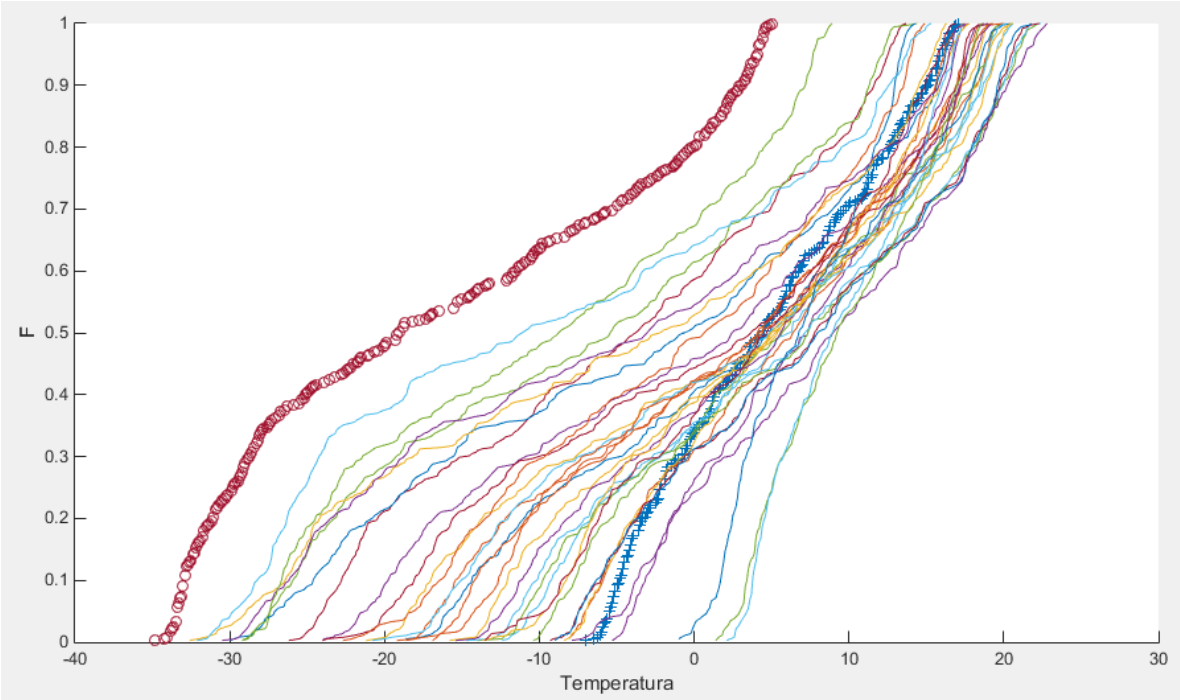
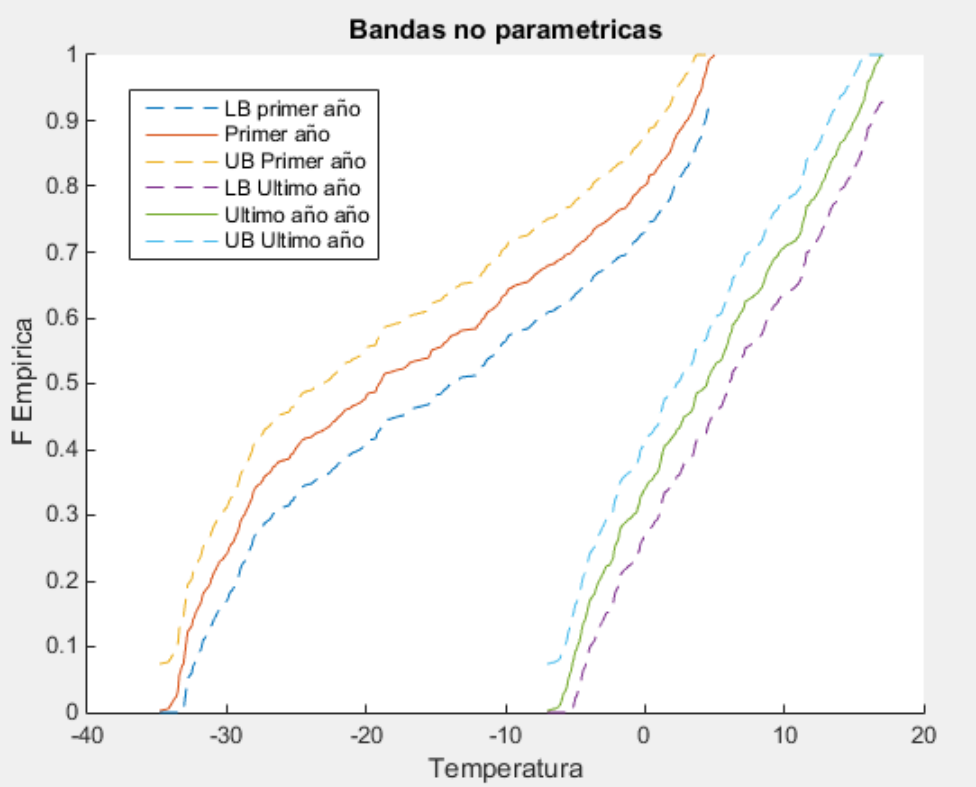
Solución Taller 1 – Pablo A. Saldarriaga Aristizabal

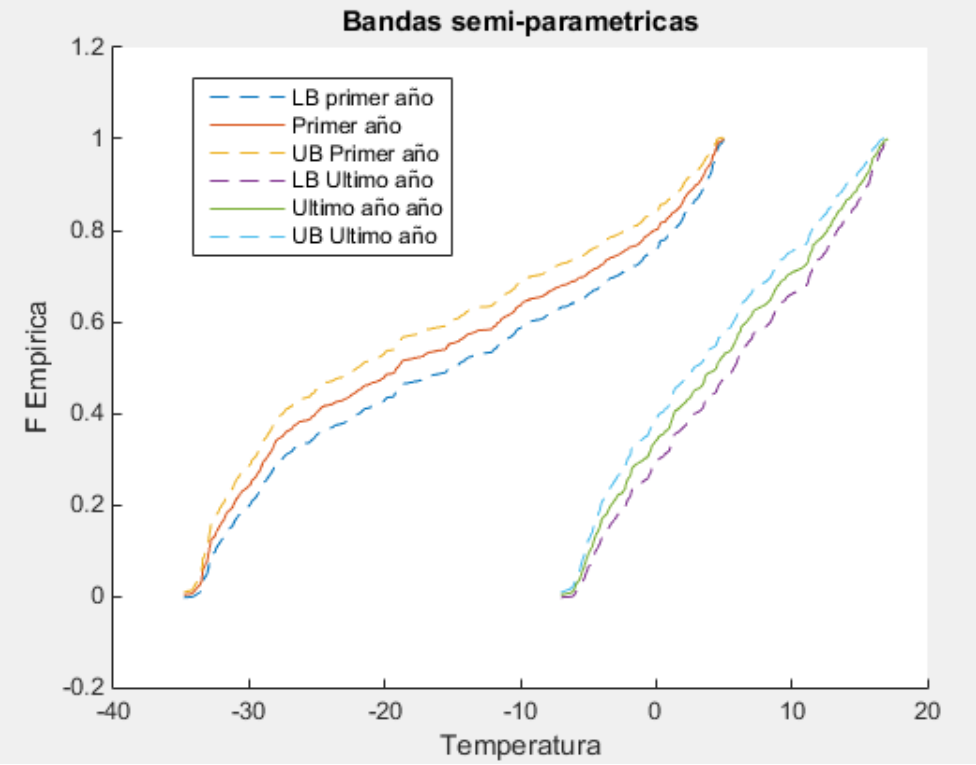
a. Grafique en un mismo plano las funciones de distribución empíricas de las temperaturas de cada año.



La gráfica representada por “o” corresponde a las temperaturas del primer año, mientras que la representada por “+” corresponde a las temperaturas del año más reciente, las demás son las temperaturas en los años intermedios. Claramente se evidencia un cambio en la temperatura en el último año a comparación del primer año, pues antes era más probable que la temperatura diaria en Canadá fuera más baja, mientras que en el último año es más probable que se presenten temperaturas más calientes. Analizando ambas distribuciones empíricas, podemos ver que antes temperaturas entre -10 a 5 grados eran más comunes en el pasado, ahora esas temperaturas no son tan probables, ya que temperaturas superiores a 10 grados son las más comunes en los últimos años.

b. **Calcule y grafique las bandas de confianza a un 95% de confianza para la función de distribución empírica del primer año y último año. ¿Hay sectores solapados?**





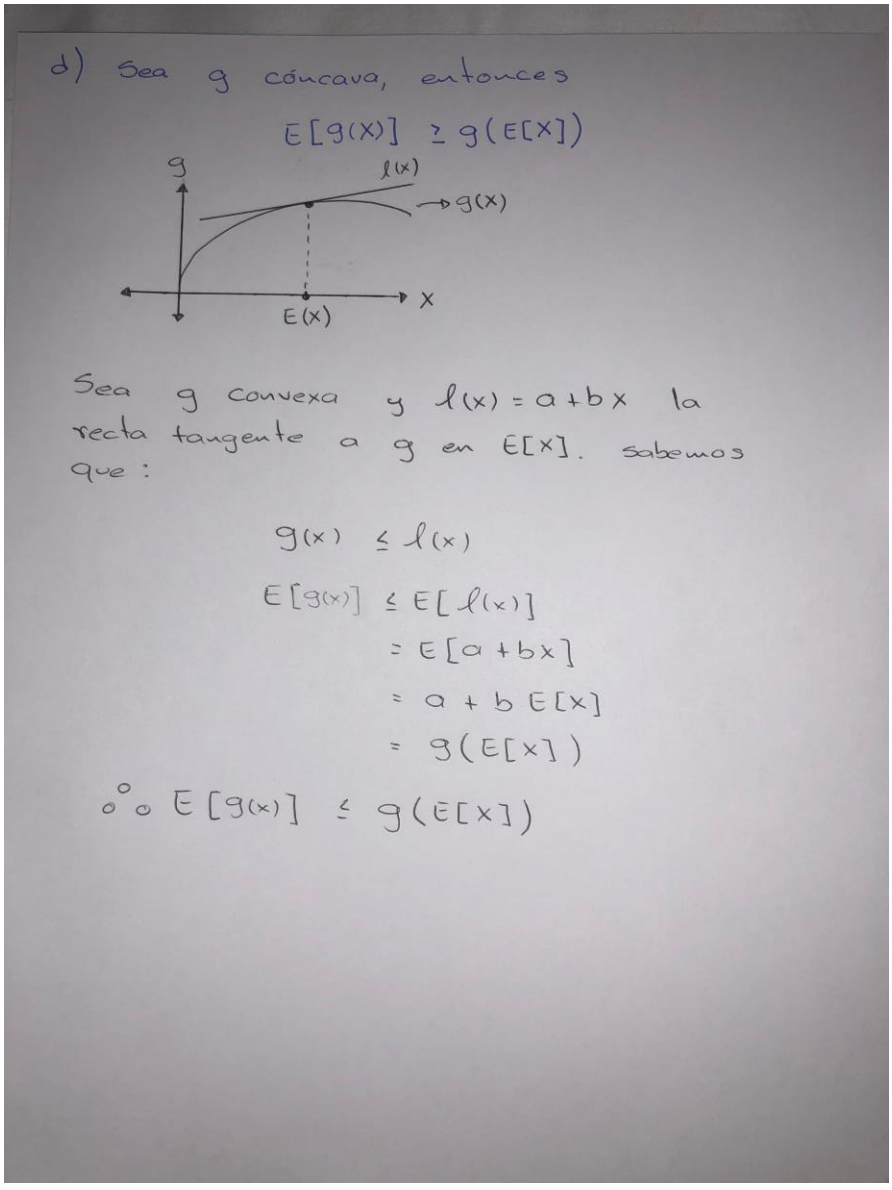
Respecto al primer año y el último año, no se presentan sectores solapados.

c. **Escriba y ejecute un código que permita visualizar el Teorema de Glivenko Cantelli para una distribución exponencial de parámetro 1.**

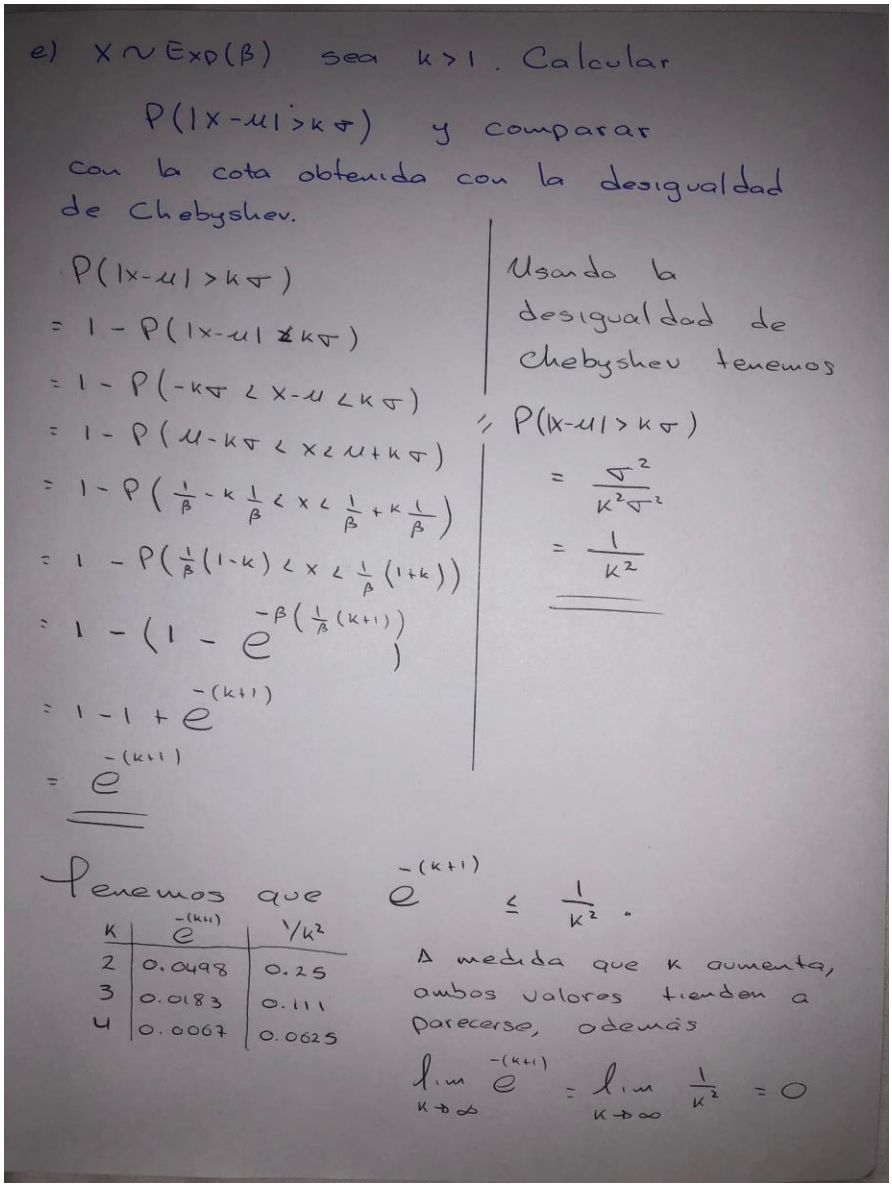
El script se corrió modificando el valor de n, los resultados están presentados en la tabla. Podemos ver que a medida que n tiene a infinito, max{|Fn(xi)-F(xi)|} tiende a 0 (cero).

n	Max{|Fn(xi)-F(xi)|}
10	0.20496
100	0.10335
1000	0.026376
10000	0.0080243
100000	0.0035879

d.



e.



e) $X \sim Exp(\beta)$ sea $k > 1$. Calcular

$P(|X-\mu| > k\sigma)$ y comparar con la cota obtenida con la desigualdad de Chebyshev.

$P(|X-\mu| > k\sigma)$

$= 1 - P(|X-\mu| \leq k\sigma)$

$= 1 - P(-k\sigma < X-\mu < k\sigma)$

$= 1 - P(\mu - k\sigma < X < \mu + k\sigma)$

$= 1 - P\left(\frac{1}{\beta} - k\frac{1}{\beta} < X < \frac{1}{\beta} + k\frac{1}{\beta}\right)$

$= 1 - P\left(\frac{1}{\beta}(1-k) < X < \frac{1}{\beta}(1+k)\right)$

$= 1 - \left(1 - e^{-\beta\left(\frac{1}{\beta}(k+1)\right)}\right)$

$= 1 - 1 + e^{-(k+1)}$

$= e^{-(k+1)}$

Usando la desigualdad de Chebyshev tenemos

$\leq P(|X-\mu| > k\sigma)$

$= \frac{\sigma^2}{k^2\sigma^2}$

$= \frac{1}{k^2}$

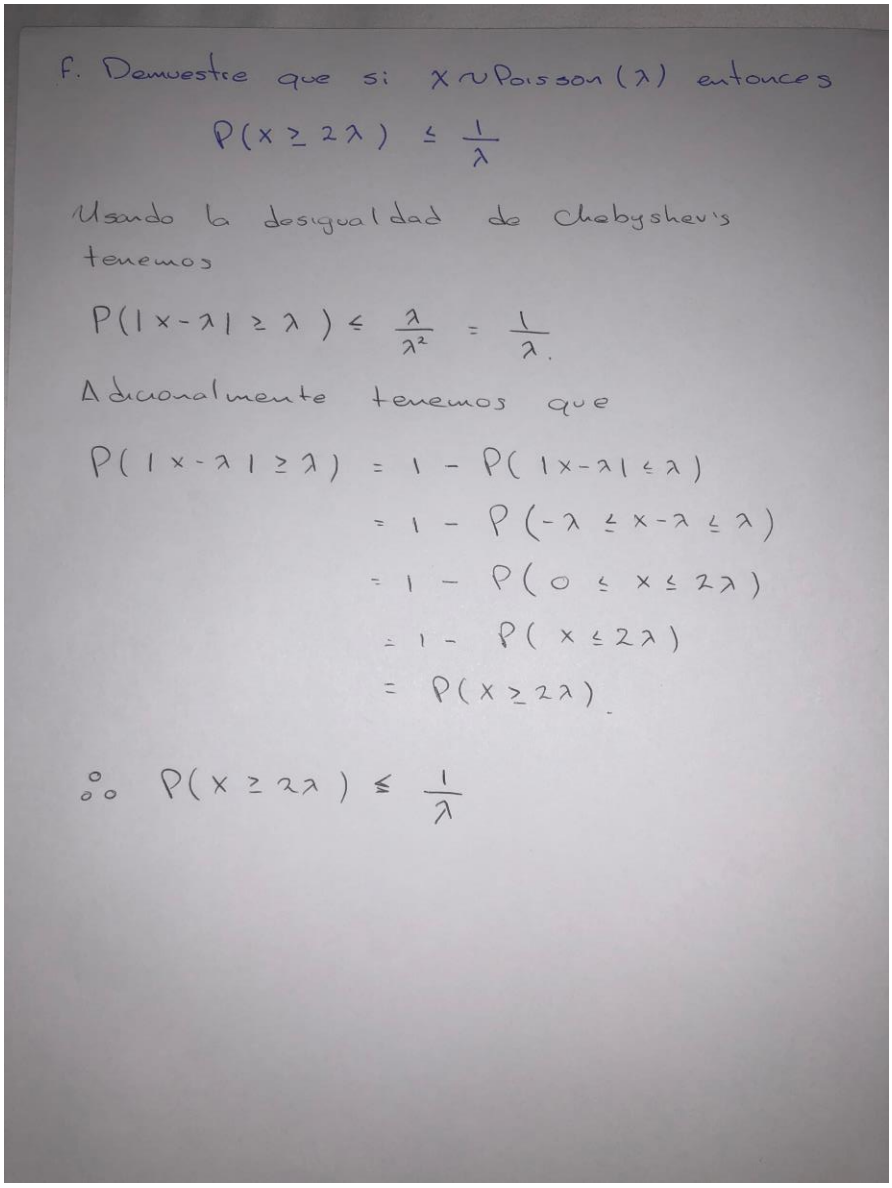
Tenemos que $e^{-(k+1)} \leq \frac{1}{k^2}$.

k	$e^{-(k+1)}$	$1/k^2$
2	0.0498	0.25
3	0.0183	0.111
4	0.0067	0.0625

A medida que k aumenta, ambos valores tienden a parecerse, además

$\lim_{k\to\infty} e^{-(k+1)} = \lim_{k\to\infty} \frac{1}{k^2} = 0$

f.



f. Demuestre que si $X \sim Poisson(\lambda)$ entonces

$$P(X \geq 2\lambda) \leq \frac{1}{\lambda}$$

Usando la desigualdad de Chebyshev's tenemos

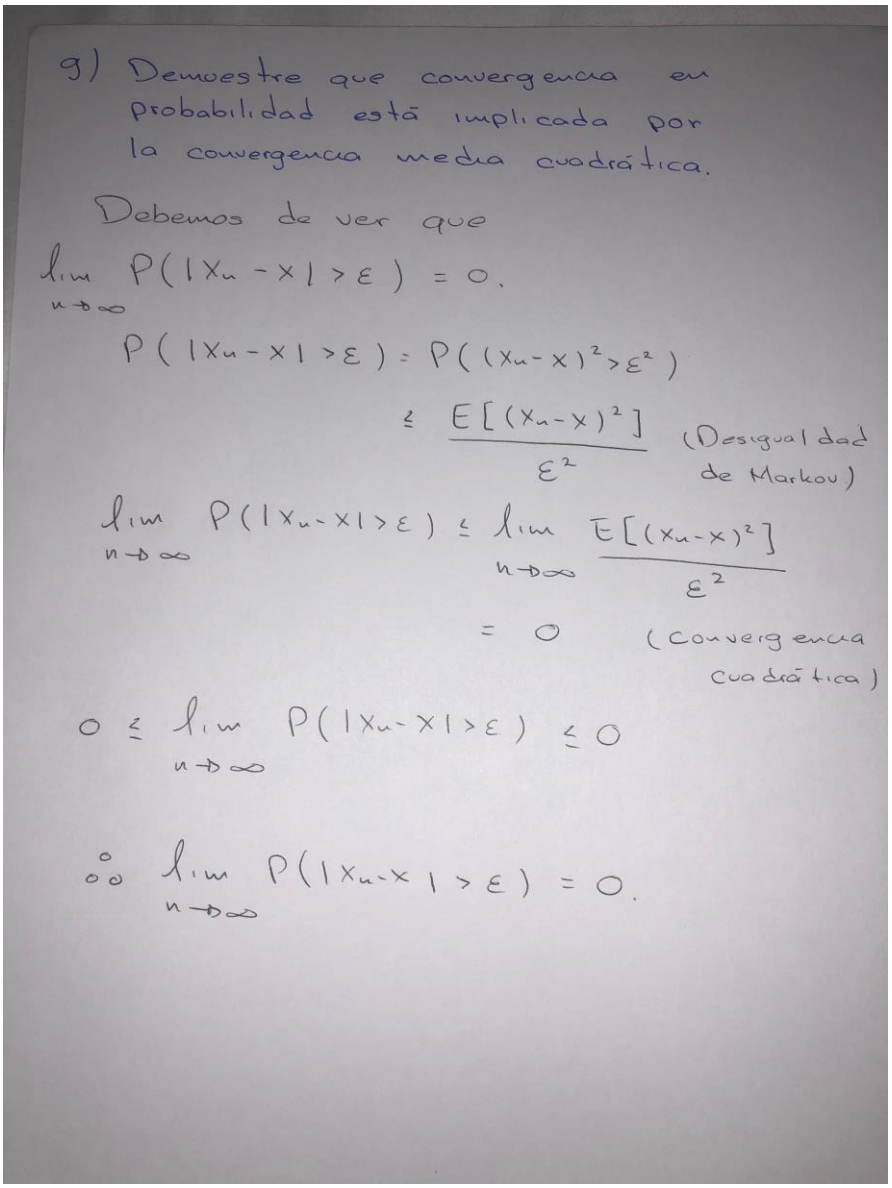
$$P(|X-\lambda| \geq \lambda) \leq \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}.$$

Adicionalmente tenemos que

$$\begin{aligned} P(|X-\lambda| \geq \lambda) &= 1 - P(|X-\lambda| \leq \lambda) \\ &= 1 - P(-\lambda \leq X-\lambda \leq \lambda) \\ &= 1 - P(0 \leq X \leq 2\lambda) \\ &= 1 - P(X \leq 2\lambda) \\ &= P(X \geq 2\lambda). \end{aligned}$$

$\therefore P(X \geq 2\lambda) \leq \frac{1}{\lambda}$

g.



g) Demuestre que convergencia en probabilidad está implicada por la convergencia media cuadrática.

Debemos de ver que

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0.$$

$$P(|X_n - X| > \varepsilon) = P((X_n - X)^2 > \varepsilon^2)$$

$$\leq \frac{E[(X_n - X)^2]}{\varepsilon^2} \quad \text{(Desigualdad de Markov)}$$

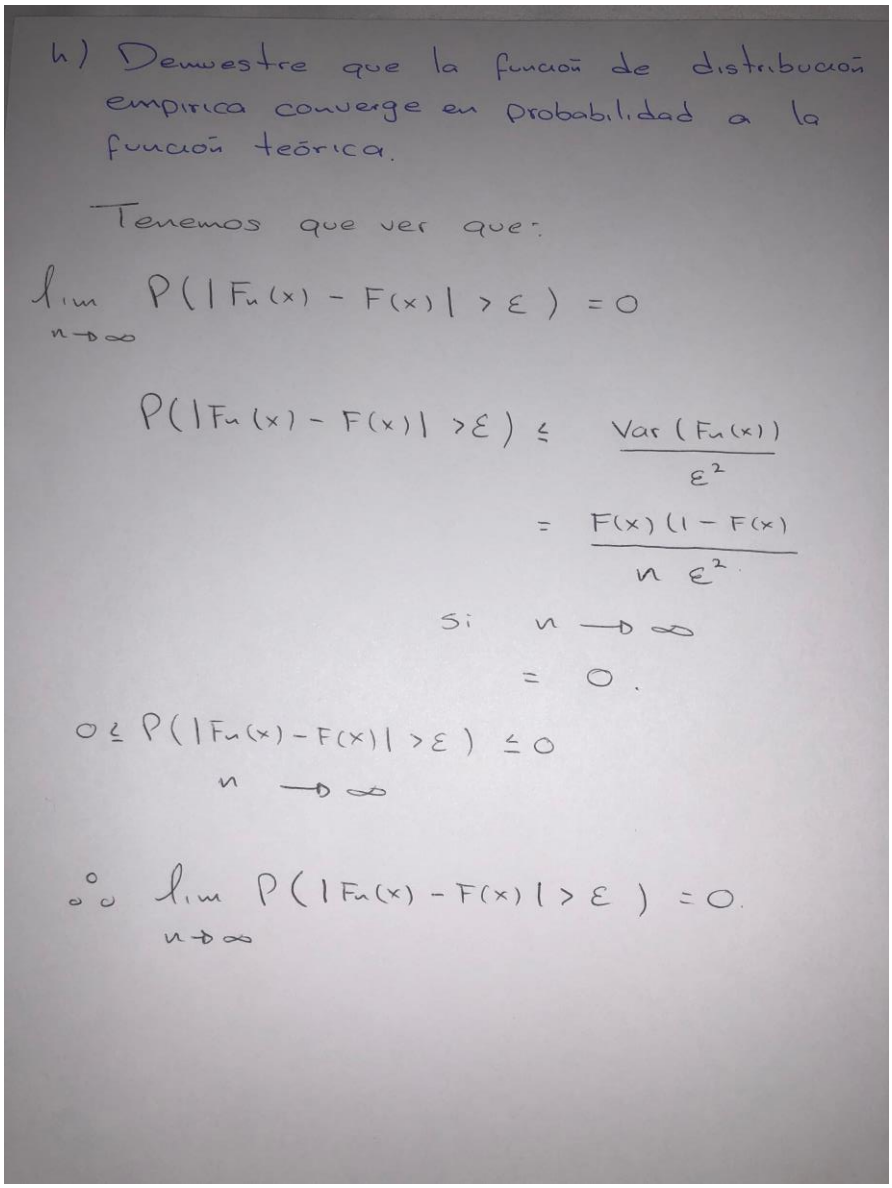
$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) \leq \lim_{n\to\infty} \frac{E[(X_n - X)^2]}{\varepsilon^2}$$

$$= 0 \quad \text{(Convergencia cuadrática)}$$

$$0 \leq \lim_{n\to\infty} P(|X_n - X| > \varepsilon) \leq 0$$

$$\therefore \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0.$$

h.



h) Demuestre que la función de distribución empírica converge en probabilidad a la función teórica.

Tenemos que ver que:

$$\lim_{n\to\infty} P(|F_n(x) - F(x)| > \varepsilon) = 0$$

$$P(|F_n(x) - F(x)| > \varepsilon) \leq \frac{Var(F_n(x))}{\varepsilon^2}$$

$$= \frac{F(x)(1 - F(x))}{n\,\varepsilon^2}$$

Si $n \to \infty$

$$= 0.$$

$$0 \leq P(|F_n(x) - F(x)| > \varepsilon) \leq 0$$

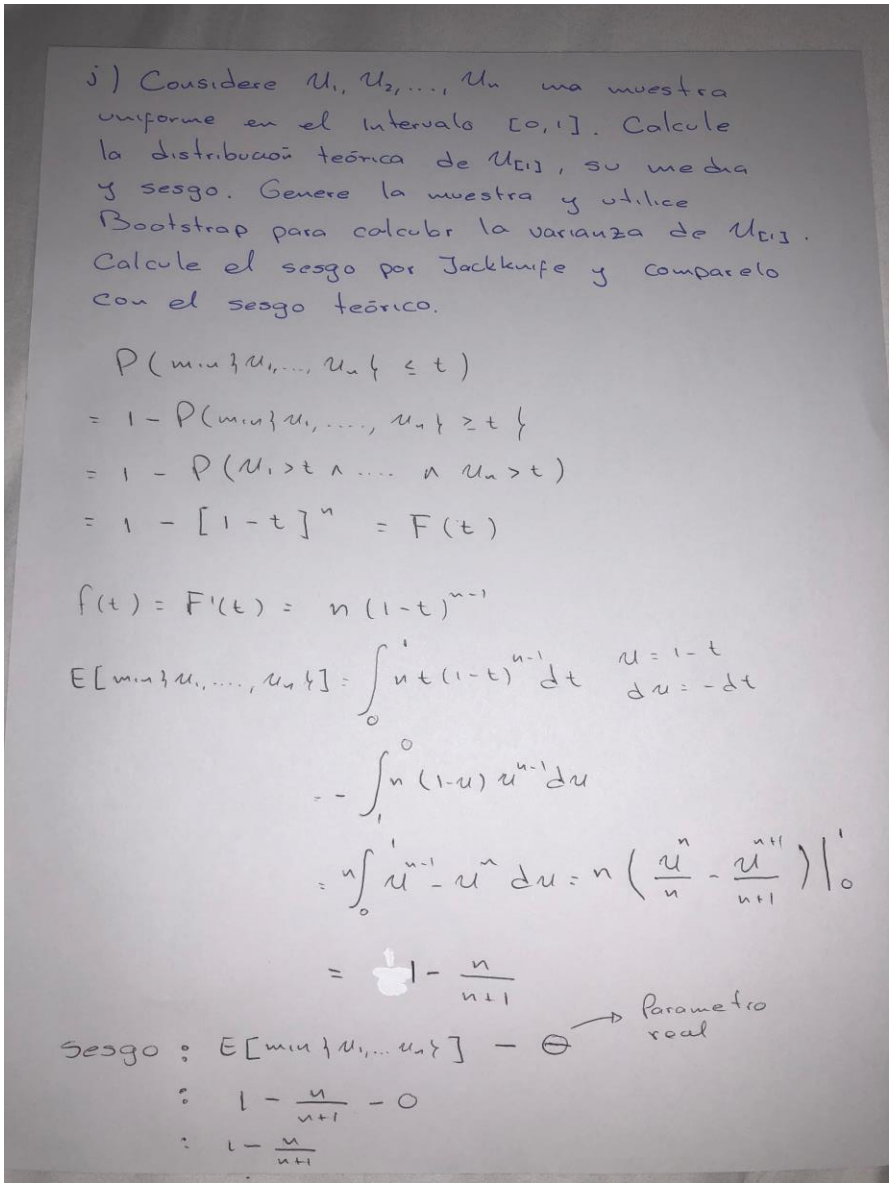
$n \to \infty$

$$\therefore \lim_{n\to\infty} P(|F_n(x) - F(x)| > \varepsilon) = 0.$$

i. **Considere las temperaturas diarias del primer año. Calcule un intervalo de confianza para la temperatura mínima. Calcule el sesgo de *T*[1] y la varianza**.

Sesgo: 0.59836
Mínimo Estimado: -34.8
Intervalo de confianza (Usando varianza de Jackknife): [-35.9728, -33.6272]

j. **Considere *U*1, *U2, …, Un* una muestra de una distribución uniforme el intervalo [0*;* 1]. Calcule la distribución teórica de *U*[1], su media y sesgo. Genere la muestra y utilice bootstrap para calcular la varianza de *U*[1]. Calcule el sesgo por Jacknife y compárelo con el sesgo teórico.**



Generando una muestra, tenemos que:

Varianza Bootstrap: 0.00034448
Sesgo Jackknife: 0.017937
Sesgo Teorico: 0.009901

k. **Explique una forma no paramétrica y robusta de calcular las componentes principales. Aplique la técnica a las temperaturas del primer año y último año. Compárelas con las componentes principales obtenidas de forma habitual.**

La técnica de componentes principales es una técnica de reducción de dimensionalidad, esta usa ya sea la matriz de covarianza de los datos, o la matriz de correlación, de forma tal que, obteniendo los vectores propios de la matriz, es posible crear una combinación lineal de los datos en múltiples dimensiones a una proyección en una dimensión menor. Teniendo

esto claro, una versión robusta de los componentes principales, es realizar el mismo análisis, pero utilizando una matriz de covarianzas robusta (el caso que fue implementado, es la matriz de covarianzas obtenido al calcular el comedian de los datos), por lo que podría generar mejores resultados con la presencia de datos contaminados.

Resultados datos originales:

Versión Paramétrica				**Versión No Paramétrica**			
% Varianza	**Valores propios**	**Coeficientes**		**% Varianza**	**Valores propios**	**Coeficientes**	
99.19	238.8732	0.4623	0.8867	98.2432	198.9327	0.4674	0.8840
0.81	1.9430	0.8867	-0.4623	1.7568	3.5573	0.8840	-0.4674

Resultados datos contaminados:

Versión Paramétrica				**Versión No Paramétrica**			
% Varianza	**Valores propios**	**Coeficientes**		**% Varianza**	**Valores propios**	**Coeficientes**	
98.1402	243.4103	0.4814	0.8765	98.5668	209.6023	0.4573	0.8893
1.8598	4.6128	0.8765	-0.4814	1.4332	3.0477	0.8893	-0.4573

Al comparar los resultados con los datos contaminados y los datos originales, ambas versiones del análisis de componentes principales presentan resultados similares, sin embargo, es más sensible la versión paramétrica al tener datos contaminados, pues el cambio que hubo al utilizar la versión no paramétrica non presenta grandes cambios.