Estadística inferencial

Daniel Maldonado

1. Prueba de hipótesis

La estadística inferencial permite generalizar los resultados de muestras hacia poblaciones con niveles razonables de confianza. Esto se realiza mediante pruebas de hipótesis.

Una hipótesis experimental podría ser que la media $\bar{X}$ del grupo al cual se aplicó una intervención es distinta de la media μ de la población de la cual se obtuvo la muestra. Es decir,

$$\begin{aligned} \bar{X} &< \mu \\ &\text{o} \\ \bar{X} &> \mu, \end{aligned}$$

o de forma más general:

$$\bar{X} \neq \mu.$$

Suponemos que la muestra tenía originalmente la misma media que la población, y que la aplicación de nuestra intervención desplazó su media en una cierta dirección. En cierto modo, suponemos que nuestra intervención tiene el efecto de crear una segunda población con una media distinta de la población original.

Si la muestra fuese perfectamente representativa de la población original, entonces se podría concluir que, si la media muestral $\bar{X}$ después del tratamiento es distinta de la media de la población μ, entonces el tratamiento es eficaz. Sin embargo, la realidad no suele ser tan bella.

Una amenaza seria a la validez de la relación encontrada entre la aplicación del tratamiento experimental y el cambio en la media de la muestra es la posibilidad de un error de muestreo, es decir, que por azar la media de la muestra no fuese representativa de la población desde el comienzo. Esto tiene base en el *teorema central del límite* o *teorema del límite central*.

De acuerdo con el teorema podemos obtener muestras de una población y calcular la media de cada una de esas muestras. Si la cantidad de medias es lo bastante grande, entonces la distribución de esas medias tenderá a ser normal (una campana de Gauss) y la media formada por estas medias muestrales será muy cercana a la media μ de la población. Sin embargo, aunque las medias de las muestras tenderán a rondar el valor de la media poblacional, habrá una minoría que se encuentre muy lejos, ya sea por encima o

por debajo. Así, existe una probabilidad nada despreciable de que la media de la muestra obtenida de la población esté muy por encima o muy por debajo del valor poblacional (Figura 1).

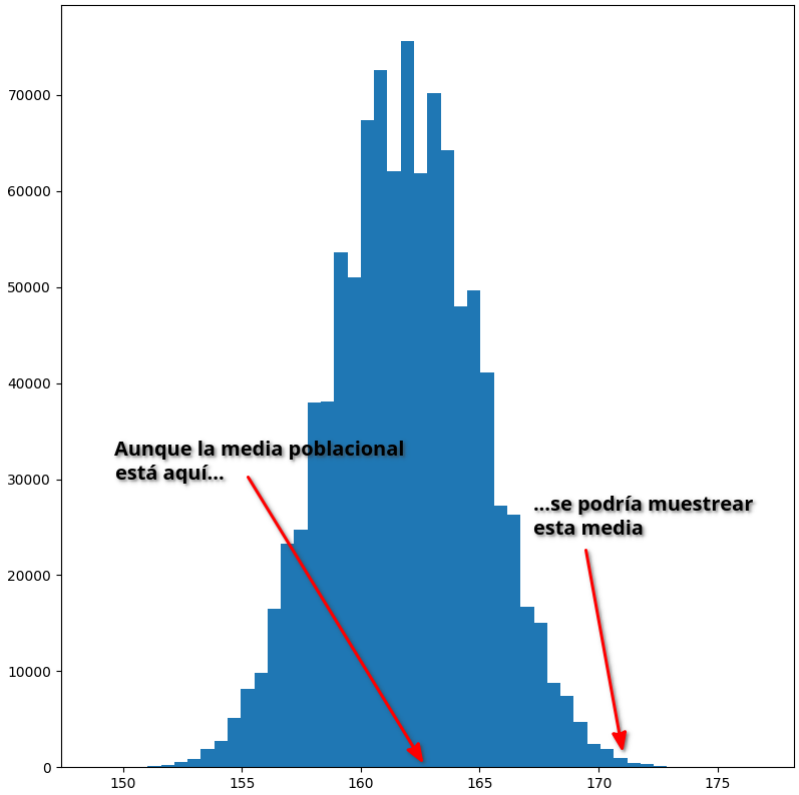


Figura 1: La media de la muestra puede estar en cualquier punto de la distribución.

Si este fuera el caso, entonces la diferencia que encontramos entre la media $\bar{X}$ de la muestra tras el tratamiento y la media de la población μ sería solo debida al azar.

El papel de la estadística inferencial es garantizar que no cometamos el error de atribuir a la variable independiente (el tratamiento) las diferencias debidas a un error de muestreo. ¿Concluimos que la relación encontrada en la muestra se cumpliría si evaluamos a toda la población, o concluimos que es una coincidencia debida al error?

El procedimiento estadístico específico a usar dependerá del diseño experimental y la hipótesis, pero de modo general se puede utilizar estadística *paramétrica* y *no paramétrica*.

La estadística paramétrica requiere que se cumplan ciertos supuestos dentro de los datos de la muestra:

- La población de puntuaciones de la variable dependiente forma una distribución normal (o aproximadamente normal)
- Las puntuaciones tienen nivel de medición de intervalos o de razón

La estadística paramétrica suele ser preferible, así que se utiliza a menos que haya flagrantes violaciones de los supuestos que tiene.

De forma general los pasos para las pruebas de hipótesis son

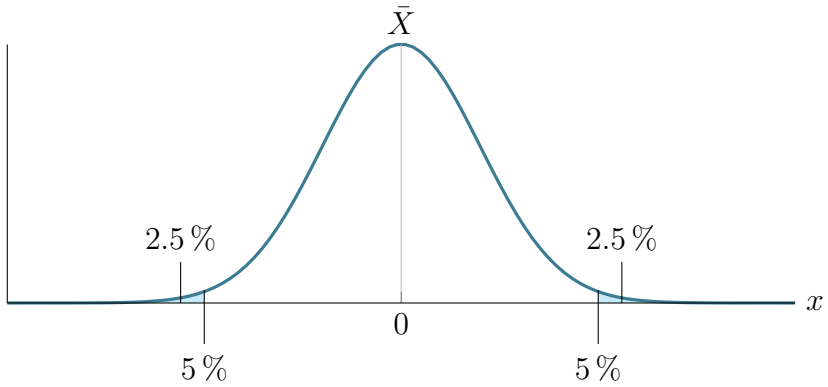
1. Establecer una hipótesis experimental
2. Diseñar y correr un experimento que permita probarla
3. Traducir la hipótesis experimental en una hipótesis estadística
4. Seleccionar y llevar a cabo el procedimiento estadístico correcto para probarla

Las hipótesis generalmente dirán cosas similares a "*a incrementos en X corresponden incrementos en Y*", o "*a incrementos en X corresponden disminuciones en Y*", es decir, habrá una relación ordenada entre las variables, y esa relación tendrá una dirección positiva o negativa.

Si la relación es positiva, esto significará que esperamos que la media de la muestra después del tratamiento esté a la *derecha* de la media poblacional. Si la hipótesis es correcta entonces esta media muestral caerá en un punto lo bastante alejado de la media poblacional para poder decir con razonables niveles de confianza que no pertenece a la misma distribución, sino que pertenece a una nueva distribución creada por la intervención.

El punto de corte estándar en psicología es el percentil 95 de la distribución poblacional. Es decir, si la media de la muestra cae en la cola derecha de la distribución poblacional en un punto tan alejado de la media que menos del 5 % de las ocasiones en que se muestree esa distribución aparecerá un valor así de elevado, entonces tendremos suficiente certeza de que la intervención es eficaz y efectivamente desplaza los puntajes hacia la derecha.

Si anticipamos una relación negativa, ocurre lo mismo pero hacia el lado izquierdo de la distribución. Si anticipamos que habrá un cambio, pero no sabemos en qué dirección, entonces el punto de corte de 5 % se repartirá entre las dos colas de la distribución, 2.5 % de cada lado. Esta es la diferencia entre pruebas de "una cola" y de "dos colas".



2. Prueba *z*

El experimento más simple es uno de una sola muestra en el cual comparamos la media de la muestra tras el tratamiento con la media de la población de origen. La idea detrás

de este diseño es comparar dos niveles de la variable independiente: el nivel experimental con otro nivel ya conocido. Un nivel conocido suele ser el de la ausencia de manipulación. Por ejemplo, si se conoce la calificación media de una población de primaria y se quiere probar una intervención que busca incrementarla entonces se tienen ya los dos niveles de la variable independiente: presencia y ausencia.

Ejemplo

Una población tiene una media de puntos de IQ de 100 y una desviación estándar de 15. Queremos probar la eficacia de un tratamiento que pretende incrementar la inteligencia media.

Tomamos una muestra aleatoria de la población con tamaño de 36, le aplicamos el tratamiento, y después hacemos una medición de su IQ. El resultado es de

$$\bar{X} = 105.$$

Concluimos entonces que, dado que la media de la muestra es más grande que la media de la población, el tratamiento funciona.

Fin.

Excepto que no funciona así. Puede haber implícito un error de muestreo, y el supuesto efecto del tratamiento podría deberse al azar. Para demostrar que verdaderamente existe un efecto es necesario determinar qué tan probable es encontrar un dato tan grande como 105 o mayor en una población con media de 100. Si la probabilidad de encontrar un dato así de grande por azar es menor al 5 %, entonces podremos concluir con relativa certeza que el cambio se debió a la intervención y no al azar (aunque nunca tendremos total seguridad de ello, porque para estar totalmente seguros deberíamos aplicar el tratamiento a la población completa y el presupuesto no alcanza para tanto).

Específicamente debemos comparar una hipótesis nula H_0 en la cual la media de la nueva población (la población creada por nuestra intervención) es igual a la media de la población original, con una hipótesis alternativa H_a en la cual la media de la nueva población es *distinta* de la media de la población original:

$$\begin{aligned} H_0: &\quad \mu = 100 \\ H_a: &\quad \mu \neq 100 \end{aligned}$$

La hipótesis nula H_0 siempre corresponderá a la ausencia de efecto o de relación entre la variable independiente y la dependiente. Si esta hipótesis es cierta, entonces concluiríamos que la media de nuestra muestra es una de las tantas medias que serían obtenidas de muestrear a la población original y no aplicar ninguna intervención.

Debe notarse que las hipótesis H_0 y H_a componen todas las posibilidades. La media de la población con tratamiento μ puede ser igual a 100 o distinta de 100, y no hay más posibilidades. Lo mismo aplicará para cualquier prueba de hipótesis que realicemos: las hipótesis propuestas deben agotar el espacio de las posibilidades.

Para este caso particular el procedimiento estadístico usado se conoce como prueba *z*. Este procedimiento consiste en calcular el puntaje z de la media de la muestra tras el tratamiento, y después determinar la localización de esa media en la distribución de la población dada por el teorema central del límite. Recordemos que el puntaje z de un dato indica cuánto éste se desvía de la media de su población o qué tan atípico es con respecto a ella, y se calcula restándole la media y dividiendo el resultado entre la desviación estándar:

$$z = \frac{X - \bar{X}}{S_X}.$$

Este diseño requiere que la distribución de la variable dependiente sea aproximadamente normal y que la muestra haya sido seleccionada de forma aleatoria. Además, es indispensable conocer la media de la población y su desviación estándar, lo que por lo general no ocurrirá. Pero supongamos que sí por esta ocasión.

Para hacer la comparación entre la media muestral y la media de la población suponemos una distribución formada por infinitas medias muestrales con tamaño de 36 (porque ese es el tamaño de la muestra dado en el ejemplo). La media de esta distribución muestral estará muy cerca de la media verdadera de la población. Entonces, esta distribución muestral indicará la frecuencia de todas las $\bar{X}$ que podrían encontrarse si se toman infinitas muestras aleatorias de tamaño 36 de la población. Suponiendo que H_0 sea correcta, toda media distinta de 100 vendrá de error de muestreo de esta distribución.

El siguiente paso será determinar el umbral de aceptación que se utilizará para la comparación, es decir, cuánto riesgo de equivocarnos estamos dispuestos a aceptar. En psicología el riesgo aceptable es del 5 % por convención, pero en otras áreas (como medicina) no se acepta más del 1 %. Este umbral se denomina α:

$$\alpha = 0.05$$

$\alpha = 0.05$ indica que estamos dispuestos a aceptar un riesgo de equivocarnos del 5 % y, por lo tanto, concluir una de cada 20 veces que existe un efecto de la variable independiente cuando esto no es real (y la variación en los datos se debió a un error de muestreo).

¿Por qué no usar un α más estricto? Porque eso podría significar arriesgarse a ignorar un efecto real, es decir, concluir que no existe un efecto por parte de la variable independiente cuando en realidad sí lo hubo.

Estos dos tipos de errores—falso positivo y falso negativo—se conocen también como errores de tipo I y tipo II o errores α y β.

	Existe Efecto	No Existe Efecto
Decimos que sí existe efecto	Correcto	Falso positivo Error tipo I Error α
Decimos que no existe efecto	Falso negativo Error tipo II Error β	Correcto

α y β son complementarios y en conjunto deben sumar 1. Cuanto mayor sea α, mayor la probabilidad de concluir que sí existe un efecto cuando no lo hay, pero también mayor la probabilidad de detectar efectos que sí existen. Cuanto mayor es β mayor será la probabilidad de pasar por alto efectos reales, pero también menor será la probabilidad de concluir que sí existe un efecto cuando no lo hay.

Una vez elegido el umbral α, éste se debe localizar dentro de la distribución de medias muestrales. El umbral estará repartido entre ambas colas de la distribución en este caso debido a que nuestra hipótesis es únicamente que el IQ medio cambiará, pero no predecimos la dirección.

La localización del umbral α dentro de la distribución está dada por tablas que pueden encontrarse en internet o en libros de estadística. La tabla de z es sencilla debido a que indica solo la ubicación en porcentaje de un punto z particular. En el caso del umbral de 0.025 % (porque al ser una prueba de dos colas dividimos α entre 2) la traducción en puntaje z es de $\pm$1,96. Es decir, el umbral de $\pm$2.5 % está localizado a $\pm$1.96 desviaciones estándar de la media de una distribución normal. Este valor será llamado z_{crit}.

Lo único que resta es convertir el valor de la media de la muestra, 105, en puntaje z y compararlo con z_{crit}. La fórmula para computar el puntaje z de los datos (z_{obt}) es:

$$z_{obt} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

donde

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{N}}$$

Es decir, se obtiene la diferencia entre la media muestral y la poblacional ($\bar{X} - \mu$), y se divide entre el error estándar de la media ($\sigma_{\bar{X}}$).

Calculamos primero el error estándar de la media. N es el tamaño de la muestra, y σ_X es la desviación estándar de la población. Para el ejemplo con $\sigma_X = 5$ y $N = 36$:

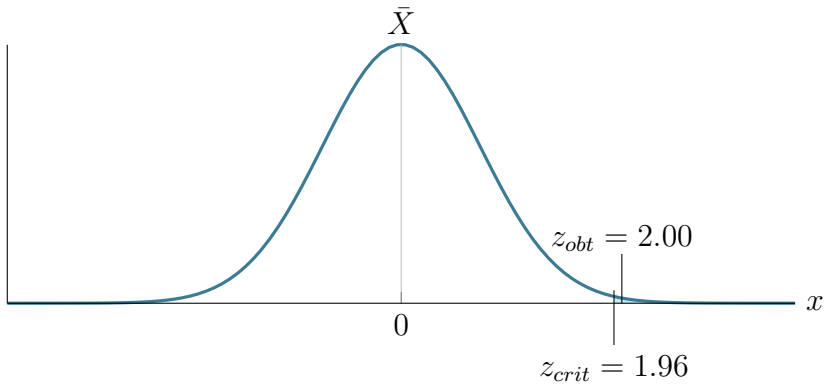
$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{N}} = \frac{15}{\sqrt{36}} = \frac{15}{6} = 2.5$$

Después computamos z_{obt}. μ será la media de la población, $\bar{X}$ se obtiene de la muestra, y $\sigma_{\bar{X}}$ es el error estándar que acabamos de calcular. Entonces:

$$z_{obt} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{105 - 100}{2.5} = \frac{+5}{2.5} = +2.00$$

Nuestro z_{obt} es $+2.00$. Para interpretar lo que esto significa es necesario compararlo con z_{crit}.

El valor de z_{crit} era de ± 1.96. Dado que el valor de $+2.00$ es más grande sabemos que la media de nuestra muestra se encuentra más allá del umbral de $\alpha = 0.025$:



Esto indica que podemos *rechazar* la hipótesis nula H_0 a favor de la hipótesis alternativa H_a, es decir, que podemos concluir con un nivel razonable de confianza que nuestra muestra con media de $\bar{X} = 105$ difícilmente podría haber venido de una población con una media de $\mu = 100$, y que lo más probable es que venga de una población distinta, probablemente la población "creada" por nuestra intervención.

Dicho de otra forma, decimos que nuestros resultados son *significativos*. Significativo no quiere decir *importante*, solo que difícilmente ocurrirían si no existiese una relación en la población. Es importante notar que no hemos demostrado que nuestra intervención funciona, sino solamente que una muestra como la que encontramos sería improbable si la intervención no funcionase, pero esto aun es una posibilidad. Y tampoco demostramos que es nuestra variable independiente lo que ocasionó el cambio. Esto aun pudo deberse a factores fuera de nuestro control. Por ello es importante replicar los estudios en ciencia y mantener riguroso control experimental.

3. Correlaciones

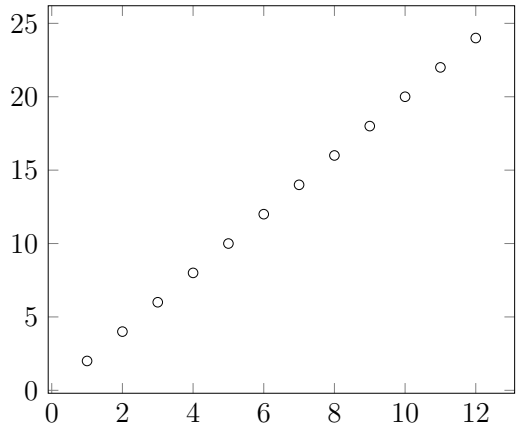
Un modo distinto de probar la existencia de relaciones en las variables son los coeficientes de correlación. Éstos nos permiten resumir de modo eficiente la forma en que los datos se organizan los unos con respecto a los otros. En un solo número nos indican tres tipos distintos de relaciones:

1. Relaciones positivas: cuando el valor de la variable X incrementa en una unidad, el valor de Y incrementa también en una cantidad relativamente constante.

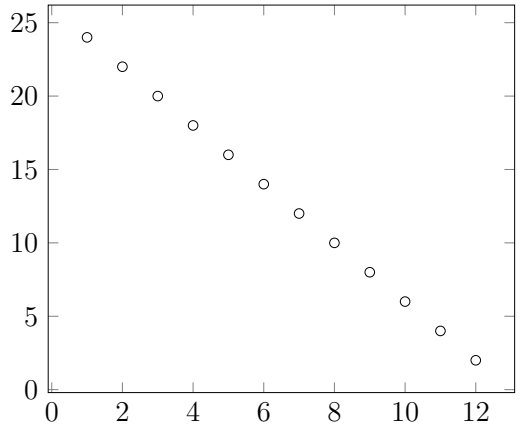
2. Negativas: cuando el valor de X incrementa en una unidad, el valor de Y disminuye en una cantidad relativamente constante.

3. Sin relación: el valor de X no permite predecir el valor de Y de manera confiable.

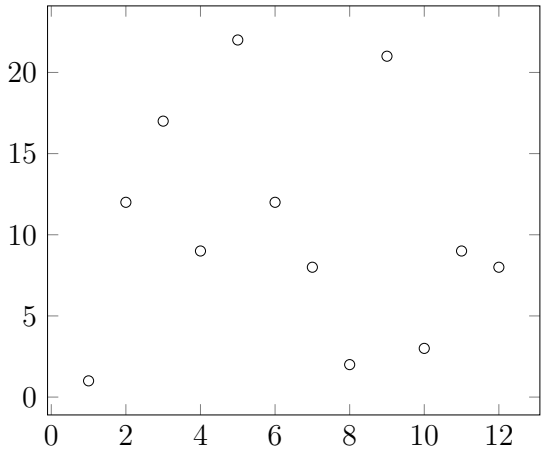
Los coeficientes de correlación van de -1 a +1. +1 indica una relación perfecta positiva:



-1 indica una relación perfecta negativa:



0 indica una falta completa de relación:



Aunque generalmente no encontraremos +1, -1 ni 0. Los datos suelen tener relaciones más imperfectas y difíciles de interpretar. Además, existirán variables con relaciones no lineales que serán más difíciles de representar mediante correlaciones.

En estudios correlacionales, a diferencia de otros estudios, no se analizan medidas de tendencia central, sino conjuntos completos de datos. Por ello la mejor manera de representar las relaciones entre variables en estudios de este tipo son los diagramas de dispersión (como los de arriba). A cada dato de la variable X siempre debe corresponderle un dato de la variable Y, y generalmente vendrán del mismo caso o sujeto.

Este tipo de estudio no permite establecer causalidad entre las variables. Esto puede hacerse solo mediante la revisión teórica.

Existen distintos coeficientes de correlación que se pueden utilizar dependiendo del tipo de dato que se quiera relacionar, pero el coeficiente más común es el de Pearson, llamado *Coeficiente de Correlación Producto-Momento de Pearson*. El símbolo de este coeficiente es r.

r compara cuán consistentemente cada valor de X va acompañado de un valor de Y. Para hacer la comparación cada valor de X y Y se transforma en un puntaje z (nombrados z_X y z_Y), y después se determina la cantidad media de correspondencia entre todos los pares. La fórmula del coeficiente es:

$$r = \frac{\sum(z_X z_Y)}{N}$$

Se multiplica cada z_X por su par z_Y, se suman todos los productos, y el resultado se divide entre N. Una manera desarrollada de la fórmula es esta:

$$r = \frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{\left[N\left(\sum X^2\right) - \left(\sum X\right)^2\right]\left[N\left(\sum Y^2\right) - \left(\sum Y\right)^2\right]}}$$

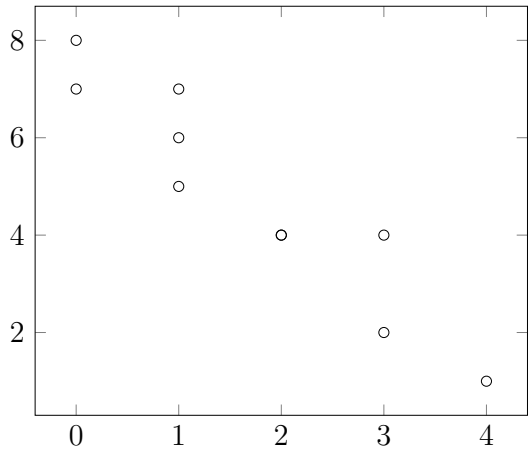
Este monstruo se deriva de reemplazar z_X y z_Y con sus fórmulas, y en cada una de ellas reemplazar los símbolos de la media y la desviación estándar con sus fórmulas. Después todo se simplifica y este es el resultado.

Ejemplo

Una manzana al día mantiene alejado al doctor. Lo mismo que el ajo con los vampiros. Un buen día le preguntamos a una población cuántas manzanas comen y cuántas veces han ido al doctor en el último año. Obtenemos estos datos:

Participante	Manzanas	Visitas al doc
1	0	8
2	0	7
3	1	7
4	1	6
5	1	5
6	2	4
7	2	4
8	3	4
9	3	2
10	4	0

Graficada, esta relación se ve así:



A simple vista parece que las manzanas son buenas como repelente de doctores. Pero para comunicar esto a alguien tenemos que darle los datos completos o mostrarle la gráfica. Para comunicar de modo más eficiente obtenemos el coeficiente de correlación. Recordando que la fórmula nos pide X, X^2, Y, Y^2, y XY, y las sumas de cada una, creamos esta tabla:

Participante	X	X^2	Y	Y^2	XY
1	0	0	8	64	0
2	0	0	7	49	0
3	1	1	7	49	7
4	1	1	6	36	6
5	1	1	5	25	5
6	2	4	4	16	8
7	2	4	4	16	8
8	3	9	4	16	12
9	3	9	2	4	6
10	4	16	0	0	0
N = 10	$\Sigma X = 17$ $(\Sigma X)^2 = 289$	$\Sigma X^2 = 45$	$\Sigma Y = 47$ $(\Sigma Y)^2 = 2209$	$\Sigma Y^2 = 275$	$\Sigma XY = 52$

Sustituyendo en la fórmula obtenemos:

$$\begin{aligned} r &= \frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{\left[N\left(\sum X^2\right) - \left(\sum X\right)^2\right]\left[N\left(\sum Y^2\right) - \left(\sum Y\right)^2\right]}} \\ &= \frac{10(52) - (17)(47)}{\sqrt{[10(45) - 289][10(275) - 2209]}} \\ &= \frac{520 - 799}{\sqrt{[161][541]}} \\ &= \frac{-279}{295.129} \\ r &= -.95 \end{aligned}$$

Esto indica que el coeficiente de correlación entre la cantidad de manzanas comidas y las visitas al doctor es de $r = -0{,}95$, una fuerte relación negativa, lo que indica que podríamos predecir confiablemente las visitas al doctor de una persona conociendo las manzanas que come, o viceversa.

Sin embargo, debe tenerse en cuenta que esta relación fue encontrada en la *muestra* a la que tenemos acceso. En última instancia, queremos generalizar la relación a toda la población, lo que requiere la aplicación de un procedimiento de inferencia.

Igual que la media de la población lleva la letra griega μ, el coeficiente de correlación de una población se denomina con la letra griega "rho", ρ.

Al igual que en el caso de r, la relación encontrada en la muestra mediante el coeficiente de correlación podría deberse a un error de muestreo.

El primer paso para probar la significancia de una correlación es revisar que se cumplan tres supuestos:

1. Hay una muestra aleatoria de pares de datos X y Y, y ambas variables son de nivel intervalar o de razón.

2. Ambas variables provienen de distribuciones aproximadamente normales. Aunque si la N es mayor a 25 este supuesto pierde importancia.

3. La hipótesis nula es que no hay correlación en la población.

Igual que en el caso de z, establecemos nuestras hipótesis estadísticas. La hipótesis nula es que no existe relación entre las variables, es decir,

$$H_0 : r = 0.$$

La hipótesis alternativa será que existe una relación. Esta puede ser positiva o negativa. Si desconocemos su dirección, podríamos decir simplemente que hay una relación entre las variables:

$$H_a : r \neq 0.$$

El teorema central del límite aplica también en este caso, pero de un modo ligeramente distinto. Suponemos que existe una población de sujetos en los cuales podemos medir nuestras dos variables de interés. En esta población podemos tomar una cantidad infinita de muestras, y para cada muestra calcular el coeficiente de correlación r. La distribución de coeficientes muestrales r tomados de infinitas muestras tiende a tomar la forma de una distribución normal. De este modo, nuestra labor estadística consiste en determinar, de entre todas las r posibles que podemos obtener al muestrear una población en la cual no existe correlación entre las variables, si la r que de hecho obtuvimos es lo bastante atípica para considerar que proviene de una población en la cual sí existe la relación.

Dicho de otro modo, suponemos una distribución normal de valores de r con media en 0. La media es cero porque suponemos que en la población no existe la relación, así que ese será el valor "real" del parámetro. Al tomar muestras de esta distribución obtendremos, ocasionalmente y por azar, valores altos y bajos de r a pesar de que la media real sea de 0. Debemos determinar cuán atípico es el valor de r que obtenemos de la muestra en comparación con esta distribución teórica. Si ese valor ocurre menos del 5 % de las ocasiones (o el umbral que hayamos elegido como punto de corte), entonces concluimos que sí existe una relación en la población.

Afortunadamente el valor medio de esta distribución siempre es de 0, por lo que la r obtenida de la muestra nos comunica directamente su posición en esta distribución teórica. Lo único que es necesario saber es el tamaño de N, debido a que la distribución tiene una forma levemente distinta para cada tamaño de muestra. Para este propósito, igual que con la prueba z, existen tablas que indican la posición de cada valor de r en la distribución con base en los *grados de libertad*—una medida obtenida de la cantidad de datos de la muestra, equivalente a $N - 2$—(Figura 2).

En este caso tenemos 8 grados de libertad debido a que el tamaño de la N es de 10. Si seguimos en la tabla el renglón de 8 grados de libertad y buscamos el valor más aproximado al resultado de $r = -0{,}95$ obtenido (pero sin pasarnos de él), encontraremos que el resultado es significativo al nivel de 0.01 para una prueba de dos colas, es decir, que en una población en la cual no existiese ninguna relación entre las manzanas y las visitas al médico encontraríamos un coeficiente de correlación de -0.95 en menos de 1 de cada 100 muestras con $N = 10$ tomadas. Un nivel muy razonable de confianza.

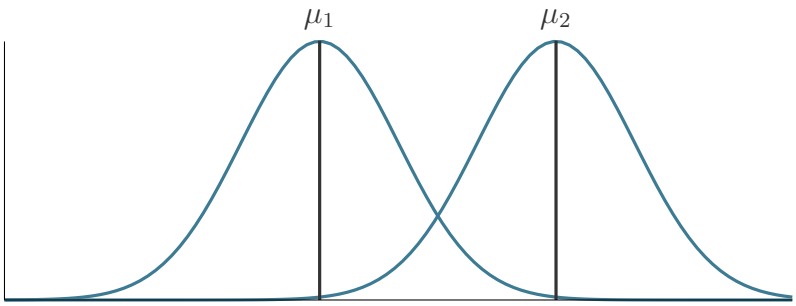
	Level of significance for a one-tailed test			
	.05	.025	.01	.005
df =N-2	Level of significance for a two-tailed test			
	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
20	.360	.423	.492	.537
30	.296	.349	.409	.449
40	.257	.304	.358	.393
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.302
80	.183	.217	.256	.284
90	.173	.205	.242	.267
100	.164	.195	.230	.254
∞	.073	.087	.103	.114

Figura 2: Tabla de valores de r en función de los grados de libertad

4. Diferencia entre grupos

Además de el cambio coordinado entre dos variables es posible determinar el cambio entre dos grupos expuestos a manipulaciones distintas, o más precisamente, a dos niveles de una manipulación. El caso más simple fue el visto con la prueba z en el cual comparamos el grupo experimental con el nivel ya conocido de la población de la que se toma la muestra. Sin embargo, casos así serán la excepción más que la regla debido a que es difícil conocer las características reales de poblaciones completas (a no ser que se obtengan datos de un censo poblacional). Lo más común será que tomemos dos muestras de una población con características desconocidas, establezcamos manipulaciones distintas, y debamos hacer una comparación entre ellas para determinar si las manipulaciones son eficaces para modificar las características de las poblaciones.

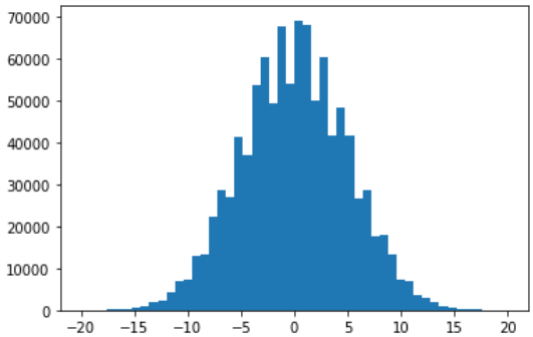
En un experimento con dos muestras medimos las puntuaciones de los sujetos bajo dos condiciones. La condición 1 produce una $\bar{X}$ que representa a μ_1, la μ que encontraríamos si se le aplicara la manipulación a toda la población. Del mismo modo, la condición 2 produce una μ_2. Un resultado posible sería el siguiente:



Idealmente podríamos concluir que, si las medias poblacionales μ representadas por las medias muestrales $\bar{X}$ son distintas, entonces la manipulación tiene un efecto. Pero, de nuevo, la realidad no es tan bella. Existe también aquí la posibilidad de un error de muestreo que genere la diferencia: quizá desde el comienzo, antes de la manipulación, las muestras ya tenían medias distintas por azar a pesar de estar representando a una misma población. Siendo así, es necesario demostrar que la diferencia encontrada entre las medias de las muestras es lo bastante grande para poder afirmar con niveles razonables de seguridad que no pertenecen a una misma población. Ese es el papel de la prueba T.

El principio estadístico es similar al principio de la prueba z: es necesario determinar la posición de un resultado dentro de una distribución muestral. Sin embargo, debido a que estamos determinando si una diferencia entre medias es *significativa*, es necesario que la distribución muestral esté también en términos de diferencias entre medias.

En esta ocasión suponemos una cantidad infinita de *pares* de muestras tomados de una población. Para cada par calculamos la media de cada uno y los restamos entre sí. El proceso se repite infinitamente y las diferencias entre las medias de las muestras se grafican en un histograma de frecuencias. Al graficar nos daremos cuenta de que las diferencias entre infinitas medias tomadas de una misma población tienden a formar una distribución normal, es decir, que también son descritas por el teorema central del límite (figura 3).



La única diferencia con el histograma que ya conocemos es que en esta ocasión, debido a que estamos calculando diferencias entre medias de una misma población, los datos se agruparán alrededor del valor de 0. Esto se debe a que frecuentemente se muestrearán valores cercanos a la media real de la población, y restarlos entre sí generalmente resultará en números pequeños. Solo en las pocas ocasiones en que se muestreen valores muy alejados de la media poblacional se encontrarán valores altos para la diferencia entre las medias.

La prueba T tiene como función determinar en qué posición de la distribución de *diferencias de medias* se encuentra el valor que encontramos en un estudio dado. Si ese valor se encuentra lo bastante lejos de la distribución como para ser muestreado en menos del 5 % de las ocasiones (o lo que determinemos con el umbral α), entonces determinaremos que existen diferencias entre nuestros grupos, pues será improbable que un valor elevado provenga de una distribución de *diferencias de medias* centrada en 0.

La prueba T tiene dos variantes:

- Para muestras independientes: se usa si los sujetos de cada grupo son muestreados sin tomar en cuenta a los del otro grupo.
- Para muestras relacionadas: se usa si los datos de ambos grupos provienen de los mismos sujetos pasando por condiciones distintas, o si los sujetos fueron seleccionados de manera propositiva mediante pareo.

La prueba T para muestras independientes tiene como requisitos:

1. Los puntajes tienen nivel de intervalo o razón
2. Las poblaciones tienen distribuciones aproximadamente normales
3. Las poblaciones tienen varianzas homogéneas, es decir, que si computamos la varianza de cada población por separado obtendríamos el mismo resultado.
4. La cantidad n de casos es similar

Las hipótesis estadísticas en este caso están dichas en términos de las diferencias entre las medias de las poblaciones de las cuales vienen las muestras. Es decir, la hipótesis alternativa indicaría que la media poblacional μ_1 es *distinta* de la media poblacional μ_2:

$$H_a : \mu_1 \neq \mu_2,$$

o en términos de su diferencia:

$$H_a : \mu_1 - \mu_2 \neq 0.$$

La hipótesis nula diría que su diferencia sí es de cero:

$$H_0 : \mu_1 = \mu_2$$
$$H_0 : \mu_1 - \mu_2 = 0$$

Las hipótesis no tienen un valor específico para μ, por lo que serán las mismas para cualquier estudio.

La manera de contrastar las hipótesis es calculando el valor de t, un estadístico que indica la posición de la diferencia entre las medias dentro de la distribución de diferencias dada por el teorema central del límite.

La fórmula del estadístico t es:

$$t_{obt} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

Se trata de una fórmula similar a la del puntaje z debido a que su significado es similar: la posición del valor obtenido con respecto a una distribución normal. Se obtiene la diferencia entre *la diferencia entre medias muestrales* y *la diferencia entre medias poblacionales* (así es, una diferencia de diferencias), y el resultado se divide entre un error estándar. El valor $s_{\bar{X}_1-\bar{X}_2}$ se refiere al *error estándar de la diferencia*, es decir, qué tanto se alejan en promedio las diferencias entre medias muestreadas del valor central de la distribución. La forma de calcularlo es la siguiente:

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\left(s^2_{pool}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

Finalmente, en esta última fórmula el valor s^2_{pool} se refiere a la *varianza agrupada*, que es la media ponderada de varianzas. Específicamente, esto se refiere a la media de las varianzas de las dos poblaciones de las que provienen los datos, pero con un peso asignado en función de la cantidad de datos que tiene cada grupo usando los grados de libertad de cada muestra:

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

Así, la forma de calcular t es:

1. Obtener la varianza agrupada
2. Obtener el error estándar de las diferencias
3. Calcular el estadístico t

El valor resultante de t_{obtenida} se compara con una tabla de valores críticos y eso permite determinar la significancia de la prueba.

Ejemplo

Se pretende determinar si acaso hay diferencias entre los perros border collie y los golden retriever en la capacidad de aprender y recordar nombres de objetos. Para ello se seleccionan perros de cada raza al azar y se les enseña, a lo largo de una semana, a traer 30 juguetes tras escuchar su nombre. Después de una semana se hace una prueba para determinar cuántos nombres recuerdan aun. Supongamos que se encontraron 17 perros collie y 15 retriever. En promedio los collie recordaron 23 comandos; y los retriever, 20. La varianza del grupo de collie es de 9, y la del grupo de retriever es de 7.5. Nuestra labor es determinar si los collie son mejores para recordar comandos con base en los datos de nuestra muestra.

Como primer paso obtenemos la diferencia entre las medias de los grupos:

$$23 - 20 = 3,$$

es decir, la diferencia en puntaje entre grupos es de 3.

Después estimamos la varianza agrupada. Recordamos que la fórmula es:

$$\begin{aligned} s^2_{pool} &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)} \\ &= \frac{(17-1)9 + (15-1)7.5}{(17-1)+(15-1)} \\ &= \frac{144+105}{30} \\ &= \frac{249}{30} \\ &= 8.3 \end{aligned}$$

Así estimamos que la varianza conjunta de la población de ambas razas de perros es de 8.3.

Después utilizamos la varianza agrupada para calcular el error estándar de la diferencia:

$$\begin{aligned} s_{\bar{X}_1-\bar{X}_2} &= \sqrt{\left(s^2_{pool}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \\ &= \sqrt{8.3\left(\frac{1}{17}+\frac{1}{15}\right)} \\ &= \sqrt{8.3(0.126)} \\ &= \sqrt{1.046} \\ &= 1.023 \end{aligned}$$

El error estándar de la diferencia es entonces de 1.023. Finalmente esto es utilizado para calcular el estadístico t:

$$\begin{aligned} t_{\text{obtenida}} &= \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{s_{\bar{X}_1-\bar{X}_2}} \\ &= \frac{(23-20)-0}{1.023} \\ &= \frac{(+3.0)-0}{1.023} \\ &= +2.93 \end{aligned}$$

El valor $t = +2.93$ nos indica la posición de la distribución de diferencias de medias en la cual se ubica nuestra diferencia medida de $+3$. Para interpretar este valor es necesario

compararlo con una tabla de valores críticos de t. Es necesaria una tabla debido a que para cada cantidad de grados de libertad, el histograma de frecuencias de diferencias de medias tiene una forma ligeramente distinta, así que para distintos grados de libertad cada punto de t estará en una localización ligeramente distinta.

En este caso tenemos un tamaño de N de 32, y la manera de calcular los grados de libertad para una prueba t es:

$$gl = N - 2$$

debido a que se tienen dos grupos. Sustituyendo con los valores del ejemplo:

$$gl = 32 - 2 = 30$$

Critical values of *t* for two-tailed tests

Significance level (α)

Degrees of freedom *(df)*	.2	.15	.1	.05	.025	.01	.005	.001
1	3.078	4.165	6.314	12.706	25.452	63.657	127.321	636.619
2	1.886	2.282	2.920	4.303	6.205	9.925	14.089	31.599
3	1.638	1.924	2.353	3.182	4.177	5.841	7.453	12.924
4	1.533	1.778	2.132	2.776	3.495	4.604	5.598	8.610
5	1.476	1.699	2.015	2.571	3.163	4.032	4.773	6.869
6	1.440	1.650	1.943	2.447	2.969	3.707	4.317	5.959
7	1.415	1.617	1.895	2.365	2.841	3.499	4.029	5.408
8	1.397	1.592	1.860	2.306	2.752	3.355	3.833	5.041
9	1.383	1.574	1.833	2.262	2.685	3.250	3.690	4.781
10	1.372	1.559	1.812	2.228	2.634	3.169	3.581	4.587
11	1.363	1.548	1.796	2.201	2.593	3.106	3.497	4.437
12	1.356	1.538	1.782	2.179	2.560	3.055	3.428	4.318
13	1.350	1.530	1.771	2.160	2.533	3.012	3.372	4.221
14	1.345	1.523	1.761	2.145	2.510	2.977	3.326	4.140
15	1.341	1.517	1.753	2.131	2.490	2.947	3.286	4.073
16	1.337	1.512	1.746	2.120	2.473	2.921	3.252	4.015
17	1.333	1.508	1.740	2.110	2.458	2.898	3.222	3.965
18	1.330	1.504	1.734	2.101	2.445	2.878	3.197	3.922
19	1.328	1.500	1.729	2.093	2.433	2.861	3.174	3.883
20	1.325	1.497	1.725	2.086	2.423	2.845	3.153	3.850
21	1.323	1.494	1.721	2.080	2.414	2.831	3.135	3.819
22	1.321	1.492	1.717	2.074	2.405	2.819	3.119	3.792
23	1.319	1.489	1.714	2.069	2.398	2.807	3.104	3.768
24	1.318	1.487	1.711	2.064	2.391	2.797	3.091	3.745
25	1.316	1.485	1.708	2.060	2.385	2.787	3.078	3.725
26	1.315	1.483	1.706	2.056	2.379	2.779	3.067	3.707
27	1.314	1.482	1.703	2.052	2.373	2.771	3.057	3.690
28	1.313	1.480	1.701	2.048	2.368	2.763	3.047	3.674
29	1.311	1.479	1.699	2.045	2.364	2.756	3.038	3.659
30	1.310	1.477	1.697	2.042	2.360	2.750	3.030	3.646
40	1.303	1.468	1.684	2.021	2.329	2.704	2.971	3.551
50	1.299	1.462	1.676	2.009	2.311	2.678	2.937	3.496
60	1.296	1.458	1.671	2.000	2.299	2.660	2.915	3.460
70	1.294	1.456	1.667	1.994	2.291	2.648	2.899	3.435
80	1.292	1.453	1.664	1.990	2.284	2.639	2.887	3.416
100	1.290	1.451	1.660	1.984	2.276	2.626	2.871	3.390
1000	1.282	1.441	1.646	1.962	2.245	2.581	2.813	3.300
Infinite	1.282	1.440	1.645	1.960	2.241	2.576	2.807	3.291

Con un $\alpha = 0.05$ y una prueba de dos colas, el valor crítico de t es de ± 2.042. El valor obtenido está más allá de este punto, por lo que podemos concluir que la diferencia encontrada en la muestra es significativa, es decir, que difícilmente sería encontrada si acaso no existiese una relación entre la raza de perro y la capacidad para recordar comandos.