Feedback — Homework 6

Help Center

Thank you. Your submission for this homework was received.

You submitted this homework on **Sun 12 Apr 2015 10:25 PM PDT**. You got a score of **8.00** out of **8.00**.

Question 1

This interactive question is intended to shed light on whether polling mechanisms can be adapted to real-life settings were answers may be unreliable because individuals do not wish to reveal their true opinions because of peer or social pressure, or perhaps because of fear of censure from authorities, or perhaps simply out of a desire to avoid embarrassment in publicly espousing an opinion that may be unfashionable or considered to be "un-intellectual". In such settings how can data be collected while preserving anonymity?

In this exercise we will set up a questionnaire by utilising randomisation in such a way as to guarantee a degree of anonymity: the polling agent (in this case a Google poll put together by your benevolent chefs) will have no way of telling whether anyone's negative opinion is really their own. The purpose of this exercise is not to suggest that the highly stylised model that we will look at is in any way implementable outside of a classroom environment (very unlikely). Rather, the purpose of this exercise is to show how statistical inferences can be made from unreliable data. After all the data are collected at the end of the week we will provide an analysis of this problem together with estimates of the proportions of the respondents in this cohort belonging to various belief groups though we will not be able to identify who has what belief individually.

We will pick a hot-button topic that is current, that of climate change. A variety of scientists weighing in on the subject have opined that there is evidence to show that climate change is accelerating. But, on the other side of the debate, there have been scientists and policy makers from a variety of fronts who have questioned the conclusion in public forums. A recent retraction of portions of a United Nations climate change report have raised additional questions in the minds of laymen who, based on their own experiences and observations, may have their own opinions on the score, for or against. So, this is the setting.,

Please follow the link below to access a questionnaire soliciting your opinions on four issues with regard to the issue of climate change. Each question has two possible answers: "YES" or "NO". But before you answer each question, we ask you to flip a coin: if the coin lands on heads then we want you to answer the question truthfully and if the coin lands on tails then we want you to answer "NO" to the question, regardless of your true opinion. Do this four times, once for each of the four issue questions. A key element in this survey is that the methodology preserves anonymity: negative answers do not necessarily indicate that the respondent has that opinion. We will provide an analysis of the class data in the solutions next week.

Click here to go to the question.

Once you've answered the questions of the survey, think about the problem for a bit. Which of the following possibilities do you think is most likely to be true? (There is no penalty for an incorrect answer.)

Your Answer		Score	Explanation
○ You can estimate the fraction of negative opinions with the same error and confidence guarantees as when all questions are answered truthfully and there are no coin flips.			
◉ You can still estimate the fraction of negative opinions but for a given error, the level of confidence would be lower than when all questions are answered truthfully and there are no coin flips.	✔	1.00	
○ It is not possible to accurately estimate the fraction of negative opinions that are held by the members of this cohort from this randomised data.			
Total		1.00 / 1.00	

Question Explanation

This problem provides a (simplistic) example of privacy-preserving statistical queries. In the setting we envisage, the opinions, positive or negative, yes or no, of a group of individuals is sought in a context where a publicly expressed negative opinion may be considered to be “embarrassing” or may even face social opprobrium. In such cases it may be hard to get truthful answers to surveys from individuals who are worried about the consequences of publicly stating their true opinions. A naïve survey which simply tabulates respondent opinions then is likely to produce estimates that are not at all accurate. In settings where there are significant policy consequences such as the prevalence of disease agents, undocumented immigration, attitudes towards science and religion, and such like, poor estimates can lead to very poor policy and possibly significant social and economic cost. The goal in such settings is to estimate the proportion of people who truly have negative opinions, while controlling the unreliability in responses in a systematic fashion by making provision for anonymity within a structure from which reliable information about the group can be extracted while concealing individual detail.

In our example the coin flip mechanism is designed as a mechanism to create ambiguity over negative answers. Of course, this is a highly stylised

experiment and you will be readily able to see difficulties with implementing it outside of a classroom environment. The objective in this exercise is to show how we can, in principle, design experiments to deal with the desire of a public for anonymity by expanding our purview to unreliable responses.

In this problem, when the coin lands head, the respondent reveals his belief truthfully (either yes or no), and when the coin lands tails, the respondent always says “no”. Therefore, if a person says “yes”, we are sure that his real answer is “yes”. However, if he says “no”, nothing can be inferred, since a “no” answer may be truthful (if the coins has landed heads) or not truthful (if the coin has landed tails).

So, let's begin with an analysis.

The sample space: In our sanitised model we're dealing with a sequence of repeated independent trials so it will suffice to understand the sample space and measure attendant on each trial. What is the sample space for a single trial? Well, there are two chance elements: the true opinion of a randomly selected individual from the population and the result of the coin toss. Let Y denote the opinion of the selected individual writing 1 for a positive opinion and 0 for a negative opinion, and let Z denote the result of the coin toss, as usual identifying 1 with heads and 0 with tails. The sample space hence has four elements with $(Y, Z) \in \{(0,0), (0,1), (1,0), (1,1)\}$. Let p_0 denote the (fixed but unknown) fraction of individuals in the population with positive opinions and $q_0 = 1 - p_0$ the corresponding fraction of individuals with negative opinions. We wish to estimate q_0; this is equivalent to estimating p_0 as the estimate error will be the same in magnitude but going in the opposite direction.

The probability measure: As Y and Z are independent variables, the mass function attached to the atoms may be written down almost by inspection by multiplying probabilities:

$$\begin{aligned}
\mathbf{P}\{Y = 0, Z = 0\} &= q_0 \cdot \tfrac{1}{2}, \\
\mathbf{P}\{Y = 1, Z = 0\} &= p_0 \cdot \tfrac{1}{2}, \\
\mathbf{P}\{Y = 0, Z = 1\} &= q_0 \cdot \tfrac{1}{2}, \\
\mathbf{P}\{Y = 1, Z = 1\} &= p_0 \cdot \tfrac{1}{2}.
\end{aligned}$$

The response X of the individual depends both on his privately held opinion and the result of the toss of the coin. The outcomes may be systematically tabulated as shown in the following table.

		Z	
	X	0	1
Y	0	0	0
	1	0	1

The response X is a Bernoulli variable which inherits its distribution from Y and Z. As $X = 1$ if, and only if, both Y and Z are equal to 1, we see that

$$\begin{aligned}
p := \mathbf{P}\{X = 1\} &= \mathbf{P}\{Y = 1, Z = 1\} = p_0 \cdot \tfrac{1}{2}, \\
q := 1 - p = \mathbf{P}\{X = 0\} &= 1 - p_0 \cdot \tfrac{1}{2}.
\end{aligned}$$

Or, in short, $X \sim \text{Bernoulli}\left(\frac{p_0}{2}\right)$. We're now ready to deal with a random sample of responses.

Suppose $X_1, \ldots, X_n$ is a random sample of responses governed by a sequence of $\text{Bernoulli}(p)$ trials where $p = p_0/2$. Given an error tolerance $\epsilon > 0$ and a desired confidence $1 - \delta$, we wish to generate an estimate $\widehat{p}_0 = \widehat{p}_)(X_1, \ldots, X_n)$ of p_0 as a function of the sample so that

$$\mathbf{P}\big\{|\widehat{p}_0 - p_0| > \epsilon\big\} < \delta.$$

This looks like it is setting up nicely for the law of large numbers. What do we know? As usual, write $S_n = X_1 + \cdots + X_n$. We know by the law of large numbers that S_n/n concentrates nicely at $p = p_0/2$. So this suggest that we should consider the estimate $\widehat{p}_0 = 2S_n/n$. Now let's massage our requirement in the form where Chebyshev's inequality (or better, the de Moivre–Laplace theorem) can weigh in. Algebraic manipulation is now all that is needed to cast the probability of the "bad' event in a more familiar form:

$$\mathbf{P}\big\{|\widehat{p}_0 - p_0| > \epsilon\big\} = \mathbf{P}\big\{\big|\tfrac{2S_n}{n} - p_0\big| > \epsilon\big\} = \mathbf{P}\big\{\big|\tfrac{S_n}{n} - \tfrac{p_0}{2}\big| > \tfrac{\epsilon}{2}\big\} = \mathbf{P}\big\{\big|\tfrac{S_n}{n} - p\big| > \tfrac{\epsilon}{2}\big\}.$$

This is all set up for Chebyshev's inequality (see Lectures 11.1: f and g): all we will need to do is identify ϵ with $\epsilon/2$. We thus have an upper bound for the probability on the right given by

$$\frac{1}{4n\left(\frac{\epsilon}{2}\right)^2} = \frac{1}{n\epsilon^2}$$

and it will suffice for this to be $\leq \delta$ to be within our error and confidence requirements. Thus, if $n \geq \frac{1}{\epsilon^2 \delta}$ then we can be assured that the absolute value of the error in estimating p_0 by $2S_n/n$ is no more than ϵ with confidence at least $1 - \delta$. *Remarkably, we can still provide reliable and accurate estimates of the underlying subpopulation proportions but we pay a price for the unreliability in the responses in a four-fold increase in sample size*. The reason, of course, is that if we want to be within a desired error tolerance for the estimate of the sub-population proportion then, the unreliability in responses forces a much tighter error estimate (two-fold better) for the estimate of the response Bernoulli parameter. Using the de Moivre–Laplace theorem improves the bottom line but we still end up paying the four-fold penalty in increased sample size; again this is because ϵ has to be replaced by $\epsilon/2$ in all our equations. The following table shows sample values.

ϵ	$1 - \delta$	n via Chebshev	n via de Moivre–Laplace
0.03	0.95	8000	1540
0.03	0.95	22 224	4272
0.01	0.99	1000 000	90 000

The moral? The price to be paid for lack of reliability is not trivial.

Survey data: 590 responses

1. Do you feel threatened by the effects of global warming / climate change? Yes: 226 (38.3%). No: 356 (60.3%).

2. Do you believe climate change is being worsened by man-made efforts, and that changing your habits can help reduce the rate of climate change? Yes: 288 (48.8%). No: 296 (50.2%).

3. Do you believe that the ozone layer is thinning due to ozone-depletion substances like CFCs which are released into the atmosphere? Yes: 285 (48.3%). No: 299 (50.7%).

4. Do you believe the polar ice caps are melting due to global temperature increases? Yes: 311 (52.7%). No: 273 (46.3%).

Based on this data, what can you conclude now, in view of our analysis, about the fraction of the class who have a privately held negative response to each of these four questions? Among other things you may wish to consider whether the sample was significant enough and whether there is a likelihood of bias.

Question 2

You are told that a coin is tossed a fixed number of times and the number of successes tracked. You do not know how many times n the coin was tossed nor the success probability p of the coin. Suppose S represents the number of successes obtained in these tosses. Using Chebyshev's inequality, estimate the probability of the event that the absolute value of the difference between S and its expected value exceeds three times the standard deviation of S. [*Hint: You will need to reinterpret terms in our derivation of Chebyshev's inequality.*]

Your Answer		Score	Explanation
$\frac{4}{9}$			
$\frac{1}{4}$			
$\frac{1}{3}$			
$\frac{1}{9}$	✔	1.00	
$\frac{1}{2}$			
0			
Total		1.00 / 1.00	

Question Explanation

Write $S = S_n$ to make explicit the dependence on n. We are interested in the event that $\big|S_n - \mathbf{E}(S_n)\big| > 3\sqrt{\mathrm{Var}(S_n)}$. In view of Lecture 10.1: c, we know that S_n has the binomial distribution with parameters n and p, whence, by Lecture 10.1: o, we know that $\mathbf{E}(S_n) = np$ and, by Lecture 10.1: p, that $\mathrm{Var}(S_n) = npq$ where $q = 1 - p$. By Chebyshev's inequality (replace $n\epsilon$ by $3\sqrt{npq}$ in Lecture 11.1: f and follow the line of argument through), we conclude that

$$\mathbf{P}\big\{\big|S_n - \mathbf{E}(S_n)\big| > 3\sqrt{\mathrm{Var}(S_n)}\,\big\} = \mathbf{P}\big\{|S_n - np| > 3\sqrt{npq}\,\big\} \leq \frac{npq}{\big(3\sqrt{npq}\big)^2} = \frac{1}{9}.$$

Interpreting the result in words, the probability that the number of successes differs from its expected value by more than three standard deviations is less than 10%. This gives us ammunition in support of the *three sigma* rule of thumb that you may have encountered in the news in discussions of statistical data: with relatively high probability, the number of outcomes lies within three standard deviations of the mean.

Can we refine our calculation a little? You may be tempted to try a normal approximation to evaluate the probability of this event using the theorem of de Moivre and Laplace (Lecture 10.3: d). If you tried it you would obtain a speculative approximation

$$\mathbf{P}\big\{|S_n - np| > 3\sqrt{npq}\,\big\} = \mathbf{P}\left\{\left|\frac{S_n - np}{\sqrt{npq}}\right| > 3\right\} = \mathbf{P}\big\{|S_n^*| > 3\big\} \stackrel{?}{\approx} \int_{-\infty}^{-3} \phi(x)\,dx + \int_{3}^{\infty} \phi(x)\,dx = 2\int_{-\infty}^{-3} \phi(x)\,dx = 2\Phi(-3) = 0.0026998\cdots.$$

We've used the notation introduced in Lecture 11.3: c for the normal distribution function $\Phi(\cdot)$ with the standardised variable $S_n^* = (S_n - np)/\sqrt{npq}$ viewing the binomial in the proper centre and scale introduced in Lectures 11.3: b and d. The region of integration is determined by the inequality $|x| > 3$ to be comprised of the two intervals $(-\infty, -3)$ and $(3, \infty)$; and the symmetry of the bell curve (formally, $\phi(x)$ is an *even* function) shows that the area under the curve of the normal density in the left tail $(-\infty, -3)$ is the same as the area under the curve in the right tail $(3, \infty)$.

The normal approximation suggests that the bound obtained by Chebyshev's inequality is much too loose. The probability that the number of successes differs from its expected value by more than three standard deviations seems to be in actuality less than 0.3% giving an even stronger basis for the *three sigma* rule of thumb! This is alluring but we cannot directly utilise it in our problem as n is not known: if the number of trials is small then normal approximation would be poor as the asymptotics of the central limit theorem would not have kicked in.

This leads to new questions on attempting to estimate the error in normal approximation when the number of trials is finite. These questions lead to fertile new directions of inquiry. Roughly speaking, the error in normal approximation decreases proportionally to $n^{-1/2}$. This is the content of the famous *Berry–Esséen theorem*.

Question 3

The following prompt should be used for **Questions 3, 4, and 5**.

In problems involving genders we typically assume that male and female genders are equally likely. In other words, we assume that births constitute Bernoulli trials corresponding to the toss of a fair coin where the probability of giving birth to a boy and the probability of giving birth to a girl are both $1/2$. In such a model one would expect that the population of men and women should be roughly the same. This is typically reflected in census data in societies but there are places where there appears to be a surprising gender imbalance.

For example, the *missing women of Asia* refers to a United Nations estimate that there are about 106 men to every 100 women in some parts of Asia. This is statistically significant and is not easily explained as due to chance fluctuations.

One hypothesis to explain this discrepancy is that in some societal groups family units feel pressured to have a boy, and so keep having children until a male child is born. Our goal in this sequence of problems is to see whether societal pressure of this nature can explain an observed discrepancy in male and female populations.

So, for these problems we consider a model where each child born has equal probability of being male or female. We will suppose that each family keeps having children indefinitely until a boy is born (at which point the parents have no more children).

What is the probability that a family has *exactly* two children?

Your Answer		Score	Explanation
$\frac{1}{8}$			
$\frac{1}{2}$			
$\frac{1}{3}$			
$\frac{3}{4}$			
$\frac{1}{4}$	✔	1.00	
$\frac{7}{8}$			
Total		1.00 / 1.00	

Question Explanation

Let us enumerate all the information in this question:

(1) There are **106 men** to every **100 women**.
(2) $\mathbf{P}$**(Girl) =** $\mathbf{P}$**(Boy) =** $\frac{1}{2}$.
(3) **Assumption**: A family keeps having children until they have a boy, at which point they stop having children.

What is the question asking for? Here the question is asking for the probability that a family would have exactly two children. First note that the question information outlined as item (1) above is irrelevant to the solution of this problem.

Now, in calculating the probability of having exactly two children, we first need to identify the sample space. Remember the assumption noted in (3), a family will keep having children until they have a boy. Let us represent a girl using the letter G and a boy using the letter B. The sample space will have the form $\{GGG \ldots B\}$ where there can be any number of G's before the first and only B. The question is interested in the event, let us call it event A, where $A = \{GB\}$.

The direct way of calculating the probability of having exactly two children is to calculate the probability that event A = $\{GB\}$ occurs. Thus, we are solving for $\mathbf{P}(A)$.

Solution 1:
$\mathbf{P}(A)$ = $\mathbf{P}$(first child is a girl) $\times$ $\mathbf{P}$(second child is a boy) = $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.
This calculation is valid since each child birth is an independent event. The topic of independence will be discussed in more depth later on in the course.

Solution 2:
Another way of calculating $\mathbf{P}(A)$, the probability of having exactly two children, uses the event's compliment A^c which is the probability of *not* having two children: $A^c = \{B, GGB, GGGB, GGGGB, GG \ldots B, \ldots\}$. This occurs when either the first child is a boy, or there are at least two girls before the first boy. Note that the event A^c includes all possible events *except* the event $A = \{GB\}$.

We know the probability that the first child is a boy is $\frac{1}{2}$. Therefore, we look at the probability of the first event listed in A^c: $\mathbf{P}(\{B\}) = \frac{1}{2}$.
The remaining events that make up A^c are events in which there are at least two girls. The probability of of this can be calculated by looking at the probability of the first two children being a girl: $\mathbf{P}(\{GGB, GGGB, GGGGB, GG \ldots B, \ldots\} = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

Since the events $\{B\}$ and $\{GGB, GGGB, GGGGB, GG \ldots B, \ldots\}$ are mutually exclusive, additivity shows us that:
$\mathbf{P}(A^c) = \mathbf{P}(\{B\}) + \mathbf{P}(\{GGB, GGGB, GGGGB, GG \ldots B, \ldots\}) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$.
From total probability, we know that $\mathbf{P}(A) + \mathbf{P}(A^c) = 1$.
Therefore, $\mathbf{P}(A) = 1 - \mathbf{P}(A^c) = 1 - \frac{3}{4} = \frac{1}{4}$.

Question 4

What is the probability that a family has *at most* four children?

Your Answer		Score	Explanation
$\frac{15}{16}$	✔	1.00	
$\frac{7}{8}$			
$\frac{3}{4}$			

$\frac{1}{8}$	
$\frac{1}{16}$	
$\frac{1}{4}$	
Total	1.00 / 1.00

Question Explanation

Again, let's identify the event of interest as: A = {four children or less}, and the event's compliment: A^c = {five or more children}.

Solution 1:
We could compute the probability of A by noting that:$\mathbf{P}(A) = \mathbf{P}$(the family has a single child) + $\mathbf{P}$(the family has exactly two children) + $\mathbf{P}$(the family has exactly three children) + $\mathbf{P}$(the family has exactly four children).
This would also give us the answer using the calculation:
$\frac{1}{2} + (\frac{1}{2})^2 + (\frac{1}{2})^3 + (\frac{1}{2})^4 = \frac{15}{16}$.

Solution 2:
Starting with A^c, we can see that the probability of A^c can be written as the probability of all events such that the first four children are girls.
Therefore, $\mathbf{P}(A^c) = \sum_{i=0}^{\infty} (\frac{1}{2})^{5+i}$, which is a geometric series that can be computed using the geometric sum formula: $\frac{(\frac{1}{2})^5}{1-\frac{1}{2}} = (\frac{1}{2})^4$.
This can be calculated more easily upon the realization that the probability of the event A^c, having five or more children, is the same as the probability of the event of having four girls. Hence:
$\mathbf{P}(A^c) = \mathbf{P}(\{\text{having four girls}\}) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$.
Thus, $\mathbf{P}(A) = 1 - \mathbf{P}(A^c) = 1 - \frac{1}{2^4} = \frac{15}{16}$.

Question 5

What is the probability that a family has *more girls* than boys?

Your Answer		Score	Explanation
$\frac{1}{3}$			
$\frac{1}{4}$	✔	1.00	
$\frac{3}{4}$			
$\frac{7}{8}$			
$\frac{1}{8}$			
$\frac{1}{2}$			
Total		1.00 / 1.00	

Question Explanation

Solution 1:
Remember that a family will continue having children until they have a boy. This means that they will have **at most** one boy. Therefore, the number of girls will definitely be greater than the number of boys as long as the first two children are girls: $\mathbf{P}\{\text{first two children are girls}\} = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

Solution 2:
We can also look at this problem from the point of view of A^c. A^c = {family does not have more girls than boys} = {family has more boys than girls OR an equal number of boys and girls}. Using this logic, you can see that $A^c = \{B, GB\}$. From earlier, we know that $\mathbf{P}(\{B\}) = \frac{1}{2}$, and that $\mathbf{P}(\{GB\}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.
Since the events are mutually exclusive (it is not possible to have only one boy AND have one girl and one boy), we can use additivity to find that:
$\mathbf{P}(A^c) = \mathbf{P}(\{B\}) + \mathbf{P}(\{GB\}) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$.
Thus from total probability, $\mathbf{P}(A) = 1 - \mathbf{P}(A^c) = 1 - \frac{3}{4} = \frac{1}{4}$.

Question 6

The following prompt should be used for **Questions 6, 7, and 8**. The results build one atop the other so do them in sequence: you will need the result of Question 6 to tackle Question 7, and the result of Question 7 to tackle Question 8.

The mechanism for copying chromosomes (the hereditary carriers of genetic information) in cells is subject to errors (breakages) some of which can be repaired by cellular processes. Suppose that there are N breakages in transcribing a given chromosome. We model N as a Poisson random variable with expectation λ. Suppose that each breakage is repaired with probability p *independently of the repair of other breakages*. Let K denote the number of breakages that are repaired.

Suppose that n and k are given non-negative integers with $k \leq n$. What is the probability that $K = k$ breakages are repaired *given* that $N = n$ breakages have occurred?

Your Answer		Score	Explanation
$\left(e^{-\lambda}\,\frac{\lambda^n}{n!}\right)^k$			
$\binom{n}{k}p^k(1-p)^{n-k}$	✔	1.00	
$\binom{n}{k}p^k(1-p)^{n-k}\cdot e^{-\lambda}\,\frac{\lambda^n}{n!}$			
$\frac{\lambda^n}{n!}\,e^{-\lambda}\cdot p^k(1-p)$			
$p^k(1-p)^{n-k}$			
$\binom{n}{k}p^k e^{-k\lambda}(1-p)^{n-k}$			
Total		1.00 / 1.00	

Question Explanation

Given that there are n breakages, each break is repaired independently with probability p. In other words, conditioned on there being n breakages, the repair of each break is a Bernoulli trial with success probability p. Given n trials, the accumulated number of successes may be identified with the number of repairs K and so, conditioned on $N = n$, the number K of repairs is binomially distributed with parameters n and p. It follows that

$$\mathbf{P}\{K = k \mid N = n\} = \binom{n}{k}p^k(1-p)^{n-k}.$$

Question 7

For any given non-negative integer k, what is the (unconditional) probability that exactly $K = k$ breakages are repaired?

[*Hint: You're going to need additivity in the guise of the theorem of total probability. If k breakages are repaired, what can you say about the possible range of values n for the total number of breakages? You will find it useful to write $\lambda^n = \lambda^k \cdot \lambda^{n-k}$ in simplifying the resultant series. The exponential series formula*

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{i=0}^{\infty}\frac{x^i}{i!}$$

will be useful in simplifying your sum.]

Your Answer		Score	Explanation
$\frac{\lambda^k}{k!}\,e^{-\lambda}$			
$\frac{\lambda^k}{k!}\,(1-p)^k e^{-\lambda}$			
$\frac{\lambda^k}{k!}$			
$\frac{p\lambda^k}{k!}\,(1-p)^k e^{-\lambda}$			
$e^{-\lambda p}\,\frac{(\lambda p)^k}{k!}$	✔	1.00	
$e^{-\lambda(1-p)}\,\frac{\lambda^k(1-p)^k}{k!}$			
Total		1.00 / 1.00	

Question Explanation

The number of breakages cannot be smaller than the number of repairs and so, for a given k, the number n of breakages in total must satisfy $n \geq k$. Write $q = 1 - p$ as usual and leverage the result of Question 6: by conditioning on the number of breakages, via total probability [see Lecture 8.2: g] we obtain

$$\mathbf{P}\{K=k\} = \sum_{n=k}^{\infty} \mathbf{P}\{K=k \mid N=n\}\mathbf{P}\{N=n\} = \sum_{n=k}^{\infty} \binom{n}{k} p^k q^{n-k} \cdot e^{-\lambda} \frac{\lambda^n}{n!}$$
$$= \sum_{n=k}^{\infty} \frac{n!}{k!(n-k)!} p^k q^{n-k} \cdot e^{-\lambda} \frac{\lambda^k \lambda^{n-k}}{n!}$$
$$= e^{-\lambda} \frac{(\lambda p)^k}{k!} \sum_{n=k}^{\infty} \frac{(\lambda q)^{n-k}}{(n-k)!}$$
$$= e^{-\lambda} \frac{(\lambda p)^k}{k!} \left[1 + (\lambda q) + \frac{(\lambda q)^2}{2!} + \frac{(\lambda q)^3}{3!} + \cdots\right]$$
$$= e^{-\lambda} \frac{(\lambda p)^k}{k!} \cdot e^{\lambda q} = e^{-\lambda+\lambda q} \frac{(\lambda p)^k}{k!} = e^{-\lambda(1-q)} \frac{(\lambda p)^k}{k!} = e^{-\lambda p} \frac{(\lambda p)^k}{k!}.$$

We've discovered that the (unconditional) distribution of the number of repairs is Poisson with parameter λp. The student may find this, in retrospect, to be very satisfyingly intuitive.

Question 8

What is the expected number of repaired breakages? [*Hint: Examine your result from Question 7. Do you recognise the mass function?*]

Your Answer		Score	Explanation
λe^{λ}			
pe^{λ}			
e^{λ}			
$\lambda e^{-\lambda}$			
$\frac{1}{p}$			
λp	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

We've seen in Question 7 that the number K of repairs is Poisson with parameter λp. It follows that $\mathbf{E}(K) = \lambda p$. [See Lecture 10.2: n for the expectation of the Poisson distribution.]