¿Qué significa que las diferencias sean estadísticamente significativas?

El 8 de mayo del 2019 la versión online de diario regional El Día de la región de Coquimbo publicó una noticia con el siguiente titular ”*Encuesta Nacional Urbana de Seguridad Ciudadana. Cae la victimización y la percepción de inseguridad en la Región de Coquimbo*” http://www.diarioeldia.cl/region/cae-victimizacion-percepcion-inseguridad-en-region-coquimbo

Más adelante, en el cuerpo del reportaje, se menciona lo siguiente: “*De acuerdo a los datos entregados por el estudio realizado por la Subsecretaría de Prevención del Delito, la victimización bajó en la Región de Coquimbo en un 4,5%, del 23,4% al 18,9%;* ***variación que es estadísticamente significativa…***”

¿Qué significa que una variación o diferencia sea estadísticamente significativa? Ya que si le preguntamos a cualquier persona esta nos responderá que ¡claro que 23,4% es diferente de 18,9%!

Debemos tener en cuenta que estos resultados provienen de encuestas que no son aplicadas a toda la población, es decir, que se aplican a una muestra o subconjunto de ella y por ende los resultados tienen asociado un margen de error. Esto significa, por un lado que lo que se obtiene es una estimación del índice de victimización y por otro, que si la victimización es del 23,4% podría haber sido, por ejemplo, 24% o 23%.

Solo se puede conocer el valor “real”, que en estadística se conoce como el parámetro, del índice de victimización si se hubiese hecho la encuesta a toda la población.

Cuando intento explicar a mis alumnos esta materia les doy el siguiente ejemplo. A final de semestre siempre existe al alumno que termina el curso con 3,9 y se acerca y te dice “¡Profe, 3,9 es lo mismo que 4,0!”. Por ahí uno un poco más atrevido te dice… “¡Profe!... ¡3,8 es lo mismo que 4,0! Y hay algunos aún más atrevidos que te dicen… "Pofe… ¿3,7 es lo mismo que 4,0? El punto es, ¿a partir de qué valor "tres coma algo" ya no es lo mismo que 4,0?

(Que quede por escrito, que para efectos prácticos, para mí a final de semestre 4 es 4, 3,9 es 3,9… y así sucesivamente.)

Retomando el ejemplo anterior, en este caso la diferencia entre la medición del 2018 y 2017 es 4,5% y es considerada estadísticamente significativa. Sin invertimos la pregunta, es decir, nos preguntamos **cuándo la diferencia no es estadísticamente significativa**, tenderíamos a responder que la diferencia no es estadísticamente significativa cuando esta se aproxima a 0… pero, ¿qué tanto?

En estadística existen un par de metodologías para detectar diferencias estadísticamente significativas entre dos estimaciones. En esta sección nos enfocaremos en los intervalos de confianza para la diferencia de proporciones (porcentajes) o medias según sea el caso. En esta oportunidad abordaremos el concepto de *Intervalo de Confianza para la Diferencia de Proporciones* o *Intervalo de Confianza para la Diferencia de Medias*.

Antes de hablar acerca del intervalo de confianza debemos entender lo que en estadística se conoce como parámetro.

Un parámetro es cualquier medida resumen de la población. La media, varianza y proporción son los parámetros más usuales. Estos se calculan consideran los datos de toda la población. De lo contrario lo que se obtiene son *estimaciones* de los parámetros poblacionales. Cuando comparamos dos poblaciones ya sea a través de proporciones o medias, los parámetros ahora son la diferencia de las proporciones o la diferencia de las medias.

Por ejemplo, sea llamamos p_1 a la proporción poblacional de victimización del 2018 y p_2 a la proporción poblacional de victimización del 2017. La estimación de p_1, $\hat{p}_1$ es 23,4% y la estimación de p_2, $\hat{p}_2$ es 18,9%. La diferencia poblacional entre el índice de victimización entre el 2018 y 2109 es $p_1 - p_2$. La estimación de la diferencia poblacional es $\hat{p}_1 - \hat{p}_2$, es decir, $23{,}4\% - 18{,}9\% = 4{,}5\%$.

Un intervalo de confianza es, como su nombre lo indica, un intervalo o rango de valores entre los cuales puede oscilar el valor de un parámetro. Un intervalo de confianza para la diferencia de proporciones será un intervalo o rango de valores entre los que oscila el valor de la diferencia de proporciones con un cierto nivel de confianza o seguridad que usualmente se toma como 95%.

En otras palabras un intervalo de confianza nos da un rango de valores plausibles para el parámetro. Si el intervalo de confianza para la diferencia de proporciones contiene al 0, podremos decir que 0 es un valor plausible para esta diferencia y por lo tanto esta diferencia NO es estadísticamente significativa. Por otro lado, si este intervalo NO contiene al 0, podremos afirma que con un 95% de confianza la diferencia entre ambas proporciones es estadísticamente significativa.

La fórmula para el intervalo de confianza para la diferencia de proporciones con un 95% de confianza esta dada por

$$(\hat{p}_1 - \hat{p}_2) - 2 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \leq p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) + 2 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Donde $n_1$es el tamaño de muestra de la primera población y n_2 es el tamaño de muestra de la segunda población.

Así, $(\hat{p}_1 - \hat{p}_2) - 2 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ es el límite inferior del intervalo de confianza del 95% y

$(\hat{p}_1 - \hat{p}_2) + 2 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ es el límite superior

del intervalo de confianza del 95%.

El término $2 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ corresponde al error de estimar $p_1 - p_2$ por $\hat{p}_1 - \hat{p}_2$ con un 95% de confianza.

Si consideremos que el número de casos encuestados el 2018 fue de 1.588 y el 2017 fue de 1.744 (Fuente: https://www.ine.cl/estadisticas/sociales/encuesta-de-seguridad-ciudadana) se obtiene que el intervalo de confianza del 95% para la diferencia del índice de victimización entre 2018 y 2017 esta dado, aproximadamente por,

$0{,}014 \leq p_1 - p_2 \leq 0{,}076$ o equivalentemente por $1{,}4\% \leq p_1 - p_2 \leq 7{,}6\%. (*)$ En otras palabras con un 95% de confianza la diferencia entre el índice de victimización del 2018 y el 2107 en la región de Coquimbo esta entre 1,4% y 7,6%. Esto quiere decir que la diferencia "real", con un 95% de confianza, esta entre 1,4% y 7,6% por lo que 0 no es un valor plausible y entonces la diferencia es estadísticamente significativa.

Si en vez de hablar de proporciones o porcentajes hablamos de medias o promedios, ¿cómo podemos saber si la diferencia entre dos medias es estadísticamente significativa?

Consideremos la nota publicada el 25 de junio de 2018 titulada así: "*INE: Gasto promedio de los hogares de las principales ciudades de Chile supera el millón de pesos*" (https://www.ahoranoticias.cl/noticias/nacional/228442-las-demandas-tras-el-ultimatum-de-funcionarios-de-gendarmeria-al-gobierno.html)

En el cuerpo del artículo se menciona lo siguiente: "*La encuesta evidenció, además, que el gasto promedio mensual del hogar* ***varía notoriamente según el nivel educacional del sustentador principal de este.*** *De esta forma, en aquellos hogares donde el nivel educacional del sustentador principal llegó hasta la primaria (sistema antiguo), el gasto asciende a $545.881, y en los que tienen un sustentador principal con enseñanza básica, el gasto es de $666.537.*"

Aquí se habla de que existen diferencias entre el gasto promedio mensual según el nivel educacional del sustentador principal, pero no menciona si estas son estadísticamente significativas.

Veamos el caso particular de la comparación de aquellos hogares donde el nivel educacional del sustentador principal llegó hasta primaria y de aquellos en que el sustentador principal llegó hasta enseñanza básica.

Llamaremos $\boldsymbol{\mu_1}$ al gasto promedio mensual de los hogares donde el nivel educacional del sustentador principal llegó hasta enseñanza básica. Llamaremos $\boldsymbol{\mu_2}$ al gasto promedio mensual de los hogares donde el nivel educacional del sustentador principal llegó hasta primaria.

Nuestros parámetros de interés son ahora $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$ y $\boldsymbol{\mu_1 - \mu_2}$.

La estimación de $\boldsymbol{\mu_1}$, $\boldsymbol{\hat{\mu}_1}$, es el gasto promedio obtenido a partir del estudio, es decir, \$666.537. De la misma manera la estimación de $\boldsymbol{\mu_2}$, $\boldsymbol{\hat{\mu}_2}$, es el gasto promedio obtenido a partir del estudio, es decir, \$545.881.

La estimación de la diferencia entre el gasto promedio de los hogares donde el nivel educacional de sustentador principal llegó hasta enseñanza básica y los hogares donde el nivel educacional del sustentador principal llegó hasta primaria es $\boldsymbol{\hat{\mu}_1 - \hat{\mu}_2 = 666.537 - 545.881 = 120.656}$. ¿Será estadísticamente significativa esta diferencia? Al igual que en el ejemplo de las proporciones ambas estimaciones tienen un error asociado y podemos calcular un intervalo de confianza para esta diferencia. En este caso la fórmula del intervalo de confianza del 95% para la diferencia de medias esta dada por :

$$(\hat{\mu}_1 - \hat{\mu}_2) - 2 \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\hat{\mu}_1 - \hat{\mu}_2) + 2 \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (**)$$

Así $(\hat{\mu}_1 - \hat{\mu}_2) - 2 \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ es el límite inferior del intervalo de confianza del 95% y $(\hat{\mu}_1 - \hat{\mu}_2) + 2 \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ es el límite superior. El término $2 \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ corresponde al error de estimar $\mu_1 - \mu_2$ por $\hat{\mu}_1 - \hat{\mu}_2$ con un 95% de confianza.

s_1^2 corresponde a la varianza muestral, es decir a la varianza calculada a partir de la muestra, del gasto por hogar donde el nivel educacional de sustentador principal llegó hasta educación básica. Análogamente s_2^2 corresponde a la varianza muestral del gasto por hogar donde el nivel educacional de sustentador principal llegó hasta educación básica. n_1 es el tamaño de muestra de la primera población y n_2 es el tamaño de muestra de la segunda población.

Usando la base de datos de la Encuesta de Presupuestos Familiares (EPF) que se encuentra en https://www.ine.cl/estadisticas/ingresos-y-gastos/epf se obtiene que $s_1^2 = 599.148$ y $s_2^2 = 461.978$. Además se obtiene que $n_1 = 401.583$ y $n_2 = 218.949$. (***).

Reemplazando se obtiene que con un 95% de confianza, la diferencia entre el gasto medio de los hogares donde el nivel educacional del sustentador principal llegó hasta educación básica y el gasto medio de los hogares donde el nivel educacional llegó hasta primaria, esta entre \$117.922 y \$123.390. Dado que el 0 no se encuentre en este intervalo, podemos concluir entonces que existe una diferencia estadísticamente significativa entre el gasto promedio de ambos grupos.

(*) Dada la metodología del ENUSC la fórmula del cálculo del intervalo de confianza para la diferencia de proporciones es más compleja que lo que se presenta en el texto. La fórmula aquí presentada es la que se usa en cursos básicos de inferencia estadística.

(**) Cuando las varianzas se desconocen, pueden darse dos casos, se asumen iguales o se asumen distintas. Por simplicidad para el lector aquí solo se aborda la situación donde se asumen varianzas distintas.

(***) Cifras calculadas utilizando el factor de expansión calculado por el INE para este estudio.