Stat 110 Strategic Practice 4, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1 Distributions and Expected Values for Discrete Random Variables

1. Find an example of two discrete random variables X and Y (on the same sample space) such that X and Y have the same distribution (i.e., same PMF and same CDF), but the event $X = Y$ *never* occurs.

2. Let X be a random day of the week, coded so that Monday is 1, Tuesday is 2, etc. (so X takes values $1, 2, \ldots, 7$, with equal probabilities). Let Y be the next day after X (again represented as an integer between 1 and 7). Do X and Y have the same distribution? What is $P(X < Y)$?

3. A coin is tossed repeatedly until it lands Heads for the first time. Let X be the number of tosses that are required (including the toss that landed Heads), and let p be the probability of Heads. Find the CDF of X, and for $p = 1/2$ sketch its graph.

4. Are there discrete random variables X and Y such that $E(X) > 100E(Y)$ but Y is greater than X with probability at least 0.99?

5. Let X be a discrete r.v. with possible values 1,2,3,.... Let $F(x) = P(X \leq x)$ be the CDF of X. Show that

$$E(X) = \sum_{n=0}^{\infty} (1 - F(n)) .$$

   Hint: organize the order of summation carefully, using the fact that, for example, $P(X > 3) = P(X = 4) + P(X = 5) + \ldots$.

6. Job candidates $C_1, C_2, \ldots$ are interviewed one by one, and the interviewer compares them and keeps an updated list of rankings (if n candidates have been interviewed so far, this is a list of the n candidates, from best to worst). Assume that there is no limit on the number of candidates available, that for any n the candidates $C_1, C_2, \ldots, C_n$ are equally likely to arrive in any order, and that there are no ties in the rankings given by the interview.

Let X be the index of the first candidate to come along who ranks as better than the very first candidate C_1 (so C_X is better than C_1, but the candidates after 1 but prior to X (if any) are worse than C_1. For example, if C_2 and C_3 are worse than C_1 but C_4 is better than C_1, then $X = 4$. All 4! orderings of the first 4 candidates are equally likely, so it could have happened that the first candidate was the best out of the first 4 candidates, in which case $X > 4$.

What is $E(X)$ (which is a measure of how long, on average, the interviewer needs to wait to find someone better than the very first candidate)? Hint: find $P(X > n)$ by interpreting what $X > n$ says about how C_1 compares with other candidates, and then apply the result of the previous problem.

2 Indicator Random Variables and Linearity of Expectation

1. A group of 50 people are comparing their birthdays (as usual, assume their birthdays are independent, are not February 29, etc.). Find the expected number of pairs of people with the same birthday, and the expected number of days in the year on which at least two of these people were born.

2. A total of 20 bags of Haribo gummi bears are randomly distributed to the 20 students in a certain Stat 110 section. Each bag is obtained by a random student, and the outcomes of who gets which bag are independent. Find the average number of bags of gummi bears that the first three students get in total, and find the average number of students who get at least one bag.

3. There are 100 shoelaces in a box. At each stage, you pick two random ends and tie them together. Either this results in a longer shoelace (if the two ends came from different pieces), or it results in a loop (if the two ends came from the same piece). What are the expected number of steps until everything is in loops, and the expected number of loops after everything is in loops? (This is a famous interview problem; leave the latter answer as a sum.)

   Hint: for each step, create an indicator r.v. for whether a loop was created then, and note that the number of free ends goes down by 2 after each step.

4. A *hash table* is a commonly used data structure in computer science, allowing for fast information retrieval. For example, suppose we want to store some people's phone numbers. Assume that no two of the people have the same

name. For each name x, a *hash function* h is used, where $h(x)$ is the location to store x's phone number. After such a table has been computed, to look up x's phone number one just recomputes $h(x)$ and then looks up what is stored in that location.

The hash function h is deterministic, since we don't want to get different results every time we compute $h(x)$. But h is often chosen to be *pseudorandom*. For this problem, assume that true randomness is used. So let there be k people, with each person's phone number stored in a random location (independently), represented by an integer between 1 and n. It then might happen that one location has more than one phone number stored there, if two different people x and y end up with the same random location for their information to be stored.

Find the expected number of locations with no phone numbers stored, the expected number with exactly one phone number, and the expected number with more than one phone number (should these quantities add up to n?).

Stat 110 Strategic Practice 4 Solutions, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1 Distributions and Expected Values for Discrete Random Variables

1. Find an example of two discrete random variables X and Y (on the same sample space) such that X and Y have the same distribution (i.e., same PMF and same CDF), but the event $X = Y$ *never* occurs.

   For a simple example, let $X \sim \text{Bernoulli}(1/2)$ (i.e., X can be thought of as a fair coin flip), and let $Y = 1 - X$. Then Y is also Bernoulli(1/2) by symmetry, but $X = Y$ is impossible. A more general example is to let $X \sim \text{Binomial}(n, 1/2)$ and $Y = n - X$, where n is any odd number (think of this as interchanging the definitions of "success" and "failure"!).

2. Let X be a random day of the week, coded so that Monday is 1, Tuesday is 2, etc. (so X takes values $1, 2, \ldots, 7$, with equal probabilities). Let Y be the next day after X (again represented as an integer between 1 and 7). Do X and Y have the same distribution? What is $P(X < Y)$?

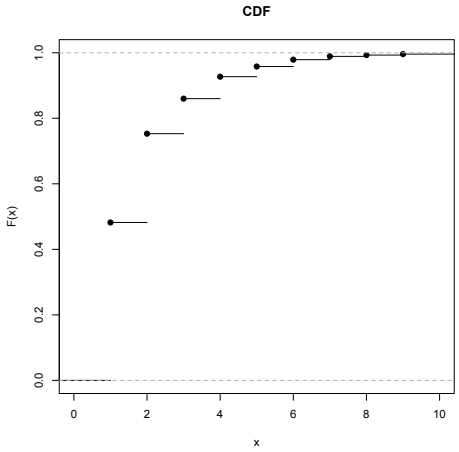
   They have the same distribution since Y is also equally likely to represent any day of the week, but $P(X < Y) = P(X \neq 7) = \frac{6}{7}$.

3. A coin is tossed repeatedly until it lands Heads for the first time. Let X be the number of tosses that are required (including the toss that landed Heads), and let p be the probability of Heads. Find the CDF of X, and for $p = 1/2$ sketch its graph.

   By the story of the Geometric, we have $X - 1 \sim \text{Geometric}(p)$. Using this or directly, the PMF is $P(X = k) = p(1-p)^{k-1}$ for $k \in \{1, 2, 3, \ldots\}$ (and 0 otherwise). The CDF can be obtained by adding up the PMF (from $k = 1$ to $k = \lfloor x \rfloor$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x). We can also see directly that

   $$P(X \leq x) = 1 - P(X > x) = 1 - (1-p)^{\lfloor x \rfloor}$$

   for $x \geq 1$, since $X > x$ says that the first $\lfloor x \rfloor$ flips land tails. The CDF is 0 for $x < 1$. For a fair coin, the CDF is $F(x) = 1 - \frac{1}{2^{\lfloor x \rfloor}}$ for $x \geq 1$, and $F(x) = 0$ for $x < 1$, as illustrated below.



4. Are there discrete random variables X and Y such that $E(X) > 100E(Y)$ but Y is greater than X with probability at least 0.99?

   Yes: consider what happens if we make X usually 0 but on rare occasions, X is extremely large (like the outcome of a lottery); Y, on the other hand, can be more moderate. For example, let X be 10^6 with probability $1/100$ and 0 with probability $99/100$, and let Y be the constant 1.

5. Let X be a discrete r.v. with possible values 1,2,3,.... Let $F(x) = P(X \leq x)$ be the CDF of X. Show that

$$E(X) = \sum_{n=0}^{\infty} (1 - F(n)) .$$

   Hint: organize the order of summation carefully, using the fact that, for example, $P(X > 3) = P(X = 4) + P(X = 5) + \ldots$.

   Note that

$$\sum_{n=0}^{\infty}(1 - F(n)) \;\; = \;\; \sum_{n=0}^{\infty} P(X > n) = \sum_{n=0}^{\infty} \sum_{k=n+1}^{\infty} P(X = k)$$

   For each k, the term $P(X = k)$ appears exactly k times: there is one $P(X = k)$ term for each nonnegative integer $n < k$. More visually, write out some terms:

$$P(X = 1) + P(X = 2) + P(X = 3) + \cdots +$$

$$P(X=2)+P(X=3)+P(X=4)+\cdots+$$
$$P(X=3)+P(X=4)+\cdots+$$
$$\cdots$$

Rearranging the terms of this series (which is allowed since the terms are non-negative),

$$\sum_{n=0}^{\infty}(1-F(n)) \;=\; \sum_{k=1}^{\infty} kP(X=k) = E[X].$$

6. Job candidates $C_1, C_2, \ldots$ are interviewed one by one, and the interviewer compares them and keeps an updated list of rankings (if n candidates have been interviewed so far, this is a list of the n candidates, from best to worst). Assume that there is no limit on the number of candidates available, that for any n the candidates $C_1, C_2, \ldots, C_n$ are equally likely to arrive in any order, and that there are no ties in the rankings given by the interview.

   Let X be the index of the first candidate to come along who ranks as better than the very first candidate C_1 (so C_X is better than C_1, but the candidates after 1 but prior to X (if any) are worse than C_1. For example, if C_2 and C_3 are worse than C_1 but C_4 is better than C_1, then $X = 4$. All 4! orderings of the first 4 candidates are equally likely, so it could have happened that the first candidate was the best out of the first 4 candidates, in which case $X > 4$.

   What is $E(X)$ (which is a measure of how long, on average, the interviewer needs to wait to find someone better than the very first candidate)? Hint: find $P(X > n)$ by interpreting what $X > n$ says about how C_1 compares with other candidates, and then apply the result of the previous problem.

   For $n \geq 2$, $P(X > n)$ is the probability that none of $C_2, C_3, \ldots, C_n$ are better candidates than C_1, i.e., the probability that the first candidate is the highest ranked out of the first n. Since any ordering of the first n candidates is equally likely, each of the first n is equally likely to be the highest ranked of the first n, so $P(X > n) = 1/n$. For $n = 0$ or $n = 1$, $P(X > n) = 1$ (note that it does not make sense to say the probability is $1/n$ when $n = 0$). Applying the result of the previous problem,

$$E(X) = \sum_{n=0}^{\infty} P(X > n) = P(X > 0) + \sum_{n=1}^{\infty} P(X > n) = 1 + \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

since the series is the *harmonic series*, which diverges.

How can the average waiting time to find someone better than the first candidate be infinite? In the real world, there are always only finitely many candidates so the expected waiting time is finite, just as in the St. Petersburg paradox there must in reality be an upper bound on the number of rounds. The harmonic series diverges very slowly, so even with millions of job candidates the average waiting time would not be very large.

2 Indicator Random Variables and Linearity of Expectation

1. A group of 50 people are comparing their birthdays (as usual, assume their birthdays are independent, are not February 29, etc.). Find the expected number of pairs of people with the same birthday, and the expected number of days in the year on which at least two of these people were born.

   Creating an indicator r.v. for each pair of people, we have that the expected number of pairs of people with the same birthday is $\binom{50}{2}\frac{1}{365}$ by linearity. Now create an indicator r.v. for each day of the year, taking the value 1 if at least two of the people were born that day (and 0 otherwise). Then the expected number of days on which at least two people were born is $365\left(1-(\frac{364}{365})^{50}-50\cdot\frac{1}{365}\cdot(\frac{364}{365})^{49}\right)$.

2. A total of 20 bags of Haribo gummi bears are randomly distributed to the 20 students in a certain Stat 110 section. Each bag is obtained by a random student, and the outcomes of who gets which bag are independent. Find the average number of bags of gummi bears that the first three students get in total, and find the average number of students who get at least one bag.

   Let X_j be the number of bags of gummi bears that the jth student gets, and let I_j be the indicator of $X_j \geq 1$. Then $X_j \sim \text{Bin}(20, \frac{1}{20})$, so $E(X_j) = 1$. So $E(X_1 + X_2 + X_3) = 3$ by linearity.

   The average number of students who get at least one bag is

$$E(I_1 + \cdots + I_{20}) = 20E(I_1) = 20P(I_1 = 1) = 20\left(1 - \left(\frac{19}{20}\right)^{20}\right).$$

3. There are 100 shoelaces in a box. At each stage, you pick two random ends and tie them together. Either this results in a longer shoelace (if the two ends came from different pieces), or it results in a loop (if the two ends came from the same piece). What are the expected number of steps until everything is in loops, and the expected number of loops after everything is in loops? (This is a famous interview problem; leave the latter answer as a sum.)

   Hint: for each step, create an indicator r.v. for whether a loop was created then, and note that the number of free ends goes down by 2 after each step.

   Initially there are 200 free ends. The number of free ends decreases by 2 each time since either two separate pieces are tied together, or a new loop is formed. So exactly 100 steps are always needed. Let I_j be the indicator r.v. for whether a new loop is formed at the jth step. At the time when there are n unlooped pieces (so $2n$ ends), the probability of forming a new loop is $\frac{n}{\binom{2n}{2}} = \frac{1}{2n-1}$ since any 2 ends are equally likely to be chosen, and there are n ways to pick both ends of 1 of the n pieces. By linearity, the expected number of loops is

$$\sum_{n=1}^{100} \frac{1}{2n-1}.$$

4. A *hash table* is a commonly used data structure in computer science, allowing for fast information retrieval. For example, suppose we want to store some people's phone numbers. Assume that no two of the people have the same name. For each name x, a *hash function* h is used, where $h(x)$ is the location to store x's phone number. After such a table has been computed, to look up x's phone number one just recomputes $h(x)$ and then looks up what is stored in that location.

   The hash function h is deterministic, since we don't want to get different results every time we compute $h(x)$. But h is often chosen to be *pseudorandom*. For this problem, assume that true randomness is used. So let there be k people, with each person's phone number stored in a random location (independently), represented by an integer between 1 and n. It then might happen that one location has more than one phone number stored there, if two different people x and y end up with the same random location for their information to be stored.

   Find the expected number of locations with no phone numbers stored, the

expected number with exactly one phone number, and the expected number with more than one phone number (should these quantities add up to n?).

Let I_j be an indicator random variable equal to 1 if the j^{th} location is empty, and 0 otherwise, for $1 \leq j \leq n$. Then $P(I_j = 1) = (1 - 1/n)^k$, since the phone numbers are stored in independent random locations. Then $I_1 + \cdots + I_n$ is the number of empty locations. By linearity of expectation, we have

$$E\left(\sum_{j=1}^{n} I_j\right) = \sum_{j=1}^{n} E(I_j) = n(1 - 1/n)^k.$$

Similarly, the probability of a specific location having exactly 1 phone number stored is $\frac{k}{n}(1-\frac{1}{n})^{k-1}$, so the expected number of such locations is $k(1-1/n)^{k-1}$. By linearity, the sum of the three expected values is n, so the expected number of locations with more than one phone number is $n-n(1-1/n)^k-k(1-1/n)^{k-1}$.

Stat 110 Homework 4, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1. Let X be a r.v. whose possible values are $0, 1, 2, \ldots$, with CDF F. In some countries, rather than using a CDF, the convention is to use the function G defined by $G(x) = P(X < x)$ to specify a distribution. Find a way to convert from F to G, i.e., if F is a known function show how to obtain $G(x)$ for all real x.

2. There are n eggs, each of which hatches a chick with probability p (independently). Each of these chicks survives with probability r, independently. What is the distribution of the number of chicks that hatch? What is the distribution of the number of chicks that survive? (Give the PMFs; also give the names of the distributions and their parameters, if they are distributions we have seen in class.)

3. A couple decides to keep having children until they have at least one boy and at least one girl, and then stop. Assume they never have twins, that the "trials" are independent with probability 1/2 of a boy, and that they are fertile enough to keep producing children indefinitely. What is the expected number of children?

4. Randomly, k distinguishable balls are placed into n distinguishable boxes, with all possibilities equally likely. Find the expected number of empty boxes.

5. A scientist wishes to study whether men or women are more likely to have a certain disease, or whether they are equally likely. A random sample of m women and n men is gathered, and each person is tested for the disease (assume for this problem that the test is completely accurate). The numbers of women and men in the sample who have the disease are X and Y respectively, with $X \sim \text{Bin}(m, p_1)$ and $Y \sim \text{Bin}(n, p_2)$. Here p_1 and p_2 are unknown, and we are interested in testing the "null hypothesis" $p_1 = p_2$.

(a) Consider a 2 by 2 table listing with rows corresponding to disease status and columns corresponding to gender, with each entry the count of how many people have that disease status and gender (so $m + n$ is the sum of all 4 entries). Suppose that it is observed that $X + Y = r$.

The *Fisher exact test* is based on conditioning on both the row and column sums, so m, n, r are all treated as fixed, and then seeing if the observed value of X is "extreme" compared to this conditional distribution. Assuming the null hypothesis, use Bayes' Rule to find the conditional PMF of X given $X + Y = r$. Is this a distribution we have studied in class? If so, say which (and give its parameters).

(b) Give an intuitive explanation for the distribution of (a), explaining how this problem relates to other problems we've seen, and why p_1 disappears (magically?) in the distribution found in (a).

6. Consider the following algorithm for sorting a list of n distinct numbers into increasing order. Initially they are in a random order, with all orders equally likely. The algorithm compares the numbers in positions 1 and 2, and swaps them if needed, then it compares the new numbers in positions 2 and 3, and swaps them if needed, etc., until it has gone through the whole list. Call this one "sweep" through the list. After the first sweep, the largest number is at the end, so the second sweep (if needed) only needs to work with the first $n-1$ positions. Similarly, the third sweep (if needed) only needs to work with the first $n-2$ positions, etc. Sweeps are performed until $n-1$ sweeps have been completed or there is a swapless sweep.

For example, if the initial list is 53241 (omitting commas), then the following 4 sweeps are performed to sort the list, with a total of 10 comparisons:

$$53241 \to 35241 \to 32541 \to 32451 \to 32415.$$

$$32415 \to 23415 \to 23415 \to 23145.$$

$$23145 \to 23145 \to 21345.$$

$$21345 \to 12345.$$

(a) An *inversion* is a pair of numbers that are out of order (e.g., 12345 has no inversions, while 53241 has 8 inversions). Find the expected number of inversions in the original list.

(b) Show that the expected number of comparisons is between $\frac{1}{2}\binom{n}{2}$ and $\binom{n}{2}$.

Hint for (b): for one bound, think about how many comparisons are made if $n-1$ sweeps are done; for the other bound, use Part (a).

7. Athletes compete one at a time at the high jump. Let X_j be how high the jth jumper jumped, with $X_1, X_2, \ldots$ i.i.d. with a continuous distribution. We say that the jth jumper set a *record* if X_j is greater than all of $X_{j-1}, \ldots, X_1$.

(a) Is the event "the 110th jumper sets a record" independent of the event "the 111th jumper sets a record"? Justify your answer by finding the relevant probabilities in the definition of independence *and* with an intuitive explanation.

(b) Find the mean number of records among the first n jumpers (as a sum). What happens to the mean as $n \to \infty$?

Stat 110 Homework 4 Solutions, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1. Let X be a r.v. whose possible values are $0, 1, 2, \ldots$, with CDF F. In some countries, rather than using a CDF, the convention is to use the function G defined by $G(x) = P(X < x)$ to specify a distribution. Find a way to convert from F to G, i.e., if F is a known function show how to obtain $G(x)$ for all real x.

Write

$$G(x) = P(X \leq x) - P(X = x) = F(x) - P(X = x).$$

If x is not a nonnegative integer, then $P(X = x) = 0$ so $G(x) = F(x)$. For x a nonnegative integer,

$$P(X = x) = F(x) - F(x - 1/2)$$

since the PMF corresponds to the lengths of the jumps in the CDF. (The $1/2$ was chosen for concreteness; we also have $F(x - 1/2) = F(x - a)$ for any $a \in (0, 1]$.) Thus,

$$G(x) = \begin{cases} F(x) \text{ if } x \notin \{0, 1, 2, \ldots\} \\ F(x - 1/2) \text{ if } x \in \{0, 1, 2, \ldots\}. \end{cases}$$

More compactly, we can also write $G(x) = \lim_{t \to x^-} F(t)$, where the $-$ denotes taking a limit from the left (recall that F is right continuous), and $G(x) = F(\lceil x \rceil - 1)$, where $\lceil x \rceil$ is the "ceiling" of x (the smallest integer greater than or equal to x).

2. There are n eggs, each of which hatches a chick with probability p (independently). Each of these chicks survives with probability r, independently. What is the distribution of the number of chicks that hatch? What is the distribution of the number of chicks that survive? (Give the PMFs; also give the names of the distributions and their parameters, if they are distributions we have seen in class.)

Let H be the number of eggs that hatch and X be the number of hatchlings that survive. Think of each egg as a Bernoulli trial, where for H we define "success" to mean hatching, while for X we define "success" to mean surviving. For example, in the picture above, where ☺ denotes an egg that hatches with the chick surviving, x denotes an egg that hatched but whose chick died, and ▢ denotes an egg that didn't hatch, the events $H = 7$, $X = 5$ occurred. By the story of the Binomial, $H \sim \text{Bin}(n, p)$, with PMF $P(H = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$.

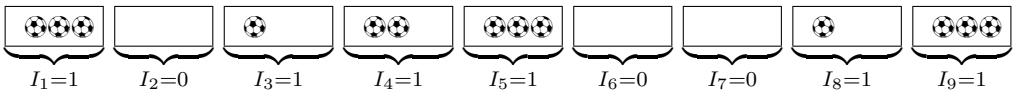
The eggs independently have probability pr each of hatching a chick that survives. By the story of the Binomial, we have $X \sim \text{Bin}(n, pr)$, with PMF $P(X = k) = \binom{n}{k} (pr)^k (1 - pr)^{n-k}$ for $k = 0, 1, \ldots, n$.

3. A couple decides to keep having children until they have at least one boy and at least one girl, and then stop. Assume they never have twins, that the "trials" are independent with probability 1/2 of a boy, and that they are fertile enough to keep producing children indefinitely. What is the expected number of children?

Let X be the number of children needed, starting with the 2nd child, to obtain one whose gender is not the same as that of the firstborn. Then $X-1$ is Geom(1/2), so $E(X)=2$. This does not include the firstborn, so the expected total number of children is $E(X+1)=E(X)+1=3$.

Miracle check: an answer of 2 or lower would be a miracle since the couple always needs to have at least 2 children, and sometimes they need more. An answer of 4 or higher would be a miracle since 4 is the expected number of children needed such that there is a boy and a girl with the boy older than the girl.

4. Randomly, k distinguishable balls are placed into n distinguishable boxes, with all possibilities equally likely. Find the expected number of empty boxes.



Let I_j be the indicator random variable for the j^{th} box being empty, so $I_1+\cdots+I_n$ is the number of empty boxes (the above picture illustrates a possible outcome with $n=9, k=13$). Then $E(I_j)=P(I_j=1)=(1-1/n)^k$. By linearity,

$$E\left(\sum_{j=1}^{n} I_j\right)=\sum_{j=1}^{n} E(I_j)=n(1-1/n)^k.$$

Miracle check: for any $k\geq 1$, there can be at most $n-1$ empty boxes, so the expected number of empty boxes must be at most $n-1$. Here, we do have $n(1-1/n)^k \leq n(1-1/n)=n-1$. And for $k=0$ the answer should reduce to n, while for $n=1, k\geq 1$ it should reduce to 0. Also, it makes sense that the expected number of empty boxes converges to 0 (without ever reaching 0, if $n\geq 2$) as $k\to\infty$.

5. A scientist wishes to study whether men or women are more likely to have a certain disease, or whether they are equally likely. A random sample of m women and n men is gathered, and each person is tested for the disease (assume for this problem that the test is completely accurate). The numbers of women and men in the sample who have the disease are X and Y respectively, with $X\sim \text{Bin}(m,p_1)$ and $Y\sim \text{Bin}(n,p_2)$. Here p_1 and p_2 are unknown, and we are interested in testing the "null hypothesis" $p_1=p_2$.

(a) Consider a 2 by 2 table listing with rows corresponding to disease status and columns corresponding to gender, with each entry the count of how many people have that disease status and gender (so $m+n$ is the sum of all 4 entries). Suppose that it is observed that $X+Y=r$.

The *Fisher exact test* is based on conditioning on both the row and column sums, so m, n, r are all treated as fixed, and then seeing if the observed value of X is "extreme" compared to this conditional distribution. Assuming the null hypothesis, use Bayes' Rule to find the conditional PMF of X given $X+Y=r$. Is this a distribution we have studied in class? If so, say which (and give its parameters).

First let us build the 2×2 table (conditioning on the totals m, n, and r).

	Women	Men	**Total**
Disease	x	$r-x$	r
No Disease	$m-x$	$n-r+x$	$m+n-r$
Total	m	n	$m+n$

Next, let us compute $P(X=x|X+Y=r)$. By Bayes' rule,

$$\begin{aligned} P(X=x|X+Y=r) &= \frac{P(X+Y=r|X=x)P(X=x)}{P(X+Y=r)} \\ &= \frac{P(Y=r-x)P(X=x)}{P(X+Y=r)}. \end{aligned}$$

Assuming the null hypothesis and letting $p=p_1=p_2$, we have $X \sim \text{Bin}(m,p)$ and $Y \sim \text{Bin}(n,p)$ with X independent of Y, so $X+Y \sim \text{Bin}(n+m,p)$. Thus,

$$\begin{aligned} P(X=x|X+Y=r) &= \frac{\binom{n}{r-x}p^{r-x}(1-p)^{n-r+x}\binom{m}{x}p^x(1-p)^{m-x}}{\binom{m+n}{r}p^r(1-p)^{m+n-r}} \\ &= \frac{\binom{m}{x}\binom{n}{r-x}}{\binom{m+n}{r}}. \end{aligned}$$

So the conditional distribution is *Hypergeometric* with parameters m, n, r.

(b) Give an intuitive explanation for the distribution of (a), explaining how this problem relates to other problems we've seen, and why p_1 disappears (magically?) in the distribution found in (a).

This problem has the same structure as the elk (capture-recapture) problem. In the elk problem, we take a sample of elk from a population, where earlier some were tagged, and we want to know the distribution of the number of tagged elk in the

sample. By analogy, think of the women as corresponding to tagged elk, and men as corresponding to untagged elk. Having r people be infected with the disease corresponds to capturing a new sample of r elk the number of women among the r diseased individuals corresponds to the number of tagged elk in the new sample. Under the null hypothesis and given that $X + Y = r$, the set of diseased people is equally likely to be any set of r people.

It makes sense that the conditional distribution of the number of diseased women does not depend on p, since once we know that $X + Y = r$, we can work directly in terms of the fact that we have a population with r diseased and $m + n - r$ undiseased people, without worrying about the value of p that originally generated the population characteristics.

6. Consider the following algorithm for sorting a list of n distinct numbers into increasing order. Initially they are in a random order, with all orders equally likely. The algorithm compares the numbers in positions 1 and 2, and swaps them if needed, then it compares the new numbers in positions 2 and 3, and swaps them if needed, etc., until it has gone through the whole list. Call this one "sweep" through the list. After the first sweep, the largest number is at the end, so the second sweep (if needed) only needs to work with the first $n - 1$ positions. Similarly, the third sweep (if needed) only needs to work with the first $n - 2$ positions, etc. Sweeps are performed until $n - 1$ sweeps have been completed or there is a swapless sweep.

For example, if the initial list is 53241 (omitting commas), then the following 4 sweeps are performed to sort the list, with a total of 10 comparisons:

$$53241 \to 35241 \to 32541 \to 32451 \to 32415.$$

$$32415 \to 23415 \to 23415 \to 23145.$$

$$23145 \to 23145 \to 21345.$$

$$21345 \to 12345.$$

(a) An *inversion* is a pair of numbers that are out of order (e.g., 12345 has no inversions, while 53241 has 8 inversions). Find the expected number of inversions in the original list.

There are $\binom{n}{2}$ pairs of numbers, each of which is equally likely to be in either order. So by symmetry, linearity, and indicator r.v.s, we immediately have that the expected number of inversions is $\frac{1}{2}\binom{n}{2}$.

(b) Show that the expected number of comparisons is between $\frac{1}{2}\binom{n}{2}$ and $\binom{n}{2}$.

Hint for (b): for one bound, think about how many comparisons are made if $n-1$ sweeps are done; for the other bound, use Part (a).

Let X be the number of comparisons and V be the number of inversions. On the one hand, $X \geq V$ since every inversion must be repaired. So $E(X) \geq E(V) = \frac{1}{2}\binom{n}{2}$. On the other hand, there are $n-1$ comparisons needed in the first sweep, $n-2$ in the second sweep (if needed), ..., and 1 in the $(n-1)$st sweep (if needed). So

$$X \leq (n-1) + (n-2) + \cdots + 2 + 1 = \frac{n(n-1)}{2} = \binom{n}{2}.$$

Hence, $\frac{1}{2}\binom{n}{2} \leq E(X) \leq \binom{n}{2}$. This algorithm is known as *bubble-sort*. See http://www.youtube.com/watch?v=lyZQPjUT5B4 for a Hungarian folk dance of the bubble-sort algorithm!

7. Athletes compete one at a time at the high jump. Let X_j be how high the jth jumper jumped, with $X_1, X_2, \ldots$ i.i.d. with a continuous distribution. We say that the jth jumper set a *record* if X_j is greater than all of $X_{j-1}, \ldots, X_1$.

(a) Is the event "the 110th jumper sets a record" independent of the event "the 111th jumper sets a record"? Justify your answer by finding the relevant probabilities in the definition of independence *and* with an intuitive explanation.

Let I_j be the indicator r.v. for the jth jumper setting a record. By symmetry, $P(I_j = 1) = 1/j$ (as all of the first j jumps are equally likely to be the highest of those jumps). Also,

$$P(I_{110} = 1, I_{111} = 1) = \frac{109!}{111!} = \frac{1}{110 \cdot 111},$$

since having the 110th and 111th jumps both being records is the same thing as having the highest of the first 111 jumps being in position 111, the second highest being in position 110, and the remaining 109 being in any order. So

$$P(I_{110} = 1, I_{111} = 1) = P(I_{110} = 1)P(I_{111} = 1),$$

which shows that the 110th jumper setting a record is independent of the 111th jumper setting a record. Intuitively, this makes sense since learning that the 111th jumper sets a record gives us no information about the "internal" matter of how the first 110 jumps are arranged amongst themselves.

(b) Find the mean number of records among the first n jumpers (as a sum). What happens to the mean as $n \to \infty$?

By linearity, the expected number of records among the first n jumpers is $\sum_{j=1}^{n} \frac{1}{j}$, which goes to ∞ as $n \to \infty$ (as this is the harmonic series).