# UNA PROPUESTA DE EVALUACION PARA CONJUNTOS DE DATOS CON INFORMACION FALTANTE EN LA ENCUESTA PERMANENTE DE HOGARES

Badler, Clara; Alsina, Sara; Pagano, Ariel H.; Puigsubirá, Cristina; Vitelleschi, Ma. Susana

*Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Fac. Cs. Económicas y Estadística. Universidad Nacional de Rosario.*[1]

## INTRODUCCION

El análisis de conjuntos de datos con información faltante requiere tratamiento de la misma, condicionado al planteo de ciertos supuestos. La metodología más generalizada supone que: " la información perdida tiene las mismas características que la observada".

Una forma estadística operacional de evaluación de este supuesto, traduce el comportamiento de la variable observada en forma completa y con pérdida en términos de distribuciones de probabilidad. A tal fin, las pruebas a distribución libre presentan características prácticas tanto para investigadores como usuarios.

Se propone la aplicación de las pruebas de Wilcoxon y Kolmogorov para evaluar las hipótesis de no existencia de diferencias entre los datos observados y los perdidos, como acercamiento a la determinación de la aleatoriedad de la pérdida.

## MATERIAL

*BASE DE DATOS*
Encuesta Permanente de Hogares
Aglomerado Gran Rosario
Ondas Octubre 97 - Octubre 98

→ *SUB-BASE* Personas Desocupadas → *VARIABLES*

- **ITF** (Monto del Ingreso Total Familiar)

  ITF = 0 no siempre implica ITF nulo, entonces se considera:

  - Grupo de Información Faltante ***(GIF)***: Desocupados con ITF = 0
  - Grupo de Información Completa ***(GIC)***: Desocupados con ITF ≠ 0

- **EDAD** (En años cumplidos). Observada completamente

## METODOLOGIA

**ANALISIS DESCRIPTIVO DE LA VARIABLE EDAD EN GIF y GIC**

Medidas resúmenes:

media (Ma)

modo (Mo)

mediana (Mna)

desvío standard (Sd)

coeficiente de asimetría (As).

**PRUEBA DE WILCOXON (W)**

*Plantea la hipótesis de que dos muestras aleatorias pueden ser pensadas como una sola muestra de una población*

$H_0$: F(t) = G(t) $\forall$ t

F y G: Funciones de distribución de las muestra (de tamaño m y n) correspondiente a las poblaciones 1 y 2

$H_1$: G (t) = F (t - $\Delta$) Modelo de traslación

Para muestras grandes y bajo normalidad asintòtica de W, se rechaza $H_0$ si:

$$|W^*| \geq Z_{\alpha/2}$$

siendo:

$$W^* = \frac{W - \{ n ( m + n + 1 ) / 2 \}}{\{ m . n ( m + n + 1 ) / 12 \}^{1/2}}$$

donde W es la suma de los rangos asignados a una de las muestras

**PRUEBA DE KOLMOGOROV (D)**

*Plantea la hipótesis de que la función de distribución de una muestra corresponde a una determinada familia paramétrica con parámetros no siempre especificados*

$H_0$: $F(x) = \hat{F}_0(x)$ $\forall$ x

$H_1$: $F(x) \neq \hat{F}_0(x)$ en al menos una x

Se rechaza $H_0$ si:

$$D > \frac{1.36}{\sqrt{n}}$$

para un nivel de significación del 5% y un n > 30

Siendo D:

$$D = \sup_{-\infty < X < +\infty} \left\{ \left| F_n(x) - \hat{F}_0(x) \right| \right\}$$

$F_n(x)$: función de distribución muestral $\sup_{-\infty<x<+\infty}$: valor máximo de las diferencias en valor absoluto entre la distribución empírica y la propuesta.

Se utilizan la distribución Normal ($\mu,\sigma^2$) y Gamma ($\alpha$=forma y $\beta$=escala).

## RESULTADOS

Cuadro 1. Distribución de los desocupados según grupo y onda

| | Grupo de Información | | |
|---|---|---|---|
| Onda | GIF | GIC | Total |
| Oct-97 | 33 (16.1%) | 172 (83.9%) | 205 (100%) |
| Oct-98 | 40 (23.5%) | 170 (76.5%) | 210 (100%) |

Cuadro 2. Medidas descriptivas de la variable EDAD según grupo y onda

| Onda | Oct - 97 | | Oct - 98 | |
|---|---|---|---|---|
| Medida | GIF | GIC | GIF | GIC |
| Ma | 37.9 | 31.7 | 34.8 | 33.9 |
| Mo | 19 y 43 | 19 y 21 | 19 | 18 |
| Mna | 43 | 29 | 29 | 33 |
| Sd | 16.5 | 13.3 | 15.6 | 15.1 |
| As | 0.16 | 0.7 | 0.59 | 0.62 |

**Octubre 97:** comportamiento variable EDAD varía entre grupos. Grupo GIC marcada asimetría positiva.
**Octubre 98:** EDAD posee similar ubicación y forma en los grupos GIF y GIC.

Cuadro 3. Aplicación de la prueba de Wilcoxon

| Onda | W* | Prob. Asociada | |
|---|---|---|---|
| Oct-97 | 1.8 | 0.07 | **(R)** |
| Oct-98 | 0.255 | 0.8 | **(NR)** |

$\alpha$=0.1 R: Rechazo de la $H_0$, NR: No rechazo de la $H_0$

**Octubre 97:** distribución variable EDAD grupo GIF desplazada con respecto al grupo GIC.
**Octubre 98:** esto no ocurre.

Cuadro 4. Aplicación de la prueba de Kolmogorov (Valores de la estadística D)

| Onda | Oct - 97 | | Oct - 98 | |
|---|---|---|---|---|
| Distribución | GIF | GIC | GIF | GIC |
| Normal | 0.159 (NR) | 0.115 (R) | 0.09 (NR) | 0.135 (R) |
| Gamma | 0.174(NR) | 0.08 (NR) | 0.062 (NR) | 0.093 (NR) |
| Valores Críticos | | | | |
| $1.36/\sqrt{n}$ | 0.237 | 0.104 | 0.215 | 0.104 |

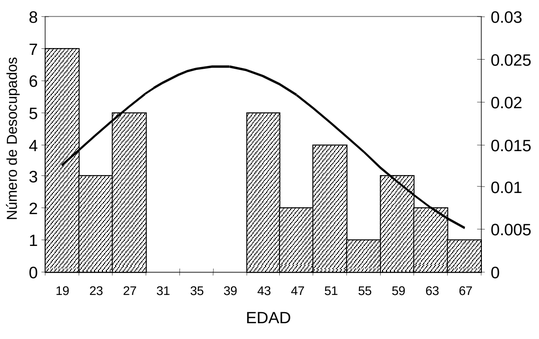
Distribuciones con menor discrepancia entre la distribución empírica y la propuesta:

**Octubre 97**
**Grupo GIF:** Normal (37.9; 16.5)
**Grupo GIC:** Gamma (6.06; 5.24)

**Octubre 98**
**Grupo GIF:** Gamma (5.1; 6.8)
**Grupo GIC:** Gamma (5.3; 6.4)

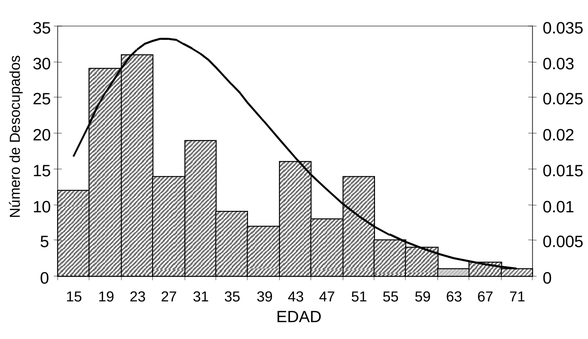
Gráfico 1. Distribución de la variable EDAD y ajustes propuestos según grupo y onda

GIF - Octubre 1997



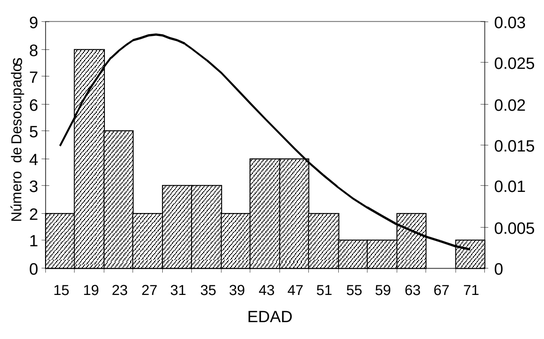

: Distribución Normal (37.9 ; 16.5)

GIC - Octubre 1997



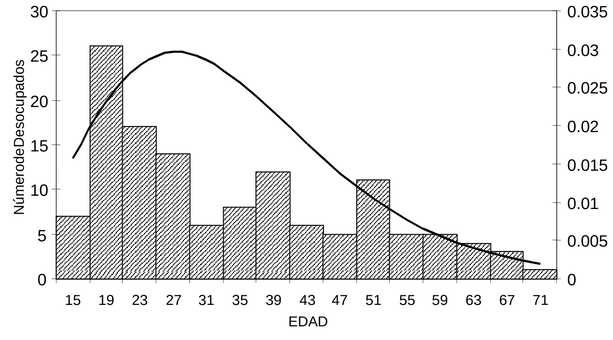

: Distribución Gamma (6.06 ; 5.24)

GIF- Octubre 1998




: Distribución Gamma (5.1 ; 6.8)

GIF- Octubre 1998




: Distribución Gamma (5.3 ; 6.4)

La EDAD de los desocupados pertenecientes al grupo GIF difiere en forma y ubicación con respecto al grupo GIC, o sea que las unidades de ambos grupos no provendrían de la misma población.

El comportamiento de la EDAD en ambos grupos es similar; las unidades de ambos grupos provendrían de la misma población.

## DISCUSION

La aplicación de herramientas basadas en métodos a distribución libre para evaluar conjuntos de datos con información perdida permite, de manera ágil y sin la necesidad de realizar supuestos distribucionales, acceder a una forma de evaluación diferente del comportamiento de los grupos con información completa y faltante, comparando tanto la ubicación como las posibles funciones de distribución de las cuales provendrían y orientando las conclusiones hacia la posibilidad de considerar la aleatoriedad de la pérdida.

El usuario de estas herramientas deberá evaluar la conveniencia de las mismas según la aplicación que se realice y ser cauteloso a la hora de extraer conclusiones ya que, en algunos casos, es posible encontrar evidencias poco claras de desplazamiento o más de una función de distribución que resulte no significativa.

[1] Proyectos:PICT Nº 0200095-01996 de la ANPyT y PID Nº 19/E045 de la SECyT