Centro de Investigación en Matemáticas (CIMAT).
Unidad Monterrey
Inferencia Estadística

Tarea 3 (Ejercicios en clase)

Gustavo Hernández Angeles

5 de octubre de 2024

1 Ejercicio 176

a) Crea una columna de 100 valores de una Unif(0,1) en el software de tu preferencia.

b) Construye una columna con la fórmula

$$x = -2\log(1-y) \tag{1.1}$$

c) Construye el histograma de esta nueva columna y concluye.

$$y = 1 - e^{-x/2} \Rightarrow x = -2\log(1-y)$$

SOLUCIÓN

Para el inciso a) utilicé la función `runif(size, min = 0, max = 1)` para obtener una muestra de Unif(0,1) y guardando estos datos en una variable y. Para el inciso b) simplemente utilicé la línea `x <- -2*log(1-y)`, ayudandome de la vectorización de `R`. Finalmente, grafiqué las frecuencias normalizadas mediante `ggplot` usando `geom_histogram()` junto a la función de densidad de probabilidad correspondiente (exponencial con parámetro $\theta = 2$) para visualizar la aproximación. El resultado se muestra a continuación.

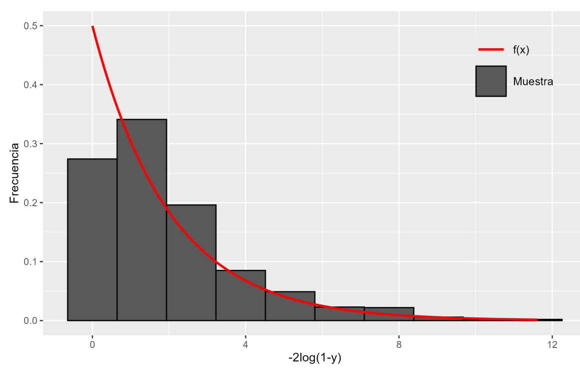


Figura 1.1: Frecuencias normalizadas de la muestra al aplicar la transformación $-2\log(1-y)$, junto a la función de densidad $f(x) = \frac{1}{2}e^{-\frac{1}{2}x}$. En este caso, el tamaño de la muestra es de 1000.

Podemos ver en la figura 1.1 que al aplicar la determinada transformación a una muestra de la distribución Unif(0,1) se puede aproximar a la distribución exponencial con parámetro $\theta = 2$. En el código se puede cambiar a voluntad el tamaño de la muestra, haciendo más visible la aproximación cuando n es muy grande. Por lo tanto, la transformación propuesta es correcta.

$$y = 1 - e^{-x/2} \Rightarrow x = -2\log(1-y)$$

2 Ejercicio 204

Se han realizado ciertas pruebas de resistencia en ladrillos obteniéndose las mediciones que a continuación se muestran, agrupadas en una tabla de frecuencias.

Interv.	De	Hasta	Frecuencia	Frec. relativa (%)
1	28.70	32.65	5	5.56
2	32.65	36.60	6	6.67
3	36.60	40.55	11	12.22
4	40.55	44.50	17	18.89
5	44.50	48.45	19	21.11
6	48.45	52.40	19	21.11
7	52.40	56.35	7	7.78
8	56.35	60.30	2	2.22
9	60.30	64.25	3	3.33
10	64.25	68.20	1	1.11

a) Traza el histograma.

b) Calcula la probabilidad de cada intervalo de clase de la tabla de frecuencias, asumiendo que las resistencias siguen una distribución normal con media 45.47 y varianza 58.19.

c) Compara las frecuencias relativas con las probabilidades bajo normalidad ¿Qué se puede concluir? Esta es la idea base de una prueba de Bondad de Ajuste conocida como χ^2 de Pearson. Si la media y varianza de la normal no se conocen de antemano, podemos usar los valores muestrales correspondientes como una aproximación de los mismos. A esto le llamamos estimación de parámetros y ya hablaremos más adelante de las cualidades de los estimadores.

SOLUCIÓN

Inciso a)

Para trazar el histograma utilicé `geom_col()` de `ggplot2`. Las frecuencias se grafican contra una nueva columna que sirve como el punto medio en el intervalo para la visualización. El resultado está en la figura 2.1.

Inciso b)

Utilicé la función implementada en `R`, `pnorm(x,mean,sd)`, para calcular las probabilidades en cada intervalo dando como entrada los valores propuestos en el problema para la media y la varianza. De esta forma obtenemos una nueva columna para los datos con la probabilidad correspondiente.

Inciso c)

Para comparar las frecuencias relativas con las probabilidades en la distribución normal, decidí realizar una visualización presentada en la figura 2.2. Se puede observar que la distribución

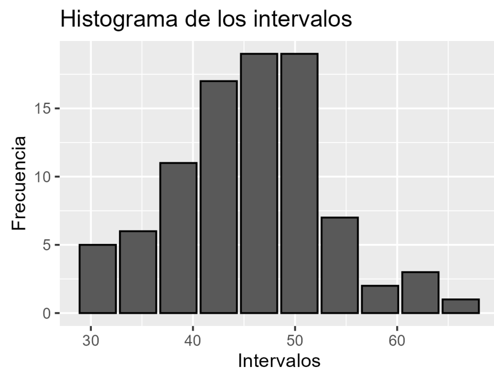


Figura 2.1: Histograma de los datos dados.

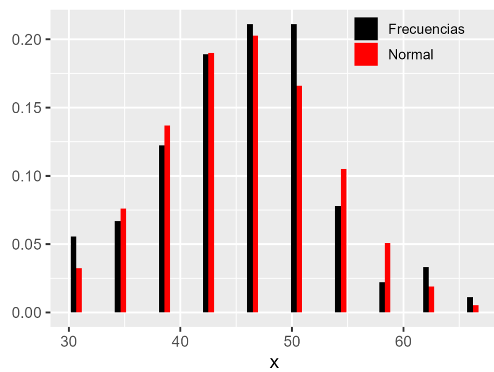


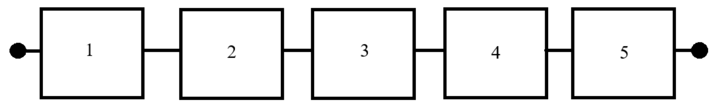
Figura 2.2: Comparación de las frecuencias relativas y las probabilidades bajo normalidad.

de los datos siguen de manera muy ligera a la distribución normal, siendo más comparables alrededor de la media. Mayor cantidad de datos, junto a intervalos más cortos, pueden ayudar a verificar la normalidad mediante visualización más facilmente.

¿

3 Ejercicio 244

Consideremos un sistema formado por 5 componentes idénticos conectados en serie tal como se muestra a continuación.



Tan pronto como un componente falla, el sistema completo falla. Supongamos que cada componente sigue un modelo de tiempo a la falla exponencial con $\theta = 100$, y que lo componentes fallan en forma independiente una de la otra. Definamos los eventos

$$A_i = \{i\text{-ésimo componente dura al menos } t \text{ horas}\}, \quad i = 1, 2, 3, 4, 5$$

Las A_i's son independientes e identicamente distribuidas. Sea X el tiempo de la falla del sistema.

a) El evento $\{X \geq t\}$, ¿a cuál evento, en términos de las A_i's, es equivalente?

b) Usando la independencia de las A_i's, calcula $P(X \geq t)$. Obtén $F(t) = P(X \leq t)$. ¿Cuál es la distribución de X?

c) Si en lugar de 5 componentes tenemos n, ¿cuál es la distribución de X?

SOLUCIÓN

Inciso a)

Si $X =$ *tiempo al que falla el sistema*, entonces para $X \geq t$ requerimos que todos los componentes duren al menos t horas, es decir, todos los eventos A_i se deben cumplir. Por lo tanto, el evento $X \geq t$ es la intersección de todos los eventos A_i.

$$\{X \geq t\} = \{A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5\} \tag{3.1}$$

Inciso b)

Asumiremos ahora que los eventos son mutuamente independientes, es decir, cada evento será independiente de cualquier intersección que se pueda formar de los otros eventos. Sea $X_i =$ *tiempo al que el i-esimo componente falla* y sabiendo que $X_i \sim$ Exponencial$(\theta = 100)$, calculemos ahora la probabilidad del evento $X \geq t$.

$$\begin{aligned}
P(X \geq t) &= P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) \text{ ; eq 3.1} \\
&= P(A_1)P(A_2)P(A_3)P(A_4)P(A_5) \text{ ; Con la independencia mutua} \\
&= P(X_1 \geq t)P(X_2 \geq t)P(X_3 \geq t)P(X_4 \geq t)P(X_5 \geq t) \\
&= \Big[\int_t^\infty \frac{1}{\theta} e^{-x/\theta} dx\Big]^5 = \Big[e^{-t/\theta}\Big] \\
&= e^{-t/(\theta/5)}
\end{aligned}$$

Obteniendo $F(t) = P(X \leq t)$, podemos obtener la función de masa de la distribución de X y podremos identificar su distribución:

$$F(t) = P(X \leq t) = 1 - P(X \geq t) = 1 - e^{-t/(\theta/5)}$$

Derivando $F(t)$

$$\frac{d}{dt} F(t) = f(t) = \frac{1}{(\theta/5)} e^{-t/(\theta/5)} \tag{3.2}$$

Por lo tanto, la distribución de X es una Exponencial con parámetro $\theta' = \theta/5$. En este caso particular, donde $\theta = 100$, $\theta' = 20$.

$$\boxed{\therefore X \sim \text{Exponencial}(20)}$$

Inciso c)

Siguiendo el mismo desarrollo que hicimos desde 3.1 hasta 3.2, con n eventos A_i y n variables aleatorias exponenciales e iid X_i, vemos que podemos simplemente reemplazar el divisor 5 por n en 3.2. Y por lo tanto, con $\theta = 100$:

$$\boxed{\therefore X \sim \text{Exponencial}(\theta = 100/n)}$$

4 Ejercicio 285

- En el ejemplo anterior usa la relación con la Binomial para hacer el cálculo de la probabilidad en b).
- En muchos proyectos se emplea un método llamado PERT para coordinar varias actividades en proyectos grandes y/o complejos (por ejemplo, fue empleado en la construcción de los Apolo). Un supuesto estándar de esta técnica es que el tiempo necesario para completar cualquier actividad una vez que ha comenzado, se puede modelar mediante una distribución Beta con A = tiempo optimista (si todo va bien) y B = tiempo pesimista (si todo va mal).

Por ejemplo, supongamos que estamos construyendo casas habitación y que el tiempo necesario, en días, para poner los cimientos de una casa sigue una distribución Beta con $A = 2$ y $B = 5$. Además, se puede calcular que $\alpha = 2$ y $\beta = 3$. Calcula la probabilidad de que el tiempo para terminar los cimientos sea menor a 3 días.

SOLUCIÓN

Primer punto.

En el ejemplo anterior se utiliza una variable aleatoria X con distribución Beta$(\alpha = 1, \beta = 4)$. Recordamos la relación entre la v.a. $X \sim$ Binomial(n, p) y la v.a. $Y \sim$ Beta$(\alpha = k, \beta = n - (k-1))$:

$$P(Y > p) = P(X \leq k - 1) \tag{4.1}$$

Teniendo que $\alpha = k = 1$, $\beta = n - (1 - 1) = n = 4$, y $p = 0.25$ para nuestro caso. Entonces

$$\begin{aligned} P(Y > 0.25) &= P(X \leq 0) = P(X = 0) \\ &= \binom{4}{0} (0.25)^0 (0.75)^4 = 0.316 \end{aligned}$$

El cual coincide con el resultado obtenido en clase.

Segundo punto.

Sea X la variable aleatoria con distribución Beta(α, β) definida en el intervalo $[A = 2, B = 5]$ y sea Y la variable aleatoria con distribución Beta(α, β) definida en el intervalo $[0, 1]$. Entonces la relación que existe entre ellas (el escalamiento en un intervalo) está dado por:

$$Y = \frac{X - A}{B - A} \tag{4.2}$$

Entonces

$$\begin{aligned}
P(X \leq 3) &= P(Y \leq \frac{3-2}{5-2}) = P(Y \leq \frac{1}{3}) \\
&= \int_0^{1/3} \frac{\Gamma(2+3)}{\Gamma(2)\Gamma(3)} x^{2-1}(1-x)^{3-1} dx \\
&= \int_0^{1/3} \frac{4!}{1!2!} x(1-x)^2 dx = 12\int_0^{1/3}(x-2x^2+x^3)dx \\
&= 12(\frac{1}{2}x^2 - \frac{2}{3}x^3 + \frac{1}{4}x^4)\Big|_0^{1/3} \\
&= 12(\frac{1}{2}(1/3)^2 - \frac{2}{3}(1/3)^3 + \frac{1}{4}(1/3)^4) \\
&= 0.407
\end{aligned}$$

$\boxed{\therefore \text{ La probabilidad de que se terminen los cimientos antes de 3 dias es de } 40.7\,\%}$

5 Ejercicio 306

1. Graficar la función *Weibull* para los siguientes valores de los parámetros:

$$\alpha = 1, \beta = 2$$

$$\alpha = 2, \beta = 2$$

$$\alpha = 3, \beta = 4$$

   Algebraicamente, o bien generando variables con ésta distribución y construyendo un histograma.

2. Grafica las funciones de confiabilidad y riesgo para cada uno de los casos anteriores.
3. Si $h(t) = a + bt$, ¿cuál es la densidad asociada?
4. Estudiar las funciones de riesgo para la distribución *Normal, Gamma y Lognormal*. ¿Cuál es la principal dificultad en estos casos?

$\boxed{\textbf{SOLUCIÓN}}$

Inciso a)

Realizaremos las graficas mediante el software `R` y con la librería `ggplot2`. Cabe destacar que la función de densidad de probabilidad Weibull difiere con la mostrada en clase en cuestión de los parámetros, ya que se definen de distinta forma; el parámetro `shape` $= \beta$ y el parámetro `scale` $= \alpha^{-1/\beta}$. Teniendo en cuenta esto, podemos utilizar las funciones integradas de `R` para graficar la densidad de probabilidad, cuyo resultado se muestra en la figura 5.1.

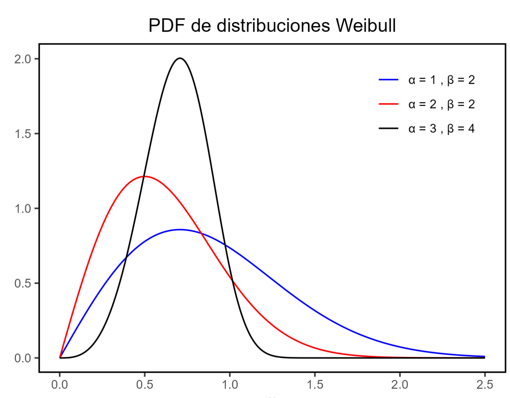


Figura 5.1: Funciones de densidad de probabilidad de la distribución Weibull con distintos parámetros α, β.

Inciso b)

De igual forma, utilizamos `R` para realizar las gráficas. Esta vez, para graficar la función de confiabilidad $R(t)$ se hace uso de su definición: $R(t) = 1 - F(t)$, obteniendo el valor de $F(t)$ con ayuda de la función `pweibull()`. Se muestra el resultado en la figura 5.2.

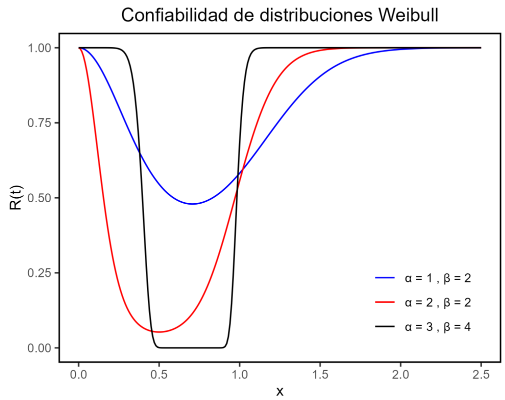
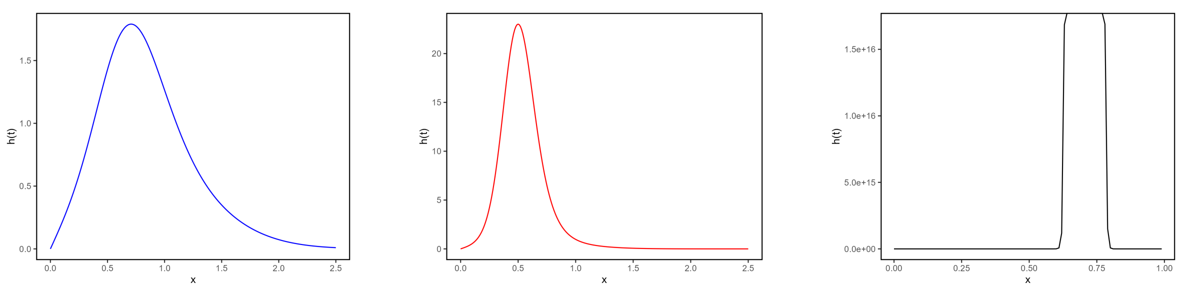


Figura 5.2: Funciones de confiabilidad $R(t)$ de la distribución Weibull con distintos parámetros α, β.

Para graficar las funciones de riesgo, utilizaremos los datos que ya hemos generado en los incisos anteriores, puesto que por definición la función de riesgo $h(t) = f(t)/R(t)$. El resultado se muestra en la figura 5.3.



(a) $\alpha = 1, \beta = 2$ (b) $\alpha = 2, \beta = 2$ (c) $\alpha = 3, \beta = 4$

Figura 5.3: Gráficas de las funciones de riesgo de la distribución Weibull para distintos valores de los parámetros α y β.

Cabe destacar que, debido a las grandes variaciones entre cada gráfica, decidí separarla en tres gráficas distintas. Se puede observar que el conjunto de parámetros $\alpha = 3, \beta = 4$ genera valores $h(t) \sim \infty$ justo en donde $R(t) \sim 0$ (visible en la gráfica 5.2) debido a la misma definición de h como inversamente proporcional a R.

Inciso c)

Para hallar la densidad asociada podemos hallar primer la confiabilidad $R(t)$ para luego hallar la función de distribución $F(t)$, Teniendo la información de que $h(t) = a + bt$, podemos aplicar la ecuación vista en clase:

$$\begin{aligned} R(t) &= Ce^{-\int h(t)dt} = Ce^{-\int (a+bt)dt} \\ &= Ce^{-(at+bt^2/2)} \end{aligned}$$

Y utilizando la condición inicial: $R(0) = 1$

$$R(0) = Ce^0 = 1$$

Y entonces $C = 1$. Por lo tanto seguimos con $F(t) = 1 - R(t)$

$$F(t) = 1 - e^{-(at+bt^2/2)}$$

Y podemos derivar para determinar la función de densidad de probabilidad $f(t)$.

$$F'(t) = f(t) = \frac{d}{dt}(1 - e^{-(at+bt^2/2)}) = (a + bt)e^{-(at+bt^2/2)} \tag{5.1}$$

$$\boxed{\therefore f(t) = (a + bt)e^{-(at+bt^2/2)}}$$

Inciso d)

Las formas matemáticas de las funciones de riesgo $h(t)$ para cada distribución mencionada (Normal, Lognormal y Gamma) contienen dentro una función que solo es evaluable computacionalmente. En las funciones de riesgo para las distribuciones normal y lognormal interviene la función de error $erf(x)$, mientras que para la distribución Gamma interviene la función gamma incompleta $\gamma(k, x)$. Por estas razones, las funciones de riesgo para estas distribuciones puede ser algo complicado de interpretar viendo únicamente su forma matemática, para la distribución Gamma puede ser incluso complicado de interpretar visualizando la función, debido a que depende fuertemente de los parámetros de la distribución.