Sponsored by

CASIO

@MEIConference #MEIConf2018

Topics in FM Statistics

Paul Chillingworth

What statistics is new in Further Maths?

- Can assume all students have the same statistics knowledge from A level… though when each topic is covered in Mathematics could impact on when they are taught in Further Mathematics

What statistics is new in Further Maths?

- Can assume all students have the same statistics knowledge from A level… though when each topic is covered in Mathematics could impact on when they are taught in Further Mathematics
- No prescribed common content, but there are some topics which are covered in all specs

What statistics is new in Further Maths?

- Look at ‘Comparing Further Mathematics Statistics Content’
- You may want to discuss these questions:
  - Which topics are common?
  - Which topics are AS?
  - Do any of these surprise you?

Randomness

- Can you define randomness?

Randomness

- Can you define randomness?

- http://www.s253053503.websitehome.co.uk/msv/msv-8.html

What do you know about the Poisson distribution?

What do you know about the Poisson distribution?

- events occur randomly over time or space
- they occur independently of each other
- constant average number per interval of time or space

Poisson probability for parameter λ is $P(X = r) = \frac{e^{-\lambda}\lambda^{r}}{r!}$

Modelling with a Poisson

- Can you think of any practical situations that can be modelled by a Poisson distribution?

Modelling with a Poisson

Many experimental situations occur in which we observe the counts of events within a set unit of time, area, volume, length etc. For example,

- The number of cases of a disease in different towns
- The number of mutations in set sized regions of a chromosome
- The number of dolphin pod sightings along a flight path through a region
- The number of particles emitted by a radioactive source in a given time
- The number of births per hour during a given day

Conditions

Let X represent the number of occurrences per interval of time or space. Then X can be modelled by a Poisson distribution provided:-

- events occur at random and independently of each other in a given interval of time or space;
- the average number of events in the given interval (λ) is fixed and finite. This rate is proportional to the interval;
- events occur singly.

Properties of the Poisson

- A discrete random variable with probability function $P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$ $for\ x = 0, 1, 2,..$
- If the interval (in time or space) is scaled up or down, the mean is correspondingly scaled.
- Additive: If X has a Poisson distribution with parameter λ and Y has a Poisson distribution with parameter μ then, if X and Y are independent, the distribution of $X + Y$, the sum of observations from the two random variables has a Poisson distribution with mean λ + μ

Mean and variance

Proof that $E(X) = \lambda$

Mean and variance

Proof that $E(X) = \lambda$

$$E(X) = \sum_{i=0}^{\infty} iP(X = i) = \sum_{i=0}^{\infty} \frac{ie^{-\lambda}\lambda^{i}}{i!}$$

Mean and variance

Proof that $E(X) = \lambda$

$$E(X) = \sum_{i=0}^{\infty} iP(X = i) = \sum_{i=0}^{\infty} \frac{ie^{-\lambda}\lambda^i}{i!}$$

$$= e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!}$$

Mean and variance

Proof that $E(X) = \lambda$

$$E(X) = \sum_{i=0}^{\infty} iP(X = i) = \sum_{i=0}^{\infty} \frac{ie^{-\lambda}\lambda^i}{i!}$$

$$= e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!}$$

$$= e^{-\lambda}\lambda \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = e^{-\lambda}\lambda \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$$

Mean and variance

Proof that $E(X) = \lambda$

$$E(X) = \sum_{i=0}^{\infty} iP(X = i) = \sum_{i=0}^{\infty} \frac{ie^{-\lambda}\lambda^i}{i!}$$

$$= e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!}$$

$$= e^{-\lambda}\lambda \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = e^{-\lambda}\lambda \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda}\lambda e^{\lambda} = \lambda$$

Mean and variance

It can also be shown that $\text{Var}(X) = \lambda$, so hence

$$E(X) = \text{Var}(X)$$

This is a fairly unique property and can be used to test informally for a Poisson

Example

- A firm investigated the number of employees suffering injuries whilst at work.
- The results recorded below were obtained for a 52-week period:

Number of employees injured in a week	0	1	2	3	≥ 4
Number of weeks	31	17	3	1	0

Give reasons why you might expect this distribution to approximate to a Poisson distribution.

Example

Number of employees injured in a week	0	1	2	3	≥ 4
Number of weeks	31	17	3	1	0

Give reasons why you might expect this distribution to approximate to a Poisson distribution.

If we can reasonably assume that injuries occur:

- Singly
- Independently
- At a uniform average rate

Then the number of employees injured will follow a Poisson.

Example

Number of employees injured in a week	0	1	2	3	≥ 4
Number of weeks	31	17	3	1	0

Evaluate the mean and variance of the data and explain why this gives further evidence in favour of a Poisson distribution.

Example

Number of employees injured in a week	0	1	2	3	≥ 4
Number of weeks	31	17	3	1	0

Evaluate the mean and variance of the data and explain why this gives further evidence in favour of a Poisson distribution.

The summary statistics are $\sum x = 26$ and $\sum x^2 = 38$

This gives $\bar{x} = 0.5$ and $s^2 = \frac{\sum x^2 - n\bar{x}^2}{n-1} = \frac{38-(52\times 0.5^2)}{51} = 0.490$

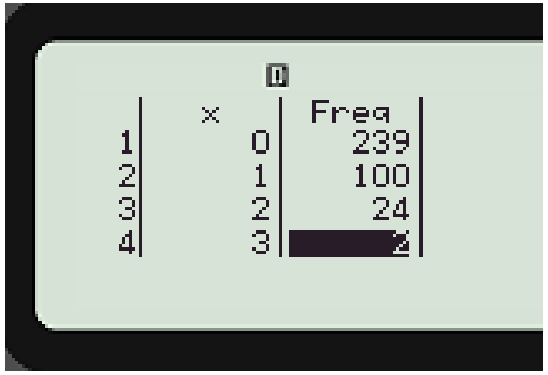
A real application

The application of a Poisson model to the annual distribution of daily mortality at hospitals.

How good a fit is the Poisson?

Daily Deaths	Frequency
0	239
1	100
2	24
3	2
4	0

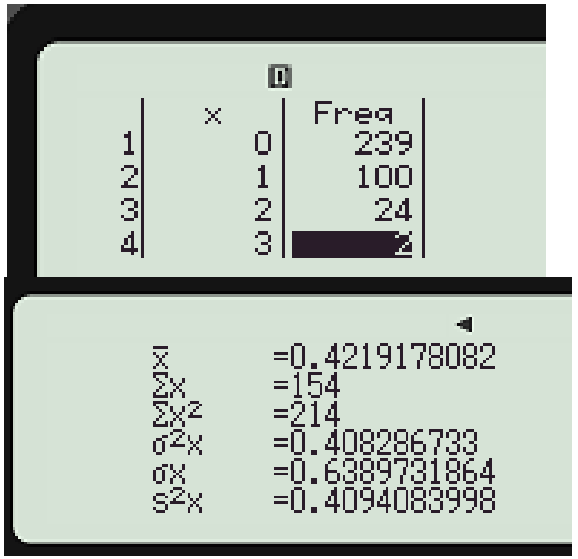
Let's look at the daily death data.



How good a fit is the Poisson?

Daily Deaths	Frequency
0	239
1	100
2	24
3	2
4	0

Let's look at the daily death data.



How good a fit is the Poisson?

Daily Deaths	Frequency
0	239
1	100
2	24
3	2
4	0

We can check what the expected frequencies should be like as we know the mean which is 0.422

How good a fit is the Poisson?

Daily Deaths	Frequency
0	239
1	100
2	24
3	2
4	0

We can check what the expected frequencies should be like as we know the mean is 0.422

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

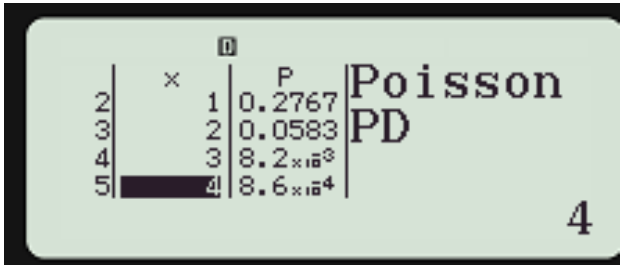
So $P(X = 0) = e^{-0.422}\frac{0.422^0}{0!} = 0.6557$

How good a fit is the Poisson?

We can check what the expected frequencies should be like as we know the mean is 0.422

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

So $P(X = 0) = e^{-0.422}\frac{0.422^0}{0!} = 0.6557$

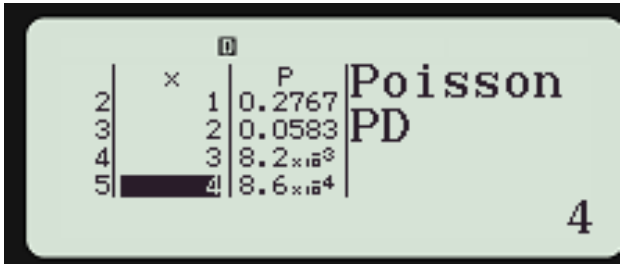


How good a fit is the Poisson?

The expected values = 365 x probability

$365 \times 0.6557 = 239$
$365 \times 0.2767 = 101$
$365 \times 0.0583 = 21$
$365 \times 0.0082 = 3$
$365 \times 0.0011 = 1$

Daily Deaths	Frequency
0	239
1	100
2	24
3	2
4	0



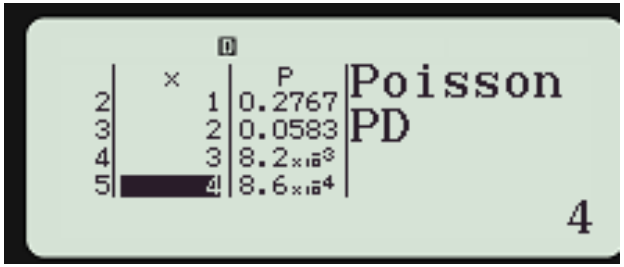
How good a fit is the Poisson?

The expected values = 365 x probability

$365 \times 0.6557 = 239$
$365 \times 0.2767 = 101$
$365 \times 0.0583 = 21$
$365 \times 0.0082 = 3$
$365 \times 0.0011 = 1$

4 or more

Daily Deaths	Frequency
0	239
1	100
2	24
3	2
4	0



How good a fit is the Poisson?

This looks to be a very good fit, but usually we'd expect more variation and this all very ad-hoc, so it is useful to have a more formal measure of how close the fit it is and whether we can read anything into it.

Making some hypotheses

- Null hypothesis H_0: the number of daily deaths at the Hospital over a 9 year period follows a Poisson distribution.

- Alternative hypothesis H_1: the number of daily deaths at the Hospital over a 9 year period does not follow a Poisson distribution.

- Assuming the null hypothesis is true gives us a way of calculating expected frequencies.

Test Statistic

- As with other hypothesis tests, we need a test statistic to test.
- e.g. like we do when testing a binomial proportion (the number of successes) or the mean of a normal distribution ($\bar{x}$ - the sample mean).
- The test statistic needs to measure the closeness between the observed and expected values.

Expected values

Daily Deaths	Observed	Expected	
0	2060	2050	
1	956	968	
2	223	228	
3	43	36	
4 or more	5	5	
	3287	3287	

Expected values

Daily Deaths	Observed	Expected	O-E
0	2060	2050	10
1	956	968	-12
2	223	228	-5
3	43	36	7
4 or more	5	5	0
	3287	3287	0

Expected values

- To get round the problem of some numbers being positive and some negative, we can square all the differences.
- Some of these squared differences could be very large. To cut them down to size, each is divided by the expected frequency for that cell.

Daily Deaths	Observed	Expected	O-E
0	2060	2050	10
1	956	968	-12
2	223	228	-5
3	43	36	7
4 or more	5	5	0
	3287	3287	0

Expected values

Daily Deaths	Observed	Expected	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	2060	2050	10	100	0.0488
1	956	968	-12	144	0.1488
2	223	228	-5	25	0.1096
3	43	36	7	49	1.3611
4 or more	5	5	0	0	0
	3287	3287	0		1.6683

So the test statistic X^2 has a value of 1.6683.

Expected values

Daily Deaths	Observed	Expected	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	2060	2050	10	100	0.0488
1	956	968	-12	144	0.1488
2	223	228	-5	25	0.1096
3	43	36	7	49	1.3611
4 or more	5	5	0	0	0
	3287	3287	0		1.6683

So the test statistic X^2 has a value of 1.6683.
If the test statistic is 'big' then this means that there was a lot of difference between expected and observed frequencies – this leads us to doubt the null hypothesis.

Chi-squared

To decide how ‘big’ is unlikely under the null hypothesis we need to know:

- The distribution of X^2
- The level of significance at which we wish to test.

Chi-squared

The observed value, O, can be thought of as (approximately) a Poisson random variable with mean E, the expected frequency.

The variance of that Poisson is also E, so $\sqrt{E}$ is its standard deviation.

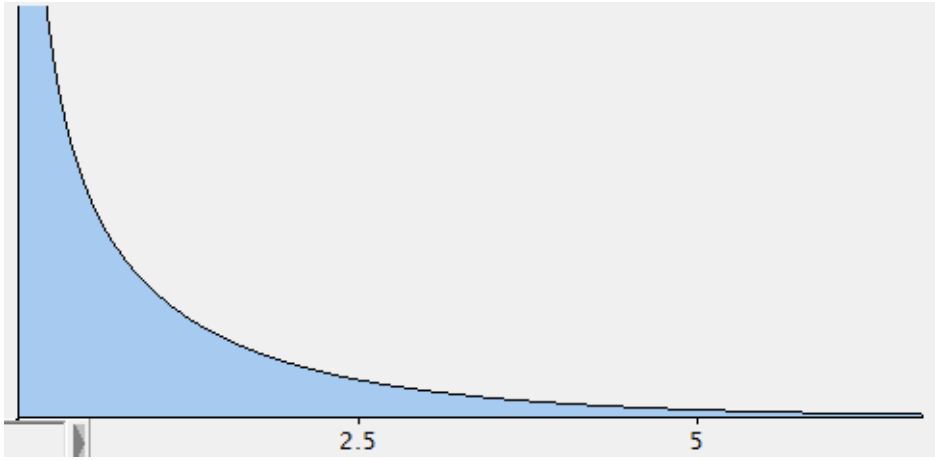
$(O-E)/\sqrt{E}$ will have the standard Normal distribution (approximately, and provided E is not too small).

Hence $\sum \frac{(O-E)^2}{E}$ is (approximately) the sum of the squares of a set of standard Normals.

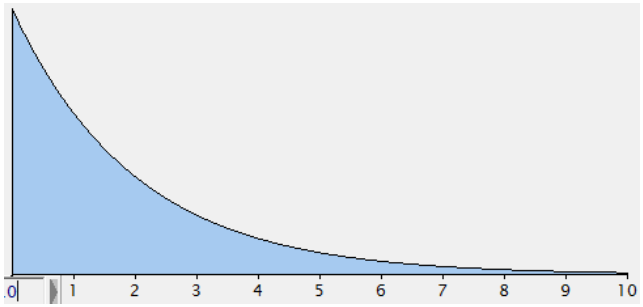
Chi-squared

- The chi-squared random variable with parameter *n* is **defined** as the sum of the squares of *n* independent standard Normals. (So there is a family of chi-squared distributions for *n* = 1, 2, 3 …)

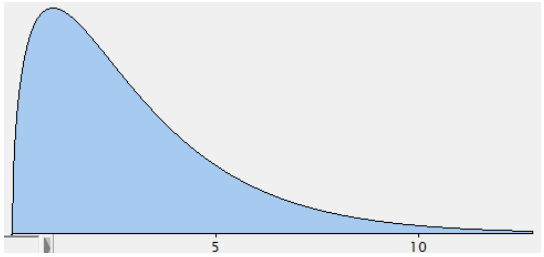
- E.g., χ_1^2 looks like this:



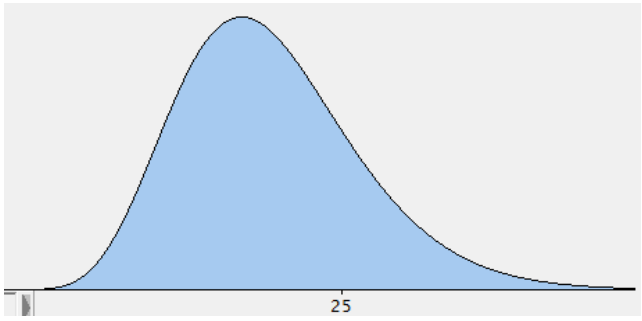
- χ_2^2



- χ_3^2



χ_{20}^2



Chi-squared

- So, we can compare the test statistic $\sum \frac{(O-E)^2}{E}$ with a chi-squared distribution.
- But …
- *Which* chi-squared distribution?
- And …
- What sort of comparison do we need to make?

Chi-squared

- Under the null hypothesis, the test statistic $\sum \frac{(O-E)^2}{E}$ will have a chi-squared distribution with *n* equal to the number of *independent* variables (This is referred to as chi-squared with *n* degrees of freedom.)

Chi-squared

- Under the null hypothesis, the test statistic $\sum \frac{(O-E)^2}{E}$ will have a chi-squared distribution with *n* equal to the number of *independent* variables (This is referred to as chi-squared with *n* degrees of freedom.)
- In this case because the totals must agree and the means must agree, so we have 2 constraints and there are only 5-2=3 independent variables.

Chi-squared

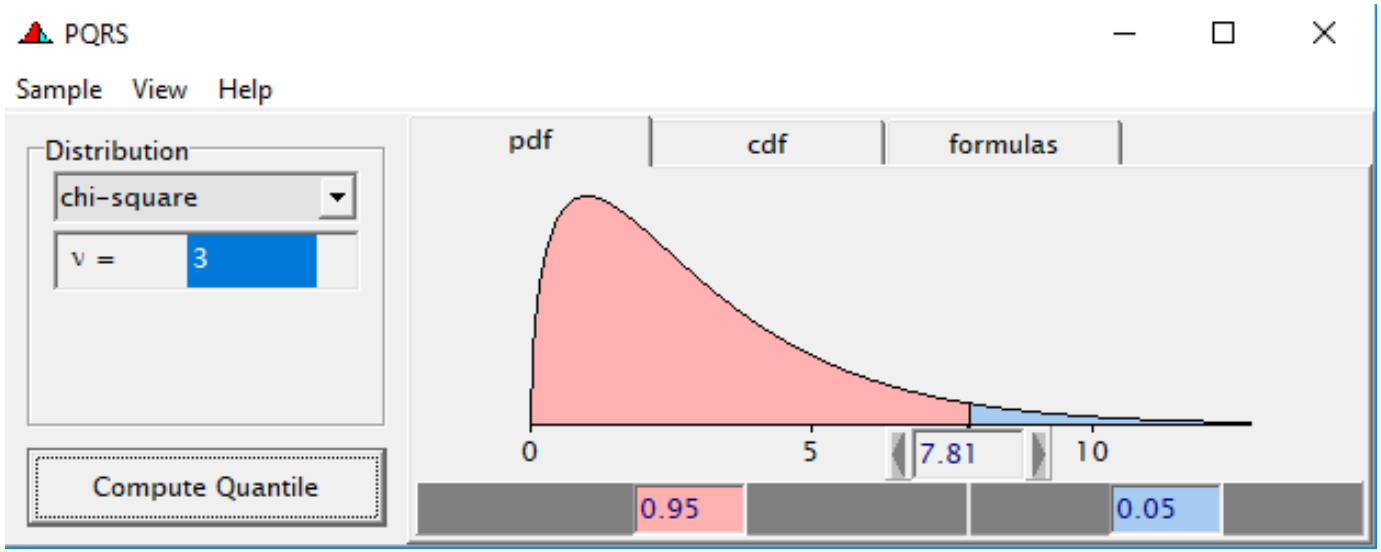
- Under the null hypothesis, the test statistic $\sum \frac{(O-E)^2}{E}$ will have a chi-squared distribution with *n* equal to the number of *independent* variables (This is referred to as chi-squared with *n* degrees of freedom.)
- In this case because the totals must agree and the means must agree, so we have 2 constraints and there are only 5-2=3 independent variables.
- Hence we need to consider a χ_3^2 distribution

Which tail?

- In referring the value of $\sum \frac{(O-E)^2}{E}$ to the appropriate chi-squared distribution, we are looking for the set of values that favour the alternative hypothesis over the null.
- That means we want to compare the value of the test statistic with the extreme right hand tail of the appropriate chi-squared distribution.

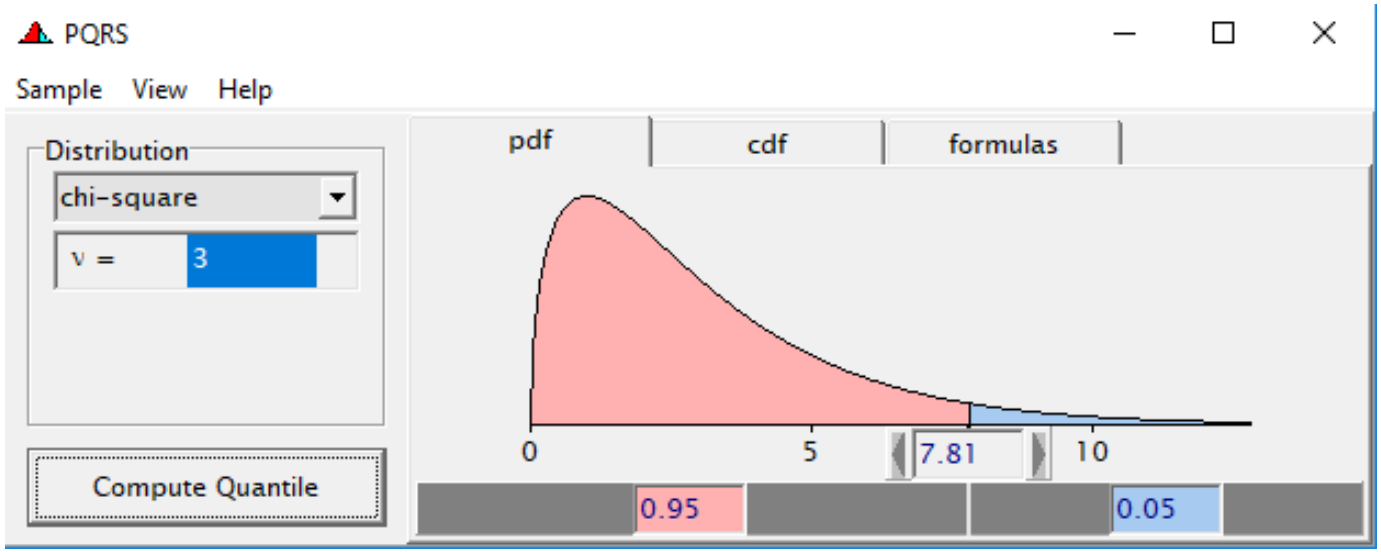
Testing at the 5% level

- So, in our example, we compare the value of $\sum \frac{(O-E)^2}{E}$ with values in the right hand tail of χ_3^2.
- If we were doing a 5% test we would be comparing the test statistic with a value of 7.81



Testing at the 5% level

As 1.6683 < 7.81, we do not reject the null hypothesis, concluding that the Poisson distribution is a good fit.



Link to Gnumeric Files

bit.ly/pressf9

MEI Conference 2018

Sponsored by CASIO

Teaching topics in A level Further Mathematics Statistics

Paul Chillingworth

paul.chillingworth@mei.org.uk

Comparing Further Mathematics Statistics Content

Content	Detail	AQA	Edexcel	MEI	OCR
Discrete random variables	PDFs, expectation and variance, Poisson, Discrete Uniform Distribution	AS	S1 AS	a	AS
	Functions of a discrete random variable	AS	S1 AS	a	A
	Binomial Distribution		S1 AS	a	AS
	Negative Binomial Distribution		S1 A		
	Geometric distribution		S1 A	a	AS
	Use of distributional approximations		S1 AS		
Continuous Random Variables	PDFs, mean and variance, mode, median and quartiles, CDFs, combining random variables, rectangular distributions.	AS & A	S2 AS	b	A
	Exponential Distribution	A			
The Normal Distribution	Normal as a model			b	
	Tests of Normality			b	
	Combining Normal variables		S2 A	b	A
Sampling and the sample mean	Experimental Design, Random Sampling			a	
	Sample estimates of μ and σ, $\bar{X}$ as a random variable	AS	S2 A	b	A
	Central limit theorem		S1 A	b	A
Bivariate Data	Scatter Diagrams, PMCC, Spearman's Rank CC, Hypothesis Tests for PMCC, Hypothesis Tests for SRCC, effect size.		S2 AS	a	AS
	Regression lines; residuals		S2 AS	a	AS
Inference: Hypothesis Tests	Chi-squared Goodness of fit		S1 AS	a	AS
	Chi-squared Contingency	AS	S1 AS	a	AS
	Mean of a Poisson Distribution	AS	S1 AS		
	Parameter of a Geometric Distribution		S1 A		
	Of the sample mean using normal	AS	S2A	b	A
	Errors and Power of Test	AS & A	S1 A		
	Of the Sample mean using t-Distribution	A	S2 A	b	
	Difference in means, Paired t-test		S2 A		
Tests of Variance	Variance of a normal Distribution, F-test on 2 samples		S2 A		
Inference: Confidence Intervals	For mean using normal, For mean using t-Distribution, Paired samples, Difference in means	AS A	S2 A	b	A
	Variance of a normal Distribution		S2 A		
Non –parametric Tests	Wilcoxon Sign Rank (T)			b	A
	Mann-Witney U Test (W), Paired and 2-sample tests				A
	Normal Approximation to W and T				A
Further Probability	Permutations and Combinations				AS
	Simulation of RVs using spreadsheets			b	
	Probability Generating Functions		S1 A		

Please refer to Awarding Organisation documentation for details and specifics