Exam 2000 Qu 2-5

>Q.1
a) The figures in table 2 are the figures in table 1 divided by the population in millions. For example, 1996 physicians in table 1 = 3609 and in table 2 = 3609/13.1 = 116.0

b) The proportion of health care workers per million have gone up between 1984 and 1966. Most noteably, the number of dentists and nurses have approximatly tripled.

agreed but I also included the totals for each year.

d) All health care professionals have increased per million of the population between 1966 and 1984; however, miwives had a decrease between 1981 and 1984 which brought it below the 1976 level. Dentists and nurses had aproximatly tripled over the period. Midwives and pharmasists had approximatly doubled. Doctors had the smallest percent increase.

I mentioned that the level of midwives had actually decreased between 1981 and 1984 and wondered if it could be due to more also receiving a general training so that they were included within the nurse category rather than pure midwives.

e) 1966 = 24.84, 1971 = 19.52, 1976 = 18.24, 1981 = 17.84 , 1984 = 15.18
The proportion of doctors has been decreasing over time even though the number of doctors per million of the population has been increasing. This is caused by doctors increasing at a slower rate than other health care professionals.

I made the proportion of the total number of healthworkers to be 0.249 etc

>Q.2.
The following data are bat-to-prey detection distances in centimetres:
23 27 34 40 42 45 52 56 62 68 83
Find the sample median and sample upper quartile of these data.

median=45 upper quart = 62

>Q.3
The scatter plot arose from UNICEF data on child mortality published in 1992. It shows the mortality rate of children under 5 yrs of age per 1000 live births against the adult female literacy rate in Central and South American countries. The literacy rate is the percentage of females aged 15 or over who can read or write. Describe 2 aspects of the relationship between the variables that are suggested by this graph.

It appears as if the higher the rate of female literacy, the lower the mortality rate. There is one outlier of literacy below 50 and mortality about 100. If that data point were taken out, then a straight line regression model could be used. There is an inverse correlation between adult literacy and the mortality rate.

>
>Q.4
>Give 2 reasons why the function
>
> P(X = n) = (n - 2) / 4, n = 1,2,3,4,5
>
>is not a valid probability mass function.
>
>0<=p(X=n) but here we have -1/4
total p(X=n)=1 here 1.25

>Q.5
You are told that the boxplot in Figure 2 represents a dataset of 150

points with mean 15. Explain why you should question this information.

The boxplot indicates data negatively skewed; as such mean < median. But here mean 15 median 12.5

Exam 2000 Qu 6-10

>Q.6
Calculate the mean and median of the discrete probability distribution given below:

x	-1	0	1	2	3
P(x)	0.05	0.5	0.05	0.2	0.2

mean = 1 median = 0

>Q.7
The histogram and boxplot in fig 3 were obtained using the same data set. State which one best highlights the main feature of these data. In one sentence, justify your choice.

The histogram best highlights the main feature of these data; the bimodal nature of the distribution. The boxplot is indicative of a normal distribution with negative skewe, and in no way is able to show that we have two peaks, one around 90, the other 140 to 160.

The question said to answer in one sentance, you used 2. I put :
Since the data is bimodal, the histogram describes the data better as the median and quartiles are not meaningful because they incorrectly describe the data assume the data as unimodal. (I admit it is very wordy, but it is hard to get all your thoughts in one sentance when you are under time pressure.)

Rewording your answer to one sentance:
The histogram best highlights the bimodal nature of the data, while the boxplot indicates a slightly negativly skewed distribution and ignores the two peaks.

>Q.8
On a hospital maternity ward, babies are born with a particular abnormality at a frequency of 1 per 500 births.
(a) What distribution would you use to model the random variable X, the number of births between one abnormal birth up to and including the next?

N¬Geometric(1/500)

(b) What is the mean and standard deviation of X?

mean= 500 std dev = 499.5

>Q.9
Let X denote a binomial B(4, 0.3) random variable. Calculate the following probabilities:
(a) P(X = 0)
(b) P(X = 1)
(c) P(X > 1)

P(X=0) 0.2401
P(X=1) 0.4116
P(X>=1) 0.3483

>Q.10
The c.d.f. of the Pareto probability distribution with parameters K and theta is given by:
F(x) = 1 – (K/theta)^theta, x >=K
Where K > 0 and theta > 1. Calculate the 0.975 quartile of the Pareto distribution with parameters K = 50 and theta = 2.

To find the .975 quantile, set the formula equal to .975. K and Θ are given. Then you need to solve for x using algebra.

.975= 1 - (50/ x)^2
.025 = (50/x)^2
sqrt(.025) = 50/x
x= 50 / sqrt(.025)
x = 316

316.228

Q.11
(a) Binomial distribution. If the drug group is X, then X ~ B(110,p)
If the placebo group is Y, then Y ~ B(102,p)

(b) Assumptions made:
1. Trials are independent
2. Outcome of each trial is either a success (in remission) or a failure (not in remission).
3. Probability of a successful outcome is the same for each trial.
4. Finite number of trials.

(c)
To find p for drug group, X: p = 42/110 = 0.382
for placebo group, Y: p = 12/102 = 0.118

Thus, X ~ B(110,0.382) and Y~ B(102,0.118)

Q.12
(a) The discrete uniform distribution would be a suitable model, as the blockage is equally likely to have occurred under any of the 20 houses.

(b) mean = (n + 1)/2 = (20 + 1) / 2 = 10.5

variance = (n^2 - 1)/12 = (20^2 +1)/12 = 33.25

(c) As there are 4 x 15 minutes in one hour, the probability is P(X <= 4):

P(X <= 4) = 1/20 + 1/20 + 1/20 + 1/20 + 1/20 = 0.25

I get .2 using F(X) = x/n. I admit I could be using the wrong forumla, so let me know what you think.

I actually think we were both wrong. I had a look in my A level book (getting well-thumbed by now) and this is my revised solution:

the distribution is uniform continuous between 1 and 20, thus X ~ U(1,20).
The formula in the handbook f(x) = 1/(b-a) gives 1/(20 - 1) = 1/19.

If 1 house is checked in 15 mins, then in an hour 4 houses are checked.
Thus P(X <= 4) = 1/19 x (4 - 1) = 1/19 x 3 = 0.159 (to 3 d.p.)

Why uniform continuous? This is uniform discrete - you cant have 2.3 houses or 2.4 houses.

No your orig answer correct

I wouldn't have said so. To my thinking the probability for the first house is 1/20 which makes the next a 1/19 chance and so on. I would have though the logical answer would be 1/20 + 1/19 +1/18 +1/17 = 0.217

Finally, if it were indeed right that you get the probability that the blockage is under one of the first 4 houses as 1/20 + 1/19 + 1/18 + 1/17, then it would also be true that the probability of the blockage being under one of the first 20 houses is 1/20 + 1/19 + …1/2 + 1, and that is clearly a lot more than 1 (I make it about 3.6). Probabilities can't be greater than 1 so this can't be right.

Q.13
(not sure about this at all!)

mean(x) = 2, mean(y)=4 var(x)=4 var(y)=16

(a) sum X + Y:
X ~ M(2) and Y ~ M(4), so X + Y ~ M(2 + 4) = M(6)

so mean = 1/lambda = 1/6 and variance = 1/lambda^2 = 1/6^2 = 1/36

E(S) = E(X)+E(Y) = 6
V(S) = V(X) + V(Y) = 20

(b) difference X - Y:
X - Y ~ M(2 - 4) = M(-2)

so mean = 1/-2 = -0.5 and variance = 1/lambda^2 = 1/4

E(S) = E(X)-E(Y) = -2
V(S) = V(X) + V(Y) = 20

ODER
E[X+Y] = 2+4=6 V[X+Y]= 4+16=22
E[X-Y] = 2-4=-2 v[X-Y]=4+16=22

Q.14
Not sure how to do this. My instinct is a 'No' answer, though.

No beacuse for a Poisson distribution the mean is equal to the variance. Here, mean = 10.2, variance = 65.61

ODER
For Poisson mean = variance ; not the case here

Q.15
X ~ Poisson(3)
P(X > 1) = 1 - P(X < 1)
= 1 - P(X = 0)
= 1 - exp(-3)
= 1 - 0.049787
= 0.9502 (to 4 d.p.)

I get .0000 as P(X>1) = 1-P(X)<=1 = 1-(P(X)=0+(P(X)=1))
You put (X>1) = 1-P(X)<1 = 1- P(X)=0

Q.16
(a) X ~ N(6,12^2)

P(X > 0) = P(Z > (x - u)/sigma) = P(Z > (0 - 6)/12) = P(Z > -0.5) = P(Z < 0.5)

= phi(0.5) (using Table 2)
= 0.6915.

(b) P(0 < X < 2) = P((0 - 6)/12 < Z < (2 - 6)/12) = P(-0.5 < Z < -0.33)

= phi(0.5) - phi(0.33) = 0.6915 - 0.6293 = 0.0622 ODER 0.0609

Q.17
Got into problems with this:
(a)
X ~ N(99, 5^2/100) n = 100, mu = 99, sigma = 5

*X~N(100*99+5^2*100)*
= X~(9900,2500)

(b) I kg = 1000 g
so mean = 99/1000 = 0.099 kg
and SD = 5/1000 = 0.005 kg

mean = 9.9 kg
sd = sqrt(2500)/1000 = .050

ODER:
N(9.9,0.05^2)

(c)
got in a tangle here and didn't finish.

P(X > 10) =.0228

Q.18 & Q.19 Went blank and couldn't work them out!!

18.a) E(S) = u(lorry) + u(other) = 5+20=25
Poisson(25)

b) S~N(25,25)
P(x>30) = .1587

*19. T~M(1/3) so T~(Poisson(1/3*4) = T~(Poisson(4/3*)*
P(X=0) = .2636

Q.20
20 out of 100 contaminated, 20/100 = 0.2, so:

p hat = 1/X bar = 1/0.2 = 5 (using a geometric distribution)

The question didn't say it was a geometric distribution.
p max lik = p sample
= 1-(1-9)^10 = 20/100
== 1-(1-9)^10 = .2
= (1-9)^10 = .8
1 p − .9779
p= .02207

Q.21
No, it is not correct. The max likelihood estimate of theta occurs at the peak of the graph and the value is read off the x axis, so theta is approx. 0.575. (Was something extra required here??)

Q.22
95% of times the parameter will lie in the confidence interval (31.45,35.66).

Q.23
Find using: (u+, u-) = (x bar - Z s/sqrt n, x bar + Z s/sqrt n)
= (1.992 - 2.576 x 1.943/sqrt 122, 1.992 + 2.576 x 1.943/sqrt 122)
= (1.992 - 0.453, 1.992 + 0.453)
= (1.539, 2.445)

you accidently used variance instead of the standard deviation in the formula, but since you showed your work I think they probably would have given you 2 out of the 3 points (judging on how they treat that kind of error on TMA's).
(1.667,2.317)

Q.24
Don't know the answer to this one, but guess that the Z point is larger.

The t one is larger. For very large samples, t approaches z. The larger the sample is, the more likely it is to be representative of the population. t is a more conservative value then z, allowing for more variation and is therefore usually more appropriate.

Q.25
(a) The means for 2 groups of data is firstly calculated. The variance for the same groups is then calculated. The variances are within a factor of 3 of each other, which meets the requirement for doing a two sample t-test. The test compares the means of 2 populations, with null hypothesis that the means are the same. The SP value is very small so the null hypothesis is rejected in favour of the alternative hypothesis, that the means are not the same.

I added that one had to assume the variation in the populations were normally distributed.

(b)
The number of degrees of freedom is 48, so the sum of the 2 samples would be:
(n1 + n2 - 2) = 48
(n1 + n2) = 48 + 2 = 50

Q.26
This sort of question just leaves my brain spinning, so didn't write anything particularly intelligible here, I'm afraid.

An SP gives the percent of experiments that would give less support for the null hypothesis that the current sample. The null hypothesis is not "impossible" and is not necessarily "false". A low SP gives evidence to reject a null hypothesis, but is not a 100% guarentee that it is wrong as samples have sample variation.

Q.27
1. The size of differences is not taken into account in the sign test, cf Wilcoxon.
2. The sign test often fails to reject the null hypothesis in all but the most obvious cases.

Q.28
This test compares the observed frequencies (Oi) with frequencies expected (Ei) under a hypothesised model (Poisson distribution in this case).

1. The data are allocated to k categories, so that the Ei value in each category is at least 5. For instance, the Defects values for groups 7, 8 and 9 each may have Ei < 5 when calculated. If they do, they are pooled into the Defects group 6, so that Ei > 5.

2. The chi squared test statistic is used (quote formula from handbook). The number of degrees of freedom is k - p - 1. where p is the number of parameters that required estimation in order to calculate the Ei values. In the example above, k = 7 (if the last 4 groups are pooled), p = 1, so the degrees of freedom = 5.

3. The SP is given by the upper tail probability of chi squared(5). A very large or very small Sp value would indicate that the null hypothesis should be rejected.

I would add use Poisson(smple mean) to estimate the expected).
Since the mean is estimated, degrees of freedom = k-2. A very small SP would lead to rejection. A high SP would not.

Q.29
(a) The value 3 does not fall within the 95% c.i. (2.56,2.98), thus it is in the rejection region for the test, and we would reject the null hypothesis and accept the alternative hypothesis that mu is not equal to 3.

(b)
No idea about this one.

According to the first researcher, the SP was less then .05 as 3 was not in the 95% confidence interval. Therefore, one of the researchers made a mistake.

Q.30
(a) A straight line drawn through the origin would seem to give an adequate fit, just by looking at these data points with the eye. Most of the data points would lie reasonably close to the line.

The straight line would not go through the origin. The graph does not start at 0 for the y axis.

(b) y = 133.56 + (21.65 x 1000) = 21,783,56 = 21,783 (to 5 s.f.)

The number of TV sets would increase to 21,783.

*since x is already in thousands, y = 133.56+21.65*1 = 155.21*

(c) Don't know how to calculate this.

I think this is right:
(Bmin, Bmax) = Bhat +or - t s/(sqrt(sum(xi=meanx)^2)*
=21.65+-2.262(32.74/9.05559)*
=(-13.47,29.03).
Notice that 0 is in the confidence interval so it is possible the true slope is 0.

Q.31
The prediction interval is longer, as it must allow for individual variation.

Q.32
(a) The message is that cases of TB are on the increase again (since 1987) and so people should be more aware of the risk of catching TB.

(b) If the points are connected by a curve, rather than 2 straight lines, it could be argued that TB cases have reached the lowest level yet.

Q.34
(a) expected value = (29 x 22) / 55 = 11.6

observed value of 8 was lower than the expected value of 11.6 so fewer Control patients were Better than would have been expected.

(b) degrees of freedom = ((r - 1)(c - 1)) = 2

From Table 6, chi squared(2) at 0.90 level = 4.61

The test statistic value of 4.175 is below this, indicating the result is below the 10% confidence level. I would conclude there is not enough evidence to give a possible association between these variables

the SP is between .1 and .5 (closer to .1)

(c) No association means that the use of treatment does not produce a better DSS value than the control.

Q.35
for t = 3, f(t) = 1 - exp(-lambda x 3) = I - exp(-0.5 x 3) = 1 - exp(-1.5) = 1 - 0.22313 = 0.78 (to 2 d.p)

The question asks for the distribution, not p
*X~Poisson(2*3)= X~Poisson(6)*

Q.36
(a) M hat = T [202/566 364/566] = T [0.357 0.643]
not T [364/1005 641/1005] not T [0.362 0.638]

(b) I would conclude that the sequence of runs in the DNS sequence was produced by a Bernoulli model, as there is no evidence to reject it.

I added that the SP was .8414

ANHANG ZU FRAGE 12 C

HI everyone and thanks, Kevin for the detailed reply,

I received a reply from my tutor, as follows:

>I don't think there is a clear cut answer here as the question may be ambiguous.
>
>In parts (a) and (b) X, is the number of the house where the blockage has occurred.
>
>X has a discrete uniform distribution
>P(X=x) = 1/20, x=1,2,3,...,20.
>
>Not X~U(1,20) which is the notation for the continuous uniform.
>
>The answer to part (c) is P(Y<=4), I think we can agree but what is the distribution of Y here. It depends I suppose on how we >interpret the question (the argument is between 'House Number' and 'Number of Houses'). If you use the uniform cdf for Y, then as before >Y=X is the 'House Number' from 1 to 20 where the blockage has occurred. i.e. we are assuming the engineers begin at house 'Number 1' and >then check 'Number 2', 'Number 3' etc in order until the blockage is found. The geometric approach makes Y the number of houses checked
>(out of 20) until the blockage is located, so we are not interested now in the 'House number', or the order in which the houses are >checked.
>
>I would guess that either approach would be OK here, subject to a correct statement of what you are assuming the question asks (always a >good idea if you are unsure of what is being asked) - but my feeling is that the second (geometric) appproach was intended.

I think the last part is sage advice - if you're not sure what's being asked, say why you're doing what you're doing! That said, I still think the answer is p=0.2!

Best,

Kevin McConway writes:
Peter J. Whitelaw,oufcnt3.open.ac.uk writes:
Regarding 2000 / Q12(c):

I did the Q2000 paper this morning and have the benefit of model answers from the London Region M246 tutors. Hopefully this may lay the contention regarding Q12 (c) to rest. It may add kindling to the fire however...

In Q12 (c) I, too, assumed X~U(1,20) i.e. discrete uniform and therefore the for P(X=<4) in 12 (c) we should apply the cdf for discrete uniform i.e. probability = 4/20 = 0.2

Can I jump in here? As some of you know, I do work for the OU (and am exam board chair for M346), but I haven't been involved in M246 since the first year of the course so I'm not saying anything 'official' here. So I might have got this wrong. But I think Peter (and others who have given the same answer as above) is actually correct.

However, the model solution takes a different approach and I think the answer is in the question. It says nothing about where the engineer starts e.g. at house no 1. The chance of each house having the blockage is Bernoulli(0.05). Consequently the waiting time before the first success is Y~Geometric(0.05). Y is the number of houses checked until the blockage is found.

The model parameters are therefore p=0.05 (blocked), q=0.95 (not blocked)

So we are seeking P(Y=<4) = 1 - (not blocked)^4 = 1 - (0.95)^4 = 0.1855.

I'm afraid I don't think this is right. In order to use the Geometric, the events that there is a blockage under each house must be independent Bernoulli trials, with the same probability of success for each one (see box on p. 119 of the course book). Here the trials do all have two possible outcomes (blockage or no blockage), and they all have the same probability of 'success', but they aren't independent. This is because the question tells you that there is one blockage in the sewer. So, for example, the events 'blockage under first house' and 'blockage under second house' are not independent. If you know the first one has occurred, then you know the second one can't occur. Thus the geometric model is wrong, I'm afraid (well, I think it is!).

Another way to see that is to note that, if the geometric model is right, then the probability that there is no blockage under any of the 20 houses is q^20 = 0.95^20 = 0.358. But that can't be right because we know there is exactly one blockage under the 20 houses. In fact, if the individual houses were independent Bernoulli trials, then the number of blockages in the street is binomial, B(20, 0.05), and there could in principle be any number of blockages in the street between 0 and 20 inclusive. But we're told that isn't the case. (I suppose one could just about interpret the question to mean "there is at least one blockage" --- but even then, the geometric can't be right because it gives nonzero probability to an event that hasn't occurred, i.e. that there is no blockage.)

(The geometric model would, I suppose, be right if the engineers inspect the houses at random, with no regard to whether they've already looked under a house. They just pick one at random, check under it for a blockage, then pick another one at random, which might or might not be be the same house again, and inspect again, and so on. In this case, the fact that they never pay any attention to what they did before makes the trials independent. But I think it's going very far out on a limb to interpret the question in this way. It's true that the question doesn't say that they start at number 1 and work up the street; but as long as they don't allow themselves to go back and inspect under a house they've already done, the order doesn't matter; you still want the probability that the blockage is under one of 4 houses out of 20, so it's 4/20 = 0.2.)

I appreciate the approach but, to be frank, I don't see why the cdf approach is wrong. I have e-mailed my tutor for further details and will post further details if/when they are forthcoming.

As I said, I don't think the cdf approach that you (and others) used is wrong. But I'll also be interested to see what your tutor said.

Oh, while I'm on, I agree with Ed Mulligan's explanation of why the 1/19, 1/18 and so on are the wrong probabilities to consider. Ed wasn't saying (I think) that you have to multiply probabilities together to answer this question. He was just showing that there's two ways to get at the probability that the second house has a blockage under it. One way, the short way, is to simply say that the blockage is equally likely to be under any of the houses (because the question says so), there are 20 of them, so the probability that it's under any given one of them (including the second) is 1/20. Another way is indeed to take account of the fact that the probability the blockage is under the second, given that it's not under the first, is 1/19. But that's not the probability we want. We want the (unconditional) probability that it's under the second. To find that, we also need the probability that there is no blockage under the first house, which is 19/20. Then:

Pr (blockage under 2nd) = pr(blockage under 2nd, given blockage not under first) x pr (blockage not under first)

= 1/19 x 19/20
=1/20.
This gives the same answer as the short way, but it's longer-winded.
That's (I think) all Ed was saying.

Finally, it it were indeed right that you get the probability that the blockage is under one of the first four houses as 1/20 + 1/19 + 1/18 + 1/17, then it would also be true that the probability of the blockage being under one of the first 20 houses is 1/20 + 1/19 + ... + 1/2 + 1, and that is clearly a lot more than 1. (I make it about 3.6.) Probabilities can't be greater than 1 so this can't be right.

Hope this helps, and good luck in the exam.

Regards,

Kevin
>
>
>All the best,
>
>Peter
>pjw@helion.net
>
>1999 MST121, MS221
>2000 M203
>2001 MST207
>2002 M246