Ravi Dayabhai

5 Stat 110 Midterm from 2007

1. Alice and Bob have just met, and wonder whether they have a mutual friend. Each has 50 friends, out of 1000 other people who live in their town. They think that it's unlikely that they have a friend in common, saying "each of us is only friends with 5% of the people here, so it would be very unlikely that our two 5%'s overlap."

Assume that Alice's 50 friends are a random sample of the 1000 people (equally likely to be any 50 of the 1000), and similarly for Bob. Also assume that knowing who Alice's friends are gives no information about who Bob's friends are.

(a) Compute the expected number of mutual friends Alice and Bob have (simplify).

Let I_j be indicator r.v. for mutual friendship. Let $X = \sum_{j=1}^{1000} I_j$, so $E(X)$ is what we want to find. By independence and symmetry, $E(I_j) = (.05)(.05)$. This is a Poisson Paradigm: large # of events (1000 people) with small $p_j = E(I_j)$ each. Therefore $X \sim \text{Pois}(\lambda)$ where $\lambda = \sum_{j}^{1000} p_j = 1000(.05)(.05) = \boxed{2.5 = E(X)}$ by expectation of a Poisson r.v.

(b) Let X be the number of mutual friends they have. Find the PMF of X.

From above, $X \sim \text{Pois}(2.5)$. The PMF of X is $P(X=k) = \frac{e^{-2.5}(2.5)^k}{k!}$. This is an approximation.

The actual PMF follows $X \sim \text{HGeom}(50, 950, 50)$ =

$$P(X=k) = \frac{\binom{50}{k}\binom{950}{50-k}}{\binom{1000}{50}}$$

(c) Is the distribution of X one of the important distributions we have looked at and if so, which one? Note: even without solving (b), you can get credit by giving clear reasons for for or against each of the important distributions.

This is an example of the hypergeometric distribution (that can be approximated using the Poisson Paradigm). It's HGeom because Alice's friends are the first tag/set and Bob's friends are the resampling/set. HGeom answers what is probability I have k of Alice's friends having sampled from Bob's friends $\Leftrightarrow$ # of mutual friends.

2. Two coins are in a hat. The coins look alike, but one coin is fair (with probability 1/2 of Heads), while the other coin is biased, with probability 1/4 of Heads. One of the coins is randomly pulled from the hat, without knowing which of the two it is. Call the chosen coin "Coin C".

(a) Coin C is tossed twice, showing Heads both times. Given this information, what is the probability that Coin C is the fair coin? (Simplify.)

$$P(C \text{ is fair} \mid HH) = \frac{P(HH \mid C \text{ is fair})\, P(C \text{ is fair})}{P(HH \mid C \text{ is fair})\, P(C \text{ is fair}) + P(HH \mid C \text{ is NOT fair}) \cdot P(C \text{ is NOT fair})}$$

(by Bayes' Theorem and independence)

$$= \frac{\left(\frac{1}{2}\cdot\frac{1}{2}\right)\frac{1}{2}}{\left(\frac{1}{2}\cdot\frac{1}{2}\right)\frac{1}{2} + \left(\frac{1}{4}\cdot\frac{1}{4}\right)\frac{1}{2}} = \frac{\frac{1}{8}\cdot\frac{4}{4}}{\frac{4}{4}\cdot\frac{1}{8} + \frac{1}{32}} = \boxed{\frac{4}{5}}$$

(b) Are the events "first toss of Coin C is Heads" and "second toss of Coin C is Heads" independent? Explain briefly.

The coin tosses are conditionally independent given a particular coin is chosen. $P(H \text{ second toss} \mid C_{fair}) = P(H \text{ first toss} \mid C_{fair}) = P(H \mid C_{fair})$.

But the tosses, unconditionally, are not independent since we gain info in subsequent tosses!

(c) Find the probability that in 10 flips of Coin C, there will be exactly 3 Heads. (The coin is equally likely to be either of the 2 coins; do not assume it already landed Heads twice as in (a). Do not simplify.)

$$\boxed{\frac{1}{2}\binom{10}{3}\left(\frac{1}{2}\right)^{3}\left(\frac{1}{2}\right)^{7} + \frac{1}{2}\binom{10}{3}\left(\frac{1}{4}\right)^{3}\left(\frac{3}{4}\right)^{7}}$$

(by LOTP.)

$\frac{1}{2}$ fair

$\frac{1}{2}$ unfair

(∴ must add up to 50)

$\underbrace{:::\cdots}_{1} + 10 \mid \underbrace{\because :}_{2} + 10 \mid \underbrace{'}_{3} + 10 \mid \underbrace{\therefore}_{4} + 10 \mid \underbrace{:}_{5} + 10$

3. Five people have just won a \$100 prize, and are deciding how to divide the \$100 up between them. Assume that whole dollars are used, not cents. Also, for example, giving \$50 to the first person and \$10 to the second is different from vice versa.

(a) How many ways are there to divide up the \$100, such that each gets at least \$10?

Hint: there are $\binom{n+k-1}{k}$ ways to put k indistinguishable balls into n distinguishable boxes; you can use this fact without deriving it. If everyone needs to get at least \$10, then we only need to count the ways to distribute the remaining \$50.

$$\boxed{\binom{5+50-1}{50} = \binom{54}{50} = \frac{\cancel{54}^{9}\cdot 53 \cdot \cancel{52}^{13} \cdot 51 \cdot \cancel{50!}}{\cancel{50!}\cdot \cancel{4}\cdot \cancel{3}\cdot \cancel{2}\cdot 1} = 53\cdot 9\cdot 13\cdot 51}$$

(b) Assume that the \$100 is randomly divided up, with all of the possible allocations counted in (a) equally likely. Find the expected amount of money that the first person receives (justify your reasoning).

$$\boxed{\frac{\sum_{k=0}^{50} \binom{4+(50-k)-1}{50-k} k}{\binom{54}{50}}}$$

We know any one person, in this case, the first, can receive anything from \$0 to \$50 of the remaining cash (\$50). This calculation gets the weights by distributing the remaining \$(50-k) among the 4 other people for all $k \in \{0,1,\ldots,50\}$.

By SYMMETRY, we know the expected value is the same across all 5 people ∴ $E(X_1) = E(X_2) = E(X_3) = E(X_4) = E(X_5) \Longrightarrow$

by LIN. OF EXPECTATION

$5 \cdot E(X_1) = 50 \Rightarrow E(X_1) = \$10 + \$10$ guaranteed $= \boxed{\$20}$.

(c) Let A_j be the event that the jth person receives more than the first person (for $2 \le j \le 5$), when the \$100 is randomly allocated as in (b). Are A_2 and A_3 independent? (No explanation needed for this.) Express $I_{A_2 \cap A_3}$ and $I_{A_2 \cup A_3}$ in terms of I_{A_2} and I_{A_3} (where I_A is the indicator random variable of any event A).

- A_2 and A_3 are not independent.
- $I_{A_2 \cap A_3} = (I_{A_2}) I_{A_3}$ is 1 only if both A_2, A_3 occur, 0 otherwise
- $I_{A_2 \cup A_3} = I_{A_2} + I_{A_3} - (I_{A_2}) I_{A_3}$ by PIE

4. (a) Let $X \sim \text{Pois}(\lambda)$, with $\lambda > 0$. Find $E(X!)$, the average factorial of X. (Simplify, and specify what condition on λ is needed to make the expectation finite.)

$E(X!) = \sum_{x=0}^{\infty} x! P(X=x) = \sum_{x=0}^{\infty} x! \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \lambda^x$ is finite when $0<\lambda<1$, since $\sum_{x=0}^{\infty} \lambda^x$ converges to $\frac{1}{1-\lambda}$; otherwise $E(X!)$ diverges to ∞. In finite case, $\boxed{E(X!) = \frac{e^{-\lambda}}{1-\lambda}}$

(b) The owner of a certain website is studying the distribution of the number of visitors to the site. Every day, a million people independently decide whether to visit the site, with probability $p = 2 \times 10^{-6}$ of visiting. Give a good, simple approximation of the probability of getting *at least two* visitors on a particular day (simplify; your answer should not involve series).

This is a Poisson Paradigm: a large number of independent events (1000 000 people) each w/ a very small probability of happening (2e-6). If $X = \sum_{j=1}^{1000000} I_j$ where I_j is indicator r.v. of whether jth person visits site, then we can approximate $X \sim \text{Pois}(\lambda)$ where $\lambda = \sum_{j=1}^{1000000} p_j = (1000\,000)(2e{-6}) = 2$.

$1 - P(X=0) - P(X=1) = 1 - \left(\frac{e^{-2}2^0}{0!} + \frac{e^{-2}2^1}{1!}\right) = 1-(e^{-2}+2e^{-2}) = \boxed{1-3e^{-2}}$.

(c) In the scenario of (b), approximately how many days will it take on average until there is a day with at least two visitors (including the day itself)?

Modeling # of days until at least 2 visitors can be done using $1+Y$ where $Y \sim \text{Geom}(1-3e^{-2})$. $E(1+Y) = 1+E(Y) = \frac{1-(1-3e^{-2})}{1-3e^{-2}} + 1$ by definition of GEOM expectation.

$$\boxed{= 1 + \frac{3e^{-2}}{1-3e^{-2}} \text{ days}}$$

5. Alice flips a fair coin n times and Bob flips another fair coin $n+1$ times, resulting in independent $X \sim \text{Bin}(n, \frac{1}{2})$ and $Y \sim \text{Bin}(n+1, \frac{1}{2})$.

(a) Let $V = \min(X, Y)$ be the smaller of X and Y, and let $W = \max(X, Y)$ be the larger of X and Y. (If $X = Y$, then $V = W = X = Y$.) Find $E(V) + E(W)$ in terms of n (simplify). By LINEARITY of EXPECTATION, $E(V) + E(W) = E(V+W) = E(X+Y)$ since when $X \neq Y$, $V+W$ must be $X+Y$ and when $X=Y$, $V+W$ must still be $X+Y$, $E(X+Y) = E(X) + E(Y) = \boxed{\frac{1}{2}n + \frac{1}{2}(n+1)}$

(b) Is it true that $P(\underset{\text{successes}}{X} < Y) = P(\underset{\text{failures}}{n - X} < n+1-Y)$? Explain why or why not.

Because the success probabilities are the same ($\frac{1}{2}$) for X and Y, counting failures yields the same PMFs, respectively.

e.g. $P(X=x) = P(n-X = n-x) = \binom{n}{x}\frac{1}{2}^n = \binom{n}{n-x}\frac{1}{2}^n$

$P(Y=y) = P(n+1-Y = n+1-y) = \binom{n+1}{y}\frac{1}{2}^{n+1} = \binom{n+1}{n+1-y}\frac{1}{2}^{n+1}$

By independence, $P(X<Y) = \sum_{x=0}^{n} P(X<Y)P(X=x)$ which is the same thing as $\sum_{x=0}^{n} P(n-x < n+1-Y)P(X=x)$

(c) Compute $P(X < Y)$ (simplify). Hint: use (b) and that X and Y are integers.

We see that $Y = X + H$ where $H \sim \text{Bin}(1, \frac{1}{2})$ since two binomials w/ the same success probability yield one binomial with the sum of trials and the same success probability. $P(X<Y) = P(X < X+H) = P(0<H) = 1 - CDF_H(0) = 1 - \cancel{\binom{1}{0}}\frac{1}{2}^0\frac{1}{2}^1 = \boxed{\frac{1}{2}}$.

CORRECT but different method from prescribed:

$P(X<Y) = P(n-X < n+1-Y)$
$= P(-X < 1-Y)$
$= P(X+1 > Y)$

$P(X<Y) = P(Y \leq X) \implies P(X<Y) = 1 - P(X<Y)$

complements!

$\boxed{= \frac{1}{2}}$ ✓

6 Stat 110 Midterm from 2008

1. The gambler de Méré asked Pascal whether it is more likely to get at least one six in 4 rolls of a die, or to get at least one double-six in 24 rolls of a pair of dice. Continuing this pattern, suppose that a group of n fair dice is rolled $4 \cdot 6^{n-1}$ times.

(a) Find the expected number of times that "all sixes" is achieved (i.e., how often among the $4 \cdot 6^{n-1}$ rolls it happens that all n dice land 6 simultaneously). (Simplify.)

The prob. of "all sixes" for n dice is $\left(\frac{1}{6}\right)^n$. By SYMMETRY and LINEARITY of EXPECTATION, $E(X) = E(I_1 + I_2 + \dots + I_{4 \cdot 6^{n-1}})$ where I_j is indicator r.v that j^{th} roll is "all sixes"; $E(X) = 4 \cdot 6^{n-1} E(I_1)$, and by FUND BRIDGE,

$E(X) = 4 \cdot 6^{n-1} \left(\frac{1}{6}\right)^n = 4 \cdot \frac{6^n 6^{-1}}{6^n} = \frac{4}{6} = \boxed{\frac{2}{3}}$.

(b) Give a simple but accurate approximation of the probability of having at least one occurrence of "all sixes", for n large (in terms of e but not n).

This is a Poisson Paradigm example. $P(X \geq 1) = 1 - P(X=0)$ where X is a r.v. counting occurrences of "all sixes". $E(X) = \frac{2}{3} = \lambda$ for $X \sim \text{Pois}(\lambda)$.

$\boxed{1 - e^{-\frac{2}{3}}}$

(c) de Méré finds it tedious to re-roll so many dice. So after one normal roll of the n dice, in going from one roll to the next, with probability 6/7 he leaves the dice in the same configuration and with probability 1/7 he re-rolls. For example, if $n = 3$ and the 7th roll is $(3, 1, 4)$, then 6/7 of the time the 8th roll remains $(3, 1, 4)$ and 1/7 of the time the 8th roll is a new random outcome. Does the expected number of times that "all sixes" is achieved stay ~~the same~~, increase, or decrease (compared with (a))? Give a short but clear explanation.

~~$\boxed{\text{increase}}$. With outcomes being more sticky ($\frac{6}{7}$ probability of persisting roll-to-roll, landing all 6's is more likely to persist!~~

STAY THE SAME! b/c of LIN. of EXP holds for dependent r.v.s too!

$P(L) = .10$
$P(L^c) = .90$

M_1, M_2 } legit label M_1^c, M_2^c } spam label

2. To battle against spam, Bob installs two anti-spam programs. An email arrives, which is either legitimate (event L) or spam (event L^c), and which program j marks as legitimate (event M_j) or marks as spam (event M_j^c) for $j \in \{1, 2\}$. Assume that 10% of Bob's email is legitimate and that the two programs are each "90% accurate" in the sense that $P(M_j|L) = P(M_j^c|L^c) = 9/10$. Also assume that given whether an email is spam, the two programs' outputs are conditionally independent.

(a) Find the probability that the email is legitimate, given that the 1st program marks it as legitimate (simplify).

$$P(L|M_1) = \frac{P(M_1|L)P(L)}{P(M_1|L)P(L) + P(M_1|L^c)P(L^c)} = \frac{(\frac{9}{10})(\frac{1}{10})}{(\frac{9}{10})(\frac{1}{10}) + (\frac{1}{10})(\frac{9}{10})}$$

$$\boxed{= \frac{1}{2}}$$

(b) Find the probability that the email is legitimate, given that both programs mark it as legitimate (simplify).

$$P(L|M_1, M_2) = \frac{P(M_1, M_2|L)P(L)}{P(M_1, M_2)}$$ (By conditional ind. between first, second spam filters)

$$= \frac{P(M_1|L)P(M_2|L)P(L)}{P(M_1|L)P(M_2|L)P(L) + P(M_1|L^c)P(M_2|L^c)P(L^c)} = \frac{(\frac{9}{10})^2(\frac{1}{10})}{(\frac{9}{10})^2(\frac{1}{10}) + (\frac{1}{10})^2(\frac{9}{10})}$$

$$= \frac{81}{81+9} = \boxed{\frac{9}{10}}$$

(c) Bob runs the 1st program and M_1 occurs. He updates his probabilities and then runs the 2nd program. Let $\tilde{P}(A) = P(A|M_1)$ be the updated probability function after running the 1st program. Explain briefly in words whether or not $\tilde{P}(L|M_2) = P(L|M_1 \cap M_2)$: is conditioning on $M_1 \cap M_2$ in one step equivalent to first conditioning on M_1, then updating probabilities, and then conditioning on M_2?

$\boxed{\text{Yes}}$

Bayes' Rule is coherent

$$P(L|M_1 \cap M_2) = \frac{P(M_1, M_2|L)P(L)}{P(M_1, M_2)} = \frac{P(M_2|L)\boxed{P(M_1|L)P(L)}}{P(M_2|M_1)\boxed{P(M_1)}}$$ $\tilde{P}(L)$

$$\tilde{P}(L|M_1 \cap M_2) = \frac{P(M_2|L)\tilde{P}(L)}{P(M_2|M_1)} \stackrel{?}{=} \frac{P(M_2|L)\tilde{P}(L)}{P(M_2)} = \tilde{P}(L|M_2)$$ ✓

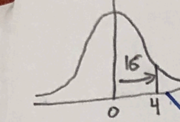
ODDS FORM BAYES

$$\frac{P(L|M_1, M_2)}{P(L^c|M_1, M_2)} = \frac{P(M_2|L)\tilde{P}(L)}{P(M_2|L^c)\tilde{P}(L^c)} = \frac{\tilde{P}(L|M_2)}{\tilde{P}(L^c|M_2)} = \frac{\tilde{P}(L|M_2)}{1 - \tilde{P}(L|M_2)}$$

divide top & bottom by $P(M_2)$

$Z \sim N(0,1)$. By scale-location transformation,

$X_i = \mu + Z\sigma \Rightarrow Z = \frac{X_i - \mu}{\sigma} = \frac{X_i - 0}{+\sqrt{4}} = \frac{1}{2}X \Rightarrow X_i = 2Z$



CORRECT LOGIC!

ALMOST!

3. (a) Let $X_1, X_2, \ldots$ be independent $\mathcal{N}(0,4)$ r.v.s., and let J be the smallest value of j such that $X_j > 4$ (i.e., the index of the first X_j exceeding 4). In terms of the standard Normal CDF Φ, find $E(J)$ (simplify).

$J \sim FS(p) \Leftrightarrow K+1 = J$, $K \sim \text{Geom}(p)$. p is probability that $X_j > 4$. $E(K) = \frac{1-p}{p}$, so $E(J) = 1 + E(K)$. $1 - \Phi(2) = p$,

so $\boxed{E(J) = 1 + \frac{\Phi(2)}{1-\Phi(2)}}$

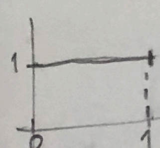
Dumb mistake! $N(0,4)$ means $\mu = 0$, $\sigma^2 = 4$, so $\sigma = 2$!

(b) Let f and g be PDFs with $f(x) > 0$ and $g(x) > 0$ for all x. Let X be a random variable with PDF f. Find the expected value of the ratio $\frac{g(X)}{f(X)}$ (simplify).

By LOTUS, $E\left(\frac{g(X)}{f(X)}\right) = \int_{-\infty}^{\infty} \frac{g(x)}{\cancel{f(x)}} \cancel{f(x)}\, dx = \boxed{1}$ since $g(x)$ is PDF.

(c) Define $F(x) = e^{-e^{-x}}$. This is a CDF (called the *Gumbel distribution*) and is a continuous, strictly increasing function. Let X have CDF F, and define $W = F(X)$. What are the mean and variance of W (simplify)?

W is the quantile function! (Plugging in a r.v. into its CDF yields it's quantile by UNIV. OF UNIFORM.)



$W \sim \text{Unif}(0,1) \Rightarrow E(W) = \frac{1-0}{2} = \boxed{\frac{1}{2}}$

$\text{Var}(W) = E(X^2) - E(X)^2$

$= E(X^2) - \left(\frac{1}{2}\right)^2$

$= \frac{1}{3} - \frac{1}{4} = \boxed{\frac{1}{12}}$

$F(X^2) = \int_0^1 x^2 f(x)\, dx$

$= \int_0^1 x^2\, dx = \frac{1}{3}x^3 \Big|_0^1 = \frac{1}{3}$

4. (a) Find $E(2^X)$ for $X \sim \text{Pois}(\lambda)$ (simplify).

By LOTUS, $E(2^x) = e^{-\lambda} \sum_x \frac{(2\lambda)^x}{x!} = e^{-\lambda} e^{2\lambda} = \boxed{e^{\lambda}}$

(b) Let X and Y be independent Pois(λ) r.v.s, and $T = X + Y$. Later in the course, we will show that $T \sim \text{Pois}(2\lambda)$; here you may use this fact. Find the conditional distribution of X given $T = n$, i.e., find the conditional PMF $P(X = k|T = n)$ (simplify). Which "important distribution" is this conditional distribution, if any?

By independence: $P(X=k \mid T=n) = \frac{P(Y=n-k)P(X=k)}{P(T=n)} =$

$$\frac{\frac{\cancel{e^{-\lambda}}\lambda^{n-k}}{(n-k)!} \cdot \frac{\cancel{e^{-\lambda}}\lambda^{k}}{k!}}{\frac{\cancel{e^{-2\lambda}}(2\lambda)^n}{n!}} = \frac{\lambda^n}{(2\lambda)^n}\binom{n}{k} = \boxed{\binom{n}{k}\left(\frac{1}{2}\right)^n \Rightarrow \text{Bin}\left(n, \frac{1}{2}\right)}$$

PMF for

(c) Again let X and Y be Pois(λ) r.v.s, and $T = X + Y$, but now assume now that X and Y are *not* independent, and in fact $X = Y$. Prove or disprove the claim that $T \sim \text{Pois}(2\lambda)$ in this scenario.

$P(T=n) = P(X+Y=n) = P(2X=n) = P(X=\frac{1}{2}n) =$

$$\frac{e^{-\lambda}\sqrt{\lambda^n}}{\left(\frac{n}{2}\right)!} \neq \frac{e^{-2\lambda}(2\lambda)^n}{n!}$$

So $T \not\sim \text{Pois}(\lambda)$ when $X=Y$.

can only take EVEN values $\Rightarrow$ not poisson!

7 Stat 110 Midterm from 2009

1. (a) Let $X \sim \text{Pois}(\lambda)$. Find $E(e^X)$ (simplify).

$E(e^x) = \sum_x e^x \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_x \frac{(\lambda e)^x}{x!} = e^{-\lambda} \cdot e^{\lambda e}$

(by LOTUS)

(by Taylor series approximation)

$\boxed{e^{\lambda(e-1)}}$

(b) The numbers $1, 2, 3, \ldots, n$ are listed in some random order (with all $n!$ permutations equally likely). An *inversion* occurs each time a pair of numbers is out of order, i.e., the larger number is earlier in the list than the smaller number. For example, 3, 1, 4, 2 has 3 inversions (3 before 1, 3 before 2, 4 before 2). Find the expected number of inversions in the list (simplify).

Let X be count of inversions. Let I_k be the indicator that the k^{th} pair of digits is inverted.

Then $X = \sum_k I_k$, so $E(X) = E(\sum_k I_k)$.

By LINEARITY OF EXPECTATION & SYMMETRY,

$E(X) = \binom{n}{2} E(I_1) = \boxed{\binom{n}{2}/2}$

2. Consider four nonstandard dice (the *Efron dice*), whose sides are labeled as follows (the 6 sides on each die are equally likely).

A: 4, 4, 4, 4, 0, 0
B: 3, 3, 3, 3, 3, 3
C: 6, 6, 2, 2, 2, 2
D: 5, 5, 5, 1, 1, 1

These four dice are each rolled once. Let A be the result for die A, B be the result for die B, etc.

(a) Find $P(A > B), P(B > C), P(C > D)$, and $P(D > A)$.

$\boxed{P(A>B) = \frac{2}{3}}$

$\boxed{P(B>C) = \frac{2}{3}}$

$\boxed{P(C>D)} = P(C>D \mid D=1)\, P(D=1) + P(C>D \mid D=5)\, P(D=5)$

$\boxed{\frac{2}{3}} = (1)(\frac{1}{2}) + (\frac{1}{3})(\frac{1}{2})$

$\boxed{P(D>A)} = P(D>A \mid A=0)\, P(A=0) + P(D>A \mid A=4)\, P(A=4)$

$\boxed{\frac{2}{3}} = (1)(\frac{1}{3}) + (\frac{1}{2})(\frac{2}{3})$

(b) Is the event $A > B$ independent of the event $B > C$? Is the event $B > C$ independent of the event $C > D$? Explain.

$A > B$ is independent of $B > C$ since B always yields a $3 \Rightarrow P(A>B)$ is affected only by A, and $P(B>C)$ is affected by C and A does not affect C, vice versa.

$$P(B>C, C>D) = P(B>D) \overset{?}{=} P(B>C)\, P(C>D)$$

$$\overset{?}{=} (\frac{2}{3})(\frac{2}{3}) = \frac{4}{9}$$

$$P(B>D) = \frac{1}{2} \neq \frac{4}{9}$$

So $B>C$ and $C>D$ are not independent events.

3. A discrete distribution has the *memoryless property* if for X a random variable with that distribution, $P(X \geq j+k|X \geq j) = P(X \geq k)$ for all nonnegative integers j, k.

(a) If X has a memoryless distribution with CDF F and PMF $p_i = P(X = i)$, find an expression for $P(X \geq j + k)$ in terms of $F(j), F(k), p_j, p_k$.

By LOTP : $P(X \geq j+k) =$

$= P(X \geq j+k \mid X \geq j)\, P(X \geq j) + \underbrace{\cancel{P(X \geq j+k \cap X < j)}}_{0}$

$= (1 - F(k) + p_k)(1 - F(j) + p_j)$

(b) Name one important discrete distribution we have studied so far which has the memoryless property. Justify your answer with a clear interpretation in words or with a computation.

$\boxed{\text{Geometric}}$; The geometric counts failures before first success for independent trials. The probability of $\geq k$ additional failures after j failures is still the same as $\geq k$ failures from the get go.

$P(X \geq j+k \mid X \geq j) = (1 - F(k) + p_k)$

$\downarrow \qquad\qquad\qquad\qquad \parallel$

$P(X - j \geq k) \longrightarrow P(X \geq k)$

$X - j$ → some non-negative integer

4. The book *Red State, Blue State, Rich State, Poor State* (by Andrew Gelman) discusses the following election phenomenon: within any U.S. state, a wealthy voter is more likely to vote for a Republican than a poor voter; yet the wealthier states tend to favor Democratic candidates! In short: rich individuals (in any state) tend to vote for Republicans, while states with a higher percentage of rich people tend to favor Democrats.

(a) Assume for simplicity that there are only 2 states (called Red and Blue), each of which has 100 people, and that each person is either rich or poor, and either a Democrat or a Republican. Make up numbers consistent with the above, showing how this phenomenon is possible, by giving a 2 by 2 table for each state (listing how many people in each state are rich Democrats, etc.).

RED STATE

	poor	rich	
D	29	1	30
R	61	9	70
	90	10	100

BLUE STATE

	poor	rich	
D	35	20	55
R	20	25	45
	55	45	100

(b) In the setup of (a) (not necessarily with the numbers you made up there), let D be the event that a randomly chosen person is a Democrat (with all 200 people equally likely), and B be the event that the person lives in the Blue State. Suppose that 10 people move from the Blue State to the Red State. Write P_{old} and P_{new} for probabilities before and after they move. Assume that people do not change parties, so we have $P_{\text{new}}(D) = P_{\text{old}}(D)$. Is it possible that *both* $P_{\text{new}}(D|B) > P_{\text{old}}(D|B)$ and $P_{\text{new}}(D|B^c) > P_{\text{old}}(D|B^c)$ are true? If so, explain how it is possible and why it does not contradict the law of total probability $P(D) = P(D|B)P(B) + P(D|B^c)P(B^c)$; if not, show that it is impossible.

$$P_n(D) = P_n(D|B)\,P_n(B) + P_n(D|B^c)\,P_n(B^c) \stackrel{?}{=}$$

$$P_o(D|B)\,P_o(B) + P_o(D|B^c)\,P_o(B^c) = P_o(D)$$

$$= P_n(D|B)(.45) + P_n(D|B^c)(.55) \stackrel{?}{=}$$

$$P_o(D|B)(.50) + P_o(D|B^c)(.50)$$

This is not possible given the conditions! Even if $P_n(D|B)$ was greater than $P_o(D|B)$ by an amount that offset the $P_n(B) < P_o(B)$ differential, the second term will always be greater for $P_n(D)$ than $P_o(D)$!

TRUE but bad logic re: first term!

8 Stat 110 Midterm from 2010

1. A family has two children. The genders of the first-born and second-born are independent (with boy and girl equally likely), and which seasons the children were born in are independent, with all 4 seasons equally likely.

(a) Find the probability that both children are girls, given that a randomly chosen one of the two is a girl who was born in winter (simplify).

Given one girl is born in winter, we just need other child to be a girl, so $\boxed{\frac{1}{2}}$.

(b) Find the probability that both children are girls, given that at least one of the two is a girl who was born in winter (simplify).

$$P(\text{both girls} \mid \text{at least 1 girl born in winter}) = \frac{P(\text{both girls} \cap \text{at least 1 girl born in winter})}{P(\text{at least 1 girl born in winter})}$$

$1 - P(\text{no girls born in winter})$

$1 - \left(\frac{7}{8}\right)^2 = \frac{15}{64}$

$$\underbrace{2\left(\frac{1}{8}\right)\left(\frac{4}{8}\right)}_{\text{one winter girl}} - \underbrace{\left(\frac{1}{8}\right)^2}_{\text{both girl(s) born in winter}} = \frac{7}{64}$$

$\boxed{\frac{7}{15}}$

$P(\text{at least 1 correct \#}) = 1 - \frac{\binom{30}{5}}{\binom{35}{5}} = \frac{\binom{35}{5} - \binom{30}{5}}{\binom{35}{5}}$

$P(\text{exactly 3 correct \#s}) = \frac{\binom{5}{3}\binom{30}{2}}{\binom{35}{5}}$

2. In each day that the "Mass Cash" lottery is run in Massachusetts, 5 of the integers from 1 to 35 are chosen (randomly and without replacement).

(a) When playing this lottery, find the probability of guessing exactly 3 numbers right, given that you guess at least 1 of the numbers right (leave your answer in terms of binomial coefficients).

$P(\text{exactly 3 correct \#s} \mid \text{at least 1 correct \#}) = \frac{P(\text{exactly 3 correct \#s})}{P(\text{at least 1 correct \#})}$

$$\boxed{\frac{\binom{5}{3}\binom{30}{2}}{\binom{35}{5} - \binom{30}{5}}}$$

(b) Find an exact expression for the expected number of days needed so that all of the $\binom{35}{5}$ possible lottery outcomes will have occurred (leave your answer as a sum, which can involve binomial coefficients).

Coupon collector problem! Let $N_i \sim FS((n-i+1)/n)$ where $n = \binom{35}{5}$ and i indexes all possible combinations of lottery numbers. By LINEARITY,

$E(\sum_i^n N_i) = E(N_1) + E(N_2) + \ldots + E(N_n) = 1 + \frac{n}{n-1} + \frac{n}{n-2} + \ldots + n =$

$$\boxed{\binom{35}{5} \sum_{i=1}^{\binom{35}{5}} \frac{1}{i}}$$

(c) Approximate the probability that after 50 days of the lottery, every number from 1 to 35 has been picked at least once (don't simplify, but your answer shouldn't involve a sum).

X

close!

Let X_i count the number of times number i shows up in 50 days, i.e., $i \in \{1, \ldots, 35\}$. $X_i = \sum_{j=1}^{50} I_j$, where I_j is indicator if number i appears on day j. $E(X_i)$ for all $i = 50 \cdot \frac{\binom{34}{4}}{\binom{35}{5}}$. This is the expected rate of number i's appearances per 50 days. Let $Y_i \sim Pois(\lambda)$ where $\lambda = E(X_i) = 50 \cdot \binom{34}{4} / \binom{35}{5}$ and Y_i is distribution of number of times number i shows up per 50 days. Since we care about $P(Y_1 > 0, Y_2 > 0, \ldots, Y_{35} > 0)$, and by weak independence, (Poisson paradigm).

not really satisfying large n, small p_j

$$\boxed{\left(1 - e^{-50\frac{\binom{34}{4}}{\binom{35}{5}}}\right)^{35}} \approx \cancel{0.9407} \; 0.972 \text{ vs. } 0.984 \text{ correct answer.}$$

$$\left(e^{-35\left(1 - \frac{\binom{34}{4}}{\binom{35}{5}}\right)^{50}}\right)$$

Let $g(x) = \left[\ln\left(\frac{x}{1-x}\right)\right]^2$; $E(X^2) = E(g(U))$.

So by LOTUS, $E(g(U)) = \int_{-\infty}^{\infty} g(u) \cdot f(U)\, du$. But U is 0 other than when $0 \le x \le 1$.

3. Let $U \sim \text{Unif}(0, 1)$, and $X = \ln\left(\frac{U}{1-U}\right)$.

(a) ~~Write down (but do not compute) an integral giving $E(X^2)$.~~

$$\boxed{\int_0^1 \ln\left(\frac{u}{1-u}\right)^2 du}$$

(b) Find the CDF of X (simplify).

$P(X < x) = P\left(\ln\left(\frac{U}{1-U}\right) < x\right) = P\left(\frac{U}{1-U} < e^x\right) =$

$P(U < (1-U)e^x) = P(U + Ue^x < e^x) = \underbrace{P\left(U < \frac{e^x}{1+e^x}\right)}_{\text{probability } \propto \text{ to "length" for } U \sim \text{Unif}(0,1)} =$

$$P(X < x) = \boxed{\frac{e^x}{1+e^x}}$$

(c) Find $E(X)$ without using calculus (simplify).
Hint: $1 - U$ has the same distribution as U.

$E(X) = E(g(U))$, where $g(U) = X = \ln\left(\frac{U}{1-U}\right)$.

By def. of exp.,

$$E(X) = \int_0^1 \ln\frac{u}{1-u}\, du = \int_0^1 [\ln u - \ln(1-u)]\, du$$

$$= \int_0^1 \ln(u)\, du - \int_0^1 \ln(1-u)\, du$$

$\boxed{= 0}$ since U and $1-U$ have same distribution.

Wow, so bad. Need to practice ch. 5 (cont. nv.).

Also, Simpson's paradox,

4. Let $X_1, X_2, X_3, \ldots, X_{10}$ be the total number of inches of rain in Boston in October of 2011, 2012, 2013, ..., 2020, with these r.v.s independent $\mathcal{N}(\mu, \sigma^2)$. (Of course, rainfall can't be negative, but μ and σ are such that it is extremely likely that all the X_j's are positive.) We say that a *record value* is set in a certain year if the rainfall is greater than all the previous years (going back to 2011; so by definition, a record is always set in the first year, 2011).

(a) On average, how many of these 10 years will set record values? (Your answer can be a sum but the terms should be simple.)

Let I_j indicate whether record was set in year j. We solve for $E(\sum_{j=1}^{10} I_j)$. By linearity of expectation and the fund. bridge,

$$E\left(\sum_{j=1}^{10} I_j\right) = 1 + P(X_2 > X_1) + P(X_3 > X_1, X_3 > X_2) \ldots + P(X_{10} > X_1, \ldots, X_{10} > X_9).$$

$$= 1 + \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^4 + \ldots + \left(\frac{1}{2}\right)^9$$

by independence

$$= \boxed{\sum_{i=0}^{9} \left(\frac{1}{2}\right)^i} \qquad = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{10}$$

(b) Is the indicator of whether the year 2013 sets a record independent of the indicator of whether the year 2014 sets a record? (Justify briefly.)

$\boxed{\text{Yes}}$ $P(I_{2013} = 1, I_{2014} = 1) \overset{?}{=} P(I_{2013} = 1) \cdot P(I_{2014} = 1)$ is true (so, they are indep) because $P(I_{2013} = 1)$ doesn't depend on the outcome in 2014 — only how many prior years there are to consider; $P(X_4 > X_1, X_2, X_3) = \frac{1}{4}$

(c) Later in the course, we will show that if $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then $Y_1 - Y_2 \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$. Using this fact, find the probability that the October 2014 rainfall will be more than double the October 2013 rainfall in Boston, in terms of Φ.

Let Y_1 = October 2013 rainfall, Y_2 = October 2014 rainfall. $Y_1, Y_2 \sim \mathcal{N}(\mu, \sigma^2)$

$P(Y_2 > 2Y_1) = P(Y_2 - Y_1 > Y_1)$. $Y_2 - Y_1 \sim \mathcal{N}(\mu_2 - \mu_1, \sigma_1^2 + \sigma_2^2)$.

So, $Y_2 - Y_1 = X \sim \mathcal{N}(0, 2\sigma^2)$. CDF of X in terms of Φ is $\Phi\left(\frac{x}{\sigma\sqrt{2}}\right)$.

So $P(X < x) = \Phi\left(\frac{x}{\sigma\sqrt{2}}\right) \Rightarrow 1 - \Phi\left(\frac{x}{\sigma\sqrt{2}}\right) = P(X > x)$. Then

$$\boxed{P(X > Y_1) = 1 - \Phi\left(\frac{Y_1}{\sigma\sqrt{2}}\right)} \qquad 1 - \Phi\left(\frac{\mu}{\sigma\sqrt{5}}\right)$$