Ing. V. Luis Foglia
Lic. Est. Griselda M. Negri

# ESTIMACION e INTERVALOS DE CONFIANZA

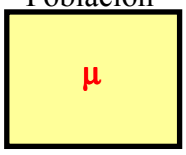
## (a) Para medias

Supóngase que en una población de jóvenes de 15 a 30 años de edad, interesa estudiar el nivel de glucosa en sangre.
Sin embargo, como los *NIVELES* de glucosa varían de joven a joven, parece lógico utilizar el promedio de todos ellos, como un indicador de nivel de glucosa en la población de jóvenes.

O sea si:

N = cantidad de jóvenes en la población
$X_1$ = nivel de glucosa en joven Nro. 1
$X_2$ = nivel de glucosa en joven Nro. 2
$X_3$ = nivel de glucosa en joven Nro. 3
´
`
$X_N$ = nivel de glucosa en joven Nro. N

Población




$$\mu = \frac{X_1 + X_2 + ..... + X_N}{N}$$

Media o promedio, poblacional

Este promedio, calculado con los niveles de glucosa de TODOS los jóvenes de la población, es la media o promedio poblacional que llamaremos μ.

Este es el valor que desearía conocer el investigador, y sobre el que debe elaborar las conclusiones de su trabajo.

Sin embargo, difícilmente se disponga de recursos económicos y de tiempo, como para determinar niveles de glucosa en todos los miembros de una población.

Pero supongamos una situación más real: el investigador decide tomar al azar una muestra de 30 jóvenes de la población, y determina en cada uno de ellos el nivel de glucosa en sangre resultando:

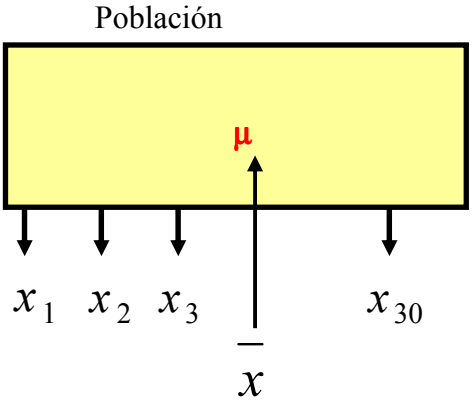
Joven Nº 1 $x_1$ =89
Joven Nº 2 $x_2$ =77
79

| | |
|---|---|
| | 85 |
| | 90 |
| | 87 |
| | 80 |
| | 81 |
| | 102 |
| | 87 |
| | 89 |
| Joven Nº 30 | $x_{30}$ =76 |

Y con estos datos calcula la media (que representamos como "$\bar{x}$" y el desvío estándar muestral, así:

$$\bar{x} = \frac{x_1 + x_2 + .... + x_{12}}{30} = \frac{89 + 77 + .....76}{30} = 85{,}17$$ Media muestral

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + .....(x_{30} - \bar{x})^2}{30 - 1}} = 7{,}23$$ Desvío Estándar Muestral




Esta media de $\bar{x}$ = 85,17, fue calculada con solo 30 jóvenes de la muestra de aquí que se la denomine media muestral para diferenciarla de μ, que se la calcula con la totalidad de jóvenes de la población.

¿Coincidiría $\bar{x}$ con μ?

En general no, ya que $\bar{x}$ = 85,17 fue calculada con solo 30 individuos de la población.

Sin embargo $\bar{x}$ es un "estimador" de μ.

¿Qué quiere decir esto?
Significa que aunque no coincida con **μ**, $\bar{x}$ estará "cerca" de **μ**.

Pero, como evaluar que tan "cerca" esta $\bar{x}$ = 85,17 de la verdadera media **μ**, que desconocemos?

Esto se logra calculando el error estándar del estimador $\bar{x}$, cuya expresión es:

$$\text{Error}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{7,23}{\sqrt{30}} = 1,32$$

Error estándar de la media muestral.

Esta cantidad debe entenderse como el error promedio de $\bar{x}$ al estimar **μ**.
Ahora bien, como las conclusiones de un trabajo deben referirse a la verdadera media de la población **μ**, y no a estimaciones que contienen errores, informaremos para **μ**, no un solo valor, sino un intervalo que lo contiene.

Si deseamos un intervalo que contiene a **μ**, con un 95% de probabilidad, este se calcula así:

$$\left[\bar{x} - 1,96\, error\ (\bar{x}) \quad ; \quad \bar{x} + 1,96\, error\ (\bar{x})\right]$$

Intervalo de confianza al 95%.

O sea : $[85,17 - 1,96 * 1,32 ; 85,17 + 1,96 + 1,32]$

$[82,6 ; 87,8]$ al 95%

Esto debe interpretarse como "la verdadera media esta entre 82,6 y 87,8 con una seguridad de 95%".

Si se quiere un intervalo que nos dé mejor seguridad de incluir a **μ**, por ejemplo del 99% solo debemos cambiar el 1,96 que figura en la fórmula anterior por 2,58, y el intervalo será:

$[85,17-2,58 * 1,32 ; 85,17+2,58 * 1,32]$ Intervalo de confianza al 99%.

[ 81,76 ; 88,58 ] al 99 %

necesariamente más amplio que el anterior ya que tiene mayor seguridad de contener a la media.

**(b) Para prevalencias**

Supóngase ahora, que en una población de adultos de una localidad, queremos investigar el consumo de tabaco.
Concretamente, en cada individuo interesa estudiar un atributo que tiene solo dos alternativas:

FUMA NO FUMA

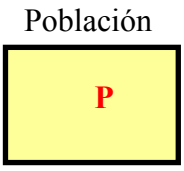
Notar que esta situación es diferente a la del ejemplo anterior donde en cada joven interesaba el nivel de glucosa, que tiene muchas alternativas:

83,2 75 104 93,5 etc.

Debido a lo anterior, como indicador del consumo de tabaco utilizaremos la prevalencia de los fumadores en la localidad así:

N ⟶ cantidad de adultos en la población
F ⟶ cantidad de fumadores en la población.

Luego: $\mathbf{P} = \frac{F}{N} * 100$ Prevalencia poblacional se representa por **P** mayúscula.



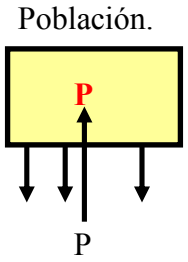

Esta prevalencia, al calcularse con la totalidad de individuos, es la "prevalencia poblacional". Sería el valor que quiere conocer el investigador, y sobre el que debe elaborar las conclusiones de su trabajo.

Pero, como ya se mencionó anteriormente, difícilmente se dispongan en la práctica de recursos suficientes como para encarar un relevamiento de toda la población.

Supóngase entonces que el investigador decide tomar una muestra al azar de 225 adultos, y encuentra que 90 son fumadores.

Entonces: $n = 225$ cantidad de individuos en la muestra.
$f = 90$ cantidad de fumadores en la muestra.

$$p = \frac{f}{n} * 100 = \frac{90}{225} * 100 = 40\ \%$$ Prevalencia Muestral




Esta prevalencia del 40%, al ser calculada con sólo los 225 adultos de la muestra se denomina "prevalencia muestral" $p$, para diferenciarla de la "prevalencia poblacional" **P**, que se calcula con toda la población.

Al igual que en el ejemplo anterior $p$ = 40% no necesariamente coincidirá con el valor **P** desconocido.

Sin embargo p = 40% es un "estimador" de **P**, en el sentido que, aunque no coincida estará "cerca".

Para evaluar que tan cerca estará $p$ = 40% de la verdadera prevalencia poblacional, calculamos el "error estándar" de $p$ así:

$$\text{error}(p) = \sqrt{\frac{p(100-p)}{n}}$$ Error estándar de la prevalencia muestral.

$$= \sqrt{\frac{40(100-40)}{225}} = \quad 3{,}26\ \%$$

Esta cantidad debe interpretarse como que el error promedio de $p$ es 3,26 %.
Ahora bien, como las conclusiones de un trabajo deben referirse a la verdadera prevalencia poblacional **P**, y no a estimaciones que contienen errores, informaremos para **P**, no un solo valor, sino un intervalo que la contiene.
Si deseamos un intervalo que contenga a P, con un 95 % de probabilidad, este se calcula así:

$$[p - 1{,}96 error(p)\ ;\ p + 1{,}96\ \text{error (p)}]$$ Intervalo de confianza al 95%

O sea: $[40 - 1{,}96 * 3{,}26\ ;\ 40 + 1{,}96 * 3{,}26]$

$[33{,}6\ \%\ ;\ 46{,}4\ \%]$

Esto debe interpretarse como que la verdadera prevalencia del consumo de tabaco esta entre

33,6 y 46,4 % con una seguridad del 95%.

Además, si se desea un intervalo pero con una seguridad de 99%, solo debe modificarse en la expresión anterior, el 1,96 por el 2,58, resultando.

$$[40 - 2{,}58 * 3{,}26 \; ; \; 40 + 2{,}58 * 3{,}26]$$

[ 31,6 % ; 48,4 % ] al 99%

Nótese que al ser un intervalo más seguro, es también más amplio.

**(c) Para riesgo relativo**

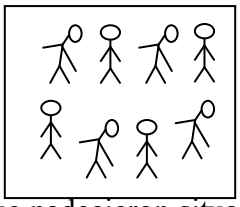
Supóngase que en una localidad se desea estudiar la relación entre situaciones de violencia familiar y el consumo de marihuana en jóvenes de 15 a 30 años.

Aquí, en cada joven interesan conocer las siguientes dos variables de atributos:

V ⟶ ¿Existen o existieron en su hogar situaciones de violencia familiar? SI – NO

M ⟶ ¿Consume o ha consumido marihuana habitualmente? SI - NO

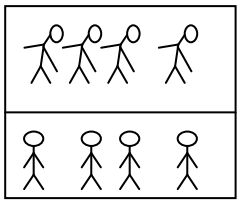
Podemos representar la población por un cuadrado, donde N es la cantidad total de jóvenes.



N (Total de jóvenes)

Como en la población hay jóvenes que padecieron situaciones de violencia familiar, y otros que no, separémosla en dos subpoblaciones, las de los:

Jóvenes **con** situaciones de Violencia Familiar ⟶ $N_v$ jóvenes.
Jóvenes **sin** situaciones de Violencia Familiar ⟶ $N_{\bar{v}}$ jóvenes.




$N_v$ (con violencia familiar)

$N_{\bar{v}}$ (sin violencia familiar)

Ahora y suponiendo que se dispone de una cantidad ilimitada de recursos, clasifiquemos los jóvenes de cada grupo según su consumo de marihuana.

Se tendrá la población particionada así:

| | $M$ | $\overline{M}$ | |
|---|---|---|---|
| $V$ | A | B | $N_v$ |
| $\overline{V}$ | C | D | $N_{\overline{v}}$ |

En la subpoblación que padeció situaciones de violencia familiar.
A⟶ Cantidad de consumidores de marihuana.
B⟶ Cantidad de NO consumidores de marihuana.

Por supuesto A+B = $N_v$

En la subpoblación que **NO** padeció violencia familiar.

C⟶ Cantidad de consumidores de marihuana.
D⟶Cantidad de NO consumidores de marihuana.

Por supuesto C+D = $N_{\overline{v}}$

Planteadas las cosas así, como indicador del consumo de marihuana, se calculará su prevalencia en cada uno de los grupos:

$P_V = \frac{A}{N_V} * 100$ Prevalencia de consumo de marihuana, en los que padecieron situaciones de violencia familiar.

$P_{\overline{V}} = \frac{C}{N_{\overline{V}}} * 100$ Prevalencia de consumo de marihuana, en los que **NO** padecieron situaciones de violencia familiar.

Estas dos prevalencias, al ser calculadas con la totalidad de jóvenes, son prevalencias poblacionales.

| | $M$ | $\overline{M}$ | | |
|---|---|---|---|---|
| V | A | B | $N_v$ | $P_V = \frac{A}{N_V} * 100$ |
| $\overline{V}$ | C | D | $N_{\overline{v}}$ | $P_{\overline{V}} = \frac{C}{N_{\overline{V}}} * 100$ |

En un caso como el presente, es de esperar una mayor prevalencia de consumo; en el grupo que padeció violencia familiar, respecto del otro. ( $P_V > P_{\overline{V}}$ ).

Como indicador del RIESGO que aportan las situaciones de violencia familiar sobre el consumo de marihuana se define el RIESGO RELATIVO así:

$$\mathbf{RR} = \frac{P_V}{P_{\overline{V}}}$$ Riesgo relativo poblacional.

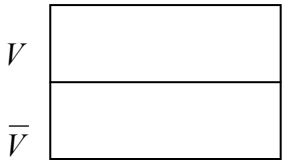
Este riesgo relativo poblacional es el que mide si hay asociación entre situación de violencia familiar y el consumo de marihuana.
Si es mayor que 1, indica una mayor prevalencia del consumo de marihuana entre los que padecieron violencia familiar respecto de las que no pasaron por esta situación. Esto se interpreta como que la violencia familiar es un **factor de riesgo** respecto del consumo de marihuana. **EXISTE ASOCIACION.**

Si **RR** = 1, indica que la prevalencia de consumo es igual en los dos grupos, interpretándose como una **NO EXISTENCIA DE ASOCIACIÓN** entre situaciones de violencia familiar y consumo.

Si **RR** <1 indicaría un menor consumo de marihuana entre quienes pasaron por situaciones de violencia familiar, respecto del otro grupo. Esto se interpretaría, aunque no es de esperar que ocurra en este caso, como que la violencia familiar es un **factor protector** contra el consumo de marihuana. **EXISTE ASOCIACIÓN.**
Todo lo anterior fue elaborado sobre la población.
Supóngase ahora que el investigador dispone de recursos limitados, y resuelve encarar el estudio a través de una muestra al azar de 1000 jóvenes, interrogando a cada uno de ellos sobre las variables mencionadas.
Ahora la muestra la representamos así

n = 1000 jóvenes

Supóngase que de estos jóvenes, 240 pasaron por situaciones de violencia familiar, quedando la muestra separada en dos grupos.




$n_v$ = 240 con violencia familiar

$n_{\overline{v}}$ = 760 sin violencia familiar

n = 1000 jóvenes

Y además supóngase que 18 del primer grupo, y 19 del segundo son o han sido consumidores de marihuana.
Queda entonces para la muestra:

| | $M$ | $\overline{M}$ | |
|---|---|---|---|
| $V$ | 18 | 222 | 240 con violencia familiar |
| $\overline{V}$ | 19 | 741 | 760 sin violencia familiar |

Las prevalencias en cada grupo son:

$$p_V = \frac{18}{240} * 100 = 7{,}5\ \%$$ (Prevalencias muestrales)

$$p_{\overline{V}} = \frac{19}{760} * 100 = 3{,}5\ \%$$

Luego el riesgo relativo "muestral" es:

$$\text{rr} = \frac{p_V}{p_{\overline{V}}} = \frac{7{,}5\%}{2{,}5\%} = 3$$ Riesgo relativo muestral

Que este riesgo valga 3, significa que en situaciones de violencia familiar se triplica la prevalencia del consumo de marihuana.
Pero este riesgo rr = 3, al ser calculado solo con los 1000 jóvenes de la muestra, es el "riesgo relativo muestral", para diferenciarlo del **RR** "riesgo relativo poblacional" que se calcula con toda la población.
Como se dijo anteriormente, las conclusiones de un trabajo deben referirse al verdadero riesgo poblacional **RR**, y no va a estimaciones como rr = 3, que contiene errores.
Por eso, se informara para **RR**, no un solo valor, sino un intervalo que lo contiene.
A diferencia de los ejemplos anteriores, el calculo de un intervalo de confianza para **RR** requiere de ciertas complicaciones matemáticas.

Por eso aquí recurrimos al EPI-INFO 6.0.

Selecionamos: 1 - Programas.
2 - Statcalc calculador
3 - Tables (2 * 2,2 * n)

Y en la cuadrícula.

| | | Disease | |
|---|---|---|---|
| E x P O S U R E | | | - |
| | | | |

Se ingresan los valores muestrales:

| | | Disease | |
|---|---|---|---|
| | | + | - |
| + | | 18 | 222 |
| - | | 19 | 741 |

y apretando F4 resulta:
Un riesgo relativo rr =3

y un intervalo de confianza para **RR**,

[1,60;5,62] al 95 %

Esto debe interpretarse como que “el verdadero riesgo relativo en la población esta entre 1,6 y 5,62 con una seguridad del 95%”.