Recuperación de Información Multimedia

Evaluación de Efectividad

(Anexo: Ciencia, Papers)

CC5213 – Recuperación de Información Multimedia
Departamento de Ciencias de la Computación
Universidad de Chile
Juan Manuel Barrios – https://juan.cl/mir/ – 2019

Efectividad

- Efectividad ("effectiveness" o eficacia) se refiere a la calidad de la respuesta de un sistema
- Evaluación de la efectividad
  - Cuantificar el grado en que un sistema logra el objetivo deseado
- Dos tipos de sistemas a evaluar:
  - Sistemas de Detección:
    - Dada una consulta, el sistema determina una respuesta (Verdadero/Falso, A/B/C, etc.) para cada objeto del dataset junto con un valor de confianza (score)
  - Sistemas de Recuperación:
    - Dada una consulta, el sistema ordena los objetos del dataset del más al menos relevante

Colecciones de referencia

- Colección de referencia (Corpus o Dataset):
  - Conjunto de documentos usados para probar y evaluar algoritmos
  - Usualmente incluye:
    - Conjunto de datos
    - Conjunto de consultas (query set)
    - Respuestas esperadas para cada consulta (Ground-Truth)
- Medida de efectividad
  - Valor que determina el parecido entre la respuesta entregada por un sistema y la respuesta esperada (definida por el ground-truth)

Evaluación de Sistemas de Detección

Matriz de Confusión

Evaluación de la efectividad

- Respuestas positivas dadas por el sistema: TP+FP
- Cantidad que en realidad son positivas: TP+FN
- Respuestas correctas del sistema: TP+TN
- Respuestas incorrectas del sistema: FP+FN
- Medidas:
  - Precision = TP / (TP + FP)
  - Recall = TP / (TP + FN)
  - Accuracy = (TP + TN) / (TP + TN + FP + FN)
  - F_1 score = Media armónica entre Precision y Recall

    2 / [(1/Precision) + (1/Recall)]
  - MCC = Correlación entre Realidad y Respuesta

    (TP*TN)-(FP*FN) / raiz[(TP+FP)*(FN+TN)*(FP+TN)*(TP+FN)]

Evaluación de la efectividad

- Medidas normalizadas:
  - True Positive Rate o Sensitivity o Recall:
    - **TPR**= TP / (TP + FN)
  - True Negative Rate o Specificity:
    - **TNR**= TN / (FP + TN)
  - False Positive Rate
    - **FPR**= FP / (FP + TN) = 1 - TNR
  - False Negative Rate
    - **FNR**= FN / (TP + FN) = 1 – TPR

Evaluación de la efectividad

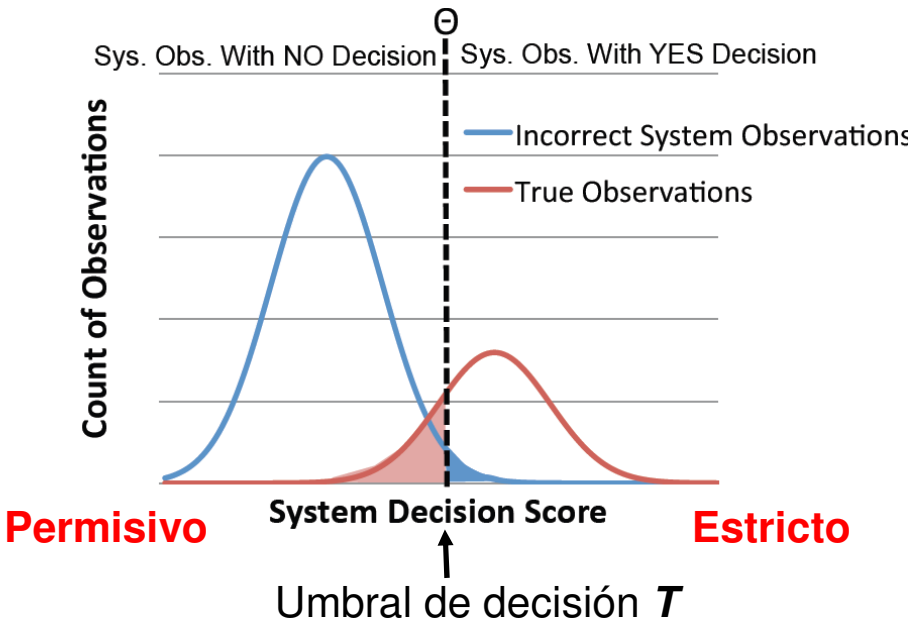
- Otros nombres usados para las medidas normalizadas:
  - CAR (Correct Acceptance Rate) (TPR)
  - CRR (Correct Rejection Rate) (TNR)
  - FAR (False Acceptance Rate) (FPR)
  - FRR (False Rejection Rate) (FNR)

Ajuste de la Efectividad

- Junto con cada respuesta se tiene un valor de confianza o score
- El score se compara con un umbral de decisión ***T***
  - Si score ≥ ***T*** → responder Verdadero
  - Si score < ***T*** → responder Falso
- Si ***T* es bajo**, el sistema es **permisivo**
  - Buen recall, mal precision
- Si ***T* es alto**, el sistema es **estricto**
  - Buen precision, mal recall
- La elección de ***T*** depende del costo asociado a errores FP y FN

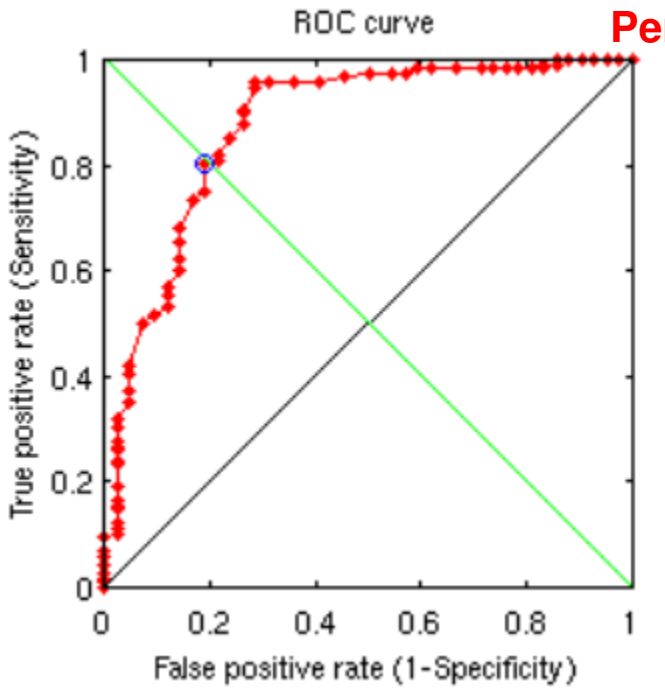
Curvas de Comportamiento

- Variando el umbral ***T*** se puede graficar el comportamiento del detector
  - Se obtienen curvas de comportamiento del sistema donde cada punto es un posible punto de operación

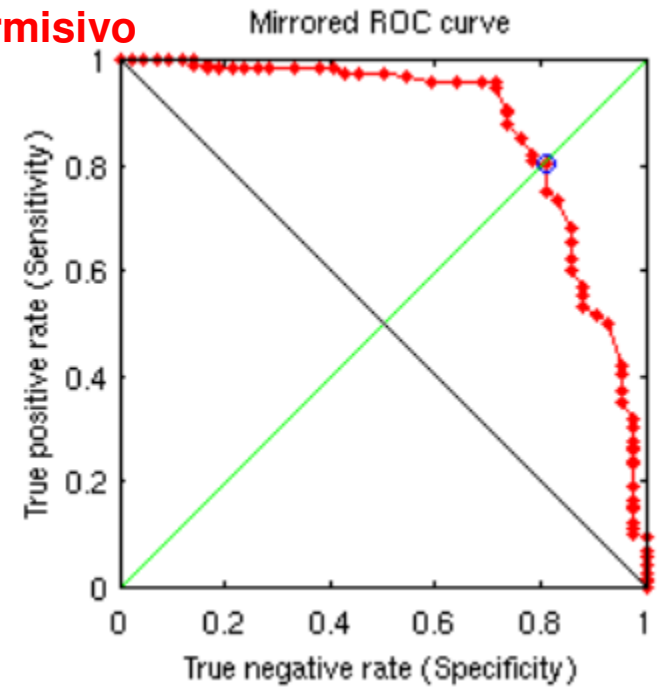


Curvas de Comportamiento

- Curva ROC (Receiver Operating Characteristic):
  - TPR vs FPR o sensitivity vs (1 – specificity)



Permisivo

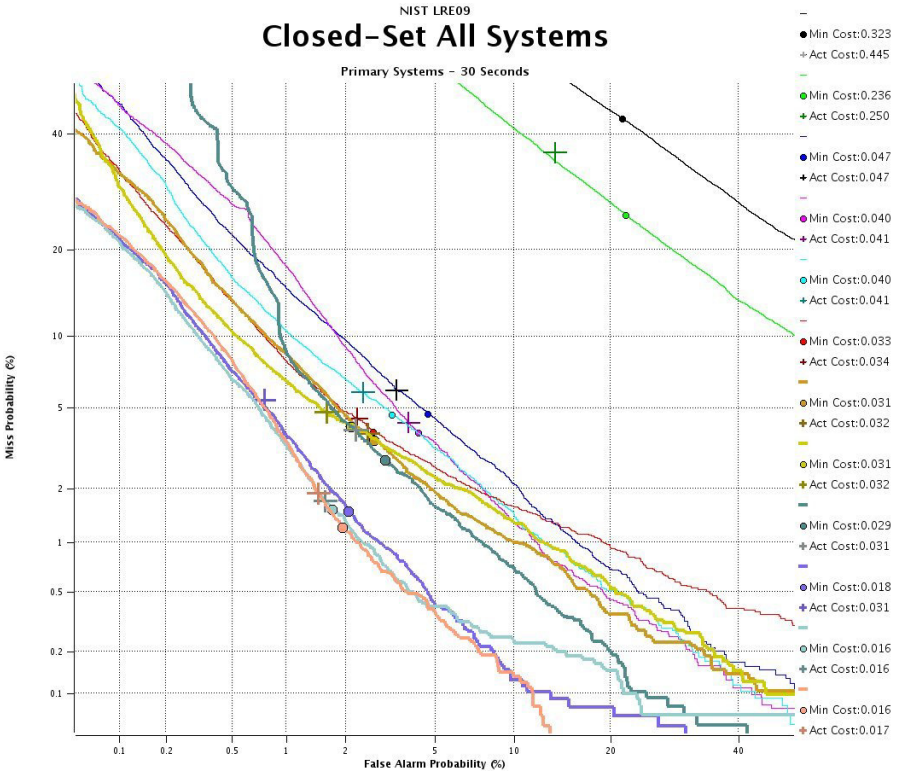


Estricto

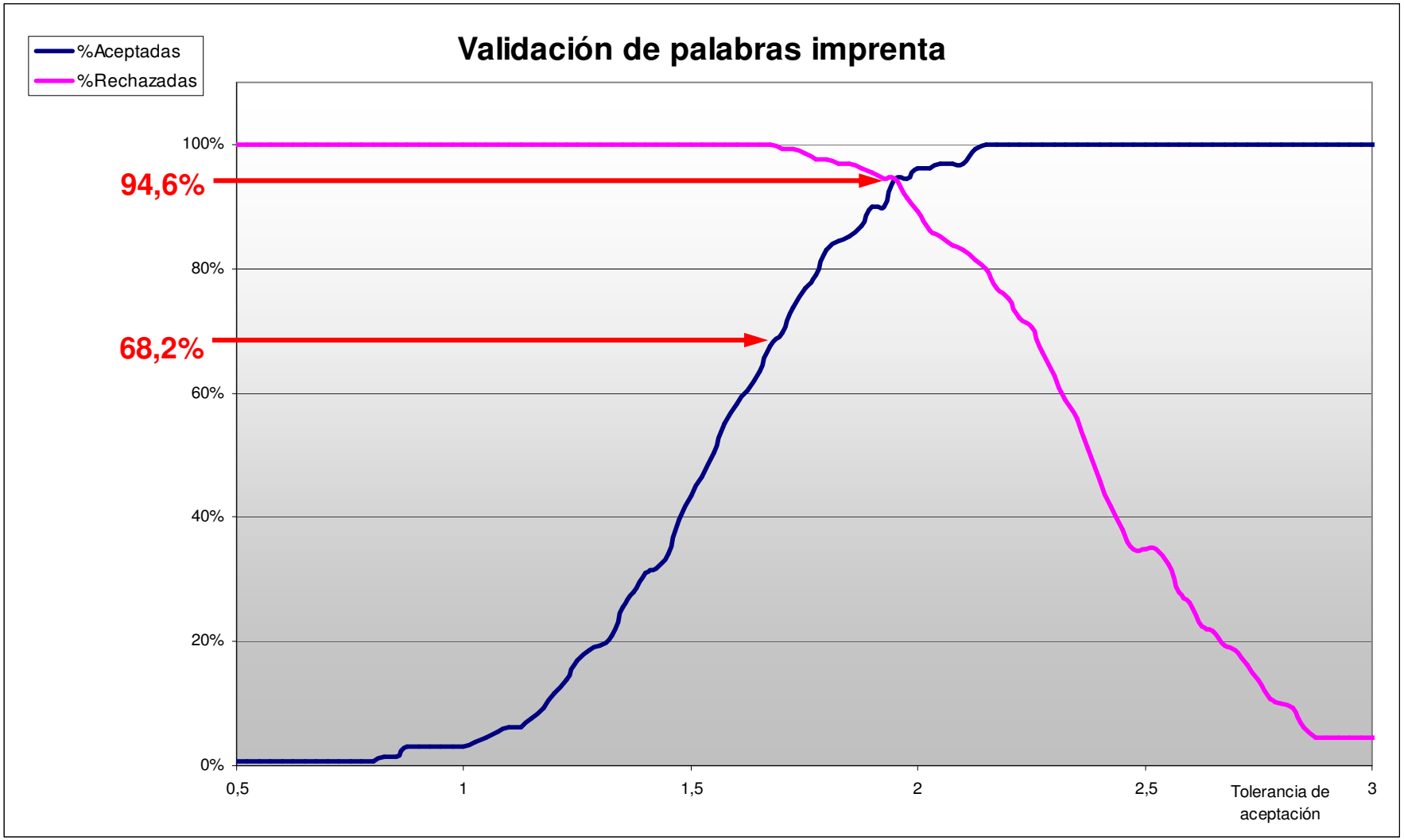
Estricto

Curvas de Comportamiento

- Detection error tradeoff (DET)
  - FNR vs FPR en un gráfico Log-Log



Curvas de Comportamiento



Ejemplo Ajuste de Efectividad

- Se tiene un detector con estos resultados:

	Dijo V	Dijo F	
Es V	900	100	1.000
Es F	2.000	7.000	9.000
	2.900	7.100	10.000

Recall = 90%
Precision = 31%
F_1 score = 0.462
Accuracy = 79%
MCC = 0.448

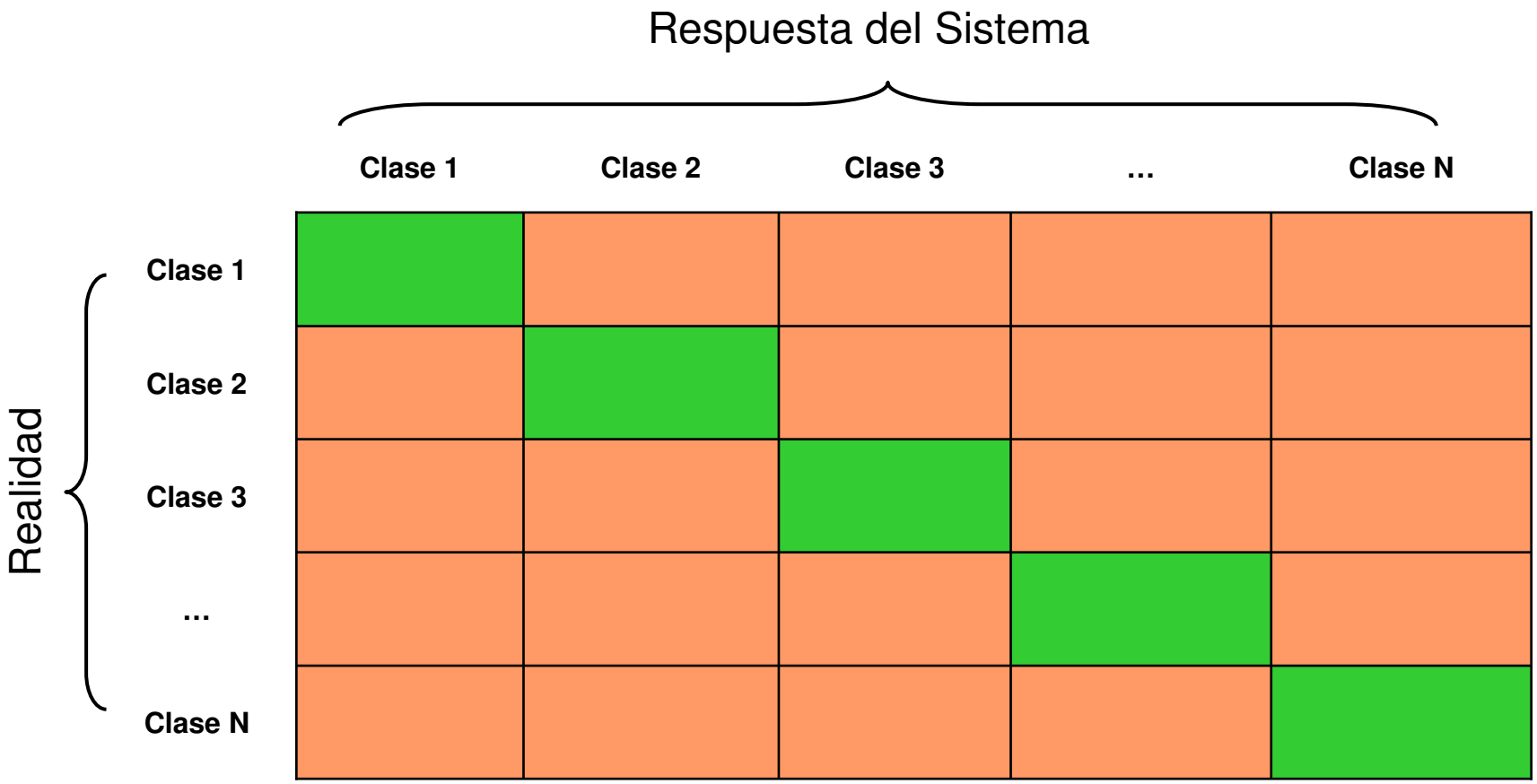
- Con un umbral de desición más estricto:

	Dijo V	Dijo F	
Es V	300	700	1.000
Es F	20	8.980	9.000
	320	9.580	10.000

Recall = 30%
Precision = 94%
F_1 score = 0.455
Accuracy = 93%
MCC = 0.508

Variando el umbral de decisión es posible ajustar FP vs FN

Matriz de Confusión N clases



En la diagonal se muestran los datos correctamente clasificados

Evaluación de Sistemas de Recuperación

Sistema de Recuperación

- Para una consulta de obtiene una lista de documentos ordenados del más al menos relevante (“ranking”)
  - Ejemplos: Buscar documentos en Internet, Buscar fotos parecidas, Buscar apariciones de un objeto

Consulta 

Respuestas correctas (ground-truth)

Im126	Im938	Im837	Im012

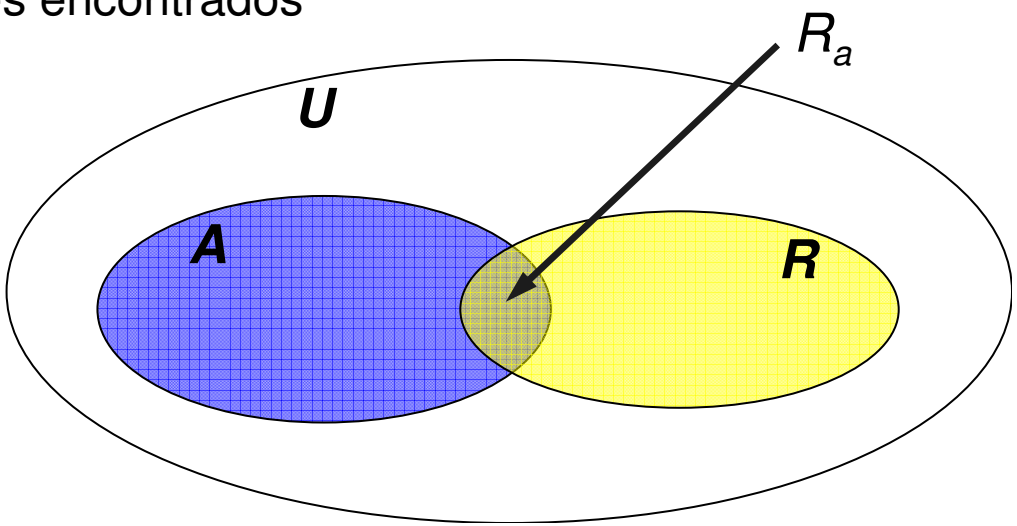
Respuestas obtenidas	Im292	Im126	Im343	Im510	Im012	Im282	Im133	Im938
Rank	1	2	3	4	5	6	7	8

Evaluación de la Efectividad

- **Precision**: La proporción de documentos relevantes con respecto al total de documentos retornados por el sistema
- **Recall**: La proporción de documentos relevantes con respecto al total de documentos que debió haber encontrado el sistema
- Gráficamente:
  - *U* es el total de documentos existentes
  - *A* es el conjunto de documentos encontrados (respuestas)
  - *R* es el conjunto de documentos relevantes (correctos)
  - R_a es el conjunto de relevantes encontrados

$$\text{Precision} = \frac{|R_a|}{|A|}$$

$$\text{Recall} = \frac{|R_a|}{|R|}$$



Evaluación de la Efectividad

- Se desea conocer como cambia Precision en función de Recall
  - Calcular Precision desde la posición 1 hasta la posición de cada respuesta correcta encontrada

Respuestas correctas (ground-truth)	**Im126**	**Im938**	**Im837**	**Im012**

Rank	1	2	3	4	5	6	7	8	∞
Respuesta del Sistema	**Im292**	**Im126**	**Im343**	**Im510**	**Im012**	**Im282**	**Im133**	**Im938**	**Im837**
	✗	✓	✗	✗	✓	✗	✗	✓	not-found
Recall		1/4=0.25			2/4=0.5			3/4=0.75	4/4=1
Precision		1/2=0.5			2/5=0.4			3/8=0.38	4/∞=0

Precisión Interpolada

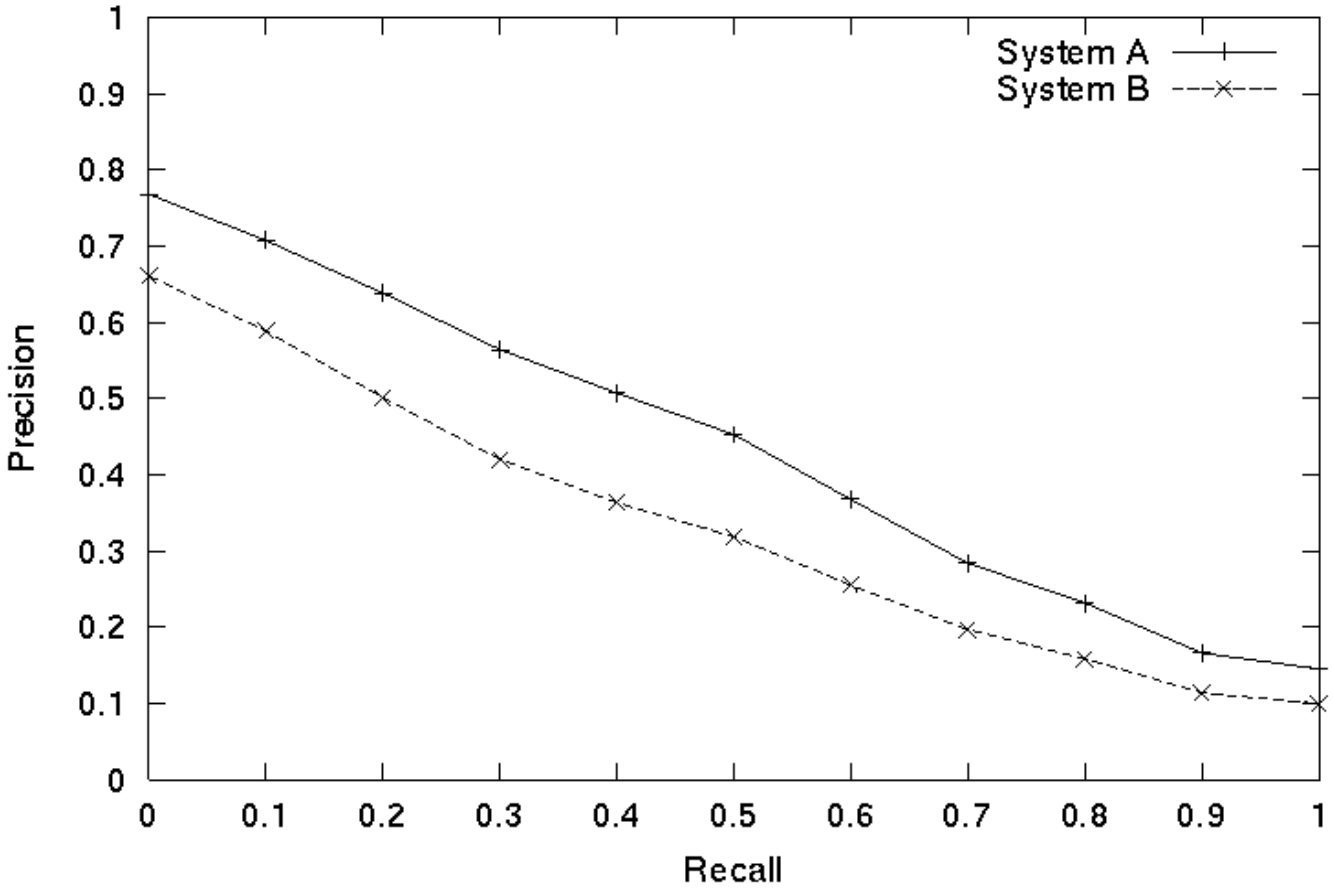
- Calcular cada valor de Recall y su Precision asociado
  - En el ejemplo: P(0.25)=0.5, P(0.5)=0.4, P(0.75)=0.38, P(1)=0
- Precisión Interpolada:
  - Máxima precisión desde cierto punto de recall en adelante:

$$PI(r) = \text{Max}_{j \geq r} \{P(j)\}$$

  - PI es una función decreciente
  - Se calcula para 11 puntos de recall: {0, 0.1, 0.2, …, 0.9, 1}
  - En el ejemplo: PI(0)=PI(0.1)=PI(0.2)=0.5, PI(0.3)=PI(0.4)=PI(0.5)=0.4, etc.
- Se promedian las precisiones interpoladas para todas las consultas en los 11 puntos de Recall y se construye un gráfico Recall vs Precision

Ver Baeza, cap 4.3

Gráfico Recall vs Precision



Sistema A es más efectivo que el Sistema B

Medidas de un solo valor

- **Average Precision (AP)**
  - Para una consulta: promedio de Precision en las posiciones donde se encontró un documento relevante
    - Si un documento relevante no está en la lista de respuestas, entonces tiene Precision 0
    - En el ej.: AP=(0.5 + 0.4 + 0.38 + 0) / 4 = 0.32
  - Notar que NO ES el promedio de precision en todas las posiciones, si no que solo considera las posiciones donde cambia el recall
  - Para un sistema: promedio el AP de todas las consultas. **Mean Average Precision (MAP)**

Ver Baeza, cap 4.3

Medidas de un solo valor

- **Precision at Rank (P@r)**
  - Para una consulta: c(r)/r
    - r es una posición (ej. 1, 5, 20,100)
    - c(r) es el número de documentos relevantes encontrados entre las posiciones 1 y r
  - Para un sistema: promediar P@r de todas las consultas
- **Recall at Rank (R@r)**
  - Para una consulta: c(r)/t
    - t es el número total de documentos relevantes para la consulta
  - Para un sistema: promediar R@r de todas las consultas

Medidas de un solo valor

- **R-precision**
  - Precisión en la posición igual a la cantidad de respuestas correctas, es decir, cuando |A|=|R|.
  - Para un sistema: promediar R-precision de todas las consultas
- **Reciprocal Rank**
  - Precision al obtener la primera respuesta correcta: 1/r
    - r es el menor rank de las respuestas correctas
  - Se puede restringir considerando sólo las primeras S respuestas, i.e.: 1/r si r ≤ S o 0 si no
  - Para un sistema: promediar Reciprocal Rank de todas las consultas. **Mean Reciprocal Rank (MRR)**

Medidas de un solo valor

- **F-measure (F_1 score) y E-measure**
  - Combinar Precision y Recall en un único valor usando la media armónica

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}}$$

$r(j)$: recall j^{th} objeto en el ranking

$$E(j) = 1 - \frac{1+b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}}$$

$P(j)$: precision j^{th} objeto en el ranking

b: parámetro del usuario

Ejercicio

Calcular para cada sistema: MAP, MRR, Recall@5 y R-Precision.

Ground-Truth:

Q_1	Im38	Im09	Im49
Q_2	Im56	Im34	
Q_3	Im36	Im53	

Sistema 1:

	Q_1	Q_2	Q_3
1	Im38	Im12	Im36
2	Im94	Im56	Im30
3	Im09	Im49	Im35
4	Im73	Im55	Im93
5	Im74	Im34	Im53
6	Im48	Im03	Im63

Sistema 2:

	Q_1	Q_2	Q_3
1	Im49	Im56	Im09
2	Im91	Im50	Im36
3	Im35	Im96	Im63
4	Im09	Im33	Im18
5	Im62	Im34	Im60
6	Im38	Im40	Im28

Ejercicio MAP

Sistema 1:

	Q_1	Q_2	Q_3
1	✓	×	✓
2	×	✓	×
3	✓	×	×
4	×	×	×
5	×	✓	✓
6	×	×	×

$AP_1=(1+2/3+0)/3=0.56$

$AP_2=(1/2+2/5)/2=0.45$

$AP_3=(1+2/5)/2=0.70$

MAP=0.57

Sistema 2:

	Q_1	Q_2	Q_3
1	✓	✓	×
2	×	×	✓
3	×	×	×
4	✓	×	×
5	×	✓	×
6	✓	×	×

$AP_1=(1+2/4+3/6)/3=0.67$

$AP_2=(1+2/5)/2=0.70$

$AP_3=(1/2+0)/2=0.25$

MAP=0.54

AP=Promediar precisión al encontrar cada relevante

MAP=Promediar AP de cada query

¿Es la diferencia de MAP (0.57 vs 0.54) suficientemente grande para concluir que un sistema es mejor que el otro?

Ejercicio MRR

Sistema 1:

	Q_1	Q_2	Q_3
1	✓	×	✓
2		✓	
3			
4			
5			
6			

MRR = (1+1/2+1)/3

MRR = **0.83**

Sistema 2:

	Q_1	Q_2	Q_3
1	✓	✓	×
2			✓
3			
4			
5			
6			

MRR = (1+1+1/2)/3

MRR = **0.83**

MRR = Promediar precisión de la posición del primer relevante encontrado

Ejercicio Recall@5

Sistema 1:

	Q_1	Q_2	Q_3
1	✓	×	✓
2	×	✓	×
3	✓	×	×
4	×	×	×
5	×	✓	✓
6			

Recall@5 = (2/3+2/2+2/2)/3

Recall@5 = **0.89**

Sistema 2:

	Q_1	Q_2	Q_3
1	✓	✓	×
2	×	×	✓
3	×	×	×
4	✓	×	×
5	×	✓	×
6			

Recall@5 = (2/3+2/2+1/2)/3

Recall@5 = **0.72**

Recall@k = Promediar recall de la posición k

Ejercicio R-Precision

Sistema 1:

	Q_1	Q_2	Q_3
1	✓	×	✓
2	×	✓	×
3	✓		
4			
5			
6			

R-Precision = (2/3+1/2+1/2)/3

R-Precision = **0.56**

Sistema 2:

	Q_1	Q_2	Q_3
1	✓	✓	×
2	×	×	✓
3	×		
4			
5			
6			

R-Precision = (1/3+1/2+1/2)/3

R-Precision = **0.44**

R-Precision = Promediar precisión de la posición igual a la cantidad de relevantes de la query

Resumen

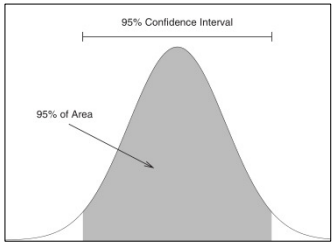
- MAP considera la posición de todas las respuestas correctas de cada consulta
  - Si MAP=1 las *n* correctas están siempre en las primeras *n* posiciones
- P@*r* y R@*r* evalúan solo las primeras *r* posiciones
  - Si AverageP@5=1 significa que las primeras 5 respuestas contienen siempre respuestas correctas
  - R@*r* no funciona bien con *r* pequeños. R-Precision es como R@*r* donde *r* se ajusta a las cantidad de correctas de cada consulta
- MRR considera la posición de una única respuesta correcta
  - Si MRR=1 la primera correcta está siempre en la primera posición
- F1 combina dos valores usando media armónica (en vez de promediar)
  - Evita que un valor muy bueno esconda uno malo

Significancia

"¿Es la diferencia de MAP (0.57 vs 0.54) suficientemente grande para concluir que un sistema es mejor que el otro?"

- MAP busca predecir el resultado que un sistema logrará en general para cualquier consulta genérica, a través de *n* observaciones (el AP de las *n* consultas del ground-truth)
- Los intervalos de confianza determinan un rango donde se encuentra un valor real, calculado como el promedio de las observaciones
  - Brevemente, el intervalo depende del número de observaciones, su varianza y riesgo α (o P-value) típicamente 5%:

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu_{\text{real}} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$



Correlación de Rankings

Correlación de Rankings

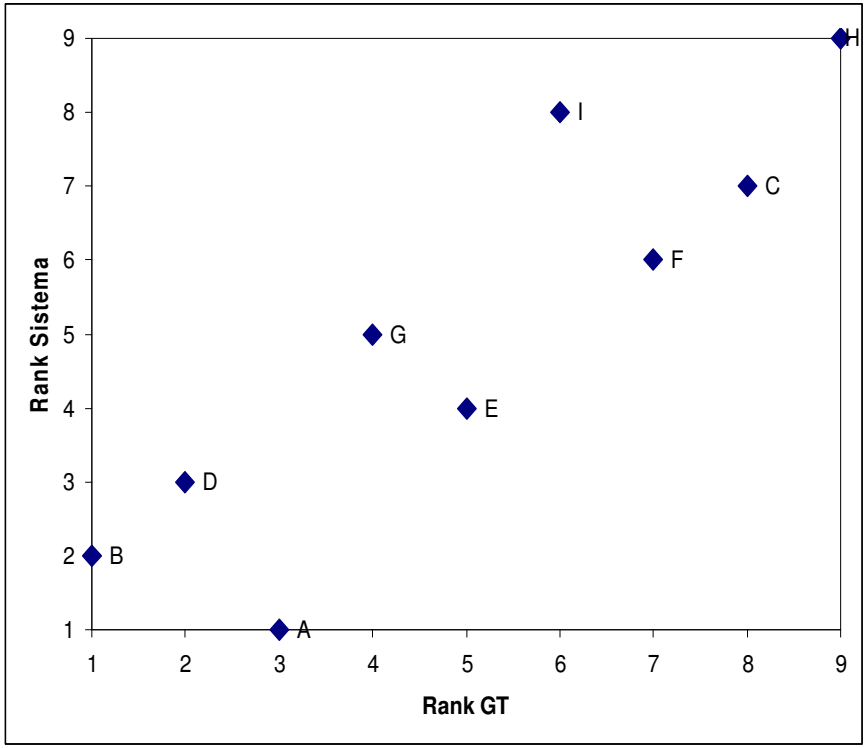
- El Ground-Truth es un ordenamiento ideal de documentos
- Se desea medir el parecido (correlación) entre la respuesta del sistema y la respuesta ideal
  - Valor entre -1 y 1: 1=idénticos, 0=sin relación, -1=inversos
- Se desea comparar permutaciones, es decir, listas de largo *n* con valores entre 1 y *n*, cada valor aparece una única vez
  - Si faltan documentos, se agregan al final en el mismo orden

Rank	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
GT (ideal)	**B**	**D**	**A**	**G**	**E**	**I**	**F**	**C**	**H**
Sistema	**A**	**B**	**D**	**E**	**G**	**F**	**C**	**I**	**H**

Coeficiente de Spearman

- Para cada documento calcular la diferencia de rank entre ambas listas

	Rank GT	Rank Sistema	Diff	Diff^2
A	3	1	2	4
B	1	2	-1	1
C	8	7	1	1
D	2	3	-1	1
E	5	4	1	1
F	7	6	1	1
G	4	5	-1	1
H	9	9	0	0
I	6	8	-2	4
***S*=Suma Diff^2**				**14**



Coeficiente de Spearman

- Calcular las diferencias de rank entre la respuesta y el ideal y calcular la suma de cuadrados (S):

$$S = \sum \text{Diff}^2$$

- El Coeficiente de Spearman se obtiene al escalar el valor *S* al rango -1 a 1
  - Para listas de largo *n* el valor máximo que puede obtener *S* es:

$$M = n(n^2 - 1)/3$$

- Coeficiente Spearman $= 1 - \frac{2S}{M} = 1 - \frac{6 \sum \text{Diff}^2}{n(n^2 - 1)}$

- En el ejemplo:
  - Spearman(GT, Sistema) = 0.883

Coeficiente de Kendall Tau

- Comparar dos listas por medio de comparar la posición relativa entre todos los pares de documentos
- Para cada par de documentos (d_i,d_j):
  - Si en ambas listas d_i esta antes que d_j, o en ambas listas d_i esta después que d_j entonces ambas listas son concordantes para el par (d_i,d_j)
  - En cambio si en una lista d_i esta antes que d_j pero en la otra lista d_i esta después que d_j las listas son discordantes para ese par
  - Es decir, el par (d_i,d_j) es concordante si:

$$\text{rank}_A(d_i) > \text{rank}_A(d_j) \text{ y } \text{rank}_B(d_i) > \text{rank}_B(d_j)$$
$$\text{o } \text{rank}_A(d_i) < \text{rank}_A(d_j) \text{ y } \text{rank}_B(d_i) < \text{rank}_B(d_j)$$

Coeficiente de Kendall Tau

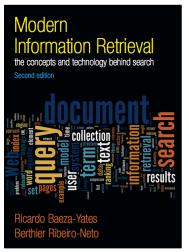
- Para todos los pares de documentos contar cuantos son concordantes y cuantos son discordantes entre ambas listas
- Para *n* documentos, total pares $M_n = n(n-1)/2$
- Coeficiente de Kendall Tau es:

τ = (#pares_concordantes – #pares_discordantes) / M_n

```
tau = 0
for (i = 2 to n)
  for (j = 1 to i-1)
    tau += signo(rank_A[i]-rank_A[j]) * signo(rank_B[i]-rank_B[j])
tau /= n * (n-1) / 2
```

Bibliografía

- **Modern Information Retrieval**. Baeza-Yates, Ribeiro-Neto, 2011.
  - Cap. 4



Anexo:
Ciencia, Papers, Test de Hipótesis

Publicaciones Científicas

- Tipos de publicaciones:
  - Paper, Short Paper, Poster, Survey, Position paper, Capítulo de Libro, Technical Report, …
- Publishers (Springer, Elsevier, IEEE, ACM, …)
- Conferencias vs Journals (ISI, Scopus, …)
- Peer Reviewed vs Non Peer Reviewed
- Lugares para buscar papers:
  - https://scholar.google.com
  - Web of Knowledge
  - arXiv.org

Papers

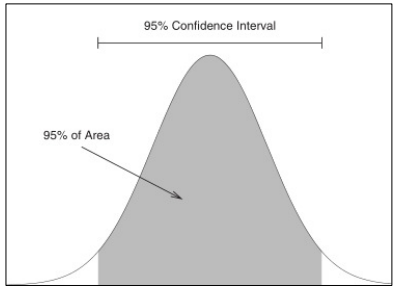
- Cada paper tiene una o más contribuciones
- Estructura típica de un paper:
  - Título, Autores (con su filiación), Abstract, 1.Introducción, 2.Estado del Arte, 3.Propuesta (idea novedosa), 4.Experimentos/Evaluaciones, 5.Conclusiones
- Idealmente los resultados deben ser replicables (aunque no siempre es posible)
  - Detalles de experimentos, Datasets usados, etc.
- Existen métricas de papers, autores y conferencias
  - Citas, Impact Factor, h-index, …
  - Publish or Perish (PoP)
  - Webometrics
    - http://www.webometrics.info/en/node/92
    - http://www.webometrics.info/en/Latin_America/Chile

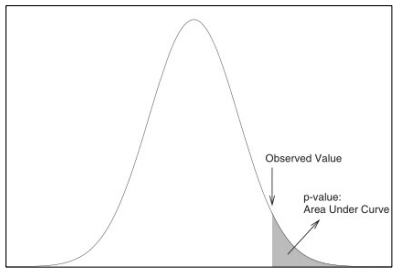
Experimentos y Test de Hipótesis

- En ciencia se desea saber si un experimento tiene efecto
- Se definen dos conjuntos, en uno se aplica el experimento y en el otro no (control group):
  - Hipótesis nula: no hay efecto (no hay diferencia entre ambos conjuntos)
  - Hipótesis alternativa: si hay efecto (hay diferencia notoria)
- Si se encuentra una diferencia significativa entre ambos conjuntos, tenemos evidencia para “rechazar la hipótesis nula”
  - Conclusión: no es verdad que no hay efecto
- Si no se encuentra diferencia entre los conjuntos ocurre un “fallo en rechazar la hipótesis nula”
- Este enfoque es muy usado en medicina, biología (no muy usado en ciencias de la computación)

Test de Hipótesis

- Se obtienen *n* observaciones de un experimento y se calcula su promedio μ y varianza σ
- Se asume una distribución normal (o t-student si n<30) se calcula un intervalo donde con probabilidad (1-α) se encuentra el promedio real (μ_{real})
  - α=5% o P-value se refiere al riesgo o a la probabilidad de rechazar la hipótesis nula cuando en realidad era verdadera





$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu_{\text{real}} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Problemas del Test de Hipótesis

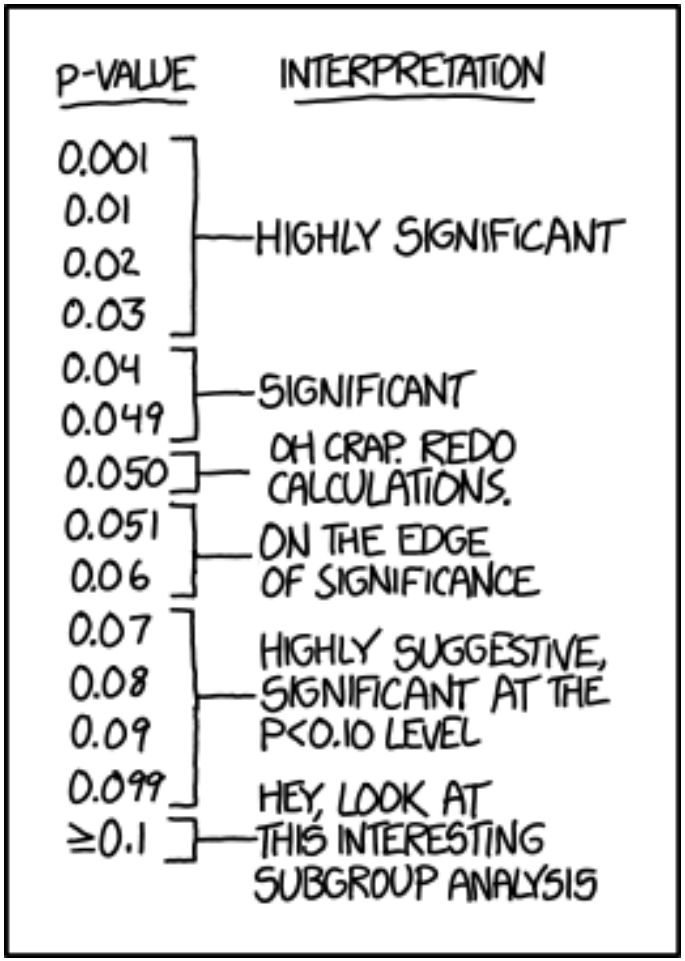
- Usualmente se usa P-value con un valor arbitrario del 5%
  - Ver https://en.wikipedia.org/wiki/Data_dredging
- La ocurrencia de un efecto (significancia estadística) se transforma artificialmente en una decisión booleana
  - No se señala la magnitud del efecto
  - Cualquier efecto (marginal o no) puede ser estadísticamente significativo
    - Basta hacer experimentos con un *n* muy grande
  - “Con una confianza de un 95% se concluye que comer X aumenta la probabilidad de cáncer”
    - Se concluye que “aumenta”, independiente de si aumenta en un 1% o en un 500%

Observación vs Experimento

- Estudio Observacional
  - Esconden causa-efecto, confounding variables, spurious correlations
    - https://www.tylervigen.com/spurious-correlations
  - No es buena idea elaborar test de hipótesis con datos históricos
  - Si se revisan los datos y se hace un test de hipótesis adhoc es muy posible encontrar patrones particulares existentes únicamente en esos datos
    - Se pueden usar distintos datasets, evaluando en un conjunto de test desconocido, pero no es recomendable
- Experimento Controlado
  - Para evaluar si existe un efecto se debe diseñar y realizar un experimento adecuado (cuidando variables confounding, sesgos, etc.)

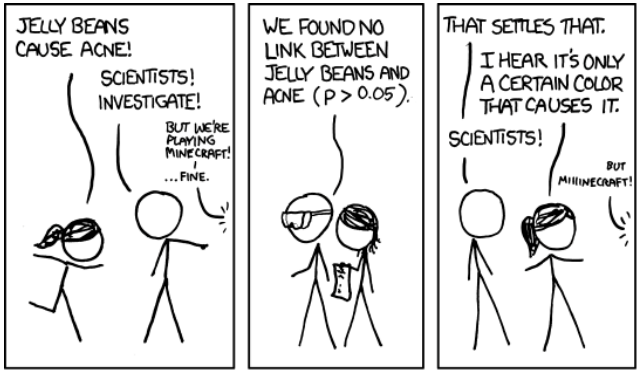
P-Value

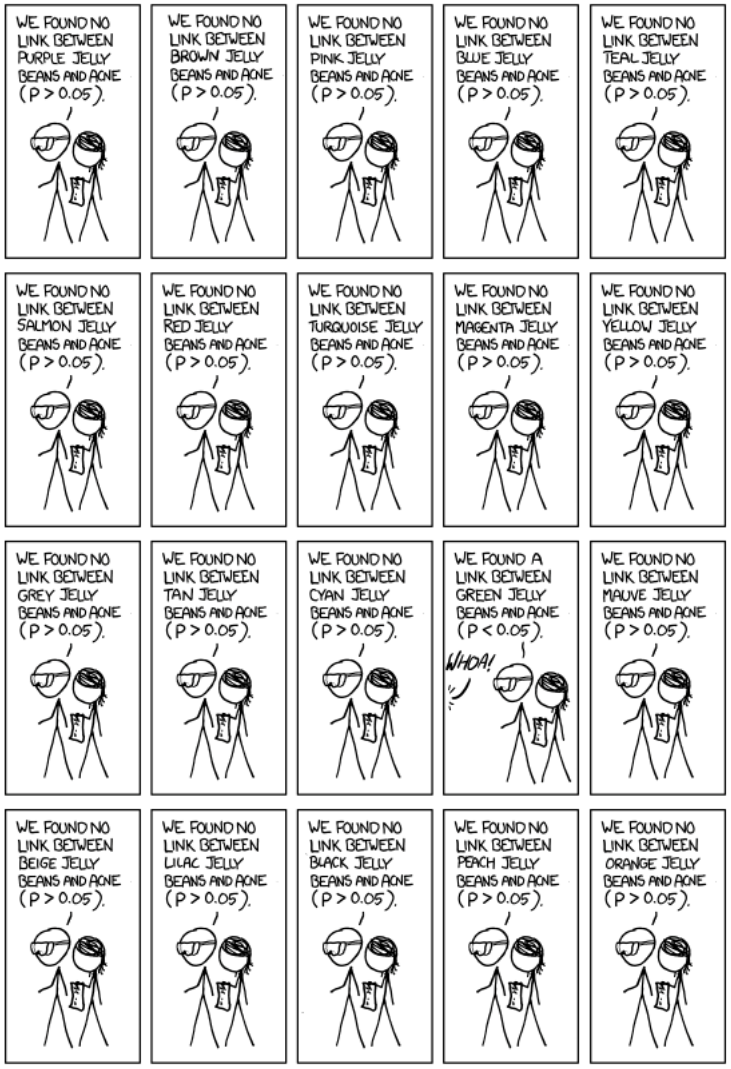
Fuente:
https://xkcd.com/1478/

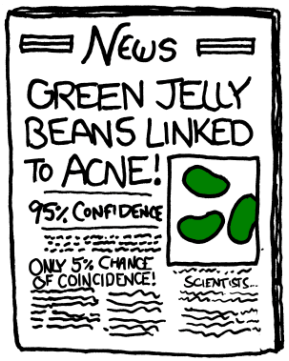


Ver discusión en: https://www.explainxkcd.com/wiki/index.php/1478

P-Hacking







Fuente: https://xkcd.com/882/

Ver discusión en:

https://www.explainxkcd.com/wiki/index.php/882

https://en.wikipedia.org/wiki/Data_dredging