

## SCORING

L'obiettivo è trovare un modo per ricavare lo score da attribuire ai documenti per trovare i top K documenti.

Bisogna trovare una score function in grado di misurare quanto bene un documento fa "match" con la query.

## JACCARD COEFFICIENT

È una misura come di quanto due insiemi A e B si sovrappongono:

$$J(A, B) = \frac{A \cap B}{A \cup B} \in [\phi, 1]$$

Possiamo usare questo concetto per vedere quanti termini della query sono nel documento.

### Esempio

Query: ides of march

Doc 1: caesar died in march

Doc 2: the long march

$$J(Q, D_1) = \frac{1}{6} = 0,167 \quad J(Q, D_2) = \frac{1}{5} = 0,2$$

Jaccard ci dice che fra i due documenti è il secondo quello più simile alla query. Ma questo è evidentemente falso, dato che semanticamente il Doc 2 fa riferimento a tutt'altro.

### Problemi Jaccard

- 1) Non tiene conto delle ripetizioni di un termine
- 2) Non distingue termini veri ma importanti dagli altri
- 3) Non considera la lunghezza di un documento

Come tener conto della frequenza?

Mettiamo le parole nella matrice di incidenza Term - Document:

|           | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth |
|-----------|----------------------|---------------|-------------|--------|---------|---------|
| Antony    | 157                  | 73            | 0           | 0      | 0       | 1       |
| Brutus    | 4                    | 157           | 0           | 1      | 0       | 0       |
| Caesar    | 232                  | 227           | 0           | 22     | 21      | 1       |
| Calpurnia | 0                    | 10            | 0           | 0      | 0       | 0       |
| Cleopatra | 57                   | 0             | 0           | 0      | 0       | 0       |
| mercy     | 2                    | 0             | 3           | 57     | 5       | 1       |
| worser    | 2                    | 0             | 1           | 1      | 12      | 0       |

Dove gli elementi della matrice indicano la frequenza di quel termine in quel documento.

## BAG OF WORDS MODEL

Fin ora abbiamo rappresentato le relazioni fra termini e documenti come un vettore, facendo una forte supposizione: non ci interessa la posizione delle parole nel documento. Però consideriamo l'esempio:

- 1) John è più veloce di Mary
- 2) Mary è più veloce di John

Questi due documenti hanno lo stesso vettore, anche se indicano uno l'opposto dell'altro.

Definizione term-frequency

$tf_{t,d}$  è il numero di occorrenze di  $t$  in  $d$ .

## WHY ASSUMPTION

Il peso di un termine è proporzionale alla sua frequenza grezza (assoluta). Questa affermazione fa riferimento al fatto che spesso gli scrittori ripetono una certa parola mentre discutono o elaborano un argomento. Le parole significative sono in grado di distinguere oggetti rilevanti.

## WEIGHTING TERM FREQUENCY

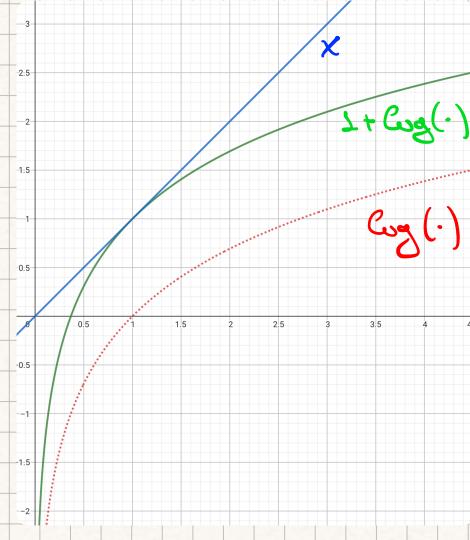
Ma la **frequenza grezza** dei termini non è proprio quello che ci serve, anche perché un documento che contiene 100 volte un termine è più rilevante di uno che lo contiene solo una volta ma non 100 volte più rilevante.

La rilevanza non cresce linearmente, perciò si usa il logarithm della frequenza:

$$w_{t,d} = \begin{cases} 1 + \log(t_{f,t,d}) & \text{per } t_{f,t,d} > 0 \\ 0 & \text{altrimenti.} \end{cases}$$

Da cui possiamo ricavare la score function come:

$$s(q,d) = \sum_{t \in q \cap d} (1 + \log(t_{f,t,d}))$$



## TERMINI RARI

I termini rari sono più significativi di termini ad alta frequenza, basti pensare alle stopword. Per esempio il termine "elettronico" è sicuramente contenuto in un documento rilevante per la query "elettronico".

L'obiettivo è dare un peso maggiore ai termini rari.

### COLLECTION FREQUENCY $F_t$

$F_t$  del termine  $t$  è il numero di occorrenze di  $t$  nella collezione.

### DOCUMENT FREQUENCY $dft$

$dft$  del termine  $t$  è il numero di documenti che contengono  $t$ .

## INVERSO DOCUMENT FREQUENCY

Del termine  $t$ :

$$idf_t = \log \frac{N}{dft} \in [0, 1]$$

dove  $N$  è il numero di documenti della collezione.

Questa matrice è quella che tutti i search engine usano.

**Exhaustivity of a document:** Ovvero quanto quel documento copre il topic principale

**Specificity of a term:** Quanto bene un termine descrive il topic principale di un documento.

## TFIDF

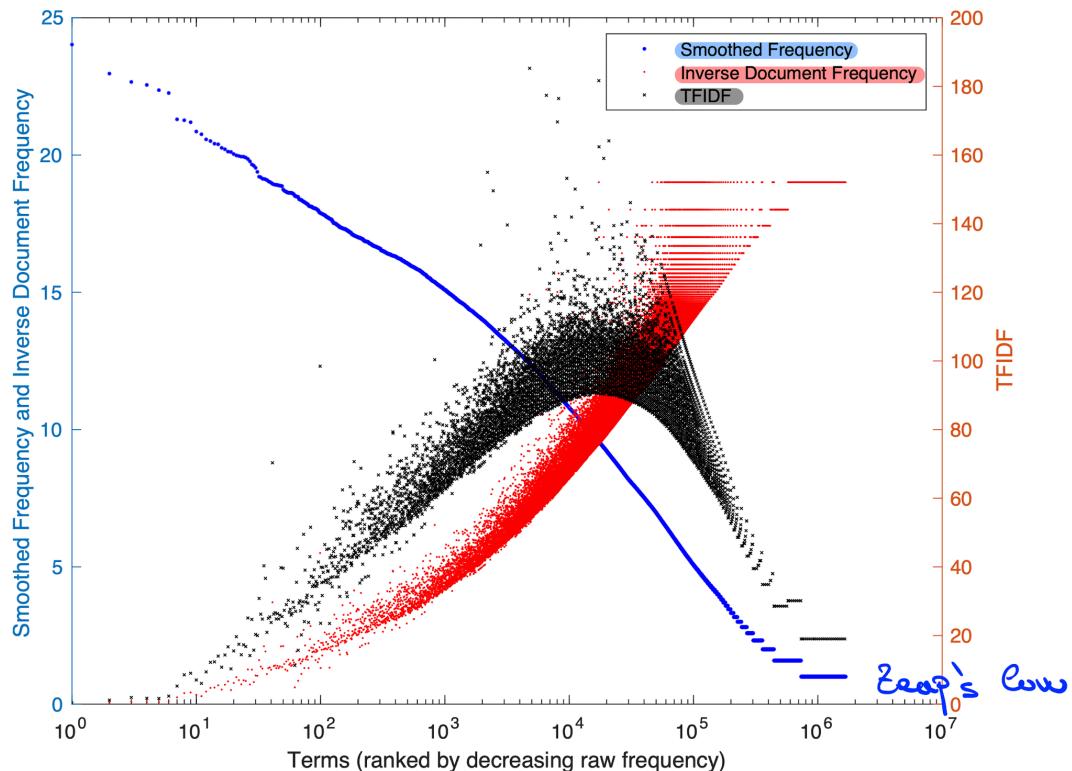
Mettiamo insieme la weighting term frequency e la inverse document frequency:

$$w_{t,d} = \begin{cases} [1 + \log(t_{ft,d})] \log\left(\frac{N}{df_t}\right) & \text{per } t_{ft,d} > 0 \\ \emptyset & \text{altrimenti.} \end{cases}$$

La score function:

$$s(q, d) = \sum_{t \in q \cap d} w_{t,d}$$

## TF, IDF and TFIDF Example



## DOCUMENTI, QUERY e VETTORI

Abbiamo uno spazio vettoriale grande  $|V|$  (la dimensione del vocabolario).

- 1) I documenti possono essere visti come vettori in questo spazio vettoriale
- 2) Come anche le query.

Sono vettori con grandezze elevate, e soprattutto sono davvero molto sparsi (pieni di elementi nulli).

Dato che lavoriamo con vettori, un'idea è usare la distanza euclidea per valutare la similitudine. **Pessima idea**, perché con high-dimensional vector tale distanza è comunque elevata.

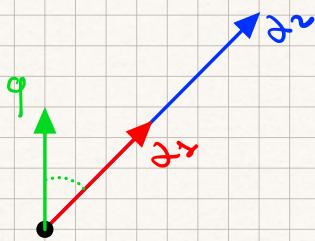
### Esempio

Se creiamo un documento in append con se stesso, avremo due documenti la cui rappresentazione vettoriale è il l'uno il doppio dell'altro:



però semanticamente sono uguali.

Possiamo considerare l'angolo fra i due vettori, quello è  $\phi$ , e rimane uguale anche se confrontato col una query:



## COSINE SIMILARITY

Metrica che compare gli angoli fra due vettori  $\vec{q}$  e  $\vec{d}$ :

$$s(\vec{q}, \vec{d}) = \frac{\langle \vec{q}, \vec{d} \rangle}{\|\vec{q}\| \|\vec{d}\|}$$

Che si può anche scrivere come:

$$s(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{|V|} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

Assumiamo che  $\|\vec{d}\| = \|\vec{q}\| = 1$  così che lo score delle cosine diventa:

$$s(q, d) = \sum_{i=1}^{|V|} q_i \cdot d_i$$

Utilizzando la matrice TF e IDF abbiamo che

$$s(q, d) = \sum_{t \in q \cap d} \text{TF}(t, d) \cdot \text{IDF}(t, d)$$

Dove entrambi  $\text{TF}(\cdot)$  e  $\text{IDF}(\cdot)$  sono vettori così fatti:

$$\text{IDF} = \begin{bmatrix} \text{IDF}(t_1) \\ \vdots \\ \text{IDF}(t_{|V|}) \end{bmatrix}$$

$$\text{TF} = \begin{bmatrix} \text{TF}(t_1, d) \\ \vdots \\ \text{TF}(t_{|V|}, d) \end{bmatrix}$$

Quindi possiamo scrivere

$$s(q, d) = \sum_{i=1}^{|V|} TF(t_i, d) \cdot IDF(i)$$

Considerando che sia la query che lo score sono vettori di dimensione  $|V|$ , possiamo scrivere che:

$$\begin{aligned} q^T s(d) &= q \cdot s(d) = \sum_{i=1}^{|V|} q_i \cdot s_i(d) = \\ &= \sum_{q_i \neq 0} s_i(d) = \sum_{t \in q \cap d} TFIDF(t, d). \end{aligned}$$

## EVALUATION

Come possiamo capire se un utente è felice? Possiamo vedere su cosa clicca, ma bisogna fare attenzione a titoli fuorviatori.

In ogni caso per capire se un search engine lavora bene, bisogna capire quanto è **rilevante** il risultato dato.

Per capire se un risultato è rilevante si usano tre elementi di benchmark:

- 1) **Benchmark document collection**
- 2) **Benchmark suite of query**
- 3) **Valutazione (Assessment)** di rilevanza e non rilevanza per ciascuna query con ciascun documento

Quindi ogni documento di benchmark deve avere un **giudizio di rilevanza** (relevance judgement) rispetto a ogni query di benchmark. Questo giudizio può essere:

- 1) **Binario** (Si o No)
- 2) **Graduale** (su una scala: 0, 1, 2, 3, 4, 5)

## QUERY DI TEST

Abbiamo comunque bisogno di query di test, esse devono essere:

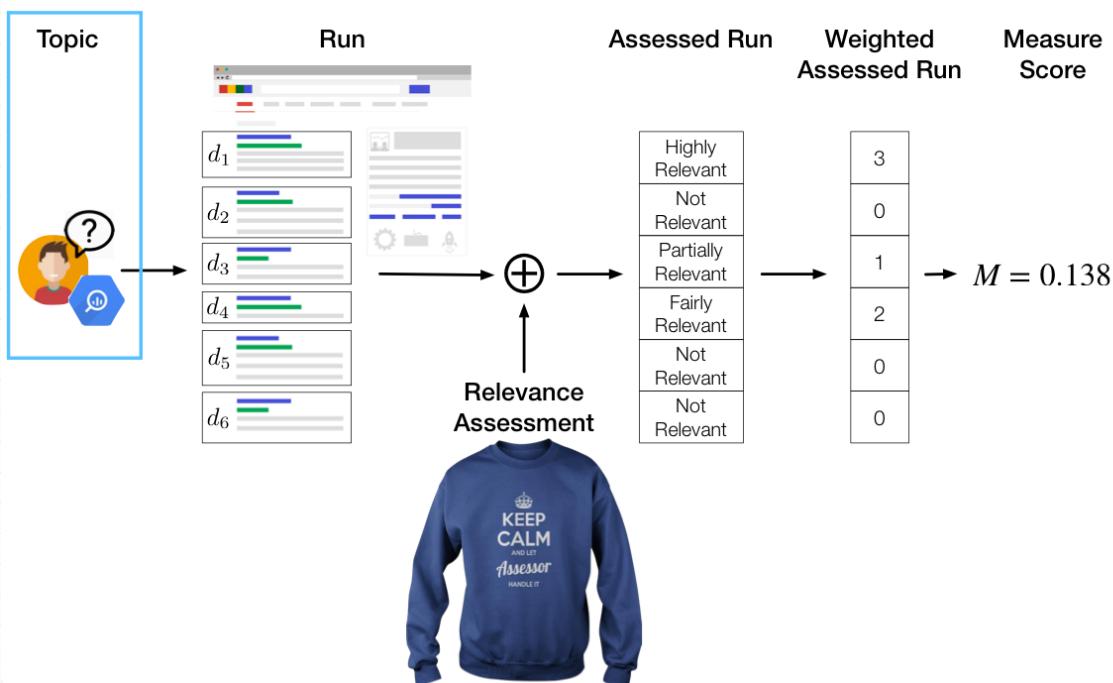
- 1) Iveranti ai documenti disponibili
- 2) Rappresentative dei bisogni degli utenti, non possono essere generate randomicamente, poiché gli utenti non fanno così.

3) Prese da un query log se disponibile.

## COLLEZIONI Sperimentali

Usare collezioni sperimentali, su determinati topic e con un relevant judgment ci permette di ripetere l'esperimento e di renderlo comparabile.

## VALUTARE CON COLLEZIONI DI TEST



Il valore  $M$  va confrontato con l' $M$  trovato per topic diversi.

Ma cosa vuol dire topic? Questo consiste in:

- 1) Titolo
- 2) Descrizione
- 3) Narrative

} Si possono trovare in un doc XML dentro il tag <topic>

## Esempio XML

```
<?xml version="1.0" encoding="UTF-8"?>
<topic>
    <identifier>41</identifier>

    <title lang="en">Pesticides in Baby Food</title>
    <title lang="fr">Des pesticides dans la nourriture pour bébés</title>
    <title lang="it">Pesticidi negli alimenti per bambini</title>
    <title lang="ru">Пестициды в детском питании</title>
    <title lang="zh">嬰兒食品中含有殺蟲劑</title>
    <title lang="ja">ベビーフード中の病害防除剤</title>
    <title lang="th">ยา รักษากัน ที่ เก็บของ กับ ยาพาร์เมลัง ใน อាណา เด็ก</title>
    <title lang="so">Sunta cayayaanka ee Cuntada Ilmaha</title>
    <title lang="sw">Dawa za kuulia wadudu katika Chakula cha Mtoto</title>

    <description lang="en">Find reports on pesticides in baby food.</description>
    <description lang="fr">
        Rechercher des documents sur les pesticides dans la nourriture pour bébés.
    </description>
    <description lang="it">
        Trova documenti che parlano dei pesticidi negli alimenti per bambini.
    </description>
    <description lang="ru">Найти статьи о пестицидах в детском питании</description>
    <description lang="zh">查詢有關嬰兒食品中含有殺蟲劑的報導。</description>
    <description lang="ja">ベビーフード中の病害防除剤に関する記事を探したい。</description>
    <description lang="th">หา รายงาน ที่ เก็บของ กับ ยาพาร์เมลัง ใน อាណา เด็ก</description>
    <description lang="so">Hel wargelinada sunta cayayaanka ee cuntada ilmaha.</description>
    <description lang="sw">
        Pata ripoti kuhusu dawa za kuulia wadudu katika chakula cha mtoto.
    </description>

    <narrative lang="en">
        Relevant documents give information on the discovery of pesticides in baby food.
        They report on different brands, supermarkets, and companies selling baby food
        which contains pesticides. They also discuss measures against the contamination
        of baby food by pesticides.
    </narrative>
    <narrative lang="fr">
        Les documents pertinents informent sur la découverte de pesticides dans la
        nourriture pour bébés. Ils contiennent des informations sur les différentes
        marques, les supermarchés et les firmes ayant mis en vente de la nourriture pour
        bébés renfermant des pesticides. Ils relatent également les mesures prises contre
        la contamination de la nourriture pour bébés par les pesticides.
    </narrative>
    <narrative lang="it">
        I documenti rilevanti forniscono informazioni sulla scoperta di pesticidi nei
        cibi per bambini. Riportano i diversi marchi, i supermercati e le ditte che hanno
        venduto alimenti per bambini con i pesticidi. Sono anche rilevanti i documenti
        che discutono le misure contro la contaminazione degli alimenti per bambini con
        i pesticidi.
    </narrative>
</topic>
```

Tipicamente le collezioni di test usano 50 topic diversi.

## Esempio di Relevance Assessment

Topic ID Fixed Document No Judgement

|     |   |               |   |
|-----|---|---------------|---|
| 41  | 0 | LA050394-0237 | 0 |
| 41  | 0 | LA112394-0177 | 0 |
| 41  | 0 | LA091294-0164 | 1 |
| 41  | 0 | LA040594-0187 | 0 |
| 41  | 0 | LA041694-0248 | 1 |
| ... |   |               |   |
| 42  | 0 | LA031694-0234 | 0 |
| 42  | 0 | LA040494-0111 | 0 |
| 42  | 0 | LA081794-0171 | 1 |

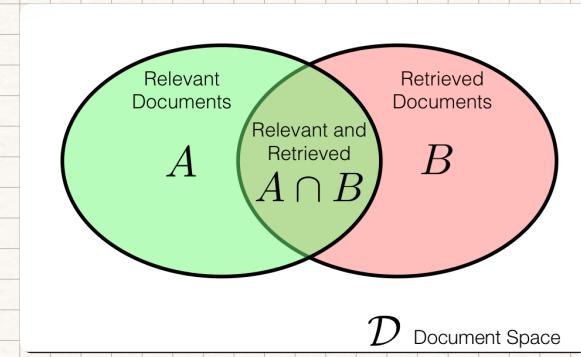
È un file di tipo tsv, quindi è un formato testuale, dove, tipicamente per ogni topic ci sono dai 300 ai 700 documenti giudicati. Questo numero può variare fra topic diversi.

## POOLING

Dato che è impossibile giudicare ogni documento di una collezione, si usa la tecnica di **pooling**. Questa tecnica è usata in TREC (Text Retrieval Conference). La tecnica consiste nel prendere i top-K risultati per diversi motori di ricerca, e unirli in un unico pool. Eliminare poi i documenti duplicati.

I documenti trovati vanno poi presentati contemporaneamente ai relevance assessors.

# Precision & Recall



Precision:

$$P = \frac{A \cap B}{B}$$

Recall

$$R = \frac{A \cap B}{A}$$

La precisione ci indica quanti documenti, fra quelli che il sistema restituisce sono davvero rilevanti.

La recall invece ci dà informazione sul numero di documenti rilevanti che il sistema ritorna rispetto a tutti i documenti rilevanti per una data query nella collezione.

Le due misure insieme ci danno l'idea di quanto un sistema IR sia efficace nel ritornare documenti rilevanti.

F-MEASURE

$$F = \frac{2(P \cdot R)}{P + R}$$

PROBLEMA

Usando queste metriche (Precision, recall, F-score) non possiamo ricavare una classifica dei top-k risultati, non sappiamo se un documento che viene messo al primo posto sia effettivamente il più rilevante, e noi vogliamo valutare anche questo aspetto dei motori di ricerca, ovvero se la classifica stilata è buona o no.

## RANK-BASED PRECISION & RECALL

Precision nel mettere documenti rilevanti nella top-K:

$$P@K = \frac{1}{K} \sum_{i=1}^K \gamma_i$$

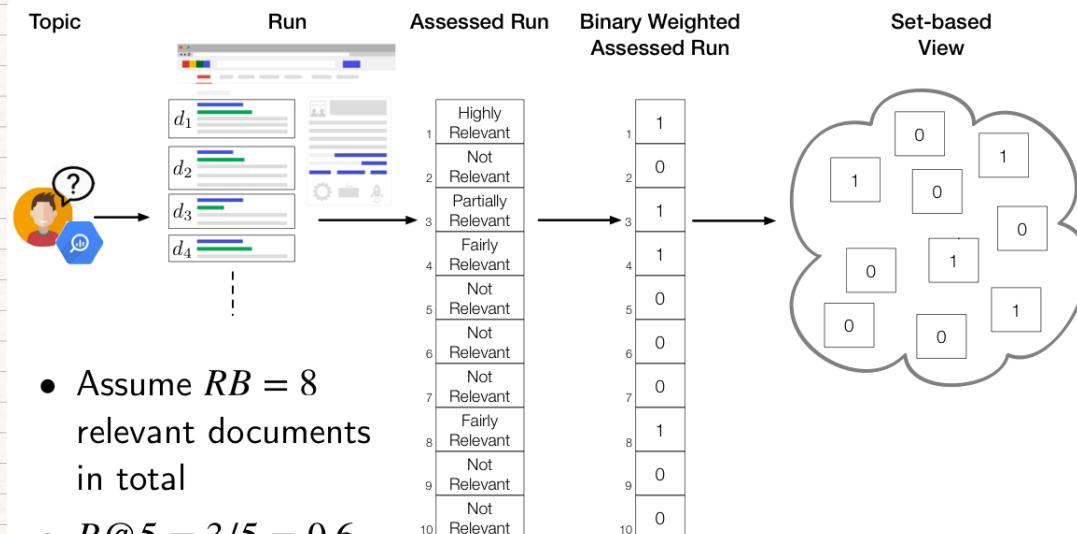
Dove con  $\gamma_i$  si intende il relevant judgment del documento  $i$ -esimo nei top-K

Recall dei top-K documenti risultanti

$$R@K = \frac{1}{RB} \sum_{i=1}^K \gamma_i$$

Dove RB sta per recall-base, ovvero il numero totale di documenti rilevanti nella collezione, dato la query.

Esempio:

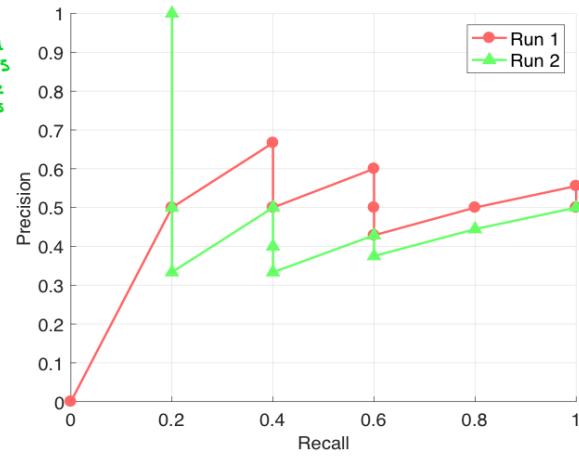


- Assume  $RB = 8$  relevant documents in total
- $P@5 = 3/5 = 0.6$
- $R@5 = 3/8 = 0.375$

RB=5

## Precision-Recall Curve

|    | Run1                   | Run2                   |
|----|------------------------|------------------------|
| 1  | 0 P = 0.00<br>R = 0.00 | 1 P = 1.00<br>R = 0.20 |
| 2  | 1 P = 0.50<br>R = 0.20 | 0 P = 0.50<br>R = 0.20 |
| 3  | 1 P = 0.66<br>R = 0.40 | 0 P = 0.33<br>R = 0.20 |
| 4  | 0 P = 0.50<br>R = 0.40 | 1 P = 0.50<br>R = 0.40 |
| 5  | 1 P = 0.60<br>R = 0.60 | 0 P = 0.40<br>R = 0.40 |
| 6  | 0 P = 0.50<br>R = 0.60 | 0 P = 0.33<br>R = 0.40 |
| 7  | 0 P = 0.42<br>R = 0.60 | 1 P = 0.42<br>R = 0.60 |
| 8  | 1 P = 0.50<br>R = 0.80 | 0 P = 0.37<br>R = 0.60 |
| 9  | 1 P = 0.55<br>R = 1.00 | 1 P = 0.44<br>R = 0.80 |
| 10 | 0 P = 0.50<br>R = 1.00 | 1 P = 0.50<br>R = 1.00 |

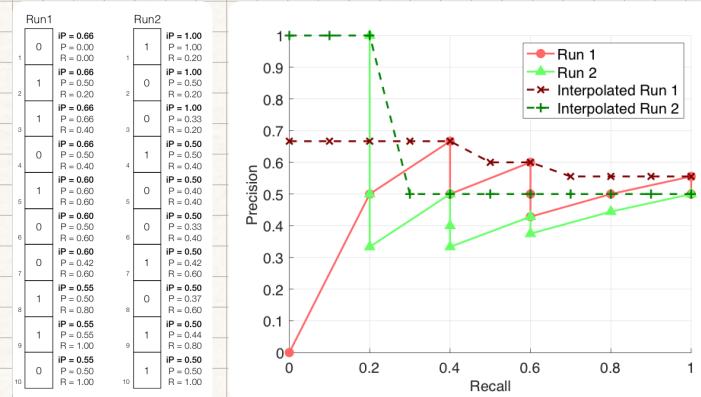


Questa curva è tipicamente a "sego", il problema è la difficoltà nel confrontare run diversi, questo perché possono avere diversi risultati in termini di RECALL. Inoltre, si dovrebbero confrontare le aree sotto quelle curve... difficile.

## Interpolated Precision-Recall Curve

Per interpolare la Precision con un valore standard di Recall  $R_f$ , usiamo il valore massimo di Precision ottenuto per ogni valore attuale di Recall  $R \geq R_f$

$$IP@R_f = \max_{R \geq R_f} P@R$$



## AVERAGE PRECISION

Consideriamo la posizione di ogni documento ( $K_1, K_2, \dots, K_{RB}$ ) e calcoliamo  $P@K$   $\forall K \in (K_1, \dots, K_{RB})$ . Per average precision non c'è altro che la media aritmetica di tutti questi  $P@K$ .

### Mean Average Precision

Ora, troviamo le varie AP per diversi topic e ne facciamo una nuova media MAP. Ovvero la media delle medie.

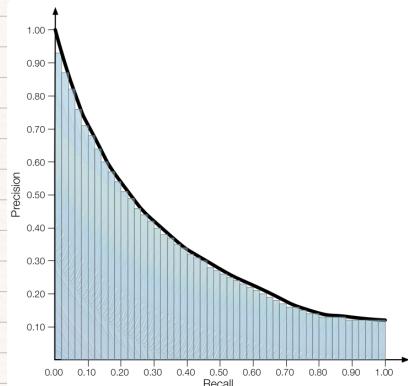
**Pro:** Un solo numero per valutare un intero sistema.

### Esempio:

| Topic | Run | Assessed Run | Binary Weighted Assessed Run       |
|-------|-----|--------------|------------------------------------|
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              |                                    |
|       |     |              | <img alt="Binary weighted assessed |

## AREA SOTTO LA CURVA P-R

Usiamo l'area sotto la curva come identificatore univoco della efficacia del sistema:



$$AUC = \sum_k P@k \times \Delta R_k$$

$$\text{Dove } \Delta R_k = R@k - R@(k-1)$$

$$AUC = \sum_{m=1}^N P@m \cdot \underline{(R@m - R@(m-1))}$$

Quando troviamo, alla posizione  $m$ , un documento non rilevante, abbiamo che  $R@m = R@(m-1)$  quindi la differenza va a 0. Difatti danno un contributo i soli documenti rilevanti, chiamiamo questo set ordinato per rilevanza  $R$ :

$$AUC = \sum_{k \in R} P@k \cdot (R@k - R@(k-1))$$

Considerando la definizione di  $R@k$ , possiamo sicuramente affermare che la differenza fra due documenti rilevanti consecutivi, in termini di  $RECALL$  è sempre pari a  $\frac{1}{kR}$  da qui:

$$AUC = \frac{1}{R} \sum_{k \in R} P@k = AP$$

## DISCOUNTED CUMULATIVE GAIN

$$DCG@K = \begin{cases} \sum_{i=1}^K \varepsilon_i & \text{per } K < b \\ DCG@(K-1) + \frac{\varepsilon_K}{\log_b K} & \text{per } K \geq b \end{cases}$$

Questa formula si può riassumere:

$$DCG@K = \sum_{i=1}^{K-1} \frac{\varepsilon_i}{\max(1, \log_b K)}$$

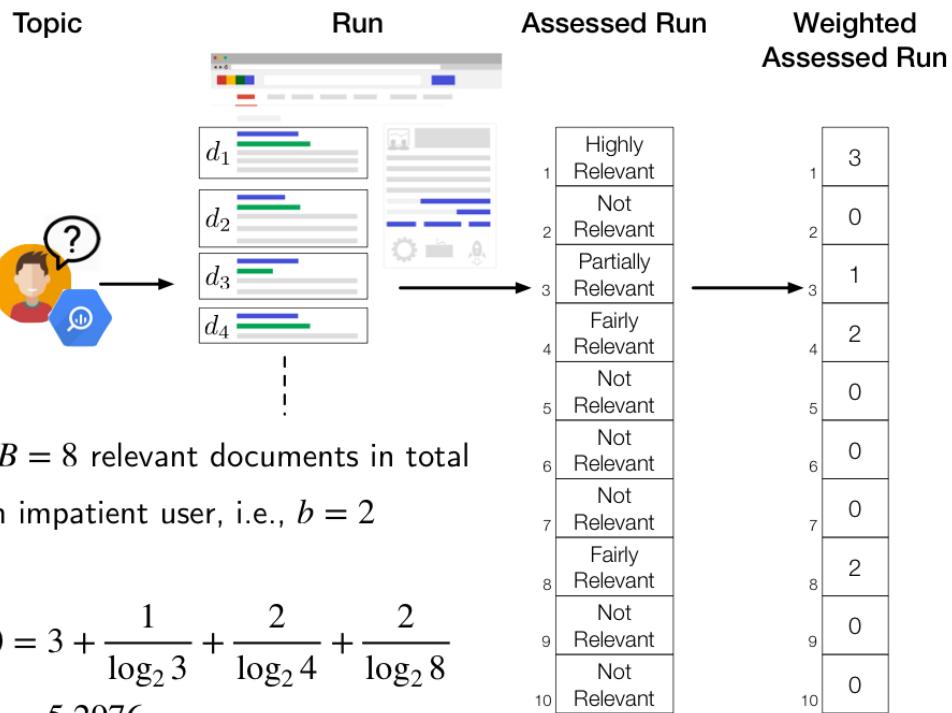
Dove  $b$  indica la "patienza" dell'utente nel scorrere i risultati, più è grande più l'utente scorre  $b$  risultati, dondoli quindi più pesa rispetto a quelli dopo  $b$ .

Gli utenti nella maggior parte dei casi non controllano tutti i risultati che un sistema IR restituisce, quindi danno più peso ai primi  $b$  risultati.

Da notare che DCG:

- 1) Gestisce naturalmente la rilevanza multi-livello.
- 2) Non dipende da una ~~recall-base~~ RB
- 3) Non è confinata fra  $[\emptyset, 1]$

# Example



- Assume  $RB = 8$  relevant documents in total
- Assume an impatient user, i.e.,  $b = 2$

$$DCG@10 = 3 + \frac{1}{\log_2 3} + \frac{2}{\log_2 4} + \frac{2}{\log_2 8} \\ = 5.2976$$

## USSR model

Ogni misura per valutare un sistema IR, ha al suo interno un modello utente, ovvero come ci si aspetta che un generico utente si comporti con il sistema:

1) **Browsing model**: Come si comporta con i risultati.

2) **Model of doc utility**: Come l'utente definisce utile un documento rilevante

3) **Utility accumulation model**: Come l'utente accumula utilità durante la ricerca.

# NORMALIZED DCG

Per normalizzare DCG@K fra [0, 1] bisogna calcolare il **run ideale**, ovvero i top-k documenti rilevanti, ordinati per rilevanza discendente. Questo ordine ottimale corrisponde al massimo valore possibile di DCG@K ed è detto **iNDCG@K**, da qui la normalizzazione:

$$mDCG@K = \frac{DCG@K}{eNDCG@K}$$

## Example

The diagram illustrates the calculation of DCG@10, iDCG@10, and nDCG@10 from a user query, a run, an assessed run, a weighted assessed run, and a weighted ideal run.

**User Query:** A user icon with a question mark is connected by an arrow to the **Run**.

**Run:** A horizontal bar chart showing document snippets  $d_1$ ,  $d_2$ ,  $d_3$ , and  $d_4$ . Each snippet has a green bar above it, indicating relevance.

**Assessed Run:** A vertical list of 10 relevance judgments for documents  $d_1$  through  $d_{10}$ :

|    |                    |
|----|--------------------|
| 1  | Highly Relevant    |
| 2  | Not Relevant       |
| 3  | Partially Relevant |
| 4  | Fairly Relevant    |
| 5  | Not Relevant       |
| 6  | Not Relevant       |
| 7  | Not Relevant       |
| 8  | Fairly Relevant    |
| 9  | Not Relevant       |
| 10 | Not Relevant       |

**Weighted Assessed Run:** A vertical list of 10 scores corresponding to the assessed run judgments:

|    |   |
|----|---|
| 1  | 3 |
| 2  | 0 |
| 3  | 1 |
| 4  | 2 |
| 5  | 0 |
| 6  | 0 |
| 7  | 0 |
| 8  | 2 |
| 9  | 0 |
| 10 | 0 |

**Weighted Ideal Run:** A vertical list of 10 scores representing the ideal run (all relevant documents are highly relevant):

|    |   |
|----|---|
| 1  | 3 |
| 2  | 3 |
| 3  | 2 |
| 4  | 2 |
| 5  | 2 |
| 6  | 1 |
| 7  | 1 |
| 8  | 1 |
| 9  | 0 |
| 10 | 0 |

**Assumptions:**

- Assume  $RB = 8$  relevant documents in total
- Assume an impatient user, i.e.,  $b = 2$

**Results:**

- $DCG@10 = 5.2976$
- $iDCG@10 = 10.1996$
- $nDCG@10 = 0.5194$

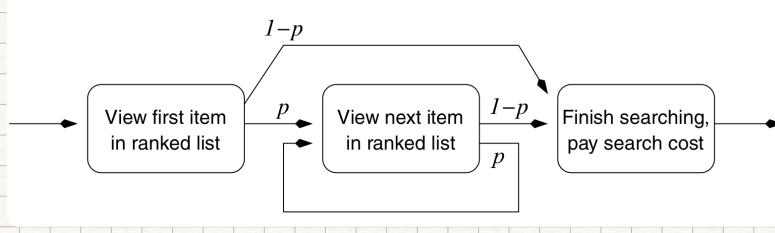
## RANK-BIASED PRECISION

Questo user model presuppone che l'utente partendo dai top- $K$  doc con probabilità  $p$  va al prossimo doc mentre con probabilità  $(1-p)$  si ferma.  $p$  è detta **persistence**.

Valori tipici per  $p$  sono 0.5 per utenti impazienti, mentre 0.8 per utenti pazienti. 0.95 estremamente pazienti.

$$RBP@K = (1-p) \sum_{i=1}^K p^{i-1} \cdot r_i$$

La prima sommatoria parte da 1 perché il modello **presuppone** che il primo documento venga sempre letto.



### Example

| Topic  | Run | Assessed Run   | Binary Weighted Assessed Run |
|--|-----|--|------------------------------|
|  |     |  |                              |
|  |     |  |                              |
|  |     | 1 Highly Relevant<br>2 Not Relevant<br>3 Partially Relevant<br>4 Fairly Relevant<br>5 Not Relevant<br>6 Not Relevant<br>7 Not Relevant<br>8 Fairly Relevant<br>9 Not Relevant<br>10 Not Relevant |                              |
| <ul style="list-style-type: none"> <li>Assume RB = 8 relevant documents in total</li> <li>Assume a patient user, i.e., <math>p = 0.8</math></li> </ul> |     | $RBP@10 = (1 - 0.8)(0.8^{1-1} + 0.8^{3-1} + 0.8^{4-1} + 0.8^{8-1}) = 0.4723$   |                              |

## MEAN RECIPROCAL RANK

Alcune query ha un solo (o quasi) documento  
davvero rilevante, sono i casi:

- 1) Quando si conosce l'oggetto della ricerca (Pagina Wikipedia)
- 2) Query navigazionale (voglio l'URL di Facebook)
- 3) Cerco un fatto o una notizia

Considerando la posizione in classifica, diciamo sia  $K$ ,  
del primo documento rilevante. Esso, in questo caso, sarà  
l'unico su cui l'utente cliccherà.

Definiamo il **Reciprocal Rank score RR**:

$$RR = \frac{1}{K}$$

Il **mean reciprocal rank** è la media degli RR  
per diverse query.

Questo però non è una buona metrica

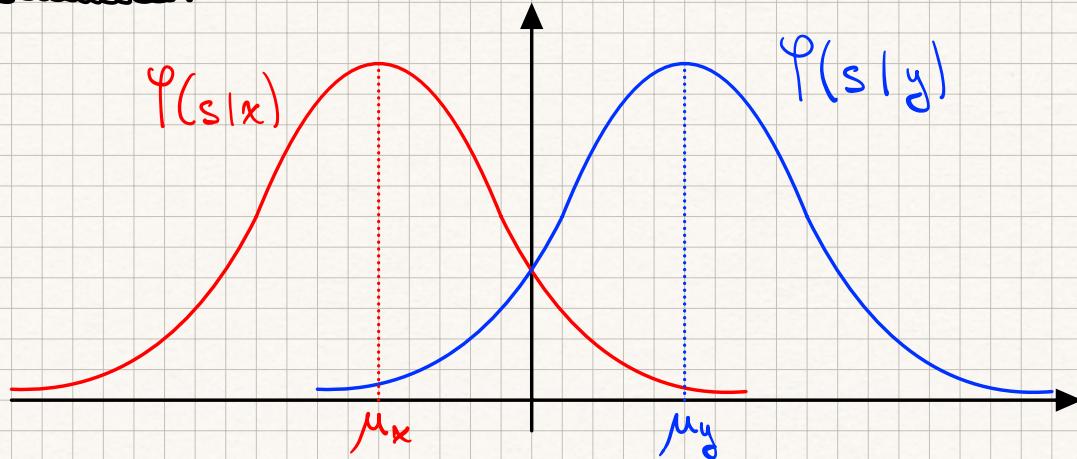
## SIGNIFICANCE TESTING

Cosa possiamo concludere, dai risultati delle stesse query per due risultati di due sistemi? Non possiamo dare conclusioni affidate senza scommettere un test statistico.

Chiamiamo  $Q$  la variabile aleatoria che rappresenta tutte le query possibili. Usiamo ora una funzione  $M$  per trasformare  $Q$  in un'altra variabile aleatoria  $S$ , che è poi lo score:



Ora processiamo due sistemi diversi  $X$  e  $Y$ , li processiamo con le stesse query randomiche  $q \in Q$  e riceviamo  $S_x$  e  $S_y$  ovvero gli score dei due sistemi.  $S_x$  e  $S_y$  creano una distribuzione di probabilità. Calcoliamola:



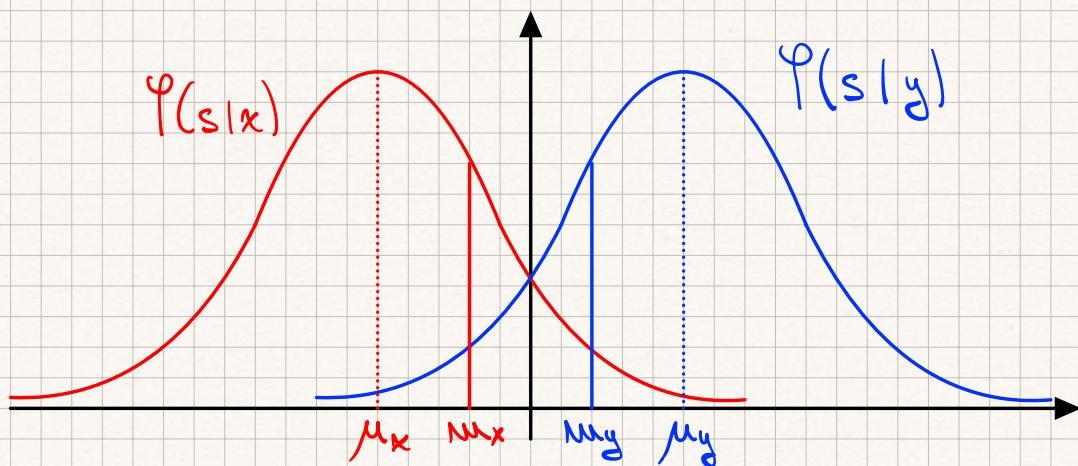
Calcolando le  $\mu_x$  e  $\mu_y$ , dette **population mean**.

Noi però non abbiamo tutte le query  $q \in Q$  possibili, ma abbiamo solo un sottoinsieme di query di benchmark  $\{q_1, \dots, q_m\} \subseteq Q$  da dove deriviamo  $S_x(q_1), \dots, S_x(q_m)$  e  $S_y(q_1), \dots, S_y(q_m)$ . Questi score possono riassumere come:

$$\bar{M}_x = \frac{1}{N} \sum_{i=1}^m S_x(q_i)$$

$$\bar{M}_y = \frac{1}{N} \sum_{i=1}^m S_y(q_i)$$

Detta sample mean che fornisce approssimazione  $M_x$  e  $M_y$  ma non è propriamente esatta, questo perché è calcolata su un sottoinsieme di  $Q$ :



Come in figura può succedere che  $\bar{M}_x$  e  $\bar{M}_y$  siano più vicini di quanto sono vicini  $M_x$  e  $M_y$ . Questo non ci permette di concludere che il sistema X sia migliore di Y o viceversa.

Definiamo ora la nuova variabile aleatoria  $\Delta$ :

$$\Delta = S_x - S_y$$

Da qui calcoliamo  $M_g$ . In realtà possiamo solo calcolare  $M_{\bar{g}}$ :

$$\{q_1 \dots q_m\} \rightarrow \{e_1 \dots e_n\} \rightarrow M_{\bar{g}}$$

Ora bisogna capire se il campione di query che stiamo considerando, è abbastanza rappresentativo oppure abbiamo avuto sfortuna e il campione ci mostra proprio il contrario di quello che poi è la realtà dei fatti, ovvero consideran  $\neq H_0 \in Q$ .

### NULL HYPOTHESIS

In generale si formula una ipotesi  $H$  su  $\mathcal{S}$ .

Questa  $H$  può avere due forme

1) **Parametrica**:  $H: "M_g = 3"$ , basato su  $\mathcal{S}$

2) **Non-Parametrica**:  $H: "\mathcal{S} \text{ segue la distribuzione normale}"$ , basato sulla forma di  $\Phi(e|H)$

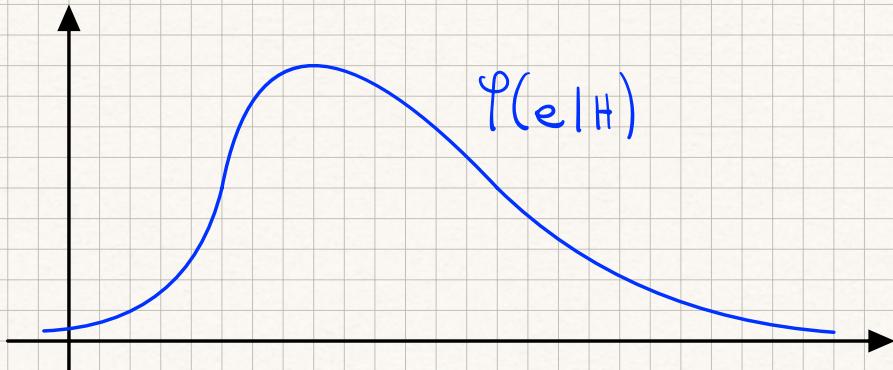
All'inizio assumiamo che  $H$  sia vera.

Ragioniamo su quello che consideriamo sia la distribuzione di  $\Phi(e|H)$ , dove consideriamo  $H$  vera.

Da qui applichiamo una procedura che ci permette di rigettare o confermare  $H$ .

PASSI

1) Assumiamo che  $\varphi(e|H)$  abbia una certa forma:

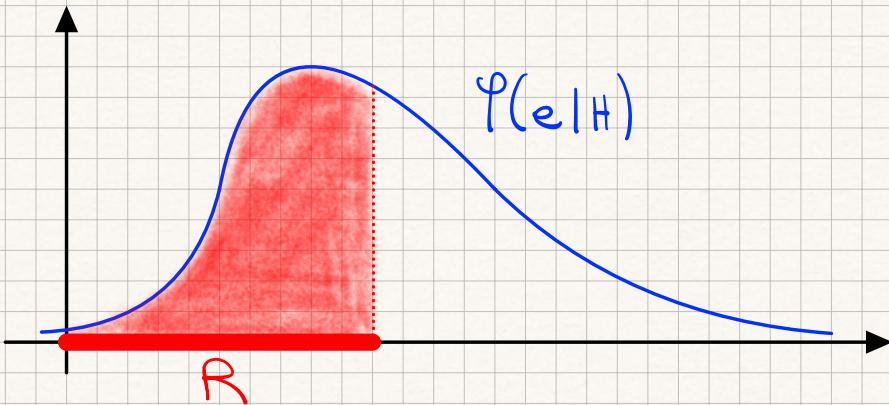


2) Assumiamo di usare un test statistico  $Z_n$

$$Z_n = Z_n(\delta_1 \dots \delta_n)$$

Il test  $Z_n$  infatti è una trasformazione di un gruppo di variabili aleatorie ( $\delta_1 \dots \delta_n$  nel nostro caso)

3) Bisogna ora definire una regione  $R$ , dove poter rifiutare  $H$ .



Quindi se il nostro valore attuale cade dentro  $R$ , allora possiamo rifiutare  $H$ .

4) Calcoliamo  $Z_m$  con i nostri campioni

5) Rigettiamo  $H_0$  se  $Z_m \in R$ . Se  $Z_m \notin R$  non siamo in grado di concludere nulla!

Dobbiamo quindi selezionare  $\varphi(\cdot)$ ,  $Z_m \in R$ .

Per convenzione la Null Hypothesis è detta  $H_0$ , mentre il suo complementare è detto  $H_1$ .

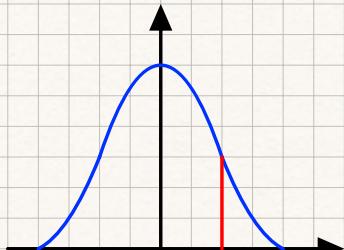
## P-VALUE

I test statisticci servono per calcolare il p-value, ovvero la probabilità che un valore statistico possa essere osservato se  $H_0$  è vera.

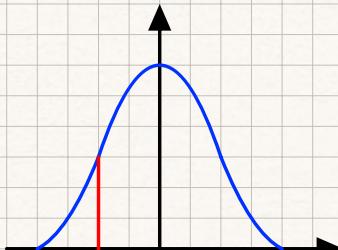
$$P(Z_m \in R \mid H_0) = \int_R \varphi(z_m \mid H_0) dz_m = \alpha$$

Se il p-value è sufficientemente piccolo, possiamo rigettare  $H_0$ .

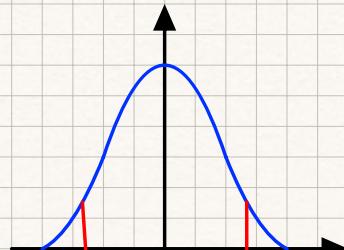
Come decido i confini di  $R$ ?



ONE SIDE



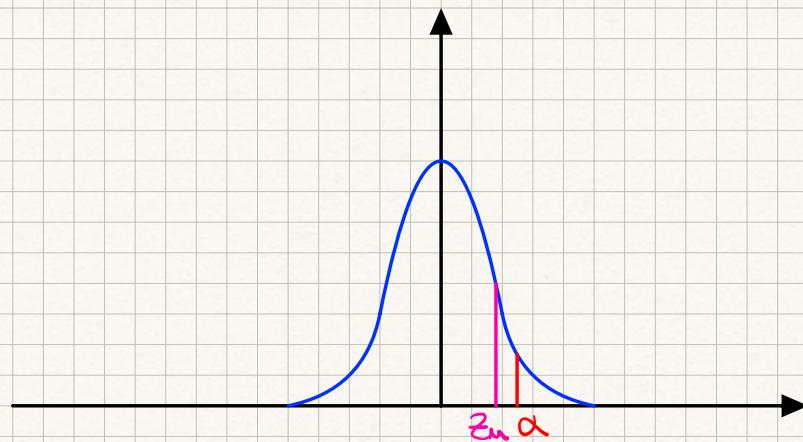
ONE SIDE



TWO SIDES

Ora valutiamo il P-value

$$P(|z_m| \geq z_m | H_0) = \int_{|z_m| \geq z_m} \varphi(e | H_0) de = \text{P-value}$$



Dove  $z_m$  è il valore del nostro test statistico.

Se il P-value  $\leq \alpha$  allora possiamo rigettare  $H_0$ ,  
in favore di  $H_1$ .

$\alpha$  è detto **significant level**.

## t-TEST

Il t-Test è un test statistico che si basa sul calcolo del parametico t:

$$t = \frac{\overline{B-A}}{\sqrt{K}}$$

Una volta trovato t, bisogna fissare il **p-value** desiderato. Una volta fissato il p-value, bisogna cercare in corrispondenza del nostro p-value e dei gradi di libertà df (Il numero di campioni usati per calcolare  $\overline{B-A}$ ), il valore di t, in una tabella:

| df | 0.25  | 0.20  | 0.15  | 0.10  | 0.05  | 0.025 | 0.02  | 0.01  | 0.005 | 0.001 |
|----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1  | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 15.89 | 31.82 | 63.66 | 318.3 |
| 2  | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 4.849 | 6.965 | 9.925 | 22.33 |
| 3  | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 3.482 | 4.541 | 5.841 | 10.21 |
| 4  | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 2.999 | 3.747 | 4.604 | 7.173 |
| 5  | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 2.757 | 3.365 | 4.032 | 5.893 |
| 6  | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 2.612 | 3.143 | 3.707 | 5.208 |
| 7  | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.517 | 2.998 | 3.499 | 4.785 |
| 8  | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.449 | 2.896 | 3.355 | 4.501 |
| 9  | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.398 | 2.821 | 3.250 | 4.297 |
| 10 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.359 | 2.764 | 3.169 | 4.144 |
| 11 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.328 | 2.718 | 3.106 | 4.025 |
| 12 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.303 | 2.681 | 3.055 | 3.930 |
| 13 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.282 | 2.650 | 3.012 | 3.852 |
| 14 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.264 | 2.624 | 2.977 | 3.787 |

Se il valore di t calcolato è maggiore uguale al t in tabella allora possiamo rifiutare  $H_0$ .

$$t \geq t_{\text{TAB}} \Rightarrow \text{Rifiuto } H_0$$

## WILCOXON TEST

Il Wilcoxon Test si basa sui segni del set  $B-A$  dove vado a cercare  $t_{i=1 \dots df} b_i - a_i$ .

- 1) Calcolo ogni  $b_i - a_i$  con  $b_i \in B$  e  $a_i \in A$ .
- 2) Prendo i valori assoluti e li ordino. Scrivo però i segni.
- 3) Faccio un Rank dei "valori assoluti" ordinati. In caso di differenze uguali in modulo aggiungo a ogni rank un  $\frac{1}{2}$ :

$$P = \frac{1}{\# \text{ differenze uguali in modulo}}$$

### Esempio

Ci sono tre numeri X alla posizione 5, allora avrò tre rank = 5, 3.

- 4) Rimetto i segni ai valori dei rank. Sono i rank.
- 5) Faccio la stessa cosa del t-Test, solo che con la tabella del Wilcoxon Test, e quindi vado alla ricerca della somma dei rank.

| alpha values |       |       |      |       |      |      |      |
|--------------|-------|-------|------|-------|------|------|------|
| n            | 0.001 | 0.005 | 0.01 | 0.025 | 0.05 | 0.10 | 0.20 |
| 5            | --    | --    | --   | --    | --   | 0    | 2    |
| 6            | --    | --    | --   | --    | 0    | 2    | 3    |
| 7            | --    | --    | --   | 0     | 2    | 3    | 5    |
| 8            | --    | --    | 0    | 2     | 3    | 5    | 8    |
| 9            | --    | 0     | 1    | 3     | 5    | 8    | 10   |
| 10           | --    | 1     | 3    | 5     | 8    | 10   | 14   |
| 11           | 0     | 3     | 5    | 8     | 10   | 13   | 17   |
| 12           | 1     | 5     | 7    | 10    | 13   | 17   | 21   |
| 13           | 2     | 7     | 9    | 13    | 17   | 21   | 26   |
| 14           | 4     | 9     | 12   | 17    | 21   | 25   | 31   |
| 15           | 6     | 12    | 15   | 20    | 25   | 30   | 36   |
| 16           | 8     | 15    | 19   | 25    | 29   | 35   | 42   |
| 17           | 11    | 19    | 23   | 29    | 34   | 41   | 48   |
| 18           | 14    | 23    | 27   | 34    | 40   | 47   | 55   |
| 19           | 18    | 27    | 32   | 39    | 46   | 53   | 62   |
| 20           | 21    | 32    | 37   | 45    | 52   | 60   | 69   |
| 21           | 25    | 37    | 42   | 51    | 58   | 67   | 77   |
| 22           | 30    | 42    | 48   | 57    | 65   | 75   | 86   |
| 23           | 35    | 48    | 54   | 64    | 73   | 83   | 94   |
| 24           | 40    | 54    | 61   | 72    | 81   | 91   | 104  |
| 25           | 45    | 60    | 68   | 79    | 89   | 100  | 113  |
| 26           | 51    | 67    | 75   | 87    | 98   | 110  | 124  |
| 27           | 57    | 74    | 83   | 96    | 107  | 119  | 134  |

| alpha values |       |       |      |       |      |      |      |
|--------------|-------|-------|------|-------|------|------|------|
| n            | 0.001 | 0.005 | 0.01 | 0.025 | 0.05 | 0.10 | 0.20 |
| 28           | 64    | 82    | 91   | 105   | 116  | 130  | 145  |
| 29           | 71    | 90    | 100  | 114   | 126  | 140  | 157  |
| 30           | 78    | 98    | 109  | 124   | 137  | 151  | 169  |
| 31           | 86    | 107   | 118  | 134   | 147  | 163  | 181  |
| 32           | 94    | 116   | 128  | 144   | 159  | 175  | 194  |
| 33           | 102   | 126   | 138  | 155   | 170  | 187  | 207  |
| 34           | 111   | 136   | 148  | 167   | 182  | 200  | 221  |
| 35           | 120   | 146   | 159  | 178   | 195  | 213  | 235  |
| 36           | 130   | 157   | 171  | 191   | 208  | 227  | 250  |
| 37           | 140   | 168   | 182  | 203   | 221  | 241  | 265  |
| 38           | 150   | 180   | 194  | 216   | 235  | 256  | 281  |
| 39           | 161   | 192   | 207  | 230   | 249  | 271  | 297  |
| 40           | 172   | 204   | 220  | 244   | 264  | 286  | 313  |
| 41           | 183   | 217   | 233  | 258   | 279  | 302  | 330  |
| 42           | 195   | 230   | 247  | 273   | 294  | 319  | 348  |
| 43           | 207   | 244   | 261  | 288   | 310  | 336  | 365  |
| 44           | 220   | 258   | 276  | 303   | 327  | 353  | 384  |
| 45           | 233   | 272   | 291  | 319   | 343  | 371  | 402  |
| 46           | 246   | 287   | 307  | 336   | 361  | 389  | 422  |
| 47           | 260   | 302   | 322  | 353   | 378  | 407  | 441  |
| 48           | 274   | 318   | 339  | 370   | 396  | 426  | 462  |
| 49           | 289   | 334   | 355  | 388   | 415  | 446  | 482  |
| 50           | 304   | 350   | 373  | 406   | 434  | 466  | 503  |