De La Cruz-Oré, Jorge Luis
¿Qué significan los grados de libertad?
Revista Peruana de Epidemiología, vol. 17, núm. 2, agosto, 2013, pp. 1-6
Sociedad Peruana de Epidemiología
Lima, Perú

Revista Peruana de Epidemiología,
ISSN (Versión electrónica): 1609-7211
revista@rpe.epiredperu.net
Sociedad Peruana de Epidemiología
Perú

REVISTA PERUANA DE EPIDEMIOLOGÍA
ISSN 1609-7211

RPEonline

ARTÍCULO DE REVISIÓN
REVIEW

¿Qué significan los grados de libertad?

What do degrees of freedom mean?

Jorge Luis De La Cruz-Oré[A,B]

RESUMEN

Los grados de libertad representan un tema central en la estadística moderna, sin embargo su concepto se explica poco en los libros de texto. A pesar de que Gauss los usó por primera vez al estimar las distancias entre las estrellas, no aparece formalmente hasta los escritos de Gosset (Student) en 1908. El concepto de grados de libertad se puede entender desde un punto de vista geométrico, algebraico e incluso intuitivo. La geometría nos describe a los grados de libertad como espacios e hiperespacios de libertad a través de los cuales una medida de resumen puede moverse y tomar diferentes valores. El punto de vista algebraico los describe como el número de ecuaciones que se establecen usando los datos. Ambos puntos de vista están relacionados y ayudan a comprender con mayor profundidad el concepto de grados de libertad. La aplicaciones de los grados de libertad están extendidas a través de toda la estadística, el cálculo de la desviación estándar y la prueba *t* de Student son solo algunos ejemplos.

PALABRAS CLAVE: Grados de libertad, Dimensiones, *t* de student.

INTRODUCCIÓN

Pese a ser un tema muy importante a lo largo de casi todos los capítulos de la estadística, el concepto de grados de libertad (*degrees of freedom* en inglés) es escasamente explicado en los textos de consulta, advirtiendo que se trata de un tema difícil de comprender, de esta manera es pasado por alto o dejado a la posterior revisión del estudiante. Para el matemático que aprende la complicada teoría de los grados de libertad no son necesarias más explicaciones. Sin embargo, para quienes tienen que usar la estadística de manera práctica como los epidemiólogos e investigadores médicos, la mejor comprensión de estos conceptos ayudará a comprender mejor uno de los temas centrales en la estadística.

En la estadística moderna, no se halla el concepto de grados de libertad antes de los trabajos de William Sealy Gosset (*"El probable error de un promedio"*, Student) en 1908,[1] y es por primera vez explicado por Sir Ronald Fisher, el fundador de la estadística moderna en 1915.[1] Sin embargo el concepto ya era familiar a Friedrich Gauss, quien generalizó la teoría de los mínimos cuadrados, muy usados en regresión lineal. Gauss determinó en palabras y en fórmula que al número de observaciones se le debe restar el número de términos desconocidos que se están estimando a partir de los datos para servir como divisor en el cálculo del error estándar de un grupo de datos.[1]

La presente revisión tiene como objetivo dar una idea intuitiva acerca de los grados de libertad y sus fundamentos. Se intenta que sea comprendido por quienes no tienen una formación rigurosa en matemáticas pero que usan la estadística para investigar, enseñar o la están aprendiendo. En las siguientes líneas se presentan primero dos ejemplos ilustrativos y posteriormente se profundiza en el verdadero significado de los grados de libertad. Así mismo, se han sacrificado profundidad y exactitud en favor de la didáctica, y a veces se recurre a repeticiones que ayudarán a una mejor comprensión del tema.

(A) Maestría en Bioestadística, Facultad de Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos. Lima, Perú. (B) CCBR - Lima, Sede Clínica Ricardo Palma.
Correspondencia a Jorge Luis De La Cruz-Oré: dj_jdo@hotmail.com
Recibido el 11 de abril de 2013 y aprobado el 2 de mayo de 2013.
Cita sugerida: De La Cruz-Oré JL. ¿Qué significan los grados de libertad? *Rev peru epidemiol* 2013; 17 (2) [6 pp.]

Ejemplo 1
Si tenemos que escoger a 10 personas de un grupo grande de modo tal que el peso promedio sea de 60 Kg, tenemos la libertad de elegir a los diez que nosotros consideremos. Obviamente pueden existir muchas muestras de diez diferentes personas, pero siempre debemos tener en cuenta que el promedio de los pesos debe ser 60 Kg. Fácilmente nos podemos dar cuenta que solo podemos elegir libremente a las primeras 9 personas, dado que para elegir al décimo este debe ser elegido de manera tal que el promedio del grupo no sea mayor ni menor de 60 Kg. Es decir, podemos elegir con libertad a los 9 primeros, y el décimo queda automáticamente restringido por la condición de que su peso debe ser tal que la media de los diez pesos debe ser 60 Kg. Por lo tanto, para una muestra de 10 personas escogidas al azar, bajo la condición de que la media de los pesos sea 60 Kg, tenemos 9 grados de libertad.

Ejemplo 2
Otro ejemplo ilustrativo podría ser el popular juego japonés Sudoku, en el cual no se puede repetir el mismo número en la misma columna o fila. Si tomamos una columna, y se nos indica que ubiquemos los números del 1 al 9, entonces podemos elegir libremente los primeros 8 números, dado que el noveno numero no puede ser elegido por nosotros, quedando restringido por los 8 números anteriores. El primer número puede ser cualquiera de los nueve, el segundo debe ser elegido de entre los ocho restantes, el tercero debe ser escogido de entre los siete que quedan, y así hasta que solo nos quede elegir entre dos números, tras lo cual, solo nos quedará un último y único número, el cual ya no puede ser elegido libremente. Por lo tanto, tenemos 8 grados de libertad.

Ahora que hemos visto estos dos ejemplos, nos queda formalizar un poco más acerca del concepto de grados de libertad. Aunque tal concepto quedará claro tras leer todo el artículo, podemos decir *"que los grados de libertad son los valores de un conjunto de datos que tienen la posibilidad de variar después de haber impuesto alguna restricción a los mismos"*. Como veremos posteriormente, los grados de libertad son una función del tamaño de la muestra así como del número de variables independientes.[1] Una segunda definición es que *"los grados de libertad son la cantidad de datos que pueden variar independientemente después de que algunas relaciones lineales han sido establecidas entre ellos"*.[2]

Definiciones de Grados de Libertad
Se presentan a continuación algunas definiciones de los grados de libertad en textos de estadística de uso común:

> *"La suma de los valores de las desviaciones de los valores individuales con respecto a su media es igual a cero, hecho que puede demostrarse. Si se conocen los n-1 valores de los valores a partir de la media, entonces se conoce el n-ésimo valor, ya que queda determinado automáticamente debido a la restricción de que todos los valores de n sumen cero"*. (Daniel Wayne 2007, p41)[3]

> *"El valor de los grados de libertad se relaciona con el número de veces que se usa la información de la muestra"*. (Dawson 2005, p91)[4]

> *"Se definen como el número de valores que podemos escoger libremente"*. (Levin 1996, p388)[5]

> *"Los grados de libertad de una prueba estadística son el número de datos que son libres de variar cuando se calcula tal prueba"*. (Pagano 2009, p321)[6]

LOS GRADOS DE LIBERTAD EXPLICADOS

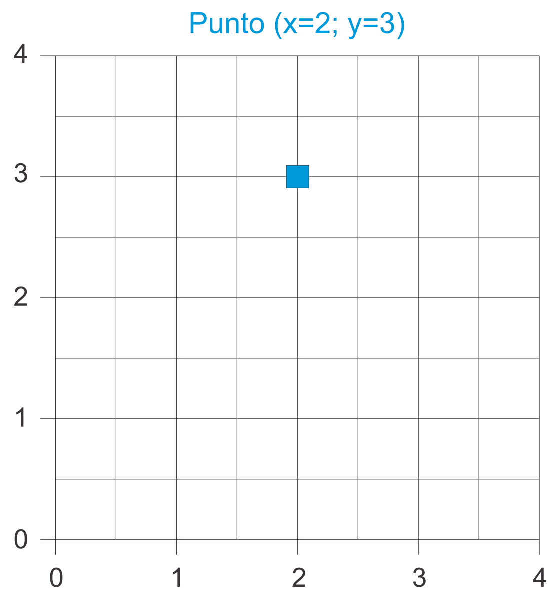
Una aproximación geométrica

Empecemos imaginando un punto fijo (un punto por definición no tiene dimensiones), tal punto no tiene ningún lugar a donde pueda moverse, y por lo tanto no tiene ningún grado de libertad.

Continuemos pensando e imaginemos un punto sobre una recta (la recta tiene una sola dimensión), ahora nuestro punto puede moverse a lo largo de la recta y decimos que tiene un grado de libertad, es decir la libertad de moverse en una dimensión tal y como lo hace un tren sobre las líneas férreas. Podemos conocer su ubicación si sabemos a qué distancia se encuentra de un punto de origen.

Si incrementamos las dimensiones en nuestro ejercicio mental y ubicamos a nuestro punto en un plano (dos dimensiones), nuestro punto puede moverse libremente a lo largo y a lo ancho de tal plano, por lo tanto tiene dos grados de libertad para recorrer sin restricciones, además podemos ubicarlo en un plano cartesiano de tal modo que para conocer su ubicación exacta necesitamos encontrar dos puntos (x; y) a partir del origen del plano. Nuestro ejemplo es similar a pensar en un automóvil sobre un campo de concreto, puede moverse a la derecha-izquierda o en sentido atrás-adelante, siendo posible conocer su ubicación exacta si sabemos las dos coordenadas desde un origen fijo (Figura 1).

FIGURA 1: Representación de un punto en un plano formado por dos dimensiones x e y. El punto puede moverse a lo largo y ancho del plano, por lo tanto tiene dos grados de libertad.



Ahora pasemos a imaginar una dimensión más y ubiquemos a nuestro punto en un espacio formado por tres dimensiones (largo, ancho y altura), de tal manera que nuestro punto se puede mover libremente en cualquiera de las dimensiones y tiene por lo tanto tres grados de libertad. Esto es equivalente a pensar en una mosca que se desplaza volando en una habitación. Podemos conocer su ubicación exacta si tenemos los valores de tres coordenadas en un momento dado. También podemos imaginar a un avión en pleno vuelo y saber que además de moverse hacia adelante se puede mover en el sentido derecha-izquierda y arriba-abajo, de tal manera que podemos conocer su ubicación si sabemos cuál es su altitud, latitud y longitud.

En matemáticas podemos pensar en muchas dimensiones más, así los espacios que determinan se llaman hiperespacios, y se representa por el símbolo R^n si se trata de espacios de números reales, en tales espacios un punto tendrá n grados de libertad o dimensiones por los que pueda moverse. Es casi imposible poder imaginar tales hiperespacios, pero al igual que en los párrafos anteriores, uno puede conocer la ubicación exacta de un punto sabiendo las n coordenadas que lo determinan.[7]

Vamos a considerar en toda la exposición siguiente que una muestra de tamaño n estará representada por un único punto en un hiperespacio cuyas coordenadas son cada uno de los n valores de la muestra. De tal manera que si tenemos una muestra de dos sujetos, éstos serán representados por un único punto que se mueve en un plano de dos dimensiones, y si tenemos una muestra de tres sujetos, estarán representados por un punto que se mueva en tres dimensiones. De igual manera, una muestra de n sujetos se representará ubicando los n valores en las n coordenadas y al unir las proyecciones quedará representado por un único punto que se mueve libremente a través de un hiperespacio de n dimensiones.

Hasta aquí hemos visto que un punto que se mueve libremente en un espacio con n dimensiones tiene n grados de libertad, pero, ¿qué sucede si de antemano se fija el valor que toma uno de los ejes cartesianos? En otras palabras, debemos preguntarnos ¿cómo quedan los grados de libertad si se restringe un eje? Volviendo al ejemplo del punto en dos dimensiones (x,y), si alguien de antemano restringe una de las dimensiones y nos dice que el valor del eje x debe ser necesariamente 3, entonces nuestro punto no podrá moverse libremente en dos dimensiones y habrá perdido un grado de libertad (n – 1 = 1 grado de libertad), es decir, sólo se podrá mover libremente en una recta (la recta que pase por x=3). Lo mismo ocurriría para el punto en tres, cuatro, o más dimensiones.

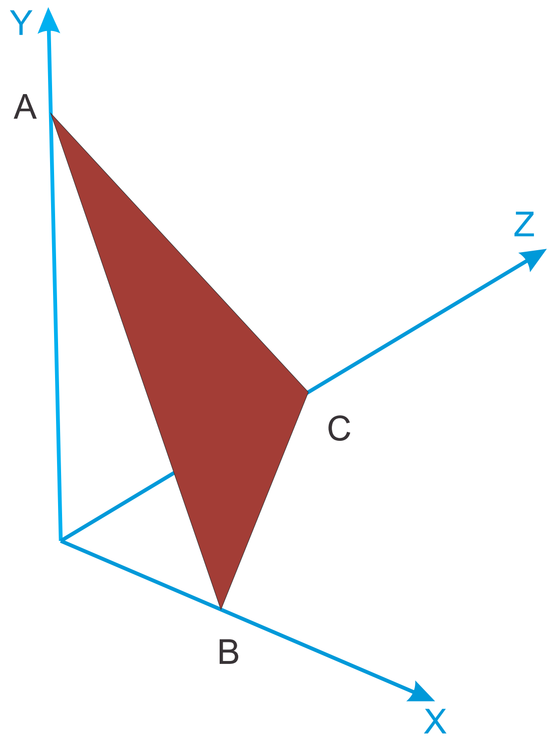
Pero no solamente se puede restringir una dimensión, sino también se pueden restringir todas las que se deseen. Así, en un espacio de tres dimensiones (x,y,z) se pueden restringir x e y de tal manera que solo tendremos una dimensión (3-2=1). En hiperespacios de más de tres dimensiones se pueden restringir más de dos dimensiones.

Generalizando, si un punto se encuentra en un hiperespacio de n dimensiones, entonces tiene n grados de libertad, a menos que se especifique de antemano algún valor de una o más dimensiones, en cuyo caso el número de grados de libertad queda determinado por la diferencia entre n y el número de dimensiones restringidas (r), por lo tanto, tendremos que los nuevos grados de libertad han disminuido hasta llegar a ser (n-r).

Pero, ¿cómo se hace para restringir un número de valores r? Hay dos maneras, la primera es simplemente dando a conocer uno de los valores que tomaría uno de los ejes y la segunda mediante alguna fórmula que relacione a todos los datos (como la fórmula para hallar el promedio).

Pensemos en un espacio definido por tres ejes (que determina un espacio de tres dimensiones). Un sistema como este es fácil de imaginar si pensamos en la esquina despejada de una habitación, allí podemos llamar a los ejes x, y, z. Si elegimos tres valores cualesquiera para los ejes y los unimos, podemos obtener un plano triangular ABC como el de la figura 2.

FIGURA 2: El plano ABC en dos dimensiones está formado a partir de la relación que existe entre tres puntos en los ejes cartesianos.



Cabe recordar que dos puntos cualesquiera al unirse forman una recta (cuando se relacionan de manera lineal), y que tres puntos cualesquiera forman un plano bidimensional cuando se unen por intermedio de alguna relación. En la figura 2, los tres puntos son:

$$B = n_1 = (x_i,0,0)$$
$$A = n_2 = (0,y_i,0)$$
$$C = n_3 = (0,0,z_i)$$

Y la función que las relaciona es:

$$\frac{x_i + y_i + z_i}{3} = \overline{X}$$

Donde:

x_i es el valor de alguna variable medida en la persona 1

y_i es el valor para la persona 2

z_i es el valor para la persona 3

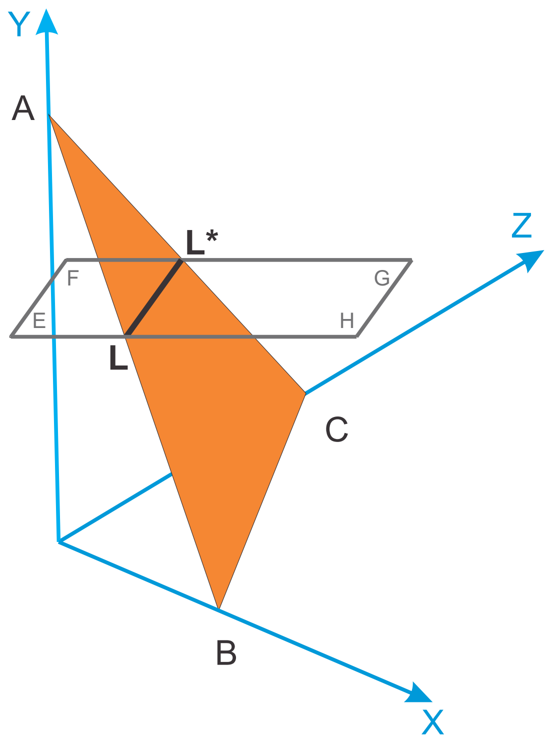
$\overline{X}$ es la media de los tres datos

Como podemos ver, cuando se estima la media de tres datos (tres puntos cardinales), el resultado se encuentra libre de moverse en un plano de dos dimensiones (triángulo ABC de la figura 2). Generalizando, cuando n datos se relacionan mediante el cálculo de la media (o de alguna otra relación lineal), tal media puede tomar diferentes valores que se mueven libremente en un plano de n-1 dimensiones. Volviendo a un ejemplo previo, si tenemos una muestra de 10 sujetos (n=10), y se calcula la media de sus pesos, tal media se mueve libremente dentro de los límites de un hiperespacio de 9 dimensiones (n-1=9) dependiendo de los posibles valores que puedan tomar los pesos de las 10 personas.

Es así que cuando dos puntos se relacional linealmente forman un espacio de una sola dimensión (una recta), cuando son tres los puntos que se relacionan forman un plano de dos dimensiones, cuando los puntos a relacionarse sean cuatro, formarán un espacio de tres dimensiones. etc. Queda entonces claro que cuando un número cualquiera de puntos en un espacio se relacionan por alguna fórmula, forman un nuevo espacio que es exactamente de una dimensión menos que la dimensión original, y cada vez que se intente explicar a los puntos mediante una nueva relación, las dimensiones se seguirán reduciendo sucesivamente.

En la figura 3, se puede apreciar que hay dos relaciones para los tres puntos. La primera relación es la fórmula de la media, que forma el triángulo ABC, la segunda relación es una ecuación que produce un paralelogramo EFGH que intersecta a nuestro primer plano. Si se exige que se cumplan ambas relaciones a la vez con los tres datos, entonces los únicos valores posibles se encuentran en cualquier punto de la recta LL* y se mueven libremente a través de ella.[8] Como se puede apreciar, al espacio inicial de tres dimensiones se le han impuesto dos restricciones, por lo tanto, los únicos valores que satisfacen ambas relaciones se pueden mover libremente en un espacio de 3-2=1 dimensión (una línea), y todos los cálculos posteriores que incluyan a éstos resultados deberán tener en cuenta que solo poseen un grado de libertad.

FIGURA 3: Dos relaciones impuestas a tres puntos en los ejes cartesianos reducen las dimensiones a una línea LL* de una sola dimensión.



Una aproximación algebraica

Si se nos da a elegir los valores de dos variables cualesquiera x, y, nosotros podemos pensar en infinitos posibles valores para cada una de ellas, pero cuando se nos muestra una relación que las explica, la situación cambia. Por ejemplo, los dos valores pueden relacionarse de la siguiente manera:

$$x + y = 9$$

En éste caso, solo tenemos la libertad de elegir a nuestro antojo a una de las variables, ya que la segunda variable queda determinada automáticamente por la relación que las une.

Si además se nos impone una segunda ecuación que las explique, de tal manera que tenemos ahora dos relaciones para nuestras variables:

$$x + y = 9 \quad ...... (1)$$

$$x - y = 1 \quad ...... (2)$$

Ya no nos queda ninguna variable que podamos elegir libremente, dado que tenemos dos variables en dos ecuaciones, y por lo tanto nuestros grados de libertad se han reducido a 2-2=0. Si resolvemos ambas ecuaciones, encontramos que x=5; y=4.

De la misma forma, si tenemos n sujetos con sus respectivos pesos, y calculamos el promedio de los mismos mediante la fórmula:

$$\overline{X} = \frac{x_1 + x_2 + ... + x_n}{n} = \sum \frac{x_i}{n}$$

Automáticamente nuestros grados de libertad quedan restringidos a (n – 1) opciones, y los cálculos que impliquen el uso de la media $\overline{X}$ en nuevas fórmulas deben ser trabajados en una dimensión menos que la original (es decir n-1) como en la aproximación geométrica. Generalmente, cada vez que calculamos un nuevo estimador, perdemos un grado de libertad.

USANDO LOS GRADOS DE LIBERTAD

Al calcular la desviación estándar muestral

La desviación estándar muestral nos da una idea del grado de dispersión de n datos alrededor de su media, y para llegar a conocer tal media se ha establecido una relación entre los datos sumándolos a todos y dividiéndolos entre el número de los mismos (es decir entre n). Al ser la media el resultado de tal relación, queda libre de moverse a través de un nuevo espacio de n-1 dimensiones, por lo tanto, la fórmula de la desviación estándar (y el de la varianza) miden la dispersión de los datos alrededor de un punto (la media) en un espacio de n-1 dimensiones, y es por esto que tienen como denominador a esa cantidad de grados de libertad tal como se evidencia en la fórmula siguiente:

$$s^2 = \frac{(x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + ... + (x_n - \overline{X})^2}{n\text{-}1}$$

Donde:	s^2	es la varianza muestral
	s	es la desviación estándar muestral
	$x_1 \ldots x_n$	representan a los *n* datos
	n	representa el número de la muestra
	$(n\text{-}1)$	son los grados de libertad

Pongamos un ejemplo. Se desea conocer la varianza y desviación estándar de la hemoglobina de una muestra de cinco pacientes procedentes de un área con alta prevalencia de parasitosis intestinal. Se tienen los siguientes datos de hemoglobina: 10; 11; 9; 7; 9.

La media muestral es:

$$\overline{X} = \frac{10 + 11 + 9 + 7 + 9}{5} = 9.2$$

Y tanto el cálculo de la varianza como el de la desviación estándar deben tener en cuenta a los grados de libertad en el denominador. Para el caso, los grados de libertad es igual al número de observaciones menos 1, es decir 5 – 1 = 4. La varianza muestral será:

$$s^2 = \frac{(10\text{-}9.2)^2 + (11\text{-}9.2)^2 + (9\text{-}9.2)^2 + (7\text{-}9.2)^2 + (9\text{-}9.2)^2}{4} = 2.2$$

Y la desviación estándar muestral será la raíz cuadrada de 2.2, es decir 1.48.

Grados de libertad para la distribución t de Student

La distribución ***t*** sirve para probar la hipótesis de igualdad de las medias de dos grupos de datos. Se empieza teniendo dos grupos, el primero de ellos tiene n_1 elementos, y el segundo n_2, se hallan las medias muestrales de ambos, $\overline{X}_1$ y $\overline{X}_2$ respectivamente. Por lo tanto, cada grupo ha reducido sus dimensiones de libertad tras calcular las medias, y los nuevos grados de libertad son entonces $(n_1\text{-}1)$ y $(n_2\text{-}2)$. Para hallar la desviación estándar (y la varianza), solamente se obtiene el promedio a partir de ambas varianzas ponderadas según sus grados de libertad:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Y cuando se busca en la tabla de valores para la distribución ***t*** de Student se considera como los grados de libertad la suma total de los grados de libertad de ambos grupos.

Podemos apreciar mejor esto si ponemos un ejemplo. Se desea conocer si la presencia de parásitos intestinales interfiere con la absorción de hierro y por consiguiente producen niveles bajos de hemoglobina. Se pone a prueba la hipótesis nula de no diferencia del nivel de hemoglobina promedio de dos grupos de personas, el primer grupo procede de un área con alta prevalencia de parasitosis intestinal y el segundo grupo procede de un área con baja prevalencia. Si el número muestral del grupo uno es de 20 sujetos, y el tamaño del segundo es de 28, entonces los grados de libertad asociados a la prueba t son:

$$GL = (20 - 1) + (28 - 1) = 46$$

Y tendremos en cuenta a los 46 grados de libertad cuando hallemos el punto crítico para el rechazo de la hipótesis nula. Si consideramos un valor alfa igual a 0.05, entonces nuestro punto crítico de rechazo será:

$$t_{(0.95\,;\,46)} = \pm 1.678$$

Y si nuestro estadístico ***t*** calculado cae a la izquierda de -1.678 o a la derecha de +1.678 se rechaza la hipótesis nula y por lo tanto hay evidencia que las medias de hemoglobina de ambos grupos son diferentes. Es importante mencionar que la distribución ***t*** siempre depende de los grados de libertad.

VOLVIENDO A EMPEZAR

Ahora tenemos un mejor entendimiento de los grados de libertad:

Los grados de libertad son iguales al número de observaciones independientes que son libres de variar (el número de sujetos en los datos) menos el número de parámetros estimados (el número de relaciones impuestas a los datos).[1,9,10] En otras palabras, están relacionados al tamaño de la muestra. Así mismo, los grados de libertad son usados para definir las distribuciones estadísticas y con ellos poder realizar las pruebas de hipótesis.

CONCLUSIONES

El número de grados de libertad constituye un tema central en la estadística moderna, sin embargo no es comprendida por la mayoría de investigadores. El número de grados de libertad se comprende mejor si es visto como el número de dimensiones espaciales en los que un punto es libre de moverse. Cada relación que se establece (es decir, cada estadístico que se calcula a partir de los datos) hace que se tengan que modificar los grados de libertad si tal estadístico va a ser usado en cálculos futuros. En el último caso, los grados de libertad quedan restringidos a la diferencia entre la cantidad de datos y el número de relaciones establecidas entre los mismos. Los grados de libertad encuentran su aplicación en una gran cantidad de modelos estadísticos, siendo la prueba ***t*** solo un ejemplo.

****

REFERENCIAS BIBLIOGRÁFICAS

1. Walker HM. Degrees of freedom. Journal of Educational Psycology. 1940;31(4):253-269.

2. Moonan W. The Statistical Interpretation of Degrees of Freedom. The Journal of Experimental Education. 1953;21(3), 259-264.

3. Daniel W. Bioestadística: Base Para el Análisis de las Ciencias de la Salud. México: Limusa Wiley. 2007.

4. Dawson B, Trapp R. Bioestadística Médica. México: Manual Moderno. 2005.

5. Levin R, Rubin D. Estadística Para Administradores. México: Prentice-Hall. 1996.

6. Pagano R. Understanding Statistics in the Behavioral Sciences. Canada: Wadsworth, Cengage Learning. 2009.

7. Strang G. Introduction to Linear Algebra. Brooks/Cole. 2006.

8. Grossman S. Álgebra Lineal. México: McGraw-Hill/Interamericana Editores. 2008.

9. Toothaker L, Miller L. Introductory Statistics for the Behavioral Sciences. Pacific Grove, CA: Brooks/Cole. 1996.

10. Yu C. Illustrating Degrees of Freedom in Terms of Sample Size and Dimensionality. 1997. Disponible en: http://www.creative-wisdom.com/computer/sas/df.html

ABSTRACT

WHAT DO DEGREES OF FREEDOM MEAN?

Degrees of freedom is a central topic in modern statistics, however is poor explained in textbooks. Gauss used it for first time in estimates of astronomic distances but it appears formally in Gosset (Student) manuscripts in 1908. Degrees of freedom can be understood from geometric, algebraic or even intuitive point of view. Geometrics shows degrees of freedom as spaces through a summary measure can move freely assuming different values. Algebra describes degrees of freedom as number of unknown variables in respect of number of equations we can establish using data. Both algebraic and geometric points of view are related and help us to understand better the meaning of degrees of freedom. Applications in statistics are widely spread but standard deviation and Student *t* are just some examples.

KEYWORDS: Degrees of freedom, Dimensions, Student t.