CUADERNOS DEL GRUPO DE ESTADÍSTICA PARA EL ESTUDIO DEL LENGUAJE

Volumen I

Julia Carden (ed.)
Ezequiel Koile (ed.)
Analí Taboh (ed.)
Federico Alvarez
Santiago Gualchi
Mercedes Güemes
Laura Tallon
Charo Zacchigna

GESEL

Edición
Julia Carden
Ezequiel Koile
Analí Taboh

Cuadernos del Grupo de Estadística para el Estudio del Lenguaje / Federico Alvarez; Santiago Gualchi; Mercedes Güemes; Laura Tallon y Charo Zacchigna.

1era edición.
Buenos Aires.
2021.
ISSN:

CUADERNOS DEL GRUPO DE ESTADÍSTICA PARA EL ESTUDIO DEL LENGUAJE

Volumen I

Federico Alvarez, Santiago Gualchi , Mercedes Güemes, Laura Tallon y Charo Zacchigna.

Editado por Julia Carden, Ezequiel Koile y Analí Taboh.

Prólogo

Sobre este cuaderno

El Cuaderno del Grupo de Estadística para el Estudio del Lenguaje - Volumen I es una obra colectiva. Los distintos capítulos que lo integran son el resultado de una recopilación bibliográfica llevada a cabo dentro del marco del subsidio FiloCyT “Métodos cuantitativos en el estudio del lenguaje”. Cada capítulo fue elaborado para acompañar una reunión del equipo en la que se trataron los distintos temas con mayor profundidad, a modo de presentación oral. Este volumen no pretende ser una obra acabada sobre los temas de estadística que se abordan, sino una guía para la lectura de bibliografía más compleja y completa sobre el tema. Por este motivo, estos capítulos deben ser acompañados de la lecturas que se sugieren en las referencias.

Estos cuadernos fueron concebidos para aportar material de consulta sobre estadística aplicada al estudio del lenguaje que, al menos en español, sigue siendo un área de vacancia bibliográfica. El objetivo es que proporcionen información, acceso a nuevos materiales y una explicación breve y simple sobre los puntos esenciales que deben ser conocidos para quienes se introduzcan en el estudio cuantitativo del lenguaje y no posean una base firme en matemática.

Capítulo IV

Estadística inferencial

María Mercedes Güemes

La estadística inferencial permite generalizar desde una serie de observaciones representativas a un universo más grande de posibles observaciones, usando pruebas de hipótesis (como los *t*-test y ANOVA). En el presente capítulo, se abordan las nociones básicas de la estadística inferencial, como la obtención de la muestra, la postulación y el testeo de hipótesis, los errores y su relación con la significancia estadística. Posteriormente, se presentan tres teorías en las que se basan las pruebas estadísticas actuales, sus ventajas y sus limitaciones (Fisher y Neyman-Pearson, NHST). Por último, se exponen las interpretaciones incorrectas más habituales que se hacen sobre los test estadísticos, valores *p*, intervalos de confianza y poder según Greenland (2016).

1. Introducción

Realizar un análisis descriptivo es un paso crucial para conocer las características de un conjunto de datos. Sin embargo, los resultados se limitan a ese set particular de datos (muestra) y no sirven para deducir propiedades de una población mayor.

A diferencia de la estadística descriptiva, la estadística inferencial posibilita derivar las conclusiones obtenidas de una muestra a un conjunto de datos más amplio. Gracias a la estadística inferencial, es posible estudiar un fenómeno sin necesidad de tener datos sobre toda la población. A modo de ejemplo, si el objetivo de una investigación es determinar si los hablantes del español del Río de la Plata presentan mayor velocidad de habla que los hablantes del español peninsular, no es necesario (ni es posible) tener información sobre todos los hablantes de cada variedad lingüística. A partir de una pequeña parte de cada población (muestra) se pueden aplicar distintos métodos y procedimientos para generar inferencias sobre toda la población.

1.1. Muestras

El punto de partida de la estadística inferencial es la obtención de una *muestra*. La muestra debe ser lo suficientemente representativa para garantizar que todo el análisis estadístico sea correcto. Es importante que la recolección de datos, es decir, el diseño de muestreo, presente un protocolo de base científica (para más información sobre el tema, ver Capítulo II).

Según Johnson (2011) siempre se debería poder balancear las muestras con datos aleatorios de la población. No obstante, en lingüística, es muy habitual que se utilicen muestras no aleatorias por varios motivos. Por ejemplo, cuando se requiere estudiar un fenómeno que sucede en contextos específicos, como en el bilingüismo, o cuando se intenta controlar la variabilidad lingüística de los socio- y cronolectos. Si bien es frecuente que en lingüística experimental se trabaje con muestras no aleatorias, desde el punto de vista estadístico esto no es adecuado. La capacidad de generalizar los resultados a la población depende de que se

obtenga una buena muestra. Se debe considerar que un buen método estadístico no compensa una mala muestra (Johnson, 2011).

Una vez que se obtuvo la muestra sobre la que se va a trabajar, el objetivo es la estimación de parámetros de la población y el testeo de hipótesis. Dado que no se cuenta con información sobre los valores reales de la población, lo que se suele hacer es estimar los parámetros a través de los estadísticos que presenta la muestra. Para ello, se van a distinguir dos tipos de medidas. Las medidas de la población se representan con letras griegas y las de la muestra con letras romanas (para más información sobre cómo se calculan estas medidas, ver Capítulo III).

Parámetros (población)

μ = media poblacional

σ = desvío estándar poblacional

σ^2 = varianza poblacional

Estadísticos (muestra)

$\bar{x}$ = media muestral

s = desvío estándar muestral

s^2 = varianza muestral

1.2. Testeo de hipótesis

Frente a las medias de dos muestras ($\bar{x}$ de la muestra x e $\bar{y}$ de la muestra y), ¿cómo se puede establecer si esas medias son diferentes entre sí? Lo más importante es, en realidad, saber si las diferencias estimadas para las medias $\bar{x}$ e $\bar{y}$ pueden representar las diferencias entre μ_1 y μ_2. De esta forma, se puede establecer el valor de confianza de $\bar{x}$ sobre μ_1.

Por ejemplo, se puede estudiar un fenómeno fonético vinculado a dos tipos de poblaciones: determinar si la emisión de la conjunción *pero* varía según el género del grupo que la emite (datos tomados de Ferrari et al., 2021[1]). La investigación toma como observación la duración de la conjunción *pero* en milisegundos en dos contextos: voces femeninas y voces masculinas. De esta forma, cuenta con dos muestras y sus medidas. La media de emisión de *pero* para las voces masculinas es de 200 ms (media $\bar{x}$ de la muestra x) con un desvío estándar de 8 ms, mientras que la media para las emisiones del grupo de voces femeninas es 172 ms (media $\bar{y}$ de la muestra y) con 32 ms de desvío estándar. A partir de estas medidas, es posible definir, luego de generar una serie de cálculos estadísticos, si esos dos tipos de emisiones (muestras) difieren en la duración de la emisión *pero,* o si esas diferencias numéricas en la duración de *pero* se deben a una variación posible de la toma de dos muestras al azar que no se relaciona con el tipo de voces que emiten la conjunción.

[1] Los datos que se toman en este capítulo corresponden a material no publicado que se obtuvo de la investigación llevada a cabo en ese trabajo.

2. Teorema del Límite Central

Una de las bases teóricas que permiten generar inferencias sobre las medias de la muestra, sin tener información sobre la población total, es el **Teorema del Límite Central** (**TLC**). Dado que es difícil conocer la distribución de toda la población, se toma como dato la distribución de las muestras. Esto significa que si se toman muchas muestras sobre la población, se pueden tener muchas medias muestrales y con eso se puede elaborar una distribución muestral. En ese caso, si se toman muchas muestras (*sampling*), ¿cómo se vería la distribución de todas las medias de esas muestras (la distribución de las muestras o *sampling distribution*)?

El TLC explica que, sin importar la distribución de la población (sea normal, asimétrica u otras), es una propiedad de las medias tener una distribución cercana a la normal. Además, cuando el número de observaciones de cada muestra (N) aumenta, la distribución de las medias se acerca más a la normalidad (Figura 1). En esto se basa el Teorema del Límite Central y sirve para hacer inferencias sobre la media, aunque no se conozca la distribución de la población total. Así, se puede usar la distribución normal o cercana a la normal para generar conclusiones de probabilidad sobre la media (como se hace con los *z-scores*) aunque la distribución de la población no sea normal.

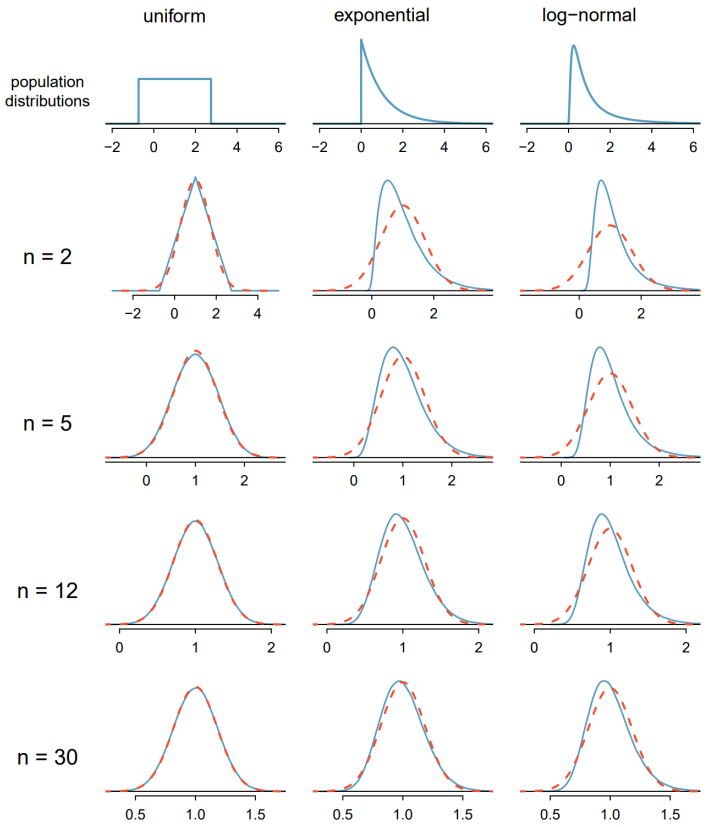

Figura 1. Distribución de las medias en relación al número de casos. Se grafica la distribución de las medias de n *muestras de variables con distribución uniforme, exponencial y log-normal. En línea punteada se superpone una distribución normal con los mismos parámetros (Gráfico obtenido de OpenIntro Statistics, Diez et al., 2012, p. 186)*

Sobre las medidas de dispersión relacionadas con el TLC, se observa que el error estándar (SE, el desvío estándar de todas las medias, ver Capítulo III) disminuye cuando el N de las muestras aumenta. El principio general que rige esto es el siguiente: las estimaciones de la población son más acertadas si se tiene una muestra más grande. Si el SE es muy alto, hay que preguntarse si la muestra es adecuada. El SE también depende de σ (desviación estándar de la población). Si la población tiene menor σ, el SE es menor. No es necesario contar con una distribución de medias para calcular el SE. Si no se conoce sigma, se usa s (desviación estándar de la muestra).

$$SE = \frac{s}{\sqrt{n}}$$

Gracias a estas propiedades de las medias y los cálculos de dispersión, se pueden hacer inferencias sobre la media de una muestra. A pesar de que se puede utilizar este principio para el testeo de hipótesis, el TLC requiere que se cumplan dos condiciones: la muestra debe ser independiente (es decir, las observaciones deben ser aleatorias) y el número de observaciones no debe ser menor a 30 (N > 30). Si bien es importante saber que la distribución muestral posibilita hacer inferencias a partir de una sola muestra, en la práctica la distribución real de las muestras es desconocida. El TLC es un concepto abstracto que permite generar deducciones sobre la muestra con la que se va a trabajar.

3. Hipótesis

Como se observó en el Capítulo III, para sacar conclusiones de probabilidad de las observaciones se las convierte en *z-scores*; de la misma manera, para hacer inferencias sobre la media de la población, se calcula el valor de *t*.

$$t = \frac{\bar{x} - \mu}{s}$$

Dado que no se conoce σ y se utiliza s en su lugar, no es del todo correcto asumir una distribución normal. Por eso, se utiliza la distribución *t*, que es muy similar y da cuenta de cuán certeras son las inferencias sobre σ. La distribución de *t* también es unimodal y simétrica (Figura 2), pero tiene una mayor área en los extremos y menor altura en el centro, lo que permite que más observaciones se sitúen a más de 2 desviaciones estándar de la media (esto compensa que σ se estime mal). Tiene un solo parámetro, los grados de libertad (gl), que son los que determinan el grosor de las colas (a mayor gl, más cercana a la normal). Con la fórmula para el valor de *t*, se puede asignar un valor a μ.

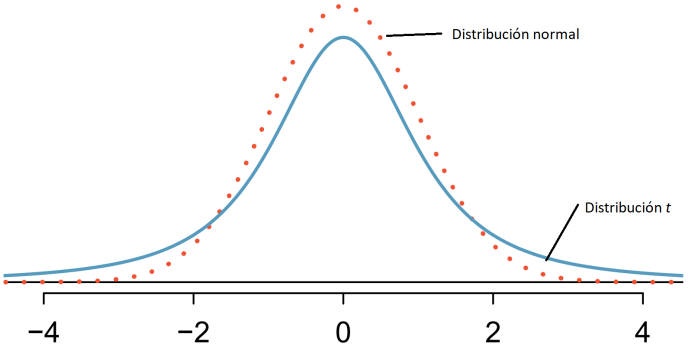

Figura 2. Distribución normal y distribución t (Gráfico obtenido de Open Intro Statistics de Diez, 2012, página 223)

En este punto, antes de realizar los cálculos, es necesario postular las hipótesis que se van a testear. Se distinguen entonces por un lado, la H_0 o hipótesis nula que es aquella que postula la "no diferencia" entre ambas poblaciones x e y. Representa la mirada escéptica sobre un fenómeno. Si en una investigación se quiere probar un nuevo fenómeno, la H_0 postula que este no ocurre. Por otra parte, la H_1 o hipótesis alternativa presenta una postura nueva que se sostiene en la investigación y va en contra de la H_0.

En el ejemplo del caso de la duración de *pero*, los datos necesarios para comparar ambos grupos son la media ($\bar{x}$), el desvío estándar (s) y el número de observaciones (n) para cada grupo.

Duración de *pero*	Media ($\bar{x}$)	Desvío (s)	Observaciones (n)
Voces femeninas	172 ms	32 ms	18
Voces masculinas	200 ms	8 ms	20

Comparando estos datos en bruto, no se puede sacar una conclusión fiable sobre la duración de *pero* en estas dos muestras. No se sabe si la diferencia de esos dos números obedece a una variabilidad esperable en la toma de datos o si esa diferencia revela que hay un fenómeno que la produce. Esto conduce a la postulación de las hipótesis de investigación:

H_0 = no hay diferencias entre esos dos grupos, ambas poblaciones tienen la misma duración en la emisión de la conjunción, que la media sea 172 o 200 no es relevante. La relación entre las variables duración de *pero* y género (f-m) es independiente.

H_0: $\mu_f - \mu_m = 0$

H_1= sí hay diferencias entre los grupos. Las medias revelan que algo está ocurriendo. Una población presenta mayor duración en la emisión de la conjunción que la otra, no se debe al azar la diferencia numérica entre ambas muestras. La relación entre las variables duración de *pero* y género es dependiente.

H_1: $\mu_f - \mu_m \neq 0$

Si bien hay otros tipos de testeo de hipótesis, uno de los más habituales es el de comparar dos medias. Podría compararse la media de una muestra con un valor determinado, o postular que la media de la población es mayor o menor a un número (en ese caso se puede usar la primera fórmula de *t*). En este caso, al comparar dos medias, se utiliza una versión adaptada para el cálculo de *t*. Antes de esto, se debe calcular el error estándar para la diferencia de dos medias. A continuación, se observa que la forma de calcular el SE para dos muestras es obtener la raíz de la suma del desvío de las dos muestras. Esta es la fórmula (a la derecha está aplicada con el ejemplo de los datos de la duración de *pero*):

$$SE = \sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}} = \sqrt{\frac{32^2}{18} + \frac{8^2}{20}} = 7{,}75$$

En el cálculo de t, se puede ver que el valor estimado es la diferencia de las medias de las muestras y se le resta el valor de la hipótesis nula, es decir 0, y esto se divide por el valor del error.

$$t = \frac{valor\ estimado - valor\ nulo}{error\ estándar} = \frac{(200 - 172) - 0}{7{,}75} = 1{,}03$$

Si se aplica la fórmula, se obtiene el valor de t = 1,03. El valor de *t* indica en qué parte de la distribución se encuentra el valor del estimado (en este caso, diferencias entre las medias). Por este motivo, al valor de t se le puede asignar una probabilidad o valor de *p* que permite interpretar el resultado en relación a las hipótesis de investigación.

3.1. El valor de *p*

El valor de *p* o el *p*-valor se define como la probabilidad de obtener el valor observado o uno más extremo asumiendo que la hipótesis nula es verdadera. Esto significa que, para hacer un testeo de hipótesis, se parte de la veracidad de la hipótesis nula. Así, la pregunta que conduce el testeo de las hipótesis es: dado que la H_0 es verdadera, ¿cuál es la probabilidad de que una muestra tenga el valor observado?

Si bien el valor de *p* asociado a un estadístico (*z*, *t*, *F*, entre otros) se puede obtener mediante ecuaciones, la forma más habitual es utilizar un método que lo compute automáticamente. Se pueden usar plataformas como R o SPSS para obtener estos cálculos estadísticos. En el próximo volumen de estos cuadernos se abordará el análisis estadístico de datos lingüísticos con R.

En el ejemplo tomado anteriormente, se contaba con un valor de *t* de 1,03 para la diferencia de las dos muestras. El valor de *p* asociado a este valor es 0.31. Esto significa que, si se asume que la diferencia de las medias es 0, la probabilidad de que la diferencia entre dos muestras aleatorias sea 28 ms es de 31 veces en 100. Si bien parece poco que las probabilidades de obtener esta diferencia de medias sea de aproximadamente el 31%, hay una manera de vincular este número con la verificación de las hipótesis de investigación.

3.2. Errores

Para rechazar o aceptar una hipótesis, se cuenta con el valor de *p* que manifiesta una probabilidad asociada a un estadístico, calculado a partir de las características de una muestra. Junto con esa probabilidad, se encuentra el porcentaje de error asumido de que la interpretación sobre las hipótesis sea incorrecta. Existen dos tipos de errores asociados al valor de *p*:

	H_0 es verdadera	H_0 es falsa
Aceptar H_0	Correcto	**Error de Tipo II**
Rechazar H_0	**Error de Tipo I**	Correcto

¿Cómo se cuantifica esta diferencia? Hay que elegir una probabilidad del Error de Tipo I (falso positivo) que estemos dispuestos a tolerar. El criterio para establecer esta probabilidad de Error de Tipo I se llama **nivel de significancia (α)**. Un nivel de α aceptable y que se usa habitualmente en lingüística y otras disciplinas similares es 0,05.

En el caso de la duración de *pero*, el valor de *p* es 0,096. Si se establece como límite aceptable α menor a 0,05, el valor *p* para las muestras no tiene significancia estadística. Es decir, ese valor de *p* no es suficiente para rechazar la H_0. Dicho de otra manera, a partir de las muestras obtenidas no hay evidencias suficientes para afirmar que H_0 no es verdadera. Como no se puede rechazar H_0, se asume que no hay diferencias en la emisión de *pero* en relación al género.

También se puede definir el margen de probabilidad del Error de Tipo II (falso negativo). Esto se conoce como **beta (β)**. Un nivel de beta aceptable es 0,2. El **poder o potencia estadística** se basa en beta y se define como 1-β. El poder estadístico se aumenta con un incremento de N, ya que hace que la prueba sea más sensible a pequeñas diferencias. Sin embargo, el poder estadístico tiene un nivel de máxima que se alcanza después de un valor determinado de N.

4. Teorías de testeo de hipótesis

4.1. La propuesta de Fisher

A partir de 1925, Fisher se encargó de desarrollar y promover tests de significancia (Perezgonzalez, 2015). La perspectiva de Fisher sobre el análisis de datos se puede resumir en cinco pasos:

1. Elegir el test correcto de acuerdo con la naturaleza de las variables con las que se trabaja.

El tipo de test lo determina la distribución elegida, otras características vienen con la muestra.

2. Establecer la hipótesis nula.

Puede ser direccional (si se espera un resultado determinado) o no direccional (no hay predicciones sobre los datos). La H_0 no tiene que ser siempre nula (= 0). Puede ser que una diferencia no supere un valor específico.

3. Calcular la probabilidad teórica de los resultados observados considerando que H_0 es cierta (valor de *p*).

Cuando los datos de la muestra se acercan a la media de la distribución nula, la probabilidad aumenta (el valor de *p* es más alto); cuanto más se aleja el valor del centro de esa distribución, menos probable resulta ese valor (el valor de *p* es más bajo). El valor de *p* no es una probabilidad de un punto exacto, es una probabilidad acumulada del área que va desde el punto observado hasta la cola de la distribución. Es importante recordar que la H_0 siempre es verdadera, no se puede falsear, porque toda la distribución del test está basada en esta hipótesis.

4. Evaluar la significancia estadística α de los resultados.

Para determinar si el valor de *p* es lo suficientemente bajo para rechazar la H_0, se debe establecer el nivel de significancia. Los niveles de confianza más elegidos son 5% ($\alpha \approx 0{,}05$) o 1% ($\alpha \approx 0{,}01$). Los test estadísticos pueden ser de una cola de dos colas. En este último caso, el nivel de significancia se divide en áreas de ambos extremos (positivo y negativo). Así, el 5%, por ejemplo, cubriría el 2,5% de cada lado. Si se realizan múltiples tests, se incrementa la probabilidad de encontrar significancia estadística en los resultados. Para ello, se utilizan las correcciones que nivelan hacia abajo el valor de *p* (p. ej., la corrección de Bonferroni, que es el que más se usa a pesar de ser muy conservadora).

5. Interpretar el nivel de significancia de los resultados.

Un resultado significativo se explica de una de las siguientes dos formas. O este resultado ocurre raramente con probabilidad *p*, o la H_0 no explica los resultados satisfactoriamente. En general, los resultados no significativos no se reportan. Sin embargo, según Fisher aportan información y pueden resultar un medio para confirmar o reforzar la H_0. [Nota: rechazar o dudar de la H_0 bajo la perspectiva de Fisher, estadísticamente, no convierte a lo opuesto en verdadero, el valor de *p* no sirve para apoyar la H_1].

Para destacar los puntos positivos de la perspectiva de Fisher, su propuesta es flexible porque puede ser utilizada para investigación exploratoria. Además, es inferencial, ya que permite ir de la muestra a la población. Tiene también puntos negativos. Fisher nunca explicitó el poder estadístico de las pruebas. Sí habló de mayor "sensibilidad" cuando se aumentaba la muestra, pero nunca hizo un cálculo matemático para controlar este poder. Otra de las críticas es que en el marco teórico de Fisher no hay declaración explícita de la H_1, esta es la negación implícita de la H_0.

4.2. La propuesta de Neyman y Pearson

Al intentar mejorar la propuesta de Fisher, los autores terminaron por elaborar una forma alternativa para el testeo de datos, más matemática que la anterior. La propuesta de Neyman y Pearson (Neyman y Pearson, 1928, 1933; Neyman, 1956) se puede resumir en los siguientes pasos:

1. Establecer el tamaño del efecto esperado para una población.

Una de las innovaciones de esta perspectiva es explicitar la H_1 cuando se están explorando los datos. La H_1 representaría una segunda población que se sitúa junto a la población principal con el mismo continuo de valores. Estas dos poblaciones difieren en el tamaño del efecto. El tamaño del efecto es el grado de posibilidad de visualizar diferencias entre poblaciones. Cuanto menor es el tamaño del efecto, menos posibilidad de identificar pequeñas diferencias entre ellas. Como, en general, se desconocen los parámetros de la población, se recurre al tamaño del efecto de la población de muestras (*sampling distribution*). Las muestras no se superponen, están separadas. El porcentaje de distribución conocido es llamado beta (β) o **MES (*Minimum Effect Size*)** y representa la parte de la hipótesis principal HM (H_0)que no va a ser rechazada por el test.

2. Seleccionar un test óptimo.

Se debe elegir un test estadístico según su poder (los paramétricos tienen más poder que los no paramétricos) y en condiciones que aumenten el poder (por ejemplo., ampliar el N).

3. Definir la hipótesis principal (HM)

De acuerdo con Neyman y Pearson (1928), siempre se deben definir al menos dos hipótesis. La principal es la que debe ser testeada y debe incorporar el MES [Nota: H_0 y HM son muy similares y se postulan igual. Neyman y Pearson la llaman hipótesis nula también. Sin embargo, HM se postula explícitamente, incorpora un MES y compite con otra hipótesis, Figura 4].

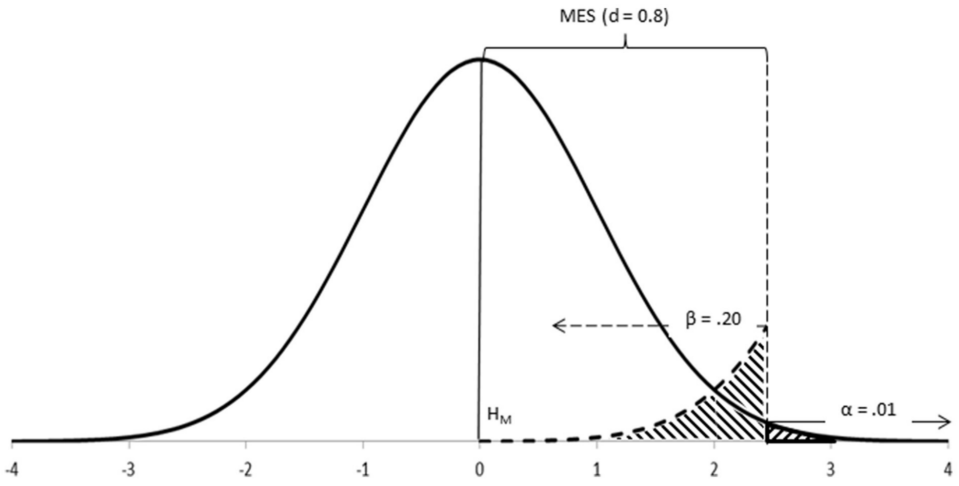

Figura 4. MES para calcular el valor de beta (Gráfico obtenido de Perezgonzalez, 2015, página 5)

Error de Tipo I: rechazar incorrectamente la HM (y aceptar H_1 incorrectamente). Para Neyman-Pearson este error tiene relevancia a largo plazo (no es identificable en un solo ensayo) y debe controlarse durante el diseño de un proyecto, no se puede controlar *a posteriori*.

Alfa (α): probabilidad de cometer el Error de Tipo I. Los niveles de alpha más utilizados son 0,05 y 0,01. A diferencia de la significancia estadística, alpha se debe incorporar en la postulación de la HM, no admite gradación y no se basa en rechazar la HM sino en aceptar una H_1.

Región crítica: alfa permite generar una región crítica, a partir de la cual se define la probabilidad de ese valor según las hipótesis de investigación. Si el valor del test cae fuera de la región, es probable dentro de la HM; en cambio, si entra en la región crítica, es un valor probable de la H_1.

Valor crítico: es el valor que delimita la región crítica y se define a priori, en la postulación de la HM.

HM : $Media_1 – Media_2 = 0 \pm MES$, $\alpha = 0.05$, $CVt = 2.38$

4. Definir la hipótesis alternativa (H_1)

No es necesario establecer valores definidos para la H_1 (incluso los autores las definen vagamente), solo es una oposición de la HM que debe incluir el MES.

Error de Tipo II: rechazar incorrectamente la H_1. Es menos grave que el Error de Tipo I.

Beta (β): probabilidad de cometer el Error de Tipo II. No puede ser más pequeña que alpha (porque lo que testeamos es la HM), pero debe reducirse lo más posible. Neyman-Pearson proponen fijar β en el máximo 0,20 y el mínimo en α. Debe incorporarse este valor en la H_1.

H_1: $Media_1 – Media_2 \neq 0 \pm MES$, $\beta = 0,20$

5. Calcular el tamaño de la muestra (N) para un buen poder estadístico (1–β).

El poder o potencia estadística es la probabilidad de rechazar correctamente HM en favor de la H_1. Es matemáticamente opuesta al error del Tipo II, es decir que es 1–β.

6. Calcular el valor crítico del test.

A partir de N y alpha podemos definir el valor crítico para delimitar los rangos desde donde evaluar nuestras hipótesis.

7. Calcular el valor del test de la investigación.

Cuando este valor está más cerca de cero, los datos están más cerca de la HM, mientras que cuanto más se alejan de cero, más se alejan de la HM. [Nota: desde la perspectiva de Neyman-Pearson se pueden usar los *p*-valores, solo que hay que recordar que los valores van en dirección opuesta].

8. Decidir a favor de la HM o la H1

Neyman (1955) dice:

- Si el resultado observado cae en la región crítica, rechazar HM y aceptar H_1.
- Si el resultado cae fuera de la región crítica y el test tiene mucha potencia, aceptar HM y rechazar H_1.

- Si el resultado cae fuera de la región crítica y el test tiene baja potencia, no se puede sacar conclusiones.

Como puntos a favor, la perspectiva de Neyman-Pearson incorpora la potencia o poder estadístico para evitar errores de Tipo II y, además, es una perspectiva mejor para investigaciones que toman diversas muestras de la misma población. Tiene en contra una menor flexibilidad y puede confundirse con la teoría de Fisher si no se tienen en cuenta el MES y beta.

4.3. NHST (Null Hypothesis Significance Testing)

Es la perspectiva que más se usa actualmente. Es una amalgama, sin criterio, de las propuestas de Fisher y Neyman-Pearson. Según el autor que la usa, puede tender más hacia una teoría o la otra. En general, se usa la perspectiva práctica de Neyman-Pearson y la filosofía de Fisher. Según Perezgonzalez (2015) se utiliza en todos los ámbitos (editores, investigadores, revisores), pero es una pseudociencia. El autor propone solucionar esto acercando la NHST a las dos teorías. Recomienda el uso de G*Power para implementar las recomendaciones para ambos casos. La teoría de Fisher es más cercana a la NHST. Muchos paquetes estadísticos la tienen como base (por ejemplo, SPSS). Se puede mejorar esto introduciendo los conceptos de poder estadístico y tamaño del efecto. No habla de H_1 ni nada por el estilo. La NHST es muy dañina para la teoría de Neyman-Pearson. Un error grave es la utilización de los valores *p* como evidencia de errores de Tipo I. Una solución es ajustar alfa y beta.

5. Malas interpretaciones sobre los test estadísticos, valores *p*, intervalos de confianza y poder (Greenland et al., 2016)

Debemos tener cuidado con la forma en que interpretamos los conceptos estadísticos en la práctica, pues estos no siempre son entendidos completamente por el investigador. Listamos aquí algunos ejemplos típicos de cómo se los suele malinterpretar.

5.1. Formas habituales de malinterpretar un único valor de *p*

1. El valor de *p* es la probabilidad de que la hipótesis principal sea cierta (p. ej., *p* = 0,01 es 1% de probabilidad de que H_0 sea verdad).

No. El test asume que H_0 es verdadera. El valor de *p* indica el grado en el cual los datos de la muestra se ajustan al modelo que predice el test. *p* = 0,01 indica que los datos no se acercan al modelo y a la hipótesis sostenida en el test (si todos los supuestos se cumplen).

2. El valor de *p* indica la probabilidad de que la asociación se deba al azar (chance).

No. Esta es una falacia que deriva de la anterior. El valor de *p* es la probabilidad de un estadístico asumiendo que los supuestos son verdaderos (un supuesto sería que la asociación se deba al azar).

3. Un resultado significativo ($p \leq 0,05$) significa que la H_0 es falsa o debe rechazarse.

No. Un valor bajo de *p* indica que los datos analizados son inusuales dados los supuestos estadísticos del test (entre ellos la H_0). Puede ser un valor causado por un gran error aleatorio o por una violación de algún otro supuesto.

4. Un resultado no significativo ($p > 0,05$) indica que la H_0 es verdadera o debe ser aceptada.

No. Un valor alto de *p* indica que el valor no es inusual dados los supuestos asumidos en el test. Puede ser un valor causado por un gran error aleatorio o por una violación de algún otro supuesto. Puede ser que los datos sean usuales para otra hipótesis también o bajo otros supuestos.

5. Un valor alto de *p* es evidencia a favor de la H_0.

No, salvo que *p* sea 1. Cualquier *p* menor a 1 es indicio de que la hipótesis del test no es compatible con los datos.

6. Un valor de *p* mayor a 0,05 significa que ningún efecto fue observado.

No. Significa que la H_0 sigue siendo una entre otras hipótesis sobre los datos. Salvo que el punto observado coincida exactamente con el punto nulo, es un error definir los resultados como "sin evidencia de un efecto".

7. La significancia estadística indica que se encontró una relación importante a nivel científico.

No. Puede deberse a una violación de supuestos o a errores de una muestra grande. El hecho de que sea inusual no significa que sea científicamente relevante.

8. La falta de significancia estadística indica que el tamaño del efecto es pequeño.

No. Especialmente en un estudio chico, grandes efectos pueden estar tapados por ruido.

9. El valor de *p* indica el porcentaje de chances de que ocurra nuestra observación si la H_0 es real.

No. Muestra la distribución de la frecuencia de la data si los supuestos son ciertos.

10. Si se rechaza H_0 porque $p < 0,05$ significa que la probabilidad del error de "falso positivo" es 5%.

No. Si rechazo H_0 y es real, el error es del 100%.

11. Un valor de $p = 0,05$ y $p \leq 0,05$ significan lo mismo.

No. Uno es un punto determinado y el otro incluye los valores más extremos.

12. Una forma correcta de reportar *p* es a partir de un valor que no incluya el = (p. ej., reportar $p < 0,02$ cuando $p = 0,015$)

No. No se pueden interpretar bien los resultados. Únicamente cuando los valores *p* son muy bajos pierden precisión (0,001).

13. La significancia estadística es una propiedad del fenómeno, entonces una prueba estadística detecta esta significancia.

Este error se manifiesta en la frase "se encontró evidencia de significancia estadística…". La significancia estadística es una descripción del valor de *p*, no una propiedad de la población.

14. Siempre se debe usar un valor de *p* bilateral.

No. Es lo más habitual y se usa si la H_0 incluye un valor específico (p. ej., 0). Si la pregunta de investigación es unilateral, es válido usar un valor de *p* unilateral.

5.2. Formas habituales de malinterpretar varios valores *p* y comparaciones

15. Cuando se testea una hipótesis y todos, o casi todos los resultados muestran $p > 0,05$, se asume que es una forma de verificar la H_0.

No. Que un grupo de muestras individuales no den significancia estadística no implica que sea una evidencia total de "falta de efecto".

16. Cuando los valores *p* obtenidos de dos muestras se encuentran en lados diferentes de 0,05, los resultados se interpretan como opuestos.

No. Los test estadísticos son sensibles a las variaciones de las muestras. Eso se puede reflejar en *p*, pero pueden presentar las mismas asociaciones. Las diferencias de resultados pueden evaluarse con pruebas (llamadas análisis de la heterogeneidad, interacción o modificación).

17. Cuando la misma hipótesis es testeada en diferentes poblaciones y se obtienen valores *p* similares, los resultados concuerdan.

No. Dos estudios pueden mostrar valores *p* similares pero no mostrar las mismas asociaciones, debido a diferencias en las poblaciones.

18. Si se observa un valor de *p* muy bajo en un estudio, es altamente probable que en el próximo estudio se obtenga un resultado similar.

No. El valor de *p* indica justamente la probabilidad de obtener un valor similar o menor. Es decir, si $p = 0,03$, la probabilidad de que sea igual o menor es 3%. La probabilidad de la réplica es exactamente el valor de *p*. Esto siempre que sean muestras independientes y que los supuestos sean ciertos en ambos estudios. Si no es así, el valor de *p* es muy sensible al tamaño de la muestra y a las violaciones de los supuestos.

5.3. Formas habituales de malinterpretar los intervalos de confianza

El problema empieza cuando se interpreta que $p > 0,05$ significa que la H_0 tiene 5% de chances de ser falsa y surgen las siguientes falacias:

19. El intervalo de confianza específico presentado en un estudio tiene 95% de probabilidad de contener el verdadero tamaño del efecto.

No. La frecuencia con la cual un intervalo contiene el efecto real es 100% si lo contiene y 0%, si no lo contiene El 95% refiere a la cantidad de intervalos de confianza que contienen el valor real. Esto significa que el 95% de los intervalos de confianza creados por muestras del mismo tamaño de la misma población van a contener el parámetro real de la población, si todos los supuestos son reales.

20. Un resultado fuera del intervalo de confianza del 95% ha sido refutado o excluido por los datos.

No. Lo que significa es que la combinación de los datos con los supuestos asumidos, en relación al criterio del 95%, es incompatible con los datos observados por algún motivo.

21. Si dos intervalos de confianza se superponen significa que las diferencias entre los valores estimados o de los estudios no es significativa.

No. Los intervalos pueden coincidir, pero sus diferencias pueden producir $p < 0,05$ por diferencias entre estudios.

22. Un intervalo de confianza de 95% observado predice que el 95% de los puntos estimados de los futuros estudios van a estar en dentro de ese intervalo.

No. El 95% se refiere a la cantidad de otros intervalos no observados que van a contener el verdadero efecto, no cuál es la frecuencia de futuros puntos estimados sobre ese intervalo.

23. Si un intervalo incluye el valor nulo y otro lo excluye, el primero es más preciso.

No. La precisión depende de la anchura del intervalo.

24. Si se rechaza la hipótesis nula porque *p* > 0,05 y el poder estadístico es de 90%, las chances de cometer un falso negativo son del 10%.

No. El 10% se refiere a la cantidad de veces que se cometería este error en este test estadístico a través de muchos estudios si la hipótesis alternativa es cierta, no se refiere a un uso singular del test.

25. Si *p* supera el 0,05 y el poder es del 90%, los resultados apoyan la hipótesis nula y no la alternativa.

Si bien es un razonamiento intuitivo, hay contraejemplos en los cuales el valor de *p* para la hipótesis nula está entre 0,05 y 0,10 y, sin embargo, las alternativas tienen un valor de *p* que excede el 0,10 y el poder es del 90%. Este tipo de cálculos pueden explicar la crisis de replicación, ya que definir los resultados de un estudio único de forma dicotómica anticipa un problema.

6. Referencias

Cetinkaya-Rundel, M. (2019). Data Analysis. Duke University. Coursera.

Diez, D. M., Barr, C. D., & Cetinkaya-Rundel, M. (2012). *OpenIntro statistics*. CreateSpace Independent Publishing Platform.

Ferrari L., Güemes M. M., Tallon L., Torres H., & Urcelay M. B. (2021). Correlatos prosódicos de los distintos valores de la conjunción pero. *Verba: Anuario Galego de Filoloxía, 48*. https://doi.org/10.15304/verba.48.6477

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological), 17*, 69–78. https://doi.org/10.1111/j.2517-6161.1955.tb00180.x

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337-350. https://doi.org/10.1007/s10654-016-0149-3

Johnson, K. (2011). *Quantitative methods in linguistics*. John Wiley & Sons.

Neyman, J. (1967). R. A. Fisher (1890-1962): an appreciation. *Science, 156*, 1456–1460. https://doi.org/10.1126/science.156.3781.1456

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika, 20A*(1/2), 175–240. https://doi.org/10.2307/2331945

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, 231*, 289–337. https://doi.org/10.1098/rsta.1933.0009

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6*, 223. https://doi.org/10.3389/fpsyg.2015.00223


El material completo se encuentra en https://gesel.github.io/