S3 Cheat Sheets

1 Combinations of random variables

1.1 Be able to combine two normally distributed variables

$$E(X \pm Y) = E(X) \pm E(Y)$$

$$Var(X \pm Y) = Var(X) + Var(Y)$$

Remember

$$Var(2X) = 2^2 Var(X)$$

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) = 2Var(X)$$

None of this works for standard deviation, remember to square beforehand

1.2 Be able to calculate $\mathbf{P(|X - Y| < k)}$ for some constant k

$$P(|X - Y| < 3) = P(-3 < X - Y < 3) = P(X - Y < 3) - P(X - Y < -3)$$

2 Sampling

2.1 Definitions

Hypothesis - A statement concerning a population parameter

Critical Region - The range of values that would lead to the rejection of H_0, and so the acceptance of H_1

Statistic - A random variable, a function of known observations from a population, no unknown parameters

Sampling unit - An individual unit of the population

Sampling frame - A list of all sampling units

Sampling distribution - All possible samples are chosen from the population

Census - When every member of the population is investigated

Hypothesis test - A mathematical procedure to examine a value of a population parameter proposed by the null hypothesis compared with an alternative hypothesis

Population - A collection of all items

Finite Population - One in which each individual member can be given a number

Infinite population - One where it is impossible to number each individual member

Significance level - The probability of incorrectly rejecting H_0

Sample - A selection of items from a population

2.2 Sampling

	What it is	When to use	Advantages	Disadvantages
Census	A collection of data from an entire population	Gives a completely accurate result	• Small population • Easy to collect data • Large variation of opinion	• Time consuming+Expensive • Can not be used when testing involves destruction • Large volume of data to process
Random Sampling	Each thing has an equal chance of being selected	• Large population • Have a sampling frame	• Numbers truly random and free from bias • Easy to use • Each number has a known equal chance of being selected	• Needs a sampling frame
Systematic sampling	Required elements are chosen at regular intervals in an ordered list	• Time constraint	• Simple to use • Suitable for large samples	• Only random if ordered list is truly random • Can introduce bias
Stratified sampling	Population is divided into groups and a simple random sample is carried out in each group	• More accurate when strata are present • Reflects population structure	• It can give more accurate estimates than simple random sampling where clear strata are present • Reflects the population structure	• Within the strata, the problems are than same as for any simple random sample • If the strata are not clearly defined they may overlap
Quota sampling	The population is divided into groups by gender etc. A quota of people in each group is set to try and reflect the group's proportion in the whole population	• There is no sampling frame	• Enables fieldwork to be done quickly because a small sample size is taken. • Costs kept to a minimum • Administering test is easy	• Not possible to estimate the sampling errors • Interviewers may not be able to judge characteristics easily • Non responses are not recorded • Can introduce interview bias

2.3 How to carry out samples

2.3.1 Simple random sampling

- Allocate numbers to each member of the population
- Use random number tables to select different numbers until enough have been selected
- Members corresponding to the numbers become the sample

2.3.2 Stratified sampling

- Number the members in each strata
- Use random numbers to select the members from each group
- Number from each group in sample to be proportional to the number in the sample - calculate

2.3.3 Systematic sampling

- Randomly select first number between 1 and $Pop/sample$
- Select every person at an interval of $Pop/sample$ after that

2.3.4 Quota sampling

- Non random sampling
- Groups of the population

3 Estimation, confidence intervals and tests

3.1 Know the definition of the central limit theorem

- If $X_1, ..., X_n$ is a random sample of size n, for large n
- Drawn from the population of any distribution with mean μ and variance σ^2
- The $\overline{X}$ is approximately $N(\mu, \frac{\sigma^2}{n})$

Most if this is under the sampling distributions header on the data sheet, just remember to include when n is large

3.2 Know the definition of a statistic

A random variable, a function of known observations from a population, no unknown parameters

3.3 Understand what $\overline{X}$ is and find the distribution of $\overline{X}$

$\overline{x}$ - Mean for a specific sample
$\overline{X}$ - A random variable allowing the sample mean to vary across different samples

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Make sure when writing the distribution to use μ, the population mean, rather than the sample mean.

3.4 Find value of n such that a value is in the confidence interval

Use the formula:

$$\overline{X} \pm \times \frac{\sigma}{\sqrt{n}}$$

By setting this equal to a value, the value of n can be found such that the value will be in the confidence interval

3.5 Find the minimum sample size required to have sufficient confidence that the sample mean lies in some range

Example
$X \sim N(40, 3^2)$. *Find the minimum sample size such that the probability of the sample mean being greater than 42 is less than 5%.*

$$\overline{X} \sim N\left(40, \frac{9}{n}\right)$$

$$P(\overline{X} > 42) = P\left(Z > \frac{42-40}{\frac{3}{\sqrt{n}}}\right)$$

Find the C.V.

$$\frac{2}{\frac{3}{\sqrt{n}}} \geqslant 1.6449 \quad \Rightarrow n \geqslant 6.087 \quad \therefore n = 7$$

3.6 Understand estimator notation

$\hat{\mu} = \overline{x}$, anything with a hat on is an estimator

3.7 Find the value of S^2

$$S^2 = \frac{\sum(X_i - \overline{X})^2}{n-1}$$

Replace the top of the fraction with S_{xx}

$$S^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

3.8 Hypothesis test for the difference between means

When performing a hypothesis test for the difference between means, use the formula on the formula book:

$$\frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

H_0 is always that the means are equal, or that they are equal with addition or subtraction to one
H_1 is always that one is greater, with the same addition or subtraction if stated in H_0

3.9 State assumptions used to carry out a hypothesis test

In general:

$$s^2 = \sigma^2$$

Sample variance=Population variance

3.10 Explain the relevance of the CLT

It allows us to assume our means are normally distributed

4 Goodness of fit and Contingency tables

4.1 Goodness of fit method

Method for testing goodness of fit:

1. Determine which distribution would conceptually be most appropriate
2. Set significance level
3. Estimate parameters (if necessary) from observed data
4. Form hypotheses H_0 and H_1
5. Calculate expected frequencies
6. Combine expected frequencies so that none are < 5
7. Find degrees of freedom
8. Calculate critical value of χ^2 from the table
9. Calculate $\sum \frac{(O_i - E_i)^2}{E_i}$
10. See if the value is significant and draw conclusion

X^2 is distributed with a chi squared distribution χ^2_ν
Where ν = degrees of freedom

$$\text{The number of degrees of freedom} = \text{Number of classes (after combining)} - 1$$

$$\sum \frac{(O_i - E_i)^2}{E_i} \text{can be rewritten as} \quad \left(\sum \frac{O_i^2}{E_i} \right) - N$$

H_0 is always that the distribution is a good model
H_1 is always that the distribution is not a good model

4.1.1 Normal distribution

When testing for a normal distribution the two ends of the table must extend to infinity to ensure that the probabilities add to 1

4.1.2 Uniform distributions

It is necessary to state the type of uniform distribution:

- **Continuous** - Can take any of the values over the range
- **Discrete** - Can only take specific values over the range (for example odd numbers)

4.1.3 Degrees of freedom

The degrees of freedom starts as the number of columns, then 1 is subtracted from it for various reasons:

- Probabilities must add to 1 so one of the probabilities could be approximated
- Approximation of p in binomial, or λ in poisson when not given

4.2 Contingency tables

- We use this test to see if two factors are independent of each other
- We describe them by: Number of rows × Number of columns
- H_0 is that they are independent
- H_1 is that they are not independent
- $$\text{Expected values} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$
- $$\nu = (\text{Number of rows-1})(\text{Number of columns-1})$$
- Remember that the same rule that expected frequencies cannot be less than 5 still applies

Contingency tables and PMCC tests both look for linear relationships, however a contingency table looks to see if the data lies on the line $y = x$, whereas the PMCC looks to see if the data lies on the line $y = kx$

5 Hypothesis tests for PMCC and Spearman's Rank Correlation Coefficient

Spearman's rank - The tendency for y to increase as x increases
PMCC - How close x and y are to a linear relationship

Spearman's rank is equivalent to PMCC if the data is ranked, this causes the data to adopt a linear relationship if y increases as x increases.

The Spearman's rank formula does not work if there are tied ranks

The critical values in the PMCC table assume that the data is jointly normally distributed, this is the assumption if asked for a hypothesis test. The spearman's rank distribution doesn't assume the data to be jointly normally distributed.

ρ is the population parameter which is the **actual** correlation between the variables.
r and r_s is the observed correlation based on the sample

Hypothesis tests can be either one or two tailed, depending if you are just looking for a correlation or specifically a positive/negative correlation.

6 Conditions for distributions

6.1 Normal Distribution

- Probabilities symmetrical about the mean

6.2 Binomial distribution

- Fixed number of trials
- Two outcomes
- Independent trials
- Constant probability of success

6.3 Poisson distribution

- Events occur at random
- Independent events
- Constant rate of occurrence
- No simultaneous events

6.4 Uniform distribution

- All outcomes have the same probability

Hypothesis tests

What is being tested	**H_0**	**H_1**
Difference between means	$\mu_a = \mu_b$	$H_1 : \mu_a > \mu_b$ or $\mu_a < \mu_b$ or $\mu_a \neq \mu_b$
Goodness of fit	A _____ Distribution is a suitable model	A _____ Distribution is not a suitable model
Contingency Tables	No association between the two data sets	Association between the two data sets
Spearman's rank	$\rho_s = 0$	$\rho_s > 0, \rho_s < 0, \rho_s \neq 0$
PMCC	$\rho = 0$	$\rho > 0, \rho < 0, \rho \neq 0$

Tables

6.5 Goodness of fit

Number		
Observed		
Expected		
$\frac{(O-E)^2}{E}$		

6.6 Contingency tables

	C			D		
	O	E	$\frac{(O-E)^2}{E}$	O	E	$\frac{(O-E)^2}{E}$
A						
B						

6.7 Difference between means

	A	B
$\overline{x}$		
s^2		
n		

6.8 Spearman's rank

	A	B	C
Rank 1			
Rank 2			
d^2			