Problem 1. A government wants to find out how many of its citizens are filling out fraudulent tax returns. It creates a survey with the single question “Do you fill out your tax return honestly?” Along with the question come the following Instructions: Toss a fair coin. If the result of the coin toss is heads, answer the question truthfully. If the result of the coin toss is tails, answer the question with NO regardless of whether you cheat or not.

Randomising the response makes it impossible to deduce for any single individual that they cheat. The result of the survey is that 45% of the respondents say YES. Suppose we interpret this as meaning that a randomly chosen member of the population will answer the survey YES with probability 0.45. Find the proportion of the population that fills out fraudulent returns.

Solution. Suppose the proportion of people who cheat is \( p \). The only people who answer YES are those who are honest and get heads. The probability that a random person is honest is \( 1 - p \), and the probability of heads is \( \frac{1}{2} \). These are independent, so \( \frac{1-p}{2} = 0.45 \), so \( p = 0.1 \).

Problem 2. Consider \( n \) flips of a coin. A run is a sequence of consecutive tosses with the same result. For \( k < n \), let \( E_k \) be the event that a run is completed at time \( k \); this means that the results of the \( k \)th and \( (k+1) \)st flips are different. For example, if \( n = 10 \) and the outcomes of the first 10 flips are HHHTTHHTTH then runs are completed at times 3, 5, 7, 9.

(a) Show that if the coin is fair, then the events \( E_k \), \( 1 \leq k \leq n \) are independent. (This requires you to show that \( P(\cap_{i=1}^{m} E_{k_i}) = \prod_{i=1}^{m} P(E_{k_i}) \) for every choice of \( m \leq n \) and \( k_1 < \ldots < k_m \).)

(b) Show that if the coin is not fair, the events are not independent. (An unfair coin gives H with probability \( p \neq \frac{1}{2} \).

Solution. (a) Consider some increasing sequence \( k_1 < k_2 < \ldots \). If we condition on \( E_{k_1} \ldots E_{k_{i-1}} \), this tells us something about the coins up to \( k_{i-1} + 1 \leq k_i \). No matter what these coins show, the probability that the \( k_i + 1 \) coin is different from the \( k_i \) coin is \( \frac{1}{2} \). Therefore, the probability of \( E_{k_i} \) conditioned on \( E_{k_1} \ldots E_{k_{i-1}} \) is \( \frac{1}{2} \). We use the formula

\[
P(\cap_{i=1}^{m} E_{k_i}) = \prod_{i=1}^{m} P(E_{k_i} | E_{k_1 \ldots k_{i-1}}).
\]

Since each term on the product is \( \frac{1}{2} \), we get

\[
P(\cap_{i=1}^{m} E_{k_i}) = \prod_{i=1}^{m} \frac{1}{2} = \prod_{i=1}^{m} P(E_{k_i}).
\]

Since this holds for any sequence of indices, the events are independent.
(b) To show the events are not independent it is enough to show that \( P(E_1E_2) \neq P(E_1)P(E_2) \). The intersection is \( \{HTH, THT\} \) with probability \( p^2(1-p) + p(1-p)^2 = p(1-p) \). However, \( P(E_1) = P(E_2) = 2p(1-p) \). Therefore \( E_1, E_2 \) are independent only if \( p(1-p) = (2p(1-p))^2 \). The solutions to this are only \( p = 0, 1, \frac{1}{3} \). Therefore the events are only independent if the coin is fair, or if it is always heads or always tails.

**Problem 3.** Eye colour is determined by a single gene, and each individual has two copies of the gene. For simplicity, we assume there are only blue and brown eyes. If both genes of an individual are blue, they have blue eyes; Otherwise they have brown eyes. (In technical terms, blue eyes are *recessive* and brown eyes are dominant.) A child inherits randomly one copy of the gene from each parent. (See e.g. season 4 of *The bridge* for the significance of these facts.)

(a) Suppose Fanny and her parents have brown eyes but her brother Alexander has blue eyes. What is the conditional probability that Fanny has two brown genes?

(b) Suppose moreover that Fanny has 3 children with Lars, who has blue eyes, and all 3 have brown eyes. What is the conditional probability of the same event?

**Solution.** (a) Fanny and Alexander’s parents must both have one brown eye and one blue eye genes. Thus Fanny has either blue-brown, brown-blue, or brown-brown. Let \( A \) The conditional probability that she has two brown genes is \( \frac{1}{3} \).

(b) Let \( A_2 \) be the event that Fanny has two brown eye genes, and \( A_1 \) that she has brown-blue genes. We have \( P(A_2) = \frac{1}{3} \) and \( P(A_1) = \frac{2}{3} \) from part (a). Let \( B_3 \) be the event that three children all have brown eyes. We have \( P(B_3|A_2) = 1 \) and \( P(B_3|A_1) = \left(\frac{1}{2}\right)^3 \). Apply Bayes’ formula to find

\[
P(A_2|B) = \frac{\frac{1}{3} \cdot 1}{\frac{1}{3} \cdot 1 + \frac{2}{3} \cdot \frac{1}{8}} = \frac{4}{5}.
\]

**Problem 4.** Two hockey teams, A and B play a series of games, until one of the teams wins 4 games. Suppose team A has probability \( p \) of winning each game, and games are independent. Let \( X \) be the total number of games that are played.

(a) Find the probability mass function of \( X \).

(b) What is the probability that team A wins the series conditioned on \( X = 4 \)?

(c) What is the probability that team A wins the series conditioned on \( X = 7 \)?

(Simplify your expressions as much as possible.)

**Solution.** (a) Let \( q = 1 - p \).

\[
p(4) = p^4 + q^4 \quad \quad p(6) = 10p^4q^2 + 10p^2q^4
\]

\[
p(5) = 4p^4q + 4pq^4 \quad \quad p(7) = 20p^4q^3 + 20p^3q^4 = 20p^3q^3.
\]

In each case the first term is the probability that A wins and the second that B wins. For example, to find \( p(6) \), in order for team A to win in 6 games, they must lead 3-2 after 5 games (\( \binom{5}{2} = 10 \) ways for this to happen) and then win the last game. Each such sequence has probability \( p^4q^2 \).
(b) This is \( \frac{p^4}{p^4 + q^4} \).
(c) This is \( \frac{20p^4q^3}{20p^4 + q^4} = p \). We can see this directly, since if the series lasts 7 games, then after 6 games the score is 3-3, and the last game determines the winner.

**Problem 5.** We wish to select between two options \( A, B \) with probability \( \frac{1}{2} \) each. We are given a coin which comes up heads with an unknown probability \( p \). Show that the following procedure works: Toss the coin twice. If the results are \((H, T)\) pick \( A \). If the results are \((T, H)\) pick \( B \). Otherwise start again.

**Solution.** In each attempt we get result \( A \) with probability \( a = p(1 - p) \) and \( B \) with probability \( a \), and otherwise we try again. It is intuitively clear that eventually we get either \( A \) or \( B \) and the two have equal probability. To make this precise, let \( A_n \) be the event that we get \( A \) in the \( n \)th attempt. Then \( P(A_n) = a(1 - 2a)^{n-1} \). Adding these up we have \( P(A) = \sum_{n=1}^{\infty} P(A_n) = \frac{1}{2} \). (Using \( \sum (1-x)^n = \frac{1}{x} \).) Similarly \( P(B) = \frac{1}{2} \).

Note that the probability that we finish on the \( n \)th attempt is \( 2a(1 - 2a)^{n-1} \), so the number of attempts is geometric with parameter \( 2a \).

**Problem 6.** A fair die is rolled four times.
(a) Let \( Y \) denote the number of distinct rolls. Find the probability mass function of \( Y \).
(b) Let \( Z \) denote the minimal result for the 4 throws. Find the probability mass function of \( Z \).

**Solution.** (a) We have

\[
p(1) = \frac{6}{6^4} \quad p(2) = \frac{14 \cdot \binom{6}{2}}{6^4} \quad p(3) = \frac{\binom{6}{3} \cdot 3 \cdot 12}{6^4} \quad p(4) = \frac{6 \cdot 5 \cdot 4 \cdot 3}{6^4}.
\]

For example, there are \( \binom{6}{2} \) ways to choose which two values appear, and there are 14 sequences where the two specific values appear and no others \( (2^4 \text{ where no others appear, minus the two where only one of them appears}) \). Similarly, if 3 values appear, there are \( \binom{6}{3} \) ways to pick which three values appear, 3 ways to specify which of them appears twice, and 12 ways to arrange the throws with given outcomes \( \{a, a, b, c\} \).
(b) We have \( p(k) = \frac{(7-k)^4 - (6-k)^4}{6^4} \) for \( k = 1, 2, \ldots, 6 \). This is since there are \( (7-k)^4 \) sequences where the minimum is at least \( k \), and we subtract \( (6-k)^4 \) where the minimum is at least \( k + 1 \).

**Problem 7.** Coding component:
(a) Simulate 100,000 geometric random variables with parameter \( p = 0.01 \) and create a histogram of the resulting values, with buckets for each of the values 1 to 1000.
(b) Next, create a plot of the probability mass function of the same geometric random variable, over the integers 1 to 1000. Briefly describe how this plot compares to the histogram from part (a).
Finally, plot the function $f(t) = e^{-t}$ for $t$ between 0 and 10. Briefly compare this plot to the two plots above.
Submit your code, plots, and written answers to the questions in (b),(c). (Depending on your computer, the plots may take a few minutes to appear.)

Solution. Sample code:

```python
import numpy as np
import matplotlib.pyplot as plt

geom = np.random.geometric
X_arr = [geom(.01) for i in range(100000)]

H = [0]*1001
for x in X_arr:
    if x<= 1000: H[x] += 1

plt.figure()
plt.plot(H)
plt.title('Histogram of 100000 Geom(.01) random variables')
plt.savefig('hw2_q7a.pdf')

plt.figure()
plt.plot([.01 * .99**n for n in range(1,1000)])
plt.title('pmf of the Geom(.01) random variable')
plt.savefig('hw2_q7b.pdf')

plt.figure()
X = np.arange(0,10,.001)
Y=np.exp(-X)
plt.plot(X,Y)
plt.title('plot of $e^{-t}$')
plt.savefig('hw2_q7c.pdf')
```

The resulting graphs look almost the same. The difference between the first two is only the y-axis, which has a ratio of 100000 (the number of variables). The difference between the second and third is a scale ratio of $100 = 1/p$ in both axes. The reason was discussed in class on Friday in detail.
Histogram of 10000 Geom(.01) random variables

PMF of the Geom(.01) random variable

Plot of $e^{-t}$