Inferencia Estadística | Tarea 03 (Ejercicios Notas)

Aguirre Calzadilla César Miguel

04 de octubre de 2024

Códigos

Todo el código escrito para esta tarea será anexado en un archivo de RStudio. Dentro se encuentran las rutinas escritas para la tarea así como comentarios sobre las mismas.

Ejercicio p. 176

Crea una columna con 100 valores de una Unif(0,1) en el software de tu preferencia. Construye otra columan con la fórmula $x = -2log(1 - y)$. Construye el histograma de esta nueva columna y concluye.

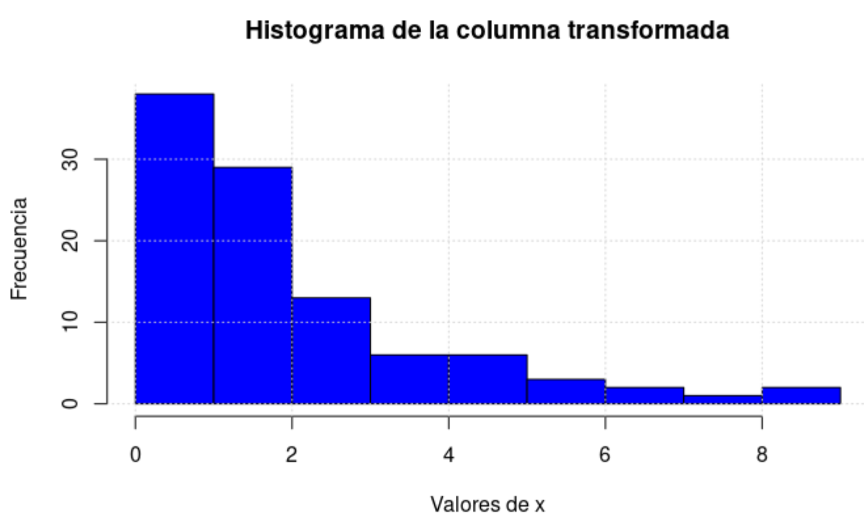


Figura 1: Caption

Podemos notar que el rango de x se concentra en valores pequeños, mientras que para valores de y cercanos a 1, la transformación produce valores grandes para x. Además, la frecuencia es alta para valores de x cercanos a 0, algo que va disminuyendo mientras x aumenta. Esto es característico de transformaciones logarítmicas.

De hecho, el perfil de este histograma podría parecer lo que se espera para una variable que sigue una distribución exponencial.

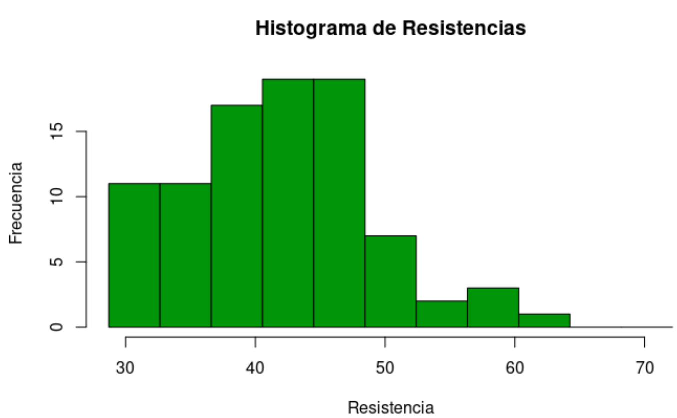
Ejercicio p. 204

Se han realizado ciertas pruebas de resistencia en ladrillos obteniéndose las mediciones que a continuación se muestran, agrupadas en una tabla de frecuencias (la tabla se muestra al final del enunciado). Lo primero que tendrás que hacer será trazar el histograma de los datos. Después, calcula la probabilidad de que 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.

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

Cuadro 1: Tabla de frecuencias y frecuencias relativas.

El histograma asociado a estos datos es el siguiente.



A primera vista, no parece haber confirmación absoluta sobre si los dato siguen o no una distribución normal. Tienes cierto perfil parecido, pero no es nada concluyente. Para ello, se nos pide realizar una prueba χ^2 de Bondad de Ajuste, pero antes de eso, debemos calcular la probabilidad de cada intervalo, asumiendo que las resistencias siguen una distribución normal con $\mu = 45.47$ y $\sigma^2 = 58.19$.

Entonces, dado que se tienen los datos muestrales, podemos calcular probabilidades para cada intervalo, de tal modo que asumieremos que los datos siguen una $Normal(45.47, 58.19)$. Hay que recordar que la desviación estándar es:

$$\sigma = \sqrt{58.19} = 7.63$$

Para cada intervalo, utilizaremos la CDF:

$$P(a \leq x \leq b) = F(b) - F(a)$$

Con $F(x)$, la función de distribución acumulada de una normal con $\mu = 45.47$ y $\sigma = 7.63$. Para esto, podemos utilizar una tabla de valores de la normal o software, por medio de RStudio calculamos los valores de en cada intervalo.

Sin embargo, no está de más recordar cómo es el cómputo para el cálculo de probabilidades por intervalo, definiendo límites a_i y b_i de tal modo que construyamos intervalos utilizando la CDF de la normal de manera directa, la cuál está definida como:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp(-\frac{(t-\mu)^2}{2\sigma^2})\ dt$$

Entonces, la probabilidad en el intervalo $[a_i, b_i]$ quedaría como:

$$P(a_i \leq x \leq b_i) = F(b_i) - F(a_i) \quad \forall i = 1, 2, ..., n$$

Los valores de $F(b_i)$ y $F(a_i)$ se pueden obtener en R mediante la función *pnorm*, que calcula la CDF de la normal estándar para un valor dado. Luego, la probabildiad para cada intervalo es:

$$Prob.[1_i, b_i] = \Phi\Big(\frac{b_i - \mu}{\sigma}\Big) - \Phi\Big(\frac{a_i - \mu}{\sigma}\Big)$$

Con Φ como la CDF de la normal estándar. Utilizando R, al final nos queda la sigueinte tabla, la cual muestra la probabilidad de cada intervalo solicitado.

Intervalo	Probabilidad
28.70 -> 32.65	0.032460595
32.65 -> 36.6	0.076037850
36.60 -> 40.55	0.137014274
40.55 -> 44.5	0.189934590
44.50 -> 48.45	0.202566209
48.45 -> 52.4	0.166210740
52.40 -> 56.35	0.104922022
56.35 -> 60.3	0.050951668
60.30 -> 64.25	0.019032086
64.25 -> 68.2	0.005467484

Cuadro 2: Probabilidades por intervalo.

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 a los mismos. A esto le llamamos estimación de parámetros y ya hablaremos más adelante de las cualidades de los estimadores.

En este caso, se nos pide realizar una prueba de "Ji-cuadrada". Esta prueba se usa para determinar si existe una diferencia significativa entre las frecuencias observadas y las frecuencias esperadas en una o más categorías. Es ampliamente utilizada para evaluar so los datos observados se ajustan a una distribución teórica esperada, siendo esto una Prueba de Bondad de Ajuste. A pesar de ello, se puede utilizar bajo otros contextos, como pruebas de independencia y homogeneidad.

En este caso, la prueba de bondad de ajuste comparan el perfil de los datos observados con la distribución teórica esperada para ver si las diferencias entre ambas son significativas o si pueden ser atribuidas al azar. Se comparan las frecuencias observadas con las frecuencias esperadas, bajo la hipótesis de que los datos siguen la distribución seleccionada.

La idea básica es revisar que las desviaciones entre lo observado y lo esperado sean losuficientemente grandes como para concluir que los datos no sigan una distribución normal. Para esta prueba, el estadísticio χ^2 es:

$$\chi^2 = \sum \frac{(O_i - \epsilon_i)^2}{\epsilon_i}$$

Con O_i := frecuencias observadas en la categoría i y ϵ_i := frecuencias esperadas en la categoría i.

Las hipótesis son: H_0 := hipótesis nula, en la cual los datos siguen una distribución Normal. Por otro lado, H_1 := hipótesis alternativa, definida como $\neg H_0$.

Otro concepto importante son los grados de libertad. Par auna prueba de bondad de ajuste, los grados de libertad de ajuste se calculan como:

$$df = grados\ de\ libertad = \#\ de\ categoras - 1 - \#\ de\ parmetros\ estimados$$

En la distribución normal tenemos dos parámetros: μ y σ.

También necesitamos el p-value, definido como:

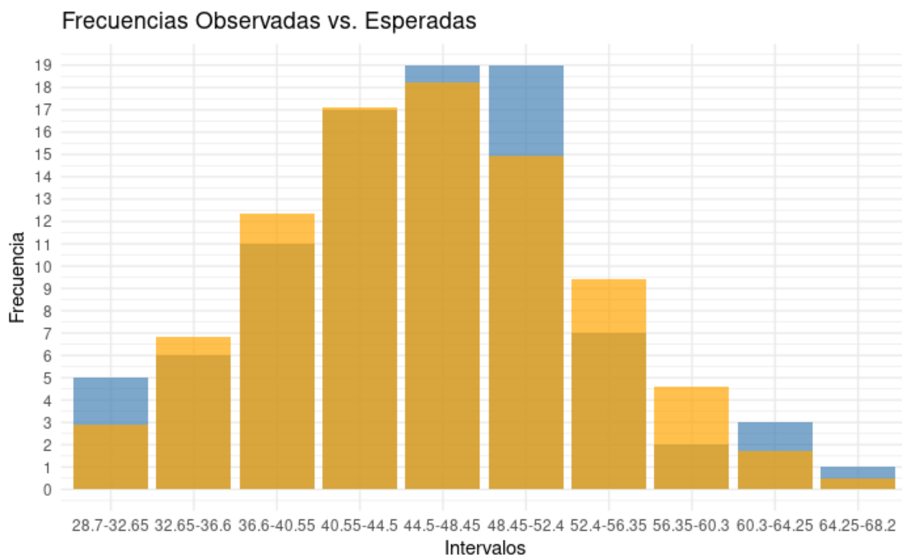
$$p - value = 1 - F(\chi^2, df)$$

Con $F(\chi^2, df)$ siendo la distribución acumulada de la distribución Ji-cuadrada. Se puede obtener con software, tal como lo haremos aquí.

En conclusión, si el p-value es menor que el valor de significancia α (con $\alpha = 0.05$ como el valor estándar), entonces se rechaza a $H - 0$, la hipótesis nula. Por lo tanto, en dicho caso, los datos no siguen una distribución normal. Si el p-value $>\alpha$, entonces no hay suficientes pruebas para rechazar H_0, la hipótesis nula.

En general, se recomienda que tengamos al menos cinco categorías a evaluar para poder realizar una prueba de este estilo y tener resultados confiables. El software en este caso arrojó un valor $\chi^2 = 6.31$, $p-value = 0.7085$ y unos $df = 9$. Como $p-value > \alpha$, entonces no se rechaza la hipótesis nula y podemos decir que los datos siguen una distribución normal de la forma $Normal(45.47, 7.63)$.

De hecho, podemos ver en la siguiente gráfica una comparación entre las frecuencias esperadas (en azul) y las frecuencias estimadas (en amarillo). Esto revela como la tendencia de los datos se acerca bastante a la normal.



Ejercicio p. 244

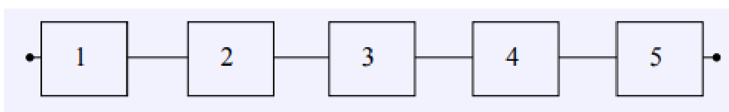
Consideremos un sisetma formado por cinco componentes idénticos conectados en serie tal como se muestra en la figura después del enunciado. Tan pronto como un componente falla, el sistema completo falla. Supongamos que cada componetne sigue un modelo de tiempo a la falla exponencial con $\theta = 100$, y que los componentes fallan en forma independiente una de la otra. Definamos los eventos: $A_i = \{i - esimo\ componente\ dura\ al\ menos\ t\ horas\}$, con $i = 1, 2, 3, 4, 5$. Las A_i son independientes e identicamente distribuidas. Sea X el tiempo de la falla del sistema, responde:

(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 cinco componentes tenemos n, ¿ Cuál es la distribución de X?

Si nuestro sistema está formado por cinco competentes de las mismas características, y distribuidos en serie, como se ve a continuación:



Entonces, ya sabemos que cada componente tiene un tiempo de vida que sigue una distribución exponencial con $\theta = 100$ y con función de supervivencia $S_i(t) = e^{-\frac{t}{100}}$. Además, podemos modelar los tiempos de vida como:

$$T_{min} = mn$$

Sin embargo, se menciona que los tiempos de vida se modelan como eventos independientes. Entonces, la probabilidad de que el sistema no falle en un tiempo t será la multiplicación de todas las probabilidades. Es decir, el producto de las probabilidades de los tiempos de vida de cada componente:

$$P(T_{min}) = \prod_{i=1}^{5} P(T_i > t)$$

Entonces:

$$P(T_{min}) = \left(e^{-\frac{t}{100}}\right)^5$$

$$P(T_{min}) = e^{-\frac{t}{20}}$$

Aquí podemos encontrarnos con que $\theta = 20$, esto significa que el tiempo de vida de nuestro sistema sigue una Distribución Exponencial con parámetro igual a 20 hrs.

Ejercicio p. 265

¿Cuál es la densidad de una variable aleatoria χ^2 con $\nu = 2$?

Sabemos que para una v.a. χ^2, la función de densidad está construida de la siguiente manera:

$$f(x) = \begin{cases} \frac{1}{\Gamma(\frac{\nu}{2})2^{\frac{\nu}{2}}} \cdot x^{(\frac{\nu}{2}-1)} \cdot exp(-\frac{x}{2}) & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Por lo tanto, es claro que lo que se recurre solamente es hacer la sustitución solicitada y evaluar. Entonces, sea $\nu = 2$:

Para $X \geq 0$:

$$f(x) = \frac{1}{\Gamma(\frac{2}{2})2^{\frac{2}{2}}} x^{\frac{2}{2}-1} exp(-\frac{x}{2}) = \frac{1}{2} x^0 e^{-\frac{x}{2}}$$

Entonces:

$$f(x) = \frac{1}{2} e^{-\frac{x}{2}}$$

Y $f(x) = 0$ para $x < 0$. Por lo tanto, cuando $\nu = 2$:

$$f(x) = \begin{cases} \frac{1}{2} e^{-\frac{x}{2}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Esta función se vería como en la figura siguiente.

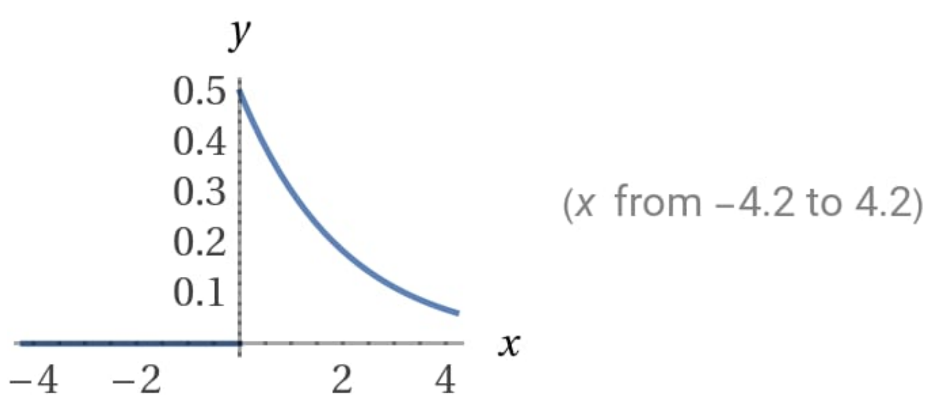


Figura 2: Gráfica de χ^2 cuando $\nu = 2$

Ejercicio p. 285

En muchos proyectos se emplea un método llamado PERT (*Program Evaluation and Review Technique*), para coordinar varias actividades en empresas grandes y/o complejas. 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 definido como "el tiempo optimista" (si todo va bien), y B como "el 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 $\alpha = 2$ y $\beta = 3$. Calcula la probabilidad de que el tiempo para terminar los cimientos sea menor a tres días. (Nota: A y B son valores de los parámetros, recuerda que primero debes codificar la v.a. para que tome valores entre (0,1).

Sabemos que si $X \sim Beta(\alpha, \beta)$, entonces la función de densidad $f(x)$ es:

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \textit{cualquier otro caso} \end{cases}$$

En nuestro problema, tenemos que:

$$f(x) = \frac{\Gamma(2+3)}{\Gamma(2)\Gamma(3)x^{2-1}(1-x)^{(3-1)}}$$

Sabemos que $\Gamma(2+3) = \Gamma(5) = 4! = 24$, $\Gamma(3) = 2! = 4$, y que $\Gamma(2) = 1! = 1$. Entonces:

$$f(x) = 12x(1-x)^2$$

Sin embargo, en este caso, el tiempo X necesario para completar la construcción tiene un intervalo de $[A, B] = [2, 5]$, no de $[0, 1]$, como está establecido para el modelo Beta base. Sin embargo, se puede realizar una transformación lineal para ajustar nuestro intervalo tal que se crea una X' con $X' \in [0, 1]$.

$$X = A + (B-A)X' \quad \textit{con } A = 2 \;\&\; B = 5$$

Entonces:

$$X = 2 + (5-2)X' \;\Rightarrow\; X = 2 + 3X'$$

Por lo tanto:

$$X' = \frac{X-2}{3}$$

De ese modo, pasamos de buscar la probabilidad de $X \leq 3$, para buscar $P(X' \leq \frac{X-2}{3})$ con $X = 3$. Entonces $P(X' \leq \frac{3-2}{3}) = P(X' \leq \frac{1}{3})$.

Entonces:

$$P(X \leq 3) = P(X' \leq \frac{1}{3}) = \int_0^{1/3} 12X'(1-X')^2 \, dX' = 12 \int_0^{1/3} X'(1-X')^2 \, dX'$$

$$\Rightarrow P(X' \leq \frac{1}{3}) = 12 \int_0^{1/3} (X' - 2X'^2 + X'^3) \, dX'$$

$$\Rightarrow P(X' \leq \frac{1}{3}) = 12 \int_0^{1/3} X' \, dX' - 24 \int_0^{1/3} X'^2 \, dX' + 12 \int_0^{1/3} X'^3 \, dX'$$

$$\Rightarrow P(X' \leq \frac{1}{3}) = 12 \left[\frac{X'^2}{2} \right]_0^{1/3} - 24 \left[\frac{X'^3}{3} \right]_0^{1/3} + 12 \left[\frac{X'^4}{4} \right]_0^{1/3}$$

$$\Rightarrow P(X' \leq \frac{1}{3}) = 6 \left[\left(\frac{1}{3} \right)^2 \right]_0^{1/3} - 8 \left[\left(\frac{1}{3} \right)^3 \right]_0^{1/3} + 3 \left[\left(\frac{1}{3} \right)^4 \right]_0^{1/3}$$

$$\Rightarrow P(X' \leq \frac{1}{3}) = \frac{11}{27} \simeq 0.407407407...$$

Por lo tanto, la probabilidad de que se termine la obra en menos de tres días es de aproximadamente el 40 %.

Este resultado tiene sentido, pues la esperanza $E(X') = \frac{\alpha}{\alpha+\beta} = \frac{2}{2+3} = \frac{2}{5} = 0.4$. Y por la transformación lineal $X = A + (B - A)X'$:

$$E(X) = A + (B - A)E(X') = 2 + 3E(X') = 2 + 3(0.4) = 3.2$$

Además, podemos recurrir a la expresión ajustada para $A \leq X \leq B$:

$$f(x) = \begin{cases} \frac{1}{B-A} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{X-A}{B-A} \right)^{\alpha-1} \left(\frac{B-X}{B-A} \right)^{\beta-1} & A < x < B \\ 0 & \textit{cualquier otro caso} \end{cases}$$

Esta expresión ya está ajustada para que podamos emplearla de manera directa, sin necesidad de la transformación lineal $X = A+(B-A)X'$. Entonces, podemos desarrollar, recordando que $\alpha = 2$ y $\beta = 3$:

$$f(x) = \frac{1}{5-2} 12 \left(\frac{X-2}{5-2} \right)^{\alpha-1} \left(\frac{5-X}{5-2} \right)^{\beta-1} = \frac{12}{3} \left(\frac{X-2}{3} \right)^{\alpha-1} \left(\frac{5-X}{3} \right)^{\beta-1}$$

Entonces:

$$f(x) = \frac{12}{81} \left(x - 2 \right) \left(5 - x \right)^2$$

$$f(x) = \frac{12}{81} \left(x^2 - 12x^2 + 45x - 50 \right)$$

En este caso, ya podríamos integrar de [2, 3], pues eso cae en nuestro intervalo [2, 5] de interés. Así:

$$f(x) = \frac{12}{81} \int_2^3 \left(x^2 - 12x^2 + 45x - 50 \right) dx$$

Después de desarrollar todo:

$$f(x) = \frac{12}{81} \frac{11}{4} \simeq 0.407407407...$$

Por lo tanto, llegamos a la misma probabilidad del 40 % de que se realice la construcción en menos de tres días.

Ejercicio p. 305

Grafica la función Weibull para los siguientes valores de los parámetros ($\alpha = 1, \beta = 2$)**,** ($\alpha = 2, \beta = 2$)**,** ($\alpha = 3, \beta = 4$)**. Puedes hacerlo de manera algebraica, o bien, generando variables con ésta distribución construyendo un histograma.Después gráficas las funciones de confiabilidad y riesgo para cada uno de los casos anteriores.**

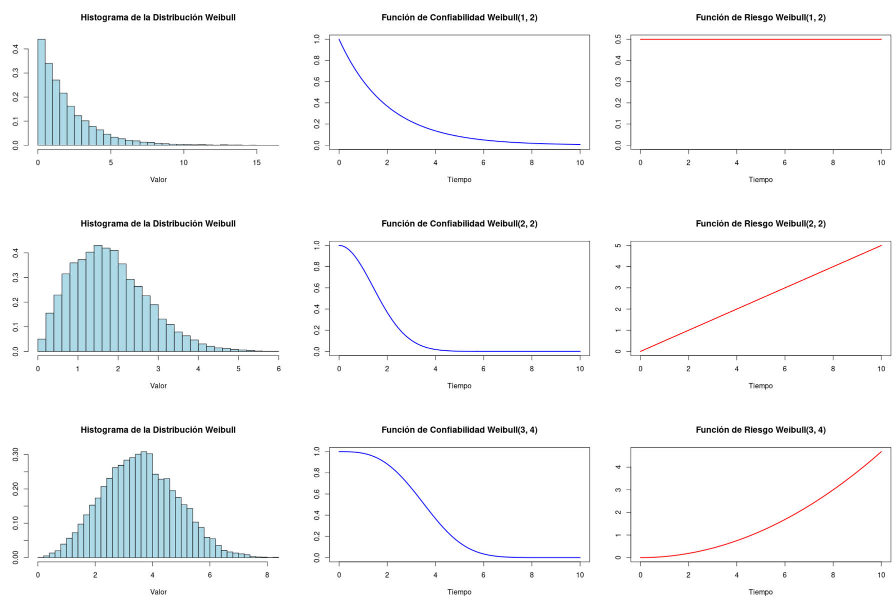


Figura 3: Comparación entre funciones Weibul | Arriba: $Weibul(1, 2)$, en medio: $Weibul(2, 2)$, abajo: $Weibul(3, 4)$

Si $h(t) = a + bt$**, ¿cuál es la densidad asociada?**
Sabemos que:

$$h(t) = \frac{f(t)}{S(t)}$$

Además, podemos encontrar que $h(t) = -\frac{d}{dt} ln\Big(S(t)\Big)$. Entonces:

$$\int h(t)\; dt = -ln\Big(S(t)\Big)$$

Entonces:

$$\int_0^t (a + bt)\; dt = \int_0^t a\; dt + \int_0^t bt\; dt = at + \frac{1}{2}bt^2$$

Por lo que:

$$ln\Big(S(t)\Big) = -(at + \frac{1}{2}bt^2) \;\Rightarrow\; exp\Big(ln(S(t))\Big) = exp\Big(-(at + \frac{1}{2}bt^2)\Big)$$

$$\Rightarrow S(t) = e^{-(at+\frac{1}{2}bt^2)}$$

Por lo tanto:

$$f(t) = h(t) \cdot S(t)$$

$$f(t) = (a + bt) \cdot e^{-(at+\frac{1}{2}bt^2)}$$