Stat 110 Strategic Practice 6, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1 Exponential Distribution and Memorylessness

1. Fred (the protagonist of HW 6 #1) wants to sell his car, after moving back to Blissville (where he is happy with the bus system). He decides to sell it to the first person to offer at least \$15,000 for it. Assume that the offers are independent Exponential random variables with mean \$10,000.

   (a) Find the expected number of offers Fred will have.

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

2. Find $E(X^3)$ for $X \sim \text{Expo}(\lambda)$, using LOTUS and the fact that $E(X) = 1/\lambda$ and $\text{Var}(X) = 1/\lambda^2$, and integration by parts at most once (see also Problem 1 in the MGF section).

3. Let $X_1, \dots, X_n$ be independent, with $X_j \sim \text{Expo}(\lambda_j)$. (They are i.i.d. if all the λ_j's are equal, but we are not assuming that.) Let $M = \min(X_1, \dots, X_n)$. Show that $M \sim \text{Expo}(\lambda_1 + \dots + \lambda_n)$, and interpret this intuitively.

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 Exponential(λ) distribution.

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

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

2 Moment Generating Functions (MGFs)

1. Find $E(X^3)$ for $X \sim \text{Expo}(\lambda)$ using the MGF of X (see also Problem 2 in the Exponential Distribution section).

2. If X has MGF $M(t)$, what is the MGF of $-X$? What is the MGF of $a + bX$, where a and b are constants?

3. Let $U_1, U_2, \ldots, U_{60}$ be i.i.d. Uniform(0,1) and $X = U_1 + U_2 + \cdots + U_{60}$. Find the MGF of X.

4. Let $X \sim \text{Pois}(\lambda)$, and let $M(t)$ be the MGF of X. The *cumulant generating function* is defined to be $g(t) = \ln M(t)$. 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 the jth cumulant of X, for all $j \geq 1$.

Stat 110 Strategic Practice 6 Solutions, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1 Exponential Distribution and Memorylessness

1. Fred (the protagonist of HW 6 #1) wants to sell his car, after moving back to Blissville (where he is happy with the bus system). He decides to sell it to the first person to offer at least \$15,000 for it. Assume that the offers are independent Exponential random variables with mean \$10,000.

   (a) Find the expected number of offers Fred will have.

   The offers on the car are i.i.d. $X_i \sim \text{Expo}(1/10000)$. We are waiting for the first success, where "success" means that an offer is at least \$15,000. So the number of offers that are too low is $\text{Geom}(p)$ with $p = P(X_i \geq 15000) = \exp(-15000/10000) \approx 0.223$. Including the successful offer, the expected number of offers is thus $(1-p)/p + 1 = 1/p \approx 4.48$.

   (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

   $$E(X_N) = E(X|X \geq 15000)$$

   for $X \sim \text{Expo}(1/10000)$, since the successful offer is an Exponential for which we have the information that the value is at least \$15,000. To compute this, remember the memoryless property of the Exponential! For any $a > 0$, if $X \sim \text{Expo}(\lambda)$ then the distribution of $X - a$ given $X > a$ is itself $\text{Expo}(\lambda)$. So

   $$E(X|X \geq 15000) = 15000 + EX = \$25000,$$

   far more than his minimum price!

   Alternatively this can be done by integration (though we prefer the memoryless property), in which case it's crucial to use the conditional PDF (found using Bayes' Rule) rather than just integrating $xf(x)$ from a to ∞ (the conditional distribution needs to be properly normalized): letting $a = 15000$,

   $$f(x|X > a) = \frac{P(X > a|X = x)f(x)}{P(X > a)} = \frac{\lambda e^{-\lambda x}}{e^{-\lambda a}} = \lambda e^{-\lambda(x-a)} \text{ for } x > a,$$

and then

$$E(X|X \geq a) = \int_a^\infty xf(x|X > a)dx$$

also works out to $\$25,000$.

2. Find $E(X^3)$ for $X \sim \text{Expo}(\lambda)$, using LOTUS and the fact that $E(X) = 1/\lambda$ and $\text{Var}(X) = 1/\lambda^2$, and integration by parts at most once (see also Problem 1 in the MGF section).

   By LOTUS, we have:

$$\begin{aligned} E(X^3) &= \int_0^\infty x^3 \lambda e^{-\lambda x} dx = -x^3 e^{-\lambda x}\Big|_0^\infty + \frac{3}{\lambda}\int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= \frac{3}{\lambda}E(X^2) = \frac{3}{\lambda}(Var(X) + (EX)^2) = \frac{6}{\lambda^3}, \end{aligned}$$

   where the second equality uses integration by parts, letting $u = x^3$ and $dv = \lambda e^{-\lambda x}dx$ and we multiply the second term by 1 written as λ/λ.

3. Let $X_1, \ldots, X_n$ be independent, with $X_j \sim \text{Expo}(\lambda_j)$. (They are i.i.d. if all the λ_j's are equal, but we are not assuming that.) Let $M = \min(X_1, \ldots, X_n)$. Show that $M \sim \text{Expo}(\lambda_1 + \cdots + \lambda_n)$, and interpret this intuitively.

   We can find the distribution of M by considering its *survival function* $P(M > t)$, since the survival function is 1 minus the CDF.

$$\begin{aligned} P(M > t) &= P(\min(X_1, \ldots, X_n) > t) = P(X_1 > t, \ldots, X_n > t) \\ &= P(X_1 > t) \ldots P(X_n > t) = e^{-\lambda_1 y} \cdot \cdots \cdot e^{-\lambda_n y} = e^{-(\lambda_1 + \cdots + \lambda_n)y}, \end{aligned}$$

   where the second equality holds since saying that the minimum of the X_j's is greater than t is the same as saying that all of the X_j's are greater than t. Thus, M has the survival function (and the CDF) of an Exponential distribution with parameter $\lambda_1 + \cdots + \lambda_n$. Intuitively, it makes sense that M should have a continuous, memoryless distribution (which implies that it's Exponential) and if we interpret λ_j as rates, it makes sense that M has a combined rate of $\lambda_1 + \cdots + \lambda_n$ since we can imagine, for example, X_1 as the waiting time for a green car to pass by, X_2 as the waiting time for a blue car to pass by, etc., assigning a color to each X_j; then M is the waiting time for a car with any of these colors to pass by.

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 Exponential(λ) distribution.

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

   Alice begins to be served when either Bob or Claire leaves. By the memoryless property, the additional time needed to serve whichever of Bob or Claire is still there is Exponential(λ). The time it takes to serve Alice is also Exponential(λ), so by symmetry the probability is 1/2 that Alice is the last to be done being served.

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

   The expected time spent waiting in line is $\frac{1}{2\lambda}$ since the minimum of two independent Exponentials is Exponential with rate parameter the sum of the two individual rate parameters. The expected time spent being served is $\frac{1}{\lambda}$. So the expected total time is

$$\frac{1}{2\lambda} + \frac{1}{\lambda} = \frac{3}{2\lambda}.$$

2 Moment Generating Functions (MGFs)

1. Find $E(X^3)$ for $X \sim \text{Expo}(\lambda)$ using the MGF of X (see also Problem 2 in the Exponential Distribution section).

   The MGF of an Exponential random variable with rate parameter λ is $M(t) = E(e^{tX}) = (1 - \frac{t}{\lambda})^{-1} = \frac{\lambda}{\lambda - t}$ for $t < \lambda$ (so there is an open interval containing 0 on which $M(t)$ is finite). To get the third moment, we can take the third derivative of the MGF and evaluate at $t = 0$:

$$E(X^3) = \frac{d^3M(t)}{dt^3}\bigg|_{t=0} = \frac{6}{\lambda^3}(1 - \frac{t}{\lambda})^{-4}\bigg|_{t=0} = \frac{6}{\lambda^3}$$

   But a much nicer way to use the MGF here is via pattern recognition: note that $M(t)$ looks like it came from a geometric series:

$$\frac{1}{1 - \frac{t}{\lambda}} = \sum_{n=0}^{\infty} \left(\frac{t}{\lambda}\right)^n = \sum_{n=0}^{\infty} \frac{n!}{\lambda^n} \frac{t^n}{n!},$$

for $|t| < \lambda$. The coefficient of $\frac{t^n}{n!}$ here is the nth moment of X, so we have $E(X^n) = \frac{n!}{\lambda^n}$ for all nonnegative integers n. So again we get $E(X^3) = \frac{6}{\lambda^3}$. This method not only avoids the need to compute the 3rd derivative of $M(t)$ directly, but also it gives us *all* the moments of X in one fell swoop!

2. If X has MGF $M(t)$, what is the MGF of $-X$? What is the MGF of $a + bX$, where a and b are constants?

   The MGF of $-X$ is $E(e^{-tX}) = M(-t)$. The MGF of $a + bX$ is

$$E(e^{t(a+bX)}) = E(e^{at+btX}) = e^{at}E(e^{btX}) = e^{at}M(bt).$$

3. Let $U_1, U_2, \dots, U_{60}$ be i.i.d. Uniform(0,1) and $X = U_1 + U_2 + \cdots + U_{60}$. Find the MGF of X.

   The MGF of U_1 is $E(e^{tU_1}) = \int_0^1 e^{tu}du = \frac{1}{t}(e^t - 1)$ for $t \neq 0$, and the MGF of U_1 is 1 for $t = 0$. Thus, the MGF of X is 1 for $t = 0$, and for $t \neq 0$ it is

$$E(e^{tX}) = E(e^{t(U_1+\cdots+U_{60})}) = \left(E(e^{tU_1})\right)^{60} = \frac{(e^t - 1)^{60}}{t^{60}}.$$

4. Let $X \sim \text{Pois}(\lambda)$, and let $M(t)$ be the MGF of X. The *cumulant generating function* is defined to be $g(t) = \ln M(t)$. 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 the jth cumulant of X, for all $j \geq 1$.

   Using the Taylor series for e^t,

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

   so $c_j = \lambda$ for all $j \geq 1$.

Stat 110 Homework 6, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1. Fred lives in Blissville, where buses always arrive exactly on time, with the time between successive buses fixed at 10 minutes. Having lost his watch, he arrives at the bus stop at a random time (assume that buses run 24 hours a day, and that the time that Fred arrives is uniformly random on a particular day).

(a) What is the distribution of how long Fred has to wait for the next bus? What is the average time that Fred has to wait?

(b) Given that the bus has not yet arrived after 6 minutes, what is the probability that Fred will have to wait at least 3 more minutes?

(c) Fred moves to Blotchville, a city with inferior urban planning and where buses are much more erratic. Now, when any bus arrives, the time until the next bus arrives is an Exponential random variable with mean 10 minutes. Fred arrives at the bus stop at a random time, not knowing how long ago the previous bus came. What is the distribution of Fred's waiting time for the next bus? What is the average time that Fred has to wait? (Hint: don't forget the memoryless property.)

(d) When Fred complains to a friend how much worse transportation is in Blotchville, the friend says: "Stop whining so much! You arrive at a uniform instant between the previous bus arrival and the next bus arrival. The average length of that interval between buses is 10 minutes, but since you are equally likely to arrive at any time in that interval, your average waiting time is only 5 minutes."

Fred disagrees, both from experience and from solving Part (c) while waiting for the bus. Explain what (if anything) is wrong with the friend's reasoning.

2. Three Stat 110 students are working independently on this pset. All 3 start at 1 pm on a certain day, and each takes an Exponential time with mean 6 hours to complete this pset. What is the earliest time when all 3 students will have completed this pset, on average? (That is, *all* of the 3 students need to be done with this pset.)

3. Consider an experiment where we observe the value of a random variable X, and estimate the value of an unknown constant θ using some random variable $T = g(X)$ that is a function of X. The r.v. T is called an *estimator*. Think of X as the data observed in the experiment, and θ as an unknown parameter related to the distribution of X.

For example, consider the experiment of flipping a coin n times, where the coin has an unknown probability θ of Heads. After the experiment is performed, we have observed the value of $X \sim \text{Bin}(n, \theta)$. The most natural estimator for θ is then X/n.

(a) The *bias* of an estimator T for θ is defined as $b(T) = E(T) - \theta$. The *mean squared error* is the average squared error when using $T(X)$ to estimate θ:

$$\text{MSE}(T) = E(T - \theta)^2.$$

Show that

$$\text{MSE}(T) = \text{Var}(T) + (b(T))^2 .$$

This implies that for fixed MSE, lower bias can only be attained at the cost of higher variance and vice versa; this is a form of the *bias-variance tradeoff*, a phenomenon which arises throughout statistics.

(b) Show without using calculus that the constant c that minimizes $E(X - c)^2$ is the expected value of X. (This means that in choosing a single number to summarize X, the mean is the best choice if the goal is to minimize the average squared error.)

(c) For the case that X is continuous with PDF $f(x)$ which is positive everywhere, show that the value of c that minimizes $E|X - c|$ is the median of X (which is the value m with $P(X \leq m) = 1/2$.

Hint: this can be done either with or without calculus. For the calculus method, use LOTUS to write $E|X - c|$ as an integral, and then split the integral into 2 pieces to get rid of the absolute values. Then use the fundamental theorem of calculus (after writing, for example, $\int_{-\infty}^{c}(c - x)f(x)dx = c\int_{-\infty}^{c} f(x)dx - \int_{-\infty}^{c} xf(x)dx$).

4. (a) Suppose that we have a list of the populations of every country in the world.

Guess, without looking at data yet, what percentage of the populations have the digit 1 as their first digit (e.g., a country with a population of 1,234,567 has first digit 1 and a country with population 89,012,345 does not).

Note: (a) is a rare problem where the only way to lose points is to find out the right answer rather than guessing!

(b) After having done (a), look through a list of populations such as

`http://en.wikipedia.org/wiki/List_of_countries_by_population`

and count how many start with a 1. What percentage of countries is this?

(c) *Benford's Law* states that in a very large variety of real-life data sets, the first digit approximately follows a particular distribution with about a 30% chance of a 1, an 18% chance of a 2, and in general

$$P(D = j) = \log_{10}\left(\frac{j+1}{j}\right), \text{ for } j \in \{1, 2, 3, \dots, 9\},$$

where D is the first digit of a randomly chosen element. Check that this is a PMF (using properties of logs, not with a calculator).

(d) Suppose that we write the random value in some problem (e.g., the population of a random country) in scientific notation as $X \times 10^N$, where N is a nonnegative integer and $1 \leq X < 10$. Assume that X is a continuous r.v. with PDF

$$f(x) = c/x, \text{ for } 1 \leq x \leq 10$$

(and 0 otherwise), with c a constant. What is the value of c (be careful with the bases of logs)? Intuitively, we might hope that the distribution of X does not depend on the choice of units in which X is measured. To see whether this holds, let $Y = aX$ with $a > 0$. What is the PDF of Y (specifying where it is nonzero)?

(e) Show that if we have a random number $X \times 10^N$ (written in scientific notation) and X has the PDF $f(x)$ from (d), then the first digit (which is also the first digit of X) has the PMF given in (c).

Hint: what does $D = j$ correspond to in terms of the values of X?

5. Customers arrive at the Leftorium store according to a Poisson process with rate λ customers per hour. The true value of λ is unknown, so we treat it as a random variable (this is called a *Bayesian* approach). Suppose that our prior beliefs about λ can be expressed as $\lambda \sim \text{Expo}(3)$. Let X be the number of customers who arrive at the Leftorium between 1 pm and 3 pm tomorrow. Given that $X = 2$ is observed, find the conditional PDF of λ (this is known as the *posterior density* of λ).

6. Let $X_n \sim \text{Bin}(n, p_n)$ for all $n \geq 1$, where np_n is a constant $\lambda > 0$ for all n (so $p_n = \lambda/n$). Let $X \sim \text{Pois}(\lambda)$. Show that the MGF of X_n converges to the MGF of X (this gives another way to see that the $\text{Bin}(n, p)$ distribution can be well-approximated by the $\text{Pois}(\lambda)$ when n is large, p is small, and $\lambda = np$ is moderate).

7. Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = e^X$. Then Y has a *Log-Normal distribution* (which means "log is Normal"; note that "log of a Normal" doesn't make sense since Normals can be negative).

Find the mean and variance of Y using the MGF of X, *without doing any integrals.* Then for $\mu = 0, \sigma = 1$, find the nth moment $E(Y^n)$ (in terms of n).

Stat 110 Homework 6 Solutions, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1. Fred lives in Blissville, where buses always arrive exactly on time, with the time between successive buses fixed at 10 minutes. Having lost his watch, he arrives at the bus stop at a random time (assume that buses run 24 hours a day, and that the time that Fred arrives is uniformly random on a particular day).

(a) What is the distribution of how long Fred has to wait for the next bus? What is the average time that Fred has to wait?

The distribution is Uniform on $[0, 10]$, so the mean is 5 minutes.

(b) Given that the bus has not yet arrived after 6 minutes, what is the probability that Fred will have to wait at least 3 more minutes?

Let T be the waiting time. Then

$$P(T \geq 6+3|T>6) = \frac{P(T \geq 9, T>6)}{P(T>6)} = \frac{P(T \geq 9)}{P(T>6)} = \frac{1/10}{4/10} = \frac{1}{4}.$$

(c) Fred moves to Blotchville, a city with inferior urban planning and where buses are much more erratic. Now, when any bus arrives, the time until the next bus arrives is an Exponential random variable with mean 10 minutes. Fred arrives at the bus stop at a random time, not knowing how long ago the previous bus came. What is the distribution of Fred's waiting time for the next bus? What is the average time that Fred has to wait? (Hint: don't forget the memoryless property.)

By the memoryless property, the distribution is Exponential with parameter $1/10$ (and mean 10 minutes) regardless of when Fred arrives (how much longer the next bus will take to arrive is independent of how long ago the previous bus arrived). The average time that Fred has to wait is 10 minutes.

(d) When Fred complains to a friend how much worse transportation is in Blotchville, the friend says: "Stop whining so much! You arrive at a uniform instant between the previous bus arrival and the next bus arrival. The average length of that interval between buses is 10 minutes, but since you are equally likely to arrive at any time in that interval, your average waiting time is only 5 minutes."

Fred disagrees, both from experience and from solving Part (c) while waiting for the bus. Explain what (if anything) is wrong with the friends reasoning.

The average length of a time interval between 2 buses is 10 minutes, but this does not imply that Fred's average waiting time is 5 minutes. This is because Fred is not equally likely to arrive at any of these intervals: Fred is more likely to arrive during a long interval between buses than to arrive during a short interval between buses. For example, if one interval between buses is 50 minutes and another interval is 5 minutes, then Fred is 10 times more likely to arrive during the 50 minute interval.

This phenomenon is known as *length-biasing*, and it comes up in many real-life situations. For example, asking randomly chosen mothers how many children they have yields a different distribution from asking randomly chosen people how many siblings they have, including themselves. Asking students the sizes of their classes and averaging those results may give a much higher value than taking a list of classes and averaging the sizes of each (this is called the *class size paradox*).

2. Three Stat 110 students are working independently on this pset. All 3 start at 1 pm on a certain day, and each takes an Exponential time with mean 6 hours to complete this pset. What is the earliest time when all 3 students will have completed this pset, on average? (That is, *all* of the 3 students need to be done with this pset.)

Label the students as $1, 2, 3$, and let X_j be how long it takes student j to finish the pset. Let T be the time it takes for all 3 students to complete the pset, so $T = T_1 + T_2 + T_3$ where $T_1 = \min(X_1, X_2, X_3)$ is how long it takes for one student to complete the pset, T_2 is the additional time it takes for a second student to complete the pset, and T_3 is the additional time until all 3 have completed the pset. Then $T_1 \sim \text{Expo}(\frac{3}{6})$ since, as shown on Strategic Practice 6, the minimum of independent Exponentials is Exponential with rate the sum of the rates. By the memoryless property, at the first time when a student completes the pset the other two students are "starting from fresh," so $T_2 \sim \text{Expo}(\frac{2}{6})$. Again by the memoryless property, $T_3 \sim \text{Expo}(\frac{1}{6})$. Thus, $E(T) = 2 + 3 + 6 = 11$, which shows that on average, the 3 students will have all completed the pset at midnight, 11 hours after they started.

3. Consider an experiment where we observe the value of a random variable X, and estimate the value of an unknown constant θ using some random variable $T = g(X)$ that is a function of X. The r.v. T is called an *estimator*. Think of X as the data observed in the experiment, and θ as an unknown parameter related to the distribution of X.

For example, consider the experiment of flipping a coin n times, where the coin has an unknown probability θ of Heads. After the experiment is performed, we have observed the value of $X \sim \text{Bin}(n, \theta)$. The most natural estimator for θ is then X/n.

(a) The *bias* of an estimator T for θ is defined as $b(T) = E(T) - \theta$. The *mean squared error* is the average squared error when using $T(X)$ to estimate θ:

$$\text{MSE}(T) = E(T - \theta)^2.$$

Show that

$$\text{MSE}(T) = \text{Var}(T) + (b(T))^2 .$$

This implies that for fixed MSE, lower bias can only be attained at the cost of higher variance and vice versa; this is a form of the *bias-variance tradeoff*, a phenomenon which arises throughout statistics.

Using the fact that adding a constant does not affect variance, we have

$$\begin{aligned}\text{Var}(T) &= \text{Var}(T-\theta)\\ &= E(T-\theta)^2 - (E(T-\theta))^2\\ &= \text{MSE}(T) - (b(T))^2,\end{aligned}$$

which proves the desired identity.

(b) Show without using calculus that the constant c that minimizes $E(X-c)^2$ is the expected value of X. (This means that in choosing a single number to summarize X, the mean is the best choice if the goal is to minimize the average squared error.)

Applying (a) (with c in place of θ and $T(X) = X$),

$$E(X-c)^2 = \text{Var}(X) + (E(X)-c)^2 \geq \text{Var}(X)$$

for all c. Equality holds if and only if $E(X) - c = 0$, which is equivalent to $c = EX$. Thus, the unique value of c that minimizes $E(X-c)^2$ is EX. Note that for $c = EX$, the value of $E(X-c)^2$ is $\text{Var}(X)$.

(c) For the case that X is continuous with PDF $f(x)$ which is positive everywhere, show that the value of c that minimizes $E|X-c|$ is the median of X (which is the value m with $P(X \leq m) = 1/2$.

Hint: this can be done either with or without calculus. For the calculus method, use LOTUS to write $E|X-c|$ as an integral, and then split the integral into 2 pieces to get rid of the absolute values. Then use the fundamental theorem of calculus (after writing, for example, $\int_{-\infty}^{c}(c-x)f(x)dx = c\int_{-\infty}^{c} f(x)dx - \int_{-\infty}^{c} xf(x)dx$).

Proof with calculus: We want to minimize

$$E|X-c| = \int_{-\infty}^{\infty} |x-c|f(x)dx = \int_{-\infty}^{c} (c-x)f(x)dx + \int_{c}^{\infty} (x-c)f(x)dx,$$

where we split the integral into 2 pieces to handle the absolute values. This becomes

$$E|X-c| = c\int_{-\infty}^{c} f(x)dx - \int_{-\infty}^{c} xf(x)dx + \int_{c}^{\infty} xf(x)dx - c\int_{c}^{\infty} f(x)dx.$$

Now differentiate both sides with respect to c, using the fundamental theorem of calculus:

$$\frac{d}{dc}(E|X-c|) = \int_{-\infty}^{c} f(x)dx + cf(c) - cf(c) - cf(c) - \int_{c}^{\infty} f(x)dx + cf(c),$$

which simplifies to $P(X \leq c) - (1 - P(X \leq c)) = 2P(X \leq c) - 1$. This is 0 when $P(X \leq c) = 1/2$. The second derivative is $f(c) + f(c) = 2f(c) > 0$, so we have found a minimum. Thus, $E|X-c|$ is minimized when c is the median.

Proof without calculus: We will show a more general result, not assuming that X is a continuous r.v. A number m is called a *median* of the distribution of X if $P(X \leq m) \geq 1/2, P(X \geq m) \geq 1/2$. (So a median may not be unique; it will be unique if the CDF is continuous and strictly increasing.) Let m be a median, and let $a \neq m$. We need to show $E|X-m| \leq E|X-a|$, which is equivalent to $E(|X-a| - |X-m|) \geq 0$. Assume $m < a$ (the case $m > a$ can be handled by the same method).

Note that $|X-a| - |X-m| = a - X - (m - X) = a - m$ if $X \leq m$, and $|X-a| - |X-m| \geq X - a - (X - m) = m - a$ if $X > m$. Splitting the definition of expected value into 2 parts based on whether $X > m$ occurs, we have

$$E(|X-a| - |X-m|) \geq (a-m)P(X \leq m) + (m-a)P(X > m),$$

which simplifies to $(a-m)(P(X \leq m) - P(X > m))$. By definition of median,

$$P(X \leq m) - P(X > m) = P(X \leq m) - (1 - P(X \leq m)) = 2P(X \leq m) - 1 \geq 0,$$

which shows that $E(|X-m|) \leq E(|X-a|)$.

4. (a) Suppose that we have a list of the populations of every country in the world.

Guess, without looking at data yet, what percentage of the populations have the digit 1 as their first digit (e.g., a country with a population of 1,234,567 has first digit 1 and a country with population 89,012,345 does not).

Note: (a) is a rare problem where the only way to lose points is to find out the right answer rather than guessing!

What did you guess for this part?

(b) After having done (a), look through a list of populations such as

`http://en.wikipedia.org/wiki/List_of_countries_by_population` and count how many start with a 1. What percentage of countries is this?

According to Wikipedia (as of October 15, 2011), 63 out of 225 countries have a total population whose first digit is 1, which is 28%. (This depends slightly on whether certain territories included in the Wikipedia list should be considered as "countries" but the purpose of this problem is not to delve into the sovereignty or nation-status of territories). It is striking that 28% of the countries have first digit 1, as this is so much higher than one would expect from guessing that the first digit is equally likely to be any of $1, 2, \ldots, 9$. This phenomenon is known as *Benford's Law* and a distribution similar to the one derived below has been observed in many diverse settings (such as lengths of rivers, physical constants, stock prices).

(c) *Benford's Law* states that in a very large variety of real-life data sets, the first digit approximately follows a particular distribution with about a 30% chance of a 1, an 18% chance of a 2, and in general

$$P(D = j) = \log_{10}\left(\frac{j+1}{j}\right), \text{ for } j \in \{1, 2, 3, \ldots, 9\},$$

where D is the first digit of a randomly chosen element. Check that this is a PMF (using properties of logs, not with a calculator).

The function $P(D = j)$ is nonnegative and the sum over all values is

$$\sum_{j=1}^{9} \log_{10} \frac{j+1}{j} = \sum_{j=1}^{9} (\log_{10}(j+1) - \log_{10}(j)).$$

All terms cancel except $\log_{10} 10 - \log_{10} 1 = 1$ (this is a *telescoping series*). Since the values add to 1 and are nonnegative, $P(D = j)$ is a PMF.

(d) Suppose that we write the random value in some problem (e.g., the population of a random country) in scientific notation as $X \times 10^N$, where N is a nonnegative integer and $1 \leq X < 10$. Assume that X is a continuous r.v. with PDF

$$f(x) = c/x, \text{ for } 1 \leq x \leq 10$$

(and 0 otherwise), with c a constant. What is the value of c (be careful with the bases of logs)? Intuitively, we might hope that the distribution of X does not depend on the choice of units in which X is measured. To see whether this holds, let $Y = aX$ with $a > 0$. What is the PDF of Y (specifying where it is nonzero)?

The PDF ($f(x) = c/x, 1 \leq x \leq 10$) must integrate to one, by definition; therefore

$$1 = c \int_{1}^{10} dx/x = c(\ln 10 - \ln 1) = c \ln 10.$$

So the constant of proportionality $c = 1/\ln 10 = \log_{10} e$). If $Y = aX$ (a *change in scale*), then Y has pdf c/y with the same value of c as before, except now $a \leq y \leq 10a$ rather than $1 \leq x \leq 10$. So the PDF takes the same form for aX as for X, but over a different range.

(e) Show that if we have a random number $X \times 10^N$ (written in scientific notation) and X has the PDF $f(x)$ from (d), then the first digit (which is also the first digit of X) has the PMF given in (c).

Hint: what does $D = j$ correspond to in terms of the values of X?

The first digit $D = d$ when $d \leq X < d + 1$. The probability of this is

$$P(D = d) = P(d \leq X < d + 1) = \int_d^{d+1} \frac{1}{x \ln 10} dx,$$

which is then $\log_{10}(d + 1) - \log_{10}(d)$, identical to our earlier PMF.

5. Customers arrive at the Leftorium store according to a Poisson process with rate λ customers per hour. The true value of λ is unknown, so we treat it as a random variable (this is called a *Bayesian* approach). Suppose that our prior beliefs about λ can be expressed as $\lambda \sim \text{Expo}(3)$. Let X be the number of customers who arrive at the Leftorium between 1 pm and 3 pm tomorrow. Given that $X = 2$ is observed, find the conditional PDF of λ (this is known as the *posterior density* of λ).

Let us write Λ (capital λ) for the r.v. and λ for a specific possible value of Λ (in practice, often both would be written as λ, but to make sure the distinction between a r.v. and its values is clear we will maintain the notational distinction here). We need to find the conditional PDF of λ given the evidence, which we write as $f_{\Lambda|X}(\lambda|2)$. By Bayes' Rule, the posterior density of λ is

$$f_{\Lambda|X}(\lambda|x) = \frac{P(X = x|\Lambda = \lambda) f_\Lambda(\lambda)}{P(X = x)},$$

where $f_\Lambda(\lambda) = 3e^{-3\lambda}$ for $\lambda > 0$, and $P(X = x|\Lambda = \lambda)$ is found using the $\text{Pois}(2\lambda)$ PMF. For $x = 2$, the numerator is

$$\frac{e^{-2\lambda}(2\lambda)^2}{2!} \cdot 3e^{-3\lambda} = 6\lambda^2 e^{-5\lambda}.$$

For the denominator when $x = 2$, note that it is a constant (not depending on λ), so it must be the constant that makes the conditional PDF integrate to 1. Equivalently, we can use the continuous version of the Law of Total Probability:

$$P(X = 2) = \int_0^\infty P(X = 2|\Lambda = \lambda) f_\Lambda(\lambda) d\lambda.$$

The conditional PDF is proportional to $\lambda^2 e^{-5\lambda}$, so we just need to find $\int_0^\infty \lambda^2 e^{-5\lambda} d\lambda$. This can be done using integration by parts, but a neater way is to recognize that

$$5\int_0^\infty y^2 e^{-5y} dy = \frac{2!}{5^2} = \frac{2}{25},$$

as this is the integral we'd get using LOTUS to find $E(Y^2)$ for $Y \sim \text{Expo}(5)$. Thus,

$$f_{\Lambda|X}(\lambda|2) = \frac{125}{2}\lambda^2 e^{-5\lambda},$$

for $\lambda > 0$. (This is known as the Gamma(3,5) distribution.)

6. Let $X_n \sim \text{Bin}(n, p_n)$ for all $n \geq 1$, where np_n is a constant $\lambda > 0$ for all n (so $p_n = \lambda/n$). Let $X \sim \text{Pois}(\lambda)$. Show that the MGF of X_n converges to the MGF of X (this gives another way to see that the $\text{Bin}(n, p)$ distribution can be well-approximated by the $\text{Pois}(\lambda)$ when n is large, p is small, and $\lambda = np$ is moderate).

Using the fact that $(1 + x/n)^n \to e^x$ as $n \to \infty$ (which was given in class when the Poisson was first introduced, and is also easy to obtain by taking logs and then using L'Hôpital's Rule), we have

$$E(e^{tX_n}) = (1 - p_n + p_n e^t)^n = (1 + \lambda(e^t - 1)/n)^n \quad \to \quad e^{\lambda(e^t-1)} = E(e^{tX}).$$

7. Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = e^X$. Then Y has a *Log-Normal distribution* (which means "log is Normal"; note that "log of a Normal" doesn't make sense since Normals can be negative).

Find the mean and variance of Y using the MGF of X, *without doing any integrals.* Then for $\mu = 0, \sigma = 1$, find the nth moment $E(Y^n)$ (in terms of n).

The mean and variance of Y can be computed as

$$\begin{aligned} E(Y) = E(e^X) = &\quad e^{\mu+\sigma^2/2} \\ E(Y^2) = E(e^{2X}) = &\quad e^{2\mu+2\sigma^2} \\ \text{Var}(Y) = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2} = &\quad e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right). \end{aligned}$$

The nth moment of Y when $\mu = 0$ and $\sigma = 1$ is

$$E(Y^n) = E(e^{nX}) \quad = \quad e^{n^2/2}.$$