Module: Machine Translation Quality Assessment
Lecturer: João Lucas Cavalheiro Camargo
Instructions - Lab 3 - Evaluation Design

By the end of this lab, you will be able to:

- Plan the evaluation of a machine translation system. You will adopt evaluation best practices, and describe the following elements: metrics, evaluators, guidelines, evaluation platforms, likert scales, test suites, and risk assessment measures.
- Reflect about the use of human evaluation metrics.

Tasks

Task 0 - What to do?

You will plan the evaluation of a machine translation (MT) system. Use this document alongside the following materials:

- MT Scenario Sheet

Task 1 - What is the content type? What is its perishability?

First, consider the content type. Content can fall under various domains, such as the legal domain for documentation, creative domains like marketing materials, or general domains if news articles are to be translated.

As such, consider the figure below:

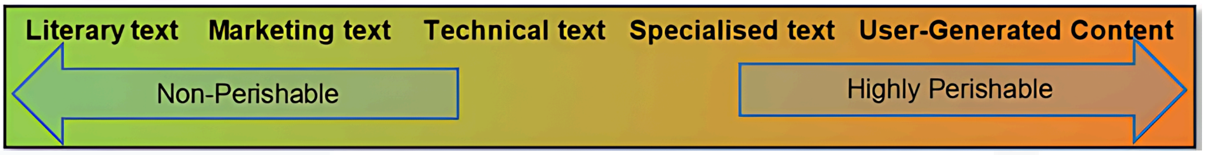


Figure 1. Perishability of content, Doherty et al. (2018)

Figure 1 illustrates a scale referring to the perishability of content. Some translated content is intended for long-term use, while other content may only be relevant for a short period, such as a week. What does this imply?

If content is to be used long-term, you will want the highest quality standards for several reasons:

1. **User Expectations**: Consider the users of this content. Will it be used to understand product or service reviews on a website? Is there a high expectation regarding the register, necessitating grammatical correctness? Or is the expectation of register lower, making grammatical correctness less critical as long as the content is sufficiently clear?
2. **Impact of Content**: Will this content be used for critical tasks, such as internal company procedures or medical information requiring utmost care? Or will it be used for routine daily tasks, such as summarising documents for internal communication?
3. **Public or Private Use**: Is the content intended for public or private consumption? Public content may have higher expectations, especially if accessible to anyone. For private content, expectations might vary depending on the seniority, role, and expertise of the system's users.

This is not a non-exhaustive list, as other contextual factors may impact your choice.

Task 1a - For this content type, what risk assessment measures will you use?

Consider the points above. How will you assess the risks associated with using your system?

- **Public End-User Risks**: If the system is used by end-users in a public setting, such as a website, does the content pose any risks?
- **Subject Matter Expert Considerations**: If used by a subject matter expert, are there instructions or warnings provided?
- **Content Modification**: Will this content be modified post-translation? If so, to what extent?

Based on your risk assessment, which human evaluation metrics will you use or recommend?

- **Expert Analysis**: Will you involve professional translators to analyse complex linguistic phenomena?
- **User Testing**: Will you test the system with users appropriate for its intended use, such as physicians using MT to read scientific articles?

- **Test Suites**: Will you design or utilise specific test suites to ensure your system can handle critical content issues? For example, if working with technical texts, you might focus on terminology.

Task 1b - Choose a test set

With the domain and content selected, you must now choose a test set.

Suppose you have **1 million aligned sentence pairs** in your chosen language pair for developing your MT system. You can set aside a portion of this corpus as the test set—for example, 10% (or 100,000 sentence pairs). Alternatively, you may opt to choose another corpus with similar content to verify your system's performance.

You want to compare your MT system's translations with the test set to evaluate performance on unseen data. As previously studied, this comparison will involve combining and correlating automatic evaluation metrics with human evaluation metrics.

Depending on the content type and after performing the risk assessment, you might also consider designing or using existing test suites.

Task 1c - Will you use a test suite?

In addition to using a test set, you may wish to verify specific linguistic phenomena. For example:

- **Medical Content**: If building a system for medical content, you might test for terminological accuracy.
- **Consistency Checks**: If developing a system for gist translations but concerned about consistency in tense and pronouns, you could test against crafted examples.

Based on the results from using a test suite, you might decide it's safer and more effective to fine-tune your model or train your system with more targeted data. This process helps identify weaknesses in your MT system.

Task 2 - Choose an MT system.

After defining the content type, risk assessment measures, test data type, and whether you are using a test suite, you should have a clear idea of the type of MT system you want to develop and evaluate.

It's generally advisable to select other MT systems capable of translating similar content types. Comparing your system against these can be useful in the evaluation, as you want to measure your system's performance relative to others in executing the task for which you were commissioned.

Task 3 - Choose the metrics

Selecting appropriate metrics for your quality evaluation is essential. Depending on the content type and MT system, you'll need to use a combination of different metrics.

- **Established MT Systems**: If working on a well-established MT system in financial records or corporate manuals, you might emphasise automatic metrics for routine checks when implementing in a similar area.
- **New Content Areas**: If implementing the system in a different area, you might place greater emphasis on human evaluation metrics combined with routine automatic metric checks.
- **Developing New MT Systems**: For new systems, it's crucial to use a comprehensive evaluation combining human and automatic evaluation metrics to ensure the system can handle the content type effectively.

Task 3a - Human Evaluation Metrics

Consider our previous lectures, and select one (or more) metrics from below, depending on your use case:

- Adequacy/Fluency
- Error Annotation
- Readability
- Comprehensibility
- Usability
- Acceptability
- Ranking

Additionally, consider how these metrics will be evaluated:

- **At Sentence Level**
- **With Specific Context Span** (e.g., preceding/succeeding sentences)
- **At Document Level**

You do not need to make the decision now, but start thinking about it, as it will be impacted by **Task 3d**.

Task 3b - Automatic Evaluation Metrics

Consider different types of automatic metrics:

Reference-based metrics

- BLEU
- chrF
- TER

Referenceless metrics

- BERTScore
- BLEURT
- COMET

Document-level approaches

- Concatenating sentences to form a document to evaluate with traditional metrics (Wong and Kit, 2012; Gong, Zhang and Zhou, 2015; Xiong et al. 2019, Liu et al., 2020, Saunders, Stahlberg and Bryne, 2020)
- BlonDe
- Pretrained metrics converted into document-level metrics, e.g. DOC-COMET (Vernikos et al. 2022)
- DiscoScore (Zhao et al., 2023)
- SLIDE (Raunkat et al., 2023)

Task 3c - Choose the evaluators

Considering the metrics, especially when using human evaluation, you need to decide on the type of evaluators involved.

It's generally recommended to employ **translators as evaluators**, particularly when:

- Developing a new MT system.
- The MT system is intended for translators as users.
- Performing risk assessments.

Moreover, consider aligning the evaluation metrics with your evaluators, which can be groups of one of the below (or more):

- **Expert Evaluators (Translators)**:
    - Skilled in examining complex linguistic phenomena.
    - Capable of identifying errors and their severity (important in risk assessments).
    - **Specialisation Matters**: Recruit translators with the appropriate expertise to match specific domains. For example, when working with medical content, employ translators specialising in medical translations.

- **End-Users**:
    - Could be experts or non-experts.
    - For experts (translators): Evaluate satisfaction or productivity with the MT system.
    - For subject matter experts: Assess the system according to their needs when subject matter expertise is part of the process.
    - For non-experts: Assess comprehensibility of translations in contexts like online product reviews.

- **Crowdsourcing**:
    - Useful when a large number of evaluators is needed (e.g., shared tasks).
    - Requires stricter quality control measures.

Be mindful that the type of human evaluation metric must match the type of evaluator with the purpose of the evaluation. It might not make sense to ask an end-user to perform a critical error annotation evaluation, you might want to check if product reviews or the instructions of a website are comprehensible when machine translated.

The choice of evaluator affects the claims you can make about your system. For instance, you cannot claim your system was checked by experts if only crowdsourced evaluations were conducted. This consideration will be discussed in more detail in the next lesson on ethics.

Task 3d - Choose an evaluation platform

The evaluation platform is crucial as it depends entirely on the evaluation metrics and evaluators you plan to use.

There are several evaluation platforms available, such as:

Human Evaluation

Cloud-based survey tools:

- Google Forms
- SurveyMonkey
- Qualtrics

Specialised Evaluation Platforms:

- Appraise
- KantanAI

Crowdsourcing Platforms:

- Amazon Mechanical Turk

Automatic Evaluation

Reference-based and referenceless metrics

- MATEO
- MutNMT

These are merely suggestions of some platforms for different purposes. Be mindful that the capabilities of each platform may limit or allow the evaluation to reach its full potential. Platforms and formats can be adapted, but be warned that they may affect the robustness of your evaluation.

Task 3e - Will you use Likert scales? If so, how?

When designing your evaluation, it's worth contemplating whether Likert scales are suitable for your needs, taking into account the evaluation platform you're using and the nature of your content.

In some cases, simple yes/no responses might be appropriate. For example, in a comprehensibility evaluation, you might ask: "Can you understand what this product review is describing?" with options like "Yes" or "No".

Or you might want to use Likert Scales. It is up to you to decide about the scale, particularly when it comes to adding a **middle point**. Including it allows evaluators to express neutrality, while excluding it encourages a more decisive response.

Or you may decide to use a Direct Assessment (DA) scale, widely used in evaluation campaigns. The DA scale typically ranges from 0 to 100, and you have several options:

- **Show the increments of the scale.**

  E.g. You may show 5 increments (0, 25, 50, 75, 100)
  You may show 10 increments (0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100).

- **Show labels.**

  E.g. You can show two labels for Adequacy, “None of it” on one end (matching the increment 0) and “All of it” (matching the increment 100).

  - You can show five labels for Fluency, “No Fluency” (matching the increment 0), “Little Fluency” (matching the increment 25), “Some Fluency” (matching the increment 50), “Near Native” (matching the increment 75) or “Native” (matching the increment 100).
  - You can show four labels for Adequacy, “None”, “Little of it”, “Most of it”, “All of it” and no increments.

- **Add or remove custom positions for the scale**

  E.g. The custom position of the scale can start at any increment (0, 25, 50, 75, 100).

- **Show value to evaluators**

  Your evaluator may see the value during their evaluation. For example, they can see if they chose the value 74.
  Or you might choose not to show to evaluators the value of their evaluation (the scale will not show any value, and it may show only increments).

- **Scale snapping**

  In the scenario you want evaluators to give values that only match the increments.
  For example, if the evaluator chooses a point in the interface that would correspond to 74, the scale snaps to the increment 75.
  Or if the evaluator chooses a point in the interface that would correspond to 21, the scale snaps to 25.

All these scale settings can impact your evaluation.

Task 3f - Write about the guidelines

Guidelines are an essential part of your evaluation, serving as a bridge between your objectives and the evaluators' understanding. They can help reduce subjectivity and promote consistency in how evaluations are conducted.

Before launching the full evaluation, conducting a pilot test can reveal any unclear areas in your guidelines or process. This allows you to make adjustments beforehand.

Be sure to strike a balance with your guidelines:

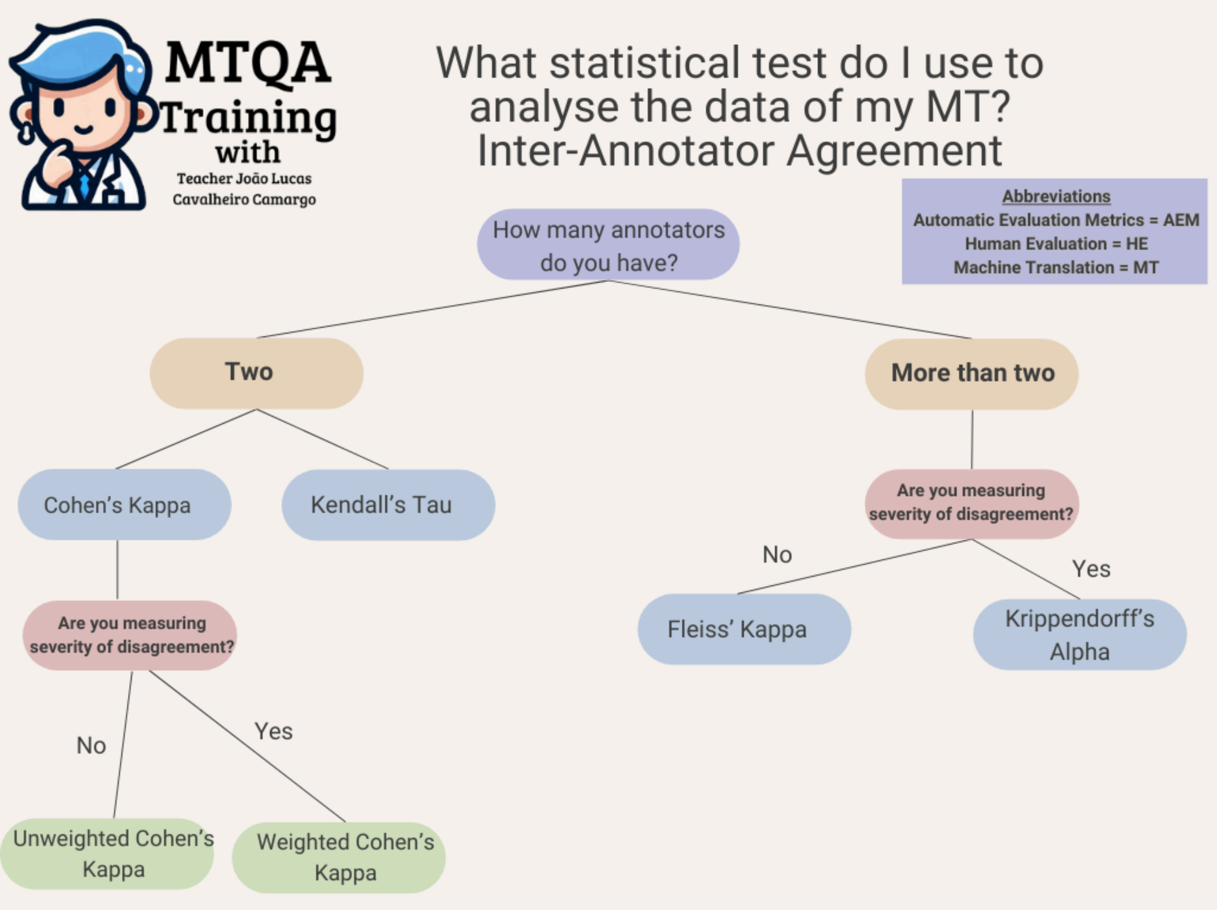
- Aim for language that is clear and straightforward. Consider the background and expertise of your evaluators to ensure the guidelines are accessible.
- Outline the evaluation process step by step, but avoid overwhelming details. Focus on what's essential for evaluators to know.
- Offer a brief tutorial on navigating the platform, including any specific functionalities they'll need. Visual aids like screenshots can be very helpful here.
- Try to foresee areas where evaluators might find ambiguous and address them proactively in the guidelines.
- When conducting pilots, invite evaluators to share feedback on the guidelines themselves. This can help you refine the instructions and improve the evaluation process.

Task 4 - How will the evaluation be analysed? What statistical tests will be employed?

After going through all these tasks, you need to systematically analyse your results. You will have to apply statistical methods, which can be separated into three types:

- **Different inter-annotator agreements measures**
  Unweighted Cohen's Kappa (κ), Weighted Cohen's Kappa, Fleiss`Kappa, Krippendorff's Alpha (a), using Kendall's Tau to check degree of agreement

When deciding what measures to refer to inter-annotator agreement, refer to Figure 2:



MTQA Training with Teacher João Lucas Cavalheiro Camargo

What statistical test do I use to analyse the data of my MT? Inter-Annotator Agreement

Abbreviations
Automatic Evaluation Metrics = AEM
Human Evaluation = HE
Machine Translation = MT

How many annotators do you have?

Two

Cohen's Kappa

Are you measuring severity of disagreement?

No

Unweighted Cohen's Kappa

Yes

Weighted Cohen's Kappa

Kendall's Tau

More than two

Are you measuring severity of disagreement?

No

Fleiss' Kappa

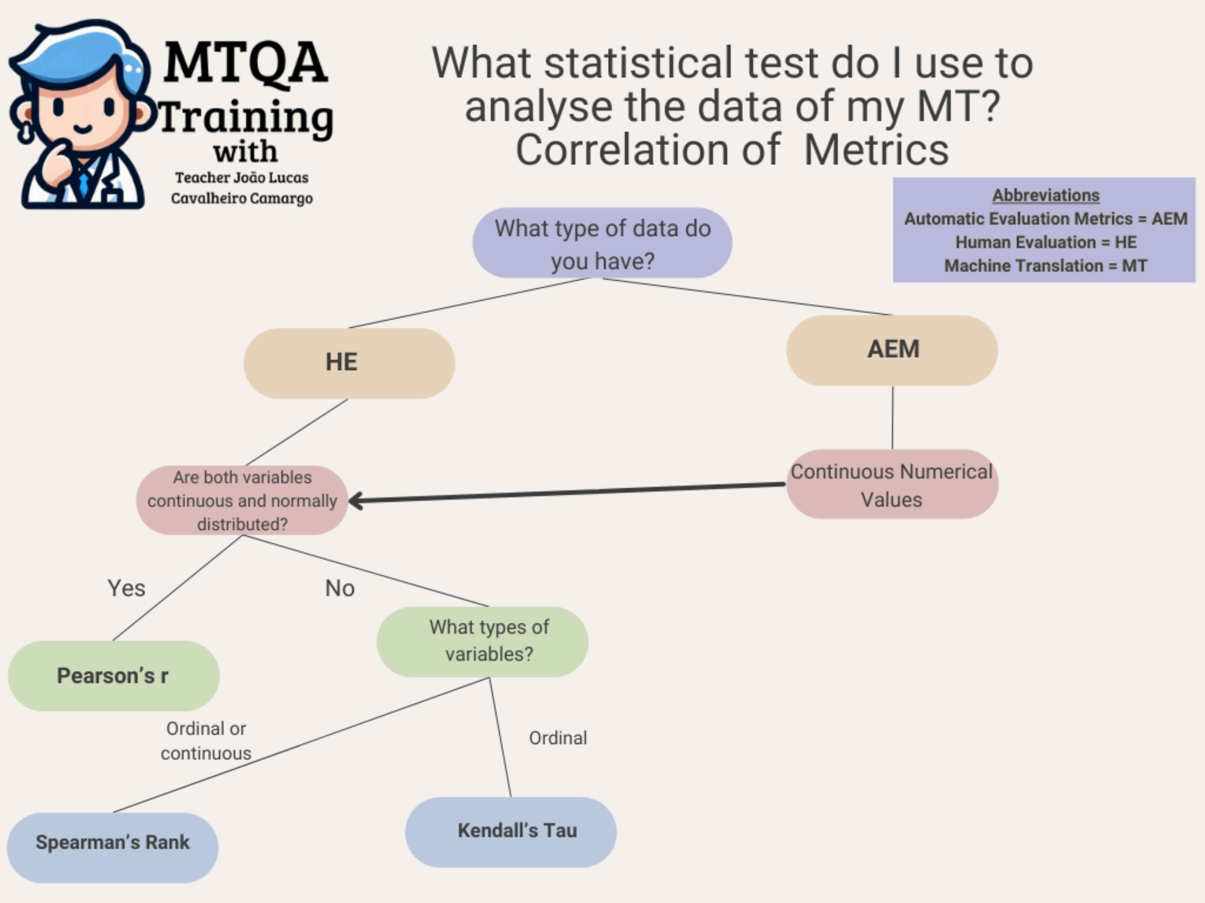
Yes

Krippendorff's Alpha

Figure 2. Statistical measures for Inter-Annotator Agreement

- **Statistical tests for Correlation**
  Pearson Correlation Coefficient (r), Spearman's Rank Correlation Coefficient (ρ)

When referring to statistical tests to correlate human evaluation with automatic evaluation metrics, consider Figure 3:



What statistical test do I use to analyse the data of my MT? Correlation of Metrics

Abbreviations
Automatic Evaluation Metrics = AEM
Human Evaluation = HE
Machine Translation = MT

MTQA Training with Teacher João Lucas Cavalheiro Camargo
What type of data do you have?
HE
AEM
Continuous Numerical Values
Are both variables continuous and normally distributed?
Yes
No
Pearson's r
What types of variables?
Ordinal or continuous
Ordinal
Spearman's Rank
Kendall's Tau

Figure 3. Statistical measures for the correlation of metrics

- **Analysing Statistical Significance**
  Checking p-values, comparing means between groups (e.g., different MT systems), Paired Samples t-tests, One-way ANOVA, Mann-Whitney U Test, Wilcoxon Signed-Rank Test, Kruskal-Wallis H Test

For analysing the statistical significance of your results, consider Figure 4:

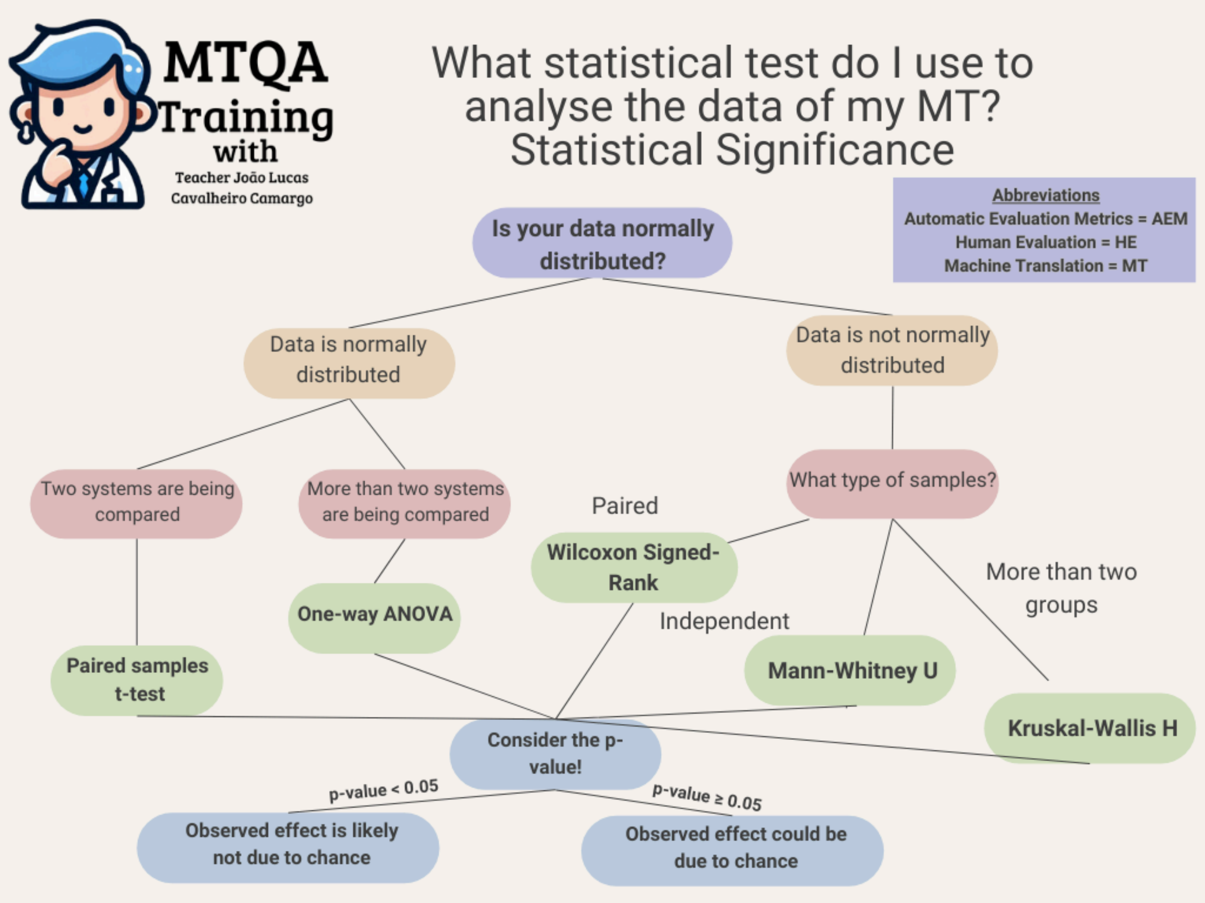

MTQA Training with Teacher João Lucas Cavalheiro Camargo
What statistical test do I use to analyse the data of my MT? Statistical Significance
Abbreviations
Automatic Evaluation Metrics = AEM
Human Evaluation = HE
Machine Translation = MT
Is your data normally distributed?
Data is normally distributed
Data is not normally distributed
Two systems are being compared
More than two systems are being compared
What type of samples?
Paired
Wilcoxon Signed-Rank
One-way ANOVA
Independent
More than two groups
Paired samples t-test
Mann-Whitney U
Kruskal-Wallis H
Consider the p-value!
p-value < 0.05
p-value ≥ 0.05
Observed effect is likely not due to chance
Observed effect could be due to chance

Figure 4. Statistical measures for the analysis of statistical significance of results