M248 2003 solutions

The original solutions, created by Kenny Laing, have been supplemented by some additional answers from several 2005 M248 students. There are some discrepancies between our answers, but only around the second decimal place.

The answers are designed to aid revision, are by no means definitive, may contain errors, and do not contain the level of detail needed for the exam. They are not endorsed or verified by the OU.

Q1a
33.3% for diabetes, 19.6 % for plasma glucose conc < 148.5

Q1b
Percentage of row totals??

Q1ci
$p = 104 / 532 \approx 0.1955$

Q1cii
$p = 78 / 177 \approx 0.4407$

Q2
Median = 60, Lower Quartile = 50

Q3
31 degrees

Q4
I would use the median. The data is right-skew with 3 outliers. The median is more resistant to extreme values than the mean.

Q5
Male guinea pigs are, on average, heavier than females, as the boxplot shows that the lightest male is heavier than all but one of the females (the point that is shown as an outlier). Median, quartiles, minimum and maximum all located higher for males.
The range and interquartile range, and spread of adjacent values of the weights are wider for the males than for the females, although the iqr is only slightly wider for males.
Note: the words median, dispersion, range and outliers should be included for the marks.

Q6
There is a positive correlation between age and length, with increased age being associated with increased length, although this tails off at higher ages (after about 18 years).
The relationship appears to follow an increasing monotonic curve.
There is some variability, as can be seen by the scatter.
The points do not pass through the origin.
(Probably only two features needed for the marks).

Q7a
Left-skew

Q7b
Yes, because the stated sample mean is smaller than the stated sample median.
For left-skew data, the mean is smaller than the median.

Q8
The normal distribution allows for negative values, which is not appropriate for time. The mean and variance = 1, so the standard deviation = 1, and so 99% of the values lie within 3 standard deviations of the mean, that is, between -2 and 4.

A normal distribution is symmetric about the mean, whereas the time between successive events will have a mode at 0 (I think!)
(Probably an exponential distribution would provide a better model, but the question does not ask for an alternative.)

Q9
Histogram (b) has the mode at 1, is right-skew and appears to be unbounded, hence it is a Geometric distribution.
Histogram (a) is also right-skew and is unimodal, but appears to have a finite range.
In a Poisson distribution, x can equal 0: in a Geometric distribution, x cannot equal 0.
Hence, Histogram (a) is Poisson(0.7) and Histogram (b) is G(0.5).

Q10a
0.963

Q10b
0.117

Q11
0.3456

Q12
Mean = 2, Variance = 1.4

Q13
Don't get caught out with this type of question. It's the variance and mean that have to be equal to be suitable for a Poisson. Here we are given the s.d. so you need to square 2.2 and conclude that Poisson is NOT suitable

Q14
Poisson (1.5) and 0.1912 for the probability

Q15a
G(0.06)

Q15bi
0.0344

Q15bii
0.5386

Q15ci
X~B(1000,0.06)

Q15cii
X~N(60,56.4) d) 0.0808
(Remember the continuity correction)

Q16a
X ~ Poisson(3)

Q16a
0.1680

Q16c
T ~ M(0.25) so P(T<2) = 0.3935

Q17
E(X – Y) = E(X) – E(Y) = 5 – 3 = 2 V(X – Y) = V(X) + V(Y) = 25 + 9 = 34

Q18
(195.2 , 213.2) based on 0.975 percentile of the t(11) distribution

Q19
Function is decreasing so (mu-,mu+) = (f(theta+),f(theta-)) = (e^-0.21,e^-0.12) = (0.81,0.89)
It is valid because the transformation function is always decreasing.

Q20a
The p's and q's (i.e. 1 - p) below are the proportions for each of the respective vaccines.
A ~ N(1200p,1200pq) B ~ N(1500p,1500pq)
PS. Not 100% about this answer to (a) but it "feels" right!

Q20b
The assumptions are:
The probability of a child getting whooping cough is independent of whether or not any other child gets whooping cough;
The probability of a child getting whooping cough is constant;
The occurrence of whooping cough cases is normally distributed.
Note: np and npq should be at least 5.

Q20c
(-0.0049,0.0149)

Q20d
This confidence interval defines a plausible range of values for the true population mean difference in the following sense. If the difference were greater than 0.0149, then the probability of observing a sample mean smaller than or equal to (x bar) would be less than 0.025. Similarly, if the difference were smaller than -0.0049, then the probability of observing a sample mean greater than or equal to (x bar) would be less than 0.025. Such values are deemed implausible at the 95% confidence level.

Q21a
The null distribution of T is t(7).
The observed value of the test statistic $\approx$ 2.840
T lies between $Q_{0.975}$ and $Q_{0.99}$
0.05$\geq$ p > 0.02
Moderate evidence against null hypothesis.
The data provides moderate evidence that the true population mean volume of orange juice in cartons filled by the machine is not 500 ml. Indeed, the data suggest that the average volume is greater than 500 ml.

Q21b
We are told that the sample was drawn from a normal distribution so the test statistic is:

$$\frac{(n-1)S^2}{s^2} \approx c^2(n-1)$$

The sample variance S^2 is 4.8 (from part (a)) and the population variance ó^2 is 2.1. Therefore, the value of the test statistic is,

$$\frac{7\times 4.8}{2.1} = 16$$

The null distribution of the test statistic is $c^2 \approx (n-1)$, which is ÷^{2} -(7)
(NB, you are not asked to evaluate a conclusion, but their is weak evidence against H_0, since 0.1 p >0.05.)

Q22
The *p* value would be less than 0.01, providing strong evidence against the null hypothesis.
Since the confidence interval does not contain the value 2, the null hypothesis would be rejected at the 1% significance level.

Q23a
The underlying population distribution is symmetric.

However, the sample size should be at least 16, whereas it is only 6 (including 1 tied value) so the result may not be very reliable. Another good assumption which could earn you brownie points is that the null distribution of the test statistic is found by assuming that absolute differences with a particular rank is just as likely to be associated with a positive difference as with a negative difference!!

Q23b
2

Q23c
Since $p > 0.10$, there is little evidence against H_0. I conclude that it is feasible that the median difference in response times is zero.

Q23d
The test is much more powerful than a sign test as it includes information on the magnitude of the differences (unlike the sign test).
The handbook says that the test relies on the assumption that the underlying population distribution is symmetric. The sample size n is less than 16 so that assumption may not be valid in this case, which would produce a less powerful test.

Remember: Decrease Significance Level - Increase Power
Increase differences (between means) - Increases Power
Increase SD - Decrease Power
Increase sample size - Increase Power

Q24a
Degrees of freedom = 16 - 1 - 1 = 14

Q24b
$0.05 \geq p > 0.025$

Q24c
Moderate evidence against null hypothesis that binomial model is a good fit to the data. Therefore, the binomial model may not be a very good fit.

Q25a
Poisson distribution, Poisson(m)
Data are discrete counts, are right-skew, unbounded range, with range starting at zero.

Q25b
Exponential distribution with rate l , that is, M(l).
Data are continuous, range starts at zero, is unbounded. Used for events that are random in time. Mode at 0???

Q26a
$S_{xx} = 30.489$
$S_{xy} = 95.425$

Q26b
$b \approx 3.130$
$a \approx 65.96$
y = 65.96 + 3.130x
Height in cm = 65.96 + 3.13 x humerus length.

Q26c
When x = 34.2, $y \approx 173.01$

Q26d
(167.74, 178.28)

Q26e
The individual response at x = 34.2 involves the additional variability of individual observations around the mean. For a confidence interval, this additional variability is excluded, so a confidence interval will be narrower than a prediction interval.

Q27
Approximate value is 0.24

Q28a
0.7

Q28b
$P^{21} (1 - p)^9$

Q29a
Spearman rank correlation coefficient is more appropriate, since points appear to lie on a monotonic decreasing curve.

Q29b
Negative, because as X increases, Y decreases.

Q30a
0.486

Q30b
$0.5 \geq p > 0.1$

Q30c
Since $p > 0.1$, there is little evidence against the null hypothesis of no association between the variables.
I conclude that the data do not provide evidence that the drug is effective in reducing the incidence of coronary artery restoriosis.