① Let X be result of fair dice roll and $Y = 7 - X$.

Ravi Dayabhai

② Yes, X & Y have the same distribution b/c both are uniform with support 1-7.

	X	Y
M	1	2
T	2	3
W	3	4
R	4	5
F	5	6
St	6	7
Su	7	1

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 $\boxed{P(X < Y) = \frac{6}{7}}$

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. (SEE BELOW)

④ Let Y be r.v. with $P(Y=1)=1$ and X be a r.v with $P(X=0)=.99$ and $P(X=10^{10})=.01$. X and Y are independent

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? $\boxed{\text{Yes}}$

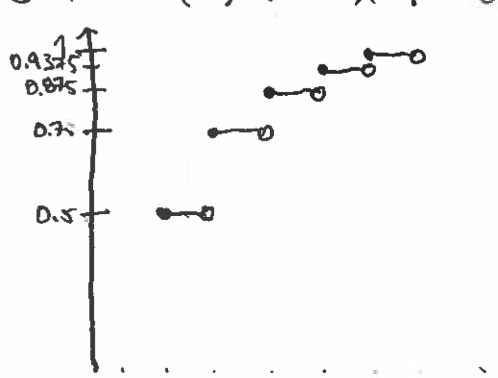
5. Let X be a discrete r.v. with possible values 1,2,3,.... Let $F(x) = P(X \le x)$ be the CDF of X. Show that

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

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. (SEE BACK)

③ $X \sim FS(\frac{1}{2}) \iff X-1 \sim Geom(\frac{1}{2})$. CDF for X is the same as CDF for $X \sim Geom(\frac{1}{2})$, but shifted right by one, i.e., $X-1 \sim Geom(\frac{1}{2})$. $P(X=k) = P(X-1 = k-1)$ shows the PMF equivalency



⑤ $x \in \{1,2,3\ldots\}$; $F(x) = P(X \le x)$. Show $E(X) = \sum_{n=0}^{\infty} 1 - F(n)$.

Starting with a few terms: $n=0$: $1-F(0) = 1$

$n=1$: $1-F(1) = P(X>1) = 1-P(X\le 1) = 1-P(X=1)$

$n=2$: $1-F(2) = P(X>2) = 1-P(X\le 2) = 1-P(X=1) - P(X=2)$

$n=3$: $1-F(3) = P(X>3) = 1-P(X\le 3) = 1-P(X=1)-P(X=2) - P(X=3)$

$$\sum_{n=0}^{\infty} 1-F(n) = 1+(1)n - P(X=1)n - P(X=2)(n-1) - P(X=3)(n-2)\ldots$$

$$= 1+n - n\cdot P(X=1) - n\cdot P(X=2) + P(X=2) - n\, P(X=3) + 2P(X=3)\ldots$$

$$= 1+n-n\,\underbrace{[P(X=1)+\cancel{P(X=2)+P(X=3)+\ldots}]}_{} + P(X=2) + 2P(X=3) + 3P(X=4)+\ldots$$

$$= [P(X=1)+P(X=2)+P(X=3)+\ldots] + P(X=2) + 2P(X=3) + 3P(X=4) + \ldots$$

$$= 1\cdot P(X=1) + 2\cdot P(X=2) + 3\cdot P(X=3) + \ldots = \boxed{\sum_{n=1}^{\infty} n\cdot P(X=n) = E(X)}$$

⑥ $\boxed{E(X) = \infty}$ $P(X>n) = \frac{1}{n}$ for $n>0$; $P(X>0)=1$. Using the identity from #5: $\sum_{n=0}^{\infty} 1-CDF(X) = \sum_{n=0}^{\infty} P(X>n)$,

(by SYMMETRY) but $1 + \underbrace{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots}_{\text{harmonic series}}$ diverges since harmonic series goes to ∞.

⑦ Average total bags obtained by first 3 students

Let X_j be # of bags obtained by j^{th} student where $j\in\{1\ldots 20\}$.

$$E(X_1+X_2+X_3) \underset{\text{LINEARITY}}{=} E(X_1)+E(X_2)+E(X_3) \underset{\text{SYMMETRY}}{=} 3E(X_1).\quad X_j \sim Bin(20, \tfrac{1}{20})$$

so $E(X_j) = 1$. $3\cdot 1 = \boxed{3}$.

Average number of students who get ≥ 1 bag

Let I_j be indicator that j^{th} student gets ≥ 1 bag.

$$E(I_1+\ldots+I_{20}) \underset{\text{LINEARITY}}{=} E(I_1)+\ldots+E(I_{20}) \underset{\text{SYMMETRY}}{=} 20\,E(I_1).\quad E(I_1) = P(A_1) \text{ where}$$

A_1 is event student #1 gets ≥ 1 bag. $P(A_1) = 1-P(A_1^c) = 1-\left(\frac{19}{20}\right)^{20}$,

so $E(I_1+\ldots+I_{20}) = \boxed{20\left(1-\left(\frac{19}{20}\right)^{20}\right)}$

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 $\boxed{E(X)}$ (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. (SEE BACK)

(SEE BACK OF PREVIOUS PAGE

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.

There are always exactly k steps where K = # of laces.

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

Let X be a rv that counts the number of loops formed. Let I_j be indicator that a loop was formed at step j.

$$X = \sum_{j=1}^{100} I_j \Rightarrow E(X) = E\Big(\sum_{j=1}^{100} I_j\Big) = \sum_{j=1}^{100} E(I_j)$$

LINEARITY of expectation

$$\boxed{= \sum_{j=1}^{100} \frac{1}{200-(2j-1)}}$$

Because there are always 2 fewer ends after each step, we can see that there is always, at each step j, $\frac{1}{n-(2j-1)}$ probability that the step will yield a loop! (e.g. consider trivial case w/ 3 laces)

number of total ends (2k=n)

(2) There are $\binom{50}{2}$ pairs of people. Set up indicator r.v.s for each pair and label $1, 2, \ldots, \binom{50}{2}$. Let X be # of pairs of people w/ same birthday: $E(X) = E(I_1 + I_2 + \ldots + I_{\binom{50}{2}})$.

By <u>SYMMETRY</u>, $I_1 = I_2 = \ldots = I_{\binom{50}{2}}$ since probability any pair has same birthday is the same. By <u>LINEARITY OF EXPECTATION</u>,

$E(X) = E(I_1) + E(I_2) + \ldots + E(I_{\binom{50}{2}}) = \binom{50}{2} E(I_1)$. By <u>FUND. BRIDGE</u>

$E(I_1) = P(\text{first pair has same birthday}) = \frac{1}{365}$. $\boxed{E(X) = \binom{50}{2}\frac{1}{365}}$.

Let Y be r.v. that counts the number of days in the year where at least 2 people were born. Follow a similar procedure from above except the indicators $I_1 \cdots I_{365}$ indicate whether 2+ people have birthdays on day j. $E(Y) = E(I_1) + \ldots + E(I_{365})$ by <u>LINEARITY OF EXPECTATION</u>. The probability of 2+ people having their birthday, $P(D_j) = 1 - P(D_j^c)$ where $P(D^c_{j(0)}) + P(D^c_{j(1)}) = P(D_j)$ where $D^c_{(0)}$ and $D^c_{(1)}$ are events where 0 and exactly 1 person has birthday on given day D_j. By <u>SYMMETRY</u> this is the same $P(D_j)$ for all j days. $1 - P(D_j^c) = 1 - \underbrace{P(D^c_{j(0)})}_{\cancel{\binom{50}{0}}\left(\frac{364}{365}\right)^{50}} - \underbrace{P(D^c_{j(1)})}_{\binom{50}{1}\frac{364^{49}}{365^{50}}} = P(D_j)$

So by <u>FUND. BRIDGE</u>, $E(I_j) = P(D_j) = 1 - \left(\frac{364}{365}\right)^{50} - \left(\frac{50}{365}\right)\left(\frac{364}{365}\right)^{49}$

and by SYMMETRY, $\boxed{E(Y) = 365\left[1 - \left(\frac{364}{365}\right)^{50} - \left(\frac{50}{365}\right)\left(\frac{364}{365}\right)^{49}\right]}$

Note to self: This appears isomorphic to the birthday problem!

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 $\begin{cases} X_0 = \text{\# of locations with 0 ph. numbers} \\ X_1 = \text{'' '' 1 '' ''} \\ X_{1+} = \text{'' '' 1+ '' ''} \end{cases}$ and I_j be indicators for $1 \leq j \leq n$. The indicator $I_j = 1$ when the condition given by $X_\square$ is met. Thus, by LINEARITY of expectation, the FUND. BRIDGE, and SYMMETRY:

$$\boxed{E(X_0) = n\left(\frac{n-1}{n}\right)^k}$$

$$\boxed{E(X_{1+}) = n - n\left(\frac{n-1}{n}\right)^k - k\left(\frac{n-1}{n}\right)^{k-1}}$$

$$\boxed{E(X_1) = \frac{\cancel{n}k}{\cancel{n}}\left(\frac{n-1}{n}\right)^{k-1}}$$

$E(X_{1+}) = n - E(X_0) - E(X_1)$ because the three conditions are MECE, so $X_0 + X_1 + X_{1+} = n$. [3] Taking expectation of both sides: $E(X_0 + X_1 + X_{1+}) = E(n) = n = E(X_0) + E(X_1) + E(X_{1+})$

① $\boxed{CDF = F(x) = P(X \leq x) = P(X < x) + P(X = x) = G(x) + PMF}$

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. (SEE ABOVE)

Let X be r.v. counting the number of chicks that hatch from n eggs each w/ prob. p of hatching (ind.). $\boxed{X \sim Bin(n,p)}$

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.) $\boxed{P(Y=m) = \binom{n}{m} r^m p^m (1-rp)^{n-m}}$

$\boxed{Y \sim Bin(n, rp)}$ models the distribution of chicks that survive.

$\boxed{P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}}$

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? (SEE BELOW)

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

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" $p_1 = p_2$.

(SEE BACK.

(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).

③ This can be modeled by a r.v. $X \sim FS(\frac{1}{2})_p$ which is the same as $Y \sim Geom(\frac{1}{2})$ where $Y = X - 1 \Leftrightarrow Y + 1 = X$

$E(Y) + 1 = E(X)$ by LINEARITY of expectation. $E(Y) = \frac{q}{p} = 1$, so $E(X) = 2$.

But this undercounts possible cases because definition of success/failure only happens after having first child, so $\boxed{E(X) + 1 = 3}$ is the expected number of children.

(4) Let X be a r.v. that counts the number of empty boxes. Then $X = \sum_{j=1}^{n} I_j$ where I_j is indicator of j^{th} box being empty.

$E(X) = \sum_{j=1}^{n} E(I_j)$ by LINEARITY, and $E(X) = n \cdot E(I_1)$ by SYMMETRY.

Thus, $\boxed{E(X) = n \cdot \frac{(n-1)^k}{n^k}}$.

(5) A.

	MEN	WOMEN	
DISEASE	$r-k$	k	r
¬ DISEASE	$m-r+k$	$n-k$	$m+n-r$
	m	n	

Null hypothesis assumption that $p_1 = p_2 = p \Rightarrow X+Y \sim Bin(m+n, p)$.

$P(X=k \mid X+Y=r) = ?$

$$P(X=k \mid X+Y=r) = \frac{P(X+Y=r \mid X=k)\ P(X=k)}{P(X+Y=r)} \quad \text{by BAYES' RULE.}$$

$$= \frac{P(Y=r-k)\ P(X=k)}{P(X+Y=r)} \quad \text{by INDEPENDENCE}$$

$$\text{PMF of HGeom}(n,m,r) = \frac{\binom{m}{r-k} \cancel{p^{r-k}} \cancel{q^{m-r+k}} \binom{n}{k} \cancel{p^k} \cancel{q^{n-k}}}{\binom{m+n}{r} \cancel{p^r} \cancel{q^{m+n-r}}} = \boxed{\frac{\binom{m}{r-k}\binom{n}{k}}{\binom{m+n}{r}}}$$

(6) A. Let X be a random variable that counts the number of inversions. by DEF. of INVERSION, there are $\binom{n}{2}$ pairs of numbers and X can be expressed as $X = \sum_{j=1}^{\binom{n}{2}} I_j$, so $E(X) = \sum_{j=1}^{\binom{n}{2}} E(I_j)$ by LINEARITY of EXP. & the FUND. BRIDGE. By SYMMETRY, all $I_j = \frac{1}{2}$ since it's equally likely the first number is greater or less than the second number for the j^{th} pair of numbers. $\boxed{E(X) = \binom{n}{2} \cdot \frac{1}{2}}$

(6) B. The expected value of any r.v. must take on a value $\leq$ the max. of support. If Y is r.v. counting # of comparisons, at most there will be $\binom{n}{2}$ comparisons (when n numbers are in reverse order) = UPPER BOUND for $E(Y)$. The LOWER BOUND of $E(Y)$ is $\frac{1}{2}\binom{n}{2} = E(X)$ because the # of inversions is always $\leq$ # of comparisons, so $E(Y) \geq E(X)$!

This is like elk, where men are the tagged elk, women are untagged elk, and r elk are sampled (read: infected). The p's, q's disappear because any collection of r people are equally likely to be infected (naive def. of probability) and we know enough to use m, n, r, k to count cases.

(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). (SEE ABOVE)

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. bubble sort!
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:

$$53241 \to 35241 \to 32541 \to 32451 \to 32415.$$
$$32415 \to 23415 \to 23415 \to 23145.$$
$$23145 \to 23145 \to 21345.$$
$$21345 \to 12345.$$

after sweep n

→ comparisons.

(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. (SEE BACK OF PREVIOUS PAGE)

(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). (SEE BACK OF PREVIOUS PAGE)

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 $n \to \infty$? (SEE BACK)

* Because each jumper's jump is i.i.d., the P(110th jumper sets record) = $\frac{1}{110}$ and P(111th jumper sets record) = $\frac{1}{111}$ because the distribution is continuous, so each jumper's height is unique. By SYMMETRY it's equally likely that any of the 110 or 111 jumpers achieves max height. The probability both events happen is $\frac{1}{110 \cdot 111}$, which by DEF. of IND, means they are independent events. This is because there is only 1 way both set records consecutively and $110 \cdot 111$ ways to give 1st & 2nd place among 111 jumpers.

④ B. Find the mean # of records among first n jumpers as $n \to \infty$.

From above/earlier, let I_j be indicator that j^{th} jumper sets a record. If X is r.v. counting records, $X = \sum_{j=1}^{\infty} I_j$. By LINEARITY of EXPECTATION, $E(X) = \sum_{j=1}^{\infty} E(I_j)$. We know $P(j^{th}$ jumper sets record$) = \frac{1}{j}$ (from above/earlier), so $E(X) = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots$ (harmonic series!) ... This diverges by COMPARISON TEST:

PROOF. Replace each term with greatest power of $\frac{1}{2}$ that is $\leq$ said term. If this new series diverges, harmonic series diverges.

$$E(X) = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \ldots$$

$$\begin{array}{cccccccc} \geq & \geq & \geq & \geq & \geq & \geq & \geq & \geq \\ \left(\frac{1}{2}\right)^0 + & \left(\frac{1}{2}\right)^1 + & \left(\frac{1}{2}\right)^2 + & \left(\frac{1}{2}\right)^2 + & \left(\frac{1}{2}\right)^3 + & \left(\frac{1}{2}\right)^3 + & \left(\frac{1}{2}\right)^3 + & \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^4 + \ldots \end{array}$$

$$\geq 1 + \frac{1}{2} + \underbrace{\frac{1}{4} + \frac{1}{4}}_{2 \cdot \frac{1}{4}} + \underbrace{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}}_{4 \cdot \frac{1}{8}} + \underbrace{\ldots}_{8 \cdot \frac{1}{16}} + \ldots$$

$$\geq 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \ldots$$

$$\boxed{E(X) \to \infty \text{, by COMPARISON TEST, when } n \to \infty}$$