Continuous Distributions Cheat Sheet

Continuous distributions are useful as they allow us to find probabilities for continuous random variables, which can take on an infinite number of values. For example, in modelling heights of people. As the random variable is continuous, the probability of some exact specific value occurring is so infinitesimally small it is 0, so we look at the probability of the random variable taking a value in a range. This makes sense since in the real world, any measurement we make as some associated uncertainty – it's impossible to measure something to an infinite number of decimal places.

Continuous random variables

Examples of continuous random variables are often to do with height or mass, for example the distribution of heights of the male population, the mass of car tyres produced in a factory, or the length of time for a seed to germinate. Although lots of the concepts between continuous random variables and discrete random variables are similar, it is important not to get them confused:

Continuous random variables	Discrete random variables
Infinite number of values that the random variable can take.	Finite number of values that the rando variables can take.
A probability density function is used to define the probability of the random variable taking values within a specific range.	A probability mass function is used to define the probability of the random variable taking a specific value in the sample space.
Integration is used to work out probabilities.	Summation is used to work out probabilities.
The graph of the random variable is a continuous shape/curve/line.	The graph of the random variables is discrete, like a bar chart.

If X is a continuous random variable with probability density function $f(x)$, then:

- $\boldsymbol{f(x) \geq 0}$ **for all** $\boldsymbol{x \in \mathbb{R}}$ (probabilities and probability densities cannot be negative!)
- $\boldsymbol{P(a < X < b) = \int_a^b f(x)\,dx}$
  (probabilities are found by finding the area under the probability density function between the given limits)
- $\boldsymbol{\int_{-\infty}^{\infty} f(x)\ dx = 1}$ (probabilities always sum to 1)

Although the conditions above take the range that the probability density function is applicable on as the set of real numbers, or equivalently from negative infinity to infinity, in practice, the probability density function may only be defined on a subset of the real numbers.

Example 1: Could the following two functions ever be probability density functions?

$$f(x) = \begin{cases} 3x+2, & 0 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}, \quad g(x) = \begin{cases} kx^2, & 0 \leq x \leq 3 \\ 0, & \text{otherwise} \end{cases}.$$

For $f(x)$, it is clear by inspection that it will always be positive, so we need to check whether the total probability is equal to 1.	$\int_0^4 3x+2\ dx = \left[\frac{3x^2}{2}+2x\right]_0^4 = 32$ $32 \neq 1$, so $f(x)$ cannot represent a probability density function.
For $g(x)$, we can see that if $k > 0$, the function will always be positive, so we need to check for what value of k the total probability is equal to 1.	$\int_0^3 kx^2\ dx = \left[\frac{kx^3}{3}\right]_0^3 = \frac{k \times 27}{3} = 9k$ Therefore if $k = \frac{1}{9}$, $g(x)$ represents a probability density function.

The cumulative distribution function

Finding probabilities from the probability density function can be time consuming, especially if you have to find the probabilities of multiple ranges. To overcome this, you can use the cumulative distribution function for a random variable.

- For a random variable X, the cumulative distribution function is $F(x) = P(X \leq x)$. This is equivalent of finding the area under the p.d.f to the left of the x value, or integrating the p,d,f from $-\infty$ to the x value required:

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

  By simply substituting the required x value into $F(x)$, we can find the area under the curve.

Notice that the probability density function (p.d.f) is written with a lowercase and the cumulative distribution function (c.d.f) is written with a capital.

Example 2: A continuous random variable x has cumulative distribution function $F(x) = \begin{cases} 0, & x < -2 \\ \frac{x+2}{6}, & -2 \leq x \leq 4. \\ 1, & x > 4 \end{cases}$

Find a) $P(X < 0)$, and b) $P(-1 \leq X \leq 3)$.

a) The definition of the c.d.f states that $F(x) = P(X \leq x)$. Therefore, as the function is continuous, $P(X < 0) = P(X \leq 0)$, and we can find the answer by substituting 0 into $F(x)$.	$P(X < 0) = \frac{0+2}{6} = \frac{1}{3}$
b) To find $P(-1 \leq X \leq 3)$, calculate $P(X \leq 3) - P(X \leq -1)$.	$P(-1 \leq X \leq 3) = \frac{3+2}{6} - \frac{-1+2}{6} = \frac{4}{6} = \frac{2}{3}$

Example 3: A random variable X has probability density function given by $f(x) = \begin{cases} \frac{x}{2}, & 0 \leq x \leq 1 \\ \frac{x^3}{5}, & 1 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$.

Fully specify the cumulative distribution function.

Integrate each section of the p.d.f separately. The c.d.f, as the name suggests, is cumulative, so will always increase, so you need to add on the value of the c.d.f at the upper limit of the previous section. We used the prime (') to distinguish our integration variable from our integration limit.	For $x < 0$: $F(x) = 0$ For $0 \leq x \leq 1$: $F(x) = \int_0^x \frac{x'}{2}\,dx' = \frac{x^2}{4}$ For $1 < x \leq 2$: $F(x) = F(1) + \int_1^x \frac{x'^3}{5}dx' = \frac{1}{4} + \frac{x^4}{20} - \frac{1}{20}$ For $2 < x$, $F(x) = 1$
Write the cumulative distribution in full, this means defining $F(x)$ for all values of $x \in \mathbb{R}$.	$F(x) = \begin{cases} 0, & x < 0 \\ \frac{x}{5}, & 0 \leq x \leq 1 \\ \frac{1}{5} + \frac{x^4}{20}, & 1 < x \leq 2 \\ 1, & x > 2 \end{cases}$

Mean and variance of a continuous distribution

The ideas of means and variance can also apply to continuous random variables.
If X is a continuous random variable with p.d.f $f(x)$:

- The **mean**, or **expectation** of X is given by: $E(X) = \mu = \int_{-\infty}^{\infty} xf(x)\ dx$
- The **variance** of X is given by:

$$\text{Var}(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\ dx = \int_{-\infty}^{\infty} x^2 f(x)\ dx - \mu^2$$
$$= E(X^2) - (E(X))^2$$

- The **mean of a function** of the continuous random variable is given by:

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

  If $g(X)$ is a **linear** function of the form $aX + b$, the following results are useful:

$$E(aX+b) = aE(X) + b$$
$$\text{Var}(aX+b) = a^2\text{Var}(X)$$

These definition for $E(X)$ and the first definition for $\text{Var}(X)$ are given in your formula booklet, and all of them are very similar to the relevant formulae for discrete random variables.

Example 4: The continuous random variable X has probability density function $f(x) = \begin{cases} \frac{3x}{56}(5-x), & 0 \leq x \leq 4 \\ 0, & \text{otherwise} \end{cases}$.

Find $E(X)$ and $\text{Var}(X)$.

Use the definition of $E(X)$ and evaluate.	$E(X) = \int_0^4 \frac{3x^2}{56}(5-x)\ dx = \frac{3}{56}\left[\frac{5x^3}{3} - \frac{x^4}{4}\right]_0^4 = \frac{16}{7}$
Use the definition of $\text{Var}(X)$. For this example, as we have already calculated $E(X)$, we will use $\text{Var}(X) = E(X^2) - (E(X))^2$.	$\text{Var}(X) = \int_0^4 \frac{3x^3}{56}(5-x)\ dx - \left(\frac{16}{7}\right)^2 = \frac{3}{56}\left[\frac{5x^4}{4} - \frac{x^5}{5}\right]_0^4 - \left(\frac{16}{7}\right)^2$ $= \frac{216}{35} - \frac{256}{49} = 8.276$

Mode, median, percentiles and skewness

The mode of a continuous random variable is the value of x that the p.d.f is at it's maximum or equivalently the value of x that the probability density function is the densest. Like in discrete random variables, it is possible for a random variable to have more than one mode.

Example 5: The random variable X has probability density function $f(x) = \begin{cases} -\frac{2}{9}x^2 + \frac{10}{9}x - \frac{8}{9}, & 1 \leq x \leq 4 \\ 0, & \text{otherwise} \end{cases}$.

Find the modal value.

The graph of the p.d.f will have the point of greatest probability, equivalently its modal point at the highest point on the graph, which will be a turning point. Differentiate the p.d.f.	$\frac{d}{dx}\left(-\frac{2}{9}x^2 + \frac{10}{9}x - \frac{8}{9}\right) = -\frac{4}{9}x + \frac{10}{9}$
To find the turning point, set the differentiated expression to 0 and solve for x.	$-\frac{4}{9}x + \frac{10}{9} = 0 \Rightarrow \frac{10}{9} = \frac{4}{9}x \Rightarrow x = \frac{5}{2}$

The cumulative density function can be used to define different statistics for a continuous random variable.
If X is a continuous random variable with c.d.f $F(x)$:

- The **median** of X is the value m such that $F(m) = 0.5$
- The **lower quartile** of X is the value Q_1 such that $F(Q_1) = 0.25$
- The **upper quartile** of X is the value Q_3 such that $F(Q_3) = 0.75$
- The ***n*th percentile** of X is the value P_n such that $F(P_n) = \frac{n}{100}$

Example 6: A random variable X has probability density function $f(x) = \begin{cases} \frac{1}{3}, & 0 \leq x \leq 1 \\ \frac{8x^3}{45}, & 1 < x \leq 2 \\ 0, & \text{otherwise} \end{cases}$.

Calculate the median of X.

Find the c.d.f	For $0 \leq x \leq 1$: $F(x) = \int_0^x \frac{1}{3}\,dx = \frac{x}{3}$ For $1 \leq x \leq 2$: $F(x) = \frac{1}{3} + \int_1^x \frac{8x^3}{45}\,dx = \frac{1}{3} + \frac{8x^4}{180} - \frac{8}{180}$ For $x \in \mathbb{R}$ $F(x) = \begin{cases} 0, & x < 0, \\ \frac{x}{3}, & 0 \leq x \leq 1, \\ \frac{13}{45} + \frac{2}{45}x^4, & 1 < x \leq 2, \\ 1, & 2 \leq x. \end{cases}$
To find the median, find the value of m such that $F(m) = 0.5$	From the c.d.f, we know that $F(1) = \frac{1}{3}$, so the median won't fall in the range $0 \leq x \leq 1$ and we look at $1 < x \leq 2$. $\frac{13}{45} + \frac{2}{45}m^4 = 0.5 \Rightarrow m^4 = 4.75 \Rightarrow m = 1.48$

You should also be able to describe the **skewness** of a distribution.

- If a p.d.f is 'bottom heavy', or the max value of $f(x)$ is in the lower values of x then it is **'positively skewed'**.
- If a p.d.f is symmetrical, it can be described as having **no skew**.
- If a p.d.f is 'top heavy', or the max value of $f(x)$ is in the higher values of x, then it is **'negatively skewed'**.

You can also decide whether a distribution is positively or negatively skewed by comparing measures or central tendency, the median and mode:

Positive skew: *mode* < *median* < *mean*
Negative skew: *mean* < *median* < *mode*

It is sufficient to compare one pair of measures of central tendency to justify skewness.

Example 7: The continuous random variable X has probability density function $f(x) = \begin{cases} \frac{x}{8}, & 0 \leq x \leq 4 \\ 0, & \text{otherwise} \end{cases}$.

Find the median and mode and comment on the skewness.

To find the median, find the value of m such that $F(m) = 0.5$. $F(x)$ can be found by inspection.	$F(x) = \begin{cases} 0, & x < 0 \\ \frac{x^2}{16}, & 0 \leq x \leq 4 \\ 1, & 4 < x \end{cases}$ $\frac{m^2}{16} = 0.5 \Rightarrow m^2 = 8 \Rightarrow m = 2\sqrt{2}$
The mode is the value of x such that the value of $f(x)$ is the greatest.	As $\frac{x}{8}$ is an increasing function, it is clear to see that the mode of $f(x)$ is at 4.
Compare the median and mode and describe the skew.	As $2\sqrt{2} < 4$, median < mode, and therefore the distribution is negatively skewed.

The continuous uniform distribution

A random variable that has a continuous uniform distribution over the interval $[a, b]$ has p.d.f $f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & otherwise \end{cases}$.

If X has the continuous uniform distribution over the interval $[a, b]$, it is denoted $X \sim U[a, b]$. For a continuous uniform distribution $U[a, b]$:

- $E(X) = \frac{a+b}{2}$
- $\text{Var}(X) = \frac{(b-a)^2}{12}$
- $F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{x-b}, & a \leq x \leq b \\ 1, & x > b \end{cases}$

These results can all be derived from the skills learned previously in the chapter.

The continuous uniform distribution can be used to model real life situations, such as waiting times (for example knowing that a bus will arrive at a station every 10 minutes, but not knowing when the next bus will arrive is modelled by $X \sim U[0,10]$), or rounding errors (for example rounding to the nearest cm means that the error, R, follows the uniform distribution $R \sim U[-0.5, 0.5]$.