Estadística inferencial – Parte I

1. Introducción

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).

2-Muestras

La estadística inferencial se realiza a partir de muestras, dado que es muy complejo trabajar con toda la población. Si queremos estudiar un fenómeno lingüístico, por ejemplo, resulta casi imposible tener datos de la población total.

Las medidas de la población son: μ σ σ^2

Las medidas de la muestra son: $\bar{x}$, s, s^2

En lingüística, en general, se utilizan muestras no aleatorias por varios motivos[1]. Según Johnson, esto no es lo adecuado, siempre deberíamos poder balancear las muestras con datos aleatorios de la población. La capacidad de generalizar los resultados a la población depende de que se obtenga una buena muestra. Un buen método estadístico no compensa una mala muestra. Hay demasiadas variables en juego en lingüística, siempre hay que ser conscientes de esta complejidad.

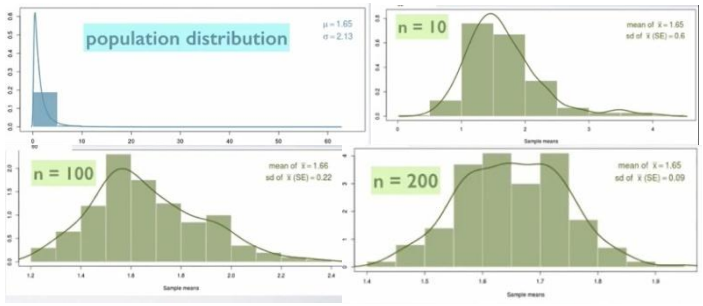
[1] Nota: porque queremos estudiar un fenómeno que ocurre en contextos específicos ej. bilingüismo, o porque queremos controlar la variablilidad lingüística de los socio y cronolectos.

2. Testeo de hipótesis

Frente las medias de dos muestras ($\bar{x}$ de la muestra x_i e $\bar{y}$ de la muestra y_i) ¿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.

3. Teorema del Límite Central

Como en muchos casos es difícil conocer la distribución de toda la población, tenemos como dato la distribución de las muestras. 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")? Sin importar la distribución de la población (sea normal, asimétrica u otras) es una propiedad de las medias caer en 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. Esto se conoce como Teorema del Límite Central y sirve para hacer inferencias sobre la media, aunque no sepamos 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.



Sobre las medidas de dispersión relacionadas con la CLT, se observa que el error estándar (el desvío estándar de todas las medias) disminuye cuando el N de las muestras aumenta. El principio general es que podemos hacer estimaciones de la población más acertadas si tenemos una muestra más grande. Si el SE es muy alto, hay que preguntarse si la muestra es adecuada. El SE también depende de σ (SD de la población), si la población tiene menos SD, el SE es menor.

No es necesario una distribución de medias para calcular el SE. Si no se conoce sigma, se usa s (SD de la media)

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

Gracias a estas condiciones de las medias y los cálculos de dispersión, podemos hacer inferencias sobre una media de una muestra.

Condiciones para CLT:

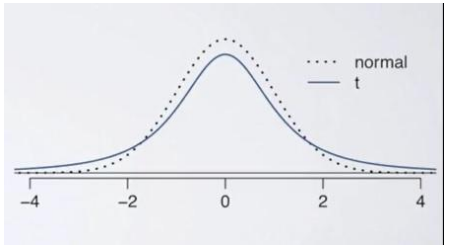
- Independencia de la muestra (observaciones aleatorias)
- N>30

4. Ho

Para sacar conclusiones de probabilidad para las observaciones se las convierte en z-scores; de la misma manera, para hacer inferencias sobre la media de la población, se utiliza un método similar para calcular el valor de t.

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

Como no conocemos σ, usamos s. Esto significa que no es del todo correcto usar la distribución normal. Por eso, se utiliza la distribución *t*, que es muy similar y da cuenta de cuán certeros estamos sobre σ.



La distribución de *t* también es unimodal y simétrica, pero tiene un área adicional en los extremos y menor altura en el centro, lo que permite que más observaciones se sitúen a más de 2SD de la media (esto compensa que sigma se estime mal). Tiene un solo parámetro, los grados de libertad, que son los que determinan el grosor de las colas (mayor gl, más cercana a la normal). Con la fórmula para el valor de *t*, podemos asignar un valor a μ, aunque no lo conozcamos.

Por ejemplo, en el caso del VOT, teníamos $\bar{x}$=84,6. Podemos decir que la media de la población, μ , es igual a 100 ms. Y ahí tenemos nuestra primera hipótesis, que sería la H_0 o hipótesis nula de la "no diferencia". En ese caso evaluamos cuánto difiere $\bar{x}$ de μ en relación al SE. Si se aplica la fórmula, obtenemos el valor de t=-2,16, que significa que si asumimos que la media de la población es 100, la

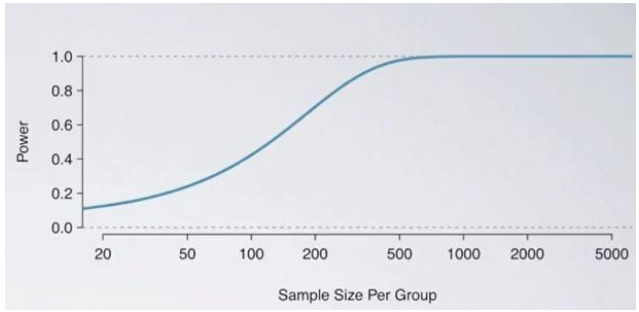
probabilidad de que $\bar{x}$ sea 84,6 es de 2 veces en 100 (muy baja). La probabilidad de que esa sea la media es menos de 2%, ya que la probabilidad de ese valor *t* de es p=0,02. Como es bastante baja, debo rechazar H_0 y admitir la H_A.

H_0 μ=100
H_A $\mu \neq 100$

Errores:

	H_0 es verdad	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 que podamos tolerar. El criterio para establecer esta probabilidad de Error de Tipo I se llama nivel de alfa (α). Un nivel de α aceptable es 0,05.
También se puede definir el margen de probabilidad del Error de Tipo II. 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.



5. Teorías de testeo de hipótesis

5.1 La propuesta de Fisher

A partir de 1925, Fisher se encargó de desarrollar y promover test de significancia. 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 trabajamos.**
   El tipo de test 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 la data de la muestra se acerca a la media de la distribución nula, la probabilidad aumenta (el valor de *p* es más alto); cuanto más se aleja nuestro 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 [Nota: 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 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% (sig ≈ 0,05) o 1% (sig ≈ 0,01). [Nota: se pueden reportar los valores del test y con eso sacar conclusiones] Los test estadísticos son de una cola (f tests) o de dos colas (t tests). 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 (ej. Bonferroni, que es el que más se usa pero es muy conservador).

5. **Interpretar el nivel de significancia de los resultados.**
   Un resultado significativo debería explicarse en una afirmación doble. O este resultado ocurre raramente con la probabilidad p o la H_0 no explica los resultados satisfactoriamente. En general los resultados no significativos no se reportan, sin embargo según Fisher, nos aporta información y puede 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_A]

Para destacar los puntos positivos de la perspectiva de Fisher, su propuesta es flexible porque pudo ser utilizada para testear muchas cosas y 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_A, la H_A es la negación implícita de la H_0.

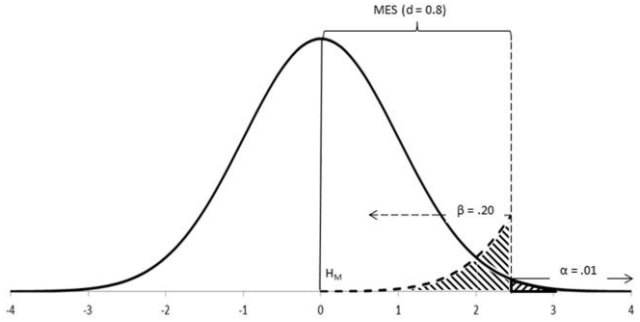
5.2 La propuesta de Neyman y Pearson

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. 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_A cuando se están explorando los datos. La H_A representaría una segunda población que se sitúa junto a la población de 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 H_M 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 (ej. ampliar el N)
3. **Definir la hipótesis principal (H_M)**
   De acuerdo con Neyman-Pearson, 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 H_M son muy similiares y se postulan igual Neyman-Pearson la llaman hipótesis nula también. Sin embargo, H_M se postula explícitamente, incorpora un MES y compite con otra hipótesis.



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

Alpha (α): probabilidad de cometer el Error de Tipo I. Los niveles de alpha más utilizados son (α = 0.05) y (α = 0.01) A diferencia de la significancia estadística, alpha se debe incorporar en la postulación de la H_M, no admite gradación y no se basa en rechazar la H_m sino en aceptar una H_A.

Región crítica: Alpha 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 H_M; en cambio, si entra en la región crítica, es un valor probable de la H_A

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

$$H_M : M1–M2 = 0 \pm MES, \alpha = 0.05, CVt = 2.38$$

4. **Definir la hipótesis alternativa (H_A)**
   No es necesario establecer valores definidos para la H_A (incluso los autores las definían vagamente) solo es una oposición de la H_M que debe incluir el MES.

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

Beta (β): probabilidad de cometer el Error de Tipo II. No puede ser más chiquita que Alpha (porque lo que testeamos es la H_M) 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_A.

$$HA: M1–M2 \neq 0 \pm MES, \beta = 0{,}20$$

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 H_M en favor de la H_A. Es matemáticamente opuesta al error del Tipo II es decir 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 H_M, mientras que cuanto más se alejan de cero, más se alejan de la H_M [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 HA**

   Neyman (1955) dice:

   - Si el resultado observado cae en la región crítica, rechazar H_M y aceptar H_A
   - Si el resultado cae fuera de la región crítica y el test tiene mucha potencia, aceptar H_M y rechazar H_A
   - Si el resultado cae fuera de la región crítica y el test tiene baja potencia, no podés sacar conclusiones. Neyman no entiende por qué elegiste ese test, hiciste todo mal.

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: menor flexibilidad y puede confundirse con la teoría de Fisher si no se tienen en cuenta el MES y Beta.

5.3 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. Según el autor que la usa, puede tender más para una teoría que otra. En general, se usa la perspectiva práctica de Neyman-Pearson y la filosofía de Fisher. La usamos todos y en todos los ámbitos (editores, investigadores, revisores), pero es una pseudociencia. Perezgonzalez 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 (SPSS). Se puede mejorar esto introduciendo los conceptos de poder estadístico y tamaño del efecto. No habla de H_A 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 de p como evidencia de errores de Tipo I. Una solución es ajustar alpha y beta.