Prof. David Draper
Department of
Applied Mathematics and Statistics
University of California, Santa Cruz

Name: Jeremy Lafond

AMS 131: Take-Home Test 2

Target due date: Wed 18 Jul 2018 *[520 total points]*

1. *[70 total points]* (the Exchange Paradox) You're playing the following game against an opponent, with a referee also taking part. The referee has two envelopes (numbered 1 and 2 for the sake of this problem, but when the game is played the envelopes have no markings on them), and (without you or your opponent seeing what she does) she puts $\$m$ in envelope 1 and $\$2\,m$ in envelope 2 for some $m > 0$ (treat m as continuous in this problem even though in practice it would have to be rounded to the nearest dollar or penny). You and your opponent each get one of the envelopes at random. You open your envelope secretly and find $\$x$ (your opponent also looks secretly in his envelope), and the referee then asks you if you want to trade envelopes with your opponent. You reason that if you trade, you will get either $\$\frac{x}{2}$ or $\$2\,x$, each with probability $\frac{1}{2}$. This makes the expected value of the amount of money you'll get if you trade equal to $\left(\frac{1}{2}\right)\left(\frac{\$x}{2}\right) + \left(\frac{1}{2}\right)(\$2\,x) = \frac{\$5x}{4}$, which is greater than the $\$x$ you currently have, so you offer to trade. The paradox is that your opponent is capable of making exactly the same calculation. How can the trade be advantageous for both of you?

The point of this problem is to demonstrate that the above reasoning is flawed from a Bayesian point of view; the conclusion that trading envelopes is always optimal is based on the assumption that there's no information obtained by observing the contents of the envelope you get, and this assumption can be seen to be false when you reason in a Bayesian way. At a moment in time before the game begins, let $p(m)$ be your prior distribution on the amount of money M the referee will put in envelope 1, and let X be the amount of money you'll find in your envelope when you open it (when the game is actually played, the observed x, of course, will be data that can be used to decrease your uncertainty about M).

(a) Explain why the setup of this problem implies that $P(X = m|M = m) = P(X = 2\,m|M = m) = \frac{1}{2}$, and use this to show that

$$P(M = x|X = x) = \frac{p(x)}{p(x) + p\left(\frac{x}{2}\right)} \quad \text{and} \quad P\Big(M = \frac{x}{2}\Big|\, X = x\Big) = \frac{p\left(\frac{x}{2}\right)}{p(x) + p\left(\frac{x}{2}\right)}\,. \tag{1}$$

Demonstrate from this that the expected value of the amount Y of money in your opponent's envelope, given than you've found $\$x$ in the envelope you've opened, is

$$E(Y|X = x) = \frac{p(x)}{p(x) + p\left(\frac{x}{2}\right)}(2\,x) + \frac{p\left(\frac{x}{2}\right)}{p(x) + p\left(\frac{x}{2}\right)}\left(\frac{x}{2}\right)\,. \tag{2}$$

[20 points]

$P(X = m|M = m) = P(X = 2\,m|M = m) = \frac{1}{2}$ because X must equal M or $2M$. When observing $X = x$, M can only take on one of two values(x or $\frac{x}{2}$). Therefore, there is a 50% chance of either

value being in the envelope; however, this must be after information about x has been observed and thus the original logic did NOT use Bayesian reasoning in the first place.

Applying Bayes theorem to each case we are left with:

$$Pr(M = x|X = x) = \frac{Pr(X = x|M = x)p(x)}{Pr(X = x|M = x)p(x) + Pr(X = x|M = \frac{x}{2})p(\frac{x}{2})} = \frac{p(x)}{p(x) + p(\frac{x}{2})}$$

and continuing this approach:

$$Pr(M = \frac{x}{2}|X = x) = \frac{Pr(X = x|M = \frac{x}{2})p(\frac{x}{2})}{Pr(X = x|M = \frac{x}{2})p(\frac{x}{2}) + Pr(X = x|M = x)p(x)} = \frac{p(\frac{x}{2})}{p(x) + p(\frac{x}{2})}$$

Now that the above is represented, we can see that if I keep the envelope I have, I win x dollars. If I trade the envelope, I win $\frac{x}{2}$ dollars, only if I have the envelope with $2M$ dollars. etc etc.

Obviously, the expected value of Y in my opponents envelope based on me finding $\$x$ in my envelope is the combined above odds of either event occurring for me since there is a 50% chance that either occurs in the first place. However we must still must consider the value of $X = x$ for each case, based on Bayesian analysis to avoid the original paradoxical conclusion.

Therefore, we are left with:

$$E(Y|X = x) = (Pr(M = x|X = x))(\text{value of x}|X = 2x) + (Pr(M = \frac{x}{2}|X = x))(\text{value of x}|X = \frac{x}{2})$$

Since $X = x$:

$$= \frac{p(x)}{p(x) + p(\frac{x}{2})}(2x) + \frac{p(\frac{x}{2})}{p(x) + p(\frac{x}{2})}(\frac{x}{2})$$

(b) Suppose that for you in this game, money and utility coincide (or at least suppose that utility is linear in money for you with a positive slope). Use Bayesian decision theory, through the principle of maximizing expected utility, to show that you should offer to trade envelopes only if

$$p\Big(\frac{x}{2}\Big) < 2\,p(x)\,. \tag{3}$$

Using the above conclusion in (a), considering that it also applies to my own odds, we see that if $p(x) = 2p(x)$.

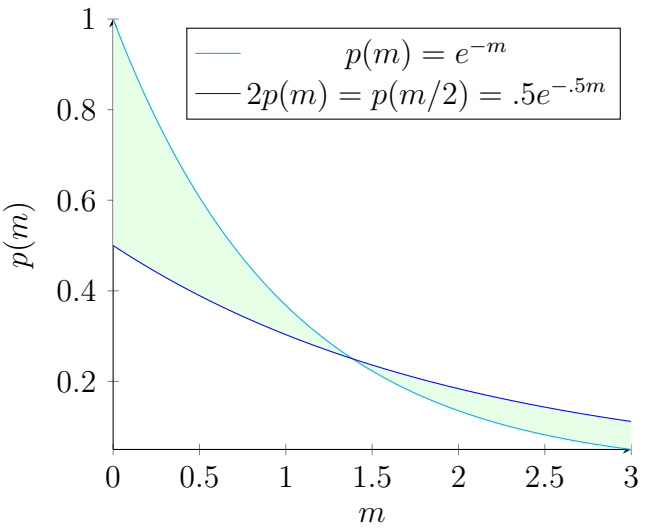
$$E(Y|X = x) = x$$

Therefore:

If $p(\frac{x}{2}) > 2p(x)$, I should keep the envelope because I stand to win less than x.

If $p(\frac{x}{2}) < 2(p(x)$, I should trade the envelope because I stand to gain more than x.

If you and two friends (one of whom would serve as the referee) were to actually play this game with real money in the envelopes, it would probably be the case that small amounts of money are more likely to be chosen by the referee than big amounts, which makes it interesting to explore condition (3) for prior distributions that are decreasing (that is, for which $p(m_2) < p(m_1)$ for $m_2 > m_1$). Make a sketch of what condition (3) implies for a decreasing p. One possible example of a continuous decreasing family of priors on M is the *exponential* distribution indexed by the parameter λ, which represents the reciprocal of the mean of the distribution. Identify the set of conditions in this family of priors, as a function of x and λ, under which it's optimal for you to trade. Does the inequality you obtain in this way make good intuitive sense (in terms of both x and λ)? Explain briefly. *[40 points]*



This sketch is a rough estimate of what is implied by condition (3) for decreasing $p(m)$ with an arbitrary decreasing exponential distribution. The implication being that the shaded region on the left is optimal for trading envelopes and the shaded region on the right is optimal for keeping envelopes.

The above sketch is an example of a single exponential distribution where $\lambda = .5$.

However, the same conditions apply for any λ such that:

If

$$p(x, \lambda) = \lambda e^{-\lambda x}$$

then the below equation still holds for optimal values to trade envelopes:

$$p(\frac{x}{2}, \lambda) < p(x, \lambda)$$

As far as making sense goes, I'm not sure it feels "intuitive"; however, it makes sense based on the previous logic as any relatively large amount of money observed in my envelope, especially when considering the referee is bias towards small amounts, is likely to be my best option and I should not trade. Conversely, any relatively small amount in my envelope encourages me to seek a trade. But again, all of this is not very intuitive just because it seems in the end that we are still dealing with dollar amounts that cannot be "relatively" quantified in the first place.

Upon deeper inspection, the functions always intersect at:

$$\frac{ln(2)}{\lambda}$$

This particular fact does provide some intuition as this marks the **median** of our distribution and some probabilistic midway point would naturally serve as the point at which one would start considering trading based on the fact that our problem was fairly symmetric in the first place.

(c) Looking carefully at the correct argument in paragraph 2 of this problem, identify precisely the point at which the argument in the first paragraph breaks down, and specify what someone who believes the argument in paragraph 1 is implicitly assuming about the prior distribution $p(m)$. *[10 points]*

The assumption seems to be made that x is the same in both cases when in reality, the 5/4 line of logic essentially ignores any information that would have been gained by looking in the envelope. In this example, improper or non-informative priors are used to inform us to make a decision about m without actually basing it off of any observations. Essentially this person must believe that $\$m$ and $\$2m$ and naturally $\$\frac{x}{2}$ and $\$2x$ are using the same m and x values respectively and thus from the same probability distribution. However, it is easy to see that one probability density must be exactly half of the other as it has twice the range given the same input.

2. *[210 total points]* (practice with joint, marginal and conditional densities) This is a toy problem designed to give you practice in working with a number of the concepts we've examined; in a course like this, every now and then you have to stop looking at real-world problems and just work on technique (it's similar to classical musicians needing to practice scales in addition to actual pieces of symphonic or chamber music).

Suppose that the continuous random vector $\boldsymbol{X} = (X_1, X_2)$ has PDF given by

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \left\{ \begin{array}{cc} 4\,x_1\,x_2 & \text{for } 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 & \text{otherwise} \end{array} \right\} \tag{4}$$

in which $\boldsymbol{x} = (x_1, x_2)$, and define the random vector $\boldsymbol{Y} = (Y_1, Y_2)$ with the transformation $(Y_1 = X_1, Y_2 = X_1\,X_2)$.

(a) Are X_1 and X_2 independent? Present any relevant calculations to support your answer. *[10 points]*

In order for X_1 and X_2 to be independent the following equation must hold true:

$$f_{\underline{X_1}}(X_1) * f_{\underline{X_2}}(X_2) = f_x(x) = f_{\underline{X_1,X_2}}(X_1, X_2)$$

The marginals are:

$$f_{\underline{X_1}}(X_1) = \int_S f_{\underline{x_1,x_2}}(x_1, x_2)\mathrm{d}x_1 = \int_0^1 4x_1x_2\mathrm{d}x_1 = 2x_2$$

and

$$f_{\underline{X_2}}(X_2) = \int_S f_{\underline{x_1,x_2}}(x_1, x_2)\mathrm{d}x_2 = \int_0^1 4x_1x_2\mathrm{d}x_2 = 2x_1$$

$$f_{\underline{x_1}}(x_1) * f_{\underline{x_2}}(x_2) = 2x_1 * 2x_2 = 4x_1x_2$$

However,

$$f_x(x) = f_{\underline{X_1,X_2}}(X_1, X_2) = 4x_1x_2$$

Thus,

$$f_{\underline{X_1,X_2}}(X_1, X_2) = f_{\underline{x_1}}(x_1) * f_{\underline{x_2}}(x_2)$$

Therefore X_1 and X_2 are independent.

(b) Either work out the correlation $\rho(X_1, X_2)$ between X_1 and X_2 or explain why no calculation is necessary in correctly identifying the value of ρ. *[10 points]*

Recall:

If X_1 and X_2 are independent and their values are $0 < X_n < \infty$ then:

$$Cov(X_1, X_2) = \rho(X_1, X_2) = 0$$

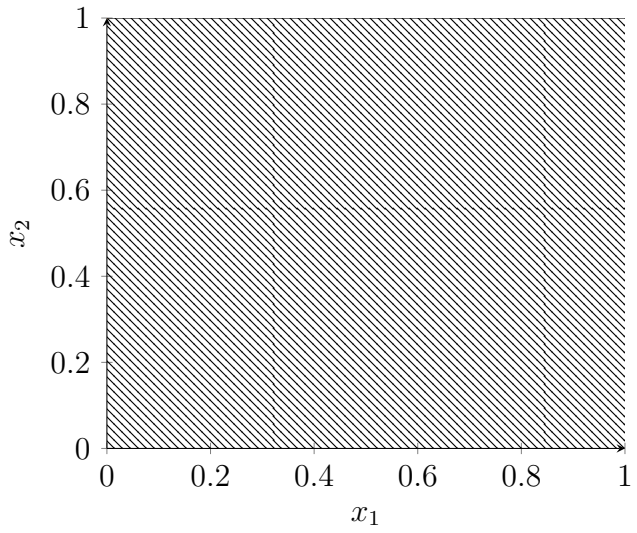
We found in part a that X_1 and X_2 are independent and we know that they range from 0 to 1.

Therefore, $\rho(X_1, X_2) = 0$

(c) Sketch the set S of possible $\boldsymbol{X}$ values and the image T of S under the transformation from $\boldsymbol{X}$ to $\boldsymbol{Y}$, and show that the joint distribution of $\boldsymbol{Y} = (Y_1, Y_2)$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \left\{ \begin{array}{cc} 4\,\frac{y_2}{y_1} & \text{for } 0 < y_1 < 1, 0 < y_2 < y_1 < 1 \\ 0 & \text{otherwise} \end{array} \right\}, \tag{5}$$

in which $\boldsymbol{y} = (y_1, y_2)$. Verify your calculation by demonstrating that $\iint_T f_{\boldsymbol{Y}}(\boldsymbol{y})\, d\boldsymbol{y} = 1$. *[50 points]*



The is the support set S of possible X values. It's essentially the unit square not including the edges.

Recall that:

$$f_y Y = f_x[h_1(y)h_2(y)] \det J$$

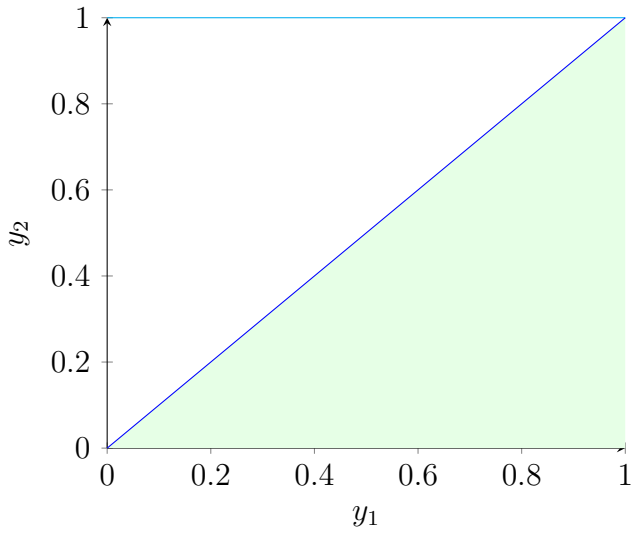
Where $\det J$ is the determinant of the Jacobian matrix of partials of each inverse.

h1 and h2 are the inverse functions of Y_1 and Y_2 respectively.

given our transformations we can solve for X_1 and X_2 respectively:
$Y_1 = X_1 \implies h_1(y) = Y_1$

$Y_2 = X_1X_2 \implies X_2 = \frac{Y_2}{X_1} \implies X_2 = \frac{Y_2}{(Y_1)} \implies h_2(y) = \frac{Y_2}{Y_1}$

We can find the set T of points using the equations above and the inequalities of the support set S. Using substitution we are left with:



This is the support set T of possible Y values of the image T of S under the transformation from X to Y. It is bounded by the line $y_1 = y_2$ and the line $y_1 = 1$.

we can now construct a Jacobian of partials:

$$J = \begin{bmatrix} \frac{\partial_{h_1}}{\partial_{Y_1}} & \frac{\partial_{h_1}}{\partial_{Y_2}} \\ \frac{\partial_{h_2}}{\partial_{Y_1}} & \frac{\partial_{h_2}}{\partial_{Y_2}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{Y_2}{Y_1} & \frac{1}{Y_1} \end{bmatrix} \implies \det J = (1)(\frac{1}{Y_1}) - (0)(-\frac{Y_2}{Y_1^2}) = \frac{1}{Y_1}$$

Extending this logic, we are left with:

$$f_y Y = f_x[h_1(y)h_2(y)]det|J| = 4[h_1(y)h_2(y)] \det J = 4[(Y_1)(\frac{Y_2}{Y_1})]\frac{1}{Y_1} = 4(\frac{Y_2}{Y_1})$$

and based on our image of T of S we find the below equation to be true:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \left\{ \begin{array}{cc} 4\frac{y_2}{y_1} & \text{for } 0 < y_1 < 1, 0 < y_2 < y_1 < 1 \\ 0 & \text{otherwise} \end{array} \right\} \tag{6}$$

as was to be shown.

Below is the verification that:

$\iint_T f_{\boldsymbol{Y}}(\boldsymbol{y})\, d\boldsymbol{y} = 1$

$$\int \int_S f_{\boldsymbol{Y}}(\boldsymbol{y}) d\boldsymbol{y} = \int_0^1 \int_{y_2}^1 4\frac{y_2}{y_1} \mathrm{d}y_1 \mathrm{d}y_2$$

$$= \int_0^1 -4y_2 \ln(y_2) \mathrm{d}y_2 = \text{Integration by parts...} = 1$$

(d) Work out

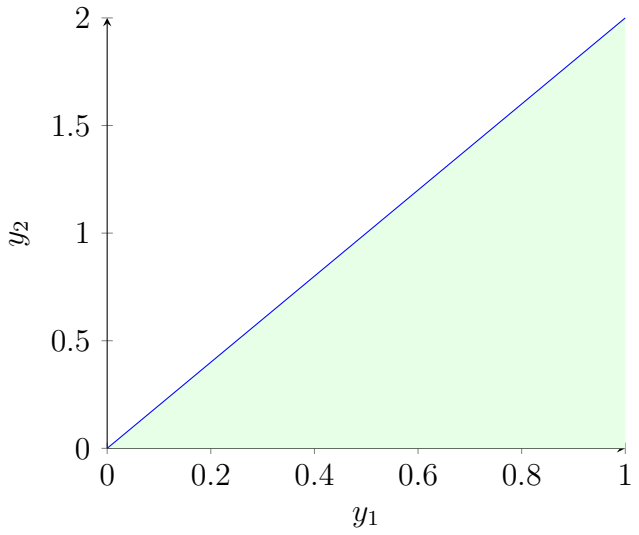
(i) the marginal distributions for Y_1 and Y_2, sketching both distributions and checking that they both integrate to 1;

The marginals are:

$$f_{\underline{Y_1}}(Y_1) = \int_T f_{\underline{y_1,y_2}}(y_1, y_2)\mathrm{d}y_2 = \int_0^{y_1} 4\frac{y_2}{y_1}\mathrm{d}y_2 = \frac{4}{y_1} * \int_0^{y_1} y_2\mathrm{d}y_2$$

$$= \frac{4}{y_1} * [\Big|_0^{y_1}(\frac{y_2^2}{2})] = \frac{4}{y_1}[(\frac{y_1^2}{2}) = 2y_1$$

$$\int_0^1 2y_1 dy_1 = \Big|_0^1(y_1^2) = 1$$

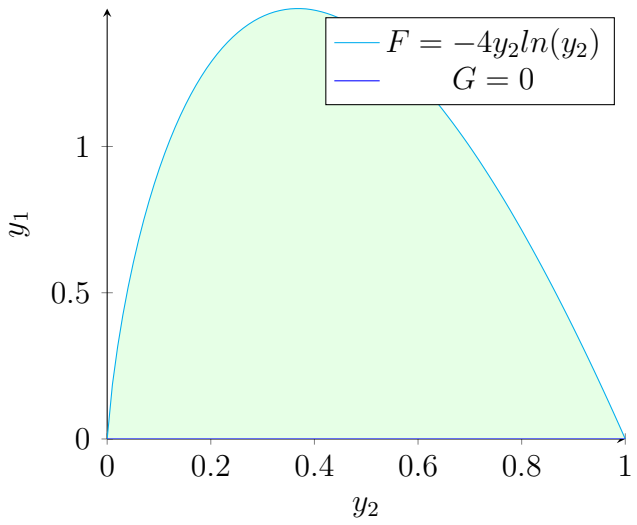


This is a sketch of the marginal:

$$f_{\underline{Y_1}}(Y_1)$$

and

$$f_{\underline{Y_2}}(Y_2) = \int_T f_{\underline{Y_1,Y_2}}(Y_1, Y_2)\mathrm{d}Y_1 = \int_{y_2}^1 4\frac{y_2}{y_1}\mathrm{d}y_1 = 4y_2 * \int_{y_2}^1 \frac{1}{y_1}\mathrm{d}y_1 = 4y_2 * \Big|_{y_2}^1 ln(y_1) = -4y_2 \ln(y_2)$$



Is the sketch of the marginal:

$$f_{\underline{Y_2}}(Y_2)$$

$$\int_0^1 -4y_2 \ln(y_2) \mathrm{d}y_2 = \text{Integration by parts...} = 1$$

(ii) the conditional distributions $f_{Y_1 \mid Y_2}(y_1 \mid y_2)$ and $f_{Y_2 \mid Y_1}(y_2 \mid y_1)$, checking that they each integrate to 1; and

Recall that:

$$f_{Y_1 \mid Y_2}(y_1 \mid y_2) = \frac{F_{Y_1 Y_2}(y_1 y_2)}{F_{Y_2}(y_2)} = \frac{\text{CDF of}(y_1 y_2)}{\text{Marginal CDF of}(y_2)}$$

Which is:

$$\frac{\int_0^1 \int_{y_2}^1 \frac{4y_2}{y_1} \mathrm{dy_1 y_2}}{\int_0^1 -4y_2 ln(y_2) \mathrm{d}y_2} = \frac{\int_0^1 -4y_2 ln(y_2) \mathrm{d}y_2}{1} = 1$$

extending this same logic:

$$f_{Y_2 \mid Y_1}(y_2 \mid y_1) = \frac{F_{Y_2 Y_1}(y_2 y_1)}{F_{Y_1}(y_1)} = \frac{\text{CDF of}(y_2 y_1)}{\text{Marginal CDF of}(y_1)} = 1$$

borrowing from the above integrals in part (i) we can easily see both conditional distributions integrate to 1 as these integrals have already been evaluated to integrate to 1 and therefore the entire fraction is also 1.

(iii) the conditional expectations $E(Y_1 \mid Y_2)$ and $E(Y_2 \mid Y_1)$; and

Recall that:

$$E(Y_1 \mid Y_2) = \frac{E(Y_2 \mid Y_1)E(Y_1)}{E(Y_2)} = \frac{\int_{-\infty}^{\infty} Y_1(2Y_1)}{\int_{-\infty}^{\infty} Y_2(-4Y_2 \ln(Y_2))}$$

(iv)the conditional variances $V(Y_1 \mid Y_2)$ and $V(Y_2 \mid Y_1)$. (*Hint:* recall that the variance of a random variable W is just $E(W^2) - [E(W)]^2$.)

[120 points]

Are Y_1 and Y_2 independent? Present any relevant calculations to support your answer. *[10 points]*

In order for Y_1 and Y_2 to be independent the following equation must hold true:

$$f_{\underline{Y_1}}(Y_1) * f_{\underline{Y_2}}(Y_2) = f_y(y) = f_{\underline{Y_1,Y_2}}(Y_1, Y_2)$$

the marginals are:

$$f_{\underline{Y_2}}(Y_2) = -4y_2 ln(y_2)$$

$$f_{\underline{Y_1}}(Y_1) = 2y_1$$

and

$$f_y(y) = 4\frac{y_2}{y_1}$$

$$f_{\underline{Y_2}}(Y_2) * f_{\underline{Y_1}}(Y_1) = -8y_1y_2 ln(y_2) \neq f_y(y)$$

Therefore Y_1 and Y_2 are not independent.

Either work out the correlation $\rho(Y_1, Y_2)$ between Y_1 and Y_2 or explain why no calculation is necessary in correctly identifying the value of ρ. *[10 points]*

$$\rho(y_1, y_2) = \frac{Cov(Y_1, Y_2)}{\sigma_{y_1}\sigma_{y_2}} = \frac{E(Y_1Y_2) - E(Y_1)E(Y_2)}{\sigma_{y_1}\sigma_{y_2}}$$

3. *[100 total points]* (moment-generating functions) Distributions may in general be skewed, but there may be conditions on their parameters that make the skewness get smaller or even disappear. This problem uses moment-generating functions (MGFs) to explore that idea for two important discrete distributions, the Binomial and the Poisson.

(a) We saw in class that if $X \sim \text{Binomial}(n, p)$, for $0 < p < 1$ and integer $n \geq 1$, then the MGF of X is given by

$$\psi_X(t) = \left[p\, e^t + (1-p)\right]^n . \tag{7}$$

for all real t, and we used this to work out the first three moments of X (note that the expression for $E\left(X^3\right)$ is only correct for $n \geq 3$):

$$E(X) = n\, p\,, \quad E\left(X^2\right) = n\, p[(1 + (n-1)p]\,, \tag{8}$$

$$E\left(X^3\right) = n\, p[1 + (n-2)(n-1)p^2 + 3\,(n-1)p]\,, \tag{9}$$

from which we also found that $V(X) = n\, p(1-p)$. Show that the above facts imply that

$$\text{skewness}(X) = \frac{1 - 2\, p}{\sqrt{n\, p(1-p)}}\,. \tag{10}$$

Under what condition on p, if any, does the skewness vanish? Under what condition on n, if any, does the skewness tend to 0? Explain briefly. *[30 points]*

We know that skewness(X) can be expressed as:

$$\text{Skewness}(X) = (\frac{E(X-\mu)}{\sigma})^3$$

or

$$\text{Skewness}(X) = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}$$

where μ_i is a central moment.

By extension:

$$\text{Skewness}(X) = \frac{np(1-p)(1-2p)}{(np(1-p))^{\frac{3}{2}}} = \frac{1-2p}{\sqrt{np(1-p)}}$$

as shown above.

If $p = \frac{1}{2}$ then the numerator is 0 and the skewness vanishes. If n grows very large, skewness tends to zero or is minimized as its role in this equation is in the denominator. This makes sense as massively increasing sample size in a binomial distribution will cause the data to balance itself. Additionally, when p is .5 we know that there is an equal probability of failure or success which implies symmetric or non-skewed data in the first place.

(b) In our brief discussion of stochastic processes we encountered the *Poisson* distribution: if $Y \sim \text{Poisson}(\lambda)$, for $\lambda > 0$, then the PF of Y is

$$f_Y(y) = \left\{ \begin{array}{cc} \frac{\lambda^y\, e^{-\lambda}}{y!} & \text{for } y = 0, 1, \ldots \\ 0 & \text{otherwise} \end{array} \right\} . \tag{11}$$

(i) Use this to show that for all real t the MGF of Y is

$$\psi_Y(t) = e^{\lambda\left(e^t-1\right)}\,. \tag{12}$$

[10 points]

$$E(e^{(\lambda)t}) = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!}$$

The summation is similiar to the exponenetial function:

$$\sum_{x=0}^{\infty}\frac{a^x}{x!} = e^a$$

Thus:

$$E(e^{(\lambda)t}) = e^{-\lambda}e^{(\lambda)e^t}$$

Therefore:

$$\psi_y(t) = e^{\lambda(e^t-1)}$$

(ii) Use $\psi_Y(t)$ to compute the first three moments of Y, the variance of Y and the skewness of Y. Under what condition on λ, if any, does the skewness either disappear or tend to 0? Explain briefly. *[60 points]*

The first moment is:

$$M_1 = \frac{d_1}{d_1 t}e^{\lambda(e^t-1)} = \lambda(e^t)e^{\lambda(e^t-1)}$$

The second moment is:

$$M_2 = \frac{d_2}{d_2 t}e^{\lambda(e^t-1)} = \lambda(e^t)e^{\lambda(e^t-1)} + (\lambda(e^t))^2e^{\lambda(e^t-1)} = \lambda(e^t)e^{\lambda(e^t-1)}[1+\lambda(e^t)]$$

The third moment shows a pattern:

$$\frac{d_3}{d_3 t}e^{\lambda(e^t-1)} = M_1(1+\lambda(e^t)) + M_2(\lambda(e^t))$$

Which can be expanded to:

$$\lambda(e^t)e^{\lambda(e^t-1)}[1+\lambda(e^t)]^2 + \lambda(e^t)^2e^{\lambda(e^t-1)}$$

Recall that:

$$Var(x) = E(Y^2) = (E(Y))^2$$

and:

$$E(Y^2) = \sum Y^2 Pr(Y = y)$$

Using these facts and what we know about exponential sum forms:

$$E(Y^2) = \sum_{t\geq 0} t^2 \frac{1}{t!}\lambda^t e^{-\lambda}$$

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

$$= \lambda(e^{-\lambda})(\sum_{t\geq 1}(t-1)\frac{1}{(t-1)!}\lambda^{t-1} + \sum_{t\geq 1}\frac{1}{(t-1)!}\lambda^{t-1})$$

$$= \lambda(e^{-\lambda})(\sum_{t\geq 1}\frac{1}{(t-2)!}\lambda^{t-2} + \sum_{t\geq 1}\frac{1}{(t-1)!}\lambda^{t-1})$$

$$= \lambda(e^{-\lambda})(\sum_{i\geq 0}\frac{1}{i!}\lambda^{i} + \sum_{j\geq 1}\frac{1}{j!}\lambda^{j})$$

$$= \lambda(e^{-\lambda})(\lambda(e^{\lambda}) + e^{\lambda})$$

$$= \lambda(\lambda + 1)$$

$$= \lambda^2 + \lambda$$

$$var(Y) = E(Y^2) - (E(Y))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

$$var(Y) = \lambda$$

The annoyingly complicated calculations of var(Y) help us with skewness below:

$$\text{Skewness}(Y) = (\frac{E(Y-\mu)}{\sigma})^3$$

or

$$\text{Skewness}(Y) = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}$$

where μ_i is a central moment.
and:
$E(Y^2) = \lambda^2 + \lambda$,
$E(Y) = \lambda$,
$\sigma = \sqrt{\lambda}$

By extension:

$$Skew(Y) = \frac{\lambda^3 + 3\lambda^2 + \lambda - 3\lambda^3 - 3\lambda^2 + 2\lambda^3}{\lambda^{\frac{3}{2}}} = \frac{\lambda}{\lambda^{\frac{3}{2}}}$$

$$= \frac{1}{\sqrt{\lambda}}$$

Assuming $\lambda > 0$, we can see that as λ grows very large, skewness tends to zero;however, the skewness can never actually be zero in this case.

4. *[140 total points]* (archaeology) Paleobotanists estimate the moment in the remote past when a given species became extinct by taking cylindrical, vertical core samples well below the earth's surface and looking for the last occurrence of the species in the fossil record, measured in meters above the point P at which the species was known to have first emerged. Letting $\{y_i, i = 1, \ldots, n\}$ denote a sample of such distances above P at a random set of locations, the model $\boxed{(Y_i|\theta) \overset{\text{IID}}{\sim} \text{Uniform}(0, \theta)\ (*)}$ emerges from simple and plausible assumptions. In this model the unknown $\theta > 0$ can be used, through carbon dating, to estimate the species extinction time.

The marginal distribution of a single observation y_i in this model may be written

$$p_{Y_i}(y_i \,|\, \theta) = \left\{ \begin{array}{cc} \frac{1}{\theta} & \text{if } 0 \le y_i \le \theta \\ 0 & \text{otherwise} \end{array} \right\} = \frac{1}{\theta}\, I\left(0 \le y_i \le \theta\right) , \tag{13}$$

where $I(A) = 1$ if A is true and 0 otherwise.

(a) Briefly explain why the statement $\{0 \le y_i \le \theta \text{ for all } i = 1, \ldots, n\}$ is equivalent to the statement $\{m = \max(y_1, \ldots y_n) \le \theta\}$, and use this to show that the joint distribution of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ in this model is

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \frac{I(m \le \theta)}{\theta^n} . \tag{14}$$

[20 points]

$\{0 \le y_i \le \theta$ for all i = 1, . . ., n$\} = \{m = max(y_1, ..., y_n) \le \theta\}$ because θ can be thought of as the last observed point above P at which the species was not extinct. Thus, y_i can be thought of as any observed vertical point in a core sample between P and θ. Every element must also follow $0 \le y_i \le \theta$ because $I(A) = 1$ when A is true, in this case $A = y_i$ for $i = \{1, ..., n\}$. Therefore, the right hand and left hand statements imply each other because $\max(y_i, ..., y_n)$ can be no larger than θ and still be true.

Extending this logic we can see that:

$$P_{Y_i}(y_i, ...y_n|\theta) = P_{Y_i}(y_1|\theta)P_{Y_i}(y_2|\theta)...P_{Y_i}(y_n|\theta) = \frac{1}{\theta^n} I(m = max(y_i, ..., y_n \le \theta)$$

given the multiplicative rule of probabilities.

(b) Letting the observed values of $(Y_1, \ldots, Y_n)$ be $\boldsymbol{y} = (y_1, \ldots, y_n)$, an important object in both frequentist and Bayesian inferential statistics is the *likelihood function* $\ell(\theta \,|\, \boldsymbol{y})$, which is obtained from the joint distribution of $(Y_1, \ldots, Y_n)$ simply by

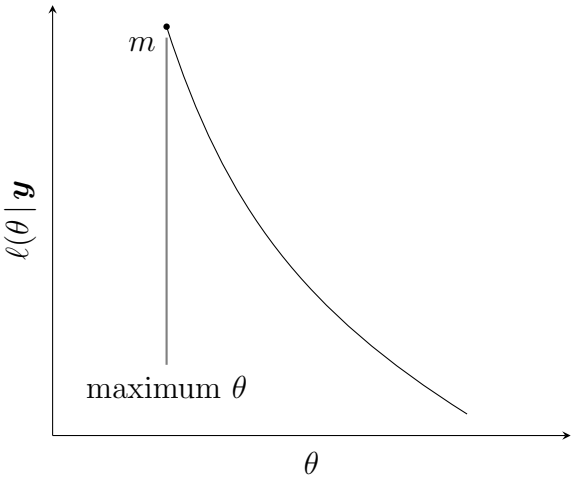
(1) thinking of $f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)$ as a function of θ for fixed $\boldsymbol{y}$, and

(2) multiplying by an arbitrary positive constant c:

$$\ell(\theta \,|\, \boldsymbol{y}) = c\, f_{\boldsymbol{Y}}(\boldsymbol{y})\,. \tag{15}$$

Using this terminology, in part (a) you showed that the likelihood function in this problem is $\ell(\theta \,|\, \boldsymbol{y}) = \theta^{-n} I(\theta \geq m)$, where m is the largest of the y_i values. Both frequentists and Bayesians are interested in something called the *maximum likelihood estimator* (MLE) $\hat{\theta}_{\text{MLE}}$, which is the value of θ that makes $\ell(\theta \,|\, \boldsymbol{y})$ as large as possible.

(i) Make a rough sketch of the likelihood function, and use your sketch to show that the MLE in this problem is $\hat{\theta}_{\text{MLE}} = m = \max(y_1, \ldots y_n)$. *[20 points]*



This isn't the best sketch but the idea is that it's a decreasing function of θ for a fixed m. $\ell(\theta \,|\, \boldsymbol{y}$ is at a maximum for a fixed Y when θ is at a minimum. The minimum value of θ is then m. If $\theta < m$, $\ell(\theta \,|\, \boldsymbol{y} = 0$. if $\theta \geq m$,$\ell(\theta \,|\, \boldsymbol{y} > 0$. To find $\ell(\theta \,|\, \boldsymbol{y}$'s maximum we would check the value of θ in $[m, \infty)$.

(ii) Maximization of a function is usually accomplished by setting its first derivative to 0 and solving the resulting equation. Briefly explain why that method won't work in finding the MLE in this case. *[10 points]*

$\ell(\theta \,|\, \boldsymbol{y}) = 0$, $\forall \theta < m$

In one case, $\frac{d}{d\theta}\ell(\theta \,|\, \boldsymbol{y}) = 0$ but $\ell(\theta \,|\, \boldsymbol{y})$ is not maximized in this interval as it is equal to 0. In another case $[m, \infty)$, $\ell(\theta \,|\, \boldsymbol{y}) > 0$ but because it's a decreasing function of θ there are no tangent lines at any θ on the curve $\ell(\theta \,|\, \boldsymbol{y})$. This is why this method fails for finding the MLE in this particular case.

(c) A positive quantity W follows the *Pareto* distribution (written $W \sim \text{Pareto}(\alpha, \beta)$) if, for parameters $\alpha, \beta > 0$, it has density

$$f_W(w) = \left\{ \begin{array}{cc} \alpha\, \beta^{\alpha}\, w^{-(\alpha+1)} & \text{if } w \geq \beta \\ 0 & \text{otherwise} \end{array} \right\}. \tag{16}$$

This distribution has mean $\frac{\alpha\beta}{\alpha-1}$ (if $\alpha > 1$) and variance $\frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$ (if $\alpha > 2$).

(i) For frequentists the likelihood function is just a function $\ell(\theta \,|\, \boldsymbol{y})$, but for Bayesians it can be regarded as an un-normalized density function for θ. Show that, from this point of view, the likelihood function in this problem corresponds to the Pareto$(n-1, m)$ distribution. *[10 points]*

In the Bayesian case, the likelihood function can be regarded as:

$$\ell(\theta \,|\, \boldsymbol{y}) = \frac{1}{\theta^n} I(\theta \geq m) = \left\{ \begin{array}{cc} \theta^{-(n-1+1)} & \text{if } \theta \geq m \\ 0 & \text{otherwise} \end{array} \right\}$$

in order to make $\ell(\theta \,|\, \boldsymbol{y})$ a normalized density function, we must calculate the below equation:

$$c \int_m^{\infty} \theta^{-(n+1-1)} d\theta = 1$$

$$c\big[\frac{\theta^{-n+1}}{-n+1}\big]\Big|_m^{\infty} = 1$$

$$c\big[0 - \frac{m^{-n+1}}{-n+1}\big] = 1$$

$$c * \frac{1}{m^{n-1}(n-1)} = 1$$

$$c = (n-1)m^{n-1}$$

Therefore, we consider the likelihood function that corresponds to the density function:

$$c\ell(\theta \,|\, \boldsymbol{y}) = \left\{ \begin{array}{cc} (n-1)m^{n-1}\theta^{-(n-1+1)} & \text{if } \theta \geq m \\ 0 & \text{otherwise} \end{array} \right\}$$

(ii) Bayes's Theorem for a one-dimensional continuous unknown (such as θ in this situation) says that the conditional density $f_{\Theta\,|\,\boldsymbol{Y}}(\theta\,|\,\boldsymbol{y})$ for θ given $\boldsymbol{Y} = \boldsymbol{y}$ — which is called the *posterior distribution* for θ given the data — is a positive (normalizing) constant c times a PDF $f_\Theta(\theta)$ — called the *prior distribution* for θ — that captures any available information about θ external to the data set, times the likelihood distribution $\ell(\theta\,|\,\boldsymbol{y})$:

$$\begin{array}{ccccccc} f_{\Theta\,|\,\boldsymbol{Y}}(\theta\,|\,\boldsymbol{y}) & = & c & \cdot & f_\Theta(\theta) & \cdot & \ell(\theta\,|\,\boldsymbol{y}) \\ \left(\begin{array}{c}\text{posterior}\\ \text{distribution}\end{array}\right) & = & \left(\begin{array}{c}\text{normalizing}\\ \text{constant}\end{array}\right) & \cdot & \left(\begin{array}{c}\text{prior}\\ \text{distribution}\end{array}\right) & \cdot & \left(\begin{array}{c}\text{likelihood}\\ \text{distribution}\end{array}\right) \end{array} \tag{17}$$

The posterior distribution is the goal of a Bayesian inferential analysis: it summarizes *all* available information, both external to and internal to your data set. Show that if the prior distribution for θ in this problem is taken to be (16), under the model $(*)$ above the posterior distribution is $f_{\Theta\,|\,\boldsymbol{Y}}(\theta\,|\,\boldsymbol{y}) = \text{Pareto}\,[\alpha + n, \max(\beta, m)]$. (Bayesian terminology: Note that what just happened was that the product of two Pareto distributions (prior, likelihood) is another Pareto distribution (posterior); a prior distribution that makes this happen is called *conjugate* to the likelihood in the model.) *[20 points]*

Prior distribution of θ is pareto to $(\alpha, \beta) = f_\theta(\theta)$.

Now suppose:

$$f_{(\theta\,|\,y)} = \text{normalizing constant} * f_\theta(\theta) * c\ell(\theta\,|\,y).$$

by extension:

$$f_{\theta\,|\,y}(\theta\,|\,y)\alpha c f_\theta(\theta)\ell(\theta\,|\,y) = \left\{ \begin{array}{cc} \alpha\beta^\alpha\theta^{-(\alpha+1)}(n-1)m^{n-1}\theta^{-(n-1+1)} & \text{if } \theta \geq max(\beta, m) \\ 0 & \text{otherwise} \end{array} \right\}$$

(k involves the constants α, β, n.)

Similarly to the above, we must set:

$$\int_\theta f_{\theta\,|\,y}(\theta\,|\,y)d\theta = 1$$

with the condition that$(k' = (n+\alpha)\big[max(\beta, m)\big]^{n+\alpha})$:

$$\int_{\max(\beta,m)}^{\infty} \theta^{-(n+\alpha+1)}d\theta = 1$$

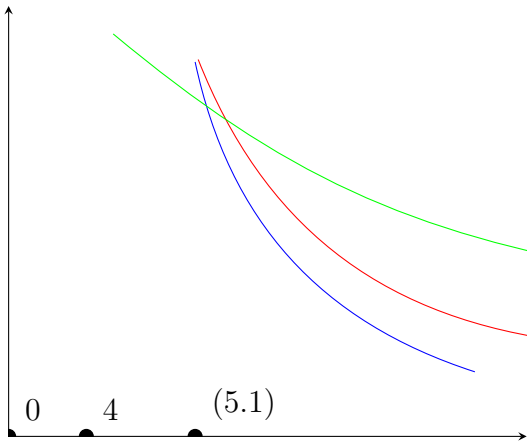
this will allow us to find an appropriate normalizing constant to express the posterior distribution:

$$f_{\theta\,|\,y}(\theta\,|\,y) = \left\{ \begin{array}{cc} (n+\alpha)max(\beta,m)^{(n+\alpha)})\theta^{-(n+\alpha+1)} & \text{if } \theta \geq max(\beta,m) \\ 0 & \text{otherwise} \end{array} \right\}$$

Which is the density Pareto to our $(n+\alpha, max(\beta,m))$ distribution

(d) In an experiment conducted in the Antarctic in the 1980s to study a particular species of fossil ammonite, the following was a linearly rescaled version of the observed data: $y = (y_1,\ldots,y_n) = (2.8, 1.7, 1.0, 5.1, 3.7, 1.5, 4.3, 2.0, 3.2, 2.1, 0.4)$. Prior information equivalent to a Pareto distribution specified by the choice $(\alpha,\beta) = (2.5, 4)$ was available.

(i) Plot the prior, likelihood, and posterior distributions arising from this data set on the same graph, explicitly identifying the three curves. *[30 points]*



again, this sketch isn't perfect but the important pieces are labeled. I'm not great at tikzpictures yet but some basic assumptions are made about the distributions visually.

green = prior = $p(\theta)$

blue = posterior = $p(\theta\,|\,5.1)$

red = likelihood = $\ell(\theta\,|\,5.1)$

(ii) Work out the posterior mean and SD (square root of the posterior variance), and use them to complete the following sentence:

On the basis of this prior and data information, the θ value for this species of fossil ammonite is about _____, give or take about _____.

[30 points]

$n = 11, \alpha = 2.5, \gamma = 5.1$

$$V(\theta\,|\,m) = \frac{(n+\alpha)(\gamma)^2}{(n+\alpha-1)(n+\alpha-2)} = \frac{13.5 * 5.1^2}{12.5 * 11.5} \approx 2.4427$$

$$E(\theta \mid m) = \frac{(\alpha + \gamma + n)(\gamma)}{n + \alpha - 1} = \frac{18.6(5.1)}{12.5} = 7.588$$

by extension:

On the basis of this prior and data information, the θ value for this species of fossil ammonite is about _____*7.59*_____*, give or take about* _____*2.44*_____*.*