Subject CS1

CMP Upgrade 2022/23

CMP Upgrade

This CMP Upgrade lists the changes to the Syllabus, Core Reading and the ActEd material since last year that might realistically affect your chance of success in the exam. It is produced so that you can manually amend your 2022 CMP to make it suitable for study for the 2023 exams. It includes replacement pages and additional pages where appropriate.

Alternatively, you can buy a full set of up-to-date Course Notes / CMP at a significantly reduced price if you have previously bought the full-price Course Notes / CMP in this subject. Please see our 2023 *Student Brochure* for more details.

We only accept the current version of assignments for marking, *ie* those published for the sessions leading to the 2023 exams. If you wish to submit your script for marking but only have an old version, then you can order the current assignments free of charge if you have purchased the same assignments in the same subject in a previous year, and have purchased marking for the 2023 session.

This CMP Upgrade contains:

- all significant changes to the Syllabus and Core Reading
- additional changes to the ActEd Course Notes and Assignments that will make them suitable for study for the 2023 exams.

1 Changes to the Syllabus

This section contains all the ***non-trivial*** changes to the Syllabus Objectives.

Objective 3.2 and 3.2.2 has been amended as follows:

3.2 Confidence intervals and prediction intervals

3.2.2 Define in general terms a prediction interval for a future observation based on a **model fitted to a** random sample.

2 Changes to the Core Reading and ActEd text

This section contains all the ***non-trivial*** changes to the Core Reading and ActEd text.

Chapter 2

Section 1

Distribution: changes to **Probability mass function:** for the discrete distributions, on pages 3, 5, 6, 7, 9 and 12.

Section 1.6

The support of the variable for the hypergeometric distribution, given in the Core Reading, has been corrected.

Probability function: $P(X = x) = \dfrac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$, $\max\{0, n-N+k\} \le x \le \min\{k, n\}$; $0 < p < 1$.

Section 2.2 and 2.3

Range changes to **support** on pages 18 and 25.

Section 3.1

The Core Reading on page 33 has been updated.

The Poisson process is an example of a counting process. Here the number of events occurring is of interest. Since the number of events is being counted over time, the number of events that take place in the time interval $(0, t]$, $\{N(t)\}_{t \ge 0}$, must satisfy the following conditions:

Chapter 4

The function $f(x, y)$ changes to $p(x, y)$ throughout for the discrete case, on pages 3, 6, 9 and 30 and $f_X(x)$ changes to $p_X(x)$ on page 6.

Chapter 6

Section 4

The Core Reading at the bottom of page 15 has been replaced.

The normal approximation appears to be reasonably good across the core of the range of values, which is consistent with the previous diagram.

Chapter 8

Section 1.2

The final paragraph of Core Reading for the two-parameter case at the bottom of page 4 has been updated.

The first order moment used is the mean. The second order moment can be either about the mean (*ie* the sample variance), which is often easier to calculate, or about the origin, *ie* $E[X^2]$. The solution does not depend on which choice is made.

Section 2.1

New Core Reading has been added to the following paragraph on page 10.

The likelihood is the probability of observing the sample in the discrete case, and is proportional to the probability of observing values in the neighbourhood of the sample in the continuous case. The maximum likelihood estimate for the unknown parameter is simply the value of the parameter at which the likelihood is greatest, if such a value exists. Recall that the stationary points of a function (such as its maximum) are those at which its derivative is zero.

Section 2.2

New Core Reading has been added to the following paragraph on page 13.

The only difference is that a partial derivative is taken with respect to each parameter, before equating each to zero and solving the resulting system of simultaneous equations for the parameters. Checking that the stationary point is a maximum is more complicated when there is more than one parameter to be estimated. This is beyond the scope of the Subject CS1 syllabus.

Section 5

The Core Reading for the CRLB formula has been updated and a new paragraph added on page 27. The definitions have been swapped around. The definition from page 23 of the *Tables* is now given first with the new paragraph below it.

where $\text{CRLB} = \dfrac{1}{-E\left[\dfrac{\partial^2}{\partial\theta^2}\log L(\theta, \underline{X})\right]}$

The CRLB is a lower bound on the variance (and hence mean squared error) of unbiased estimators of the unknown parameter.

The last paragraph of Core Reading for the CRLB have been updated.

Two alternative expressions for the CRLB are:

$$\text{CRLB} = \frac{1}{E\left\{\left[\dfrac{\partial}{\partial\theta}\log L(\theta, \underline{X})\right]^2\right\}} \quad \text{and} \quad \text{CRLB} = \frac{1}{nE\left\{\left[\dfrac{\partial}{\partial\theta}\log f(X, \theta)\right]^2\right\}}$$

Chapter 9

Section 4.1

New Core Reading has been added on page 17 and θ has been changed to p throughout this section. The inequalities have been updated. Replacement pages 17 to 20 are attached.

Section 4.2

New Core Reading has been added on to the end of page 22 to the normal approximation section.

The normal approximation

Again, it is easy for a computer to calculate an exact confidence interval for λ even for a large sample from Poisson(λ), or a single observation from Poisson(λ) where λ is large. However, on a piece of paper a normal approximation can be used either from

$\sum X_i \sim \text{Poi}(n\lambda) \to N(n\lambda, n\lambda)$ or from the Central Limit Theorem as $\bar{X} \to N\left(\lambda, \frac{\lambda}{n}\right)$. (Note that limiting distributions require more careful treatment than is given here: rescaling is needed to eliminate n from the right-hand sides before sense can be made of the convergence in distribution. These informal statements can, nonetheless, be more intuitive.)

Section 5.3

θ has been changed to p throughout this section on pages 28 and 29.

Chapter 10

Section 1.5

The Core Reading has been updated at the bottom of page 5.

The level of significance of the test, denoted α, is the probability of committing a Type I error, *ie* it is the probability of rejecting H_0 when it is in fact true. The probability of committing a Type II error, denoted β, is the probability of failing to reject H_0 when it is false. An ideal test would be one which simultaneously minimises α and β. This ideal however is not attainable in practice.

Section 2.1

The second paragraph of Core Reading has been updated on page 11.

Here it is only in certain special cases (usually one-sided cases) that a single test is available which is best (*ie* uniformly most powerful) for all parameter values. Another approach is used to derive sensible tests in cases where a single best test (in the sense of the Neyman-Pearson lemma) is unavailable. This approach, which is a generalisation of the lemma, produces tests that are referred to as *likelihood ratio tests*.

Section 2.2

The Core Reading has been updated under the probability calculation on page 12.

So, the probability of observing 82 or fewer heads from an unbiased coin is extremely low. Therefore, there is very strong evidence against H_0 and in favour of H_1. A good way of expressing the result is: 'we have very strong evidence against the hypothesis that the coin is fair (p-value 0.007) and conclude that it is biased against heads'.

Testing does not prove that any hypothesis is true or untrue. Failure to detect a departure from H_0 means that there is not enough evidence to justify rejecting H_0. Failure to reject H_0 may therefore reflect a paucity of evidence – it should not usually be taken as confirmation of H_0 being true. Indeed, H_0 is usually a precise statement, which is almost certainly not exactly true.

Section 4.1 (b)

The Core Reading has been updated on page 24.

Large samples: use S_i^2 to estimate σ_i^2 and then proceed as in (a) using a z statistic.

Chapter 11

Section 1.2

The Core Reading has been added to on page 13, under the Kendall rank correlation coefficient section.

Kendall's rank correlation coefficient τ measures the strength of monotonic relationship between two variables (more specifically, it is a measure of dependence that relies only on the ranks of the observations).

Under the formula for Kendall's tau Core Reading has been added on page 14.

When there are ties, the formula can be adjusted to account for this (not examinable).

Section 1.3

The R code in the Core Reading has been updated at the bottom of page 19 to include a null hypothesis.

The R code for carrying out a hypothesis test of the form:

H_0 : Pearson correlation coefficient = 0

using data sets x and y is:

```
cor.test(x, y, method = "pearson")
```

There is no built-in function to perform tests of the form:

H_0 : Pearson correlation coefficient = ρ_0

The R code in the Core Reading has been updated near the top of page 21.

The R code for carrying out a hypothesis test using the Kendall rank correlation coefficient for variables *x* and *y* is:

```
cor.test(x, y, method = "kendall")
```

Note that `cor.test` will determine exact *p*-values if $n < 50$ (assuming no tied values); for larger samples the test statistic is approximately normally distributed.

Section 3

The Core Reading has been updated in to the second paragraph of Core Reading page 26.

Principal component analysis (PCA), also called factor analysis, provides a method for reducing the dimensionality of the data set, $\underline{X}$ – in other words, it allows the user to identify a parsimonious set of components that can then be used to model and understand the data.

New Core Reading has been added on to the third paragraph of Core Reading page 26.

These components are chosen to be uncorrelated linear combinations of the variables of the data that maximise the variance. PCA identifies a set of uncorrelated linear combinations of the original data – geometrically, it represents the data on a different set of axes. The hope is that by retaining the most important axes and ignoring the rest, the dimension of the data set can be reduced, without sacrificing too much information.

New Core Reading has been added to the first sentence on page 30.

Consider our set of equity returns, $\underline{X}$, from four different markets across 12 time periods. From the scatterplot matrix in Section 2.1, there is correlation between many of the variables.

Chapter 12

Section 4.6

The first sentence of Core Reading has been replaced on page 37.

When there is a large number of potential explanatory variables, it can be difficult to select a subset that delivers a close model fit without succumbing to over-fitting. This can occur if too many variables are included: a good 'in-sample' fit may produce poor 'out-of-sample' performance.

In-sample refers to the data that you use to fit the model and out-of-sample refers to the data you use to predict future values.

We highlight here two procedures that are commonly observed, partly because they can be readily automated:

New Core Reading has been added to the end of this section, at the bottom of page 38, before Section 4.7.

It should be admitted that these approaches are not beyond criticism. A stepwise approach may fail to find good subsets of the explanatory variables. For example, a forward stepwise approach can fail to select variables that work well in combination but not individually. On the other hand, since the number of potential models may be large, there is a greater risk of overfitting. Holding back some of the original data for testing the final fitted model can help control this risk.

Chapter 13

Section 0

The third paragraph of Core Reading has been updated on page 3.

This is particularly important in actuarial work where the data very often do not have a normal distribution. For example, in mortality, the Poisson distribution is used in modelling numbers of deaths and the exponential is used for survival analysis.

Section 3.3

New Core Reading has been added on page 24, before the paragraph that begins 'In the following table…'.

This encoding of factors into integer-valued levels (1 to k) is carried out automatically by R when it is given a categorical variable with k levels. More than this, it creates $k-1$ binary variables (*ie* variables taking the values 0 or 1) corresponding to all but one of these levels. These are used in R's model matrix when it fits models. It creates $k-1$ rather than k to avoid the degeneracy that would otherwise be implicit in the linear predictors. In our example, adding the same constant to α_1 and α_2 would be equivalent to adding this constant to the intercept term. Reported results from fitting a model will therefore have one fewer level than may be expected for each factor.

Section 5.5

The last paragraph of Core Reading before the R Code has been updated on page 44.

When comparing two models, the smaller the AIC, the better the balance between model fit and model complexity according to this measure. So if the change in deviance is more than twice the change in the number of parameters then it would give a smaller AIC.

Chapter 14

Section 3.1

The R code for the Monte Carlo Bayesian estimate at the bottom of page 17 has been removed from the Core Reading.

Chapter 15

Section 3.4

The paragraph of Core Reading after equation 15.3.5 has been updated on page 19.

Equation (15.3.4) is a credibility estimate of $E(X \mid \theta)$ since it is a weighted average of two estimates: the first, $\bar{x}$, is a maximum likelihood estimate based solely on data from the risk itself, and the second, μ is the best available estimate if no data were available from the risk itself.

The R code at the top of page 20 has been removed from the Core Reading.

Section 3.5

The last paragraph of Core Reading in this section has been updated on page 22.

This shows that X_1 and X_2 are not unconditionally independent. The relationship between X_1 and X_2 is that their common mean is a random variable. If this mean, θ, were known, then this relationship would be broken and we would have unconditional independence.

Chapter 16

Section 1.4

The first paragraph of Core Reading on page 11 has been updated.

Each row of the earlier table (Table 1) corresponds to a fixed value of θ. Bearing this and the definitions of $m(\theta_i)$ and $s^2(\theta_i)$ in mind, obvious estimators for $m(\theta_i)$ and $s^2(\theta_i)$ are:

$$\bar{X}_i \qquad \text{and} \qquad \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2$$

respectively.

3 Changes to the X Assignments

Overall

There have been minor changes throughout the assignments, including changes to mark allocations.

Additional solutions to mathematical parts of questions have been added in order to illustrate how they could be answered in the exam, using Word.

More significant changes are listed below.

Assignment X1

Solution X1.5

The solution to part (ii) has been expanded and the marks changed to 2, ½ and ½:

(ii) ***Simulation***

Setting $u = F(x)$ and rearranging:

$$u = 1 - \left(\frac{5}{5+x}\right)^2 \Rightarrow \left(\frac{5}{5+x}\right)^2 = 1-u$$

$$\Rightarrow \frac{5}{5+x} = \sqrt{1-u}$$

$$\Rightarrow 5+x = \frac{5}{\sqrt{1-u}}$$

$$\Rightarrow x = \frac{5}{\sqrt{1-u}} - 5 \qquad [2]$$

Setting $u = 0.656$ gives:

$$x = \frac{5}{\sqrt{1-0.656}} - 5 = 3.52 \qquad [½]$$

Setting $u = 0.285$ gives:

$$x = \frac{5}{\sqrt{1-0.285}} - 5 = 0.913 \qquad [½]$$

[Total 3]

Solution X1.12

The solution to part (ii) has been expanded to show the first step of the chain rule:

$$C_S''(t) = 5,000(-4)(-26)(1-4t)^{-27} = 520,000(1-4t)^{-27}$$

$$\Rightarrow \quad \text{var}(S) = C_S''(0) = 520,000 \qquad [1]$$

$$C_S'''(t) = 520,000(-4)(-27)(1-4t)^{-28} = 56,160,000(1-4t)^{-28}$$

$$\Rightarrow \quad \text{skew}(S) = C_S'''(0) = 56,160,000 \qquad [½]$$

Assignment X3

Question X3.14

Part (iii) has been reworded:

(iii) Calculate:

(a) a 90% confidence interval for the average cost of a job lasting 4 hours

(b) a 90% prediction interval for the cost of an individual job lasting 6 hours. [6]

Solution X3.14

The missing ½ mark has been added to the solution in part (ii):

Under H_0: $\dfrac{\hat{b}-b}{\sqrt{\hat{\sigma}^2/s_{xx}}} \sim t_{n-2}$ [½]

The solution to part (iii)(b) and (iv) has been reworded:

(iii)(b) ***Prediction interval for an individual job lasting 6 hours***

This gives a prediction interval of:

$$134.50 \pm 1.943 \times 6.970 = 134.5 \pm 13.54 = (£120.96, £148.04) \qquad [1]$$

(iv) ***Comment***

The prediction interval for the individual job is wider (£27) than the confidence interval for the average cost (£9). So there is greater uncertainty over an individual result than an average one. [1]

Assignment X4

Solution X4.6

In part (ii) the value of the fourth sample variance has been corrected.

The sample variances for each of the risks are:

$$\frac{1}{2}\sum_{j=1}^{3}(x_{1j}-\bar{x}_1)^2=\frac{22,344}{2}=11,172 \qquad \frac{1}{2}\sum_{j=1}^{3}(x_{2j}-\bar{x}_2)^2=\frac{20,294}{2}=10,147$$

$$\frac{1}{2}\sum_{j=1}^{3}(x_{3j}-\bar{x}_3)^2=\frac{21,800}{2}=10,900 \qquad \frac{1}{2}\sum_{j=1}^{3}(x_{4j}-\bar{x}_4)^2=\frac{23,994}{2}=11,997 \quad [½]$$

Solution X4.8

The solution to part (ii) includes null hypotheses and the marks have been changed to give ½ for the first null hypothesis and the last mark has been reduced to ½ mark.

(ii) ***Comparing models***

Comparing SA+PT with SA

H_0 : SA + PT is not a significant improvement over the SA model [½]

The difference in the scaled deviances is $238.4-206.7=31.7$.

This is greater than 16.92, the upper 5% point of the $\chi^2_{11-2}=\chi^2_9$ distribution. So we have sufficient evidence at the 5% significance level to reject the null hypothesis. We conclude that SA + PT **is** a significant improvement over the SA model. [1]

Comparing SA∗PT with SA+PT

H_0 : SA ∗ PT is not a significant improvement over the SA + PT model

The difference in the scaled deviances is $206.7-178.3=28.4$.

This is greater than 16.92, the upper 5% point of the $\chi^2_{20-11}=\chi^2_9$ distribution. So we have sufficient evidence at the 5% significance level to reject the null hypothesis. We conclude that SA ∗ PT **is** a significant improvement over the SA + PT model. [1]

Comparing SA∗PT+NB with SA∗PT

H_0 : SA ∗ PT + NB is not a significant improvement over the SA ∗ PT model

The difference in the scaled deviances is $178.3-166.2=12.1$.

This is greater than 11.07, the upper 5% point of the $\chi^2_{25-20}=\chi^2_5$ distribution. So we have sufficient evidence at the 5% significance level to reject the null hypothesis. We conclude that SA ∗ PT + NB **is** a significant improvement over the SA ∗ PT model. [1]

Comparing SA*PT*NB with SA*PT+NB

H_0 : SA * PT * NB is not a significant improvement over the SA * PT + NB model

The difference in the scaled deviances is $166.2 - 58.9 = 107.3$.

This is less than 118.7, the upper 5% point of the $\chi^2_{120-25} = \chi^2_{95}$ distribution.

We could interpolate in the tables between χ^2_{90} and χ^2_{100} to get the figure of 118.7 or we could use Excel or R. The value of χ^2_{95} can be calculated using the command `qchisq(0.95,95)` *in R.*

So we have insufficient evidence at the 5% significance level to reject the null hypothesis. We conclude that SA * PT * NB is **not** a significant improvement over the SA * PT + NB model. [1]

So the analyst should choose the SA * PT+NB model. [½]

[Total 5]

Solution X4.10

The solution to part (iii)(a) has been updated and includes a null hypothesis and the marks have been changed to allow for this.

(iii)(a) ***Significant improvement?***

We are testing:

H_0 : the original model is not a significant improvement over the simplified model

H_1 : the original model is a significant improvement over the simplified model [½]

The observed value of the test statistic is the difference in the scaled deviances:

$0.135 - 0.0120 = 0.123$ [½]

The original model has 2 parameters (α and β) and the simplified model has 1 parameter (α). Under the null hypothesis, the test statistic is (approximately) a realisation of a χ^2_1 random variable. [½]

The upper 5% point of the χ^2_1 distribution is 3.841. Since $0.123 < 3.841$, there is insufficient evidence to reject the null hypothesis at the 5% significance level. So we conclude that the original model is not a significant improvement over the simplified model, *ie* including the parameter β does not reduce the scaled deviance enough to justify the additional model complexity. [1]

4 Changes to the Y Assignments

Overall

There have been minor changes throughout the assignments, including changes to mark allocations.

More significant changes are listed below.

Assignment Y2

Question Y2.1 is new and the old Question Y2.1 is now Question Y2.3. The old Question Y2.3 is now Question Y2.4. New pages for Question Y2.1 and Solution Y2.1 are attached.

Question Y2.2

Parts (iv) and (v) have been removed.

Solution Y2.2

(ii)(a) ***Plot the prior density of μ***

The coding for yvals has been corrected.

```
yvals <- dnorm(xvals,pmean,sqrt(pvar))
```
[1]

The marks for (ii) (b) are now all ½ marks to give a total of 6 marks for part (ii).

Parts (iv) and (v) have been removed.

5 Other tuition services

In addition to the CMP you might find the following services helpful with your study.

5.1 Study material

We also offer the following study material in Subject CS1:

- Flashcards
- Revision Notes
- ASET (ActEd Solutions with Exam Technique) and Mini-ASET
- Mock Exam and AMP (Additional Mock Pack).

For further details on ActEd's study materials, please refer to the 2023 *Student Brochure*, which is available from the ActEd website at **ActEd.co.uk**.

5.2 Tutorials

We offer the following (face-to-face and/or online) tutorials in Subject CS1:

- a set of Regular Tutorials (four days or eight half days)
- a Block (or Split Block) Tutorial (lasting four full days)
- a Paper B preparation day
- a five-day Bundle (four days of regular tutorials plus a Paper B preparation day)
- an Online Classroom.

For further details on ActEd's tutorials, please refer to our latest *Tuition Bulletin*, which is available from the ActEd website at **ActEd.co.uk**.

5.3 Marking

You can have your attempts at any of our assignments or mock exams marked by ActEd. When marking your scripts, we aim to provide specific advice to improve your chances of success in the exam and to return your scripts as quickly as possible.

For further details on ActEd's marking services, please refer to the 2023 *Student Brochure*, which is available from the ActEd website at **ActEd.co.uk**.

5.4 Feedback on the study material

ActEd is always pleased to receive feedback from students about any aspect of our study programmes. Please let us know if you have any specific comments (*eg* about certain sections of the notes or particular questions) or general suggestions about how we can improve the study material. We will incorporate as many of your suggestions as we can when we update the course material each year.

If you have any comments on this course, please send them by email to **CS1@bpp.com**.

4 Confidence intervals for binomial & Poisson parameters

Both these situations involve a discrete distribution, which introduces the difficulty of probabilities not being exactly 0.95, and so 'at least 0.95' is used instead. Also, when not using the large-sample normal approximations, the pivotal quantity method must be adjusted.

One approach is to use a quantity $h(\underline{X})$ whose distribution involves θ such that:

$$P\left(h_1(\theta) < h(\underline{X}) < h_2(\theta)\right) \geq 0.95$$

Then if both $h_1(\theta)$ and $h_2(\theta)$ are monotonic increasing (or both decreasing), the inequalities can be inverted to obtain a confidence interval as before.

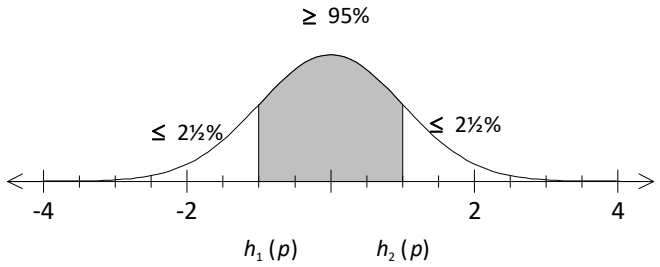
4.1 The binomial distribution

If X is a single observation from $Bin(n, p)$, where n is known and p is unknown, the maximum likelihood estimator is:

$$\hat{p} = \frac{X}{n}$$

What follows is a slight diversion from our aim of obtaining a confidence interval for p. It is just demonstrating that the method is sound.

Using X as the quantity $h(\underline{X})$, it is necessary to find if $h_1(p)$ and $h_2(p)$ exist such that $P\left(h_1(p) < X < h_2(p)\right) \geq 0.95$, where with equal tails $P\left(X \leq h_1(p)\right) \leq 0.025$ and $P\left(X \geq h_2(p)\right) \leq 0.025$.



We can have *at most* 2.5% in the lower (or upper) tail, so we need to be very careful about finding the values of h_1 and h_2.

There is no explicit expression for the quantities $h(p_1)$ and $h_2(p)$.

For example, for the $Bin(20, 0.3)$ case:

$$P(X \leq 1) = 0.0076 \text{ and } P(X \leq 2) = 0.0355 \quad \therefore h_1(p) = 1$$

Also:

$$P(X \geq 11) = 0.0171,\ P(X \geq 10) = 0.0480 \quad \therefore h_2(p) = 11$$

Question

Calculate the values of h_1 and h_2 for the binomial distribution with parameters $n = 20$ and $p = 0.4$.

Solution

If $X \sim Bin(20, 0.4)$, then (using page 188 of the *Tables*) $P(X \leq 3) = 0.0160$ and $P(X \leq 4) = 0.0510$. So $h_1 = 3$.

Also $P(X \geq 13) = 0.0210$ and $P(X \geq 12) = 0.0565$. So $h_2 = 13$.

h_1 and h_2 have higher values than for the $Bin(20, 0.3)$ case.

So $h_1(p)$ and $h_2(p)$ do exist and increase with p.

We're back on track. We can move on to obtain our confidence interval for p.

Therefore the inequalities from $P(h_1(p) < X < h_2(p)) \geq 0.95$ can be inverted as follows:

$$X \leq h_2(p) \Rightarrow p \geq p_2(X)$$

$$X \geq h_1(p) \Rightarrow p \leq p_1(X)$$

This gives a 95% confidence interval of the form $p_2(X) < p < p_1(X)$.

Note: The lower limit $p_2(X)$ comes from the upper tail probabilities and the upper limit $p_1(X)$ from the lower tail probabilities.

We'll see this is the case in the question on the next page.

However since there are no explicit expressions for $h_1(p)$ and $h_2(p)$, there are no expressions for $p_1(X)$ and $p_2(X)$ and they will have to be calculated numerically.

So, adopting the convention of including the observed x in the tails, p_1 and p_2 can be found by solving:

$$\sum_{r=x}^{n} b(r; n, p_2) = 0.025 \quad \text{and} \quad \sum_{r=0}^{x} b(r; n, p_1) = 0.025$$

where $b(r; n, p)$ is the binomial probability of observing r successes from n trials.

These can be expressed in terms of the distribution function $F(x;p)$:

$$1-F(x-1;p_2)=0.025 \quad \text{and} \quad F(x;p_1)=0.025$$

Note: Equality can be attained as p has a continuous range $(0,1)$ and the 'discrete' problem does not arise.

The R function for an exact 95% confidence interval for the proportion is:

```
binom.test(x,n, conf=0.95)
```

Question

We have obtained a value of 1 from the binomial distribution with parameters $n=20$ and p. Construct a 95% symmetrical confidence interval for p.

Solution

We need the upper limit, p_1, such that $P(X\le 1)=0.025$ under $Bin(20,p_1)$. We also need the lower limit, p_2, such that $P(X\ge 1)=0.025$ under $Bin(20,p_2)$.

For the first equation, we have $(1-p_1)^{20}+20(1-p_1)^{19}p_1=0.025$.

Solving this we obtain $p_1=0.249$.

A numerical method will be needed here, or trial and improvement. One approach would be to write the equation in the form $(1-p_1)^{19}(1+19p_1)=0.025$*, then iterate using*

$p_{n+1}=1-\left(\dfrac{0.025}{1+19p_n}\right)^{\frac{1}{19}}$ *starting with* $p_1=0.5$.

For the second equation we have $(1-p_2)^{20}=0.975$.

Solving this we obtain $p_2=0.00127$.

Our confidence interval is then $(0.00127, 0.249)$.

The normal approximation

It is easy for a computer to calculate an exact confidence interval for the binomial parameter p even if n is 'large'. However, on a piece of paper we use the normal approximation to the binomial distribution.

$\dfrac{X-np}{\sqrt{np(1-p)}}$ **can be used as a pivotal quantity.**

Solving the resulting equations for p would not be easy.

However $\dfrac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}}$, with $\hat{p}$ in place of p (in the denominator only)**, can be used in a simpler way and yields the standard 95% confidence interval used in practice, namely:**

$$\frac{X \pm 1.96\sqrt{n\hat{p}(1-\hat{p})}}{n}$$

or $\hat{p} \pm 1.96\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$, **where** $\hat{p} = \dfrac{X}{n}$.

Question

In a one-year mortality investigation, 45 of the 250 ninety-year-olds present at the start of the investigation died before the end of the year. Assuming that the number of deaths has a binomial distribution with parameters $n = 250$ and q, calculate a symmetrical 90% confidence interval for the unknown mortality rate q.

Solution

Since 250 is a large sample, we know that $\dfrac{X - nq}{\sqrt{nq(1-q)}} \sim N(0,1)$ approximately.

Since $P(-1.6449 < Z < 1.6449) = 0.90$, we can say that:

$$P\left(-1.6449 < \frac{X - 250q}{\sqrt{250q(1-q)}} < 1.6449\right) = 0.90$$

Rearranging this:

$$P\left(\frac{X}{250} - 1.6449\sqrt{\frac{q(1-q)}{250}} < q < \frac{X}{250} + 1.6449\sqrt{\frac{q(1-q)}{250}}\right) = 0.90$$

Replacing X by the observed value of 45 gives $q = \dfrac{45}{250}$.

Therefore a symmetrical 90% confidence interval for q is $(0.140, 0.220)$.

Y2.1 A market research organisation conducted a survey in a large city. A random sample of 400 households showed that 68 were such that at least one person in the household was a member of a health/fitness club.

(i) Calculate a 90% confidence interval for the true proportion of households in the city with at least one person being a member of a health/fitness club. [3]

A 2×2 contingency table was set up to investigate whether membership of a health/fitness club was dependent on the distance to the nearest club (split into <10 mins away and >10 mins away).

	Member	Not member
< 10 mins	32	118
> 10 mins	36	214

The organisation carries out an appropriate χ^2 test of:

H_0 : membership and distance are independent

H_1 : membership and distance are not independent

(ii) Carry out this test at the 5% level, stating your conclusion clearly. [4]

[Total 7]

Assignment Y2 – Solutions

Markers: This document sets out one approach to solving each of the questions (sometimes with alternatives). Please give credit for any other valid approaches.

Solution Y2.1

(i) ***Confidence interval***

The code is:

```
binom.test(68,400, alternative = "two.sided",
     conf.level = 0.90)
```

[1½]

Since the alternative hypothesis is the default option, students can omit this argument and still receive full marks. Students may also abbreviate the arguments. For example, `alternative` *to* `alt` *and* `conf.level` *to* `conf`*.*

```
        Exact binomial test

data:  68 and 400
number of successes = 68, number of trials = 400, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
90 percent confidence interval:
 0.1397511 0.2039428
sample estimates:
probability of success
                  0.17
```

[½]

Some kind of output is required – though students may choose to extract the confidence interval only, by using `$conf.int` *(or any abbreviation).*

The 90% confidence interval for *p* is (0.1397511, 0.2039428) or (0.140,0.204) to 3 SF. [1]

Students must explicitly give the confidence interval separately from the full output. This can be manually written or extracted from the output using `$conf.int`*.*

[Total 3]

(ii) ***Contingency table test***

Storing the data as a matrix:

```
obs <- matrix(c(32,118,36,214), nrow = 2, ncol = 2,
           byrow = TRUE)
```

[1]

Students only need to include one of nrow or ncol for full marks. Students may also enter this matrix by columns instead of rows as follows:

```
obs <- matrix(c(32,36,118,214), nrow = 2, ncol = 2)
```


Carrying out the test:

```
chisq.test(obs)
```
[½]

No other arguments are necessary for the mark. Students may remove the default continuity correction using the option `correct = FALSE` *but whilst this is consistent with the paper version it is less accurate and so should be discouraged.*

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  obs
X-squared = 2.8406, df = 1, p-value = 0.09191
```
[½]

Some kind of output is required – though students may choose to just extract the p-value only, by using `$p.value` *(or an abbreviation of at least two characters).*

This gives a *p*-value of 0.09191. [½]

Students must explicitly give the p-value separate from the full output. This can be manually written or extracted from the output using `$p.value`*.*

This is less than 5%. [½]

Hence, we have insufficient evidence to reject H_0 at the 5% level. [½]

Therefore, it is reasonable to assume that membership and distance are independent. [½]

[Total 4]

If students remove the continuity correction using `correct = FALSE` *they will obtain a p-value of 0.06831 with the same conclusion.*