Clase semana 2 - Análisis cuantitativo para la toma de decisiones

Fernando Cortés Tejada

31/03/2023

Agenda

1. Distribuciones multivariadas
   1.1. Vector aleatorio
   1.2. Distribuciones conjuntas, marginales y condicionales
   1.3. Valor esperado y matrices de varianza y covarianza
   1.4. Correlación

Distribuciones Multivariadas

Ejemplo

Sea $X =$ Nivel de producción semanal en litros de dos marcas A y B de un solvente industrial y $Y =$ Nivel de producción semanal en litros de la marca A del solvente. Si (X, Y) es un vector aleatorio con la siguiente función de densidad conjunta:

$$f_{XY}(x, y) = \begin{cases} Cy^2 e^{-x} & , \quad \text{si } 0 < y < x < \infty \\ 0 & , \quad \text{en caso contrario} \end{cases}$$

a. Halle la constante normalizadora C
b. Halle la probabilidad de que en una semana la producción del solvente A supere el 75% de toda la producción.
c. ¿Son X e Y v.a's independientes?

Sugerencia: utilizar

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

Solución

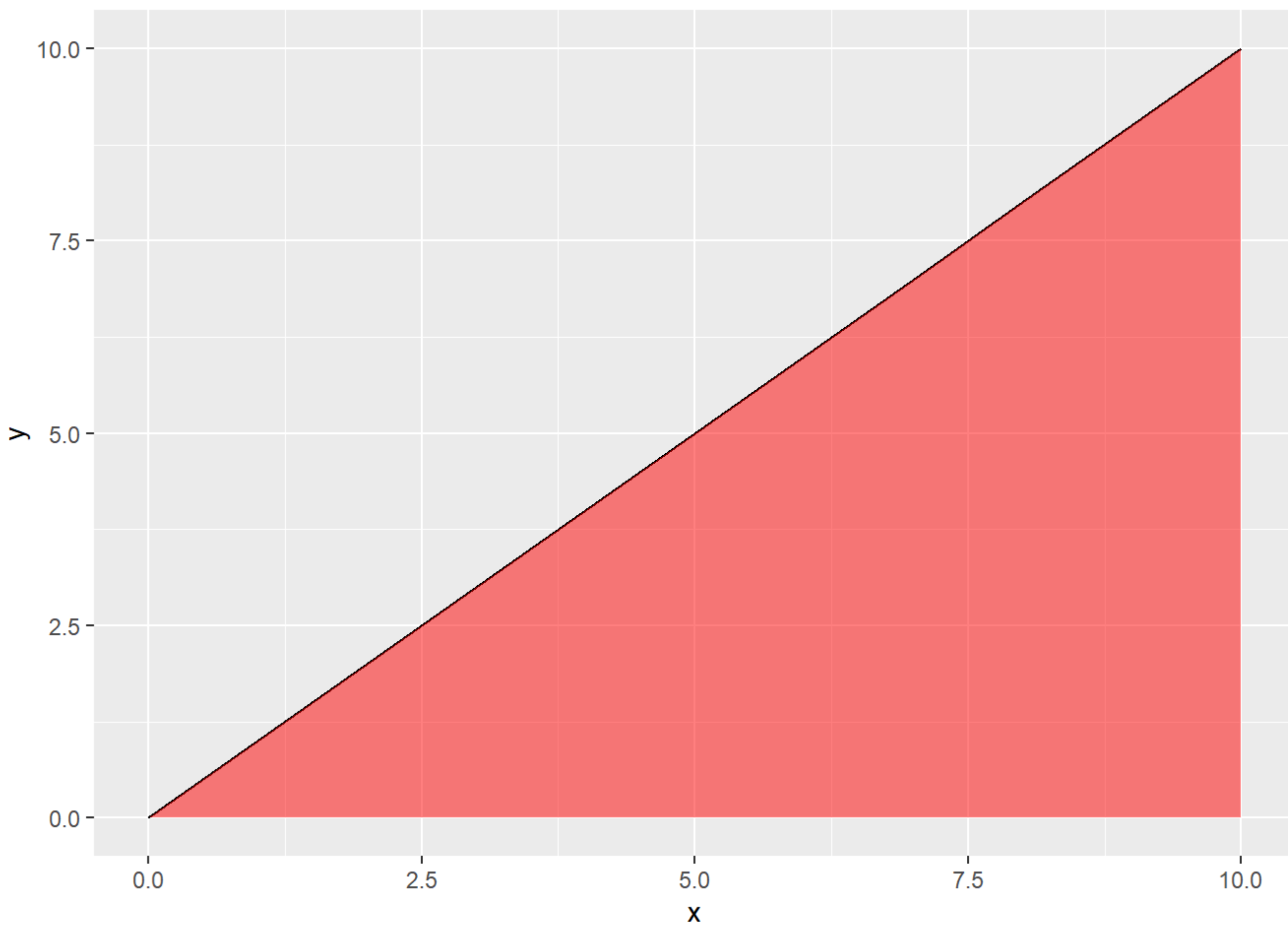
[a] Para hallar C, tomemos en cuenta que

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1$$

entonces

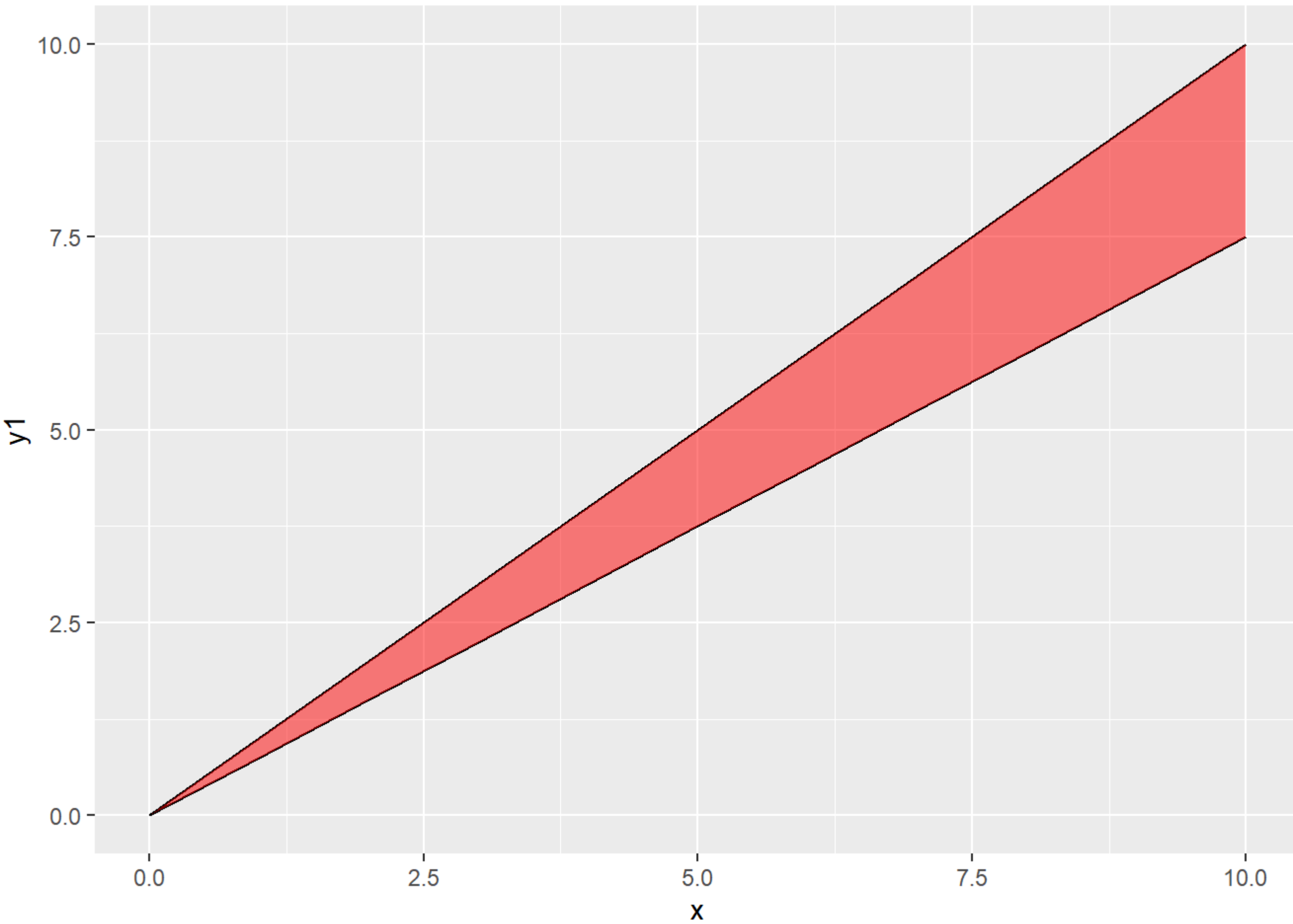
$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy &= C\int_0^{\infty}\left(\int_0^x y^2 e^{-x}dy\right)dx \\
&= C\int_0^{\infty} e^{-x}\left(\int_0^x y^2 dy\right)dx \\
&= C\int_0^{\infty}\frac{x^3}{3}e^{-x}dx \\
&= \frac{C}{3}\Gamma(4) \\
&= 2C = 1 \\
\Rightarrow C &= \frac{1}{2}
\end{aligned}$$

Por tanto, $C = \frac{1}{2}$. Geométricamente, lo de arriba nos dice que el volumen bajo la función f_{XY} sobre su rango es 1. El rango utilizado lo podemos ver en la siguiente imagen:



Es recomendable graficar la región de integración, pues ella nos permitirá establecer fácilmente los límites de integración.

[b] Acá nos piden $P(Y > 0.75X)$. Esto quiere decir que, adicionalmente a la restricción de $0 < y < x < \infty$, tenemos que agregarle la de $y > 0.75x$. La gráfica de esta región (a la que llamaremos A) es



y por tanto, tomando primero un diferencial de y se tiene que

$$
\begin{aligned}
P(Y > 0.75X) &= \iint_A f_{XY}(x, y)dxdy \\
&= \int_0^\infty \frac{e^{-x}}{2}\left(\int_{0.75x}^{x} y^2 dy\right)dx \\
&= \int_0^\infty \frac{e^{-x}}{2}\left(\frac{x^3}{3} - \frac{9x^3}{64}\right)dx \\
&= \frac{37}{384}\int_0^\infty x^3 e^{-x}dx \\
&= \frac{37}{384}\Gamma(4) = 0.578125
\end{aligned}
$$

Por lo que la probabilidad de que el solvente A supere el 75% de toda la producción es 57.8%.

[c] Para saber si X y Y son independientes, debemos hallar las distribuciones marginales de cada una. Estas vienen dadas por

$$\begin{aligned} f_X(x) &= \int_0^x \frac{1}{2} y^2 e^{-x} dy \\ &= \frac{x^3}{6} e^{-x}, \qquad \forall x > 0, \end{aligned}$$

y

$$\begin{aligned} f_Y(y) &= \int_y^\infty \frac{1}{2} y^2 e^{-x} dx \\ &= \frac{1}{2} y^2 e^{-y}, \qquad \forall y > 0. \end{aligned}$$

Entonces, podemos tomar un valor cualquiera, como $(x, y) = (2, 1)$ por ejemplo, lo que nos da para la distribución conjunta

$$f_{XY}(2, 1) = \frac{1}{2}(1)^2 e^{-2} = \frac{1}{2} e^{-2}$$

y para la multiplicación de marginales

$$f_X(2) f_Y(1) = \frac{2^3}{6} e^{-2} \times \frac{1}{2}(1)^2 e^{-1} = \frac{2}{3} e^{-3},$$

por lo tanto X y Y no son variables aleatorias independientes.

Valor esperado y matrices de varianza y covarianza

Si transformamos un vector aleatorio $\mathbf{X}$ en una variable aleatoria $g(\mathbf{X})$, a través de una función $g : \mathrm{R}^k \to \mathrm{R}$, se define el valor esperado de la variable aleatoria $g(\mathbf{X})$ mediante

$$E(g(\mathbf{X})) = \begin{cases} \sum_{\mathbf{x}} g(\mathbf{x}) P_{X_1 X_2 \ldots X_k}(\mathbf{x}) \\ \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} g(\mathbf{x}) f_{X_1 X_2 \ldots X_k}(\mathbf{x}) dx_1 dx_2 \ldots dx_k \end{cases}$$

dependiendo, respectivamente si $\mathbf{X}$ es un vector aleatorio discreto o continuo.

Esta noción de valor esperado, puede también extenderse a funciones que toman valores en R^k e incluso matrices, dando pie a dos de las más distintivas características de un vector aleatorio (que ahora escribiremos como un vector columna): su vector de medias y su matriz de varianzas-covarianzas. Ellas están definidas por

$$\boldsymbol{\mu} = E(\mathbf{X}) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_k \end{bmatrix} = [\mu_1, \mu_2, \ldots, \mu_k]^\top$$

y

$$\boldsymbol{\Sigma} = E\left((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top\right) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k} & \sigma_{2k} & \cdots & \sigma_k^2 \end{bmatrix},$$

respectivamente, donde $\mu_i = E(X_i)$, la diagonal de $\boldsymbol{\Sigma}$ contiene a las varianzas $\sigma_i^2 = V(X_i)$ de las variables aleatorias componentes y las entradas (i, j) no diagonales denotan a la covarianza entre la variable X_i y X_j.

La covarianza σ_{ij}, que también acostumbraremos denotarla por $Cov(X_i, X_j)$, es una medida de asociación lineal entre estas variables aleatorias componentes y se define por

$$\sigma_{ij} = E((X_i - \mu_i)(X_j - \mu_j)) = E(X_i X_j) - \mu_i \mu_j$$

Correlación

Una medida similar a la covarianza es la correlación de Pearson, la cual tiene la ventaja de ser adimensional y acotada. Ella, que la denotaremos por ρ_{ij} o $Cor(X_i, X_j)$, se define por

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

y tiene las siguientes propiedades importantes:

Proposición

- Si X_i y X_j son variables independientes, entonces $\rho_{ij} = 0$.
- $|\rho_{ij}| \leq 1$.
- $|\rho_{ij}| = 1 \Leftrightarrow P(X_j = a + bX_i) = 1$, donde $a = \mu_j - b\mu_i \quad \text{y} \quad b = \frac{\sigma_{ij}}{\sigma_i^2}$.

las cuales nos dicen que mientras más cercana este la correlación a 1 (o -1), más lineal y directamente (inversamente) estarán relacionadas las variables X_i y X_j.

Al igual que la matriz de varianza-covarianza podemos también definir la matriz de correlaciones de $\mathbf{X}$ mediante

$$Cor(\mathbf{X}) = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{12} & 1 & \cdots & \rho_{2k} \\ \cdots & \cdots & \ddots & \cdots \\ \rho_{1k} & \rho_{2k} & \cdots & 1 \end{bmatrix}.$$

La proposición siguiente nos indica como calcular el vector de medias y la matriz de varianzas-covarianzas de cualquier transformación multilineal de un vector aleatorio.

Proposición

Sea $\mathbf{X}$ un vector aleatorio (columna) k-dimensional con vector de medias $\boldsymbol{\mu}$ y matriz de varianzas-covarianzas $\boldsymbol{\Sigma}$, $\mathbf{A}$ una matriz $m \times k$ de constantes y $\mathbf{b}$ un vector $m \times 1$ no aleatorio. Si definimos el vector aleatorio $m -$ dimensional $\mathbf{Y}$, mediante la transformación $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$, entonces el vector de medias y la matriz de varianzas-covarianzas de $\mathbf{Y}$ vienen dadas respectivamente por $\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ y $\Sigma_{\mathbf{Y}} = \mathbf{A\Sigma A}^\top$.

Ejemplo

Volvamos al ejemplo del control de calidad de las cajas. Donde se producían 24 cajas, con probabilidad 0.1 de que cada una resulte defectuosa, y se tomaba un control de 6 cajas.

Bajo dicho contexto, ahora tenemos que

- El costo unitario de producción es de 8 soles y el de venta 12 soles
- Cada inspección cuesta 0.5 soles
- Toda unidad defectuosa encontrada en el control es reemplazada
- Toda unidad defectuosa que un consumidor encuentre es reemplazada y se indenmiza al consumidor con 3 soles por unidad.

Halle la utilidad esperada y la desviación estándar de la utilidad.

Solución

La utilidad por la venta de una caja viene dada por

$$\begin{aligned} U(X_1, X_2) = U &= 24 \times 12 - 24 \times 8 - 6 \times 0.5 - 8X_2 - 11(X_1 - X_2) \\ &= 93 - 11X_1 + 3X_2 \\ &= [-11, 3]\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + 93. \end{aligned}$$

Así, como las distribuciones marginales de X_1 y X_2 son respectivamente, $X_1 \sim B(24, 0.1)$ y $X_2 \sim B(6, 0.1)$, la media y varianza de este costo vendrá dada por

$$E(U) = [-11, 3]\begin{bmatrix} 24 \times 0.1 \\ 6 \times 0.1 \end{bmatrix} + 93 = 68.4$$

y

$$V(U) = [-11, 3]\begin{bmatrix} 24 \times 0.1 \times (1 - 0.1) & \sigma_{12} \\ \sigma_{12} & 6 \times 0.1 \times (1 - 0.1) \end{bmatrix}\begin{bmatrix} -11 \\ 3 \end{bmatrix},$$

donde en la covarianza $\sigma_{12} = E(X_1X_2) - E(X_1)E(X_2)$, entre X_1 y X_2, nos falta hallar el término $E(X_1X_2)$, el cual se puede hallar en R mediante

```
N <- 24
n <- 6
p <- 0.1

Pxy <- function(x,y) {
  dhyper(y, x, N-x, n)*dbinom(x, N, p)
}

E_x1x2 <- 0
for(x in 0:24) {
  for(y in 0:min(x, 6)) {
    E_x1x2 <- E_x1x2 + (x*y)*Pxy(x, y)
  }
}

(sigma12 <- E_x1x2 - 2.4*0.6)
```

```
## [1] 0.54
```

Así,

$$V(U) = [-11, 33]\begin{bmatrix} 2.16 & 0.54 \\ 0.54 & 0.54 \end{bmatrix}\begin{bmatrix} -11 \\ 3 \end{bmatrix} = 230.58$$

y la desviación estándar de los costos será 15.18 soles.

Ejemplo

Volvamos al ejemplo de la producción semanal de solvente industrial A y B, donde teníamos la distribución conjunta

$$f_{X_1X_2}(x,y) = \begin{cases} \frac{1}{2}y^2e^{-x} & , \quad \text{si } 0 < y < x < \infty \\ 0 & , \quad \text{en caso contrario} \end{cases}$$

y las distribuciones marginales

$$f_{X_1}(x) = \frac{x^3}{6}e^{-x}, \qquad \forall x > 0,$$

y

$$f_{X_2}(y) = \frac{1}{2}y^2e^{-y}, \qquad \forall y > 0.$$

Nos piden

a. Halle e interprete la correlación entre X_1 y X_2.

b. Obtenga la función $m(x) = E(X_2 \mid X_1 = x)$, conocida como la regresión de X_2 sobre X_1. Interprete.

c. Halle la proporción esperada de la producción total del solvente que es destinada a la marca A.

Solución

[a] La correlación está definida por

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i\sigma_j}$$

por lo tanto, necesitamos hallar la covarianza σ_{12} y las desviaciones estándar σ_1 y σ_2. Las formas de las ecuaciones que utilizaremos, por simplicidad serán:

$$\sigma_{12} = E(X_1X_2) - E(X_1)E(X_2)$$

y

$$\sigma_i^2 = E(X_i^2) - E(X_i)^2.$$

Nos damos cuenta que tenemos que hallar los términos: $E(X_1)$, $E(X_2)$, $E(X_1X_2)$, $E(X_1^2)$ y $E(X_2^2)$ para poder construir la correlación entre X_1 y X_2

$$\rho_{12} = \frac{E(X_1X_2) - E(X_1)E(X_2)}{\sqrt{E(X_1^2) - E(X_1)^2}\sqrt{E(X_2^2) - E(X_2)^2}}.$$

Comencemos con $E(X_1)$

$$
\begin{aligned}
E(X_1) &= \int_0^\infty x f_{X_1}(x) dx \\
&= \int_0^\infty \frac{x^4}{6} e^{-x} dx \\
&= \frac{1}{6}\Gamma(5) = 4.
\end{aligned}
$$

Seguimos con $E(X_2)$

$$
\begin{aligned}
E(X_2) &= \int_0^\infty y f_{X_2}(y) dy \\
&= \int_0^\infty \frac{1}{2} y^3 e^{-y} dy \\
&= \frac{1}{2}\Gamma(4) = 3.
\end{aligned}
$$

Ahora $E(X_1 X_2)$

$$
\begin{aligned}
E(X_1 X_2) &= \int_0^\infty \int_0^x xy f_{X_1 X_2}(x, y) dy dx \\
&= \int_0^\infty \left(\int_0^x x \frac{y^3}{2} e^{-x} dy \right) dx \\
&= \frac{1}{2} \int_0^\infty x e^{-x} \left(\int_0^x y^3 dy \right) dx \\
&= \frac{1}{8} \int_0^\infty x^5 e^{-x} dx \\
&= \frac{1}{8}\Gamma(6) = 15.
\end{aligned}
$$

Seguimos con $E(X_1^2)$

$$
\begin{aligned}
E(X_1^2) &= \int_0^\infty x^2 f_{X_1}(x) dx \\
&= \int_0^\infty \frac{x^5}{6} e^{-x} dx \\
&= \frac{1}{6}\Gamma(6) = 20.
\end{aligned}
$$

Finalmente para $E(X_2^2)$

$$
\begin{aligned}
E(X_2^2) &= \int_0^\infty y^2 f_{X_2}(y) dy \\
&= \int_0^\infty \frac{1}{2} y^4 e^{-y} dy \\
&= \frac{1}{2}\Gamma(5) = 12.
\end{aligned}
$$

Por lo tanto, la correlación entre X_1 y X_2 viene dada por

$$\rho_{12} = \frac{15 - 4 \times 3}{\sqrt{20 - 4^2}\sqrt{12 - 3^2}} = 0.8660254$$

y se tiene una asociación fuerte y directa entre X_1 y X_2.

[b] La regresión de X_2 sobre X_1 viene dada por

$$m(x) = E(X_2 \mid X_1 = x) = \int_0^x y f_{X_2 \mid X_1 = x}(y) dy,$$

la cual se lee como “el valor esperado que tomará X_2 dado que sabemos que X_1 tomó el valor de x”.

Ahora, la función de densidad condicional de la anterior ecuación la podemos obtener de la siguiente forma:

$$\begin{aligned} f_{X_2 \mid X_1 = x}(y) &= \frac{f_{X_1, X_2}(x, y)}{f_{X_1}(x)} \\ &= \frac{\frac{1}{2} y^2 e^{-x}}{\frac{1}{6} x^3 e^{-x}} \\ &= 3 \frac{y^2}{x^3} \end{aligned}$$

la cual representa la distribución de X_2 cuando X_1 toma el valor x.

Así,

$$m(x) = \int_0^x 3 \frac{y^3}{x^3} dy = \frac{3}{4} x$$

lo que nos dice que el nivel de producción de la marca A se incrementará en 0.75 litros por cada incremento de un litro que experimente la producción total del solvente.

[c] Se nos pide

$$\begin{aligned} E\left(\frac{X_2}{X_1}\right) &= \int_0^\infty \frac{y}{x} \frac{y^2}{2} e^{-x} dy dx \\ &= \int_0^\infty \frac{e^{-x}}{2x} \left(\int_0^x y^3 dy\right) dx \\ &= \int_0^\infty \frac{e^{-x}}{2x} \frac{x^4}{4} dx \\ &= \frac{\Gamma(4)}{8} = 0.75, \end{aligned}$$

es decir, se espera que el 75% de la producción del solvente se destine a la marca A.