S3 Cheat Sheet

Chapter	Usual types of questions	Tips	What can go ugly
1 – Combinations of Random Variables	• Be able to combine two Normally Distributed variables: $E(X \pm Y) = E(X) \pm E(Y)$ $Var(X \pm Y) = Var(X) + Var(Y)$ (note we always add variances) • Find the distribution of some new variable, e.g. $A = 2B + 3C$ where the distributions of B and C are known. • Find $P(B > A)$ for two normally distributed variables. • Be able to turn a worded description into some expression involving normally distributed variables, e.g. "Find the probability that a cat is twice as tall as two dogs combined": $C > 2D_1 + 2D_2$ • Be able to calculate $P(\|X - Y\| < k)$ for some constant k.	• The key is to appreciate that difference between $2X$ and $X_1 + X_2$, where in the first we're doubling the value of a single random variable, while in the latter we're adding the values of two distinct variables (each with distribution X). In the former: $Var(2X) = 4Var(X)$ In the latter: $Var(X_1 + X_2) = Var(X_1) + Var(X_2) = 2Var(X)$ To appreciate the difference, consider doubling the thickness of a piece of paper versus putting two separate pieces of paper on top of each other. Because one piece's thickness deviating above the mean thickness can 'cancel out' the thickness of another piece below the mean, we can see in the former case the variance will be greater. • If $A = \sum_{i=0}^{3} B_i$, be sure you interpret this as $B_1 + B_2 + B_3$ not $3B$ (see above!). • To deal with $P(B > A)$, use the fact that $P(B - A > 0)$. We can combine $B - A$ into a single normally distributed variable, say C, and can then calculate $P(C > 0)$ in the usual S1 way. • Note that $P(\|X - Y\| < 3)$ means "there is at most 3 difference between X and Y. The difference could be either positive or negative thus: $P(\|X - Y\| < 3) = P(-3 < X - Y < 3)$ As before, find the distribution of the combined variable $A = X - Y$ then calculate the probability $P(-3 < A < 3) = P(A < 3) - P(A < -3)$. • Note that it's the variances that get added, not the standard deviation. The standard deviation of two variables added would be $\sqrt{\sigma_1^2 + \sigma_2^2}$, i.e. we square the standard deviations to get variances, then add them to get new variance, and square root result to get back to s.d.	• Confusing $2X$ and $X_1 + X_2$, particularly when processing word problems, e.g. "two dogs" should be $D_1 + D_2$ and NOT $2D$, as two separate dogs are being considered. • Not correctly handling modulus | … | function. • Accidentally subtracting variances in $X - Y$
2 – Sampling	• Determine the most suitable sampling method given the context. • Explain why a sampling method may not be suitable given the context. • Explain how to carry out a given sampling method (whether describing more generically or in a given context). • Give the advantages and disadvantages of each sampling method. • Be able to use the random number table in the formula	Just know your mark scheme. I've taken all these advantages/disadvantages from mark schemes rather than the textbook. Note that for advantages/disadvantages when comparing two different sampling methods, mark schemes sometimes say "no repetitions/opposites", i.e. you can't have the advantage of one as the disadvantage of the other! • Simple Random Sampling: a. "Explain how to sample 15 out of 120 students using simple random sampling": *Allocate a number between 1 and N (or equiv) to each pupil. (1 mark)* *Use random number tables, computer or calculator to select 15 different numbers between 1 and 120 (or equiv). (1 mark)* *Pupils corresponding to these numbers become the sample. (1 mark)* Note that the important parts here are allocating some kind of identifier to each object, describing a method to generate a random number, and then matching the number against the identifiers. b. Advantages: "Random process so possible to estimate sampling errors", "Free from bias".	• Just revise all these verbatim!

	booklet for simple random sampling.	c. Disadvantages: "Not suitable when sample size is large", "Sampling frame required which may not exist or difficult to construct for a large population". • Stratified Sampling: a. Advantages (over random sampling): "Enables estimation of statistics for each strata/Reduce variability" or "More representative" or "Reflects population structure". b. Disadvantages: "Sampling frame is required", "Strata may not be clear as may be overlap between groups" or "not suitable for large populations". c. Explain how to carry out stratified sampling of 200 people from 6000 full time staff and 4000 part time staff: "(1) Label full time staff 1-6000, part time staff 1-4000" (2) Use random numbers to select staff from each group (3) Simple random sample of 120 full time staff and 80 part time" Notice that marks come from identifying the number from each group, and the two usual marks for the description of random sampling. d. "Explain why stratified sampling in this context could not be used." (Note this question is also applicable to simple random sampling) Answer: "*Because the sampling frame is impossible to maintain."* This question came up in the context of a lake of fish – it's obviously very difficult to know how many fish there are in the lake and how many of each type. • Systematic Sampling: a. Explain how you could sampling 100 names from 50 000 in a phone directory.": *"(1) Randomly select number from 1 to 500* *(2) Select that person and every 500th person after that"* Notice it's important that the person is selected randomly b. Advantages: "Simple/Easy to carry out" (they don't want "quick/cheap") "Suitable for large samples" (not allowed to say 'large populations') c. Disadvantages: "Only random if original list truly random" "Requires a list of the population/Must assign number to each member of population" • Quota Sampling: a. "Explain how to carry out": Non-random sampling (1 mark) from groups of the population (1 mark). b. Advantages: "field work can be done quickly", "can be done cheaply", (unlike stratified sampling) "sampling frame not required", "administration easy", "non-response not an issue". c. The last advantage is important because quota sampling is an alternative to simple stratified sampling when the sampling frame isn't possible to obtain. This has been tested on in the past. d. Disadvantages: "Not possible to estimate sampling errors", "Process not random" (and liable to the surveyor's bias), "non-response not recorded".	

3 – Estimation, Confidence Intervals and Test	- Know the definition of the Central Limit Theorem (3 marks) - State the relevance of the Central Limit Theorem in context. - S2: Understand the definition of a statistic and identify whether certain functions are statistics or not. - S2: Find the sampling distribution of a range of statistics where it is possible to enumerate all possible samples. - Understand what $\bar{X}$ is and find the distribution of $\bar{X}$. - Make calculations about the mean of a sample given $\bar{X}$. - Find the minimum sample size required to have sufficient confidence that the sample mean lies in some range. - State assumptions used to carry out a hypothesis test. - Find a confidence interval and understand what it means. - Carry out a hypothesis test on the difference of means.	- See my S2 notes with regards to how to find the sampling distribution for statistics such as the mode, range and median. When the sampling units are discrete values it's possible to list all the possibilities, and produce a sampling distribution directly. - Don't get confused between the population distribution and the sampling distribution of the sample mean. The population distribution could be any distribution (e.g. normal, Binomial, or just any arbitrarily defined one) with mean μ and variance σ^2. It might seem odd thinking of the population as a 'distribution' rather than a list, but it allows us to consider possible samples that could be generated. e.g. If your population was just 3 people with ages 14, 14, 20, then the population distribution would be a random variable X where $P(X = 14) = \frac{2}{3}$ and $P(X = 20) = \frac{1}{3}$. Meanwhile the sampling distribution of the sample mean $\bar{X}$ (note that $\bar{x}$ refers to the mean for a specific sample whereas $\bar{X}$ is a random variable allowing the sample mean to vary across different samples) is approximately: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ by the Central Limit Theorem. Note that $\bar{X}$ is normally distributed even if the original distribution wasn't! You do not need to know proofs for $Var(\bar{X}) = \frac{\sigma^2}{n}$ or $E(\bar{X}) = \mu$. - $X_1, \ldots, X_n$ represents each possible choice of item in the sample of size n, and have the same distribution as the population (since we're choosing each item for the sample from the population). They don't each represent a sample but the elements in the sample. A statistic is a function of the items in the sample, e.g. $f(X_1, X_2, X_3) = \frac{X_1 X_2 + X_3}{\sqrt{5}}$ The function must not depend on any population parameters, e.g. μ. - **Central Limit Theorem**: If $X_1, \ldots, X_n$ is a random sample of size n , for large n, (1 mark) drawn from a population of any distribution with mean μ and variance σ^2 (1 mark) then $\bar{X}$ is (approximately) $N\left(\mu, \frac{\sigma^2}{n}\right)$ (1 mark) - The **standard error** is simply the standard deviation of $\bar{X}$, i.e. $\frac{\sigma}{\sqrt{n}}$ - An **estimator** is a statistic used as an 'estimate' of a population parameter. $\hat{\theta}$ is the estimator of the population parameter θ. $\hat{\mu} = \bar{x}$ $\hat{\sigma} = s$ An estimator is 'unbiased' if across samples, the average value of the estimator (i.e. the mean of the sampling distribution for that statistic) is equal to the population parameter. $\bar{x}$ and s^2 are unbiased because $E(\bar{X}) = \mu$ and $E(s^2) = \sigma^2$. It can be seen that the 'regular variance' calculated from the sample will not give the	- Mixing up $Var(\bar{X}) = \frac{\sigma^2}{n}$ and the standard error $\frac{\sigma}{\sqrt{n}}$, e.g. by using some blend of the two such as $\frac{\sigma^2}{\sqrt{n}}$. - Miscalculating z value for difference of means test, e.g. because you don't understand what to do with the $\mu_A - \mu_B$. - Accidentally using the sample mean rather than μ when stating the distribution of $\bar{X}$. - Calculating normal variance when you should have calculated s^2. Note that the STATS mode on your calculator will find s^2.

population variance σ^2 if we consider a sample size of 1: in our sample the variance is 0, but the population variance is probably not, thus the variance of the sample underestimates σ^2.

- $\boldsymbol{s^2 = \frac{1}{n-1}(\Sigma X^2 - n\bar{X}^2)}$
  This simplified form of the unbiased variance formula is NOT in the formula booklet. Practice using it!
- **To find the distribution of $\bar{X}$,** just find the mean and variance of the population distribution then use $N\left(\mu, \frac{\sigma^2}{n}\right)$.
  Be careful: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ is based on the population parameters, not means/variances we've calculated from the sample. Example:
  "*The continuous random variable X is uniformly distributed over the interval [a – 1, a + 5] where a is a constant. Fifty observations of X are taken, giving a sample mean of 17.2. Use the Central Limit Theorem to give an approximate distribution of $\bar{X}$.*"
  You may be tempted to put $\bar{X} \sim N\left(17.2, \frac{6^2 \div 12}{20}\right)$. However, you would have just given $N(\bar{x}, \sigma^2)$ not $N(\mu, \sigma^2)$. The correct solution is $N\left(a+2, \frac{6^2 \div 12}{20}\right)$.
- **95% confidence interval**: $\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$
  This makes sense when you consider that the 'top 2.5%' in a normal distribution is $z = 1.96$, i.e. the middle 95% is 1.96 standard deviations within the mean.
  You need to know what this confidence interval means: i.e. that there is a 95% chance that the population mean μ lies in this interval.
- For doing hypothesis tests with regards to your sampling distribution $\bar{X}$:
  $$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$
  You shouldn't need to 'memorise' this: it's just using the S1 definition of z (i.e. $z = \frac{x-\mu}{\sigma}$) but used on your $\bar{X}$ distribution $N\left(\mu, \frac{\sigma^2}{n}\right)$
- **Hypothesis test on difference of means:**
  Remember that under the null hypothesis: $\mu_x - \mu_y = 0$ (because there was no change in mean), and if we're trying to find when the mean of A is 1 greater than the mean of B: $\mu_A - \mu_B = 1$
  Example: "A farmer monitored the amount of lead in soil in a field next to a factory. He took 100 samples of soil, randomly selected from different parts of the field, and found the mean weight of lead to be 67 mg/kg with standard deviation 25 mg/kg. After the factory closed, the farmer took 150 samples of soil, randomly selected from different parts of the field, and found the mean weight of lead to be 60 mg/kg with standard deviation 10 mg/kg. Test at the 5% level of significance whether or not the mean weight of lead in the soil decreased after the factory closed. State your hypotheses clearly."

a. State null and alternate hypothesis: (1 mark)
$H_0: \mu_A = \mu_B$ (i.e. mean of soil hasn't changed)
$H_1: \mu_A < \mu_B$ (i.e. mean of soil has decreased)

b. Determine z value (3 marks) – Formula is in formula booklet but you need to know how to deal with $\mu_x - \mu_y$ as per discussion above (which in this example is 0).

$$z = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \frac{67 - 60}{\sqrt{\frac{25^2}{100} + \frac{10^2}{150}}} = 2.6616$$

c. Find critical value associated with 5% using second z-table (1 mark)
If your test is two tailed (i.e. $H_1: \mu_A \neq \mu_B$) then you need to half this.

$$1.6449$$

d. As always, 2 mark conclusion.
$2.6616 > 1.6449$ so reject H_0 (1 mark)
There is evidence that the amount of lead in the soil has increased. (1 mark)

- **"Explain the relevance of the Central Limit Theorem**" (in context)
"It allows us to assume our mean(s) are normally distributed" (we have multiple means when comparing means)
- **"State what assumptions you made" (when doing a hypothesis test)**
"That $s^2 = \sigma^2$" or "Sample variance = population variance".
Since σ^2 is never generally available when doing hypothesis tests (i.e. we're just using data from samples), but our calculations for z require σ^2, we use s^2 instead.
- You need to be able to make calculations based on the sample mean, including:

a. "The mean IQ in Buttersworth Academy is 102 with standard deviation 5. A sample of 20 students is taken. Calculate the probability that their mean IQ is above 105."

$$z = \frac{105 - 102}{\frac{5}{\sqrt{20}}} = 2.68$$
$$P(\bar{X} > 105) = P(Z > 2.68) = 0.00368 \ldots$$

b. "A random sample of size n is to be taken from a population that is normally distributed with mean 40 and standard deviation 3. Find the **minimum sample size** such that the probability of the sample mean being greater than 42 is less than 5%."

$X \sim N(40, 3^2) \therefore \;\; \bar{X} \sim N\left(40, \frac{9}{n}\right)$

$$P(\bar{X} > 42) = P\left(Z > \frac{42 - 40}{\frac{3}{\sqrt{n}}}\right)$$
$$\frac{2}{\frac{3}{\sqrt{n}}} \geq 1.6449 \qquad n \geq 6.087 \quad \therefore \quad n = 7$$

<table>
<tr>
<td>4 – Goodness of Fits and Contingency Tables</td>
<td>

- Carry out a hypothesis test that determines whether a particular model/distribution fits some observed data. The distribution may be Binomial, Poisson, Discrete Uniform, or grouped data tested against continuous distributions: normal or continuous uniform.
- When testing a Binomial model where p is not given, calculate p and be able to adjust ν appropriately.
- The same as above for Poisson Distributions when λ is not given.
- Be able to carry out hypothesis tests on whether two variables have an association (i.e. contingency tables).

</td>
<td>

- **Fitting a Binomial Distribution**

"A total of 100 random samples of 6 items are selected from a production line in a factory and the number of defective items in each sample is recorded. The results are summarised in the table below."

<table>
<tr><td><i>Number of defective items</i></td><td><i>0</i></td><td><i>1</i></td><td><i>2</i></td><td><i>3</i></td><td><i>4</i></td><td><i>5</i></td><td><i>6</i></td></tr>
<tr><td><i>Number of samples</i></td><td><i>6</i></td><td><i>16</i></td><td><i>20</i></td><td><i>23</i></td><td><i>17</i></td><td><i>10</i></td><td><i>8</i></td></tr>
</table>

"A factory manager suggests that the data can be modelled by a binomial distribution with n = 6. He uses the mean from the sample above and calculates expected frequencies as shown in the table below"

<table>
<tr><td><i>Number of defective items</i></td><td><i>0</i></td><td><i>1</i></td><td><i>2</i></td><td><i>3</i></td><td><i>4</i></td><td><i>5</i></td><td><i>6</i></td></tr>
<tr><td><i>Expected frequency</i></td><td><i>1.87</i></td><td><i>10.54</i></td><td><i>24.82</i></td><td><i>a</i></td><td><i>22.01</i></td><td><i>8.29</i></td><td><i>b</i></td></tr>
</table>

"Calculate the value of a and the value of b, giving your answers to 2 decimal places. Hence test, at the 5% level, whether or not the binomial distribution is a suitable model for the number of defective items in samples of 6 items. State your hypotheses clearly."

Note first that exam questions tend to calculate most of the expected frequencies for you, to avoid too much calculation.

a. State null and alternative hypotheses:
H_0: Binomial distribution is a good fit.
H_1: Binomial distribution is not a good fit.

b. We aren't given p, so we need to calculate it.
First calculate mean number of defective items in usual S1 way: 2.91
Divide mean by number of trials n to get p: $p = \frac{2.91}{6}$

c. In general, expected frequency $=$ probability of outcome $\times$ total sample size
Use distribution (in this case Binomial):

$$a = 100 \times P(X = 3) = 100 \times \left(\binom{6}{3} \times 0.485^3 \times 0.515^3\right) = 31.17$$

$$b = 100 \times 0.485^6 = 1.30$$

d. Make a new table, conflating any expected counts where the count is less than 5. This usually happens at the tails (i.e. left and right end of table), so there won't be any ambiguity over what to merge. You will need the observed frequencies again too (adding again where you've conflated any columns).

<table>
<tr><td>Number of defective items</td><td>0 or 1</td><td>2</td><td>3</td><td>4</td><td>5 or 6</td></tr>
<tr><td><i>O</i></td><td>22</td><td>20</td><td>23</td><td>17</td><td>18</td></tr>
<tr><td><i>E</i></td><td>12.41</td><td>24.82</td><td>31.17</td><td>22.01</td><td>9.59</td></tr>
</table>

e. Calculate the goodness of fit $X^2 = \left(\sum \frac{O^2}{E}\right) - N$ or $X^2 = \sum \frac{(O-E)^2}{E}$

$$X^2 = \frac{(22 - 12.41)^2}{12.41} + \cdots + \frac{(18 - 9.59)^2}{9.59} = 18.998 \ldots$$

f. Determine the degrees of freedom ν. Start with number of cells in your conflated table. (i.e. 5). Subtract 1 (because the count in the last column could have been determined using 100 subtracting the others). Subtract 1 because we calculated p ourselves.

</td>
<td>

- Not adjusting ν when you calculate p or λ yourself.
- For Binomial distribution, don't confuse n (number of trials) and N (sample size). For p divide by n not N.
- Not observing whether there's 'gaps' or not for grouped data.
- Not including the whole tail at each end of the table when finding probabilities for a normal distribution goodness of fit.
- Getting your null and alternative hypothesis the wrong way round: remember for goodness of fit the null hypothesis is that the Poisson/uniform/normal/Binomial distribution IS a good model.
- Accidentally counting the 'total' column/row in the number of columns/rows, when determining ν.
- Confusing notation for X^2 and χ^2. The former is a statistic which gives a measure of fit. χ^2 is a distribution which adds a number of normal distributions and is a good approximation of X^2 when the expected counts are large.

</td>
</tr>
</table>

$$\nu = 5 - 2 = 3$$

g. Use the χ^2 table with your calculated X^2. It is always one-tailed.

$$\chi_3^2(0.05) = 7.815$$

h. Make your conclusion, ensuring you put it in context of problem. (2 marks)
"18.998 > 7.815 so reject null hypothesis.
Binomial is not a good model for number of defective items."

- **Fitting a Poisson Distribution:**
  Exactly the same method. If λ needs to be estimated, just find the mean of the frequency table, and adjust ν by subtracting 1.
- **Fitting a Uniform Distribution:**
  Again, the same method. No parameters need to be estimated.
- **Fitting continuous distributions:**
  If you have gaps, ensure you 'fill them in' first by padding out boundaries with 0.5 each side.
- **Fitting the normal distribution**
  For the normal distribution, since your probabilities need to add up to 1, the two ends of your table must be extended to infinity. When calculating $P(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma})$ where $[a, b]$ is each class interval, it's helpful to write out the two probabilities you're subtracting in $P\left(Z < \frac{b-\mu}{\sigma}\right) - P\left(Z < \frac{a-\mu}{\sigma}\right)$ so that you can reuse them.
  The last probability can be calculated using 1 minus the others.

  Example: Suppose we want to carry out a hypothesis test on whether a normal distribution is suitable for the waiting time of customers given a sample of 50:

Waiting time in minutes (x)	Frequency
0–3	8
3–5	12
5–6	13
6–8	9
8–12	8

We need to calculate $\bar{x} = 5.49$ and $s_x^2 = 6.88$ (which reduces ν by 2)
Then a good way to structure table is: (Note, I've used a calculator to calculate z-probabilities, which gives slightly different values to your tables – the mark scheme allows both)

Waiting times	$z = \frac{b-\mu}{s}$	$P(a < X < b)$	Estimated Freq
< 3	$\frac{3-5.49}{\sqrt{6.88}} = 1.144$	$P(Z < -0.9493) = 0.1712$	$50 \times 0.1712 = 8.56$
$3 - 5$	$\frac{5-5.49}{\sqrt{6.88}} = -0.187$	$0.4295 - 0.1712 = 0.2583$	$50 \times 0.2583 = 12.92$
…			
$6 - 8$	$\frac{8-5.49}{\sqrt{6.88}} = 0.9569$	$0.8307 - 0.5771 = 0.2536$	$50 \times 0.2536 = 12.68$
> 8		$1 - 0.8307 = 0.1693$	$50 \times 0.1693 = 8.465$

- If using the $X^2 = \left(\sum \frac{O^2}{E}\right) - N$ formula, forgetting to subtract N at the end.

Then just conflate the table as before and calculate goodness of fit.

- **Fitting a continuous uniform distribution**
  Example: A total of 228 items are collected from an archaeological site. The distance from the centre of the site is recorded for each item. The results are summarised in the table below.

Distance from the centre of the site (m)	0–1	1–2	2–4	4–6	6–9	9–12
Number of items	22	15	44	37	52	58

  The probabilities this time are simply what fraction of the total range each interval occupies, e.g. Expected frequency for 0-1 is $228 \times \frac{1}{12}$ and for 9-12 is $228 \times \frac{3}{12}$.
  This was a 12 mark question: 3 marks for calculating probabilities and expected frequencies, 2 for calculating $\frac{O^2}{E}$ or $\frac{(O-E)^2}{E}$ for each interval, 1 for stating hypotheses, 2 for calculating X^2, 1 for ν (in this case $6 - 1 = 5$), 1 for χ^2 table lookup, 2 for conclusion.
- For **contingency tables,** remember the null hypothesis is that the variables have NO association. Structure of proof:

		Cholesterol Level	
		High	Low
Saturated fat intake	High	12	8
	Low	26	54

  a. State null and alternative hypotheses. (1 mark)
  H_0: *Cholesterol level independent of intake of saturated fats (no association)*
  H_1: *Cholesterol level dependent on intake of saturated fats (association)*

  Calculated expected values under null hypothesis, by first adding a total column/row if not already given. (2 marks) Each value calculated using:

$$Expected\ value = \frac{row\ total \times column\ total}{grand\ total}$$

Cholesterol Level	High	Low	TOTAL
High	7.6	12.4	20
Low	30.4	49.6	80
TOTAL	38	62	100

  c. Use a separate table to calculate X^2 by comparing expected and observed frequencies. The mark scheme allows either formula for X^2. (3 marks)

O	E	$\frac{(O-E)^2}{E}$	$\frac{O^2}{E}$
12	7.6	2.547… or $\frac{242}{95}$	18.947… or $\frac{360}{19}$
8	12.4	1.56129… or $\frac{242}{155}$	5.161… or $\frac{160}{31}$
26	30.4	0.6368… or $\frac{121}{190}$	22.236… or $\frac{845}{38}$
54	49.6	0.3903… or $\frac{121}{310}$	58.790… or $\frac{3645}{62}$

		$X^2 = 5.14$ Determine $\nu = (r-1)(c-1)$. (1 mark) The reason for this is that the values in the row and column can be determined if the counts are fixed, and hence cannot ‘vary’. Be careful, the totals columns do NOT count towards the totals. $\nu = 2 \times 2 = 4$ d. Use the χ^2 table with your threshold. It is always one-tailed. (1 mark) $\chi_1^2(0.05) = 3.84$ e. As usual, a two-pointer conclusion put in context (2 marks). *5.14 > 3.841 so sufficient evidence to reject [Condone “accept H_1 ”]* *Association between cholesterol level and saturated fat intake.*	
5 – Hypothesis tests for PMCC and Spearman’s Rank Correlation Coefficient	• Carry out a hypothesis test to determine if two sets of data are correlated by rank (i.e. using Spearman’s Rank Correlation Coefficient) • Appreciate why we might use Spearman’s Rank rather than PMCC. • Interpret a value of r_s. • Know the assumption we’re making when carrying out a hypothesis test on correlation.	• The main advantage of Spearman’s Rank Correlation over the Product Moment Correlation Coefficient is that it doesn’t assume the data is linear: as long as the data has some overall increasing or decreasing trend (e.g. a quadratic shape), Spearman’s Rank will identify this correlation, whereas the PMCC would be closer to 0 if the data points don’t form a straight line. When carrying out a hypothesis test, there’s another advantage which we’ll see below. • Spearman’s Rank is equivalent to PMCC if the data is first ranked. Thus if you first rank your data (it doesn’t matter if 1 is the largest value or the smallest, as long as you are consistent with your two variables), you could always calculate r directly using the STATS mode on your Casio calculator (see my S1 revision notes on how to do this). • If the data is ranked, a quicker way to calculate the PMCC is using the following formula (given in your formula booket): $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$ where d is the difference between the ranks for each data point (since these will be squared, it doesn’t matter if the different is positive or negative). The proof is not required. The formula does not work for tied ranks (use PMCC on ranked data instead). Note use of r for PMCC and r_s for Spearman’s Rank Correlation Coefficient. • $r_s = 1$ means the rankings are in perfect agreement. $r_s = 0$ means there is no correlation in the ranks. $r_s = -1$ means the rankings are exactly reversed. • (In the context of hypothesis testing) “Give a reason to support the use of the rank correlation coefficient rather than the product moment correlation coefficient with these data.” The critical values in the PMCC table assume that the data is (jointly) normally distributed. The Spearman’s table doesn’t make this assumption (since ranks clearly can’t be normally distributed). Write “The variables cannot be assumed to be normally distributed”. Another potential reason to justify use of the rank correlation coefficient is if the data is already ranked. • ρ (pronounced “rho”) is a population parameter which is the actual correlation between variables X and Y. For hypothesis tests in S3, the null hypothesis is always $\rho = 0$, that there is no actual correlation, and that any observed correlation was by chance. r and r_s meanwhile is the observed correlation based on the sample (i.e. the test statistic).	• When using the r_s formula, always check your value by calculated the PMCC on your ranked data using the STATS mode of your calculator! • Incorrectly using the PMCC table instead of the Spearman’s Rank table for hypothesis testing. • Not distinguishing between one and two-tailed correlation hypothesis tests. • As before, mixing up your null and alternate hypotheses. • Accidentally doing: $r_s = \frac{1 - 6\sum d^2}{n(n^2-1)}$

- The hypothesis test can be either be one-tailed, e.g. a claim there is positive correlation, or two-tailed, i.e. a claim there is correlation.
- Example test: “A county councillor is investigating the level of hardship, h, of a town and the number of calls per 100 people to the emergency services, c. He collects data for 7 randomly selected towns in the county. The results are shown in the table below. After collecting the data, the councillor thinks there is no correlation between hardship and the number of calls to the emergency services. Test, at the 5% level of significance, the councillor’s claim. State your hypotheses clearly.”

Town	*A*	*B*	*C*	*D*	*E*	*F*	*G*
h	14	20	16	18	37	19	24
c	52	45	43	42	61	82	55

- Suppose we have already calculated $r_s = 0.5$. Then:
  a. State hypotheses (1 mark):
     $H_0: \rho = 0$ *(variables are not correlated)*
     $H_1: \rho \neq 0$ *(variables are correlated)*
  b. Using Spearman’s Rank table with significance level (in this case halved as two-tailed). Note that if two-tailed, you have two critical values. (1 mark)
     Critical values of r_s: ± 0.7857
  c. Usual two-mark conclusion:
     “0.5 < 0.7857 therefore insufficient evidence to reject H_0” (1 mark)
     “Councillor’s claim is supported” (1 mark)

Summary of different hypothesis tests: (see respective chapter notes)

Chapter	Test description/name	Null/alternative hypotheses	One/two-tailed?
3	Difference of means	$H_0: \mu_A = \mu_B$ or $\mu_A - \mu_B = k$ $H_1: \mu_A > \mu_B$ or $\mu_A \neq \mu_B$ or $\mu_A < \mu_B$ or $\mu_A - \mu_B > k$	One or two
4	Test of goodness of fit of distribution	H_0: Binomial/Poisson/… is a good fit H_1: Binomial/… is not a good fit	One only
4	Contingency table	H_0: No association between cholesterol level and fitness. H_1: Association between cholesterol level and fitness.	One only
5	PMCC is 0 or Spearman's rank correlation coefficient is 0	$H_0: \rho = 0$ (i.e. no correlation between variables) $H_1: \rho \neq 0$ or $\rho < 0$ or $\rho > 0$	One or two

Note that a contingency table test and $\rho = 0$ test are very different: in the former, there is perfect association only if the values in each set of data is exactly the same (i.e. we're testing how different the values are from each other), whereas for the latter, there is perfect correlation (in the PMCC test) if the data perfectly fits a straight line, but the X and Y values don't need to be the same. A contingency table also allows more than 2 categories (e.g. English score, Maths score, Science score) whereas this doesn't make sense for correlation coefficients.