $P(X \geq 0) = 1$.
3. The p.d.f. of $f_X(x)$ is $\frac{1}{\beta}$ at $x = 0$, and grows exponentially downwards as $x$ grows larger. The probability of an interval of length $L$ decreases as $L$ moves

down the graph (e.g. the probability between 2 and 4 is greater than the probability between 4 and 6). The probability is concentrated towards the origin.

4. The c.d.f. of $X$ is:

$$\begin{aligned} F_X(x) &= 0 \quad \forall x < 0 \\ &= 0 + \int_0^x \frac{a}{\beta} e^{\frac{-y}{\beta}} dy \quad \text{for } x \geq 0 \\ &= 1 - e^{\frac{-x}{\beta}} \end{aligned}$$

5. $\mu = E(X) = \int_{-\infty}^{\infty} x f_X(x) dx = 0 + \int_0^{\infty} x \frac{1}{\beta} e^{-\frac{x}{\beta}} dx = \beta$ You can do this using integration by parts, or we'll see a trick for this later. Also, $\sigma^2 = Var(X) = \beta^2$.

6. Sometimes, the exponential distribution will be "parameterized" in a different way, i.e. the parameter will be written in a different form. The alternative form for the p.d.f. is:

$$\begin{aligned} f_X(x) &= \lambda e^{\lambda x} \text{ for} x \geq 0 \\ &= 0 \text{ elsewhere} \end{aligned}$$

   Watch out how the writer is writing in the parameter for the distribution! In this case, $E(X) = \frac{1}{\lambda}$ and $Var(X) = \frac{1}{\lambda^2}$.

7. **The memoryless property:**
   **Theorem:** Let $X \sim Exp(\beta)$. Then, $P(x \leq X < x + h | X \geq x) = P(0 \leq X < h)$.
   In other words, the information that $X \geq x$ is "forgotten".
   Important note: the memoryless property *does not* assert that $P(0 \leq X < h) = P(x \leq X < x + h)$!

**Proof:** Let $x \leq X < x+h$ be $B$, and $X \geq x$ be $A$. Then,

$$\begin{aligned} P(x \leq X < x+h | X \geq x) &= P(B \cap A)/P(A) \\ &= P(B)/P(B) \text{since B is a subset of A} \\ &= \frac{F_X(x+h) - F_X(x)}{1 - P(X < x)} \\ &= \frac{1 - e^{\frac{-(x+h)}{\beta}} - (1 - e^{\frac{-x}{\beta}})}{1 - (1 - e^{\frac{-x}{\beta}})} \\ &= 1 - e^{\frac{-h}{\beta}} \\ &= P(0 < X \leq h) \end{aligned}$$

There is an interesting converse – the *only* continuous distribution with the memoryless property is the exponential. The geometric discrete distribution also has this property.

8. The exponential distribution is used when you believe that $X$ has a constant "hazard" (i.e. $P(x \leq X < x+h | X \geq x)$ is roughly constant in $x$ for small $h$). It is also used to model the times between events that occur according to a Poisson Process (explained in later statistic courses).

# 16 20 March

## 16.1 The Gamma Distribution

Before defining the gamma distribution, we need to define the *gamma function*.

### 16.1.1 The Gamma Function

Let $\alpha > 0$. We denote the gamma function by $\Gamma(\alpha)$ and define

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx$$

The gamma function has the following two important properties:

1. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

2. $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$

### 16.1.2 The Gamma Distribution

**Definition:** A random variable $X$ has a gamma density with parameters $\alpha, \beta$ if its p.d.f. is given by:

$$f_X(x) = \frac{1}{\Gamma(\alpha)} \frac{x^{\alpha-1} e^{\frac{-x}{\beta}}}{\beta^\alpha} \text{for } x \geq 0$$

$f_X(x) = 0$ when $x < 0$.

**Notes:**

**(1)**

$$\int_0^\infty \frac{1}{\Gamma(\alpha)} \frac{x^{\alpha-1} e^{\frac{-x}{\beta}}}{\beta^\alpha}\, dx = 1$$

This is as it should be – set $y = \frac{x}{\beta}$ to get $\frac{1}{\Gamma(\alpha)} \int_0^\infty y^{\alpha-1} e^{-y}\, dy = 1$. Set $y = \frac{x}{\beta}$ and $dx = \beta\, dy$ to get the same answer.

This just proves that this given formula is a density, as it integrates out to 1.

**(2)** The gamma distribution is said to "flexible", meaning that many different shapes for the p.d.f. can be induced by changing the two parameters $\alpha$ and $\beta$.

For $\alpha > 1$, the p.d.f. grows rapidly immediately after $x > 0$, then drops off with a tail as $x$ grows larger. $f_X$ is said to be skewed to the right.

For $\alpha = 1$, $f_X$ grows exponentially downwards, similar to the exponential distribution.

For $\alpha < 1$, the p.d.f. also grows exponentially downwards, but more steeply.

**(3)** The gamma density can be used to model the waiting time for the $n$th event if the times between events are independent exponential random variables.

**(4)** There are two important special cases of the gamma distribution:

1. If we set $\alpha = 1$, we get an exponential distribution with parameter $\beta$.
2. If we set $\alpha = \frac{\nu}{2}$ and $\beta = 2$, then the density becomes:

$$\frac{1}{\Gamma(\frac{\nu}{2})} \frac{x^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}}} e^{\frac{-x}{2}}$$

For $x \geq 0$ (0 otherwise).
This particular p.d.f. plays an important role in statistics, and is called a *Chi-square* p.d.f. "with $\nu$ degrees of freedom". $\nu$ is just a parameter with this peculiar name.
We write $X \sim \chi^2_\nu$ to mean "$X$ has a Chi-square distribution with $\nu$ degrees of freedom".

**(5)** It is not too difficult to derive $E(X)$ and $Var(X)$ from the definition. It will be easier, however, once we know about moment-generating functions.

In the end, $E(X) = \alpha\beta$. Write $x^\alpha$ as $x^{\alpha+1-1}$ and let $y = \frac{x}{\beta}$, and carry out the integration.

We get, similarly, $Var(X) = \alpha\beta^2$.

In particular, if $X \sim \chi^2_\nu$, then $E(X) = \nu$ (set $\alpha = \frac{\nu}{2}$ and $\beta = 2$) and $Var(X) = 2\nu$).

**(6)** The c.d.f. $F_X$ is not known in closed form: $F(x) = 0$ for $x < 0$ and:

$$F_X(x) = \int_0^x \frac{1}{\Gamma(\alpha)} \frac{y^{\alpha-1}}{\beta^\alpha} e^{\frac{-y}{\beta}}\, dy$$

For $x > 0$.

**(7)** Notation:

We write $X \sim Gamma(\alpha, \beta)$ to mean "$X$ has a gamma distribution with parameters $\alpha, \beta$".

## 16.2 The Normal (or Gaussian) Distribution

The normal or Gaussian distribution is easily the most important distribution in probability and statistics! The distribution seems to occur naturally all over.

### 16.2.1 Definition

The random variable $X$ has a normal distribution with parameters $\mu, \sigma^2$ if its p.d.f. is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

for $-\infty < x < \infty$.

### 16.2.2 Notes

**(1)** We write $X \sim N(\mu, \sigma^2)$.

**(2)** It is possible to show directly that $E(X) =$ the parameter $\mu$ and $Var(X) =$ the parameter $\sigma^2$.

Note that if $X \sim N(1.2, 7.8)$, we mean that $\mu = 1.2$ and $\sigma^2 = 7.8$ *not* $\sigma = 7.8$. We'll derive $\mu$ and $\sigma^2$ by using the so-called moment generating function later, as the current integration would be a bit tricky (but not impossible).

**(3)** The p.d.f. has the famous bell shape. The features are:

1. $f_X$ is symmetric about $\mu$.
2. Changing $\mu$ changes the location of the p.d.f., i.e. where it is centred on the $x$-axis.
3. Increasing $\sigma^2$ increases the spread of the p.d.f. and decreasing $\sigma^2$ decreases the spread.

**(4)** The c.d.f. is not known in closed form, similar to the gamma distribution. Probabilities of intervals need to be done using numerical integration.

However, unlike the gamma density, it is possible to use a single table to find any normal probability. The idea is to reduce the general problem to what is called a *standard normal problem.*

### 16.2.3 Standard Normal Problem

**Background:** if I give you *any* random variable with mean $\mu$ and standard deviation $\sigma$, then

$$Y = \frac{X - \mu}{\sigma}$$

has $E(Y) = 0$ and $Var(Y) = 1$.

We are said to have *standardized* $X$.

However, if $X \sim N(\mu, \sigma^2)$, we have the following:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

This is called a *standard normal* random variable or distribution.

# 17 22 March

## 17.1 Standardizing continued

Our main result from Tuesday: if $X \sim N(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

*Note*: For any random variable with mean $\mu$ and variance $\sigma^2$, $\frac{X-\mu}{\sigma}$ will have mean 0 and variance 1. The proof of this is simple - just plug in the fraction for $X$ in $E(X)$ and $Var(X)$.

The point of the main result is that, after standardizing, we still get a normal random variable If you standardize a random variable, you don't always get a random variable of the same type (unless you're standardizing a random variable with a normal distribution).

### 17.1.1 Example 1

**Problem:** If $X \sim N(-1.2, 4)$, find $P(-1.9 \leq X < 2.2)$.

**Solution:** The idea is to reduce the problem to a $N(0, 1)$ problem, and then use $N(0, 1)$ tables.

*Step 1:* (do not draw a sketch now)

$$\begin{aligned} P(-1.9 \leq X < 2.2) &= P(\frac{-1.9 - (-1.2)}{2} \leq \frac{X - (-1.2)}{2} < \frac{2.2 - (-1.2)}{2}) \\ &= P(-.35 \leq Z < 1.7) \text{ our main result from before} \end{aligned}$$

*Step 2:* draw a sketch: (draw a standard normal distribution with mean = 0, shade in area A between $-.35$ and 1.7)

Tables will give you areas to the right of a value $z$. Areas to the right of $z = 3$ is essentially 0, so tables will usually not give values of $z > 3$. Recall that the areas will be the probabilities – e.g. the area to the right of $z = 1.78$ will equal $P(Z \geq 1.78)$.

We'll call $A_1$ the area between $-.35$ and 0, and $A_2$ the area between 0 and 1.7. We'll get these values by using the symmetry of $N(0, 1)$ about $\mu = 0$. Tables only give positive values of $z$, so to get $A_1$, subtract $P(Z \geq .35)$ from .5.

From the table values, $A_2 = .5 - .0446$ and $A_1 = .5 - .3632$, so $A = A_1 + A_2 = .5922$.

### 17.1.2 Example 2

**Problem:** Use the $N(0, 1)$ tables inversely here. Suppose that a car battery is known to have a lifetime that is approximately normally distributed with a mean of 36 months and a standard deviation of 6 months. What should the warranty period be set at so that only 5% of batteries will need to be replaced?

*Notice:* Batteries cannot have a negative lifetime, so our true normal distribution will not work. However, our lifetime is so skewed to the right that $2\sigma$ is still way to the right of 0, so we can shift our model. Strictly speaking, modelling anything that cannot hold negative values is not correctly, but for almost all cases, the normal distribution will work just fine.

**Solution:** We have that $X \sim N(36, 6^2)$. Let $x_0$ be the required warranty period. We want that $x_0$ such that $P(X < x_0) = .05$, i.e. such that $P(X \geq x_0) = .95$.

Reduce this distribution to a standard normal distribution. So, we seek $x_0$ such that:

$$P(\frac{X - 36}{6} \geq \frac{x_0 - 36}{6}) = .95$$

i.e. such that $P(Z \geq \frac{x_0 - 46}{6}) = .95$.

Draw a sketch – we're looking for a $z_0$ from standard normal tables such that the area to the right is .95, then set that equal to our above probability, and we can solve our problem.

This $z_0$ must be to the left of the mean 0, since the area to the right of the mean is .5. Find a $z_1$ such that the area to the right of it is .05, and according to our tables,

$z_1 = 1.64$. Take the negative, so the area to the right of $z_0 = -1.64$ is .95 (using the symmetry of the normal distribution).

From the $N(0, 1)$ tables, we know that $P(Z \leq -1.64) = .95$. Finally, we must have that

$$\frac{x_0 - 36}{6} = -1.64$$

We get $x_0 = 26.16$ months.

## 17.2 Moment Generating Functions

### 17.2.1 Definition

Let $X$ be a random variable with p.d.f. $f_X$ (respectively, probability $P_X$ for the discrete case). We define the moment generating function (denoted m.g.f.) to be that function of t, such that

$$M_X(t) = E(e^{tx})$$

### 17.2.2 Notes

**(1)** The m.g.f. is a function of the real values $t$.

**(2)** In the continuous case,

$$E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx$$

In the discrete case,

$$E(e^{tx}) = \sum_x e^{tx} P(X = x)$$

**(3)** For some distributions, the m.g.f. does not exist because the integral (or sum) does not converge. We say that the m.g.f. exists if it exists in some interval containing 0.

**(4)** If the m.g.f. exists, then it is possible to recover the p.d.f. or probability function, i.e. there is a one-to-one correspondence between a p.d.f. (pf) and a m.g.f.. Recovering it, although possible, is a bit complicated, and we will not be expected to do so.

**(5) Uses of the m.g.f.** – the m.g.f. can be used to find the *moments* of a random variable, and is often easier than finding the moments ($E(X^k)$) by using the definition.

**Theorem:** $E(X^k) = M^{(k)}(0)$.

**Proof (continuous case):** (discrete case – replace integral by sum) We have, by definition:

$$\begin{aligned} M_X(t) &= E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx \\ M_X^{(1)}(t) &= \frac{d}{dt} \int \dots \\ &= \int \frac{d}{dt} \dots \\ &= \int x e^{tx} f_X(x)\, dx \\ \text{now set } t &= 0 \\ M_X^{(1)}(0) &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\ &= E(X) \end{aligned}$$

In general, we get

$$M^{(k)}(t) = \int_{-\infty}^{\infty} x^k e^{tx} f_X(x)\, dx$$

This gives $M^{(k)}(0) = E(X^l)$, as advertised.

# 18 27 March

## 18.1 Recall

$$\begin{aligned} M_X(t) &= E(e^{tx}) \\ M_X^{(k)}(0) &= E(X^k) \end{aligned}$$

## 18.2 The m.g.f.s of some important distributions

### 18.2.1 Gamma distribution

$X \sim Gamma(\alpha, \beta)$

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \frac{1}{\Gamma(\alpha)} \frac{x^{\alpha-1}}{\beta^\alpha} e^{\frac{-x}{\beta}}\, dx \\ &= \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} \int_0^\infty x^{\alpha-1} e^{-x(\frac{1}{\beta}-t)}\, dx \end{aligned}$$

Set $y = x(\frac{1}{\beta} - t)$ with $dy = (\frac{1}{\beta}) - t\, dx$ to get

$$M_X(t) = \frac{1}{(1-\beta t)^\alpha} \quad \text{for } |\beta t| < 1$$

From $M_X(t)$, we immediately get

$$M_X'(0) = \alpha\beta(1-\beta t)^{-\alpha-1}\Big|_{t=0}$$

by the chain rule. Therefore, $M_X'(0) = \alpha\beta$. Similarly,

$$M_X''(t)\Big|_{t=0} = \alpha\beta^2 + \alpha^2\beta^2$$

as $Var(Y) = E(Y^2) - (E(Y))^2$. Therefore, $Var(X) = \alpha\beta^2$.

So, in particular, for $\alpha = 1$ (i.e. the exponential distribution), we have

$$M_X(t) = \frac{1}{(1-\beta t)}$$

with $E(X) = \beta$ and $Var(X) = \beta^2$.

For $\alpha = \frac{\nu}{2}$ and $\beta = 2$ (i.e. a chi-square distribution with $\nu$ degrees of freedom), we get

$$M_X(t) = \frac{1}{(1-2t)^{\frac{\nu}{2}}}$$

and $E(X) = \nu$ and $Var(X) = 2\nu$.

### 18.2.2 Binomial distribution

$X \sim Bin(n, p)$

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= (1 - p + pe^t)^n \ \forall t \in (-\infty, \infty) \end{aligned}$$

(note: $1 - p = a$ and $pe^t = b$)

We get $M'_X(0) = np$ and $M''_X(0) = np(1-p) + n^2p^2$. Therefore, $Var(X) = np(1-p)$.

### 18.2.3 Poisson distribution

$X \sim Po(\lambda)$

$$\begin{aligned} M_X(t) &= E(e^{tx}) \\ &= \sum_{x=0}^{\infty} \frac{e^{tx}\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t\lambda)^x}{x!} \\ &= e^{-\lambda} e^{e^t\lambda} \\ &= e^{\lambda(e^t-1)} \end{aligned}$$

Therefore, $E(X) = M'_X(0) = \lambda$ and $E(X^2) = \lambda + \lambda^2$, so therefore $Var(X) = \lambda$.

### 18.2.4 Normal distribution

Let $X \sim N(0, 1)$ to start with.

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}x^2} \ \forall x \in (-\infty, \infty)$$

Therefore,

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx \\ &= \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2-2tx+t^2)} e^{\frac{t^2}{2}} \frac{1}{\sqrt{2\pi}}\, dx \\ &= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2}\, dx \\ &= e^{\frac{t^2}{2}} \text{ since the integrand is just a } N(t,1) \text{ p.d.f..} \end{aligned}$$

To get the m.g.f. of a $N(\mu, \sigma^2)$ random variable for arbitrary $\mu$ and $\sigma^2$, we use the following property of an m.g.f.: let $a$ and $b$ be constants. Then,

$$M_{aX+b}(t) = e^{bt} M_X(at)$$

Now, recall that if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. Therefore, we can always write a $N(\mu, \sigma^2)$ random variable $X$ as $X = \sigma Z + \mu$.

Putting these two results together, we get

$$M_X(t) = M_{\sigma Z+\mu}(t) = e^{\mu t} e^{\frac{\sigma^2 t^2}{2}}$$

and

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \; \forall t \in (-\infty, \infty)$$

We find $M_X'(0) = \mu$, $M_X''(0) = \sigma^2 + \mu^2$, and $Var(X) = \sigma^2$.

## 18.3 Transformations of random variables

Often, we're given the distribution of a random variable $X$, but we're more interested in some function $Y = g(X)$ of this random variable, e.g. maybe we have the distribution of the velocity $V$, and we are interested in the distribution of the kinetic energy $Y = \frac{1}{2}mV^2$. In general, we're concerned with finding the distribution of $g(X)$ knowing the distribution of $X$.

First, consider the following two examples to illustrate the eventual formula for the continuous case.

### 18.3.1 Example 1

Let $X \sim N(\mu, \sigma^2)$. Find the p.d.f. of $Z = \frac{X-\mu}{\sigma}$. We know the answer to this, but we don't know the proof for it.

**Step 1:** write down the c.d.f. of $Z$.

$$\begin{aligned} F_Z(z) &= P(Z \leq z) \\ &= P(\frac{X-\mu}{\sigma} \leq z) \end{aligned}$$

**Step 2:** write the $F_Z$ in terms of $F_X$.

$$\begin{aligned} P(\frac{X-\mu}{\sigma} \leq z) &= P(X \leq \sigma z + \mu) \\ &= F_X(\sigma z + \mu) \end{aligned}$$

**Step 3:** differentiate in terms of $z$.

$$\begin{aligned} f_z(z) &= \frac{d}{dz} F_Z(z) \\ &= \sigma f_X(\sigma z + \mu) \text{ by chain rule} \end{aligned}$$

**Finally:**

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \Big|_{\sigma z + \mu} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \ \forall z \in (-\infty, \infty) \end{aligned}$$

### 18.3.2 Example 2 (Careful!)

Let $Z \sim N(0, 1)$. Find the p.d.f. of $Y = Z^2$.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(Z^2 \leq y) \\ &= P(|Z| \leq \sqrt{y}) \\ &= P(-\sqrt{y} \leq Z \leq \sqrt{y}) \\ &= F_Z(\sqrt{y}) - F_Z(-\sqrt{y}) \text{ for } y \geq 0 \end{aligned}$$

Finally,

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}F_Y(y) \\ &= \frac{1}{2}y^{-\frac{1}{2}}f_Z(\sqrt{y}) + \frac{1}{2}y^{-\frac{1}{2}}f_Z(-\sqrt{y}) \ \forall y \geq 0 \\ &= y^{-\frac{1}{2}}f_Z(\sqrt{y}) \ (N(0,1) \text{ density symmetric about } 0) \end{aligned}$$

# 19 29 March

## 19.1 Transformations continued

If $Z \sim N(0,1)$, then $Y = Z^2 \sim X^2$. We have

$$\begin{aligned} F_Y(z) &= P(-\sqrt{z} \leq Z \leq \sqrt{z}) \\ & F_Z(\sqrt{z}) - F_Z(-\sqrt{z}) \\ f_Y(z) &= \frac{d}{dz}F_Y(z) \\ &= \frac{1}{2}z^{-\frac{1}{2}}f_Z(\sqrt{z}) + f_Z(-\sqrt{z})\frac{1}{2}z^{-\frac{1}{2}} \end{aligned}$$

But $f_Z$ is a $N(0,1)$ p.d.f. which is symmetric about 0, and therefore $f_Z(\sqrt{z}) = f_Z(-\sqrt{z})$.

Therefore, we have

$$\begin{aligned} f_Y(z) &= z^{-\frac{1}{2}}f_Z(\sqrt{z}) \text{ for } 0 < z < \infty \\ &= 0 \text{ for } z \leq 0 \end{aligned}$$

Finally, we use the following facts: (with $\alpha = \frac{\nu}{2} = \frac{1}{2}$ and $\beta = 2$)

$$f_Z(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2} \text{ for } -\infty < u < \infty$$

If $W \sim \chi^2_1$, then

$$f_W(w) = \frac{1}{\sqrt{\pi}}\frac{1}{2^{\frac{1}{2}}}w^{-\frac{1}{2}}e^{-\frac{w}{2}} \text{ for } w \geq 0$$

with $f_W(w) = 0$ elsewhere.

We have

$$\begin{aligned} f_Y(z) &= z^{-\frac{1}{2}} \frac{1}{\sqrt{2}\sqrt{\pi}} e^{-\frac{1}{2}z} \text{ for } z > 0 \\ &= 0 \text{ for } z \leq 0 \end{aligned}$$

We're done!

### 19.1.1 General formula and theorem

We can now give a general formula that allows us to go from the p.d.f. of a given random variable $X$ to the p.d.f. of a transformed random variable $Y = g(X)$.

**Theorem:** Let $X$ have p.d.f. $f_X$. Let $y = g(x)$ be either strictly increasing or strictly decreasing as a function of $x$, and $X$ is continuous. Define $Y = g(X)$. Then, the p.d.f. of $Y$ be

$$f_Y(y) = f_X(g^{-1}(y))|\frac{dx}{dy}| \text{ for the appropriate range of values of } Y.$$

Note that $|\frac{dx}{dy}| = \frac{1}{|\frac{dy}{dx}|}$.

**Proof:** First,

$$\begin{aligned} F_Y(y) &= P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) \text{ (if } g \text{ increasing)} \\ \text{while } &= P(X \geq g^{-1}(y) \text{ (if } g \text{ decreasing)} \end{aligned}$$

Thus, we have,

$$\begin{aligned} F_Y(y) &= F_X(g^{-1}(y)) \text{ (if } g \text{ increasing)} \\ &= 1 - F_X(g^{-1}(y)) \text{ (if } g \text{ decreasing)} \end{aligned}$$

Finally, as $g(x) = y \Rightarrow x = g^{-1}(y)$,

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= f_X(g^{-1}(y)) \frac{dx}{dy} \text{ (if } g \text{ increasing)} \\ &= -f_X(g^{-1}(y)) \frac{dx}{dy} \text{ (if } g \text{ decreasing)} \end{aligned}$$

But if $g$ is decreasing, then $\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}} < 0$, so that the two situations (where $g$ is increasing and $g$ is decreasing) can be combined into a single formula.

$$f_Y(y) = f_X(g^{-1}(y)) |\frac{dx}{dy}|$$

*Note:* do not apply this formula unless $g$ is either strictly increasing or strictly decreasing.

### 19.1.2 The probability integral transformation

The following is a very important result that allows one to simulate observations from a given probability distribution by knowing only how to simulate observations from a $U(0,1)$ distribution. This famous result is called *the probability integral transformation*.

Let $X$ be a continuous random variable with a strictly increasing $F_X$. Let $Y = F_X(X)$. Then $Y \sim U(0,1)$.

*Note:* here, our $g$ is $F_X$, i.e. $Y = g(X) = F_X(X)$. Also, you must use $F_X$ and not some other $F$.

**Proof:** by our formula,

$$f_Y(y) = f_X(F_X^{-1}(y)) \frac{dx}{dy}$$

($|\frac{dx}{dy}| = \frac{dx}{dy}$ since $F_X$ is increasing)

$$f_Y(y) = f_X(F_X^{-1}(y)) \frac{1}{\frac{dy}{dx}}$$

Recall $y = F_X(x)$, $\frac{dy}{dx} = f_X(x)$, and $x = F_X^{-1}(y)$. Therefore,

$$f_Y(y) = f_X(F_X^{-1}(y))\frac{1}{f_X(F_X^{-1}(y))} = 1 \text{ for } 0 < y < 1$$

and $f_Y(y) = 0$ elsewhere.

We recognize the above as a $U(0, 1)$ p.d.f.

## 19.2 Joint probability distributions

Very often, we're interested in the simultaneous behaviour of several random variables, rather than one at a time, as we have considered up until now, e.g. if $X =$ number of kilometres traveled by a tire and $Y =$ tread depth, we may wish to know about the simultaneous or joint distribution of $X$ and $Y$. This leads to so-called *multivariate distributions.*

We shall considered bivariate (i.e. pairs of random variables) distributions, and indicate the general extensions at the end.

### 19.2.1 Definition

The random variables $X$ and $Y$ have joint c.d.f., denoted $F_{X,Y}$, if

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

Which we denote as $P(X \leq x, Y \leq y)$.

### 19.2.2 Notes

**(1)** This definition holds for both continuous and discrete random variables.

**(2)** It is possible to show (in an advanced probability course) that the joint c.d.f. uniquely determines the probability distribution in two-dimensional space.

# 20 03 April

## 20.1 Properties of the joint c.d.f.

**(1)** $F_{X,Y}(x,y)$ uniquely determines the joint probability distribution in two-dimensional space, i.e. in theory, given any event $B$ in $\mathbb{R}^2$, once we know $F_{X,Y}(x,y)$ for all $-\infty < x < \infty$, $-\infty < x < \infty$, then $P((X,Y) \in B)$ is uniquely determined.

**(2)** We define the so-called *marginal c.d.f.* $F_X$ and $F_Y$ of $F_{X,Y}$ as follows:

$$F_X(x) = P(X \leq x) = F_{X,Y}(x, +\infty) = \lim_{y \to \infty} F_{X,Y}(x,y)$$

and, similarly,

$$F_Y(y) = F_{X,Y}(+\infty, y)$$

**(3)** $F_{X,Y}(x,y)$ is non-decreasing in $x$ and $y$ (e.g. $F_{X,Y}(x,y) \leq F_{X,Y}(x',y)$ for $x' > x$).

**(4)** $F_{X,Y}(-\infty,-\infty) = 0$ (c.f. $F_X(-\infty) = 0$) and $F_{X,Y}(\infty,\infty) = 1$.

**(5)** $F_{X,Y}(x,y)$ is jointly continuous from the right (c.f. $F_X(x)$ is continuous from the right).

## 20.2 The role of the p.d.f. and probability functions in joint distributions

**Definition:** We call $f_{X,Y}(x,y)$ the *joint p.d.f.* of $(X,Y)$ if

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

and $f_{X,Y}(x,y) \geq 0$.

We call $P_{X,Y}(x,y)$ the *joint probability function* of $(X,Y)$ if

$$P((X,Y) \in A) = \sum \sum_{(x,y) \in A} P_{X,Y}(x,y)$$

for all events $A$ in $\mathbb{R}^2$.

In particular, for event $A = (-\infty, x] \times (-\infty, y]$ in the continuous case:

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

For the discrete case:

$$F_{X,Y}(x,y) = \sum_{v \leq y} \sum_{u \leq x} P_{X,Y}(u,v)$$

It follows that

$$f_{X,Y}(x,y)\, dx\, dy \approx P(x < X \leq x + dx, y < Y \leq y + dy)$$

and that

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

By the Fundamental Theorem of Calculus,

$$\frac{\delta^2}{\delta x\, \delta y} F_{X,Y}(x,y) = f_{X,Y}(x,y)$$

Given $f_{X,Y}(x,y)$, to find the marginal p.d.f., integrate out the variable you wish to get rid of:

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

and

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

(beware of times when the p.d.f. changes its form – see next example for details)

### 20.2.1 Example

Let

$$f_{X,Y}(x,y) = \frac{2}{3}(x + 2y) \text{ for } 0 < x < 1, 0 < y < 1$$

and $f_{X,Y}(x,y) = 0$ elsewhere.

**(1)** Find the marginal p.d.f.s $f_X$ and $f_Y$.

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy \\ &= 0 + \int_0^1 \frac{2}{3}(x+2y)\, dy + 0 \\ &= \frac{2}{3}(x+1) \text{ for } 0 < x < 1 \\ &= 0 \text{ elsewhere} \end{aligned}$$

and

$$\begin{aligned} f_Y(y) &= 0 + \int_0^1 \frac{2}{3}(x+2y)\, dx \\ &= \frac{1}{3}(1+4y) \text{ for } 0 < y < 1 \\ &= 0 \text{ elsewhere} \end{aligned}$$

**(2)** Find the joint c.d.f. of $(X, Y)$.

We need $F_{X,Y}(x,y)$ for all $(x,y) \in \mathbb{R}^2$.

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = 0 \text{ for either } x \leq 0 \text{ or } y \leq 0$$

For $0 < x < 1$ and $0 < y < 1$:

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\, du\, dv \\ &= \int_0^y \int_0^x \frac{2}{3}(u+2v)\, du\, dv \\ &= \frac{1}{3}x^2y + \frac{2}{3}xy^2 \end{aligned}$$

For $0 < x < 1$ and $y \geq 1$:

$$\begin{aligned} F_{X,Y}(x,y) &= \int_0^1 \int_0^x \frac{2}{3}(u+2v)\, du\, dv + \int_1^y \int_0^x 0\, du\, dv \\ &= \frac{x^2}{3} + \frac{2}{3}x \end{aligned}$$

For $x \geq 1$ and $0 < y < 1$:

$$\begin{aligned} F_{X,Y}(x,y) &= \int_0^y \int_0^1 \frac{2}{3}(u+2v)\, du\, dv + \int_0^y \int_1^x 0\, du\, dv \\ &= \frac{y}{3} + \frac{2}{3}y^2 \end{aligned}$$

For $x \geq 1$ and $y \geq 1$:

$$F_{X,Y}(x,y) = 1$$

**(3)** Find the marginal c.d.f. $F_X$.

The first possible way involves using the marginal p.d.f. that we got from part 1 of this example.

$$F_X(x) = 0 \text{ for } x \leq 0$$

and

$$\begin{aligned} F_X(x) &= \int_0^x \frac{2}{3}(u+1)\, du \text{ for } 0 < x < 1 \\ &= \frac{x^2}{3} + \frac{2}{3}x \end{aligned}$$

and, for $x \geq 1$, $F_X(x) = 1$.

The second possible way:

$$\begin{aligned} F_X(x) = F_{X,Y}(x, +\infty) & \\ &= 0 \text{ for } x \leq 0 \\ &= \frac{x^2}{3} + \frac{2}{3}x \text{ for } 0 < x < 1 \\ &= 1 \text{ for } x \geq 1 \end{aligned}$$

*Note*: identify the part of the range of $F_{X,Y}(x,y)$ where $0 < x < 1$ and where you can let $y \to +\infty$. In this case, the part is $0 < x < 1$ and $y \geq 1$. Finally, for such $y$, $F_{X,Y}(x,y)$ **does not** change with $y$.

Thus, we get the same marginal c.d.f. for $X$ by two different but equivalent methods.

## 20.3 Conditional Distributions

Conditioning plays a huge part in probability and statistics. Hence, we need to consider conditional distributions.

**Definition:** Given the joint probability function of $(X, Y)$ $P_{X,Y}(x, y) = P(X = x, Y = y)$, we define the conditional probability function of $Y$ given $X = x$ to be:

$$P(Y = y \mid X = x) = \frac{P_{X,Y}(x, y)}{P_X(x)} \; \forall x : P_X(x) \neq 0$$

This is denoted as $F_{Y \mid X \leq x}(y \mid X \leq x)$.

# 21 05 April

## 21.1 Conditional distributions continued

### 21.1.1 Conditional probability function

Also straight-forward is the *conditional probability function*:

$$P_{Y \mid X=x}(y) = P(Y = y \mid X = x)$$

Again, by the definition of conditional probability, the right hand side is equal to

$$\frac{P(X = x, Y = y)}{P(X = x)} = \frac{P_{X,Y}(x, y)}{P_X(x)}$$

(provided that $P(X = x) \neq 0$).

### 21.1.2 Conditional probability density function

Something more interesting is how we deal with, say, $P(Y \leq y \mid X = x)$ when $X$ and $Y$ are jointly continuous. We cannot define this as the ratio of the joint divided by the $P(X = x)$ since the latter is 0 for *all* $x$. Because of this, we need to take a slightly different route.

First, define the *conditional p.d.f.* of $Y$ given $X = x$ denoted by

$$f_{Y \mid X=x}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

for all $x$ such that $f_X(x) \neq 0$.

Now, since we know that if we integrate a p.d.f. over a region $A$, we get the probability of that region. Therefore,

$$\begin{aligned} F_{Y \mid X=x}(y \mid x) &= P(Y \leq y \mid X = x) \\ &= \int_{-\infty}^{y} f_{Y \mid X=x}(u \mid x)\, du \end{aligned}$$

It follows that

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

Note that this is the usual definition of expected value, except that we use the conditional p.d.f..

In the discrete case:

$$P(Y \leq y \mid X = x) = \sum_{\forall u : u \leq y} P_{Y \mid X=x}(u \mid x)$$

It's only in the continuous case where things get a bit more complicated.

## 21.2 The Law of Total Probability for Random Variables

The following theorems are very useful analogues of the Law of Total Probability for events.

Recall:

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

### 21.2.1 Discrete case

For discrete random variables:

$$\begin{aligned} P(Y = y) &= \sum_{\forall x} P(Y = y \mid X = x)P(X = x) \\ &= \sum_{\forall x} P_{Y \mid X=x} P_X(x) \end{aligned}$$

The proof of this is the same as the proof for sets.

### 21.2.2 Continuous case

For continuous random variables, we have the following theorem:

**Theorem:** Let $X, Y$ have joint p.d.f $X, Y$ with conditional p.d.f. $f_{Y \mid X=x}(y \mid x)$. Then,

$$\begin{aligned} \text{(a) } f_Y(y) &= \int_{-\infty}^{\infty} f_{Y \mid X=x}(y \mid x) f_X(x)\, dx \\ \text{(b) and } F_Y(y) &= \int_{-\infty}^{\infty} F_{Y \mid X=x}(y \mid x) f_X(x)\, dx \end{aligned}$$

**Proof of (a):** We have

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx \\ &= \int_{-\infty}^{\infty} f_{Y \mid X}(y \mid x) f_X(x)\, dx \end{aligned}$$

Try part (b) yourself. Or not. Doesn't matter to me.

## 21.3 Example

From Tuesday's example, we had

$$f_{X,Y}(x, y) = \frac{2}{3}(x + 2y) \text{ for } 0 < x < 1, 0 < y < 1$$

and $f_{X,Y}(x,y) = 0$ elsewhere.

**(4)** Find $f_{Y \mid X=x}(y \mid x)$.

We must find

$$\frac{f_{X,Y}(x,y)}{f_X(x)}$$

By (1), we have:

For $0 < y < 1$ and $0 < x < 1$:

$$f_{Y \mid X=x}(y \mid x) = \frac{\frac{2}{3}(x+2y)}{\frac{2}{3}(x+1)}$$

Otherwise, for $y \notin (0,1)$,

$$f_{Y \mid X=x}(y \mid x) = 0$$

**(5)** Find $E(Y \mid X = x)$ for $0 < x < 1$ and, in particular, find $E(Y \mid X = \frac{1}{2})$.

By definition,

$$\begin{aligned} E(Y \mid X = x) &= \int_{-\infty}^{\infty} y \; f_{Y \mid X=x}(y \mid x) \; dy \\ &= \int_0^1 \frac{y(x+2y)}{x+1} \; dy \\ &= \frac{3}{2(x+1)}(\frac{x}{3} + \frac{2}{3}) \text{ for } 0 < x < 1 \end{aligned}$$

In particular,

$$E(Y \mid X = \frac{1}{2}) = \frac{3}{2(\frac{1}{2}+1)}(\frac{1}{6} + \frac{2}{3}) = \frac{5}{6}$$

## 21.4 Bivariate analogues

We seek a summary of a bivariate distribution – a sort of analogue to $E(X)$ or $Var(X)$ for a univariate distribution. Before doing this, however, we need another definition.

**Definition:** Let $g(x, y)$ be a real-valued function of $(x, y)$. Then, we define for the continuous case:

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

In the discrete case:

$$E(g(X,Y)) = \sum_{y}\sum_{x} g(x,y)\ P_{X,Y}(x,y)$$

### 21.4.1 Covariance

An important special case is when

$$g(X,Y) = (X-\mu_X)(Y-\mu_Y)$$

(i.e. $E(X) = \mu_X$ and $E(Y) = \mu_Y$)

Thus, in this case, we're talking about

$$E((X-\mu_X)(Y-\mu_Y))$$

This is given a special name – the *covariance* between $X$ and $Y$, the $Cov(X, Y)$, and is denoted by $\sigma_{XY}$.

### 21.4.2 Notes

**(1)** $Cov(X, Y)$ is a measure of how $X$ and $Y$ vary about their means simultaneously. If $Cov(X, Y) > 0$, then this tells us that as $X$ increases, so does $Y$, and also as $X$ decreases, so does $Y$. Conversely, if $Cov(X, Y) < 0$, then as $X$ increases, $Y$ tends to decrease, and vice versa.

**(2) "little theorem"**

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

Proof: by definition,

$$\begin{aligned} Cov(X,Y) &= E((X-\mu_X)(Y-\mu_Y)) \\ &= E(XY - \mu_Y X - \mu_X Y + \mu_X\mu_Y) \\ &= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X\mu_Y \\ &= E(XY) - \mu_X\mu_Y \end{aligned}$$

# 22 10 April

## 22.1 Covariance continued

### 22.1.1 Covariance and Correlation

While the sign of $Cov(X,Y)$ tells you whether or not $X$ and $Y$ tend to vary in the same direction together (positive if they do, negative if in opposite directions), the magnitude of $Cov(X,Y)$ depends on the scale of measurement. Thus,

$$\begin{aligned} Cov(aX,bY) &= E((aX-a\mu_X)(bY-b\mu_Y)) \\ &= E(ab(X-\mu_X)(Y-\mu_Y)) \\ &= abE((X-\mu_X)(Y-\mu_Y)) \\ &= abCov(X,Y) \end{aligned}$$

In other words, $Cov(aX,bY) \neq Cov(X,Y)$. Therefore, we define a new quantity that has the same sign as $Cov(X,Y)$, but which is *scale invariant*. This way, it does not matter what scale we take our measurements in (Celsius vs Fahrenheit, kilometres vs miles, etc.). Thus, we define the *correlation* between $X$ and $Y$, written as $Corr(X,Y)$ (also $\rho(X,Y)$), and is defined as

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

Note that the sign of $\rho(X,Y)$ is the same as the sign of $Cov(X,Y)$, and

$$|\rho(aX,bY)| = |\frac{Cov(aX,bY)}{\sqrt{Var(aX)Var(aY)}}| = \frac{|ab|Cov(X,Y)}{|ab|\sqrt{Var(X)Var(Y)}} = \rho(X,Y)$$

(i.e. $\rho$ is scale invariant).

### 22.1.2 Important remarks on the correlation coefficient

**(1)** It is not difficult to show that $|\rho(X,Y)| \leq 1$ (i.e. $-1 \leq \rho(X,Y) \leq 1$) using the Cauchy-Schwartz inequality ($E(XY) \leq (E(X^2)E(Y^2))^{\frac{1}{2}}$), or using the fact that $0 \leq E((X-Y)^2) = E(X^2) - 2E(XY) + E(Y^2) \Rightarrow 2E(XY) \leq E(X^2) + E(Y^2)$. The proof was given in class, but will not be on the final.

**(2)** $|\rho| = 1$ if and only if $Y = aX + b$ for constants $a, b$, i.e. if and only if there is a perfect linear relationship. Further, the above proof from **(1)** gives us the claim for **(2)**, since $E(X \pm Y)^2 = 0 \Leftrightarrow Y = \pm X$.

**(3)** It is important to note that the correlation between two random variables is a measure of **linear** dependence between them, and *nothing else.* Avoid using the term "correlation" to describe dependence in general.

### 22.1.3 Example

(Same numbers from Thursday)

$$f_{X,Y}(x, y) = \frac{2}{3}(x + 2y) \text{ for } 0 < x < 1, 0 < y < 1$$

and $f_{X,Y}(x, y) = 0$ elsewhere.

**(6)** Find $Cov(X, Y)$.

We need $\mu_X$ and $\mu_Y$. We then need $E(XY)$.

$$\begin{aligned}
\mu_X &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\
&= \int_0^1 x\frac{2}{3}(x+1)\, dx \\
&= \frac{5}{9} \\
\mu_Y &= \int_{-\infty}^{\infty} y f_Y(y)\, dy \\
&= \int_0^1 y\frac{1}{3}(1+4y)\, dy \\
&= \frac{11}{18} \\
E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\, dx\, dy \\
&= \int_0^1\int_0^1 xy\frac{2}{3}(x+2y)\, dx\, dy \\
&= \frac{1}{3} \\
Cov(X,Y) &= \frac{1}{3} - \frac{5}{9}\frac{11}{18} \\
&= \frac{-3}{486}
\end{aligned}$$

**(7)** Find $Corr(X,Y) = \rho(X,Y)$.

We need $Var(X)$ and $Var(Y)$.

$$\begin{aligned} E(X^2) &= \int_0^1 x^2 \frac{2}{3}(x+1)\, dx \\ &= \frac{7}{18} \\ Var(X) &= \frac{7}{18} - (\frac{5}{9})^2 \\ &= .0803 \\ \sigma_X &= \sqrt{.0803} \\ &= .2833 \\ \ldots &= \ldots \\ Var(Y) &= .6821 \\ \sigma_Y &= \sqrt{.6821} \\ &= .8259 \end{aligned}$$

So, at last, we can find $Corr(X, Y)$.

$$\begin{aligned} Corr(X, Y) &= \frac{\frac{-3}{486}}{.2833 \times .8259} \\ &= -0.0319 \end{aligned}$$

### 22.1.4 Linking variance and covariance

**Theorem:** $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$.

**Proof:**

$$\begin{aligned} Var(X+Y) &= E((X+Y)^2) - (\mu_X + \mu_Y)^2 \\ &= E(X^2) - \mu_X^2 + E(Y^2) - \mu_Y^2 + 2E(XY) - 2\mu_X\mu_Y \\ &= Var(X) + Var(Y) + 2Cov(X, Y) \end{aligned}$$

For $Var(X - Y)$, we get $Var(X) + Var(Y) - 2Cov(X, Y)$.

**Corollary:** if $Cov(X, Y) = 0$, then $Var(X + Y) = Var(X) + Var(Y)$. In general:

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$$

if $X_i, X_j$ are uncorrelated and $i \neq j$.

## 22.2 Independence between random variables

We talked about independence of events, and so now it is natural to discuss the notion of independence between random variables.

### 22.2.1 Definition

Note: This definition is valid whether the random variables are continuous or discrete.

The random variables $X_1, X_2, \ldots, X_n$ are said to be *independent* if and only if

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\ldots F_{X_n}(x_n)$$

for all $-\infty < x_i < \infty$.

If $X_1, X_2, \ldots, X_n$ are jointly continuous, then it is easy to see that they are independent if and only if

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\ldots f_{X_n}(x_n)$$

for all $-\infty < x_i < \infty$.

If $X_1, X_2, \ldots, X_n$ are jointly discrete, then they are independent if and only if

$$P_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P_{X_1}(x_1)P_{X_2}(x_2)\ldots P_{X_n}(x_n)$$

for all $-\infty < x_i < \infty$.

### 22.2.2 Example continued

**(8)** Are $X$ and $Y$ independent?

Use $f_{X,Y}(x, y)$ and $f_X(x)f_Y(y)$. Try $x = \frac{1}{4}, y = \frac{1}{4}$.

$$\begin{aligned} f_{X,Y}(\frac{1}{4}, \frac{1}{4}) &= \frac{2}{3}(\frac{1}{4} + \frac{2}{4}) \\ f_X(\frac{1}{4}) &= \frac{2}{3}\frac{5}{4} \\ f_Y(\frac{1}{4}) &= \frac{1}{3}(1 + \frac{4}{4}) \end{aligned}$$

So that

$$f_X(\frac{1}{4})f_Y(\frac{1}{4}) = \frac{5}{9} \neq \frac{1}{2}$$

Therefore, $X$ and $Y$ are not independent.

### 22.2.3 Independence and covariance

What is the relationship between independence and covariance?

**Theorem:** If $X$ and $Y$ are independent, then $Cov(X, Y) = 0$. Note that the converse is not true.

**Proof: (continuous case)** Assume $X \perp Y$. We must show that $E(XY) = E(X)E(Y)$.

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} yf_Y(y)\,dy \int_{-\infty}^{\infty} xf_X(x)\,dx \\ &= \mu_X\mu_Y \end{aligned}$$

# 23 12 April

## 23.1 Sums of independent random variables

Sums of random variables are particularly important in probability and statistics since we frequently encounter averages of random variables, apart from the divisor $n$,

$$\frac{1}{n}\sum_{i=1}^{n}$$

denoted $\bar{x}$, is just a sum.

We'll do three things:

First off, use the m.g.f. method to find the *exact* distribution of a sum of independent random variables, under certain circumstances.

Secondly, use the Central Limit Theorem to use the approximate distribution of a sum of a "large" number of independent random variables under general conditions.

Thirdly, we'll discuss the Weak Law of Large Numbers that enables to us to interpret probability as a limiting relative frequency.

**The moment generating function method for finding the distribution of a sum of independent random variables:** Recall from last class that if $X \perp Y$, then $E(XY) = E(X)E(Y)$ (i.e. $Cov(X, Y) = 0$). In general, if $X_1, X_2, \ldots, X_n$ are independent, then

$$E(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} E(X_i).$$

The following extended result is true: if $X \perp Y$, then $g_1(x) \perp g_2(y)$ for all functions $g_1, g_2$.

### 23.1.1 Setup

We have independent r.v.s $X_1, X_2, \ldots, X_n$ that are assumed to come from some known distribution (e.g. Poisson, Normal, etc.). We want to find the distribution of

$$S_n = \sum_{i=1}^{n} X_i.$$

This is how the m.g.f. method works:

**Step 1:** find the m.g.f.s $M_{X_i}(t)$.

**Step 2:** find the m.g.f. of $S_n$, $M_{S_n}(t)$ as follows:

$$\begin{aligned} M_{S_n}(t) &= M_{\sum_{i=1}^n X_i}(t) \\ &= E(e^{t\sum_{i=1}^n X_i}) \\ &= E(e^{tX_1}e^{tX_2}\dots e^{tX_n}) \\ &= E(g(X_1)g(X_2)\dots g(X_n)) \text{ where } g = e^{tX_i}. \\ &= \prod_{i=1}^n E(e^{tX_i}) \end{aligned}$$

The last step above is valid because functions of independent random variables are themselves independent, and $E(g(X_1)g(X_2)) = E(g(X_1))E(g(X_2))$. Then,

$$\begin{aligned} \dots &= \dots \\ M_{S_n}(t) &= \prod_{i=1}^n M_{X_i}(t) \end{aligned}$$

*Remark:* if the $X_i$'s all have the same distribution (termed *identically distributed*), then

$$M_{S_n}(t) = (M_{X_i}(t))^n.$$

**Step 3:** having found $\prod_{i=1}^n M_{X_i}(t)$, we hope to recognize its form as the m.g.f. of a familiar distribution. If so, then by the Uniqueness Theorem of M.G.F.s, that distribution **must be** the distribution of $S_n$.

### 23.1.2 In practice

**Theorem:** Let $X_1, X_2, \dots, X_n$ be independent random variables.

**(a)** Let $X_i \sim Poisson(\lambda_i)$.

**(b)** Let $X_i \sim N(\mu_i, \sigma_i^2)$.

**(c)** Let $X_i \sim Binomial(n_i, p)$.

**(d)** Let $X_i \sim \chi^2_{\nu_i}$.

Find the distributions of of $S_n$ in **(a)** through **(d)**.

**Solution:** (easy!)

**(a)** We have $M_{X_i}(t) = e^{\lambda_i(e^t-1)}$. Therefore,

$$M_{S_n}(t) = \prod_{i=1}^{n} e^{\lambda_i(e^t-1)} = e^{\sum_{i=1}^n \lambda_i(e^t-1)}$$

which we recognize as the m.g.f. of a $Poisson(\sum_{i=1}^n \lambda_i)$ random variable. Therefore, by the Uniqueness Theorem, $S_n \sim Poisson(\sum_{i=1}^n \lambda_i)$. In particular, if $\lambda_1 = \lambda_2 = \cdots = \lambda_n = \lambda$, then $S_n \sim Poisson(n\lambda)$.

**(b)** $M_{X_i}(t) = e^{\mu_i t + \frac{\sigma_i^2 t^2}{2}}$

$$\begin{aligned} M_{S_n}(t) &= \prod_{i=1}^{n} e^{\mu_i t + \frac{\sigma_i^2 t^2}{2}} \\ &= e^{\sum_{i=1}^n \mu_i t + \sum_{i=1}^n \sigma_i^2 \frac{t^2}{2}} \end{aligned}$$

which we recognize as the m.g.f. of a $N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$ random variable.

**(c)** (try it yourself)

**(d)**

$$M_{X_i}(t) = \frac{1}{(1-2t)^{\frac{\nu_i}{2}}}$$

Therefore,

$$M_{S_n}(t) = \frac{1}{(1-2t)^{\sum_{i=1}^n \frac{\nu_i}{2}}}$$

Which is the m.g.f. of a

$$\chi^2_{\sum_{i=1}^n \nu_i} \text{ r.v.}$$

## 23.2 The Central Limit Theorem

Roughly, the Central Limit Theorem says: sums of a large number of independent random variables are approximately normally distributed.

**Theorem:** Let $X_1, X_2, \ldots$ be independent and identically distributed (i.i.d.) random variables with mean $\mu$ and variance $\sigma^2$. Then,

$$P(\frac{S_n - n\mu}{\sqrt{n}\sigma}) \leq x) \to P(Z \leq x) \; \forall x \text{ as } n \to \infty$$

where $Z$ is a Standard Normal distribution (i.e. $Z \sim N(0,1)$).

Remember this is talking about the *sums* of the r.v.s, not the r.v.s themselves!

This helps explains why, in the real world, a lot of factors seem to be normally distributed – IQ scores, heights, weights, etc.. This does not prove why, say, heights are normally distributed, however – it's just an observation and a plausibility argument.

### 23.2.1 Notes

**(1)** $Var(S_n) = \sum_{i=1}^{n} Var(X_i) = n\sigma^2$, while $E(S_n) = n\mu$. Therefore, the l.h.s. of the Central Limit Theorem is just $S_n$ standardized to have mean 0 and standard deviation 1.

**(2)** The C.L.T. can be written in the form

$$P(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x) \to P(Z \leq x)$$

where $Z \sim N(0,1)$. Just divide top and bottom by $n$.

**(3)** Note that the C.L.T. gives us a Normal distribution as the approximate distribution as the sum of **any** i.i.d. random variables.

### 23.2.2 Application

Suppose that it is known that the survival time for patients with Alzheimer's disease from onset of symptoms has a mean of 8 years and a standard deviation of 4 years. If a sample of 30 patients with the disease is taken, what is the approximate probability that their average survival will be less than seven years?

**Solution:** the C.L.T. generally works well with $n \geq 30$. We'll let $\bar{X} = \frac{\sum_{i=1}^{30} X_i}{30}$. We

want $P(\bar{X} < 7)$. Use the C.L.T. as follows:

$$\begin{aligned} P(\bar{X} < 7) &= P(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{7 - \mu}{\frac{\sigma}{\sqrt{n}}}) \\ &= P(\frac{\bar{X} - 8}{\frac{4}{\sqrt{30}}} < \frac{7 - 8}{\frac{4}{\sqrt{30}}}) \\ &\approx P(Z < -1.37) \\ &= .0853 \end{aligned}$$