S2 Cheat Sheet

Chapter	Usual types of questions	Tips	What can go ugly
1 – Binomial Distribution	• Finding the probability of a certain number of successes. • Finding the probability of some range of successes, e.g. $P(1 < X \leq 3)$, $P(X > 5), P(X \leq 2)$ • Be able to list the assumptions made in order to model a scenario using a Binomial Distribution. • Calculating mean $= np$ and variance $= npq$ (where $q = 1 - p$) • Sneaky Geometric distribution questions (see right) • Calculating an unknown value of p from context. • Calculating an unknown value of n from context. • Solve problems in which tables have to be used, but the probability of success is > 0.5 (i.e. not in table), by instead counting the number of failures. • Solving "double inequalities", e.g. "smallest value of k such that $P(X \geq k) \leq 0.1$"	• $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ • Remember the two edge cases where you don't need to use the full formula: a. 0 successes: $(1-p)^n$ b. n successes: p^n c. Thus "at least 1 success": $1 - (1-p)^n$ • For worded questions, always start your working by writing out your distribution, e.g. $X \sim B(20, 0.1)$ • Sometimes you require the probability of a range when the cumulative table can't be used, because the value of p is not a nice number. This involves subtracting the opposite cases from 1: e.g. $P(X \geq 2) = 1 - P(X = 1) - P(X = 0)$ • Remember your table requires $\leq$, so if you have $P(X < 5)$, use $P(X \leq 4)$. See on the right regarding problems of 'flipping' your inequality. • Assumptions of a Binomial Distribution: a. Fixed number of trials. b. Probability of success constant. c. Each trial is independent (ensure you put in context!) d. Each trial has two outcomes ('success' and 'failure') • When p is unknown: "An unfair coin with probability of heads p is tossed 20 times. The probability of seeing no heads is 0.1. Determine p". Since this is an 'edge case': $P(X = 0) = (1-p)^{20}$. Thus: $(1-p)^{20} = 0.1$ $1 - p = \sqrt[20]{0.1}$ $p = 1 - \sqrt[20]{0.1} = 0.109$ • When n is unknown: "I play a game for which the probability of winning is 0.7. If I win every game, what is the smallest number of times I play such that the probability of winning every game is less than 0.01?" Again an edge case so: $P(X = n) = 0.7^n$ $0.7^n < 0.01$ $n \log 0.7 < \log 0.01$ $n > \frac{\log 0.01}{\log 0.7} = 12.9$ Thus at least 13 games required. Notice that the direction of the inequality reversed because we divided by a negative number ($\log 0.07$). • Sometimes you'll get a part of a question which requires some non-Binomial probabilistic calculation, particularly involving some number of failures before a success is obtained, e.g. "Bob keeps firing arrows at a target until he gets a Bullseye. The probability he gets a Bullseye is 0.4. What's the probability he hits the Bullseye on the 4th shot": $0.6^3 \times 0.4$	• Misreading terms like "at least" or "more than". Make sure you get $<$ vs $\leq$ right. • Incorrectly 'flipping' a probability for > and $\geq$, i.e. being one off when you like up a value in the cumulative Binomial table. $P(X > 1) = 1 - P(X \leq 1)$ $P(X \geq 1) = 1 - P(X = 0)$ Just think logically about what the opposite of "more than 1" is and so on. • Similarly, incorrect switching from the number of successes to the number of failures, usually by forgetting to replace the value with n minus it, or not preserving the strictness/non-strictness of the inequality.

This is related to something called the Geometric Distribution which isn't formally covered in the syllabus.

- When switching from the number of successes X to the number of failures Y (so that the probability is less than 0.5 for the purposes of using tables), flip the inequality (but preserve $<$ vs $\leq$) and if the number of successes was k, use $n-k$ for number of failures:

$$P(X < k) = P(Y > n - k)$$
$$P(X \geq k) = P(Y \leq n - k)$$

e.g. "In Joe's café 70% of customers buy a cup of tea. In a random sample of 20 customers find the probability that more than 15 buy a cup of tea."

$$X \sim B(20,0.7)$$
$$\therefore Y \sim B(20,0.3)$$
$$P(X > 15) = P(Y < 20 - 15)$$
$$= P(Y < 5)$$
$$= P(Y \leq 4)$$

- "Given $X \sim B(50,0.6)$, find the smallest value of k such that $P(X < k) \geq 0.9$"

We can only use the table if the probability is less than 0.5. This question requires a great deal of care, particularly with the effect of switching from X to Y and getting $<$ vs $\leq$ right!

$$X \sim B(50,0.6)$$
$$Y \sim B(50,0.4)$$
$$P(X < k) = P(Y > 50 - k) \geq 0.9$$
$$1 - P(Y \leq 50 - k) \geq 0.9$$
$$P(Y \leq 50 - k) \leq 0.1$$
$$50 - k \leq 15$$
$$k \geq 35 \quad \therefore \quad k = 35$$

2 – Poisson Distribution	• Be able to state the conditions under which a Poisson distribution may be used. • As with the Binomial Distribution, find the probability of a particular number of events happening within a given time frame, or within some range using tables. • State the mean and variance of a Poisson distribution. • Be able to scale a Poisson distribution to a different time period. • Feed the probability obtained from a Poisson distribution into a Binomial Distribution, and make subsequent calculations. • Approximate a Binomial Distribution using a Poisson Distributions, and be able to state the conditions under which we can do so. • Again, forming and solving 'double inequalities': e.g. "smallest value of k such that $P(X \geq k) \leq 0.1$"	• $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ While this is in the formula booklet, the easy way to remember it is that reading left to right and then down, the λ repeats consecutively, as does the x. • $E(X) = \lambda$ and $Var(X) = \lambda$. The fact these are the same sometimes provides a justification for why a Poisson distribution would be suitable to model certain data. See 'Wordy Questions' page. • As with the Binomial Distribution, ensure you state the distribution for any wordy question, e.g. $X \sim Po(5)$. • Conditions required for Poisson: a. Events occur independently (e.g. volcano eruptions might not be modelled using Poisson because volcano less likely to erupt immediately after previous eruption, thus eruptions not independent). b. Events occur singly in time. c. A fixed rate for which events occur. • A very common occurrence is that you will need to scale to another time period. e.g. "A printer jams on average 0.3 times an hour. Find the probability over a 5 hour period the printer jams at least 4 times." *Just scaling the 3 times an hour to a 15 times every 5 hours:* $X \sim Po(0.3 \times 5) \rightarrow X \sim Po(1.5)$ $P(X \geq 4) = 1 - P(X \leq 3)$ $= 1 - 0.9344 = 0.0.0656$ • Another common question is to feed the value calculated from a Poisson question, into a Binomial Distribution. e.g. "Defects occur in planks of wood with rate 0.5 per 10cm. If Bob buys 6 blanks each of length 100cm, find prob that fewer than 2 of planks contain at most 3 defects." *First find probability a plank of 100cm contains at most 3 defects:* $X \sim Po(5) \qquad P(X \leq 3) = 0.2650$ *Then feed into a Binomial Distribution:* *Let Y be the number planks with at most 3 defects.* $Y \sim B(6, 0.2650)$ $P(Y < 2) = P(Y \leq 1) = P(Y = 0) + P(Y = 1)$ $= 0.735^6 + 6 \times 0.265 \times 0.735^5$ *Notice that we couldn't use tables for the Binomial here because of the non-nice p value.* • As with the Binomial Distribution, we can get nasty 'double inequality' questions: "While a popcorn bag is in the microwave, an average of 5 pops can be heard per second. What's the minimum number of pops heard such that there's less than a 10% chance of hearing more than this number of pops?" $X \sim Po(5)$ $P(X > k) < 0.1$ $1 - P(X \leq k) < 0.1$ $P(X \leq k) > 0.9$ $k = 8$ Note we had to take care with $X > k$ vs $X \geq k$ and < 0.1 vs ≤ 0.1. • For Binomial → Poisson approximations, see notes on Normal Approximations.	• Since these questions are very similar to Binomial questions (except we use a different table and evaluate the probabilities differently), the same potential problems can arise, e.g. not correctly flipping an inequality from $P(X > k)$ to $1 - P(X \leq k)$ and so on. • Not realising a Poisson question is a Poisson question! Remember that any mention of 'rate' implies Poisson rather than Binomial. • Also not realising when both a Poisson Distribution and a Binomial Distribution have to be used within the same question!

3 – Continuous Random Variables	• Find the probability of some range of values given a probability density function or cumulative distribution function. e.g. $P(X > 2), P(1 < X < 2)$ • Appreciate that $P(X = k) = 0$ if X is a continuous random variable. • State the probability density function given a graph. • Comment on the skew of a distribution. • Calculate $E(X), Var(X)$, the median/quartiles and the mode of a probability density function. • Convert from $f(x)$ to $F(x)$ and vice versa, potentially involving multiple ranges. • Be able to calculate $E(X^2)$ (surprisingly common!) • Be able to calculate the median and quartiles. • Be able to calculate the mode, either by finding the turning point, or by inspection of the graph.	• Remember that $f(x)$ is the probability density function for continuous variables, and that this is not a 'probability' as such: we only get a probability when we integrate $f(x)$ over some range. • Relatedly, if X is a continuous variable, then $P(X = k) = 0$ because the probability of a specific value is infinitely small (e.g. no one has an 'exact' height of 1.5m). • Think of $F(x)$ as "the running total of the probability up to x". $$F(x) = P(X < x)$$ • To find a probability over a range: $$P(a < X < b) = \int_a^b f(x)\,dx$$ $$P(X > a) = \int_a^\infty f(x)\,dx$$ For the latter, if you know the probability is 0 after some value b (this will always be the case in exams), we can use b instead of ∞. • If $F(x)$ is known, then you could calculate say $P(X > 2)$ using $P(X > 2) = 1 - P(X \le 2)$ $$= 1 - F(2)$$ • When asked to find the value of some constant k used in a p.d.f., use the fact that $\int_{-\infty}^{\infty} f(x)\,dx = 1$, i.e. the area under the whole probability function is 1. • However, if the cumulative distribution function is given, then there is no need to integrate, just use $F(b) = 1$ where b is the highest possible value (since $P(X \le b) = 1$). • Remember that it doesn't matter for continuous variables if you use $X > k$ or $X \ge k$ (but it does matter for discrete variables!). • $E(X) = \int_{-\infty}^{\infty} x\,f(x)\,dx$ • $Var(X) = E(X^2) - E(X)^2$ $$= \int_{-\infty}^{\infty} x^2\,f(x)\,dx - \mu^2$$ • To go from $f(x)$ to $F(x)$, integrate. e.g. $$f(x) = \begin{cases} \frac{2}{3}x & 1 < x \le 2 \\ 0 & otherwise \end{cases}$$ Then in the $1 < x \le 2$ range: $$F(x) = \int_1^x \frac{2}{3}t\,dt = \left[\frac{1}{3}t^2\right]_1^x = \frac{1}{3}x^2 - \frac{1}{3}$$ Therefore the full cumulative distribution function is: $$F(x) = \begin{cases} 0 & x \le 1 \\ \frac{1}{3}x^2 - \frac{1}{3} & 1 < x \le 2 \\ 1 & x > 2 \end{cases}$$ Note the extra '1' row required since the running total of the probability by the time you get to 2 is 1. Note that the use of t was to avoid a clash with the x used as the upper limit of the integral. But the mark scheme permits $\int_2^x \frac{2}{3}x\,dx$, so do this way if you find it less confusing.	• A common error is doing $P(X \ge 10) = 1 - P(X \le 9)$ for continuous variables. This is true for discrete variables, but the opposite of "more than 10m tall" is not "under 9m tall"! $$P(X \ge 10) = 1 - P(X < 10)$$ Note that the $<$ vs $\le$ does not matter since X is continuous. • Note paying attention to whether the question gives the cumulative distribution function $F(x)$ or the probability distribution function $f(x)$. This will completely change the approach to use for answering a question! If you're finding the median/quartiles and you're already given $F(x)$, don't integrate! • When finding the cumulative distribution function, forgetting the rows for the two 'ends' of the ranges. You should have one more row in $F(x)$ than $f(x)$. The same applies when going from $F(x)$ to $f(x)$: you should have one less row. • When finding the cumulative distribution function, then in the example on the left, it might have been tempted to go straight from $\frac{2}{3}x$ to

When $f(x)$ has multiple rows, see the note on the right about ensuring you add the running total up to that range.

- To go from $F(x)$ to $f(x)$ just differentiate. Don't forget that the '1' row disappears.
- To find median or quartiles: use $F(Q_1) = 0.25, F(Q_2) = 0.5, F(Q_3) = 0.75$. Either $F(x)$ will already be given, or you will have to determine it from $f(x)$ first.
  Sometimes you have to determine which range the quartile/median occurs in first by evaluating the borderline values (although this is not necessary if you only have one range).
  e.g. If:

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{4}x^2 & 0 \le x \le 1 \\ \frac{1}{20}x^4 + \frac{1}{5} & 1 < x \le 2 \\ 1 & x > 2 \end{cases}$$

  and we wished to find the median Q_2, then it might be in the $0 \le x < 1$ range or $1 < x \le 2$ range. However $F(1) = \frac{1}{4}$, i.e. the running total of the probability up to 1 is 0.25, thus the median wouldn't have yet occurred, and thus it's in the $1 < x \le 2$ range.
  Then using $F(Q_2) = 0.5$:

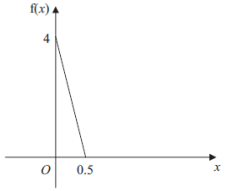
$$\frac{1}{20}x^4 + \frac{1}{5} = \frac{1}{2}$$

  and so on.
- The mode is the value of x such that $f(x)$ is at its maximum. The mode can be calculated in two different ways: (and usually you can only use one of the two)
  a. For curved graphs, finding the turning points using

$$\frac{df(x)}{dx} = 0$$

  b. Using the graph. e.g. If $f(x) = \begin{cases} 1 - \frac{1}{2}x & 0 < x < 2 \\ 0 & otherwise \end{cases}$, then we can see from the sketch that the probability is greatest when $x = 0$, so the mode is 0.
- If asked to find the probability density function of a given graph, ensure you don't just give the equation of the line you see: you need to use the full curly brace construction covering all values:
  Then $f(x) = 4 - 8x$ is not enough, we need to write:

$$f(x) = \begin{cases} 4 - 8x & 0 \le x \le 4 \\ 0 & otherwise \end{cases}$$

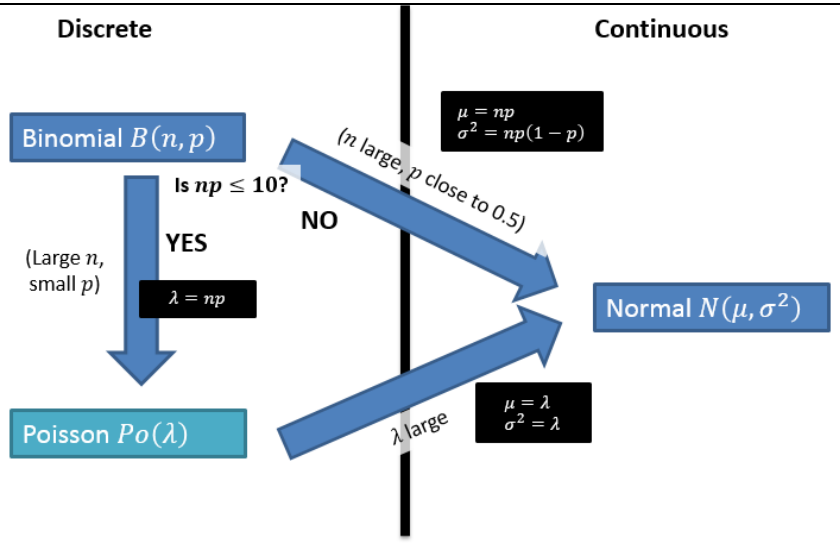


- When asked to find skew of a probability density function, if two of the mode, median and mean have been found, compare in the usual S1 way (remembering if positive skew that $mean > median > mode$). If you only have the graph look at the shape: a 'positive tail' mean positive skew.

$\frac{1}{3}x^2$ without properly evaluating $\int_1^x \frac{2}{3}t\,dt$
This will result in a missing constant.

- When finding the cumulative distribution function from $f(x)$ where there are multiple ranges, forgetting to add on the running total up to the start of the range being considered. e.g. If you had ranges $0 < x \le 1$ and $1 < x \le 2$ in $f(x)$, then when finding $F(x)$ in the latter range, $F(x) = F(1) + \int_1^x f(t)\,dt$. This is because you want the area up to 1 and then the area between 1 and x.
  i.e. Don't forget the $F(1)$!
- When finding the mode, accidentally giving the probability density of the mode as the answer rather than the mode itself (e.g. Jan 2011 Q5d: answer is 0 not 4).

4 – Continuous Uniform Distribution	• Find the probability of a range for a continuous uniform distribution, e.g. $P(2 < X < 3)$ or $P(X > 3)$. • Sketch the probability function of a continuous uniform distribution. • Find the mean or variance of a continuous uniform distribution: • Find the a and b of $X \sim U(a, b)$ when the mean and variance of the distribution is given. • Be able to calculate $E(X^2)$ • The probability calculated from a uniform distribution may be fed into a Binomial distribution.	• If $X \sim U(a, b)$ then: $$E(X) = \frac{a+b}{2}$$ $$Var(X) = \frac{(b-a)^2}{12}$$ I remember the variance as "a twelfth of the squared difference". • Suppose $E(X) = 2$ and $Var(X) = 27$ and $X \sim U(a, b)$ where a and b are unknown. Then: $$\frac{a+b}{2} = 2 \quad \rightarrow \quad a + b = 4$$ $$\frac{(a-b)^2}{12} = 27 \quad \rightarrow a - b = 18$$ We can then solve these simultaneously. • The key is just remembering the area of the rectangle (when you sketch the p.d.f.) is 1. Therefore if $X \sim U(3,5)$ then since the width is 2, the height is clearly $\frac{1}{2}$. We would specify the probability density function as: $$f(x) = \begin{cases} \frac{1}{2} & 3 \le x \le 5 \\ 0 & otherwise \end{cases}$$ • For the example above, we could then find $P(X > 4.2)$ by just considering the rectangular area involved: $$P(X > 4.2) = 0.8 \times 0.5 = 0.4$$ • $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx$ For the above example, $E(X^2) = \int_3^5 \frac{1}{2} x^2\, dx = \left[\frac{1}{6}x^3\right]_3^5 = \frac{49}{3}$ • "I pick 10 real numbers randomly from 12 to 17. Find the probability that at least 5 of these numbers are greater than 15.5." If X is the number picked each time, $P(X > 15.5) = (17 - 15.5) \times \frac{1}{5} = 0.3$. Then if Y is the number of times a number greater than 15.5 was picked, $Y \sim B(10, 0.3)$, and calculate $P(Y \ge 5)$. • Don't forget your rules of coding from S1: $$Var(7X) = 49Var(X)$$ $$Var(2 - 5X) = 25Var(X)$$	• Suppose that $X \sim U(3,5)$. Then what is $P(4.2 < X < 6)$? You might be tempted to calculate $(6 - 4.2) \times \frac{1}{2}$, but the probability above a value of 5 is 0, thus: $P(4.2 < X < 6) = P(4.2 < X < 5) = (5 - 4.2) \times \frac{1}{2}$. i.e. we 'truncate' any part of the range which is outside the range of the uniform distribution.

5 – Normal Approximations	• Be able to approximate a Binomial Distribution using a Normal Distribution. • Be able to approximate a Poisson Distribution using a Normal Distribution. • Be able to approximate a Binomial Distribution using a Poisson Distribution. • Be able to give the conditions under which such approximations can be made. • Justify why we need continuity corrections.	 Continuity correction • This diagram may seem like quite a lot to memorise, but all you need to memorise for carrying out the majority of approximations is this: If you have a Binomial Distribution, is $np \leq 10$? If yes use Poisson, else use Normal. In terms of converting between the distributions, the mean and variance of the Normal/Poisson approximation is just the mean and variance of the original distribution. • If asked why a continuity correction is needed (and suppose the original distribution is Poisson), say: "Poisson is discrete, but Normal is continuous". • For continuity corrections, we want to go from a discrete X to a continuous version of it Y. You will never get a continuity correction wrong if you carry out these two simple steps: a. Make sure your inequality uses $\leq$ or $\geq$ instead of < or >. i.e. Ensure inequality is non-strict. b. 'Extend' your range by 0.5 at each end. i.e. If you visualise your inequality as a line on the number line, it should be 0.5 longer each end. Examples: $P(X < 4) = P(X \leq 5) = P(Y \leq 5.5)$ $P(3 < X \leq 5) = P(4 \leq X \leq 5) = P(3.5 \leq Y \leq 5.5)$ $P(X = 5) = P(4.5 \leq Y \leq 5.5)$ • I prefer to do the continuity correction immediately, i.e. before you either reverse the direction of the inequality or standardise. e.g. $P(X > 5) = P(X \geq 6) = P(Y \geq 5.5) = 1 - P(Y < 5.5) = 1 - P(Z < \cdots)$ • The number of marks effectively tells you what approximation you are using: If at least 6 marks, it's a Normal Approximation (because of the many steps of converting the distribution, standardising and continuity corrections), otherwise it's Binomial → Poisson.	• If asked to give the conditions under which a Binomial Distribution can be approximated using a Poisson Distribution, do NOT say $np \leq 10$. This condition is a rule of thumb only: the actual condition is "n is large, p is small" (from which $np \leq 10$ stems). • All manner of things can go wrong with continuity corrections. This might be forgetting to convert to $\geq$ or $\leq$ first (i.e. incorrectly going from $P(X < 4)$ to $P(Y < 4.5)$), or making your range 0.5 smaller rather than 0.5 larger, e.g. incorrectly from $P(X \leq 4)$ to $P(Y \leq 3.5)$. • See note on the left about the perils of scaling the value instead of λ in the case of the Poisson Distribution. • In the formula for Z, accidentally dividing by the variance rather than the standard deviation.

		• Example Normal Approximation: "*The number of houses sold by an estate agent follows a Poisson distribution, with a mean of 2 per week. The estate agent will receive a bonus if he sells more than 25 houses in the next 10 weeks. Use a suitable approximation to estimate the probability that the estate agent receives a bonus*." a. Note first that you might be tempted to scale the 25 houses in 10 weeks to 5 houses in 2 weeks and stick with the original $\lambda = 2$. The catastrophic flaw in doing this is that the continuity correction affects the range differently depending on whether you're using the original or scaled value. If not scaling: $P(X > 25) = P(X \geq 26) = P(Y \geq 25.5)$ If scaling: $P(X > 5) = P(X \geq 6) = P(Y \geq 5.5)$ In the latter incorrect case the 0.5 has a greater effect on the smaller value of 6 compared with the larger value of 26, so the probability will be too high. b. Step 1: Determine what approximation to use. In this example we have a Poisson Distribution, which always goes to Normal. If it were Binomial, you'd first determine if $np \leq 10$. c. Step 2: Identify original distribution. As discussed, we scale λ (rather than the 25), so: $X \sim Po(20)$ d. Step 3: Write the approximation, potentially with reference to a new continuous variable Y which is the continuous version of X, i.e.: $Y \sim N(20,20)$ As discussed, use the mean and variance of the original distribution. e. Step 4: If necessary, carry out continuity correction to get a probability in terms of Y: $P(X > 25) = P(X \geq 26) = P(Y \geq 25.5)$ f. Step 5: Use your S1 knowledge and find the probability by first standardising. Don't forget that you're dividing by the standard deviation, not the variance: $P(Y \geq 25.5) = P\left(Z \geq \frac{25.5 - 20}{\sqrt{20}}\right) = P(Z \geq 1.23)$ $= 1 - P(Z \leq 1.23) = 1 - 0.8907 = 0.1093$	
6 – Populations and Samples	• Be able to define key terms such as 'sampling distribution', 'sampling frame', 'population', 'sample', 'statistic'. • Be able to describe what the population is or sampling frame is given the context, and identify reasons for differences between the two. • Be able to list possible samples. • Be able to calculate the	• Key definitions: a. **Statistic**: "*A random variable (1) which is some function of the sample and not dependent on any population parameters (1)"* - I think the 'random variable' bit is a bit pernicious (as does Wikipedia), but c'est la vie! If 1 mark, the second part is important. *b.* '**Population**': *The collection of all items.* *c.* '**Sample'**: *Some subset of the population which is intended to be representative of the population.* *d.* '**Census'**: *When the entire population is sampled.* e. '**Sampling unit'**: *Individual member or element of the population or sampling frame.* f. '**Sampling frame'**: *A list of all sampling units or all the population.* *g.* **Sampling distribution***: All possible samples are chosen from a population (1); the values of a statistic and the associated probabilities is a sampling distribution (1).* • It's important you get your head around what the sampling distribution actually is: It gives the distribution over possible values of the statistic as we take different samples. So if for example	• When finding the sampling distribution, forgetting that different orderings are different possibilities, e.g. (1,1,2) and (1,2,1) should both be considered. • Even though you can subtract from 1 to find the last probability in your sampling distribution, if you have time, you might want to calculate it 'the long way' to check your answer, as probabilities

	sampling distribution for a variety of statistics, such as median (very common!), range, maximum and mode. • Be able to identify when the sampling distribution is a Binomial Distribution or otherwise, and specify this distribution.	the statistic was the 'range' of the sample, then this range is likely to vary as we take different samples. As these ranges <u>vary across samples</u>, it forms a distribution. • The sampling frame is the list of things in the population that are available for sampling, e.g. "The ID numbers", "The list of car registration numbers". The mark scheme seems to particularly like it when you refer to some identifying property of the things in the sampling frame. The sampling frame may be different from the population, because some things in the population may not be available for sampling. e.g. If sampling people who've visited a medical practice, "some people may have left the area but hadn't deregistered". • When listing outcomes, it helps to be systematic in listing them so you don't miss any. Note that <u>different orderings count as distinct possibilities</u>. e.g. "You have a large collection of 1p, 2p and 5p coins, and take 3 coins. Find all samples in which the maximum is 5." We may want to first list the possibilities where 5 appears once, 5 appears twice, and so on…" (5,1,1), (1,5,1), (1,1,5), (5,1,2), (5,2,1), (1,5,2), (2,5,1), (1,2,5), (2,1,5), (5,2,2), (2,5,2), (2,2,5), (5,5,1), (5,1,5), (1,5,5), (5,5,2), (5,2,5), (2,5,5), (5,5,5) • e.g. "A bag contains a large number of 1p and 2p coin, of which 40% are 1p and 60% are 2p. A sample of 2 coins. Find the sampling distribution of the sample maximum." When finding the sampling distribution, it may help to have a table as follows to organise your working, such that the outcomes for each possible value of the statistic are grouped: Notice that we didn't need to do any complicated calculation for the last probability, because it was just 1 minus the others! Had we had to calculate it fully, then $2 \times 0.4 \times 0.6 + 0.6^2 = 0.84$ • On the rare occasion you get a question asking for the sampling distribution, where you don't actually have to do any calculation, but just have to consider what well-known distribution you get as the sample varies: "*A factory produces components. Each component has a unique identity number and it is assumed that 2% of the components are faulty. On a particular day, a quality control manager wishes to take a random sample of 50 components. A statistic F represents the number of faulty components in the sample. Specifying the sampling distribution of F.*" We know a sampling distribution is the possible values of the statistic as we take different samples of 50 light bulbs. If the statistic is the count of light bulbs, we can see this count varies Binomially between 0 and 50. Thus $F \sim B(50, 0.02)$	should obviously all add up to 1.

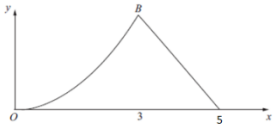
Possibilities	Statistic (Maximum)	Probability
(1,1)	1	$0.4^2 = 0.16$
(1,2), (2,1), (2,2)	2	$1 - 0.16 = 0.84$

7 – Hypothesis Testing	• Be able to define key terms such as 'critical region', 'hypothesis test', 'significance level of a test'. • Determine a critical region (one or two ranges depending on whether one or two tailed) • Understand when we need the probability to be strictly within the significance level, and when to be close to it as possible either side. • Be able to calculate the 'actual level of significance'. • Carry out a hypothesis test involving Binomial and Poisson distributions. • Carry out a hypothesis test where a normal approximation is required.	• Key terms whose definitions you need to remember: a. **Critical Region**: The range of values such that the null hypothesis is rejected. b. **Hypothesis Test**: a procedure to examine a value of a population parameter proposed by the null hypothesis c. **Significance level**: "the probability of rejecting H_0 if H_0 is true", or "the probability of incorrectly rejecting H_0". • The "or more extreme" thing often confuses people: is "more extreme" below the value or above it? We can always tell this by what side of the mean we're on. If $\lambda = 7$ and we're interested in seeing 10 hits to a website "or more extreme", then since 10 is above the mean of 7, clearly we want "10 or above" to get the tail. Ensure "10 or above" is $X \geq 10$ not $X > 10$. • **Identifying the critical values from a table**: Firstly, note if the test is one-tailed (involving $<$ or $>$) or two-tailed (involving $\neq$), as you need to halve the significance level each side in the latter case. At the left tail, find the closest value under the significance level (or half it) in the table, or if you are explicitly told to use the closest value to the significance level, do that. This is the lower end of the critical region. However, at the right tail, first find the closest value above it (or again if explicitly told, the closest value either side), but then go one above it. This is the critical value. Example: *Under the null hypothesis $X \sim B(10,0.35)$. For a two-tailed test with significance level 5%, what are the critical values?* Looking at the tables, at left end $X = 0$, and at right end, closest value with probability above 0.975 is 7, thus going one above $X = 8$. If we had been asked for the 'closest value' to 0.025 and 0.975, then we'd then get $X = 0$ and $X = 7$ instead. It is vitally important you specify the probability of being in each tail to evidence that you have used the table, e.g. "$P(X \geq 8) = 0.9952$ • For the critical region, don't forget to provide a lower limit or upper limit in the case of the Binomial Distribution, as the outcomes are finite. For the previous example: $X = 0 \text{ or } 8 \leq X \leq 10$. Mark schemes usually condone the lack of ≤ 10, but don't take any chances. • The actual level of significance is the actual probability of being in the critical region. You should have already written out the probabilities of being in each part of the critical region, so it's then just a case of adding the two probabilities. • The mark scheme for a hypothesis test without a normal approximation is as follows: a. Specifying H_0 and H_1 (1 mark) b. Specifying the distribution for X under the null hypothesis, e.g. $X \sim B(10,0.2)$ (1 mark) The p or λ will be your population parameter under the null hypothesis. c. 2 marks for either: Determining the probability of the observed value or more extreme (e.g. $P(X \geq 5) = 1 - P(X \leq 4) = \cdots$ or determining the critical region.	• Forgetting to halve the significance level for two-tailed tests. • Being one off the critical value (particularly at the right tail). See notes on the left. Pay careful attention to whether it says "use the closest value", or not. • Getting your continuity correction wrong (see notes for Chapter 5).

		d. Using your probability to state whether H_0 is rejected or not, ensuring you directly compare your probability with significance level. e.g. "0.0723 > 0.05, so not significant (H_0 is not rejected)" (1 mark) e. Put this conclusion in context. "Bob is not justified in his claim that the rate of flamingo attacks has increased." • The mark scheme for a hypothesis test with a normal approximation is usually broken down as the following: a. Specifying H_0 and H_1 (1 mark) b. Possibly a mark for specifying your distribution of X, i.e. $X \sim B(\dots)$ or $X \sim Po(\dots)$ c. Specifying the distribution of Y for your normal approximation, $Y \sim N(\dots, \dots)$ (1 mark) d. Doing the continuity correction: $P(X \geq 40) = P(Y \geq 39.5)$ (1 mark) e. Standardising to get $P(Z \dots)$. Note at this stage $>$ vs $\geq$ does not matter as variable is continuous. (1 mark) f. As above, 2 marks for your two-part conclusion. • Note that it is possible for the actual level of significance to be greater than the level of significance, if you were asked to find the closest value to 0.025/0.975, etc rather than those strictly below/above these values.	

Wordy/interpretation questions:

- Definitions (see the respective chapter notes above):
  Statistic, population, census, sampling frame, sampling unit, sampling distribution, critical region, hypothesis test, significance level.
- Modelling assumptions:
  - "List the assumptions made when modelling as a Binomial distribution". See Chapter 1.
  - "State two conditions under which a Poisson distribution is a suitable model to use in statistical work". See Chapter 2.
- (Surprisingly common!) "Describe the skewness of X, giving a reason for your answer." *The mode and mean had previously been calculated in this question so "Positive skew as $mean > mode$". 1 mark for skew type, 1 for reason.*
- Approximation related:
  - "Write down the conditions under which the Poisson distribution can be used as an approximation to the Binomial distribution."
    λ is large
  - "Write down the two conditions needed to approximate the Binomial distribution by the Poisson distribution."
    n is large, p is small (<u>NOT</u> $np \leq 10$)
  - "Write down which of the approximations used in part (a) is the most accurate estimate of the probability [One was Poisson, one Binomial]. You must give a reason for your answer."
    Normal approximation (1) because n is large and p close to half (1). <u>OR</u> Normal because $mean \neq variance$
- "State the probability of incorrectly rejecting H_0 using this critical region."
  Add the probabilities of your critical region(s).
- [Given that X is a continuous variable] "Determine $P(X = 3)$". *Answer is 0.*
- "Identify a sampling frame."
  See notes on Chapter 6. Remember the mark scheme likes reference to an 'identifier', e.g. "list of unique identification numbers of the cookers."
- [In context of cookers being tested] "Give one reason, other than to save time and cost, why a sample is taken rather than a census."
  There would be no cookers left to sell (i.e. the idea that testing something in a sample destroys it/makes it unsellable).
- "Identify the sampling units". *(Using above example) A cooker*
- "A researcher took a sample of 100 voters from a certain town and asked them who they would vote for in an election. The proportion who said they would vote for Dr Smith was 35%. State the population and the statistic in this case. What do you understand by the sampling distribution of this statistic."
  Population is the residents of the town. Statistic is percentage/proportion who vote for Dr Smith. Sampling distribution here is number of people who voted for Dr Smith in all possible samples of 100.
- [List of possible statistics given] "State, giving a reason which of the following is not a statistic based on this sample."
  The one involving the μ and σ, because these are population parameters, whereas a statistic must be based on the sample only.
- "Suggest a suitable model to describe the number of vehicles passing the fixed point in a 15 s interval." $X \sim Po(9)$ *(depending on what the average rate in the question is, which often you have to scale)*
- [Using distribution pictured] "state, giving your reason, whether E(X) < 3, E(X) = 3 or E(X) > 3."
  $E(X) < 3$ *because the graph tails to the left (i.e. negative skew).*



- "Find, to 2 decimal places, the value of k so that $P(\mu - k\sigma < X < \mu + k\sigma) = 0.5$." *This means within k standard deviations either side of the mean, so be definition, we want $P(-k < Z < k)$. Since this is middle 50%, upper end is upper quartile so find when $P(Z = k) = 0.75$.*
- "Comment on this finding in the light of your critical region found in part (*a*)." *11 is in the critical region (1 mark) therefore there is evidence of a change/increase in the proportion/number of customers buying single tins (1 mark).*
- [Mean and variance of some data was calculated] "Explain how the answers from part (c) support the choice of a Poisson distribution as a model."
  For a Poisson model , Mean = Variance ; For these data $3.69 \approx 3.73$