Solutions Homework 5

October 10, 2018

Solution to Exercise 1.5.11: Let

$$p(B,x) \;=\; \delta_x(B) \;=\; I_B(x),$$

for $x \in I\!R$ and $B \in \mathcal{B}$. Clearly $p(\cdot,x)$ is a Borel probability measure for each x, so condition (ii) in the definition of conditional distributions holds (Definition 1.5.2). Turning to condition (i), of course $P[X \in B|X] = E[I_B(X)|X]$ by the definition of conditional probability, and by Theorem 1.5.7(f), $E[I_B(X)|X] = I_B(X)$, a.s., so $E[I_B(X)|X = x] = I_B(x)$, P_X-a.s. But, as noted above, this is just $p(B,x)$.

This is intuitively the right answer since given $X = x$, we should have a conditional probability totally concentrated on x, which is what δ_x does.

Solution to Exercise 1.5.12: We are looking for a conditional distribution of a two dimensional random vector, so our conditional distribution has to "live" on two dimensional space. Note that we are given that the first component $X_1 = x_1$, so its marginal conditional distribution should be δ_{x_1}. See the result in Exercise 1.5.11. Now, the marginal conditional distribution for X_2 should be the obvious answer: $P_{X_2|X_1}(\cdot|x_1) = \text{Law}[X_2|X_1 = x_1]$. Note that any r.v. is independent of a degenerate r.v., so there is only one way to make a joint distribution with these marginals:

$$p(\cdot,x_1) \;=\; \delta_{x_1} \times \text{Law}[X_2|X_1 = x_1].$$

Let's check that this satisfies the requisite properties as spelled out in Remark 1.5.7(a):

(1) For all $x_1 \in I\!R$, $p(\cdot,x_1)$ is a Borel p.m. on $(I\!R^2, \mathcal{B}^2)$ since it is the product of two Borel p.m.'s on $I\!R$.

(2) We want to check that for all $B \in \mathcal{B}^2$, $p(B,x_1)$ is a measurable function of x_1. Note that

$$\begin{aligned} p(B,x_1) &= \int_{R^2} I_B(\xi_1,\xi_2)\, d\left[\delta_{x_1} \times P_{X_2|X_1}(\cdot|x_1)\right](\xi_1,\xi_2) \\ &= \int_R \left[\int_R I_B(\xi_1,\xi_2)\, d\delta_{x_1}(\xi_1)\right] dP_{X_2|X_1}(\cdot|x_1)(\xi_2) \end{aligned}$$

$$\begin{aligned} &= \int_R I_B(x_1, \xi_2)\, dP_{X_2|X_1}(\cdot|x_1)(\xi_2) \\ &= P_{X_2|X_1}(B_1(x_1)|x_1) \\ &= P\left[X_2 \in B_1(x_1)|X_1 = x_1\right], \end{aligned}$$

where

$$B_1(x_1) \;=\; \{x_2 \in I\!R : (x_1, x_2) \in B\}.$$

Note that for each x_1, this set is measurable – this is implicitly one of the conclusions of Fubini's theorem, that we can fix the value of one variable and then the function is measurable in the other variable. Now, the last expression in our little calculation above, namely $P\left[X_2 \in B_1(x_1)|X_1 = x_1\right]$, is a Borel function of x_1 by definition of (this kind of) conditional probability (expectation).

(3) Now, we want to show that for all Borel sets $A \in I\!R$, $B \in I\!R^2$,

$$P[X_1 \in A \,\&\, (X_1, X_2) \in B] \;=\; \int_R I_A(x_1) p(B, x_1) dP_{X_1}(x_1).$$

If we write out $p(B, x_1)$ as an integral and carry out the calculation as in the previous step, then we obtain

$$\begin{aligned} &\int_R I_A(x_1) p(B, x_1) dP_{X_1}(x_1) \\ &\quad = \int_R I_A(x_1) \left[\int_R I_B(x_1, x_2)\, dP_{X_2|X_1}(\cdot|x_1)(x_2)\right] dP_{X_1}(x_1) \\ &\quad = \int_R \int_R I_A(x_1) I_B(x_1, x_2)\, dP_{X_2|X_1}(\cdot|x_1)(x_2) dP_{X_1}(x_1) \\ &\quad = \int_R E\left[I_A(X_1) I_B(X_1, X_2)|X_1 = x_1\right]\, dP_{X_1}(x_1) \\ &\qquad\quad \text{(by equation (1.71) in Theorem 1.5.6)} \\ &\quad = E\left[E\left[I_A(X_1) I_B(X_1, X_2)|X_1\right]\right] \\ &\quad = E\left[I_A(X_1) I_B(X_1, X_2)\right] \\ &\quad = P[X_1 \in A \,\&\, (X_1, X_2) \in B], \end{aligned}$$

which is the desired result.

Solution to Borel Handout Problem 2. Problem Statement: *Here is another example: Let (U, V) be uniformly distributed on the unit square, i.e. they have joint (Lebesgue) probability density*

$$f_{UV}(u, v) \;=\; 1, \quad 0 < u < 1,\; 0 < v < 1.$$

(a) *Let $X = V - U$. Find the joint density of U and X and the conditional density of U given $X = x$.*

Solution: Let $Z = U$. This is a fairly simple linear transformation. The inverse transform is

$$\begin{aligned} h_1(x,z) &= z, \\ h_2(x,z) &= x + z. \end{aligned}$$

So, $(U,V) = (h_1(X,Z), h_2(X,Z)) = h(X,Z)$. The Jacobian is easy to compute:

$$J(x,z) \;=\; |\det Dh(x,z)| \;=\; \left|\det \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}\right| \;=\; 1.$$

Hence,

$$\begin{aligned} f_{XZ}(x,z) &= f_{UV}(h_1(x,z), h_2(x,z)) J(x,z) \\ &= I_{(0,1)}(z) I_{(0,1)}(x+z) \\ &= I_{(-1,0]}(x) I_{(-x,1)}(z) \;+\; I_{(0,1)}(x) I_{(0,1-x)}(z). \end{aligned}$$

One may check that the last equation is true by logically verifying that the inequalities $0 < z < 1$ and $0 < x + z < 1$ are equivalent to

$$(-1 < x \le 0 \,\&\, -x < z < 1) \quad \text{or} \quad (0 < x < 1 \,\&\, 0 < z < 1 - x).$$

It is easiest to see this by looking a the picture of $\{(x,z) : 0 < x + z < 1$ and $0 < z < 1\}$ which is a parallelogram with vertices at $(0,0)$, $(-1,1)$, $(0,1)$, and $(1,0)$. Now we can obtain the conditional density for Z given X by inspection:

$$f_{Z|X}(z|x) \;=\; C(x) I_{(-1,0]}(x) I_{(-x,1)}(z) \;+\; C(x) I_{(0,1)}(x) I_{(0,1-x)}(z),$$

where $C(x)$ is determined by $\int f_{Z|X}(z|x)dz = 1$ for each $x \in (-1,1)$. Now for each $x \in (-1,1)$, only one of the terms is positive, and by trivial calculus we get

$$f_{Z|X}(z|x) \;=\; I_{(-1,0]}(x)(1+x)^{-1} I_{(-x,1)}(z) \;+\; I_{(0,1)}(x)(1-x)^{-1} I_{(0,1-x)}(z).$$

Put differently,

$$Z|X = x \;\sim\; \begin{cases} \text{Unif}(-x, 1) & \text{if } -1 < x \le 0, \\ \text{Unif}(0, 1-x) & \text{if } 0 < x < 1. \end{cases}$$

In conclusion, since $Z = U$ we get that the conditional distribution of U given $X = 0$ is Unif(0,1), at least based on this family of conditional distributions.

(b) *Let $Y = V/U$. Find the joint density of U and Y and the conditional density of U given $Y = y$.*

Solution: Let $Z = U$ as before. Then the inverse transforms are

$$\begin{aligned} h_1(y,z) &= z \\ h_2(y,z) &= yz. \end{aligned}$$

The Jacobian is

$$J(y,z) \;=\; \left|\det \begin{bmatrix} 0 & 1 \\ z & y \end{bmatrix}\right| \;=\; |z| \;=\; z,$$

noting that $z > 0$. Hence, the joint density for (Y, Z) is

$$\begin{aligned} f_{YZ}(y,z) &= f_{UV}(h_1(y,z), h_2(y,z))J(y,z) \\ &= I_{(0,1)}(z)I_{(0,1)}(yz)z. \end{aligned}$$

It helps to draw a picture of the region where this joint density is non-zero, but let's reason it out "algebraically". We have the inequalities

$$0 < z < 1, \quad 0 < yz < 1.$$

Since we want to get $f_{Z|Y}$, we need to fix values of y and see the dependence on z, so we want to get bounds on z where the bounds depend on y, i.e. solve for z. The inequalities give us $0 < z < \min\{1, 1/y\}$, and of course $0 < y$. Hence, we have

$$\begin{aligned} y \le 1 &\;\Rightarrow\; 0 < z < 1, \\ y > 1 &\;\Rightarrow\; 0 < z < 1/y. \end{aligned}$$

Hence, we can write

$$f_{YZ}(y,z) \;=\; I_{(0,1]}(y)I_{(0,1)}(z)z \,+\, I_{(1,\infty)}(y)I_{(0,1/y)}(z)z.$$

As before,

$$\begin{aligned} f_{Z|Y}(z|y) &= C(y)f_{YZ}(y,z) \\ &= 2zI_{(0,1)}(z)I_{(0,1]}(y) \,+\, 2y^2 zI_{(0,1/y)}(z)I_{(1,\infty)}(y). \end{aligned}$$

Putting in $y = 1$ we get

$$f_{Z|Y}(z|1) \;=\; 2zI_{(0,1)}(z).$$

(c) *Note that the events* $[U = V]$ *and* $[X = 0]$ *and* $[Y = 1]$ *are all the same. However,*

$$f_{U|X}(u|0) \;\neq\; f_{U|Y}(u|1).$$

Explain.

Solution: The easiest and probably best way to explain this is to assume that we don't observe exact values for the random variables, but rather a rounded off version. So, thinking of part (a), we observe $(\tilde{X}, \tilde{Z})$ which are obtained by rounding off (X, Z) to a high number of decimal places, and then we have discrete random variables. If they are rounded off to k decimal places, then $\tilde{X} = X + \Delta X$ where $|\Delta X| \leq .5 * 10^{-k} = \epsilon$, and

$$\begin{aligned} P[\tilde{X} = \tilde{x}\,\&\,\tilde{Z} = \tilde{z}] &\doteq f_{XZ}(x, z)\epsilon^2, \\ P[\tilde{X} = \tilde{x}] &\doteq f_X(x)\epsilon, \\ P[\tilde{Z} = \tilde{z}|\tilde{X} = \tilde{x}] &\doteq \frac{f_{XZ}(x, z)\epsilon^2}{f_X(x)\epsilon} \;=\; f_{Z|X}(z|x)\epsilon. \end{aligned}$$

Of course, the same calculations hold for $Z|Y$.

None of this gives a good intuitive explanation for me, yet. I think the most informative way of looking at it is to "plot" the calculations in the original coordinates where area is (proportional to) probability. So, for the first transformation in part (a), we have illustrated the results with $\epsilon = 0.05$, i.e., rounding off to one decimal place. This appears in Figure 1. The shaded area shows the values of (U, V) where $X = V - U$ would be rounded off to 0.0, i.e. $-0.05 \leq X \leq 0.05$. The darker shaded area shows the intersection of the events $[-.05 \leq V - U \leq .05]$ and $[.15 \leq U \leq .25]$, where the latter event corresponds to a rounded off value of U which is .2. Then the conditional probability

$$P[.15 \leq U \leq .25\,|\,-.05 \leq X \leq .05]$$

is the ratio of the lower left darker area to the area of the entire shaded strip. The upper right darker shaded area corresponds to the intersection of $[-.05 \leq V - U \leq .05]$ and $[.75 \leq U \leq .85]$, where the latter event corresponds to a rounded off value of U which is .8. We see that the two

conditional probabilities (for U near .2 and U near .8) are the same. In fact, all the conditional probabilities for different values of U will be the same, except when U is near the endpoints 0 and 1, when the analogue of the darker areas will be cut off by the limits on U. Thus, the conditional distribution of a rounded off U given a rounded off value of X will be nearly uniform, with some deviations at the boundaries.

Now we consider the situation when we replace X with $Y = V/U$. The shaded area represents the event $.95 \leq V/U \leq 1.05$, which is the event that the rounded off value of $Y = V/U$ is 1 (when rounding off to one decimal place after the decimal point). Again, the lower left darker shaded area corresponds to the intersection of this given event with $|U - .2| \leq .05$, and the upper darker area to $|U - .8| \leq .05$. It is easy to check (using the formula for area of a trapezoid) that the area corresponding to a rounded off value of U intersected with $.95 \leq V/U \leq 1.05$ will be proportional to the rounded off value, except at the endpoints. This makes the conditional density $f_{U|Y}(u|1) = 2uI_{(0,1)}(u)$ seem intuitively correct.

Going back to the conditional densities for the continuous random variables, we could make $f_{U|X}(u|0)$ and $f_{U|Y}(u|1)$ the same since the events $[X = 0]$ and $[Y = 1]$ have probability 0, but then at least one of the conditional densities would not be continuous (in the given variable, x or y), which seems artificial. In any event, conditional distributions satisfy defining properties as a family and are unique only up to sets of probability 0. There is no way to say whether particular distributions given $[V - U = 0]$ and $[V/U = 1]$ are correct or not, but only if a family of distributions given $V - U$ or V/U is correct.

Solution to Exercise 2.1.7 Using linearity of the expectation operator and basic matrix algebra,

$$\begin{aligned}
\mathrm{Cov}[\underline{X}, \underline{Y}] &= E\left[(\underline{X} - E[\underline{X}])(\underline{Y} - E[\underline{Y}])^T\right] \\
&\mathit{speq} \quad E[\underline{X}\underline{Y}^T] - E\left[E[\underline{X}]\underline{Y}^T\right] - E\left[\underline{X}E[\underline{Y}]^T\right] + E[\underline{X}]E[\underline{Y}]^T \\
&\mathit{speq} \quad E[\underline{X}\underline{Y}^T] - E[\underline{X}]E[\underline{Y}]^T - E[\underline{X}]E[\underline{Y}]^T + E[\underline{X}]E[\underline{Y}]^T \\
&\mathit{speq} \quad E[\underline{X}\underline{Y}^T] - E[\underline{X}]E[\underline{Y}]^T.
\end{aligned}$$

Solution to Exercise 2.1.17

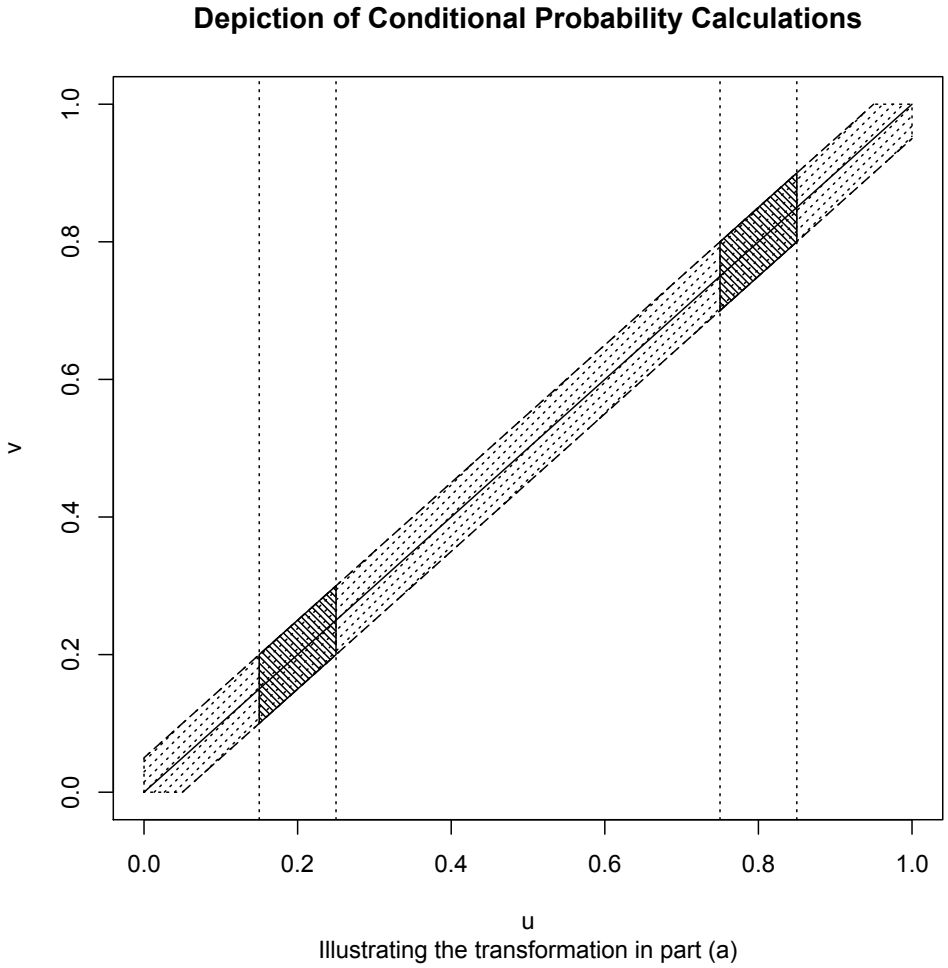

Figure 1: Depiction of conditional probabilities for U given $V - U$ is close to 0.

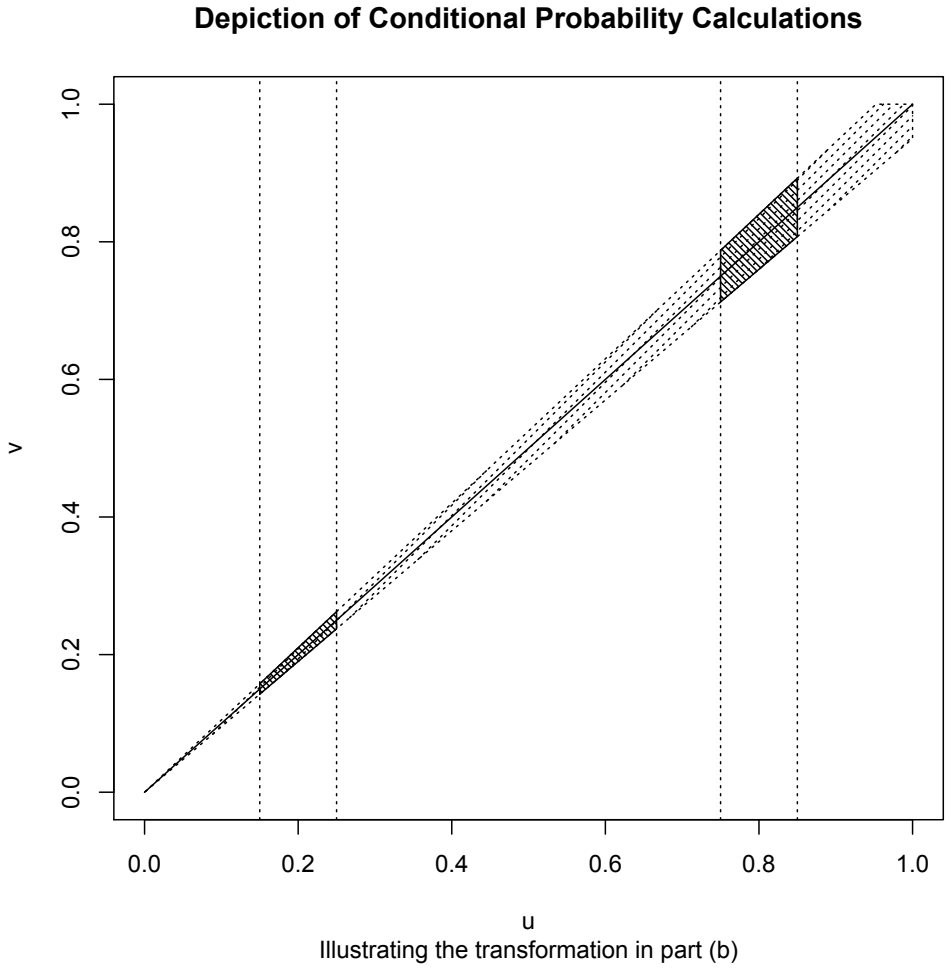

Figure 2: Depiction of conditional probabilities for U given V/U is close to 1.

If $r \geq 1$, then for $x \geq 0$, the mapping $f(x) = x^r$ is convex (the first derivative is $f'(x) = rx^{r-1}$ which is nondecreasing on $[0, \infty)$). Taking $r = q/p$ with $1 \leq p \leq q$, we have by Jensen's inequality

$$\begin{aligned} E[|X|^p]^{q/p} &= f\left(E[|X|^p]\right) \\ &\leq E\left[f(|X|^p)\right] \\ &= E\left[|X|^q\right]. \end{aligned}$$

Now, take the $1/q$ power of both sides and use the fact that power functions are nondecreasing for positive powers.

Solution to Exercise 2.1.20: **(a)** The claim is that $\psi(t) = t \log t \geq t - 1$, which we should verify.

The figure shows a plot of $\psi(t)$ and $t-1$ and it appears that the following are true:

(i) $t - 1$ is tangent to $\psi(t)$ at $t = 1$;

(ii) $\psi(t)$ is a strictly convex function;

(iii) it is always above its tangent.

Let's verify these three claims. We claim that the function $\psi(t) = t \log t$ is strictly convex. The first and second derivatives are

$$\begin{aligned} \psi'(t) &= \log t + 1, \\ \psi''(t) &= 1/t. \end{aligned}$$

As the second derivative is strictly positive, the claim (ii) follows. Also, $\psi(1) = 0$ and $\psi'(1) = 1$, so the line $y = t - 1$ is the tangent at $t = 1$ (since the values and slopes agree). Claim (iii) is a general property of convex functions, but it is clear in case the function is differentiable (as it is here) since then we know the first derivative is increasing, so $\zeta(t) = \psi(t) - (t - 1)$ satisfies $\zeta'(1) = 0$, $\zeta'(t) < 0$ for $t < 1$, and $\zeta'(t) > 0$ for $t > 1$, so ζ has a unique minimum at $t = 1$.

To show the integral defining $K(Q, P)$ exists, we will show that the integral of the negative part is finite. Now we can't do anything with "$\int \log\left(\frac{dQ}{dP}\right) dQ$" until we know it exists because none of our theorems apply. As mentioned in class, we can show the integral of the negative part is finite. There are a couple of trivial little facts about negative parts we need:

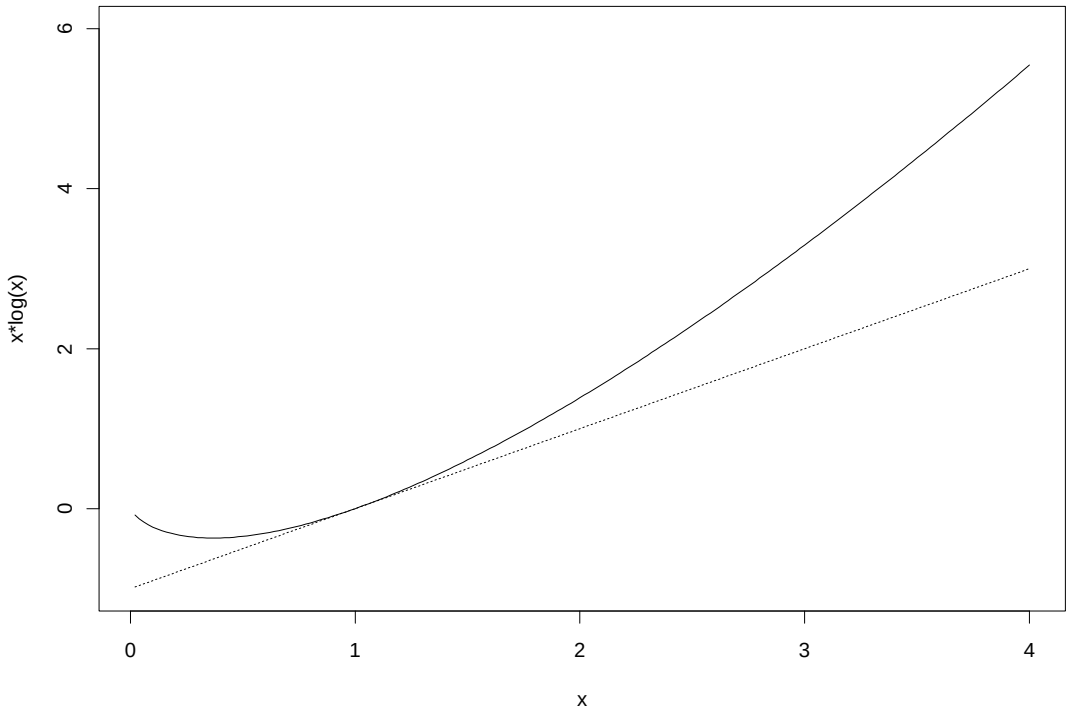

Figure 3: Plot of $y = x \log(x)$ with line $y = x - 1$ overlaid.

Trivial Fact #1. $f \leq g$ then $f_- \geq g_-$.

Trivial Fact #2. if $f \geq 0$ then $(fg)_- = fg_-$.

So, we have

$$\begin{aligned} \int \left(\log \left[\frac{dQ}{dP} \right] \right)_- dQ &= \int \left(\log \left[\frac{dQ}{dP} \right] \right)_- \frac{dQ}{dP} \, dP \\ &\qquad \text{by Proposition 1.4.2 (a)} \\ &= \int \left(\log \left[\frac{dQ}{dP} \right] \frac{dQ}{dP} \right)_- dP \\ &\qquad \text{using Trivial Fact \#2 since } \tfrac{dQ}{dP} \geq 0 \end{aligned}$$

$$\begin{aligned} &\leq \int \left(\frac{dQ}{dP} - 1\right)_- dP \\ &\qquad \text{using Trivial Fact \#1 and the inequality on } t\log t \\ &\leq \int (-1)_- \, dP \\ &\qquad \text{since } \tfrac{dQ}{dP} - 1 \geq 1 \\ &= 1. \end{aligned}$$

(b) Now we can apply the usual results on integration as in Proposition 1.2.5 since we have shown the integral is well defined.

$$\begin{aligned} \int \log\left(\frac{dQ}{dP}\right) dQ &= \int \log\left(\frac{dQ}{dP}\right)\frac{dQ}{dP}\, dP \\ &\geq \int \left(\frac{dQ}{dP} - 1\right) dP \\ &\qquad \text{by the inequality proved above} \\ &= \int \frac{dQ}{dP}\, dP \; - \int 1\, dP \\ &= \int 1\, dQ \; - \; 1 \\ &= 1 - 1 \; = \; 0. \end{aligned}$$

(c) Suppose $K(Q,P) = 0$, and we will show $P = Q$. Since ψ is strictly convex, by Jensen's inequality

$$\int \psi\left(\frac{dQ}{dP}\right) dP \; \geq \; \psi\left(\int \frac{dQ}{dP}\, dP\right) \; = \; \psi(1) \; = \; 0,$$

with strict inequality if and only if $\frac{dQ}{dP}$ is constant, P-a.s. One observes that the l.h.s. of the displayed inequality is $K(Q,P)$, and the r.h.s. is $\psi\left(\int 1 dQ\right)$ $= 1\log 1 = 0$. Also, if $\frac{dQ}{dP}$ is constant, P-a.s., then the constant must be 1 because that's the integral of $\frac{dQ}{dP}$ w.r.t. P. But $\frac{dQ}{dP} = 1$, P-a.s., means $Q = P$.

Solution to Exercise 2.2.1

For any complex number $z = z_1 + iz_2$ it is clear that the modulus $|z| = \|\underline{z}\|$ where $\underline{z} = (z_1, z_2)$ is a 2-D vector. Therefore, the result will follow if we show that for any $f : (\Omega, \mathcal{F}) \longrightarrow (I\!R^d, \mathcal{B}_d)$, we have

$$\left\| \int f d\mu \right\| \; \leq \; \int \|f\| d\mu.$$

It will be useful to have this more general result.

If μ is a probability measure, then the result follows easily from Jensen's inequality. Note that the map $\psi(x) = \|x\|$ is convex since if $t \in [0,1]$, $\psi(tx + (1-t)y) = \|tx + (1-t)y\| \leq \|tx\| + \|(1-t)y\| = t\|x\| + (1-t)\|y\| = t\psi(x) + (1-t)\psi(y)$. Thus, writing $E[\cdots]$ in place of $\int \cdots d\mu$ and X in place of f we have

$$\psi(E[X]) \;=\; \|E[X]\| \;\leq\; E[\psi(X)] \;=\; E[\|X\|].$$

However, if μ is not a probability measure, this argument doesn't apply, and we are asked to prove the more general result. I have tried to find a simple proof, but it doesn't seem to exist. When you are stumped on some general result about integrals, think of simple functions.

Let ϕ be a vector valued simple function, i.e. $\phi = (\phi_1, \ldots \phi_d)$ where each component function ϕ_j is a simple function. For convenience, we can assume the sets A_i in the representations of each ϕ_j are the same:

$$\phi_j \;=\; \sum_i a_{ij} I_{A_i}.$$

In a vector notation:

$$\phi \;=\; \sum_i a_i I_{A_i}.$$

Now by the triangle inequality,

$$\begin{aligned} \left\| \int \phi d\mu \right\| &= \left\| \sum_i a_i \mu(A_i) \right\| \\ &\leq \sum_i \|a_i\| \, \mu(A_i) \\ &= \int \|\phi\| \, d\mu, \end{aligned}$$

which is the desired result.

Now we consider a general vector valued function f. Writing $f = (f_1, \ldots, f_d)$, we can find simple functions $\phi_{jn} \to f_j$ as $n \to \infty$ for each j, $1 \leq j \leq d$, and $|\phi_{jn}| \leq |f_j|$ for all j and n. This latter inequality implies $\|\phi_n\| \leq \|f\|$ where $\phi_n = (\phi_{1n}, \ldots, \phi_{dn})$ is a vector valued simple function. Note that we can assume $\int \|f\| d\mu < \infty$, since otherwise the inequality we are trying to prove is trivial (assuming $\int f d\mu$ is defined; which for vector valued functions would require all components are finite, which would imply $\int \|f\| d\mu < \infty$ anyway).

Thus by DCT we have $\int \phi_{jn} d\mu \to \int f_j d\mu$ for $1 \le j \le d$, hence $\int \phi_n d\mu \to \int f d\mu$, and since $\psi(x) = \|x\|$ is a continuous function, we have

$$\left\| \int \phi_n d\mu \right\| \to \left\| \int f d\mu \right\|.$$

Now it is easy to see that $\|\phi_n\|$ is a simple function and $\|\phi_n\| \le \|f\|$, so by DCT

$$\int \|\phi_n\| d\mu \to \int \|f\| d\mu.$$

By the result we just proved for simple functions, we have $\|\int \phi_n d\mu\| \le \int \|\phi_n\| d\mu$, so by the limiting results above we conclude $\|\int f d\mu\| \le \int \|f\| d\mu$.

Solution to Exercise 2.3.3 **(a)** First of all, we need to derive the m.g.f. for $\underline{Z}$ to verify that $\underline{Z}$ has the the $N(0, I)$ distribution. Some students simply jumped to the conclusion that $\underline{Z} \sim N(0, I)$ without proving it. We have

$$\begin{aligned}
\psi_{\underline{Z}}(\underline{u}) &= E\left[\exp\left(\underline{u}^t \underline{Z}\right)\right] \\
&= E\left[\exp\left(\sum_i u_i Z_i\right)\right] \\
&= E\left[\prod_i \exp\left(Z_i\right)\right] \\
&= \prod_i E\left[\exp\left(Z_i\right)\right] \\
&\qquad \text{by Theorem 1.3.4(b). since the components of } \underline{Z} \text{ are independent} \\
&= \prod_i \psi_{Z_i}(u_i).
\end{aligned}$$

It is easy to check that

$$\psi_{Z_i}(u_i) = \exp\left(u_i^2/2\right).$$

Hence,

$$\psi_{\underline{Z}}(\underline{u}) = \exp\left(\underline{u}^t \underline{u}/2\right) = \exp\left(\|\underline{u}\|^2/2\right).$$

Apply Theorem 2.2.1 (c) to obtain

$$\begin{aligned}
\psi_{\underline{X}}(\underline{v}) &= \exp\left(\underline{v}^t \underline{\mu}\right) \psi_{\underline{Z}}\left(A^t \underline{v}\right) \\
&= \exp\left(\underline{v}^t \underline{\mu}\right) \exp\left((A^t \underline{v})^t (A^t \underline{v})/2\right) \\
&= \exp\left(\underline{v}^t \underline{\mu} + \underline{v}^t (AA^t) \underline{v}\right),
\end{aligned}$$

which is the m.g.f. of a $N(\underline{\mu}, AA^t)$ distribution as defined in Definition 2.3.5.

(b) Looking at equation (2.2.43), we will need to work with the exponent to get it into the correct form. Write

$$W \;=\; V^{-1}.$$

Then one can check

$$\begin{aligned}(\underline{x}-\underline{\mu})^t V^{-1}(\underline{x}-\underline{\mu}) &= \underline{x}^t W \underline{x} \;-\; 2\underline{\mu}^2 W \underline{x} \;+\; \underline{\mu}^t W \mu \\ &= \sum_{i=1}^{n} W_{ii} x_i^2 \;+\; 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} W_{ij} x_i x_j \;-\; 2\sum_{i=2}^{n}\sum_{j=i}^{i-1} W_{ij}\mu_i x_j \;+\; \underline{\mu}^t W \underline{\mu}.\end{aligned}$$

Note that we have been careful to at least not duplicate terms. Thus, we could have combined the first two summations into $\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}x_i x_j$, but for $i \neq j$ there would be two identical terms. While it is not a requirement in order to put the density in the exponential family form, it is always desirable to avoid obvious linear constraints on the sufficient statistics and natural parameters. Anyway, we have then that the density has the form

$$\begin{aligned}f(\underline{x}) &= \frac{1}{(2\pi)^{n/2}\det(V)^{1/2}} \exp\left[-\frac{1}{2}\underline{\mu}^t W \underline{\mu}\right] \\ &\qquad \exp\left[-\frac{1}{2}\sum_{i=1}^{n} W_{ii}x_i^2 \;-\; \sum_{i=2}^{n}\sum_{j=1}^{i-1} W_{ij}x_i x_j \;+\; \sum_{j=1}^{n}\left(\sum_{i=1}^{n} W_{ij}\mu_i\right) x_j\right],\end{aligned}$$

which is an exponential family with

$$\begin{aligned}\eta_k &= \begin{cases} -\frac{1}{2}W_{kk} & \text{if } 1 \le k \le n, \\ W_{ij} & \text{if } k = n + (i-1)(i-2)/2 + j, \text{ for some } 2 \le i \le n, \\ & \quad 1 \le j \le i-1 \\ \sum_{i=1}^{n} W_{ij}\mu_i & \text{if } k = n(n+1)/2 + j, \text{ for some } 1 \le j \le n; \end{cases} \\ T_k &= \begin{cases} x_{kk}^2 & \text{if } 1 \le k \le n, \\ x_i x_j & \text{if } k = n + (i-1)(i-2)/2 + j, \text{ for some } 2 \le i \le n, \\ & \quad 1 \le j \le i-1 \\ x_i & \text{if } k = n(n+1)/2 + j, \text{ for some } 1 \le j \le n; \end{cases} \\ B(\underline{\mu}, V) &= \frac{1}{2}\log\det(V) \;+\; \frac{1}{2}\underline{\mu}^t W \underline{\mu} \\ h(\underline{x}) &= \frac{1}{(2\pi)^{n/2}}\end{aligned}$$

(c) The family is more or less in canonical form: one has to make $d\nu = (2\pi)^{-n/2} dm$ to eliminate the $h(\underline{x})$ factor and express the normalizing constant in terms of the $\underline{\eta}$. The latter is not quite trivial but purely algebraic manipulation. Checking to see if it is minimal and full rank is more important and not just purely algebraic manipulations. We will follow the usual procedure for checking full rank (which implies minimal) by checking that the sufficient statistic does not satisfy a linear constraint and then that the space of the natural parameter has nonempty interior. Now the sufficient statistic $\underline{T}(\underline{x})$ is basically the $n(n+3)/2$ vector of all linear and quadratic terms that can be made with the components of $\underline{x}$, so a linear constraint on $\underline{T}$ amounts to a quadratic constraint on $\underline{x}$. Since $\underline{x}$ varies over all of $\mathbb{R}^n$, it is intuitively clear it doesn't satisfy any such constraint, m-a.e. To see this rigorously, consider the set

$$A \;=\; \{\underline{x} : \sum_i a_i x_i^2 \;+\; \sum_{i<j} b_{ij} x_i x_j \;+\; \sum_i c_i x_i \;+\; d \;=\; 0\},$$

where the a_i, b_{ij}, c_i, and d are constants. Then fix all but a single component of $\underline{x}$, say x_1 is allowed to vary but x_2, ..., x_n are fixed. Then the set of x_1 satisfying the constraint is at most 2 points (the number of real solutions of a quadratic equation), which has Lebesgue measure 0. Then we apply a Fubini argument as in Exercise 1.3.17 to conclude $m^n(A) = 0$.

To see that the space of natural parameter values has nonempty interior, consider the map $\eta(\underline{\mu}, V)$ from $\mathbb{R}^n \times \mathcal{P}_n$ where $\mathcal{P}_n$ is the set of $n \times n$ strictly positive definite matrices. An element $V \in \mathcal{P}_n$ may be treated as an $n(n+1)/2$ dimensional vector of the independent entries in the matrix (recall such a V is symmetric, so there are not n^2 independent entries). We claim $\mathcal{P}_n$ is an open subset of $\mathbb{R}^{n(n+1)/2}$. Let $V \in \mathcal{P}_n$ and we will show there is a neighborhood of V which is a subset of $\mathcal{P}_n$. Now $S = \{\underline{x} \in \mathbb{R}^n : \|\underline{x}\| = 1\}$ is a compact set, and the mappin $\underline{x} \mapsto \underline{x}^t V \underline{x}$ is continuous so it achieves its minimum m on S, which is positive since V is strictly positive definite. If $T \in \mathcal{P}_n$ then $|\underline{x}^t V \underline{x} - \underline{x}^t T \underline{x}| \le n^2 \max_{ij} |V_{ij} - T_{ij}|$ for $\underline{x} \in X$ since $\max_i |x_i| < 1$ and $\max_{ij} |V_{ij} - T_{ij}| \le \|V - T\|$ when considered as vectors in $\mathbb{R}^{n(n+1)/2}$. Thus, if $\|V - T\| < m/(2n^2)$, it follows that $\underline{x}^t T \underline{x} > m/2 > 0$ for all $\underline{x} \in S$, and hence for any $\underline{x} \neq \underline{0}$ we have $\underline{x}^t T \underline{x} = \|\underline{x}\|^2 (\underline{x}/\|\underline{x}\|)^t T (\underline{x}/\|\underline{x}\|) > (\underline{x}/\|\underline{x}\|) m/2 > 0$, and hence T is strictly positive definite. This shows that the neighborhood of radius $m/(2n^2)$ about V is contained in $\mathcal{P}_n$, and hence $\mathcal{P}_n$ is open.

As V ranges over all of $\mathcal{P}_n$, so does it's inverse W. Thus, the range the first $n(n+1)/2$ components of $\underline{\eta}$ is essentially $\mathcal{P}_n$, an open subset of

$I\!R^{n(n+1)/2}$. (The map which multiplies the first n components by $-1/2$ is continuous and continuously invertible, so it preserves open sets). Now the last n components of $\underline{\eta}$ are obtained by multiplying an arbitrary $\underline{\mu} \in I\!R^n$ by W, so these vary over all of $I\!R^n$. Thus, the space of natural parameter values is $\mathcal{P}_n \times I\!R^n$, which is a nonempty open set, and so has nonempty interior.

Solution to Exercise 2.3.10:

(a) For x a nonnegative integer, we have

$$f_\lambda(x) \quad = \quad \exp\left[(\log\lambda)(x) - \lambda\right](x!)^{-1}.$$

The quantities in the exponential type family formulation are

$$\begin{aligned} \eta(\lambda) &= \log\lambda, \\ T(x) &= x, \\ B(\lambda) &= \lambda, \\ h(x) &= (x!)^{-1}. \end{aligned}$$

Switching to the natural parameter and dropping the $h(x)$, we have the canonical form

$$f(x;\eta) \;=\; \exp\left[\eta x - e^\eta\right],$$

and in particular, $A(\eta) = e^\eta$. As λ ranges over $(0,\infty)$, $\log\lambda$ ranges over all of $I\!R$, which is clearly the natural parameter space. This clearly has nonempty interior (any neighborhood of any point is contained in this natural parameter space). Also, $T(X)$ is not a degenerate r.v. (in one dimension, satisfying a linear constraint is the same as being a degenerate r.v.), so the family is full rank.

(b) We have for $x \in \{0, 1, \ldots, n\}$,

$$\begin{aligned} f_p(x) &= \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \\ &= \exp\left\{x \log[p/(1-p)] - n\log[1/(1-p)]\right\} \left(\begin{array}{c} n \\ x \end{array} \right), \end{aligned}$$

and the exponential family components are

$$\begin{aligned} \eta(p) &= \log[p/(1-p)], \\ T(x) &= x, \\ B(p) &= -n\log(1-p), \\ h(x) &= \left(\begin{array}{c} n \\ x \end{array} \right). \end{aligned}$$

In order to obtain $A(\eta)$ we have to invert the mapping $p \mapsto \eta(p)$, which is clearly,

$$p \;=\; \frac{e^\eta}{1+e^\eta},$$

and then

$$-n\log(1-p) \;=\; n\log(1+e^\eta) \;=\; A(\eta).$$

Thus, the canonical form is

$$f(x;\eta) \;=\; \exp\left[\eta x - (-n\log(1+e^\eta))\right].$$

As p ranges over $(0,1)$, η ranges over $\mathbb{R}$, which has nonempty interior. Note that we have to rule out $p=0$ and $p=1$ when we put this into exponential family type form. Clearly, for $0<p<1$, X is a nondegenerate r.v., so we have a full rank family.

Remarks: The transformation to the natural parameter $p \mapsto \log[p/(1-p)]$ is sometimes referred to as the *logit* or *log-odds*.

(c) Assuming $0<x<1$, we may write the density as

$$f_{\alpha,\beta}(x) \;=\; \exp\left[\alpha\log x + \beta\log(1-x) - \log\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right][x(1-x)]^{-1}.$$

The components of the exponential family are clearly

$$\begin{aligned}
\eta(\alpha,\beta) &= (\alpha,\beta) \;=\; \begin{bmatrix}\alpha\\ \beta\end{bmatrix},\\
T(x) &= (\log x, \log(1-x)),\\
B(\alpha,\beta) &= \log\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},\\
h(x) &= [x(1-x)]^{-1}.
\end{aligned}$$

One could also have chosen to express the density as

$$f_{\alpha,\beta}(x) \;=\; \exp\left[(\alpha-1)\log x + (\beta-1)\log(1-x) - \log\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right],$$

which would have resulted in $\eta=(\alpha-1,\beta-1)$ and $h(x)=1$ (or $h(x)=I_{(0,1)}(x)$). Since we can "throw away" $h(x)$, we generally prefer to make it more complicated and simplify the natural parameter map, but it obviously makes little difference.

Setting $A(\eta) = B(\eta_1, \eta_2)$ and using the dominating measure ν given by

$$d\nu(x) \;=\; [x(1-x)]^{-1} I_{(0,1)}(x) dm(x),$$

we have the canonical form

$$f(x;\eta) \;=\; \exp[\eta' T(x) - A(\eta)].$$

Clearly η ranges over $(0,\infty) \times (0,\infty)$ as α and β range over $(0,\infty)$. Is this the largest possible set of values for η? We will assume that the reader knows from first year calculus that

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx \;=\; \infty, \quad \text{if either } \alpha < 0 \text{ or } \beta < 0.$$

So $(0,\infty) \times (0,\infty)$ is the natural parameter space. This is a nonempty open set, so it has nonempty interior (an open set is equal to its interior). To show that the sufficient statistic $T(X) = (\log X, \log(1-X))$ doesn't satisfy any linear constraint, we will show that the family is identifiable. To this end, we will show that given $(\alpha, \beta) \in (0,\infty) \times (0,\infty)$ and $(a, b) \in (0,\infty) \times (0,\infty)$, with $(\alpha, \beta) \neq (a, b)$, there are only finitely many values of $x \in (0,1)$ where

$$x^{\alpha-1}(1-x)^{\beta-1} \;=\; x^{a-1}(1-x)^{b-1}.$$

Note that if two Beta densities gave the same probability measure, these two expressions would have to be equal for Lebesgue almost all $x \in (0,1)$. By dividing one side of one equation by the other, this is equivalent to showing that given $(a, b) \neq (1, 1)$ there are only finitely many solutions to

$$x^{a-1}(1-x)^{b-1} \;=\; 1.$$

By considering the cases $0 < a < 1$, $1 < a < \infty$, etc., one can easily check that there are only one or two solutions to the last equation for $x \in (0,1)$. Thus, the family is identifiable and based on the remarks, the sufficient statistic $T(X)$ does not satisfy any linear constraints, and hence the family is full rank.

(d) I believe that here the version of $\mathbb{N}$ we are using is the one that includes 0. I think this version of the negative binomial is the total number of "failures" until we observe m "successes" in a sequence of independent Bernoulli trials with "success" probability p. The author does have some

confusion about whether 0 is a natural number or not. Anyway, with this caveat, for x a nonnegative integer, the density (w.r.t. counting measure on the nonnegative integers) can be written as

$$f_p(x) \;=\; \exp\left[x \log(1-p) - m \log(1/p)\right] \left(\begin{array}{c} m+x-1 \\ m-1 \end{array} \right),$$

and the exponential family components are

$$\begin{aligned} \eta(p) &= \log(1-p), \\ T(x) &= x, \\ B(p) &= -m \log p, \\ h(x) &= \left(\begin{array}{c} m+x-1 \\ m-1 \end{array} \right) \end{aligned}$$

Solving for p, we obtain

$$p = 1 - e^{\eta}.$$

Note that as p ranges over $(0, 1)$, η ranges over $(-\infty, 0)$. Note that we can allow $p = 1$ in the Negative Binomial p.m.f. (probability mass function) formula, but we have to delete this value when putting it in exponential family form. Clearly $(-\infty, 0)$ is the natural parameter space since $\eta \geq 0$ leads to $1 - p \geq 1$ and

$$\sum_{x=0}^{\infty} \left(\begin{array}{c} m+x-1 \\ m-1 \end{array} \right) (1-p)^x \;=\; \infty.$$

To see this, note that this is increasing in $q = 1 - p$, and for $q < 1$ we have

$$\sum_{x=0}^{\infty} \left(\begin{array}{c} m+x-1 \\ m-1 \end{array} \right) q^x \;=\; (1-q)^{-m} \;\to\; \infty, \text{ as } q \to 1.$$

So we have the canonical form

$$f(x; \eta) \;=\; \exp[\eta x - A(\eta)],$$

with

$$A(\eta) \;=\; -m \log(1 - e^{\eta}), \quad -\infty < \eta < 0.$$

Now the natural parameter space is a nonempty open set, so has nonempty interior. The sufficient statistic $T(X) = X$ clearly has a non-degenerate

distribution, so does not satisfy any linear constraints, and thus the family is full rank.

Solution to Exercise 2.3.14: The density w.r.t. counting measure on $I\!N^n$ is

$$f_\lambda(\underline{y}) \;=\; \exp\left[-\lambda \sum_{i=1}^n t_i \;+\; \log\lambda \sum_{i=1}^n y_i\right] \prod_{i=1}^n \frac{1}{y_i!}.$$

This is clearly an exponential family where the sufficient statistic is

$$T \;=\; \sum_{i=1}^n Y_i,$$

the natural parameter is

$$\eta(\lambda) \;=\; \log\lambda,$$

and the negative log of the normalizing constant is

$$A(\eta) \;=\; e^\eta \sum_{i=1}^n t_i.$$

Of course the other (more or less irrelevant) factor is

$$h(\underline{y}) \;=\; \prod_{i=1}^n \frac{1}{y_i!}.$$

In canonical form it looks like

$$f_\eta(\underline{y}) \;=\; \exp\left[-e^\eta \sum_{i=1}^n t_i \;+\; \eta \sum_{i=1}^n y_i\right] \prod_{i=1}^n \frac{1}{y_i!}.$$

For part (c), So using Proposition 3.2.2 (b) we compute the m.g.f. for T is

$$\begin{aligned} \psi_\eta(u) &= \exp\left[\exp(\eta+u)\sum_{i=1}^n t_i \;-\; \exp(\eta)\sum_{i=1}^n t_i\right] \\ &= \exp\left[(e^u-1)e^\eta \sum_{i=1}^n t_i\right]. \end{aligned}$$

Re-expressing this in terms of the original parameter λ gives

$$\psi_\lambda(u) \;=\; \exp\left[(e^u-1)\lambda \sum_{i=1}^n t_i\right].$$

Do we recognize this m.g.f.? Well, the m.g.f. of the $Poisson(\theta)$ is

$$\begin{aligned}
&\sum_{k=0}^{\infty} e^{uk} \frac{\theta^k}{k!} e^{-\theta} \\
&\quad = \quad \sum_{k=0}^{\infty} \frac{(e^u\theta)^k}{k!} e^{-\theta} \\
&\quad = \quad \exp\left[e^u\theta\right] e^{-\theta} \\
&\quad = \quad \exp\left[(e^u - 1)\theta\right],
\end{aligned}$$

so we see that $\sum_i Y_i$ has a $Poisson\left(\lambda \sum_{i=1}^n t_i\right)$ distribution.