Stat 110 Final

Prof. Joe Blitzstein

December 15, 2011

This exam is closed book and closed notes, except for four standard-sized sheets of paper (8.5" by 11") which can have notes on both sides. No calculators, computers, or cell phones are allowed. Show your work, and try to check whether your answers make sense. Answers should be exact unless an approximation is asked for, and fully simplified unless otherwise specified.

There are 8 problems, each of which is worth 12 points. Within a problem, all parts are weighted equally. The last page contains a table of important distributions. The three pages before that can be used for scratch work or for extra space. If you want work done there or on backs of pages to be graded, mention where to look *in big letters with a box around them, on the page with the question.* Good luck!

Name ___________________________________

Harvard ID ______________________________

1. ____________ (/12 points)

2. ____________ (/12 points)

3. ____________ (/12 points)

4. ____________ (/12 points)

5. ____________ (/12 points)

6. ____________ (/12 points)

7. ____________ (/12 points)

8. ____________ (/12 points)

Free _____4_____ (/4 points)

Total ____________ (/100 points)

1. A Pois(λ) number of people vote in a certain election. Each voter votes for Candidate A with probability p and for Candidate B with probability $q = 1 - p$, independently of all the other voters. Let V be the difference in votes, defined as the number of votes for A minus the number of votes for B.

(a) Find $E(V)$ (simplify).

(b) Find $\text{Var}(V)$ (simplify).

2. Let X and Y be i.i.d. Gamma$(\frac{1}{2}, \frac{1}{2})$, and let $Z \sim \mathcal{N}(0,1)$ (note that X and Z may be dependent, and Y and Z may be dependent). For (a),(b),(c), write the most appropriate of $<$, $>$, $=$, or ? in each blank; for (d),(e),(f), write the most appropriate of $\leq$, $\geq$, $=$, or ? in each blank. It is *not* necessary to justify your answers for full credit, but partial credit may be available for justified answers that are flawed but on the right track.

(a) $P(X < Y)$ ____ $1/2$

(b) $P(X = Z^2)$ ____ 1

(c) $P(Z \geq \frac{1}{X^4+Y^4+7})$ ____ 1

(d) $E(\frac{X}{X+Y})E((X+Y)^2)$ ____ $E(X^2) + (E(X))^2$

(e) $E(X^2 Z^2)$ ____ $\sqrt{E(X^4)E(X^2)}$

(f) $E((X+2Y)^4)$ ____ 3^4

3. Ten million people enter a certain lottery. For each person, the chance of winning is one in ten million, independently.

(a) Find a simple, good approximation for the PMF of the number of people who win the lottery.

(b) Congratulations! You won the lottery. However, there may be other winners. Assume now that the number of winners other than you is $W \sim \text{Pois}(1)$, and that if there is more than one winner, then the prize is awarded to one randomly chosen winner. Given this information, find the probability that you win the prize (simplify).

4. A drunken man wanders around randomly in a large space. At each step, he moves one unit of distance North, South, East, or West, with equal probabilities. Choose coordinates such that his initial position is $(0, 0)$ and if he is at (x, y) at some time, then one step later he is at $(x, y+1), (x, y-1), (x+1, y)$, or $(x-1, y)$. Let (X_n, Y_n) and R_n be his position and distance from the origin after n steps, respectively.

General hint: note that X_n is a sum of r.v.s with possible values $-1, 0, 1$, and likewise for Y_n, but be careful throughout the problem about independence.

(a) Determine whether or not X_n is independent of Y_n (explain clearly).

(b) Find $\text{Cov}(X_n, Y_n)$ (simplify).

(c) Find $E(R_n^2)$ (simplify).

5. Each of 111 people names his or her 5 favorite movies out of a list of 11 movies.

(a) Alice and Bob are 2 of the 111 people. Assume *for this part only* that Alice's 5 favorite movies out of the 11 are random, with all sets of 5 equally likely, and likewise for Bob, independently. Find the expected number of movies in common to Alice's and Bob's lists of favorite movies (simplify).

(b) Show that there are 2 movies such that at least 21 of the people name both of these movies as favorites.

6. (a) A woman is pregnant, with a due date of January 10, 2012. Of course, the actual date on which she will give birth is not necessarily the due date. On a timeline, define time 0 to be the instant when January 10, 2012 begins. Suppose that the time T when the woman gives birth has a Normal distribution, centered at 0 and with standard deviation 8 days. What is the probability that she gives birth on her due date? (Your answer should be in terms of Φ, and simplified.)

(b) Another pregnant woman has the same due date as the woman from (a). Continuing with the setup of (a), let T_0 be the time of the first of the two births. Assume that the two birth times are i.i.d. Find the variance of T_0 (in terms of integrals, which do not need to be fully simplified).

7. Fred wants to sell his car, after moving back to Blissville. He decides to sell it to the first person to offer at least \$12,000 for it. Assume that the offers are independent Exponential random variables with mean \$6,000.

(a) Find the expected number of offers Fred will have (including the offer he accepts).

(b) Find the expected amount of money that Fred gets for the car.

8. Let G be an undirected network with nodes labeled $1, 2, \ldots, M$ (edges from a node to itself are not allowed), where $M \geq 2$ and random walk on this network is irreducible. Let d_j be the degree of node j for each j. Create a Markov chain on the state space $1, 2, \ldots, M$, with transitions as follows. From state i, generate a "proposal" j by choosing a uniformly random j such that there is an edge between i and j in G; then go to j with probability $\min(d_i/d_j, 1)$, and stay at i otherwise.

(a) Find the transition probability q_{ij} from i to j for this chain, for all states i, j (be sure to specify when this is 0, and to find q_{ii}, which you can leave as a sum).

(b) Find the stationary distribution of this chain (simplify).

Extra Page 1. Don't panic.

Extra Page 2. Remember the memoryless property!

Extra Page 3. Conditioning is the soul of statistics.

Table of Important Distributions

Name	Param.	PMF or PDF	Mean	Variance
Bernoulli	p	$P(X=1)=p, P(X=0)=q$	p	pq
Binomial	n,p	$\binom{n}{k}p^k q^{n-k},$ for $k \in \{0,1,\dots,n\}$	np	npq
Geometric	p	$q^k p,$ for $k \in \{0,1,2,\dots\}$	q/p	q/p^2
NegBinom	r,p	$\binom{r+n-1}{r-1}p^r q^n, n \in \{0,1,2,\dots\}$	rq/p	rq/p^2
Hypergeom	w,b,n	$\frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}},$ for $k \in \{0,1,\dots,n\}$	$\mu = \frac{nw}{w+b}$	$(\frac{w+b-n}{w+b-1})n\frac{\mu}{n}(1-\frac{\mu}{n})$
Poisson	λ	$\frac{e^{-\lambda}\lambda^k}{k!},$ for $k \in \{0,1,2,\dots\}$	λ	λ
Uniform	$a<b$	$\frac{1}{b-a},$ for $x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	μ,σ^2	$\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$	μ	σ^2
Exponential	λ	$\lambda e^{-\lambda x},$ for $x>0$	$1/\lambda$	$1/\lambda^2$
Gamma	a,λ	$\Gamma(a)^{-1}(\lambda x)^a e^{-\lambda x}x^{-1},$ for $x>0$	a/λ	a/λ^2
Beta	a,b	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1},$ for $0<x<1$	$\mu = \frac{a}{a+b}$	$\frac{\mu(1-\mu)}{a+b+1}$
χ^2	n	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2},$ for $x>0$	n	$2n$
Student-t	n	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+x^2/n)^{-(n+1)/2}$	0 if $n>1$	$\frac{n}{n-2}$ if $n>2$

Solutions to Stat 110 Final

Prof. Joe Blitzstein

Fall 2011

1. A Pois(λ) number of people vote in a certain election. Each voter votes for Candidate A with probability p and for Candidate B with probability $q = 1 - p$, independently of all the other voters. Let V be the difference in votes, defined as the number of votes for A minus the number of votes for B.

(a) Find $E(V)$ (simplify).

Let X and Y be the number of votes for A and B respectively, and let $N = X + Y$. Then $X|N \sim \text{Bin}(N, p)$ and $Y|N \sim \text{Bin}(N, q)$. By Adam's Law or the chicken-egg story, $E(X) = \lambda p$ and $E(Y) = \lambda q$. So

$$E(V) = E(X - Y) = E(X) - E(Y) = \lambda(p - q).$$

(b) Find Var(V) (simplify).

By the chicken-egg story, $X \sim \text{Pois}(\lambda p)$ and $Y \sim \text{Pois}(\lambda q)$ are independent. So

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) = \lambda p + \lambda q = \lambda.$$

2. Let X and Y be i.i.d. Gamma$(\frac{1}{2}, \frac{1}{2})$, and let $Z \sim \mathcal{N}(0,1)$ (note that X and Z may be dependent, and Y and Z may be dependent). For (a),(b),(c), write the most appropriate of $<$, $>$, $=$, or ? in each blank; for (d),(e),(f), write the most appropriate of $\leq$, $\geq$, $=$, or ? in each blank. It is *not* necessary to justify your answers for full credit, but partial credit may be available for justified answers that are flawed but on the right track.

(a) $P(X < Y) = 1/2$

This is since X and Y are i.i.d. continuous r.v.s.

(b) $P(X = Z^2)$? 1

This is since the probability is 0 if X and Z are independent, but it is 1 if X and Z^2 are the same r.v., which is possible since $Z^2 \sim \chi_1^2$, so $Z^2 \sim$ Gamma$(\frac{1}{2}, \frac{1}{2})$.

(c) $P(Z \geq \frac{1}{X^4+Y^4+7}) < 1$

This is since Z may be negative, and $\frac{1}{X^4+Y^4+7}$ is positive.

(d) $E(\frac{X}{X+Y})E((X+Y)^2) = E(X^2) + (E(X))^2$

By the bank-post office story, $X/(X+Y)$ and $(X+Y)^2$ are independent (and thus uncorrelated). So since X and Y are i.i.d., the lefthand side becomes

$$E(X(X+Y)) = E(X^2 + XY) = E(X^2) + E(XY) = E(X^2) + (E(X))^2.$$

(e) $E(X^2Z^2) \leq \sqrt{E(X^4)E(X^2)}$

By Cauchy-Schwarz, $E(X^2Z^2) \leq \sqrt{E(X^4)E(Z^4)}$. And $E(Z^4) = E(X^2)$ since X and Z^2 are χ_1^2, or since $E(Z^4) = 3$ (as shown in class) and $E(X^2) = \text{Var}(X) + (E(X))^2 = 2 + 1 = 3$.

(f) $E((X+2Y)^4) \geq 3^4$

This is true by Jensen's inequality, since $E(X+2Y) = 1 + 2 = 3$.

3. Ten million people enter a certain lottery. For each person, the chance of winning is one in ten million, independently.

(a) Find a simple, good approximation for the PMF of the number of people who win the lottery.

Let X be the number of people who win. Then $E(X) = \frac{10^7}{10^7} = 1$. A Poisson approximation is very good here since X is the number of "successes" for a very large number of independent trials where the probability of success on each trial is very low. So X is approximately Pois(1), and for k a nonnegative integer,

$$P(X = k) \approx \frac{1}{e \cdot k!}.$$

(b) Congratulations! You won the lottery. However, there may be other winners. Assume now that the number of winners other than you is $W \sim \text{Pois}(1)$, and that if there is more than one winner, then the prize is awarded to one randomly chosen winner. Given this information, find the probability that you win the prize (simplify).

Let A be the event that you win the prize, and condition on W:

$$P(A) = \sum_{k=0}^{\infty} P(A|W = k)P(W = k) = \frac{1}{e}\sum_{k=0}^{\infty} \frac{1}{k+1}\frac{1}{k!} = \frac{1}{e}\sum_{k=0}^{\infty} \frac{1}{(k+1)!} = \frac{e-1}{e} = 1 - \frac{1}{e}.$$

4. A drunken man wanders around randomly in a large space. At each step, he moves one unit of distance North, South, East, or West, with equal probabilities. Choose coordinates such that his initial position is $(0,0)$ and if he is at (x,y) at some time, then one step later he is at $(x,y+1),(x,y-1),(x+1,y)$, or $(x-1,y)$. Let (X_n,Y_n) and R_n be his position and distance from the origin after n steps, respectively.

General hint: note that X_n is a sum of r.v.s with possible values $-1,0,1$, and likewise for Y_n, but be careful throughout the problem about independence.

(a) Determine whether or not X_n is independent of Y_n (explain clearly).

They are *not* independent, as seen by considering an *extreme case* such as the event that the drunk headed East for the entire time: note that $P(Y_n=0|X_n=n)=1$.

(b) Find $\text{Cov}(X_n,Y_n)$ (simplify).

Write $X_n=\sum_{i=1}^n Z_i$ and $Y_n=\sum_{j=1}^n W_j$, where Z_i is -1 if his ith step is Westward, 1 if his ith step is Eastward, and 0 otherwise, and similarly for W_j. Then Z_i is independent of W_j for $i\neq j$. But Z_i and W_i are highly dependent: exactly one of them is 0 since he moves in one direction at a time. Then $\text{Cov}(Z_i,W_i)=E(Z_iW_i)-E(Z_i)E(W_i)=0$ since Z_iW_i is always 0, and Z_i and W_i have mean 0. So

$$\text{Cov}(X_n,Y_n)=\sum_{i,j}\text{Cov}(Z_i,W_j)=0.$$

(c) Find $E(R_n^2)$ (simplify).

We have $R_n^2=X_n^2+Y_n^2$, and $E(Z_iZ_j)=0$ for $i\neq j$. So

$$E(R_n^2)=E(X_n^2)+E(Y_n^2)=2E(X_n^2)=2nE(Z_1^2)=n,$$

since $Z_1^2\sim\text{Bern}(1/2)$.

5. Each of 111 people names his or her 5 favorite movies out of a list of 11 movies.

(a) Alice and Bob are 2 of the 111 people. Assume *for this part only* that Alice's 5 favorite movies out of the 11 are random, with all sets of 5 equally likely, and likewise for Bob, independently. Find the expected number of movies in common to Alice's and Bob's lists of favorite movies (simplify).

Let I_j be the indicator for the jth movie being on both lists, for $1 \leq j \leq 11$. By symmetry and linearity, the desired expected value is

$$11\left(\frac{5}{11}\right)^2 = \frac{25}{11}.$$

(This is essentially the same problem as the "mutual friends" of Alice and Bob problem from a previous midterm.)

(b) Show that there are 2 movies such that at least 21 of the people name both of these movies as favorites.

Choose 2 *random* movies (one at a time, without replacement). Let X be the number of people who name both of these movies. Creating an indicator r.v. for each person,

$$E(X) = 111P(\text{Alice names both random movies}) = 111\left(\frac{5}{11}\cdot\frac{4}{10}\right) = \left(\frac{111}{110}\right)20 > 20,$$

since the first chosen movie has a 5/11 chance of being on Alice's list and given that it is, the second chosen movie has a 4/10 chance of being on the list (or we can use the Hypergeometric PMF after "tagging" Alice's favorite movies).

Thus, there must exist 2 movies such that at least 21 of the people name both of them as favorites. (This is essentially the same problem as the "committees" existence problem from class.)

6. (a) A woman is pregnant, with a due date of January 10, 2012. Of course, the actual date on which she will give birth is not necessarily the due date. On a timeline, define time 0 to be the instant when January 10, 2012 begins. Suppose that the time T when the woman gives birth has a Normal distribution, centered at 0 and with standard deviation 8 days. What is the probability that she gives birth on her due date? (Your answer should be in terms of Φ, and simplified.)

We want to find $P(0 \le T < 1)$, where $T \sim \mathcal{N}(0, 64)$ is measured in days. This is

$$P(0 \le T/8 < 1/8) = \Phi(1/8) - \Phi(0) = \Phi(1/8) - 1/2.$$

(b) Another pregnant woman has the same due date as the woman from (a). Continuing with the setup of (a), let T_0 be the time of the first of the two births. Assume that the two birth times are i.i.d. Find the variance of T_0 (in terms of integrals, which do not need to be fully simplified).

Write $T_0 = \min(T_1, T_2)$ where T_j is the birth time for the jth woman. First find the PDF of T_0 (this can also be done using order statistics results):

$$P(T_0 > t) = P(T_1 > t, T_2 > t) = (1 - \Phi(t/8))^2,$$

so the PDF of t is the derivative of $1 - P(T_0 > t)$, which is

$$f_0(t) = \frac{1}{4}\left(1 - \Phi(t/8)\right)\varphi(t/8),$$

where $\varphi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$ is the $\mathcal{N}(0,1)$ PDF. Thus,

$$\text{Var}(T_0) = \int_{-\infty}^{\infty} t^2 f_0(t)dt - \left(\int_{-\infty}^{\infty} t f_0(t)dt\right)^2,$$

with f_0 as above. (With some additional work and using symmetry, it can be shown that this simplifies to $64(1 - \frac{1}{\pi})$, a neat-looking answer!)

7. Fred wants to sell his car, after moving back to Blissville. He decides to sell it to the first person to offer at least \$12,000 for it. Assume that the offers are independent Exponential random variables with mean \$6,000.

(a) Find the expected number of offers Fred will have (including the offer he accepts).

The offers on the car are i.i.d. $X_i \sim \text{Expo}(1/6000)$. So the number of offers that are too low is Geom(p) with $p = P(X_i \geq 12000) = \exp(-12000/6000) = e^{-2}$. Including the successful offer, the expected number of offers is thus $(1-p)/p + 1 = 1/p = e^2$.

(b) Find the expected amount of money that Fred gets for the car.

Let N be the number of offers, so the sale price of the car is X_N. Note that

$$E(X_N) = E(X|X \geq 12000)$$

for $X \sim \text{Expo}(1/6000)$, since the successful offer is an Exponential for which our information is that the value is at least \$12,000. To compute this, remember the memoryless property! For any $a > 0$, if $X \sim \text{Expo}(\lambda)$ then the distribution of $X - a$ given $X > a$ is itself Expo(λ). So

$$E(X|X \geq 12000) = 12000 + E(X) = 18000,$$

which shows that Fred's expected sale price is \$18,000.

8. Let G be an undirected network with nodes labeled $1, 2, \ldots, M$ (edges from a node to itself are not allowed), where $M \geq 2$ and random walk on this network is irreducible. Let d_j be the degree of node j for each j. Create a Markov chain on the state space $1, 2, \ldots, M$, with transitions as follows. From state i, generate a "proposal" j by choosing a uniformly random j such that there is an edge between i and j in G; then go to j with probability $\min(d_i/d_j, 1)$, and stay at i otherwise.

(a) Find the transition probability q_{ij} from i to j for this chain, for all states i, j (be sure to specify when this is 0, and to find q_{ii}, which you can leave as a sum).

First let $i \neq j$. If there is no $\{i, j\}$ edge, then $q_{ij} = 0$. If there is an $\{i, j\}$ edge, then

$$q_{ij} = (1/d_i)\min(d_i/d_j, 1) = \begin{cases} 1/d_i \text{ if } d_i \geq d_j, \\ 1/d_j \text{ if } d_i < d_j \end{cases},$$

since the proposal to go to j must be made and then accepted. For $i = j$, we have $q_{ii} = 1 - \sum_{j \neq i} q_{ij}$ since each row of the transition matrix must sum to 1.

(b) Find the stationary distribution of this chain (simplify).

Note that $q_{ij} = q_{ji}$ for all states i, j. This is clearly true if $i = j$ or $q_{ij} = 0$, so assume $i \neq j$ and $q_{ij} > 0$. If $d_i \geq d_j$, then $q_{ij} = 1/d_i$ and $q_{ji} = (1/d_j)(d_j/d_i) = 1/d_i$, while if $d_i < d_j$, then $q_{ij} = (1/d_i)(d_i/d_j) = 1/d_j$ and $q_{ji} = 1/d_j$.

Thus, the chain is reversible with respect to the uniform distribution over the states, and the stationary distribution is uniform over the states, i.e., state j has stationary probability $1/M$ for all j. (This is an example of the *Metropolis algorithm*, which was discussed in the ESP session and another example of which is in SP 11.)