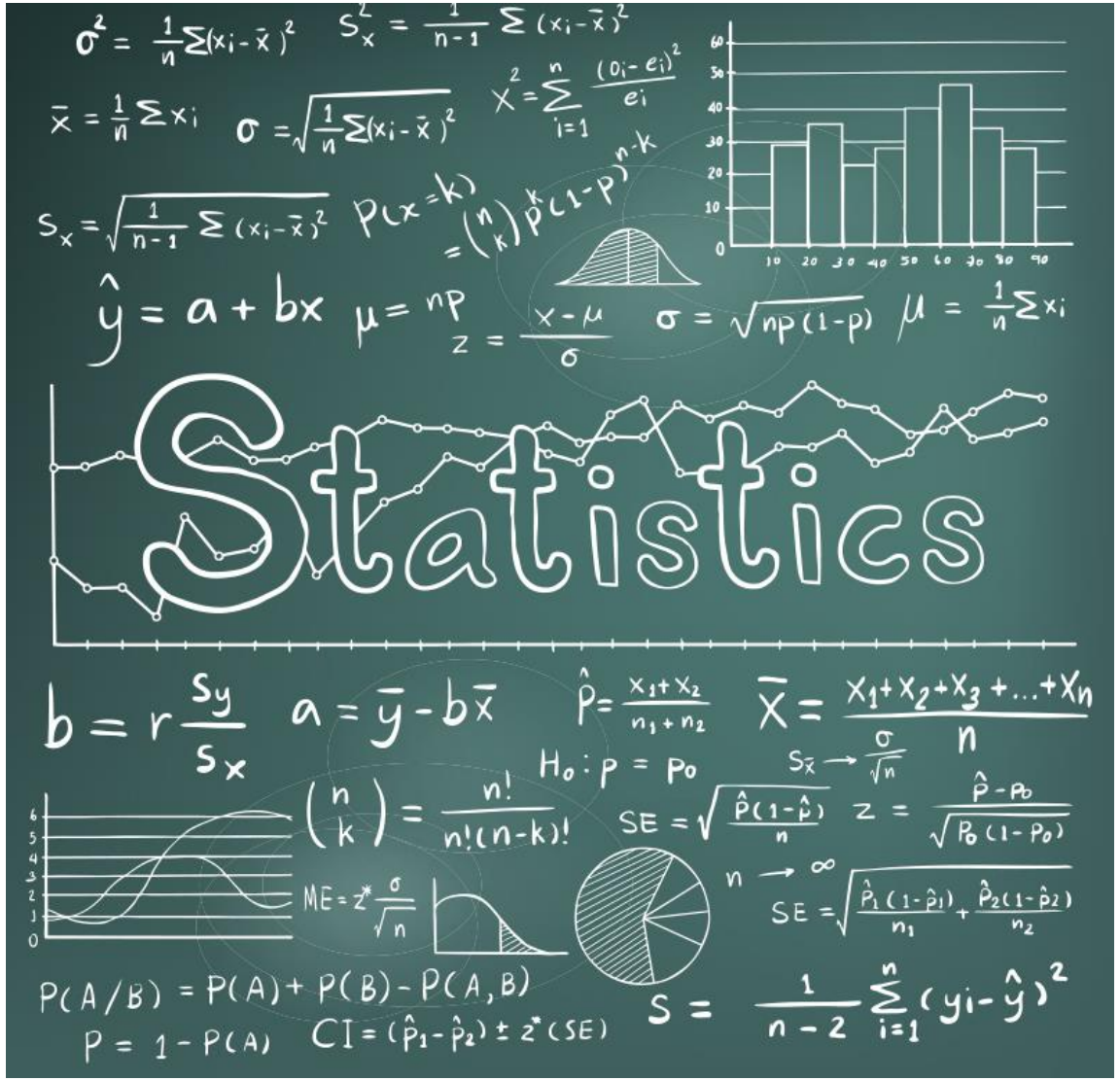
Expresión Diferencial de Genes y su Estadística

7 de Agosto 2024

Estadística

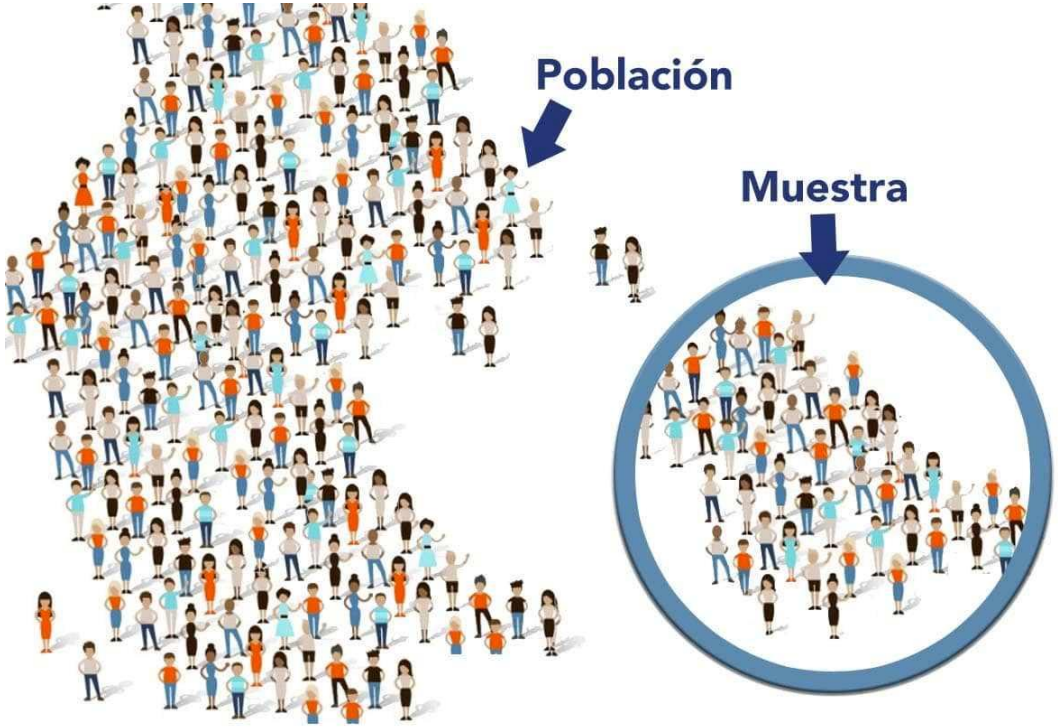
En el siglo XVIII el término "estadística" designaba la colección sistemática de datos demográficos y económicos por los estados. A principios del siglo XIX, el significado de "estadística" fue ampliado para incluir la disciplina ocupada de recolectar, resumir y analizar los datos.



Muestra vs Población

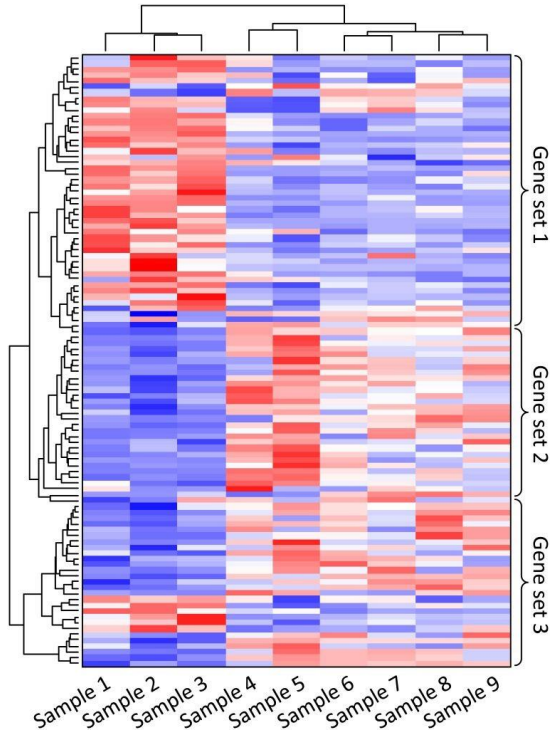
La población en investigación es un conjunto completo de elementos que poseen un parámetro común entre sí.

Una muestra es la parte más pequeña del total, es decir, un subconjunto de toda la población.

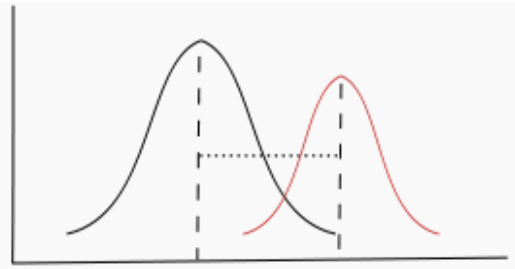


Pruebas de Hipótesis

Análisis de expresión diferencial



$$t = \frac{\bar{X} - \bar{Y}}{\sigma(X - Y)} = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{\sigma^2_X}{n_X} - \frac{\sigma^2_Y}{n_Y}\right)}}$$



Análisis de expresión diferencial

¿Cuál es la probabilidad de sacar 10 caras tirando 10 veces una moneda?

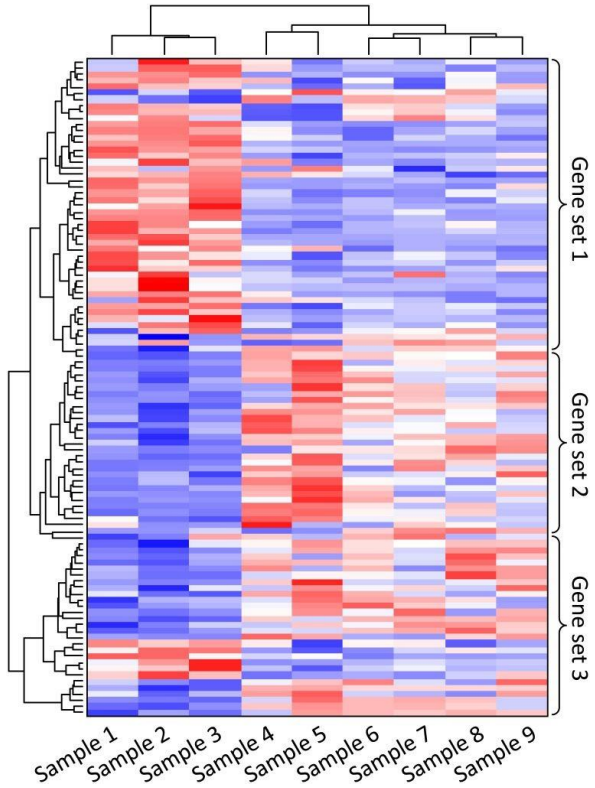
- P= (1/2)^10= 0,00098

¿Cuál es la probabilidad de sacar 10 caras tirando 10 veces pero si tiramos a la vez 1000 monedas?

- P= 1-(1-0.5)^1000=0,62

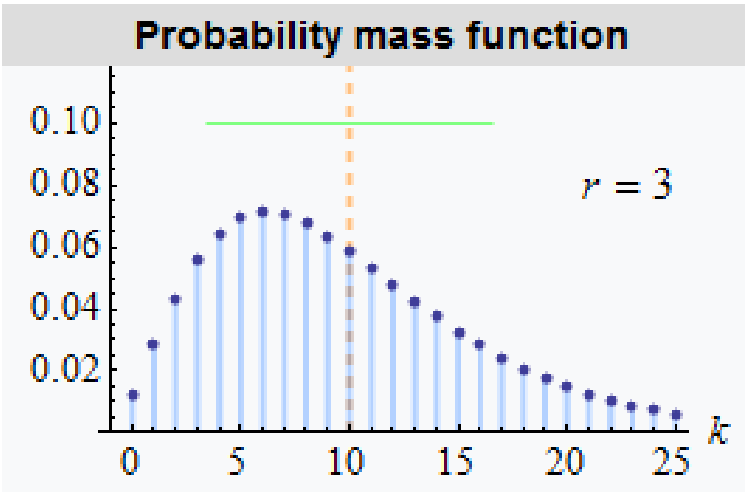
Sea cual sea el estadístico que gastamos tenemos que tener en cuenta este factor, estamos analizando miles de genes a la vez. Un p-value del 0,05 significa que el 5% de los genes no se expresan realmente diferencialmente (falso positivo).

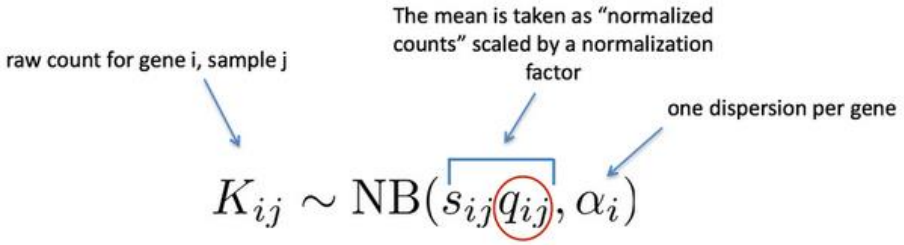
Si pruebas 100 genes tienes 5 falsos positivos; pero si pruebas 20000 genes tendrás 1000.



DESEQ2

La distribución binomial negativa es una distribución de probabilidad discreta que describe el número de ensayos necesarios para obtener un número fijo de éxitos en una serie de ensayos de Bernoulli independientes, cada uno con la misma probabilidad de éxito. Esta distribución es útil en situaciones donde estamos interesados en modelar el número de fracasos antes de alcanzar un número determinado de éxitos.



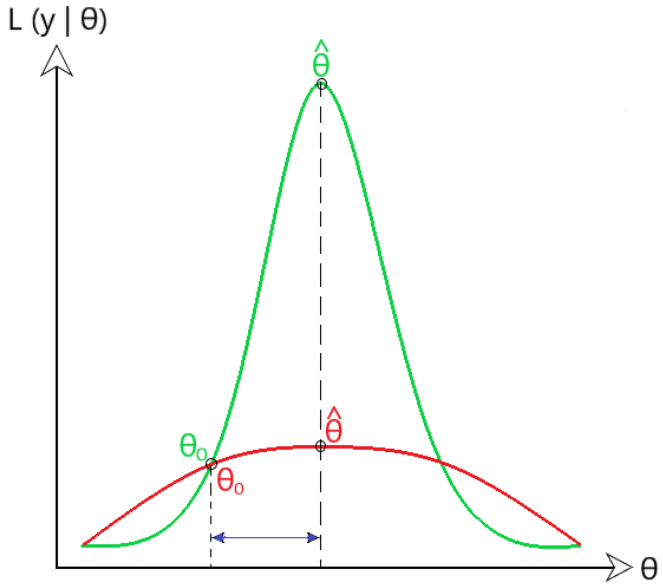


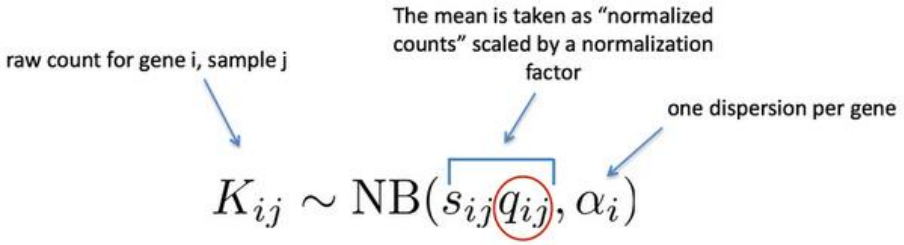
DESEQ2 (Wald test)

La prueba de Wald es una prueba estadística utilizada para evaluar la significancia de los coeficientes en modelos de regresión, como la regresión logística, regresión lineal, y modelos lineales generalizados. La prueba de Wald es particularmente útil para determinar si un parámetro particular de un modelo es significativamente diferente de cero.

Fundamentos de la Prueba de Wald

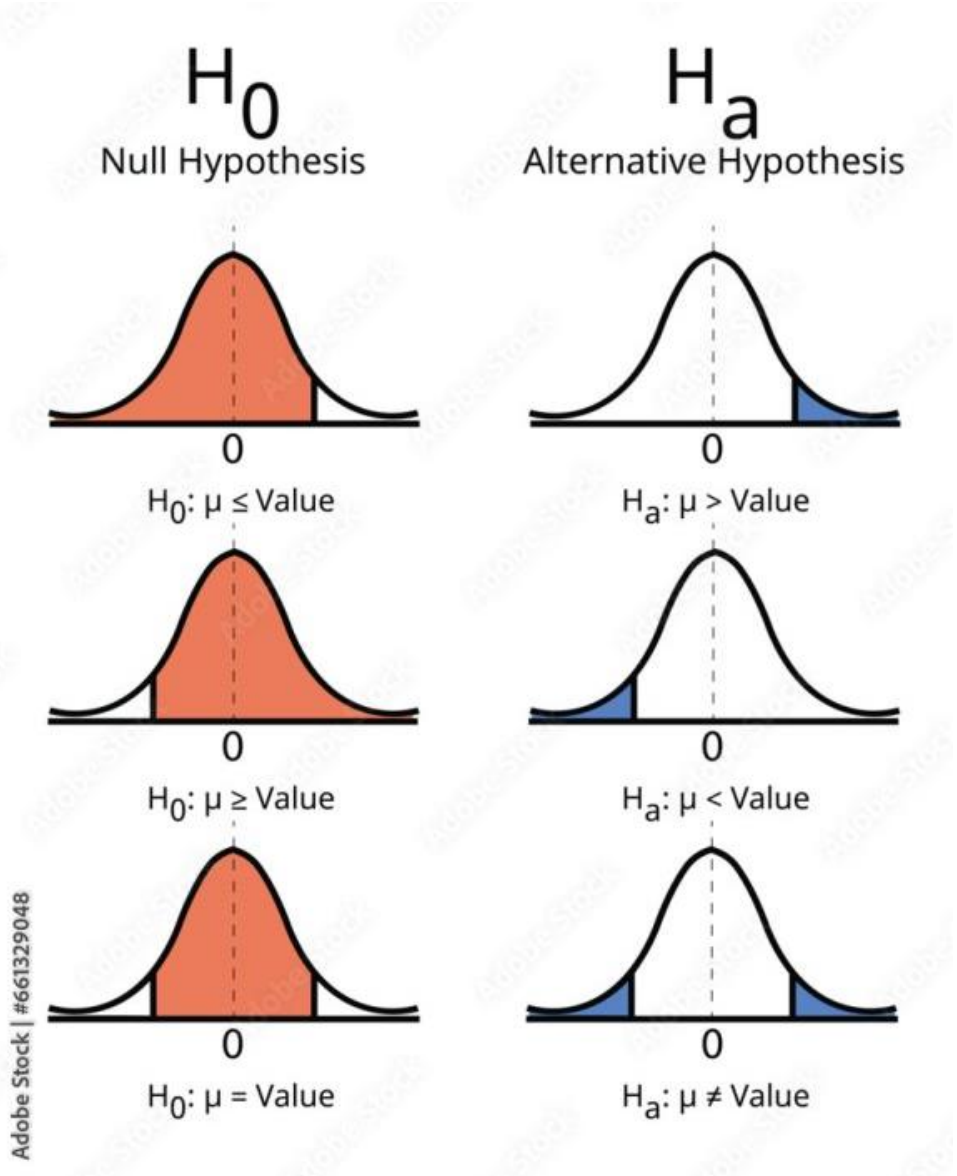
Dada una estimación del parámetro β y su error estándar SE(β), la prueba de Wald calcula un estadístico de prueba W que sigue una distribución normal estándar bajo la hipótesis nula.





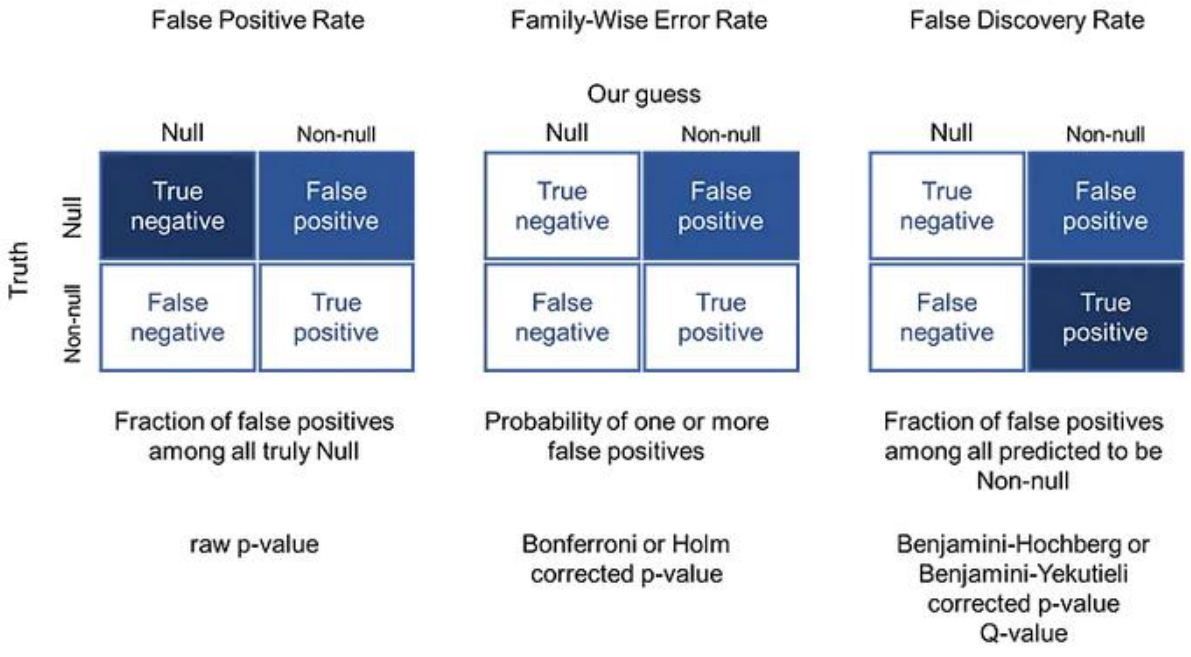
Valor de P

La probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta. En términos simples, el valor p ayuda a diferenciar resultados que son producto del azar del muestreo, de resultados que son estadísticamente significativos.



P-value ajustado

En análisis estadísticos, cuando se realizan múltiples pruebas, es importante ajustar los valores p para controlar la tasa de falsos positivos (errores tipo I). Esto es especialmente relevante en contextos como el Over Representation Analysis (ORA), donde se prueban muchos términos funcionales simultáneamente.

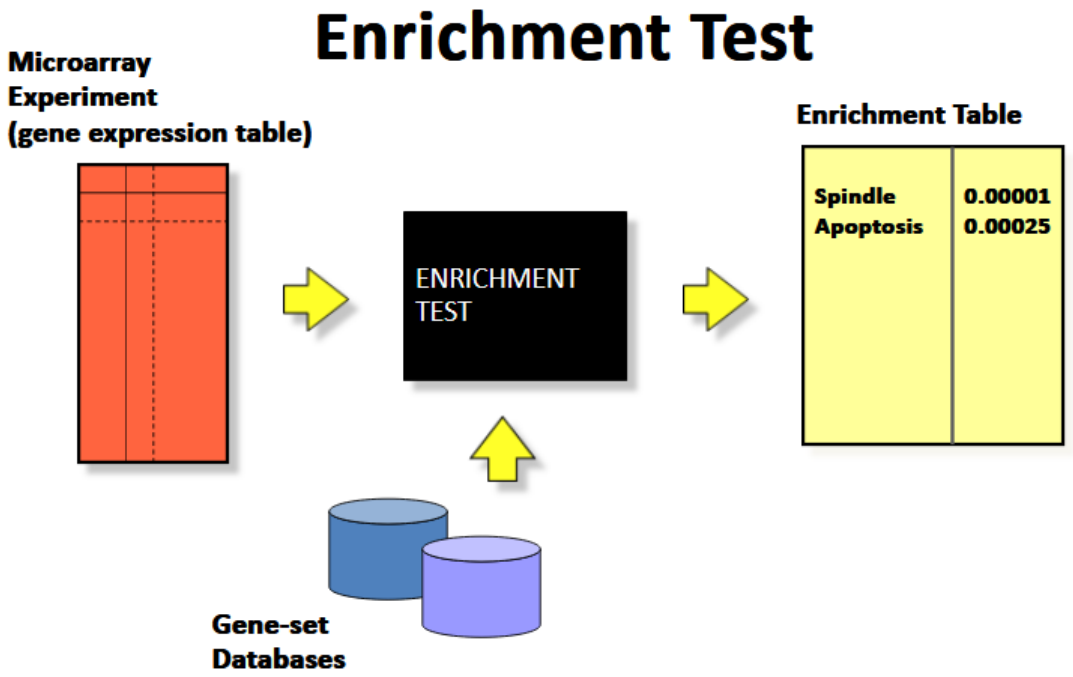


https://towardsdatascience.com/why-and-how-to-adjust-p-values-in-multiple-hypothesis-testing-2ccf174cdbf8

Enriqueciemiento

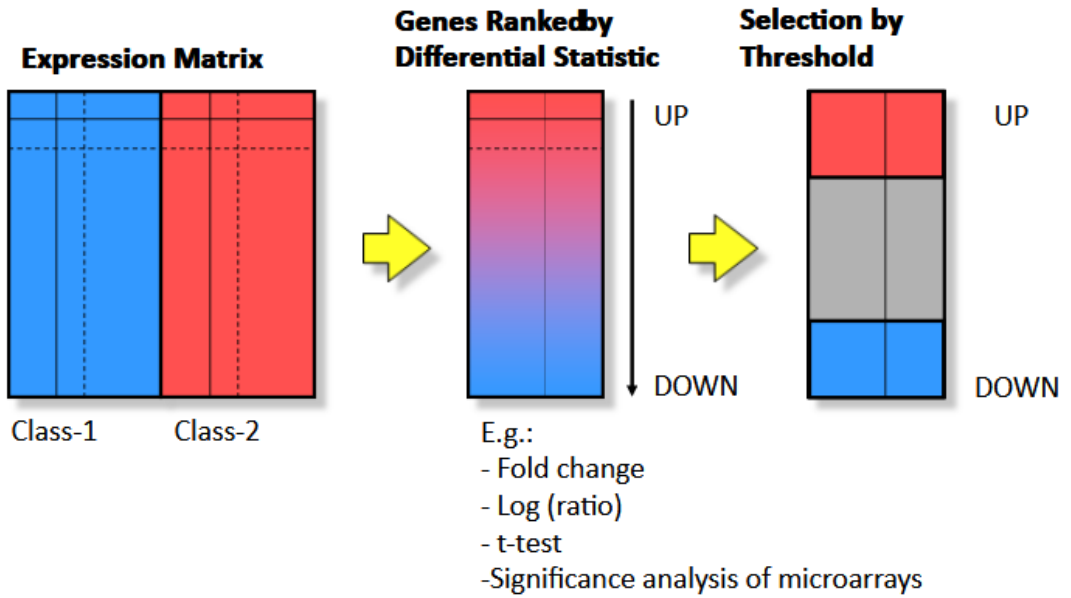
- **Lista de Genes** (Valor de logFC)
  - Los genes se encuentran en las listas de referencia
  - Prueba Hipergeométrica

- **Ranked list** (expresión diferencial)
  - Por valor, se encuentra significativo la lista de referencia en mis datos
  - GSEA



Over Representation Analysis (ORA)

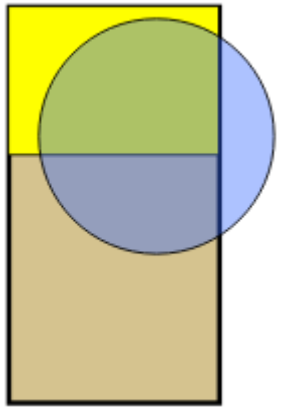
Identificar si ciertos conjuntos de genes o términos funcionales están sobrerrepresentados en un grupo específico de genes en comparación con lo que se esperaría por azar.

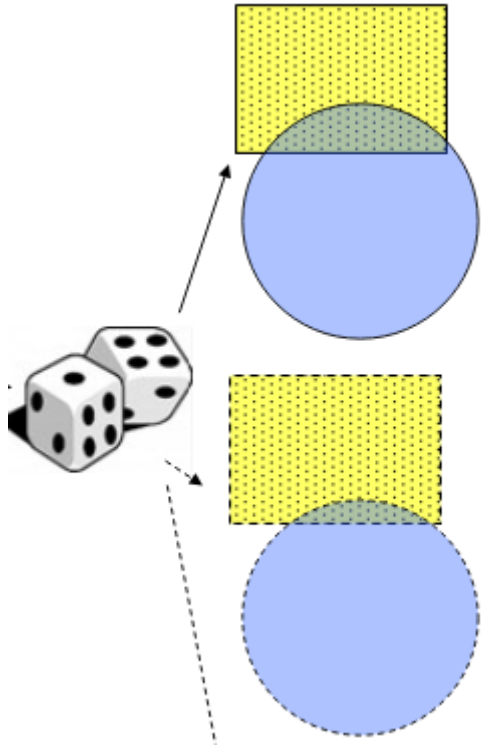


Over Representation Analysis (ORA)

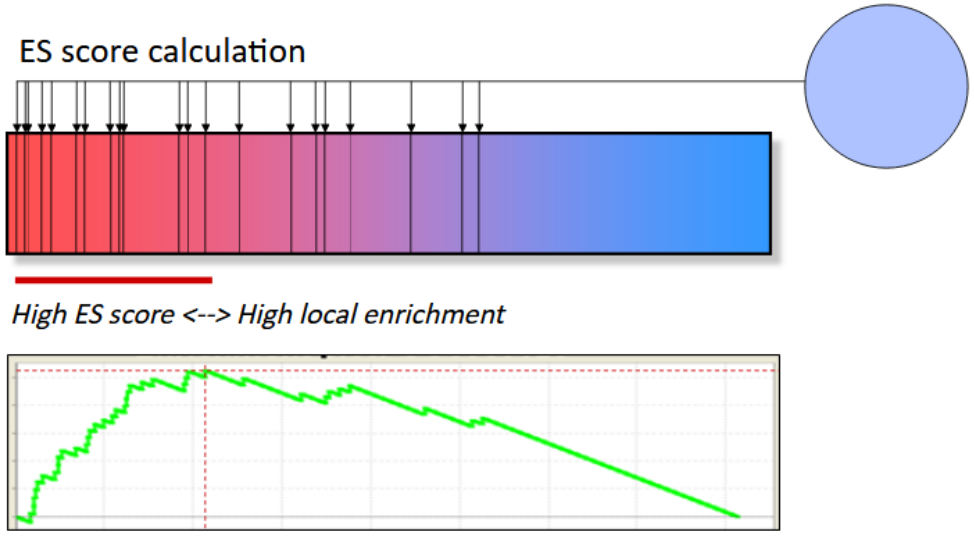
Supongamos que tienes un universo de 1000 genes (N = 1000) y 100 de ellos están asociados con un término funcional específico (K = 100). En tu conjunto de genes de interés hay 50 genes (n = 50) y 10 de ellos están asociados con el término funcional (k = 10).

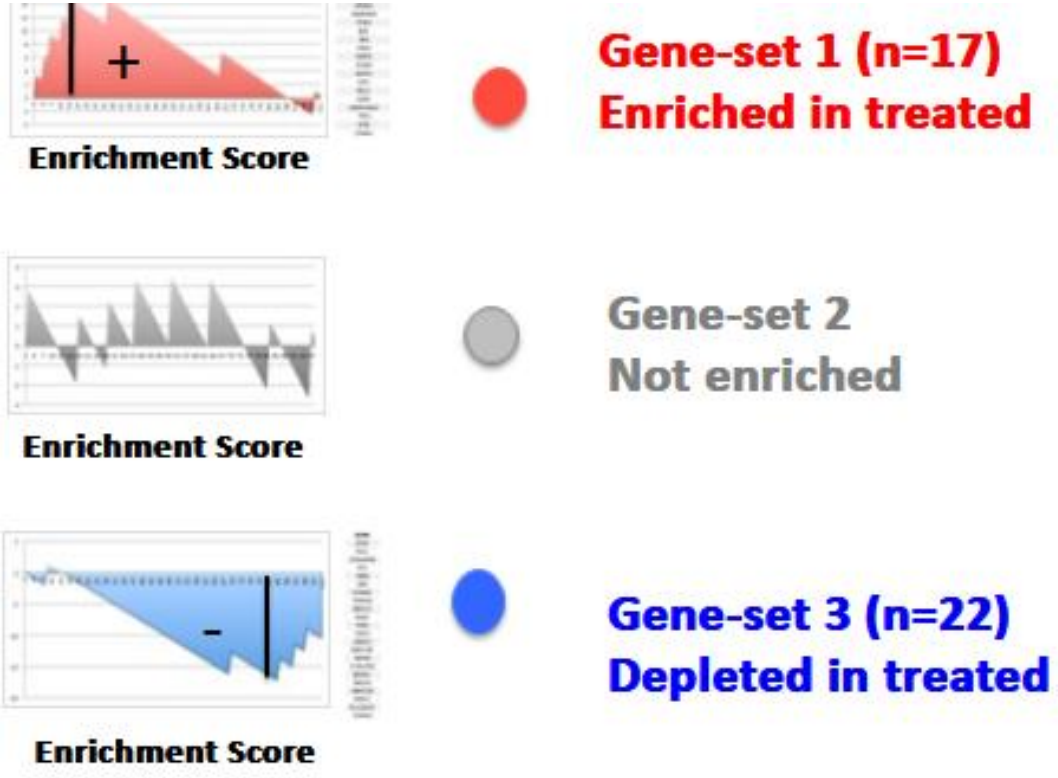
La probabilidad de observar al menos 10 genes asociados con el término funcional en una muestra de 50 genes se puede calcular utilizando la distribución hipergeométrica. Si esta probabilidad (valor p) es baja, puedes concluir que el término funcional está sobrerrepresentado en tu conjunto de genes de interés.





GSEA


ES score calculation
High ES score <--> High local enrichment


+
Enrichment Score
Gene-set 1 (n=17)
Enriched in treated
Enrichment Score
Gene-set 2
Not enriched
-
Enrichment Score
Gene-set 3 (n=22)
Depleted in treated

Enriquecimiento

	Input	Output	Pro	Cons
ORA	A list of gene IDs (no stats needed)	A per-pathway hypergeometric test result	- Simple	- Requires arbitrary thresholds and ignores any statistics associated with a gene
			- Inexpensive computationally to calculate p-values	- Assumes independence of genes and pathways
GSEA	A list of genes IDs with gene-level summary statistics	A per-pathway enrichment score	- Includes all genes (no arbitrary threshold!)	- Permutations can be expensive
			- Attempts to measure coordination of genes	- Does not account for pathway overlap