Information Retrieval

Exercise – Winter term 2025/2026

klara.gutekunst@uni-kassel.de

Agenda

1. Research Questions
2. Hypothesis Testing
3. Assignment
4. Inspiration

Research Questions

What is a good research question?

Research Questions

- A good research question. . . [Bartos 1992]
  - . . . asks about the relationship between two or more variables.
  - . . . is testable (i.e., it is possible to collect data to answer the question).
  - . . . is stated clearly and in the form of a question.
  - . . . does not pose an ethical or moral problem for implementation.
  - . . . is specific and restricted in scope.
  - . . . identifies exactly what is to be solved.

- Examples:
  - *Poor:*
    “What is the effectiveness of parent education when given problem children?”
  - *Good:*
    “What is the effect of the STEP parenting program on the ability of parents to use natural, logical consequences (as opposed to punishment) with their child who has been diagnosed with bipolar disorder?”

Hypothesis Testing

What is a good hypothesis?
How to test a hypothesis?

Hypothesis Testing

- A good hypothesis...
  - ...is founded in a problem statement and supported by research.
  - ...is testable.
  - ...states an expected relationship between variables.
  - ...is stated as simply and concisely as possible.

- Hypothesis testing:
  - Step 1: What are your variables? (nominal, ordinal, scale, ratio)
  - Step 2: Measure the variables (Are aggregated measures enough?)
  - Step 3: Significance test (Null hypothesis? Which α level? Which significance test?) [lecture video 2024]

Hypothesis Testing

For the following hypothesis:

We hypothesize that the retrieval pipeline using query expansion combined with the MonoT5 reranker achieves a significantly different nDCG@10 compared to BM25 paired only with the MonoT5 reranker.

Hypothesis Testing

For the following hypothesis:

`We hypothesize that the retrieval pipeline using query expansion combined with the MonoT5 reranker achieves a significantly different nDCG@10 compared to BM25 paired only with the MonoT5 reranker.`

Person X has formulated the following null hypothesis H_0:

`The retrieval pipeline using query expansion with the MonoT5 reranker yields an nDCG@10 that is not significantly different from the nDCG@10 achieved by BM25 with the MonoT5 reranker.`

$\rightarrow$ Our goal is to falsify H_0.

Hypothesis Testing

For the following hypothesis:

```
We hypothesize that the retrieval pipeline using query
expansion combined with the MonoT5 reranker achieves a
significantly different nDCG@10 compared to BM25 paired only
with the MonoT5 reranker.
```

Person X has formulated the following null hypothesis H_0:

```
The retrieval pipeline using query expansion with the MonoT5
reranker yields an nDCG@10 that is not significantly
different from the nDCG@10 achieved by BM25 with the MonoT5
reranker.
```

Do you see any problems?

Hypothesis Testing

Person X has formulated the following null hypothesis H_0:

`The retrieval pipeline using query expansion with the MonoT5 reranker yields an nDCG@10 that is not significantly different from the nDCG@10 achieved by BM25 with the MonoT5 reranker.`

Do you see any problems?

- Which retrieval pipeline is being referred to?
- Which query expansion method is used?
- What re-ranking depth is applied?
- What significance level α is assumed?
- Which dataset is being evaluated?
- ...

The vaguer the hypothesis, the harder it is to reject H_0.

Hypothesis Testing

Person X has formulated the following null hypothesis H_0:

The retrieval pipeline using query expansion with the MonoT5 reranker yields an nDCG@10 that is not significantly different from the nDCG@10 achieved by BM25 with the MonoT5 reranker.

How could a better null hypothesis H_0 sound?

There is no statistically significant difference in nDCG@10 on **MS MARCO** between (i) the **BM25-based retrieval pipeline** described in Section X with **RM3 query expansion** and followed by **re-ranking the top-100** initial retrieval results with a monoT5 reranker[**footnote** with model], and (ii) the same pipeline without RM3 query expansion **(**$\alpha = 0.05$**).**

Multiple Hypotheses

Consider the research question: `Does the MonoT5 reranker improve rankings?`

You may define multiple hypotheses of different strength:

1. `There is a statistically significant difference ...`
2. `There is a statistically significant improvement ...`

You may define multiple null hypotheses of different strength:

- H_0^0: `There is no statistically significant difference in nDCG@10 on ...`
- H_0^1: `There is no statistically significant improvement in nDCG@10 on ...`

Each hypothesis requires its own significance test to attempt to falsify it.

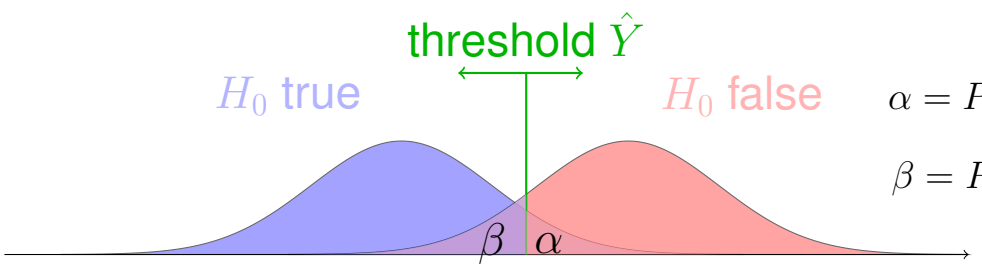
Using both a weaker and a stronger hypothesis is advantageous: even if you cannot show a performance improvement (i.e., cannot reject H_0^1), you may still be able to reject the weaker hypothesis (i.e., H_0^0), which still supports a meaningful conclusion.

Significance test

Decide whether the data provide sufficient evidence to reject a particular null hypothesis H_0.

- Probabilities α, p
- $\alpha = P(\text{reject } H_0 | H_0 \text{ true})$
- $p = P(X = x | H_0 \text{ true})$: Obtain a result at least as extreme, given H_0 is true
- Result is statistically significant, if $p \leq \alpha$

	H_0 is true	H_0 is false
$\neg$ Reject H_0	True Negative	Type II error
	probability $1 - \alpha$	probability β
Reject H_0	Type I error	True Positive
	probability α	probability $1 - \beta$

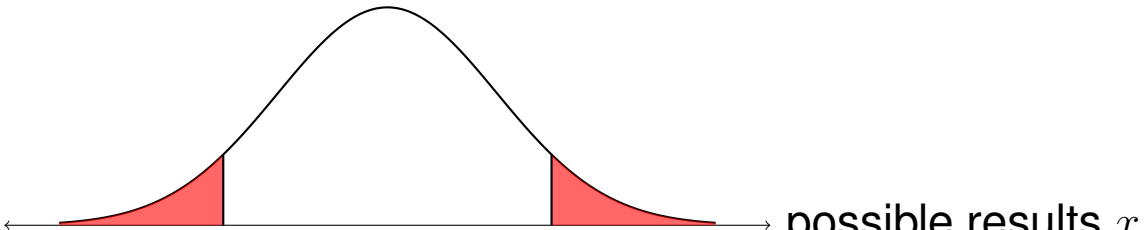


$$\alpha = P(\text{reject } H_0 | H_0 \text{ true}) = P(X > \hat{Y} | H_0 \text{ true})$$

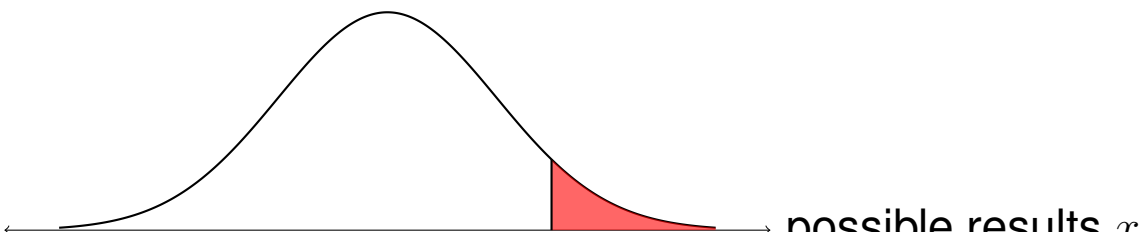
$$\beta = P(\neg \text{ reject } H_0 | H_0 \text{ false}) = P(X < \hat{Y} | H_0 \text{ false})$$

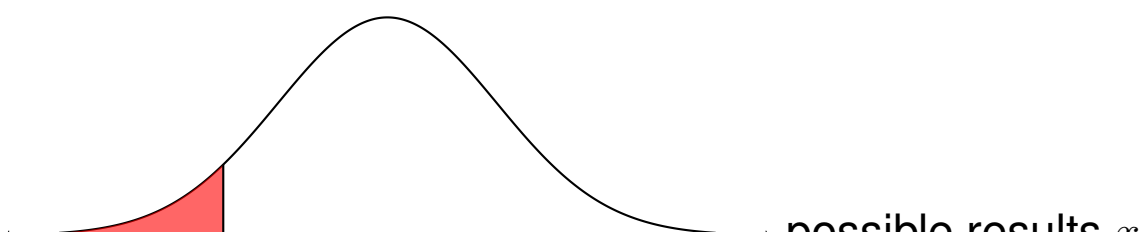
Significance test

Two-tailed test tests whether the observed statistic X is *either* unusually small or unusually large under H_0.



One-tailed test tests whether X is unusually large (right tail) or unusually small (left tail) under H_0.





Significance test

- The choice of statistical test depends on, among others, the scaling of the data [Wikipedia]
  - Interval data for which the sampling distribution of the test statistic is approximately normal . . .
    - . . . with *unknown* variance → Student's t-test
    - . . . with *known* population mean and variance → Z-test
  - Ordinal data → Sign test / . . .
  - Nominal data → McNemar's test / Chi-squared test / . . .
  - . . .

Example:

Suppose you have nDCG scores s_A, s_B for different topics t_A, t_B from two retrieval systems A and B. You might hypothesize that the mean nDCG score of system A is higher than that of system B, i.e., $\overline{s_A} \geq \overline{s_B}$.

To test this, you could define the null hypothesis as: $H_0 : \overline{s_A} \leq \overline{s_B}$ . . . and use a test statistic such as: $t = \frac{\overline{s_A} - \overline{s_B}}{\frac{\sigma_A}{\sqrt{|t_A|}}}$ (Student's t-test [Formula])

Assignment

1. Develop a research question
   - Grounded in exploratory analysis and literature review
   - Not too complex
   - Focus on effectiveness rather than efficiency
2. Formulate ≥ 1 hypotheses for your research question
3. Test your hypotheses
   - Apply appropriate statistical tests to evaluate and potentially reject each null hypothesis
   - Use the previously annotated topics for evaluation (via TIRA)
   - You may use the final effectiveness scores of the 10 baseline systems and all submitted approaches (i.e., the full leaderboard)
   - Test results will be provided only *after* you have formulated your hypotheses
4. Briefly interpret your results and answer your research question in a written report

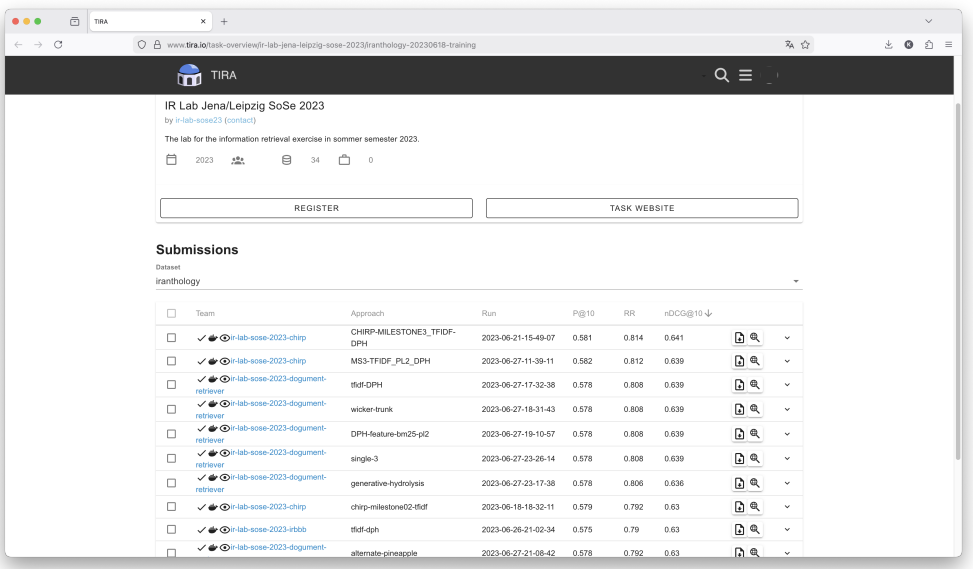
Next Steps

- Exercise sheet on temir.org

- Assignment
  - Due Date: Monday, 19.01.2025, 23:59
  - Deliverable:
    - TIRA submission(s)
    - Short report (1.5–2 pages) written in LaTeX
      - Include the sections: `Introduction`, `Related Work`, `Method`, `Results`, `Conclusion`, `References` (optionally `Appendix`)
      - For formatting examples (tables, figures, etc.), see past Webis publications [example]
      - Use the WOWS 2025 paper template [WOWS 2025]
      - Submit the report via email, CC your team member, and include both a PDF version and the LaTeX source files as a separate zipped archive

Inspiration

- System effectiveness from last semesters [summer ’23][winter ’23/’24][summer ’24]
  - Which systems performed well?
  - Which topics were difficult?
  - Where were “good” retrieval systems fooled?



Leaderboard of past approaches.

If a { . . . } button is present, the code for that approach is available.

- TIREx components overview [link]

Appendix: WOWS

- International Workshop on Open Web Search (WOWS)
- Held at ECIR 2026, 29.03-02.04.2026, Delft, Netherlands
- More information: [WOWS 2026 website]
- Optional participation: Submit your work (call for papers opens soon)

Appendix: Variables

Scale (Operation)	Categories (no order or direction)	Natural Order	Equal Intervals	True Zero	Example
Nominal (=)	👍	👎	👎	👎	Marital status, sex, gender
Ordinal (median)	👍	👍	👎	👎	Student grade
Interval ($a+b$,$a-b$, $\frac{a+b}{2}$)	👍	👍	👍	👎	Temperature in °C or °F, year
Ratio ($a \cdot b$, $\frac{a}{b}$, $\sqrt{a \cdot b}$)	👍	👍	👍	👍	Temperature in K, age, height, weight