# Subject CS1

## CMP Upgrade 2021/22

### CMP Upgrade

This CMP Upgrade lists the changes to the Syllabus objectives, 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 2021 CMP to make it suitable for study for the 2022 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 2022 Student Brochure for more details.

We only accept the current version of assignments for marking, ie those published for the sessions leading to the 2022 exams. If you wish to submit your script for marking but have only 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 2022 session.

This CMP Upgrade contains:

- all significant changes to the Syllabus objectives and Core Reading
- additional changes to the ActEd Course Notes and Assignments that will make them suitable for study for the 2022 exams.

# 0 Changes to the Syllabus

This section contains all the non-trivial changes to the syllabus objectives.

Prediction intervals have been added to Objective 3.2 as follows:

3.2 Confidence intervals

3.2.1 Define in general terms a confidence interval for an unknown parameter of a distribution based on a random sample.

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

3.2.3 Derive a confidence interval for an unknown parameter using a given sampling distribution.

3.2.4 Calculate confidence intervals for the mean and the variance of a normal distribution.

3.2.5 Calculate confidence intervals for a binomial probability and a Poisson mean, including the use of the normal approximation in both cases.

3.2.6 Calculate confidence intervals for two-sample situations involving the normal distribution, and the binomial and Poisson distributions using the normal approximation.

3.2.7 Calculate confidence intervals for a difference between two means from paired data.

3.2.8 Use the bootstrap method to obtain confidence intervals.

Sensitivity and specificity have been added to Objective 3.3.1 as follows:

3.3 Hypothesis testing and goodness of fit

3.3.1 Explain what is meant by the following terms: null and alternative hypotheses, simple and composite hypotheses, type I and type II errors, sensitivity, specificity, test statistic, likelihood ratio, critical region, level of significance, probability-value and power of a test.

3.3.2 Apply basic tests for the one-sample and two-sample situations involving the normal, binomial and Poisson distributions, and apply basic tests for paired data.

3.3.3 Apply the permutation approach to non-parametric hypothesis tests.

3.3.4 Use a chi-square test to test the hypothesis that a random sample is from a particular distribution, including cases where parameters are unknown.

3.3.5 Explain what is meant by a contingency (or two-way) table, and use a chi-square test to test the independence of two classification criteria.

Credible intervals have been added to Objective 5.1 as follows:

5.1 Explain the fundamental concepts of Bayesian statistics and use these concepts to calculate Bayesian estimates.

5.1.1 Use Bayes' theorem to calculate simple conditional probabilities.

5.1.2 Explain what is meant by a prior distribution, a posterior distribution and a conjugate prior distribution.

5.1.3 Derive the posterior distribution for a parameter in simple cases.

5.1.4 Explain what is meant by a loss function.

5.1.5 Use simple loss functions to derive Bayesian estimates of parameters.

5.1.6 Derive credible intervals in simple cases.

# 1 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 1.6

The range of values taken by the hypergeometric distribution, given in the Core Reading, has been corrected to start at zero. Replacement pages are attached.

**Distribution:** $P(X=x)=\dfrac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}\ ,\ x=0,1,2,3,\ldots\ ;\ 0<p<1.$

## Chapter 3

### Section 2

An error in the derivation of the skewness has been corrected. Replacement pages are attached.

$$=\frac{E[X^3](1)^3-3(1)^2E[X]E[X^2]+2(1)(E[X])^3}{1^4}$$

## Chapter 4

### Solutions

The solutions to parts (i) and (ii) of Question 4.5 should include the input values for which they are valid. Replacement pages are attached.

## Chapter 6

### Solutions

The solution to Question 6.9, the mean and variance of the approximate normal distribution (200 and 300) are the wrong way around in the standardisations (although the resulting numbers are correct). Replacement pages are attached.

## Chapter 9

Section 1

New Core Reading and ActEd text has been added to give a general overview of prediction intervals. Replacement pages are attached.

Section 2.1

A new formula and an example of prediction intervals have been included in the case when sampling from a normal distribution with known variance. Replacement pages are attached.

Section 3.3

A new section has been added to cover prediction intervals in the case when sampling from a normal distribution with unknown variance. Replacement pages are attached.

Summary

This has been updated to include the new prediction interval formulae. Replacement pages are attached.

## Chapter 10

Section 1.5

New Core Reading and ActEd text has been added to cover the new sensitivity and specificity objective. Replacement pages are attached.

Summary

This has been updated to include sensitivity and specificity. Replacement pages are attached.

## Chapter 11

Section 1.3

The final sentence on page 22 is missing the word 'evidence'. It should read:

So we have very strong evidence to reject $H_0$, and we conclude that the mock scores in CS1 and CS2 are positively correlated.

## Chapter 12

Section 2.4

The value of $\hat{\alpha}$ stated at the end of the question on page 20 is negative when it should be positive. The sentence should say:

Recall that $\hat{\alpha}=0.164, \hat{\beta}=0.88231, \hat{\sigma}^2=0.0732$ and $S_{xx}=8.444$.

# Chapter 13

## Section 5.4

The following has been added to the end of this section on page 42:

The residual deviance outputted by the `glm()` function is a measure of fit, similar to the scaled deviance and deviance defined earlier. However, this output won't necessarily match the scaled deviance or deviance calculated from first principles using the formulae in this section.

## Section 5.6

The first line of R code on page 47 should be updating `modelA` instead of `model1`. It should read:

**We remove the interaction term `wt:disp`, as this is the least significant.**

```
      modelB <- update(modelA, ~.-wt:disp)
```

## Section 6.1

The following has been added to the Pearson residual's R box on page 50:

The Pearson residuals returned by R are calculated slightly differently from the definition given in this section. Therefore, this output won't necessarily match the Pearson residuals calculated from first principles using $\dfrac{y-\hat{\mu}}{\sqrt{\text{var}(\hat{\mu})}}$.

## Section 6.2

The following has been added to the deviance residual's R box on page 51:

The deviance residuals returned by R are calculated slightly differently from the definition given in this section. Therefore, this output won't necessarily match the deviance residuals calculated from first principles using the formulae in this section.

## Section 6.3

The Core Reading under the graph of the residuals on page 52 has been corrected to read:

**There does appear to be some pattern here and the three named points on the graph might be outliers.**

## Summary

The summary paragraph for backward selection should say that covariates are removed until the AIC reaches a minimum, not until it reaches a maximum.

Replacement pages are attached for all of the above changes.

## Chapter 14

### Section 5

A new section has been added to cover the new credible intervals objective.  Replacement pages are attached.

### Summary

This has been updated to include the new credible interval objective.  Replacement pages are attached.

## Chapter 15

### Section 3.4

Two typos in the Core Reading R code on page 20 have been corrected.  Replacement pages are attached.

## Chapter 16

### Section 2.2

The formulae on page 22 have been corrected to include their conditionality on theta.  Replacement pages are attached.

### Section 2.4

Part d of the question on page 28 has been deleted.

# 2 Changes to the X Assignments

## Overall

The X Assignments have been changed significantly to reflect the online nature of the exams. We have not detailed all of the changes in this upgrade.

If you would like the new assignments without marking, then retakers can purchase an updated CMP or standalone X Assignments at a significantly reduced price. Further information on retaker discounts can be found at:

www.acted.co.uk/paper_reduced_prices.html

If you wish to submit your scripts for marking but have only 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 2022 session. We only accept the current version of assignments for marking, ie those published for the sessions leading to the 2022 exams.

.

# 3 Changes to the Y Assignments

The Y2 Assignment has been changed significantly to better reflect the CS1 Paper B exams. We have not detailed all of the changes in this upgrade.

The new assignment available on the PBOR.

# 4 Other tuition services

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

## 4.1 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 2022 Student Brochure, which is available from the ActEd website at www.ActEd.co.uk.

## 4.2 Tutorials

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

- a set of Regular Tutorials (lasting a total of four days plus one day for R)
- a Split Block Tutorial (lasting four full days plus one day for R)
- 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 www.ActEd.co.uk.

## 4.3 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 2022 Student Brochure, which is available from the ActEd website at www.ActEd.co.uk.

## 4.4 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.

# 2 Cumulant generating functions

**For many random variables the cumulant generating function (CGF) is easier to use than the MGF in evaluating the mean and variance.**

## Definition

**The cumulant generating function, $C_X(t)$, of a random variable $X$ is given by:**

$$\boldsymbol{C_X(t) = \ln M_X(t)}$$

We can treat this as the definition of the CGF.

## Question

The MGF of the $Bin(n,p)$ distribution is given by:

$$M(t) = \left(q + pe^t\right)^n$$

State the CGF of the $Bin(n,p)$ distribution.

## Solution

$$C_X(t) = \ln M_X(t) = \ln(q + pe^t)^n = n\ln(q + pe^t)$$

**As a result, if $C_X(t)$ is known, it is easy to determine $M_X(t)$.**

We have $M_X(t) = e^{C_X(t)}$.

## Calculating moments

**The first three derivatives of $C_X(t)$ evaluated at $t = 0$ give the mean, variance and skewness of $X$ directly.**

**These results can be proved as follows:**

$$C'_X(t) = \frac{M'_X(t)}{M_X(t)}$$

$$C''_X(t) = \frac{M''_X(t)M_X(t) - (M'_X(t))^2}{(M_X(t))^2}$$

and $$C'''_X(t) = \frac{M'''_X(t)(M_X(t))^3 - 3(M_X(t))^2 M'_X(t)M''_X(t) + 2M_X(t)(M'_X(t))^3}{(M_X(t))^4}$$

**Now $M_X(0) = 1$, so:**

$$C'_X(0) = \frac{M'_X(0)}{M_X(0)} = \frac{E[X]}{1}$$

$$C''_X(0) = \frac{M''_X(0)M_X(0) - (M'_X(0))^2}{M_X^2(0)} = \frac{E[X^2](1) - (E[X])^2}{1^2} = \text{var}[X]\,;$$

**and**
$$C'''_X(0) = \frac{M'''_X(0)(M_X(0))^3 - 3(M_X(0))^2 M'_X(0)M''_X(0) + 2M_X(0)(M'_X(0))^3}{(M_X(0))^4}$$

$$= \frac{E[X^3](1)^3 - 3(1)^2 E[X]E[X^2] + 2(1)(E[X])^3}{1^4}$$

$$= skew(X)$$

## Question

State the CGF of $X$ where $X \sim Gamma(\alpha, \lambda)$. Hence prove that $E(X) = \frac{\alpha}{\lambda}$, $\text{var}(X) = \frac{\alpha}{\lambda^2}$ and $skew(X) = \frac{2\alpha}{\lambda^3}$.

## Solution

$$M_X(t) = \frac{\lambda^\alpha}{(\lambda - t)^\alpha} = \left(1 - \frac{t}{\lambda}\right)^{-\alpha} \Rightarrow C_X(t) = -\alpha \ln\left(1 - \frac{t}{\lambda}\right) \qquad t < \lambda$$

Differentiating with respect to $t$:

$$C'_X(t) = -\alpha \times \frac{-\frac{1}{\lambda}}{\left(1 - \frac{t}{\lambda}\right)} = \frac{\alpha}{\lambda}\left(1 - \frac{t}{\lambda}\right)^{-1} \qquad \Rightarrow E(X) = C'_X(0) = \frac{\alpha}{\lambda}$$

$$C''_X(t) = -\frac{\alpha}{\lambda}\left(1 - \frac{t}{\lambda}\right)^{-2} \times -\frac{1}{\lambda} = \frac{\alpha}{\lambda^2}\left(1 - \frac{t}{\lambda}\right)^{-2} \Rightarrow \text{var}(X) = C''_X(0) = \frac{\alpha}{\lambda^2}$$

$$C'''_X(t) = -\frac{2\alpha}{\lambda^2}\left(1 - \frac{t}{\lambda}\right)^{-3} \times -\frac{1}{\lambda} = \frac{2\alpha}{\lambda^3}\left(1 - \frac{t}{\lambda}\right)^{-3} \Rightarrow skew(X) = C'''_X(0) = \frac{2\alpha}{\lambda^3}$$

**The coefficient of $\frac{t^r}{r!}$ in the Maclaurin series of $C_X(t) = \ln M_X(t)$ is called the $r$ th cumulant and is denoted by $\kappa_r$.**

Similarly:

$$P_{N|M=m}(m,n)=\frac{P(N=n,M=m)}{P(M=m)}=\left(\frac{m}{35\times 2^{n-2}}\right)\div\frac{m}{10}=\frac{1}{7\times 2^{n-3}},\quad n=1,\,2,\,3$$

is the conditional probability function of $N$ given $M=m$.

These are identical to the marginal distributions obtained in the chapter text.

4.5 (i) ***Marginal density***

$$f_X(x)=\int_{y=0}^{1}\frac{4}{5}\left(3x^2+xy\right)dy=\left[\frac{4}{5}\left(3x^2y+\frac{1}{2}xy^2\right)\right]_{y=0}^{1}=\frac{4}{5}\left(3x^2+\frac{1}{2}x\right)\quad\text{for } 0<x<1\qquad [2]$$

(ii) ***Conditional density***

$$f_{Y|X=x}(x,y)=\frac{f_{X,Y}(x,y)}{f_X(x)}=\frac{\frac{4}{5}\left(3x^2+xy\right)}{\frac{4}{5}\left(3x^2+\frac{1}{2}x\right)}=\frac{3x^2+xy}{3x^2+\frac{1}{2}x}=\frac{3x+y}{3x+\frac{1}{2}}\quad\text{for } 0<y<1\qquad [1]$$

(iii) ***Covarianc***e

Using the marginal density function of $X$:

$$E(X)=\int_{x=0}^{1}\frac{4}{5}\left(3x^3+\frac{1}{2}x^2\right)dx=\frac{4}{5}\left[\frac{3}{4}x^4+\frac{1}{6}x^3\right]_{x=0}^{1}=\frac{11}{15}\qquad [1]$$

Obtaining the marginal density function of $Y$:

$$f_Y(y)=\int_{x=0}^{1}\frac{4}{5}\left(3x^2+xy\right)dx=\frac{4}{5}\left[x^3+\frac{1}{2}x^2y\right]_{x=0}^{1}=\frac{4}{5}\left(1+\frac{1}{2}y\right)\quad\text{for } 0<y<1$$

So:

$$E(Y)=\int_{y=0}^{1}\frac{4}{5}\left(y+\frac{1}{2}y^2\right)dy=\frac{4}{5}\left[\frac{1}{2}y^2+\frac{1}{6}y^3\right]_{y=0}^{1}=\frac{8}{15}\qquad [1]$$

Now:

$$E(XY)=\int_{x=0}^{1}\int_{y=0}^{1}\frac{4}{5}\left(3x^3y+x^2y^2\right)dy\,dx$$

$$=\int_{x=0}^{1}\frac{4}{5}\left[\frac{3}{2}x^3y^2+\frac{1}{3}x^2y^3\right]_{y=0}^{1}dx$$

$$=\int_{x=0}^{1}\frac{4}{5}\left(\frac{3}{2}x^3+\frac{1}{3}x^2\right)dx$$

$$=\frac{4}{5}\left[\frac{3}{8}x^4+\frac{1}{9}x^3\right]_{x=0}^{1}$$

$$=\frac{7}{18}$$ [2]

Hence:

$$\text{cov}(X,Y)=\frac{7}{18}-\frac{11}{15}\times\frac{8}{15}=-\frac{1}{450}$$ [1]

4.6 The covariance of $X$ and $Y$ was obtained in Section 2.4 to be $\text{cov}(X,Y)=0.02$. The variances of the marginal distributions are:

$$\text{var}(X)=E(X^2)-\left[E(X)\right]^2=0^2\times0.4+1^2\times0.3+2^2\times0.3-(0.9)^2=0.69$$

and: $$\text{var}(Y)=E(Y^2)-\left[E(Y)\right]^2=1^2\times0.2+2^2\times0.4+3^2\times0.4-\left(2.2\right)^2=0.56$$

So the correlation coefficient is:

$$corr(X,Y)=\frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}=\frac{0.02}{\sqrt{0.69\times0.56}}=0.0322$$

4.7 Let $X$ be the amount of a home insurance claim and $Y$ the amount of a car insurance claim. Then:

$$X\sim N(800,100^2)\quad\text{and}\quad Y\sim N(1200,300^2)$$

We require:

$$P\left((Y_1+Y_2+Y_3)>(X_1+X_2+X_3+X_4)+800\right)$$

$$=P\left((Y_1+Y_2+Y_3)-(X_1+X_2+X_3+X_4)>800\right)$$

6.6 Let $X$ be the number of individuals with blood group A.

$$X \sim Bin(300, 0.45) \dot{\sim} N(135, 74.25) \qquad [1]$$

Using a continuity correction $P(X > 115)$ becomes $P(X > 115.5)$: [1]

$$P\left(Z > \frac{115.5 - 135}{\sqrt{74.25}}\right) = P(Z > -2.263) = P(Z < 2.263) = 0.988 \qquad [1]$$

6.7 If our population is normal, we do not need the central limit theorem. The distribution of $\overline{X}$ is exactly normal:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad [1]$$

Hence:

$$P(\overline{X} > 26) = P\left(Z > \frac{26 - 25}{2/\sqrt{16}}\right) = P(Z > 2) = 1 - 0.97725 = 0.02275 \qquad [2]$$

6.8 Let $X_i$ be the sum assured under the $i$ th policy.

We require:

$$P\left[\sum_{i=1}^{100} X_i > 845{,}000\right]$$

Now, according to the Central Limit Theorem:

$$\sum_{i=1}^{100} X_i \sim N\left(100 \times 8000,\ 100 \times 3000^2\right) \text{ (approximately)} \qquad [1]$$

Therefore:

$$P\left[\sum_{i=1}^{100} X_i > 845{,}000\right] \approx P\left(Z > \frac{845{,}000 - 800{,}000}{30{,}000}\right) = P\left(Z > 1.5\right)$$

$$= 1 - 0.93319 = 0.06681 \qquad [2]$$

6.9 We have the sum of 100 discrete uniform random variables, $X_i$ $(i = 1, 2, \cdots, 100)$. Using the formulae from page 10 of the *Tables*, with $a = 1$, $b = 5$ and $h = 1$, we get:

$$E(X_i) = \frac{1+5}{2} = 3$$

$$\text{var}(X_i) = \frac{1}{12}(5-1)(5-1+2) = 2 \qquad [1]$$

Using the Central Limit Theorem:

$$S=\sum_{i=1}^{100} X_i \dot{\sim} N(300,200) \qquad [1]$$

Using a continuity correction, the probability is:

$$P(280 \le S \le 320) \approx P(279.5 < S < 320.5) \qquad [1]$$

Standardising this:

$$\begin{aligned} P(279.5<S<320.5) &= P(S<320.5)-P(S<279.5) \\ &= P\left(Z<\frac{320.5-300}{\sqrt{200}}\right)-P\left(Z<\frac{279.5-300}{\sqrt{200}}\right) \\ &= P(Z<1.44957)-P(Z<-1.44957) \\ &= P(Z<1.44957)-[1-P(Z<1.44957)] \\ &= 2P(Z<1.44957)-1 \\ &= 2\times 0.92641-1 \\ &= 0.85282 \end{aligned} \qquad [2]$$

6.10 (i)(a) ***Mode***

The mode is the maximum of the PDF $f(y)$:

$$f(y)=\frac{\lambda^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\lambda y} \qquad y>0$$

Differentiating and setting the derivative equal to zero gives:

$$\frac{d}{dy}f(y)=\frac{\lambda^\alpha}{\Gamma(\alpha)}\left[(\alpha-1)y^{\alpha-2}e^{-\lambda y}-\lambda y^{\alpha-1}e^{-\lambda y}\right]$$

$$\Rightarrow \quad y^{\alpha-2}e^{-\lambda y}[(\alpha-1)-\lambda y]=0$$

*Alternatively, we could differentiate the log of the PDF.*

This gives:

$$y=0 \qquad \text{or} \qquad y=\frac{\alpha-1}{\lambda}$$

Since $f(y)\ge 0$ and $f(0)=0$, the first solution of zero **must** be a minimum and therefore the second solution **must** be a maximum.

# 1 Confidence intervals in general

**A confidence interval provides an 'interval estimate' of an unknown parameter (as opposed to a 'point estimate'). It is designed to contain the parameter's value with some stated probability. The width of the interval provides a measure of the precision of the estimator involved.**

**A $100(1-\alpha)\%$ confidence interval for $\theta$ is defined by specifying random variables $\hat{\theta}_1(\underline{X})$, $\hat{\theta}_2(\underline{X})$ depending on the sample $\underline{X} = (X_1, \dots, X_n)$ such that $P(\hat{\theta}_1(\underline{X}) < \theta < \hat{\theta}_2(\underline{X})) = 1-\alpha$ .**

**Rightly or wrongly, $\alpha = 0.05$ leading to a 95% confidence interval, is by far the most common case used in practice and we will tend to use this in most of our illustrations.**

**Thus $P\left(\hat{\theta}_1(\underline{X}) < \theta < \hat{\theta}_2(\underline{X})\right) = 0.95$ specifies $\left(\hat{\theta}_1(\underline{X}), \hat{\theta}_2(\underline{X})\right)$ as a 95% confidence interval for $\theta$ . This emphasises the fact that it is the interval and not $\theta$ that is random. In the long run, 95% of the realisations of such intervals will include $\theta$ and 5% of the realisations will not include $\theta$ .**

**Confidence intervals are not unique. In general they should be obtained via the sampling distribution of a good estimator, in particular the maximum likelihood estimator. Even then there is a choice between one-sided and two-sided intervals and between equal-tailed and shortest-length intervals although these are often the same, *eg* for sampling distributions that are symmetrical about the unknown value of the parameter.**

We will see some examples of these shortly.

**Often, we are more interested in statements about future observations than about the parameters underlying the distribution of these observations.**

**This arises in the context of regression models, for example, when a fitted model is being used to make predictions about future observations. Even if the parameter $\theta$ equals the unknown mean of the distribution, it will not be the case that a future observation will fall within a 95% confidence interval with probability 95%. For this, a prediction interval is required.**

A confidence interval gives us some information about the value of a fixed *parameter*, $\theta$ , from a particular distribution, but a prediction interval gives us information about the next future *value* from that distribution, $X$ .

**A $100(1-\alpha)\%$ prediction interval for $X_{n+1}$ is defined by random variables $l(\underline{X})$ , $h(\underline{X})$ such that $P(l(\underline{X}) < X_{n+1} < h(\underline{X})) = 1-\alpha$ . Prediction intervals are, like confidence intervals, not unique but typical choices are one-sided or symmetric. Prediction intervals can be defined more generally for functions of one or more future observations.**

For example, in Chapter 12, we will predict the output of the function $\alpha + \beta x$ .

# 2 Derivation of confidence and prediction intervals

## 2.1 The pivotal method

**There is a general method of constructing confidence intervals called the pivotal method.**

**This method requires a pivotal quantity of the form** $g(\underline{X},\theta)$ **to be found with the following properties:**

**(1) it is a function of the sample values and the unknown parameter** $\theta$

**(2) its distribution is completely known**

**(3) it is monotonic in** $\theta$ **.**

The distribution in condition (2) must not depend on $\theta$. 'Monotonic' means that the function either consistently increases or decreases with $\theta$.

**The equation:**

$$\int_{g_1}^{g_2} f(t)\,dt = 0.95, \quad \text{(where } f(t) \text{ is the known probability (density) of } g(\underline{X},\theta))$$

**defines two values,** $g_1$ **and** $g_2$**, such that:**

$$P\big(g_1 < g(\underline{X},\theta) < g_2\big) = 0.95$$

$g_1$ and $g_2$ are usually constants.

We are assuming here that $X$ has a continuous distribution. We will look shortly at examples based on discrete distributions.

**If** $g(\underline{X},\theta)$ **is monotonic increasing in** $\theta$**, then:**

$$g(\underline{X},\theta) < g_2 \Leftrightarrow \theta < \theta_2 \quad \text{for some number } \theta_2$$

$$g_1 < g(\underline{X},\theta) \Leftrightarrow \theta_1 < \theta \quad \text{for some number } \theta_1$$

**and if** $g(\underline{X},\theta)$ **is monotonic decreasing in** $\theta$**, then:**

$$g(\underline{X},\theta) < g_2 \Leftrightarrow \theta_1 < \theta$$

$$g_1 < g(\underline{X},\theta) \Leftrightarrow \theta < \theta_2$$

**resulting in** $(\theta_1,\theta_2)$ **being a 95% confidence interval for** $\theta$ **.**

**Fortunately, in most practical situations such quantities** $g(\underline{X},\theta)$ **do exist, although an approximation to the method is needed for the binomial and Poisson cases.**

With prediction intervals, we are predicting a single future value from the distribution. Since we already have a sample of values $(X_1, \dots, X_n)$ from this distribution, we'll call this new predicted value $X_{n+1}$.

**A similar approach can be used for prediction intervals. In the example above, of sampling from a normal distribution with known variance, $\bar{X} - X_{n+1}$ has a distribution that does not depend on $\mu$, and in fact:**

$$\frac{\bar{X} - X_{n+1}}{\sigma\sqrt{1 + 1/n}} \sim N(0,1)$$

The predicted value comes from a normal distribution, $X_{n+1} \sim N(\mu, \sigma^2)$. The Central Limit Theorem tells us that for samples from a normal distribution, $\bar{X} \sim N(\mu, \sigma^2/n)$. Hence, using the linear combination of normal distributions result from Chapter 4:

$$\bar{X} - X_{n+1} \sim N(\mu - \mu, \sigma^2/n + \sigma^2) = N(0, \sigma^2(1/n + 1))$$

Standardising this gives the result above.

**The previous derivations therefore give prediction intervals for $X_{n+1}$ if we replace $\sigma/\sqrt{n}$ with $\sigma\sqrt{1 + 1/n}$ : a 95% prediction interval for the random sample of size 20 above is:**

$$\bar{X} \pm 1.96\sigma\sqrt{1 + 1/20} = 62.75 \pm 20.08$$

A less formal way to consider this is as follows. The predicted value comes from a $N(\mu, \sigma^2)$ distribution. Since $P(-1.96 < X < 1.96) = 0.95$, we know that 95% of the values from that distribution lies between $\mu \pm 1.96\sigma$.

However, we do not know the true value of $\mu$ but a 95% confidence interval for it is given by $\bar{X} \pm 1.96\sigma/\sqrt{n}$. Putting these two together, a 95% confidence interval for a predicted value $X_{n+1}$ is:

$$\bar{X} \pm \left(1.96\sigma/\sqrt{n} + 1.96\sigma\right) = \bar{X} \pm 1.96\sigma\sqrt{\frac{1}{n} + 1}$$

## Question

The average IQ of a random sample of 50 university students is found to be 132. Calculate a symmetrical 99% prediction interval for the average IQ of university students, assuming that IQs are normally distributed. It is known from previous studies that the standard deviation of IQs among students is approximately 20.

## Solution

Since the distribution is normal, we use $\dfrac{\overline{X}-X_{n+1}}{\sigma\sqrt{1+1/n}} \sim N(0,1)$, when $\sigma$ is known.

From the *Tables* we know that $0.99 = P(-2.5758 < Z < 2.5758)$, so:

$$0.99 = P(-2.5758 < \frac{\overline{X}-X_{n+1}}{\sigma\sqrt{1+1/n}} < 2.5758)$$

Rearranging to obtain limits for $X_{n+1}$:

$$0.99 = P(\overline{X} - 2.5758\sigma\sqrt{1+1/n} < X_{n+1} < \overline{X} + 2.5758\sigma\sqrt{1+1/n})$$

Using $n = 50$, $\sigma = 20$ and $\overline{X} = 132$ from the question, we obtain the interval $132 \pm 52.03$, or $(80.0, 184.0)$.

So a symmetrical 99% prediction interval for the average IQ is $(80.0, 184.0)$.

## 3.3 Prediction interval for normal distribution

We've already seen that:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

**Replacing $\mu$ with $X_{n+1}$ and adjusting the denominator produces a pivotal quantity with the same distribution:**

$$\frac{\bar{X}-X_{n+1}}{S\sqrt{1+1/n}} \sim t_{n-1}$$

**A prediction interval for $X_{n+1}$ can therefore take the form:**

$$\bar{X} \pm t_{0.025,n-1}\, S\sqrt{1+1/n}$$

### Question

The heights of 10-year-old children are normally distributed. The heights of a random sample of five children (in cm) are: 124cm, 122cm, 130cm, 125cm and 132cm.

Calculate a 90% confidence interval for the predicted height of a 10-year-old child based on these data values.

### Solution

Since the sample comes from a normal distribution, we know that $\frac{\bar{X}-X_{n+1}}{S\sqrt{1+1/n}}$ has a $t_{n-1}$ distribution, where $S^2$ is the sample variance.

From the *Tables*, we find that $t_{0.05,4}=2.132$, *ie* $0.90=P(-2.132<t_4<2.132)$. So:

$$0.90=P(-2.132<\frac{\bar{X}-X_{n+1}}{S\sqrt{1+1/n}}<2.132)$$

Rearranging the inequality to isolate $X_{n+1}$ gives:

$$0.90=P(\bar{X}-2.132S\sqrt{1+1/n}<X_{n+1}<\bar{X}+2.132S\sqrt{1+1/n})$$

For this sample, we have $n=5$, $\bar{x}=126.6$, and $s^2=17.8$. Using these values gives a 90% prediction interval of:

$$(116.7,\ 136.5)$$

**There is no simple function for calculating prediction intervals for fitted distributions in R. Prediction intervals can either be calculated by implementing the above formula from scratch or alternatively by leveraging the functionality in R for calculating prediction intervals for linear regressions (by regressing on a constant):**

```
# create random sample

      set.seed(23)
      x<- rnorm(10) # 10 observations in sample

# calculate confidence and prediction intervals from scratch

      mu <- mean(x) # sample mean
      sigma <- sqrt(var(x)) # square root of sample variance

      confidence_interval <- c(mu + sigma * sqrt(1/10) * qt(0.025,9),
            mu + sigma * sqrt(1/10) * qt(0.975,9))

      prediction_interval <- c(mu + sigma * sqrt(1+1/10) * qt(0.025,9),
            mu + sigma * sqrt(1+1/10) * qt(0.975,9))

# calculate confidence and prediction intervals using linear regression
functionality (lm)

# data.frame(1) is just dummy data in formulae below

      predict(lm(x~1),data.frame(1),interval = "confidence")
      predict(lm(x~1),data.frame(1),interval = "prediction")
```

The linear regression approach is covered in Chapter 12.

# Chapter 9 Summary

## Confidence intervals

A confidence interval gives us a range of values in which we believe the true parameter value lies, together with an associated probability. There are a number of different situations for which we can find confidence intervals.

For a single sample from a normal distribution:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \sigma^2 \text{ known} \qquad\qquad \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \quad \sigma^2 \text{ unknown}$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

For samples from two independent normal distributions:

$$\frac{\left(\bar{X}_1-\bar{X}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1) \qquad\qquad \frac{\left(\bar{X}_1-\bar{X}_2\right)-\left(\mu_1-\mu_2\right)}{S_p\sqrt{1/n_1+1/n_2}} \sim t_{n_1+n_2-2}$$

$\sigma^2$ known

$\sigma^2$ unknown
Assuming equal variances

where:

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

To compare the variances of two independent normal populations:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

For a sample from a binomial distribution:

$$\frac{\hat{p}-p}{\sqrt{\hat{p}\hat{q}/n}} \sim N(0,1) \quad \text{or} \quad \frac{X-np}{\sqrt{n\hat{p}\hat{q}}} \sim N(0,1) \quad \text{(approximately)}$$

For samples from two independent binomial distributions:

$$\frac{\left(\hat{p}_1-\hat{p}_2\right)-(p_1-p_2)}{\sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1}+\dfrac{\hat{p}_2\hat{q}_2}{n_2}}} \sim N(0,1) \quad \text{(approximately)} \quad \text{where} \quad \hat{p}_1=\frac{X_1}{n_1}, \hat{p}_2=\frac{X_2}{n_2}$$

For a sample from a Poisson distribution:

$$\frac{\hat{\lambda}-\lambda}{\sqrt{\hat{\lambda}/n}} \sim N(0,1) \quad \text{or} \quad \frac{\sum X - n\lambda}{\sqrt{n\hat{\lambda}}} \sim N(0,1) \quad \text{(approximately)}$$

For samples from two independent Poisson distributions:

$$\frac{(\hat{\lambda}_1-\hat{\lambda}_2)-(\lambda_1-\lambda_2)}{\sqrt{\dfrac{\hat{\lambda}_1}{n_1}+\dfrac{\hat{\lambda}_2}{n_2}}} \sim N(0,1) \quad \text{(approximately)} \quad \text{where} \quad \hat{\lambda}_1=\overline{X}_1, \hat{\lambda}_2=\overline{X}_2$$

General confidence intervals for parameters can be found, using the pivotal method, and the formulae given above.

For paired data we subtract the paired values to come up with a new variable, $D$ , and then follow one of the other standard confidence interval calculations:

$$\frac{\overline{X}_D-\mu_D}{S_D/\sqrt{n}} \sim t_{n-1} \quad \sigma_D^2 \text{ unknown}$$

## Prediction Intervals

A prediction interval gives us a range of values for a future predicted value, together with an associated probability.

For a single sample from a normal distribution:

$$\frac{\overline{X}-X_{n+1}}{\sigma\sqrt{1+1/n}} \sim N(0,1) \quad \sigma^2 \text{ known} \qquad \frac{\overline{X}-X_{n+1}}{S\sqrt{1+1/n}} \sim t_{n-1} \quad \sigma^2 \text{ unknown}$$

## 1.3 One-sided and two-sided tests

In a test of whether smoking reduces life expectancies, the hypotheses are:

$H_0$ : smoking makes no difference to life expectancy

$H_1$ : smoking reduces life expectancy

This is an example of a one-sided test, since we are only considering the possibility of a reduction in life expectancy, *ie* a change in one direction. However we could have specified the hypotheses as follows:

$H_0$ : smoking makes no difference to life expectancy

$H_1$ : smoking affects life expectancy

This is a two-sided test since the alternative hypothesis considers the possibility of a change in either direction, *ie* an increase or a decrease.

## 1.4 Test statistics

**The actual decision is based on the value of a suitable function of the data, the test statistic. The set of possible values of the test statistic itself divides into two subsets, a region in which the value of the test statistic is judged consistent with $H_0$, and its complement, the critical region (or rejection region), in which the value of the test statistic is judged inconsistent with $H_0$. If the test statistic has a value in the critical region, $H_0$ is rejected. The test statistic (like any statistic) must be such that its distribution is completely specified when the value of the parameter itself is specified (and in particular 'under $H_0$' *ie* when $H_0$ is true).**

In exam questions the test statistic is generally calculated from data given in the question. For details of how to reach a conclusion in practice, see Section 3.1.

## 1.5 Errors

**It is rare for data to enable discrimination with certainty between the two hypotheses. The result of performing a test may be the correct decision, but two kinds of error could arise:**

**Type I error: reject $H_0$ when it is true; and**
**Type II error: fail to reject $H_0$ when it is false.**

**The level of significance of the test, denoted $\alpha$ , 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 $\beta$, is the probability of accepting $H_0$ when it is false. An ideal test would be one which simultaneously minimises $\alpha$ and $\beta$. This ideal however is not attainable in practice.**

## Question

A random variable $X$ is believed to follow an $Exp(\lambda)$ distribution. In order to test the null hypothesis $\mu = 20$ against the alternative hypothesis $\mu = 30$, where $\mu = 1/\lambda$, a single value is observed from the distribution. If this value is less than 28, $H_0$ is not rejected, otherwise $H_0$ is rejected.

Calculate the probabilities of:

(i) a Type I error

(ii) a Type II error.

## Solution

(i) The probability of a Type I error is given by:

$$P(\text{reject } H_0 \text{ when } H_0 \text{ true}) = P\left(X > 28 \text{ when } X \sim Exp\left(1/20\right)\right)$$

$$= 1 - F_X(28) = e^{-28/20} = 0.2466$$

*The CDF of the exponential distribution is given on page 11 of the Tables.*

(ii) The probability of a Type II error is given by:

$$P(\text{do not reject } H_0 \text{ when } H_0 \text{ false}) = P\left(X < 28 \text{ when } X \sim Exp\left(1/30\right)\right)$$

$$= F_X(28) = 1 - e^{-28/30} = 0.6068$$

*In this case we were forced to choose between $H_0 : \mu = 20$ and $H_1 : \mu = 30$. So saying that $H_0$ is false is the same as saying that $\mu = 30$.*

*Since we've only got one value in our sample here, not surprisingly, the probabilities of Type I and Type II errors are quite big.*

The probability of a Type I error is also referred to as the 'size' of the test, which will normally be a small number such as 0.05 (say).

**The power of a test is the probability of rejecting $H_0$ when it is false, so that the power equals $1-\beta$.**

In general, this will be a function of the unknown parameter value.

**For simple hypotheses the power is a single value, but for composite hypotheses it is a function being defined at all points in the alternative hypothesis.**

A test with a high power is said to be 'powerful' as it is very effective at demonstrating a positive result.

## Question

Give an expression in terms of $\mu$ for the power of the test in the question on the previous page. Comment on how the power is affected by the value of $\mu$.

## Solution

The power is the probability of rejecting $H_0$ when the true value of the parameter $\mu$ is some value other than $\mu = 20$. In terms of $\mu$ this is:

$$P\left(X > 28 \mid X \sim Exp\left(1/\mu\right)\right) = 1 - F_X(28) = e^{-28/\mu}$$

If $\mu$ is large (1,000, say), then the power will be close to 1, since the test will reject $H_0 : \mu = 20$ very easily. Conversely if $\mu$ is small (10, say), then the power will be close to 0, since the test will not reject $H_0 : \mu = 20$ very easily.

**Type I and II errors can also arise in the context of binary classification, a common situation in healthcare as well as in machine learning contexts. Here, rather than gathering a data sample consisting of multiple observations to assess whether a (population-level) hypothesis holds, a decision is required for each individual observation.**

**In a medical context, the classification is into healthy and diseased based on a binary test result. In these contexts:**

**A Type I error, known as a false positive, occurs when a healthy individual receives a positive test result; and**

**A Type II error, known as a false negative, occurs when a diseased individual tests negative for the disease.**

The equivalent null hypothesis in this case is that the individual is healthy, and we are carrying out a test to ascertain whether this is the case. If the null hypothesis is true (*ie* the individual is actually healthy) but the test is positive (indicating that the individual has the disease), then we would be rejecting a true hypothesis and making a Type I error.

If the null hypothesis is false (*ie* the individual is sick) but the test is negative (indicating that the individual does not have the disease), then we would be failing to reject a false hypothesis and making a Type II error.

The table below shows all the possible outcomes from a medical test result:

| | | Test result predicts patient as having disease | |
|---|---|---|---|
| | | YES | NO |
| Patient actually has disease | YES | True positive (TP) | False negative (FN)<br>Type II error |
| | NO | False positive (FP)<br>Type I error | True negative (TN) |

**The probability of a diseased individual testing positive for the disease** (*ie* a true positive rate)**, is the *sensitivity* of the test:**

$$\begin{aligned}\text{Sensitivity} &= \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false negatives}}\\ &= \frac{\text{Number of true positives}}{\text{Total number of people with the disease}}\\ &= P(\text{positive test} \mid \text{individual has the disease})\\ &= 1 - P(\text{negative test} \mid \text{individual has the disease})\\ &= 1 - P(\text{Type II error})\\ &= \text{Power of the test}\end{aligned}$$

**The probability of a healthy individual testing negative** (*ie* a true negative rate)**, which is 1 minus the probability of a false positive, is called the *specificity* of the test.**

$$\begin{aligned}\text{Specificity} &= \frac{\text{Number of true negatives}}{\text{Number of true negatives} + \text{Number of false positives}}\\ &= \frac{\text{Number of true negatives}}{\text{Total number of people who do not have the disease}}\\ &= P(\text{negative test} \mid \text{individual does not have the disease})\\ &= 1 - P(\text{positive test} \mid \text{individual does not have the disease})\\ &= 1 - P(\text{Type I error})\end{aligned}$$

## Question

A short screening test has just been developed for depression. An independent blind comparison was made with a gold-standard test for diagnosis of depression among 200 psychiatric outpatients.

Among the 50 outpatients found to be depressed according to the gold-standard test, 35 patients tested positive under the new short test. Among 150 patients found not to be depressed according to the gold-standard test, 30 patients tested positive under the new short test.

Calculate the sensitivity and specificity of the short screening test, assuming that the gold-standard test correctly classifies each individual.

## Solution

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Total number of people with depression}} = \frac{35}{50} = 70\%$$

$$\text{Specificity} = \frac{\text{Number of true negatives}}{\text{Total number of individuals without depression}} = \frac{150-30}{150} = 80\%$$

Examples of binary classifications in machine learning contexts include:

- classifying emails according to whether they are spam
- assessing whether claims received by an insurance company are fraudulent.

One method of making such predictions is to use a generalised linear model with a binomial distribution. We'll cover this in Chapter 13. Other methods are covered in Subject CS2.

**Although the contexts are different in important respects (*eg* hypothesis testing seeks to make inferences, classifiers seek to make predictions; the true state is usually known with certainty, at least for a training set, in classification problems), understanding the trade-offs of minimising Type I versus Type II errors play an important role in test selection in both cases.**

For example, in the case of using a smear test to identify cervical cancer, it is vital to have a test with a high sensitivity (currently it's 86%-100%), as cervical cancer is a serious but treatable condition if caught early. However, smear tests have a much lower specificity (currently 30%-87%), which means that a high proportion of women with a positive cervical smear test who go on to have further investigation subsequently find that there is no cause for concern. This is considered a small price to pay compared to the alternative.

**R can calculate the power of a one-sample *t* test (covered in Section 3.1) using the function:**

```
power.t.test
```

# 2 Classical testing, significance and *p*-values

## 2.1 'Best' tests

**The classical approach to finding a 'good' test (called the Neyman-Pearson theory) fixes the value of $\alpha$, *ie* the level of significance required and then tries to find such a test for which the other error probability, $\beta$, is as small as possible for every value of the parameter specified by the alternative hypothesis. This can also be described as finding the 'most powerful' test.**

**The key result in the search for such a test is the Neyman-Pearson lemma, which provides the 'best' test (smallest $\beta$) in the case of two simple hypotheses. For a given level, the critical region (and in fact the test statistic) for the best test is determined by setting an upper bound on the likelihood ratio $L_0/L_1$, where $L_0$ and $L_1$ are the likelihood functions of the data under $H_0$ and $H_1$ respectively.**

### The Neyman-Pearson lemma

**Formally, if $C$ is a critical region of size $\alpha$ and there exists a constant $k$ such that $L_0/L_1 \leq k$ inside $C$ and $L_0/L_1 \geq k$ outside $C$, then $C$ is a most powerful critical region of size $\alpha$ for testing the simple hypothesis $\theta = \theta_0$ against the simple alternative hypothesis $\theta = \theta_1$.**

So a Neyman-Pearson test rejects $H_0$ if:

$$\frac{\text{Likelihood under } H_0}{\text{Likelihood under } H_1} < \text{critical value}$$

### Question

A random variable $X$ is believed to follow an $Exp(\lambda)$ distribution. In order to test the null hypothesis $\mu = 20$ against the alternative hypothesis $\mu = 30$, where $\mu = 1/\lambda$, a single value is observed from the distribution. If this value is less than 28, $H_0$ is not rejected, otherwise $H_0$ is rejected.

Show that this is a Neyman-Pearson test.

### Solution

Given a single value from an exponential distribution, the Neyman-Pearson criterion is 'reject $H_0$ if $L_0/L_1 <$ critical value'. Using the null and alternative hypotheses, the test becomes:

$$\frac{\frac{1}{20}e^{-\frac{x}{20}}}{\frac{1}{30}e^{-\frac{x}{30}}} < \text{constant}$$

## Chapter 10 Summary

Statistical tests can be used to test assertions about populations.

The process of statistical testing involves setting up a null hypothesis and an alternative hypothesis, calculating a test statistic and using this to determine a *p*-value.

The probability of a Type I error is the probability of rejecting $H_0$ when it is true. This is also called the size (or level) of the test. The probability of a Type II error is the probability of not rejecting $H_0$ when it is false. The power of a test is the probability of rejecting $H_0$ when it is false.

Errors can also occur in the context of binary classifications, for example when an individual is classified as testing positive or negative for a particular disease. The null hypothesis is that the individual does not have the disease. A Type I error is a false positive and a Type II error is a false negative. The sensitivity of this test is the true positive rate (which is $1-P(\text{Type II error})=\text{power of the test}$). The specificity of this test is the true negative rate (which is $1-P(\text{Type I error})$).

The 'best' test can be found using the likelihood ratio criterion. This leads to the tests detailed overleaf.

The test for two normal means (unknown variances) requires that the variances are the same and uses the pooled sample variance:

$$s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

$\chi^2$ tests can be carried out to test for goodness of fit or to test whether two factors are independent (using contingency tables).

The statistic is $\sum\frac{(O_i-E_i)^2}{E_i}$.

To find the number of degrees of freedom for the goodness of fit test, take the number of cells, subtract 1 if the total of the observed figures has been used in the calculation of the expected numbers (which is usually the case), and then subtract the number of parameters estimated.

To find the number of degrees of freedom for a contingency table calculate $(r-1)(c-1)$. If the expected numbers in some cells are small, these should be grouped. One degree of

## One-sample normal distribution

$$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \quad \sigma^2 \text{ known} \qquad \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1} \quad \sigma^2 \text{ unknown}$$

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

## Two-sample normal distribution

$$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1) \qquad \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_p\sqrt{1/n_1+1/n_2}} \sim t_{n_1+n_2-2}$$

$\sigma^2$ known $\qquad$ $\sigma^2$ unknown

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

## One-sample binomial

$$\frac{\hat{p}-p_0}{\sqrt{p_0q_0/n}} \mathrel{\dot{\sim}} N(0,1) \quad \text{or} \quad \frac{X-np_0}{\sqrt{np_0q_0}} \mathrel{\dot{\sim}} N(0,1)$$

with continuity correction

## Two-sample binomial

$$\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{p}\hat{q}}{n_1}+\frac{\hat{p}\hat{q}}{n_2}}} \mathrel{\dot{\sim}} N(0,1) \qquad \hat{p}=\frac{x_1+x_2}{n_1+n_2}$$

is the overall sample proportion

## One-sample Poisson

$$\frac{\bar{X}-\lambda_0}{\sqrt{\lambda_0/n}} \mathrel{\dot{\sim}} N(0,1) \quad \text{or} \quad \frac{\sum X-n\lambda_0}{\sqrt{n\lambda_0}} \mathrel{\dot{\sim}} N(0,1)$$

with continuity correction

## Two-sample Poisson

$$\frac{(\hat{\lambda}_1-\hat{\lambda}_2)-(\lambda_1-\lambda_2)}{\sqrt{\frac{\hat{\lambda}}{n_1}+\frac{\hat{\lambda}}{n_2}}} \mathrel{\dot{\sim}} N(0,1) \qquad \hat{\lambda}=\frac{n_1\hat{\lambda}_1+n_2\hat{\lambda}_2}{n_1+n_2}$$

is the overall sample mean

## Scaled deviance

The scaled deviance for a particular model $M$ is defined as:

$$SD_M = 2\left(\ell_S - \ell_M\right)$$

**The deviance for the current model, $D_M$, is defined such that:**

$$\textbf{scaled deviance} = \frac{D_M}{\varphi}$$

Remember that $\varphi$ is a scale parameter, so it seems sensible that it should be used to connect the deviance with the scaled deviance. For the Poisson and exponential distributions, $\varphi = 1$, so the scaled deviance and the deviance are identical.

**The smaller the deviance, the better the model from the point of view of model fit.**

However, there will be a trade-off here. A model with many parameters will fit the data well. However a model with too many parameters will be difficult and complex to build, and will not necessarily lead to better prediction in the future. It is possible for models to be 'over-parameterised', *ie* factors are included that lead to a slightly, but not significantly, better fit. When choosing linear models, we will usually need to strike a balance between a model with too few parameters (which will not take account of factors that have a substantial impact on the data, and will therefore not be sensitive enough) and one with too many parameters (which will be too sensitive to factors that really do not have much effect on the results). We use the principle of parsimony here, *ie* we choose the simplest model that does the job.

**This can be illustrated by considering the case when the data are normally distributed.**

**In this case, the log-likelihood for a sample of size $n$ is:**

$$\begin{aligned}\ell(y;\theta,\varphi) &= \sum_{i=1}^{n} \log f_Y(y_i;\theta_i,\varphi) \\ &= -\frac{n}{2}\log 2\pi\sigma^2 - \sum_{i=1}^{n}\frac{(y_i-\theta_i)^2}{2\sigma^2}\end{aligned}$$

The likelihood function for a random sample of size $n$ is $f(y_1)f(y_2)\ldots f(y_n)$. When we take logs, we add the logs of the individual PDF. Recall that for the normal distribution the natural parameter is the mean, *ie* $\theta_i = \mu_i$.

**For the saturated model, the parameter $\theta_i$ is estimated by $y_i$, and so the second term disappears. Thus, the scaled deviance (twice the difference between the values of the log-likelihood under the current and saturated models) is**

$$\sum_{i=1}^{n}\frac{(y_i-\hat{\theta}_i)^2}{\sigma^2}$$

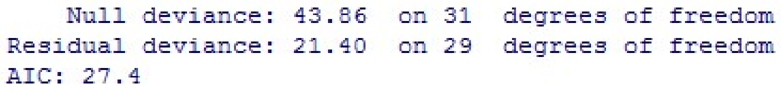
**where $\hat{\theta}_i$ is the fitted value for the current model.**

**The deviance (remembering that the scale parameter $\varphi = \sigma^2$), is the well-known residual sum of squares:**

$$\sum_{i=1}^{n}(y_i - \hat{\theta}_i)^2$$

This is why the deviance is defined with a factor of two in it, so that for the normal model the deviance is equal to the residual sum of squares that we met in linear regression.

**The residual deviance (*ie* the deviance after all the covariates have been included) is displayed as part of the results from `summary(model)`. For example:**



```
    Null deviance: 43.86  on 31  degrees of freedom
Residual deviance: 21.40  on 29  degrees of freedom
AIC: 27.4
```

**In R we can obtain a breakdown of how the deviance is reduced by each covariate added sequentially by using `anova(model)`. However, unlike for linear regression, this command does not automatically carry out a test.**

Also recall that the smaller the residual (left over) deviance, the better the fit of the model.

The residual deviance outputted by the `glm()` function is a measure of fit, similar to the scaled deviance and deviance defined earlier. However, this output won't necessarily match the scaled deviance or deviance calculated from first principles using the formulae in this section.

## 5.5 Using scaled deviance and Akaike's Information Criterion to choose between models

**Adding more covariates will always improve the fit and thus decrease the deviance, however we need to determine whether adding a particular covariate leads to a significant decrease in the deviance.**

**For normally distributed data, the scaled deviance has a $\chi^2$ distribution. Since the scale parameter for the normal $\varphi = \sigma^2$ must be estimated, we compare models by taking ratios of sum-of-squares and using $F$ tests (as in the analysis of variance for linear regression models).**

We covered this in Section 4.3 from the previous chapter.

**Thus, if we want to decide if Model 2 (which has $p+q$ parameters and scaled deviance $S_2$) is a significant improvement over Model 1 (which has $p$ parameters and scaled deviance $S_1$), we see if $\dfrac{(S_1 - S_2)/q}{S_2/(n-(p+q))}$ is greater than the 5% value for the $F_{q,n-p-q}$ distribution.**

**The code for comparing two normally distributed models, `model1` and `model2`, in R is:**

```
anova(model1, model2, test="F")
```

**In the case of data that are not normally distributed, the scale parameter may be known (for example, for the Poisson distribution $\varphi = 1$), and the deviance is only asymptotically a $\chi^2$ distribution. For these reasons, the common procedure is to compare two models by looking at the difference in the scaled deviance and comparing with a $\chi^2$ distribution.**

Since the distributions are only asymptotically normal, the $F$ test will not be very accurate. We get a better result by comparing two approximate $\chi^2$ distributions.

To be more precise, it's the absolute difference between the scaled deviances that is compared with $\chi^2$.

**Thus, if we want to decide if Model 2 (which has $p+q$ parameters and scaled deviance $S_2$) is a significant improvement over Model 1 (which has $p$ parameters and scaled deviance $S_1$), we see if $S_1 - S_2$ is greater than the 5% value for the $\chi^2_q$ distribution.**

Recall that we subtract one degree of freedom for each extra parameter introduced. So it's the difference between $p$ and $p+q$ that matters.

Since $\chi^2_p + \chi^2_q \sim \chi^2_{p+q}$ (provided the random variables are independent), it makes sense to say that the difference in the scaled deviances has a $\chi^2_q$ distribution.

What we are trying to do here is to decide whether the added complexity results in significant additional accuracy. If not, then it would be preferable to use the model with fewer parameters.

Alternatively, we could express this test in terms of the log-likelihood functions. If we let $\ell_p$ and $\ell_{p+q}$ denote the log-likelihoods of the models with $p$ and $p+q$ parameters respectively, then the test statistic can be written as:

$$\begin{aligned} S_1 - S_2 &= 2\left(\ell_S - \ell_p\right) - 2\left(\ell_S - \ell_{p+q}\right) \\ &= -2\left(\ell_p - \ell_{p+q}\right) \end{aligned}$$

This is the format given on page 23 of the *Tables* and will be used in Subject CS2 to compare Cox regression models.

## Question

Explain why the test statistic will always be positive.

## Solution

As we have mentioned before, adding more parameters will improve the fit of the model to the data. Therefore we would expect the value of the likelihood function to be larger for models with more parameters. Hence, $\ell_{p+q} > \ell_p$ and so the statistic will be positive.

**The code for comparing these two** (non-normally distributed) **models,** `model1` **and** `model2`**, in R is:**

```
      anova(model1, model2, test="Chi")
```

**A very important point is that this method of comparison can only be used for *nested* models. In other words, Model 1 must be a submodel of Model 2. Thus, we can compare two models for which the distribution of the data and the link function are the same, but the linear predictor has one extra parameter in Model 2. For example $\beta_0 + \beta_1 x$ and $\beta_0 + \beta_1 x + \beta_2 x^2$. But we could not compare in this way if the distribution of the data or the link function are different, or, for example, when the linear predictors are $\beta_0 + \beta_1 x + \beta_2 x^2$ and $\beta_0 + \beta_3 \log x$. It should be clear that we *can* gauge the importance of factors by examining the scaled deviances, but we cannot use the testing procedure outlined above.**

In the first case, the difference between the models is $\beta_2 x^2$, and so a significant difference between the models tells us that the quadratic term should be included. In the second case, the difference between the models is $\beta_3 \log x - \beta_2 x^2$, and so a significant difference doesn't tell us *which* parameter is significant.

**An alternative method of comparing models is to use Akaike's Information Criterion (AIC). Since the deviance will always decrease as more covariates are added to the model, there will always be a tendency to add more covariates. However this will increase the complexity of the model which is generally considered to be undesirable. To take account of the undesirability of increased complexity, computer packages will often quote the AIC, which is a penalised log-likelihood:**

$$\text{AIC} = -2 \times \log L_M + 2 \times \text{number of parameters}$$

**where $\log L_M$ is the log-likelihood of the model under consideration.**

**When comparing two models, the smaller the AIC, the better the fit. So if the change in deviance is more than twice the change in the number of parameters then it would give a smaller AIC.**

This is approximately equivalent to checking whether the difference in deviance is greater than the 5% value of the $\chi^2$ distribution for degrees of freedom between 5 and 15. However, it has the added advantage of being a simple way to compare GLMs without formal testing. This is similar to comparing the adjusted $R^2$ for multiple linear regression models in the previous chapter and hence is displayed as part of the output of a computer fitted GLM.

**In R the AIC is displayed as part of the results from** `summary(model)`**.**

An example of this is given in the R box at the end of Section 5.4.

## 5.6 The process of selecting explanatory variables

**As for multiple linear regression the process of selecting the optimal set of covariates for a GLM is not always easy. Again, we could use one of the two following approaches:**

**(1) Forward selection. Add the covariate that reduces the AIC the most or causes a significant decrease in the deviance. Continue in this way until adding any more causes the AIC to rise or does not lead to a significant improvement in the deviance. Note we should start with main effects before interaction terms and linear terms before polynomial.**

Suppose we are modelling the number of claims on a motor insurance portfolio and we have data on the driver's age, sex and vehicle group. We would start with the null model (*ie* a single constant equal to the sample mean). Then we would try each of single covariate models (linear function of age or the factors sex or vehicle group) to see which produces the most significant improvement in a $\chi^2$ test or reduces the AIC the most. Suppose this was sex. Then we would try adding a second covariate (linear function of age or the factor vehicle group). Suppose this was age. Then we would try adding the third covariate (vehicle group). We might then try a quadratic function of the variable age (and maybe higher powers) or each of 2 term interactions (eg sex*age or sex*group or age*group). Finally we would try the 3 term interaction (*ie* sex*age*group).

**(2) Backward selection. Start by adding all available covariates and interactions. Then remove covariates one by one starting with the least significant until the AIC reaches a minimum or there is no significant improvement in the deviance, and all the remaining covariates have a statistically significant impact on the response.**

So with the last example we would start with the 3 term interaction sex*age*group and look at which parameter has the largest *p*-value (in a test of it being zero) and remove that. We should see a significant improvement in a $\chi^2$ test and the AIC should fall. Then we remove the next parameter with the largest *p*-value and so on.

The Core Reading uses R to demonstrate this procedure. Whilst this will be covered in the CS1 PBOR, it's important to understand the process here.

## Example

**We demonstrate both of these methods in R using a binomial model on the `mtcars` dataset from the MASS package to determine whether a car has a V engine or an S engine (`vs`) using weight in 1000 lbs (`wt`) and engine displacement in cubic inches (`disp`) as covariates.**

**Forward selection**

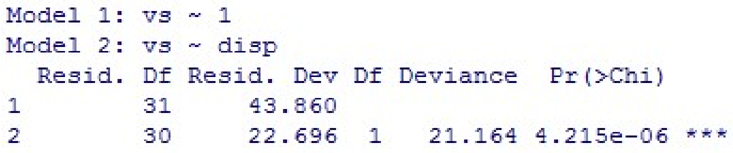
**Starting with the null model:**

```
model0 <- glm(vs ~ 1, data=mtcars, family=binomial)
```

The AIC of this model (which would be displayed using `summary(model0)`) is 45.86.

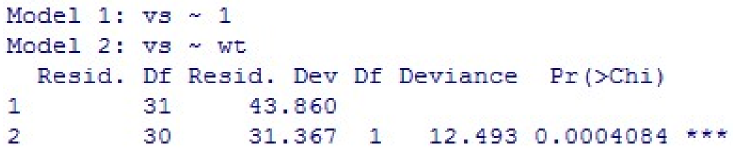
**We have to choose whether we add `disp` or `wt` first. We try each and see which has the greatest improvement in the deviance.**

```
model1 <- update(model0, ~.+ disp)
anova(model0, model1, test="Chi")
```



```
Model 1: vs ~ 1
Model 2: vs ~ disp
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        31     43.860
2        30     22.696  1   21.164 4.215e-06 ***
```

```
      model2 <- update(model0, ~.+ wt)
      anova(model0, model2, test="Chi")
```



```
Model 1: vs ~ 1
Model 2: vs ~ wt
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        31     43.860
2        30     31.367  1   12.493 0.0004084 ***
```

**So we can see that `disp` has produced the more significant result – so we add that covariate first.**

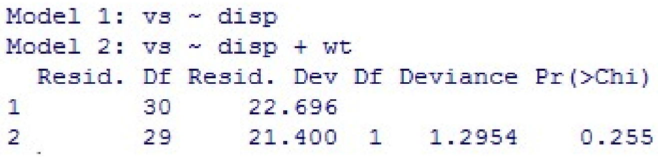
R always calls the models we are comparing 'Model 1' and 'Model 2', irrespective of how we have named them. This can lead to confusion if we are not careful.

The AIC of model 1 (adding `disp`) is 26.7 whereas the AIC of model 2 (adding `wt`) is 35.37. Therefore adding disp reduces the AIC more from model 0's value of 45.86.

**Let us now see if adding `wt` to `disp` produces a significant improvement:**

```
      model3 <- update(model1, ~.+ wt)

      anova(model1, model3, test="Chi")
```



```
Model 1: vs ~ disp
Model 2: vs ~ disp + wt
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        30     22.696
2        29     21.400  1   1.2954    0.255
```

**This has not led to a significant improvement in the deviance so we would not add `wt` (and therefore we definitely would not add an interaction term between `disp` and `wt`).**

The AIC of model 3 (adding `wt`) is 27.4 which is worse than model 1's AIC of 26.7. Therefore we would not add it.

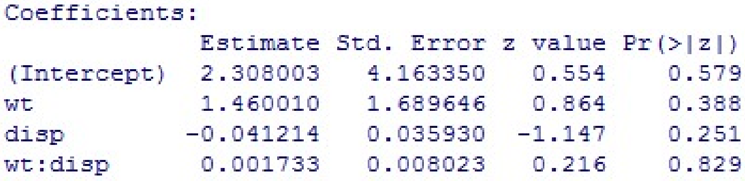
**Incidentally the AIC for models 0, 1, 2, 3 are 45.86, 26.7, 35.37 and 27.4. So using these would have given the same results (as Model 1 produces a smaller AIC than Model 2, and then Model 3 increases the AIC and so we would not have selected it).**

**<u>Backward selection</u>**

**Starting with all the possibilities:**

```
      modelA <- glm(vs ~ wt * disp, data=mtcars, family=binomial)
```

**The output is:**



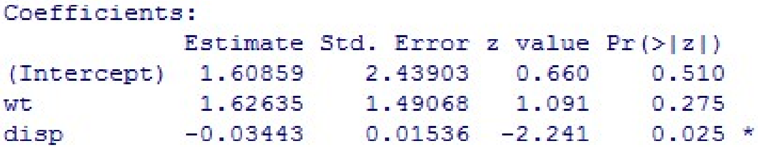
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.308003   4.163350   0.554    0.579
wt           1.460010   1.689646   0.864    0.388
disp        -0.041214   0.035930  -1.147    0.251
wt:disp      0.001733   0.008023   0.216    0.829
```

**None of these covariates are significant.**

The parameter of the interaction term has the highest *p*-value (0.829), and so is most likely to be zero.

**We first remove the interaction term `wt:disp`, as this is the least significant parameter:**

```
      modelB <- update(model1, ~.-wt:disp)
```



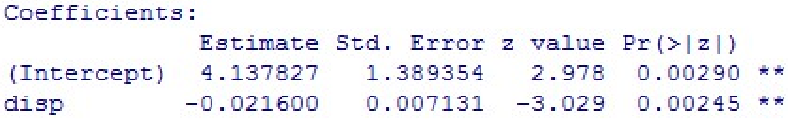
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.60859    2.43903   0.660    0.510
wt           1.62635    1.49068   1.091    0.275
disp        -0.03443    0.01536  -2.241    0.025 *
```

**The AIC has fallen from 29.361 to 27.4.**

Alternatively, carrying out a $\chi^2$ test using `anova(modelA, modelB, test="Chi")` would show that there is no significant difference between the models (*p*-value of 0.8417) and therefore we are correct to remove the interaction term between `wt` and `disp`.

**The `wt` term is not significant so removing that:**

```
      modelC <- update(modelB, ~.-wt)
```



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.137827   1.389354   2.978  0.00290 **
disp        -0.021600   0.007131  -3.029  0.00245 **
```

**Both of these coefficients are significant and the AIC has fallen from 27.4 to 26.696.**

Alternatively, carrying out a $\chi^2$ test using `anova(modelB, modelC, test="Chi")` would show that there is no significant difference between the models (*p*-value of 0.255) and therefore we are correct to remove the `wt` covariate.

**We would stop at this model. If we remove the `disp` term (to give the null model), the AIC *increases* to 45.86.**

Alternatively, carrying out a $\chi^2$ test between these two models would show a very significant difference (*p*-value of less than 0.001) and therefore we should not remove the `disp` covariate.

**We can see that both forward and backward selection lead to the same model being chosen in this case.**

## 5.7 Estimating the response variable

**Once we have obtained our model and its estimates, we are then able to calculate the value of the linear predictor, $\eta$, and by using the inverse of the link function we can calculate our estimate of the response variable $\hat{\mu} = g^{-1}(\hat{\eta})$.**

Substituting the estimated parameters into the linear predictor gives the estimated value of the linear predictor for different individuals. The link function links the linear predictor to the mean of the distribution. Hence we can obtain an estimate for the mean of the distribution of $Y$ for that individual.

Let's now return to the Core Reading example on page 45.

**Suppose, we wish to estimate the probability of having a V engine for a car with weight 2,100 lbs and displacement 180 cubic inches.**

**Using our linear predictor $\beta_0 + \beta_1 \times \text{disp}$ (*ie* `vs ~ disp`), we obtained estimates of $\hat{\beta}_0 = 4.137827$ and $\hat{\beta}_1 = -0.021600$.**

These coefficients displayed as part of the summary output of Model C in the example above.

**Hence, for displacement 180 we have $\hat{\eta} = 4.137827 - 0.021600 \times 180 = 0.24983$. We did not specify the link function so we shall use the canonical binomial link function which is the logit function.**

$$0.24983 = \log\left(\frac{\hat{\mu}}{1-\hat{\mu}}\right) \quad \Rightarrow \quad \hat{\mu} = \frac{e^{0.24983}}{1+e^{0.24983}} = 0.562$$

Recall that the mean for a binomial model is the probability. So the probability of having a V engine for a car with weight 2,100 lbs and displacement 180 cubic inches is 56.2%.

The figure 2,100 does not enter the calculation because we removed the weight covariate.

**In R we can obtain this as follows:**

```
newdata <-data.frame(disp=180)
predict(model,newdata,type="response")
```

# 6 Residuals analysis and assessment of model fit

**Once a possible model has been found it should be checked by looking at the residuals. The residuals are based on the differences between the observed responses, $y$ , and the fitted responses, $\hat{\mu}$. The fitted responses are obtained by applying the inverse of the link function to the linear predictor with the fitted values of the parameters.**

We looked at how we could obtain predicted responses values in the previous section. The fitted values are the predicted $Y$ values for the observed data set, $x$.

**The R code for obtaining the fitted values of a GLM is:**

```
fitted(model)
```

For example, in the actuarial pass rates model detailed on page 6, we could calculate from the model what the pass rate ought to be for students who have attended tutorials, submitted three assignments and scored 60% on the mock exam.

The difference between this theoretical pass rate and the actual pass rate observed for students who match the criteria exactly will give us the residuals.

## Question

Draw up a table showing the differences between the actual and expected values of the truancy rates in the example on page 9.

## Solution

Recall that the expected number of unexplained absences in a year were modelled by:

$\eta = \alpha_i + \beta_j + \gamma x$ where $x = \text{age}$, and $\alpha$ and $\beta$ are as follows:

$$\alpha_{WC} = -2.64 \qquad \alpha_{OC} = -1.14 \qquad \beta_M = -3.26 \qquad \beta_F = -3.54 \qquad \gamma = 0.64$$

where $WC = \text{Within catchment}$, $OC =$ Outside catchment, $M = \text{Male}$, $F = \text{Female}$.

This gives expected values of:

| | | Age last birthday | | | |
|---|---|---|---|---|---|
| | | 8 | 10 | 12 | 14 |
| Within catchment area | Male | 0.46 | 1.65 | 5.93 | 21.33 |
| | Female | 0.35 | 1.25 | 4.48 | 16.12 |
| Outside catchment area | Male | 2.05 | 7.39 | 26.58 | 95.58 |
| | Female | 1.55 | 5.58 | 20.09 | 72.24 |

So the differences between the actual values (given on page 9) and expected values are:

| | | Age last birthday | | | |
|---|---|---|---|---|---|
| | | 8 | 10 | 12 | 14 |
| Within catchment area | Male | 1.34 | 0.35 | 0.37 | –7.23 |
| | Female | 0.15 | 0.35 | 0.52 | 0.08 |
| Outside catchment area | Male | 0.05 | 0.11 | –1.08 | –23.58 |
| | Female | 1.25 | 0.62 | –0.49 | –4.04 |

The procedure here is a natural extension of the way we calculated residuals for linear regression models covered in the previous chapter. However, because of the different distributions used, we need to transform these 'raw' residuals so we are able to interpret them meaningfully.

**There are two kinds of residuals: Pearson and deviance.**

## 6.1 Pearson residuals

**The Pearson residuals are defined as:**

$$\frac{y-\hat{\mu}}{\sqrt{\text{var}(\hat{\mu})}}$$

The $\text{var}(\hat{\mu})$ in the denominator refers to the variance of the response distribution, $\text{var}(Y)$ using the fitted values, $\hat{\mu}$, in the formula. For example, since the variance of the exponential distribution is $\mu^2$, we have $\text{var}(\hat{\mu})=\hat{\mu}^2$ in that case.

**The Pearson residual, which is often used for normally distributed data, has the disadvantage that its distribution is often skewed for non-normal data. This makes the interpretation of residual plots difficult.**

**The R code for obtaining the Pearson residuals is:**

```
residuals(model, type= "pearson")
```

The Pearson residuals returned by R are calculated slightly differently from the definition given in this section. Therefore, this output won't necessarily match the Pearson residuals calculated from first principles using $\dfrac{y-\hat{\mu}}{\sqrt{\text{var}(\hat{\mu})}}$.

If the data come from a normal distribution, then the Pearson residuals will follow the standard normal distribution. By comparing these residuals to a standard normal (*eg* by using a Q-Q plot), we can determine whether the model is a good fit.

However, for non-normal data the Pearson residuals will not follow the standard normal distribution and won't even be symmetrical. This makes it difficult to determine whether the model is a good fit. Hence we will need to use a different type of residual.

## 6.2 Deviance residuals

**Deviance residuals are defined as the product of the sign of $y - \hat{\mu}$ and the square root of the contribution of $y$ to the scaled deviance. Thus, the deviance residual is:**

$$\boldsymbol{sign(y - \hat{\mu})d_i}$$

**where the scaled deviance is $\sum d_i^2$ .**

Recall that:

$$sign(x) = \begin{cases} +1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}$$

**Deviance residuals are usually more likely to be symmetrically distributed and to have approximately normal distributions, and are preferred for actuarial applications.**

**The R code for obtaining the deviance residuals is:**

```
    residuals(model)
```

The deviance residuals returned by R are calculated slightly differently from the definition given in this section. Therefore, this output won't necessarily match the deviance residuals calculated from first principles using the formulae in this section.

We can see that deviance residuals are more likely to be symmetrically distributed by considering the following result: If $\{X_i\}$ is a set of independent normal random variables, then $Y = \sum X_i^2$ will have a $\chi^2$ distribution. Therefore, since $\sum d_i^2$ (*ie* the scaled deviance) is approximately $\chi^2$, it follows that $d_i$ (and also the deviance residual) is likely to be approximately normal.

**Note that for normally distributed data, the Pearson and deviance residuals are identical.**

## Question

Show that, for normally distributed data, the Pearson and deviance residuals are identical.

## Solution

If $Y_i \sim N(\mu_i, \sigma^2)$, then from Section 6.1, the Pearson residuals are:

$$\frac{y_i - \hat{\mu}_i}{\sqrt{\text{var}(\hat{\mu}_i)}} = \frac{y_i - \hat{\mu}_i}{\sigma}$$

In Section 5.4, we saw that the scaled deviance was:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma^2} = \sum_{i=1}^{n} d_i^2$$

So the deviance residuals are given by:

$$sign(y_i - \mu_i)d_i = sign(y_i - \mu_i)\left|\frac{y_i - \mu_i}{\sigma}\right| = \frac{y_i - \mu_i}{\sigma}$$

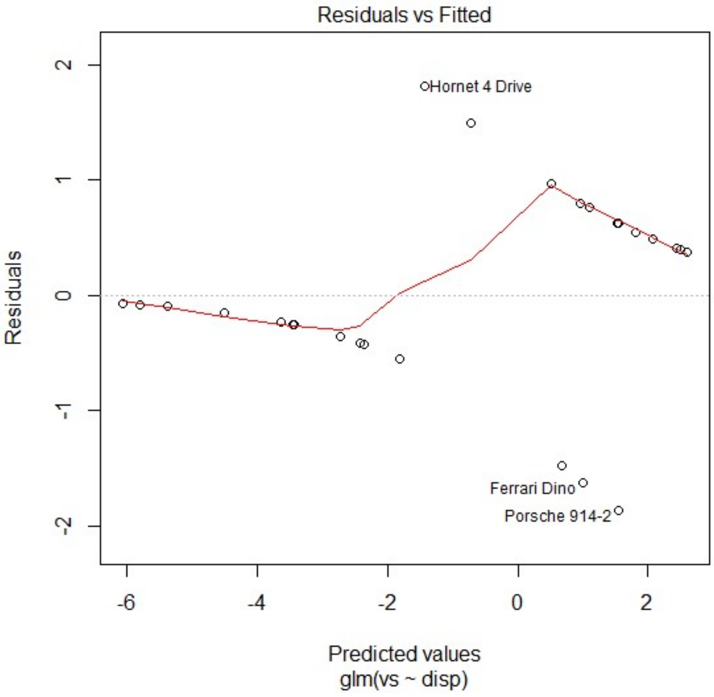
Hence the Pearson residuals and the deviance residuals are the same.

## 6.3 Using residual plots to check the fit

**The assumptions of a GLM require that the residuals should show no patterns. The presence of a pattern implies that something has been missed in the relationship between the predictors and the response. If this is the case, other model specifications should be tried.**

So, in addition to the residuals being symmetrical, we would expect no connection between the residuals and the explanatory covariates. Rather than plotting the residuals against each of the covariates, we could just see if there is a pattern when plotted against the fitted values.

**For our model above** (on the `mtcars` dataset), **a plot of the residuals against the fitted values is as follows:**



**There does appear to be some pattern here and the three named points on the graph might be outliers.**

## Scaled deviance

The scaled deviance (or likelihood ratio) is used to compare the fit of the saturated model with the fit of another model. The scaled deviance of Model 1 is defined as:

$$SD_1 = 2\left(\ln L_S - \ln L_1\right)$$

where $L_S$ is the likelihood of the saturated model.

The poorer the fit of Model 1, the bigger the scaled deviance will be.

## Comparing models

Where the data are normally distributed, it can be shown that, for two *nested* models, Models 1 and 2 where Model 1 has $p$ parameters and Model 2 has $p+q$ parameters:

$$SD_1 - SD_2 \sim \chi_q^2$$

For other distributions, the difference in the scaled deviances has an approximate (asymptotic) chi-square distribution with $q$ degrees of freedom.

Alternatively, we can compare the reduction in the AIC of the two models.

## The process of selecting explanatory variables

(1) Forward selection. Add the covariate that reduces the AIC the most or causes a significant decrease in the deviance. Continue in this way until adding any more causes the AIC to rise or does not lead to a significant improvement in the deviance. It is usual to consider main effects before interaction terms and linear terms before polynomials.

(2) Backward selection. Start by adding all available covariates and interactions. Then remove covariates one by one starting with the least significant until the AIC reaches a minimum or there is no significant improvement in the deviance, and all the remaining covariates have a statistically significant impact on the response.

## Rules for determining the number of parameters in a model

The constant model has 1 parameter.

A model consisting of one main effect that is a variable (*eg* age) has two parameters (*eg* $\beta_0$ and $\beta_1$).

A model consisting of one main effect that is a factor (*eg* sex) has as many parameters as there are categories (*eg* $\alpha_i$, *i* = 1 (male) and *i* = 2 (female)).

When a new main effect is added to a model (*eg* age + sex), we add $n-1$ parameters where $n$ is the number of parameters if the main effect were on its own (*eg* for age + sex, the number of parameters is 2 + (2 – 1) = 3).

When an interactive effect (a dot term) is added to a model (*eg* age + sex + age.sex), we add $(m-1)(n-1)$ parameters for the interactive effect (*eg* for age + sex + age.sex, the number of parameters is 2 + (2 – 1) + (2 – 1)(2 – 1) = 4).

A model consisting of a star term only (*eg* age*sex) has $mn$ parameters where *m* and *n* are the number of parameters if the main effects were on their own (*eg* for age*sex, the number of parameters is $2\times 2=4$ ).

## Residuals

A residual is a measure of the difference between the observed values $y_i$ and the fitted values $\hat{\mu}_i$. Two commonly used residuals for GLMs are the Pearson residual and the deviance residual.

***Pearson residuals***

These are $\dfrac{y-\hat{\mu}}{\sqrt{\text{var}(\hat{\mu})}}$ where $\text{var}\left(\hat{\mu}\right)$ is $\text{var}\left(Y\right)$ with $\mu$ replaced by the corresponding fitted value $\hat{\mu}$.

The Pearson residual, which is often used for normally distributed data, has the disadvantage that its distribution is often skewed for non-normal data. This makes the interpretation of residuals plots difficult.

***Deviance residuals***

These are $sign(y-\hat{\mu})d_i$ where $\sum d_i^2$ is the scaled deviance of the model.

Deviance residuals are usually more likely to be symmetrically distributed and to have approximately normal distributions, and are preferred for actuarial applications.

For normally distributed data, the Pearson and deviance residuals are identical.

## Testing whether a parameter is significantly different from zero

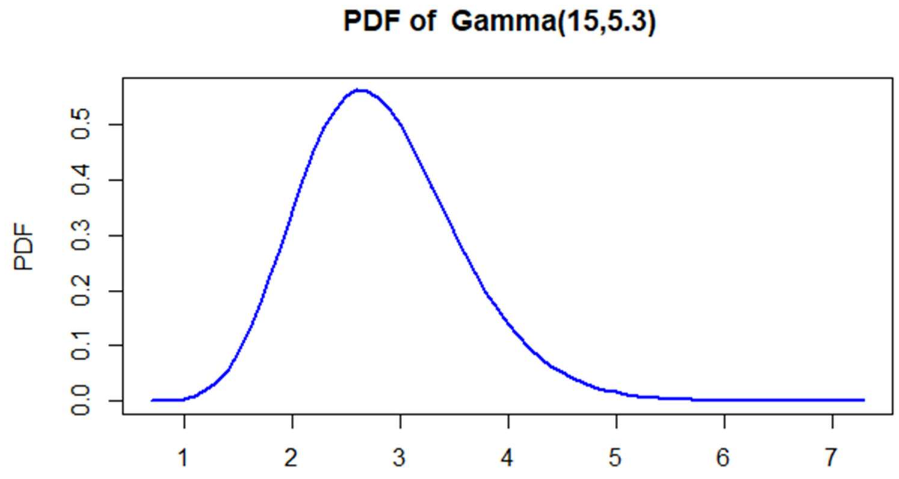
As a general rule, we can conclude that a parameter is significantly different from zero if it is at least twice as big in absolute terms as its standard error, *ie* if:

$$\left|\beta\right| > 2\, s.e\left(\beta\right)$$

# 5 Credible Intervals

**Having derived the posterior distribution of a parameter $\theta$, there are several ways in which we can summarise inferences about $\theta$. For single parameters, a plot of the posterior density is very informative and shows clearly the range of values consistent with our posterior beliefs.**

In Section 5.1 below, the Core Reading considers a numerical example where the posterior distribution is $Gamma(15,5.3)$. A plot of the PDF of this distribution is given below:



**As described earlier, we can also quote quantities such as the posterior mean of a parameter or the posterior variance.**

For the $Gamma(15,5.3)$ distribution pictured above, the mean is 2.83, the variance is 0.534 and the standard deviation is 0.731.

**For expressing and quantifying uncertainty about the values of $\theta$, a natural analogue of the classical confidence interval is the Bayesian credible interval.**

In Chapter 8, we saw how to estimate parameters using the method of moments and the method of maximum likelihood. In Chapter 9, we used confidence intervals to express the uncertainty in these estimates. Earlier in this chapter, we estimated a parameter using the mean, mode or median of its posterior distribution. We will now explain how to express the uncertainty in these estimates.

**Suppose that, given data $\underline{x}$, we derive the posterior density of $\theta$ as $f(\theta \mid \underline{x})$. Then, for $0<\alpha<1$, a $100(1-\alpha)\%$ credible interval for $\theta$ is a region of $\theta$, say $A$, which is such that:**

$$P(\theta \in A \mid \underline{x}) = \int_A f(\theta \mid \underline{x})\, d\theta = 1-\alpha$$

**So, a $100(1-\alpha)\%$ credible interval is an interval whose posterior probability of containing $\theta$ is $1-\alpha$.**

## 5.1 Equal-tailed credible intervals

**Often, we quote an equal-tailed credible interval, obtained by using the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ critical points of the posterior distribution. For example, with $\alpha = 0.05$, the 2.5% and 97.5% critical points of the posterior distribution would give a 95% credible interval.**

This is similar to the approach we used in Chapter 9 to calculate confidence intervals. If we want a two-sided 95% confidence interval, we split the remaining 5% equally between the two tails.

By definition, an equal-tailed credible interval must contain the median of the posterior distribution, *ie* the posterior estimate for $\theta$ under absolute loss.

**To calculate** equal-tailed **credible intervals for a parameter we need the cumulative distribution function of its posterior distribution. When the posterior distribution has a convenient form, such as a normal, beta or gamma distribution, we can usually use statistical tables, or standard functions from a computer package such as R to do the calculations.**

There are no tables for the beta distribution in the *Tables*, so we have to use R to obtain credible intervals based on a beta posterior distribution. We can, however, use the standard normal tables for a normal posterior, and the chi-square tables, along with the gamma-chi relationship, for a gamma posterior.

**Example**

**Suppose that, given data $\underline{x}$, the posterior distribution of the parameter $\theta$ is a gamma distribution with parameters 15 and 5.3, *ie* $\theta \mid \underline{x} \sim Gamma(15, 5.3)$. For an equal-tailed 90% credible interval of $\theta$, we need the 5% and 95% critical points of the $Gamma(15, 5.3)$ distribution.**

**In R we can use:**

```
qgamma(0.05,15,5.3)
qgamma(0.95,15,5.3)
```

**to obtain the 90% equal-tailed credible interval as (1.74,4.13).**

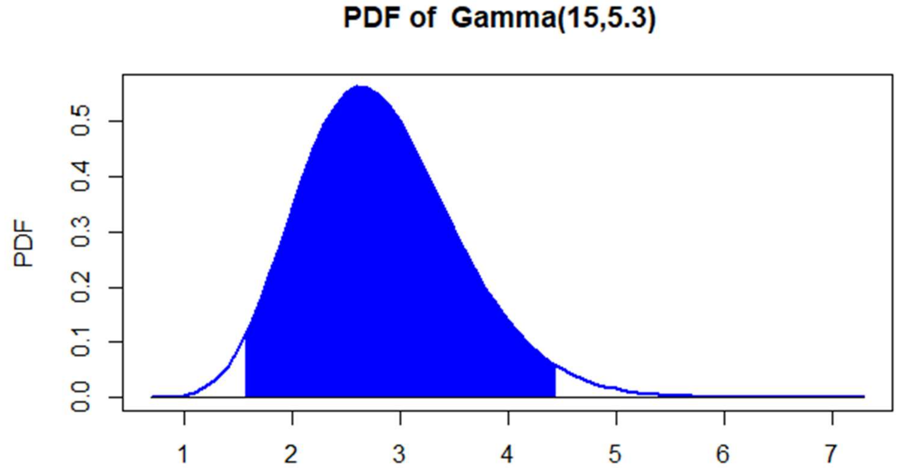
**Notice that, in this case, we can also use the relationship between the gamma and the chi-square distribution to calculate the interval. In particular, we have:**

$$(2 \times 5.3)\theta \mid \underline{x} = 10.6\theta \mid \underline{x} \sim Gamma(15, 1/2), \quad \textit{ie } 10.6\theta \mid \underline{x} \sim \chi^2_{30}$$

**From statistical tables, we have that the 5% and 95% critical points of the $\chi^2_{30}$ distribution are 18.49 and 43.77, respectively. So a 90% equal-tailed credible interval for $10.6\theta \mid \underline{x}$ is (18.49, 43.77), and therefore a 90% equal-tailed credible interval for $\theta \mid \underline{x}$ is:**

$$\left(\frac{18.49}{10.6}, \frac{43.77}{10.6}\right) = (1.74, 4.13) \text{, exactly as before.}$$

We can similarly obtain a 95% equal-tailed credible interval for $\theta | \underline{x}$ :



The credible interval is (1.58, 4.43). 95% of the distribution (the shaded area in the diagram above) lies between these values, with 2.5% on either side. The areas under the graph in the two tails are equal, *ie* $P(\theta < 1.58 | \underline{x}) = P(\theta > 4.43 | \underline{x}) = 0.025$ .

## Question

A random sample of size 15 from a normal distribution with mean $\mu$ and standard deviation 3 yields the following data values:

10.75 −0.29 5.37 6.68 8.77 1.69 7.12 4.89 6.45 4.27 9.37 5.68 3.87 7.70 6.98

The prior distribution of $\mu$ is $N(5,2^2)$ .

Calculate an equal-tailed 95% Bayesian credible interval for $\mu$ based on these data values. You are given that the posterior distribution of $\mu$ is $N(5.83,0.722^2)$ .

## Solution

From the *Tables*, we have $P(-1.96 < Z < 1.96) = 0.95$ . So the lower and upper 2.5% points of $N(5.83,0.722^2)$ are:

$$5.83 - 1.96 \times 0.772 = 4.41$$

and:

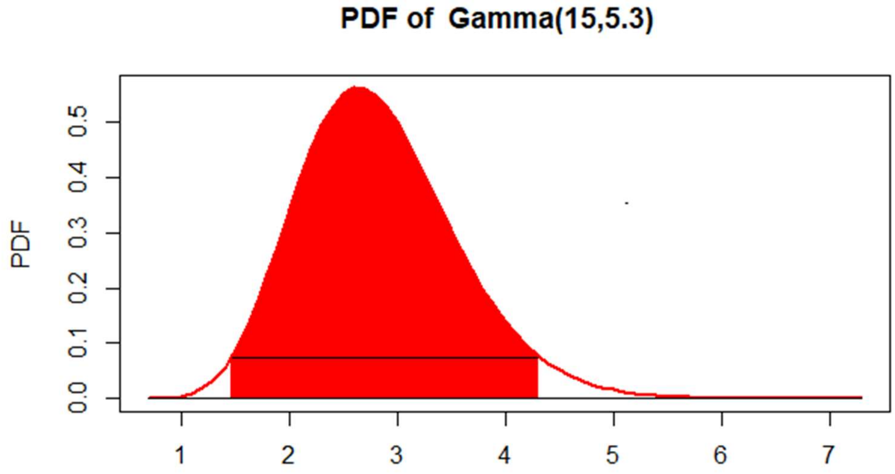
$$5.83 + 1.96 \times 0.772 = 7.24$$

So an equal-tailed 95% credible interval for $\mu$ is $(4.41, 7.24)$ .

## 5.2 Highest posterior density intervals

**As an alternative to an equal-tailed credible interval, a $100(1-\alpha)\%$ highest posterior density interval for $\theta$ could be quoted. In addition to satisfying $P(\theta \in A \mid \underline{x}) = 1-\alpha$, this interval is such that the minimum density of any point within the interval $A$ is equal to or higher than the density outside that interval.**

The following diagram shows a 95% highest posterior density interval for $\theta \mid \underline{x}$ :



Calculating highest posterior density intervals for non-symmetrical distributions is not straightforward. In R, the package `bayestestR` has the function `hdi` that calculates the highest density interval for a parameter. This is beyond the scope of Subject CS1, but for interested students, the code used to generate the 95% highest posterior density interval in this example is given below:

```
install.packages("bayestestR")
library("bayestestR")

set.seed(3)
x <- rgamma(100000,15,5.3)
hdi(x,ci=0.95)
```

The credible interval is (1.48, 4.29). The areas under the graph in the two tails are *not* equal, *ie* $P(\theta < 1.48 \mid \underline{x}) \neq P(\theta > 4.29 \mid \underline{x}) \neq 0.025$, although the probabilities do sum to 5%.

For unimodal distributions (such as the gamma distribution), the two endpoints of a highest posterior density interval have the same height (*ie* density). In the example above:

$$f(1.48) = f(4.29) = 0.80$$

The densities of all the values in a higher posterior density interval are larger than the densities of those outside the interval (*ie* the graph is higher in the interval). So, a higher posterior density interval contains a collection of most likely values of the parameter $\theta$, which is a desirable property. By definition, a higher posterior density interval must contain the mode, *ie* the posterior estimate for $\theta$ under 0-1 loss.

For a unimodal distribution, the highest posterior density interval is the shortest interval amongst all Bayesian credible intervals. For symmetrical distributions, such as a normal posterior distribution, the equal-tailed credible interval and highest posterior density interval are identical when based on the same data set. For skewed distributions, such as the gamma and most beta posterior distributions, the highest posterior density interval is not the same as the equal-tailed interval (as we have seen in the example above involving the $Gamma(15, 5.3)$ distribution).

The chapter summary starts on the next page so that you can keep all the chapter summaries together for revision purposes.

# Chapter 14 Summary

## Bayesian estimation *v* classical estimation

A common problem in statistics is to estimate the value of some unknown parameter $\theta$ .

The classical approach to this problem is to treat $\theta$ as a fixed, but unknown, constant and use sample data to estimate its value. For example, if $\theta$ represents some population mean then its value may be estimated by a sample mean.

The Bayesian approach is to treat $\theta$ as a random variable.

## Prior distribution

The prior distribution of $\theta$ represents the knowledge available about the possible values of $\theta$ before the collection of any sample data.

## Likelihood function

A likelihood function, $L$, is then determined, based on a random sample $\underline{X}=\left(X_1, X_2, \ldots, X_n\right)$. The likelihood function is the joint PDF (or, in the discrete case, the joint probability) of $X_1, X_2, \ldots, X_n \mid \theta$.

## Posterior distribution

The prior distribution and the likelihood function are combined to obtain the posterior distribution of $\theta$.

When $\theta$ is a continuous random variable:

$$f_{post}(\theta) \propto f_{prior}(\theta) \times L$$

When $\theta$ is a discrete random variable, the posterior distribution is a set of conditional probabilities.

## Conjugate distributions

For a given likelihood, if the prior distribution leads to a posterior distribution belonging to the same family as the prior, then this prior is called the conjugate prior for this likelihood.

## Uninformative prior distributions

If we have no prior knowledge about $\theta$ , a uniform prior distribution should be used. This is sometimes referred to as an uninformative prior distribution. When the prior distribution is uniform, the posterior PDF is proportional to the likelihood function.

## Loss functions

A loss function, such as quadratic (or squared) error loss, absolute error loss or all-or-nothing (0/1) loss gives a measure of the loss incurred when $\hat{\theta}$ is used as an estimator of the true value of $\theta$ . In other words, it measures the seriousness of an incorrect estimator.

Under squared error loss, the mean of the posterior distribution minimises the expected loss function.

Under absolute error loss, the median of the posterior distribution minimises the expected loss function.

Under all-or-nothing loss, the mode of the posterior distribution minimises the expected loss function.

## Credible intervals

A Bayesian credible interval quantifies uncertainty about the values of parameter $\theta$. A $100(1-\alpha)\%$ credible interval is an interval whose posterior probability of containing $\theta$ is $1-\alpha$.

These can be equal-tailed intervals or highest posterior density intervals.

The endpoints of an equal-tailed 95% credible interval for $\theta$ are the lower and upper 2.5% points of the posterior distribution of $\theta$. If the posterior distribution is a standard distribution with tabulated values, we can calculate equal-tailed confidence intervals algebraically.

The densities of all points within a highest posterior density interval are greater than or equal to the densities of all points that lie outside the interval. We can use R to calculate highest posterior density intervals.

**So the posterior distribution of $\theta$ given $\underline{x}$ is:**

$$N\left(\frac{\frac{n\bar{x}}{\sigma_1^2}+\frac{\mu}{\sigma_2^2}}{\frac{n}{\sigma_1^2}+\frac{1}{\sigma_2^2}}, \frac{1}{\frac{n}{\sigma_1^2}+\frac{1}{\sigma_2^2}}\right)$$

**where:**

$$\bar{x}=\sum_{i=1}^{n} x_i / n$$

The Bayesian estimate of $\theta$ under quadratic loss is the mean of this posterior distribution:

$$E(\theta \mid \underline{x})=\frac{\frac{n\bar{x}}{\sigma_1^2}+\frac{\mu}{\sigma_2^2}}{\frac{n}{\sigma_1^2}+\frac{1}{\sigma_2^2}}$$

$$=\frac{\frac{n}{\sigma_1^2}}{\frac{n}{\sigma_1^2}+\frac{1}{\sigma_2^2}}\bar{x}+\frac{\frac{1}{\sigma_2^2}}{\frac{n}{\sigma_1^2}+\frac{1}{\sigma_2^2}}\mu$$

**or:**

$$E(\theta \mid \underline{x})=Z\,\bar{x}+(1-Z)\mu \qquad \textbf{(15.3.4)}$$

**where:**

$$Z=\frac{n}{n+(\sigma_1^2/\sigma_2^2)} \qquad \textbf{(15.3.5)}$$

**Equation (15.3.4) is a credibility estimate of $E(\theta \mid \underline{x})$ 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, $\mu$ is the best available estimate if no data were available from the risk itself.**

**Notice that, as for the Poisson/gamma model, the estimate based solely on data from the risk itself is a linear function of the observed data values.**

**There are some further points to be made about the credibility factor, $Z$, given by (15.3.5):**

- **It is always between zero and one.**
- **It is an increasing function of $n$, the amount of data available.**
- **It is an increasing function of $\sigma_2$, the standard deviation of the prior distribution.**

**These features are all exactly what would be expected for a credibility factor.**

Notice also that, as $\sigma_1^2$ increases, the denominator increases, and so $Z$ decreases. $\sigma_1^2$ denotes the variance of the distribution of the sample values. If this is large, then the sample values are likely to be spread over a wide range, and they will therefore be less reliable for estimation.

**The R code to obtain the Monte Carlo credibility premiums for the above based on *M* simulations is:**

```
Z <- n/(n+sigma1^2/sigma2^2)
cp <- rep(0,M)
for (i in 1:M)
    {theta <- rnorm(1,mu,sigma2)
    x <- rnorm(n,theta,sigma1)
    cp[i] <- Z*mean(x)+(1-Z)*mu
    }
```

**The average of these credibility estimates is given by:**

```
mean(cp)
```

## 3.5 Further remarks on the normal/normal model

**In Section 3.4 the normal/normal model for the estimation of a pure premium was discussed within the framework of Bayesian statistics. In this section the same model will be considered, without making any different assumptions, but in a slightly different way.**

**The reason for doing this is that some of the observations will be helpful when empirical Bayes credibility theory is considered in the next chapter.**

**In this section, as in Section 3.4, the problem is to estimate the expected aggregate claims produced each year by a risk. Let:**

$$X_1, X_2, \ldots, X_n, X_{n+1}, \ldots$$

**be random variables representing the aggregate claims in successive years. The following assumptions are made.**

**The distribution of each $X_j$ depends on the value of a fixed, but unknown, parameter, $\theta$.**

**The conditional distribution of $X_j$ given $\theta$ is $N(\theta, \sigma_1^2)$.**

**Given $\theta$, the random variables $\{X_j\}$ are independent.**

**The prior distribution of $\theta$ is $N(\mu, \sigma_2^2)$.**

**The values of $X_1, X_2, \ldots, X_n$ have already been observed and the expected aggregate claims in the coming, *ie* $(n+1)$th, year need to be estimated.**

So, depending on the context of the problem, $X_j$ represents either:

- the aggregate claim amount in Year $j$ per unit of risk volume, or
- the total number of claims in Year $j$ per unit of risk volume.

In Model 1, we assume that $P_j$ is always equal to 1, *ie* the volume of business is the same for each risk group.

## Assumptions for EBCT Model 2

**The assumptions that specify EBCT Model 2 are as follows.**

**Assumption 7: The distribution of each $X_j$ depends on the value of a parameter, $\theta$, whose value is the same for each $j$ but is unknown.**

**Assumption 8: Given $\theta$, the $X_j$'s are independent (but not necessarily identically distributed).**

**Assumption 9: $E(X_j \mid \theta)$ does not depend on $j$.**

**Assumption 10: $P_j \operatorname{var}(X_j \mid \theta)$ does not depend on $j$.**

**As in previous sections, $\theta$ is known as the risk parameter for the risk, and, as for EBCT Model 1, it could be just a single real valued number or a more general quantity such as a vector of real valued numbers. Assumption 7 is the standard assumption for all credibility models considered here. Assumption 8 corresponds to Assumption 2 in EBCT Model 1, but notice that Assumption 8 is *slightly weaker* than Assumption 2. Assumption 8 does not require the $X_j$'s to be conditionally (given $\theta$) identically distributed, but only to be conditionally independent. There is no assumption in EBCT Model 2 that the $X_j$'s are unconditionally, or conditionally given $\theta$, identically distributed.**

**If all the $P_j$'s are equal to 1, then Assumptions 7-10, taken together, become the same as Assumptions 4, 5 and 6 (taken together) in EBCT Model 1. Thus, if all the $P_j$'s are equal to 1, EBCT Model 2 is exactly the same as EBCT Model 1.**

**Having made Assumptions 9 and 10, $m(\theta)$ and $s^2(\theta)$ can be defined as follows:**

$$m(\theta) = E(X_j \mid \theta)$$

$$s^2(\theta) = P_j \operatorname{var}(X_j \mid \theta)$$

**The definition of $m(\theta)$ corresponds exactly to the definition for EBCT Model 1 in Section 1 but the definition of $s^2(\theta)$ is slightly different.**

In Model 2, there is a factor of $P_j$ in the definition of $s^2(\theta)$. In Model 1, $P_j = 1$ and so $s^2(\theta) = \operatorname{var}(X_j \mid \theta)$.

**To gain a little more insight into Assumptions 9 and 10, consider the following example. Suppose the risk being considered is made up of a different number of independent policies each year and that the number of policies in Year $j$ is $P_j$.**

It is important to realise that $P_j$ is a known quantity, not a random variable.

**Suppose also that the aggregate claims in a single year from a single policy have mean $m(\theta)$ and variance $s^2(\theta)$, where $m(\ )$ and $s^2(\ )$ are functions of $\theta$, and $\theta$ is the fixed, but unknown, risk parameter for all these policies. Now let $Y_j$ denote the aggregate claims from all the policies in force in Year $j$.**

Then $E(Y_j\,|\,\theta)$ is the expected aggregate claim amount from all policies in year $j$, and:

$$E(Y_j\,|\,\theta)=\sum_{k=1}^{P_j}\text{expected aggregate claim amount for policy } k=\sum_{k=1}^{P_j}m(\theta)$$

So:

$$\boldsymbol{E(Y_j\,|\,\theta)=P_j\,m(\theta)}$$

Also, since the policies are assumed to be independent:

$$\text{var}(Y_j\,|\,\theta)=\sum_{k=1}^{P_j}\text{variance of aggregate claim amount for policy } k=\sum_{k=1}^{P_j}s^2(\theta)$$

*ie*:

$$\boldsymbol{\text{var}(Y_j\,|\,\theta)=P_j\,s^2(\theta)}$$

Then, since $X_j=\dfrac{Y_j}{P_j}$:

$$E(X_j\,|\,\theta)=\frac{1}{P_j}E(Y_j\,|\,\theta)=m(\theta)\quad\text{and}\quad\text{var}(X_j\,|\,\theta)=\frac{1}{P_j^2}\text{var}(Y_j\,|\,\theta)=\frac{s^2(\theta)}{P_j}$$

So:

$$\boldsymbol{E(X_j\,|\,\theta)=m(\theta)\quad\text{and}\quad P_j\,\text{var}(X_j\,|\,\theta)=s^2(\theta)}$$