For extra practice, several additional review problems are printed below. Solutions to these problems can be found on the exams page of the course website. I've tried to sort the problems by whether I think they are more appropriate as an in-class question or as a take-home question. While these questions are representative of the typical scope and difficulty of individual exam questions, this review is not comprehensive, nor does it necessarily represent the total amount of time available for the exam.

In-class Exam.

(1) Determine whether the following statement is always true or sometimes false (If true, explain why. If false, give a counter-example):

If A, B, C are events so that A and B are conditionally independent given C, then A and B are independent.

Solution. This statement is false. Suppose we have two coins: a fair coin, and a double headed coin. A coin is selected uniformly at random and flipped twice. Let A be the event that the first flip is heads, let B be the event that the second flip is heads, and let C be the event that the chosen coin is fair. Then A and B are conditionally independent given C (since once the coin has been chosen, flips are independent), but A and B are not independent, since if A occurs, it makes it more likely that C did not occur, and therefore, that B occurs as well.

(2) Suppose $X \sim \text{Bin}(8, 0.25)$ and let $Y = 2X + 1$. Find and simplify a formula for the PMF of Y. Then compute the expected value of Y.

Solution. The PMF for Y is given by

$$p_Y(k) = P(Y = k) = P(2X + 1 = k) = P(X = (k-1)/2) = p_X((k-1)/2)$$

which is valid for $k = \{1, 3, 5, \ldots, 17\}$. Using linearity of expectation,

$$E[Y] = E[2X + 1] = 2E[X] + 1 = 2 \cdot 8 \cdot 0.25 + 1 = 5$$

(3) A box has three coins in it. One has heads on both sides, one has tails on both sides, and one is a fair coin. A coin is selected uniformly at random and flipped twice.

(a) If the result of the first flip is heads, what is the probability that the coin is two-headed?
(b) If the result of the first flip is heads, what is the probability that the second flip is also heads?
(c) If the result of both flips are heads, what is the probability that the coin is two-headed?

Solution. (a) Let A be the event that the first flip is heads and let C be the event that the double-headed coin is chosen, let D be the event that the doubled-tailed coin is chosen, and let F be the event that the fair coin is chosen. Then by Bayes' Rule and LoTP

$$\begin{aligned} P(C|A) &= P(A|C)\frac{P(C)}{P(A)} \\ &= \frac{P(A|C)P(C)}{P(A|C)P(C) + P(A|D)P(D) + P(A|F)P(F)} \\ &= \frac{1 \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} \\ &= \frac{2}{3} \end{aligned}$$

(b) Let B be the event that the second flip is heads. Then by LoTP and the result of part (a):

$$\begin{aligned} P(B|A) &= P(B|A, C)P(C|A) + P(B|A, D)P(D|A) + P(B|A, F)P(F|A) \\ &= P(B|C)P(C|A) + P(B|D)P(D|A) + P(B|F)P(F|A) \\ &= 1 \cdot \frac{2}{3} + 0 \cdot 0 + \frac{1}{2} \cdot \frac{1}{3} \\ &= \frac{5}{6} \end{aligned}$$

where $P(C^c|A) = 1 - P(C|A)$, since conditional probabilities are probabilities.

(c) Using Bayes' Rule,

$$\begin{aligned} P(C|A,B) =& P(A,B|C)\frac{P(C)}{P(A,B)} \\ =& \frac{P(A,B|C)P(C)}{P(A,B|C)P(C) + P(A,B|D)P(D) + P(A,B|F)P(F)} \\ =& \frac{1 \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot 0} \\ =& \frac{4}{5} \end{aligned}$$

Since C implies the event $A \cap B$, and since A and B are conditionally independent given C.

(4) Suppose A and B are events. Use axioms of probability to show that if $A \subseteq B$, then $P(A) \leq P(B)$.

Solution. Write the event B as the union of disjoint events:

$$B = (B \cap A) \cup (B \cap A^c) = A \cup (B \cap A^c)$$

since $A \subseteq B$. Then, as the probability of the union of disjoint events is the sum of the probabilities of those events,

$$P(B) = P(A) + P(B \cap A^c)$$

Since $P(B \cap A^c) \geq 0$, then $P(B) \geq P(A)$.

(5) Two teams, the Alligators and the Bears, play a series of games, where the Alligators have probability p of winning each game (independently). The winner of the series is the first team to win 2 more games than their opponent. Compute the expected number of games in the series.

Solution. Let $q = 1 - p$ and consider a sequence of independent Bernoulli trials corresponding to pairs of consecutive games (i.e. pair the first and second game, and the third and fourth game, and so on), where success occurs if one team wins both games. Note that the series is over exactly when the first success occurs. The probability that the Alligators win two consecutive games is p^2 and the probability that the Bears win two consecutive games is q^2, and so the probability that either team wins two consecutive games is $p^2 + q^2$.

The number of pairs of games in the series is therefore $\text{FS}(p^2 + q^2)$ and so the expected number of **pairs** of games is $1/(p^2 + q^2)$. Since every pair consists of two games, then the expected number of games is $2/(p^2 + q^2)$.

(6) Suppose A and B are events with $0 < P(B) < 1$. Show that if $P(A|B) > P(A|B^c)$, then

$$P(A|B) > P(A) > P(A|B^c).$$

Solution. Suppose $P(A|B) > P(A|B^c)$. By LoTP

$$\begin{aligned} P(A) =& P(A|B)P(B) + P(A|B^c)P(B^c) \\ <& P(A|B)P(B) + P(A|B)P(B^c) \\ =& P(A|B)(P(B) + P(B^c)) \\ =& P(A|B) \end{aligned}$$

Similarly,

$$\begin{aligned} P(A) =& P(A|B)P(B) + P(A|B^c)P(B^c) \\ >& P(A|B^c)P(B) + P(A|B^c)P(B^c) \\ =& P(A|B^c)(P(B) + P(B^c)) \\ =& P(A|B^c) \end{aligned}$$

Hence, $P(A|B) > P(A) > P(A|B^c)$ as desired.

(7) Define a function p on the set $D = \{1, 2, \ldots, 9\}$ by

$$p(k) = \log_{10}(k+1) - \log_{10}(k)$$

(a) Verify that p is a valid PMF function.
(b) Suppose X is a random variable with PMF p. Find a formula for the CDF F of X.
(c) Let $Y \sim \text{DUnif}(D)$, independent of X. What is the probability that $Y = X$?

Solution. (a) In order for p to be a valid PMF:

- $p(k) \geq 0$ for all k
- $\sum_k p(k) = 1$

To verify the first property, note that $\log_{10}(x)$ is an increasing function for $x > 0$, and so $\log_{10}(k+1) > \log_{10}(k)$, and so $p(k) > 0$.
To see that the second property holds, note the following sum of values of p is telescoping, and so

$$\sum_{k=1}^{9} p(k) = \sum_{k=1}^{9} \log_{10}(k+1) - \log_{10}(k) = \log_{10}(10) - \log_{10}(1) = 1 - 0 = 1$$

(b) Again, using the fact that the sum of values of p is telescoping, if $x \in \{1, \ldots, 9\}$, then

$$F_X(x) = P(X \leq x) = \sum_{k=1}^{x} p(x) = \log_{10}(x) - \log_{10}(1) = \log_{10}(x)$$

More generally, for arbitrary real numbers x, then

$$F_X(x) = \begin{cases} \log_{10}(\lfloor x \rfloor), & \text{if } 1 \leq x \leq 9, \\ 1, & \text{if } x \geq 9, \\ 0, & \text{if } x \leq 1. \end{cases}$$

where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x.
(c) Using LoTP and the fact that X, Y are independent, then

$$\begin{aligned} P(X = Y) &= \sum_{k=1}^{9} P(X = Y | Y = k) P(Y = k) \\ &= \sum_{k=1}^{9} P(X = k | Y = k) P(Y = k) \\ &= \sum_{k=1}^{9} P(X = k) P(Y = k) \\ &= \frac{1}{9} \sum_{k=1}^{9} P(X = k) \\ &= \frac{1}{9} \end{aligned}$$

(8) On a certain statistics test (not necessarily this one!), 10 out of 100 key terms will be randomly selected to appear on an exam. A student then must choose 7 of these 10 to define. Since the student knows the format of the exam in advance, the student is trying to decide how many of these terms to memorize.
(a) Suppose the student studies s key terms, where s is an integer between 0 and 100. What is the distribution of X? Give the name and parameters of the distribution, in terms of s.

(b) Write an expression involving the explicit formula for the PMF or CDF of the named distribution for the probability that the student knows at least 7 of the 10 key terms, assuming the student studies $s = 75$ key terms.

Solution. (a) Treat the terms you study as "tagged" terms, and the terms you do not study as "untagged" terms. Then the number of key terms on the exam that you've studied is $\text{HGeom}(s, 100 - s, 10)$.

(b) The probability that you know at least 7 of the terms is

$$P(X \geq 7) = 1 - P(X \leq 6) = 1 - \sum_{k=0}^{6} \frac{\binom{75}{k}\binom{25}{10-k}}{\binom{100}{10}}$$

which is approximately 0.785 (although you weren't asked to numerically approximate for this problem).

Take-home Exam.

(1) A box contains a total of N marbles, colored blue, white and orange. Let b, w, o denote the proportions of blue, white and orange marbles, respectively.
   (a) Marbles are drawn 1-by-1 randomly with replacement. Find the probability that the first time a blue marble is drawn is before the first time a white marble is drawn.
   (b) Suppose instead that marbles are drawn 1-by-1 randomly without replacement. Find the probability that the first time a blue marble is drawn is before the first time a white marble is drawn.

Solution. (a) Label marbles as either orange or non-orange (i.e. blue or white). Then the event "the first time a blue marble is drawn is before the first time a white marble is drawn" is the same as the event "the first non-orange marble drawn is blue". Since we are sampling with replacement, then

$$\begin{aligned} P(\text{ first non-orange is blue }) &= P(\text{marble is blue} \mid \text{marble is non-orange }) \\ &= \frac{P(\text{marble is blue })}{P(\text{ marble is non-orange })} \\ &= \frac{b}{b+w} \end{aligned}$$

(b) Observe that there bN blue marbles, wN white marbles and oN orange marbles. Line up the marbles in the order they are drawn and identify the first non-range marble. By symmetry, it is equally likely to be any of the $bN + wN$ non-orange marbles, and so the probability that it is blue is

$$\frac{bN}{bN + wN} = \frac{b}{b+w}$$

which is identical to the answer in the previous part.

(2) Suppose you are answering a multiple-choice problem on an exam, and have to choose one of n options (exactly one of which is correct). Let K be the event that you know the answer and let R be the event that you get the problem right (either through knowledge or through lucky guessing). Suppose that if you know the right answer, you will definitely get the problem right, but if you do not know the answer, you will guess completely randomly. Let $P(K) = p$.
   (a) Find $P(K|R)$ in terms of p and n.
   (b) Show that $P(K|R) \geq p$ and explain why this makes sense intuitively. When (if ever) does $P(K|R) = p$?

Solution. (a) By Bayes' Rule and the Law of Total Probability,

$$P(K|R) = \frac{P(R|K)P(K)}{P(R|K)P(K) + P(R|K^c)P(K^c)} = \frac{p}{p + (1-p)/n}$$

(b) By the previous part, the statement $P(K|R) \geq p$ is equivalent to the statement

$$p + (1-p)/n \leq 1$$

But for any n, $(1-p)/n \leq 1-p$, and so

$$p + (1-p)/n \leq p + (1-p) = 1$$

as desired.

Intuitively, since you always get the correct answer when you know the answer, and only sometimes get the correct answer when you don't know the correct answer, getting the correct answer should make it more likely that you were in the case when you knew the answer than when you didn't.

By the preceding calculation, $P(K|R) = p$ when $n = 1$, or when $p = 0$, or when $p = 1$.

(3) Suppose n individuals in a large population have a rare disease. Two medical tests are available to detect the disease. Suppose the first test detects the disease with probability p_1, while the second test detects it with probability p_2. Assume that detection is independent between individuals, as well as between tests. Let X_1 be the number of individuals the first test detects, let X_2 be the number the second test detects, and let X be the number detected by at least one of the tests.
(a) Find the distribution of X.
(b) Assume for this part that $p_1 = p_2$. Find the conditional distribution of X_1 given that $X_1+X_2 = t$.

Solution. (a) Give each individual an ID number from 1 to n and consider the ordered list of individuals as a sequence of Bernoulli trials, where success in each trial indicates that at least 1 of the tests detected the disease in that individual. The probability of success in each trial is $1-(1-p_1)(1-p_2)$ (since the trial is a failure only when both tests do not detect the disease). Then X is the number of successes in this sequence of n Bernoulli trials, and so $X \sim \text{Bin}(n, 1-(1-p_1)(1-p_2))$.

(b) By Theorem 3.9.2, the variable $X_1|(X_1+X_2=t)$ is $\text{HGeom}(n,n,t)$. Alternatively, by Bayes' Rule, let $p = p_1 = p_2$ and let $T = X_1 + X_2$, which is $\text{Bin}(2n,p)$. Then

$$\begin{aligned} P(X_1 = k|T=t) &= \frac{P(T=t|X_1=k)(X_1=k)}{P(T=t)} \\ &= \frac{\binom{n}{t-k}p^t q^{n-t+k} \cdot \binom{n}{k}p^k q^{n-k}}{\binom{2n}{t}p^t q^{2n-t}} \\ &= \frac{\binom{n}{t-k}\binom{n}{k}}{\binom{2n}{t}} \end{aligned}$$

for $0 \le k \le t$.

(4) Consider a daily lottery in which 5 of the integers from 1 to 50 are chosen without replacement.
(a) Find the probability that you guess exactly 3 numbers correctly, given that you guess at least 1 of the numbers correctly.
(b) Find an exact expression for the expected number of lotteries needed so that all $\binom{50}{5}$ possible lottery outcomes have occurred.

Solution. (a) The distribution of the number of correct guesses is $\text{HGeom}(5,45,5)$; think of the winning lotto numbers as the white balls and the non-winning numbers as the black balls. A person's guesses represent a sample of 5 balls from the 50 balls total. Then

$$P(\text{ exactly 3 correct } | \text{ at least 1 correct }) = \frac{P(\text{ exactly 3 correct })}{1-P(\text{ exactly 0 correct })} = \frac{\binom{5}{3}\binom{45}{2}/\binom{50}{5}}{1-\binom{5}{0}\binom{50}{5}/\binom{50}{5}} \approx 0.011$$

(b) Let $n = \binom{50}{5}$. Let T_1 be the number of days until the first new set of numbers appears (which is always 1), let T_2 be the number of days until the second new set of numbers appears, and so on. Then the total number of days required is $T = T_1 + T_2 + \cdots + T_n$. Observe that by the story of the first-success distribution,

$$N_k \sim \text{FS}((n-k+1)/n) \qquad 2 \le k \le n$$

and $E[N_k] = \frac{n}{n-k+1}$. Then by linearity of expectation,

$$E[N] = E[N_1] + E[N_2] + \cdots + E[N_n] = 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1} = n\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right)$$

Using R, this is approximately 32,085,564 days (or 87,906 years).

(5) A fair 6-sided die is rolled 6 times. What is the probability that at least 1 of the 6 values never appears? First find an exact formula for the solution, and then verify by writing an R function and performing a simulation to estimate the probability.

Solution. This problem is equivalent to the birthday problem, with a group of 6 students choosing birthdays from a list of 6 dates. Think of the 1st roll as the birthday of the 1st student, the 2nd roll as the birthday of the 2nd, and so on. The event "at least one number never appears" is the same as the event "at least 1 number appears more than once" which is equivalent to the event "there is at least one birthday match". The probability of at least one birthday match is

$$P(\text{ at least one birthday match }) = 1 - \frac{6!}{6^6} = 1 - \frac{5!}{6^5} \approx 0.9845679$$

In R, the following code can be used to simulate 1 experiment:

```
die_roll <- function(){
  n_unique <- length(unique(sample(6, size = 6, replace = T)))
  return(n_unique<6)
  }
```

Simulating this experiment 10,000 times and computing the number of times the result occurs gives an approximate probability of

```
mean(replicate(10000,die_roll()))
0.9822
```