SP 4.

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 $P(X < Y)$?

1) Let X be the indicator that a coin flipped heads.
Let Y be the indicator that the coin flipped tails.

2) Since $P(Y=2 \mid X=1) = 1$

$$\frac{P(Y=2 \cap X=1)}{P(X=1)} = 1$$

$$\frac{P(X=1 \mid Y=2)\,P(Y=2)}{P(X=1)} = 1$$

$$P(Y=2) = P(X=1) = 1/7$$

$\therefore P(X=1) = \ldots = P(X=7) = P(Y=1) = \ldots = P(Y=7)$

$\Rightarrow Y$ has same distribution as X.

$$P(X<Y) = P(X<Y=1)P(Y=1) + \ldots + P(X<Y=7)P(Y=7)$$
$$= 0 \cdot 1/7 + 1/7 \cdot 1/7 + \ldots + 6/7 \cdot 1/7$$
$$= 1/7\,(1/7 + 2/7 + \ldots + 6/7)$$
$$= \frac{1}{49}(1+2+\ldots+6)$$
$$= \frac{1}{49} \cdot \frac{6 \cdot 7}{2} = \frac{1}{49} \cdot 21 = 3/7 \;//$$

$E(Y) = 4 \qquad P[X < E(Y)] = P(X<4) = 3/7 \;//$ ✗

Since $Y = X+1$, for $1 \le X \le 6$ and $Y = 1$ for $X = 7$,

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

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.

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?

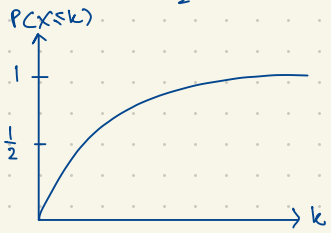
3) $P(X \le k) = P(X=1) + P(X=2) + \dots + P(X=k)$

$= p + qp + q^2p + \dots + q^{k-1}p$

$= p(1 + q + q^2 + \dots + q^{k-1})$

$= p\,\frac{1-q^k}{1-q}$ //

Let $p = \frac{1}{2}$, $\frac{1}{2}\left[\frac{1-\frac{1}{2}^k}{\frac{1}{2}}\right] = 1 - \frac{1}{2}^k$



4) $E(X) > 100E(Y)$

$P(X = 10^{10}) = 10^{-5}$ $\quad P(Y=1) = 1$

$P(X=0) = 1 - 10^{-5}$

$P(Y > X) = P(Y=1 \cap X=0)$

$= 1 \cdot (1 - 10^{-5})$

$= 0.99999 > 0.99$

$E(X) > 10^5 > E(Y) = 1$

17.

57

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, $P(X > 3) = P(X = 4) + P(X = 5) + \ldots$.

$F(x) = P(X \leq x)$

$F(0) = P(X \leq 0) = 0$

$F(1) = P(X \leq 1)$

$E(X) = \sum_{x=1}^{\infty} x P(X = x)$

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

$= (1 - P(X \leq 0)) + (1 - P(X \leq 1)) + \ldots$

$= P(X > 0) + P(X > 1) + \ldots$

$P(X > 0) = P(X = 1) + P(X = 2) + \ldots$

$P(X > 1) = P(X = 2) + P(X = 3) + \ldots$

$\vdots$

$P(X > n) = P(X = n+1) + P(X = n+2) + \ldots$

$\Rightarrow P(X > 0) + P(X > 1) + P(X > 2) + \ldots$

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

$= E(X)$

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

6)

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.

$P(X>0) = 1$

$P(X>1) = 1$

$P(X>2) = P(C_2 \text{ better than at least 1 candidates}) = \frac{1}{2}$

$P(X>3) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$

$P(X>4) = \frac{1}{4}$

$\vdots$

$P(X>n) = \frac{1}{n}$

$P(C_n \text{ worse than } C_1)$

$C_1, C_2, \ldots, C_{n-1}$

There are n positions for C_n to occupy

$\therefore P(C_n \text{ worse than } C_1) = \frac{n-1}{n}$

From (5), $E(X) = \sum_{n=0}^{\infty} (1 - F(n))$ where $F(n) = P(X \le n)$

$= \sum_{n=0}^{\infty} 1 - P(X \le n)$

$= 1 + \sum_{n=1}^{\infty} \frac{1}{n}$

$= 1 + H_\infty$

$= \infty$ //

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.

1) Let X be a r.v representing the number of pairs of people with the same birthday.
Let X_k be a r.v representing the number of people who has the same birthday as person k.

$$X = (X_1 + X_2 + \dots + X_n)\frac{1}{2}$$

Let B_j^k be the indicator r.v such that $B_j^k = 1$ if person j has the same birthday as person k, and 0 otherwise.

$$X_k = B_1^k + \dots + B_n^k - B_k^k$$

$$E(B_j^k) = 1/365$$

$$\begin{aligned} E(X_k) &= E(B_1^k + \dots + B_n^k) \\ &= E(B_1^k) + \dots + E(B_n^k) \\ &= n-1/365 \end{aligned}$$

$$\begin{aligned} E(X) &= \tfrac{1}{2} E(X_1 + \dots + X_n) \\ &= \tfrac{n}{2} E(X_1) \\ &= \tfrac{n}{2} \cdot \tfrac{n-1}{365} \\ &= \tfrac{n(n-1)}{730} \end{aligned}$$

When $n=50$, $E(X) \approx 3.3562$ //

Let D be the r.v that represents the no. of days in a year in which at least 2 people were born.
Let D_k be an indicator r.v such that $D_k = 1$ if at least 2 people were born on day k, and 0 otherwise.

$$D = D_1 + \dots + D_{365}$$

$$\begin{aligned} E(D_k) &= P(\text{at least 2 people were born on day } k) \\ &= 1 - P(\text{at most 1 person was born on day } k) \\ &= 1 - P(\text{no one was born on } k) - P(\text{only 1 was born on } k) \\ &= 1 - \left(\tfrac{364}{365}\right)^{50} - \left(\tfrac{364}{365}\right)^{49} \cdot \tfrac{1}{365} \cdot 50 \end{aligned}$$

$$\begin{aligned} E(D) &= E(D_1 + \dots + D_{365}) \\ &= E(D_1) + \dots + E(D_{365}) \\ &= 365\, E(D_1) \\ &= 365 \left[1 - \left(\tfrac{364}{365}\right)^{50} - \left(\tfrac{364}{365}\right)^{49} \cdot \tfrac{50}{365} \right] \end{aligned}$$

//

2) Let S_k be an r.v that represents the no. of bags that student k gets, with $1 \le k \le 20$.
Let G_j^k be an indicator r.v s.t $G_j^k = 1$ if student k gets bag j, else $G_j^k = 0$.

$E(S_k) = E(G_1^k + \dots + G_{20}^k)$
$= 1/20 + \dots + 1/20 = 1$
$E(S_1 + S_2 + S_3) = 3$ //

Let N be the number of students who get at least 1 bag.
Let B_k be an indicator r.v s.t $B_k = 1$ if student k has at least 1 bag, and 0 otherwise.

$N = B_1 + \dots + B_{20}$

$E(B_k) = P(\text{student k has at least 1 bag})$
$= 1 - P(\text{student has no bags})$
$= 1 - (19/20)^{20}$

$\Rightarrow E(N) = 20[1 - (19/20)^{20}]$ //

$P(\text{student k gets no bags}) = \frac{\binom{20+18}{18}}{\binom{20+19}{19}} = \frac{19}{39}$

20 bags to 19 people

$\binom{20+18}{18}$

$E(B_k) = 1 - \frac{19}{39} = \frac{20}{39}$

$E(N) = E(B_1) + \dots + E(\frac{20}{39})$
$= 20(\frac{20}{39})$
≈ 10.256

$(\frac{19}{20})^{20}$

$S_k = 1 \cdot 20 \cdot \frac{1}{20} \cdot (\frac{19}{20})^{k}$

→ coz choice of bag matters
e.g if person gets bag 1, 3, 5 is diff from getting bag 2, 4 11 even though same amt

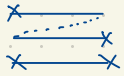
3)

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.)

[redacted]

Suppose there are k ends left, then if loop is formed there will be $k-2$ ends left

Suppose there are k ends left, then if string is extended there will be $k-2$ ends left



$\therefore$ each stage decreases the no. of ends by 2.
Hence no. of steps needed for everything to be a loop is $\frac{\text{no. of ends}}{2} = 100$ //

b) $P(\text{C stage forms a loop} \mid k \text{ ends left}) = \frac{1}{k-1}$

$P(\text{C stage extends a string} \mid k \text{ ends left}) = \frac{k-2}{k-1}$

Let L_i be a r.v for the no. of loops at step i.
Let $B_i \sim \text{Bern}(p)$ s.t $B_i = 1$ if a loop is formed at step i, and 0 otherwise.

$P(\text{C stage } i \text{ forms a loop}) = P(\text{C stage forms a loop} \mid 2i \text{ end left})$

$= \frac{1}{2i-1}$

$\Rightarrow B_i \sim \text{Bern}(\frac{1}{2i-1})$

$L_i = B_i + L_{i-1}$
$= B_i + B_{i-1} + L_{i-2}$
$= B_i + B_{i-1} + \dots + B_2 + B_1$
$= \sum_{j=1}^{i} \frac{1}{2j-1}$

$\Rightarrow L_{100} = \sum_{i=1}^{100} \frac{1}{2i-1}$

//

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 L be a r.v for the number of locations without phone numbers
Let $L_1, \dots, L_n$ be a Bernoulli r.v s.t $L_i = 1$ if location i has no number stored, 0 otherwise

Let N_i^j be the event when position i has person j number.

$$P(\text{location } i \text{ has no phone number}) = P(N_i^{1c} \cap N_i^{2c} \cap \dots \cap N_i^{kc})$$
$$= P(N_i^{1c}) \cdot P(N_i^{2c}) \cdot \dots \cdot P(N_i^{kc})$$
$$= P(N_i^{jc})^k$$

$$P(N_i^{jc}) = 1 - P(N_i^j) = 1 - \frac{1}{n}$$
$$= \frac{n-1}{n}$$

$$\Rightarrow P(\text{location } i \text{ has no phone number}) = \left(\frac{n-1}{n}\right)^k$$
$$\Rightarrow E(L_i) = \left(\frac{n-1}{n}\right)^k$$

$$L = L_1 + L_2 + \dots + L_n$$
$$E(L) = E(L_1 + \dots + L_n)$$
$$= E(L_1) + \dots + E(L_n)$$
$$= nE(L_i)$$
$$= n\left(\frac{n-1}{n}\right)^k \quad //$$ ✓

Let X be an r.v for the no. of locations with exactly one phone number.
Let $X_1, \dots, X_n$ be indicator r.v representing with $X_i = 1$ if position i has exactly 1 number.

$$X = X_1 + \dots + X_n$$

$$E(X_i) = P(\text{location } i \text{ has exactly 1 phone number})$$
$$= \binom{k}{1} \frac{1}{n} \left(\frac{n-1}{n}\right)^{k-1}$$
$$= k \frac{1}{n} \left(\frac{n-1}{n}\right)^{k-1}$$

$$\Rightarrow E(X) = E(X_1 + \dots + X_n)$$
$$= E(X_1) + \dots + E(X_n)$$
$$= nE(X_i)$$
$$= k\left(\frac{n-1}{n}\right)^{k-1} \quad //$$ ✓

Let M be the no. of locations with more than 1 number.

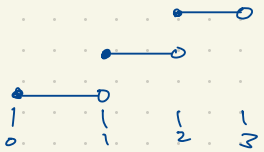
Since each location can either have no numbers, exactly one number, or more than one number, then $M + X + L = n$

$$E(M+X+L) = n$$
$$E(M) + E(X) + E(L) = n$$
$$E(M) = n - E(X) - E(L)$$
$$= n - k\left(\frac{n-1}{n}\right)^{k-1} - n\left(\frac{n-1}{n}\right)^{k}$$

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.

1) $F(x) = P(X \le x)$

For $G(x)$,



$$G(x) = P(X \le x) - P(X = x)$$
$$= P(X \le x) - [P(X \le x) - P(X \le x-1)]$$
$$= P(X \le x-1)$$

2) 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 C be a r.v for the no. of chicks that hatch.

$$P(C=c) = \binom{n}{c} p^c (1-p)^{n-c} \Rightarrow Bin(n,p)$$

Let S be a r.v for the no of chicks that survive.

$$P(S=s) = P(S=s \mid C=s) P(C=s) + P(S=s \mid C=s+1) P(C=s+1) + \ldots + P(S=s \mid C=n) P(C=n)$$

$$P(S=s \mid C=c) = \binom{c}{s} r^s (1-r)^{c-s}$$

$$P(S=s) = \sum_{i=s}^{n} \binom{i}{s} r^s (1-r)^{i-s} \cdot \binom{n}{i} p^i (1-p)^{n-i} \quad \text{or}$$

$$P(S=s) = \binom{n}{s} (rp)^s (1-rp)^{n-s}$$

3)

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 C be a r.v for the no. of children until get at least 1 girl and 1 boy.

$P(C=k) = (1-p)^{k-2} p$ where p is the probability to get different gender from first child.

$= \frac{1}{2^{k-1}}$

$E(C) = P(C=0)\cdot 0 + P(C=1)\cdot 1 + P(C=2)\cdot 2 + \dots$

$= 0 + 0 + \frac{1}{2}\cdot 2 + \frac{1}{4}\cdot 3 + \frac{1}{8}\cdot 4 + \dots + \frac{1}{2^{n-1}}\cdot n + \dots$

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

$\int_2^{\infty} \frac{n}{2^{n-1}} \le \sum_{n=2}^{\infty} \frac{n}{2^{n-1}} \le 1 + \int_2^{\infty} \frac{n}{2^{n-1}}$

$\int \frac{x}{2^{x-1}} dx = \int x 2^{1-x} dx$

Let $u = x$, $dv = 2^{1-x} dx$

$du = dx$, $dv = e^{(1-x)\ln 2} dx$

$v = -\frac{1}{\ln 2} e^{(1-x)\ln 2}$

$uv - \int v\,du$

$-\frac{x}{\ln 2} 2^{1-x} + \frac{1}{\ln 2} \int e^{(1-x)\ln 2} dx$

$-\frac{x}{\ln 2} 2^{1-x} - \frac{1}{(\ln 2)^2} 2^{1-x} \Big|_2^{\infty}$

$= \frac{x}{\ln 2 \, 2^{x-1}} - \frac{1}{(\ln 2)^2 2^{x-1}} \Big|_2^{\infty}$ $\qquad \frac{x}{2^{x-1}}$

$= 0 - 0 - \left(\frac{2}{\ln 2 \cdot 2} - \frac{1}{(\ln 2)^2 2}\right)$

$= \frac{1}{(\ln 2)^2 \cdot 2} + \frac{1}{\ln 2} = \frac{1}{\ln 2}\left[\frac{1}{2\ln 2} + 1\right]$

$= 2.48838$

$2.483 \le E(C) \le 3.483$

$\sum_{n=2}^{\infty} \frac{n}{2^{n-1}} = \sum_{n=2}^{\infty} \left(\frac{n-1}{2^{n-1}} + \frac{1}{2^{n-1}}\right)$

$= \sum_{n=1}^{\infty} \frac{n}{2^n} + \frac{1}{2^n}$

$= E(\text{Geom}(\tfrac{1}{2})) + \sum_{n=1}^{\infty} \frac{1}{2^n}$

$= 1 + \frac{1}{1-\frac{1}{2}} = 1 + 2 = 3$ //

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

Let B be a r.v for the no. of empty boxes.
Let $B_1, \dots, B_n$ be an indicator r.v s.t $B_k = 1$ if box k is empty, $B_k = 0$ otherwise.

$$B = B_1 + \dots + B_n$$

$$\begin{aligned} E(B) &= E(B_1 + \dots + B_n) \\ &= E(B_1) + \dots + E(B_n) \\ &= P(B_1 = 1) + \dots + P(B_n = 1) \end{aligned}$$

Let $X_1^i, X_2^i, \dots, X_k^i$ be indicator r.v s.t $X_j^i = 1$ if ball j is in box i, 0 otherwise.

$$\begin{aligned} P(B_i = 1) &= P(X_1^i = 0) P(X_2^i = 0) \dots P(X_k^i = 0) \\ &= \prod_{j=1}^{k} \frac{n-1}{n} = \left(\frac{n-1}{n}\right)^k \end{aligned}$$

$$\Rightarrow E(B) = \sum_{i=1}^{n} \left(\frac{n-1}{n}\right)^k = n\left(\frac{n-1}{n}\right)^k \quad //$$

5) 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).

a)

	M	F	
T	X	Y	$= r$
F	$m-X$	$n-Y$	$= m+n-r$

Assume null hypothesis is true, that $p_1 = p_2$

Let $p = p_1 = p_2$

$$P(X=k \mid X+Y=r) = \frac{P(X+Y=r \mid X=k) \cdot P(X=k)}{P(X+Y=r)}$$

$$= \frac{P(Y=r-k) \cdot P(X=k)}{P(X+Y=r)}$$

$P(Y=r-k) = 0$ if $r-k > m$

$$P(Y=r-k) = \binom{n}{r-k} p^{r-k} (1-p)^{n-r+k}$$

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

$$P(X=k)P(Y=r-k) = \binom{n}{r-k}\binom{m}{k} p^r (1-p)^{n+m-r}$$

Since $p_1 = p_2$, $X+Y \sim \text{Bin}(n+m, p)$

$$P(X+Y=r) = \binom{n+m}{r} p^r (1-p)^{n+m-r}$$

$$\Rightarrow \frac{P(Y=r-k) \cdot P(X=k)}{P(X+Y=r)} = \frac{\binom{n}{r-k}\binom{m}{k}}{\binom{n+m}{r}} \sim \text{HGeom}(n, m, r)$$

b) Since $p_1 = p_2$, then both men and women have the same probability of getting the disease.

$X+Y=r \Rightarrow$ pick a sample of size r in the men and women population.

$P(X=k \mid X+Y=r) \Rightarrow$ what is the probability of picking exactly k men in a sample size of r.

Since $p_1 = p_2$, all combinations of men and women s.t. $X+Y=r$ have the same probability.

Thus the problem is equivalent to the HGeom problem.

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).

10 342

a) Let I be a r.v for the no. of inversions in the original list.

Let $I_{i,j}$ where $i<j$, $1 \le i,j \le n$, be an indicator r.v s.t $I_{i,j}=1$ if numbers at pos i and j are inverted.

$$I = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I_{i,j}$$

$$E(I) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E(I_{i,j})$$

$$\frac{1}{10}\cdot\frac{10}{10} + \frac{1}{10}\cdot\frac{9}{10} + \frac{1}{10}\cdot\frac{8}{10} + \dots + \frac{1}{10}\cdot\frac{1}{10}$$

$$= \frac{1}{10}\cdot\frac{10(11)}{2} = \frac{11}{20} \quad 1-\frac{11}{20} = \frac{9}{20}$$

$$E(I_{i,j}) = P(N_i = 9 \text{ and } N_j < 9) + P(N_i = 8 \text{ and } N_j < 8) + \dots + P(N_i = 2 \text{ and } N_j < 2) + P(N_i = 1 \text{ and } N_j < 1) + P(N_i = 0 \text{ and } N_j < 0)$$

$$= \frac{1}{10}\cdot\frac{9}{10} + \dots + \frac{1}{10}\cdot\frac{1}{10}$$

$$= \frac{1}{10^2}(9 + 8 + \dots + 1) = \frac{1}{10^2}\left(\frac{9\cdot 10}{2}\right) = \frac{9}{20}$$

Correct if not DISTINCT. But we're looking for DISTINCT numbers $\Rightarrow \frac{1}{2}$

$$\Rightarrow E(I) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{4}{9} = \sum_{j=n}^{n} \frac{4}{9} + \sum_{j=n-1}^{n} \frac{4}{9} + \dots + \sum_{j=2}^{n} \frac{4}{9}$$

$$= \frac{4}{9} + 2\left(\frac{4}{9}\right) + \dots + (n-1)\frac{4}{9}$$

$$= \frac{4}{9}[1 + 2 + \dots + (n-1)]$$

$$= \frac{4}{9}\cdot\frac{n(n-1)}{2} = \frac{2n(n-1)}{9} \quad //$$

$$\frac{1}{2}\binom{n}{2} = \frac{1}{2}\cdot\frac{n!}{2!(n-2)!} = \frac{1}{2}\cdot\frac{n(n-1)}{2!} = \frac{n(n-1)}{4}$$

If distinct, $E(I) = \frac{1}{2}\left(\frac{n(n-1)}{2}\right) = \frac{n(n-1)}{4} = \frac{1}{2}\binom{n}{2}$

b) Let C_n be a r.v for the amount of comparisons done on a list of length n.

$$C_n \leq n-1 + n-2 + \ldots + 1$$
$$= \frac{n(n-1)}{2}$$
$$= \binom{n}{2}$$

$$E(I) \leq E(C_n) \leq \binom{n}{2}$$

$$\frac{1}{2}\binom{n}{2} \leq E(C_n) \leq \binom{n}{2}$$ //

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 $n \to \infty$?

a) Let R_i be the event that the i^{th} jumper set a record.

For independence,

$$P(R_{110} \mid R_{111}) = P(R_{110})$$

$$P(R_{110} \mid R_{111}) = \frac{P(R_{111} \mid R_{110})\,P(R_{110})}{P(R_{111})}$$

$$\begin{aligned} P(R_{111} \mid R_{110}) &= P(X_{111} > X_{110}) \\ &= P(X_{111} > X_{110} \cap X_{111} > X_{109} \cap \ldots \cap X_{111} > X_1) \\ &= P(R_{111}) \end{aligned}$$

$$\Rightarrow P(R_{110} \mid R_{111}) = P(R_{110})$$

$$\Rightarrow \text{Independent.}$$

Knowing that 111 sets a record does not tell us anything about 110 setting a record. 111 setting a record means 111 beats $1, \ldots, 110$. Regardless of whether 110 sets a record or not, 111 will still beat 110 hence they are independent.

$$P(R_{110} \cap R_{111}) = \frac{109!}{111!} = \frac{1}{110 \cdot 111} = \frac{1}{12210}$$

$$P(R_{111}) = \frac{110!}{111!} = \frac{1}{111}$$

$$P(R_{110}) = \frac{109!}{110!} = \frac{1}{110}$$

b) $$E(R) = E(R_1) + E(R_2) + \ldots + E(R_n)$$

$$= 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$$

$$= \sum_{i=0}^{n} \frac{1}{i}$$

$$\lim_{n \to \infty} E(R) = \sum_{i=0}^{\infty} \frac{1}{i} = \infty \quad //$$