Stat 110 Final Review, Fall 2012

Prof. Joe Blitzstein

1 General Information

The final will be on Friday 12/14, from 9 am to noon. Students with last names A-Q should go to Hall B, and students with last names R-Z should go to Hall C. No books, notes, copying, collaboration, computers, cell phones, or calculators are allowed, except that you may bring four pages of standard-sized paper (8.5" x 11") with anything you want written or typed on both sides. There will be 8 problems, equally weighted. The material covered will be cumulative since probability *is* cumulative.

To study, I recommend solving and re-solving lots and lots of practice problems! Often the best way to resolve a problem is to re-solve it. It's a good idea to work through as many of the problems on this handout as possible without looking at solutions (and then discussing with others and looking at solutions to check your answers and for any problems where you were really stuck), and to take at least two of the practice finals under timed conditions using only four pages of notes. Carefully studying class notes, homework solutions, and course handouts is also very important, and should be done *actively* (interweaving reading, reflecting, thinking, solving problems, and asking questions).

2 Topics

- Counting: multiplication rule, tree diagrams, permutations, binomial coefficients, sampling table, story proofs.
- Basic Probability: sample spaces, events, naive definition of probability, axioms of probability, odds, inclusion-exclusion, unions, intersections, complements.
- Conditional Probability: definition and meaning, probability of an intersection, Bayes' Rule, Law of Total Probability, thinking conditionally, wishful thinking, conditioning on the first step, independence vs. conditional independence, prior vs. posterior probability.
- Random Variables: definition and meaning, stories, discrete vs. continuous, distributions, CDFs, PMFs, PDFs, MGFs, indicator r.v.s, memoryless property, Universality of the Uniform, Poisson processes, Poisson approximation,

Normal approximation to Binomial, 68-95-99.7% Rule, Beta as conjugate prior for the Binomial, sums (convolutions), location and scale, order statistics.

- Expected Value: linearity, fundamental bridge, variance, standard deviation, covariance, correlation, LOTUS.

- Conditional Expectation: definition and meaning, taking out what's known, conditional variance, Adam's Law (iterated expectation), Eve's Law.

- Important Discrete Distributions: Bernoulli, Binomial, Geometric (and First Success), Negative Binomial, Hypergeometric, Poisson.

- Important Continuous Distributions: Uniform, Normal, Exponential, Gamma, Beta, Chi-Square (χ^2), Student-t.

- Jointly Distributed Random Variables: joint, conditional, and marginal distributions, independence, change of variables, Multinomial (story, joint PMF, lumping, covariance between components), Multivariate Normal (subvectors, linear combinations, uncorrelated implies independent within an MVN).

- Convergence: Law of Large Numbers, Central Limit Theorem.

- Inequalities: Cauchy-Schwarz, Markov, Chebyshev, Jensen.

- Markov Chains: Markov property, transition matrix, irreducibility, periodicity, stationary distributions, reversibility.

- General Concepts and Strategies: conditioning, stories, symmetry, linearity, indicator r.v.s, giving relevant objects names, pattern recognition, checking for category errors, checking simple and extreme cases.

- Some Important Examples and Stories: birthday problem, matching problem, Monty Hall, gambler's ruin, prosecutor's fallacy, testing for a disease, elk problem (capture-recapture), happy-sad men-women (2 by 2 table), chicken-egg problem, coupon (toy) collector, St. Petersburg paradox, Simpson's paradox, two envelope paradox, waiting time for HH vs. for HT, store with a random number of customers, bank-post office story, Bayes' billiards, random walk on an undirected network, Metropolis algorithm.

3 Important Distributions

3.1 Table of Distributions

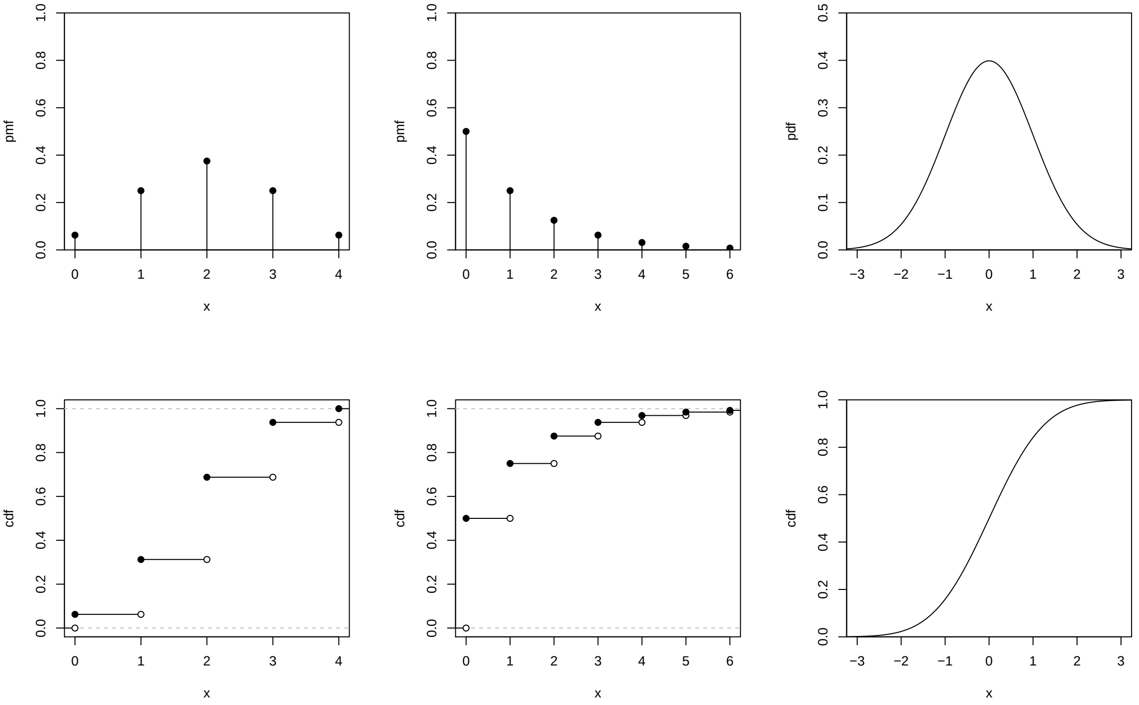
The table below *will be provided on the final.* Let $0 < p < 1$ and $q = 1 - p$.

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
FS	p	pq^{k-1}, for $k \in \{1, 2, \dots\}$	$1/p$	q/p^2
Geom	p	pq^k, for $k \in \{0, 1, 2, \dots\}$	q/p	q/p^2
NBinom	r, p	$\binom{r+n-1}{r-1}p^r q^n, n \in \{0, 1, 2, \dots\}$	rq/p	rq/p^2
HGeom	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
Expo	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 $n > 2$

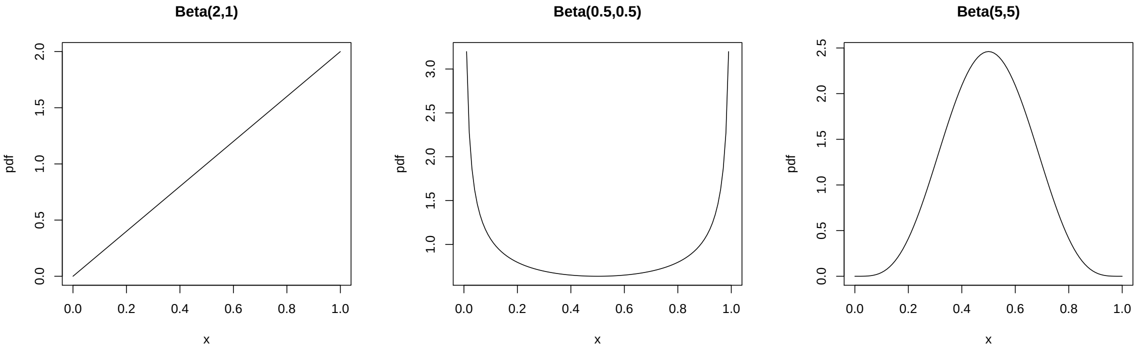
The function Γ is given by $\Gamma(a) = \int_0^\infty x^a e^{-x}\frac{dx}{x}$ for all $a > 0$. For any $a > 0$, $\Gamma(a+1) = a\Gamma(a)$. We have $\Gamma(n) = (n-1)!$ for n a positive integer, and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

3.2 Pretty PMF/PDF/CDF Plots

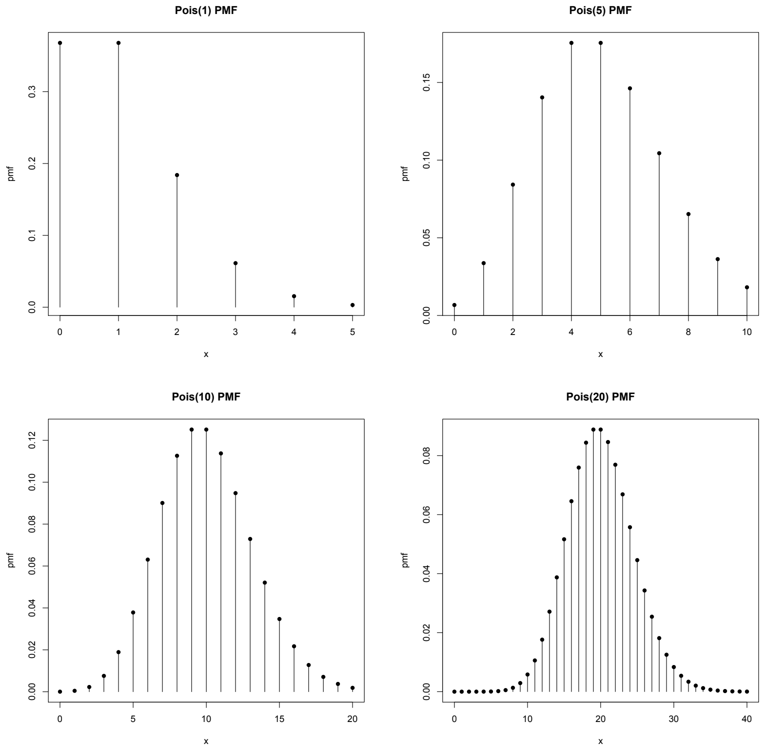
Here are plots of the Bin$(4, 1/2)$, Geom$(1/2)$, and $\mathcal{N}(0,1)$ PMFs/PDFs, with a plot of the corresponding CDF below each. Note that the CDFs are increasing, right continuous, converge to 0 as $x \to -\infty$, and converge to 1 as $x \to \infty$.



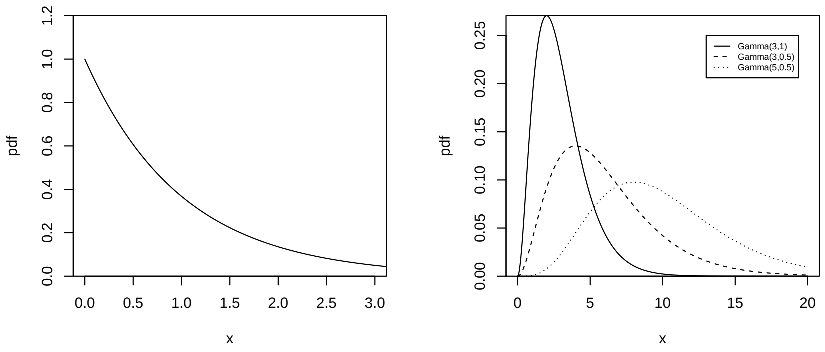
Next, we show a few of the many possible shapes of the PDF of a Beta (depicted are the Beta(2,1), Beta(0.5,0.5), and Beta(5,5) PDFs).



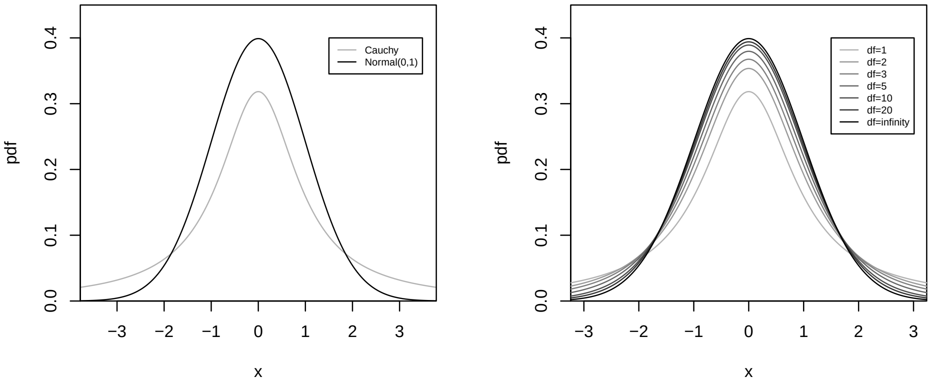
Here are Poisson PMFs with parameter values $1, 5, 10, 20$. By the Central Limit Theorem, for large n a Pois(n) r.v. is approximately $\mathcal{N}(n, n)$ in distribution, since Pois(n) is the distribution of the sum of n i.i.d. Pois(1) r.v.s.



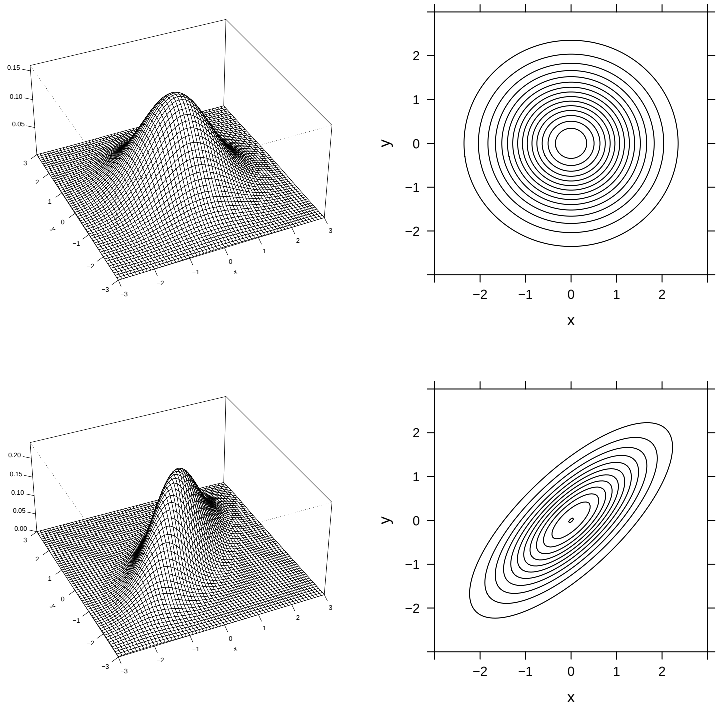
Next, we show the Expo(1) (i.e., Gamma$(1, 1)$) PDF on the left, and three other Gamma PDFs (Gamma$(3, 1)$, Gamma$(3, 0.5)$, Gamma$(5, 0.5)$) on the right.



The "evil" Cauchy PDF is shown below (lighter curve on the left), in comparison with the standard Normal PDF (darker curve); note that it has much heavier tails than the Normal. The Cauchy distribution is Student-t with 1 degree of freedom. On the right, we show Student-t distributions with various degrees of freedom. They get closer and closer to $\mathcal{N}(0,1)$ as the number of degrees of freedom grows.



To visualize the Bivariate Normal with $\mathcal{N}(0,1)$ marginals and correlation ρ, joint PDF plots (left) and contour plots (right) are shown for $\rho = 0$ and then for $\rho = 0.75$.



3.3 Connections Between Distributions

The table in Section 3.1 summarizes the PMFs/PDFs of the important distributions, and their means and variances, but it does not say where each distribution comes from (stories), or how the distributions are interconnected. Some of these connections between distributions are listed below.

Note that some of the important distributions are special cases of others: Bernoulli is a special case of Binomial; Geometric is a special case of Negative Binomial; Unif(0,1) is a special case of Beta; and Expo and χ^2 are both special cases of Gamma.

1. **Binomial**: If $X_1, \dots, X_n$ are i.i.d. Bern(p), then $X_1 + \dots + X_n \sim \text{Bin}(n, p)$.

2. **First Success**: The First Success (FS) distribution is just a shifted Geometric, adding 1 to include the success: $T \sim \text{FS}(p)$ is equivalent to $T - 1 \sim \text{Geom}(p)$.

3. **Neg. Binom.**: If $G_1, \dots, G_r$ are i.i.d. Geom(p), then $G_1 + \dots + G_r \sim \text{NBin}(r, p)$.

4. **Location and Scale**: If $Z \sim \mathcal{N}(0, 1)$, then $\mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$.

   If $U \sim \text{Unif}(0, 1)$ and $a < b$, then $a + (b - a)U \sim \text{Unif}(a, b)$.

   If $X \sim \text{Expo}(1)$, then $\lambda^{-1}X \sim \text{Expo}(\lambda)$.

   If $Y \sim \text{Gamma}(a, \lambda)$, then $\lambda Y \sim \text{Gamma}(a, 1)$.

5. **Symmetry**: If $X \sim \text{Bin}(n, 1/2)$, then $n - X \sim \text{Bin}(n, 1/2)$.

   If $U \sim \text{Unif}(0, 1)$, then $1 - U \sim \text{Unif}(0, 1)$.

   If $Z \sim \mathcal{N}(0, 1)$, then $-Z \sim \mathcal{N}(0, 1)$.

6. **Universality of the Uniform**: Let F be the CDF of a continuous r.v., such that F^{-1} exists. If $U \sim \text{Unif}(0, 1)$, then $F^{-1}(U)$ has CDF F. Conversely, if $X \sim F$, then $F(X) \sim \text{Unif}(0, 1)$.

7. **Uniform and Beta**: $\text{Unif}(0, 1)$ is the same distribution as $\text{Beta}(1, 1)$. The jth order statistic of n i.i.d. $\text{Unif}(0, 1)$ r.v.s is $\text{Beta}(j, n - j + 1)$.

8. **Beta and Binomial**: Beta is the conjugate prior to Binomial, in the sense that if $X|p \sim \text{Bin}(n, p)$ and the prior is $p \sim \text{Beta}(a, b)$, then the posterior is $p|X \sim \text{Beta}(a + X, b + n - X)$.

9. **Gamma**: If $X_1, \dots, X_n$ are i.i.d. Expo(λ), then $X_1 + \dots + X_n \sim \text{Gamma}(n, \lambda)$.

10. **Gamma and Poisson**: In a Poisson process of rate λ, the number of arrivals in a time interval of length t is $\text{Pois}(\lambda t)$, and the time of the nth arrival is $\text{Gamma}(n, \lambda)$.

11. **Gamma and Beta**: If $X \sim \text{Gamma}(a, \lambda)$, $Y \sim \text{Gamma}(b, \lambda)$ are independent, then $X/(X+Y) \sim \text{Beta}(a, b)$ is independent of $X + Y \sim \text{Gamma}(a+b, \lambda)$.

12. **Chi-Square**: χ^2_n is the same distribution as $\text{Gamma}(n/2, 1/2)$.

13. **Student-*t***: If $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_n$ are independent, then $\frac{Z}{\sqrt{Y/n}}$ has the Student-t distribution with n degrees of freedom. For $n = 1$, this is the Cauchy distribution, which is also the distribution of Z_1/Z_2 with Z_1, Z_2 i.i.d. $\mathcal{N}(0,1)$.

4 Sums of Independent Random Variables

Let $X_1, X_2, \dots, X_n$ be *independent* r.v.s. Here is the distribution of the sum in some nice cases.

X_i	$\sum_{i=1}^n X_i$
$\text{Bern}(p)$	$\text{Bin}(n, p)$
$\text{Bin}(m_i, p)$	$\text{Bin}(\sum_{i=1}^n m_i, p)$
$\text{Geom}(p)$	$\text{NBin}(n, p)$
$\text{NBin}(r_i, p)$	$\text{NBin}(\sum_{i=1}^n r_i, p)$
$\text{Pois}(\lambda_i)$	$\text{Pois}(\sum_{i=1}^n \lambda_i)$
Unif(0,1)	Triangle(0,1,2) ($n = 2$)
$\mathcal{N}(\mu_i, \sigma_i^2)$	$\mathcal{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$
$\text{Expo}(\lambda)$	$\text{Gamma}(n, \lambda)$
$\text{Gamma}(a_i, \lambda)$	$\text{Gamma}(\sum_{i=1}^n a_i, \lambda)$
Z_i^2, for $Z_i \sim \mathcal{N}(0,1)$	χ^2_n

5 Review of Some Useful Results

5.1 De Morgan's Laws

$$(A_1 \cup A_2 \cdots \cup A_n)^c = A_1^c \cap A_2^c \cdots \cap A_n^c,$$
$$(A_1 \cap A_2 \cdots \cap A_n)^c = A_1^c \cup A_2^c \cdots \cup A_n^c.$$

5.2 Complements

$$P(A^c) = 1 - P(A).$$

5.3 Unions

$$P(A \cup B) = P(A) + P(B) - P(A \cap B);$$

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i), \text{ if the } A_i \text{ are disjoint;}$$

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} P(A_i);$$

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{k=1}^{n} \left((-1)^{k+1} \sum_{i_1 < i_2 < \cdots < i_k} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) \right) \text{ (Inclusion-Exclusion).}$$

5.4 Intersections

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B),$$
$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \ldots, A_{n-1}).$$

5.5 Law of Total Probability

If $B_1, B_2, \ldots, B_n$ are a partition of the sample space S (i.e., they are disjoint and their union is all of S) and $P(B_i) > 0$ for all i, then

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

An analogous formula holds for conditioning on a continuous r.v. X with PDF $f(x)$:

$$P(A) = \int_{-\infty}^{\infty} P(A|X = x)f(x)dx.$$

Similarly, to go from a joint PDF $f(x, y)$ for (X, Y) to the marginal PDF of Y, integrate over all values of x:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx.$$

5.6 Bayes' Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Often the denominator $P(B)$ is then expanded by the Law of Total Probability. For continuous r.v.s X and Y, Bayes' Rule becomes

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}.$$

5.7 Expected Value, Variance, and Covariance

Expected value is *linear*: for any random variables X and Y and constant c,

$$E(X + Y) = E(X) + E(Y),$$

$$E(cX) = cE(X).$$

Variance can be computed in two ways:

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

Constants come out from variance as the constant squared:

$$\text{Var}(cX) = c^2\text{Var}(X).$$

For the variance of the sum, there is a covariance term:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y),$$

where

$$\text{Cov}(X, Y) = E((X - EX)(Y - EY)) = E(XY) - (EX)(EY).$$

So if X and Y are uncorrelated, then the variance of the sum is the sum of the variances. Recall that *independent implies uncorrelated but not vice versa.*
Covariance is symmetric:

$$\text{Cov}(Y, X) = \text{Cov}(X, Y).$$

Covariances of sums can be expanded as

$$\text{Cov}(X + Y, Z + W) = \text{Cov}(X, Z) + \text{Cov}(X, W) + \text{Cov}(Y, Z) + \text{Cov}(Y, W).$$

Note that for c a constant,

$$\text{Cov}(X, c) = 0,$$

$$\text{Cov}(cX, Y) = c\text{Cov}(X, Y).$$

The correlation of X and Y, which is between -1 and 1, is

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)}.$$

This is also the covariance of the standardized versions of X and Y.

5.8 Law of the Unconscious Statistician (LOTUS)

Let X be a discrete random variable and h be a real-valued function. Then $Y = h(X)$ is a random variable. To compute EY using the definition of expected value, we would need to first find the PMF of Y and use $EY = \sum_y yP(Y = y)$. The Law of the Unconscious Statistician says we can use the PMF of X directly:

$$Eh(X) = \sum_x h(x)P(X = x).$$

Similarly, for X a continuous r.v. with PDF $f_X(x)$, we can find the expected value of $Y = h(X)$ by integrating $h(x)$ times the PDF of X, without first finding $f_Y(y)$:

$$Eh(X) = \int_{-\infty}^{\infty} h(x)f_X(x)dx.$$

5.9 Indicator Random Variables

Let A and B be events. Indicator r.v.s bridge between probability and expectation: $P(A) = E(I_A)$, where I_A is the indicator r.v. for A. It is often useful to think of a “counting” r.v. as a sum of indicator r.v.s. Indicator r.v.s have many pleasant properties. For example, $(I_A)^k = I_A$ for any positive number k, so it's easy to handle moments of indicator r.v.s. Also note that

$$I_{A \cap B} = I_A I_B,$$

$$I_{A \cup B} = I_A + I_B - I_A I_B.$$

5.10 Symmetry

There are many beautiful and useful forms of symmetry in statistics. For example:

1. If X and Y are i.i.d., then $P(X < Y) = P(Y < X)$. More generally, if $X_1, \dots, X_n$ are i.i.d., then all orderings of $X_1, \dots, X_n$ are equally likely; in the continuous case it follows that $P(X_1 < X_2 < \cdots < X_n) = \frac{1}{n!}$, while in the discrete case we also have to consider ties.

2. If we shuffle a deck of cards and deal the first two cards, then the probability is 1/52 that the second card is the Ace of Spades, since by symmetry it's equally likely to be any card; it's not necessary to do a law of total probability calculation conditioning on the first card.

3. Consider the Hypergeometric, thought of as the distribution of the number of white balls, where we draw n balls from a jar with w white balls and b black balls (without replacement). By symmetry and linearity, we can immediately get that the expected value is $n\frac{w}{w+b}$, even though the trials are not independent, as the jth ball is equally likely to be any of the balls, and linearity still holds with dependent r.v.s.

4. By symmetry we can see immediately that if T is Cauchy, then $1/T$ is also Cauchy (since if we flip the ratio of two i.i.d. $\mathcal{N}(0,1)$ r.v.s, we still have the ratio of two i.i.d. $\mathcal{N}(0,1)$ r.v.s!).

5. $E(X_1|X_1 + X_2) = E(X_2|X_1 + X_2)$ by symmetry if X_1 and X_2 are i.i.d. So by linearity, $E(X_1|X_1 + X_2) + E(X_2|X_1 + X_2) = E(X_1 + X_2|X_1 + X_2) = X_1 + X_2$, which gives $E(X_1|X_1 + X_2) = (X_1 + X_2)/2$.

5.11 Change of Variables

Let $\mathbf{Y} = g(\mathbf{X})$, where g is an invertible function from $\mathbb{R}^n$ to such that all the first order partial derivatives exist and are continuous, and $\mathbf{X} = (X_1, \dots, X_n)$ is a continuous random vector with PDF $f_{\mathbf{X}}$. The PDF of $\mathbf{Y}$ can be found using a Jacobian as follows:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right|,$$

where $\mathbf{x} = g^{-1}(\mathbf{y})$ and $|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}|$ is the absolute value of the Jacobian determinant of g^{-1} (here $|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}|$ can either be found directly or by taking the reciprocal of $|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}|$).

In the case $n = 1$, this implies that if $Y = g(X)$ where g is differentiable with $g'(x) > 0$ everywhere, then

$$f_Y(y) = f_X(x) \frac{dx}{dy},$$

which is easily remembered if written in the form

$$f_Y(y)dy = f_X(x)dx.$$

Remember when using this that $f_Y(y)$ is a function of y (found by solving for x in terms of y), and the bounds for y should be specified. For example, if $y = e^x$ and x ranges over $\mathbb{R}$, then y ranges over $(0, \infty)$.

5.12 Order Statistics

Let $X_1, \dots, X_n$ be i.i.d. continuous r.v.s with PDF f and CDF F. The order statistics are obtained by sorting the X_i's, with $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$. The marginal PDF of the jth order statistic is

$$f_{X_{(j)}}(x) = n \binom{n-1}{j-1} f(x) F(x)^{j-1} (1 - F(x))^{n-j}.$$

5.13 Moment Generating Functions

The moment generating function of X is the function

$$M_X(t) = E(e^{tX}),$$

if this exists for all t in some open interval containing 0. For $X_1, \dots, X_n$ independent, the MGF of the sum $S_n = X_1 + \dots + X_n$ is

$$M_{S_n}(t) = M_{X_1}(t) \cdots M_{X_n}(t),$$

which is often much easier to deal with than a convolution sum or integral. Note that $M_X(0) = 1$. The name "moment generating function" comes from the fact that the derivatives of M_X at 0 give the moments of X:

$$M'_X(0) = E(X), M''_X(0) = E(X^2), M'''_X(0) = E(X^3), \dots.$$

Sometimes we can find *all* the moments simultaneously, without explicitly taking derivatives, by finding the Taylor series for $M_X(t)$. For example, the MGF of $X \sim \text{Expo}(1)$ is $\frac{1}{1-t}$ for $t < 1$, which is the geometric series $\sum_{n=0}^{\infty} t^n = \sum_{n=0}^{\infty} n! \frac{t^n}{n!}$. So the nth moment of X is $n!$ (since this is the coefficient of $\frac{t^n}{n!}$ in the series).

5.14 Conditional Expectation

The conditional expected value $E(Y|X = x)$ is a number (for each x) which is the average value of Y, given the information that $X = x$. The definition is analogous to the definition of EY: just replace the PMF or PDF by the conditional PMF or conditional PDF.

It is often very convenient to just directly condition on X to obtain $E(Y|X)$, which is a random variable (it is a function of X). This intuitively says to average Y, treating X as if it were a known constant: $E(Y|X = x)$ is a function of x, and $E(Y|X)$ is obtained from $E(Y|X = x)$ by "changing x to X". For example, if $E(Y|X = x) = x^3$, then $E(Y|X) = X^3$.

Important properties of conditional expectation:

$$E(Y_1 + Y_2|X) = E(Y_1|X) + E(Y_2|X) \text{ (linearity)};$$

$$E(Y|X) = E(Y) \text{ if } X \text{ and } Y \text{ are independent};$$

$$E(h(X)Y|X) = h(X)E(Y|X) \text{ (taking out what's known)};$$

$$E(Y) = E(E(Y|X)) \text{ (Adam's Law (iterated expectation))};$$

$$\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X)) \text{ (Eve's Law)}.$$

The latter two identities are often useful for finding the mean and variance of Y: first condition on some choice of X where the conditional distribution of Y given X is easier to work with than the unconditional distribution of Y, and then account for the randomness of X.

5.15 Convergence

Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance σ^2. The sample mean of the first n of these r.v.s is defined as

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The Strong Law of Large Numbers says that with probability 1, the sample mean converges to the true mean:

$$\bar{X}_n \to \mu \text{ with probability 1.}$$

The Weak Law of Large Numbers (which follows from Chebyshev's Inequality) says that $\bar{X}_n$ will be very close to μ with very high probability: for any $\epsilon > 0$,

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0 \text{ as } n \to \infty.$$

The Central Limit Theorem says that the sum of a large number of i.i.d. random variables is approximately Normal in distribution. More precisely, standardize the sum $X_1 + \cdots + X_n$ (by subtracting its mean and dividing by its standard deviation); then the standardized sum approaches $\mathcal{N}(0,1)$ in distribution (i.e., the CDF of the standardized sum converges to Φ). So

$$\frac{(X_1 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1) \text{ in distribution.}$$

In terms of the sample mean,

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \to \mathcal{N}(0,1) \text{ in distribution.}$$

5.16 Inequalities

When probabilities and expected values are hard to compute exactly, it is useful to have inequalities. One simple but handy inequality is Markov's Inequality:

$$P(X \geq a) \leq \frac{E|X|}{a},$$

for any $a > 0$. Let X have mean μ and variance σ^2. Using Markov's Inequality with $(X-\mu)^2$ in place of X gives Chebyshev's Inequality:

$$P(|X - \mu| \geq a) \leq \sigma^2/a^2.$$

For convex functions g (convexity of g is equivalent to $g''(x) \geq 0$ for all x, assuming this exists), there is Jensen's Inequality (the reverse inequality holds for concave g):

$$E(g(X)) \geq g(E(X)) \text{ for } g \text{ convex.}$$

The Cauchy-Schwarz inequality bounds the expected product of X and Y:

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}.$$

If X and Y have mean 0 and variance 1, this reduces to saying that the correlation is between -1 and 1. It follows that correlation is *always* between -1 and 1.

5.17 Markov Chains

Consider a Markov chain $X_0, X_1, \dots$ with transition matrix $Q = (q_{ij})$, and let $\mathbf{v}$ be a row vector listing the initial probabilities of being in each state. Then $\mathbf{v}Q^n$ is the row vector listing the probabilities of being in each state after n steps, i.e., the jth component is $P(X_n = j)$.

A row vector $\mathbf{s}$ of probabilities (adding to 1) is *stationary* for the chain if $\mathbf{s}Q = \mathbf{s}$; by the above, if a chain starts out with a stationary distribution then the distribution stays the same forever. Any irreducible Markov chain has a unique stationary distribution $\mathbf{s}$, and the chain converges to it: $P(X_n = i) \to s_i$ as $n \to \infty$.

If $\mathbf{s}$ is a row vector of probabilities (adding to 1) that satisfies the reversibility condition $s_i q_{ij} = s_j q_{ji}$ for all states i, j, then it automatically follows that $\mathbf{s}$ is a stationary distribution for the chain; not all chains have this condition hold, but for those that do it is often easier to show that $\mathbf{s}$ is stationary using the reversibility condition than by showing $\mathbf{s}Q = \mathbf{s}$.

6 Common Mistakes in Probability

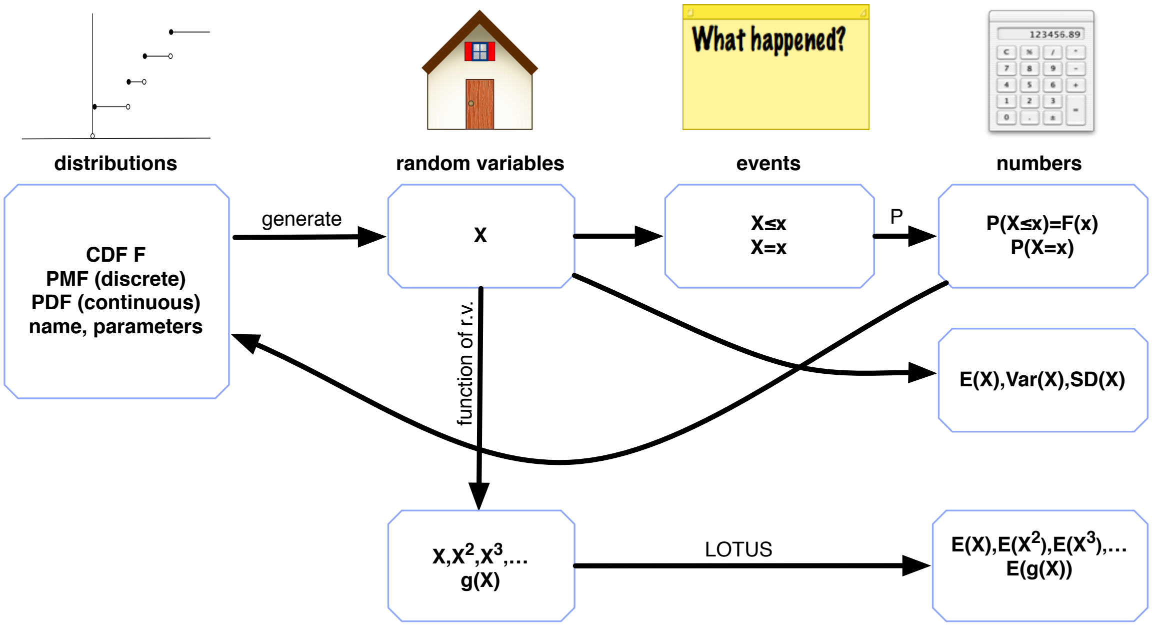


Figure 1: Four fundamental objects in probability: distributions (blueprints), random variables, events, and numbers. From a CDF F we can generate an r.v. $X \sim F$. From X, we can generate many other r.v.s by taking functions of X, and we can use LOTUS to find their expected values. The mean, variance, and standard deviation of X express the "average" and "spread" of the distribution of X (in particular, they only depend on F, not directly on X itself). There are various events describing the behavior of X. Most notably, for any constant x the events $X \leq x$ and $X = x$ are of interest. Knowing the probabilities of these events for all x gives us the CDF and (in the discrete case) the PMF, taking us full circle.

6.1 Category errors

A category error is a mistake that not only happens to be wrong, but also is wrong in *every possible universe.* If someone answers the question "How many students are in Stat 110?" with "10, since it's one ten," that is wrong (and a very bad approximation to the truth); but there is no *logical* reason the enrollment couldn't be 10, aside from the logical necessity of learning probability for reasoning about uncertainty in the world. But answering the question with "-42" or "π" or "pink elephants" would be a category error. To help avoid being categorically wrong, always think about what type an answer should have. Should it be an integer? A positive integer? A number between 0 and 1? A random variable? A distribution? An event? See Figure 1 and the midterm solutions for diagrams exploring the distinction and connections between distributions, r.v.s, events, and numbers.

- Probabilities must be between 0 and 1.

  Example: When asked for an approximation to $P(X > 5)$ for a certain r.v. X with mean 7, writing "$P(X > 5) \approx E(X)/5$." This makes two mistakes: Markov's inequality gives $P(X > 5) \leq E(X)/5$, but this is an *upper bound*, not an approximation; and here $E(X)/5 = 1.4$, which is silly as an approximation to a probability since $1.4 > 1$.

- Variances must be nonnegative.

  Example: For X and Y independent r.v.s, writing that "$\text{Var}(X - Y) = \text{Var}(X) - \text{Var}(Y)$", which can immediately be seen to be wrong from the fact that it becomes negative if $\text{Var}(Y) > \text{Var}(X)$ (and 0 if X and Y are i.i.d.). The correct formula is $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(-Y) - 2\text{Cov}(X, Y)$, which is $\text{Var}(X) + \text{Var}(Y)$ if X and Y are uncorrelated.

- Correlations must be between -1 and 1.

  Example: It is common to confuse covariance and correlation; they are related by $\text{Corr}(X, Y) = \text{Cov}(X, Y)/(\text{SD}(X)\text{SD}(Y))$, which is between -1 and 1.

- The range of possible values for an answer must make sense (e.g., when working with an r.v. keep in mind its support).

  Example: Two people each have 100 friends, and we are interested in the distribution of $X =$ (number of mutual friends). Then writing "$X \sim \mathcal{N}(\mu, \sigma^2)$" doesn't make sense since X is an *integer* (sometimes we use the Normal as an *approximation* to, say, Binomials, but exact answers should be given unless an approximation is specifically asked for); "$X \sim \text{Pois}(\lambda)$" or "$X \sim \text{Bin}(500, 1/2)$" don't make sense since X has possible values $0, 1, \ldots, 100$.

- Units must make sense, if there are units in a problem.

  Example: A common careless mistake is to divide by the variance rather than the standard deviation when standardizing. Thinking of X as having units makes it clear whether to divide by variance or standard deviation, e.g., if X is measured in light years, then $E(X)$ and $\text{SD}(X)$ are also measured in light years (whereas $\text{Var}(X)$ is measured in squared light years), so the standardized r.v. $\frac{X-E(X)}{\text{SD}(X)}$ is unitless (as desired).

Thinking about units also helps explain the change of variables formula,

$$f_X(x)dx = f_Y(y)dy,$$

when $Y = g(X)$ with g differentiable and $g'(x) > 0$ for all x. Probabilities must be unitless, and the dx and dy make this the case. For example, if X is measured in nanoseconds and $Y = X^3$, then the units of $f_X(x)$ are inverse nanoseconds and the units of $f_Y(y)$ are inverse cubed nanoseconds, and we need the dx and dy to make both sides unitless (we can think of $f_X(x)dx$ intuitively as the probability that X is in a tiny interval of length dx, centered at x).

- A number can't equal a random variable (unless the r.v. is actually a constant). Quantities such as $E(X), P(X > 1), F_X(1), \text{Cov}(X, Y)$ are *numbers*. We often use the notation "$X = x$", but this is shorthand for an *event* (it is the set of all possible outcomes of the experiment where X takes the value x).

  Example: A store has $N \sim \text{Pois}(\lambda)$ customers on a certain day, each of whom spends an average of μ dollars. Let X be the total amount spent by the customers. Then "$E(X) = N\mu$" doesn't make sense, since $E(X)$ is a number, while the righthand side is an r.v.

  Example: Writing something like "$\text{Cov}(X, Y) = 3$ if $Z = 0$ and $\text{Cov}(X, Y) = 1$ if $Z = 1$" doesn't make sense, as $\text{Cov}(X, Y)$ is just one number. Similarly, students sometimes write "$E(Y) = 3$ when $X = 1$" when they mean $E(Y|X = 1) = 3$. This is both conceptually wrong since $E(Y)$ is a number, the overall average of Y, and careless notation that could lead, e.g., to getting "$E(X) = 1$ if $X = 1$, and $E(X) = 0$ if $X = 0$" rather than $EX = p$ for $X \sim \text{Bern}(p)$.

- Don't replace an r.v. by its mean, or confuse $E(g(X))$ with $g(EX)$.

  Example: On the bidding for an unknown asset problem (#6 on the final from 2008), a common mistake is to replace the random asset value V by its mean, which completely ignores the *variability* of V.

  Example: If $X - 1 \sim \text{Geom}(1/2)$, then $2^{E(X)} = 4$, but $E(2^X)$ is infinite (as in the St. Petersburg Paradox), so confusing the two is infinitely wrong. In general, if g is convex then Jensen's inequality says that $E(g(X)) \geq g(EX)$.

- An event is not a random variable.

Example: If A is an event and X is an r.v., it does not make sense to write "$E(A)$" or "$P(X)$". There is of course a deep connection between events and r.v.s, in that for any event A there is a corresponding indicator r.v. I_A, and given an r.v. X and a number x, we have events $X = x$ and $X \leq x$.

- Dummy variables in an integral can't make their way out of the integral.

  Example: In LOTUS for an r.v. X with PDF f, the letter x in $E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$ is a dummy variable; we could just as well write $\int_{-\infty}^{\infty} g(t)f(t)dt$ or $\int_{-\infty}^{\infty} g(u)f(u)du$ or even $\int_{-\infty}^{\infty} g(\Box)f(\Box)d\Box$, but the x (or whatever this dummy variable is called) can't migrate out of the integral.

- After taking a limit as $n \to \infty$, n can't appear in the answer!

  Example: Let $X_1, X_2, \dots$ be i.i.d. with mean μ and variance σ^2, and let $\bar{X}_n$ be the sample mean of $X_1, \dots, X_n$. The Central Limit Theorem says that the distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converges to $\mathcal{N}(0, 1)$ as $n \to \infty$. This implies that for large n, the distribution of $\bar{X}_n$ is *approximately* $\mathcal{N}(\mu, \sigma^2/n)$; note that n can appear in the approximation but can't appear after taking the limit.

- A function must have the correct number and types of arguments.

  Example: The joint PMF of two discrete r.v.s X and Y is the function g given by $g(x, y) = P(X = x, Y = y)$, for all real x and y (this will be 0 if x is not in the support of Y or y is not in the support of Y). We could just as well have written $g(a, b) = P(X = a, Y = b)$, but it would not make sense to give a function of 1 variable or a function of 3 variables or a function of X.

  Example: Let X and Y be independent continuous r.v.s and $T = X + Y$. The PDF of T is given by the convolution $f_T(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t - x)dx$. In this integral, x is a dummy variable. Since T is a continuous r.v., it must have a valid PDF (which is a nonnegative function of one variable, integrating to 1).

- A random variable is not the same thing as its distribution! See Section 6.4.

- The conditional expectation $E(Y|X)$ must be a function of X (possibly a constant function, but it must be computable just in terms of X). See Section 6.6.

6.2 Notational paralysis

Another common mistake is a reluctance to introduce notation. This can be both a symptom and a cause of not seeing the structure of a problem. Be sure to define your notation clearly, carefully distinguishing between distributions, random variables, events, and numbers.

- Give objects names if you want to work with them.

  Example: Suppose that we are interested in a LogNormal r.v. X (so $\log(X) \sim \mathcal{N}(\mu, \sigma^2)$ for some μ, σ^2). Then $\log(X)$ is clearly an important object, so we should give it a name, say $Y = \log(X)$. Then, $X = e^Y$ and, for example, we can easily obtain the moments of X using the MGF of Y.

  Example: Suppose that we want to show that

  $$E(\cos^4(X^2+1)) \geq (E(\cos^2(X^2+1)))^2.$$

  The essential pattern is that there is an r.v. on the right and its square on the left; so let $Y = \cos^2(X^2+1)$, which turns the desired inequality into the statement $E(Y^2) \geq (EY)^2$, which we know is true because variance is nonnegative (and by Jensen's Inequality).

- Introduce clear notation for events and r.v.s of interest.

  Example: In the Calvin and Hobbes problem (from the 2010 final), clearly the event "Calvin wins the match" is important (so give it a name) and the r.v. "how many of the first two games Calvin wins" is important (so give it a name). Make sure that events you define really are events (they are subsets of the sample space, and it must make sense to talk about whether the event occurs) and that r.v.s you define really are r.v.s (they are functions mapping the sample space to the real line, and it must make sense to talk about their distributions and talk about them as a numerical summary of some aspect of the random experiment).

- People are not indistinguishable particles.

  Example: Consider the following problem: "You are invited to attend 6 weddings next year, independently with all months of the year equally likely. What is the probability that no two weddings are in the same month?" A common mistake is to treat the weddings as indistinguishable. But no matter how generic weddings can be sometimes, there must be *some* way to distinguish two weddings!

  Instead of treating people as indistinguishable particles or as a faceless mob, give the people names (or ID numbers). This also induces names for the weddings (unless a couple can get married and later get remarried, which can be handled by augmenting the notation). For example, we can say "ok, let's look at the possible schedulings of the weddings of Daenerys and Drogo, of Cersei and Robert, ...". There are 12^6 equally likely possibilities. Note that it is $6! = 720$ times more likely to have 1 wedding per month in January through June than to have all 6 weddings in January (whereas treating weddings as indistinguishable would say that these possibilities are equally likely).

- Think about location and scale when applicable.

  Example: If $Y_j \sim \text{Expo}(\lambda)$, it may be very convenient to work with $X_j = \lambda Y_j$, which is Expo(1). In studying $X \sim \mathcal{N}(\mu, \sigma^2)$, it may be very convenient to write $X = \mu + \sigma Z$ where $Z \sim \mathcal{N}(0,1)$ is the standardized version of X.

6.3 Common sense and checking answers

Whenever possible (i.e., when not under severe time pressure), look for simple ways to check your answers, or at least to check that they are plausible. This can be done in various ways, such as using the following methods.

1. *Miracle checks.* Does your answer seem plausible? Is there a category error? Did asymmetry appear out of nowhere when there should be symmetry?

2. *Checking simple and extreme cases.* What is the answer to a simpler version of the problem? What happens if $n = 1$ or $n = 2$, or as $n \to \infty$, if the problem involves showing something for all n?

3. *Looking for alternative approaches and connections with other problems.* Is there another natural way to think about the problem? Does the problem relate to other problems we've seen?

- Probability is full of counterintuitive results, but not impossible results!

  Example: Suppose that $P(\text{snow on Saturday}) = P(\text{snow on Sunday}) = 1/2$. Then we can't say "$P(\text{snow over the weekend}) = 1$"; clearly there is *some* chance of no snow. Of course, the mistake is to ignore the need for disjointness.

  Example: In finding $E(e^X)$ for $X \sim \text{Pois}(\lambda)$, obtaining an answer that can be negative, or an answer that isn't an increasing function of λ (intuitively, it is clear that larger λ should give larger average values of e^X).

- Check simple and extreme cases whenever possible.

  Example: Suppose we want to derive the mean and variance of a Hypergeometric, which is the distribution of the number of white balls if we draw n balls without replacement from a bag containing w white balls and b black balls. Suppose that using indicator r.v.s, we (correctly) obtain that the mean is $\mu = \frac{nw}{w+b}$ and the variance is $(\frac{w+b-n}{w+b-1})n\frac{\mu}{n}(1-\frac{\mu}{n})$.

  Let's check that this makes sense for the simple case $n = 1$: then the mean and variance reduce to those of a Bern$(w/(w+b))$, which makes sense since with only 1 draw, it doesn't matter whether sampling is with replacement.

  Now let's consider an extreme case where the total number of balls $(w + b)$ is extremely large compared with n. Then it shouldn't matter much whether the sampling is with or without replacement, so the mean and variance should be very close to those of a Bin$(n, w/(b + w))$, and indeed this is the case. If we had an answer that did not make sense in simple and extreme cases, we could then look harder for a mistake or explanation.

  Example: Let $X_1, X_2, \dots, X_{1000}$ be i.i.d. with a continuous distribution, and consider the question of whether the event $X_1 < X_2$ is independent of the event $X_1 < X_3$. Many students guess intuitively that they are independent. But now consider the more extreme question of whether $P(X_1 < X_2|X_1 < X_3, X_1 < X_4, \dots, X_1 < X_{1000})$ is $P(X_1 < X_2)$. Here most students guess intuitively (and correctly) that

  $$P(X_1 < X_2|X_1 < X_3, X_1 < X_4, \dots, X_1 < X_{1000}) > P(X_1 < X_2),$$

  since the evidence that X_1 is less than all of $X_3, \dots, X_{1000}$ suggests that X_1 is very small. Yet this more extreme case is the same in principle, just different in degree. Similarly, the Monty Hall problem is easier to understand with 1000 doors than with 3 doors. To show algebraically that $X_1 < X_2$ is not independent of $X_1 < X_3$, note that $P(X_1 < X_2) = 1/2$, while

  $$P(X_1 < X_2|X_1 < X_3) = \frac{P(X_1 < X_2, X_1 < X_3)}{P(X_1 < X_3)} = \frac{1/3}{1/2} = \frac{2}{3},$$

  where the numerator is $1/3$ since by symmetry, the smallest of X_1, X_2, X_3 is equally likely to be any of them.

  Example: Let M be the MGF of a r.v. X. Then $M(0) = E(e^0) = 1$. For example, we can immediately see that $2e^t$ is not a valid MGF.

- Check that PMFs are nonnegative and sum to 1, and PDFs are nonnegative and integrate to 1 (or that it is at least plausible), when it is not too messy.

  Example: Writing that the PDF of X is "$f(x) = \frac{1}{5}e^{-5x}$ for all $x > 0$ (and 0 otherwise)" is immediately seen to be wrong by integrating (the constant in front should be 5, which can also be seen by recognizing this as an Expo(5) PDF). Writing that the PDF is "$f(x) = \frac{1+e^{-x}}{1+x}$ for all $x > 0$ (and 0 otherwise)" doesn't make sense since even though the integral is hard to do directly, clearly $\frac{1+e^{-x}}{1+x} > \frac{1}{1+x}$, and $\int_0^\infty \frac{1}{1+x}dx$ is infinite.

6.4 Random variables vs. distributions

A random variable is not the same thing as its distribution! We call this confusion *sympathetic magic*, and the consequences of this confusion are often disastrous. Every random variable has a distribution (which can always be expressed using a CDF, which can be expressed by a PMF in the discrete case, and which can be expressed by a PDF in the continuous case).

Every distribution can be used as a blueprint for generating r.v.s (for example, one way to do this is using Universality of the Uniform). But that doesn't mean that doing something to an r.v. corresponds to doing it to the distribution of the r.v. Confusing a distribution with an r.v. with that distribution is like confusing a map of a city with the city itself, or a blueprint of a house with the house itself.

The word is not the thing, the map is not the territory.

- A function of an r.v. is an r.v.

  Example: Let X be discrete with possible values $0, 1, 2, \dots$ and PMF $p_j = P(X = j)$, and let $Y = X + 3$. Then Y is discrete with possible values $3, 4, 5, \dots$, and its PMF is given by $P(Y = k) = P(X = k - 3) = p_{k-3}$ for $k \in \{3, 4, 5, \dots\}$. In the continuous case, if $Y = g(X)$ with g differentiable and strictly increasing, then we can use the change of variables formula to find the PDF of Y from the PDF of X. If we only need $E(Y)$ and not the distribution of Y, we can use LOTUS. A common mistake is not seeing why these transformations of X are themselves r.v.s and how to handle them.

  Example: Let $Z \sim \mathcal{N}(0, 1), U = \Phi(Z)$, and $V = Z^2$. Then U and V are r.v.s since they are functions of Z. If after doing the experiment it turns out that $Z = z$ occurs, then the events $U = \Phi(z)$ and $V = z^2$ will occur. By Universality of the Uniform, we have $U \sim \text{Unif}(0, 1)$. By definition of the Chi-Square, we have $V \sim \chi^2_1$.

Example: For X, Y i.i.d., writing that "$E(\max(X,Y)) = E(X)$ since $\max(X,Y)$ is either X or Y, both of which have mean $E(X)$"; this misunderstands how and why $\max(X,Y)$ is an r.v. Of course, we should have $E(\max(X,Y)) \geq E(X)$ since $\max(X,Y) \geq X$.

- Avoid sympathetic magic.

  Example: Is it possible to have two r.v.s X, Y which have the same distribution but are *never* equal, i.e., the event $X = Y$ never occurs?

  Example: In finding the PDF of XY, writing something like "$f_X(x)f_Y(y)$." This is a category error since if we let $W = XY$, we want a function $f_W(w)$, not a function of two variables x and y. The mistake is in thinking that the PDF of the product is the product of the PDFs, which comes from not understanding well what a distribution really is.

  Example: For r.v.s X and Y with PDFs f_X and f_Y respectively, the event $X < Y$ is very different conceptually from the inequality $f_X < f_Y$. In fact, it is impossible that for all t, $f_X(t) < f_Y(t)$, since both sides integrate to 1.

- A CDF $F(x) = P(X \leq x)$ is a way to specify the distribution of X, and is a function defined for all real values of x. Here X is the r.v., and x is any number; we could just as well have written $F(t) = P(X \leq t)$.

  Example: Why must a CDF $F(x)$ be defined for all x and increasing everywhere, and why is it *not* true that a CDF integrates to 1?

6.5 Discrete vs. continuous r.v.s

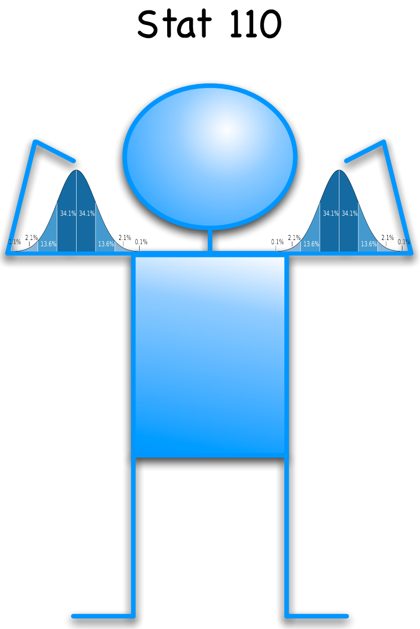
There are close connections between discrete r.v.s and continuous r.v.s, but there are key differences too.

- PMFs are for discrete r.v.s and PDFs are for continuous r.v.s. If X is a discrete r.v., then the derivative of the CDF is 0 everywhere, except at points where there is a "jump" (the jumps happen at points x with $P(X = x) > 0$, and the height of the jump is $P(X = x)$). At jump points the derivative does not exist. So there is not a useful notion of PDF for a discrete r.v. If Y is a continuous r.v., then $P(Y = y) = 0$ for all y, so there is not a useful notion of PMF for a continuous r.v.

- A PDF (evaluated at some point) is not a probability, e.g., a PDF can be greater than 1 at some points. We *integrate* a PDF to get a probability. Many results for PDFs are directly analogous to results for probability, e.g., we have the continuous and hybrid versions of Bayes' Rule, and convolution in the continuous case is analogous to in the discrete case. And as mentioned above, we can think of of $f_X(x)dx$ intuitively as the probability that X is in a tiny interval of length dx, centered at x. But care is needed since a density is not the same thing as a probability.

  Example: Let X and Y be independent positive r.v.s, with PDFs f_X and f_Y respectively, and let $T = XY$. On 2011 HW 8 #5, someone named Jacobno argues that "it's like a convolution, with a product instead of a sum. To have $T = t$ we need $X = x$ and $Y = t/x$ for some x; that has probability $f_X(x)f_Y(t/x)$, so summing up these possibilities we get that the PDF of T is $\int_0^\infty f_X(x)f_Y(t/x)dx$." But a more careful analysis shows that a Jacobian is needed: $f_T(t) = \int_0^\infty f_X(x)f_Y(t/x)\frac{dx}{x}$.

6.6 Conditioning



Conditioning is the soul of statistics.

It is easy to make mistakes with conditional probability, so it is important to think carefully about what to condition on and how to carry that out.

Conditioning is the soul of statistics.

- Condition on *all* the evidence!

  Example: In the Monty Hall problem, if Monty opens door 2 then we can't just use "$P(\text{door 1 has car}|\text{door 2 doesn't have car}) = 1/2$," since this does not condition on all the evidence: we know not just that door 2 does not have the car, but also that Monty opened door 2. *How the information was collected is itself information.* Why is this additional information relevant?

  To see this, contrast the problem as stated with the variant where Monty randomly chooses to open one of the 2 doors not picked by the contestant (so there is a chance of revealing the car and spoiling the game): different information is obtained in the two scenarios. This is another example where looking at an extreme case helps (consider the analogue of the Monty Hall problem with a billion doors).

  Example: In the murder problem (2011 HW 2 #5) a common mistake (often made by defense attorneys, intentionally or otherwise) is to focus attention on $P(\text{murder}|\text{abuse})$, which is irrelevant since we know the woman has been murdered, and we are interested in the probability of guilt given all the evidence (including the fact that the murder occurred).

- Don't destroy information.

  Example: Let $X \sim \text{Bern}(1/2)$ and $Y = 1 + W$ with $W \sim \text{Bern}(1/2)$ independent of X. Then writing "$E(X^2|X = Y) = E(Y^2) = 2.5$" is wrong (in fact, $E(X^2|X = Y) = 1$ since if $X = Y$, then $X = 1$), where the mistake is destroying the information $X = Y$, thinking we're done with that information once we have plugged in Y for X. A similar mistake is easy to make in the two envelope paradox.

  Example: On the bidding for an unknown asset problem (#6 on the final from 2008), a very common mistake is to forget to condition on the bid being accepted. In fact, we should have $E(V|\text{bid accepted}) < E(V)$ since if the bid is accepted, it restricts how much the asset could be worth (intuitively, this is similar to"buyer's remorse": it is common (though not necessarily rational) for someone to regret making an offer if the offer is accepted immediately, thinking that is a sign that a lower offer would have sufficed).

- Independence shouldn't be assumed without justification, and it is important to be careful not to implicitly assume independence without justification.

  Example: For $X_1, \dots, X_n$ i.i.d., we have $\text{Var}(X_1 + \cdots + X_n) = n\text{Var}(X_1)$, but this is not equal to $\text{Var}(X_1 + \cdots + X_1) = \text{Var}(nX_1) = n^2\text{Var}(X_1)$. For example, if X and Y are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, then $X + Y \sim \mathcal{N}(2\mu, 2\sigma^2)$, while $X + X = 2X \sim \mathcal{N}(2\mu, 4\sigma^2)$.

  Example: Is it always true that if $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\lambda)$, then $X + Y \sim \text{Pois}(2\lambda)$? What is an example of a sum of Bern(p)'s (with the same parameter p) which is not Binomial?

  Example: In the two envelope paradox, it is not true that the amount of money in the first envelope is independent of the indicator of which envelope has more money.

- Independence is completely different from disjointness!

  Example: Sometimes students try to visualize independent events A and B with two non-overlapping ovals in a Venn diagram. Such events in fact *can't* be independent (unless one has probability 0), since learning that A happened gives a great deal of information about B: it implies that B did not occur.

- Independence is a symmetric property: if A is independent of B, then B is independent of A. *There's no such thing as unrequited independence.*

  Example: If it is non-obvious whether A provides information about B but obvious that B provides information about A, then A and B can't be independent. Also, if $P(A) > 0$ and $P(B) > 0$, we are free to consider both $P(A|B)$ and $P(B|A)$, and then $P(A|B) = P(A)$ is equivalent to A and B being independent, and so is $P(B|A) = P(B)$. For example, in drawing balls from a jar we can look at $P(\text{first ball white}|\text{second ball white})$ even though chronologically the first ball is drawn before the second ball.

- The marginal distributions can be extracted from the joint distribution, but knowing the marginal distributions does not determine the joint distribution.

  Example: Calculations that are purely based on the marginal CDFs F_X and F_Y of dependent r.v.s X and Y may not shed much light on events such as $X < Y$ which involve X and Y jointly.

- Don't confuse a prior probability $P(A)$ with a posterior probability $P(A|B)$, a prior distribution with a posterior distribution, or a marginal distribution with a conditional distribution.

Example: Suppose that we observe that event B occurred. Then writing "$P(B) = 1$ since we know for sure that B happened" is careless; we have $P(B|B) = 1$, but $P(B)$ is the *prior probability* of B occurring.

Example: Suppose we are interested in whether someone has a hereditary disease, and we collect two pieces of information (e.g., data about two relatives, as on HW 2 #5, or the results of two independent diagnostic tests). Let D be the event that the person has the disease, and let B_1, B_2 be the pieces of evidence. Then $P(D)$ is the prior probability of D, before taking into account the evidence, and $P(D|B_1, B_2)$ is the posterior probability of D after conditioning on the evidence.

As shown on SP 2 #3.3 and discussed further on HW 2 #5, conditioning is coherent in the sense that we can update our beliefs in one step by conditioning on $B_1 \cap B_2$, or in two steps (e.g., we observe that B_1 occurred, so update to $P_{\text{new}}(A) = P(A|B_1)$ for all A, and then a few days later we learn that B_2 occurred, so we update again based on this information), with both ways yielding the same posterior probability $P(D|B_1, B_2)$. Writing "$P(D|B_1|B_2)$" would be invalid notation; there can be only one conditioning bar!

Example: On #3 of the Penultimate Homework, there was a Poisson process with an unknown rate λ and it was observed that $Y = y$ buses arrived in a time interval of length t. The prior was $\lambda \sim \text{Gamma}(r_0, b_0)$ (this must not involve y). The posterior also had a Gamma distribution, but with parameters updated based on the data: $\lambda|Y = y \sim \text{Gamma}(r_0 + y, b_0 + t)$. The conditional distribution of Y given λ was $\text{Pois}(\lambda t)$. But the marginal distribution of Y turned out to be $\text{NBin}(r_0, b_0/(b_0 + t))$ (this must not involve λ).

- Don't confuse $P(A|B)$ with $P(B|A)$.

  Example: This mistake is also known as the *prosecutor's fallacy* since it is often made in legal cases (but not always by the prosecutor!). For example, the prosecutor may argue that the probability of guilt given the evidence is very high by attempting to show that the probability of the evidence given innocence is very low, but in and of itself this is insufficient since it does not use the prior probability of guilt. Bayes' Rule thus becomes Bayes' Ruler, measuring the weight of the evidence by relating $P(A|B)$ to $P(B|A)$ and showing us how to update our beliefs based on evidence.

- Don't confuse $P(A|B)$ with $P(A, B)$.

Example: The law of total probability is often wrongly written without the weights as "$P(A) = P(A|B) + P(A|B^c)$" rather than $P(A) = P(A,B) + P(A,B^c) = P(A|B)P(B) + P(A|B^c)P(B^c)$.

- The expression $Y|X$ does not denote an r.v., when X and Y are r.v.s; it is notation indicating that in working with Y, we should use the conditional distribution of Y given X (i.e., treat X as a known constant). The expression $E(Y|X)$ *is* an r.v., and is a function of X (we have summed or integrated over the possible values of Y).

  Example: Writing "$E(Y|X) = Y$" is wrong, except if Y is a function of X, e.g., $E(X^3|X) = X^3$; by definition, $E(Y|X)$ must be $g(X)$ for some function g, so any answer for $E(Y|X)$ that is not of this form is a category error.

7 Stat 110 Final from 2006

(The mean was 84 out of 120 (70%), with a standard deviation of 20.)

1. The number of fish in a certain lake is a Pois(λ) random variable. Worried that there might be no fish at all, a statistician adds one fish to the lake. Let Y be the resulting number of fish (so Y is 1 plus a Pois(λ) random variable).

(a) Find $E(Y^2)$ (simplify).

(b) Find $E(1/Y)$ (in terms of λ; do not simplify yet).

(c) Find a simplified expression for $E(1/Y)$. Hint: $k!(k+1) = (k+1)!$.

2. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general.) It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

In (c) through (f), X and Y are i.i.d. (independent identically distributed) positive random variables. Assume that the various expected values exist.

(a) (probability that a roll of 2 fair dice totals 9) ____ (probability that a roll of 2 fair dice totals 10)

(b) (probability that 65% of 20 children born are girls) ____ (probability that 65% of 2000 children born are girls)

(c) $E(\sqrt{X})$ ____ $\sqrt{E(X)}$

(d) $E(\sin X)$ ____ $\sin(EX)$

(e) $P(X+Y>4)$ ____ $P(X>2)P(Y>2)$

(f) $E\left((X+Y)^2\right)$ ____ $2E(X^2)+2(EX)^2$

3. A fair die is rolled twice, with outcomes X for the 1st roll and Y for the 2nd roll.

(a) Compute the covariance of $X + Y$ and $X - Y$ (simplify).

(b) Are $X + Y$ and $X - Y$ independent? Justify your answer clearly.

(c) Find the moment generating function $M_{X+Y}(t)$ of $X + Y$ (your answer should be a function of t and can contain unsimplified finite sums).

4. A post office has 2 clerks. Alice enters the post office while 2 other customers, Bob and Claire, are being served by the 2 clerks. She is next in line. Assume that the time a clerk spends serving a customer has the Expo(λ) distribution.

(a) What is the probability that Alice is the last of the 3 customers to be done being served? (Simplify.) Justify your answer. Hint: no integrals are needed.

(b) Let X and Y be independent Expo(λ) r.v.s. Find the CDF of $\min(X, Y)$.

(c) What is the expected total time that Alice needs to spend at the post office?

5. Bob enters a casino with $X_0 = 1$ dollar and repeatedly plays the following game: with probability 1/3, the amount of money he has increases by a factor of 3; with probability 2/3, the amount of money he has decreases by a factor of 3. Let X_n be the amount of money he has after playing this game n times. Thus, X_{n+1} is $3X_n$ with probability 1/3 and is $3^{-1}X_n$ with probability 2/3.

(a) Compute $E(X_1)$, $E(X_2)$ and, in general, $E(X_n)$. (Simplify.)

(b) What happens to $E(X_n)$ as $n \to \infty$? Let Y_n be the number of times out of the first n games that Bob triples his money. What happens to Y_n/n as $n \to \infty$?

(c) Does X_n converge to some number c as $n \to \infty$ (with probability 1) and if so, what is c? Explain.

6. Let X and Y be independent standard Normal r.v.s and let $R^2 = X^2 + Y^2$ (where $R > 0$ is the distance from (X, Y) to the origin).

(a) The distribution of R^2 is an example of three of the "important distributions" listed on the last page. State which three of these distributions R^2 is an instance of, specifying the parameter values.

(b) Find the PDF of R. (Simplify.) Hint: start with the PDF $f_W(w)$ of $W = R^2$.

(c) Find $P(X > 2Y + 3)$ in terms of the standard Normal CDF Φ. (Simplify.)

(d) Compute $\text{Cov}(R^2, X)$. Are R^2 and X independent?

7. Let $U_1, U_2, \ldots, U_{60}$ be i.i.d. Unif(0,1) and $X = U_1 + U_2 + \cdots + U_{60}$.

(a) Which important distribution is the distribution of X very close to? Specify what the parameters are, and state which theorem justifies your choice.

(b) Give a simple but accurate approximation for $P(X > 17)$. Justify briefly.

(c) Find the moment generating function (MGF) of X.

8. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with $E(X_1) = 3$, and consider the sum $S_n = X_1 + X_2 + \cdots + X_n$.

(a) What is $E(X_1 X_2 X_3 | X_1)$? (Simplify. Your answer should be a function of X_1.)

(b) What is $E(X_1|S_n) + E(X_2|S_n) + \cdots + E(X_n|S_n)$? (Simplify.)

(c) What is $E(X_1|S_n)$? (Simplify.) Hint: use (b) and symmetry.

9. An urn contains red, green, and blue balls. Balls are chosen randomly with replacement (each time, the color is noted and then the ball is put back.) Let r, g, b be the probabilities of drawing a red, green, blue ball respectively ($r + g + b = 1$).

(a) Find the expected number of balls chosen before obtaining the first red ball, not including the red ball itself. (Simplify.)

(b) Find the expected number of different *colors* of balls obtained before getting the first red ball. (Simplify.)

(c) Find the probability that at least 2 of n balls drawn are red, given that at least 1 is red. (Simplify; avoid sums of large numbers of terms, and $\sum$ or $\cdots$ notation.)

10. Let $X_0, X_1, X_2, \dots$ be an irreducible Markov chain with state space $\{1, 2, \dots, M\}$, $M \geq 3$, transition matrix $Q = (q_{ij})$, and stationary distribution $\mathbf{s} = (s_1, \dots, s_M)$. The initial state X_0 is given the stationary distribution, i.e., $P(X_0 = i) = s_i$.

(a) On average, how many of $X_0, X_1, \dots, X_9$ equal 3? (In terms of $\mathbf{s}$; simplify.)

(b) Let $Y_n = (X_n - 1)(X_n - 2)$. For $M = 3$, find an example of Q (the transition matrix for the *original* chain $X_0, X_1, \dots$) where $Y_0, Y_1, \dots$ is Markov, and another example of Q where $Y_0, Y_1, \dots$ is not Markov. Mark which is which and briefly explain. In your examples, make $q_{ii} > 0$ for at least one i and make sure it is possible to get from any state to any other state eventually.

(c) If each column of Q sums to 1, what is $\mathbf{s}$? Verify using the definition of *stationary*.

8 Stat 110 Final from 2007

(The mean was 65 out of 108 (60%), with a standard deviation of 18.)

1. Consider the birthdays of 100 people. Assume people's birthdays are independent, and the 365 days of the year (exclude the possibility of February 29) are equally likely.

(a) Find the expected number of birthdays represented among the 100 people, i.e., the expected number of days that at least 1 of the people has as his or her birthday (your answer can involve unsimplified fractions but should not involve messy sums).

(b) Find the covariance between how many of the people were born on January 1 and how many were born on January 2.

2. Let X and Y be positive random variables, *not necessarily independent.* Assume that the various expected values below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general.) It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

(a) $(E(XY))^2$ ____ $E(X^2)E(Y^2)$

(b) $P(|X+Y|>2)$ ____ $\frac{1}{10}E((X+Y)^4)$

(c) $E(\ln(X+3))$ ____ $\ln(E(X+3))$

(d) $E(X^2e^X)$ ____ $E(X^2)E(e^X)$

(e) $P(X+Y=2)$ ____ $P(X=1)P(Y=1)$

(f) $P(X+Y=2)$ ____ $P\left(\{X\geq 1\}\cup\{Y\geq 1\}\right)$

3. Let X and Y be independent Pois(λ) random variables. Recall that the moment generating function (MGF) of X is $M(t) = e^{\lambda(e^t-1)}$.

(a) Find the MGF of $X + 2Y$ (simplify).

(b) Is $X + 2Y$ also Poisson? Show that it is, or that it isn't (whichever is true).

(c) Let $g(t) = \ln M(t)$ be the log of the MGF of X. Expanding $g(t)$ as a Taylor series

$$g(t) = \sum_{j=1}^{\infty} \frac{c_j}{j!} t^j$$

(the sum starts at $j = 1$ because $g(0) = 0$), the coefficient c_j is called the jth *cumulant* of X. Find c_j in terms of λ, for all $j \geq 1$ (simplify).

4. Consider the following conversation from an episode of *The Simpsons*:

> Lisa: *Dad, I think he's an ivory dealer! His boots are ivory, his hat is ivory, and I'm pretty sure that check is ivory.*
>
> Homer: *Lisa, a guy who's got lots of ivory is less likely to hurt Stampy than a guy whose ivory supplies are low.*

Here Homer and Lisa are debating the question of whether or not the man (named Blackheart) is likely to hurt Stampy the Elephant if they sell Stampy to him. They clearly disagree about how to use their observations about Blackheart to learn about the probability (conditional on the evidence) that Blackheart will hurt Stampy.

(a) Define clear notation for the various events of interest here.

(b) Express Lisa's and Homer's arguments (Lisa's is partly implicit) as conditional probability statements in terms of your notation from (a).

(c) Assume it is true that someone who has a lot of a commodity will have less desire to acquire more of the commodity. Explain what is wrong with Homer's reasoning that the evidence about Blackheart makes it less likely that he will harm Stampy.

5. Empirically, it is known that 49% of children born in the U.S. are girls (and 51% are boys). Let N be the number of children who will be born in the U.S. in March 2009, and assume that N is a Pois(λ) random variable, where λ is known. Assume that births are independent (e.g., don't worry about identical twins).

Let X be the number of girls who will be born in the U.S. in March 2009, and let Y be the number of boys who will be born then (note the importance of choosing good notation: boys have a Y chromosome).

(a) Find the joint distribution of X and Y. (Give the joint PMF.)

(b) Find $E(N|X)$ and $E(N^2|X)$.

6. Let X_1, X_2, X_3 be independent with $X_i \sim \text{Expo}(\lambda_i)$ (so with possibly different rates). A useful fact (which you may use) is that $P(X_1 < X_2) = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

(a) Find $E(X_1 + X_2 + X_3 | X_1 > 1, X_2 > 2, X_3 > 3)$ in terms of $\lambda_1, \lambda_2, \lambda_3$.

(b) Find $P(X_1 = \min(X_1, X_2, X_3))$, the probability that the first of the three Exponentials is the smallest. Hint: re-state this in terms of X_1 and $\min(X_2, X_3)$.

(c) For the case $\lambda_1 = \lambda_2 = \lambda_3 = 1$, find the PDF of $\max(X_1, X_2, X_3)$. Is this one of the "important distributions"?

7. Let $X_1, X_2, \ldots$ be i.i.d. random variables with CDF $F(x)$. For every number x, let $R_n(x)$ count how many of $X_1, \ldots, X_n$ are less than or equal to x.

(a) Find the mean and variance of $R_n(x)$ (in terms of n and $F(x)$).

(b) Assume (for this part only) that $X_1, \ldots, X_4$ are known constants. Sketch an example showing what the graph of the function $\frac{R_4(x)}{4}$ might look like. Is the function $\frac{R_4(x)}{4}$ necessarily a CDF? Explain briefly.

(c) Show that $\frac{R_n(x)}{n} \to F(x)$ as $n \to \infty$ (with probability 1).

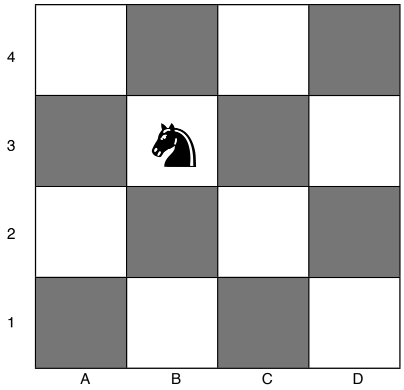
8. (a) Let T be a Student-t r.v. with 1 degree of freedom, and let $W = 1/T$. Find the PDF of W (simplify). Is this one of the "important distributions"?

Hint: no calculus is needed for this (though it can be used to check your answer).

(b) Let $W_n \sim \chi^2_n$ (the Chi-Square distribution with n degrees of freedom), for each $n \geq 1$. Do there exist a_n and b_n such that $a_n(W_n - b_n) \to \mathcal{N}(0,1)$ in distribution as $n \to \infty$? If so, find them; if not, explain why not.

(c) Let $Z \sim \mathcal{N}(0,1)$ and $Y = |Z|$. Find the PDF of Y, and approximate $P(Y < 2)$.

9. Consider a knight randomly moving around on a 4 by 4 chessboard:



The 16 squares are labeled in a grid, e.g., the knight is currently at the square B3, and the upper left square is A4. Each move of the knight is an L-shape: two squares horizontally followed by one square vertically, or vice versa. For example, from B3 the knight can move to A1, C1, D2, or D4; from A4 it can move to B2 or C3. Note that from a white square, the knight always moves to a gray square and vice versa.

At each step, the knight moves randomly, each possibility equally likely. Consider the stationary distribution of this Markov chain, where the states are the 16 squares.

(a) Which squares have the highest stationary probability? Explain very briefly.

(b) Compute the stationary distribution (simplify). Hint: random walk on a graph.

9 Stat 110 Final from 2008

(The mean was 77 out of 100, with a standard deviation of 13.)

1. Joe's iPod has 500 different songs, consisting of 50 albums of 10 songs each. He listens to 11 random songs on his iPod, with all songs equally likely and chosen independently (so repetitions may occur).

(a) What is the PMF of how many of the 11 songs are from his favorite album?

(b) What is the probability that there are 2 (or more) songs from the same album among the 11 songs he listens to? (Do not simplify.)

(c) A pair of songs is a "match" if they are from the same album. If, say, the 1st, 3rd, and 7th songs are all from the same album, this counts as 3 matches. Among the 11 songs he listens to, how many matches are there on average? (Simplify.)

2. Let X and Y be *positive* random variables, *not necessarily independent.* Assume that the various expressions below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general.) It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

(a) $P(X+Y>2)$ ____ $\frac{EX+EY}{2}$

(b) $P(X+Y>3)$ ____ $P(X>3)$

(c) $E(\cos(X))$ ____ $\cos(EX)$

(d) $E(X^{1/3})$ ____ $(EX)^{1/3}$

(e) $E(X^Y)$ ____ $(EX)^{EY}$

(f) $E\left(E(X|Y)+E(Y|X)\right)$ ____ $EX+EY$

3. (a) A woman is pregnant with twin boys. Twins may be either identical or fraternal (non-identical). In general, 1/3 of twins born are identical. Obviously, identical twins must be of the same sex; fraternal twins may or may not be. Assume that identical twins are equally likely to be both boys or both girls, while for fraternal twins all possibilities are equally likely. Given the above information, what is the probability that the woman's twins are identical?

(b) A certain genetic characteristic is of interest. For a random person, this has a numerical value given by a $\mathcal{N}(0, \sigma^2)$ r.v. Let X_1 and X_2 be the values of the genetic characteristic for the twin boys from (a). If they are identical, then $X_1 = X_2$; if they are fraternal, then X_1 and X_2 have correlation ρ. Find $\text{Cov}(X_1, X_2)$ in terms of ρ, σ^2.

4. (a) Consider i.i.d. Pois(λ) r.v.s $X_1, X_2, \ldots$. The MGF of X_j is $M(t) = e^{\lambda(e^t-1)}$. Find the MGF $M_n(t)$ of the sample mean $\bar{X}_n = \frac{1}{n}\sum_{j=1}^n X_j$. (Hint: it may help to do the $n = 2$ case first, which itself is worth a lot of partial credit, and then generalize.)

(b) Find the limit of $M_n(t)$ as $n \to \infty$. (You can do this with almost no calculation using a relevant theorem; or you can use (a) and that $e^x \approx 1 + x$ if x is very small.)

5. A post office has 2 clerks. Alice enters the post office while 2 other customers, Bob and Claire, are being served by the 2 clerks. She is next in line. Assume that the time a clerk spends serving a customer has the Expo(λ) distribution.

(a) What is the probability that Alice is the last of the 3 customers to be done being served? Justify your answer. Hint: no integrals are needed.

(b) Let X and Y be independent Expo(λ) r.v.s. Find the CDF of $\min(X, Y)$.

(c) What is the expected total time that Alice needs to spend at the post office?

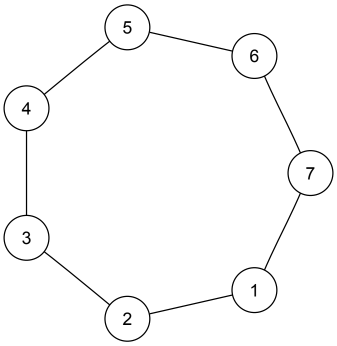
6. You are given an amazing opportunity to bid on a mystery box containing a mystery prize! The value of the prize is completely unknown, except that it is worth at least nothing, and at most a million dollars. So the true value V of the prize is considered to be Uniform on [0,1] (measured in millions of dollars).

You can choose to bid any amount b (in millions of dollars). You have the chance to get the prize for considerably less than it is worth, but you could also lose money if you bid too much. Specifically, if $b < \frac{2}{3}V$, then the bid is rejected and nothing is gained or lost. If $b \geq \frac{2}{3}V$, then the bid is accepted and your net payoff is $V - b$ (since you pay b to get a prize worth V). What is your optimal bid b (to maximize the expected payoff)?

7. (a) Let $Y = e^X$, with $X \sim \text{Expo}(3)$. Find the mean and variance of Y (simplify).

(b) For $Y_1, \dots, Y_n$ i.i.d. with the same distribution as Y from (a), what is the approximate distribution of the sample mean $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$ when n is large? (Simplify, and specify all parameters.)

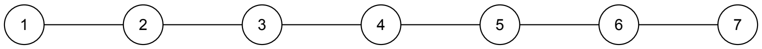
8.



(a) Consider a Markov chain on the state space $\{1, 2, \ldots, 7\}$ with the states arranged in a "circle" as shown above, and transitions given by moving one step clockwise or counterclockwise with equal probabilities. For example, from state 6, the chain moves to state 7 or state 5 with probability 1/2 each; from state 7, the chain moves to state 1 or state 6 with probability 1/2 each. The chain starts at state 1.

Find the stationary distribution of this chain.

(b) Consider a new chain obtained by "unfolding the circle." Now the states are arranged as shown below. From state 1 the chain always goes to state 2, and from state 7 the chain always goes to state 6. Find the new stationary distribution.



10 Stat 110 Final from 2009

(The mean was 75 out of 100, with a standard deviation of 15.)

1. A group of n people play "Secret Santa" as follows: each puts his or her name on a slip of paper in a hat, picks a name randomly from the hat (without replacement), and then buys a gift for that person. Unfortunately, they overlook the possibility of drawing one's own name, so some may have to buy gifts for themselves (on the bright side, some may like self-selected gifts better). Assume $n \geq 2$.

(a) Find the expected number of people who pick their own names (simplify).

(b) Find the expected number of pairs of people, A and B, such that A picks B's name and B picks A's name (where $A \neq B$ and order doesn't matter; simplify).

(c) Let X be the number of people who pick their own names. Which of the "important distributions" are conceivable as the distribution of X, just based on the possible values X takes (you do not need to list parameter values for this part)?

(d) What is the *approximate* distribution of X if n is large (specify the parameter value or values)? What does $P(X = 0)$ converge to as $n \to \infty$?

2. Let X and Y be positive random variables, *not necessarily independent.* Assume that the various expected values below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general.) It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

(a) $E(X^3)$ ____ $\sqrt{E(X^2)E(X^4)}$

(b) $P(|X+Y|>2)$ ____ $\frac{1}{16}E((X+Y)^4)$

(c) $E(\sqrt{X+3})$ ____ $\sqrt{E(X+3)}$

(d) $E(\sin^2(X)) + E(\cos^2(X))$ ____ 1

(e) $E(Y|X+3)$ ____ $E(Y|X)$

(f) $E(E(Y^2|X))$ ____ $(EY)^2$

3. Let $Z \sim \mathcal{N}(0,1)$. Find the 4th moment $E(Z^4)$ in the following two different ways:

(a) using what you know about how certain powers of Z are related to other distributions, along with information from the table of distributions.

(b) using the MGF $M(t) = e^{t^2/2}$, by writing down its Taylor series and using how the coefficients relate to moments of Z, *not* by tediously taking derivatives of $M(t)$. Hint: you can get this series immediately from the Taylor series for e^x.

4. A chicken lays n eggs. Each egg independently does or doesn't hatch, with probability p of hatching. For each egg that hatches, the chick does or doesn't survive (independently of the other eggs), with probability s of survival. Let $N \sim \text{Bin}(n, p)$ be the number of eggs which hatch, X be the number of chicks which survive, and Y be the number of chicks which hatch but don't survive (so $X + Y = N$).

(a) Find the distribution of X, preferably with a clear explanation in words rather than with a computation. If X has one of the "important distributions," say which (including its parameters).

(b) Find the joint PMF of X and Y (simplify).

(c) Are X and Y independent? Give a clear explanation in words (of course it makes sense to see if your answer is consistent with your answer to (b), but you can get full credit on this part even without doing (b); conversely, it's not enough to just say "by (b), . . ." without further explanation).

5. Suppose we wish to approximate the following integral (denoted by b):

$$b = \int_{-\infty}^{\infty} (-1)^{\lfloor x \rfloor} e^{-x^2/2} dx,$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x (e.g., $\lfloor 3.14 \rfloor = 3$).

(a) Write down a function $g(x)$ such that $E(g(X)) = b$ for $X \sim \mathcal{N}(0,1)$ (your function should *not* be in terms of b, and should handle normalizing constants carefully).

(b) Write down a function $h(u)$ such that $E(h(U)) = b$ for $U \sim \text{Unif}(0,1)$ (your function should *not* be in terms of b, and can be in terms of the function g from (a) and the standard Normal CDF Φ).

(c) Let $X_1, X_2, \dots, X_n$ be i.i.d. $\mathcal{N}(0,1)$ with n large, and let g be as in (a). What is the approximate distribution of $\frac{1}{n}\left(g(X_1) + \cdots + g(X_n)\right)$? Simplify the parameters fully (in terms of b and n), and mention which theorems you are using.

6. Let X_1 be the number of emails received by a certain person today and let X_2 be the number of emails received by that person tomorrow, with X_1 and X_2 i.i.d.

(a) Find $E(X_1|X_1 + X_2)$ (simplify).

(b) For the case $X_j \sim \text{Pois}(\lambda)$, find the conditional distribution of X_1 given $X_1 + X_2$, i.e., $P(X_1 = k|X_1 + X_2 = n)$ (simplify). Is this one of the "important distributions"?

7. Let X_1, X_2, X_3 be independent with $X_i \sim \text{Expo}(\lambda_i)$ (so with possibly different rates). A useful fact (which you may use) is that $P(X_1 < X_2) = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

(a) Find $E(X_1 + X_2 + X_3 | X_1 > 1, X_2 > 2, X_3 > 3)$ in terms of $\lambda_1, \lambda_2, \lambda_3$.

(b) Find $P(X_1 = \min(X_1, X_2, X_3))$, the probability that the first of the three Exponentials is the smallest. Hint: re-state this in terms of X_1 and $\min(X_2, X_3)$.

(c) For the case $\lambda_1 = \lambda_2 = \lambda_3 = 1$, find the PDF of $\max(X_1, X_2, X_3)$. Is this one of the "important distributions"?

8. Let X_n be the price of a certain stock at the start of the nth day, and assume that $X_0, X_1, X_2, \dots$ follows a Markov chain with transition matrix Q (assume for simplicity that the stock price can never go below 0 or above a certain upper bound, and that it is always rounded to the nearest dollar).

(a) A lazy investor only looks at the stock once a year, observing the values on days $0, 365, 2 \cdot 365, 3 \cdot 365, \dots$. So the investor observes $Y_0, Y_1, \dots$, where Y_n is the price after n years (which is $365n$ days; you can ignore leap years). Is $Y_0, Y_1, \dots$ also a Markov chain? Explain why or why not; if so, what is its transition matrix?

(b) The stock price is always an integer between \$0 and \$28. From each day to the next, the stock goes up or down by \$1 or \$2, all with equal probabilities (except for days when the stock is at or near a boundary, i.e., at \$0, \$1, \$27, or \$28).

If the stock is at \$0, it goes up to \$1 or \$2 on the next day (after receiving government bailout money). If the stock is at \$28, it goes down to \$27 or \$26 the next day. If the stock is at \$1, it either goes up to \$2 or \$3, or down to \$0 (with equal probabilities); similarly, if the stock is at \$27 it either goes up to \$28, or down to \$26 or \$25. Find the stationary distribution of the chain (simplify).

11 Stat 110 Final from 2010

(The mean was 62 out of 100, with a standard deviation of 20.)

1. Calvin and Hobbes play a match consisting of a series of games, where Calvin has probability p of winning each game (independently). They play with a "win by two" rule: the first player to win two games more than his opponent wins the match.

(a) What is the probability that Calvin wins the match (in terms of p)?

Hint: condition on the results of the first k games (for some choice of k).

(b) Find the expected number of games played.

Hint: consider the first two games as a pair, then the next two as a pair, etc.

2. A DNA sequence can be represented as a sequence of letters, where the "alphabet" has 4 letters: A,C,T,G. Suppose such a sequence is generated randomly, where the letters are independent and the probabilities of A,C,T,G are p_1, p_2, p_3, p_4 respectively.

(a) In a DNA sequence of length 115, what is the expected number of occurrences of the expression "CATCAT" (in terms of the p_j)? (Note that, for example, the expression "CATCATCAT" counts as 2 occurrences.)

(b) What is the probability that the first A appears earlier than the first C appears, as letters are generated one by one (in terms of the p_j)?

(c) For this part, assume that the p_j are unknown. Suppose we treat p_2 as a $\text{Unif}(0, 1)$ r.v. before observing any data, and that then the first 3 letters observed are "CAT". Given this information, what is the probability that the next letter is C?

3. Let X and Y be i.i.d. *positive* random variables. Assume that the various expressions below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general). It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

(a) $E(e^{X+Y})$ ____ $e^{2E(X)}$

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

(c) $E(X|3X)$ ____ $E(X|2X)$

(d) $E(X^7Y)$ ____ $E(X^7E(Y|X))$

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

(f) $P(|X-Y|>2)$ ____ $\frac{\mathrm{Var}(X)}{2}$

4. Let X be a discrete r.v. whose distinct possible values are $x_0, x_1, \dots,$ and let $p_k = P(X = x_k)$. The *entropy* of X is defined to be $H(X) = -\sum_{k=0}^{\infty} p_k \log_2(p_k)$.

(a) Find $H(X)$ for $X \sim \text{Geom}(p)$.

Hint: use properties of logs, and interpret part of the sum as an expected value.

(b) Find $H(X^3)$ for $X \sim \text{Geom}(p)$, in terms of $H(X)$.

(c) Let X and Y be i.i.d. discrete r.v.s. Show that $P(X = Y) \geq 2^{-H(X)}$.

Hint: Consider $E(\log_2(W))$, where W is an r.v. taking value p_k with probability p_k.

5. Let $Z_1, \dots, Z_n \sim \mathcal{N}(0,1)$ be i.i.d.

(a) As a function of Z_1, create an Expo(1) r.v. X (your answer can also involve the standard Normal CDF Φ).

(b) Let $Y = e^{-R}$, where $R = \sqrt{Z_1^2 + \cdots + Z_n^2}$. Write down (but do not evaluate) an integral for $E(Y)$.

(c) Let $X_1 = 3Z_1 - 2Z_2$ and $X_2 = 4Z_1 + 6Z_2$. Determine whether X_1 and X_2 are independent (being sure to mention which results you're using).

6. Let $X_1, X_2, \dots$ be i.i.d. positive r.v.s. with mean μ, and let $W_n = \frac{X_1}{X_1 + \dots + X_n}$.

(a) Find $E(W_n)$.

Hint: consider $\frac{X_1}{X_1+\dots+X_n} + \frac{X_2}{X_1+\dots+X_n} + \dots + \frac{X_n}{X_1+\dots+X_n}$.

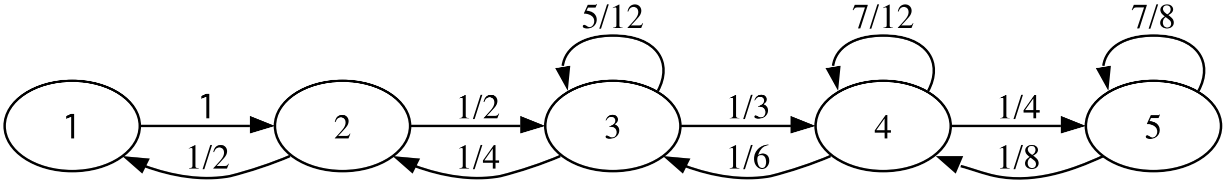
(b) What random variable does nW_n converge to as $n \to \infty$?

(c) For the case that $X_j \sim \text{Expo}(\lambda)$, find the distribution of W_n, preferably without using calculus. (If it is one of the "important distributions" state its name and specify the parameters; otherwise, give the PDF.)

7. A task is randomly assigned to one of two people (with probability 1/2 for each person). If assigned to the first person, the task takes an Expo(λ_1) length of time to complete (measured in hours), while if assigned to the second person it takes an Expo(λ_2) length of time to complete (independent of how long the first person would have taken). Let T be the time taken to complete the task.

(a) Find the mean and variance of T.

(b) Suppose instead that the task is assigned to *both* people, and let X be the time taken to complete it (by whoever completes it first, with the two people working independently). It is observed that after 24 hours, the task has not yet been completed. Conditional on this information, what is the expected value of X?



8. Find the stationary distribution of the Markov chain shown above, *without using matrices.* The number above each arrow is the corresponding transition probability.

12 Stat 110 Final from 2011

(The mean was 71 out of 100, with a standard deviation of 20.)

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

(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 $\mathrm{Cov}(X_n,Y_n)$ (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.

[Hint (there was not a hint here in the actual final, since that year I did a very similar problem in class): show that for 2 *random* movies, chosen without replacement, the expected number of people who name both movies is greater than 20. This implies the desired result since if all the numbers in a certain list of integers are at most 20, then the average of the list is at most 20.]

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).