Retrieval Evaluation

[DAT640] Information Retrieval and Text Mining

Ivica Kostric & Krisztian Balog
University of Stavanger

September 5, 2022

CC BY 4.0

In this module

1. Retrieval evaluation
2. Statistical significance testing

Retrieval evaluation

Evaluation

"To measure is to know.
If you can not measure it,
you can not improve it."

—Lord Kelvin

What to measure?

- **Effectiveness** $\Leftarrow$ our focus
  - How accurate are the search results?
  - I.e., the system's capability of ranking relevant documents ahead of non-relevant ones
- Efficiency
  - How quickly can a user get the results?
  - I.e., the response time of the system
- Usability
  - How useful is the system for real user tasks?

Evaluation in IR

- Search engine evaluation must rely on users!
- Core question: How can we get users involved?

Types of evaluation

- **Offline** (test collection based) $\Leftarrow$ our focus
- **Online** (live evaluation) $\Leftarrow$ our focus
- User studies
- Simulation of users
- ...

Retrieval evaluation

- Offline evaluation
- Online evaluation

Test collection based evaluation

- *Cranfield evaluation methodology*
- Basic idea: Build reusable test collections
- Ingredients of an IR test collection
  - Dataset (corpus of documents or *information objects*)
  - Test queries (set of *information needs*)
  - Relevance assessments
  - Evaluation measures

Relevance assessments

- Ground truth labels for query-item pairs
- **Binary**
  - 0: non-relevant
  - 1: relevant
- **Graded**, for example,
  - -1: spam / junk
  - 0: non-relevant
  - 1: somewhat relevant
  - 2: relevant
  - 3: highly relevant / perfect match

query 1	item 11	0
	item 12	1
	item 13	1
	item 14	0
	item 15	0
	...	
query 2	item 21	1
	item 22	1
	item 23	0
	...	

ground truth with binary assessments

Obtaining relevance assessments

- Obtaining relevance judgments is an expensive, time-consuming process
  - Who does it?
  - What are the instructions?
  - What is the level of agreement?
- Two approaches
  - Expert judges
  - Crowdsourcing

Text Retrieval Conference (TREC)

- Organized by the US National Institute of Standards and Technology (NIST)
- Yearly benchmarking cycle
- Developing test collections for various information retrieval tasks
- Relevance judgments created by expert judges, i.e., retired information analysts (CIA)

Examples of TREC document collections

Name	#Documents	Size
CACM	3k	2.2 MB
AP	242k	0.7 GB
GOV2	25M	426 GB
ClueWeb09	1B	25 TB

TREC topic example

<top>
<num> Number: 794

<title> pet therapy

<desc> Description:
How are pets or animals used in therapy for humans and what are the benefits?

<narr> Narrative:
Relevant documents must include details of how pet- or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it.

</top>

Crowdsourcing

- Obtain relevance judgments on a crowdsourcing platform
    - Often branded as "human intelligence platforms"
- "Microtasks" are performed in parallel by large, paid crowds

amazon mechanical turk

figure eight

CrowdTruth
The framework for crowdsourcing ground truth data

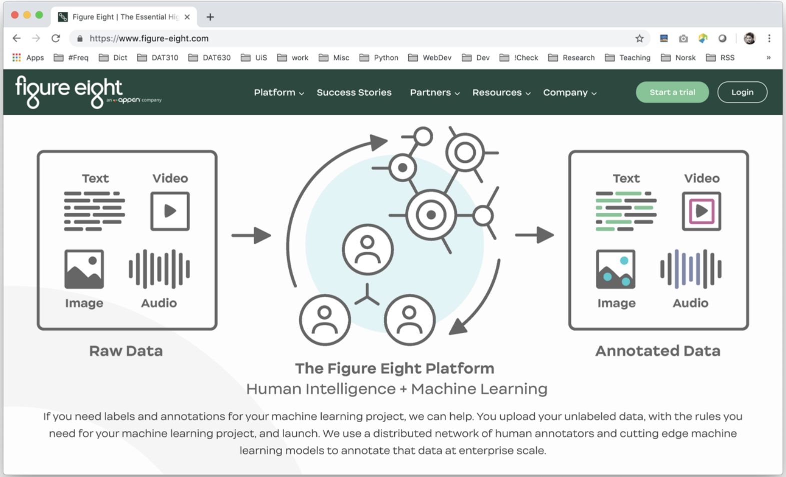


Figure Eight | The Essential Hi
https://www.figure-eight.com
Apps
#Freq
Dict
DAT310
DAT630
UiS
work
Misc
Python
WebDev
Dev
!Check
Research
Teaching
Norsk
RSS
figure eight
an appen company
Platform
Success Stories
Partners
Resources
Company
Start a trial
Login
Text
Video
Image
Audio
Raw Data
Text
Video
Image
Audio
Annotated Data
The Figure Eight Platform
Human Intelligence + Machine Learning
If you need labels and annotations for your machine learning project, we can help. You upload your unlabeled data, with the rules you need for your machine learning project, and launch. We use a distributed network of human annotators and cutting edge machine learning models to annotate that data at enterprise scale.

Example microtask

Query: button down shirt

Click here to search on Google

Result Title: Men's Essential Poplin Button-down Shirt

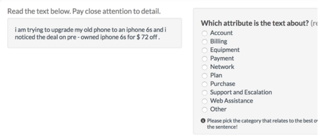
Result Image:

Click here to look at the result page

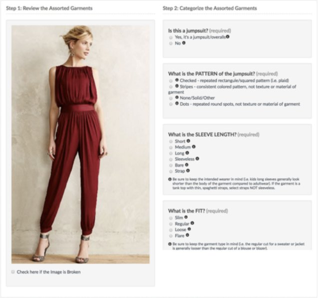
Rate how well 'Men's Essential Poplin Button-down Shirt' matches the query (required)

Irrelevant	Somewhat relevant	Relevant	Perfect Match
○	○	○	○

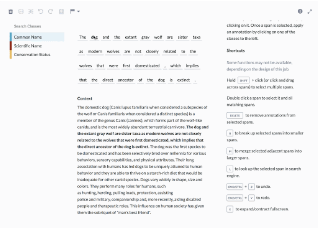
Other search related annotation tasks



Intent classification



Content categorization



Text annotation

Expert judges vs. crowdsourcing

- Expert judges
  - Each query-item pair is commonly assessed by a single person
  - Agreement is good because of "narrative"
- Crowdsourcing
  - Assessments are more noisy
  - Commonly, majority vote is taken
    - The number of labels collected for an item may be adjusted dynamically such that a majority decision is reached
- **Data is only as good as the guidelines!**

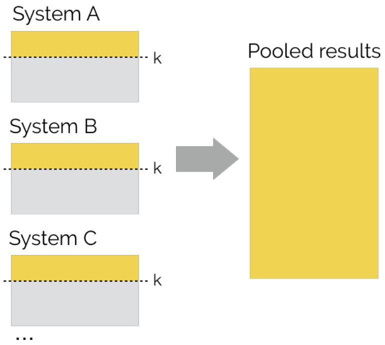
Discussion

Question

How can the relevance of all items be assessed in a large dataset for a given query?

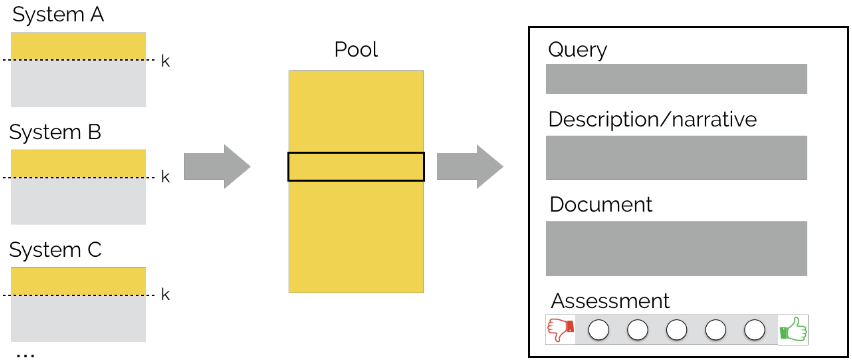
Pooling

- Exhaustive judgments for all documents in a collection is not practical
- Top-k results from different systems (algorithms) are merged into a pool
  - Duplicates are removed
  - Item order is randomized
- Produces a large number of relevance judgments for each query, although still incomplete
  - Not assessed items are assumed to be non-relevant



Pooling

- Relevance assessments are collected for all documents in the pool
  - Either using expert judges or crowd workers



Test collection based evaluation

- Ingredients of an IR test collection
  - ~~Dataset (corpus of documents or *information objects*)~~
  - ~~Test queries (set of *information needs*)~~
  - ~~Relevance assessments~~
  - **Evaluation measures**

IR evaluation measures

- Assessing the quality of a ranked list against the ground truth relevance labels
    - Commonly, a real number between 0 and 1
- **Important**: All measures are based on a (simplified) model of user needs and behavior
    - That is, the right measure depends on the particular task

Effectiveness measures

- A is the set of **relevant** documents
- B is the set of **retrieved** documents

	Relevant	Non-relevant
Retrieved	$\|A \cap B\|$	$\|\overline{A} \cap B\|$
Not retrieved	$\|A \cap \overline{B}\|$	$\|\overline{A} \cap \overline{B}\|$

Precision and recall analogously to before:

$$P = \frac{|A \cap B|}{|B|} \qquad R = \frac{|A \cap B|}{|A|}$$

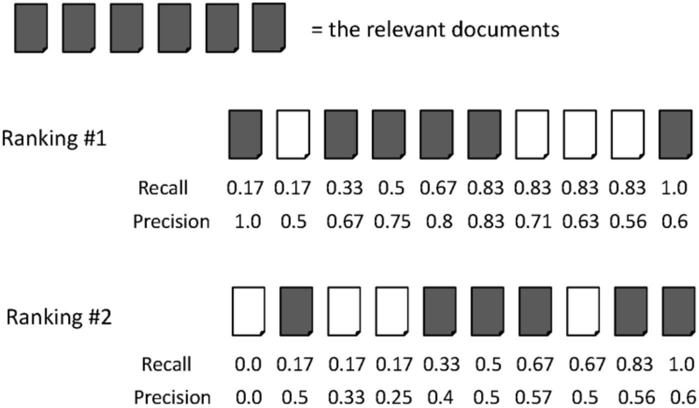
Discussion

Question

Precision and Recall are set-based metrics. How can we use them to evaluate ranked lists?

Evaluating rankings

Calculate recall and precision values at every rank position



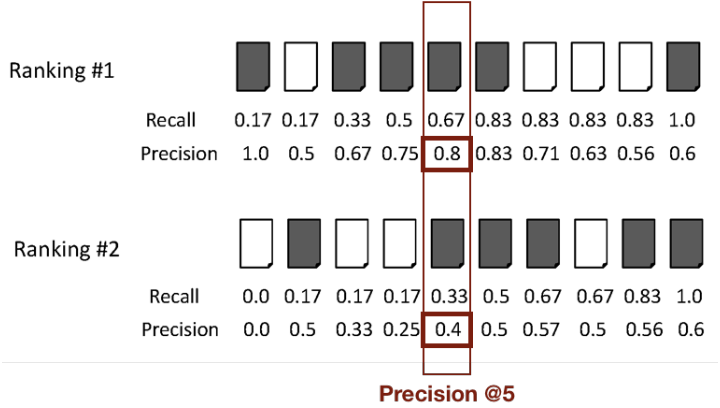
Evaluating rankings

- Calculating recall and precision values at every rank position produces a long list of numbers (see previous slide)
- Need to **summarize** the effectiveness of a ranking
- Various alternatives
  - Calculate recall and precision at fixed rank positions (P@k, R@k)
  - Calculate precision at standard recall levels, from 0.0 to 1.0 (requires interpolation)
  - Averaging the precision values from the rank positions where a relevant document was retrieved (AP)

Fixed rank positions

Compute precision/recall at a given rank position k (P@k, R@k)

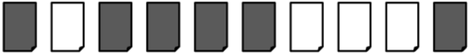
- This measure does not distinguish between differences in the rankings at positions 1 to k



Standard recall levels

Calculate precision at standard recall levels, from 0.0 to 1.0

- Each ranking is then represented using 11 numbers
- Values of precision at these standard recall levels are often not available, for example:



Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

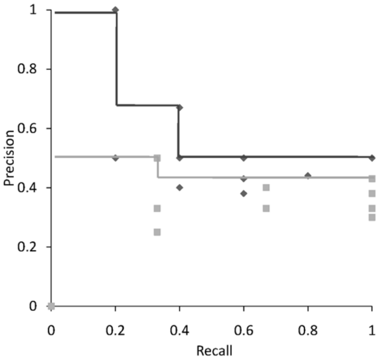
- *Interpolation* is needed

Interpolation

- To average graphs, calculate precision at standard recall levels:

  $$P(R) = \max\{P' : R' \geq R \wedge (R', P') \in S\}$$

  - where S is the set of observed (R, P) points
- Defines precision at any recall level as the maximum precision observed in any recall-precision point at a higher recall level
- Produces a step function



Average Precision

- Average the precision values from the rank positions where a relevant document was retrieved

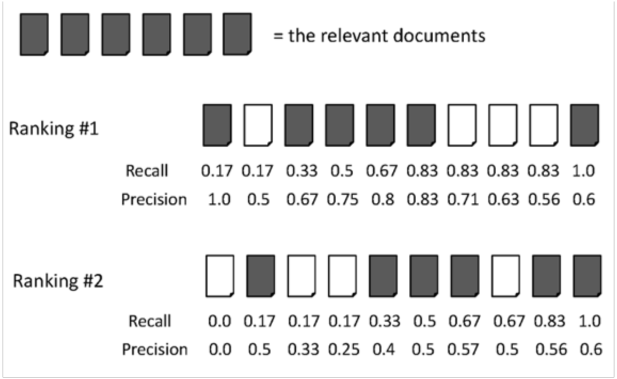
$$AP = \frac{1}{|Rel|} \sum_{\substack{i = 1, \ldots, n \\ d_i \in Rel}} P(i) \longrightarrow \text{Precision at rank i}$$

Total number of relevant documents
According to the ground truth

Only relevant documents contribute to the sum

- If a relevant document is not retrieved (in the top k ranks, e.g, $k = 1000$) then its contribution is 0.0
- AP is single number that is based on the ranking of all the relevant documents
- The value depends heavily on the highly ranked relevant documents

Average Precision



$$\text{Ranking \#1:}\ \ (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

$$\text{Ranking \#2:}\ \ (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

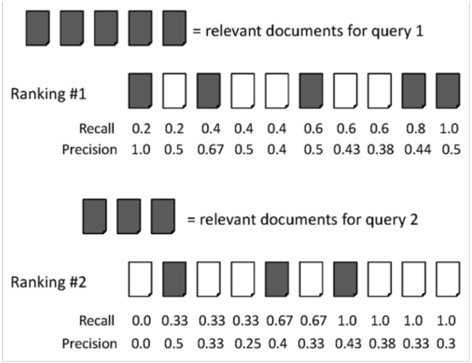
Averaging across queries

- So far: measuring ranking effectiveness on a **single query**
- Need: measure ranking effectiveness on a **set of queries**
- Average is computed over the set of queries

Mean Average Precision (MAP)

- Summarize rankings from multiple queries by averaging Average Precision
- Very succinct summary
- Most commonly used measure in research papers
- Assumes user is interested in finding many relevant documents for each query
- Requires many relevance judgments

Mean Average Precision



$$\textit{average precision query 1} = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$\textit{average precision query 2} = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$\textit{mean average precision} = (0.62 + 0.44)/2 = 0.53$$

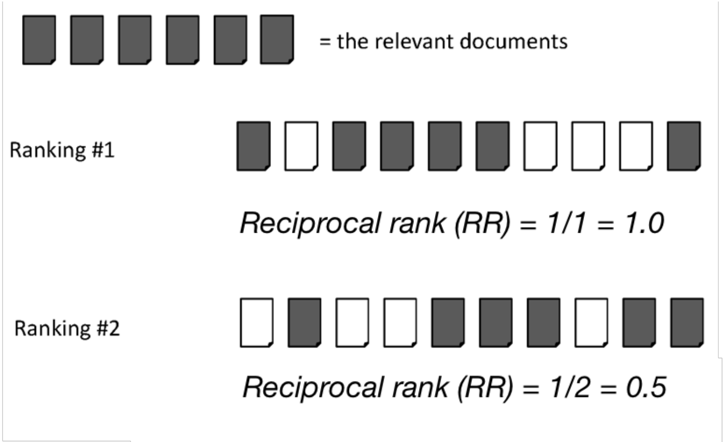
Focusing on top documents

- Users tend to look at only the top part of the ranked result list to find relevant documents
- Some search tasks have only one relevant document
    - E.g., navigational search, question answering
- Recall in those cases is not appropriate
    - Instead need to measure how well the search engine does at retrieving relevant documents at very high ranks

Focusing on top documents

- Precision at rank k (P@k)
  - k is typically 5, 10, 20
  - Easy to compute, average, understand
  - Not sensitive to rank positions less than k
- Reciprocal Rank (RR)
  - Reciprocal of the rank at which the first relevant document is retrieved
  - Mean Reciprocal Rank (MRR) is the average of the reciprocal ranks over a set of queries
  - Very sensitive to rank position

Mean Reciprocal Rank



Mean reciprocal rank (MRR) = (1.0 + 0.5) /2 = 0.75

Exercise

E5-1 Evaluation measures

Exercise

E5-2 Interpolated precision

Graded relevance

- So far: relevance in binary
- What about graded relevance levels?

Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks
- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant document
  - The lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined

Discounted Cumulative Gain (DCG)

- DCG is the total gain accumulated at a particular rank p:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2 i}$$

  - rel_i is the graded relevance level of the item retrieved at rank i
- Gain is accumulated starting at the top of the ranking and discounted by 1/log (rank)
  - E.g., discount at rank 4 is 1/2, and at rank 8 it is 1/3
- Average over the set of test queries
- Note: search engine companies have their own (secret) variants

Discounted Cumulative Gain

Rank (i)	1	2	3	4	5	6	7	8	9	10
Gain	3	2	3	0	0	1	2	2	3	0
Discounted gain	3	2/1	3/1.59	0	0	1/2.59	2/2.81	2/3	3/3.17	0
Discounted cumulative gain (DCG@i)	3	5	6.89	6.89	6.89	7.28	7.99	8.66	9.61	9.61

How good is a DCG@10 value of 9.61?

Normalized Discounted Cumulative Gain (NDCG)

- DCG values are often normalized by comparing the DCG at each rank with the DCG value for the perfect (ideal) ranking
  - I.e., divide DCG@i value with the ideal DCG value at rank i
  - Yields value between 0 and 1

Ideal ranking

Rank (i)	1	2	3	4	5	6	7	8	9	10
Gain	3	3	3	2	2	2	1	0	0	0
Discounted cumulative gain (DCG@i)	3	6	7.89	8.89	9.75	10.52	10.88	10.88	10.88	10.88

Exercise

E5-3 NDCG

Retrieval evaluation

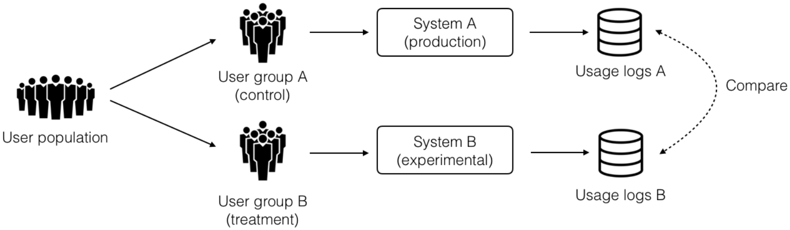
- Offline evaluation
- Online evaluation

Online evaluation

- **Idea**: See how normal users interact with a live retrieval system ("living lab") when just using it
- Observe implicit behavior
    - Clicks, skips, saves, forwards, bookmarks, likes, etc.
- Try to infer differences in behavior from different flavors of the live system
    - A/B testing, interleaving

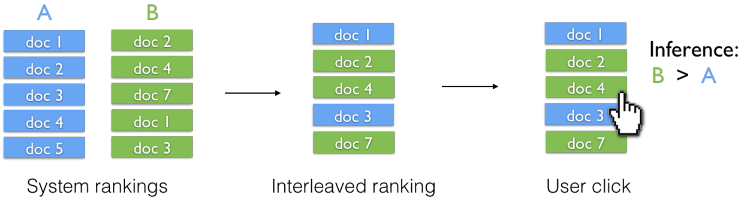
A/B testing

- Users are divided into two control (**A**) and treatment (**B**) groups
  - **A** uses the production system
  - **B** uses an experimental system
- Measure relative system performance based on usage logs



Interleaving

- Combine two rankings (A and B) into a single list
- Determine a winner on each query impression
  - Can be a draw too
- Aggregate wins on a large number of impressions to determine which ranker is better

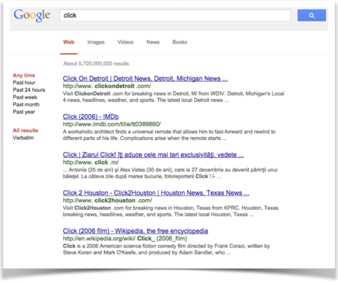


A/B testing vs. interleaving

- A/B testing
  - Between subject design
  - Can be used for evaluating any feature (new ranking algorithms, new features, UI design changes, etc.)
- Interleaving
  - Within subject design
  - Reduces variance (same users/queries for both A and B)
  - Needs 1 to 2 orders of magnitude less data
    - $\sim$100K queries for interleaving in a mature web search engine ($\gg$1M for A/B testing)
  - Limited to evaluating ranked lists

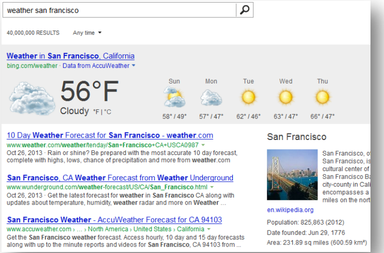
Measures in online evaluation

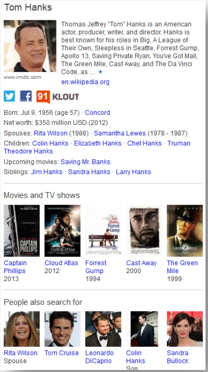
- **Inferred from observable user behavior**
- Clicks
- Mouse movement
- Browser action
  - Bookmark, save, print, ...
- Time
  - Dwell time, time on SERP, ...
- Explicit judgment
  - Likes, favorites, ...
- Query reformulations
- ...



Challenges in online evaluation

- **Simple measures break!**



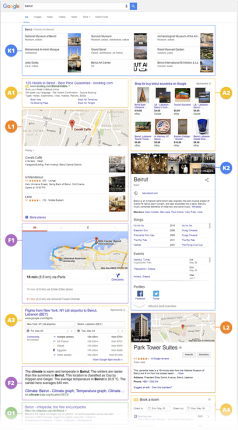


Instant answers
(satisfaction not observable)

Exploration
(more time/queries is not necessarily bad effort)

Challenges in online evaluation

- **Whole page relevance**
- Page is composed by a layered stack of modules
  - Web result ranking
  - $\Rightarrow$ Result caption generation
  - $\Rightarrow$ Answer triggering/ranking
  - $\Rightarrow$ Knowledge panel composition
  - $\Rightarrow$ Whole page composition
- Changes in modules lower in the stack have upstream effects



Pros and cons of online evaluation

- Advantages
  - No need for expensive dataset creation
  - Perfectly realistic setting: (most) users are not even aware that they are guinea pigs
  - Scales very well: can include millions of users
- Disadvantages
  - Requires a service with lots of users
  - Can be highly nontrivial how to interpret implicit feedback signals
  - Experiments are difficult to repeat

Offline vs. online evaluation

	Offline	Online
Basic assumption	Assessors tell you what is relevant	Observable user behavior can tell you what is relevant
Quality	Data is only as good as the guidelines	Real user data, real and representative information needs
Realisticity	Simplified scenario, cannot go beyond a certain level of complexity	Perfectly realistic setting (users are not aware that they are guinea pigs)
Assessment cost	Expensive	Cheap
Scalability	Doesn't scale	Scales very well
Repeatability	Repeatable	Not repeatable
Throughput	High	Low
Risk	None	High

Statistical significance testing

Statistical significance testing

- Comparison between two systems.
- Inherent noise in evaluation (*e.g.*, variations in topics, assessors' behavior).
- Question: Is a result likely due to chance?

Essential ingredients of a significance test

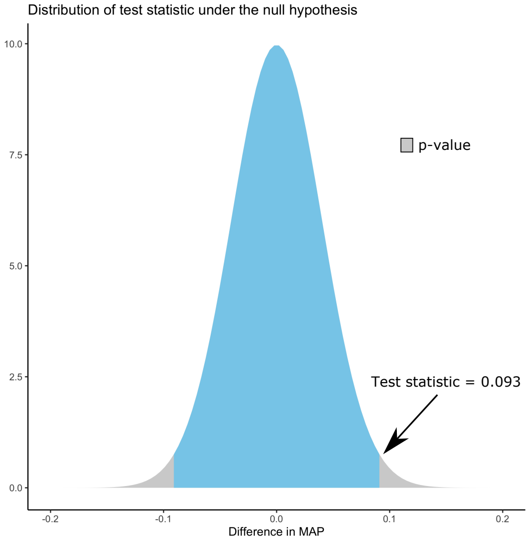
- A **test statistic or criterion** used to compare two systems. The difference in mean of an IR metric is commonly used.
- A **null hypothesis** H_0 and an **alternative hypothesis** H_1.
- A distribution of the test statistic given H_0.
- A **significance level** α used to determine if the comparison is statistically significant.
- **p-value** probability which determine whether there is evidence to reject H_0.

$$\text{p-value} = P(T(X^*) \leq T(X_0) \mid H_0) + P(T(X^*) \geq T(X_0) \mid H_0) \tag{1}$$

Test statistic example

	System A	System B	Difference (A-B)
	0.2215	0.0765	0.145
	0.3924	0.0426	0.3498
	0.6540	0.5738	0.0802
	0.5611	0.1571	0.404
	0.9186	0.9881	-0.0695
	0.1104	0.7164	-0.606
	0.6086	0.7507	-0.1421
	0.5062	0.4350	0.0712
	0.9688	0.3959	0.5729
	0.9950	0.8709	0.1241
Mean Average Precision (MAP)	0.5937	0.5007	**0.093**

p-value example



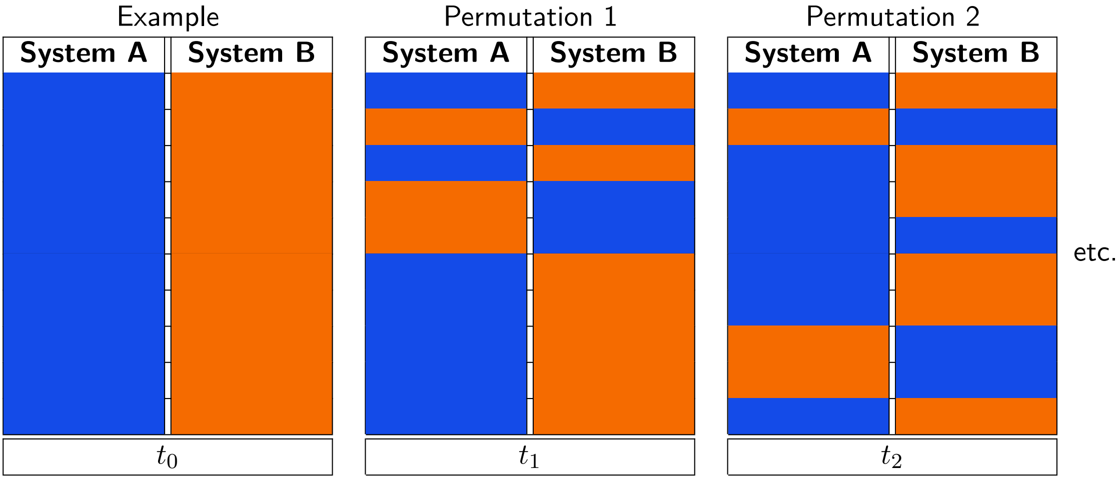
Commonly-used tests

- **Randomization (permutation) test** $\Leftarrow$ our focus
- **Student's paired sample t-test** $\Leftarrow$ our focus
- Wilcoxon signed rank test
- Bootstrap test
- Sign test
- ...

Randomization (permutation) test

- Null hypothesis H_0: systems A and B are identical.
- alternative hypothesis H_1: systems A and B are not identical.
- Test statistic of your choice. For this example, we use the mean difference.
- Distribution of test statistic T
  - Create all or a sample of permutations.
  - Record test statistic for each permutation.
- Compute p-value

Randomization (permutation) test



$T = [t_0, t_1, t_2, etc., t_p]$, with p as number of permutations.

Student's paired sample t-test

- Hypothesis for paired sample t-test
  - $H_0 : \bar{x}_A = \bar{x}_B$, systems A and B are random samples from the same normal distribution.
  - $H_1 : \bar{x}_A \neq \bar{x}_B$
- $t = \frac{\bar{x}_D}{\frac{s_D}{\sqrt{n}}}$, with $\bar{x}_D$ and s_D as the average and standard deviation of the differences between all pairs.
- Compute p-value

Randomization test vs. Student's t-test

	Randomization test	Student's t-test
Test statistic	Any	Difference of means
Normality assumption	No	Yes

Exercise

E5-4 Statistical significance testing

Summary

- Ingredients of offline test collections
- Collecting relevance assessments (expert judges/crowdsourcing, pooling, binary vs. graded relevance)
- Set retrieval measures (precision, recall, F1)
- Ranked retrieval measures (AP, RR, NDCG)
- Evaluating rankings for multiple queries
- Online evaluation (A/B testing and interleaving)
- Statistical significance testing

Reading

- Text Data Management and Analysis (Zhai&Massung)
    - Chapter 9
- Smucker, Mark D., James Allan, and Ben Carterette. "A comparison of statistical significance tests for information retrieval evaluation." *Proceedings of the sixteenth ACM Conference on information and knowledge management*. 2007.