Stats Pack

Corrections to 2014 study material

Comment

This document contains details of any errors and ambiguities in the Stats Pack study materials for 2014 that have been brought to our attention. We will incorporate these changes in the study material each year. We are always happy to receive feedback from students, particularly details concerning any errors, contradictions or unclear statements in the courses. If you have any such comments on this course please email them to StatsPack@bpp.com.

Important note

This document was last revised on 3rd June 2014. The date on which any corrections have been added is noted at the start of each section.

Chapter 1 (updated on 22nd April 2014)

Page 24

There is an error in the first paragraph after the stem and leaf diagram. It should read "whereas type B claims are located between $300 and $400"

Replacement pages can be found at the end.

Page 37

The axes for the bar chart in solution 1.6 are the wrong way round. "Frequency" should be on the vertical axis and "Mock Results" should be on the horizontal axis.

Replacement pages can be found at the end.

Chapter 2 (updated on 30th January 2014)

P2.7 Page 34 and page 47

There is a typo in the table on the question and the solution. The figure of 3300 should be £300.

Replacement pages can be found at the end.

Chapter 7 (updated on 30th January 2014)

Solution 7.9 Page 58

The second row of the table it should read $P(V=v)$ not $F(v)$.

Replacement pages can be found at the end.

Chapter 7 (updated on 30th January 2014)

Solution P7.10(ii) Page 80

There is a typo in the first line of part (ii) of the solution. It should read "using $\text{var}(aX+b)=a^2\,\text{var}(X)$".

Replacement pages can be found at the end.

Chapter 8 (updated on 2nd June 2014)

Mode Page 27

There is a typo in the first line under the graph of the Poisson distribution. It should read "It is easy to see that **1 and 2** are both modes as these have the greatest probability of occurring."

Replacement pages can be found at the end.

Question 1.15

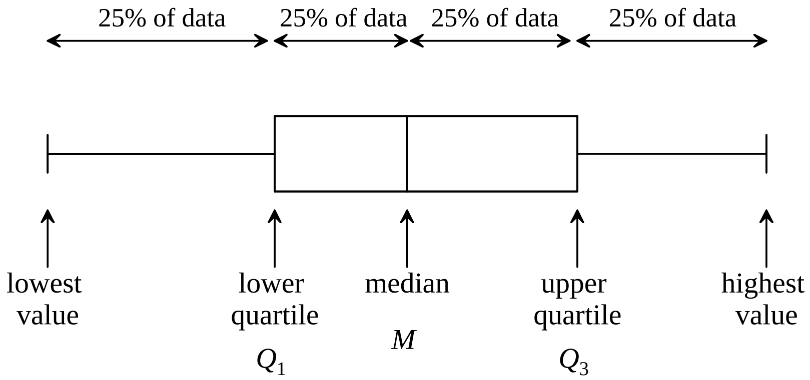
Use your cumulative frequency curve from Question 1.14 to estimate:

(i) how many policyholders are aged 32 or less

(ii) the age under which 75% of the policyholders lie.

3.6 Boxplot

A boxplot (also called a box and whisker plot) is another way of showing data:



The rectangle (box) in the middle represents the middle 50% of the data (between the values that are a ¼ and ¾ of the way through the data). The lines (whiskers) extend from the box to the smallest and largest values. The diagram also shows the middle value (called the **median**).

A boxplot is particularly effective when comparing two sets of data, however to draw the diagram we need to calculate the median and the quartiles. Since the median will be covered in Chapter 2 and the quartiles will be covered in Chapter 3 we will deal with this type of diagram at the end of Chapter 3.

In the exam it is expected that you would draw a boxplot accurately on graph paper.

4 Using diagrams to compare data

Once we have drawn our diagrams we can use them to interpret the patterns in the data or compare two or more data sets. In Subject CT3 we will be looking at three features of any data set: the location, the spread and the skewness.

4.1 Location

The location of a data set is simply where the data is *located* – *ie* where is the centre of the data or about what values is it grouped. In everyday language you may use ‘average’ to describe the location.

The stem and leaf diagrams below show the claim amounts (in $’s) under two different types of insurance:

Type A

0	2 7
1	1 1 3 6 8 9
2	3 4 4 4 7
3	0 5
4	1
5	2

Type B

0	8
1	0 2 3
2	1 4 6 8
3	2 3 3 6 9 9
4	0 1 5
5	4

Key: 2|5 represents $250

Type A claims are mostly located between $100 and $200 whereas type B claims are located between $300 and $400. So we could say the type B claims are greater *on average* than type A claims.

In Chapter 2, we will use the mean, median and mode to measure the location of a set of data.

Solution 1.5

The cumulative frequency table is:

Time (t)	**Cumulative Frequency**
$t < 0.5$	2
$t < 1$	7
$t < 2$	14
$t < 5$	26
$t < 10$	30

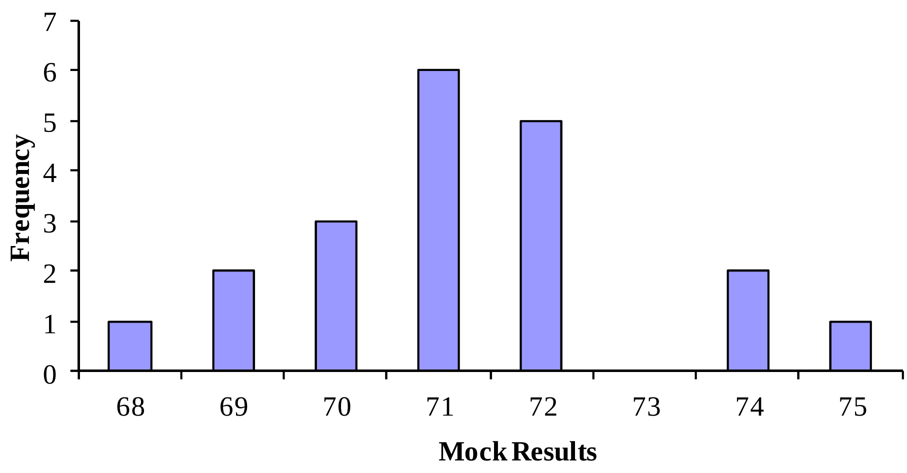
Or we could use "up to 0.5 mins", "up to 1 min", etc as the groups.

Solution 1.6

Putting this data into a frequency table:

Mock result	**Frequency**
68	1
69	2
70	3
71	6
72	5
73	0
74	2
75	1

It is now easy to draw the bar chart:

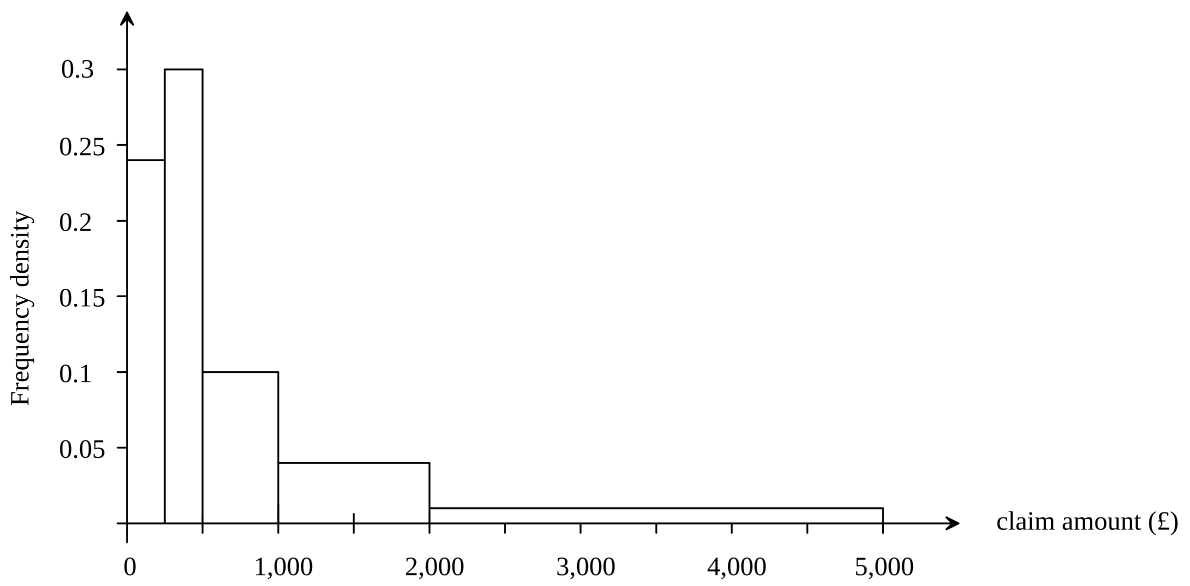


Solution 1.7

(i) Using $frequency\ density = \dfrac{frequency}{class\ width}$ we get:

Claim amount (x)	Frequency	Frequency density
$0 \le x < 250$	60	$60 \div 250 = 0.24$
$250 \le x < 500$	75	$75 \div 250 = 0.3$
$500 \le x < 1,000$	50	$50 \div 500 = 0.1$
$1,000 \le x < 2,000$	40	$40 \div 1,000 = 0.04$
$2,000 \le x < 5,000$	30	$30 \div 3,000 = 0.01$

(ii) The histogram is:



P2.4 Subject 101, September 2000, Q2 (part)

Consider a random sample of 47 white-collar workers and a random sample of 24 blue-collar workers from the workforce of a large company. The mean salary for the sample of white-collar workers is £28,470; whereas the mean salary for the sample of blue-collar workers is £21,420.

Calculate the mean of the salaries in the combined sample of 71 employees. [1]

Section 3: Sample median

P2.5 Subject C1, September 1995, Q1 (adjusted)

A random sample of 15 motor windscreen claim amounts (in £) is given by:

121	107	139	72	123
114	215	156	100	136
169	89	115	153	111

What is the median claim amount? [2]

P2.6 Subject 101, September 2001, Q1 (part)

Data were collected on 100 consecutive days for the number of claims, x, arising from a group of policies. This resulted in the following frequency distribution:

x	0	1	2	3	4	≥ 5
f	14	25	26	18	12	5

Calculate the median for these data. [1]

P2.7 Subject C1, September 1997, Q9 (part)

The table below shows a grouped frequency distribution for 100 claim amounts on a certain class of insurance policy.

Claim Amount	Frequency
under £100	4
£100 – 149.99	10
£150 – 199.99	25
£200 – 249.99	30
£250 – 299.99	15
£300 – 349.99	12
£350 – 399.99	4
£400 or over	0

Determine an approximate value for the median of these claim amounts. [4]

Section 5: Location and skewness

P2.8 Subject C1, Specimen 1993, Q2

For a particular class of insurance policy the distribution of claim amounts is positively skewed. Which of the following statements about the claim amount distribution is true?

A mode > median > mean

B mean > median > mode

C median > mode > mean

D mean > mode > median [2]

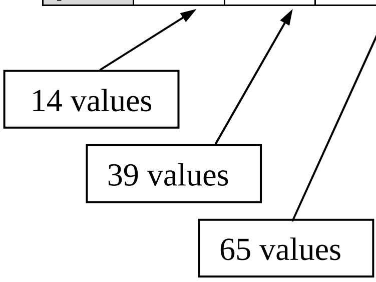
P2.5 First of all we need to put the claim amounts in order:

72 89 100 107 111 114 115 121 123 136 139 153 156 169 215

The median is the $\tfrac{1}{2}\times 15+\tfrac{1}{2}=8\text{th}$ value which is £121.

P2.6 The median is the $\tfrac{1}{2}\times 100+\tfrac{1}{2}=50\tfrac{1}{2}$ th value. So counting through the frequencies:

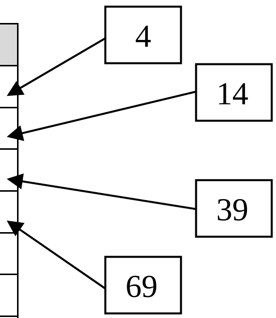
x	0	1	2	3	4	≥ 5
f	14	25	26	18	12	5



So the 50½th value is 2 claims.

P2.7 The median is the $\tfrac{1}{2}\times 100=50$ th value. So counting through the frequencies:

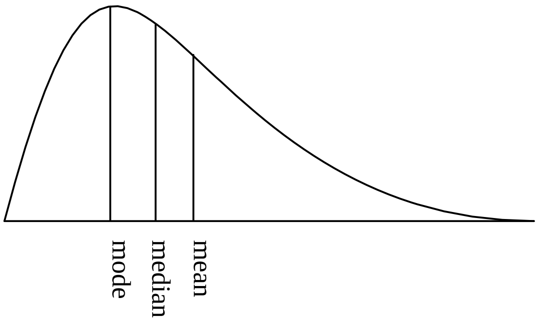
Claim Amount	**Frequency**
under £100	4
£100 – 149.99	10
£150 – 199.99	25
£200 – 249.99	30
£250 – 299.99	15
£300 – 349.99	12
£350 – 399.99	4
£400 or over	0



The 50th value is the 11th value in the $£200-£249.99$ group. Using interpolation, we get the median to be:

$$200+\frac{11}{30}\times 49.99=£218.33$$

P2.8 In Section 5 we had the following diagram:



Hence, we can see that answer B is correct.

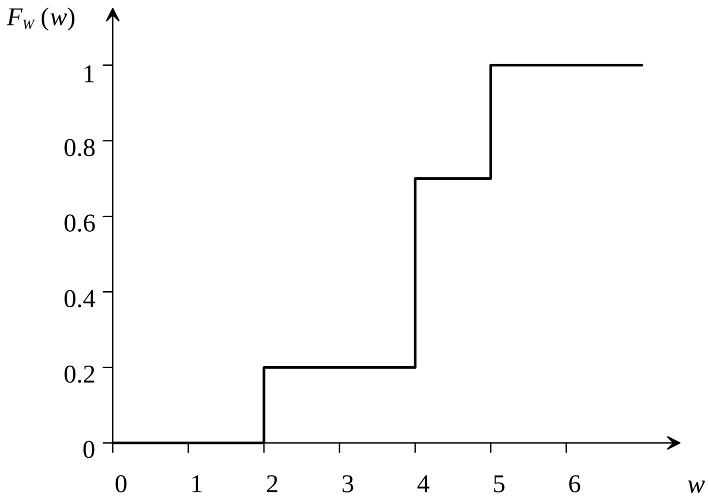
P2.9 Using the sum of the fifty amounts given in the question:

$$\bar{x} = \frac{£92,780}{50} = £1,855.60$$

The median is the $½ \times 50 + ½ = 25½$ th value. Counting through the leaves, we see that this lies between the 3 and the 4 on the 17 stem (*ie* between 1,730 and 1,740). Hence, the median is £1,735.

Solution 7.8

(i) The graph will be:



(ii) From this we get the cumulative distribution function to be:

$$F_W(w) = \begin{cases} 0 & w < 2 \\ 0.2 & 2 \le w < 4 \\ 0.7 & 4 \le w < 5 \\ 1 & 5 \le w \end{cases}$$

(iii) Either reading off the graph or using the CDF we get:

$$F_W(1) = 0$$

$$F_W(4.5) = 0.7$$

$$F_W(10) = 1$$

Solution 7.9

We can solve this by obtaining the probability function or by using the cumulative function directly.

Obtaining the probability function

First we notice that the CDF jumps up at $x = 1, 2, 3$ and 4. To obtain the probabilities of obtaining each of these values, we simply subtract the cumulative probabilities:

v	1	2	3	4
P(V = v)	0.216	0.432	0.288	0.064

We can now obtain the probabilities:

(i) $P(V = 2) = 0.432$

(ii) $P(V > 1) = P(V = 2) + P(V = 3) + P(V = 4) = 0.432 + 0.288 + 0.064 = 0.784$

(iii) $P(V < 3) = P(V = 1) + P(V = 2) = 0.216 + 0.432 = 0.648$

Using the cumulative distribution function directly.

(i) Using the fact that subtracting cumulative probabilities gives the original probabilities:

$$P(V = 2) = F_V(2) - F_V(1) = 0.648 - 0.216 = 0.432$$

(ii) Since probabilities of all possible values sum to 1, we get:

$$P(V > 1) = 1 - P(V \leq 1) = 1 - F_V(1) = 1 - 0.216 = 0.784$$

(iii) Reading directly from the cumulative distribution function:

$$P(V < 3) = 0.648$$

P7.9 The mean of C is given by:

$$E(C) = E(7.00 + 0.0742N)$$

$$= 7.00 + 0.0742E(N) \qquad \text{using } E(aX + b) = aE(X) + b$$

$$= 7.00 + 0.0742 \times 600 \qquad \text{since } E(N) = 600$$

$$= 51.52$$

The variance of C is given by:

$$\text{var}(C) = \text{var}(7.00 + 0.0742N)$$

$$= 0.0742^2 \text{ var}(N) \qquad \text{using var}(aX + b) = a^2 \text{ var}(X)$$

$$= 0.0742^2 \times 250 \qquad \text{since var}(N) = 250$$

$$= 1.37641$$

P7.10 (i) $E(3+6U)=3+6E(U)$ using $E(aX+b)=aE(X)+b$

$=3+6\times 8$ since $E(N)=600$

$=51$

(ii) $\text{var}(8-2U)=(-2)^2\ \text{var}(U)$ using $\text{var}(aX+b)=a^2\ \text{var}(X)+b$

$=4\times 3^2$ since $\text{var}(U)=3^2$

$=36$

So the standard deviation is $\sqrt{36}=6$.

Alternatively, we could have used $sd(aX+b)=a\times sd(X)$*:*

(iii) $\text{var}\left(\dfrac{U-8}{3}\right)=\text{var}\left(\tfrac{1}{3}U-\tfrac{8}{3}\right)$

$=\left(\tfrac{1}{3}\right)^2\ \text{var}(U)$ using $\text{var}(aX+b)=a^2\ \text{var}(X)$

$=\tfrac{1}{9}\times 3^2$ since $\text{var}(U)=3^2$

$=1$

(iv) $E(U^2-4U+7)=E(U^2)-4E(U)+7$ splitting up the expectation

$=E(U^2)-4\times 8+7$ since $E(U)=8$

$=E(U^2)-25$

To work this out we need $E(U^2)$. We don't have the distribution, so we can't work it out from first principles. We use the trick of rearranging the variance formula:

$$\text{var}(U)=E(U^2)-E^2(U)$$

$$\Rightarrow\quad E(U^2)=\text{var}(U)+E^2(U)=3^2+8^2=73$$

Hence:

$$E(U^2-4U+7)=73-25=48$$

Median

There is no easy way to get the median value other than counting through the probabilities to find which value is half way through the distribution. For example, when $X \sim Poi(2)$ we have:

x	0	1	2	3	4	5	…
P(X = x)	0.135	0.271	0.271	0.180	0.090	0.036	…

0.135

$0.135+0.271=0.406$

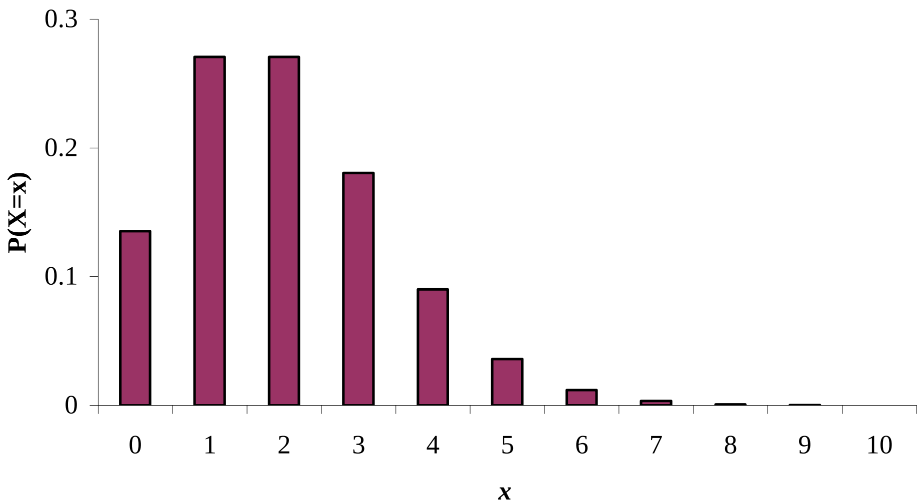
so 0.5 is in here!

$0.135+0.271+0.271=0.677$

So we can see that the median is 2.

Mode

The mode is the value that has the greatest probability. Looking at either the probability distribution of $X \sim Poi(2)$ given above or its graph below:



It is easy to see that 1 and 2 are both modes as these have the greatest probability of occurring.

Question 8.15

Calculate the median and mode of $X \sim Poi(1)$.

In summary:

Poisson distribution, $Poi(\lambda)$

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

$$E(X) = \lambda$$

$$\text{var}(X) = \lambda$$

These results are given in the *Tables* on page 7 and so do not need to be memorised.

4.4 Probabilities of a Poisson distribution

We can use the probability function to calculate probabilities. Suppose $X \sim Poi(3)$, our probability function is:

$$P(X = x) = \frac{3^x}{x!}e^{-3} \qquad x = 0,1,2,\dots$$

To calculate a single probability, $P(X = 2)$, we just substitute the value into our probability function:

$$P(X = 2) = \frac{3^2}{2!}e^{-3} = \frac{9}{2}e^{-3} = 0.22404$$

What about calculating $P(X \geq 2)$? Since $X \sim Poi(3)$ can take values $0,1,2,\dots$ this means:

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + \cdots$$

Eek! There's no way we can work it out this way! So we need to use the fact that probabilities of all the values X can take sum to 1:

$$P(X < 2) + P(X \geq 2) = 1 \quad \Rightarrow \quad P(X \geq 2) = 1 - P(X < 2)$$