Taller 1 Metodos No Parametricos

Mosquera Galvis, Liceth Cristina - lcmosquerg@eafit.edu.co
Rios Naranjo, Johan Steward - jriosna1@eafit.edu.co
Estrada Pérez, Juan Diego - jestra15@eafit.edu.co
Cuscagua López, Juan Mauricio - jcuscagu@eafit.edu.co
Rozo Alzate Javier Arturo - jarozoa@eafit.edu.co
Programa: Técnicas robustas y no paramétricas
Docente:
Henry Laniado Rodas

5 de septiembre de 2019

1. Grafique en un mismo plano las funciones de distribución empíricas de las temperaturas de cada año. Explique con base en las gráficas de las funciones, si es observable un efecto de cambio climático. Interprete los intervalos donde la distribución empírica de la temperaturas del año más reciente es mayor a la del año más antiguo.

El código en el anexo 11.1 se usa para graficar de manera iterativa la distribución empírica de las temperaturas. La figura 1 muestra las temperaturas para cada día del año a lo largo de 35 años. Por otro lado, la figura 2 muestra el comportamiento de las distribuciones empíricas para cada año.

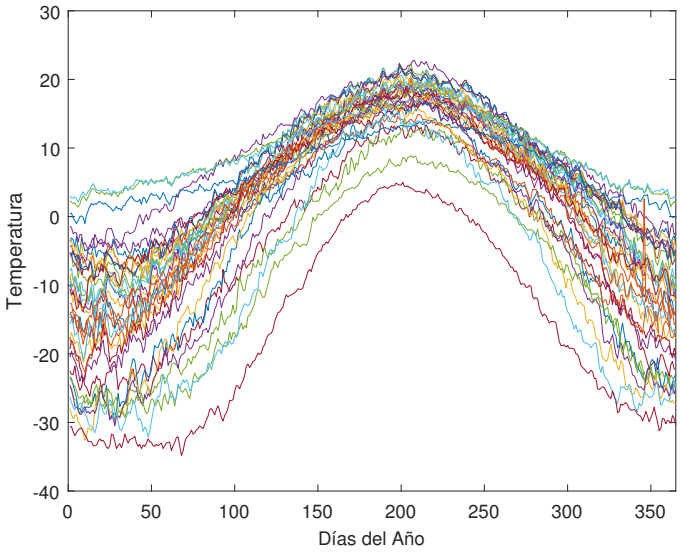


Figura 1: Temperaturas 35 años.

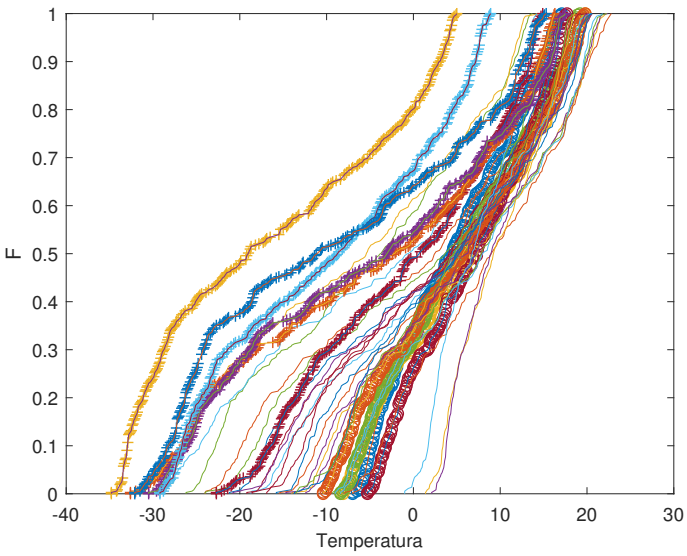


Figura 2: FDA Empíricas.

al comparar los 5 años mas antiguos (lineas punteadas en '+') con los 5 años mas recientes (lineas punteadas en 'o') se observa que las temperaturas en los primeros años eran menores a las temperaturas en los años mas recientes. En los primeros años se presentaban temperaturas entre -35 y 15. Este intervalo se ve más acotado en años posteriores donde hay temperaturas entre -10 y 20.

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

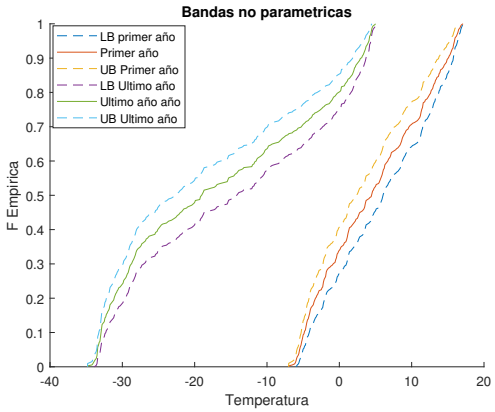
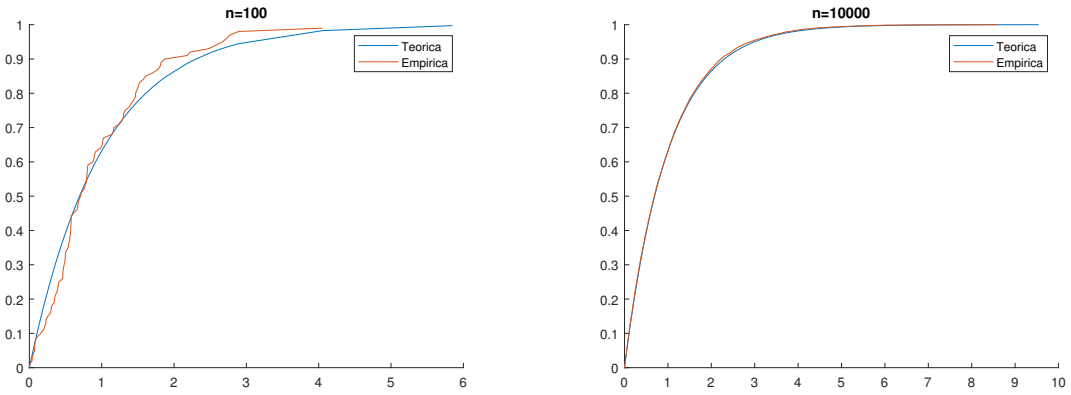


Figura 3: Bandas de confianza de funciones de distribución empíricas para el primer y último año a un 95 % de confianza.

Al comparar las funciones empíricas del primer y último año, se observa que no hay solapamiento entre las funciones o las bandas de confianza al 95 %.

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



(a) Distribución exponencial teórica vs estimada, $n = 100$

(b) Distribución exponencial teórica vs estimada, $n = 10000$

Figura 4: Comparación del ajuste ante el incremento de n

n	$max(\|F_n(x_i) - F(x_i)\|)$
100	0.11986
1000	0.018922
10000	0.0066632
100000	0.002496

Cuadro 1: Convergencia ante incremento de n

El código adjunto en el anexo 11.3 permite calcular, para un tamaño n determinado, el valor $max(|Fn(xi) - F(xi)|)$ asociado al teorema de Glivenko Cantelli. Para diferentes valores de n, la tabla 1 muestra la convergencia a 0 de la expresión $max(|Fn(xi) - F(xi)|)$. Las figuras 4a y 4b muestran cómo el ajuste a la distribución teórica mejora a medida que n incrementa.

4. Enuncie y demuestre la desigualdad de Jensen para funciónes cóncavas.

Para cualquier función cóncava f, se cumple que

$$E[f(X)] \leq f(E[X])$$

prueba: Suponga que f es diferenciable. La función f es cóncava si, para cada x y y se cumple

$$f(x) \leq f(y) + (x - y)f'(y)$$

Sea $x = X$ y $y = E[X]$, por lo tanto se puede escribir

$$f(X) \leq f(E[X]) + (X - E[X])f'(E[X])$$

Esta desigualdad se cumple para todo X, por lo que al tomar valor esperado a ambos lados, se tiene que

$$E[f(X)] \leq f(E[X]) + f'(E[X])E[(X - E[X])] = f(E[X])$$

5. Suponga que X es una variable aleatoria exponencial de parámetro β. Calcule $P(|X - \mu| > k\sigma)$ para $k > 1$. Compare el resultado con la cota obtenida de la desiguladad de Chebyshev.

Dada la variable aleatoria $X \sim \exp(\beta)$ y tomando $\mu = \frac{1}{\beta}$, $\sigma = \frac{1}{\beta}$ se tiene que:

$$\begin{aligned}
P(|X - \mu| > k\sigma) &= 1 - P(|X - \mu| < k\sigma) \\
&= 1 - P(-k\sigma < X - \mu < k\sigma) \\
&= 1 - P(\mu - k\sigma < X < \mu + k\sigma) \\
&= 1 - P(\frac{1}{\beta} - k\frac{1}{\beta} < X < \frac{1}{\beta} + k\frac{1}{\beta}) \\
&= 1 - P(\frac{1}{\beta}(1 - k) < X < \frac{1}{\beta}(1 + k)) \\
&= 1 - (1 - e^{-\beta(\frac{1}{\beta}(k+1))} \\
&= 1 - (1 - e^{-(k+1)}) \\
&= e^{-(k+1)}
\end{aligned}$$

Al usar la desigualdad de Chebyshev se tiene que:

$$P(|X - \mu| > k\sigma) = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

de esta manera, se tiene que $e^{-(k+1)} \leq \frac{1}{k^2}$. Por lo tanto, a medida que k aumenta, ambos valores tienden a converger, esto se ve al tomar límite en ambos lados de la desigualdad.

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

6. Demuestre que si X es Poisson de parámetro λ entonces $P(x \geq 2\lambda) \leq \frac{1}{\lambda}$

De la desigualdad de Chebyshev tenemos que

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

Por otro lado, se tiene que:

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

Por lo tanto,

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

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

Se debe probar que

$$\lim_{n\to\infty} P(|x_n - x| > \epsilon) = 0$$

Consideremos entonces,

$$\begin{aligned} P(|X_n - x| > \epsilon) &= P((X_n - x)^2 > \epsilon^2) \\ &\leq \frac{E[(x_n - x)^2]}{\epsilon^2} \end{aligned}$$

Esto se cumple debido a la desigualdad de Markov. Al tomar límites en ambos lados de la desigualdad

$$\begin{aligned} \lim_{n\to\infty} P(|X_n - x| > \epsilon) &\leq \lim_{n\to\infty} P((X_n - x)^2 > \epsilon^2) \\ &= 0 \end{aligned}$$

de esta manera

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

por lo tanto

$$\lim_{n\to\infty} P(|X_n - x| > \epsilon) = 0$$

8. Considere las temperaturas diarias del primer año. Calcule el intervalo de confianza para la temperatura mínima. Calcule el sesgo de $T_{[1]}$ y la varianza

En los códigos relacionados en el anexo 11.4, se relacionan los cálculos para el sesgo, el mínimo estimado y un intervalo de confianza usando la varianza de Jack-knife. Los resultados son los siguientes:

- Sesgo: 0.49863
- Mínimo Estimado: -7
- Intervalo de confianza: [-7.9773, -6.0227]

9. Considere $U_1, U_2, ..., U_n$ una muestra de una distribución uniforme en 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

Para encontrar el valor teórico del sesgo, hay que considerar $P(min\{U_1, ..., U_n\} \leq t)$:

$$\begin{aligned} P(min\{U_1, ..., U_n\} \leq t) &= 1 - P(min\{U_1, ..., U_n\} \geq t) \\ &= 1 - P(u_1 > t \wedge ... \wedge u_n > t) \\ &= 1 - [1-t]^n \\ &= F(t) \end{aligned}$$

por lo tanto, se tiene que $f(t) = F'(t) = n(1-t)^{n-1}$. Por otro lado,

$$\begin{aligned} E[min\{u_1, ..., u_n\}] &= \int_0^1 nt(1-t)^{n-1}dt \\ &= -\int_1^0 n(1-u)u^{n-1}du \\ &= n\int_0^1 u^{n-1}u^n du \\ &= n(\frac{u^n}{n} - \frac{u^{n+1}}{n+1})\Big|_0^1 = 1 - \frac{n}{n+1} \end{aligned}$$

Considerando θ como el parámetro real, se tiene que el sesgo teórico es

$$E[min\{u_1, ..., u_n\}] - \theta = 1 - \frac{n}{n+1}$$

Usando el código del anexo 11.5, computacionalmente se obtienen los siguientes resultados:

- Varianza Bootstrap: 0.00019173
- Sesgo Jack-knife: 0.020724
- Sesgo Teórico: 0.0099014

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

Entre las principales técnicas de reducción de dimensionalidad, análisis de componentes principales es uno de los más comunes en implementación. Esta metodología usa bien sea la matriz de covarianzas o la matriz de correlaciones de tal forma que, al obtener los vectores propios de la matriz, se puede crear una combinación lineal de los datos en múltiples dimensiones a una proyección en una dimensión menor. En este orden de ideas, una versión robusta del análisis de componentes principales se daría al implementar una matriz de covarianza robusta la cual podría generar resultados más estables ante la presencia de datos atípicos.

La tabla 2 muestra los resultados del análisis de componentes principales paramétrico y no paramétrico para los datos originales. La tabla **??** muestra los mismos resultados para datos contaminados.

Paramétrico				No Paramétrico			
% Var	Valores Propios	Coeficientes		% Var	Valores Propios	Coeficientes	
96.8418	535.9271	0.6286	0.7777	98.5668	209.6023	0.4573	0.8893
3.1582	17.4777	0.7777	-0.6286	1.4332	3.0477	0.8893	-0.4573

Cuadro 2: Resultados de análisis de componentes principales sobre los datos originales.

Paramétrico				No Paramétrico			
% Var	Valores Propios	Coeficientes		% Var	Valores Propios	Coeficientes	
95.8948	322.2904	0.5721	0.8201	98.2432	198.9327	0.4674	0.8840
4.1052	13.7972	0.8201	-0.5721	1.7568	3.5573	0.8840	-0.4674

Cuadro 3: Resultados de análisis de componentes principales sobre los datos contaminados.

Se puede ver como incluir datos contaminados en la muestra altera la capacidad explicativa de los componentes principales. La implementación se encuentra en el código adjunto en el anexo 11.6

11. Anexos

11.1. Anexo punto 1

```
load('temperaturas.mat')
A= temperatura;
[rows,cols] = size(A);
for i = 1:cols
    if i <= 5
        C = A(:,i);
        [f,x] = ecdf(C);
        plot(x,f, 'o')
        hold on
    end
    if i >= (cols-5)
        C = A(:,i);
        [f,x] = ecdf(C);
        plot(x,f,'+')
        hold on
    end
    C = A(:,i);
    [f,x] = ecdf(C);
    plot(x,f)
    hold on
end

figure
for i = 1:cols
    plot(A(:,i))
    hold on
end
```

```
xlim([0  365])
```

11.2. Anexo punto 2

```
load('temperaturas.mat')

[f1,x1] = ecdf(temperatura(:,1));
[f35,x35] = ecdf(temperatura(:,end));

alfa = 0.05;

%% first year

error1 = sqrt(f1.*(1-f1)/length(f1)); %Error Estandar

%Calculo de las bandas semi-parametrico
L1 = f1-1.96.*error1;
U1 = f1+1.96.*error1;

%Calculo de las bandas No parametrico

eps = sqrt((1/(2*length(f1)))*log(2/alfa));
for i=1:length(f1)
 L1np(i) = max([f1(i)-eps,0]);
 U1np(i) = min([f1(i)+eps,1]);
end

%% last year

error35 = sqrt(f35.*(1-f35)/length(f35)); %Error Estandar

%Calculo de las bandas semi-parametrico
L35 = f35-1.96.*error35;
U35 = f35+1.96.*error35;

%Calculo de las bandas No parametrico
eps = sqrt((1/(2*length(f35)))*log(2/alfa));
for i=1:length(f35)
 L35np(i) = max([f35(i)-eps,0]);
 U35np(i) = min([f35(i)+eps,1]);
end

%% Plots

%Bandas Semiparametricas
figure
hold on
plot(x1,L1,'--')
plot(x1,f1)
plot(x1,U1,'--')

```

```
plot(x35,L35,'--')
plot(x35,f35)
plot(x35,U35,'--')
legend('LB primer a o','Primer a o','UB Primer a o','LB Ultimo a o','
    Ultimo a o a o','UB Ultimo a o')
ylabel('F Empirica')
xlabel('Temperatura')
ylim([0 1])
title('Bandas semi-parametricas')
hold off

%Bandas no parametricas
figure
hold on
plot(x1,L1,'--')
plot(x1,f1)
plot(x1,U1,'--')

plot(x35,L35,'--')
plot(x35,f35)
plot(x35,U35,'--')
legend('LB primer a o','Primer a o','UB Primer a o','LB Ultimo a o','
    Ultimo a o a o','UB Ultimo a o')
ylabel('F Empirica')
xlabel('Temperatura')
ylim([0 1])
title('Bandas no parametricas')
hold off
```

11.3. Anexo punto 3

```
n = 10000;
lambda = 1;
X = sort(-(1/lambda)*log(rand(n,1)/lambda));

F = (1-exp(-lambda*X));
[F_est, x_est] = ecdf(X);
F_est = F_est(1:end-1);
x_est = x_est(1:end-1);

figure
hold on
plot(X,F)
plot(x_est,F_est)
legend('Teorica','Empirica')
title(strcat('n=',num2str(n)))
hold off

GlivCant = max(abs(F_est-F));
disp(strcat('Glivenko-Cantelli: ',num2str(GlivCant)))
```

11.4. Anexo punto 8

```
load('temperaturas.mat')

n = 365;
X = temperatura(:,1);
minimo = min(X);

%Sesgo
Bias =Bias_min(X);
disp(strcat('Sesgo:',num2str(Bias)))

%Varianza
Ssq_tilde = 0;
for i=1:n
    T_tilde_barr = 0;
    for j=1:n
        T_tilde_barr = T_tilde_barr + Ti_tilde(X,j);
    end
    T_tilde_barr = T_tilde_barr/n;
    Ssq_tilde = Ssq_tilde + (Ti_tilde(X,i) - T_tilde_barr)^2;
end
Ssq_tilde = Ssq_tilde/(n-1);

Varianza = Ssq_tilde/n;
SE = sqrt(Varianza);

%Intervalo de Confianza:

Lx = minimo-1.96.*SE;
Ux = minimo+1.96.*SE;

disp(strcat('Minimo Estimado: ',num2str(minimo)))
disp(strcat('Intervalo de confianza (Usando varianza de Jackknife): ',
    num2str(Lx),', ',num2str(Ux)))
```

```
function bias = Bias_min(X)
Tn_barr = 0;
[n y] = size(X);
Tn = min(X);

for i=1:n
Xi = X;
Xi(i) = [];
Ssq = min(Xi);
Tn_barr = Tn_barr + Ssq;
end

Tn_barr = Tn_barr/n;
bias = (n-1)*(Tn_barr-Tn);
end
```

11.5. Anexo punto 9

```
rng(19)
n=100;
U = rand(n,1);

%Varianza Boostrap
B = 1000;
EB = bootstrp(B,@min,U);

Vboots = 0;
for b=1:B
    Vboots = Vboots + (EB(b,1)-mean(EB))^2;
end

Vboots = Vboots/B;

%Sesgo
Bias = Bias_min(U);
Bias_teorico = 1-(n/(n+1));

disp(strcat('Varianza Bootstrap:',num2str(Vboots)))
disp(strcat('Sesgo Jackknife:',num2str(Bias)))
disp(strcat('Sesgo Teorico:',num2str(Bias_teorico)))
```

11.6. Anexo punto 10

```
%% Comedian y Datos
comedian = @(X,Y) median((X-median(X)).*(Y-median(Y)));
load('temperaturas.mat')
%% Componentes principales
X = [temperatura(:,1) temperatura(:,35)];

v = cov(X);
[Coef,latent,explained] = pcacov(v);

z1 = Coef(:,1);
z2 = Coef(:,2);

%Version robusta de la matriz de covarianzas

covR = zeros(2,2);
for i=1:2
    for j=1:2
        covR(i,j) = Comedian(X(:,i),X(:,j));
    end
end

%PCA Robusto

[CoefR,latentR,explainedR] = pcacov(covR);

```

```
zr1 = CoefR(:,1);
zr2 = CoefR(:,2);

%% Contaminaci n de los datos
X = [X ; (X(28:35,:)+ 100) ];

v = cov(X);
[Coef,latent,explained] = pcacov(v);

z1 = Coef(:,1);
z2 = Coef(:,2);

%Version robusta de la matriz de covarianzas

covR = zeros(2,2);
for i=1:2
    for j=1:2
        covR(i,j) = Comedian(X(:,i),X(:,j));
    end
end

%PCA Robusto

[CoefR,latentR,explainedR] = pcacov(covR);
```